aws

開発者ガイド

AWS Data Pipeline

API バージョン 2012-10-29

AWS Data Pipeline: 開発者ガイド

Table of Contents

..... x
とは AWS Data Pipeline 1
からのワークロードの移行 AWS Data Pipeline 2
ワークロードの への移行 AWS Glue 3
AWS Step Functions へのワークロードの移行 4
Amazon MWAAへのワークロードの移行 5
概念のマッピング 6
サンプル 7
関連サービス 8
アクセス AWS Data Pipeline 9
料金 9
パイプラインの作業アクティビティ用にサポートされているインスタンスタイプ 10
AWSリージョン別のデフォルトのAmazon EC2インスタンス 10
その他のサポートされるAmazon EC2インスタンス 12
Amazon EMR クラスターでサポートされるAmazon EC2インスタンス 13
AWS Data Pipeline 概念 14
パイプライン定義 14
パイプラインのコンポーネント、インスタンス、試行 15
Task Runner 17
データノード 18
データベース 19
アクティビティ 19
前提条件 20
システム管理型の前提条件 21
ユーザー管理型の前提条件 21
リソース 21
リソースの制限 22
サポートされているプラットフォーム 22
Amazon EMR クラスターと AWS Data Pipelineを使用した Amazon EC2 スポットインスタンス 23
アクション 24
パイプラインの事前モニタリング 25
セットアップ 26
にサインアップする AWS 26

にサインアップする AWS アカウント 26
管理アクセスを持つユーザーを作成する 27
AWS Data Pipeline およびパイプラインリソースの IAM ロールを作成する 28
必要なアクションを実行するための IAM プリンシパル (ユーザーとグループ) の許可 28
プログラムによるアクセス権を付与する 30
AWS Data Pipeline の使用の開始 32
パイプラインの作成 33
実行中のパイプラインのモニタリング 34
出力の表示 35
パイプラインの削除 35
パイプラインの使用 36
パイプラインの作成 36
CLI を使用して Data Pipeline テンプレートからパイプラインを作成する 37
パイプラインの表示 56
パイプラインのステータスコードの解釈 56
パイプラインとコンポーネントのヘルス状態の解釈 58
パイプライン定義の表示 60
パイプラインインスタンスの詳細の表示 60
パイプラインのログの表示 61
パイプラインの編集 63
制限事項 63
AWS CLI を使用したパイプラインの編集 64
パイプラインのクローン作成 64
パイプラインのタグ付け 65
パイプラインの非アクティブ化 66
AWS CLI を使用したパイプラインの非アクティブ化 67
パイプラインの削除 67
アクティビティによるデータとテーブルのステージング 68
ShellCommandActivity によるデータのステージング 69
Hive およびステージングをサポートするデータノードによるテーブルのステージング 71
Hive およびステージングをサポートしないデータノードによるテーブルのステージング 72
複数リージョンにあるリソースの使用 73
カスケードの失敗と再実行 76
アクティビティ 76
データノードと前提条件 77
リソース 77

カスケードが失敗したオブジェクトの再実行 77
カスケードの失敗とバックフィル 77
パイプライン定義ファイルの構文 78
ファイル構造 78
パイプラインのフィールド 79
ユーザー定義フィールド 80
API の使用 81
AWS SDK をインストールする 81
AWS Data Pipeline に対する HTTP リクエストの実行 82
セキュリティ 87
データ保護 88
Identity and Access Management 89
AWS Data Pipeline の IAM ポリシー 90
AWS Data Pipeline のポリシー例 95
IAM ロール 98
ログ記録とモニタリング 106
AWS Data PipelineCloudTrailでの情報 106
AWS Data Pipeline ログファイルエントリの概要 107
インシデントへの対応 108
コンプライアンス検証 108
耐障害性 109
インフラストラクチャセキュリティ 109
AWS Data Pipeline での設定と脆弱性の分析 109
チュートリアル 110
Hadoop Streaming EMRで Amazon を使用してデータを処理 110
開始する前に 111
CLI の使用 111
Amazon S3 から Amazon S3 への CSV データのコピー 115
開始する前に 117
CLI の使用 117
Amazon S3 への MySQL データのエクスポート 124
開始する前に 125
CLI の使用 126
Amazon Redshift へのデータのコピー 135
開始する前に: COPY オプションの設定 136
開始する前に: パイプライン、セキュリティ、およびクラスターを設定する 137

CLI の使用 139
パイプラインの式と関数 149
単純データ型 149
DateTime 149
数値 149
オブジェクト参照 149
[Period] (期間) 150
文字列 150
表現 150
フィールドとオブジェクトの参照 151
入れ子式 152
リスト 152
ノード式 153
式の評価 154
数学関数 154
文字列関数 155
日付および時刻関数 156
特殊文字 164
パイプラインオブジェクトリファレンス 165
データノード 166
D ynamoDBDataノード 167
MySqlDataNode 174
RedshiftDataNode 182
S3DataNode 190
SqlDataNode 198
アクティビティ 205
CopyActivity 206
EmrActivity 214
HadoopActivity 224
HiveActivity 235
HiveCopyActivity 245
PigActivity 255
RedshiftCopyActivity 270
ShellCommandActivity 285
SqlActivity 295
リソース 304

Ec2Resource 304
EmrCluster 316
HttpProxy 348
前提条件 351
D ynamoDBDataが存在する 352
D ynamoDBTableが存在する 356
存在する 360
S3KeyExists 364
S3PrefixNotEmpty 369
ShellCommandPrecondition 373
データベース 378
JdbcDatabase 379
RdsDatabase 381
RedshiftDatabase 383
データ形式 386
CSV データ形式 386
Custom データ形式 388
D ynamoDBData形式 389
DynamoDBExportDataFormat 392
RegEx データ形式 395
TSV データ形式 397
アクション 399
SnsAlarm 399
終了 401
スケジュール 403
例 403
構文 408
ユーティリティ 410
ShellScriptConfig 410
EmrConfiguration 412
プロパティ 417
Task Runnerの操作 420
AWS Data Pipelineマネージドリソースの Task Runner 420
Task Runnerを使用した既存のリソースでの作業の実行 422
Task Runnerのインストール 423
(オプション) Task Runner に Amazon へのアクセス権を付与する RDS 424

Task Runnerの起動 426
Task Runnerのログの確認 427
Task Runnerのスレッドと前提条件 427
Task Runnerの設定オプション 427
プロキシ経由によるTask Runnerの使用 430
Task Runner とカスタム AMIs 430
トラブルシューティング 432
パイプラインのエラーを見つける 432
パイプラインを処理する Amazon EMR クラスターの特定 433
パイプラインのステータスの詳細を解釈する 434
エラーログを見つける 435
パイプラインのログ 436
Hadoop ジョブと Amazon EMR ステップログ 436
一般的な問題を解決する 437
ステータスが保留中のままパイプラインが先に進まない 437
Runner のステータスを待機してパイプラインコンポーネントが先に進まない 438
WAITING_ON_DEPENDENCIES ステータスでパイプラインコンポーネントが先に進まない 438
予定した時期に実行が開始されない 439
パイプラインコンポーネントが間違った順番で実行される 440
EMR クラスターが「リクエストに含まれているセキュリティトークンが無効です」というエラーで失敗する 440
リソースにアクセスするアクセス許可が不十分である 440
ステータスコード: 400 エラーコード: PipelineNotFoundException 440
パイプラインを作成するとセキュリティトークンエラーが発生する 441
コンソールでパイプラインの詳細を表示できない 441
リモートランナーのエラー - ステータスコード: 404, AWS サービス: Amazon S3 441
アクセスが拒否される - datapipeline 関数を実行する権限がない 441
古い Amazon EMR AMI では、大きい CSV ファイルに対して間違ったデータが作成される可能性がある 442
AWS Data Pipeline の上限を引き上げる 442
Limits 443
アカウントの制限 443
ウェブサービス呼び出しの制限 444
スケーリングに関する考慮事項 446
AWS Data Pipeline リソース 447

ドキュメント履歴 449

AWS Data Pipeline は、新規顧客には利用できなくなりました。の既存のお客様は、通常どおりサービスを AWS Data Pipeline 引き続き使用できます。詳細はこちら

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。

とは AWS Data Pipeline

> **Note**
>
> AWS Data Pipeline サービスはメンテナンスモードであり、新機能やリージョンの拡張は予定されていません。詳細および既存のワークロードの移行方法については、「からのワークロードの移行 AWS Data Pipeline」を参照してください。

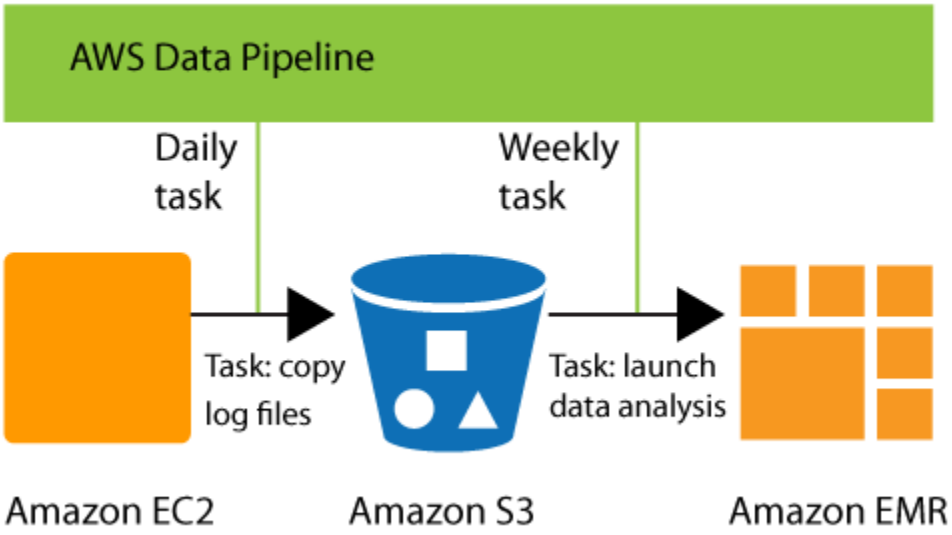
AWS Data Pipeline は、データの移動と変換を自動化するために使用できるウェブサービスです。を使用すると AWS Data Pipeline、データ駆動型のワークフローを定義して、以前のタスクが正常に完了したかどうかにタスクを依存させることができます。データ変換のパラメータを定義し、設定したロジック AWS Data Pipeline を適用します。

の以下のコンポーネント AWS Data Pipeline が連携してデータを管理します。

- パイプライン定義とは、データ管理のビジネスロジックを指定したものです。詳細については、「パイプライン定義ファイルの構文」を参照してください。
- パイプラインは、タスクをスケジュールして実行するために、Amazon EC2インスタンスを作成して定義済みの作業アクティビティを実行します。パイプラインにパイプライン定義をアップロードし、パイプラインをアクティブ化します。実行中のパイプラインのパイプライン定義を編集し、有効にするパイプラインを再度アクティブ化することができます。パイプラインを非アクティブ化し、データソースを修正して、パイプラインを再度アクティブ化することができます。パイプラインの処理が終了したら、パイプラインを削除できます。
- Task Runnerは、タスクをポーリングし、それらのタスクを実行します。例えば、Task RunnerはAmazon S3にログファイルをコピーしてAmazon EMR クラスターを起動できます。Task Runnerは、パイプラインの定義によって作成されたリソースに自動的にインストールおよび実行されます。カスタムタスクランナーアプリケーションを記述することも、 が提供する Task Runner アプリケーションを使用することもできます AWS Data Pipeline。詳細については、「Task Runner」を参照してください。

例えば、 AWS Data Pipeline を使用してウェブサーバーのログを毎日 Amazon Simple Storage Service (Amazon S3) にアーカイブし、それらのログに対して毎週 Amazon EMR (Amazon EMR) クラスターを実行してトラフィックレポートを生成できます。 は、データをコピーする日次タスクと Amazon EMR クラスターを起動する週次タスクを AWS Data Pipeline スケジュールします。 AWS

Data Pipeline また、ログのアップロードに予期しない遅延がある場合でも、Amazon EMR は分析を開始する前に最終日のデータが Amazon S3 にアップロードされるのを待機します。



内容

- からのワークロードの移行 AWS Data Pipeline
- 関連サービス
- アクセス AWS Data Pipeline
- 料金
- パイプラインの作業アクティビティ用にサポートされているインスタンスタイプ

からのワークロードの移行 AWS Data Pipeline

AWS は 2012 年に AWS Data Pipeline サービスを開始しました。当時は、さまざまなコンピューティングオプションを使用して、異なるデータソース間でデータを確実に移動できるサービスが求められていました。現在は、お客様により優れたエクスペリエンスを提供するサービスが他にもあります。例えば、 AWS Glue を使用して Apache Spark アプリケーションを実行およびオーケストレーションしたり、 AWS Step Functions を使用して AWS サービスコンポーネントをオーケストレーションしたり、Amazon Managed Workflows for Apache Airflow (Amazon MWAA) を使用して Apache Airflow のワークフローオーケストレーションを管理したりできます。

このトピックでは、 から AWS Data Pipeline 代替オプションに移行する方法について説明します。選択するオプションは、 AWS Data Pipelineの現在のワークロードによって異なります。の一般的な

ユースケース AWS Data Pipeline を AWS Glue、、 AWS Step Functions、または Amazon MWAA に移行できます。

ワークロードの への移行 AWS Glue

AWS Glueは、分析を行うユーザーが複数のソースからのデータを簡単に検出、準備、移動、統合できるようにするサーバーレスのデータ統合サービスです。これには、ジョブの作成、実行、ワークフローのオーケストレーションのためのツールが含まれています。を使用すると AWS Glue、70 を超える多様なデータソースを検出して接続し、一元化されたデータカタログでデータを管理できます。抽出、変換、ロード (ETL) パイプラインを視覚的に作成、実行、モニタリングして、データをデータレイクにロードできます。また、Amazon Athena、Amazon EMR、Amazon Redshift Spectrumを使用して、カタログ化されたデータをすぐに検索し、クエリできます。

次の AWS Glue 場合は、 AWS Data Pipeline ワークロードを に移行することをお勧めします。

- さまざまなデータソース、ビジュアルエディタやノートブックなどのオーサリングインターフェイス、データ品質や機密データ検出などの高度なデータ管理機能をサポートするサーバーレスデータ統合サービスを探している。
- ワークロードは、 AWS Glue ワークフロー、ジョブ (Python または Apache Spark 内) 、クローラ (既存のパイプラインが Apache Spark 上に構築されているなど) に移行できます。
- 取り込み、処理、転送、整合性テスト、品質チェックなど、データパイプラインのあらゆる側面を処理できる単一のプラットフォームを必要としている。
- 既存のパイプラインは、DynamoDB テーブルを Amazon S3 にエクスポートするなど、 AWS Data Pipeline コンソールの定義済みテンプレートから作成されており、同じ目的テンプレートを探しています。
- ワークロードがApache Hiveのような特定のHadoopエコシステムアプリケーションには依存しない。
- ワークロードにオンプレミスサーバーのオーケストレーションが必要ない。

AWS は、クローラ (データの検出) および ETL ジョブ (データの処理とロード) に対して、秒単位で課金される時間単位の料金を請求します。 AWS Glue Studio は AWS Glue リソース用の組み込みオーケストレーションエンジンであり、追加料金なしで提供されます。料金の詳細については、「AWS Glue の料金」を参照してください。

AWS Step Functions へのワークロードの移行

AWS Step Functions は、ビジネスクリティカルなアプリケーションのワークフローを構築できるサーバーレスオーケストレーションサービスです。Step Functions では、ビジュアルエディタを使用してワークフローを構築し、AWS Lambda 、Amazon EMR、DynamoDB など、250 を超える AWS サービスの 11,000 を超えるアクションと直接統合できます。Step Functions を使用して、データ処理パイプラインのオーケストレーション、エラーの処理、基盤となる AWS サービスのスロットリング制限の操作を行うことができます。機械学習モデルを処理して公開するワークフローを作成し、マイクロサービスをオーケストレーションし、 AWS などの サービスを制御して AWS Glue、抽出、変換、ロード (ETL) ワークフローを作成できます。また、手動による介入が必要なアプリケーション用に実行時間が長い自動化されたワークフローを作成することもできます。

と同様に AWS Data Pipeline、 AWS Step Functions は が提供するフルマネージドサービスです AWS。インフラストラクチャの管理、ワーカーのパッチ適用、OSバージョン更新の管理などを行う必要はありません。

次の場合は、 AWS Data Pipeline ワークロードを AWS Step Functions に移行することをお勧めします。

- サーバーレスで可用性の高いワークフローオーケストレーションサービスを探している。
- 1つのタスク実行の粒度で課金される、費用対効果の高いソリューションを探している。
- ワークロードは、Amazon EMR、Lambda、DynamoDB など AWS Glue、他の複数の AWS サービスのタスクをオーケストレーションしています。
- ワークフロー作成用の drag-and-drop ビジュアルデザイナーが付属するローコードソリューションを探しており、新しいプログラミング概念を学ぶ必要はありません。
- 11,000 を超えるアクションをカバーする 250 を超える他の AWS サービスとの統合を提供し out-of-the-box、カスタムの非AWS サービスおよびアクティビティとの統合を可能にするサービスを探しています。

AWS Data Pipeline と Step Functions はどちらも JSON 形式を使用してワークフローを定義します。これにより、ワークフローをソース管理に保存し、バージョンを管理し、アクセスを制御し、CI/CDで自動化することができます。Step Functionsは、完全にJSONに基づいたAmazon State Language と呼ばれる構文を使用しており、ワークフローのテキスト表現と視覚表現をシームレスに切り替えることができます。

Step Functionsでは、現在 AWS Data Pipelineで使用しているのと同じバージョンのAmazon EMRを選択できます。

AWS Data Pipeline マネージドリソースでのアクティビティを移行するには、Step Functions での AWS SDK サービス統合を使用して、リソースのプロビジョニングとクリーンアップを自動化できます。

オンプレミスサーバー、ユーザーマネージドEC2インスタンス、またはユーザーマネージドEMRクラスター上のアクティビティを移行する場合、SSMエージェントをインスタンスにインストールできます。コマンドは、Step FunctionsからAWS Systems Managerの実行コマンドを使用して開始できます。Amazon EventBridge で定義されたスケジュールからステートマシンを開始することもできます。

AWS Step Functions には、標準ワークフローと Express ワークフローの 2 種類のワークフローがあります。標準ワークフローでは、アプリケーションの実行に必要な状態遷移の回数に基づいて課金されます。Expressワークフローでは、ワークフローのリクエスト数とその期間に基づいて課金されます。料金の詳細については、「AWS Step Functionsの料金」を参照してください。

Amazon MWAAへのワークロードの移行

Amazon MWAA (Managed Workflows for Apache Airflow) は、Apache Airflow 用のマネージドオーケストレーションサービスで、クラウドでの end-to-end 大規模なデータパイプラインのセットアップと運用を容易にします。Apache Airflowは、「ワークフロー」と呼ばれる一連のプロセスとタスクをプログラムで作成、スケジュール、監視するためのオープンソースのツールです。Amazon MWAA では、AirflowとPythonプログラミング言語を使用して、スケーラビリティ、可用性、セキュリティの基盤となるインフラストラクチャを管理せずにワークフローを作成できます。Amazon MWAA は、ワークフロー実行容量をニーズに合わせて自動的にスケーリングし、 AWS セキュリティサービスと統合して、データへの高速かつ安全なアクセスを提供します。

と同様に AWS Data Pipeline、Amazon MWAA は が提供するフルマネージドサービスです AWS。これらのサービスに固有の新しい概念をいくつか学ぶ必要がありますが、インフラストラクチャの管理、ワーカーのパッチ適用、OSバージョン更新の管理などを行う必要はありません。

次の場合は、 AWS Data Pipeline ワークロードを Amazon MWAA に移行することをお勧めします。

- Pythonで記述されたワークフローをオーケストレーションするための、マネージド型の可用性の高いサービスを探している。
- 移植性を最大限に高めるに、フルマネージドで広く採用されているオープンソーステクノロジであるApache Airflowに移行したいと考えている。
- 取り込み、処理、転送、整合性テスト、品質チェックなど、データパイプラインのあらゆる側面を処理できる単一のプラットフォームを必要としている。

- オブザーバビリティのための優れたUI、失敗したワークフローの再起動、バックフィル、タスクの再試行などの機能を備えた、データパイプラインオーケストレーション向けに設計されたサービスを探している。
- 800 を超える事前構築済みのオペレータとセンサーを備えたサービスを探しており、 だけでなく AWS 以外のサービス AWS もカバーしています。

Amazon MWAAワークフローはPythonを使用するDirected Acyclic Graphs (DAG) として定義されるため、ソースコードとして扱うこともできます。Airflowの拡張可能なPythonフレームワークにより、事実上あらゆるテクノロジーと接続するワークフローを構築できます。ワークフローを表示および監視するための優れたユーザーインターフェイスが付属しており、バージョン管理システムと簡単に統合してCI/CDプロセスを自動化できます。

Amazon MWAAでは、現在 AWS Data Pipelineで使用しているのと同じバージョンのAmazon EMRを選択できます。

AWS は、Airflow 環境の実行時間に加えて、ワーカーまたはウェブサーバーの容量を増やすための追加の自動スケーリングに対して課金します。料金の詳細については、「Amazon Managed Workflows for Apache Airflowの料金」を参照してください。

概念のマッピング

次の表には、サービスで使用される主要な概念のマッピングが示されています。Data Pipelineに精通している人がStep FunctionsとMWAAの用語を理解するのに役立ちます。

Data Pipeline	接着語	Step Functions	Amazon MWAA
パイプライン	ワークフロー	ワークフロー	Direct acylic graphs
パイプライン定義JSON	ワークフロー定義またはPythonベースのブループリント	Amazon State Language JSON	Pythonベース
アクティビティ	ジョブ	ステートとタスク	タスク（オペレータとセンサー）
インスタンス	ジョブ実行	実行	DAG実行
Attempts	再試行回数	Catcherとretrier	再試行

Data Pipeline	接着語	Step Functions	Amazon MWAA
パイプラインスケジュール	スケジュールトリガー	EventBridge スケジューラタスク	Cron、タイムテーブル、データ対応
パイプラインの式と関数	ブループリントライブラリ	Step Functions組み込み関数とAWS Lambda	拡張可能なPythonフレームワーク

サンプル

以下のセクションでは、 から個々のサービスへの移行を参照できる公開例 AWS Data Pipeline を示します。それらを例として参照し、ユースケースに基づいて更新してテストすることで、個々のサービスで独自のパイプラインを構築できます。

AWS Glue サンプル

次のリストには、 の最も一般的な AWS Data Pipeline ユースケースのサンプル実装が含まれています AWS Glue。

- Sparkジョブの実行
- JDBCからAmazon S3へのデータのコピー（Amazon Redshiftを含む）
- Amazon S3からJDBCへのデータのコピー（Amazon Redshiftを含む）
- Amazon S3からDynamoDBへのデータのコピー
- Amazon Redshiftとの間でのデータの移動
- DynamoDBテーブルへのクロスアカウントおよびクロスリージョンでのアクセス

AWS Step Functions のサンプル

次のリストには、 AWS Step Functions AWS Data Pipeline の最も一般的なユースケースのサンプル実装が含まれています。

- Amazon EMRジョブの管理
- Amazon EMR Serverlessでのデータ処理ジョブの実行
- Hive/Pig/Hadoopジョブの実行
- 大規模なデータセットのクエリ (Amazon AthenaAmazon S3、 AWS Glue)

- Amazon Redshiftを使用したETLワークフローの実行
- AWS Glue クローラーのオーケストレーション

AWS Step Functions を使用するための追加のチュートリアルとサンプルプロジェクトを参照してください。

Amazon MWAAのサンプル

次のリストには、Amazon MWAA AWS Data Pipeline の最も一般的なユースケースのサンプル実装が含まれています。

- Amazon EMRジョブの実行
- Apache HiveとHadoop用のカスタムプラグインの作成
- Amazon S3からRedshiftへのデータのコピー
- リモートEC2インスタンスでのシェルスクリプトの実行
- ハイブリッド (オンプレミス) ワークフローのオーケストレーション

Amazon MWAAの使用については、追加のチュートリアルとサンプルプロジェクトを参照してください。

関連サービス

AWS Data Pipeline は、以下の サービスと連携してデータを保存します。

- Amazon DynamoDB–完全マネージドNoSQLデータベースを高パフォーマンス、低コストで提供します。詳細については、Amazon DynamoDB開発者ガイドを参照してください。
- Amazon RDS–大規模なデータセットにスケールする完全マネージド型のリレーショナルデータベースを提供します。詳細については、Amazon Relational Database Serviceデベロッパーガイドを参照してください。
- Amazon Redshift–高速、完全マネージド型、ペタバイト規模のデータウェアハウスサービスを提供し、大量のデータの分析を容易で費用対効果の高いものにします。詳細については、Amazon Redshiftデータベース開発者ガイドを参照してください。
- Amazon S3–セキュリティ、耐久性、およびスケーラビリティに優れたオブジェクトストレージを提供します。詳細については、Amazon Simple Storage Serviceユーザーガイドを参照してください。

AWS Data Pipeline は、次のコンピューティングサービスと連携してデータを変換します。

- Amazon EC2–ユーザーがソフトウェアシステムを構築しホストするための、自在に拡張および縮小できるコンピューティング能力（実際にはAmazonのデータセンター内のサーバー）です。詳細については、Amazon EC2 ユーザーガイド」を参照してください。
- Amazon EMR–Apache HadoopやApache Sparkなどのフレームワークを使用して、Amazon EC2サーバー間で大量のデータを簡単、迅速、費用効率よく配布および処理できます。詳細については、Amazon EMR開発者ガイドを参照してください。

アクセス AWS Data Pipeline

次のインターフェイスのいずれかを使用して、パイプラインの作成、アクセス、管理を行うことができます。

- AWS Management Console–へ AWS Data Pipelineのアクセスに使用できるウェブインターフェイスを提供します。
- AWS Command Line Interface（AWS CLI）— を含む幅広い AWS サービスのコマンドを提供し AWS Data Pipeline、Windows、macOS、Linux でサポートされています。のインストールの詳細については、 AWS CLI「」を参照してくださいAWS Command Line Interface。のコマンドのリストについては AWS Data Pipeline、「datapipeline」を参照してください。
- AWS SDK—言語固有のAPIを提供し、署名の計算、リクエストの再試行処理、エラー処理など、接続のさまざまな詳細を処理します。詳細については、AWS SDKを参照してください。
- クエリ API–HTTPSリクエストを使用して呼び出す低レベル APIを提供します。クエリ APIの使用は、 AWS Data Pipelineの最も直接的なアクセス方法ですが、リクエストに署名するハッシュの生成やエラー処理など、低レベルの詳細な作業をアプリケーションで処理する必要があります。詳細については、「AWS Data Pipeline APIリファレンス」を参照してください。

料金

Amazon Web Servicesと併せて、使用した分に応じてお支払いください。の場合 AWS Data Pipeline、アクティビティと前提条件の実行がスケジュールされている頻度と実行場所に基づいて、パイプラインの料金を支払います。詳細については、「AWS Data Pipeline の料金」を参照してください。

AWSアカウントを作成してから12か月未満の場合、無料利用枠を使用できます。無料利用枠には、1か月あたり3つの低頻度の前提条件と5つの低頻度のアクティビティが無料で含まれています。詳細については、「AWS無料利用枠」を参照してください。

パイプラインの作業アクティビティ用にサポートされているインスタンスタイプ

がパイプライン AWS Data Pipeline を実行すると、パイプラインコンポーネントをコンパイルして、実用的な Amazon EC2 インスタンスのセットを作成します。各インスタンスには、特定のタスクを実行するためのすべての情報が含まれています。完全なインスタンスのセットは、パイプラインのTo-Doリストです。 AWS Data Pipeline は、Task Runnerに処理するインスタンスを渡します。

EC2インスタンスの構成はさまざまであり、インスタンスタイプと呼ばれています。各インスタンスタイプには、異なるCPU、入力/出力、およびストレージ容量があります。アクティビティ用のインスタンスタイプを指定するほかに、複数の異なる購入オプションを選択できます。すべてのインスタンスタイプがすべてのAWSリージョンで使用できるわけではありません。インスタンスタイプが使用できない場合、パイプラインでプロビジョニングが失敗または停止することがあります。インスタンスが使用できるかどうかについては、Amazon EC2の料金表のページを参照してください。インスタンスの購入オプションのリンクを開き、リージョンでフィルタして、インスタンスタイプが当該リージョンで使用できるかどうかを確認してください。これらのインスタンスタイプ、ファミリー、仮想化タイプの詳細については、Amazon EC2インスタンスおよびAmazon Linux AMIインスタンスタイプマトリックスを参照してください。

次の表は、 が AWS Data Pipeline サポートするインスタンスタイプを示しています。を使用して AWS Data Pipeline 、 がサポートされていないリージョンを含め、任意のリージョンで Amazon EC2 AWS Data Pipeline インスタンスを起動できます。 AWS Data Pipeline がサポートされているリージョンの詳細については、「AWS リージョンとエンドポイント」を参照してください。

内容

- AWSリージョン別のデフォルトのAmazon EC2インスタンス
- その他のサポートされるAmazon EC2インスタンス
- Amazon EMR クラスターでサポートされるAmazon EC2インスタンス

AWSリージョン別のデフォルトのAmazon EC2インスタンス

パイプライン定義にインスタンスタイプを指定しなかった場合、 AWS Data Pipeline はデフォルトでインスタンスを起動します。

次の表は、 がサポートされているリージョンで がデフォルトで AWS Data Pipeline 使用する Amazon EC2 インスタンスの一覧 AWS Data Pipeline です。

リージョン名	リージョン	インスタンスタイプ
米国東部（バージニア北部）	us-east-1	m1.small
米国西部（オレゴン）	us-west-2	m1.small
アジアパシフィック（シドニー）	ap-southeast-2	m1.small
アジアパシフィック（東京）	ap-northeast-1	m1.small
欧州（アイルランド）	eu-west-1	m1.small

次の表は、 がサポートされていないリージョンでデフォルトで AWS Data Pipeline 起動される Amazon EC2 インスタンスの一覧 AWS Data Pipeline です。

リージョン名	リージョン	インスタンスタイプ
米国東部（オハイオ）	us-east-2	t2.small
米国西部（北カリフォルニア）	us-west-1	m1.small
アジアパシフィック（ムンバイ）	ap-south-1	t2.small
アジアパシフィック（シンガポール）	ap-southeast-1	m1.small
アジアパシフィック（ソウル）	ap-northeast-2	t2.small
カナダ（中部）	ca-central-1	t2.small
欧州（フランクフルト）	eu-central-1	t2.small

リージョン名	リージョン	インスタンスタイプ
欧州（ロンドン）	eu-west-2	t2.small
欧州（パリ）	eu-west-3	t2.small
南米（サンパウロ）	sa-east-1	m1.small

その他のサポートされるAmazon EC2インスタンス

パイプライン定義でインスタンスタイプを指定しなかった場合にデフォルトで作成されるインスタンスに加えて、以下のインスタンスがサポートされます。

次の表に、 が AWS Data Pipeline サポートし、指定すれば作成できる Amazon EC2 インスタンスを示します。

インスタンスクラス	インスタンスタイプ
汎用	t2.nano | t2.micro | t2.small | t2.medium | t2.large
コンピューティングの最適化	c3.large | c3.xlarge | c3.2xlarge | c3.4xlarge | c3.8xlarge | c4.large | c4.xlarge | c4.2xlarge | c4.4xlarge | c4.8xlarge | c5.xlarge | c5.9xlarge | c5.2xlarge | c5.4xlarge | c5.9xlarge | c5.18xlarge | c5d.xlarge | c5d.2xlarge | c5d.4xlarge | c5d.9xlarge | c5d.18xlarge
メモリ最適化	m3.medium | m3.large | m3.xlarge | m3.2xlarge | m4.large | m4.xlarge | m4.2xlarge | m4.4xlarge | m4.10xlarge | m4.16xlarge | m5.xlarge | m5.2xlarge | m5.4xlarge | m5.12xlarge | m5.24xlarge | m5d.xlarge | m5d.2xlarge | m5d.4xlarge | m5d.12xlarge | m5d.24xlarge r3.large | r3.xlarge | r3.2xlarge | r3.4xlarge | r3.8xlarge | r4.large | r4.xlarge | r4.2xlarge | r4.4xlarge | r4.8xlarge | r4.16xlarge
ストレージの最適化	i2.xlarge | i2.2xlarge | i2.4xlarge | i2.8xlarge | hs1.8xlarge | g2.2xlarge | g2.8xlarge | d2.xlarge | d2.2xlarge | d2.4xlarge | d2.8xlarge

Amazon EMR クラスターでサポートされるAmazon EC2インスタンス

この表は、 がサポートし、Amazon EMR クラスター用に作成できる Amazon EC2 インスタンスを示しています。 AWS Data Pipeline 詳細については、Amazon EMR管理ガイドのサポートされるインスタンスタイプを参照してください。

インスタンスクラス	インスタンスタイプ
汎用	m1.small | m1.medium | m1.large | m1.xlarge | m3.xlarge | m3.2xlarge
コンピューティングの最適化	c1.medium | c1.xlarge | c3.xlarge | c3.2xlarge | c3.4xlarge | c3.8xlarge | cc1.4xlarge| cc2.8xlarge | c4.large | c4.xlarge | c4.2xlarge| c4.4xlarge | c4.8xlarge | c5.xlarge | c5.9xlarge | c5.2xlarge | c5.4xlarge | c5.9xlarge | c5.18xlarge | c5d.xlarge | c5d.2xlarge | c5d.4xlarge | c5d.9xlarge | c5d.18xlarge
メモリ最適化	m2.xlarge | m2.2xlarge | m2.4xlarge | r3.xlarge | r3.2xlarge | r3.4xlarge | r3.8xlarge | cr1.8xlarge | m4.large | m4.xlarge | m4.2xlarge | m4.4xlarge | m4.10xlarge | m4.16large | m5.xlarge | m5.2xlarge | m5.4xlarge | m5.12xlarge | m5.24xlarge | m5d.xlarge | m5d.2xlarge | m5d.4xlarge | m5d.12xlarge | m5d.24xlarge | r4.large | r4.xlarge | r4.2xlarge | r4.4xlarge | r4.8xlarge | r4.16xlarge
ストレージの最適化	h1.4xlarge | hs1.2xlarge | hs1.4xlarge| hs1.8xlarge | i2.xlarge | i2.2xlarge | i2.4large | i2.8xlarge | d2.xlarge | d2.2xlarge| d2.4xlarge | d2.8xlarge
高速コンピューティング	g2.2xlarge | cg1.4xlarge

AWS Data Pipeline 概念

開始する前に、 の主要な概念とコンポーネントについてお読みください AWS Data Pipeline。

内容

- パイプライン定義
- パイプラインのコンポーネント、インスタンス、試行
- Task Runner
- データノード
- データベース
- アクティビティ
- 前提条件
- リソース
- アクション

パイプライン定義

パイプライン定義は、ビジネスロジックを に伝達する方法です AWS Data Pipeline。これには、以下の情報が含まれています。

- データソースの名前、場所、形式
- データを変換するアクティビティ
- これらのアクティビティのスケジュール
- アクティビティおよび前提条件を実行するリソース
- アクティビティをスケジュールする前に完了する必要がある前提条件
- パイプライン実行に伴うステータスの更新を警告する方法

パイプライン定義から、 はタスク AWS Data Pipeline を決定し、スケジュールして、タスクランナーに割り当てます。タスクが正常に完了しなかった場合、 は指示に従ってタスクを AWS Data Pipeline 再試行し、必要に応じて別のタスクランナーに再割り当てします。タスクが繰り返し失敗する場合は、通知するようにパイプラインを設定できます。

例えば、パイプライン定義で、アプリケーションによって生成されたログファイルを、2013 年の各月に Amazon S3 バケットにアーカイブすることを指定できます。 AWS Data Pipeline は、月の日数が 30 日、31 日、28 日、29 日のいずれであるかに関係なく、それぞれが 1 か月分のデータをコピーする 12 個のタスクを作成します。

パイプライン定義は、次のような方法で作成できます。

- AWS Data Pipeline コンソールを使用してグラフィックで表示
- コマンドラインインターフェイスで使用される形式の JSON ファイルを記述することによってテキストで作成する
- いずれかの AWS SDK または AWS Data Pipeline API でウェブサービスを呼び出すことによってプログラムで作成する

パイプライン定義には、次のタイプのコンポーネントを含めることができます。

パイプラインコンポーネント

データノード

タスクの入力データの場所または出力データが保存される場所。

アクティビティ

コンピューティングリソースと通常、入出力データノードを使用して、スケジュールに従って実行する作業の定義。

前提条件

アクションを実行する前に true である必要がある条件ステートメント。

リソース

パイプラインで定義する作業を実行するコンピューティングリソース。

アクション

アクティビティの失敗など、指定された条件が満たされた場合にトリガーされるアクション。

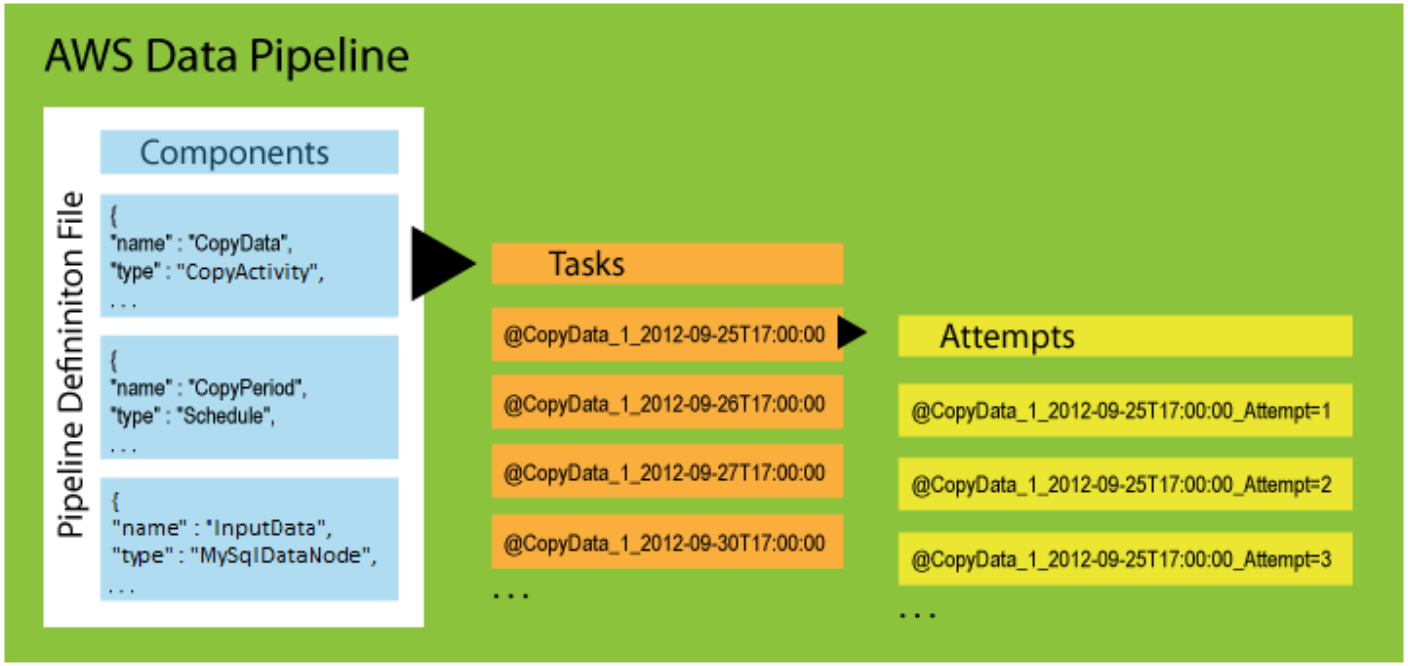
詳細については、「パイプライン定義ファイルの構文」を参照してください。

パイプラインのコンポーネント、インスタンス、試行

スケジュールされたパイプラインに関連付けられる項目には 3 つのタイプがあります。

- パイプラインコンポーネント – パイプラインコンポーネントはパイプラインのビジネスロジックを表し、パイプライン定義のさまざまなセクションによって表されます。パイプラインコンポーネントは、ワークフローのデータソース、アクティビティ、スケジュール、および前提条件を指定します。これらは親コンポーネントからプロパティを継承できます。コンポーネント間の関係は参照によって定義されます。パイプラインコンポーネントは、データ管理のルールを定義します。
- インスタンス — がパイプライン AWS Data Pipeline を実行すると、パイプラインコンポーネントをコンパイルして、一連の実用的なインスタンスを作成します。各インスタンスには、特定のタスクを実行するためのすべての情報が含まれています。インスタンスの完全なセットは、pipeline の To-Do リストです。 は、処理するタスクランナーにインスタンスを AWS Data Pipeline 渡します。
- 試行 – 堅牢なデータ管理を提供するために、 AWS Data Pipeline は失敗したオペレーションを再試行します。この処理は、タスクが最大許容再試行回数に到達するまで続行されます。試行オブジェクトは、さまざまな試行、結果、および失敗の理由（該当する場合）を追跡します。基本的に、カウンター AWS Data Pipeline パフォーマンスを持つインスタンスは、Amazon EMR クラスターや EC2 インスタンスなど、以前の試行と同じリソースを使用して再試行します。

> **Note**
>
> 失敗したタスクの再実行は耐障害性戦略の重要な部分であり、 AWS Data Pipeline のパイプライン定義では再試行を制御するための条件としきい値を提供します。ただし、 AWS Data Pipeline では指定されたすべての再試行回数に到達するまで失敗がレポートされないため、再試行回数が多すぎると、復旧不可能な障害の検出が遅れる可能性があります。再試行が AWS リソースで実行されている場合、余分な再試行によって追加料金が発生することがあります。そのため、再試行や関連する設定を制御するために使用する AWS Data Pipeline デフォルト設定を超えるのが適切である場合は、慎重に検討してください。

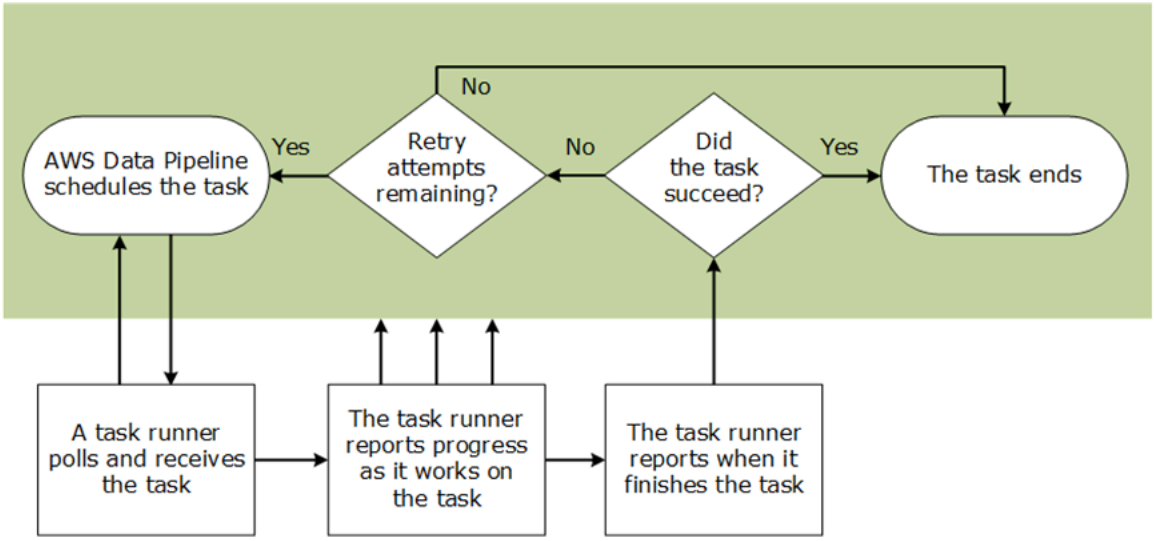


Task Runner

タスクランナーは、タスク AWS Data Pipeline をポーリングし、それらのタスクを実行するアプリケーションです。

Task Runner は、 AWS Data Pipelineで提供されるタスクランナーのデフォルトの実装です。Task Runner をインストールして設定すると、アクティブ化したパイプラインに関連付けられた AWS Data Pipeline タスクがポーリングされます。タスクが Task Runner に割り当てられている場合、そのタスクを実行し、そのステータスを AWS Data Pipelineにレポートします。

次の図は、 AWS Data Pipeline とタスクランナーがやり取りしてスケジュールされたタスクを処理する方法を示しています。タスクは、 AWS Data Pipeline サービスがタスクランナーと共有する個別の作業単位です。通常、複数のタスクを生成するアクティビティやリソースの一般的な定義であるパイプラインとは異なります。



Task Runner を使用してパイプラインを処理するには、次の 2 とおりの方法があります。

- AWS Data Pipeline は、 AWS Data Pipeline ウェブサービスによって起動および管理されるリソースに Task Runner をインストールします。
- 長期間実行されている EC2 インスタンスやオンプレミスサーバーなど、お客様が管理するコンピューティングリソースに、お客様が Task Runner をインストールします。

Task Runner の操作方法の詳細については、「Task Runnerの操作」を参照してください。

データノード

では AWS Data Pipeline、データノードは、パイプラインアクティビティが入力または出力として使用するデータの場所とタイプを定義します。 は次のタイプのデータノード AWS Data Pipeline をサポートします。

D ynamoDBDataノード

使用する HiveActivity または EmrActivity のデータを格納する DynamoDB テーブル。

SqlDataNode

パイプラインアクティビティで使用するデータを表す SQL テーブルとデータベースクエリ。

> **Note**
>
> 以前は、 が使用され MySqlDataNode ていました。 SqlDataNode 代わりに を使用します。

RedshiftDataNode

使用する RedshiftCopyActivity のデータを格納する Amazon Redshift テーブル。

S3DataNode

パイプラインアクティビティで使用する 1 つ以上のファイルを格納する Amazon S3 の場所。

データベース

AWS Data Pipeline は、次のタイプのデータベースをサポートしています。

JdbcDatabase

JDBC データベース。

RdsDatabase

Amazon RDS データベース。

RedshiftDatabase

Amazon Redshift データベース。

アクティビティ

では AWS Data Pipeline、アクティビティは実行する作業を定義するパイプラインコンポーネントです。 AWS Data Pipeline は、ある場所から別の場所へのデータの移動、Hive クエリの実行など、一般的なシナリオに対応するいくつかのパッケージ済みのアクティビティを提供します。アクティビティは拡張可能であるため、独自のカスタムスクリプトを実行して、無限の組み合わせをサポートできます。

AWS Data Pipeline では、次のタイプのアクティビティがサポートされています。

CopyActivity

ある場所から別の場所にデータをコピーします。

EmrActivity

Amazon EMR クラスターを実行します。

HiveActivity

Amazon EMR クラスターで Hive クエリを実行します。

HiveCopyActivity

高度なデータフィルタリングのサポートおよび S3DataNode と D ynamoDBDataノード のサポートを使用して、Amazon EMR クラスターで Hive クエリを実行します。

PigActivity

Amazon EMR クラスターで Pig スクリプトを実行します。

RedshiftCopyActivity

Amazon Redshift テーブルとの間でデータをコピーします。

ShellCommandActivity

アクティビティとしてカスタム UNIX/Linux シェルコマンドを実行します。

SqlActivity

データベースに対する SQL クエリを実行します。

一部のアクティビティでは、データとデータベーステーブルのステージングについて特別なサポートが提供されています。詳細については、「パイプラインのアクティビティによるデータとテーブルのステージング」を参照してください。

前提条件

では AWS Data Pipeline、前提条件は、アクティビティを実行する前に true となる必要がある条件ステートメントを含むパイプラインコンポーネントです。例えば、前提条件は、パイプラインアクティビティがコピーを試みる前にソースデータが存在するかどうかを確認できます。 は、データベーステーブルが存在するかどうか、Amazon S3 キーが存在するかどうかなど、一般的なシナリオに対応するいくつかのパッケージ済みの前提条件 AWS Data Pipeline を提供します。ただし、前提条件は拡張可能であるため、独自のカスタムスクリプトを実行して、無限の組み合わせをサポートできます。

前提条件には、システム管理型の前提条件とユーザー管理型の前提条件の 2 つの種類があります。システム管理の前提条件は、ユーザーに代わって AWS Data Pipeline ウェブサービスによって実行され、計算リソースは必要ありません。ユーザー管理型の前提条件は、runsOn または workerGroup フィールドを使用して指定したコンピューティングリソースでのみ実行されます。workerGroup リソースは、前提条件を使用するアクティビティから派生します。

システム管理型の前提条件

D ynamoDBDataが存在する

データが特定の DynamoDB テーブルに存在するかどうかを確認します。

D ynamoDBTableが存在する

DynamoDB テーブルが存在するかどうかを確認します。

S3KeyExists

Amazon S3 キーが存在するかどうかを確認します。

S3PrefixNotEmpty

Amazon S3 プレフィックスが空であるかどうかを確認します。

ユーザー管理型の前提条件

存在する

データノードが存在するかどうかを確認します。

ShellCommandPrecondition

前提条件としてカスタム Unix/Linux シェルコマンドを実行します。

リソース

では AWS Data Pipeline、リソースは、パイプラインアクティビティが指定する作業を実行する計算リソースです。 は次のタイプのリソース AWS Data Pipeline をサポートします。

Ec2Resource

パイプラインアクティビティによって定義された作業を実行する EC2 インスタンス。

EmrCluster

パイプラインアクティビティ (EmrActivity など) によって定義される作業を実行する Amazon EMR クラスター。

リソースは、作業データセットと同じリージョンで実行できます。 AWS Data Pipelineと異なるリージョンであってもかまいません。詳細については、「複数のリージョンにあるリソースとパイプラインの使用」を参照してください。

リソースの制限

AWS Data Pipeline は、大量の同時タスクに対応するようにスケールし、大規模なワークロードを処理するために必要なリソースを自動的に作成するように設定できます。これらの自動的に作成されたリソースは、お客様の管理下にあり、AWS アカウントのリソースに対する制限の対象となります。例えば、データを処理 AWS Data Pipeline するために 20 ノードの Amazon EMR クラスターを自動的に作成するように を設定し、AWS アカウントの EC2 インスタンス制限が 20 に設定されている場合、使用可能なバックフィルリソースが誤って使い果たされる可能性があります。そのため、設計時にこれらのリソースの制限を考慮するか、またはアカウントの制限を適宜増やします。サービスの制限に関する詳細については、AWS 全般リファレンスの「AWS サービスの制限」を参照してください。

> **Note**
>
> 制限は `Ec2Resource` コンポーネントオブジェクトごとに 1 個のインスタンスです。

サポートされているプラットフォーム

パイプラインは、以下のプラットフォームでリソースを起動できます。

EC2-Classic

お客様のリソースは他のユーザー様と共有する単一のフラットネットワーク内で稼働します。

EC2-VPC

お客様のリソースはご自分の AWS アカウントから論理的に独立した Virtual Private Cloud (VPC) 内で稼働します。

AWS アカウントは、リソースを両方のプラットフォームで起動できるか、EC2-VPC だけで起動できるかが、リージョンごとに決まっています。詳細については、Amazon EC2 ユーザーガイド」の「サポートされているプラットフォーム」を参照してください。

AWS アカウントで EC2-VPC のみがサポートされている場合は、各 AWS リージョンにデフォルトの VPC が作成されます。デフォルトでは、リソースはデフォルト VPC 内のデフォルトサブネットで起動されます。または、リソースの設定時に、デフォルト以外の VPC を作成し、サブネットのいずれかを指定することもできます。その場合、リソースは、指定されたデフォルト以外の VPC 内のサブネットで起動されます。

VPC でインスタンスを起動する場合は、その VPC 用に作成されたセキュリティグループを指定する必要があります。VPC でインスタンスを起動する場合、EC2-Classic 用に作成したセキュリティグループは指定できません。また、VPC のセキュリティグループを識別するセキュリティグループの名前ではなく、セキュリティグループの ID を使用する必要があります。

Amazon EMR クラスターと AWS Data Pipelineを使用した Amazon EC2 スポットインスタンス

パイプラインでは、Amazon EMR クラスターリソースのタスクノードとして Amazon EC2 スポットインスタンスを使用できます。デフォルトでは、パイプラインはオンデマンドインスタンスを使用します。スポットインスタンスでは、予備の EC2 インスタンスを使用して実行できます。スポットインスタンスという価格モデルは、オンデマンドインスタンス価格モデルやリザーブドインスタンス価格モデルを補完するものであり、アプリケーションによっては、コンピューティング性能を調達するうえで最もコスト効果の高い選択肢となる可能性があります。詳細については、Amazon EC2 スポットインスタンスの製品ページを参照してください。

スポットインスタンスを使用する場合、クラスターの起動時に はスポットインスタンス の上限価格を Amazon EMR AWS Data Pipeline に送信します。クラスターの作業は、`taskInstanceCount`フィールドを使用して定義したスポットインスタンスタスクノードの数に自動的に割り当てられます。 は、オンデマンドコアノードがパイプラインを実行できるように、タスクノードのスポットインスタンス AWS Data Pipeline を制限します。

失敗または完了したパイプラインリソースインスタンスを編集してスポットインスタンスを追加できます。パイプラインがクラスターを再起動したときには、タスクノードとしてスポットインスタンスが使用されます。

スポットインスタンスに関する考慮事項

でスポットインスタンスを使用する場合 AWS Data Pipeline、次の考慮事項が適用されます。

- スポットインスタンスは、スポットインスタンスの価格がインスタンスの上限価格を超過したときや、Amazon EC2 の容量の理由で、終了することがあります。ただし、 は、常にオンデマンドインスタンスであり、終了の対象ではないコアノードを持つクラスター AWS Data Pipeline を採用するため、データが失われることはありません。
- スポットインスタンス、容量を非同期的に満たすため、開始までに時間がかかることがあります。したがって、スポットインスタンスのパイプラインは、同等のオンデマンドインスタンスのパイプラインより自動速度が遅くなる可能性があります。
- 上限価格が低すぎるときなど、スポットインスタンスを取得できない場合、クラスターが実行されないことがあります。

アクション

AWS Data Pipeline アクションは、成功、失敗、遅延アクティビティなど、特定のイベントが発生したときにパイプラインコンポーネントが実行するステップです。アクティビティのイベントフィールドでアクションを参照します。たとえば、`snsalarm` の `onLateAction` フィールドで `EmrActivity` を参照します。

AWS Data Pipeline は、パイプラインとそのコンポーネントのステータスを無人で示す主な方法として Amazon SNS 通知に依存しています。詳細については、「Amazon SNS」を参照してください。SNS 通知に加えて、 AWS Data Pipeline コンソールと CLI を使用してパイプラインのステータス情報を取得できます。

AWS Data Pipeline は、次のアクションをサポートします。

SnsAlarm

`onSuccess`、`OnFail`、および `onLateAction` イベントに基づいて、トピックに SNS 通知を送信するアクション。

終了

保留中または未完了のアクティビティ、リソース、またはデータノードの取り消しをトリガーするアクション。`onSuccess`、`OnFail`、または `onLateAction` を含むアクションを終了することはできません。

パイプラインの事前モニタリング

問題を検出する最善の方法は、パイプラインを最初から積極的にモニタリングすることです。パイプラインコンポーネントが失敗した場合や、スケジュールされた開始時刻までに開始されない場合など、特定の状況やイベントを通知するようにパイプラインコンポーネントを設定できます。 AWS Data Pipeline は、onSuccess、、 などの Amazon SNS 通知に関連付けることができるパイプラインコンポーネントにイベントフィールドを提供することでOnFail、通知を簡単に設定できるようにしますonLateAction。

のセットアップ AWS Data Pipeline

AWS Data Pipeline を初めて使用する場合は、事前に以下のタスクを完了してください。

タスク

- にサインアップする AWS
- AWS Data Pipeline およびパイプラインリソースの IAM ロールを作成する
- 必要なアクションを実行するための IAM プリンシパル (ユーザーとグループ) の許可
- プログラムによるアクセス権を付与する

これらのタスクを完了したら、 の使用を開始できます AWS Data Pipeline。基本的なチュートリアルについては、「AWS Data Pipeline の使用の開始」を参照してください。

にサインアップする AWS

Amazon Web Services (AWS) にサインアップすると、 を含む AWS のすべてのサービスに AWS アカウントが自動的にサインアップされます AWS Data Pipeline。料金は、使用するサービスの料金のみが請求されます。 AWS Data Pipeline 使用率の詳細については、AWS Data Pipeline」を参照してください。

にサインアップする AWS アカウント

がない場合は AWS アカウント、次のステップを実行して作成します。

にサインアップするには AWS アカウント

1. https://portal.aws.amazon.com/billing/signup を開きます。
2. オンラインの手順に従います。

   サインアップ手順の一環として、通話呼び出しを受け取り、電話キーパッドで検証コードを入力するように求められます。

   にサインアップすると AWS アカウント、 AWS アカウントのルートユーザーが作成されます。ルートユーザーには、アカウントのすべての AWS サービス とリソースへのアクセス権があります。セキュリティのベストプラクティスとして、ユーザーに管理アクセスを割り当て、ルートユーザーのみを使用してルートユーザーアクセスが必要なタスクを実行してください。

AWS サインアッププロセスが完了すると、 から確認メールが送信されます。https://aws.amazon.com/ の [マイアカウント] を選んで、いつでもアカウントの現在のアクティビティを表示し、アカウントを管理できます。

管理アクセスを持つユーザーを作成する

にサインアップしたら AWS アカウント、 を保護し AWS アカウントのルートユーザー、 を有効にして AWS IAM Identity Center、日常的なタスクにルートユーザーを使用しないように管理ユーザーを作成します。

のセキュリティ保護 AWS アカウントのルートユーザー

1. ルートユーザーを選択し、 AWS アカウント E メールアドレスを入力して、アカウント所有者AWS Management Consoleとして にサインインします。次のページでパスワードを入力します。

   ルートユーザーを使用してサインインする方法については、AWS サインイン ユーザーガイドの「ルートユーザーとしてサインインする」を参照してください。
2. ルートユーザーの多要素認証 (MFA) を有効にします。

   手順については、「IAM ユーザーガイド」の AWS アカウント 「ルートユーザーの仮想 MFA デバイスを有効にする (コンソール) 」を参照してください。

管理アクセスを持つユーザーを作成する

1. IAM アイデンティティセンターを有効にします。

   手順については、「AWS IAM Identity Center ユーザーガイド」の「AWS IAM Identity Centerの有効化」を参照してください。
2. IAM アイデンティティセンターで、ユーザーに管理アクセスを付与します。

   を ID ソース IAM アイデンティティセンターディレクトリ として使用する方法のチュートリアルについては、「 ユーザーガイド」の「デフォルト でユーザーアクセス IAM アイデンティティセンターディレクトリを設定するAWS IAM Identity Center 」を参照してください。

管理アクセス権を持つユーザーとしてサインインする

- IAM アイデンティティセンターのユーザーとしてサインインするには、IAM アイデンティティセンターのユーザーの作成時に E メールアドレスに送信されたサインイン URL を使用します。

IAM Identity Center ユーザーを使用してサインインする方法については、「 AWS サインインユーザーガイド」の AWS 「 アクセスポータルにサインインする」を参照してください。

追加のユーザーにアクセス権を割り当てる

1. IAM アイデンティティセンターで、最小特権のアクセス許可を適用するというベストプラクティスに従ったアクセス許可セットを作成します。

   手順については、「AWS IAM Identity Center ユーザーガイド」の「権限設定を作成する」を参照してください。
2. グループにユーザーを割り当て、そのグループにシングルサインオンアクセス権を割り当てます。

   手順については、「AWS IAM Identity Center ユーザーガイド」の「グループの参加」を参照してください。

AWS Data Pipeline およびパイプラインリソースの IAM ロールを作成する

AWS Data Pipeline には、アクションを実行し、 AWS リソースにアクセスするためのアクセス許可を決定する IAM ロールが必要です。パイプラインロールは が AWS Data Pipeline 持つアクセス許可を決定し、リソースロールは EC2 インスタンスなどのパイプラインリソースで実行されているアプリケーションが持つアクセス許可を決定します。これらのロールは、パイプラインを作成するときに指定します。カスタムロールを指定せず、デフォルトのロール `DataPipelineDefaultRole` および `DataPipelineDefaultResourceRole` を使用する場合でも、まずロールを作成し、アクセス許可ポリシーをアタッチする必要があります。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。

必要なアクションを実行するための IAM プリンシパル (ユーザーとグループ) の許可

パイプラインを操作するには、アカウント内の IAM プリンシパル (ユーザーまたはグループ) が、必須のAWS Data Pipeline アクションと、パイプラインで定義されている他のサービスに対するアクションを実行できる必要があります。

アクセス許可を簡素化するために、 AWSDataPipeline_FullAccessマネージドポリシーを IAM プリンシパルにアタッチできます。この管理ポリシーにより、プリンシパルは、 AWS Data Pipeline カスタムロールが指定されていない場合に、ユーザーが必要とするすべてのアクションと、 で使用されるデフォルトのロールに対する iam:PassRole アクションを実行できます。

この管理ポリシーを慎重に評価し、ユーザーが必要とするものだけに許可を制限することを強くお勧めします。必要に応じて、このポリシーを開始点として使用し、許可を削除して、IAM プリンシパルにアタッチできる、より制限の厳しいインラインアクセス許可ポリシーを作成してください。詳細およびアクセス許可ポリシーの例については、AWS Data Pipeline のポリシー例を参照してください。

パイプラインを使用するすべての IAM プリンシパルにアタッチされたポリシーに、次の例のようなポリシーステートメントを含める必要があります。このステートメントにより、IAM プリンシパルは、パイプラインが使用するロールで PassRole アクションを実行できます。デフォルトのロールを使用しない場合は、*MyPipelineRole* および *MyResourceRole* を、作成したカスタムロールに置き換えてください。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": "iam:PassRole",
            "Effect": "Allow",
            "Resource": [
                "arn:aws:iam::*:role/MyPipelineRole",
                "arn:aws:iam::*:role/MyResourceRole"
            ]
        }
    ]
}
```

次の手順では、IAM グループを作成し、グループにAWSDataPipeline_FullAccess管理ポリシーをアタッチしてから、グループにユーザーを追加する方法を示します。この手順は、任意のインラインポリシーで使用できます。

ユーザーグループを作成してAWSDataPipeline_FullAccessポリシーを**DataPipelineDevelopers**アタッチするには

1. IAM コンソール (https://console.aws.amazon.com/iam/) を開きます。
2. ナビゲーションペインで [Groups]、[Create New Group] の順に選択します。

3. [Group Name] (グループ名)(例: **DataPipelineDevelopers**) を入力し、[Next Step] (次のステップ) を選択します。
4. [Filter] (フィルタ) に **AWSDataPipeline_FullAccess** を入力し、リストから選択します。
5. [Next Step]、[Create Group] の順に選択します。
6. ユーザーをグループに追加するには

   a. グループのリストから、作成したグループを選択します。
   b. [Group Actions] で、[Add Users to Group] を選択します。
   c. 追加するユーザーをリストから選択し、[Add Users to Group] (グループにユーザーを追加) を選択します。

プログラムによるアクセス権を付与する

ユーザーが の AWS 外部で を操作する場合は、プログラムによるアクセスが必要です AWS Management Console。プログラムによるアクセスを許可する方法は、 にアクセスするユーザーのタイプによって異なります AWS。

ユーザーにプログラマチックアクセス権を付与するには、以下のいずれかのオプションを選択します。

プログラマチックアクセス権を必要とするユーザー	目的	方法
ワークフォースアイデンティティ (IAM Identity Center で管理されているユーザー)	一時的な認証情報を使用して、、 AWS SDKs AWS CLI、または AWS APIs。	使用するインターフェイス用の手引きに従ってください。 • については AWS CLI、「 ユーザーガイド」の AWS CLI 「 を使用するための の設定 AWS IAM Identity CenterAWS Command Line Interface 」を参照してください。 • AWS SDKs 、ツール、AWS APIs「 SDK とツールのリファレンスガイド」

プログラマチックアクセス権を必要とするユーザー	目的	方法
		の「IAM Identity Center 認証」を参照してください。 AWS SDKs
IAM	一時的な認証情報を使用して、、 AWS SDKs AWS CLI、または AWS APIs。	「IAM ユーザーガイド」の「 AWS リソースでの一時的な認証情報の使用」の手順に従います。
IAM	(非推奨) 長期認証情報を使用して、、 AWS SDKs AWS CLI、または AWS APIs。	使用するインターフェイス用の手引きに従ってください。 • については AWS CLI、「 AWS Command Line Interface ユーザーガイド」の「IAM ユーザー認証情報を使用した認証」を参照してください。 • AWS SDKs「 SDK とツールのリファレンスガイド」の「長期的な認証情報を使用した認証」を参照してください。 AWS SDKs • AWS APIsユーザーガイド」の「IAM ユーザーのアクセスキーの管理」を参照してください。

AWS Data Pipeline の使用の開始

AWS Data Pipeline を使用すると、繰り返し発生するデータ処理のワークロードを確実に、優れたコスト効率で、順序付け、スケジュール、実行、および管理することができます。このサービスを使用することで、ビジネスロジックに基き、オンプレミスとクラウドの両方で、構造化データと非構造化データを使用して、ETL (抽出、変換、ロード) アクティビティを容易に設計できます。

AWS Data Pipeline を使用するには、データ処理のためのビジネスロジックを指定するパイプライン定義を作成します。一般的なパイプライン定義は、実行する作業を定義するアクティビティと、入力データと出力データの場所と型を定義するデータノードで構成されます。

このチュートリアルでは、Apache ウェブサーバーログに含まれる GET リクエストの数をカウントするシェルコマンドスクリプトを実行します。このパイプラインは、15 分ごとに 1 時間実行され、各イテレーションで出力を Amazon S3 に書き込みます。

前提条件

開始する前に、「のセットアップ AWS Data Pipeline」のタスクを完了します。

パイプラインオブジェクト

このパイプラインでは以下のオブジェクトを使用します。

ShellCommandActivity

入力ログファイルを読み取り、エラー数をカウントします。

S3DataNode (input)

入力ログファイルが含まれる S3 バケット。

S3DataNode (output)

出力用の S3 バケット。

Ec2Resource

アクティビティを実行するために AWS Data Pipeline で使用されるコンピューティングリソース。

大量のログファイルデータがある場合は、EC2 インスタンスではなく EMR クラスターを使用してファイルを処理できるように、パイプラインを設定することができます。

スケジュール

アクティビティを 15 分ごとに 1 時間実行することを定義します。

タスク

- パイプラインの作成
- 実行中のパイプラインのモニタリング
- 出力の表示
- パイプラインの削除

パイプラインの作成

AWS Data Pipeline の使用を開始するための最も簡単な方法は、テンプレートと呼ばれるパイプライン定義を使用することです。

パイプラインを作成するには

1. AWS Data Pipeline コンソール (https://console.aws.amazon.com/datapipeline/) を開きます。
2. ナビゲーションバーから、リージョンを選択します。お客様は場所に関係なく、使用できるリージョンをどれでも選択できます。多くの AWS リソースはリージョンに固有ですが、AWS Data Pipeline ではパイプラインと異なるリージョンにあるリソースを使用することができます。
3. 表示される最初の画面は、現在のリージョンにパイプラインを作成しているかどうかによって異なります。

   a. このリージョンにパイプラインを作成していない場合は、コンソールに初期画面が表示されます。[Get started now] を選択します。

   b. このリージョンにパイプラインを既に作成している場合、コンソールには、リージョンのパイプラインを一覧表示するページが表示されます。[Create new pipeline] (新しいパイプラインの作成) を選択します。
4. [Name] (名前) に、パイプラインの名前を入力します。
5. (オプション) [Description] (説明) に、パイプラインの説明を入力します。
6. [Source] で、[Build using a template] を選択し、[Getting Started using ShellCommandActivity] というテンプレートを選択します。

7. テンプレートを選択すると開く [Parameters] セクションで、[S3 input folder] と [Shell command to run] の値をデフォルトのままにしておきます。[S3 output folder] の横のフォルダーアイコンをクリックし、いずれかのバケットまたはフォルダーを選択して、[Select] をクリックします。
8. [Schedule] で、値をデフォルトのままにしておきます。パイプラインをアクティブ化すると、パイプライン実行が開始され、15 分ごとに 1 時間続きます。

   代わりに、[Run once on pipeline activation] を選択することもできます。
9. [Pipeline Configuration] (パイプラインの設定)で、ログ記録を有効のままにします。[S3 location for logs] (S3 ログの場所) の下のフォルダーアイコンを選択します。いずれかのバケットまたはフォルダーを選択して、[Select] (選択) を選択します。

   必要に応じて、代わりにログ記録を無効にすることができます。
10. [Security/Access] (セキュリティ/アクセス) で、[IAM roles] (IAM ロール) が [Default] (デフォルト) に設定されたままにします。
11. [Activate] (アクティブにする) をクリックします。

    必要に応じて、[Edit in Architect] (アーキテクトで編集する) を選択し、このパイプラインを変更できます。例えば、前提条件を追加できます。

実行中のパイプラインのモニタリング

パイプラインをアクティブ化すると、パイプラインの進捗状況をモニタリングできる [Execution details] ページに移動します。

パイプラインの進捗状況をモニタリングするには

1. [Update] をクリックするか F5 キーを押して、ステータスの表示を更新します。

   > **Tip**
   > 実行が表示されない場合は、パイプラインのスケジュールされた開始日時と終了日時が [Start (in UTC)] から [End (in UTC)] までの期間に含まれていることを確認し、[Update] をクリックします。
2. パイプラインのすべてのオブジェクトのステータスが `FINISHED` であれば、スケジュールされたタスクは正常に完了しています。

3. パイプラインが正常に完了していない場合は、問題に関するパイプラインの設定を確認してください。インスタンスがパイプラインの実行に失敗した場合または実行を完了できなかった場合のトラブルシューティングに関する詳細については、「一般的な問題を解決する」を参照してください。

出力の表示

Amazon S3 コンソールを開き、バケットに移動します。パイプラインを 15 分ごとに 1 時間実行した場合は、4 つのタイムスタンプ付きサブフォルダーが表示されます。各サブフォルダーには、出力が含まれた `output.txt` という名前のファイルがあります。毎回同じ入力ファイルでスクリプトを実行したため、出力ファイルも同じになります。

パイプラインの削除

料金の発生を停止するには、パイプラインを削除します。パイプラインを削除すると、パイプライン定義および関連するすべてのオブジェクトが削除されます。

パイプラインを削除するには

1. [List Pipelines] (パイプラインの一覧表示) ページで、パイプラインを選択します。
2. [Actions] (アクション) をクリックしてから、[Delete] (削除) を選択します。
3. 確認を求めるメッセージが表示されたら、[削除] を選択します。

このチュートリアルからの出力を完了したら、Amazon S3 バケットから出力フォルダーを削除します。

パイプラインの使用

コマンドラインインターフェイス (CLI)、または AWS SDKを使用して、パイプラインを管理、作成、変更することができます。以下のセクションでは、AWS Data Pipeline の基本概念を紹介し、パイプラインの使用方法を示します。

> **Important**
>
> 開始する前に、「のセットアップ AWS Data Pipeline」を参照してください。

目次

- パイプラインの作成
- パイプラインの表示
- パイプラインの編集
- パイプラインのクローン作成
- パイプラインのタグ付け
- パイプラインの非アクティブ化
- パイプラインの削除
- パイプラインのアクティビティによるデータとテーブルのステージング
- 複数のリージョンにあるリソースとパイプラインの使用
- カスケードの失敗と再実行
- パイプライン定義ファイルの構文
- API の使用

パイプラインの作成

AWS Data Pipeline でパイプラインを作成する方法は複数あります。

- AWS Command Line Interface (CLI) を使用して、すぐに使えるように提供されているテンプレートからパイプラインを作成する。詳細については、「CLI を使用して Data Pipeline テンプレートからパイプラインを作成する」を参照してください。
- AWS Command Line Interface (CLI) で JSON 形式のパイプライン定義ファイルを使用する。

- AWS SDK と言語固有の API を使用する。詳細については、「API の使用」を参照してください。

CLI を使用して Data Pipeline テンプレートからパイプラインを作成する

Data Pipeline では、設定済みのパイプライン定義 (テンプレート) がいくつか提供されています。テンプレートを使用すると、AWS Data Pipeline の使用を迅速に開始することができます。これらのテンプレートは、Amazon S3 ロケーションのパブリックバケット (`s3://datapipeline-us-east-1/templates/`) にあります。これらの定義済みテンプレートは、特定のユースケースを実現するために作成されたもので、パイプラインの作成にも使用できます。`aws s3 ls --recursive "s3://datapipeline-us-east-1/templates/"` を使用して、使用可能なすべてのテンプレートを一覧表示できます。

CLI を使用してテンプレートからパイプラインを作成する

DynamoDB テーブルを Amazon S3 にエクスポートするパイプラインを作成するとします。この場合に使用するテンプレートは、`s3://datapipeline-us-east-1/templates/DynamoDB Templates/Export DynamoDB table to S3.json` にあります。

テンプレート JSON をダウンロードし、CLI を使用してパイプラインを作成するには

1. `aws s3 cp` CLI または curl を使用してテンプレートをダウンロードします。例:

```
aws s3 cp "s3://datapipeline-us-east-1/templates/DynamoDB Templates/Export DynamoDB
 table to S3.json" <destination directory>
```

2. ダウンロードしたテンプレートを必要に応じて変更します。例えば、最新の EMR リリースバージョンを使用するには、`EmrClusterForBackup` オブジェクトの `releaseLabel` フィールドを変更し、マスターインスタンスタイプとコアインスタンスタイプを変更して、テンプレート内のパラメーターのデフォルト値を変更します。
3. `create-pipeline` CLI を使用してパイプラインを作成します。例:

```
aws datapipeline create-pipeline --name my-ddb-backup-pipeline --unique-id my-ddb-
backup-pipeline --region ap-northeast-1
```

4. 作成したパイプライン ID を書き留めておきます。
5. `put-pipeline-definition` を使用して定義をアップロードします。`--parameter-values` オプションを使用してデフォルト値をオーバーライドしたいパラメータの値を指定します。

テンプレートの詳細については、「テンプレートを選択する」を参照してください。

テンプレートを選択する

以下のテンプレートは、Amazon S3 バケット (`s3://datapipeline-us-east-1/templates/`) からダウンロードできます。

テンプレート

- Getting started using ShellCommandActivity
- AWS CLI コマンドを実行します。
- S3 への DynamoDB テーブルのエクスポート
- S3 からの DynamoDB バックアップデータのインポート
- [Run Job on an Amazon EMR Cluster] (Amazon EMR クラスターでのジョブの実行)
- Amazon S3 への Amazon RDS MySQL テーブルの完全コピー
- Amazon S3 への Amazon RDS MySQL テーブルの増分コピー
- Amazon RDS MySQL テーブルへの S3 データのロード
- [Full Copy of Amazon RDS MySQL Table to Amazon Redshift] (Amazon RedshiftへのAmazon RDS MySQL テーブルの完全コピー)
- [Incremental Copy of an Amazon RDS MySQL Table to Amazon Redshift] (Amazon Redshift への Amazon RDS MySQL テーブルの増分コピー)
- Amazon S3 から Amazon Redshift へのデータのロード

Getting started using ShellCommandActivity

Getting Started using ShellCommandActivity テンプレートでは、ログファイル内の GET リクエストの数をカウントするシェルコマンドスクリプトを実行します。出力は、スケジュールされたパイプライン実行ごとに、タイムスタンプ付きで Amazon S3 の場所に書き込まれます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- ShellCommandActivity
- S3InputNode
- S3OutputNode
- Ec2Resource

AWS CLI コマンドを実行します。

このテンプレートでは、スケジュールされた間隔でユーザー指定の AWS CLI コマンドを実行します。

S3 への DynamoDB テーブルのエクスポート

[Export DynamoDB table to S3] (S3 への DynamoDB テーブルのエクスポート) テンプレートでは、DynamoDB テーブルから Amazon S3 バケットにデータがエクスポートされるように Amazon EMR クラスターをスケジュールします。このテンプレートでは Amazon EMR クラスターを使用します。このクラスターは、DynamoDB テーブルで利用可能なスループットの値に比例してサイズ調整されます。テーブルの IOPS は増やすことができますが、その場合はインポートおよびエクスポート時に追加コストが発生する可能性があります。これまでは、エクスポートで HiveActivity が使用されましたが、現在はネイティブ MapReduce が使用されます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- EmrActivity
- EmrCluster
- D ynamoDBDataノード
- S3DataNode

S3 からの DynamoDB バックアップデータのインポート

[Import DynamoDB backup data from S3] (S3 からの DynamoDB バックアップデータのインポート) テンプレートでは、Amazon S3 にある作成済みの DynamoDB バックアップが DynamoDB テーブルにロードされるように、Amazon EMR クラスターをスケジュールします。DynamoDB テーブル内の既存の項目はバックアップデータ内の該当データで更新され、新しい項目はテーブルに追加されます。このテンプレートでは Amazon EMR クラスターを使用します。このクラスターは、DynamoDB テーブルで利用可能なスループットの値に比例してサイズ調整されます。テーブルの IOPS は増やすことができますが、その場合はインポートおよびエクスポート時に追加コストが発生する可能性があります。これまでは、インポートで HiveActivity が使用されましたが、現在はネイティブ MapReduce が使用されます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- EmrActivity
- EmrCluster

- D ynamoDBDataノード
- S3DataNode
- S3PrefixNotEmpty

[Run Job on an Amazon EMR Cluster] (Amazon EMR クラスターでのジョブの実行)

[Run Job on an Elastic MapReduce Cluster] (Elastic MapReduce クラスターでのジョブの実行) テンプレートでは、指定されたパラメータに基づいて Amazon EMR クラスターを起動し、指定されたスケジュールに基づいてステップの実行を開始します。ジョブが完了すると、EMR クラスターは終了します。オプションでブートストラップアクションを指定することにより、追加ソフトウェアのインストールや、クラスター上にあるアプリケーションの設定変更を行うことができます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- EmrActivity
- EmrCluster

Amazon S3 への Amazon RDS MySQL テーブルの完全コピー

[Full Copy of RDS MySQL Table to S3] (S3 への RDS MySQL テーブルの完全コピー) テンプレートでは、Amazon RDS MySQL テーブル全体をコピーし、Amazon S3 の場所に出力を保存します。出力は CSV ファイル形式で、指定された Amazon S3 の場所のタイムスタンプ付きサブフォルダーに保存されます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- Ec2Resource
- SqlDataNode
- S3DataNode

Amazon S3 への Amazon RDS MySQL テーブルの増分コピー

[Incremental Copy of RDS MySQL Table to S3] (S3 への RDS MySQL テーブルの増分コピー) テンプレートでは、Amazon RDS MySQL テーブルのデータの増分コピーを作成し、出力を Amazon S3 の場所に保存します。Amazon RDS MySQL テーブルには [Last Modified] (最終更新日時) 列が存在する必要があります。

このテンプレートでは、スケジュールされた開始時刻から始まるスケジュールされた間隔で、テーブルに行われた変更をコピーします。スケジュールタイプは時系列形式であるため、特定の時刻から始まる 1 時間についてコピーがスケジュールされている場合、AWS Data Pipeline は、Last Modified 列のタイムスタンプがその 1 時間に含まれるテーブル行をコピーします。テーブルへの物理的な削除はコピーされません。出力は、スケジュールされた実行ごとに、Amazon S3 の場所にあるタイムスタンプ付きサブフォルダに書き込まれます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- Ec2Resource
- SqlDataNode
- S3DataNode

Amazon RDS MySQL テーブルへの S3 データのロード

[Load S3 Data into RDS MySQL Table] (RDS MySQL テーブルへの S3 データのロード) テンプレートでは、下で指定された Amazon S3 ファイルパスから Amazon RDS MySQL テーブルに CSV ファイルがコピーされるように、Amazon EC2 インスタンスをスケジュールします。CSV ファイルにはヘッダー行を含めないようにします。このテンプレートでは、Amazon RDS MySQL テーブルの既存のエントリが、Amazon S3 データに含まれる該当項目で更新され、Amazon S3 データに含まれる新しいエントリが Amazon RDS MySQL テーブルに追加されます。既存のテーブルにデータをロードすることも、新しいテーブルを作成する SQL クエリを指定することもできます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- Ec2Resource
- SqlDataNode
- S3DataNode

Amazon RDS から Amazon Redshift へのテンプレート

以下の 2 つのテンプレートでは、変換スクリプトを使用して、Amazon RDS MySQL から Amazon Redshift にテーブルをコピーします。変換スクリプトでは、ソーステーブルのスキーマを使用して Amazon Redshift テーブルが作成されます。以下の点に注意してください。

- 分散キーを指定しなかった場合は、Amazon RDS テーブルの最初のプライマリキーが分散キーとして設定されます。
- Amazon Redshift へのコピーを実行するときに、Amazon RDS MySQL のテーブルにある列を省略することはできません。
- (オプション) テンプレートに含まれるパラメータの 1 つとして、Amazon RDS MySQL から Amazon Redshift への列データ型マッピングを指定することができます。指定した場合、スクリプトではこれを使用して Amazon Redshift テーブルが作成されます。

`Overwrite_Existing` Amazon Redshift 挿入モードが使用されている場合、次のようになります。

- 分散キーを指定しなかった場合は、Amazon RDS MySQL テーブルのプライマリキーが使用されます。
- テーブルに複合プライマリキーが存在し、分散キーが指定されていない場合は、最初の複合プライマリキーが分散キーとして使用されます。Amazon Redshift テーブルでは、最初の複合キーのみがプライマリキーとして設定されます。
- 分散キーが指定されず、Amazon RDS MySQL テーブルにプライマリキーが存在しない場合、コピー操作は失敗します。

Amazon Redshift の詳細については、次のトピックを参照してください。

- Amazon Redshift クラスター
- Amazon Redshift COPY
- 分散スタイルと DISTKEY の例
- ソートキー

次の表では、スクリプトでデータ型がどのように変換されるかを示しています。

MySQL と Amazon Redshift の間のデータ型変換

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
TINYINT, TINYINT (size)	SMALLINT	MySQL: -128 ~ 127。最大桁数を括弧内に指定できます。 Amazon Redshift: INT2。符号付き 2 バイト整数

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
TINYINT UNSIGNED、 TINYINT (size) UNSIGNED	SMALLINT	MySQL: 0 ~ 255 UNSIGNED。最大桁数を括弧内に指定できます。 Amazon Redshift: INT2。符号付き 2 バイト整数
SMALLINT, SMALLINT(size)	SMALLINT	MySQL: -32768 ~ 32767 normal。最大桁数を括弧内に指定できます。 Amazon Redshift: INT2。符号付き 2 バイト整数
SMALLINT UNSIGNED、 SMALLINT(size) UNSIGNED、	INTEGER	MySQL: 0 ~ 65535 UNSIGNED*。最大桁数を括弧内に指定できます。 Amazon Redshift: INT4。符号付き 4 バイト整数
MEDIUMINT、 MEDIUMINT(size)	INTEGER	MySQL: 388608 ~ 8388607。最大桁数を括弧内に指定できます。 Amazon Redshift: INT4。符号付き 4 バイト整数
MEDIUMINT UNSIGNED、 MEDIUMINT(size) UNSIGNED	INTEGER	MySQL: 0 ~ 16777215。最大桁数を括弧内に指定できます。 Amazon Redshift: INT4。符号付き 4 バイト整数

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
INT, INT(size)	INTEGER	MySQL: 147483648 ~ 2147483647 Amazon Redshift: INT4。符号付き 4 バイト整数
INT UNSIGNED、 INT(size) UNSIGNED	BIGINT	MySQL: 0 ~ 4294967295 Amazon Redshift: INT8。符号付き 8 バイト整数
BIGINT BIGINT(size)	BIGINT	Amazon Redshift: INT8。符号付き 8 バイト整数
BIGINT UNSIGNED BIGINT(size) UNSIGNED	VARCHAR(20*4)	MySQL: 0 ~ 18446744073709551615 Amazon Redshift: 同等のネイティブ型がないため、char 配列を使用します。
FLOAT FLOAT(size,d) FLOAT(size,d) UNSIGNED	REAL	最大桁数を size パラメータで指定できます。小数点以下の最大桁数は d パラメータで指定します。 Amazon Redshift: FLOAT4。
DOUBLE(size,d)	DOUBLE PRECISION	最大桁数を size パラメータで指定できます。小数点以下の最大桁数は d パラメータで指定します。 Amazon Redshift: FLOAT8。

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
DECIMAL(size,d)	DECIMAL(size,d)	固定小数点数を使用するための、文字列として格納される DOUBLE。最大桁数を size パラメータで指定できます。小数点以下の最大桁数は d パラメータで指定します。 Amazon Redshift: 同等のネイティブ型はありません。
CHAR(size)	VARCHAR(size*4)	固定長の文字列を格納します。これには、文字、数字、特殊文字を含めることができます。固定サイズは括弧内パラメータで指定します。255 文字まで格納できます。 右側にスペースが埋め込まれます。 Amazon Redshift: CHAR データ型ではマルチバイト文字がサポートされていないため、VARCHAR が使用されます。 1 文字あたりの最大バイト数は 4 (RFC3629 を参照) で、文字テーブルは U+10FFFF に制限されます。

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
VARCHAR(size)	VARCHAR(size*4)	255 文字まで格納できます。 VARCHAR では、以下の無効な UTF-8 コードポイントがサポートされていません: 0xD800 ~ 0xDFFF (バイトシーケンス: ED A0 80 ~ ED BF BF)、0xFDD0 ~ 0xFDEF、0xFFFE、0xFFFF (バイトシーケンス: EF B7 90 ~ EF B7 AF、EF BF BE、EF BF BF)
TINYTEXT	VARCHAR(255*4)	最大長 255 文字の文字列を格納します
TEXT	VARCHAR(max)	最大長 65,535 文字の文字列を格納します。
MEDIUMTEXT	VARCHAR(max)	0 ~ 16,777,215 文字
LONGTEXT	VARCHAR(max)	0 ~ 4,294,967,295 文字
BOOLEAN BOOL TINYINT(1)	BOOLEAN	MySQL: これらの型は TINYINT(1) と同義です。値ゼロは false と見なされます。ゼロ以外の値は true と見なされます。
BINARY[(M)]	varCHAR(255)	M は 0 ~ 255 バイト、FIXED
VARBINARY(M)	VARCHAR(max)	0 ~ 65,535 バイト
TINYBLOB	VARCHAR(255)	0 ~ 255 バイト
BLOB	VARCHAR(max)	0 ~ 65,535 バイト

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
MEDIUMBLOB	VARCHAR(max)	0 ~ 16,777,215 バイト
LONGBLOB	VARCHAR(max)	0 ~ 4,294,967,295 バイト
ENUM	VARCHAR(255*2)	リテラル ENUM 文字列の長さではなく、テーブル定義に含まれる ENUM 値の数に制限があります。
SET	VARCHAR(255*2)	ENUM と同じ。
DATE	DATE	(YYYY-MM-DD) 「1000-01-01」から「9999-12-31」
TIME	VARCHAR(10*4)	(hh:mm:ss) "-838:59:59" ~ "838:59:59"
DATETIME	TIMESTAMP	(YYYY-MM-DD hh:mm:ss) "1000-01-01 00:00:00" ~ "9999-12-31 23:59:59"
TIMESTAMP	TIMESTAMP	(YYYYMMDDhhmmss) 19700101000000 ~ 2037+
YEAR	VARCHAR(4*4)	(YYYY) 1900 ~ 2155

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
column SERIAL	ID 生成 / この列はコピーされるため、この属性は OLAP データウェアハウスでは必要ありません。 変換時に SERIAL キーワードは追加されません。	SERIAL は、実際には SEQUENCE というエンティティです。テーブルの残りの部分からは独立して存在しています。 column GENERATED BY DEFAULT これは下記と同等です: CREATE SEQUENCE name; CREATE TABLE table (column INTEGER NOT NULL DEFAULT nextval(name));
column BIGINT UNSIGNED NOT NULL AUTO_INCREMENT UNIQUE	ID 生成 / この列はコピーされるため、この属性は OLAP データウェアハウスでは必要ありません。 このため、変換時に SERIAL キーワードは追加されません。	SERIAL は、実際には SEQUENCE というエンティティです。テーブルの残りの部分からは独立して存在しています。 column GENERATED BY DEFAULT これは下記と同等です: CREATE SEQUENCE name; CREATE TABLE table (column INTEGER NOT NULL DEFAULT nextval(name));

MySQL のデータ型	Amazon Redshift のデータ型	Notes (メモ)
ZEROFILL	変換時に ZEROFILL キーワードは追加されません。	INT UNSIGNED ZEROFILL NOT NULL ZEROFILL では、フィールドに表示される値に、列定義で指定された表示幅までゼロが埋め込まれます。値の桁数が表示幅を超えても、切り捨ては行われません。ZEROFILL の使用は、UNSIGNED を暗黙に示しています。

[Full Copy of Amazon RDS MySQL Table to Amazon Redshift] (Amazon RedshiftへのAmazon RDS MySQL テーブルの完全コピー)

[Full copy of Amazon RDS MySQL table to Amazon Redshift] (Amazon Redshift への Amazon RDS MySQL テーブルの完全コピー) テンプレートでは、データを Amazon S3 フォルダにステージングすることによって、Amazon RDS MySQL テーブル全体を Amazon Redshift テーブルにコピーします。Amazon S3 ステージングフォルダが、Amazon Redshift クラスターと同じリージョンに存在する必要があります。ソースの Amazon RDS MySQL テーブルと同じスキーマを使って、Amazon Redshift テーブルが (まだ存在しない場合は) 作成されます。Amazon Redshift テーブルの作成時に適用する、Amazon RDS MySQL から Amazon Redshift への列データ型オーバーライドがあれば、指定します。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- RedshiftCopyActivity
- S3DataNode
- SqlDataNode
- RedshiftDataNode
- RedshiftDatabase

[Incremental Copy of an Amazon RDS MySQL Table to Amazon Redshift] (Amazon Redshift への Amazon RDS MySQL テーブルの増分コピー)

[Incremental copy of Amazon RDS MySQL table to Amazon Redshift] (Amazon Redshift への Amazon RDS MySQL テーブルの増分コピー) テンプレートでは、データを Amazon S3 フォルダにステージングすることによって、Amazon RDS MySQL テーブルのデータを Amazon Redshift テーブルにコピーします。

Amazon S3 ステージングフォルダが、Amazon Redshift クラスターと同じリージョンに存在する必要があります。

AWS Data Pipeline では、変換スクリプトによって、ソースの Amazon RDS MySQL テーブルと同じスキーマを使って、Amazon Redshift テーブルが (まだ存在しない場合は) 作成されます。Amazon Redshift テーブルの作成時に適用する、Amazon RDS MySQL から Amazon Redshift への列データ型オーバーライドがあれば、指定する必要があります。

このテンプレートでは、スケジュールされた開始時刻から始まるスケジュールされた間隔で、Amazon RDS MySQL テーブルに行われた変更をコピーします。Amazon RDS MySQL テーブルへの物理的な削除はコピーされません。最終更新日時の値を格納する列名を指定する必要があります。

デフォルトのテンプレートを使用して、Amazon RDS の増分コピーのパイプラインを作成する場合は、デフォルト名 `RDSToS3CopyActivity` というアクティビティが作成されます。名前は変更できます。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- RedshiftCopyActivity
- S3DataNode
- SqlDataNode
- RedshiftDataNode
- RedshiftDatabase

Amazon S3 から Amazon Redshift へのデータのロード

[Load data from S3 into Redshift] (S3 から Redshift へのデータのロード) テンプレートでは、Amazon S3 フォルダから Amazon Redshift テーブルにデータをコピーします。既存のテーブルにデータをロードすることも、テーブルを作成する SQL クエリを指定することもできます。

データは、Amazon Redshift COPY オプションに基づいてコピーされます。Amazon Redshift テーブルのスキーマは、Amazon S3 のデータと同じである必要があります。COPY オプションについては、Amazon Redshift データベース開発者ガイドの「COPY」を参照してください。

このテンプレートは以下のパイプラインオブジェクトを使用します。

- CopyActivity
- RedshiftCopyActivity
- S3DataNode
- RedshiftDataNode
- RedshiftDatabase
- Ec2Resource

パラメータ化されたテンプレートを使用したパイプラインの作成

パラメータ化されたテンプレートを使用して、パイプライン定義をカスタマイズできます。これにより、共通のパイプライン定義を作成しておき、このパイプライン定義を新しいパイプラインに追加するときに、異なるパラメータを指定することができます。

目次

- パイプライン定義への myVariables の追加
- パラメータオブジェクトの定義
- パラメータ値の定義
- パイプライン定義の送信

パイプライン定義への myVariables の追加

パイプライン定義ファイルを作成する場合、#{my##} という構文を使用して変数を指定します。変数の前にプレフィックス my を付ける必要があります。たとえば、次のパイプライン定

義ファイル (pipeline-definition.json) には、*myShellCmd*、*myS3InputLoc*、および *myS3OutputLoc* という変数が含まれています。

> **Note**
>
> パイプライン定義には、50 パラメーターという上限があります。

```
{
  "objects": [
    {
      "id": "ShellCommandActivityObj",
      "input": {
        "ref": "S3InputLocation"
      },
      "name": "ShellCommandActivityObj",
      "runsOn": {
        "ref": "EC2ResourceObj"
      },
      "command": "#{myShellCmd}",
      "output": {
        "ref": "S3OutputLocation"
      },
      "type": "ShellCommandActivity",
      "stage": "true"
    },
    {
      "id": "Default",
      "scheduleType": "CRON",
      "failureAndRerunMode": "CASCADE",
      "schedule": {
        "ref": "Schedule_15mins"
      },
      "name": "Default",
      "role": "DataPipelineDefaultRole",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "id": "S3InputLocation",
      "name": "S3InputLocation",
      "directoryPath": "#{myS3InputLoc}",
      "type": "S3DataNode"
```

```
    },
    {
      "id": "S3OutputLocation",
      "name": "S3OutputLocation",
      "directoryPath": "#{myS3OutputLoc}/#{format(@scheduledStartTime, 'YYYY-MM-dd-HH-
mm-ss')}",
      "type": "S3DataNode"
    },
    {
      "id": "Schedule_15mins",
      "occurrences": "4",
      "name": "Every 15 minutes",
      "startAt": "FIRST_ACTIVATION_DATE_TIME",
      "type": "Schedule",
      "period": "15 Minutes"
    },
    {
      "terminateAfter": "20 Minutes",
      "id": "EC2ResourceObj",
      "name": "EC2ResourceObj",
   "instanceType":"t1.micro",
      "type": "Ec2Resource"
    }
  ]
}
```

パラメータオブジェクトの定義

パイプライン定義に含まれる変数を定義したパラメータオブジェクトのファイルは、個別に作成することができます。たとえば、次の JSON ファイル (parameters.json) には、上のパイプライン定義例の *myShellCmd*、*myS3InputLoc*、および *myS3OutputLoc* という変数に対するパラメータオブジェクトが含まれています。

```
{
  "parameters": [
    {
      "id": "myShellCmd",
      "description": "Shell command to run",
      "type": "String",
      "default": "grep -rc \"GET\" ${INPUT1_STAGING_DIR}/* > ${OUTPUT1_STAGING_DIR}/
output.txt"
    },
    {
```

```
      "id": "myS3InputLoc",
      "description": "S3 input location",
      "type": "AWS::S3::ObjectKey",
      "default": "s3://us-east-1.elasticmapreduce.samples/pig-apache-logs/data"
    },
    {
      "id": "myS3OutputLoc",
      "description": "S3 output location",
      "type": "AWS::S3::ObjectKey"
    }
  ]
}
```

> **Note**
>
> 個別のファイルを使用する代わりに、これらのオブジェクトをパイプライン定義ファイルに直接追加することもできます。

次の表では、パラメータオブジェクトの属性を説明しています。

パラメータ属性

属性	タイプ	説明
id	文字列	パラメータの一意な識別子。入力中または表示中に値を隠すには、プレフィックスとしてアスタリスク (*) を追加します 例: *myVariable —。この場合、AWS Data Pipeline によって保存される前に値が暗号化される点にも注意してください。
description	文字列	パラメータの説明。
type	String、Integer、Double、または AWS::S3::ObjectKey	パラメータの型。入力値の許容範囲と検証規則を定義しま

属性	タイプ	説明
		す。デフォルト値は String です。
オプション	ブール値	パラメータがオプションか必須かを示します。デフォルトは `false` です。
allowedValues	Strings のリスト	パラメータに許可されている値をすべて列挙します。
デフォルト	文字列	パラメータのデフォルト値。パラメータ値を使用して、このパラメータの値を指定すると、デフォルト値を上書きします。
isArray	ブール値	パラメータが配列かどうかを示します。

パラメータ値の定義

個別のファイルを作成し、パラメータ値を使用して変数を定義することができます。たとえば、次に示す JSON ファイル (`file://values.json`) には、上のパイプライン定義例で使用した *`myS3OutputLoc`* 変数の値が含まれています。

```
{
  "values":
    {
      "myS3OutputLoc": "myOutputLocation"
    }
}
```

パイプライン定義の送信

パイプライン定義を送信する際には、パラメータ、パラメータオブジェクト、パラメータ値を指定できます。例えば、以下のように put-pipeline-definition AWS CLI コマンドを使用できます。

```
$ aws datapipeline put-pipeline-definition --pipeline-id id --pipeline-definition
 file://pipeline-definition.json \
--parameter-objects file://parameters.json --parameter-values-uri file://values.json
```

> **Note**
>
> パイプライン定義には、50 パラメーターという上限があります。parameter-values-uri 用のファイルサイズには、15 KB という上限があります。

パイプラインの表示

コマンドラインインターフェイス (CLI) を使用して、パイプラインを表示できます。

AWS CLI を使用してパイプラインを表示するには

- パイプラインを一覧表示するには、次の list-pipelines コマンドを使用します。

  ```
  aws datapipeline list-pipelines
  ```

パイプラインのステータスコードの解釈

AWS Data Pipeline コンソールと CLI に表示されるステータスレベルは、パイプラインとそのコンポーネントの状態を示します。パイプラインのステータスは、単にパイプラインの概要を表します。詳細を確認するには、個別のパイプラインコンポーネントのステータスを表示します。

パイプラインが準備を完了している (パイプライン定義が検証に合格した) か、現在、作業の実行中であるか、作業を完了した場合、パイプラインのステータスは SCHEDULED です。パイプラインがアクティブ化されていない、または作業を実行できない (たとえば、パイプライン定義が検証に失敗した) 場合、パイプラインのステータスは PENDING です。

ステータスが PENDING、INACTIVE、または FINISHED の場合、パイプラインは非アクティブと見なされます。非アクティブなパイプラインには料金が発生します (詳細については、「料金表」を参照してください)。

ステータスコード

ACTIVATING

EC2 インスタンスなど、コンポーネントまたはリソースが開始されています。

CANCELED

コンポーネントは、実行前にユーザーまたは AWS Data Pipeline によってキャンセルされました。これは、このコンポーネントが依存する別のコンポーネントまたはリソースで障害が発生した場合に自動的に発生する可能性があります。

CASCADE_FAILED

コンポーネントまたはリソースが、依存関係の 1 つによるカスケード障害の結果としてキャンセルされましたが、そのコンポーネントはおそらく障害の元の原因ではありませんでした。

DEACTIVATING

パイプラインが非アクティブ化されています。

FAILED

コンポーネントまたはリソースでエラーが発生し、動作を停止しました。コンポーネントまたはリソースに障害が発生すると、キャンセルや障害がそれに依存する他のコンポーネントにカスケードする可能性があります。

FINISHED

コンポーネントは割り当てられた作業を完了しました。

INACTIVE

パイプラインが非アクティブ化されました。

PAUSED

コンポーネントは一時停止され、現在作業を実行していません。

PENDING

パイプラインを初めてアクティブ化する準備が整いました。

RUNNING

リソースは実行中であり、作業を受け取る準備が整いました。

SCHEDULED

リソースの実行がスケジュールされています。

SHUTTING_DOWN

リソースは、作業が正常に完了した後にシャットダウンしています。

SKIPPED

現在のスケジュールより後のタイムスタンプを使用して、パイプラインがアクティブ化された後、コンポーネントは実行間隔をスキップしました。

TIMEDOUT

リソースが terminateAfter しきい値を超過したため、AWS Data Pipeline によって停止されました。リソースがこのステータスに達すると、AWS Data Pipeline によってそのリソースの actionOnResourceFailure、retryDelay、および retryTimeout の値が無視されます。このステータスは、リソースにのみ適用されます。

VALIDATING

パイプラインの定義が AWS Data Pipeline によって検証されています。

WAITING_FOR_RUNNER

コンポーネントは、ワーカークライアントが作業項目を取得するのを待っています。コンポーネントとワーカークライアントの関係は、そのコンポーネントによって定義されている runsOn または workerGroup フィールドによって制御されます。

WAITING_ON_DEPENDENCIES

コンポーネントは、作業を実行する前に、デフォルトおよびユーザー設定の前提条件が満たされていることを確認しています。

パイプラインとコンポーネントのヘルス状態の解釈

各パイプラインと、そのパイプライン内のコンポーネントから返されるヘルスステータスには、HEALTHY、ERROR、"-"、No Completed Executions、No Health Information Available があります。パイプラインのコンポーネントが最初の実行を完了した後、またはコンポーネントの前提条件が失敗した場合、パイプラインのヘルス状態は 1 つのみです。各コンポーネントのヘルスステータスは、パイプラインのヘルスステータスに集約され、パイプラインの実行詳細を表示すると、エラー状態が最初に表示されます。

パイプラインのヘルス状態

HEALTHY

すべてのコンポーネントの集約ヘルスステータスが HEALTHY です。これは、少なくとも 1 つのコンポーネントが正常に完了したことを意味します。HEALTHY ステータスをクリックすると、直近に正常完了したパイプラインコンポーネントインスタンスが [Execution Details] ページに表示されます。

ERROR

パイプラインの少なくとも 1 つのコンポーネントのヘルスステータスが ERROR です。ERROR ステータスをクリックすると、直近に失敗したパイプラインコンポーネントインスタンスが [Execution Details] ページに表示されます。

No Completed Executions-または-No Health Information Available

このパイプラインのヘルスステータスは、報告されていません。

> **Note**
>
> 各コンポーネントのヘルスステータスは、ほぼ即時に更新されますが、パイプラインのヘルスステータスが更新されるまでに最大で 5 分かかることがあります。

コンポーネントのヘルスステータス

HEALTHY

コンポーネント (Activity または DataNode) のヘルスステータスが HEALTHY になるのは、FINISHED または MARK_FINISHED のステータスがマークされた状態で実行が正常完了した場合です。コンポーネント名または HEALTHY ステータスをクリックすると、直近に正常完了したパイプラインコンポーネントインスタンスが [Execution Details] ページに表示されます。

ERROR

コンポーネントレベルでエラーが発生したか、いずれかの前提条件が失敗しました。FAILED、TIMEOUT、または CANCELED のステータスによってこのエラーがトリガーされます。コンポーネント名または ERROR ステータスをクリックすると、直近に失敗したパイプラインコンポーネントインスタンスが [Execution Details] ページに表示されます。

`No Completed Executions`、または `No Health Information Available`

このコンポーネントのヘルスステータスは、報告されていません。

パイプライン定義の表示

コマンドラインインターフェイス (CLI) を使用して、パイプライン定義を表示します。CLI では JSON 形式のパイプライン定義ファイルが表示されます。パイプライン定義ファイルの構文と使用方法については、「パイプライン定義ファイルの構文」を参照してください。

CLI を使用する場合は、変更を送信する前に、パイプライン定義を取得することをお勧めします。これは、パイプライン定義を最後に使用した後に、別のユーザーまたはプロセスによってそのパイプライン定義が変更された可能性があるためです。現在の定義のコピーをダウンロードし、変更のベースとして使用することで、確実に最新のパイプライン定義を使用できます。また、更新が正常に行われたことを確認できるように、変更後にパイプライン定義を再取得することをお勧めします。

CLI を使用する場合は、パイプラインの 2 つの異なるバージョンを取得できます。`active` バージョンは、現在実行中のパイプラインです。`latest` バージョンは、実行中のパイプラインを編集した際に作成されたコピーです。編集したパイプラインをアップロードすると、それが `active` バージョンになり、それまでの `active` バージョンは使用できなくなります。

AWS CLI を使用してパイプライン定義を取得するには

完全なパイプライン定義を取得するには、get-pipeline-definition コマンドを使用します。パイプライン定義は標準出力 (stdout) に表示されます。

次の例では、指定したパイプラインのパイプライン定義を取得します。

```
aws datapipeline get-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE
```

パイプラインの特定のバージョンを取得するには、`--version` オプションを使用します。次の例では、指定したパイプラインの `active` バージョンを取得します。

```
aws datapipeline get-pipeline-definition --version active --id df-00627471SOVYZEXAMPLE
```

パイプラインインスタンスの詳細の表示

パイプラインの進捗状況を監視することができます。インスタンスステータスの詳細については、「パイプラインのステータスの詳細を解釈する」を参照してください。インスタンスがパイプライン

の実行に失敗した場合または実行を完了できなかった場合のトラブルシューティングに関する詳細については、「一般的な問題を解決する」を参照してください。

AWS CLI を使用してパイプラインの進捗状況を監視するには

パイプラインの実行時間の履歴など、パイプラインインスタンスの詳細を取得するには、list-runs コマンドを使用します。このコマンドを使用すると、パイプラインの現在のステータスまたは起動された日付範囲に基づいて、返された実行のリストをフィルタリングすることができます。パイプラインの作成時期とスケジュールによっては実行履歴が大きくなる場合があるため、結果のフィルタリングは有用です。

次の例では、すべての実行の情報を取得します。

```
aws datapipeline list-runs --pipeline-id df-00627471SOVYZEXAMPLE
```

次の例では、完了したすべての実行の情報を取得します。

```
aws datapipeline list-runs --pipeline-id df-00627471SOVYZEXAMPLE --status finished
```

次の例では、指定された時間枠に起動されたすべての実行について、情報を取得します。

```
aws datapipeline list-runs --pipeline-id df-00627471SOVYZEXAMPLE --start-interval
 "2013-09-02","2013-09-11"
```

パイプラインのログの表示

Amazon S3 の場所をコンソールで指定するか、SDK/CLI を使用してデフォルトオブジェクト内の pipelineLogUri で指定することによって、パイプライン作成時のログ記録がパイプラインレベルでサポートされます。その URI 内の各パイプラインのディレクトリ構造は次のようになります。

```
pipelineId
    -componentName
        -instanceId
            -attemptId
```

df-00123456ABC7DEF8HIJK というパイプラインのディレクトリ構造は次のようになります。

```
df-00123456ABC7DEF8HIJK
```

```
    -ActivityId_fXNzc
        -@ActivityId_fXNzc_2014-05-01T00:00:00
            -@ActivityId_fXNzc_2014-05-01T00:00:00_Attempt=1
```

ShellCommandActivity では、アクティビティに関連付けられた stderr および stdout のログは、各試行のディレクトリに保存されます。

EmrCluster などのリソースで emrLogUri が設定されている場合、その値が優先されます。それ以外の場合、リソース (そのリソースの TaskRunner ログを含む) には、上に示したパイプラインのログ記録構造が使用されます。

特定のパイプライン実行のログを表示するには

1. query-objects を呼び出して ObjectId を取得し、正確なオブジェクト ID を取得します。例:

   ```
   aws datapipeline query-objects --pipeline-id <pipeline-id> --sphere ATTEMPT --region
    ap-northeast-1
   ```

   query-objects はページ分割 CLI であり、指定された pipeline-id に対してさらに実行がある場合はページ分割トークンを返すことがあります。このトークンを使うと、期待するオブジェクトが見つかるまで、すべての試行を繰り返すことができます。例えば、返される ObjectId は、@TableBackupActivity_2023-05-020T18:05:18_Attempt=1 のようになります。
2. ObjectId を使用して、以下を使用してログの場所を取得します。

   ```
   aws datapipeline describe-objects –pipeline-id <pipeline-id> --object-ids <object-id>
    --query "pipelineObjects[].fields[?key=='@logLocation'].stringValue"
   ```

失敗したアクティビティのエラーメッセージ

エラーメッセージを取得するには、まず query-objects を使用して ObjectId を取得します。

失敗した ObjectId を取得したら、describe-objects CLI を使用して実際のエラーメッセージを取得します。

```
aws datapipeline describe-objects --region ap-northeast-1 --pipeline-id
 <pipeline-id> --object-ids <object-id> --query "pipelineObjects[].fields[?
key=='errorMessage'].stringValue"
```

オブジェクトをキャンセル、再実行、または終了としてマークする

set-status CLI を使用して、実行中のオブジェクトをキャンセルするか、失敗したオブジェクトを再実行するか、実行中のオブジェクトに Finished のマークを付けます。

まず、query-objects CLI を使用してオブジェクト ID を取得します。例:

```
aws datapipeline query-objects --pipeline-id <pipeline-id> --sphere INSTANCE --region
 ap-northeast-1
```

set-status CLI を使用して、目的のオブジェクトのステータスを変更します。例:

```
aws datapipeline set-status –pipeline-id <pipeline-id> --region ap-northeast-1 --status
 TRY_CANCEL --object-ids <object-id>
```

パイプラインの編集

いずれかのパイプラインの一部を変更するには、対応するパイプライン定義を更新することができます。実行中のパイプラインを変更した後は、変更を有効にするためにパイプラインを再びアクティブ化する必要があります。また、パイプラインの 1 つ以上のコンポーネントを再実行できます。

目次

- 制限事項
- AWS CLI を使用したパイプラインの編集

制限事項

パイプラインが [PENDING] 状態であり、まだアクティブ化されていなければ、変更をパイプラインに対して加えることができません。パイプラインをアクティブ化した後は、パイプラインの編集には以下の制限が適用されます。変更は、保存して、パイプラインを再度アクティブ化した後、パイプラインオブジェクトの実行に適用されます。

- オブジェクトを削除することはできません
- 既存のオブジェクトのスケジュール期間は変更できません
- 既存のオブジェクトの参照フィールドの追加、削除、変更はできません
- 新しいオブジェクトの出力フィールドで既存のオブジェクトを参照できません
- オブジェクトの予定された開始日を変更できません (代わりに、特定の日時でパイプラインをアクティブ化してください)

AWS CLI を使用したパイプラインの編集

コマンドラインツールを使用して、パイプラインを編集できます。

まず、get-pipeline-definition コマンドを使用して、現在のパイプライン定義のコピーをダウンロードします。これにより、確実に最新のパイプライン定義を編集できます。次の例では、標準出力（stdout）にパイプライン定義を表示します。

```
aws datapipeline get-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE
```

パイプライン定義をファイルに保存し、必要に応じて編集します。put-pipeline-definition コマンドを使用してパイプライン定義を更新します。次の例では、更新されたパイプライン定義ファイルをアップロードします。

```
aws datapipeline put-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE --
pipeline-definition file://MyEmrPipelineDefinition.json
```

get-pipeline-definition コマンドを使用してパイプライン定義を再度取得することで、更新が成功したことを確認できます。パイプラインをアクティブ化するには、次の activate-pipeline コマンドを使用します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

必要に応じて、次のように --start-timestamp オプションを使用して、特定の日時からパイプラインをアクティブ化できます。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE --start-
timestamp YYYY-MM-DDTHH:MM:SSZ
```

1 つ以上のパイプラインコンポーネントを再実行するには、set-status コマンドを使用します。

パイプラインのクローン作成

クローン作成では、パイプラインのコピーが作成され、新しいパイプラインの名前を指定することができます。パイプラインの状態を問わず、エラーがあるパイプラインでもクローンを作成できますが、新しいパイプラインは、手動でアクティブ化しない限り PENDING 状態のままです。クローン操作では、新しいパイプラインはアクティブバージョンではなく、元のパイプライン定義の最新バージョンになります。クローン操作では、元のパイプラインから新しいパイプラインに完全なスケジュールはコピーされず、期間設定のみがコピーされます。

AWS CLI を使用してパイプラインのクローンを作成するには

1. 新しい名前と一意の ID を使用して新しいパイプラインを作成します。返されたパイプライン ID を書き留めておきます。
2. get-pipeline-definition CLI を使用して、クローンを作成する既存のパイプラインのパイプライン定義を取得し、それを一時ファイルに書き込みます。ファイルの絶対パスを書き留めておきます。
3. put-pipeline-definition CLI を使用して、パイプライン定義を既存のパイプラインから新しいパイプラインにコピーします。
4. get-pipeline-definition CLI を使用して新しいパイプラインの定義を取得し、パイプライン定義を確認します。

```
# Create Pipeline (returns <new-pipeline-id>)
aws datapipeline create-pipeline --name my-cloned-pipeline --unique-id my-cloned-
pipeline --region ap-northeast-1

#Get pipeline definition of existing pipeline
aws datapipeline get-pipeline-definition --pipeline-id <existing-pipeline-id> --
region ap-northeast-1 > existing_pipeline_definition.json

# Put pipeline definition to new pipeline
aws datapipeline put-pipeline-definition --pipeline-id <new-
pipeline-id> --region ap-northeast-1 --pipeline-definition file://
<absolute_path_to_existing_pipeline_definition.json>

# get pipeline definition of new pipeline
aws datapipeline get-pipeline-definition --pipeline-id <new-pipeline-id> --region
 ap-northeast-1
```

パイプラインのタグ付け

タグは、キーとオプションの値で構成される、大文字小文字を区別するキーと値のペアです。キーと値はどちらも、ユーザー定義です。各パイプラインには、最大で 10 個のタグを適用できます。タグキーは、各パイプラインで一意である必要があります。すでにパイプラインに関連付けられているキーを持つタグを追加すると、そのキーの値が更新されます。

タグをパイプラインに適用すると、基盤となるリソース (Amazon EMR クラスターや Amazon EC2 インスタンスなど) にもそのタグが適用されます。ただし、これらのタグは、さかのぼって

FINISHED 状態やその他の終了状態のリソースには適用されません。これらのリソースにタグを適用する必要がある場合は、CLI を使用できます。

タグが不要になったら、パイプラインからタグを削除できます。

AWS CLI を使用したパイプラインのタグ付け

新しいパイプラインにタグを追加するには、--tags オプションを create-pipeline コマンドに追加します。たとえば、次のオプションではパイプラインを作成して 2 つのタグを追加します。environment タグ (production という値) とowner タグ (salesという値) です。

```
--tags key=environment,value=production key=owner,value=sales
```

既存のパイプラインにタグを追加するには、次のように add-tags コマンドを使用します。

```
aws datapipeline add-tags --pipeline-id df-00627471SOVYZEXAMPLE --tags
 key=environment,value=production key=owner,value=sales
```

既存のパイプラインからタグを削除するには、次のように remove-tags コマンドを使用します。

```
aws datapipeline remove-tags --pipeline-id df-00627471SOVYZEXAMPLE --tag-keys
 environment owner
```

パイプラインの非アクティブ化

実行中のパイプラインを非アクティブ化すると、パイプライン実行が一時停止します。パイプライン実行を再開するには、パイプラインをアクティブ化します。これにより、変更を加えることができます。たとえば、メンテナンスの実施が予定されているデータベースにデータを書き込む場合、パイプラインを非アクティブ化し、メンテナンスが完了するのを待って、パイプラインをアクティブ化することができます。

パイプラインを非アクティブ化したとき、実行中のアクティビティをどうするか指定できます。デフォルトでは、これらのアクティビティはすぐにキャンセルされます。または、パイプラインを非アクティブ化する前に、アクティビティが終了するまで AWS Data Pipeline を待機させることができます。

非アクティブ化したパイプラインをアクティブ化するとき、いつ再開するのかを指定できます。デフォルトでは、AWS CLI または API を使用して、最後の実行が完了してからパイプラインが再開します。または、特定の日時を指定して、パイプラインを再開させることができます。

AWS CLI を使用したパイプラインの非アクティブ化

パイプラインを非アクティブ化するには、次の deactivate-pipeline コマンドを使用します。

```
aws datapipeline deactivate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

すべての実行中のアクティビティが終了した後にのみ、パイプラインを非アクティブ化するには、次のように --no-cancel-active オプションを追加してください。

```
aws datapipeline deactivate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE --no-cancel-active
```

準備ができたら、次の activate-pipeline コマンドを使用して、停止したパイプライン実行を再開できます。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

特定の日時からパイプラインを開始するには、次のように --start-timestamp オプションを追加します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE --start-timestamp YYYY-MM-DDTHH:MM:SSZ
```

パイプラインの削除

アプリケーションのテスト中に作成したパイプラインなど、パイプラインが必要なくなったときは、アクティブな使用から削除します。パイプラインを削除すると、削除状態になります。パイプラインが削除状態になると、そのパイプライン定義と実行履歴は失われます。したがって、パイプラインの記述も含め、パイプラインに対する操作は実行できなくなります。

> **Important**
>
> パイプラインを削除した後にリストアすることはできないため、削除する前に、今後はそのパイプラインが必要ないことを確認してください。

AWS CLI を使用してパイプラインを削除するには

パイプラインを削除するには、delete-pipeline コマンドを使用します。次のコマンドでは、指定したパイプラインを削除します。

```
aws datapipeline delete-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

パイプラインのアクティビティによるデータとテーブルのステージング

AWS Data Pipeline では、パイプラインの入出力データをステージングすることで、ShellCommandActivity や HiveActivity などの特定のアクティビティの使用が容易になります。

データのステージングを使用して、入力データノードからアクティビティを実行するリソースにデータをコピーできます。同様に、リソースから出力データノードにデータをコピーすることもできます。

Amazon EMR または Amazon EC2 のリソースでステージングされたデータは、アクティビティのシェルコマンドまたは Hive スクリプトで特別な変数を用いることで使用可能です。

テーブルのステージングはデータのステージングと似ていますが、具体的には、ステージングされたデータがデータベーステーブルの形式である点が異なります。

AWS Data Pipeline では、以下のシナリオのステージングがサポートされます。

- ShellCommandActivity によるデータのステージング
- Hive およびステージングをサポートするデータノードによるテーブルのステージング
- Hive およびステージングをサポートしないデータノードによるテーブルのステージング

> **Note**
>
> ステージングは、ShellCommandActivity などのアクティビティで stage フィールドが true に設定されている場合にのみ機能します。詳細については、「ShellCommandActivity」を参照してください。

また、データノードとアクティビティの関係には、次の 4 種類があります。

リソースでローカルにデータをステージング

入力データはリソースのローカルファイルシステムに自動的にコピーされます。出力データはリソースのローカルファイルシステムから出力データノードに自動的にコピーされます。たとえば、ShellCommandActivity の入出力で staging = true に設定すると、入力データは INPUTx_STAGING_DIR、出力データは OUTPUTx_STAGING_DIR として使用可能になります (x は入力または出力の番号です)。

アクティビティの入出力定義のステージング

入力データ形式 (列名、テーブル名) がアクティビティのリソースに自動的にコピーされます。たとえば、HiveActivity で staging = true に設定する場合です。Hive テーブルからテーブル定義をステージングするために、入力 S3DataNode で指定されたデータ形式が使用されます。

ステージングが有効ではない

アクティビティで入出力オブジェクトとそのフィールドは使用できますが、データそのものは使用できません。たとえば、デフォルトが EmrActivity であるか、それ以外のアクティビティで staging = false と設定する場合です。この設定では、AWS Data Pipeline の式構文を使用してデータフィールドを参照するためにアクティビティでデータフィールドを使用できますが、これは依存関係が満たされている場合にのみ参照されます。これは、依存関係のチェックとしてのみ機能します。アクティビティを実行するリソースに対する入力からデータをコピーする責任は、アクティビティのコードにあります。

オブジェクト間の依存関係

2 個のオブジェクト間に依存関係があり、結果としてステージングが有効でない場合と同じような状況になります。このため、データノードまたはアクティビティは、別のアクティビティを実行するための前提条件として機能します。

ShellCommandActivity によるデータのステージング

ShellCommandActivity でデータ入出力として S3DataNode オブジェクトを使用するシナリオを検討します。AWS Data Pipeline はデータノードを自動的にステージングし、以下の例に示すように、それらのデータノードがローカルファイルフォルダであるかのように環境変数 ${INPUT1_STAGING_DIR} および ${OUTPUT1_STAGING_DIR} を使用してシェルコマンドにアクセスできるようにします。INPUT1_STAGING_DIR という名前の変数の数値部分と、アクティビティが参照するデータノード数に応じた OUTPUT1_STAGING_DIR 増分。

> **Note**
>
> このシナリオは、記述されているように、データ入出力が S3DataNode オブジェクトの場合にのみ機能します。さらに、出力データのステージングは、出力 S3DataNode で directoryPath が設定されている場合にのみ許可されます。

```
{
  "id": "AggregateFiles",
  "type": "ShellCommandActivity",
  "stage": "true",
  "command": "cat ${INPUT1_STAGING_DIR}/part* > ${OUTPUT1_STAGING_DIR}/aggregated.csv",
  "input": {
    "ref": "MyInputData"
  },
  "output": {
    "ref": "MyOutputData"
  }
},
{
  "id": "MyInputData",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "filePath": "s3://my_bucket/source/#{format(@scheduledStartTime,'YYYY-MM-
dd_HHmmss')}/items"
  }
},
{
  "id": "MyOutputData",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "directoryPath": "s3://my_bucket/destination/#{format(@scheduledStartTime,'YYYY-MM-
dd_HHmmss')}"
  }
},
...
```

Hive およびステージングをサポートするデータノードによるテーブルのステージング

HiveActivity でデータ入出力として S3DataNode オブジェクトを使用するシナリオを検討します。AWS Data Pipeline はデータノードを自動的にステージングし、以下の例の HiveActivity に示すように、それらのデータノードが変数 ${input1} と ${output1} を使用する Hive テーブルであるかのように Hive スクリプトにアクセスできるようにします。input という名前の変数の数値部分と、アクティビティが参照するデータノード数に応じた output 増分。

> **Note**
>
> このシナリオは、記述されているように、データ入出力が S3DataNode オブジェクトまたは MySqlDataNode オブジェクトの場合にのみ機能します。テーブルのステージングは DynamoDBDataNode ではサポートされません。

```
{
  "id": "MyHiveActivity",
  "type": "HiveActivity",
  "schedule": {
    "ref": "MySchedule"
  },
  "runsOn": {
    "ref": "MyEmrResource"
  },
  "input": {
    "ref": "MyInputData"
  },
  "output": {
    "ref": "MyOutputData"
  },
  "hiveScript": "INSERT OVERWRITE TABLE ${output1} select * from ${input1};"
},
{
  "id": "MyInputData",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "directoryPath": "s3://test-hive/input"
  }
```

```
},
{
  "id": "MyOutputData",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "directoryPath": "s3://test-hive/output"
  }
},
...
```

Hive およびステージングをサポートしないデータノードによるテーブルのステージング

HiveActivity で入力データとして DynamoDBDataNode、出力として S3DataNode を使用するシナリオを検討します。DynamoDBDataNode ではデータのステージングは使用できないため、まず、DynamoDB テーブルを参照するために変数名 #{input.tableName} を使用して、Hive スクリプト内でテーブルを手動で作成する必要があります。この DynamoDB テーブルが出力となる場合も同様の命名法が適用されますが、変数が #{output.tableName} である点が異なります。この例の出力 S3DataNode オブジェクトではステージングを使用できるため、出力データノードを ${output1} として参照できます。

> **Note**
>
> AWS Data Pipeline では tableName または directoryPath にアクセスするために式を使用するため、この例では、テーブル名変数に # (ハッシュ) 文字プレフィックスが付きます。AWS Data Pipeline で式の評価がどのように機能するかの詳細については、「式の評価」を参照してください。

```
{
  "id": "MyHiveActivity",
  "type": "HiveActivity",
  "schedule": {
    "ref": "MySchedule"
  },
  "runsOn": {
    "ref": "MyEmrResource"
```

```
  },
  "input": {
    "ref": "MyDynamoData"
  },
  "output": {
    "ref": "MyS3Data"
  },
  "hiveScript": "-- Map DynamoDB Table
SET dynamodb.endpoint=dynamodb.us-east-1.amazonaws.com;
SET dynamodb.throughput.read.percent = 0.5;
CREATE EXTERNAL TABLE dynamodb_table (item map<string,string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "#{input.tableName}");
INSERT OVERWRITE TABLE ${output1} SELECT * FROM dynamodb_table;"
},
{
  "id": "MyDynamoData",
  "type": "DynamoDBDataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "tableName": "MyDDBTable"
},
{
  "id": "MyS3Data",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "directoryPath": "s3://test-hive/output"
  }
},
...
```

複数のリージョンにあるリソースとパイプラインの使用

デフォルトでは、Ec2Resource および EmrCluster のリソースは AWS Data Pipeline と同じリージョンで実行されますが、AWS Data Pipeline では、あるリージョンで実行するリソースで別のリージョンからの入力データを統合するなど、複数のリージョンにまたがるデータフローのオーケストレーションをサポートしています。指定したリージョンでリソースを実行できることで、依存するデータセットとリソースを同じ場所に配置する柔軟性がもたらされます。また、レイテンシーを減らすとともにクロスリージョンのデータ転送料金を回避することで、パフォーマンスを最大化するこ

とができます。region および Ec2Resource で EmrCluster フィールドを使用することで、AWS Data Pipeline とは別のリージョンで実行するようにリソースを設定できます。

次の例のパイプライン JSON ファイルは、ヨーロッパ (アイルランド) リージョンの EmrCluster リソースを実行する方法を示します。このとき、作業するクラスターの大量のデータが同じリージョンに存在することを前提としています。この例では、一般的なパイプラインとの唯一の違いは、EmrCluster の region フィールドの値が eu-west-1 に設定されていることです。

```
{
  "objects": [
    {
      "id": "Hourly",
      "type": "Schedule",
      "startDateTime": "2014-11-19T07:48:00",
      "endDateTime": "2014-11-21T07:48:00",
      "period": "1 hours"
    },
    {
      "id": "MyCluster",
      "type": "EmrCluster",
      "masterInstanceType": "m3.medium",
      "region": "eu-west-1",
      "schedule": {
        "ref": "Hourly"
      }
    },
    {
      "id": "MyEmrActivity",
      "type": "EmrActivity",
      "schedule": {
        "ref": "Hourly"
      },
      "runsOn": {
        "ref": "MyCluster"
      },
      "step": "/home/hadoop/contrib/streaming/hadoop-streaming.jar,-input,s3n://
elasticmapreduce/samples/wordcount/input,-output,s3://eu-west-1-bucket/wordcount/
output/#{@scheduledStartTime},-mapper,s3n://elasticmapreduce/samples/wordcount/
wordSplitter.py,-reducer,aggregate"
    }
  ]
}
```

次の表は、選択できるリージョンと `region` フィールドで使用する関連リージョンコードの一覧です。

> **Note**
>
> 以下のリストに示すリージョンでは、AWS Data Pipeline でワークフローを調整し、Amazon EMR や Amazon EC2 のリソースを起動できます。リージョンによっては AWS Data Pipeline がサポートされていない場合もあります。AWS Data Pipeline がサポートされているリージョンの詳細については、「AWS のリージョンとエンドポイント」を参照してください。

リージョン名	リージョンコード
米国東部 (バージニア北部)	us-east-1
米国東部 (オハイオ)	us-east-2
米国西部 (北カリフォルニア)	us-west-1
米国西部 (オレゴン)	us-west-2
カナダ (中部)	ca-central-1
欧州 (アイルランド)	eu-west-1
欧州 (ロンドン)	eu-west-2
欧州 (フランクフルト)	eu-central-1
アジアパシフィック (シンガポール)	ap-southeast-1
アジアパシフィック (シドニー)	ap-southeast-2
アジアパシフィック (ムンバイ)	ap-south-1
アジアパシフィック (東京)	ap-northeast-1
アジアパシフィック (ソウル)	ap-northeast-2

リージョン名	リージョンコード
南米 (サンパウロ)	sa-east-1

カスケードの失敗と再実行

AWS Data Pipeline では、依存関係がエラーになるかユーザーによってキャンセルされたときのパイプラインオブジェクトの動作を設定できます。他のパイプラインオブジェクト (コンシューマー) に確実に失敗をカスケードすることで、無限に保留状態になるのを防ぐことができます。すべてのアクティビティ、データノード、および前提条件には、failureAndRerunMode という名前のフィールドがあり、デフォルト値は none です。失敗のカスケードを有効にするには、failureAndRerunMode フィールドを cascade に設定します。

このフィールドを有効にしているときに、パイプラインオブジェクトが WAITING_ON_DEPENDENCIES 状態でブロックされ、保留コマンドがない状態で依存関係がエラーになった場合、カスケードの失敗が発生します。カスケードが失敗すると、以下のイベントが発生します。

- オブジェクトで障害が発生すると、そのコンシューマーは CASCADE_FAILED に設定され、元のオブジェクトとコンシューマーの前提条件が CANCELED に設定されます。
- 既に FINISHED、FAILED、または CANCELED の状態にあるオブジェクトは無視されます。

カスケードの失敗は、エラーになった元のオブジェクトに関連付けられている前提条件を除き、エラーになったオブジェクトの依存関係 (上流) に対しては作用しません。カスケードの失敗による影響を受けるパイプラインオブジェクトでは、再試行や、onFail などの後処理がトリガーされることがあります。

カスケードの失敗による詳細な影響は、オブジェクトのタイプによって異なります。

アクティビティ

依存関係のいずれかがエラーになり、それがアクティビティのコンシューマーにおけるカスケードの失敗の原因となった場合、アクティビティは CASCADE_FAILED に変更されます。アクティビティが依存するリソースでエラーが発生した場合、アクティビティは CANCELED になり、そのすべてのコンシューマーは CASCADE_FAILED になります。

データノードと前提条件

失敗したアクティビティの出力としてデータノードが設定されている場合、そのデータノードは CASCADE_FAILED 状態になります。データノードに関連する前提条件がある場合、データノードのエラーが前提条件に伝達され、それらの前提条件は CANCELED 状態になります。

リソース

リソースに依存するオブジェクトが FAILED 状態になり、リソースそのものが WAITING_ON_DEPENDENCIES 状態である場合、そのリソースは FINISHED 状態になります。

カスケードが失敗したオブジェクトの再実行

デフォルトでは、アクティビティまたはデータノードを再実行すると、関連するリソースのみが再実行されます。ただし、パイプラインオブジェクトで failureAndRerunMode フィールドを cascade に設定することで、以下の条件下でターゲットオブジェクトでの再実行コマンドをすべてのコンシューマーに伝達することができます。

- ターゲットオブジェクトのコンシューマーが CASCADE_FAILED 状態です。
- ターゲットオブジェクトの依存関係に、保留中の再実行コマンドがない。
- ターゲットオブジェクトの依存関係が、FAILED、CASCADE_FAILED、CANCELED のいずれの状態でもありません。

CASCADE_FAILED 状態のオブジェクトを再実行したが、その依存関係のいずれかが FAILED、CASCADE_FAILED、CANCELED のいずれかの状態である場合、再実行は失敗し、オブジェクトは CASCADE_FAILED の状態に戻ります。エラーになったオブジェクトを正常に再実行するには、依存関係の連鎖をさかのぼってエラーをトレースし、エラーの根本原因を特定して、そのオブジェクトを再実行する必要があります。リソースに対して再実行コマンドを発行するときは、それに依存するオブジェクトの再実行も試行することになります。

カスケードの失敗とバックフィル

失敗のカスケードを有効にしており、多くのバックフィルを生じるパイプラインがある場合、パイプラインのランタイムエラーによって、有用な処理が行われることなく、リソースの作成と削除が短時間のうちに連続して発生することがあります。AWS Data Pipeline は、パイプラインを保存するときに以下の警告メッセージを発行して、この状況についてのアラートを試みます。 *Pipeline_object_name* has 'failureAndRerunMode' field set to 'cascade' and

you are about to create a backfill with scheduleStartTime *start_time*. This can result in rapid creation of pipeline objects in case of failures. これは、カスケードの失敗によって下流のアクティビティが CASCADE_FAILED に設定されるために発生し、不要になった EMR クラスターと EC2 リソースがシャットダウンされます。この状況の影響を制限するために、短時間の範囲でパイプラインをテストすることをお勧めします。

パイプライン定義ファイルの構文

このセクションの手順は、AWS Data Pipeline コマンドラインインターフェイス (CLI) を使用してパイプライン定義ファイルを手動で操作するためのものです。これは、AWS Data Pipeline コンソールを使用してパイプラインをインタラクティブに設計する方法に代わる手段です。

パイプライン定義ファイルは、UTF-8 ファイル形式でのファイルの保存をサポートする任意のテキストエディタを使用して手動で作成し、AWS Data Pipeline コマンドラインインターフェイスを使用してファイルを送信することができます。

また、AWS Data Pipeline はパイプライン定義内で様々な複合式と関数もサポートします。詳細については、「パイプラインの式と関数」を参照してください。

ファイル構造

パイプラインの作成の最初のステップは、パイプライン定義ファイルでパイプライン定義オブジェクトを構成することです。次の例では、パイプライン定義ファイルの一般的な構造について説明します。このファイルでは、2 個のオブジェクトを '{' と '}' で囲み、カンマで区切って定義しています。

次の例では、最初のオブジェクトで名前と値のペアを 2 組定義しています。このペアをフィールドと呼びます。2 番目のオブジェクトでは、3 個のフィールドを定義しています。

```
{
  "objects" : [
    {
       "name1" : "value1",
       "name2" : "value2"
    },
    {
       "name1" : "value3",
       "name3" : "value4",
       "name4" : "value5"
    }
  ]
```

```
}
```

パイプライン定義ファイルを作成するときは、必要なパイプラインオブジェクトのタイプを選択してパイプライン定義ファイルに追加してから、適切なフィールドを追加する必要があります。パイプラインオブジェクトの詳細については、「パイプラインオブジェクトリファレンス」を参照してください。

たとえば、入力データノード用にパイプライン定義オブジェクトを 1 個作成し、出力データノード用にパイプライン定義オブジェクトをもう 1 個作成することができます。その後で、Amazon EMR を使用して入力データを処理するなどのアクティビティ用に、別のパイプライン定義オブジェクトを作成します。

パイプラインのフィールド

パイプライン定義ファイルに含めるオブジェクトタイプを決めたら、各パイプラインオブジェクトの定義にフィールドを追加します。次の例のように、フィールド名を引用符で囲み、スペース、コロン、スペースでフィールド値と区切ります。

```
"name" : "value"
```

フィールド値は、文字列、別のオブジェクトの参照、関数呼び出し、式、または前述のいずれかのタイプの整列済みリストとして指定できます。フィールド値に使用できるデータ型についての詳細は、「単純データ型」を参照してください。フィールド値の検証に使用できる関数の詳細については、「式の評価」を参照してください。

フィールドは、2048 文字に制限されています。オブジェクトのサイズは 20 KB までであるため、1 個のオブジェクトにサイズの大きいフィールドを多数追加することはできません。

各パイプラインオブジェクトには、次の例に示すように、id および type フィールドを含める必要があります。オブジェクトのタイプによっては、それ以外のフィールドが必須になる場合があります。id の値には、パイプライン定義内でユニークな、分かりやすい値を選択します。type の値は、オブジェクトのタイプを指定します。トピック「パイプラインオブジェクトリファレンス」にリストされている、サポートされるパイプライン定義オブジェクトのタイプのうち 1 つを指定します。

```
{
  "id": "MyCopyToS3",
  "type": "CopyActivity"
}
```

各オブジェクトの必須フィールドおよびオプションフィールドの詳細については、オブジェクトのドキュメントを参照してください。

あるオブジェクトのフィールドを別のオブジェクトで使用するには、そのオブジェクトの参照とともに parent フィールドを使用します。たとえば、オブジェクト "B" に、フィールド "B1" と "B2" に加えて、オブジェクト "A" のフィールド "A1" と "A2" を含めます。

```
{
  "id" : "A",
  "A1" : "value",
  "A2" : "value"
},
{
  "id" : "B",
  "parent" : {"ref" : "A"},
  "B1" : "value",
  "B2" : "value"
}
```

オブジェクトの共通のフィールドは、ID "Default" を使用して定義できます。これらのフィールドは、parent フィールドで明示的に別のオブジェクトへの参照を設定していない限り、パイプライン定義ファイル内のすべてのオブジェクトに自動的に含まれます。

```
{
  "id" : "Default",
  "onFail" : {"ref" : "FailureNotification"},
  "maximumRetries" : "3",
  "workerGroup" : "myWorkerGroup"
}
```

ユーザー定義フィールド

パイプラインコンポーネントでは、ユーザー定義フィールドまたはカスタムフィールドを作成し、式を使用してそれらを参照することができます。次の例では、S3DataNode オブジェクトに追加された myCustomField および my_customFieldReference というカスタムフィールドを示しています。

```
{
  "id": "S3DataInput",
  "type": "S3DataNode",
  "schedule": {"ref": "TheSchedule"},
```

```
  "filePath": "s3://bucket_name",
  "myCustomField": "This is a custom value in a custom field.",
  "my_customFieldReference": {"ref":"AnotherPipelineComponent"}
  },
```

ユーザー定義フィールドの名前には、すべて小文字で "my" というプレフィックスを付け、続けて大文字またはアンダースコア文字を使用する必要があります。さらに、ユーザー定義フィールドでは、前述の myCustomField の例のような文字列値、または前述の my_customFieldReference の例のような別のパイプラインコンポーネントへの参照を使用できます。

> **Note**
>
> ユーザー定義フィールドについては、AWS Data Pipeline は、別のパイプラインコンポーネントへの有効な参照のみをチェックし、ユーザーが追加するカスタムフィールドの文字列値は一切チェックしません。

API の使用

> **Note**
>
> AWS Data Pipeline を操作するプログラムを作成しているのではない場合、どの AWS SDK もインストールする必要はありません。コンソールまたはコマンドラインインターフェイスを使用してパイプラインを作成し、実行できます。詳細については、のセットアップ AWS Data Pipelineを参照してください。

AWS Data Pipeline を操作するアプリケーションを作成する場合や、カスタムタスクランナーを実装する場合、最も簡単な方法は AWS SDK のいずれかを使用する方法です。AWS SDK は、任意のプログラミング環境からのウェブサービス API の呼び出しを簡素化する機能を提供します。詳細については、「AWS SDK をインストールする」を参照してください。

AWS SDK をインストールする

AWS SDK には、API をラップして、署名の計算、リクエストの再試行の処理、エラー処理など、接続のさまざまな詳細を処理する関数が用意されています。また SDK には、AWS を呼び出すアプリケーションの作成を開始するのに役立つ、サンプルコード、チュートリアルなどのリソースも含まれています。SDK のラッパー関数を呼び出すと、AWS アプリケーションを作成するプロセスを大幅に

簡素化できます。AWS SDK をダウンロードして使用する方法の詳細については、「サンプルコードとライブラリ」を参照してください。

AWS Data Pipeline のサポートは、以下のプラットフォーム用の SDK で使用できます。

- AWS SDK for Java
- AWS SDK for Node.js
- AWS SDK for PHP
- AWS SDK for Python (Boto)
- AWS SDK for Ruby
- AWS SDK for .NET

AWS Data Pipeline に対する HTTP リクエストの実行

AWS Data Pipeline のプログラムオブジェクトの詳細については、AWS Data Pipeline API リファレンスを参照してください。

どの AWS SDK も使用しない場合は、POST リクエストメソッドを使用して、HTTP 経由で AWS Data Pipeline オペレーションを実行できます。POST メソッドでは、リクエストのヘッダーでオペレーションを指定し、リクエストの本文に、オペレーションのデータを JSON 形式で入力します。

HTTP ヘッダーの内容

AWS Data Pipeline では、HTTP リクエストのヘッダーに次の情報を入力する必要があります。

- `host` AWS Data Pipeline エンドポイント。

  エンドポイントの詳細については、「リージョンとエンドポイント」を参照してください。
- `x-amz-date` HTTP の Date ヘッダーまたは AWS の x-amz-date ヘッダーにタイムスタンプを入力する必要があります（一部の HTTP クライアントライブラリでは、Date ヘッダーを設定することができません)。x-amz-date ヘッダーがある場合には、リクエスト認証時に Date ヘッダーが無視されます。

  日付は、HTTP/1.1 RFC で規定されている次の 3 つ形式のいずれかで指定する必要があります。
  - Sun, 06 Nov 1994 08:49:37 GMT (RFC 822、RFC 1123 により更新)
  - Sunday, 06-Nov-94 08:49:37 GMT (RFC 850、RFC 1036 により廃止)
  - Sun Nov 6 08:49:37 1994（ANSI C asctime() 形式）

- Authorization AWS がリクエストの有効性と正当性を確保するために使用する、認可パラメータのセット。このヘッダーの作成方法の詳細については、「署名バージョン 4 の署名プロセス」を参照してください。
- x-amz-target 次の形式で指定する、リクエストの送信先サービスおよびデータのオペレーション: <<serviceName>>_<<API version>>.<<operationName>>

  例えば、DataPipeline_20121129.ActivatePipeline などです。
- content-type JSON とバージョンを指定します。例えば、Content-Type: application/x-amz-json-1.0 などです。

次に、パイプラインをアクティブ化する HTTP リクエストのサンプルヘッダーの例を示します。

```
POST / HTTP/1.1
host: https://datapipeline.us-east-1.amazonaws.com
x-amz-date: Mon, 12 Nov 2012 17:49:52 GMT
x-amz-target: DataPipeline_20121129.ActivatePipeline
Authorization: AuthParams
Content-Type: application/x-amz-json-1.1
Content-Length: 39
Connection: Keep-Alive
```

HTTP 本文の内容

HTTP リクエストの本文には、HTTP リクエストのヘッダーで指定されたオペレーションのデータが含まれます。このデータは、各 AWS Data Pipeline API の JSON データスキーマに従って形式が設定されている必要があります。AWS Data Pipeline の JSON データスキーマは、各オペレーションで使用できる、データとパラメータの型 (比較演算子や列挙定数など) を定義します。

HTTP リクエストの本文の形式の設定

JSON データ形式を使用すると、データ値とデータ構造を同時に送信できます。エレメントは、ブラケット表記を使用することで他のエレメント内にネストすることができます。次の例は、3 つのオブジェクトおよび対応するスロットから成るパイプライン定義を送信するためのリクエストを示しています。

```
{
 "pipelineId": "df-00627471SOVYZEXAMPLE",
 "pipelineObjects":
```

```
  [
    {"id": "Default",
     "name": "Default",
     "slots":
      [
        {"key": "workerGroup",
         "stringValue": "MyWorkerGroup"}
      ]
    },
    {"id": "Schedule",
     "name": "Schedule",
     "slots":
      [
       {"key": "startDateTime",
         "stringValue": "2012-09-25T17:00:00"},
        {"key": "type",
         "stringValue": "Schedule"},
        {"key": "period",
         "stringValue": "1 hour"},
        {"key": "endDateTime",
         "stringValue": "2012-09-25T18:00:00"}
      ]
    },
    {"id": "SayHello",
     "name": "SayHello",
     "slots":
      [
        {"key": "type",
         "stringValue": "ShellCommandActivity"},
        {"key": "command",
         "stringValue": "echo hello"},
        {"key": "parent",
         "refValue": "Default"},
        {"key": "schedule",
         "refValue": "Schedule"}

      ]
    }
  ]
}
```

HTTP 応答の処理

HTTP 応答の重要なヘッダーと、それらをアプリケーション内で扱う方法を示します。

- HTTP/1.1 – このヘッダーにはステータスコードが続きます。コードの値 200 は、オペレーションが成功したことを示します。その他の値はエラーを示します。
- x-amzn-RequestId – このヘッダーには、AWS Data Pipeline を使用してリクエストのトラブルシューティングを行う場合に使用するリクエスト ID が含まれています。リクエスト ID は、たとえば K2QH8DNOU907N97FNA2GDLL8OBVV4KQNSO5AEMVJF66Q9ASUAAJG のようになります。
- x-amz-crc32 – AWS Data Pipeline が HTTP ペイロードの CRC32 チェックサムを計算し、x-amz-crc32 ヘッダーでこのチェックサムを返します。クライアント側で独自の CRC32 チェックサムを計算して、x-amz-crc32 ヘッダーと比較することをお勧めします。チェックサムが一致しない場合は、データが送信中に壊れた可能性があります。その場合は、リクエストを再試行する必要があります。

Amazon DynamoDB からの各返信のチェックサムを SDK が計算し、不一致が検出された場合には自動的に再試行されるため、AWS SDK のユーザーは、この確認を手動で行う必要はありません。

AWS Data Pipeline JSON のリクエストと応答のサンプル

以下の例は新しいパイプラインを作成するためのリクエストを示しています。次に、新しく作成されたパイプラインの ID を含む AWS Data Pipeline 応答を示します。

HTTP POST リクエスト

```
POST / HTTP/1.1
host: https://datapipeline.us-east-1.amazonaws.com
x-amz-date: Mon, 12 Nov 2012 17:49:52 GMT
x-amz-target: DataPipeline_20121129.CreatePipeline
Authorization: AuthParams
Content-Type: application/x-amz-json-1.1
Content-Length: 50
Connection: Keep-Alive

{"name": "MyPipeline",
 "uniqueId": "12345ABCDEFG"}
```

AWS Data Pipeline 応答

```
HTTP/1.1 200
x-amzn-RequestId: b16911ce-0774-11e2-af6f-6bc7a6be60d9
x-amz-crc32: 2215946753
Content-Type: application/x-amz-json-1.0
Content-Length: 2
Date: Mon, 16 Jan 2012 17:50:53 GMT

{"pipelineId": "df-00627471SOVYZEXAMPLE"}
```

AWS Data Pipeline のセキュリティ

AWS では、クラウドセキュリティが最優先事項です。AWS のユーザーは、セキュリティを最も重視する組織の要件を満たすように構築されたデータセンターとネットワークアーキテクチャを利用できます。

セキュリティは、AWS とユーザーの間の責任共有です。責任共有モデルでは、これをクラウドのセキュリティおよびクラウド内のセキュリティと説明しています。

- クラウドのセキュリティ - AWS は、AWS クラウドで AWS のサービスを実行するインフラストラクチャを保護する責任を負います。また、AWS は、使用するサービスを安全に提供します。AWS コンプライアンスプログラムの一環として、サードパーティーの監査が定期的にセキュリティの有効性をテストおよび検証しています。AWS Data Pipeline に適用されるコンプライアンスプログラムの詳細については、「コンプライアンスプログラムによる AWS 対象範囲内のサービス」を参照してください。
- クラウド内のセキュリティ - ユーザーの責任は、使用する AWS のサービスに応じて異なります。また、お客様は、データの機密性、お客様の会社の要件、および適用される法律および規制など、その他の要因についても責任を負います。

このドキュメントは、AWS Data Pipeline を使用する際に責任共有モデルを適用する方法を理解するのに役立ちます。次のトピックでは、セキュリティおよびコンプライアンスの目的を達成するために AWS Data Pipeline を設定する方法を示します。また、AWS Data Pipeline リソースのモニタリングや保護に役立つ他の AWS サービスの使用方法についても説明します。

トピック

- AWS Data Pipeline でのデータ保護
- AWS Data Pipeline 向けの Identity and Access Management
- AWS Data Pipelineでのログ記録とモニタリング
- AWS Data Pipeline のインシデントへの対応
- AWS Data Pipeline のコンプライアンス検証
- AWS Data Pipeline での耐障害性
- AWS Data Pipeline でのインフラストラクチャセキュリティ
- AWS Data Pipeline での設定と脆弱性の分析

AWS Data Pipeline でのデータ保護

AWS責任共有モデルは、AWS Data Pipeline でのデータ保護に適用されます。このモデルで説明されているように、AWS は、AWS クラウド のすべてを実行するグローバルインフラストラクチャを保護する責任を負います。顧客は、このインフラストラクチャでホストされているコンテンツに対する管理を維持する責任があります。このコンテンツには、使用される AWS サービス のセキュリティ構成と管理タスクが含まれます。データプライバシーの詳細については、「データプライバシーのよくある質問」を参照してください。欧州でのデータ保護の詳細については、AWS セキュリティブログに投稿された「AWS 責任共有モデルおよび GDPR」のブログ記事を参照してください。

データを保護するため、AWS アカウント の認証情報を保護し、AWS IAM Identity Center または AWS Identity and Access Management (IAM) を使用して個々のユーザーをセットアップすることをお勧めします。この方法により、それぞれのジョブを遂行するために必要なアクセス許可のみを各ユーザーに付与できます。また、次の方法でデータを保護することをお勧めします。

- 各アカウントで多要素認証 (MFA) を使用します。
- SSL/TLS を使用して AWS リソースと通信します。TLS 1.2 以降が推奨されます。
- AWS CloudTrail で API とユーザーアクティビティログをセットアップします。
- AWS サービス 内でデフォルトである、すべてのセキュリティ管理に加え、AWS の暗号化ソリューションを使用します。
- Amazon Macie などの高度なマネージドセキュリティサービスを使用します。これらは、Amazon S3 に保存されている機密データの検出と保護を支援します。
- コマンドラインインターフェイスまたは API を使用して AWS にアクセスするときに FIPS 140-2 検証済みの暗号化モジュールが必要な場合は、FIPS エンドポイントを使用します。利用可能な FIPS エンドポイントの詳細については、連邦情報処理規格 (FIPS) 140-2 を参照してください。
- AWS Data Pipeline は、Amazon EMR および Amazon EC2 リソースの IMDSv2 をサポートします。Amazon EMR で IMDSv2 を使用するには、バージョン 5.23.1、5.27.1、5.32 以降、またはバージョン 6.2 以降を使用してください。詳細については、「Amazon EC2 インスタンスへのメタデータサービスリクエストを設定する」と「IMDSv2 の使用」を参照してください。

お客様の E メールアドレスなどの極秘または機密情報は、タグ、または名前フィールドなどの自由形式のテキストフィールドに配置しないことを強くお勧めします。これは、コンソール、API、AWS CLI、または AWS SDK で AWS Data Pipeline または他の AWS サービス を使用する場合も同様です。タグ、または名前に使用される自由形式のテキストフィールドに入力されるデータは、請求または診断ログに使用される場合があります。外部サーバーに URL を提供する場合、そのサーバーへのリクエストを検証できるように、認証情報を URL に含めないことを強くお勧めします。

AWS Data Pipeline 向けの Identity and Access Management

セキュリティ認証情報により、AWS のサービスに対してお客様の身分が証明され、パイプラインなどの AWS リソースを使用するアクセス許可が付与されます。AWS Data Pipeline および AWS Identity and Access Management (IAM) の機能を使用して、AWS Data Pipeline ユーザーとその他のユーザーが AWS Data Pipeline のリソースを使用できるようにできます。その際、お客様のセキュリティ認証情報は共有されません。

組織はパイプラインへのアクセスを共有することができ、これによって組織内の個人がパイプラインを共同で開発および管理できます。ただし、次のような措置を講じる必要がある場合があります。

- 特定のパイプラインにどのユーザーがアクセスできるかを制御する
- 誤って編集されないように実稼働のパイプラインを保護する
- 監査人に対してパイプラインへの読み取り専用アクセスは許可し、変更は許可しない。

AWS Data Pipeline は、幅広い機能を提供する AWS Identity and Access Management (IAM) と統合されます。

- AWS アカウント にユーザーとグループを作成します。
- AWS アカウント 内のユーザー間で AWS リソースを簡単に共有できます。
- 各ユーザーに一意のセキュリティ認証情報を割り当てます。
- サービスとリソースへの各ユーザーのアクセス権限を制御します。
- AWS アカウント 内のすべてのユーザーに対する単一の請求書を受け取ります。

AWS Data Pipeline と組み合わせて IAM を使用すると、組織のユーザーが特定の API アクションを使用してタスクを実行できるかどうか、また、特定の AWS リソースを使用できるかどうかを制御できます。パイプラインのタグとワーカーグループに基づく IAM ポリシーを使用して、パイプラインを他のユーザーと共有し、ユーザーのアクセスレベルを制御することができます。

目次

- AWS Data Pipeline の IAM ポリシー
- AWS Data Pipeline のポリシー例
- AWS Data Pipeline の IAM ロール

AWS Data Pipeline の IAM ポリシー

デフォルトでは、IAM エンティティには AWS リソースを作成または変更するためのアクセス許可はありません。IAM エンティティがリソースを作成または変更、およびタスクを実行できるようにするには、IAM ポリシーを作成する必要があります。これによって、必要な特定のリソースおよび API アクションを使用するためのアクセス許可を IAM エンティティに付与し、その後、ポリシーをそのアクセス許可が必要な IAM エンティティにアタッチします。

ポリシーをユーザーまたはユーザーのグループにアタッチする場合、ポリシーによって特定リソースの特定タスクを実行するユーザーの権限が許可または拒否されます。IAM ポリシーの一般的な情報については、IAM ユーザーガイドのアクセス許可とポリシーを参照してください。カスタム IAM ポリシーの管理と作成の詳細については、IAM ポリシーの管理を参照してください。

目次

- ポリシー構文
- タグを使用したパイプラインへのアクセスの制御
- ワーカーグループを使用したパイプラインへのアクセスの制御

ポリシー構文

IAM ポリシーは 1 つ以上のステートメントで構成される JSON ドキュメントです。各ステートメントは次のように構成されます。

```
{
  "Statement":[{
    "Effect":"effect",
    "Action":"action",
    "Resource":"*",
    "Condition":{
      "condition":{
        "key":"value"
        }
      }
    }
  ]
}
```

ポリシーステートメントは以下の要素によって構成されます。

- [Effect]:effect は、AllowまたはDenyにすることができます。デフォルトでは、IAM エンティティはリソースおよび API アクションを使用するアクセス許可がないため、リクエストはすべて拒否されます。明示的な許可はデフォルトに上書きされます。明示的な拒否はすべての許可に優先します。
- [Action]: action は、アクセス許可を付与または拒否する対象とする、特定の API アクションです。AWS Data Pipeline のアクションのリストについては、AWS Data Pipeline API リファレンスのアクションを参照してください。
- [Resource] (リソース): アクションによって影響を及ぼされるリソースです。ここで唯一の有効な値は "*" です。
- [Condition] (条件): condition はオプションです。ポリシーの発効条件を指定するために使用します。

  AWS Data Pipeline は、AWS 全体のコンテキストキー (「条件に利用可能なキー」を参照) に加え、以下のサービス固有のキーを実装します。

  - datapipeline:PipelineCreator – パイプラインを作成したユーザーにアクセスを許可します。これの例については、「パイプライン所有者にフルアクセスを付与する」を参照してください。
  - datapipeline:Tag – パイプラインのタグ付けに基づいてアクセスを許可します。詳細については、「タグを使用したパイプラインへのアクセスの制御」を参照してください。
  - datapipeline:workerGroup – ワーカーグループの名前に基づいてアクセスを許可します。詳細については、「ワーカーグループを使用したパイプラインへのアクセスの制御」を参照してください。

タグを使用したパイプラインへのアクセスの制御

パイプラインのタグを参照する IAM ポリシーを作成できます。これにより、パイプラインのタグ付けを使用して以下の操作を行うことができます。

- パイプラインへの読み取り専用アクセス権限の付与
- パイプラインへの読み取り/書き込みアクセス権限の付与
- パイプラインへのアクセスのブロック

たとえば、管理者が実稼働用と開発用の 2 つのパイプライン環境を使用しており、それぞれの環境に対して IAM グループを設定しているとします。管理者は、実稼働環境のパイプライン環境に対して、実稼働用 IAM グループのユーザーには読み取り/書き込みアクセス権限を付与し、開発者用 IAM

グループのユーザーには読み取り専用アクセス権限を付与します。管理者は、開発環境のパイプライン環境に対して、実稼働用と開発用 IAM グループのユーザーには読み取り/書き込みアクセス権限を付与します。

このシナリオを達成するため、管理者は実稼働用パイプラインに "environment=production" というタグを付け、開発者用 IAM グループに次のポリシーをアタッチします。最初のステートメントでは、すべてのパイプラインに対する読み取り専用アクセスを付与しています。2 番目のステートメントでは、"environment=production" タグがないパイプラインへの読み取り/書き込みアクセスを付与しています。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:Describe*",
        "datapipeline:ListPipelines",
        "datapipeline:GetPipelineDefinition",
        "datapipeline:QueryObjects"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "datapipeline:*",
      "Resource": "*",
      "Condition": {
        "StringNotEquals": {"datapipeline:Tag/environment": "production"}
      }
    }
  ]
}
```

さらに管理者は、実稼働の IAM グループに次のポリシーをアタッチします。このステートメントは、すべてのパイプラインに対するフルアクセスを付与しています。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
```

```
      "Action": "datapipeline:*",
      "Resource": "*"
    }
  ]
}
```

その他の例については、「タグに基づいてユーザーに読み取り専用アクセスを付与する」と「タグに基づいてユーザーにフルアクセスを付与する」を参照してください。

ワーカーグループを使用したパイプラインへのアクセスの制御

参照ワーカーグループ名を作成する IAM ポリシーを作成できます。

たとえば、管理者が実稼働用と開発用の 2 つのパイプライン環境を使用しており、それぞれの環境に対して IAM グループを設定しているとします。3 台のデータベースサーバーがあり、それぞれ実稼働環境、実稼働準備環境、および開発者環境に対して Task Runner が設定されているものとします。管理者は、実稼働用 IAM グループのユーザーは実稼働用リソースにタスクをプッシュするパイプラインを作成でき、開発用 IAM グループのユーザーは実稼働準備用と開発者用の両方のリソースにタスクをプッシュするパイプラインを作成できるようにする必要があります。

このシナリオを達成するため、管理者は実稼働用認証情報を使用して実稼働用リソースに Task Runner をインストールし、`workerGroup` を "prodresource" に設定します。さらに、開発用認証情報を使用して開発用リソースに Task Runner をインストールし、`workerGroup` を "pre-production" と "development" に設定します。管理者は、開発者用 IAM グループに次のポリシーをアタッチして、"prodresource" リソースへのアクセスをブロックします。最初のステートメントでは、すべてのパイプラインに対する読み取り専用アクセスを付与しています。2 番目のステートメントは、ワーカーグループの名前に "dev" または "pre-prod" というプレフィックスが含まれていれば、パイプラインへの読み取り/書き込みアクセスを付与します。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:Describe*",
        "datapipeline:ListPipelines",
        "datapipeline:GetPipelineDefinition",
        "datapipeline:QueryObjects"
      ],
      "Resource": "*"
```

```
    },
    {
      "Action": "datapipeline:*",
      "Effect": "Allow",
      "Resource": "*",
      "Condition": {
        "StringLike": {
          "datapipeline:workerGroup": ["dev*","pre-prod*"]
        }
      }
    }
  ]
}
```

さらに、管理者は、実稼働用 IAM グループに次のポリシーをアタッチして、"prodresource" リソースへのアクセス権限を付与します。最初のステートメントでは、すべてのパイプラインに対する読み取り専用アクセスを付与しています。2 番目のステートメントは、ワーカーグループの名前に "prod" というプレフィックスが含まれていれば、読み取り/書き込みアクセスを付与します。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:Describe*",
        "datapipeline:ListPipelines",
        "datapipeline:GetPipelineDefinition",
        "datapipeline:QueryObjects"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "datapipeline:*",
      "Resource": "*",
      "Condition": {
        "StringLike": {"datapipeline:workerGroup": "prodresource*"}
      }
    }
  ]
}
```

AWS Data Pipeline のポリシー例

次の例では、パイプラインへのフルアクセスまたは限定的なアクセスをユーザーに許可する方法を示します。

目次

- 例 1: タグに基づいてユーザーに読み取り専用アクセスを付与する
- 例 2: タグに基づいてユーザーにフルアクセスを付与する
- 例 3: パイプライン所有者にフルアクセスを付与する
- 例 4: ユーザーに AWS Data Pipeline コンソールへのアクセスを許可する

例 1: タグに基づいてユーザーに読み取り専用アクセスを付与する

次のポリシーでは、ユーザーは、"environment=production" タグの付いたパイプラインのみを実行する読み取り専用の AWS Data Pipeline API アクションを使用できます。

ListPipelines API アクションでは、タグに基づいた権限付与はサポートされていません。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:Describe*",
        "datapipeline:GetPipelineDefinition",
        "datapipeline:ValidatePipelineDefinition",
        "datapipeline:QueryObjects"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "datapipeline:Tag/environment": "production"
        }
      }
    }
  ]
}
```

例 2: タグに基づいてユーザーにフルアクセスを付与する

以下のポリシーでは、ユーザーは、"environment=test" タグの付いたパイプラインのみを実行するすべての AWS Data Pipeline API アクション (ListPipelines 以外) を使用できます。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:*"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "datapipeline:Tag/environment": "test"
        }
      }
    }
  ]
}
```

例 3: パイプライン所有者にフルアクセスを付与する

次のポリシーでは、ユーザーは、独自のパイプラインのみを実行するすべての AWS Data Pipeline API アクションを使用できます。

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "datapipeline:*"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
```

```
          "datapipeline:PipelineCreator": "${aws:userid}"
        }
      }
    }
  ]
}
```

例 4: ユーザーに AWS Data Pipeline コンソールへのアクセスを許可する

次のポリシーでは、ユーザーは、AWS Data Pipeline コンソールを使用してパイプラインを作成および管理できます。

このポリシーには、AWS Data Pipeline に必要な `roleARN` に紐付けられた特定のリソースに対する `PassRole` アクセス権限が含まれます。アイデンティティベース (IAM) の `PassRole` アクセス権限の詳細については、ブログ記事「IAM ロール (PassRole アクセス権限) を使用した EC2 インスタンスの起動アクセス権限の付与」を参照してください。

```
{
 "Version": "2012-10-17",
 "Statement": [{
   "Action": [
    "cloudwatch:*",
    "datapipeline:*",
    "dynamodb:DescribeTable",
    "elasticmapreduce:AddJobFlowSteps",
    "elasticmapreduce:ListInstance*",
    "iam:AddRoleToInstanceProfile",
    "iam:CreateInstanceProfile",
    "iam:GetInstanceProfile",
    "iam:GetRole",
    "iam:GetRolePolicy",
    "iam:ListInstanceProfiles",
    "iam:ListInstanceProfilesForRole",
    "iam:ListRoles",
    "rds:DescribeDBInstances",
    "rds:DescribeDBSecurityGroups",
    "redshift:DescribeClusters",
    "redshift:DescribeClusterSecurityGroups",
    "s3:List*",
    "sns:ListTopics"
   ],
   "Effect": "Allow",
   "Resource": [
```

```
    "*"
   ]
  },
  {
   "Action": "iam:PassRole",
   "Effect": "Allow",
   "Resource": [
    "arn:aws:iam::*:role/DataPipelineDefaultResourceRole",
    "arn:aws:iam::*:role/DataPipelineDefaultRole"
   ]
  }
 ]
}
```

AWS Data Pipeline の IAM ロール

AWS Data Pipeline は AWS Identity and Access Management ロールを使用します。IAM ロールにアタッチされたアクセス許可ポリシーにより、AWS Data Pipeline およびアプリケーションが実行できるアクション、およびそれらがアクセスできる AWS リソースが決定されます。詳細については、「IAM ユーザーガイド」の「IAM ロール」を参照してください。

AWS Data Pipeline では次の 2 つの IAM ロールが必要です。

- パイプラインロールは、AWS リソースへの AWS Data Pipeline アクセス権限を制御します。パイプラインオブジェクト定義では、`role` フィールドによってこのロールが指定されます。
- EC2 インスタンスロールは、EC2 インスタンス (Amazon EMR クラスター内の EC2 インスタンスを含む) で実行されているアプリケーションが AWS リソースに対して備えるアクセス権限を制御します。パイプラインオブジェクト定義では、`resourceRole` フィールドによってこのロールが指定されます。

> **Important**
>
> デフォルトロールで AWS Data Pipeline コンソールを使用して 2022 年 10 月 3 日より前にパイプラインを作成した場合、AWS Data Pipeline によって `DataPipelineDefaultRole` が作成され、`AWSDataPipelineRole` マネージドポリシーがロールにアタッチされています。2022 年 10 月 3 日から、`AWSDataPipelineRole` マネージドポリシーは廃止され、コンソールを使用するときにパイプラインのパイプラインロールを指定する必要があります。既存のパイプラインを確認し、`DataPipelineDefaultRole` がパイプラインと関連付けられているかどうか、および `AWSDataPipelineRole` がそのロールにアタッチされているか

どうかを判別することをお勧めします。満たされている場合は、このポリシーで許可されているアクセス権限を確認して、セキュリティ要件に適したものになっていることを確認してください。必要に応じて、このロールにアタッチされたポリシーおよびポリシーステートメントを追加、更新、または置換します。または、パイプラインを更新して、異なるアクセス許可ポリシーで作成したロールを使用することもできます。

AWS Data Pipeline ロールのアクセス許可ポリシーの例

各ロールには、そのロールがアクセスできる AWS リソースおよびそのロールが実行できるアクションを決定する 1 つ以上のアクセス許可ポリシーがアタッチされています。このトピックでは、パイプラインロールのアクセス許可ポリシーの例を示します。また、デフォルトの EC2 インスタンスロール DataPipelineDefaultResourceRole のマネージドポリシーである AmazonEC2RoleforDataPipelineRole のコンテンツも示します。

パイプラインロールのアクセス許可ポリシーの例

以下のポリシー例は、Amazon EC2 および Amazon EMR リソースを使用してパイプラインを実行するために AWS Data Pipeline で必要な必須の関数を許可するようにスコープ設定されています。また、多くのパイプラインで必要となる他の AWS リソース (Amazon Simple Storage Service や Amazon Simple Notification Service など) にアクセスするためのアクセス許可も指定しています。パイプラインで定義されているオブジェクトで AWS サービスのリソースが必要ない場合、そのサービスにアクセスするためのアクセス許可を削除することを強くお勧めします。例えば、パイプラインで D ynamoDBDataノード が定義されておらず、SnsAlarm アクションを使用していない場合、それらのアクションの許可ステートメントを削除することをお勧めします。

- *111122223333* を AWS アカウント ID に置き換えます。
- *NameOfDataPipelineRole* を、パイプラインロール (このポリシーがアタッチされているロール) の名前に置き換えます。
- *NameOfDataPipelineResourceRole* を、EC2 インスタンスロールの名前に置き換えます。
- *us-west-1* を、アプリケーションに適したリージョンに置き換えます。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
```

```
                "iam:GetInstanceProfile",
                "iam:GetRole",
                "iam:GetRolePolicy",
                "iam:ListAttachedRolePolicies",
                "iam:ListRolePolicies",
                "iam:PassRole"
            ],
            "Resource": [
                "arn:aws:iam::111122223333:role/NameOfDataPipelineRole",
                "arn:aws:iam::111122223333 :role/NameOfDataPipelineResourceRole"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "ec2:AuthorizeSecurityGroupEgress",
                "ec2:AuthorizeSecurityGroupIngress",
                "ec2:CancelSpotInstanceRequests",
                "ec2:CreateNetworkInterface",
                "ec2:CreateSecurityGroup",
                "ec2:CreateTags",
                "ec2:DeleteNetworkInterface",
                "ec2:DeleteSecurityGroup",
                "ec2:DeleteTags",
                "ec2:DescribeAvailabilityZones",
                "ec2:DescribeAccountAttributes",
                "ec2:DescribeDhcpOptions",
                "ec2:DescribeImages",
                "ec2:DescribeInstanceStatus",
                "ec2:DescribeInstances",
                "ec2:DescribeKeyPairs",
                "ec2:DescribeLaunchTemplates",
                "ec2:DescribeNetworkAcls",
                "ec2:DescribeNetworkInterfaces",
                "ec2:DescribePrefixLists",
                "ec2:DescribeRouteTables",
                "ec2:DescribeSecurityGroups",
                "ec2:DescribeSpotInstanceRequests",
                "ec2:DescribeSpotPriceHistory",
                "ec2:DescribeSubnets",
                "ec2:DescribeTags",
                "ec2:DescribeVpcAttribute",
                "ec2:DescribeVpcEndpoints",
                "ec2:DescribeVpcEndpointServices",
```

```
                "ec2:DescribeVpcs",
                "ec2:DetachNetworkInterface",
                "ec2:ModifyImageAttribute",
                "ec2:ModifyInstanceAttribute",
                "ec2:RequestSpotInstances",
                "ec2:RevokeSecurityGroupEgress",
                "ec2:RunInstances",
                "ec2:TerminateInstances",
                "ec2:DescribeVolumeStatus",
                "ec2:DescribeVolumes",
                "elasticmapreduce:TerminateJobFlows",
                "elasticmapreduce:ListSteps",
                "elasticmapreduce:ListClusters",
                "elasticmapreduce:RunJobFlow",
                "elasticmapreduce:DescribeCluster",
                "elasticmapreduce:AddTags",
                "elasticmapreduce:RemoveTags",
                "elasticmapreduce:ListInstanceGroups",
                "elasticmapreduce:ModifyInstanceGroups",
                "elasticmapreduce:GetCluster",
                "elasticmapreduce:DescribeStep",
                "elasticmapreduce:AddJobFlowSteps",
                "elasticmapreduce:ListInstances",
                "iam:ListInstanceProfiles",
                "redshift:DescribeClusters"
            ],
            "Resource": [
                "*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "sns:GetTopicAttributes",
                "sns:Publish"
            ],
            "Resource": [
                "arn:aws:sns:us-west-1:111122223333:MyFirstSNSTopic",
                "arn:aws:sns:us-west-1:111122223333:MySecondSNSTopic",
                "arn:aws:sns:us-west-1:111122223333:AnotherSNSTopic"
            ]
        },
        {
            "Effect": "Allow",
```

```
            "Action": [
                "s3:ListBucket",
                "s3:ListMultipartUploads"
            ],
            "Resource": [
              "arn:aws:s3:::MyStagingS3Bucket",
              "arn:aws:s3:::MyLogsS3Bucket",
              "arn:aws:s3:::MyInputS3Bucket",
              "arn:aws:s3:::MyOutputS3Bucket",
              "arn:aws:s3:::AnotherRequiredS3Buckets"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectMetadata",
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::MyStagingS3Bucket/*",
                "arn:aws:s3:::MyLogsS3Bucket/*",
                "arn:aws:s3:::MyInputS3Bucket/*",
                "arn:aws:s3:::MyOutputS3Bucket/*",
                "arn:aws:s3:::AnotherRequiredS3Buckets/*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "dynamodb:Scan",
                "dynamodb:DescribeTable"
            ],
            "Resource": [
                "arn:aws:dynamodb:us-west-1:111122223333:table/MyFirstDynamoDBTable",
                "arn:aws:dynamodb:us-west-1:111122223333:table/MySecondDynamoDBTable",
                "arn:aws:dynamodb:us-west-1:111122223333:table/AnotherDynamoDBTable"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "rds:DescribeDBInstances"
            ],
```

```
            "Resource": [
                "arn:aws:rds:us-west-1:111122223333:db:MyFirstRdsDb",
                "arn:aws:rds:us-west-1:111122223333:db:MySecondRdsDb",
                "arn:aws:rds:us-west-1:111122223333:db:AnotherRdsDb"
            ]
        }
    ]
}
```

EC2 インスタンスロールのデフォルトマネージドポリシー

AmazonEC2RoleforDataPipelineRole のコンテンツを以下に示します。これは、AWS Data Pipeline のデフォルトリソースロール (DataPipelineDefaultResourceRole) にアタッチされているマネージドポリシーです。パイプラインのリソースロールを定義するときは、このアクセス許可ポリシーから開始し、必要ない AWS サービスアクションのアクセス許可を削除することをお勧めします。

ポリシーのバージョン 3 を示します。これは、この執筆時点での最新のバージョンです。IAM コンソールを使用して、ポリシーの最新バージョンを表示します。

```
{
  "Version": "2012-10-17",
  "Statement": [{
      "Effect": "Allow",
      "Action": [
        "cloudwatch:*",
        "datapipeline:*",
        "dynamodb:*",
        "ec2:Describe*",
        "elasticmapreduce:AddJobFlowSteps",
        "elasticmapreduce:Describe*",
        "elasticmapreduce:ListInstance*",
        "elasticmapreduce:ModifyInstanceGroups",
        "rds:Describe*",
        "redshift:DescribeClusters",
        "redshift:DescribeClusterSecurityGroups",
        "s3:*",
        "sdb:*",
        "sns:*",
        "sqs:*"
      ],
      "Resource": ["*"]
    }]
```

```
}
```

AWS Data Pipeline の IAM ロールの作成およびロールアクセス許可の編集

次の手順に従って、IAM コンソールを使用して AWS Data Pipeline のロールを作成します。このプロセスは次の 2 つのステップで構成されています。まず、ロールにアタッチするアクセス許可ポリシーを作成します。次に、ロールを作成して、ポリシーをアタッチします。ロールを作成した後、アクセス許可ポリシーをアタッチおよびデタッチして、ロールのアクセス許可を変更できます。

> **Note**
>
> 以下の説明に従ってコンソールを使用して AWS Data Pipeline のロールを作成した場合、IAM は、ロールに必要な適切な信頼ポリシーを作成してアタッチします。

AWS Data Pipeline のロールで使用するアクセス許可ポリシーを作成するには

1. https://console.aws.amazon.com/iam/ で IAM コンソール を開きます。
2. ナビゲーションペインで [Policies] (ポリシー) を選択してから [Create policy] (ポリシーの作成) を選択します。
3. [JSON] タブを選択します。
4. パイプラインロールを作成する場合は、パイプラインロールのアクセス許可ポリシーの例のポリシーの例のコンテンツをコピーして貼り付け、セキュリティ要件に応じて適宜編集します。または、カスタム EC2 インスタンスロールを作成する場合は、EC2 インスタンスロールのデフォルトマネージドポリシーの例と同様にします。
5. [ポリシーの確認] を選択します。
6. ポリシーの名前 (例えば、`MyDataPipelineRolePolicy`) とオプションの [Description] (説明) を入力してから、[Create policy] (ポリシーの作成) を選択します。
7. ポリシーの名前をメモします。これは、ロールを作成するときに必要になります。

AWS Data Pipeline 用に IAM ロールを作成するには

1. https://console.aws.amazon.com/iam/IAMコンソールを開きます。
2. ナビゲーションペインで [Roles] (ロール) を選択し、続いて [Create Role] (ロールの作成) を選択します。
3. [Choose a use case] (ユースケースの選択) で、[Data Pipeline] を選択します。

4. [Select your use case] (ユースケースの選択) で、次のいずれかを実行します。
   - `Data Pipeline` を選択して、パイプラインロールを作成します。
   - `EC2 Role for Data Pipeline` を選択して、リソースロールを作成します。
5. [Next: Permissions] (次へ: 許可) を選択します。
6. AWS Data Pipeline のデフォルトポリシーがリストされている場合、以下のステップに進んでロールを作成してから、次の手順の説明に従ってそのロールを編集します。それ以外の場合は、上記の手順で作成したポリシーの名前を入力し、リストからそのポリシーを選択します。
7. [Next: Tags] (次へ: タグ) を選択し、ロールに追加するタグを入力してから、[Next: Review] (次へ: 確認) を選択します。
8. ロールの名前 (例えば、`MyDataPipelineRole`) とオプションの [Description] (説明) を入力してから、[Create role] (ロールの作成) を選択します。

AWS Data Pipeline の IAM ロールのアクセス許可ポリシーをアタッチまたはデタッチするには

1. https://console.aws.amazon.com/iam/IAMコンソールを開きます。
2. ナビゲーションペインで [Roles] (ロール) を選択します。
3. 検索ボックスで、編集するロールの名前 (例えば、DataPipelineDefaultRole や MyDataPipelineRole) を入力していき、リストから [Role name] (ロール名) を選択します。
4. [Permissions] (アクセス許可) タブで、以下を実行します。
   - アクセス許可ポリシーをデタッチするには、[Permissions policies] (アクセス許可ポリシー) で、ポリシーエントリの右端にある削除ボタンを選択します。確認を求められたら、[Detach] (デタッチ) を選択します。
   - 前に作成したポリシーをアタッチするには、[Attach policies] (ポリシーをアタッチします) を選択します。検索ボックスで、編集するポリシーの名前を入力していき、リストからポリシーを選択し、[Attach policy] (ポリシーのアタッチ) を選択します。

既存のパイプラインのロールの変更

パイプラインに別のパイプラインロールまたはリソースロールを割り当てる場合は、AWS Data Pipeline コンソールでアーキテクトエディタを使用できます。

コンソールを使用してパイプラインに割り当てられているロールを編集するには

1. AWS Data Pipeline コンソール (https://console.aws.amazon.com/datapipeline/) を開きます。

2. リストからパイプラインを選択し、[Actions] (アクション)、[Edit] (編集) を選択します。
3. アーキテクトエディタの右ペインで、[Others] (その他) を選択します。
4. [Resource Role] (リソースロール) および [Role] (ロール) のリストで、割り当てる AWS Data Pipeline のロールを選択してから、[Save] (保存) を選択します。

AWS Data Pipelineでのログ記録とモニタリング

AWS Data PipelineはAWS CloudTrailという、AWS Data Pipelineのユーザーやロール、またはAWSのサービスによって実行されたアクションを記録するサービスと統合しています。CloudTrailは、AWS Data PipelineのすべてのAPIコールをイベントとしてキャプチャします。キャプチャされた呼び出しには、AWS Data Pipelineコンソールの呼び出しと、AWS Data Pipeline API オペレーションへのコード呼び出しが含まれます。証跡を作成する場合は、AWS Data Pipelineのイベントなど、Amazon S3バケットへのCloudTrailイベントの継続的な配信を有効にすることができます。証跡を設定しない場合でも、CloudTrailコンソールの[Event history (イベント履歴)]で最新のイベントを表示できます。CloudTrailによって収集された情報を使用して、AWS Data Pipelineに対して行われた要求、要求が行われたIPアドレス、要求を行った人、要求が行われた日時、および追加の詳細を判別できます。

CloudTrailの詳細については、「AWS CloudTrailユーザーガイド」を参照してください。

AWS Data PipelineCloudTrailでの情報

CloudTrailは、アカウントを作成するとAWSアカウントで有効になります。AWS Data Pipelineでアクティビティが発生すると、そのアクティビティは[Event history (イベント履歴)]でAWSのその他のサービスのイベントと共にCloudTrailイベントに記録されます。最近のイベントは、AWSアカウントで表示、検索、ダウンロードできます。詳細については、「CloudTrailイベント履歴でのイベントの表示」を参照してください。

AWSのイベントなど、AWS Data Pipelineアカウントのイベントの継続的なレコードについては、追跡を作成します。追跡により、CloudTrailはログファイルをSimple Storage Service (Amazon S3) バケットに配信できます。デフォルトでは、コンソールで作成した追跡がすべてのAWSリージョンに適用されます。追跡は、AWSパーティションのすべてのリージョンからのイベントを記録し、指定したSimple Storage Service (Amazon S3) バケットにログファイルを配信します。さらに、CloudTrailログで収集したイベントデータをより詳細に分析し、それに基づく対応するためにその他のAWSのサービスを設定できます。詳細については、次を参照してください。

- 追跡を作成するための概要

- CloudTrailのサポート対象サービスと統合
- Amazon SNSのCloudTrailの通知の設定
- 「複数のリージョンからCloudTrailログファイルを受け取る」および「複数のアカウントからCloudTrailログファイルを受け取る」

ログには、CloudTrailによってすべてのAWS Data Pipelineアクションが記録されます。これらのアクションについては、AWS Data Pipeline APIリファレンスのアクションの章を参照してください。例えば、CreatePipelineアクションを呼び出すと、CloudTrailログファイルにエントリが生成されます。

各イベントまたはログエントリには、リクエストの生成者に関する情報が含まれます。同一性情報は次の判断に役立ちます。

- リクエストが、ルートとIAMロール認証情報のどちらを使用して送信されたか。
- リクエストがロールまたはフェデレーションユーザーの一時的なセキュリティ認証情報を使用して行われたかどうか。
- リクエストが、別のAWSのサービスによって送信されたかどうか。

詳細については、「CloudTrail userIdentity要素」を参照してください。

AWS Data Pipeline ログファイルエントリの概要

追跡は、指定したAmazon S3バケットにイベントをログファイルとして配信するように設定できます。CloudTrailのログファイルには、単一か複数のログエントリがあります。イベントはあらゆるソースからの単一のリクエストを表し、リクエストされたアクション、アクションの日時、リクエストのパラメータなどの情報が含まれます。CloudTrailログファイルは、パブリックAPIコールの順序付けられたスタックトレースではないため、特定の順序では表示されません。

`CreatePipeline`オペレーションを示すCloudTrailログエントリの例は、次のとおりです。

```
    {
"Records": [
  {
    "eventVersion": "1.02",
    "userIdentity": {
      "type": "Root",
```

```
        "principalId": "123456789012",
        "arn": "arn:aws:iam::aws-account-id:role/role-name",
        "accountId": "role-account-id",
        "accessKeyId": "role-access-key"
      },
      "eventTime": "2014-11-13T19:15:15Z",
      "eventSource": "datapipeline.amazonaws.com",
      "eventName": "CreatePipeline",
      "awsRegion": "us-east-1",
      "sourceIPAddress": "72.21.196.64",
      "userAgent": "aws-cli/1.5.2 Python/2.7.5 Darwin/13.4.0",
      "requestParameters": {
        "name": "testpipeline",
        "uniqueId": "sounique"
      },
      "responseElements": {
        "pipelineId": "df-06372391ZG65EXAMPLE"
      },
      "requestID": "65cbf1e8-6b69-11e4-8816-cfcbadd04c45",
      "eventID": "9f99dce0-0864-49a0-bffa-f72287197758",
      "eventType": "AwsApiCall",
      "recipientAccountId": "role-account-id"
    },
      ...additional entries
  ]
}
```

AWS Data Pipeline のインシデントへの対応

AWS Data Pipeline のインシデントへの対応は、AWS の責任事項です。AWS には、インシデントへの対応を管理する正式な文書化されたポリシーとプログラムがあります。

広範な影響を与える AWS の運用上の問題は AWS Service Health Dashboard に投稿されます。運用上の問題は Personal Health Dashboard を介して個々のアカウントにも投稿されます。

AWS Data Pipeline のコンプライアンス検証

AWS Data Pipeline は AWS コンプライアンスプログラムの対象範囲ではありません。特定のコンプライアンスプログラムの範囲内の AWS のサービスのリストについては、「コンプライアンスプログラム対象範囲内の AWS のサービス」を参照してください。一般的な情報については、「AWS コンプライアンスプログラム」を参照してください。

AWS Data Pipeline での耐障害性

AWS のグローバルインフラストラクチャは AWS リージョンとアベイラビリティーゾーンを中心に構築されます。AWSリージョンには、低レイテンシー、高いスループット、そして高度の冗長ネットワークで Connect されている複数の物理的に独立・隔離されたアベイラビリティーゾーンがあります。アベイラビリティーゾーンでは、ゾーン間で中断することなく自動的にフェイルオーバーするアプリケーションとデータベースを設計および運用することができます。アベイラビリティーゾーンは、従来の単一または複数のデータセンターインフラストラクチャよりも可用性、耐障害性、および拡張性が優れています。

AWS リージョンとアベイラビリティーゾーンの詳細については、「AWS グローバルインフラストラクチャ」を参照してください。

AWS Data Pipeline でのインフラストラクチャセキュリティ

マネージドサービスである AWS Data Pipeline は、Amazon Web Services のセキュリティプロセスの概要ホワイトペーパーに記載されている AWS グローバルネットワークセキュリティの手順で保護されています。

AWS 公開版 API コールを使用して、ネットワーク経由で AWS Data Pipeline にアクセスします。クライアントで Transport Layer Security (TLS) 1.0 以降がサポートされている必要があります。TLS 1.2 以降が推奨されています。また、Ephemeral Diffie-Hellman (DHE) や Elliptic Curve Ephemeral Diffie-Hellman (ECDHE) などの Perfect Forward Secrecy (PFS) を使用した暗号スイートもクライアントでサポートされている必要があります。これらのモードは、Java 7 以降など、最近のほとんどのシステムでサポートされています。

また、リクエストは、アクセスキー ID と、IAM プリンシパルに関連付けられているシークレットアクセスキーを使用して署名する必要があります。または、AWS Security Token Service (AWS STS) を使用して、一時的なセキュリティ認証情報を生成し、リクエストに署名することもできます。

AWS Data Pipeline での設定と脆弱性の分析

構成および IT 管理は、AWS とお客様の間で共有される責任です。詳細については、AWS 責任共有モデルを参照してください。

チュートリアル

以下のチュートリアルは、AWS Data Pipeline でパイプラインを作成して使用する処理について順を追って説明しています。

チュートリアル

- Hadoop Streaming EMRで Amazon を使用してデータを処理
- AWS Data Pipeline を使用した Amazon S3 バケット間での CSV データのコピー
- AWS Data Pipeline を使用した Amazon S3 への MySQL データのエクスポート
- AWS Data Pipeline を使用した Amazon Redshift へのデータのコピー

Hadoop Streaming EMRで Amazon を使用してデータを処理

AWS Data Pipeline を使用して Amazon EMRクラスターを管理できます。 AWS Data Pipeline を使用すると、クラスターの起動前に満たす必要がある前提条件 (例えば、今日のデータが Amazon S3 にアップロードされていることを確認する) 、クラスターを繰り返し実行するためのスケジュール、使用するクラスター設定を指定できます。以下のチュートリアルでは、簡単なクラスターの起動について順を追って説明します。

このチュートリアルでは、シンプルな Amazon EMRクラスターのパイプラインを作成して、Amazon が提供する既存の Hadoop Streaming ジョブを実行しEMR、タスクが正常に完了した後に Amazon SNS通知を送信します。このタスクには、 が提供する Amazon AWS Data Pipeline EMRクラスターリソースを使用します。サンプルアプリケーションは と呼ばれ WordCount、Amazon EMRコンソールから手動で実行することもできます。ユーザー AWS Data Pipeline に代わって によって生成されたクラスターは Amazon EMRコンソールに表示され、AWSアカウントに請求されることに注意してください。

パイプラインオブジェクト

このパイプラインでは以下のオブジェクトを使用します。

EmrActivity

パイプラインで実行する作業を定義します (Amazon が提供する既存の Hadoop Streaming ジョブを実行しますEMR) 。

EmrCluster

リソースはこのアクティビティを実行するために AWS Data Pipeline を使用します。

クラスターは Amazon EC2 instances. AWS Data Pipeline のセットです。 はクラスターを起動し、タスクの完了後にクラスターを終了します。

スケジュール

このアクティビティの開始日、時刻、および期間。オプションで終了日時を指定できます。

SnsAlarm

タスクが正常に終了した後に、指定したトピックに Amazon SNS通知を送信します。

内容

- 開始する前に
- コマンドラインを使用してクラスターを起動する

開始する前に

次の手順を完了したことを確認してください。

- のセットアップ AWS Data Pipeline の各タスクを完了する。
- (オプション) クラスターVPCの と のセキュリティグループを設定しますVPC。
- E メール通知を送信するためのトピックを作成し、トピック Amazon リソースネーム () を書き留めますARN。詳細については、Amazon Simple Notification Service 入門ガイドの「トピックの作成」を参照してください。

コマンドラインを使用してクラスターを起動する

Amazon EMRクラスターを定期的に実行してウェブログの分析や科学データの分析を行う場合は、 AWS Data Pipeline を使用して Amazon EMRクラスターを管理できます。では AWS Data Pipeline、クラスターの起動前に満たす必要がある前提条件を指定できます (例えば、今日のデータが Amazon S3 にアップロードされていることを確認するなど)。このチュートリアルでは、シンプルな Amazon EMRベースのパイプラインのモデルとして、またはより複雑なパイプラインの一部としてクラスターを起動する手順を説明します。

前提条件

を使用する前にCLI、次のステップを完了する必要があります。

1. コマンドラインインターフェイス () をインストールして設定しますCLI。詳細については、「アクセス AWS Data Pipeline」を参照してください。
2. DataPipelineDefaultRole および という名前のIAMロールDataPipelineDefaultResourceRoleが存在することを確認します。 AWS Data Pipeline コンソールによって、これらのロールが自動的に作成されます。 AWS Data Pipeline コンソールを一度も使用したことがない場合は、これらのロールを手動で作成する必要があります。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。

タスク

- パイプライン定義ファイルの作成
- パイプライン定義のアップロードとアクティブ化
- パイプライン実行の監視

パイプライン定義ファイルの作成

次のコードは、Amazon が提供する既存の Hadoop ストリーミングジョブを実行するシンプルな Amazon EMRクラスターのパイプライン定義ファイルですEMR。このサンプルアプリケーションは と呼ばれ WordCount、Amazon EMRコンソールを使用して実行することもできます。

このコードをテキストファイルにコピーし、`MyEmrPipelineDefinition.json` として保存します。Amazon S3 バケットの場所は、所有している Amazon S3 バケットの名前に置き換える必要があります。また、開始日および終了日も置き換える必要があります。クラスターをすぐに起動するには、 `startDateTime`を 1 日前の日付に設定し、 `endDateTime`を 1 日先の日付に設定します。AWS Data Pipeline その後、 は、作業のバックログとして認識されるものに対処しようとして、「期日を過ぎた」クラスターをすぐに起動します。このバックフィルは、最初のクラスターの AWS Data Pipeline 起動を 1 時間待つ必要がないことを意味します。

```
{
  "objects": [
    {
      "id": "Hourly",
      "type": "Schedule",
      "startDateTime": "2012-11-19T07:48:00",
```

```
      "endDateTime": "2012-11-21T07:48:00",
      "period": "1 hours"
    },
    {
      "id": "MyCluster",
      "type": "EmrCluster",
      "masterInstanceType": "m1.small",
      "schedule": {
        "ref": "Hourly"
      }
    },
    {
      "id": "MyEmrActivity",
      "type": "EmrActivity",
      "schedule": {
        "ref": "Hourly"
      },
      "runsOn": {
        "ref": "MyCluster"
      },
      "step": "/home/hadoop/contrib/streaming/hadoop-streaming.jar,-input,s3n://
elasticmapreduce/samples/wordcount/input,-output,s3://myawsbucket/wordcount/
output/#{@scheduledStartTime},-mapper,s3n://elasticmapreduce/samples/wordcount/
wordSplitter.py,-reducer,aggregate"
    }
  ]
}
```

このパイプラインには 3 個のオブジェクトがあります。

- Hourly。作業のスケジュールを表します。アクティビティのフィールドの 1 つとしてスケジュールを設定できます。スケジュールを設定すると、アクティビティはそのスケジュールに従って (この例では 1 時間ごとに) 実行されます。
- MyCluster。クラスターの実行に使用される Amazon EC2インスタンスのセットを表します。クラスターとして実行するEC2インスタンスのサイズと数を指定できます。インスタンスの数を指定しなかった場合、クラスターはマスターノードとタスクノードの 2 つを使用して起動されます。クラスターを起動するサブネットも指定できます。Amazon EMRが提供する に追加のソフトウェアをロードするブートストラップアクションなど、クラスターに追加の設定を追加できますAMI。
- MyEmrActivity。クラスタで処理する計算を表します。Amazon は、ストリーミング、カスケード、スクリプト化された Hive など、いくつかのタイプのクラスターEMRをサポートしています。runsOn フィールドは を参照し MyCluster、クラスターの基盤となる仕様として使用します。

パイプライン定義のアップロードとアクティブ化

パイプライン定義をアップロードし、パイプラインをアクティブ化する必要があります。次のコマンド例では、*pipeline_name* パイプラインのラベルと *pipeline_file* パイプライン定義.jsonファイルの完全修飾パスを持つ 。

AWS CLI

パイプライン定義を作成してパイプラインをアクティブ化するには、以下の create-pipeline コマンドを使用します。ほとんどのCLIコマンドでこの値を使用するため、パイプラインの ID を書き留めます。

```
aws datapipeline create-pipeline --name pipeline_name --unique-id token
{
    "pipelineId": "df-00627471SOVYZEXAMPLE"
}
```

パイプライン定義をアップロードするには、次のput-pipeline-definitionコマンドを使用します。

```
aws datapipeline put-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE --
pipeline-definition file://MyEmrPipelineDefinition.json
```

パイプラインが正常に検証された場合、validationErrors フィールドは空です。警告を確認する必要があります。

パイプラインをアクティブ化するには、以下の activate-pipeline コマンドを使用します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

以下の list-pipelines コマンドを使用して、パイプラインリストにパイプラインが表示されていることを確認できます。

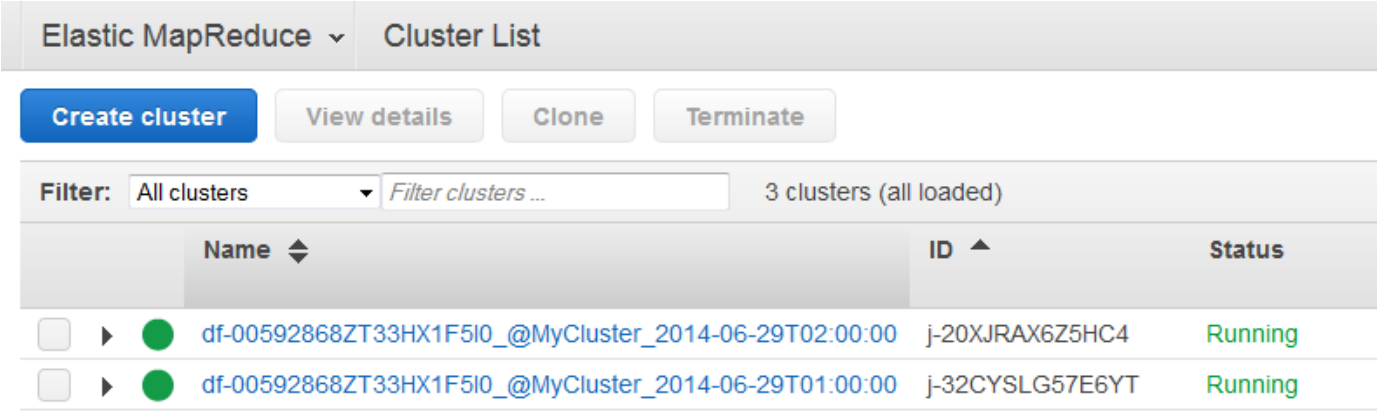
```
aws datapipeline list-pipelines
```

パイプライン実行の監視

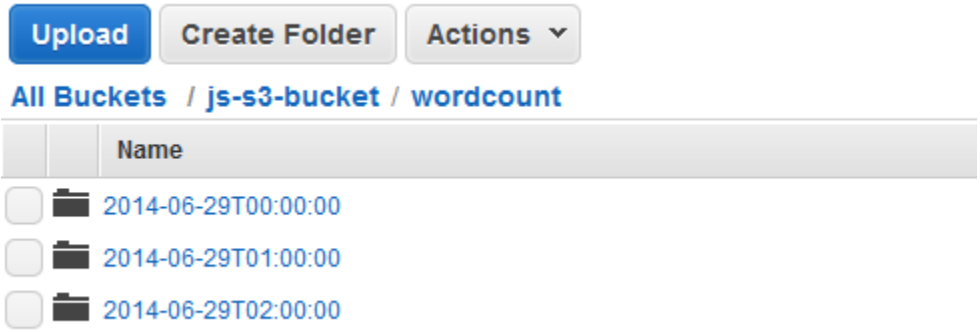
Amazon EMRコンソール AWS Data Pipeline を使用して起動したクラスターを表示したり、Amazon S3 コンソールを使用して出力フォルダを表示したりできます。

によって起動されたクラスターの進行状況を確認するには AWS Data Pipeline

1. Amazon EMRコンソールを開きます。
2. によって生成されたクラスター AWS Data Pipeline の名前の形式は次のとおりです。*<pipeline-identifier>*_@*<emr-cluster-name>*_*<launch-time>*.



3. いずれかの実行が完了した後、Amazon S3 コンソールを開いて、タイムスタンプ付きの出力フォルダーがあり、クラスターの結果が予想したとおり含まれていることを確認します。



AWS Data Pipeline を使用した Amazon S3 バケット間での CSV データのコピー

「とは AWS Data Pipeline」を読み、AWS Data Pipeline を使用してデータの移動と変換を自動化することを決定したら、データパイプラインの作成を開始します。AWS Data Pipeline での処理の意味を理解しやすくするために、単純なタスクを使って手順を説明します。

このチュートリアルでは、Amazon S3 バケット間でデータをコピーし、コピーアクティビティが正常に完了した後、Amazon SNS 通知を送信するデータパイプラインを作成するプロセスについて順

を追って説明します。このコピーアクティビティでは、AWS Data Pipeline によって管理される EC2 インスタンスを使用します。

パイプラインオブジェクト

このパイプラインでは以下のオブジェクトを使用します。

CopyActivity

このパイプラインのために AWS Data Pipeline が実行するアクティビティ (Amazon S3 バケット間での CSV データのコピー)。

> **Important**
>
> CopyActivity および S3DataNode での CSV ファイル形式の使用には、制限事項があります。詳細については、「CopyActivity」を参照してください。

スケジュール

このアクティビティの開始日、時刻、および繰り返し。オプションで終了日時を指定できます。

Ec2Resource

このアクティビティを実行するために AWS Data Pipeline が使用するリソース (EC2 インスタンス) 。

S3DataNode

このパイプラインの入力ノードと出力ノード (Amazon S3 バケット)。

SnsAlarm

指定された条件が満たされたときに AWS Data Pipeline が実行する必要があるアクション (タスクが正常に終了した後、トピックに Amazon SNS 通知を送信する)。

目次

- 開始する前に
- コマンドラインを使用して CSV データをコピーする

開始する前に

次の手順を完了したことを確認してください。

- のセットアップ AWS Data Pipeline の各タスクを完了する。
- （オプション）インスタンス用の VPC をセットアップし、この VPC 用のセキュリティグループをセットアップします。
- データソースとして Amazon S3 バケットを作成します。

  詳細については、Amazon Simple Storage Service ユーザーガイドの「バケットの作成」を参照してください。
- データを Amazon S3 バケットにアップロードします。

  詳細については、Amazon Simple Storage Service ユーザーガイドのバケットへのオブジェクトの追加を参照してください。
- データターゲットとして別の Amazon S3 バケットを作成します。
- E メール通知を送信するためのトピックを作成し、トピックの Amazon リソースネーム（ARN）をメモしておきます。詳細については、Amazon Simple Notification Service 入門ガイドの「トピックの作成」を参照してください。
- （オプション）このチュートリアルでは、AWS Data Pipeline によって作成されたデフォルトの IAM ロールポリシーを使用します。独自の IAM ロールポリシーと信頼関係を作成して設定する場合は、AWS Data Pipeline の IAM ロールで説明されている手順に従います。

コマンドラインを使用して CSV データをコピーする

パイプラインを作成し、これを使用して、Amazon S3 バケット間でデータをコピーできます。

前提条件

開始する前に、次のステップを完了しておく必要があります。

1. コマンドラインインターフェイス (CLI) をインストールして設定します。詳細については、「アクセス AWS Data Pipeline」を参照してください。
2. DataPipelineDefaultRole と DataPipelineDefaultResourceRole という名前の IAM ロールが存在していることを確認します。AWS Data Pipeline コンソールにより、自動的にこれらのロールが作成されます。AWS Data Pipeline コンソールをまだ 1 回も使用したことがない場合、これらの

ロールを手動で作成する必要があります。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。

タスク

- JSON 形式でパイプラインを定義する
- パイプライン定義をアップロードし、アクティブ化する

JSON 形式でパイプラインを定義する

この例のシナリオでは、JSON パイプライン定義と AWS Data Pipeline CLI を使用して、2 つの Amazon S3 バケット間でデータを特定の間隔でコピーするようにスケジュールする方法を説明します。これは、完全なパイプライン定義の JSON ファイルであり、その後に各セクションの説明を示します。

> **Note**
>
> JSON 形式のファイルの構文を検証できるテキストエディタを使用し、.json ファイル拡張子を使用してファイルに名前を付けることをお勧めします。

この例では、わかりやすくするために、オプションフィールドを省略し、必須フィールドのみを示しています。この例の完全なパイプラインの JSON ファイルは次のようになります。

```
{
  "objects": [
    {
      "id": "MySchedule",
      "type": "Schedule",
      "startDateTime": "2013-08-18T00:00:00",
      "endDateTime": "2013-08-19T00:00:00",
      "period": "1 day"
    },
    {
      "id": "S3Input",
      "type": "S3DataNode",
      "schedule": {
        "ref": "MySchedule"
      },
      "filePath": "s3://example-bucket/source/inputfile.csv"
```

```
    },
    {
      "id": "S3Output",
      "type": "S3DataNode",
      "schedule": {
        "ref": "MySchedule"
      },
      "filePath": "s3://example-bucket/destination/outputfile.csv"
    },
    {
      "id": "MyEC2Resource",
      "type": "Ec2Resource",
      "schedule": {
        "ref": "MySchedule"
      },
      "instanceType": "m1.medium",
      "role": "DataPipelineDefaultRole",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "id": "MyCopyActivity",
      "type": "CopyActivity",
      "runsOn": {
        "ref": "MyEC2Resource"
      },
      "input": {
        "ref": "S3Input"
      },
      "output": {
        "ref": "S3Output"
      },
      "schedule": {
        "ref": "MySchedule"
      }
    }
  ]
}
```

スケジュール

パイプラインでは、このパイプラインのアクティビティを実行する頻度を決定する期間と共に、開始日と終了日を持つスケジュールを定義します。

```
{
  "id": "MySchedule",
  "type": "Schedule",
  "startDateTime": "2013-08-18T00:00:00",
  "endDateTime": "2013-08-19T00:00:00",
  "period": "1 day"
},
```

Amazon S3 データノード

次に、入力 S3DataNode パイプラインコンポーネントによって、入力ファイルの場所を定義します。この例では、Amazon S3 バケットの場所です。入力 S3DataNode コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "S3Input",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "filePath": "s3://example-bucket/source/inputfile.csv"
},
```

ID

入力場所のユーザー定義名 (参照字にのみ使用されるラベル) 。

タイプ

パイプラインコンポーネントの型。Amazon S3 バケット内のデータが存在する場所と一致する "S3DataNode" です。

スケジュール

JSON ファイルの先行部分で作成したスケジュールコンポーネント "MySchedule" 参照します。

[Path] (パス)

データノードに関連付けられたデータへのパス。データノードの構文はそのタイプによって決まります。例えば、Amazon S3 パスの構文は、データベーステーブルに適した構文とは別の構文に従います。

次に、出力 S3DataNode コンポーネントで、データの出力先の場所を定義します。これは、入力 S3DataNode コンポーネントと同じ形式に従います。ただし、コンポーネントの名前とターゲットファイルを指定するパスは異なります。

```
{
  "id": "S3Output",
  "type": "S3DataNode",
  "schedule": {
    "ref": "MySchedule"
  },
  "filePath": "s3://example-bucket/destination/outputfile.csv"
},
```

リソース

これは、コピー操作を実行するコンピューティングリソースの定義です。この例では、AWS Data Pipeline は、コピータスクを実行するための EC2 インスタンスを自動的に作成し、コピータスクが完了するとリソースを終了します。ここで定義されているフィールドが、作業を行う EC2 インスタンスの作成と機能を制御します。EC2Resource は、次のフィールドで定義されます。

```
{
  "id": "MyEC2Resource",
  "type": "Ec2Resource",
  "schedule": {
    "ref": "MySchedule"
  },
  "instanceType": "m1.medium",
  "role": "DataPipelineDefaultRole",
  "resourceRole": "DataPipelineDefaultResourceRole"
},
```

ID

パイプラインスケジュールのユーザー定義の名前。これは参照時にのみ使用されるラベルです。

タイプ

作業を実行するコンピューティングリソースの種類。この例では、 EC2 インスタンス。EmrCluster タイプなど、その他のリソースタイプも使用できます。

スケジュール

このコンピューティングリソースを作成するスケジュール。

instanceType

作成する EC2 インスタンスのサイズ。AWS Data Pipeline で実行する作業の負荷に最適な EC2 インスタンスのサイズを設定します。ここでは、m1.medium の EC2 インスタンスを設定します。さまざまなインスタンスタイプとそれぞれの用途の詳細については、Amazon EC2 インスタンスタイプに関するトピック (http://aws.amazon.com/ec2/instance-types/) を参照してください。

ロール

リソースにアクセスするアカウントの IAM ロール (データを取得するための Amazon S3 バケットへのアクセスなど)。

resourceRole

リソースを作成するアカウントの IAM ロール (お客様に代わって EC2 インスタンスを作成および設定するなど) 。Role と ResourceRole は同じロールにすることもできますが、個別に設定することによって、セキュリティ設定での詳細度が向上します。

アクティビティ

この JSON ファイルの最後のセクションは、実行する作業を表すアクティビティの定義です。この例では、CopyActivity を使用して、http://aws.amazon.com/ec2/instance-types/ のバケット間で CSV ファイルのデータをコピーします。CopyActivity コンポーネントは、以下のフィールドで定義されます。

```
{
  "id": "MyCopyActivity",
  "type": "CopyActivity",
  "runsOn": {
    "ref": "MyEC2Resource"
  },
  "input": {
    "ref": "S3Input"
  },
  "output": {
    "ref": "S3Output"
  },
  "schedule": {
    "ref": "MySchedule"
  }
}
```

ID

アクティビティのユーザー定義の名前。これは参照時にのみ使用されるラベルです。

タイプ

MyCopyActivity など、実行するアクティビティのタイプ。

runsOn

このアクティビティが定義する作業を実行するコンピューティングリソース。この例では、先に定義した EC2 インスタンスへの参照を指定します。`runsOn` フィールドを使用すると、AWS Data Pipeline が EC2 インスタンスを自動的に作成します。`runsOn` フィールドは、リソースが AWS インフラストラクチャに存在することを示し、一方、`workerGroup` 値は、独自のオンプレミスリソースを使用して作業を実行することを示しています。

入力

コピーするデータの場所。

出力

ターゲットデータの場所。

スケジュール

このアクティビティを実行するスケジュール。

パイプライン定義をアップロードし、アクティブ化する

パイプライン定義をアップロードし、パイプラインをアクティブ化する必要があります。以下のコマンド例では、*`pipeline_name`* をパイプラインのラベルに置き換え、*`pipeline_file`* をパイプライン定義 `.json` ファイルの完全修飾パスに置き換えます。

AWS CLI

パイプライン定義を作成してパイプラインをアクティブ化するには、以下の create-pipeline コマンドを使用します。パイプラインの ID をメモします。この値は、ほとんどの CLI コマンドで使用するからです。

```
aws datapipeline create-pipeline --name pipeline_name --unique-id token
{
```

```
    "pipelineId": "df-00627471SOVYZEXAMPLE"
}
```

パイプライン定義を更新するには、以下の put-pipeline-definition コマンドを使用します。

```
aws datapipeline put-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE --
pipeline-definition file://MyEmrPipelineDefinition.json
```

パイプラインが正常に検証された場合、validationErrors フィールドは空です。警告を確認する必要があります。

パイプラインをアクティブ化するには、以下の activate-pipeline コマンドを使用します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

以下の list-pipelines コマンドを使用して、パイプラインリストにパイプラインが表示されていることを確認できます。

```
aws datapipeline list-pipelines
```

AWS Data Pipeline を使用した Amazon S3 への MySQL データのエクスポート

このチュートリアルでは、データパイプラインを作成して、MySQL データベースのテーブルからデータ (行) を Amazon S3 バケットの CSV (Comma Separated Value) ファイルにコピーし、コピー作業が正常に完了した後、Amazon SNS 通知を送信する処理について順を追って説明します。このコピーアクティビティでは、AWS Data Pipeline から提供される EC2 インスタンスを使用します。

パイプラインオブジェクト

このパイプラインでは以下のオブジェクトを使用します。

- CopyActivity
- Ec2Resource
- MySqlDataNode
- S3DataNode

- SnsAlarm

目次

- 開始する前に
- コマンドラインを使用して MySQL データをコピーする

開始する前に

次の手順を完了したことを確認してください。

- のセットアップ AWS Data Pipeline の各タスクを完了する。
- （オプション）インスタンス用の VPC をセットアップし、この VPC 用のセキュリティグループをセットアップします。
- データ出力として Amazon S3 バケットを作成します。

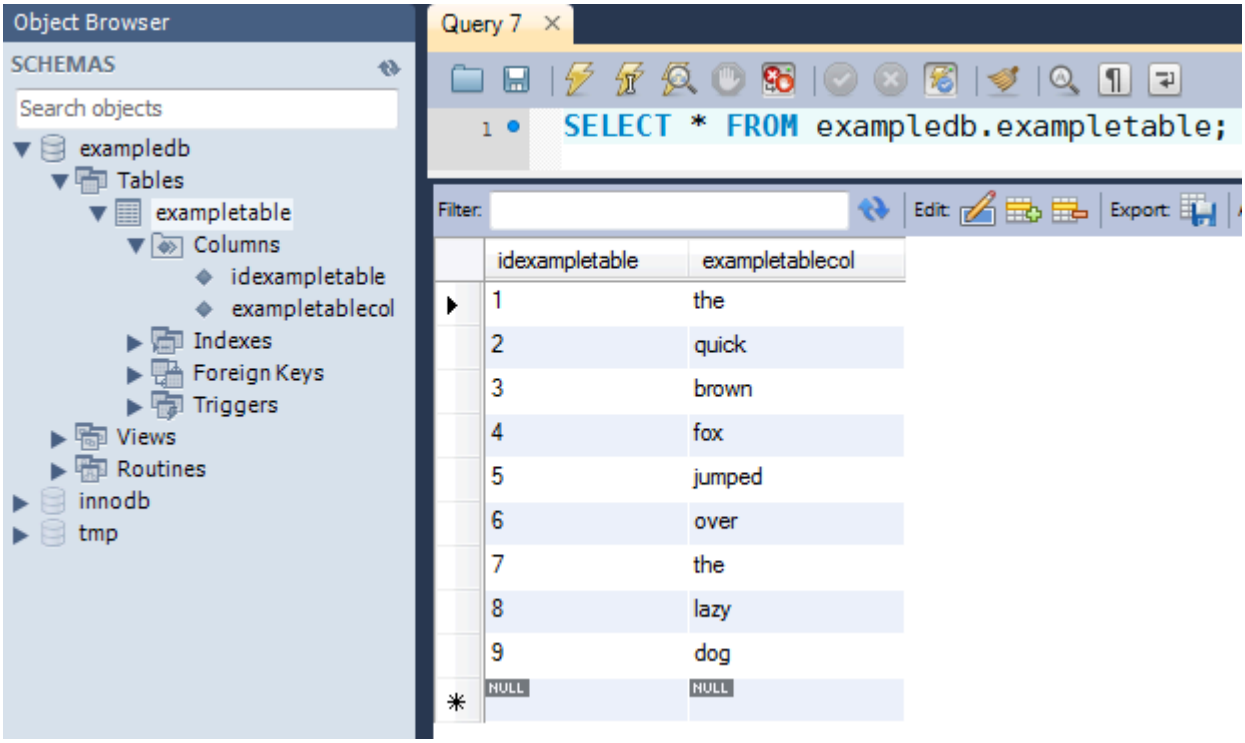
  詳細については、Amazon Simple Storage Service ユーザーガイドのバケットの作成を参照してください。
- データソースとして MySQL データベースインスタンスを作成し起動します。

  詳細については、Amazon RDS 入門ガイドの Launch a DB Instance を参照してください。Amazon RDS インスタンスを起動したら、MySQL のドキュメントのテーブルの作成を参照してください。

  > **Note**
  >
  > MySQL インスタンスの作成時に使用したユーザー名とパスワードを書き留めてください。MySQL データベースインスタンスを起動した後、インスタンスのエンドポイントを書き留めてください。この情報は後で必要になります。

- MySQL データベースインスタンスに接続し、テーブルを作成して、新しく作成したテーブルにテストデータの値を追加します。

  実例として、このチュートリアルでは、以下の設定およびサンプルデータの MySQL テーブルを使用しています。次の画面は、MySQL Workbench 5.2 CE のものです。



詳細については、MySQL ドキュメントの「テーブルの作成」と、MySQL Workbench 製品ページを参照してください。

- E メール通知を送信するためのトピックを作成し、トピックの Amazon リソースネーム（ARN）をメモしておきます。詳細については、Amazon Simple Notification Service 入門ガイドのトピックの作成を参照してください。
- （オプション）このチュートリアルでは、AWS Data Pipeline によって作成されたデフォルトの IAM ロールポリシーを使用します。代わりに IAM ロールポリシーと信頼関係を作成して設定する場合は、AWS Data Pipeline の IAM ロールで説明されている手順に従います。

コマンドラインを使用して MySQL データをコピーする

MySQL のテーブルから Amazon S3 バケットのファイルにデータをコピーするパイプラインを作成することができます。

前提条件

開始する前に、次のステップを完了しておく必要があります。

1. コマンドラインインターフェイス (CLI) をインストールして設定します。詳細については、「アクセス AWS Data Pipeline」を参照してください。

2. DataPipelineDefaultRole と DataPipelineDefaultResourceRole という名前の IAM ロールが存在していることを確認します。AWS Data Pipeline コンソールにより、自動的にこれらのロールが作成されます。AWS Data Pipeline コンソールをまだ 1 回も使用したことがない場合、これらのロールを手動で作成する必要があります。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。
3. Amazon S3 バケットと Amazon RDS インスタンスを設定します。詳細については、「開始する前に」を参照してください。

タスク

- JSON 形式でパイプラインを定義する
- パイプライン定義をアップロードし、アクティブ化する

JSON 形式でパイプラインを定義する

この例では、JSON パイプライン定義と AWS Data Pipeline CLI を使用して、指定した時間間隔で MySQL データベーステーブルのデータ (行) を Amazon S3 バケットの CSV (Comma Separated Value) ファイルにコピーする方法を示します。

これは、完全なパイプライン定義の JSON ファイルであり、その後に各セクションの説明を示します。

> **Note**
>
> JSON 形式のファイルの構文を検証できるテキストエディタを使用し、.json ファイル拡張子を使用してファイルに名前を付けることをお勧めします。

```
{
  "objects": [
    {
      "id": "ScheduleId113",
      "startDateTime": "2013-08-26T00:00:00",
      "name": "My Copy Schedule",
      "type": "Schedule",
      "period": "1 Days"
    },
    {
      "id": "CopyActivityId112",
```

```
      "input": {
        "ref": "MySqlDataNodeId115"
      },
      "schedule": {
        "ref": "ScheduleId113"
      },
      "name": "My Copy",
      "runsOn": {
        "ref": "Ec2ResourceId116"
      },
      "onSuccess": {
        "ref": "ActionId1"
      },
      "onFail": {
        "ref": "SnsAlarmId117"
      },
      "output": {
        "ref": "S3DataNodeId114"
      },
      "type": "CopyActivity"
    },
    {
      "id": "S3DataNodeId114",
      "schedule": {
        "ref": "ScheduleId113"
      },
      "filePath": "s3://example-bucket/rds-output/output.csv",
      "name": "My S3 Data",
      "type": "S3DataNode"
    },
    {
      "id": "MySqlDataNodeId115",
      "username": "my-username",
      "schedule": {
        "ref": "ScheduleId113"
      },
      "name": "My RDS Data",
      "*password": "my-password",
      "table": "table-name",
      "connectionString": "jdbc:mysql://your-sql-instance-name.id.region-
name.rds.amazonaws.com:3306/database-name",
      "selectQuery": "select * from #{table}",
      "type": "SqlDataNode"
    },
```

```
    {
      "id": "Ec2ResourceId116",
      "schedule": {
        "ref": "ScheduleId113"
      },
      "name": "My EC2 Resource",
      "role": "DataPipelineDefaultRole",
      "type": "Ec2Resource",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "message": "This is a success message.",
      "id": "ActionId1",
      "subject": "RDS to S3 copy succeeded!",
      "name": "My Success Alarm",
      "role": "DataPipelineDefaultRole",
      "topicArn": "arn:aws:sns:us-east-1:123456789012:example-topic",
      "type": "SnsAlarm"
    },
    {
      "id": "Default",
      "scheduleType": "timeseries",
      "failureAndRerunMode": "CASCADE",
      "name": "Default",
      "role": "DataPipelineDefaultRole",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "message": "There was a problem executing #{node.name} at for period
 #{node.@scheduledStartTime} to #{node.@scheduledEndTime}",
      "id": "SnsAlarmId117",
      "subject": "RDS to S3 copy failed",
      "name": "My Failure Alarm",
      "role": "DataPipelineDefaultRole",
      "topicArn": "arn:aws:sns:us-east-1:123456789012:example-topic",
      "type": "SnsAlarm"
    }
  ]
}
```

MySQL データノード

入力 MySqlDataNode パイプラインコンポーネントは、入力データの場所を定義します。この例では、Amazon RDS インスタンスです。入力 MySqlDataNode コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "MySqlDataNodeId115",
  "username": "my-username",
  "schedule": {
    "ref": "ScheduleId113"
  },
  "name": "My RDS Data",
  "*password": "my-password",
  "table": "table-name",
  "connectionString": "jdbc:mysql://your-sql-instance-name.id.region-
name.rds.amazonaws.com:3306/database-name",
  "selectQuery": "select * from #{table}",
  "type": "SqlDataNode"
},
```

ID

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

ユーザーネーム

データベーステーブルからデータを取得するのに十分なアクセス許可を持つデータベースアカウントのユーザー名。*my-username* をユーザーアカウントの名前に置き換えます。

スケジュール

JSON ファイルの先行部分で作成したスケジュールコンポーネントを参照します。

名前

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

*Password

データベースアカウントのパスワード。先頭のアスタリスク (*) は、AWS Data Pipeline がパスワード値を暗号化する必要があることを示します。*my-password*をユーザーの正しいパスワードに置き換えます。パスワードフィールド名の先頭には、特殊文字のアスタリスク (*) が付きます。詳細については、「特殊文字」を参照してください。

テーブル

コピーするデータを含むデータベーステーブルの名前。*table-name* をデータベーステーブルの名前に置き換えます。

connectionString

データベースに接続する CopyActivity オブジェクト用の JDBC 接続文字列。

selectQuery

データベーステーブルからコピーするデータを指定する有効な SQL の SELECT クエリ。#{table} は、JSON ファイルの先行部分の table 変数によって指定されたテーブル名を再利用する式です。

タイプ

SqlDataNode 型。この例の場合、これは MySQL を使用する Amazon RDS インスタンスです。

> **Note**
>
> MySqlDataNode 型は廃止されました。現時点では MySqlDataNode も使用できますが、SqlDataNode の使用をお勧めします。

Amazon S3 データノード

次に、S3Output パイプラインコンポーネントで出力ファイルの場所を定義します。出力ファイルは、この例の場合、Amazon S3 バケットの場所にある CSV ファイルです。出力 S3DataNode コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "S3DataNodeId114",
  "schedule": {
    "ref": "ScheduleId113"
  },
  "filePath": "s3://example-bucket/rds-output/output.csv",
  "name": "My S3 Data",
  "type": "S3DataNode"
},
```

ID

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

スケジュール

JSON ファイルの先行部分で作成したスケジュールコンポーネントを参照します。

filePath

データノードに関連付けられているデータへのパス。この例では CSV 出力ファイルです。

名前

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

タイプ

パイプラインオブジェクトの型。Amazon S3 バケット内のデータが存在する場所と一致する S3DataNode です。

リソース

これは、コピー操作を実行するコンピューティングリソースの定義です。この例では、AWS Data Pipeline は、コピータスクを実行するための EC2 インスタンスを自動的に作成し、コピータスクが完了するとリソースを終了します。ここで定義されているフィールドが、作業を行う EC2 インスタンスの作成と機能を制御します。EC2Resource は、次のフィールドで定義されます。

```
{
  "id": "Ec2ResourceId116",
  "schedule": {
    "ref": "ScheduleId113"
  },
  "name": "My EC2 Resource",
  "role": "DataPipelineDefaultRole",
  "type": "Ec2Resource",
  "resourceRole": "DataPipelineDefaultResourceRole"
},
```

ID

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

スケジュール

このコンピューティングリソースを作成するスケジュール。

名前

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

ロール

リソースにアクセスするアカウントの IAM ロール (データを取得するための Amazon S3 バケットへのアクセスなど)。

タイプ

作業を実行するコンピューティングリソースの種類。この例では、 EC2 インスタンス。EmrCluster タイプなど、その他のリソースタイプも使用できます。

resourceRole

リソースを作成するアカウントの IAM ロール (お客様に代わって EC2 インスタンスを作成および設定するなど) 。Role と ResourceRole は同じロールにすることもできますが、個別に設定することによって、セキュリティ設定での詳細度が向上します。

アクティビティ

この JSON ファイルの最後のセクションは、実行する作業を表すアクティビティの定義です。この例では、CopyActivity コンポーネントを使用して、Amazon S3 バケットのファイルから別のファイルにデータをコピーします。CopyActivity コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "CopyActivityId112",
  "input": {
    "ref": "MySqlDataNodeId115"
  },
  "schedule": {
    "ref": "ScheduleId113"
  },
  "name": "My Copy",
  "runsOn": {
    "ref": "Ec2ResourceId116"
  },
  "onSuccess": {
    "ref": "ActionId1"
  },
  "onFail": {
    "ref": "SnsAlarmId117"
  },
  "output": {
    "ref": "S3DataNodeId114"
  },
```

```
  "type": "CopyActivity"
},
```

ID

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

入力

コピーする MySQL データの場所。

スケジュール

このアクティビティを実行するスケジュール。

名前

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

runsOn

このアクティビティが定義する作業を実行するコンピューティングリソース。この例では、先に定義した EC2 インスタンスへの参照を指定します。runs0n フィールドを使用すると、AWS Data Pipeline が EC2 インスタンスを自動的に作成します。runs0n フィールドは、リソースが AWS インフラストラクチャに存在することを示し、一方、workerGroup 値は、独自のオンプレミスリソースを使用して作業を実行することを示しています。

onSuccess

アクティビティが正常終了した場合に送信する SnsAlarm。

onFail

アクティビティが失敗した場合に送信する SnsAlarm。

出力

CSV 出力ファイルの Amazon S3 での場所。

タイプ

実行するアクティビティのタイプ。

パイプライン定義をアップロードし、アクティブ化する

パイプライン定義をアップロードし、パイプラインをアクティブ化する必要があります。以下のコマンド例では、*pipeline_name* をパイプラインのラベルに置き換え、*pipeline_file* をパイプライン定義 .json ファイルの完全修飾パスに置き換えます。

AWS CLI

パイプライン定義を作成してパイプラインをアクティブ化するには、以下の create-pipeline コマンドを使用します。パイプラインの ID をメモします。この値は、ほとんどの CLI コマンドで使用するからです。

```
aws datapipeline create-pipeline --name pipeline_name --unique-id token
{
    "pipelineId": "df-00627471SOVYZEXAMPLE"
}
```

パイプライン定義を更新するには、以下の put-pipeline-definition コマンドを使用します。

```
aws datapipeline put-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE --
pipeline-definition file://MyEmrPipelineDefinition.json
```

パイプラインが正常に検証された場合、validationErrors フィールドは空です。警告を確認する必要があります。

パイプラインをアクティブ化するには、以下の activate-pipeline コマンドを使用します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

以下の list-pipelines コマンドを使用して、パイプラインリストにパイプラインが表示されていることを確認できます。

```
aws datapipeline list-pipelines
```

AWS Data Pipeline を使用した Amazon Redshift へのデータのコピー

このチュートリアルでは、AWS Data Pipeline コンソールで [Copy to Redshift] (Redshift へのコピー) テンプレートを使用するか、AWS Data Pipeline CLI でパイプライン定義ファイルを使用し

て、Amazon S3 から Amazon Redshift にデータを定期的に移動するパイプラインを作成するプロセスについて順を追って説明します。

Amazon S3 は、クラウドにデータを保存できるウェブサービスです。詳細については、Amazon Simple Storage Serviceユーザーガイドを参照してください。

Amazon Redshift は、クラウド内のデータウェアハウスサービスです。詳細については、「Amazon Redshift管理ガイド」を参照してください。

このチュートリアルにはいくつかの前提条件があります。コンソールまたは CLI を使用してチュートリアルを続行する前に、以下のステップを完了している必要があります。

目次

- 開始する前に: COPY オプションの設定とデータのロード
- パイプラインのセットアップ、セキュリティグループの作成、および Amazon Redshift クラスターの作成
- コマンドラインを使用した Amazon Redshift へのデータのコピー

開始する前に: COPY オプションの設定とデータのロード

AWS Data Pipeline 内で Amazon Redshift にデータをコピーする前に、以下の点を確認します。

- Amazon S3 からデータをロードします。
- Amazon Redshift で COPY アクティビティを設定します。

これらのオプションが機能しておりデータのロードが正常に完了したら、これらのオプションを AWS Data Pipeline に転送して、そこでコピーを実行します。

COPY オプションについては、Amazon Redshift データベース開発者ガイドの「COPY」を参照してください。

Amazon S3 からデータをロードする手順については、Amazon Redshift データベース開発者ガイドの Amazon S3 からデータをロードするを参照してください。

例えば、Amazon Redshift で以下の SQL コマンドを実行すると、LISTING という名前の新しいテーブルが作成され、Amazon S3 の公開バケットからサンプルデータがコピーされます。

<iam-role-arn> およびリージョンを独自の値に置き換えます。

この例の詳細については、Amazon Redshift 入門ガイドの Amazon S3 のサンプルデータをロードするを参照してください。

```
create table listing(
 listid integer not null distkey,
 sellerid integer not null,
 eventid integer not null,
 dateid smallint not null  sortkey,
 numtickets smallint not null,
 priceperticket decimal(8,2),
 totalprice decimal(8,2),
 listtime timestamp);

copy listing from 's3://awssampledbuswest2/tickit/listings_pipe.txt'
credentials 'aws_iam_role=<iam-role-arn>'
delimiter '|' region 'us-west-2';
```

パイプラインのセットアップ、セキュリティグループの作成、および Amazon Redshift クラスターの作成

チュートリアル用のセットアップを行うには

1. のセットアップ AWS Data Pipeline の各タスクを完了する。
2. セキュリティグループを作成します。
   a. Amazon EC2 コンソールを開きます。
   b. ナビゲーションペインで、[Security Groups] をクリックします。
   c. [Create Security Group] をクリックします。
   d. セキュリティグループの名前と説明を指定します。
   e. [EC2-Classic] [VPC] に No VPC を選択します。
   f. [EC2-VPC] [VPC] で、VPC の ID を選択します。
   g. [Create] (作成) をクリックします。
3. [EC2-Classic] Amazon Redshift クラスターセキュリティグループを作成し、Amazon EC2 セキュリティグループを指定します。
   a. Amazon Redshift コンソールを開きます。
   b. ナビゲーションペインで、[Security Groups] をクリックします。

c. [Create Cluster Security Group] をクリックします。

d. [Create Cluster Security Group] ダイアログボックスでクラスターセキュリティグループの名前と説明を指定します。

e. 新しいクラスターセキュリティグループの名前をクリックします。

f. [Add Connection Type] をクリックします。

g. [Add Connection Type] ダイアログボックスで、[Connection Type] から [EC2 Security Group] を選択し、[EC2 Security Group Name] から作成したセキュリティグループを選択して, [Authorize] をクリックします。

4. [EC2-VPC] Amazon Redshift クラスターセキュリティグループを作成し、VPC セキュリティグループを指定します。

a. Amazon EC2 コンソールを開きます。

b. ナビゲーションペインで、[Security Groups] をクリックします。

c. [Create Security Group] をクリックします。

d. [Create Security Group] ダイアログボックスで、セキュリティグループの名前と説明を指定し、[VPC] で VPC の ID を選択します。

e. [Add Rule] をクリックします。[Source] で、タイプ、プロトコル、およびポート範囲を指定し、セキュリティグループの ID を入力します。2 番目の手順で作成したセキュリティグループを選択します。

f. [Create] (作成) をクリックします。

5. このステップの概要を以下に示します。

既存の Amazon Redshift クラスターがある場合は、クラスター ID を記録します。

新しいクラスターを作成してサンプルデータをロードするには、Amazon Redshift の開始方法のステップに従います。クラスター作成の詳細については、「Amazon Redshift管理ガイド」の「クラスターの作成」を参照してください。

a. Amazon Redshift コンソールを開きます。

b. [Launch Cluster] をクリックします。

c. クラスターに必要な詳細を入力し、[Continue] をクリックします。

d. ノード設定を入力し、[Continue] をクリックします。

e. 追加の設定情報のページで、作成したクラスターセキュリティグループを選択し、[Continue] をクリックします。

f. クラスターの仕様を確認してから [Launch Cluster] をクリックします。

コマンドラインを使用した Amazon Redshift へのデータのコピー

このチュートリアルでは、データを Amazon S3 から Amazon Redshift にコピーする方法を示します。Amazon Redshift に新しいテーブルを作成してから、AWS Data Pipeline を使用して、CSV 形式の入力データのサンプルが含まれているパブリックの Amazon S3 バケットからこのテーブルにデータを転送します。ログはお客様が所有する Amazon S3 バケットに保存されます。

Amazon S3 は、クラウドにデータを保存できるウェブサービスです。詳細については、Amazon Simple Storage Serviceユーザーガイドを参照してください。Amazon Redshift は、クラウド内のデータウェアハウスサービスです。詳細については、「Amazon Redshift 管理ガイド」を参照してください。

前提条件

開始する前に、次のステップを完了しておく必要があります。

1. コマンドラインインターフェイス (CLI) をインストールして設定します。詳細については、「アクセス AWS Data Pipeline」を参照してください。
2. DataPipelineDefaultRole と DataPipelineDefaultResourceRole という名前の IAM ロールが存在していることを確認します。AWS Data Pipeline コンソールにより、自動的にこれらのロールが作成されます。AWS Data Pipeline コンソールをまだ 1 回も使用したことがない場合、これらのロールを手動で作成する必要があります。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。
3. Amazon Redshift で COPY コマンドをセットアップします。これは AWS Data Pipeline でコピーを実行するときに同じオプションが機能する必要があるためです。詳細については、開始する前に: COPY オプションの設定とデータのロードを参照してください。
4. Amazon Redshift データベースを設定します。詳細については、「パイプラインのセットアップ、セキュリティグループの作成、および Amazon Redshift クラスターの作成」を参照してください。

タスク

- JSON 形式でパイプラインを定義する
- パイプライン定義をアップロードし、アクティブ化する

JSON 形式でパイプラインを定義する

この例のシナリオは、データを Amazon S3 バケットから Amazon Redshift にコピーする方法を示しています。

これは、完全なパイプライン定義の JSON ファイルであり、その後に各セクションの説明を示します。JSON 形式のファイルの構文を検証できるテキストエディタを使用し、.json ファイル拡張子を使用してファイルに名前を付けることをお勧めします。

```
{
  "objects": [
    {
      "id": "CSVId1",
      "name": "DefaultCSV1",
      "type": "CSV"
    },
    {
      "id": "RedshiftDatabaseId1",
      "databaseName": "dbname",
      "username": "user",
      "name": "DefaultRedshiftDatabase1",
      "*password": "password",
      "type": "RedshiftDatabase",
      "clusterId": "redshiftclusterId"
    },
    {
      "id": "Default",
      "scheduleType": "timeseries",
      "failureAndRerunMode": "CASCADE",
      "name": "Default",
      "role": "DataPipelineDefaultRole",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "id": "RedshiftDataNodeId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "tableName": "orders",
      "name": "DefaultRedshiftDataNode1",
      "createTableSql": "create table StructuredLogs (requestBeginTime CHAR(30)
 PRIMARY KEY DISTKEY SORTKEY, requestEndTime CHAR(30), hostname CHAR(100), requestDate
 varchar(20));",
```

```
      "type": "RedshiftDataNode",
      "database": {
        "ref": "RedshiftDatabaseId1"
      }
    },
    {
      "id": "Ec2ResourceId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "securityGroups": "MySecurityGroup",
      "name": "DefaultEc2Resource1",
      "role": "DataPipelineDefaultRole",
      "logUri": "s3://myLogs",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "type": "Ec2Resource"
    },
    {
      "id": "ScheduleId1",
      "startDateTime": "yyyy-mm-ddT00:00:00",
      "name": "DefaultSchedule1",
      "type": "Schedule",
      "period": "period",
      "endDateTime": "yyyy-mm-ddT00:00:00"
    },
    {
      "id": "S3DataNodeId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "filePath": "s3://datapipeline-us-east-1/samples/hive-ads-samples.csv",
      "name": "DefaultS3DataNode1",
      "dataFormat": {
        "ref": "CSVId1"
      },
      "type": "S3DataNode"
    },
    {
      "id": "RedshiftCopyActivityId1",
      "input": {
        "ref": "S3DataNodeId1"
      },
      "schedule": {
        "ref": "ScheduleId1"
```

```
      },
      "insertMode": "KEEP_EXISTING",
      "name": "DefaultRedshiftCopyActivity1",
      "runsOn": {
        "ref": "Ec2ResourceId1"
      },
      "type": "RedshiftCopyActivity",
      "output": {
        "ref": "RedshiftDataNodeId1"
      }
    }
  ]
}
```

これらのオブジェクトの詳細については、以下のドキュメントを参照してください。

オブジェクト

- データノード
- リソース
- アクティビティ

データノード

この例では、入力データノード、出力データノード、およびデータベースが使用されています。

入力データノード

入力 S3DataNode のパイプラインコンポーネントは、Amazon S3 における入力データの場所と入力データのデータ形式を定義します。詳細については、「S3DataNode」を参照してください。

この入力コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "S3DataNodeId1",
  "schedule": {
    "ref": "ScheduleId1"
  },
  "filePath": "s3://datapipeline-us-east-1/samples/hive-ads-samples.csv",
  "name": "DefaultS3DataNode1",
  "dataFormat": {
    "ref": "CSVId1"
```

```
  },
  "type": "S3DataNode"
},
```

id

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

schedule

スケジュールコンポーネントへの参照。

filePath

データノードに関連付けられているデータへのパス。この例では CSV 入力ファイルです。

name

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

dataFormat

処理するアクティビティのデータの形式への参照。

出力データノード

出力 RedshiftDataNode パイプライン コンポーネントは、出力データの場所を定義します。この例では Amazon Redshift データベース内のテーブルです。詳細については、「RedshiftDataNode」を参照してください。この出力コンポーネントは、次のフィールドで定義されます。

```
{
  "id": "RedshiftDataNodeId1",
  "schedule": {
    "ref": "ScheduleId1"
  },
  "tableName": "orders",
  "name": "DefaultRedshiftDataNode1",
  "createTableSql": "create table StructuredLogs (requestBeginTime CHAR(30) PRIMARY
 KEY DISTKEY SORTKEY, requestEndTime CHAR(30), hostname CHAR(100), requestDate
 varchar(20));",
  "type": "RedshiftDataNode",
  "database": {
    "ref": "RedshiftDatabaseId1"
  }
```

```
},
```

`id`

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

`schedule`

スケジュールコンポーネントへの参照。

`tableName`

Amazon Redshift テーブルの名前。

`name`

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

`createTableSql`

データベースにテーブルを作成する SQL 式。

`database`

Amazon Redshift データベースへの参照。

データベース

`RedshiftDatabase` コンポーネントは、以下のフィールドで定義されます。詳細については、「RedshiftDatabase」を参照してください。

```
{
  "id": "RedshiftDatabaseId1",
  "databaseName": "dbname",
  "username": "user",
  "name": "DefaultRedshiftDatabase1",
  "*password": "password",
  "type": "RedshiftDatabase",
  "clusterId": "redshiftclusterId"
},
```

`id`

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

`databaseName`

論理データベースの名前。

`username`

データベースに接続するためのユーザー名。

`name`

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

`password`

データベースに接続するためのパスワード。

`clusterId`

Redshift クラスターの ID。

リソース

これは、コピー操作を実行するコンピューティングリソースの定義です。この例では、AWS Data Pipeline は、コピータスクを実行するための EC2 インスタンスを自動的に作成し、コピー タスクが完了するとインスタンスを終了します。ここで定義されているフィールドが、作業を行うインスタンスの作成と機能を制御します。詳細については、「Ec2Resource」を参照してください。

`Ec2Resource` は、以下のフィールドで定義されます。

```
{
  "id": "Ec2ResourceId1",
  "schedule": {
    "ref": "ScheduleId1"
  },
  "securityGroups": "MySecurityGroup",
  "name": "DefaultEc2Resource1",
  "role": "DataPipelineDefaultRole",
  "logUri": "s3://myLogs",
  "resourceRole": "DataPipelineDefaultResourceRole",
  "type": "Ec2Resource"
},
```

`id`

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

`schedule`

このコンピューティングリソースを作成するスケジュール。

`securityGroups`

リソースプールのインスタンスに使用するセキュリティグループ。

`name`

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

`role`

リソースにアクセスするアカウントの IAM ロール (データを取得するための Amazon S3 バケットへのアクセスなど)。

`logUri`

`Ec2Resource` から Task Runner ログをバックアップする先の Amazon S3 の場所へのパス。

`resourceRole`

リソースを作成するアカウントの IAM ロール (お客様に代わって EC2 インスタンスを作成および設定するなど) 。Role と ResourceRole は同じロールにすることもできますが、個別に設定することによって、セキュリティ設定での詳細度が向上します。

アクティビティ

この JSON ファイルの最後のセクションは、実行する作業を表すアクティビティの定義です。ここでは、データを Amazon S3 から Amazon Redshift にコピーするために `RedshiftCopyActivity` コンポーネントを使用します。詳細については、「RedshiftCopyActivity」を参照してください。

`RedshiftCopyActivity` コンポーネントは、以下のフィールドで定義されます。

```
{
  "id": "RedshiftCopyActivityId1",
  "input": {
    "ref": "S3DataNodeId1"
  },
  "schedule": {
    "ref": "ScheduleId1"
  },
  "insertMode": "KEEP_EXISTING",
```

```
  "name": "DefaultRedshiftCopyActivity1",
  "runsOn": {
    "ref": "Ec2ResourceId1"
  },
  "type": "RedshiftCopyActivity",
  "output": {
    "ref": "RedshiftDataNodeId1"
  }
},
```

`id`

ユーザー定義の ID。これは参照時にのみ使用されるラベルです。

`input`

Amazon S3 ソース ファイルへの参照。

`schedule`

このアクティビティを実行するスケジュール。

`insertMode`

挿入のタイプ (`KEEP_EXISTING`、`OVERWRITE_EXISTING`、または `TRUNCATE`)。

`name`

ユーザー定義の名前。これは参照時にのみ使用されるラベルです。

`runsOn`

このアクティビティが定義する作業を実行するコンピューティングリソース。

`output`

コピー先の Amazon Redshift のテーブルへの参照。

パイプライン定義をアップロードし、アクティブ化する

パイプライン定義をアップロードし、パイプラインをアクティブ化する必要があります。以下のコマンド例では、*`pipeline_name`* をパイプラインのラベルに置き換え、*`pipeline_file`* をパイプライン定義 `.json` ファイルの完全修飾パスに置き換えます。

AWS CLI

パイプライン定義を作成してパイプラインをアクティブ化するには、以下の create-pipeline コマンドを使用します。パイプラインの ID をメモします。この値は、ほとんどの CLI コマンドで使用するからです。

```
aws datapipeline create-pipeline --name pipeline_name --unique-id token
{
    "pipelineId": "df-00627471SOVYZEXAMPLE"
}
```

パイプライン定義を更新するには、以下の put-pipeline-definition コマンドを使用します。

```
aws datapipeline put-pipeline-definition --pipeline-id df-00627471SOVYZEXAMPLE --
pipeline-definition file://MyEmrPipelineDefinition.json
```

パイプラインが正常に検証された場合、validationErrors フィールドは空です。警告を確認する必要があります。

パイプラインをアクティブ化するには、以下の activate-pipeline コマンドを使用します。

```
aws datapipeline activate-pipeline --pipeline-id df-00627471SOVYZEXAMPLE
```

以下の list-pipelines コマンドを使用して、パイプラインリストにパイプラインが表示されていることを確認できます。

```
aws datapipeline list-pipelines
```

パイプラインの式と関数

このセクションでは、関連するデータ型を含め、パイプラインで式と関数を使用するための構文について説明します。

単純データ型

以下のデータ型をフィールド値として設定できます。

タイプ

- DateTime
- 数値
- オブジェクト参照
- [Period] (期間)
- 文字列

DateTime

AWS Data Pipeline では、"YYYY-MM-DDTHH:MM:SS" という形式で表される UTC/GMT の日時のみをサポートしています。以下の例では、`Schedule` オブジェクトの `startDateTime`フィールドを、UTC/GMT タイムゾーンの `1/15/2012, 11:59 p.m.` に設定します。

```
"startDateTime" : "2012-01-15T23:59:00"
```

数値

AWS Data Pipeline は、整数と浮動小数点値の両方をサポートします。

オブジェクト参照

パイプライン定義内のオブジェクト。これは、現在のオブジェクト、パイプライン内の他の場所で定義されているオブジェクトの名前、または node キーワードで参照されるフィールド内の現在のオブジェクトをリストするオブジェクトのいずれかです。node の詳細については、「フィールドとオブジェクトの参照」を参照してください。パイプラインオブジェクトのタイプについては、「パイプラインオブジェクトリファレンス」を参照してください。

[Period] (期間)

予定されているイベントを実行する頻度を示します。これは、"N [years|months|weeks|days|hours|minutes]" という形式で表されます。ここで、N は正の整数値です。

最小間隔は 15 分で、最大間隔は 3 年です。

次の例では、Schedule オブジェクトの period フィールドを 3 時間に設定します。つまり、3 時間ごとに実行されるスケジュールが設定されます。

```
"period" : "3 hours"
```

文字列

標準文字列値。文字列は、二重引用符 (") で囲む必要があります。バックスラッシュ文字 (\ ) を使用して、文字列内の文字をエスケープすることができます。複数行の文字列はサポートされていません。

次の例では、id フィールドの有効な文字列値の例を示します。

```
"id" : "My Data Object"

"id" : "My \"Data\" Object"
```

文字列には、文字列値に評価される式を含めることもできます。これらの式は文字列内に挿入され、"#{" と "}" で区切られます。次の例では、式を使用してパスに現在のオブジェクト名を挿入します。

```
"filePath" : "s3://myBucket/#{name}.csv"
```

式の使用の詳細については、「フィールドとオブジェクトの参照」および「式の評価」を参照してください。

表現

式を使用すると、関連するオブジェクト間で値を共有できます。式は実行時に AWS Data Pipeline ウェブサービスによって処理され、すべての式に式の値が代入されます。

式は "#{" と "}" で区切られます。文字列が有効である、任意のパイプライン定義オブジェクトで式を使用できます。スロットが参照であるか、タイプ ID、NAME、TYPE、SPHERE のいずれかであれば、値は評価されず、逐語的に使用されます。

次の式は、AWS Data Pipeline 関数の 1 つを呼び出します。詳細については、「式の評価」を参照してください。

```
#{format(myDateTime,'YYYY-MM-dd hh:mm:ss')}
```

フィールドとオブジェクトの参照

式では、式が含まれる現在のオブジェクトのフィールドを使用したり、参照によってリンクされている別のオブジェクトのフィールドを使用したりすることができます。

スロットの形式は、作成時間に続いてオブジェクトの作成時間 (@S3BackupLocation_2018-01-31T11:05:33 など) で構成されます。

パイプライン定義で指定された正確なスロット ID (Amazon S3 バックアップ場所のスロット ID など) を参照することもできます。スロット ID を参照するには、#{parent.@id} を使用します。

次の例では、filePath フィールドは同じオブジェクトの id フィールドを参照してファイル名を形成します。値 filePath は「s3://mybucket/ExampleDataNode.csv」に評価されます。

```
{
  "id" : "ExampleDataNode",
  "type" : "S3DataNode",
  "schedule" : {"ref" : "ExampleSchedule"},
  "filePath" : "s3://mybucket/#{parent.@id}.csv",
  "precondition" : {"ref" : "ExampleCondition"},
  "onFail" : {"ref" : "FailureNotify"}
}
```

参照によってリンクされる別のオブジェクトに存在するフィールドを使用するには、node キーワードを使用します。このキーワードはアラームおよび前提条件のオブジェクトでのみ使用できます。

引き続き前の例で説明すると、SnsAlarm 内の式で Schedule 内の日時の範囲を参照できます。これは、S3DataNode が両方を参照しているためです。

具体的には、FailureNotify の message フィールドでExampleSchedule の @scheduledStartTime および @scheduledEndTime 実行時フィールドを使用できます。これ

は、ExampleDataNode の onFail フィールドが FailureNotify を参照し、その schedule フィールドが ExampleSchedule を参照するためです。

```
{
    "id" : "FailureNotify",
    "type" : "SnsAlarm",
    "subject" : "Failed to run pipeline component",
    "message": "Error for interval
 #{node.@scheduledStartTime}..#{node.@scheduledEndTime}.",
    "topicArn":"arn:aws:sns:us-east-1:28619EXAMPLE:ExampleTopic"
},
```

> **Note**
>
> 他のシステムまたはタスクの作業に依存するパイプライン内のタスクなど、依存関係を持つパイプラインを作成できます。パイプラインで特定のリソースが必要な場合、データノードやタスクに関連付ける前提条件を使用して、それらの依存関係をパイプラインに追加します。これにより、パイプラインはデバッグが容易になり、柔軟性が高くなります。さらに、複数のパイプラインにまたがるトラブルシューティングは困難であるため、可能な場合は依存関係を 1 つのパイプライン内に維持します。

入れ子式

AWS Data Pipeline では、より複雑な式を作成するために値を入れ子にすることができます。たとえば、時間の計算 (scheduledStartTime から 30 分を引く) を行い、結果の書式を設定してパイプライン定義で使用するには、アクティビティで次の式を使用できます。

```
#{format(minusMinutes(@scheduledStartTime,30),'YYYY-MM-dd hh:mm:ss')}
```

さらに、式が SnsAlarm または前提条件の一部である場合は、node プレフィックスを使用します。

```
#{format(minusMinutes(node.@scheduledStartTime,30),'YYYY-MM-dd hh:mm:ss')}
```

リスト

式はリストやリストを指定した関数に対して評価できます。たとえば、"myList": ["one","two"] のようにリストが定義されているとします。このリストが #{'this is ' + myList} という式で使用された場合、["this is one", "this is two"] に評価されま

す。2 つのリストがある場合、Data Pipeline による評価では、最終的に平坦化されます。たとえば、myList1 が [1,2] として定義され、myList2 が [3,4] として定義されている場合、式 [#{myList1}, #{myList2}] は、[1,2,3,4] に評価されます。

ノード式

AWS Data Pipeline では、パイプラインコンポーネントの親オブジェクトへの後方参照として、SnsAlarm または PreCondition で #{node.*} 式を使用します。SnsAlarm および PreCondition は、後方参照なしにアクティビティやリソースから参照されるため、node はリファラーを参照する方法を提供します。たとえば、次のパイプライン定義は、失敗通知で node を使用してその親 (この場合は ShellCommandActivity) を参照し、親の予定された開始時刻と終了時刻を SnsAlarm メッセージに含める方法を示しています。ShellCommandActivity の scheduledStartTime 参照は、scheduledStartTime がそれ自体を参照するため、node プレフィックスは必要ではありません。

> **Note**
>
> 先頭にアットマーク (@) がついているフィールドは、そのフィールドが実行時フィールドであることを示しています。

```
{
  "id" : "ShellOut",
  "type" : "ShellCommandActivity",
  "input" : {"ref" : "HourlyData"},
  "command" : "/home/userName/xxx.sh #{@scheduledStartTime} #{@scheduledEndTime}",
  "schedule" : {"ref" : "HourlyPeriod"},
  "stderr" : "/tmp/stderr:#{@scheduledStartTime}",
  "stdout" : "/tmp/stdout:#{@scheduledStartTime}",
  "onFail" : {"ref" : "FailureNotify"},
},
{
  "id" : "FailureNotify",
  "type" : "SnsAlarm",
  "subject" : "Failed to run pipeline component",
  "message": "Error for interval
 #{node.@scheduledStartTime}..#{node.@scheduledEndTime}.",
  "topicArn":"arn:aws:sns:us-east-1:28619EXAMPLE:ExampleTopic"
},
```

AWS Data Pipeline は、ユーザー定義フィールドについては推移的参照をサポートしていますが、実行時フィールドについてはサポートしていません。推移的参照は、仲介として別のパイプラインコンポーネントに依存する 2 つのパイプラインコンポーネント間の参照です。次の例は、推移的なユーザー定義フィールドへの参照と、推移的ではない実行時フィールドへの参照を示しています。これらの参照はいずれも有効です。詳細については、「ユーザー定義フィールド」を参照してください。

```
{
  "name": "DefaultActivity1",
  "type": "CopyActivity",
  "schedule": {"ref": "Once"},
  "input": {"ref": "s3nodeOne"},
  "onSuccess": {"ref": "action"},
  "workerGroup": "test",
  "output": {"ref": "s3nodeTwo"}
},
{
  "name": "action",
  "type": "SnsAlarm",
  "message": "S3 bucket '#{node.output.directoryPath}' succeeded at
 #{node.@actualEndTime}.",
  "subject": "Testing",
  "topicArn": "arn:aws:sns:us-east-1:28619EXAMPLE:ExampleTopic",
  "role": "DataPipelineDefaultRole"
}
```

式の評価

AWS Data Pipeline には、フィールドの値を計算するために使用できる一連の関数が用意されています。次の例では、makeDate 関数を使用して、Schedule オブジェクトの startDateTime フィールドを、"2011-05-24T0:00:00" GMT/UTC に設定します。

```
"startDateTime" : "makeDate(2011,5,24)"
```

数学関数

以下の関数は、数値を操作するために使用できます。

関数	説明
+	加算。

関数	説明
	例: #{1 + 2} 結果: 3
-	減算。 例: #{1 - 2} 結果: -1
*	乗算。 例: #{1 * 2} 結果: 2
/	除算。2 個の整数を除算する場合、結果は切り捨てられます。 例: #{1 / 2}、結果: 0 例: #{1.0 / 2}、結果: .5
^	指数。 例: #{2 ^ 2} 結果: 4.0

文字列関数

以下の関数は、文字列値を操作するために使用できます。

関数	説明
+	連結。非文字列値は最初に文字列に変換されます。 例: #{"hel" + "lo"}

関数	説明
	結果: "hello"

日付および時刻関数

以下の関数は、DateTime 値を操作するために使用できます。例では、myDateTime の値は、May 24, 2011 @ 5:10 pm GMT です。

> **Note**
>
> AWS Data Pipeline の日付/時刻形式は Joda Time で、これは Java の日付と時刻のクラスを置き換えます。詳細については、「Joda Time - DateTimeFormat クラス」を参照してください。

関数	説明
int day(DateTime myDateTime)	DateTime 値の日付を整数として取得します。 例: #{day(myDateTime)} 結果: 24
int dayOfYear(DateTime myDateTime)	DateTime 値の年初からの日数を整数として取得します。 例: #{dayOfYear(myDateTime)} 結果: 144
DateTime firstOfMonth(DateTime myDateTime)	指定された DateTime の月初の DateTime オブジェクトを作成します。 例: #{firstOfMonth(myDateTime)}

関数	説明
	結果: "2011-05-01T17:10:00z"
String format(DateTime myDateTime,String format)	指定された書式文字列を使用して、指定された DateTime を変換した結果である String オブジェクトを作成します。 例: #{format(myDateTime,'YYYY-MM-dd HH:mm:ss z')} 結果: "2011-05-24T17:10:00 UTC"
int hour(DateTime myDateTime)	DateTime 値の時間を整数として取得します。 例: #{hour(myDateTime)} 結果: 17
DateTime makeDate(int year,int month,int day)	指定された年、月、日の深夜 0 時で、UTC の DateTime オブジェクトを作成します。 例: #{makeDate(2011,5,24)} 結果: "2011-05-24T0:00:00z"

関数	説明
`DateTime makeDateTime(int year,int month,int day,int hour,int minute)`	指定された年、月、日、時、分で、UTC の DateTime オブジェクトを作成します。 例: `#{makeDateTime(2011,5,24,14,21)}` 結果: `"2011-05-24T14:21:00z"`
`DateTime midnight(DateTime myDateTime)`	指定された DateTime を基準として、現在の深夜 0 時の DateTime オブジェクトを作成します。たとえば、`MyDateTime` が `2011-05-25T17:10:00z` である場合、結果は次のとおりです。 例: `#{midnight(myDateTime)}` 結果: `"2011-05-25T0:00:00z"`
`DateTime minusDays(DateTime myDateTime,int daysToSub)`	指定された DateTime から指定された日数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusDays(myDateTime,1)}` 結果: `"2011-05-23T17:10:00z"`

関数	説明
`DateTime minusHours(DateTime myDateTime,int hoursToSub)`	指定された DateTime から指定された時間数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusHours(myDateTime,1)}` 結果: `"2011-05-24T16:10:00z"`
`DateTime minusMinutes(DateTime myDateTime,int minutesToSub)`	指定された DateTime から指定された分数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusMinutes(myDateTime,1)}` 結果: `"2011-05-24T17:09:00z"`
`DateTime minusMonths(DateTime myDateTime,int monthsToSub)`	指定された DateTime から指定された月数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusMonths(myDateTime,1)}` 結果: `"2011-04-24T17:10:00z"`

関数	説明
`DateTime minusWeeks(DateTime myDateTime,int weeksToSub)`	指定された DateTime から指定された週数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusWeeks(myDateTime,1)}` 結果: `"2011-05-17T17:10:00z"`
`DateTime minusYears(DateTime myDateTime,int yearsToSub)`	指定された DateTime から指定された年数を引いた結果の DateTime オブジェクトを作成します。 例: `#{minusYears(myDateTime,1)}` 結果: `"2010-05-24T17:10:00z"`
`int minute(DateTime myDateTime)`	DateTime 値の分を整数として取得します。 例: `#{minute(myDateTime)}` 結果: `10`
`int month(DateTime myDateTime)`	DateTime 値の月を整数として取得します。 例: `#{month(myDateTime)}` 結果: `5`

関数	説明
`DateTime plusDays(DateTime myDateTime,int daysToAdd)`	指定された DateTime に指定された日数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusDays(myDateTime,1)}` 結果: `"2011-05-25T17:10:00z"`
`DateTime plusHours(DateTime myDateTime,int hoursToAdd)`	指定された DateTime に指定された時間数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusHours(myDateTime,1)}` 結果: `"2011-05-24T18:10:00z"`
`DateTime plusMinutes(DateTime myDateTime,int minutesToAdd)`	指定された DateTime に指定された分数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusMinutes(myDateTime,1)}` 結果: `"2011-05-24 17:11:00z"`

関数	説明
`DateTime plusMonths(DateTime myDateTime,int monthsToAdd)`	指定された DateTime に指定された月数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusMonths(myDateTime,1)}` 結果: `"2011-06-24T17:10:00z"`
`DateTime plusWeeks(DateTime myDateTime,int weeksToAdd)`	指定された DateTime に指定された週数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusWeeks(myDateTime,1)}` 結果: `"2011-05-31T17:10:00z"`
`DateTime plusYears(DateTime myDateTime,int yearsToAdd)`	指定された DateTime に指定された年数を足した結果の DateTime オブジェクトを作成します。 例: `#{plusYears(myDateTime,1)}` 結果: `"2012-05-24T17:10:00z"`

関数	説明
`DateTime sunday(DateTime myDateTime)`	指定された DateTime を基準として、前の日曜日の DateTime オブジェクトを作成します。指定された DateTime が日曜日である場合、結果は指定された DateTime です。 例: `#{sunday(myDateTime)}` 結果: `"2011-05-22 17:10:00 UTC"`
`int year(DateTime myDateTime)`	DateTime 値の年を整数として取得します。 例: `#{year(myDateTime)}` 結果: `2011`
`DateTime yesterday(DateTime myDateTime)`	指定された DateTime を基準として、前の日の DateTime オブジェクトを作成します。結果は minusDays(1) と同じです。 例: `#{yesterday(myDateTime)}` 結果: `"2011-05-23T17:10:00z"`

特殊文字

AWS Data Pipeline では、以下の表に示されているような、パイプライン定義で特別な意味を持つ特定の文字を使用します。

特殊文字	説明	例
@	実行時フィールド。この文字は、パイプラインの実行時にのみ使用できるフィールドのフィールド名プレフィックスです。	@actualStartTime @failureReason @resourceStatus
#	式 式は "#{" と "}" によって区切られ、中括弧で囲まれた内容は AWS Data Pipeline によって評価されます。詳細については、「表現」を参照してください。	#{format(myDateTime,'YYYY-MM-dd hh:mm:ss')} s3://mybucket/#{id}.csv
*	暗号化されたフィールド。この文字は、コンソールや CLI と AWS Data Pipeline サービスとの間で送信する際に、AWS Data Pipeline でこのフィールドの内容を暗号化する必要があることを示すフィールド名プレフィックスです。	*パスワード

パイプラインオブジェクトリファレンス

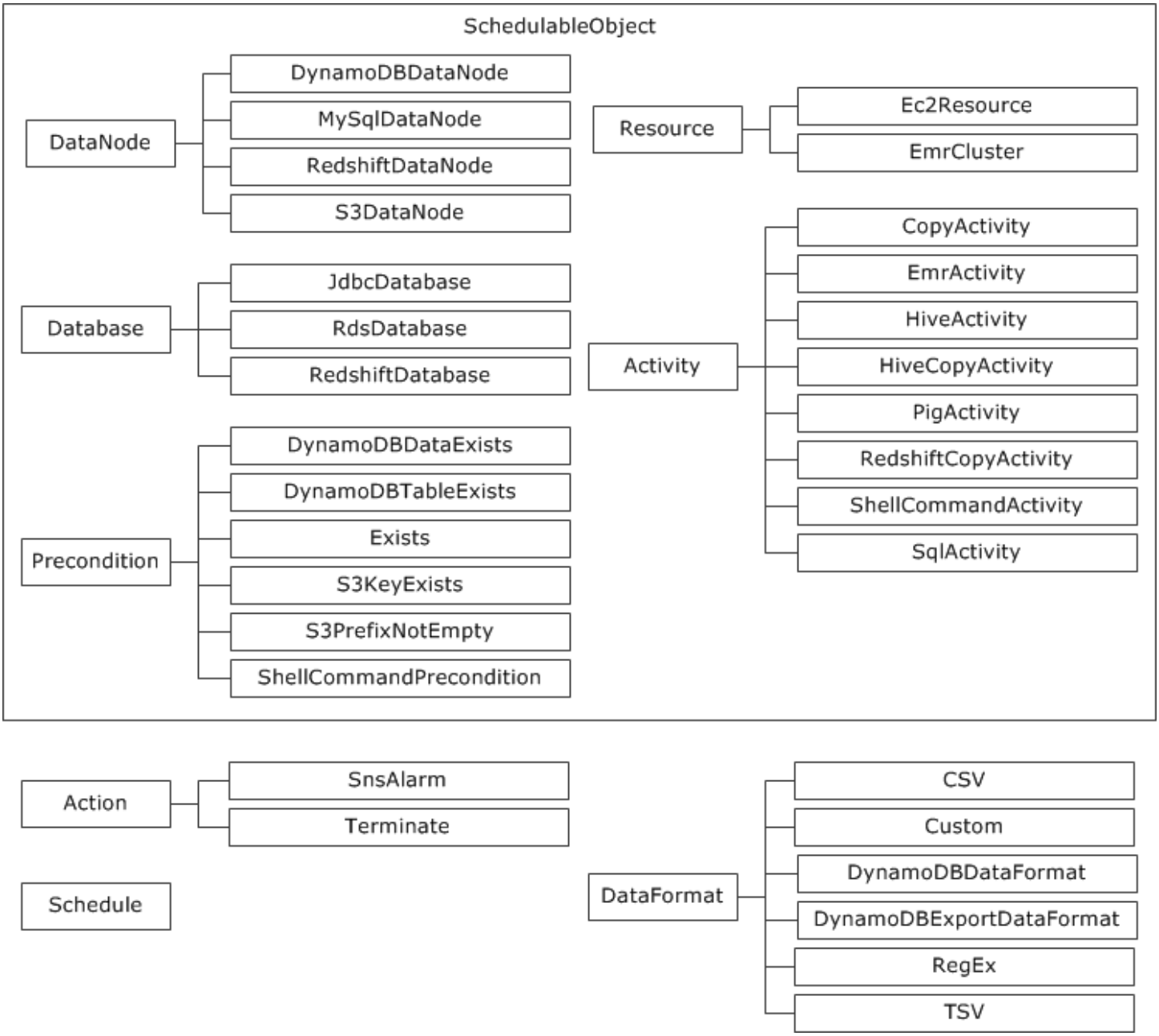
パイプライン定義では、以下のパイプラインオブジェクトとコンポーネントを使用できます。

内容

- データノード
- アクティビティ
- リソース
- 前提条件
- データベース
- データ形式
- アクション
- スケジュール
- ユーティリティ

> **Note**
>
> AWS Data Pipeline Java を使用するアプリケーションの例についてはSDK、「 での Data Pipeline DynamoDB Export Java Sample」を参照してください GitHub。

以下は、 のオブジェクト階層です AWS Data Pipeline。



データノード

AWS Data Pipeline データノードオブジェクトは次のとおりです。

オブジェクト

- D ynamoDBDataノード
- MySqlDataNode
- RedshiftDataNode

- S3DataNode
- SqlDataNode

D ynamoDBDataノード

DynamoDB を使用してデータノードを定義します。データノードは、HiveActivity または EMRActivity オブジェクトの入力として指定されます。

> **Note**
>
> DynamoDBDataNode オブジェクトは、Exists 前提条件をサポートしていません。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 2 つ参照します。CopyPeriod は Schedule オブジェクトで、Ready は前提条件オブジェクトです。

```
{
  "id" : "MyDynamoDBTable",
  "type" : "DynamoDBDataNode",
  "schedule" : { "ref" : "CopyPeriod" },
  "tableName" : "adEvents",
  "precondition" : { "ref" : "Ready" }
}
```

構文

必須フィールド	説明	スロットタイプ
tableName	DynamoDB テーブル。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「スケジュール」を参照してください。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。このフィールドを設定すると、開始後の設定時間内に完了しなかったリモートアクティビティを再試行できます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
dataFormat	DataFormat このデータノードで記述されるデータ用の 。現在、 HiveActivity および でサポートされています HiveCopyActivity。	リファレンスオブジェクト、"dataFormat":{"ref"myDynamoDBDataFormatId}
dependsOn	実行可能な別のオブジェクトで依存関係を指定します	リファレンスオブジェクト、例: "dependsOn":{"ref"myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref" myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref'myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref" myPreconditionId」}
readThroughputPercent	DynamoDB のプロビジョニングされたスループットレートが、表の割り当てられた範囲内に収まるように、読み取りオペレーションのレートを設定します。値は 0.1 ~ 1.0 の範囲の double 値です。	ダブル
region	DynamoDB テーブルがあるリージョンのコード。例えば、us-east-1 などです。これは、Hive で DynamoDB テーブルのステージングを実行する HiveActivity ときに によって使用されます。	一覧表
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトの オブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、on demand、および timeseries です。	一覧表
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
writeThroughputPercent	DynamoDB のプロビジョニングされたスループットレートが、表の割り当てられた範囲内に収まるように、書き込みオペレーションのレートを設定します。値は 0.1 ~ 1.0 の範囲の double 値です。	ダブル

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列

実行時フィールド	説明	スロットタイプ
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime

実行時フィールド	説明	スロットタイプ
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

MySqlDataNode

My を使用してデータノードを定義しますSQL。

> **Note**
>
> `MySqlDataNode` 型は廃止されました。代わりに SqlDataNode を使用することをお勧めします。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 2 つ参照します。CopyPeriod は Schedule オブジェクトで、Ready は前提条件オブジェクトです。

```
{
  "id" : "Sql Table",
  "type" : "MySqlDataNode",
  "schedule" : { "ref" : "CopyPeriod" },
  "table" : "adEvents",
  "username": "user_name",
  "*password": "my_password",
  "connectionString": "jdbc:mysql://mysqlinstance-rds.example.us-
east-1.rds.amazonaws.com:3306/database_name",
  "selectQuery" : "select * from #{table} where eventTime >=
 '#{@scheduledStartTime.format('YYYY-MM-dd HH:mm:ss')}' and eventTime <
 '#{@scheduledEndTime.format('YYYY-MM-dd HH:mm:ss')}'",
  "precondition" : { "ref" : "Ready" }
}
```

構文

必須フィールド	説明	スロットタイプ
テーブル	My データベース内のテーブルSQLの名前。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。	リファレンスオブジェクト、例: "schedule":{"ref"mayScheduleId"}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
createTableSql	テーブルSQLを作成するテーブル作成式。	文字列
データベース	データベースの名前。	リファレンスオブジェクト、例：「database」:{「ref"myDatabaseId」}

オプションのフィールド	説明	スロットタイプ
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
insertQuery	テーブルにデータを挿入するSQLステートメント。	文字列
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref" myActionId"}

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトの オブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、on demand、および timeseries です。	一覧表
schemaName	テーブルを保持するスキーマの名前	文字列
selectQuery	テーブルからデータを取得するSQLステートメント。	文字列
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- S3DataNode

RedshiftDataNode

Amazon Redshift を使用するデータノードを定義します。RedshiftDataNode は、パイプラインで使用される、データベース内のデータのプロパティ (データテーブルなど) を表します。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyRedshiftDataNode",
  "type" : "RedshiftDataNode",
  "database": { "ref": "MyRedshiftDatabase" },
  "tableName": "adEvents",
  "schedule": { "ref": "Hour" }
}
```

構文

必須フィールド	説明	スロットタイプ
データベース	テーブルが存在するデータベース。	リファレンスオブジェクト、例：「database」:{「ref"myRedshiftDatabaseId」}
tableName	Amazon Redshift テーブルの名前。テーブルがまだ存在せず、 を指定している場合、テーブルが作成されます createTableSql。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGu	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	ide/dp-object-schedule.html」を参照してください。	

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
createTableSql	データベースにテーブルを作成するSQL式。テーブルを作成するスキーマを指定することをお勧めしますmySchema。例えば、.myTable （ bestColumn varchar(25) CREATE TABLE primary key distkey, numberOfWins integer ですsortKey。 で指定されたテーブルが createTableSql フィールドで指定されたスキーマに存在しない場合tableName、 は schemaName フィールドでスクリプト AWS Data Pipeline を実行します。例えば、 を schemaName として指定し、 createTableSql フィールド mySchema には含め mySchema ない場合、テーブルは間違ったスキーマで作成されます (デフォルトでは で作成されます PUBLIC ）。これは、AWSData Pipeline が CREATETABLEステートメントを解析しないために発生します。	文字列

オプションのフィールド	説明	スロットタイプ
dependsOn	実行可能な別のオブジェクトで依存関係を指定します	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref" myActionId"}

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
primaryKeys	primaryKeys の送信先テーブルに を指定しない場合`RedShiftCopyActivity` 、 primaryKeys が として機能する列のリストを指定できますmergeKey。ただし、Amazon Redshift テーブルに既存の が primaryKey 定義されている場合、この設定は既存のキーを上書きします。	文字列
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表
schemaName	このオプションフィールドは、Amazon Redshift テーブルのスキーマ名を指定します。指定しない場合、スキーマ名は です。PUBLIC これは Amazon Redshift のデフォルトスキーマです。詳細については、「Amazon Redshift データベースデベロッパーガイド」を参照してください。	文字列
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

S3DataNode

Amazon S3 を使用するデータノードを定義します。デフォルトでは、S3DataNode はサーバー側の暗号化を使用します。これを無効にする場合は、s3 を EncryptionType に設定しますNONE。

> **Note**
>
> への入力S3DataNodeとして を使用する場合CopyActivity、 CSVおよび TSV データ形式のみがサポートされます。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した別のオブジェクトを参照します。CopyPeriod は Schedule オブジェクトです。

```
{
  "id" : "OutputData",
  "type" : "S3DataNode",
  "schedule" : { "ref" : "CopyPeriod" },
  "filePath" : "s3://myBucket/#{@scheduledStartTime}.csv"
}
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかっ	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
	たリモートアクティビティを再試行することができます。	
compression	S3DataNode によって記述されるデータの圧縮のタイプ。「none」は圧縮されず、「gzip」は gzip アルゴリズムで圧縮されます。このフィールドは、Amazon Redshift での使用と、 で S3DataNode を使用する場合にのみサポートされます CopyActivity。	一覧表
dataFormat	DataFormat この S3DataNode で説明されているデータの 。	リファレンスオブジェクト、例: "dataFormat":{"ref "myDataFormatId"}
dependsOn	実行可能な別のオブジェクトで依存関係を指定します	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
directoryPath	としての Amazon S3 ディレクトリパスURI: s3://my-bucket/my-key-for-directory。 filePath または directoryPath 値を指定する必要があります。	文字列
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
filePath	としての Amazon S3 内のオブジェクトへのパスURI。例: s3://my-bucket/my-key-for-file。 filePath または directoryPath 値を指定する必要があります。これらの値は、フォルダとファイル名を表します。 directoryPath 値を使用して、ディレクトリ内の複数のファイルに対応します。	文字列

オプションのフィールド	説明	スロットタイプ
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
manifestFilePath	Amazon Redshift. でサポートされている形式のマニフェストファイルへの Amazon S3 パスは、マニフェストファイル AWS Data Pipeline を使用して、指定された Amazon S3 ファイルをテーブルにコピーします。このフィールドは、 が S3DataNode RedShiftCopyActivity を参照する場合にのみ有効です。	文字列
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}
s3EncryptionType	Amazon S3 の暗号化タイプをオーバーライドします。値は SERVER_SIDE_ENCRYPTION または ですNONE。デフォルトではサーバー側の暗号化が有効化されます。	一覧表

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}

実行時フィールド	説明	スロットタイプ
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime

実行時フィールド	説明	スロットタイプ
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列

システムフィールド	説明	スロットタイプ
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- MySqlDataNode

SqlDataNode

SQL を使用するデータノードを定義します。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 2 つ参照します。CopyPeriod は Schedule オブジェクトで、Ready は前提条件オブジェクトです。

```
{
  "id" : "Sql Table",
  "type" : "SqlDataNode",
  "schedule" : { "ref" : "CopyPeriod" },
  "table" : "adEvents",
  "database":"myDataBaseName",
  "selectQuery" : "select * from #{table} where eventTime >=
 '#{@scheduledStartTime.format('YYYY-MM-dd HH:mm:ss')}' and eventTime <
 '#{@scheduledEndTime.format('YYYY-MM-dd HH:mm:ss')}'",
  "precondition" : { "ref" : "Ready" }
}
```

構文

必須フィールド	説明	スロットタイプ
テーブル	SQL データベース内のテーブルの名前。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかっ	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
	たリモートアクティビティを再試行することができます。	
createTableSql	テーブルSQLを作成するテーブル作成式。	文字列
データベース	データベースの名前。	リファレンスオブジェクト、例：「database」:{「ref"myDatabaseId」}
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref"myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
insertQuery	テーブルにデータを挿入するSQLステートメント。	文字列
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表
schemaName	テーブルを保持するスキーマの名前	文字列
selectQuery	テーブルからデータを取得するSQLステートメント。	文字列
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- S3DataNode

アクティビティ

AWS Data Pipeline アクティビティオブジェクトは次のとおりです。

オブジェクト

- CopyActivity
- EmrActivity
- HadoopActivity
- HiveActivity
- HiveCopyActivity
- PigActivity
- RedshiftCopyActivity
- ShellCommandActivity
- SqlActivity

CopyActivity

ある場所から別の場所にデータをコピーします。 は入力および出力SqlDataNodeとして S3DataNode と CopyActivityをサポートし、コピーオペレーションは通常 で実行されます record-by-record。ただし、以下のすべての条件が満たされた場合、CopyActivity は Amazon S3 から Amazon S3 への高速コピーを実行します。

- 入力と出力は S3DataNodes
- dataFormat フィールドが入力と出力で同じであること

入力として圧縮データファイルを指定した場合に、それを S3 データノードの compression フィールドを使用して示していなければ、CopyActivity は失敗することがあります。この場合、CopyActivity がレコード終了文字を適切に検出できないために、操作が失敗しています。さらに、 はディレクトリから別のディレクトリへのコピーとディレクトリへのファイルのコピーCopyActivityをサポートしますが、 record-by-record コピーはディレクトリをファイルにコピーするときに行われます。また、CopyActivity では、マルチパートの Amazon S3 ファイルのコピーはサポートされません。

CopyActivity には、そのCSVサポートに対する特定の制限があります。に S3DataNode as 入力を使用する場合CopyActivity、Amazon S3 の入力フィールドと出力フィールドには、CSVデータファイル形式の Unix/Linux バリアントのみを使用できます。 Amazon S3 この Unix/Linux 形式では、次の条件が満たされる必要があります。

- 区切り文字はカンマ (,) 文字です。
- レコードが引用符で囲まれることはありません。
- デフォルトのエスケープ文字はASCII値 92 (バックスラッシュ) です。
- レコード終了識別子はASCII値 10 (または "\n") です。

Windows ベースのシステムは通常、キャリッジリターンとラインフィードを一緒に使用するという異なる end-of-record 文字シーケンスを使用します (ASCII値 13 とASCII値 10)。入力データを変更するコピー前スクリプトなどを利用して、この違いを吸収し、CopyActivity が適切にレコード終端を検出できるようにする必要があります。そうしないと、CopyActivity が繰り返し失敗します。

CopyActivity を使用して PostgreSQL RDS オブジェクトからTSVデータ形式にエクスポートする場合、デフォルトのNULL文字は \n です。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 3 つ参照します。CopyPeriod は Schedule オブジェクトで、InputData と OutputData はデータノードオブジェクトです。

```
{
  "id" : "S3ToS3Copy",
  "type" : "CopyActivity",
  "schedule" : { "ref" : "CopyPeriod" },
  "input" : { "ref" : "InputData" },
  "output" : { "ref" : "OutputData" },
  "runsOn" : { "ref" : "MyEc2Resource" }
}
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref"myActivityId"}

オプションのフィールド	説明	スロットタイプ
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データソース。	リファレンスオブジェクト、例：「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
output	出力データソース。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime

実行時フィールド	説明	スロットタイプ
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列

実行時フィールド	説明	スロットタイプ
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクト	文字列

システムフィールド	説明	スロットタイプ
	により、試行オブジェクトを実行するインスタンスオブジェクトが発生します	

以下の資料も参照してください。

- ShellCommandActivity
- EmrActivity
- AWS Data Pipeline を使用した Amazon S3 への MySQL データのエクスポート

EmrActivity

EMR クラスターを実行します。

AWS Data Pipeline は、Amazon とは異なる形式のステップを使用しますEMR。たとえば、`EmrActivity`ステップフィールドJARの名前の後にカンマ区切りの引数 AWS Data Pipeline を使用します。次の例は、Amazon 用にフォーマットされたステップとEMR、それに AWS Data Pipeline 続く同等のステップを示しています。

```
s3://example-bucket/MyWork.jar arg1 arg2 arg3
```

```
"s3://example-bucket/MyWork.jar,arg1,arg2,arg3"
```

例

以下は、このオブジェクト型の例です。この例では、古いバージョンの Amazon を使用していますEMR。この例が正しいかどうかは、使用している Amazon EMRクラスターのバージョンで確認してください。

このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 3 つ参照します。`MyEmrCluster` は `EmrCluster` オブジェクトで、`MyS3Input` と `MyS3Output` は `S3DataNode` オブジェクトです。

> **Note**
>
> この例では、 stepフィールドを目的のクラスター文字列に置き換えることができます。これは、Pig スクリプト、Hadoop ストリーミングクラスター、パラメータJARを含む独自のカスタムなどです。

Hadoop 2.x (AMI 3.x)

```
{
  "id" : "MyEmrActivity",
  "type" : "EmrActivity",
  "runsOn" : { "ref" : "MyEmrCluster" },
  "preStepCommand" : "scp remoteFiles localFiles",
  "step" : ["s3://mybucket/myPath/myStep.jar,firstArg,secondArg,-files,s3://mybucket/
myPath/myFile.py,-input,s3://myinputbucket/path,-output,s3://myoutputbucket/path,-
mapper,myFile.py,-reducer,reducerName","s3://mybucket/myPath/myotherStep.jar,..."],
  "postStepCommand" : "scp localFiles remoteFiles",
  "input" : { "ref" : "MyS3Input" },
  "output" : { "ref" : "MyS3Output" }
}
```

> **Note**
>
> 引数を 1 ステップでアプリケーションに渡すには、以下の例のように、スクリプトのパスに Region を指定する必要があります。さらに、渡す引数をエスケープすることが必要になる場合があります。たとえば、script-runner.jar を使ってシェルスクリプトを実行する場合、スクリプトに引数を渡すには、引数を区切るカンマをエスケープする必要があります。次のステップスロットはそれを行う方法を示しています。
>
> ```
> "step" : "s3://eu-west-1.elasticmapreduce/libs/script-runner/script-
> runner.jar,s3://datapipeline/echo.sh,a\\\\,b\\\\,c"
> ```
>
> このステップでは、script-runner.jar を使って echo.sh シェルスクリプトを実行し、a、b、c を単一の引数としてスクリプトに渡します。最初のエスケープ文字は結果として生じる引数から削除されるので再度エスケープする必要があります。例えば、 で引数File\.gzとして がある場合JSON、 を使用してエスケープできますFile\\\\.gz。ただし、最初のエスケープは破棄されるため、File\\\\\\\\.gz を使う必要があります。

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。このオブジェクトの依存関係の実行順序を設定するために、別のオブジェクトへのスケジュール参照を指定します。この要件を満たすには、オブジェクトのスケジュールを明示的に設定できます。たとえば、`"schedule": {"ref": "DefaultSchedule"}` と指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マススケジュール内のスケジュール) がある場合は、スケジュール参照がある親オブジェクトを作成できます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	リファレンスオブジェクト、例えば「schedule」:{「ref"myScheduleId」}

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	このジョブを実行する Amazon EMRクラスター。	リファレンスオブジェクト、例：runsOn「」:{「ref"myEmrClusterId」}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。値 `runsOn` を指定し	文字列

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
	て、workerGroup がある場合、workerGroup は無視されます。	

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例：dependsOn「」：{「ref"myActivityId」}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データの場所。	リファレンスオブジェクト、例えば「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例：onFail「」:{「ref"myActionId」}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例：onLateAction「」:{「ref"myActionId」}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例：onSuccess「」:{「ref"myActionId」}
output	出力データの場所。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするための 's3:///Prefix/' などの Amazon S3 。 URIBucketName	文字列

オプションのフィールド	説明	スロットタイプ
postStepCommand	すべてのステップが完了した後に実行するシェルスクリプト。スクリプトを複数 (最大 255 個) 指定するには、postStepCommand フィールドを複数追加します。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref" myPreconditionId」}
preStepCommand	ステップを実行する前に実行するシェルスクリプト。スクリプトを複数 (最大 255 個) 指定するには、preStepCommand フィールドを複数追加します。	文字列
reportProgressTimeout	reportProgress へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
resizeClusterBefore実行中	入力または出力として指定された DynamoDB テーブルに対応するため、このアクティビティを実行する前にクラスターのサイズを変更します。 **Note** `EmrActivity` が を入力データノードまたは出力データノード`DynamoDBDataNode` として使用し、`resizeClusterBeforeRunning` を に設定すると`TRUE`、 は`m3.xlarge` インスタンスタイプを使用して AWS Data Pipeline 開始します。これによりインスタンスタイプの選択が `m3.xlarge` に上書きされるため、月別コストが増加することがあります。	ブール値
resizeClusterMaxインスタンス	サイズ変更アルゴリズムによってリクエストできるインスタンスの最大数の制限。	整数
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプでは、パイプライン定義のオブジェクトを間隔の最初にスケジュールするか、間隔の最後にスケジュールするかを指定できます。値は、cron、ondemand、および timeseries です。timeseries スケジューリングでは、インスタンスを各間隔の最後にスケジュールします。cron スケジューリングでは、インスタンスを各間隔の最初にスケジュールします。ondemand スケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。ondemand スケジュールを使用する場合は、デフォルトオブジェクトで指定し、パイプラインのオブジェクトに対して指定される唯一の scheduleType である必要があります。ondemand パイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。	一覧表
ステップ	クラスターが実行する 1 つ以上のステップ。ステップを複数 (最大 255 個) 指定するには、ステップフィールドを複数追加します。JAR 名前の後にカンマ区切りの引数を使用します。例：s3://example-bucket/MyWork.jar,arg1,arg2,arg3 「」。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeIn

実行時フィールド	説明	スロットタイプ
		stances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例：cascadeFailedOn「」：{「ref"myRunnableObjectId」}
emrStepLog	EMR アクティビティの試行でのみ使用できる Amazon EMR ステップログ	文字列
errorId	このオブジェクトが失敗した場合は `errorId`。	文字列
errorMessage	このオブジェクトが失敗した場合は `errorMessage` 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトを作成したパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例：waitingOn「」：{「ref"myRunnableObjectId」}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandActivity
- CopyActivity
- EmrCluster

HadoopActivity

クラスターで MapReduce ジョブを実行します。クラスターは、 によって管理される EMR クラスター AWS Data Pipeline でも、 を使用する場合は別のリソースでもかまいません TaskRunner。作業を並行して実行 HadoopActivity する場合に使用します。これにより、Hadoop 1 のYARNフレームワークまたはリソースネゴシエーターのスケジューリング MapReduce リソースを使用できます。Amazon EMR Step アクションを使用して作業を順番に実行する場合でも、 を使用できますEmrActivity。

例

HadoopActivity によって管理される EMRクラスターの使用 AWS Data Pipeline

次の HadoopActivity オブジェクトは、 EmrCluster リソースを使用してプログラムを実行します。

```
{
  "name": "MyHadoopActivity",
  "schedule": {"ref": "ResourcePeriod"},
  "runsOn": {"ref": “MyEmrCluster”},
```

```
   "type": "HadoopActivity",
   "preActivityTaskConfig":{"ref":"preTaskScriptConfig”},
   "jarUri": "/home/hadoop/contrib/streaming/hadoop-streaming.jar",
   "argument": [
     "-files",
     “s3://elasticmapreduce/samples/wordcount/wordSplitter.py“,
     "-mapper",
     "wordSplitter.py",
     "-reducer",
     "aggregate",
     "-input",
     "s3://elasticmapreduce/samples/wordcount/input/",
     "-output",
     “s3://test-bucket/MyHadoopActivity/#{@pipelineId}/
#{format(@scheduledStartTime,'YYYY-MM-dd')}"
   ],
   "maximumRetries": "0",
   "postActivityTaskConfig":{"ref":"postTaskScriptConfig”},
   "hadoopQueue" : “high”
 }
```

対応する を次に示します。*MyEmrCluster*。Hadoop 2 ベースの YARN の で FairScheduler およびキューを設定しますAMIs。

```
{
  "id" : "MyEmrCluster",
  "type" : "EmrCluster",
   "hadoopSchedulerType" : "PARALLEL_FAIR_SCHEDULING",
  “amiVersion” : “3.7.0”,
  "bootstrapAction" : ["s3://Region.elasticmapreduce/bootstrap-
actions/configure-hadoop,-z,yarn.scheduler.capacity.root.queues=low
\,high\,default,-z,yarn.scheduler.capacity.root.high.capacity=50,-
z,yarn.scheduler.capacity.root.low.capacity=10,-
z,yarn.scheduler.capacity.root.default.capacity=30”]
}
```

これは、Hadoop 1 FairScheduler で設定 EmrCluster するために使用する です。

```
{
      "id": "MyEmrCluster",
      "type": "EmrCluster",
      "hadoopSchedulerType": "PARALLEL_FAIR_SCHEDULING",
      "amiVersion": "2.4.8",
```

```
      "bootstrapAction": "s3://Region.elasticmapreduce/bootstrap-
actions/configure-hadoop,-m,mapred.queue.names=low\\\\,high\\\\,default,-
m,mapred.fairscheduler.poolnameproperty=mapred.job.queue.name"
          }
```

以下は EmrCluster 、Hadoop 2 ベースの CapacityScheduler の設定ですAMIs。

```
{
      "id": "MyEmrCluster",
      "type": "EmrCluster",
      "hadoopSchedulerType": "PARALLEL_CAPACITY_SCHEDULING",
      "amiVersion": "3.7.0",
      "bootstrapAction": "s3://Region.elasticmapreduce/bootstrap-
actions/configure-hadoop,-z,yarn.scheduler.capacity.root.queues=low
\\\\,high,-z,yarn.scheduler.capacity.root.high.capacity=40,-
z,yarn.scheduler.capacity.root.low.capacity=60"
    }
```

HadoopActivity 既存のEMRクラスターの使用

この例では、ワーカーグループと TaskRunner を使用して、既存のEMRクラスターでプログラムを実行します。次のパイプライン定義では、 を使用して次の HadoopActivity 操作を行います。

- でのみ MapReduce プログラムを実行する *myWorkerGroup* リソースの使用料金を見積もることができます。ワーカーグループの詳細については、「Task Runnerを使用した既存のリソースでの作業の実行」を参照してください。
- preActivityTaskConfig と postActivityTaskConfig を実行する

```
{
  "objects": [
    {
      "argument": [
        "-files",
        "s3://elasticmapreduce/samples/wordcount/wordSplitter.py",
        "-mapper",
        "wordSplitter.py",
        "-reducer",
        "aggregate",
        "-input",
        "s3://elasticmapreduce/samples/wordcount/input/",
        "-output",
```

```
        "s3://test-bucket/MyHadoopActivity/#{@pipelineId}/
#{format(@scheduledStartTime,'YYYY-MM-dd')}"
      ],
      "id": "MyHadoopActivity",
      "jarUri": "/home/hadoop/contrib/streaming/hadoop-streaming.jar",
      "name": "MyHadoopActivity",
      "type": "HadoopActivity"
    },
    {
      "id": "SchedulePeriod",
      "startDateTime": "start_datetime",
      "name": "SchedulePeriod",
      "period": "1 day",
      "type": "Schedule",
      "endDateTime": "end_datetime"
    },
    {
      "id": "ShellScriptConfig",
      "scriptUri": "s3://test-bucket/scripts/preTaskScript.sh",
      "name": "preTaskScriptConfig",
      "scriptArgument": [
        "test",
        "argument"
      ],
      "type": "ShellScriptConfig"
    },
    {
      "id": "ShellScriptConfig",
      "scriptUri": "s3://test-bucket/scripts/postTaskScript.sh",
      "name": "postTaskScriptConfig",
      "scriptArgument": [
        "test",
        "argument"
      ],
      "type": "ShellScriptConfig"
    },
    {
      "id": "Default",
      "scheduleType": "cron",
      "schedule": {
        "ref": "SchedulePeriod"
      },
      "name": "Default",
      "pipelineLogUri": "s3://test-bucket/logs/2015-05-22T18:02:00.343Z642f3fe415",
```

```
      "maximumRetries": "0",
      "workerGroup": "myWorkerGroup",
      "preActivityTaskConfig": {
        "ref": "preTaskScriptConfig"
      },
      "postActivityTaskConfig": {
        "ref": "postTaskScriptConfig"
      }
    }
  ]
}
```

構文

必須フィールド	説明	スロットタイプ
jarUri	で実行する Amazon S3 またはクラスターのローカルファイルシステムJAR内の の場所 HadoopActivity。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マススケジュール内の	リファレンスオブジェクト、例: "schedule":{"ref"myScheduleId"}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	スケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	EMR このジョブを実行するクラスター。	リファレンスオブジェクト、例: "runsOn": {"ref"myEmrClusterId"}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。 runsOn 値を指定して workerGroup 存在する場合、 workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
argument	に渡す引数JAR。	文字列
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかっ	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
	たリモートアクティビティを再試行することができます。	
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
hadoopQueue	アクティビティを送信する先の Hadoop スケジューラーのキュー名。	文字列
input	入力データの場所。	リファレンスオブジェクト、例：「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
mainClass	で実行JARしている のメインクラス HadoopActivity。	文字列
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
output	出力データの場所。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列
postActivityTask設定	実行するポストアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトの と引数のリストで構成されます。 Amazon S3	リファレンスオブジェクト、例：postActivityTask「Config」:{「ref"myShellScriptConfigId」}
preActivityTask設定	実行するプリアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトのと引数のリストで構成されます。 Amazon S3	リファレンスオブジェクト、例：preActivityTask「Config」:{「ref"myShellScriptConfigId」}

オプションのフィールド	説明	スロットタイプ
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref" myPreconditionId」}
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトの オブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみです。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、on demand、および timeseries です。	一覧表

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandActivity
- CopyActivity
- EmrCluster

HiveActivity

EMR クラスターで Hive クエリを実行します。 `HiveActivity` を使用すると、Amazon EMRアクティビティの設定が容易になり、Amazon S3 または Amazon から受信した入力データに基づいて Hive テーブルが自動的に作成されますRDS。指定する必要があるのは、ソースデータで実行する HiveQL だけです。 は`${input2}`、 `HiveActivity` オブジェクトの入力フィールドに基づいて、`${input1}`、 などを使用して Hive テーブル AWS Data Pipeline を自動的に作成します。

Amazon S3 入力の場合、`dataFormat` フィールドを使用して Hive の列名が作成されます。

MySQL (Amazon RDS) 入力の場合、SQLクエリの列名を使用して Hive 列名が作成されます。

> **Note**
>
> このアクティビティでは、Hive CSV Serde を使用します。

例

以下は、このオブジェクト型の例です。このオブジェクトは、同じパイプライン定義ファイルで定義した他のオブジェクトを 3 つ参照します。MySchedule は Schedule オブジェクトで、MyS3Input と MyS3Output はデータノードオブジェクトです。

```
{
  "name" : "ProcessLogData",
  "id" : "MyHiveActivity",
  "type" : "HiveActivity",
  "schedule" : { "ref": "MySchedule" },
  "hiveScript" : "INSERT OVERWRITE TABLE ${output1} select
 host,user,time,request,status,size from ${input1};",
  "input" : { "ref": "MyS3Input" },
  "output" : { "ref": "MyS3Output" },
  "runsOn" : { "ref": "MyEmrCluster" }
}
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。このオブジェクトの依存関係の実行順序を設定するために、別のオブジェクトへのスケジュール参照を指定します。この要件を満たすには、オブジェクトに明示的にスケジュールを設定します。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マススケジュール内のスケジュール) がある場合は、スケジュール参照がある親オブジェクトを作成できます。オプションのスケジュール設定の例については、「https://	リファレンスオブジェクト、例: "schedule":{"ref"mScheduleId"}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
hiveScript	実行する Hive スクリプト。	文字列
scriptUri	実行する Hive スクリプトの場所 (s3:// など scriptLocation) 。	文字列

必須のグループ	説明	スロットタイプ
runsOn	これがHiveActivity 実行されるEMRクラスター。	リファレンスオブジェクト、例: "runsOn": {"ref"myEmrClusterId"}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。値 runsOn を指定して、workerGroup がある場合、workerGroup は無視されます。	文字列
input	入力データソース。	「input」:{「ref"myDataNodeId」} などのリファレンスオブジェクト
output	出力データソース。	「output」:{「ref"myDataNodeId

必須のグループ	説明	スロットタイプ
		」} などのリファレンスオブジェクト

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	dependsOn「」: {「ref"myActivityId」} などのリファレンスオブジェクト
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
hadoopQueue	ジョブを送信する先の Hadoop スケジューラーのキュー名。	文字列
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数

オプションのフィールド	説明	スロットタイプ
onFail	現在のオブジェクトが失敗したときに実行するアクション。	onFail「」:{「ref"myActionId」} などのリファレンスオブジェクト
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	onLateAction「」:{「ref"myActionId」} などのリファレンスオブジェクト
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	onSuccess「」:{「ref"myActionId」} などのリファレンスオブジェクト
parent	スロットの継承元となる現在のオブジェクトの親。	「parent」:{「ref"myBaseObjectId」} などのリファレンスオブジェクト
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など)。	文字列
postActivityTask設定	実行するポストアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトの と引数のリストで構成されます。 Amazon S3	postActivityTask「Config」:{「ref"myShellScriptConfigId」} などのリファレンスオブジェクト
preActivityTask設定	実行するプリアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトのと引数のリストで構成されます。 Amazon S3	preActivityTask「Config」:{「ref"myShellScriptConfigId」} などのリファレンスオブジェクト

オプションのフィールド	説明	スロットタイプ
precondition	オプションで前提条件を定義します。すべての前提条件が満たされREADYるまで、データノードは「」とマークされません。	「precondition」:{「ref"myPreconditionId」} などのリファレンスオブジェクト
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
resizeClusterBefore実行中	入力または出力として指定された DynamoDB データノードに対応するため、このアクティビティを実行する前にクラスターのサイズを変更します。 Note アクティビティで を入力データノードまたは出力データノードDynamoDBDataNode として使用し、`resizeClusterBeforeRunning` を に設定するとTRUE、 はm3.xlarge インスタンスタイプを使用して AWS Data Pipeline 開始します。これによりインスタンスタイプの選択が `m3.xlarge` に上書きされるため、月別コストが増加することがあります。	ブール値
resizeClusterMaxインスタンス	サイズ変更アルゴリズムによってリクエストできるインスタンスの最大数の制限。	整数
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみである必要があります。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表

オプションのフィールド	説明	スロットタイプ
scriptVariable	スクリプトの実行中に Amazon EMRが Hive に渡すスクリプト変数を指定します。例えば、次のスクリプト変数の例では、 SAMPLEおよび FILTER_DATE 変数を Hive SAMPLE=s3://elasticmapreduce/samples/hive-ads に渡します。 FILTER_DATE=#{format(@scheduledStartTime,'YYYY-MM-dd')}% このフィールドでは、複数の値を指定でき、script フィールドと scriptUri フィールドの両方で機能します。さらに scriptVariable は stage の設定が true か false かに関係なく機能します。このフィールドは、 AWS Data Pipeline の式と関数を利用して Hive に動的な値を送信するときに特に便利です。	文字列
ステージ	スクリプトの実行前または後に、ステージングが有効かどうかを決定します。Hive 11 では許可されていないため、Amazon EMRAMIバージョン 3.2.0 以降を使用してください。	ブール値

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	activeInstances「」:{「ref"myRunnableObjectId」} などのリファレンスオブジェクト
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime

実行時フィールド	説明	スロットタイプ
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	cascadeFailedOn「」: {「ref"myRunnableObjectId」} などのリファレンスオブジェクト
emrStepLog	Amazon EMR ステップログは、EMRアクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime

実行時フィールド	説明	スロットタイプ
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	waitingOn「」:{「ref" myRunnableObjectId 」} などのリファレンスオブジェクト

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー。	文字列

システムフィールド	説明	スロットタイプ
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandActivity
- EmrActivity

HiveCopyActivity

EMR クラスターで Hive クエリを実行します。 HiveCopyActivityを使用すると、DynamoDB テーブル間でデータを簡単にコピーできます。 は、列および行レベルで DynamoDB からの入力データをフィルタリングする HiveQL ステートメントHiveCopyActivityを受け入れます。

例

以下の例は、HiveCopyActivity と DynamoDBExportDataFormat を使用して、タイムスタンプに基づいてデータをフィルタリングしながら DynamoDBDataNode 間でデータをコピーする方法を示します。

```
{
  "objects": [
    {
      "id" : "DataFormat.1",
      "name" : "DataFormat.1",
      "type" : "DynamoDBExportDataFormat",
      "column" : "timeStamp BIGINT"
    },
    {
      "id" : "DataFormat.2",
      "name" : "DataFormat.2",
      "type" : "DynamoDBExportDataFormat"
```

```
    },
    {
      "id" : "DynamoDBDataNode.1",
      "name" : "DynamoDBDataNode.1",
      "type" : "DynamoDBDataNode",
      "tableName" : "item_mapped_table_restore_temp",
      "schedule" : { "ref" : "ResourcePeriod" },
      "dataFormat" : { "ref" : "DataFormat.1" }
    },
    {
      "id" : "DynamoDBDataNode.2",
      "name" : "DynamoDBDataNode.2",
      "type" : "DynamoDBDataNode",
      "tableName" : "restore_table",
      "region" : "us_west_1",
      "schedule" : { "ref" : "ResourcePeriod" },
      "dataFormat" : { "ref" : "DataFormat.2" }
    },
    {
      "id" : "EmrCluster.1",
      "name" : "EmrCluster.1",
      "type" : "EmrCluster",
      "schedule" : { "ref" : "ResourcePeriod" },
      "masterInstanceType" : "m1.xlarge",
      "coreInstanceCount" : "4"
    },
    {
      "id" : "HiveTransform.1",
      "name" : "Hive Copy Transform.1",
      "type" : "HiveCopyActivity",
      "input" : { "ref" : "DynamoDBDataNode.1" },
      "output" : { "ref" : "DynamoDBDataNode.2" },
      "schedule" :{ "ref" : "ResourcePeriod" },
      "runsOn" : { "ref" : "EmrCluster.1" },
      "filterSql" : "`timeStamp` > unix_timestamp(\"#{@scheduledStartTime}\", \"yyyy-
MM-dd'T'HH:mm:ss\")"
    },
    {
      "id" : "ResourcePeriod",
      "name" : "ResourcePeriod",
      "type" : "Schedule",
      "period" : "1 Hour",
      "startDateTime" : "2013-06-04T00:00:00",
      "endDateTime" : "2013-06-04T01:00:00"
```

```
    }
  ]
}
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	リファレンスオブジェクト、例: "schedule":{"ref"mayScheduleId"}

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	実行するクラスターを指定します。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。値 `runsOn` を指定して、`workerGroup` がある場合、`workerGroup` は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
filterSql	コピーする DynamoDB または Amazon S3 データのサブセットをフィルタリングする Hive SQLステートメントフラグメント。によって自動的に追加されるため、フィルター	文字列

オプションのフィールド	説明	スロットタイプ
	には述語のみが含まれ、 WHERE句で始 AWS Data Pipeline まることはできません。	
input	入力データソース。S3DataNode または DynamoDBDataNode を指定する必要があります。DynamoDBNode を使用する場合は、DynamoDBExportDataFormat を指定します。	リファレンスオブジェクト、例：「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
output	出力データソース。入力が S3DataNode の場合、これは DynamoDBDataNode を指定する必要があります。それ以外の場合は、S3DataNode または DynamoDBDataNode を指定できます。DynamoDBNode を使用する場合は、DynamoDBExportDataFormat を指定します。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロード 's3://BucketName/Key/' するための URIなどの Amazon S3 。	文字列
postActivityTask設定	実行するポストアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトの と引数のリストで構成されます。 Amazon S3	リファレンスオブジェクト、例：postActivityTask「Config」:{「ref"myShellScriptConfigId」}
preActivityTask設定	実行するプリアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトのと引数のリストで構成されます。 Amazon S3	リファレンスオブジェクト、例：preActivityTask「Config」:{「ref"myShellScriptConfigId」}
precondition	オプションで前提条件を定義します。すべての前提条件が満たされREADYるまで、データノードは「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}

オプションのフィールド	説明	スロットタイプ
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
resizeClusterBefore実行中	入力または出力として指定された DynamoDB データノードに対応するため、このアクティビティを実行する前にクラスターのサイズを変更します。 **Note** アクティビティで を入力データノードまたは出力データノードDynamoDBDataNode として使用し、 `resizeClusterBeforeRunning` を に設定するとTRUE、 は`m3.xlarge` インスタンスタイプを使用して AWS Data Pipeline 開始します。これによりインスタンスタイプの選択が `m3.xlarge` に上書きされるため、月別コストが増加することがあります。	ブール値
resizeClusterMaxインスタンス	サイズ変更アルゴリズムによってリクエストできるインスタンスの最大数の制限	整数
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみである必要があります。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime

実行時フィールド	説明	スロットタイプ
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	Amazon EMR ステップログは、EMRアクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列

実行時フィールド	説明	スロットタイプ
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列

システムフィールド	説明	スロットタイプ
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandActivity
- EmrActivity

PigActivity

PigActivity は、 ShellCommandActivity または を使用する AWS Data Pipeline 必要なく、 で Pig スクリプトをネイティブにサポートしますEmrActivity。さらに、 はデータステージング PigActivity をサポートします。stage フィールドが true に設定されていると、 AWS Data Pipeline は Pig のスキーマとして入力データをステージングします (ユーザーがコードを追加する必要はありません) 。

例

次のパイプライン例は PigActivity の使い方を示します。パイプライン例では、以下のステップを実行します。

- MyPigActivity1 は Amazon S3 からデータをロードし、いくつかの列のデータを選択して Amazon S3 にアップロードする Pig スクリプトを実行します。
- MyPigActivity2 は最初の出力をロードし、数列と 3 行のデータを選択し、2 番目の出力として Amazon S3 にアップロードします。
- MyPigActivity3 は 2 番目の出力データをロードし、2 行のデータと「fifth」という名前の列のみを Amazon に挿入しますRDS。
- MyPigActivity4 は Amazon RDSデータをロードし、データの最初の行を選択して Amazon S3 にアップロードします。

```
{
  "objects": [
```

```
    {
      "id": "MyInputData1",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "directoryPath": "s3://example-bucket/pigTestInput",
      "name": "MyInputData1",
      "dataFormat": {
        "ref": "MyInputDataType1"
      },
      "type": "S3DataNode"
    },
    {
      "id": "MyPigActivity4",
      "scheduleType": "CRON",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "input": {
        "ref": "MyOutputData3"
      },
      "pipelineLogUri": "s3://example-bucket/path/",
      "name": "MyPigActivity4",
      "runsOn": {
        "ref": "MyEmrResource"
      },
      "type": "PigActivity",
      "dependsOn": {
        "ref": "MyPigActivity3"
      },
      "output": {
        "ref": "MyOutputData4"
      },
      "script": "B = LIMIT ${input1} 1; ${output1} = FOREACH B GENERATE one;",
      "stage": "true"
    },
    {
      "id": "MyPigActivity3",
      "scheduleType": "CRON",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "input": {
        "ref": "MyOutputData2"
```

```
    },
    "pipelineLogUri": "s3://example-bucket/path",
    "name": "MyPigActivity3",
    "runsOn": {
      "ref": "MyEmrResource"
    },
    "script": "B = LIMIT ${input1} 2; ${output1} = FOREACH B GENERATE Fifth;",
    "type": "PigActivity",
    "dependsOn": {
      "ref": "MyPigActivity2"
    },
    "output": {
      "ref": "MyOutputData3"
    },
    "stage": "true"
  },
  {
    "id": "MyOutputData2",
    "schedule": {
      "ref": "MyEmrResourcePeriod"
    },
    "name": "MyOutputData2",
    "directoryPath": "s3://example-bucket/PigActivityOutput2",
    "dataFormat": {
      "ref": "MyOutputDataType2"
    },
    "type": "S3DataNode"
  },
  {
    "id": "MyOutputData1",
    "schedule": {
      "ref": "MyEmrResourcePeriod"
    },
    "name": "MyOutputData1",
    "directoryPath": "s3://example-bucket/PigActivityOutput1",
    "dataFormat": {
      "ref": "MyOutputDataType1"
    },
    "type": "S3DataNode"
  },
  {
    "id": "MyInputDataType1",
    "name": "MyInputDataType1",
    "column": [
```

```
        "First STRING",
        "Second STRING",
        "Third STRING",
        "Fourth STRING",
        "Fifth STRING",
        "Sixth STRING",
        "Seventh STRING",
        "Eighth STRING",
        "Ninth STRING",
        "Tenth STRING"
      ],
      "inputRegEx": "^(\\S+) (\\S+) (\\S+) (\\S+) (\\S+) (\\S+) (\\S+) (\
\S+) (\\S+) (\\S+)",
      "type": "RegEx"
    },
    {
      "id": "MyEmrResource",
      "region": "us-east-1",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "keyPair": "example-keypair",
      "masterInstanceType": "m1.small",
      "enableDebugging": "true",
      "name": "MyEmrResource",
      "actionOnTaskFailure": "continue",
      "type": "EmrCluster"
    },
    {
      "id": "MyOutputDataType4",
      "name": "MyOutputDataType4",
      "column": "one STRING",
      "type": "CSV"
    },
    {
      "id": "MyOutputData4",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "directoryPath": "s3://example-bucket/PigActivityOutput3",
      "name": "MyOutputData4",
      "dataFormat": {
        "ref": "MyOutputDataType4"
      },
```

```
      "type": "S3DataNode"
    },
    {
      "id": "MyOutputDataType1",
      "name": "MyOutputDataType1",
      "column": [
        "First STRING",
        "Second STRING",
        "Third STRING",
        "Fourth STRING",
        "Fifth STRING",
        "Sixth STRING",
        "Seventh STRING",
        "Eighth STRING"
      ],
      "columnSeparator": "*",
      "type": "Custom"
    },
    {
      "id": "MyOutputData3",
      "username": "___",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "insertQuery": "insert into #{table} (one) values (?)",
      "name": "MyOutputData3",
      "*password": "___",
      "runsOn": {
        "ref": "MyEmrResource"
      },
      "connectionString": "jdbc:mysql://example-database-instance:3306/example-
database",
      "selectQuery": "select * from #{table}",
      "table": "example-table-name",
      "type": "MySqlDataNode"
    },
    {
      "id": "MyOutputDataType2",
      "name": "MyOutputDataType2",
      "column": [
        "Third STRING",
        "Fourth STRING",
        "Fifth STRING",
        "Sixth STRING",
```

```
        "Seventh STRING",
        "Eighth STRING"
      ],
      "type": "TSV"
    },
    {
      "id": "MyPigActivity2",
      "scheduleType": "CRON",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
      "input": {
        "ref": "MyOutputData1"
      },
      "pipelineLogUri": "s3://example-bucket/path",
      "name": "MyPigActivity2",
      "runsOn": {
        "ref": "MyEmrResource"
      },
      "dependsOn": {
        "ref": "MyPigActivity1"
      },
      "type": "PigActivity",
      "script": "B = LIMIT ${input1} 3; ${output1} = FOREACH B GENERATE Third, Fourth,
Fifth, Sixth, Seventh, Eighth;",
      "output": {
        "ref": "MyOutputData2"
      },
      "stage": "true"
    },
    {
      "id": "MyEmrResourcePeriod",
      "startDateTime": "2013-05-20T00:00:00",
      "name": "MyEmrResourcePeriod",
      "period": "1 day",
      "type": "Schedule",
      "endDateTime": "2013-05-21T00:00:00"
    },
    {
      "id": "MyPigActivity1",
      "scheduleType": "CRON",
      "schedule": {
        "ref": "MyEmrResourcePeriod"
      },
```

```
      "input": {
        "ref": "MyInputData1"
      },
      "pipelineLogUri": "s3://example-bucket/path",
      "scriptUri": "s3://example-bucket/script/pigTestScipt.q",
      "name": "MyPigActivity1",
      "runsOn": {
        "ref": "MyEmrResource"
      },
      "scriptVariable": [
        "column1=First",
        "column2=Second",
        "three=3"
      ],
      "type": "PigActivity",
      "output": {
        "ref": "MyOutputData1"
      },
      "stage": "true"
    }
  ]
}
```

pigTestScript.q の内容は次のとおりです。

```
B = LIMIT ${input1} $three; ${output1} = FOREACH B GENERATE $column1, $column2, Third,
 Fourth, Fifth, Sixth, Seventh, Eighth;
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。ユーザーは、このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。ユーザーは、オブジェクトに明示的にスケジュールを設定することで、この要件を満たすことができます。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	例えば、「スケジュール」: {"ref": "DefaultSchedule"} を指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合、ユーザーは、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
script	実行する Pig スクリプト。	文字列
scriptUri	実行する Pig スクリプトの場所 (s3:// など scriptLocation) 。	文字列

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	EMR これが PigActivity 実行されるクラスター。	リファレンスオブジェクト、例 : runsOn「」:{「ref"myEmrClusterId」}

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。値 runsOn を指定して、workerGroup がある場合、workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例：dependsOn「」：{「ref"myActivityId」}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データソース。	リファレンスオブジェクト、例えば「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例：onFail「」:{「ref"myActionId」}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例：onLateAction「」:{「ref"myActionId」}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例：onSuccess「」:{「ref"myActionId」}
output	出力データソース。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのAmazon S3 URI (「s3://BucketName/Key/」など)。	文字列

オプションのフィールド	説明	スロットタイプ
postActivityTask設定	実行するポストアクティビティ設定スクリプト。これは、Amazon S33 URIのシェルスクリプトの と引数のリストで構成されます。	リファレンスオブジェクト、例：postActivityTask「Config」:{「ref"myShellScriptConfigId」}
preActivityTask設定	実行するプリアクティビティ設定スクリプト。これは、Amazon S3 URIのシェルスクリプトのと引数のリストで構成されます。 Amazon S3	リファレンスオブジェクト、例：preActivityTask「Config」:{「ref"myShellScriptConfigId」}
precondition	オプションで前提条件を定義します。すべての前提条件が満たされREADYるまで、データノードは「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref"myPreconditionId」}
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
resizeClusterBefore実行中	入力または出力として指定された DynamoDB データノードに対応するため、このアクティビティを実行する前にクラスターのサイズを変更します。 **Note** アクティビティで を入力データノードまたは出力データノードDynamoDBDataNode として使用し、 resizeClusterBeforeRunning を に設定するとTRUE、 はm3.xlarge インスタンスタイプを使用して AWS Data Pipeline 開始します。これによりインスタンスタイプの選択が m3.xlarge に上書きされるため、月別コストが増加することがあります。	ブール値
resizeClusterMaxインスタンス	サイズ変更アルゴリズムによってリクエストできるインスタンスの最大数の制限。	整数
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。[Time Series Style Scheduling] は、インスタンスが各間隔の最後にスケジュールされることを意味し、[Cron Style Scheduling] は、インスタンスが各間隔の最初にスケジュールされることを意味します。オンデマンドスケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定する必要があり、パイプライン内のオブジェクトに対して scheduleType 指定されるのは のみである必要があります。オンデマンドパイプラインを使用するには、後続の実行ごとに ActivatePipeline オペレーションを呼び出すだけです。値は、cron、ondemand、および timeseries です。	一覧表
scriptVariable	Pig スクリプトに渡す引数。スクリプトまたは scriptVariable で を使用できますscriptUri。	文字列
ステージ	ステージングを有効にするかどうかを決定し、Pig スクリプトが ${INPUT1} や ${OUTPUT1} などのステージングされたデータテーブルにアクセスできるようにします。	ブール値

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例：activeInstances「」:{「ref"myRunnableObjectId」}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例：cascadeFailedOn「」:{「ref"myRunnableObjectId」}
emrStepLog	Amazon EMR ステップログは、EMRアクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

実行時フィールド	説明	スロットタイプ
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例：waitingOn「」：{「ref"myRunnableObjectId」}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandActivity
- EmrActivity

RedshiftCopyActivity

DynamoDB または Amazon S3 から Amazon Redshift にデータをコピーします。新しいテーブルにデータをロードすることも、既存のテーブルにデータを簡単にマージすることもできます。

以下に、`RedshiftCopyActivity` を使用するユースケースの概要を示します。

1. まず AWS Data Pipeline 、 を使用して Amazon S3 でデータをステージングします。

2. RedshiftCopyActivity を使用して、Amazon RDSと Amazon から Amazon Redshift EMRにデータを移動します。

   これにより、Amazon Redshift にデータをロードして分析を行うことができます。
3. を使用してSqlActivity、Amazon Redshift にロードしたデータに対してSQLクエリを実行します。

さらに、RedshiftCopyActivity はマニフェストファイルをサポートするため、S3DataNode を操作できます。詳細については、「S3DataNode」を参照してください。

例

以下は、このオブジェクト型の例です。

形式を確実に変換するために、この例では で EMPTYASNULLとIGNOREBLANKLINES特別な変換パラメータを使用しますcommandOptions。詳細については、Amazon Redshift データベース開発者ガイドのデータ変換パラメータを参照してください。

```
{
  "id" : "S3ToRedshiftCopyActivity",
  "type" : "RedshiftCopyActivity",
  "input" : { "ref": "MyS3DataNode" },
  "output" : { "ref": "MyRedshiftDataNode" },
  "insertMode" : "KEEP_EXISTING",
  "schedule" : { "ref": "Hour" },
  "runsOn" : { "ref": "MyEc2Resource" },
  "commandOptions": ["EMPTYASNULL", "IGNOREBLANKLINES"]
}
```

以下のパイプライン定義の例では、APPEND 挿入モードを使用するアクティビティを示しています。

```
{
  "objects": [
    {
      "id": "CSVId1",
      "name": "DefaultCSV1",
      "type": "CSV"
    },
    {
      "id": "RedshiftDatabaseId1",
      "databaseName": "dbname",
      "username": "user",
```

```
      "name": "DefaultRedshiftDatabase1",
      "*password": "password",
      "type": "RedshiftDatabase",
      "clusterId": "redshiftclusterId"
    },
    {
      "id": "Default",
      "scheduleType": "timeseries",
      "failureAndRerunMode": "CASCADE",
      "name": "Default",
      "role": "DataPipelineDefaultRole",
      "resourceRole": "DataPipelineDefaultResourceRole"
    },
    {
      "id": "RedshiftDataNodeId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "tableName": "orders",
      "name": "DefaultRedshiftDataNode1",
      "createTableSql": "create table StructuredLogs (requestBeginTime CHAR(30)
PRIMARY KEY DISTKEY SORTKEY, requestEndTime CHAR(30), hostname CHAR(100), requestDate
varchar(20));",
      "type": "RedshiftDataNode",
      "database": {
        "ref": "RedshiftDatabaseId1"
      }
    },
    {
      "id": "Ec2ResourceId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "securityGroups": "MySecurityGroup",
      "name": "DefaultEc2Resource1",
      "role": "DataPipelineDefaultRole",
      "logUri": "s3://myLogs",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "type": "Ec2Resource"
    },
    {
      "id": "ScheduleId1",
      "startDateTime": "yyyy-mm-ddT00:00:00",
      "name": "DefaultSchedule1",
```

```
      "type": "Schedule",
      "period": "period",
      "endDateTime": "yyyy-mm-ddT00:00:00"
    },
    {
      "id": "S3DataNodeId1",
      "schedule": {
        "ref": "ScheduleId1"
      },
      "filePath": "s3://datapipeline-us-east-1/samples/hive-ads-samples.csv",
      "name": "DefaultS3DataNode1",
      "dataFormat": {
        "ref": "CSVId1"
      },
      "type": "S3DataNode"
    },
    {
      "id": "RedshiftCopyActivityId1",
      "input": {
        "ref": "S3DataNodeId1"
      },
      "schedule": {
        "ref": "ScheduleId1"
      },
      "insertMode": "APPEND",
      "name": "DefaultRedshiftCopyActivity1",
      "runsOn": {
        "ref": "Ec2ResourceId1"
      },
      "type": "RedshiftCopyActivity",
      "output": {
        "ref": "RedshiftDataNodeId1"
      }
    }
  ]
}
```

APPEND オペレーションは、プライマリキーまたはソートキーにかかわらず、テーブルに項目を追加します。たとえば、以下のテーブルがあるとすると、同じ ID とユーザー値のレコードを追加できます。

```
ID(PK)     USER
1          aaa
```

```
2           bbb
```

以下のように、同じ ID とユーザー値のレコードを追加できます。

```
ID(PK)     USER
1          aaa
2          bbb
1          aaa
```

> **Note**
>
> APPEND オペレーションが中断されて再試行される場合、結果として再実行されるパイプラインは、最初から追加される可能性があります。これにより、重複するレコードが追加される可能性があるため、行数をカウントするロジックがある場合は、この動作に注意する必要があります。

チュートリアルについては、「AWS Data Pipeline を使用した Amazon Redshift へのデータのコピー」を参照してください。

構文

必須フィールド	説明	スロットタイプ
insertMode	ロードするデータ内の行と重複するターゲットテーブル内の AWS Data Pipeline 既存のデータの処理を決定します。 有効な値は、KEEP_EXISTING 、OVERWRITE_EXISTING 、TRUNCATE、APPEND です。 KEEP_EXISTING を指定すると、既存の行を変更することなく、新しい行がテーブルに追加されます。 KEEP_EXISTING および OVERWRITE_EXISTING はプライマリキー、ソート、およびディストリビューションキーを使用して、既存の行と一致する受信行を識別しま	一覧表

必須フィールド	説明	スロットタイプ
	す。Amazon Redshift データベース開発者ガイドの新しいデータの更新と挿入を参照してください。 TRUNCATE は、ターゲットテーブルのデータをすべて削除した後、新しいデータを書き込みます。 APPEND は Redshift テーブルの末尾にすべてのレコードを追加します。APPEND にはプライマリキーも、分散キーも、ソートキーも不要なため、重複する項目が追加される可能性があります。	

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。 このオブジェクトの依存関係の実行順序を設定するために、別のオブジェクトへのスケジュール参照を指定します。 ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。たとえば、`"schedule": {"ref": "DefaultSchedule"}` を指定することで、オブジェクトでスケジュールを明示的に設定できます。 パイプラインのマスタースケジュールにネストされたスケジュールがある場合、スケジュール参照がある親オブジェクトを作成します。	参照オブジェクト (`"schedule":{"ref": "myScheduleId"}` など)

オブジェクト呼び出しフィールド	説明	スロットタイプ
	オプションのスケジュール設定の例については、「スケジュール」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。runsOn 値を指定してworkerGroup 存在する場合、 workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
commandOptions	COPY オペレーションの実行時に Amazon Redshift データノードに渡すパラメータを受け取ります。パラメータの詳細については、「Amazon Redshift データベースデベロッパーガイドCOPY」の「」を参照してください。	文字列

オプションのフィールド	説明	スロットタイプ
	テーブルをロードする際に、COPY は暗黙的に文字列をターゲット列のデータ型に変換しようとします。自動的に発生するデフォルトのデータ変換に加えて、エラーが表示される場合や他の変換を必要とする場合は、追加の変換パラメータを指定できます。詳細については、Amazon Redshift データベース開発者ガイドのデータ変換パラメータを参照してください。 入力データノードまたは出力データノードにデータ形式が関連付けられている場合、指定されたパラメータは無視されます。 コピーオペレーションがまず COPY を使用して、ステージングテーブルにデータを挿入してから、INSERT コマンドを使用して、ステージングテーブルからターゲットテーブルにデータをコピーするため、COPY の一部のパラメータは適用されません。たとえば、COPY コマンドでテーブルの自動圧縮を有効にするパラメーターは適用されません。圧縮が必要な場合は、CREATE TABLE ステートメントに列エンコードの詳細を追加します。 また、場合によっては、Amazon Redshift クラスターからデータをアンロードし、Amazon S3 でファイルを作成する必要があるときに、RedshiftCopyActivity は Amazon Redshift からの UNLOAD オペレーションに依存します。 コピーおよびアンロード中のパフォーマンスを向上させるには、UNLOAD コマンドで PARALLEL OFF パラメータを指定します。	

オプションのフィールド	説明	スロットタイプ
	パラメータの詳細については、「Amazon Redshift データベースデベロッパーガイドUNLOAD」の「」を参照してください。	
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	参照オブジェクト: `"dependsOn": {"ref":"myActivityId"}`
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データノード。データソースは、Amazon S3、DynamoDB、または Amazon Redshift を使用できます。	参照オブジェクト: `"input": {"ref":"myDataNodeId"}`
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	参照オブジェクト: `"onFail": {"ref":"myActionId"}`

オプションのフィールド	説明	スロットタイプ
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	参照オブジェクト: `"onLateAction": {"ref":"myActionId"}`
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	参照オブジェクト: `"onSuccess": {"ref":"myActionId"}`
output	出力データノード。出力場所は、Amazon S3 または Amazon Redshift を使用できます。	参照オブジェクト: `"output": {"ref":"myDataNodeId"}`
parent	スロットの継承元となる現在のオブジェクトの親。	参照オブジェクト: `"parent": {"ref":"myBaseObjectId"}`
pipelineLogUri	パイプラインのログをアップロードするための S3 URI (「s3://BucketName/Key/」など) 。	文字列
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされるまで「READY」とマークされません。	参照オブジェクト: `"precondition": {"ref":"myPreconditionId"}`

オプションのフィールド	説明	スロットタイプ
キュー	同時発生した複数アクティビティの割り当てと優先順位付けをキュー内の位置に基づいて行うことができる、Amazon Redshift の `query_group` 設定に相当します。 Amazon Redshift では、同時接続数が 15 に制限されています。詳細については、「Amazon RDS Database デベロッパーガイド」の「キューへのクエリの割り当て」を参照してください。	文字列
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。 設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	パイプライン内のオブジェクトのスケジュールを指定できます。値は、cron、ondemand、および timeseries です。 timeseries スケジューリングは、インスタンスが各間隔の最後にスケジュールされることを意味します。 Cron スケジューリングは、インスタンスが各間隔の最初にスケジュールされることを意味します。 ondemand スケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。 ondemand パイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。 ondemand スケジュールを使用する場合は、デフォルトオブジェクトで指定し、パイプラインのオブジェクトに対して指定される唯一の scheduleType である必要があります。	一覧表

オプションのフィールド	説明	スロットタイプ
transformSql	入力データの変換に使用される SQL SELECT 式。 transformSql 式を staging という名前のテーブルで実行します。 DynamoDB または Amazon S3 からデータをコピーすると、 AWS Data Pipeline によって「staging」という名前のテーブルが作成され、データがあらかじめロードされます。このテーブルのデータは、ターゲットテーブルの更新に使用されます。 transformSql の出力スキーマは最終的なターゲットテーブルのスキーマと一致する必要があります。 transformSql オプションを指定すると、指定されたSQLステートメントから 2 番目のステージングテーブルが作成されます。この 2 番目のステージングテーブルからのデータが、最終的なターゲットテーブルで更新されます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	参照オブジェクト: "activeInstances": {"ref":"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime

実行時フィールド	説明	スロットタイプ
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	参照オブジェクト: `"cascadeFailedOn": {"ref":"myRunnableObjectId"}`
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime

実行時フィールド	説明	スロットタイプ
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	参照オブジェクト: `"waitingOn": {"ref":"myRunnableObjectId"}`

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー。	文字列

システムフィールド	説明	スロットタイプ
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球です。ライフサイクルにおける場所を示します。たとえば、コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

ShellCommandActivity

コマンドまたはスクリプトを実行します。ShellCommandActivity を使用して、時系列で、または cron 風に、スケジュールしたタスクを実行できます。

stage フィールドが true に設定されS3DataNode、 で使用されると、 はデータのステージングの概念ShellCommandActivityをサポートします。つまり、Amazon S3 から Amazon EC2やローカル環境などのステージロケーションにデータを移動し、スクリプトと を使用してデータに対して作業を実行しShellCommandActivity、Amazon S3 に戻すことができます。

この場合、シェルコマンドを入力 S3DataNode に接続すると、シェルスクリプトは、ShellCommandActivity の入力フィールドを参照する ${INPUT1_STAGING_DIR}、${INPUT2_STAGING_DIR} などのフィールドを使用して直接データを操作できます。

同様に、シェルコマンドの出力も、${OUTPUT1_STAGING_DIR}、${OUTPUT2_STAGING_DIR} などで参照される出力ディレクトリにステージングして、自動的に Amazon S3 に戻すことができます。

これらの式は、コマンドライン引数としてシェルコマンドに渡して、データ変換ロジックで使用することができます。

ShellCommandActivity では、Linux 形式のエラーコードと文字列が返されます。ShellCommandActivity の結果がエラーであれば、返される error はゼロ以外の値になります。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "CreateDirectory",
  "type" : "ShellCommandActivity",
  "command" : "mkdir new-directory"
}
```

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、schedule 期間の実行中に呼び出されます。 このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへの schedule 参照を指定します。 この要件を満たすには、オブジェクトで明示的に schedule を設定します。たとえば、"schedule": {"ref": "DefaultSchedule"}　を指定します。 ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、schedule 参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。パイプラインがスケジュールのツリー (マススケジュール内のスケジュール) で構成されている場合は、スケジュール参照がある親オブジェクトを作成します。 負荷を分散するために、 は物理オブジェクトをスケジュールより少し前に AWS Data Pipeline 作成しますが、スケジュールに従って実行します。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
コマンド	実行するコマンド。$ を使用して位置指定パラメータを参照し、scriptArgument を使用してコマンドのパラメータを指定します。この値および関連するパラメーターは、Task Runner を実行している環境で機能する必要があります。	文字列
scriptUri	シェルコマンドとしてダウンロードして実行するファイルの Amazon S3 URIパス。指定できる scriptUri または command フィールドは 1 つだけです。scriptUri はパラメータを使用できません。代わりに command を使用します。	文字列

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	Amazon EC2インスタンスや Amazon EMRクラスターなど、アクティビティやコマンドを実行する計算リソース。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
workerGroup	ルーティングタスクに使用されます。値 runsOn を指定して、workerGroup がある場合、workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、指定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データの場所。	リファレンスオブジェクト、例：「input」:{「ref"myDataNodeId」}
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
output	出力データの場所。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードする `'s3://BucketName/Key/'` URIなどの Amazon S3 。	文字列

オプションのフィールド	説明	スロットタイプ
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされるまで「」とマークされません。READY	リファレンスオブジェクト、例：「precondition」:{「ref" myPreconditionId」}
reportProgressTimeout	リモートアクティビティによる `reportProgress` への連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。 値は、cron、ondemand、および timeseries です。 timeseries に設定された場合、インスタンスは各間隔の最後にスケジュールされます。 Cron に設定された場合、インスタンスは各間隔の最初にスケジュールされます。 ondemand に設定された場合、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。ondemand スケジュールを使用する場合は、パイプラインのオブジェクトに対する唯一の scheduleType として、デフォルトオブジェクトで指定します。ondemand パイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。	一覧表

オプションのフィールド	説明	スロットタイプ
scriptArgument	コマンドで指定されたコマンドに渡す文字列の JSON形式の配列。たとえば、コマンドが echo $1 $2 であれば、scriptArgument を "param1", "param2" のように指定できます。複数の引数およびパラメータがある場合は、次のように scriptArgument を渡すことができます。 "scriptArgument":"arg1","scriptArgument":"param1","scriptArgument":"arg2","scriptArgument":"param2" scriptArgument は command でのみ使用できます。scriptUri で使用するとエラーが発生します。	文字列
ステージ	ステージングが有効かどうか決定し、ステージングされたデータ変数 (${INPUT1_STAGING_DIR} や ${OUTPUT1_STAGING_DIR} など) にシェルコマンドがアクセスできるようにします。	ブール値
stderr	コマンドからリダイレクトされたシステムエラーメッセージを受け取る パス。runsOn フィールドを使用する場合、アクティビティを実行するリソースが一時的であるため、これは Amazon S3 パスにする必要があります。ただし、workerGroup フィールドを指定した場合は、ローカルファイルパスを指定できます。	文字列

オプションのフィールド	説明	スロットタイプ
stdout	コマンドからリダイレクトされた出力を受け取る Amazon S3 パス。runsOn フィールドを使用する場合、アクティビティを実行するリソースが一時的であるため、これは Amazon S3 パスにする必要があります。ただし、workerGroup フィールドを指定した場合は、ローカルファイルパスを指定できます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクトがキャンセルされた場合の cancellationReason 。	文字列
@cascadeFailedOn	オブジェクトの失敗の原因となった依存関係の説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	Amazon EMR ステップログは、Amazon EMR アクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗した場合は errorId。	文字列

実行時フィールド	説明	スロットタイプ
errorMessage	このオブジェクトが失敗した場合は `errorMessage` 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	Amazon EMRベースのアクティビティの試行で利用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime

実行時フィールド	説明	スロットタイプ
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	オブジェクトのステータス。	文字列
@version	オブジェクトの作成に使用される AWS Data Pipeline バージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	ライフサイクル内のオブジェクトの場所です。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- CopyActivity
- EmrActivity

SqlActivity

データベースでSQLクエリ (スクリプト) を実行します。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MySqlActivity",
  "type" : "SqlActivity",
  "database" : { "ref": "MyDatabaseID" },
  "script" : "SQLQuery" | "scriptUri" : s3://scriptBucket/query.sql,
  "schedule" : { "ref": "MyScheduleID" },
}
```

構文

必須フィールド	説明	スロットタイプ
データベース	指定されたSQLスクリプトを実行するデータベース。	リファレンスオブジェクト、例：「database」:{「ref"myDatabaseId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定する必要があります。たとえば `"schedule": {"ref": "DefaultSchedule"}` を指定することで、オブジェクトでスケジュールを明示的に設定できます。 ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。	リファレンスオブジェクト、例：「schedule」:{「ref"myScheduleId」}

オブジェクト呼び出しフィールド	説明	スロットタイプ
	マスタースケジュール内にネストされたスケジュールのツリーがパイプラインにある場合、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
script	実行するSQLスクリプト。スクリプトまたはを指定する必要がありますscriptUri。スクリプトが Amazon S3 に保存されている場合、スクリプトは式として評価されません。スクリプトが Amazon S3 に保存されている場合、 に複数の値を指定する scriptArgument と便利です。	文字列
scriptUri	このアクティビティで実行するSQLスクリプトの場所URIを指定する 。	文字列

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
runsOn	アクティビティまたはコマンドを実行するコンピューティングリソース。例えば、Amazon EC2インスタンスや Amazon EMRクラスターなどです。	リファレンスオブジェクト、例: "runsOn": {"ref"myResourceId"}

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
workerGroup	ワーカーグループ。これはルーティングタスクに使用されます。値 runsOn を指定して、workerGroup がある場合、workerGroup は無視されます。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
dependsOn	実行可能な別のオブジェクトで依存関係を指定します。	リファレンスオブジェクト、例: "dependsOn":{"ref" myActivityId"}
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
input	入力データの場所。	リファレンスオブジェクト、例：「input」:{「ref"myDataNodeId」}
lateAfterTimeout	パイプラインのスケジュールされた開始までの期間。この期間内にオブジェクトの実行が開始されている必要があります。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトがまだスケジュールされていないか、「」で指定されたパイプラインのスケジュールされた開始からの期間内にまだ完了していない場合にトリガーされるアクションIateAfterTimeout。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
output	出力データの場所。これは、スクリプト内から参照する場合 (など#{output.tablename}) 、および出力データノードcreateTableSqlで「」を設定して出力テーブルを作成する場合にのみ便利です。SQL クエリの出力は出力データノードに書き込まれません。	リファレンスオブジェクト、例：「output」:{「ref"myDataNodeId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
pipelineLogUri	パイプラインのログをアップロードするためのS3 URI (「s3://BucketName/Key/」など) 。	文字列

オプションのフィールド	説明	スロットタイプ
precondition	オプションで前提条件を定義します。データノードは、すべての前提条件が満たされREADYるまで「」とマークされません。	リファレンスオブジェクト、例：「precondition」:{「ref" myPreconditionId」}
キュー	[Amazon Redshift のみ] 同時発生した複数アクティビティの割り当てと優先順位付けをキュー内の位置に基づいて行うことができる、Amazon Redshift クエリグループ設定に相当します。Amazon Redshift では、同時接続数が 15 に制限されています。詳細については、Amazon Redshift データベース開発者ガイドの「キューへのクエリの割り当て」を参照してください。	文字列
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプによって、パイプライン定義のオブジェクトを、期間の最初にスケジュールするか、最後にスケジュールするかを指定できます。値は、cron、ondemand、および timeseries です。 timeseries スケジューリングは、インスタンスが各間隔の最後にスケジュールされることを意味します。 cron スケジューリングは、インスタンスが各間隔の最初にスケジュールされることを意味します。 ondemand スケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。つまり、パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。ondemand スケジュールを使用する場合は、デフォルトオブジェクトで指定し、パイプラインのオブジェクトに対して指定される唯一の scheduleType である必要があります。ondemand パイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。	一覧表
scriptArgument	スクリプトの変数のリスト。または、式を直接スクリプトフィールドに指定することもできます。スクリプト scriptArgument が Amazon S3 に保存されている場合、 の複数の値が役立ちます。例: #{format(@scheduledStartTime, "YY-MM-DD HH:MM:SS"}\n#{formatplusPeriod(@scheduledStartTime, "1 day"), "YY-MM-DD HH:MM:SS"}	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列

実行時フィールド	説明	スロットタイプ
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

リソース

AWS Data Pipeline リソースオブジェクトは次のとおりです。

オブジェクト

- Ec2Resource
- EmrCluster
- HttpProxy

Ec2Resource

パイプラインアクティビティで定義された作業を実行する Amazon EC2インスタンス。

AWS Data Pipeline は、Amazon EC2インスタンスIMDSv2の をサポートするようになりました。Amazon インスタンスは、セッション指向のメソッドを使用して、インスタンスからメタデータ情報を取得するときに認証をより適切に処理します。セッションは、Amazon EC2インスタンスで実行されているソフトウェアがローカルに保存された Amazon EC2インスタンスのメタデータと認証情報にアクセスするために使用する一連のリクエストを開始および終了します。ソフトウェアは、への簡単なHTTPPUTリクエストでセッションを開始しますIMDSv2。IMDSv2 は、Amazon EC2インスタンスで実行されているソフトウェアにシークレットトークンを返します。このトークンをパスワードとして使用して、メタデータと認証情報IMDSv2を にリクエストします。

> **Note**
>
> Amazon EC2インスタンスIMDSv2で を使用するには、デフォルトの AMIは と互換性がないため、設定を変更する必要がありますIMDSv2。次のSSMパラメータを使用して取得できる新しいAMIバージョンを指定できます: `/aws/service/ami-amazon-linux-latest/amzn-ami-hvm-x86_64-ebs`。

EC2 インスタンスを指定しない場合に が AWS Data Pipeline 作成するデフォルトの Amazon インスタンスについては、「」を参照してください AWSリージョン別のデフォルトのAmazon EC2インスタンス。

例

EC2-クラシック

> **Important**
>
> 2013 年 12 月 4 日より前に作成された AWS アカウントのみが EC2-Classic プラットフォームをサポートしています。これらのアカウントのいずれかがある場合は、 ではなく EC2-Classic ネットワークでパイプラインのEC2Resourceオブジェクトを作成するオプションがありますVPC。ですべてのパイプラインのリソースを作成することを強くお勧めしますVPCs。さらに、EC2-Classic に既存のリソースがある場合は、それらを に移行することをお勧めしますVPC。

次のオブジェクト例では、いくつかのオプションフィールドが設定された EC2-Classic にEC2インスタンスを起動します。

```
{
  "id" : "MyEC2Resource",
  "type" : "Ec2Resource",
  "actionOnTaskFailure" : "terminate",
  "actionOnResourceFailure" : "retryAll",
  "maximumRetries" : "1",
  "instanceType" : "m5.large",
  "securityGroups" : [
    "test-group",
    "default"
  ],
```

```
  "keyPair" : "my-key-pair"
}
```

EC2-VPC

次のオブジェクト例では、デフォルト以外の にEC2インスタンスを起動しVPC、いくつかのオプションフィールドを設定します。

```
{
  "id" : "MyEC2Resource",
  "type" : "Ec2Resource",
  "actionOnTaskFailure" : "terminate",
  "actionOnResourceFailure" : "retryAll",
  "maximumRetries" : "1",
  "instanceType" : "m5.large",
  "securityGroupIds" : [
    "sg-12345678",
    "sg-12345678"
  ],
  "subnetId": "subnet-12345678",
  "associatePublicIpAddress": "true",
  "keyPair" : "my-key-pair"
}
```

構文

必須フィールド	説明	スロットタイプ
resourceRole	Amazon EC2インスタンスがアクセスできるリソースを制御するIAMロール。	文字列
ロール	がEC2インスタンスの作成 AWS Data Pipeline に使用するIAMロール。	文字列

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。	参照オブジェクト ("schedule

オブジェクト呼び出しフィールド	説明	スロットタイプ
	このオブジェクトの依存関係の実行順序を設定するには、別のオブジェクトへのスケジュール参照を指定します。これには以下の 2 つの方法があります。 • パイプライン内のすべてのオブジェクトで確実にスケジュールが継承されるようにするには、オブジェクトでスケジュールを明示的に設定します: `"schedule": {"ref": "DefaultSchedule"}` 。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置すると便利です。 • マスタースケジュール内にネストされたスケジュールがパイプラインにある場合、スケジュール参照がある親オブジェクトを作成することができます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	`":{"ref": "mySchedu leId"}` など)

オプションのフィールド	説明	スロットタイプ
actionOnResource失敗	このリソースに対するリソースの失敗後に実行されるアクション。有効な値は、`"retryall"` および `"retrynone"` です。	文字列

オプションのフィールド	説明	スロットタイプ
actionOnTask失敗	このリソースに対するタスクの失敗後に実行されるアクション。有効な値は "continue" または "terminate" です。	文字列
associatePublicIpアドレス	インスタンスにパブリック IP アドレスを割り当てるかどうかを示します インスタンスがAmazon EC2または Amazon にある場合VPC、デフォルト値は ですtrue。それ以外の場合、デフォルト値は false です。	ブール値
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、指定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
availabilityZone	Amazon EC2インスタンスを起動するアベイラビリティーゾーン。	文字列
disableIMDSv1	デフォルト値は false で、 IMDSv1と の両方を有効にしますIMDSv2。true に設定するIMDSv1と、 は無効になり、 のみが提供されます。IMDSv2s	ブール値
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
httpProxy	クライアントが AWS サービスへの接続に使用するプロキシホスト。	参照オブジェクト ("httpProxy": {"ref":"myHttpProxyId"} など)

オプションのフィールド	説明	スロットタイプ
imageId	インスタンスAMIに使用する の ID。デフォルトでは、 はHVMAMI仮想化タイプ AWS Data Pipeline を使用します。AMI IDs 使用される特定の は、リージョンに基づいています。選択した を指定AMIすることで、デフォルトを上書きHVMAMIできます。AMI タイプの詳細については、「Amazon EC2ユーザーガイド」の「Linux AMI仮想化タイプ」および「Linux AMI の検索」を参照してください。	文字列
initTimeout	リソースが起動するまでの待機時間。	[Period] (期間)
instanceCount	廃止済み。	整数
instanceType	開始する Amazon EC2インスタンスのタイプ。	文字列
keyPair	キーペアの名前。キーペアを指定せずに Amazon EC2インスタンスを起動した場合、そのインスタンスにログオンすることはできません。	文字列
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
minInstanceCount	廃止済み。	整数

オプションのフィールド	説明	スロットタイプ
onFail	現在のオブジェクトが失敗したときに実行するアクション。	参照オブジェクト("onFail": {"ref":"myActionId"} など)
onLateAction	オブジェクトが予定されていないか、まだ実行中の場合にトリガーされるアクション。	参照オブジェクト("onLateAction": {"ref":"myActionId"} など)
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	参照オブジェクト("onSuccess": {"ref":"myActionId"} など)
parent	スロットの継承元となる現在のオブジェクトの親。	参照オブジェクト("parent": {"ref":"myBaseObjectId"} など)
pipelineLogUri	パイプラインのログをアップロードするためのAmazon S3 URI (など's3://BucketName/Key/')。	文字列
region	Amazon EC2インスタンスを実行するリージョンのコード。デフォルトでは、インスタンスはパイプラインと同じリージョンで実行されます。依存するデータセットと同じリージョンでインスタンスを実行することもできます。	一覧表

オプションのフィールド	説明	スロットタイプ
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
runAsUser	を実行するユーザー TaskRunner。	文字列
runsOn	このフィールドはこのオブジェクトでは使用できません。	参照オブジェクト (`"runsOn": {"ref":"myResourceId"}` など)

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプでは、パイプライン定義のオブジェクトを間隔の最初にスケジュールするか、間隔の最後に、またはオンデマンドでスケジュールするかを指定できます。 値は次のとおりです。 • timeseries 。インスタンスは各間隔の最後にスケジュールされます。 • cron。インスタンスは各間隔の最初にスケジュールされます。 • ondemand。アクティベーションごとに 1 回パイプラインを実行することができます。パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。オンデマンドスケジュールを使用する場合は、デフォルトオブジェクトで指定し、パイプラインのオブジェクトに対して指定される唯一の scheduleType である必要があります。オンデマンドパイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。	一覧表
securityGroupIds	リソースプールIDs内のインスタンスに使用する 1 つ以上の Amazon EC2 セキュリティグループの 。	文字列
securityGroups	リソースプールのインスタンスに使用する 1 つ以上の Amazon EC2 セキュリティグループ。	文字列
spotBidPrice	スポットインスタンスの 1 時間あたりの最大金額 (ドル) であり、0 より大きく 20.00 より小さい 10 進値です。	文字列

オプションのフィールド	説明	スロットタイプ
subnetId	インスタンスを起動する Amazon EC2サブネットの ID。	文字列
terminateAfter	リソースを終了するまでの時間数。	[Period] (期間)
useOnDemandOnLastAttempt	スポットインスタンスをリクエストする最後の試行では、スポットインスタンスではなくオンデマンドインスタンスのリクエストを作成します。これにより、以前の試行がすべて失敗した場合に、最後の試行は中断されません。	ブール値
workerGroup	このフィールドはこのオブジェクトでは使用できません。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	参照オブジェクト ("activeInstances": {"ref":"myRunnableObjectId"} など)
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクトがキャンセルされた場合の `cancellationReason` 。	文字列
@cascadeFailedOn	オブジェクトが失敗した依存関係のチェーンの説明。	参照オブジェクト ("cascadeFailedOn": {"ref":"m

実行時フィールド	説明	スロットタイプ
		yRunnable ObjectId"} など)
emrStepLog	ステップログは、Amazon EMRアクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗した場合はエラーID。	文字列
errorMessage	このオブジェクトが失敗した場合はエラーメッセージ。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@failureReason	リソースの失敗の理由。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	Amazon EMRアクティビティの試行で利用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime

実行時フィールド	説明	スロットタイプ
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	参照オブジェクト (`"waitingOn": {"ref":"myRunnableObjectId"}` など)

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	ライフサイクル内のオブジェクトの場所です。コンポーネントオブジェクトにより、試行オブ	文字列

システムフィールド	説明	スロットタイプ
	ジェクトを実行するインスタンスオブジェクトが発生します。	

EmrCluster

Amazon EMRクラスターの設定を表します。このオブジェクトは、EmrActivity および HadoopActivity によってクラスターを起動するために使用されます。

内容

- スケジューラ
- Amazon EMR リリースバージョン
- Amazon アクセスEMR許可
- 構文
- 例
- 以下の資料も参照してください。

スケジューラ

スケジューラーは、Hadoop クラスター内のリソースの割り当てとジョブの優先順位付けを指定する方法を提供します。管理者またはユーザーは、ユーザーおよびアプリケーションのクラス別のスケジューラーを選択できます。スケジューラーは、キューを使用してユーザーおよびアプリケーションにリソースを割り当てることができます。キューは、クラスターを作成する際に設定します。次に、特定のタイプの作業やユーザー間の優先順位を設定できます。これにより、クラスターのリソースが効率的に使用され、複数のユーザーから作業をクラスターに送信できるようになります。次の 3 種類のスケジューラーを使用できます。

- FairScheduler — 長期間にわたってリソースを均等にスケジュールしようとします。
- CapacityScheduler — キューを使用して、クラスター管理者がさまざまな優先度とリソース割り当てのキューにユーザーを割り当てることができます。
- Default – クラスターで使用されます。サイトで設定可能です。

Amazon EMR リリースバージョン

Amazon EMRリリースは、ビッグデータエコシステムからの一連のオープンソースアプリケーションです。各リリースは、クラスターの作成時に Amazon がEMRインストールおよび設定するように選択した、さまざまなビッグデータアプリケーション、コンポーネント、および機能で構成されます。リリースラベルを使用してリリースバージョンを指定します。リリースラベルの形式は `emr-x.x.x` です。例えば、`emr-5.30.0` と指定します。リリースラベル 以降に基づく Amazon EMRクラスターは、 `releaseLabel`プロパティ`emr-4.0.0`を使用して`EmrCluster`オブジェクトのリリースラベルを指定します。以前のバージョンでは、`amiVersion` プロパティを使用しています。

> **Important**
>
> リリースバージョン 5.22.0 以降を使用して作成されたすべての Amazon EMRクラスターは、署名バージョン 4 を使用して Amazon S3 へのリクエストを認証します。以前の一部のリリースバージョンでは、署名バージョン 2 を使用しています。署名バージョン 2 のサポートは中止されています。詳細については、「Amazon S3 更新 - SigV2 の非推奨期間の延長および変更」を参照してください。Signature Version 4 をサポートする Amazon EMRリリースバージョンを使用することを強くお勧めします。4.7EMR.x 以降の以前のバージョンリリースでは、 シリーズの最新リリースが Signature Version 4 をサポートするように更新されました。以前のバージョンEMRリリースを使用する場合は、 シリーズの最新リリースを使用することをお勧めします。さらに、4.7.0 EMR より前のリリースは避けてください。

考慮事項と制約事項

最新バージョンの Task Runner の使用

リリースラベルを使用して自己管理型 `EmrCluster` オブジェクトを使用している場合は、最新の Task Runner を使用します。Task Runner の詳細については、「Task Runnerの操作」を参照してください。すべての Amazon EMR設定分類にプロパティ値を設定できます。詳細については、「Amazon リリースガイド」の「アプリケーションの設定」、the section called “EmrConfiguration”「」、およびthe section called “プロパティ”「 オブジェクトリファレンス」を参照してください。 EMR

のサポート IMDSv2

以前は、 のみが AWS Data Pipeline サポートされていましたIMDSv1。IMDSv2 では、Amazon EMR 5.23.1、5.27.1、5.32 以降、および Amazon 6.2 EMR 以降 AWS Data Pipeline がサポートされるよ

うになりました。IMDSv2 は、セッション指向の方法を使用して、インスタンスからメタデータ情報を取得するときに認証をより適切に処理します。 TaskRunner-2.0 を使用してユーザー管理リソースを作成して、 をIMDSv2呼び出すようにインスタンスを設定する必要があります。

Amazon EMR 5.32 以降および Amazon EMR 6.x

Amazon 5EMR.32 以降および 6.x リリースシリーズでは、Hadoop バージョン 3.x が使用されており、Hadoop バージョン 2.x と比較して Hadoop のクラスパスの評価方法に重大な変更が加えられています。Joda-Time のような一般的なライブラリがクラスパスから削除されました。

EmrActivity または HadoopActivity が Hadoop 3.x で削除されたライブラリに依存関係を持つ Jar ファイルを実行すると、ステップがエラー `java.lang.NoClassDefFoundError` または `java.lang.ClassNotFoundException` で失敗します。これは、Amazon EMR5.x リリースバージョンを使用して問題なく実行された Jar ファイルで発生する可能性があります。

この問題を解決するには、`EmrActivity` または `HadoopActivity` を開始する前に、`EmrCluster` オブジェクトで Hadoop クラスパスに Jar ファイルの依存関係をコピーする必要があります。これを行うために bash スクリプトが提供されています。bash スクリプトは次の場所で使用できます。*`MyRegion`* は、`EmrCluster`オブジェクトが実行される AWS リージョンです。例えば、 ですus-west-2。

```
s3://datapipeline-MyRegion/MyRegion/bootstrap-actions/latest/TaskRunner/copy-jars-to-hadoop-classpath.sh
```

スクリプトを実行する方法は、 が管理するリソースで `EmrActivity`または `HadoopActivity` を実行するか、セルフマネージドリソースで を実行する AWS Data Pipeline かによって異なります。

によって管理されるリソースを使用する場合は AWS Data Pipeline、 `EmrCluster` オブジェクトbootstrapActionに を追加します。bootstrapAction は、引数としてコピーするスクリプトと Jar ファイルを指定します。`EmrCluster` オブジェクトごとに最大 255 個の bootstrapAction フィールドを追加できます。また、ブートストラップアクションがすでに存在する `EmrCluster` オブジェクトに bootstrapAction フィールドを追加できます。

このスクリプトをブートストラップアクションとして指定するには、次の構文を使用します。ここで、 `JarFileRegion` は Jar ファイルが保存されているリージョンであり、各 は *`MyJarFilen`* は、Hadoop クラスパスにコピーされる Jar ファイルの Amazon S3 の絶対パスです。デフォルトで Hadoop クラスパスにある Jar ファイルを指定しないでください。

```
s3://datapipeline-MyRegion/MyRegion/bootstrap-actions/latest/TaskRunner/copy-jars-to-hadoop-classpath.sh,JarFileRegion,MyJarFile1,MyJarFile2[, ...]
```

次の例では、Amazon S3 の 2 つの Jar ファイル (my-jar-file.jar と emr-dynamodb-tool-4.14.0-jar-with-dependencies.jar) をコピーするブートストラップアクションを指定します。この例で使用しているリージョンは、us-west-2 です。

```
{
  "id" : "MyEmrCluster",
  "type" : "EmrCluster",
  "keyPair" : "my-key-pair",
  "masterInstanceType" : "m5.xlarge",
  "coreInstanceType" : "m5.xlarge",
  "coreInstanceCount" : "2",
  "taskInstanceType" : "m5.xlarge",
  "taskInstanceCount": "2",
  "bootstrapAction" : ["s3://datapipeline-us-west-2/us-west-2/bootstrap-actions/
latest/TaskRunner/copy-jars-to-hadoop-classpath.sh,us-west-2,s3://path/to/my-jar-
file.jar,s3://dynamodb-dpl-us-west-2/emr-ddb-storage-handler/4.14.0/emr-dynamodb-
tools-4.14.0-jar-with-dependencies.jar"]
}
```

新しい bootstrapAction への変更を有効にするには、パイプラインを保存してアクティブ化する必要があります。

セルフマネージドリソースを使用する場合は、スクリプトをクラスターインスタンスにダウンロードし、 を使用してコマンドラインから実行できますSSH。スクリプトは、/etc/hadoop/conf/shellprofile.d という名前のディレクトリを作成し、そのディレクトリ内に datapipeline-jars.sh という名前のファイルを作成します。コマンドライン引数として指定された jar ファイルは、スクリプトで作成された /home/hadoop/datapipeline_jars という名前のディレクトリにコピーされます。クラスターの設定が異なる場合は、ダウンロード後にスクリプトを適切に変更してください。

コマンドラインでスクリプトを実行するための構文は、前の例で示した bootstrapAction の使用と若干異なります。次の例に示すように、引数の間にはカンマではなくスペースを使用します。

```
./copy-jars-to-hadoop-classpath.sh us-west-2 s3://path/to/my-jar-file.jar s3://
dynamodb-dpl-us-west-2/emr-ddb-storage-handler/4.14.0/emr-dynamodb-tools-4.14.0-jar-
with-dependencies.jar
```

Amazon アクセスEMR許可

カスタムIAMロールを作成するときは、クラスターが作業を実行するために必要な最小限のアクセス許可を慎重に検討してください。Amazon S3 内のファイルや Amazon 、Amazon

RedshiftRDS、DynamoDB 内のデータなど、必要なリソースへのアクセスを必ず許可してください。visibleToAllUsers を False に設定する場合は、そのために必要なアクセス権限がロールに必要です。DataPipelineDefaultRole には、これらのアクセス権限がないことに注意してください。EmrCluster オブジェクトのロールとして DefaultDataPipelineResourceRole と DataPipelineDefaultRole のロールの結合を指定するか、この目的で独自のロールを作成する必要があります。

構文

オブジェクト呼び出しフィールド	説明	スロットタイプ
schedule	このオブジェクトは、スケジュール期間の実行中に呼び出されます。このオブジェクトの依存関係の実行順序を設定するために、別のオブジェクトへのスケジュール参照を指定します。この要件を満たすには、オブジェクトのスケジュールを明示的に設定できます。たとえば、"schedule": {"ref": "DefaultSchedule"} と指定します。ほとんどの場合、すべてのオブジェクトがそのスケジュールを継承するように、スケジュール参照をデフォルトのパイプラインオブジェクトに配置することをお勧めします。または、パイプラインにスケジュールのツリー (マスタースケジュール内のスケジュール) がある場合は、スケジュール参照がある親オブジェクトを作成できます。オプションのスケジュール設定の例については、「https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-object-schedule.html」を参照してください。	参照オブジェクト ("schedule":{"ref": "myScheduleId"} など)

オプションのフィールド	説明	スロットタイプ
actionOnResource失敗	このリソースに対するリソースの失敗後に実行されるアクション。有効な値は `retryall` (指定した期間内にクラスターに対してすべてのタスクを再試行する) と `retrynone` です。	文字列
actionOnTask失敗	このリソースに対するタスクの失敗後に実行されるアクション。有効な値は "continue" (クラスターを終了しない) と "terminate" です。	文字列
additionalMasterSecurityGroupIds	EMR クラスターの追加のマスターセキュリティグループの識別子で、sg-01 の形式に従いますXXXX6a。詳細については、「Amazon 管理ガイド」の「Amazon EMR Additional Security Groups」を参照してください。 EMR	文字列
additionalSlaveSecurityGroupIds	EMR クラスターの追加のスレーブセキュリティグループの識別子で、 の形式に従いますsg-01XXXX6a 。	文字列
amiVersion	Amazon がクラスターノードのインストール EMRに使用する Amazon マシンイメージ (AMI) バージョン。詳細については、「Amazon EMR管理ガイド」を参照してください。	文字列
applications	カンマ区切りの引数を指定してクラスターにインストールするアプリケーション。Hive と Pig がデフォルトでインストールされます。このパラメータは Amazon EMRバージョン 4.0 以降にのみ適用されます。	文字列
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかっ	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
	たリモートアクティビティを再試行することができます。	
availabilityZone	クラスターを実行するアベイラビリティーゾーン。	文字列
bootstrapAction	クラスターの開始時に実行するアクション。カンマで区切って引数を指定できます。アクションを複数 (最大 255 個) 指定するには、`bootstrapAction` フィールドを複数追加します。ブートストラップアクションを使用しないでクラスターを開始するのが、デフォルトの動作です。	文字列
設定	Amazon EMRクラスターの設定。このパラメータは Amazon EMRバージョン 4.0 以降にのみ適用されます。	参照オブジェクト (`"configuration":{"ref":"myEmrConfigurationId"}` など)
coreInstanceBid料金	が Amazon EC2インスタンスに対して支払うことができる最大スポット料金。入札価格が指定されている場合、Amazon はインスタンスグループにスポットインスタンスEMRを使用します。で指定しますUSD。	文字列
coreInstanceCount	クラスターに使用するコアノードの数。	整数
coreInstanceType	コアノードに使用する Amazon EC2インスタンスのタイプ。「Amazon EMR クラスターでサポートされるAmazon EC2インスタンス」を参照してください。	文字列

オプションのフィールド	説明	スロットタイプ
coreGroupConfiguration	Amazon EMRクラスターコアインスタンスグループの設定。このパラメータは Amazon EMR バージョン 4.0 以降にのみ適用されます。	参照オブジェクト (“configuration”: {“ref”: “myEmrConfigurationId”} など)
coreEbsConfiguration	Amazon EMRクラスターのコアグループ内の各コアノードにアタッチされる Amazon EBSボリュームの構成。詳細については、「Amazon EC2 ユーザーガイド」のEBS「最適化をサポートするインスタンスタイプ」を参照してください。	参照オブジェクト (“coreEbsConfiguration”: {“ref”: “myEbsConfiguration”} など)
customAmiId	Amazon EMRリリースバージョン 5.7.0 以降にのみ適用されます。Amazon が Amazon EC2インスタンスをEMRプロビジョニングするときにAMI使用するカスタムの AMI ID を指定します。クラスターノード設定をカスタマイズするために、ブートストラップアクションの代わりに使用することもできます。詳細については、「Amazon EMR管理ガイド」の次のトピックを参照してください。カスタム の使用 AMI	文字列

オプションのフィールド	説明	スロットタイプ
EbsBlockDeviceConfig	インスタンスグループに関連付けられたリクエストされた Amazon EBSブロックデバイスの構成。指定したボリューム数をインスタンスグループ内の各インスタンスに関連付けるための `volumesPerInstance` と `volumeSpecification` が含まれます。 • `volumesPerInstance` は、インスタンスグループ内の各インスタンスに関連付けられる特定のボリューム設定EBSを持つボリュームの数です。 • `volumeSpecification` は、Amazon EMRクラスター内のEC2インスタンスにアタッチされたEBSボリュームに対してリクエストされる、ボリュームタイプIOPS、、およびギガバイト (GiB) 単位のサイズなどのAmazon EBSボリューム仕様です。	参照オブジェクト (`“EbsBlockDeviceConfig”: {“ref”: “myEbsBlockDeviceConfig”}` など)
emrManagedMasterSecurityGroupId	Amazon EMRクラスターのマスターセキュリティグループの識別子で、 の形式に従います`sg-01XXXX6a` 。詳細については、「Amazon EMR 管理ガイド」の「セキュリティグループの設定」を参照してください。	文字列
emrManagedSlaveSecurityGroupId	Amazon EMRクラスターのスレーブセキュリティグループの識別子で、 の形式に従います`sg-01XXXX6a` 。	文字列
enableDebugging	Amazon EMRクラスターでデバッグを有効にします。	文字列
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表

オプションのフィールド	説明	スロットタイプ
hadoopSchedulerType	クラスターのスケジューラータイプ。有効なタイプは、 PARALLEL_FAIR_SCHEDULING 、 PARALLEL_CAPACITY_SCHEDULING 、 DEFAULT_SCHEDULER です。	一覧表
httpProxy	クライアントが AWSサービスへの接続に使用するプロキシホスト。	リファレンスオブジェクト、例：httpProxy「」：{「ref"myHttpProxyId」}
initTimeout	リソースが起動するまでの待機時間。	[Period] (期間)
keyPair	Amazon EMRクラスターのマスターノードへのログオンに使用する Amazon EC2キーペア。	文字列
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
masterInstanceBid料金	が Amazon EC2インスタンスに対して支払うことができる最大スポット料金。0～20.00 の 10 進数のみ (排他的) です。で指定しますUSD。この値を設定すると、Amazon EMRクラスターマスターノードのスポットインスタンスが有効になります。入札価格が指定されている場合、Amazon はインスタンスグループにスポットインスタンスEMRを使用します。	文字列
masterInstanceType	マスターノードに使用する Amazon EC2インスタンスのタイプ。「Amazon EMR クラスターでサポートされるAmazon EC2インスタンス」を参照してください。	文字列

オプションのフィールド	説明	スロットタイプ
masterGroupConfiguration	Amazon EMRクラスターマスターインスタンスグループの設定。このパラメータは Amazon EMRバージョン 4.0 以降にのみ適用されます。	参照オブジェクト (“configuration”: {“ref”: “myEmrConfigurationId”} など)
masterEbsConfiguration	Amazon EMRクラスターのマスターグループ内の各マスターノードにアタッチされる Amazon EBSボリュームの設定。詳細については、「Amazon ユーザーガイド」のEBS「最適化をサポートするインスタンスタイプ」を参照してください。 EC2	参照オブジェクト (“masterEbsConfiguration”: {“ref”: “myEbsConfiguration”} など)
maxActiveInstances	コンポーネントで同時にアクティブになるインスタンスの最大数。再実行はアクティブなインスタンスの数にはカウントされません。	整数
maximumRetries	失敗時の最大再試行回数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	参照オブジェクト ("onFail": {"ref":"myActionId"} など)
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	参照オブジェクト ("onLateAction": {"ref":"myActionId"} など)

オプションのフィールド	説明	スロットタイプ
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	参照オブジェクト ("onSuccess": {"ref":"myActionId"} など)
parent	スロットの継承元となる現在のオブジェクトの親。	参照オブジェクト ("parent": {"ref":"myBaseObjectId"} など)
pipelineLogUri	パイプラインのログをアップロードするための Amazon S3 URI (「s3://BucketName/Key/」など) 。	文字列
region	Amazon EMRクラスターを実行するリージョンのコード。デフォルトでは、クラスターはパイプラインと同じリージョンで実行されます。依存するデータセットと同じリージョンでクラスターを実行することもできます。	一覧表
releaseLabel	EMR クラスターのリリースラベル。	文字列
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
resourceRole	が Amazon EMRクラスターの作成 AWS Data Pipeline に使用するIAMロール。デフォルトのロールは `DataPipelineDefaultRole` です。	文字列
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
ロール	EC2 ノードを作成EMRするために Amazon に渡されるIAMロール。	文字列
runsOn	このフィールドはこのオブジェクトでは使用できません。	参照オブジェクト ("runsOn": {"ref":"myResourceId"} など)
securityConfiguration	クラスターに適用されるEMRセキュリティ設定の識別子。このパラメータは Amazon EMR バージョン 4.8.0 以降にのみ適用されます。	文字列
serviceAccessSecurityGroupId	Amazon EMRクラスターのサービスアクセスセキュリティグループの識別子。	文字列。形式は sg-01XXXX6a です (例: sg-1234abcd)。

オプションのフィールド	説明	スロットタイプ
scheduleType	スケジュールタイプでは、パイプライン定義のオブジェクトを間隔の最初にスケジュールするか、間隔の最後にスケジュールするかを指定できます。値は、cron、ondemand、および timeseries です。timeseries スケジューリングでは、インスタンスを各間隔の最後にスケジュールします。cron スケジューリングでは、インスタンスを各間隔の最初にスケジュールします。ondemand スケジュールにより、アクティベーションごとに 1 回パイプラインを実行することができます。パイプラインを再実行するために、クローンしたり再作成したりする必要はありません。ondemand スケジュールを使用する場合は、デフォルトオブジェクトで指定し、パイプラインのオブジェクトに対して指定される唯一の scheduleType である必要があります。ondemand パイプラインを使用するには、それ以降の実行ごとに、ActivatePipeline オペレーションを呼び出します。	一覧表
subnetId	Amazon EMRクラスターを起動するサブネットの識別子。	文字列
supportedProducts	Hadoop のサードパーティーディストリビューションなど、Amazon EMRクラスターにサードパーティーソフトウェアをインストールするパラメータ。	文字列
taskInstanceBid料金	EC2 がインスタンスに対して支払うことができる最大スポット料金。0 ~ 20.00 の 10 進数のみ。で指定しますUSD。入札価格が指定されている場合、Amazon はインスタンスグループにスポットインスタンスEMRを使用します。	文字列

オプションのフィールド	説明	スロットタイプ
taskInstanceCount	Amazon EMRクラスターに使用するタスクノードの数。	整数
taskInstanceType	タスクノードに使用する Amazon EC2インスタンスのタイプ。	文字列
taskGroupConfiguration	Amazon EMRクラスタータスクインスタンスグループの設定。このパラメータは Amazon EMRバージョン 4.0 以降にのみ適用されます。	参照オブジェクト (“configuration”: {“ref”: “myEmrConfigurationId”} など)
taskEbsConfiguration	Amazon EMRクラスターのタスクグループ内の各タスクノードにアタッチされる Amazon EBSボリュームの構成。詳細については、「Amazon ユーザーガイド」のEBS「最適化をサポートするインスタンスタイプ」を参照してください。 EC2	参照オブジェクト (“taskEbsConfiguration”: {“ref”: “myEbsConfiguration”} など)
terminateAfter	これらの多くの時間が経過した後でリソースを終了します。	整数

オプションのフィールド	説明	スロットタイプ
VolumeSpecification	Amazon EMRクラスター内の Amazon EC2インスタンスにアタッチされた Amazon EBSボリュームに対してリクエストされるボリュームタイプIOPS、、およびサイズ (GiB) などの Amazon EBSボリューム仕様。ノードは、コア、マスター、またはタスクノードです。 `VolumeSpecification` には以下が含まれます。 • `iops( )` 整数。Amazon EBSボリュームがサポートする 1 秒あたりの I/O オペレーションの数 (IOPS) 。例えば、1000 です。詳細については、「Amazon ユーザーガイド」のEBS「I/O 特性」を参照してください。 EC2 • `sizeinGB( )` 。整数。Amazon EBSボリュームサイズ、ギビバイト (GiB)、例えば 500。ボリュームタイプとハードドライブサイズの有効な組み合わせについては、「Amazon ユーザーガイド」のEBS「ボリュームタイプ」を参照してください。 EC2 • `volumetType` 。文字列。Amazon EBSボリュームタイプ。例えば、gp2。サポートされているボリュームタイプには、標準、gp2、io1、st1、sc1 などがあります。詳細については、「Amazon ユーザーガイド」のEBS「ボリュームタイプ」を参照してください。 EC2	参照オブジェクト (`“VolumeSpecification”: {“ref”: “myVolumeSpecification”}` など)

オプションのフィールド	説明	スロットタイプ
useOnDemandOnLastAttempt	リソースをリクエストする最後の試行で、スポットインスタンスではなくオンデマンドインスタンスのリクエストを作成します。これにより、以前の試行がすべて失敗した場合に、最後の試行は中断されません。	ブール値
workerGroup	このオブジェクトで使用できないフィールド。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例：activeInstances「」:{「ref"myRunnableObjectId」}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した依存関係のチェーンの説明。	リファレンスオブジェクト、例：cascadeFailedOn「」:{「ref"myRunnableObjectId」}
emrStepLog	ステップログは、Amazon EMRアクティビティの試行でのみ使用できます。	文字列
errorId	このオブジェクトが失敗した場合はエラーID。	文字列

実行時フィールド	説明	スロットタイプ
errorMessage	このオブジェクトが失敗した場合はエラーメッセージ。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
@failureReason	リソースの失敗の理由。	文字列
@finishedTime	このオブジェクトが実行を終了した時刻。	DateTime
hadoopJobLog	Amazon EMRアクティビティの試行で利用可能な Hadoop ジョブログ。	文字列
@healthStatus	終了状態に達した最後のオブジェクトインスタンスの成功または失敗を反映する、オブジェクトのヘルスステータス。	文字列
@healthStatusFromInstanceId	終了状態に達した最後のインスタンスオブジェクトの ID。	文字列
@healthStatusUpdatedTime	ヘルス状態が最後に更新された時間。	DateTime
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
@lastDeactivatedTime	このオブジェクトが最後に非アクティブ化された時刻。	DateTime
@latestCompletedRunTime	実行が完了した最後の実行の時刻。	DateTime
@latestRunTime	実行がスケジュールされた最後の実行の時刻。	DateTime
@nextRunTime	次回にスケジュールされた実行の時刻。	DateTime
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime

実行時フィールド	説明	スロットタイプ
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例：waitingOn「」：{「ref"myRunnableObjectId」}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	ライフサイクル内のオブジェクトの場所です。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

例

以下は、このオブジェクト型の例です。

内容

- で Amazon EMRクラスターを起動する hadoopVersion

- リリースラベル emr-4.x 以降の Amazon EMRクラスターを起動する
- Amazon EMRクラスターに追加のソフトウェアをインストールする
- 3.xリリースでのサーバー側暗号化の無効化
- 4.xリリースでのサーバー側暗号化の無効化
- で Hadoop を設定しKMSACLs、暗号化ゾーンを作成する HDFS
- カスタムIAMロールを指定する
- for Java で EmrCluster リソースAWSSDKを使用する
- プライベートサブネットで Amazon EMRクラスターを設定する
- クラスターノードにEBSボリュームをアタッチする

で Amazon EMRクラスターを起動する hadoopVersion

Example

次の例では、AMIバージョン 1.0 と Hadoop 0.20 を使用して Amazon EMRクラスターを起動します。

```
{
  "id" : "MyEmrCluster",
  "type" : "EmrCluster",
  "hadoopVersion" : "0.20",
  "keyPair" : "my-key-pair",
  "masterInstanceType" : "m3.xlarge",
  "coreInstanceType" : "m3.xlarge",
  "coreInstanceCount" : "10",
  "taskInstanceType" : "m3.xlarge",
  "taskInstanceCount": "10",
  "bootstrapAction" : ["s3://Region.elasticmapreduce/bootstrap-actions/configure-
hadoop,arg1,arg2,arg3","s3://Region.elasticmapreduce/bootstrap-actions/configure-
hadoop/configure-other-stuff,arg1,arg2"]
}
```

リリースラベル emr-4.x 以降の Amazon EMRクラスターを起動する

Example

次の例では、新しい `releaseLabel`フィールドを使用して Amazon EMRクラスターを起動します。

```
{
```

```
  "id" : "MyEmrCluster",
  "type" : "EmrCluster",
  "keyPair" : "my-key-pair",
  "masterInstanceType" : "m3.xlarge",
  "coreInstanceType" : "m3.xlarge",
  "coreInstanceCount" : "10",
  "taskInstanceType" : "m3.xlarge",
  "taskInstanceCount": "10",
  "releaseLabel": "emr-4.1.0",
  "applications": ["spark", "hive", "pig"],
  "configuration": {"ref":"myConfiguration"}
}
```

Amazon EMRクラスターに追加のソフトウェアをインストールする

Example

EmrCluster には、Amazon EMRクラスターにサードパーティーソフトウェアをインストールするsupportedProductsフィールドが用意されています。たとえば、MapR などの Hadoop のカスタムディストリビューションをインストールできます。サードパーティソフトウェアは、カンマ区切りの引数リストとして指定します。以下の例は、EmrCluster の supportedProducts フィールドを使用して、Karmasphere Analytics をインストールしたカスタム MapR M3 エディションクラスターを作成し、そこで EmrActivity オブジェクトを実行する方法を示しています。

```
{
    "id": "MyEmrActivity",
    "type": "EmrActivity",
    "schedule": {"ref": "ResourcePeriod"},
    "runsOn": {"ref": "MyEmrCluster"},
    "postStepCommand": "echo Ending job >> /mnt/var/log/stepCommand.txt",
    "preStepCommand": "echo Starting job > /mnt/var/log/stepCommand.txt",
    "step": "/home/hadoop/contrib/streaming/hadoop-streaming.jar,-input,s3n://
elasticmapreduce/samples/wordcount/input,-output, \
     hdfs:///output32113/,-mapper,s3n://elasticmapreduce/samples/wordcount/
wordSplitter.py,-reducer,aggregate"
  },
  {
    "id": "MyEmrCluster",
    "type": "EmrCluster",
    "schedule": {"ref": "ResourcePeriod"},
    "supportedProducts": ["mapr,--edition,m3,--version,1.2,--key1,value1","karmasphere-
enterprise-utility"],
    "masterInstanceType": "m3.xlarge",
```

```
    "taskInstanceType": "m3.xlarge"
}
```

3.xリリースでのサーバー側暗号化の無効化

Example

によって作成された Hadoop バージョン 2.x のEmrClusterアクティビティは、デフォルトでサーバー側の暗号化 AWS Data Pipeline を有効にします。サーバー側の暗号化を無効にするには、クラスターオブジェクト定義でブートストラップアクションを指定する必要があります。

次の例では、サーバー側の暗号化を無効にして EmrCluster アクティビティを作成します。

```
{
   "id":"NoSSEEmrCluster",
   "type":"EmrCluster",
   "hadoopVersion":"2.x",
   "keyPair":"my-key-pair",
   "masterInstanceType":"m3.xlarge",
   "coreInstanceType":"m3.large",
   "coreInstanceCount":"10",
   "taskInstanceType":"m3.large",
   "taskInstanceCount":"10",
   "bootstrapAction":["s3://Region.elasticmapreduce/bootstrap-actions/configure-
hadoop,-e, fs.s3.enableServerSideEncryption=false"]
}
```

4.xリリースでのサーバー側暗号化の無効化

Example

EmrConfiguration オブジェクトを使用してサーバー側の暗号化を無効にする必要があります。

次の例では、サーバー側の暗号化を無効にして EmrCluster アクティビティを作成します。

```
   {
      "name": "ReleaseLabelCluster",
      "releaseLabel": "emr-4.1.0",
      "applications": ["spark", "hive", "pig"],
      "id": "myResourceId",
      "type": "EmrCluster",
      "configuration": {
        "ref": "disableSSE"
```

```
      }
    },
    {
      "name": "disableSSE",
      "id": "disableSSE",
      "type": "EmrConfiguration",
      "classification": "emrfs-site",
      "property": [{
        "ref": "enableServerSideEncryption"
      }
      ]
    },
    {
      "name": "enableServerSideEncryption",
      "id": "enableServerSideEncryption",
      "type": "Property",
      "key": "fs.s3.enableServerSideEncryption",
      "value": "false"
    }
```

で Hadoop を設定しKMSACLs、暗号化ゾーンを作成する HDFS

Example

次のオブジェクトは、 Hadoop ACLs用に KMSを作成し、 で暗号化ゾーンと対応する暗号化キーを作成しますHDFS。

```
{
      "name": "kmsAcls",
      "id": "kmsAcls",
      "type": "EmrConfiguration",
      "classification": "hadoop-kms-acls",
      "property": [
        {"ref":"kmsBlacklist"},
        {"ref":"kmsAcl"}
      ]
    },
    {
      "name": "hdfsEncryptionZone",
      "id": "hdfsEncryptionZone",
      "type": "EmrConfiguration",
      "classification": "hdfs-encryption-zones",
      "property": [
        {"ref":"hdfsPath1"},
```

```
        {"ref":"hdfsPath2"}
      ]
    },
    {
      "name": "kmsBlacklist",
      "id": "kmsBlacklist",
      "type": "Property",
      "key": "hadoop.kms.blacklist.CREATE",
      "value": "foo,myBannedUser"
    },
    {
      "name": "kmsAcl",
      "id": "kmsAcl",
      "type": "Property",
      "key": "hadoop.kms.acl.ROLLOVER",
      "value": "myAllowedUser"
    },
    {
      "name": "hdfsPath1",
      "id": "hdfsPath1",
      "type": "Property",
      "key": "/myHDFSPath1",
      "value": "path1_key"
    },
    {
      "name": "hdfsPath2",
      "id": "hdfsPath2",
      "type": "Property",
      "key": "/myHDFSPath2",
      "value": "path2_key"
    }
```

カスタムIAMロールを指定する

Example

デフォルトでは、 は Amazon EMRサービスロールDataPipelineDefaultRoleとして を渡し、ユーザーに代わってリソースを作成するための Amazon EC2インスタンスプロファイルDataPipelineDefaultResourceRoleとして を AWS Data Pipeline 渡します。ただし、カスタム Amazon EMRサービスロールとカスタムインスタンスプロファイルを作成して、代わりに使用できます。 AWS Data Pipeline には、カスタムロールを使用してクラスターを作成するのに十分なアクセス許可が必要であり、 を信頼されたエンティティ AWS Data Pipeline として追加する必要があります。

次のオブジェクト例では、Amazon EMRクラスターのカスタムロールを指定します。

```
{
   "id":"MyEmrCluster",
   "type":"EmrCluster",
   "hadoopVersion":"2.x",
   "keyPair":"my-key-pair",
   "masterInstanceType":"m3.xlarge",
   "coreInstanceType":"m3.large",
   "coreInstanceCount":"10",
   "taskInstanceType":"m3.large",
   "taskInstanceCount":"10",
   "role":"emrServiceRole",
   "resourceRole":"emrInstanceProfile"
}
```

for Java で EmrCluster リソースAWSSDKを使用する

Example

次の例は、 EmrClusterと を使用して Amazon 4.x EMR クラスターEmrActivityを作成し、Java を使用して Spark ステップを実行する方法を示していますSDK。

```
public class dataPipelineEmr4 {

  public static void main(String[] args) {

 AWSCredentials credentials = null;
 credentials = new ProfileCredentialsProvider("/path/to/
AwsCredentials.properties","default").getCredentials();
 DataPipelineClient dp = new DataPipelineClient(credentials);
 CreatePipelineRequest createPipeline = new
 CreatePipelineRequest().withName("EMR4SDK").withUniqueId("unique");
 CreatePipelineResult createPipelineResult = dp.createPipeline(createPipeline);
 String pipelineId = createPipelineResult.getPipelineId();

 PipelineObject emrCluster = new PipelineObject()
     .withName("EmrClusterObj")
     .withId("EmrClusterObj")
     .withFields(
   new Field().withKey("releaseLabel").withStringValue("emr-4.1.0"),
   new Field().withKey("coreInstanceCount").withStringValue("3"),
   new Field().withKey("applications").withStringValue("spark"),
```

```
    new Field().withKey("applications").withStringValue("Presto-Sandbox"),
    new Field().withKey("type").withStringValue("EmrCluster"),
    new Field().withKey("keyPair").withStringValue("myKeyName"),
    new Field().withKey("masterInstanceType").withStringValue("m3.xlarge"),
    new Field().withKey("coreInstanceType").withStringValue("m3.xlarge")
    );

 PipelineObject emrActivity = new PipelineObject()
     .withName("EmrActivityObj")
     .withId("EmrActivityObj")
     .withFields(
    new Field().withKey("step").withStringValue("command-runner.jar,spark-submit,--
executor-memory,1g,--class,org.apache.spark.examples.SparkPi,/usr/lib/spark/lib/spark-
examples.jar,10"),
    new Field().withKey("runsOn").withRefValue("EmrClusterObj"),
    new Field().withKey("type").withStringValue("EmrActivity")
    );

 PipelineObject schedule = new PipelineObject()
     .withName("Every 15 Minutes")
     .withId("DefaultSchedule")
     .withFields(
    new Field().withKey("type").withStringValue("Schedule"),
    new Field().withKey("period").withStringValue("15 Minutes"),
    new Field().withKey("startAt").withStringValue("FIRST_ACTIVATION_DATE_TIME")
    );

 PipelineObject defaultObject = new PipelineObject()
     .withName("Default")
     .withId("Default")
     .withFields(
    new Field().withKey("failureAndRerunMode").withStringValue("CASCADE"),
    new Field().withKey("schedule").withRefValue("DefaultSchedule"),
    new
 Field().withKey("resourceRole").withStringValue("DataPipelineDefaultResourceRole"),
    new Field().withKey("role").withStringValue("DataPipelineDefaultRole"),
    new Field().withKey("pipelineLogUri").withStringValue("s3://myLogUri"),
    new Field().withKey("scheduleType").withStringValue("cron")
    );

 List<PipelineObject> pipelineObjects = new ArrayList<PipelineObject>();

 pipelineObjects.add(emrActivity);
 pipelineObjects.add(emrCluster);
```

```
 pipelineObjects.add(defaultObject);
 pipelineObjects.add(schedule);

 PutPipelineDefinitionRequest putPipelineDefintion = new PutPipelineDefinitionRequest()
     .withPipelineId(pipelineId)
     .withPipelineObjects(pipelineObjects);

 PutPipelineDefinitionResult putPipelineResult =
 dp.putPipelineDefinition(putPipelineDefintion);
 System.out.println(putPipelineResult);

 ActivatePipelineRequest activatePipelineReq = new ActivatePipelineRequest()
     .withPipelineId(pipelineId);
 ActivatePipelineResult activatePipelineRes = dp.activatePipeline(activatePipelineReq);

      System.out.println(activatePipelineRes);
      System.out.println(pipelineId);

    }

}
```

プライベートサブネットで Amazon EMRクラスターを設定する

Example

この例には、クラスターを のプライベートサブネットに起動する設定が含まれていますVPC。詳細については、「Amazon 管理ガイド」の「 で Amazon EMR クラスターを起動VPCする」を参照してください。 EMR この設定はオプションです。この設定は、EmrCluster オブジェクトを使用する任意のパイプラインで使用できます。

プライベートサブネットで Amazon EMRクラスターを起動するには、EmrCluster設定serviceAccessSecurityGroupIdでSubnetId、emrManagedMasterSecurityGroupId、emrManagedSlaveSecurityGroupId、および を指定します。

```
{
  "objects": [
    {
      "output": {
        "ref": "S3BackupLocation"
      },
      "input": {
```

```
        "ref": "DDBSourceTable"
      },
      "maximumRetries": "2",
      "name": "TableBackupActivity",
      "step": "s3://dynamodb-emr-#{myDDBRegion}/emr-ddb-storage-handler/2.1.0/emr-
ddb-2.1.0.jar,org.apache.hadoop.dynamodb.tools.DynamoDbExport,#{output.directoryPath},#{input.t
      "id": "TableBackupActivity",
      "runsOn": {
        "ref": "EmrClusterForBackup"
      },
      "type": "EmrActivity",
      "resizeClusterBeforeRunning": "false"
    },
    {
      "readThroughputPercent": "#{myDDBReadThroughputRatio}",
      "name": "DDBSourceTable",
      "id": "DDBSourceTable",
      "type": "DynamoDBDataNode",
      "tableName": "#{myDDBTableName}"
    },
    {
      "directoryPath": "#{myOutputS3Loc}/#{format(@scheduledStartTime, 'YYYY-MM-dd-HH-
mm-ss')}",
      "name": "S3BackupLocation",
      "id": "S3BackupLocation",
      "type": "S3DataNode"
    },
    {
      "name": "EmrClusterForBackup",
      "coreInstanceCount": "1",
      "taskInstanceCount": "1",
      "taskInstanceType": "m4.xlarge",
      "coreInstanceType": "m4.xlarge",
      "releaseLabel": "emr-4.7.0",
      "masterInstanceType": "m4.xlarge",
      "id": "EmrClusterForBackup",
      "subnetId": "#{mySubnetId}",
      "emrManagedMasterSecurityGroupId": "#{myMasterSecurityGroup}",
      "emrManagedSlaveSecurityGroupId": "#{mySlaveSecurityGroup}",
      "serviceAccessSecurityGroupId": "#{myServiceAccessSecurityGroup}",
      "region": "#{myDDBRegion}",
      "type": "EmrCluster",
      "keyPair": "user-key-pair"
    },
```

```
    {
      "failureAndRerunMode": "CASCADE",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "role": "DataPipelineDefaultRole",
      "pipelineLogUri": "#{myPipelineLogUri}",
      "scheduleType": "ONDEMAND",
      "name": "Default",
      "id": "Default"
    }
  ],
  "parameters": [
    {
      "description": "Output S3 folder",
      "id": "myOutputS3Loc",
      "type": "AWS::S3::ObjectKey"
    },
    {
      "description": "Source DynamoDB table name",
      "id": "myDDBTableName",
      "type": "String"
    },
    {
      "default": "0.25",
      "watermark": "Enter value between 0.1-1.0",
      "description": "DynamoDB read throughput ratio",
      "id": "myDDBReadThroughputRatio",
      "type": "Double"
    },
    {
      "default": "us-east-1",
      "watermark": "us-east-1",
      "description": "Region of the DynamoDB table",
      "id": "myDDBRegion",
      "type": "String"
    }
  ],
  "values": {
     "myDDBRegion": "us-east-1",
      "myDDBTableName": "ddb_table",
      "myDDBReadThroughputRatio": "0.25",
      "myOutputS3Loc": "s3://s3_path",
      "mySubnetId": "subnet_id",
      "myServiceAccessSecurityGroup":  "service access security group",
      "mySlaveSecurityGroup": "slave security group",
```

```
      "myMasterSecurityGroup": "master security group",
      "myPipelineLogUri": "s3://s3_path"
  }
}
```

クラスターノードにEBSボリュームをアタッチする

Example

パイプライン内のEMRクラスター内の任意のタイプのノードにEBSボリュームをアタッチできます。EBS ボリュームをノードにアタッチするには、EmrCluster設定TaskEbsConfigurationで coreEbsConfiguration、masterEbsConfiguration、および を使用します。

この Amazon EMRクラスターの例では、マスターノード、タスクノード、コアノードに Amazon EBSボリュームを使用します。詳細については、「Amazon 管理ガイド」の「Amazon の Amazon EBSボリュームEMR」を参照してください。 EMR

これら設定はオプションです。これらの設定は、EmrCluster オブジェクトを使用する任意のパイプラインで使用できます。

パイプラインで、EmrClusterオブジェクト設定をクリックし、マスターEBS設定 、コアEBS設定 、またはタスクEBS設定 を選択し、次の例のような設定の詳細を入力します。

```
{
  "objects": [
    {
      "output": {
        "ref": "S3BackupLocation"
      },
      "input": {
        "ref": "DDBSourceTable"
      },
      "maximumRetries": "2",
      "name": "TableBackupActivity",
      "step": "s3://dynamodb-emr-#{myDDBRegion}/emr-ddb-storage-handler/2.1.0/emr-
ddb-2.1.0.jar,org.apache.hadoop.dynamodb.tools.DynamoDbExport,#{output.directoryPath},#{input.t
      "id": "TableBackupActivity",
      "runsOn": {
        "ref": "EmrClusterForBackup"
      },
      "type": "EmrActivity",
      "resizeClusterBeforeRunning": "false"
```

```
    },
    {
      "readThroughputPercent": "#{myDDBReadThroughputRatio}",
      "name": "DDBSourceTable",
      "id": "DDBSourceTable",
      "type": "DynamoDBDataNode",
      "tableName": "#{myDDBTableName}"
    },
    {
      "directoryPath": "#{myOutputS3Loc}/#{format(@scheduledStartTime, 'YYYY-MM-dd-HH-
mm-ss')}",
      "name": "S3BackupLocation",
      "id": "S3BackupLocation",
      "type": "S3DataNode"
    },
    {
      "name": "EmrClusterForBackup",
      "coreInstanceCount": "1",
      "taskInstanceCount": "1",
      "taskInstanceType": "m4.xlarge",
      "coreInstanceType": "m4.xlarge",
      "releaseLabel": "emr-4.7.0",
      "masterInstanceType": "m4.xlarge",
      "id": "EmrClusterForBackup",
      "subnetId": "#{mySubnetId}",
      "emrManagedMasterSecurityGroupId": "#{myMasterSecurityGroup}",
      "emrManagedSlaveSecurityGroupId": "#{mySlaveSecurityGroup}",
      "region": "#{myDDBRegion}",
      "type": "EmrCluster",
      "coreEbsConfiguration": {
        "ref": "EBSConfiguration"
      },
      "masterEbsConfiguration": {
        "ref": "EBSConfiguration"
      },
      "taskEbsConfiguration": {
        "ref": "EBSConfiguration"
      },
      "keyPair": "user-key-pair"
    },
    {
       "name": "EBSConfiguration",
        "id": "EBSConfiguration",
        "ebsOptimized": "true",
```

```
        "ebsBlockDeviceConfig" : [
            { "ref": "EbsBlockDeviceConfig" }
        ],
        "type": "EbsConfiguration"
    },
    {
        "name": "EbsBlockDeviceConfig",
        "id": "EbsBlockDeviceConfig",
        "type": "EbsBlockDeviceConfig",
        "volumesPerInstance" : "2",
        "volumeSpecification" : {
            "ref": "VolumeSpecification"
        }
    },
    {
      "name": "VolumeSpecification",
      "id": "VolumeSpecification",
      "type": "VolumeSpecification",
      "sizeInGB": "500",
      "volumeType": "io1",
      "iops": "1000"
    },
    {
      "failureAndRerunMode": "CASCADE",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "role": "DataPipelineDefaultRole",
      "pipelineLogUri": "#{myPipelineLogUri}",
      "scheduleType": "ONDEMAND",
      "name": "Default",
      "id": "Default"
    }
  ],
  "parameters": [
    {
      "description": "Output S3 folder",
      "id": "myOutputS3Loc",
      "type": "AWS::S3::ObjectKey"
    },
    {
      "description": "Source DynamoDB table name",
      "id": "myDDBTableName",
      "type": "String"
    },
    {
```

```
      "default": "0.25",
      "watermark": "Enter value between 0.1-1.0",
      "description": "DynamoDB read throughput ratio",
      "id": "myDDBReadThroughputRatio",
      "type": "Double"
    },
    {
      "default": "us-east-1",
      "watermark": "us-east-1",
      "description": "Region of the DynamoDB table",
      "id": "myDDBRegion",
      "type": "String"
    }
  ],
  "values": {
     "myDDBRegion": "us-east-1",
      "myDDBTableName": "ddb_table",
      "myDDBReadThroughputRatio": "0.25",
      "myOutputS3Loc": "s3://s3_path",
      "mySubnetId": "subnet_id",
      "mySlaveSecurityGroup": "slave security group",
      "myMasterSecurityGroup": "master security group",
      "myPipelineLogUri": "s3://s3_path"
  }
}
```

以下の資料も参照してください。

- EmrActivity

HttpProxy

HttpProxy では、独自のプロキシを設定し、Task Runner にそのプロキシを介して AWS Data Pipeline サービスにアクセスさせることができます。この情報を使用して、実行中の Task Runner を設定する必要はありません。

の HttpProxy の例 TaskRunner

次のパイプライン定義は、HttpProxy オブジェクトを示しています。

```
{
  "objects": [
```

```
    {
      "schedule": {
        "ref": "Once"
      },
      "pipelineLogUri": "s3://myDPLogUri/path",
      "name": "Default",
      "id": "Default"
    },
    {
      "name": "test_proxy",
      "hostname": "hostname",
      "port": "port",
      "username": "username",
      "*password": "password",
      "windowsDomain": "windowsDomain",
      "type": "HttpProxy",
      "id": "test_proxy",
    },
    {
      "name": "ShellCommand",
      "id": "ShellCommand",
      "runsOn": {
        "ref": "Resource"
      },
      "type": "ShellCommandActivity",
      "command": "echo 'hello world' "
    },
    {
      "period": "1 day",
      "startDateTime": "2013-03-09T00:00:00",
      "name": "Once",
      "id": "Once",
      "endDateTime": "2013-03-10T00:00:00",
      "type": "Schedule"
    },
    {
      "role": "dataPipelineRole",
      "httpProxy": {
        "ref": "test_proxy"
      },
      "actionOnResourceFailure": "retrynone",
      "maximumRetries": "0",
      "type": "Ec2Resource",
      "terminateAfter": "10 minutes",
```

```
      "resourceRole": "resourceRole",
      "name": "Resource",
      "actionOnTaskFailure": "terminate",
      "securityGroups": "securityGroups",
      "keyPair": "keyPair",
      "id": "Resource",
      "region": "us-east-1"
    }
  ],
  "parameters": []
}
```

構文

必須フィールド	説明	スロットタイプ
hostname	クライアントが AWSサービスへの接続に使用するプロキシのホスト。	文字列
port	クライアントが AWSサービスへの接続に使用するプロキシホストのポート。	文字列

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
*パスワード	プロキシ用のパスワード。	文字列
s3NoProxy	Amazon S3 に接続するときにHTTPプロキシを無効にする	ブール値
username	プロキシ用のユーザー名。	文字列
windowsDomain	NTLM Proxy の Windows ドメイン名。	文字列

オプションのフィールド	説明	スロットタイプ
windowsWorkgroup	NTLM Proxy の Windows ワークグループ名。	文字列

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

前提条件

AWS Data Pipeline 前提条件オブジェクトは次のとおりです。

オブジェクト

- D ynamoDBDataが存在する
- D ynamoDBTableが存在する
- 存在する
- S3KeyExists

- S3PrefixNotEmpty
- ShellCommandPrecondition

D ynamoDBDataが存在する

DynamoDB テーブルにデータが存在することを確認する前提条件。

構文

必須フィールド	説明	スロットタイプ
ロール	前提条件を実行するために使用するロールを指定します。	文字列
tableName	確認する DynamoDB テーブル。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	失敗時の最大再試行回数	整数

オプションのフィールド	説明	スロットタイプ
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間	[Period] (期間)
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeIn

実行時フィールド	説明	スロットタイプ
		stances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
currentRetryCount	この試行で前提条件が試みられた回数。	文字列
emrStepLog	EMR EMRアクティビティの試行でのみ使用できるステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
lastRetryTime	この試行内で最後に前提条件が試みられた時刻。	文字列

実行時フィールド	説明	スロットタイプ
node	この前提条件が実行されている対象ノード	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

D ynamoDBTableが存在する

DynamoDB テーブルが存在することを確認する前提条件。

構文

必須フィールド	説明	スロットタイプ
ロール	前提条件を実行するために使用するロールを指定します。	文字列
tableName	確認する DynamoDB テーブル。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}

オプションのフィールド	説明	スロットタイプ
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間	[Period] (期間)
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}

実行時フィールド	説明	スロットタイプ
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
currentRetryCount	この試行で前提条件が試みられた回数。	文字列
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
lastRetryTime	この試行内で最後に前提条件が試みられた時刻。	文字列

実行時フィールド	説明	スロットタイプ
node	この前提条件が実行されている対象ノード	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

存在する

データノードオブジェクトが存在するか確認します。

> **Note**
>
> 代わりに、システムで管理される前提条件を使用することをお勧めします。詳細については、「前提条件」を参照してください。

例

以下は、このオブジェクト型の例です。InputData オブジェクトは、このオブジェクト (Ready) と、同じパイプライン定義ファイルで定義した別のオブジェクトを参照します。CopyPeriod は Schedule オブジェクトです。

```
{
  "id" : "InputData",
  "type" : "S3DataNode",
  "schedule" : { "ref" : "CopyPeriod" },
  "filePath" : "s3://example-bucket/InputData/#{@scheduledStartTime.format('YYYY-MM-dd-hh:mm')}.csv",
  "precondition" : { "ref" : "Ready" }
},
{
  "id" : "Ready",
  "type" : "Exists"
}
```

構文

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかっ	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
	たリモートアクティビティを再試行することができます。	
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列

実行時フィールド	説明	スロットタイプ
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
node	この前提条件が実行されている対象ノード。	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- ShellCommandPrecondition

S3KeyExists

Amazon S3 データノードにキーがあるかどうかを確認します。

例

以下は、このオブジェクト型の例です。前提条件は、s3Key パラメータで参照されるキー (s3://mybucket/mykey) が存在する場合にトリガーされます。

```
{
"id" : "InputReady",
"type" : "S3KeyExists",
"role" : "test-role",
"s3Key" : "s3://mybucket/mykey"
}
```

最初のパイプラインが完了するまで待機する 2 番目のパイプラインで、S3KeyExists を前提条件として使用することもできます。そのためには、次の操作を行います。

1. 最初のパイプラインの完了の最後に Amazon S3 にファイルを書き込みます。
2. S3KeyExists 前提条件を 2 番目のパイプラインに作成します。

構文

必須フィールド	説明	スロットタイプ
ロール	前提条件を実行するために使用するロールを指定します。	文字列
s3Key	Amazon S3 キー。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業の完了をもう 1 回試みるまでのタイムアウト。設定された場合、開始後に設定された時間内に完了しなかったリモートアクティビティが再試行されます。	[Period] (期間)
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	障害時に開始される最大試行数。	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例:

オプションのフィールド	説明	スロットタイプ
		"onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間。	[Period] (期間)
reportProgressTimeout	`reportProgress` へのリモート作業の連続した呼び出しのタイムアウト。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の連続した再試行の間のタイムアウト期間。	[Period] (期間)

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime

実行時フィールド	説明	スロットタイプ
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
currentRetryCount	この試行で前提条件が試みられた回数。	文字列
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
lastRetryTime	この試行内で最後に前提条件が試みられた時刻。	文字列

実行時フィールド	説明	スロットタイプ
node	この前提条件が実行されている対象ノード	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- ShellCommandPrecondition

S3PrefixNotEmpty

指定されたプレフィックス (として表されますURI) を持つ Amazon S3 オブジェクトが存在することを確認する前提条件。

例

次は、必須フィールド、任意フィールド、式フィールドを使用した、このオブジェクト型の例です。

```
{
  "id" : "InputReady",
  "type" : "S3PrefixNotEmpty",
  "role" : "test-role",
  "s3Prefix" : "#{node.filePath}"
}
```

構文

必須フィールド	説明	スロットタイプ
ロール	前提条件を実行するために使用するロールを指定します。	文字列
s3Prefix	オブジェクトの存在を確認する Amazon S3 プレフィックス。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列

オプションのフィールド	説明	スロットタイプ
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	失敗時の最大再試行回数	整数
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間	[Period] (期間)

オプションのフィールド	説明	スロットタイプ
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例: "cascadeFailedOn": {"ref"myRunnableObjectId"}
currentRetryCount	この試行で前提条件が試みられた回数。	文字列
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列

実行時フィールド	説明	スロットタイプ
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
lastRetryTime	この試行内で最後に前提条件が試みられた時刻。	文字列
node	この前提条件が実行されている対象ノード。	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻。	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻。	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- ShellCommandPrecondition

ShellCommandPrecondition

前提条件として実行できる Unix/Linux シェルコマンド。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "VerifyDataReadiness",
  "type" : "ShellCommandPrecondition",
  "command" : "perl check-data-ready.pl"
}
```

構文

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
コマンド	実行するコマンド。この値および関連するパラメーターは、Task Runner を実行している環境で機能する必要があります。	文字列
scriptUri	シェルコマンドとしてダウンロードして実行するファイルの Amazon S3 URIパス。1 つの scriptUri または コマンドフィールドのみが存在する必要があります。パラメータ scriptUri は使用できません。代わりに コマンドを使用してください。	文字列

オプションのフィールド	説明	スロットタイプ
attemptStatus	リモートアクティビティから最も最近報告されたステータス。	文字列
attemptTimeout	リモートの作業完了のタイムアウト。設定された場合、設定された開始時間内に完了しなかったリモートアクティビティを再試行することができます。	[Period] (期間)
failureAndRerunモード	依存関係が失敗または再実行されたときのコンシューマーノードの動作を示します。	一覧表
lateAfterTimeout	オブジェクトが完了しなければならない、パイプライン開始からの経過時間。スケジュールタイプが ondemand に設定されていない場合にのみトリガーされます。	[Period] (期間)
maximumRetries	失敗時の最大再試行回数	整数

オプションのフィールド	説明	スロットタイプ
onFail	現在のオブジェクトが失敗したときに実行するアクション。	リファレンスオブジェクト、例: "onFail": {"ref"myActionId"}
onLateAction	オブジェクトが予定されていないか、まだ完了していない場合にトリガーされるアクション。	リファレンスオブジェクト、例: "onLateAction":{"ref"myActionId"}
onSuccess	現在のオブジェクトが成功したときに実行するアクション。	リファレンスオブジェクト、例: "onSuccess":{"ref"myActionId"}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
preconditionTimeout	まだ満たされていない場合に失敗として前提条件がマークされた後の、起動からの期間	[Period] (期間)
reportProgressTimeout	へのリモートワークの連続呼び出しのタイムアウトreportProgress。設定された場合、指定された期間の進捗状況を報告しないリモートアクティビティは停止されたと見なし、再試行できます。	[Period] (期間)
retryDelay	2 回の再試行の間のタイムアウト期間。	[Period] (期間)
scriptArgument	シェルスクリプトに渡される引数	文字列

オプションのフィールド	説明	スロットタイプ
stderr	コマンドからリダイレクトされたシステムエラーメッセージを受け取る Amazon S3 パス。runsOn フィールドを使用する場合、アクティビティを実行するリソースが一時的であるため、これは Amazon S3 パスにする必要があります。ただし、workerGroup フィールドを指定した場合は、ローカルファイルパスを指定できます。	文字列
stdout	コマンドからリダイレクトされた出力を受け取る Amazon S3 パス。runsOn フィールドを使用する場合、アクティビティを実行するリソースが一時的であるため、これは Amazon S3 パスにする必要があります。ただし、workerGroup フィールドを指定した場合は、ローカルファイルパスを指定できます。	文字列

実行時フィールド	説明	スロットタイプ
@activeInstances	現在スケジュールされているアクティブなインスタンスオブジェクトのリスト。	リファレンスオブジェクト、例: "activeInstances":{"ref"myRunnableObjectId"}
@actualEndTime	このオブジェクトの実行が終了した時刻。	DateTime
@actualStartTime	このオブジェクトの実行が開始された時刻。	DateTime
cancellationReason	このオブジェクト cancellationReason がキャンセルされた場合の 。	文字列
@cascadeFailedOn	オブジェクトが失敗した際の依存関係チェーンの説明。	リファレンスオブジェクト、例:

実行時フィールド	説明	スロットタイプ
		"cascadeFailedOn": {"ref"myRunnableObjectId"}
emrStepLog	EMR EMRアクティビティの試行でのみ使用できる ステップログ	文字列
errorId	このオブジェクトが失敗 errorId した場合の 。	文字列
errorMessage	このオブジェクトが失敗 errorMessage した場合の 。	文字列
errorStackTrace	このオブジェクトが失敗した場合は、エラースタックトレース。	文字列
hadoopJobLog	EMRベースのアクティビティの試行で使用可能な Hadoop ジョブログ。	文字列
hostname	タスクの試行を取得したクライアントのホスト名。	文字列
node	この前提条件が実行されている対象ノード	リファレンスオブジェクト、例：「node」:{「ref "myRunnableObjectId」}
reportProgressTime	リモートアクティビティで進捗状況が報告された最新の時刻。	DateTime
@scheduledEndTime	オブジェクトの予定された終了時刻	DateTime
@scheduledStartTime	オブジェクトの予定された開始時刻	DateTime
@status	このオブジェクトのステータス。	文字列
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

実行時フィールド	説明	スロットタイプ
@waitingOn	このオブジェクトが待機している依存関係のリストの説明。	リファレンスオブジェクト、例: "waitingOn":{"ref"myRunnableObjectId"}

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- ShellCommandActivity
- 存在する

データベース

AWS Data Pipeline データベースオブジェクトは次のとおりです。

オブジェクト

- JdbcDatabase
- RdsDatabase
- RedshiftDatabase

JdbcDatabase

JDBC データベースを定義します。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyJdbcDatabase",
  "type" : "JdbcDatabase",
  "connectionString" : "jdbc:redshift://hostname:portnumber/dbname",
  "jdbcDriverClass" : "com.amazon.redshift.jdbc41.Driver",
  "jdbcDriverJarUri" : "s3://redshift-downloads/drivers/RedshiftJDBC41-1.1.6.1006.jar",
  "username" : "user_name",
  "*password" : "my_password"
}
```

構文

必須フィールド	説明	スロットタイプ
connectionString	データベースにアクセスするためのJDBC接続文字列。	文字列
jdbcDriverClass	JDBC 接続を確立する前にロードするドライバークラス。	文字列
*パスワード	指定するパスワード。	文字列
username	データベースに接続するときに指定するユーザー名。	文字列

オプションのフィールド	説明	スロットタイプ
databaseName	アタッチする論理データベースの名前	文字列

オプションのフィールド	説明	スロットタイプ
jdbcDriverJarURI	データベースへの接続に使用されるJDBCドライバーJARファイルの Amazon S3 内の場所。AWS Data Pipeline には、このJARファイルを読み取るためのアクセス許可が必要です。	文字列
jdbcProperties	このデータベースJDBCの接続でプロパティとして設定されるフォーム A=B のペア。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明する エラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

RdsDatabase

Amazon RDS データベースを定義します。

> **Note**
>
> RdsDatabase は Aurora をサポートしていません。代わりに、Aurora には the section called "JdbcDatabase" を使用してください。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyRdsDatabase",
  "type" : "RdsDatabase",
  "region" : "us-east-1",
  "username" : "user_name",
  "*password" : "my_password",
  "rdsInstanceId" : "my_db_instance_identifier"
}
```

Oracle エンジンの場合は jdbcDriverJarUri フィールドが必須であり、次のドライバーを指定できます: http://www.oracle.com/technetwork/database/features/jdbc/jdbc-drivers-12c-download-1958347.html。SQL サーバーエンジンの場合、 jdbcDriverJarUriフィールドは必須です。次のドライバーを指定できます: https://www.microsoft.com/en-us/download/details.aspx?displaylang=en&id=11774。MySQL エンジンと PostgreSQL エンジンの場合、jdbcDriverJarUriフィールドはオプションです。

構文

必須フィールド	説明	スロットタイプ
*パスワード	指定するパスワード。	文字列
rdsInstanceId	DB インスタンスの DBInstanceIdentifier プロパティ。	文字列

必須フィールド	説明	スロットタイプ
username	データベースに接続するときに指定するユーザー名。	文字列

オプションのフィールド	説明	スロットタイプ
databaseName	アタッチする論理データベースの名前	文字列
jdbcDriverJarURI	データベースへの接続に使用されるJDBCドライバーJARファイルの Amazon S3 内の場所。AWS Data Pipeline には、このJARファイルを読み取るためのアクセス許可が必要です。MySQL エンジンと PostgreSQL エンジンでは、このフィールドが指定されていない場合にデフォルトのドライバーが使用されますが、このフィールドを使用してデフォルトを上書きできます。Oracle エンジンと SQL Server エンジンの場合、このフィールドは必須です。	文字列
jdbcProperties	このデータベースJDBCの接続でプロパティとして設定されるフォーム A=B のペア。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
region	データベースがあるリージョンのコード。例えば、us-east-1 などです。	文字列

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

RedshiftDatabase

Amazon Redshift データベースを定義します。RedshiftDatabase は、パイプラインで使用されるデータベースのプロパティを表します。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyRedshiftDatabase",
  "type" : "RedshiftDatabase",
  "clusterId" : "myRedshiftClusterId",
  "username" : "user_name",
  "*password" : "my_password",
  "databaseName" : "database_name"
}
```

デフォルトでは、このオブジェクトでは clusterId フィールドを必要とする Postgres ドライバーが使用されます。Amazon Redshift ドライバーを使用するには、代わりに connectionString フィールドで Amazon Redshift コンソールの Amazon Redshift データベース接続文字列 ("jdbc:redshift:" で始まる) を指定します。

構文

必須フィールド	説明	スロットタイプ
*パスワード	指定するパスワード。	文字列
username	データベースに接続するときに指定するユーザー名。	文字列

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
clusterId	Amazon Redshift クラスターの作成時にユーザーによって指定された識別子。例えば、Amazon Redshift クラスターのエンドポイントが mydb.example.us-east-1.redshift.amazonaws.com の場合、正しい識別子は mydb です。この値は、Amazon Redshift コンソールでクラスター識別子またはクラスター名から取得できます。	文字列
connectionString	パイプラインとは異なるアカウントが所有する Amazon Redshift インスタンスに接続するためのJDBCエンドポイント。connectionString と clusterId の両方を指定することはできません。	文字列

オプションのフィールド	説明	スロットタイプ
databaseName	アタッチする論理データベースの名前。	文字列
jdbcProperties	このデータベースJDBCの接続のプロパティとして設定されるフォーム A=B のペア。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
region	データベースがあるリージョンのコード。例えば、us-east-1 などです。	一覧表

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

データ形式

データ AWS Data Pipeline 形式オブジェクトは次のとおりです。

オブジェクト

- CSV データ形式
- Custom データ形式
- D ynamoDBData形式
- DynamoDBExportDataFormat
- RegEx データ形式
- TSV データ形式

CSV データ形式

列区切り文字がカンマで、レコード区切り文字が改行文字である、カンマ区切りデータ形式。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyOutputDataType",
  "type" : "CSV",
  "column" : [
    "Name STRING",
    "Score INT",
    "DateOfBirth TIMESTAMP"
  ]
}
```

構文

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータに対して各フィールドで指定されたデータ型を持つ	文字列

オプションのフィールド	説明	スロットタイプ
	列名。例: STRING hostname 複数の値の場合は、スペースで区切られた列名とデータ型を使用します。	
escapeChar	後続の 1 文字を無視することをパーサーに指示する文字 (たとえば "\") 。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

Custom データ形式

特定の列区切り文字、レコード区切り文字、エスケープ文字を組み合わせて定義するカスタムデータ形式。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyOutputDataType",
  "type" : "Custom",
  "columnSeparator" : ",",
  "recordSeparator" : "\n",
  "column" : [
    "Name STRING",
    "Score INT",
    "DateOfBirth TIMESTAMP"
  ]
}
```

構文

必須フィールド	説明	スロットタイプ
columnSeparator	データファイルの列の終端を示す文字。	文字列

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータに対して各フィールドで指定されたデータ型を持つ列名。例: STRING hostname 複数の値の場合は、スペースで区切られた列名とデータ型を使用します。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「pa

オプションのフィールド	説明	スロットタイプ
		rent」:{「ref"myBaseObjectId」}
recordSeparator	データファイルの行の終端を示す文字 (たとえば "\n") 。単一の文字のみがサポートされます。	文字列

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

D ynamoDBData形式

DynamoDB テーブルにスキーマを適用して、Hive クエリでアクセスできるようにします。`DynamoDBDataFormat` は、`HiveActivity` オブジェクトおよび `DynamoDBDataNode` の入出力と共に使用されます。`DynamoDBDataFormat` では、Hive クエリのすべての列を指定する必要があります。Hive クエリで特定の列を柔軟に指定する方法の詳細または Amazon S3 サポートについては、DynamoDBExportDataFormat を参照してください。

> **Note**
>
> DynamoDB のブール型は、Hive のブール型にマッピングされません。ただし、Hive のブール型に、0 または 1 の DynamoDB 整数値をマッピングすることができます。

例

以下の例は、DynamoDBDataFormat を使用して DynamoDBDataNode 入力にスキーマを割り当てる方法を示します。こうすることで、HiveActivity オブジェクトがデータに列名でアクセスし、データを DynamoDBDataNode 出力にコピーできるようになります。

```
{
  "objects": [
    {
      "id" : "Exists.1",
      "name" : "Exists.1",
      "type" : "Exists"
    },
    {
      "id" : "DataFormat.1",
      "name" : "DataFormat.1",
      "type" : "DynamoDBDataFormat",
      "column" : [
         "hash STRING",
        "range STRING"
      ]
    },
    {
      "id" : "DynamoDBDataNode.1",
      "name" : "DynamoDBDataNode.1",
      "type" : "DynamoDBDataNode",
      "tableName" : "$INPUT_TABLE_NAME",
      "schedule" : { "ref" : "ResourcePeriod" },
      "dataFormat" : { "ref" : "DataFormat.1" }
    },
    {
      "id" : "DynamoDBDataNode.2",
      "name" : "DynamoDBDataNode.2",
      "type" : "DynamoDBDataNode",
      "tableName" : "$OUTPUT_TABLE_NAME",
      "schedule" : { "ref" : "ResourcePeriod" },
```

```
      "dataFormat" : { "ref" : "DataFormat.1" }
    },
    {
      "id" : "EmrCluster.1",
      "name" : "EmrCluster.1",
      "type" : "EmrCluster",
      "schedule" : { "ref" : "ResourcePeriod" },
      "masterInstanceType" : "m1.small",
      "keyPair" : "$KEYPAIR"
    },
    {
      "id" : "HiveActivity.1",
      "name" : "HiveActivity.1",
      "type" : "HiveActivity",
      "input" : { "ref" : "DynamoDBDataNode.1" },
      "output" : { "ref" : "DynamoDBDataNode.2" },
      "schedule" : { "ref" : "ResourcePeriod" },
      "runsOn" : { "ref" : "EmrCluster.1" },
      "hiveScript" : "insert overwrite table ${output1} select * from ${input1} ;"
    },
    {
      "id" : "ResourcePeriod",
      "name" : "ResourcePeriod",
      "type" : "Schedule",
      "period" : "1 day",
      "startDateTime" : "2012-05-04T00:00:00",
      "endDateTime" : "2012-05-05T00:00:00"
    }
  ]
}
```

構文

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータに対して各フィールドで指定されたデータ型を持つ列名。例えば、hostname STRING と指定します。複数の値の場合は、列名とデータ型をスペースで区切って使用します。	文字列

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	「parent」: {「ref"myBaseObjectId」} などのリファレンスオブジェクト

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成するために使用されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

DynamoDBExportDataFormat

DynamoDB テーブルにスキーマを適用して、Hive クエリでアクセスできるようにします。`DynamoDBExportDataFormat` は、`HiveCopyActivity` オブジェクトと、`DynamoDBDataNode` または `S3DataNode` の入出力と共に使用します。`DynamoDBExportDataFormat` には以下の利点があります。

- DynamoDB と Amazon S3 の両方のサポートを提供します。

- Hive クエリで特定の列によってデータをフィルタリングすることができます。
- スパースなスキーマである場合でも、DynamoDB のすべての属性をエクスポートします。

> **Note**
>
> DynamoDB のブール型は、Hive のブール型にマッピングされません。ただし、Hive のブール型に、0 または 1 の DynamoDB 整数値をマッピングすることができます。

例

以下の例は、HiveCopyActivity と DynamoDBExportDataFormat を使用して、タイムスタンプに基づいてデータをフィルタリングしながら DynamoDBDataNode 間でデータをコピーする方法を示します。

```
{
  "objects": [
    {
      "id" : "DataFormat.1",
      "name" : "DataFormat.1",
      "type" : "DynamoDBExportDataFormat",
      "column" : "timeStamp BIGINT"
    },
    {
      "id" : "DataFormat.2",
      "name" : "DataFormat.2",
      "type" : "DynamoDBExportDataFormat"
    },
    {
      "id" : "DynamoDBDataNode.1",
      "name" : "DynamoDBDataNode.1",
      "type" : "DynamoDBDataNode",
      "tableName" : "item_mapped_table_restore_temp",
      "schedule" : { "ref" : "ResourcePeriod" },
      "dataFormat" : { "ref" : "DataFormat.1" }
    },
    {
      "id" : "DynamoDBDataNode.2",
      "name" : "DynamoDBDataNode.2",
      "type" : "DynamoDBDataNode",
      "tableName" : "restore_table",
```

```
      "region" : "us_west_1",
      "schedule" : { "ref" : "ResourcePeriod" },
      "dataFormat" : { "ref" : "DataFormat.2" }
    },
    {
      "id" : "EmrCluster.1",
      "name" : "EmrCluster.1",
      "type" : "EmrCluster",
      "schedule" : { "ref" : "ResourcePeriod" },
      "masterInstanceType" : "m1.xlarge",
      "coreInstanceCount" : "4"
    },
    {
      "id" : "HiveTransform.1",
      "name" : "Hive Copy Transform.1",
      "type" : "HiveCopyActivity",
      "input" : { "ref" : "DynamoDBDataNode.1" },
      "output" : { "ref" : "DynamoDBDataNode.2" },
      "schedule" : { "ref" : "ResourcePeriod" },
      "runsOn" : { "ref" : "EmrCluster.1" },
      "filterSql" : "`timeStamp` > unix_timestamp(\"#{@scheduledStartTime}\", \"yyyy-
MM-dd'T'HH:mm:ss\")"
    },
    {
      "id" : "ResourcePeriod",
      "name" : "ResourcePeriod",
      "type" : "Schedule",
      "period" : "1 Hour",
      "startDateTime" : "2013-06-04T00:00:00",
      "endDateTime" : "2013-06-04T01:00:00"
    }
  ]
}
```

構文

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータに対して各フィールドで指定されたデータ型を持つ列名。例: hostname STRING	文字列

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

RegEx データ形式

正規表現によって定義されるカスタムデータ形式。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyInputDataType",
```

```
  "type" : "RegEx",
  "inputRegEx" : "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^ \"]*|\"[^\"]*\") (-|
[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\"[^\"]*\") ([^ \"]*|\"[^\"]*\"))?",
  "outputFormat" : "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s",
  "column" : [
    "host STRING",
    "identity STRING",
    "user STRING",
    "time STRING",
    "request STRING",
    "status STRING",
    "size STRING",
    "referer STRING",
    "agent STRING"
  ]
}
```

構文

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータに対して各フィールドで指定されたデータ型を持つ列名。例: STRING hostname 複数の値の場合は、スペースで区切られた列名とデータ型を使用します。	文字列
inputRegEx	S3 入力 file. を解析する正規表現 inputRegEx は、ファイル内の比較的非構造化データから列を取得する方法を提供します。	文字列
outputFormat	によって取得された列フィールドは inputRegEx、Java フォーマット構文を使用して %1$s %2$s として参照されます。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

TSV データ形式

列区切り文字がタブ文字で、レコード区切り文字が改行文字である、タブ区切りデータ形式。

例

以下は、このオブジェクト型の例です。

```
{
  "id" : "MyOutputDataType",
  "type" : "TSV",
  "column" : [
    "Name STRING",
    "Score INT",
    "DateOfBirth TIMESTAMP"
  ]
}
```

構文

オプションのフィールド	説明	スロットタイプ
列	このデータノードで記述されたデータの列名とデータ型。たとえば、"Name STRING" は、列名が Name で、データ型が STRING であることを示します。複数の列名とデータ型のペアはカンマで区切ります (例を参照)。	文字列
columnSeparator	1 つの列のフィールドと次の列のフィールドを区切る文字。デフォルトは '\t' です。	文字列
escapeChar	後続の 1 文字を無視することをパーサーに指示する文字 (たとえば "\") 。	文字列
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
recordSeparator	レコードを区切る文字。デフォルトは '\n' です。	文字列

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列

システムフィールド	説明	スロットタイプ
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

アクション

AWS Data Pipeline アクションオブジェクトは次のとおりです。

オブジェクト

- SnsAlarm
- 終了

SnsAlarm

アクティビティが正常に失敗または終了すると、Amazon SNS通知メッセージを送信します。

例

以下は、このオブジェクト型の例です。node.input および node.output の値は、onSuccess フィールドでこのオブジェクトを参照するデータノードまたはアクティビティから得られます。

```
{
  "id" : "SuccessNotify",
  "name" : "SuccessNotify",
  "type" : "SnsAlarm",
  "topicArn" : "arn:aws:sns:us-east-1:28619EXAMPLE:ExampleTopic",
  "subject" : "COPY SUCCESS: #{node.@scheduledStartTime}",
  "message" : "Files were copied from #{node.input} to #{node.output}."
}
```

構文

必須フィールド	説明	スロットタイプ
message	Amazon SNS通知の本文テキスト。	文字列
ロール	Amazon SNSアラームの作成に使用するIAMロール。	文字列
subject	Amazon SNS通知メッセージの件名。	文字列
topicArn	メッセージの宛先 Amazon ARN SNSトピック。	文字列

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
node	このアクションが実行されている対象ノード。	リファレンスオブジェクト、例：「node」:{「ref"myRunnableObjectId」}
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

終了

保留中または未完了のアクティビティ、リソース、またはデータノードのキャンセル AWS Data Pipeline をトリガーするアクション。 は、アクティビティ、リソース、またはデータノードが `lateAfterTimeout`値で始まらない場合、 CANCELLED 状態になるように試みます。

`onSuccess`、`OnFail`、または `onLateAction` リソースを含むアクションを終了することはできません。

例

以下は、このオブジェクト型の例です。この例では、`MyActivity` の `onLateAction` フィールドは、アクション `DefaultAction1` の参照を含みます。`onLateAction` のアクションを指定するときは、アクティビティが遅れていると見なされてからパイプラインのスケジュールされた開始までの期間を示す `lateAfterTimeout` 値も指定する必要があります。

```
{
  "name" : "MyActivity",
  "id" : "DefaultActivity1",
  "schedule" : {
    "ref" : "MySchedule"
  },
  "runsOn" : {
    "ref" : "MyEmrCluster"
  },
```

```
  "lateAfterTimeout" : "1 Hours",
  "type" : "EmrActivity",
  "onLateAction" : {
    "ref" : "DefaultAction1"
  },
  "step" : [
    "s3://myBucket/myPath/myStep.jar,firstArg,secondArg",
    "s3://myBucket/myPath/myOtherStep.jar,anotherArg"
  ]
},
{
  "name" : "TerminateTasks",
  "id" : "DefaultAction1",
  "type" : "Terminate"
}
```

構文

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
node	このアクションが実行されている対象ノード。	リファレンスオブジェクト、例：「node」:{「ref"myRunnableObjectId」}
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインの ID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

スケジュール

アクティビティの実行など、予定されているイベントのタイミングを定義します。

> **Note**
>
> スケジュールの開始時刻が過去の場合、 はパイプラインを AWS Data Pipeline バックフィルし、指定された開始時刻からすぐに実行のスケジューリングを開始します。テストまたは開発の場合は、比較的短い期間を使用してください。そう AWS Data Pipeline しないと、パイプラインのアクティベーションをブロックして、パイプラインコンポーネント`scheduledStartTime`が 1 日より前である場合に誤ってバックフィルされないように、その interval. AWS Data Pipeline attempts に対してパイプラインのすべての実行をキューに入れ、スケジュールしようとします。

例

以下は、このオブジェクト型の例です。これは、2012-09-01 の 00 00:00 に開始し 2012-10-01 の 00:00:00 に終了する毎時間のスケジュールを定義します。最初の期間は 2012-09-01 の 01:00:00 に終了します。

```
{
  "id" : "Hourly",
  "type" : "Schedule",
  "period" : "1 hours",
```

```
  "startDateTime" : "2012-09-01T00:00:00",
  "endDateTime" : "2012-10-01T00:00:00"
}
```

次のパイプラインは、FIRST_ACTIVATION_DATE_TIME に開始され、2014 年 4 月 25 日の 22:00:00 まで 1 時間ごとに実行されます。

```
{
     "id": "SchedulePeriod",
     "name": "SchedulePeriod",
     "startAt": "FIRST_ACTIVATION_DATE_TIME",
     "period": "1 hours",
     "type": "Schedule",
     "endDateTime": "2014-04-25T22:00:00"
   }
```

次のパイプラインは、FIRST_ACTIVATION_DATE_TIME に開始されます。1 時間ごとに実行され、3 回の実行後、終了されます。

```
{
     "id": "SchedulePeriod",
     "name": "SchedulePeriod",
     "startAt": "FIRST_ACTIVATION_DATE_TIME",
     "period": "1 hours",
     "type": "Schedule",
     "occurrences": "3"
   }
```

次のパイプラインは、2014 年 4 月 25 日の 22:00:00 に開始されます。1 時間ごとに実行され、3 回の実行後、終了されます。

```
{
     "id": "SchedulePeriod",
     "name": "SchedulePeriod",
     "startDateTime": "2014-04-25T22:00:00",
     "period": "1 hours",
     "type": "Schedule",
     "occurrences": "3"
   }
```

Default オブジェクトを使用したオンデマンド

```
{
  "name": "Default",
  "resourceRole": "DataPipelineDefaultResourceRole",
  "role": "DataPipelineDefaultRole",
  "scheduleType": "ondemand"
}
```

明示的な Schedule オブジェクトを使用したオンデマンド

```
{
  "name": "Default",
  "resourceRole": "DataPipelineDefaultResourceRole",
  "role": "DataPipelineDefaultRole",
  "scheduleType": "ondemand"
},
{
  "name": "DefaultSchedule",
  "type": "Schedule",
  "id": "DefaultSchedule",
  "period": "ONDEMAND_PERIOD",
  "startAt": "ONDEMAND_ACTIVATION_TIME"
},
```

次の例は、Default オブジェクトから Schedule を継承する方法、そのオブジェクトに対して明示的に設定する方法、または親参照から指定する方法を示します。

Default オブジェクトから継承されたスケジュール

```
{
  "objects": [
  {
      "id": "Default",
      "failureAndRerunMode":"cascade",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "role": "DataPipelineDefaultRole",
      "pipelineLogUri": "s3://myLogsbucket",
      "scheduleType": "cron",
      "schedule": {
        "ref": "DefaultSchedule"
      }
   },
   {
```

```
      "type": "Schedule",
      "id": "DefaultSchedule",
      "occurrences": "1",
      "period": "1 Day",
      "startAt": "FIRST_ACTIVATION_DATE_TIME"
    },
    {
      "id": "A_Fresh_NewEC2Instance",
      "type": "Ec2Resource",
      "terminateAfter": "1 Hour"
    },
    {
      "id": "ShellCommandActivity_HelloWorld",
      "runsOn": {
        "ref": "A_Fresh_NewEC2Instance"
      },
      "type": "ShellCommandActivity",
      "command": "echo 'Hello World!'"
    }
  ]
}
```

オブジェクトの明示的なスケジュール

```
{
  "objects": [
  {
      "id": "Default",
      "failureAndRerunMode":"cascade",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "role": "DataPipelineDefaultRole",
      "pipelineLogUri": "s3://myLogsbucket",
      "scheduleType": "cron"

   },
   {
      "type": "Schedule",
      "id": "DefaultSchedule",
      "occurrences": "1",
      "period": "1 Day",
      "startAt": "FIRST_ACTIVATION_DATE_TIME"
    },
    {
```

```
      "id": "A_Fresh_NewEC2Instance",
      "type": "Ec2Resource",
      "terminateAfter": "1 Hour"
    },
    {
      "id": "ShellCommandActivity_HelloWorld",
      "runsOn": {
        "ref": "A_Fresh_NewEC2Instance"
      },
      "schedule": {
        "ref": "DefaultSchedule"
      },
      "type": "ShellCommandActivity",
      "command": "echo 'Hello World!'"
    }
  ]
}
```

親参照からのスケジュール

```
{
  "objects": [
  {
      "id": "Default",
      "failureAndRerunMode":"cascade",
      "resourceRole": "DataPipelineDefaultResourceRole",
      "role": "DataPipelineDefaultRole",
      "pipelineLogUri": "s3://myLogsbucket",
      "scheduleType": "cron"

   },
   {
      "id": "parent1",
      "schedule": {
        "ref": "DefaultSchedule"
      }

   },
   {
      "type": "Schedule",
      "id": "DefaultSchedule",
      "occurrences": "1",
      "period": "1 Day",
```

```
      "startAt": "FIRST_ACTIVATION_DATE_TIME"
    },
    {
      "id": "A_Fresh_NewEC2Instance",
      "type": "Ec2Resource",
      "terminateAfter": "1 Hour"
    },
    {
      "id": "ShellCommandActivity_HelloWorld",
      "runsOn": {
        "ref": "A_Fresh_NewEC2Instance"
      },
      "parent": {
        "ref": "parent1"
      },
      "type": "ShellCommandActivity",
      "command": "echo 'Hello World!'"
    }
  ]
}
```

構文

必須フィールド	説明	スロットタイプ
period	パイプラインの実行頻度。形式は、"N [minutes|hours|days|weeks|months]" です。ここで N は数字で、その後に時間指定子の 1 つを続けます。たとえば、"15 minutes" は、15 分ごとにパイプラインを実行します。最小間隔は 15 分で、最大間隔は 3 年です。	[Period] (期間)

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
startAt	スケジュールされたパイプライン実行を開始する日時。有効な値は FIRST_ACTIVATIONDA	一覧表

必須のグループ (次のいずれかが必要です)	説明	スロットタイプ
	TE_ でTIME、オンデマンドパイプラインの作成を優先して廃止されます。	
startDateTime	スケジュールした実行を開始する日時。startDateTime または のいずれかを使用する必要があります startAt が、両方は使用できません。	DateTime

オプションのフィールド	説明	スロットタイプ
endDateTime	スケジュールした実行が終了する日時。startDateTime または の値より後の日付と時刻である必要がありますstartAt。デフォルトの動作では、パイプラインがシャットダウンされるまで実行をスケジューします。	DateTime
occurrences	パイプラインをアクティブ化してから実行する回数。で出現を使用することはできません endDateTime。	整数
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref'myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列
@firstActivationTime	オブジェクト作成の時刻。	DateTime
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

ユーティリティ

次のようなユーティリティオブジェクトでは、その他のパイプラインオブジェクトを設定します。

トピック

- ShellScriptConfig
- EmrConfiguration
- プロパティ

ShellScriptConfig

アクティビティで を使用して、 preActivityTaskConfig および postActivityTaskConfig のシェルスクリプトを実行します。このオブジェクトは、HadoopActivity、、HiveActivity、HiveCopyActivityおよび で使用できますPigActivity。スクリプトの S3 URIと引数のリストを指定します。

例

引数 ShellScriptConfig を持つ :

```
{
   "id" : "ShellScriptConfig_1”,
   "name" : “prescript”,
```

```
  "type" : "ShellScriptConfig",
  "scriptUri": “s3://my-bucket/shell-cleanup.sh”,
  "scriptArgument" : ["arg1","arg2"]
}
```

構文

このオブジェクトは次のフィールドを含みます。

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
scriptArgument	シェルスクリプトで使う引数のリスト。	文字列
scriptUri	ダウンロードして実行する必要がある Amazon S3 URIのスクリプト。	文字列

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列

システムフィールド	説明	スロットタイプ
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

EmrConfiguration

EmrConfiguration オブジェクトは、リリース 4.0.0 以降のEMRクラスターに使用される設定です。設定 (リストとして) は RunJobFlow API呼び出しのパラメータです。Amazon APIの設定EMRでは、分類と properties. AWS Data Pipeline uses を対応する Property オブジェクト EmrConfiguration とともに使用して、パイプラインの実行で起動されたEMRクラスターで Hadoop、Hive、Spark、Pig などのEmrClusterアプリケーションを設定します。設定は新しいクラスターに対してのみ変更できるため、既存のリソースに EmrConfiguration オブジェクトを指定することはできません。詳細については、「https://docs.aws.amazon.com/ElasticMapReduce/latest/ReleaseGuide/」を参照してください。

例

次の設定オブジェクトは、`io.file.buffer.size` で `fs.s3.block.size` および `core-site.xml` プロパティを設定します。

```
[
   {
      "classification":"core-site",
      "properties":
      {
         "io.file.buffer.size": "4096",
         "fs.s3.block.size": "67108864"
      }
   }
]
```

対応するパイプラインオブジェクト定義は、 EmrConfiguration オブジェクトと `property`フィールド内のプロパティオブジェクトのリストを使用します。

```
{
  "objects": [
```

```
    {
      "name": "ReleaseLabelCluster",
      "releaseLabel": "emr-4.1.0",
      "applications": ["spark", "hive", "pig"],
      "id": "ResourceId_I1mCc",
      "type": "EmrCluster",
      "configuration": {
        "ref": "coresite"
      }
    },
    {
      "name": "coresite",
      "id": "coresite",
      "type": "EmrConfiguration",
      "classification": "core-site",
      "property": [{
        "ref": "io-file-buffer-size"
      },
      {
        "ref": "fs-s3-block-size"
      }
      ]
    },
    {
      "name": "io-file-buffer-size",
      "id": "io-file-buffer-size",
      "type": "Property",
      "key": "io.file.buffer.size",
      "value": "4096"
    },
    {
      "name": "fs-s3-block-size",
      "id": "fs-s3-block-size",
      "type": "Property",
      "key": "fs.s3.block.size",
      "value": "67108864"
    }
  ]
}
```

次の例は、hadoop-env 分類で Hadoop 環境を設定するためのネスト設定です。

```
[
```

```
  {
    "classification": "hadoop-env",
    "properties": {},
    "configurations": [
      {
        "classification": "export",
        "properties": {
          "YARN_PROXYSERVER_HEAPSIZE": "2396"
        }
      }
    ]
  }
]
```

この設定を使用する、対応するパイプライン定義オブジェクトを次に示します。

```
{
  "objects": [
    {
      "name": "ReleaseLabelCluster",
      "releaseLabel": "emr-4.0.0",
      "applications": ["spark", "hive", "pig"],
      "id": "ResourceId_I1mCc",
      "type": "EmrCluster",
      "configuration": {
        "ref": "hadoop-env"
      }
    },
    {
      "name": "hadoop-env",
      "id": "hadoop-env",
      "type": "EmrConfiguration",
      "classification": "hadoop-env",
      "configuration": {
        "ref": "export"
      }
    },
    {
      "name": "export",
      "id": "export",
      "type": "EmrConfiguration",
      "classification": "export",
      "property": {
```

```
        "ref": "yarn-proxyserver-heapsize"
      }
    },
    {
      "name": "yarn-proxyserver-heapsize",
      "id": "yarn-proxyserver-heapsize",
      "type": "Property",
      "key": "YARN_PROXYSERVER_HEAPSIZE",
      "value": "2396"
    },
  ]
}
```

次の例では、 EMRクラスターの Hive 固有のプロパティを変更します。

```
{
    "objects": [
        {
            "name": "hivesite",
            "id": "hivesite",
            "type": "EmrConfiguration",
            "classification": "hive-site",
            "property": [
                {
                    "ref": "hive-client-timeout"
                }
            ]
        },
        {
            "name": "hive-client-timeout",
            "id": "hive-client-timeout",
            "type": "Property",
            "key": "hive.metastore.client.socket.timeout",
            "value": "2400s"
        }
    ]
}
```

構文

このオブジェクトは次のフィールドを含みます。

必須フィールド	説明	スロットタイプ
分類	設定の分類。	文字列

オプションのフィールド	説明	スロットタイプ
設定	この設定のサブ設定。	リファレンスオブジェクト、例：「configuration」:{「ref "myEmrConfigurationId」}
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}
property	設定のプロパティ。	リファレンスオブジェクト、例：「プロパティ」:{「ref"myPropertyId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトが作成されたパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー	文字列

システムフィールド	説明	スロットタイプ
@pipelineId	このオブジェクトが属するパイプラインの ID	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します	文字列

以下の資料も参照してください。

- EmrCluster
- プロパティ
- Amazon EMR リリースガイド

プロパティ

EmrConfiguration オブジェクトで使用する単一のキーと値のプロパティ。

例

次のパイプライン定義は、 を起動するための EmrConfiguration オブジェクトと対応する プロパティオブジェクトを示しています EmrCluster。

```
{
  "objects": [
    {
      "name": "ReleaseLabelCluster",
      "releaseLabel": "emr-4.1.0",
      "applications": ["spark", "hive", "pig"],
      "id": "ResourceId_I1mCc",
      "type": "EmrCluster",
      "configuration": {
        "ref": "coresite"
      }
    },
    {
      "name": "coresite",
      "id": "coresite",
```

```
      "type": "EmrConfiguration",
      "classification": "core-site",
      "property": [{
        "ref": "io-file-buffer-size"
      },
      {
        "ref": "fs-s3-block-size"
      }
      ]
    },
    {
      "name": "io-file-buffer-size",
      "id": "io-file-buffer-size",
      "type": "Property",
      "key": "io.file.buffer.size",
      "value": "4096"
    },
    {
      "name": "fs-s3-block-size",
      "id": "fs-s3-block-size",
      "type": "Property",
      "key": "fs.s3.block.size",
      "value": "67108864"
    }
  ]
}
```

構文

このオブジェクトは次のフィールドを含みます。

必須フィールド	説明	スロットタイプ
キー	キー	文字列
value	value	文字列

オプションのフィールド	説明	スロットタイプ
parent	スロットの継承元となる現在のオブジェクトの親。	リファレンスオブジェクト、例：「parent」:{「ref"myBaseObjectId」}

実行時フィールド	説明	スロットタイプ
@version	オブジェクトを作成したパイプラインのバージョン。	文字列

システムフィールド	説明	スロットタイプ
@error	形式が正しくないオブジェクトを説明するエラー。	文字列
@pipelineId	このオブジェクトが属するパイプラインのID。	文字列
@sphere	オブジェクトの球は、ライフサイクルにおける場所を示します。コンポーネントオブジェクトにより、試行オブジェクトを実行するインスタンスオブジェクトが発生します。	文字列

以下の資料も参照してください。

- EmrCluster
- EmrConfiguration
- Amazon EMR リリースガイド

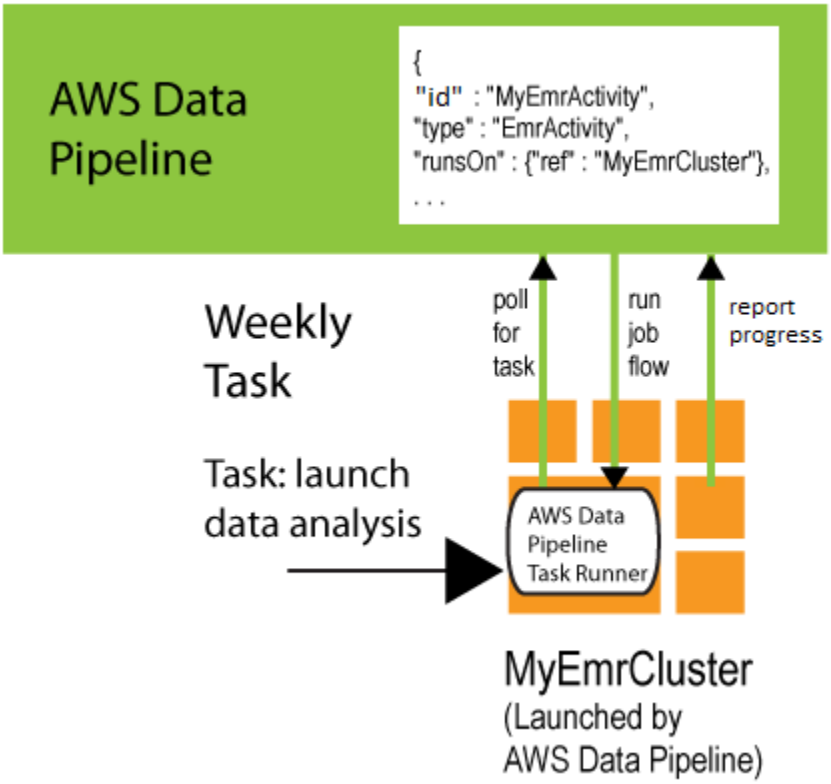
Task Runnerの操作

Task Runner は、スケジュールされたタスク AWS Data Pipeline をポーリングして Amazon EC2インスタンス、Amazon EMRクラスター、またはその他の計算リソースで実行し、ステータスを報告するタスクエージェントアプリケーションです。アプリケーションによっては、以下を行うことができます。

- AWS Data Pipeline に 1 つ以上の Task Runner アプリケーションのインストールと管理を許可します。パイプラインがアクティブ化されると、アクティビティ runsOn フィールドによって参照されるデフォルトの `Ec2Instance`または `EmrCluster` オブジェクトが自動的に作成されます。 は、EC2インスタンスまたは EMRクラスターのマスターノードに Task Runner AWS Data Pipeline をインストールします。このパターンでは、 AWS Data Pipeline はほとんどのインスタンスまたはクラスター管理を実行できます。
- パイプラインの全体または一部を、お客様が管理するリソースで実行する。潜在的なリソースには、長時間実行される Amazon EC2インスタンス、Amazon EMRクラスター、または物理サーバーが含まれます。タスクランナー (Task Runner または独自のデバイスのカスタムタスクエージェントのいずれか) は、 AWS Data Pipeline ウェブサービスと通信できる限り、ほぼどこでもインストールできます。このパターンでは、どのリソースが使用されどのように管理されるかをお客様がほぼ完全に制御できますが、Task Runnerは手動でインストールおよび設定する必要があります。これを行うには、「Task Runnerを使用した既存のリソースでの作業の実行」に記載されている手順を使用します。

AWS Data Pipelineマネージドリソースの Task Runner

リソースが によって起動および管理されると AWS Data Pipeline、ウェブサービスは、そのリソースに Task Runner を自動的にインストールして、パイプライン内のタスクを処理します。アクティビティオブジェクトの `runsOn`フィールドに計算リソース (Amazon EC2インスタンスまたは Amazon EMRクラスター) を指定します。 AWS Data Pipeline は、このリソースを起動するときに、そのリソースにTask Runnerをインストールし、`runsOn`フィールドがそのリソースに設定されている、すべてのアクティビティオブジェクトを処理するよう設定します。がリソース AWS Data Pipeline を終了すると、Task Runner ログはシャットダウンする前に Amazon S3 の場所に発行されます。



たとえば、パイプラインでEmrActivityを使用するばあいは、runsOnフィールドでEmrClusterリソースを指定します。がそのアクティビティ AWS Data Pipeline を処理すると、Amazon EMRクラスターを起動し、Task Runner をマスターノードにインストールします。次に、このTask Runnerは、runsOnフィールドがそのEmrClusterオブジェクトに設定されている、アクティビティのタスクを処理します。次のパイプライン定義の抜粋は、2つのオブジェクト間のこの関係を示します。

```
{
  "id" : "MyEmrActivity",
  "name" : "Work to perform on my data",
  "type" : "EmrActivity",
  "runsOn" : {"ref" : "MyEmrCluster"},
  "preStepCommand" : "scp remoteFiles localFiles",
  "step" : "s3://myBucket/myPath/myStep.jar,firstArg,secondArg",
  "step" : "s3://myBucket/myPath/myOtherStep.jar,anotherArg",
  "postStepCommand" : "scp localFiles remoteFiles",
  "input" : {"ref" : "MyS3Input"},
  "output" : {"ref" : "MyS3Output"}
},
{
```

```
  "id" : "MyEmrCluster",
  "name" : "EMR cluster to perform the work",
  "type" : "EmrCluster",
  "hadoopVersion" : "0.20",
  "keypair" : "myKeyPair",
  "masterInstanceType" : "m1.xlarge",
  "coreInstanceType" : "m1.small",
  "coreInstanceCount" : "10",
  "taskInstanceType" : "m1.small",
  "taskInstanceCount": "10",
  "bootstrapAction" : "s3://elasticmapreduce/libs/ba/configure-hadoop,arg1,arg2,arg3",
  "bootstrapAction" : "s3://elasticmapreduce/libs/ba/configure-other-stuff,arg1,arg2"
}
```

このアクティビティの実行の詳細および例については、「EmrActivity」を参照してください。

パイプラインに複数の AWS Data Pipelineマネージドリソースがある場合、Task Runner は各リソースにインストールされ、それらはすべて処理するタスク AWS Data Pipeline をポーリングします。

Task Runnerを使用した既存のリソースでの作業の実行

Task Runner は、Amazon EC2インスタンス、物理サーバー、ワークステーションなど、管理する計算リソースにインストールできます。Task Runner は、 AWS Data Pipeline ウェブサービスと通信できる場合に限り、互換性のあるハードウェアまたはオペレーティングシステムのどこにでもインストールできます。

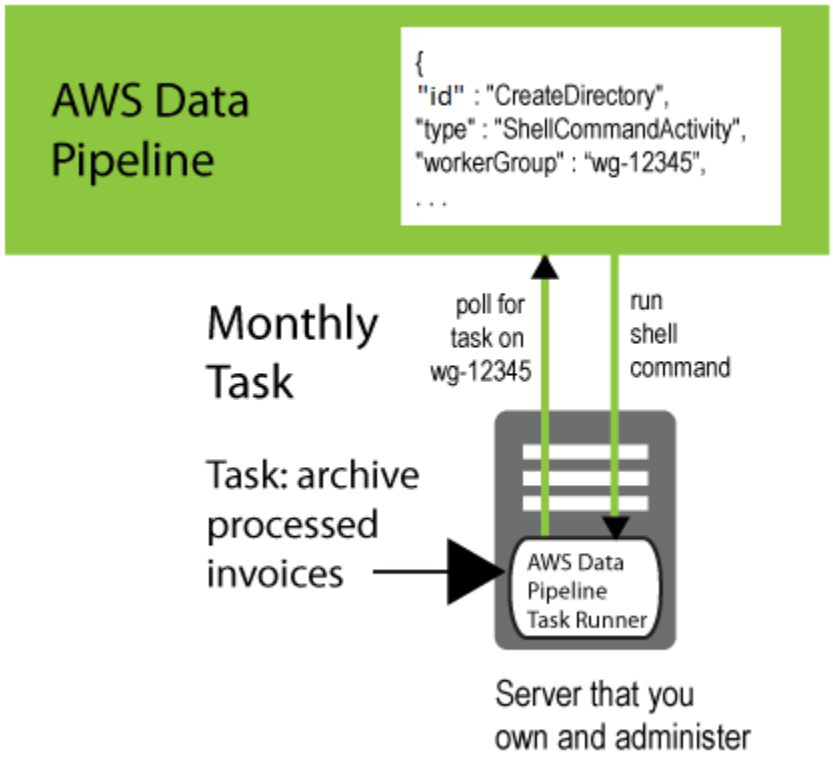
このアプローチは、例えば、 AWS Data Pipeline を使用して組織のファイアウォール内に保存されているデータを処理したい場合に役立ちます。ローカルネットワークのサーバーに Task Runner をインストールすることで、ローカルデータベースに安全にアクセスし、 AWS Data Pipeline 次のタスクを実行するようにポーリングできます。がパイプラインの処理 AWS Data Pipeline を終了するか、パイプラインを削除すると、Task Runner インスタンスは、手動でシャットダウンするまで計算リソースで実行されたままになります。Task Runnerのログはパイプライン実行の完了後も維持されます。

お客様が管理するリソースでTask Runnerを使用するには、まずTask Runnerをダウンロードし、このセクションの手順に従って、お客様のコンピューティングリソースにインストールします。

> **Note**
>
> Task Runner は、Linux、UNIX、または macOS にのみインストールできます。Task Runner はWindowsオペレーティングシステムではサポートされていません。

Task Runner 2.0を使用するのに必要なJavaの最小バージョンは1.7です。

処理する必要のあるパイプラインのアクティビティに、インストールしたTask Runnerを接続するには、オブジェクトにworkerGroupフィールドを追加し、そのワーカーグループの値をポーリングするようにTask Runnerを設定します。そのためには、Task Runner JAR ファイルを実行するときに、ワーカーグループ文字列をパラメータ (など--workerGroup=wg-12345) として渡します。




```
{
  "id" : "CreateDirectory",
  "type" : "ShellCommandActivity",
  "workerGroup" : "wg-12345",
  "command" : "mkdir new-directory"
}
```

Task Runnerのインストール

このセクションでは、Task Runnerをインストールして設定する方法とその前提条件について説明します。インストールは、手動の簡単なプロセスです。

Task Runnerをインストールするには

1. Task RunnerはJavaバージョン1.6または1.8を必要とします。Javaがインストールされているかどうか、およびどのバージョンが実行されているかを確認するには、次のコマンドを使用します。

   ```
   java -version
   ```

   コンピューターにJava 1.6または1.8がインストールされていない場合は、http://www.oracle.com/technetwork/java/index.htmlからこれらのいずれかのバージョンをダウンロードします。Javaをダウンロードしてインストールし、次のステップに進みます。

2. https://s3.amazonaws.com/datapipeline-us-east-1/us-east-1/software/latest/TaskRunner/TaskRunner-1.0.jar `TaskRunner-1.0.jar`からダウンロードし、ターゲットコンピューティングリソースのフォルダにコピーします。`EmrActivity` タスクを実行している Amazon EMRクラスターの場合は、クラスターのマスターノードに Task Runner をインストールします。

3. Task Runner を使用して AWS Data Pipeline ウェブサービスに接続してコマンドを処理する場合、ユーザーはデータパイプラインを作成または管理する権限を持つロールにプログラムでアクセスする必要があります。詳細については、「プログラムによるアクセス権を付与する」を参照してください。

4. Task Runner は、 を使用して AWS Data Pipeline ウェブサービスに接続しますHTTPS。 AWS リソースを使用している場合HTTPSは、適切なルーティングテーブルとサブネット で が有効になっていることを確認しますACL。ファイアウォールまたはプロキシを使用している場合は、ポート443が開いていることを確認します。

(オプション) Task Runner に Amazon へのアクセス権を付与する RDS

Amazon RDS では、データベースセキュリティグループ (DB セキュリティグループ) を使用して DB インスタンスへのアクセスを制御できます。DBセキュリティグループは、DBインスタンスへのネットワークアクセスをコントロールするファイアウォールのように動作します。デフォルトでは、DBインスタンスへのネットワークアクセスは無効です。Task Runner が Amazon RDSインスタンスにアクセスできるようにするには、DB セキュリティグループを変更する必要があります。Task Runner は、実行されるインスタンスから Amazon RDS アクセスを取得するため、Amazon RDSインスタンスに追加するアカウントとセキュリティグループは、Task Runner のインストール場所によって異なります。

EC2-Classic で Task Runner へのアクセスを許可するには

1. Amazon RDSコンソールを開きます。
2. ナビゲーションペインで [Instances]を選択し、DBインスタンスを選択します。
3. [Security and Network] (セキュリティとネットワーク) で、セキュリティグループを選択します。このDBセキュリティグループが選択された状態で[Security Groups] (セキュリティグループ) ページが開きます。DBセキュリティグループの詳細アイコンを選択します。
4. [Security Group Details]で、適切な[Connection Type]と[Details]を指定してルールを作成します。ここで説明したように、これらのフィールドは、Task Runnerが実行されている場所によって異なります。

   - `Ec2Resource`
     - 接続タイプ：`EC2 Security Group`

       詳細: *`my-security-group-name`* (EC2インスタンス用に作成したセキュリティグループの名前)
   - `EmrResource`
     - 接続タイプ：`EC2 Security Group`

       詳細：`ElasticMapReduce-master`
     - 接続タイプ：`EC2 Security Group`

       詳細：`ElasticMapReduce-slave`
   - ローカル環境 (オンプレミス)
     - 接続タイプ：`CIDR/IP`：

       詳細: *`my-ip-address`* (コンピュータの IP アドレス、またはコンピュータがファイアウォールの内側にある場合はネットワークの IP アドレス範囲)
5. [Add] (追加) をクリックします。

EC2- で Task Runner へのアクセスを許可するにはVPC

1. Amazon RDSコンソールを開きます。
2. ナビゲーションペインで、[インスタンス] を選択します。
3. DBインスタンスの詳細アイコンを選択します。セキュリティとネットワーク で、セキュリティグループへのリンクを開き、Amazon EC2コンソールに移動します。セキュリティグループで使

用しているコンソールデザインが古い場合は、コンソールページの上部に表示されるアイコンを選択して、新しいコンソールデザインに切り替えます。

4. [Inbound]タブで、[Edit]、[Add Rule]を選択します。DBインスタンスを起動したときに使用するデータベースポートを指定します。ここで説明したように、ソースは、Task Runnerが実行されている場所によって異なります。

   - Ec2Resource
     - *my-security-group-id*（EC2インスタンス用に作成したセキュリティグループの ID)
   - EmrResource
     - *master-security-group-id*（ElasticMapReduce-masterセキュリティグループの ID)
     - *slave-security-group-id*（ElasticMapReduce-slaveセキュリティグループの ID)
   - ローカル環境（オンプレミス）
     - *ip-address*（コンピュータの IP アドレス、またはコンピュータがファイアウォールの内側にある場合はネットワークの IP アドレス範囲）

5. [Save]（保存）をクリックします。

Task Runnerの起動

Task Runnerをインストールしたディレクトリに設定されている新しいコマンドプロンプトウィンドウで次のコマンドを実行してTask Runnerを起動します。

```
java -jar TaskRunner-1.0.jar --config ~/credentials.json --workerGroup=myWorkerGroup --
region=MyRegion --logUri=s3://mybucket/foldername
```

--configオプションは認証情報ファイルを指します。

--workerGroupオプションはワーカーグループの名前を指定します。これは、パイプラインで指定されている処理対象のタスクの値と同じである必要があります。

--regionオプションは、実行するタスクをプルする対象のサービスリージョンを指定します。

--logUriオプションは、圧縮済みログをAmazon S3の場所にプッシュするために使用されます。

Task Runnerは、アクティブなとき、ログファイルが書き込まれる場所へのパスをターミナルウィンドウに出力します。次に例を示します。

```
Logging to /Computer_Name/.../output/logs
```

Task Runnerはログインシェルから切り離されて実行される必要があります。ターミナルアプリケーションを使用してコンピューターに接続している場合は、nohupやscreenのようなユーティリティを使用して、ログアウト時にTask Runnerアプリケーションが終了しないようにする必要があります。コマンドラインオプションの詳細については、「Task Runnerの設定オプション」を参照してください。

Task Runnerのログの確認

Task Runnerが動作していることを確認する最も簡単な方法は、ログファイルを書き込んでいるかどうかをチェックすることです。Task Runnerは、Task Runnerがインストールされているディレクトリの下にあるディレクトリoutput/logsに時間ごとのログファイルを書き込みます。ファイル名はでTask Runner.log.YYYY-MM-DD-HH、HH は 00 から 23 まで実行されますUDT。ストレージ領域を節約するために、8 時間以上経過したログファイルは で圧縮されますGZip。

Task Runnerのスレッドと前提条件

Task Runnerでは、タスク、アクティビティ、前提条件ごとに1つのスレッドプールが使用されます。のデフォルト設定--tasksは 2 です。つまり、タスクプールから 2 つのスレッドが割り当てられ、各スレッドが新しいタスクのために AWS Data Pipeline サービスをポーリングします。つまり、--tasksはパフォーマンスチューニング用の属性であり、パイプラインのスループット最適化のために使用できます。

前提条件のパイプライン再試行ロジックはTask Runnerで実行されます。前提条件オブジェクトのポーリングには、2 AWS Data Pipeline つの前提条件スレッドが割り当てられます。Task Runnerは、前提条件で定義した前提条件オブジェクトretryDelayとpreconditionTimeoutフィールドを尊重します。

多くの場合、前提条件のポーリングのタイムアウトと再試行数の値を小さくすると、アプリケーションのパフォーマンスが向上する場合があります。同様に、前提条件が長時間実行されるアプリケーションでは、タイムアウトと再試行数の値を大きくする必要がある可能性があります。前提条件オブジェクトの詳細については、「前提条件」を参照してください。

Task Runnerの設定オプション

Task Runnerの起動時にコマンドラインから使用できる設定オプションを以下に示します。

コマンドラインパラメータ	説明
`--help`	コマンドラインのヘルプ。例: `Java -jar TaskRunner-1.0.jar --help`
`--config`	`credentials.json` ファイルのパスとファイル名。
`--accessId`	リクエストを行うときに使用する Task Runner の AWS アクセスキー ID。 `--accessID` オプションと`--secretKey` オプションを使用すると、credentials.jsonファイルを使用する代わりになります。`credentials.json` ファイルも指定した場合は、`--accessID` オプションと`--secretKey` オプションが優先されます。
`--secretKey`	Task Runner がリクエストを行うときに使用するシー AWS クレットキー。詳細については、「`--accessID` 」を参照してください。
`--endpoint`	エンドポイントは、ウェブサービスのURLエントリポイントである です。リクエストを行うリージョンのサービス AWS Data Pipeline エンドポイント。オプション。一般的にはリージョンの指定で充分であり、エンドポイントを設定する必要はありません。 AWS Data Pipeline リージョンとエンドポイントのリストについては、「」のAWS「データパイプラインのリージョンとエンドポイント」を参照してくださいAWS 全般のリファレンス。
`--workerGroup`	Task Runnerが作業を取得する対象のワーカーグループの名前。必須。 Task Runnerは、ウェブサービスに対するポーリングを行う場合、お客様によって指定された

コマンドラインパラメータ	説明
	認証情報とのworkerGroup 値を使用して、取得するタスク (あれば) を選択します。この値には、わかりやすい任意の名前を使用できます。唯一の要件は、Task Runnerと、対応するパイプラインアクティビティで、この文字列が一致することです。ワーカーグループ名はリージョンにバインドされます。同一のワーカーグループ名が他のリージョン内に存在しても、Task Runnerは常に--regionで指定されたリージョンからタスクを取得します。
--taskrunnerId	進捗状況をレポートするときに使用するTask RunnerのID。オプション。
--output	ログ出力ファイルのTask Runnerディレクトリ。オプション。ログファイルは、Amazon S3にプッシュされるまでローカルディレクトリに保管されます。このオプションは、デフォルトディレクトリを上書きします。
--region	使用するリージョン。オプションですが、常にリージョンを設定することをお勧めします。リージョンを指定しなかった場合、Task Runnerでは、デフォルトのサービスリージョンus-east-1 からタスクが取得されます。 その他のサポートされているリージョンとして、eu-west-1 、ap-northeast-1 、ap-southeast-2 、us-west-2 があります。
--logUri	Task Runnerが1時間おきにログファイルをバックアップする先のAmazon S3の場所へのパス。Task Runnerが終了すると、ローカルディレクトリ内のアクティブなログがAmazon S3にあるバックアップ先フォルダにプッシュされます。

コマンドラインパラメータ	説明
`--proxyHost`	Task Runnerクライアントが AWSサービスに接続するために使用するプロキシのホスト。
`--proxyPort`	Task Runnerクライアントが AWSサービスに接続するために使用するプロキシホストのポート。
`--proxyUsername`	プロキシ用のユーザー名。
`--proxyPassword`	プロキシ用のパスワード。
`--proxyDomain`	NTLM Proxy の Windows ドメイン名。
`--proxyWorkstation`	NTLM Proxy の Windows ワークステーション名。

プロキシ経由によるTask Runnerの使用

プロキシホストを使用している場合は、Task Runner を呼び出すときにその設定を指定するか、環境変数 HTTPS_ を設定できますPROXY。Task Runner で使用される環境変数は、AWSコマンドラインインターフェイス に使用されるのと同じ設定を受け入れます。

Task Runner とカスタム AMIs

パイプラインのEc2Resourceオブジェクトを指定すると、 は Task Runner をインストールして設定AMIする を使用してEC2インスタンス AWS Data Pipeline を作成します。この場合、PV互換のインスタンスタイプが必要です。または、Task Runner AMI でカスタム を作成し、 `Ec2Resource` オブジェクトの `imageId`フィールドAMIを使用してこの ID を指定することもできます。詳細については、「Ec2Resource」を参照してください。

カスタムを Task Runner で正常に使用するには AWS Data Pipeline 、 の次の要件を満たしているAMI必要があります。

- インスタンスが実行されるAMIのと同じリージョンに を作成します。詳細については、「Amazon ユーザーガイド」の「独自の の作成AMI」を参照してください。 EC2

- の仮想化タイプAMIが、使用する予定のインスタンスタイプでサポートされていることを確認します。例えば、I2 および G2 インスタンスタイプには が必要HVMAMIで、T1, C1, M1、および M2 インスタンスタイプには PV が必要ですAMI。詳細については、「Amazon ユーザーガイド」の「Linux AMI 仮想化タイプ」を参照してください。 EC2
- 次のソフトウェアをインストールしてください。
  - Linux
  - Bash
  - wget
  - unzip
  - Java 1.6または1.8
  - cloud-init
- `ec2-user`という名前のユーザーを作成して設定します。

トラブルシューティング

AWS Data Pipeline に問題がある場合、最もよくみられる症状はパイプラインが実行されないことです。コンソールと CLI から提供されるデータを使用すると、問題を特定し、解決方法を見つけることができます。

目次

- パイプラインのエラーを見つける
- パイプラインを処理する Amazon EMR クラスターの特定
- パイプラインのステータスの詳細を解釈する
- エラーログを見つける
- 一般的な問題を解決する

パイプラインのエラーを見つける

AWS Data Pipeline コンソールは、パイプラインの状態を視覚的に監視し、失敗または未完了のパイプラインの実行に関連するエラーを簡単に見つけることができる便利なツールです。

コンソールを使用して、失敗または未完了の実行に関するエラーを見つけるには

1. [List Pipelines] ページで、パイプラインインスタンスの [Status] 列に [FINISHED] 以外の値が表示されている場合は、前提条件が満たされるのをパイプラインが待っているか、または、パイプラインが失敗してパイプラインの問題を解決する必要があります。
2. [List Pipelines] (パイプラインの一覧表示) ページで、インスタンスパイプラインを見つけ、その左側にある三角形を選択して、詳細を展開します。
3. このパネルの下部にある [View execution details] (実行の詳細の表示) を選択します。[Instance summary] (インスタンスの概要) パネルが開き、指定したインスタンスの詳細が表示されます。
4. [Instance summary (インスタンスの概要)] パネルで、インスタンスの横にある三角形を選択して、インスタンスのその他の詳細を表示し、[詳細]、[More... (さらに...)] の順に選択します。選択したインスタンスの状態が [失敗] である場合は、詳細ボックスにエラーメッセージ、`errorStackTrace`、その他の情報が表示されます。この情報をファイルに保存することができます。[OK] をクリックします。
5. [Instance summary] (インスタンスの概要) ペインで、[Attempts] (試行) タブを選択し、試行の行ごとの詳細を確認します。

6. 未完了のインスタンスや失敗したインスタンスに対してアクションを実行するには、インスタンスの横にあるチェックボックスをオンにします。これによりアクションがアクティブ化されます。次に、アクション (`Rerun|Cancel|Mark Finished`) を選択します。

パイプラインを処理する Amazon EMR クラスターの特定

`EMRCluster` または `EMRActivity` が失敗して、AWS Data Pipeline コンソールに表示されるエラー情報が明確でない場合は、Amazon EMR コンソールを使用して、パイプラインを処理する Amazon EMR クラスターを特定することができます。これは、発生したエラーの詳細について記録した Amazon EMR のログを見つけるのに役立ちます。

詳細な Amazon EMR エラー詳細情報を表示するには

1. AWS Data Pipeline コンソールで、パイプラインインスタンスの横にある三角形を選択し、インスタンスの詳細を展開します。

2. [View execution details] (実行の詳細の表示) を選択し、コンポーネントの横にある三角形を選択します。
3. [Details] (詳細) 列の [More...] (さらに...) を選択します。情報画面が開き、コンポーネントの詳細情報がリストされます。画面から [instanceParent] 値 (`@EmrActivityId_xiFDD_2017-09-30T21:40:13` など) を見つけてコピーします。
4. Amazon EMR コンソールに移動して、名前が [instanceParent] の値と一致するクラスターを探し、[Debug] (デバッグ) を選択します。

   > **Note**
   >
   > [Debug] ボタンが機能するには、パイプライン定義で EmrActivity の `enableDebugging` オプションを `true` に設定し、`EmrLogUri` オプションに有効なパスを設定しておく必要があります。

5. これで、どの Amazon EMR クラスターにパイプラインの失敗の原因となったエラーが含まれるかが判明しました。Amazon EMR 開発者ガイドに記載されている[トラブルシューティングのヒント]に従ってください。

パイプラインのステータスの詳細を解釈する

AWS Data Pipeline コンソールと CLI に表示されるさまざまなステータスレベルは、パイプラインとそのコンポーネントの状態を示します。パイプラインのステータスは、単にパイプラインの概要を表します。詳細を確認するには、個別のパイプラインコンポーネントのステータスを表示します。これは、コンソールでパイプラインをクリックするか、または CLI を使用してパイプラインコンポーネントの詳細を取得することで、実行できます。

ステータスコード

`ACTIVATING`

EC2 インスタンスなど、コンポーネントまたはリソースが開始されています。

`CANCELED`

コンポーネントは、実行前にユーザーまたは AWS Data Pipeline によってキャンセルされました。これは、このコンポーネントが依存する別のコンポーネントまたはリソースで障害が発生した場合に自動的に発生する可能性があります。

`CASCADE_FAILED`

コンポーネントまたはリソースが、依存関係の 1 つによるカスケード障害の結果としてキャンセルされましたが、そのコンポーネントはおそらく障害の元の原因ではありませんでした。

`DEACTIVATING`

パイプラインが非アクティブ化されています。

`FAILED`

コンポーネントまたはリソースでエラーが発生し、動作を停止しました。コンポーネントまたはリソースに障害が発生すると、キャンセルや障害がそれに依存する他のコンポーネントにカスケードする可能性があります。

`FINISHED`

コンポーネントは割り当てられた作業を完了しました。

`INACTIVE`

パイプラインが非アクティブ化されました。

`PAUSED`

コンポーネントは一時停止され、現在作業を実行していません。

PENDING

パイプラインを初めてアクティブ化する準備が整いました。

RUNNING

リソースは実行中であり、作業を受け取る準備が整いました。

SCHEDULED

リソースの実行がスケジュールされています。

SHUTTING_DOWN

リソースは、作業が正常に完了した後にシャットダウンしています。

SKIPPED

現在のスケジュールより後のタイムスタンプを使用して、パイプラインがアクティブ化された後、コンポーネントは実行間隔をスキップしました。

TIMEDOUT

リソースが terminateAfter しきい値を超過したため、AWS Data Pipeline によって停止されました。リソースがこのステータスに達すると、AWS Data Pipeline によってそのリソースの actionOnResourceFailure、retryDelay、および retryTimeout の値が無視されます。このステータスは、リソースにのみ適用されます。

VALIDATING

パイプラインの定義が AWS Data Pipeline によって検証されています。

WAITING_FOR_RUNNER

コンポーネントは、ワーカークライアントが作業項目を取得するのを待っています。コンポーネントとワーカークライアントの関係は、そのコンポーネントによって定義されている runsOn または workerGroup フィールドによって制御されます。

WAITING_ON_DEPENDENCIES

コンポーネントは、作業を実行する前に、デフォルトおよびユーザー設定の前提条件が満たされていることを確認しています。

エラーログを見つける

このセクションでは、AWS Data Pipeline が書き込む各種ログを見つける方法を説明しています。ログを使用して、特定の障害およびエラーの原因を判断することができます。

パイプラインのログ

永続的な場所にログファイルを作成するようにパイプラインを設定することをお勧めします。以下の例では、パイプラインの Default オブジェクトの pipelineLogUri フィールドを使用して、すべてのパイプラインコンポーネントがデフォルトで Amazon S3 ログの場所を使用するようにしています (特定のパイプラインコンポーネントでログの場所を設定することにより、デフォルトから変更することができます)。

> **Note**
>
> Task Runner はデフォルトで別の場所にログを保存します。このログは、パイプラインが終了し、Task Runner を実行するインスタンスが終了すると、使用できないことがあります。詳細については、「Task Runnerのログの確認」を参照してください。

パイプライン JSON ファイルで AWS Data Pipeline CLI を使用してログの場所を設定するには、パイプラインファイルを次のテキストで開始します。

```
{ "objects": [
{
  "id":"Default",
  "pipelineLogUri":"s3://mys3bucket/error_logs"
},
...
```

パイプラインログディレクトリを設定すると、Task Runner は設定されたディレクトリに、Task Runner ログに関する以前のセクションで説明されているのと同じ書式とファイル名で、ログのコピーを作成します。

Hadoop ジョブと Amazon EMR ステップログ

HadoopActivity、HiveActivity、PigActivity などの Hadoop ベースのアクティビティを使用することによって、ランタイムスロット hadoopJobLog に返された場所で Hadoop ジョブログを確認できます。EmrActivity には独自のログ記録機能があり、これらのログは、Amazon EMR によって選択されランタイムスロット emrStepLog によって返された場所を使用して保存されます。詳細については、Amazon EMR 開発者ガイドのログファイルを表示するを参照してください。

一般的な問題を解決する

このトピックでは、AWS Data Pipeline で発生する問題のさまざまな症状と、それを解決するお勧めの手順を示します。

目次

- ステータスが保留中のままパイプラインが先に進まない
- Runner のステータスを待機してパイプラインコンポーネントが先に進まない
- WAITING_ON_DEPENDENCIES ステータスでパイプラインコンポーネントが先に進まない
- 予定した時期に実行が開始されない
- パイプラインコンポーネントが間違った順番で実行される
- EMR クラスターが「リクエストに含まれているセキュリティトークンが無効です」というエラーで失敗する
- リソースにアクセスするアクセス許可が不十分である
- ステータスコード: 400 エラーコード: PipelineNotFoundException
- パイプラインを作成するとセキュリティトークンエラーが発生する
- コンソールでパイプラインの詳細を表示できない
- リモートランナーのエラー - ステータスコード: 404, AWS サービス: Amazon S3
- アクセスが拒否される - datapipeline 関数を実行する権限がない
- 古い Amazon EMR AMI では、大きい CSV ファイルに対して間違ったデータが作成される可能性がある
- AWS Data Pipeline の上限を引き上げる

ステータスが保留中のままパイプラインが先に進まない

PENDING ステータスのまま停止していると思われるパイプラインは、パイプラインがまだアクティブ化されていないか、パイプライン定義のエラーのためにアクティブ化に失敗したことを示しています。AWS Data Pipeline CLI を使用してパイプラインを送信したとき、または AWS Data Pipeline コンソールを使用してパイプラインを保存またはアクティブ化しようとしたときに、エラーが発生しなかったことを確認します。さらに、パイプライン定義が有効であることも確認します。

CLI を使用して画面にパイプライン定義を表示するには、次のように入力します。

```
aws datapipeline --get-pipeline-definition --pipeline-id df-EXAMPLE_PIPELINE_ID
```

パイプライン定義が完全であることを確認します。閉じる中括弧があること、必要なカンマがあること、不明な参照がないこと、その他の構文エラーがないことを確認してください。JSON ファイルの構文を視覚的に検証できるテキストエディタを使用するのが最善です。

Runner のステータスを待機してパイプラインコンポーネントが先に進まない

SCHEDULED 状態のパイプラインに、WAITING_FOR_RUNNER 状態で停止しているように思われるタスクがある場合は、そのタスクの runsOn または workerGroup フィールドに有効な値が設定されていることを確認します。両方の値が空か、指定されていない場合、タスクと、タスクを実行するワーカーとの間に関連付けがないため、タスクを開始できません。これは、作業を定義したが、その作業を実行するコンピュータを定義していないという状況です。該当する場合は、パイプラインコンポーネントに割り当てた workerGroup 値が、Task Runner 用に設定した workerGroup 値と、大文字小文字も含めて厳密に同じ名前であることを確認します。

> **Note**
>
> runsOn 値を指定して、workerGroup がある場合、workerGroup は無視されます。

この問題のもう一つ考えられる原因は、Task Runner に渡されたエンドポイントおよびアクセスキーが、AWS Data Pipeline コンソール、または AWS Data Pipeline CLI ツールがインストールされたコンピューターと同じでないことです。エラーなしで新しいパイプラインが作成されたように見えても、認証情報の違いにより、Task Runner が間違った場所をポーリングしているか、正しい場所をポーリングしていても、アクセス許可が不十分で、パイプライン定義によって指定された作業を特定して実行することができません。

WAITING_ON_DEPENDENCIES ステータスでパイプラインコンポーネントが先に進まない

パイプラインが SCHEDULED 状態で、1 つ以上のタスクが WAITING_ON_DEPENDENCIES 状態で停止しているような場合は、パイプラインの初期前提条件が満たされていることを確認します。ロジックチェーンの最初のオブジェクトの前提条件が満たされていない場合、その最初のオブジェクトに依存するオブジェクトはどれも WAITING_ON_DEPENDENCIES 状態を抜けることができません。

たとえば、以下のようなパイプライン定義があるとします。この場合、InputData オブジェクトには前提条件として 'Ready' があり、InputData オブジェクトが完了する前にデータが存在する必要があります。データが存在しない場合、InputData オブジェクトは、WAITING_ON_DEPENDENCIES 状

態のまま、パスフィールドで指定したデータが使用可能になるまで待ちます。同様に、InputData に依存するすべてのオブジェクトが WAITING_ON_DEPENDENCIES 状態のまま、InputData オブジェクトが FINISHED 状態になるまで待ちます。

```
{
    "id": "InputData",
    "type": "S3DataNode",
    "filePath": "s3://elasticmapreduce/samples/wordcount/wordSplitter.py",
    "schedule":{"ref":"MySchedule"},
    "precondition": "Ready"
},
{
    "id": "Ready",
    "type": "Exists"
...
```

また、オブジェクトにデータにアクセスする適切なアクセス許可があることを確認します。前の例で、パスフィールドに指定されたデータへのアクセス許可が認証情報フィールドに含まれていない場合、InputData オブジェクトはパスフィールドに指定されたデータにアクセスできないため、たとえデータが存在していても、WAITING_ON_DEPENDENCIES 状態のまま先に進みません。

また、Amazon S3 と通信するリソースに、パブリック IP アドレスが関連付けられていない可能性もあります。たとえば、パブリックサブネット内の Ec2Resource には、パブリック IP アドレスが関連付けられている必要があります。

最後に、状況によっては、関連アクティビティの開始予定よりもはるか前に、リソースインスタンスが WAITING_ON_DEPENDENCIES 状態になる場合があります。この場合、リソースやアクティビティが失敗しているように見えることがあります。

予定した時期に実行が開始されない

タスクがスケジュール間隔の開始時に開始されるか（Cron スタイルのスケジュールタイプ）、または、スケジュール間隔の終了時に開始されるか（時系列のスケジュールタイプ）を決定する正しいスケジュールタイプを選択していることを確認します。

さらに、スケジュールオブジェクトで適切な日付を指定していること、および startDateTime と endDateTime の値が以下の例のように UTC 形式であることを確認します。

```
{
    "id": "MySchedule",
    "startDateTime": "2012-11-12T19:30:00",
```

```
    "endDateTime":"2012-11-12T20:30:00",
    "period": "1 Hour",
    "type": "Schedule"
},
```

パイプラインコンポーネントが間違った順番で実行される

パイプラインコンポーネントの開始時間と終了時間を見ると、間違った順序で実行されているか、または、想定とは異なる順序になっている場合があります。起動時に前提条件が満たされていれば、パイプラインコンポーネントは同時に実行を開始できることを理解することが重要です。つまり、パイプラインコンポーネントはデフォルトでは順次実行されません。特定の順序で実行する場合は、前提条件と dependsOn フィールドを使用して実行順序を制御する必要があります。

dependsOn フィールドに正しい必須パイプラインコンポーネントへの参照が設定済みであること、および、目的の順序を実現するために必要なコンポーネント間のポインタがすべて存在することを確認します。

EMR クラスターが「リクエストに含まれているセキュリティトークンが無効です」というエラーで失敗する

IAM ロール、ポリシー、信頼関係が、AWS Data Pipeline の IAM ロールで説明されているとおりになっているか確認します。

リソースにアクセスするアクセス許可が不十分である

IAM ロールに設定したアクセス許可によって、AWS Data Pipeline が EMR クラスターおよび EC2 インスタンスにアクセスしてパイプラインを実行できるかどうかが決まります。さらに、IAM には信頼関係の概念があり、お客様に代わってリソースを作成することも許可されます。たとえば、EC2 インスタンスを使用するパイプラインを作成して、データを移動するコマンドを実行するときに、AWS Data Pipeline はこの EC2 インスタンスをお客様の代わりにプロビジョニングすることができます。問題が発生した場合、特に、手動ではアクセスできるのに、AWS Data Pipeline ではアクセスできないリソースが関連する場合は、IAM ロール、ポリシー、信頼関係がAWS Data Pipeline の IAM ロールで説明されているとおりになっているか確認します。

ステータスコード: 400 エラーコード: PipelineNotFoundException

このエラーは、IAM のデフォルトロールに必要なアクセス許可がなく、AWS Data Pipeline が適切に機能しないことを意味します。詳細については、「AWS Data Pipeline の IAM ロール」を参照してください。

パイプラインを作成するとセキュリティトークンエラーが発生する

パイプラインを作成しようとすると、次のエラーが発生します。

Failed to create pipeline with 'pipeline_name'。Error: UnrecognizedClientException - The security token included in the request is invalid。

コンソールでパイプラインの詳細を表示できない

AWS Data Pipeline コンソールのパイプラインフィルタは、パイプラインが送信された時期に関係なく、パイプラインの予定された開始日に対して適用されます。予定された開始日を過去の日付にして新しいパイプラインを送信することは可能です。この場合、デフォルトの日付フィルタでは表示されないことがあります。パイプラインの詳細を表示するには、日付フィルターを変更して、予定されたパイプライン開始日が日付フィルタの範囲におさまるようにします。

リモートランナーのエラー - ステータスコード: 404, AWS サービス: Amazon S3

このエラーは、Task Runner が Amazon S3 ファイルにアクセスできないことを意味します。以下を確認してください。

- 認証情報を正しく設定していること
- アクセスしようとしている Amazon S3 バケットが存在すること
- Amazon S3 バケットにアクセスする権限がお客様にあること

アクセスが拒否される - datapipeline 関数を実行する権限がない

Task Runner ログに、次のようなエラーが記録されることがあります。

- ERROR Status Code: 403
- AWS Service: DataPipeline
- AWS Error Code: AccessDenied
- AWS Error Message: User: arn:aws:sts::XXXXXXXXXXXX:federated-user/i-XXXXXXXX is not authorized to perform: datapipeline:PollForTask。

> **Note**
>
> このエラーメッセージで、PollForTask は、他の AWS Data Pipeline アクセス権限と置き換わっている場合があります。

このエラーメッセージは、指定した IAM ロールが AWS Data Pipeline とやり取りするには、追加のアクセス許可が必要であることを示します。IAM ロールポリシーに、以下の行が含まれていることを確認してください。PollForTask は、追加するアクセス許可の名前に置き換えます (すべてのアクセス許可を付与する場合は、* を使用します)。新しい IAM ロールを作成し、それにポリシーを適用する方法については、IAM の使用のIAM ポリシーを管理するを参照してください。

```
{
"Action": [ "datapipeline:PollForTask" ],
"Effect": "Allow",
"Resource": ["*"]
}
```

古い Amazon EMR AMI では、大きい CSV ファイルに対して間違ったデータが作成される可能性がある

3.9 より前 (3.8 以前) の Amazon EMR AMI では、AWS Data Pipeline はカスタム InputFormat を使用して、MapReduce ジョブで使用する CSV ファイルを読み書きします。これは、サービスがテーブルを Amazon S3 との間でステージングするときに使用されます。この InputFormat の問題は、大きな CSV ファイルからレコードを読み取ると、正しくコピーされていないテーブルが作成される場合があることから発見されました。この問題は将来の Amazon EMR のリリースで修正されます。Amazon EMR AMI 3.9 または Amazon EMR リリース 4.0.0 以降を使用してください。

AWS Data Pipeline の上限を引き上げる

ときどき、AWS Data Pipeline システムの特定の上限を超えることがあります。たとえば、パイプラインのデフォルトの上限は、パイプライン数が 20 個、オブジェクト数がそれぞれ 50 個です。上限を超えるパイプラインが必要になった場合は、複数のパイプラインを統合してパイプラインの数を減らし、それぞれのオブジェクト数を増やすことを検討してください。AWS Data Pipeline の制限の詳細については、「AWS Data Pipeline の制限」を参照してください。ただし、パイプラインの統合という技を使っても制限を回避できない場合は、このフォームを使用して、容量の増加を依頼します: Data Pipeline の上限の引き上げ

AWS Data Pipeline の制限

AWS Data Pipeline は、リソースの割り当て量と割り当てレートを制限することで、すべてのユーザーが容量を利用できるようにします。

目次

- アカウントの制限
- ウェブサービス呼び出しの制限
- スケーリングに関する考慮事項

アカウントの制限

AWS アカウントごとに以下の制限が適用されます。追加の容量が必要な場合は、Amazon Web Services サポートセンターの申請フォームを使用して容量を増やすことができます。

属性	制限	引き上げ可能
パイプラインの数	100	Yes
パイプラインあたりのオブジェクトの数	100	Yes
オブジェクトあたりのアクティブなインスタンスの数	5	はい
オブジェクトあたりのフィールドの数	50	No
フィールド名前または ID あたりの UTF8 バイトの数	256	No
フィールドあたりの UTF8 バイトの数	10,240	No

属性	制限	引き上げ可能
オブジェクトあたりの UTF8 バイトの数	15,360 (フィールド名を含む)	No
オブジェクトからのインスタンス作成レート	5 分に 1 回	No
パイプラインアクティビティの再試行	タスクにつき 5 回	No
再試行間の最小遅延間隔	2 分	No
最小スケジュール間隔	15 分	No
単一のオブジェクトへのロールアップの最大数	32	No
Ec2Resource オブジェクトあたりの EC2 インスタンスの最大数	1	No

ウェブサービス呼び出しの制限

AWS Data Pipeline は、お客様がウェブサービス API を呼び出すことのできるレートを制限しています。これらの制限は、お客様に代わってウェブサービス API を呼び出すコンソール、CLI、Task Runner などの AWS Data Pipeline エージェントにも適用されます。

AWS アカウントごとに以下の制限が適用されます。これは、 ユーザーによる使用も含むアカウント全体の使用量がこれらの制限を超えてはならないことを意味します。

バーストレートを使用すると、アイドル状態の期間においてウェブサービス呼び出しを節約し、短期間にそれらすべてを消費できます。たとえば、CreatePipeline の通常レートは 5 秒に 1 回の呼び

出しです。サービスを 30 秒間呼び出さなければ、6 回の呼び出しが節約されます。その後、ウェブサービスを 1 秒間に 6 回呼び出すことができます。これはバースト制限を超えず、呼び出しの平均間隔が通常のレートの制限内であるため、呼び出しはスロットリングされません。

レートの制限を超えると、ウェブサービス呼び出しは失敗し、スロットリング例外が返されます。Task Runner (ワーカー) のデフォルト実装では、スロットリング例外で失敗した API 呼び出しを自動的に再試行します。Task Runner にはバックオフがあるため、以降の API 呼び出しの試行間隔は徐々に長くなります。ワーカーを記述する場合は、同じような再試行ロジックを実装することをお勧めします。

これらの制限は、個々の AWS アカウントに対して適用されます。

API	通常のレートの制限	バースト制限
ActivatePipeline	1 秒につき呼び出し 1 回	呼び出し 100 回
CreatePipeline	1 秒につき呼び出し 1 回	呼び出し 100 回
DeletePipeline	1 秒につき呼び出し 1 回	呼び出し 100 回
DescribeObjects	1 秒につき呼び出し 2 回	呼び出し 100 回
DescribePipelines	1 秒につき呼び出し 1 回	呼び出し 100 回
GetPipelineDefinition	1 秒につき呼び出し 1 回	呼び出し 100 回
PollForTask	1 秒につき呼び出し 2 回	呼び出し 100 回
ListPipelines	1 秒につき呼び出し 1 回	呼び出し 100 回
PutPipelineDefinition	1 秒につき呼び出し 1 回	呼び出し 100 回
QueryObjects	1 秒につき呼び出し 2 回	呼び出し 100 回
ReportTaskProgress	1 秒につき呼び出し 10 回	呼び出し 100 回
SetTaskStatus	1 秒につき呼び出し 10 回	呼び出し 100 回
SetStatus	1 秒につき呼び出し 1 回	呼び出し 100 回

API	通常のレートの制限	バースト制限
ReportTaskRunnerHeartbeat	1 秒につき呼び出し 1 回	呼び出し 100 回
ValidatePipelineDefinition	1 秒につき呼び出し 1 回	呼び出し 100 回

スケーリングに関する考慮事項

AWS Data Pipeline は、多くの同時実行タスクに対応するように拡張でき、大量の作業負荷を処理するために必要なリソースを自動的に作成するよう設定できます。これらの自動的に作成されたリソースは、お客様の管理下にあり、AWS アカウントのリソースに対する制限の対象となります。例えば、AWS アカウントの EC2 インスタンスの制限が 20 個に設定されている場合に、20 個のノードで構成される Amazon EMR クラスターを自動的に作成してデータを処理するように AWS Data Pipeline を設定すると、使用可能なバックフィルリソースを意図せずに使い切ってしまう可能性があります。そのため、設計時にこれらのリソースの制限を考慮するか、またはアカウントの制限を適宜増やします。

追加の容量が必要な場合は、Amazon Web Services サポートセンターの申請フォームを使用して容量を増やすことができます。

AWS Data Pipeline リソース

AWS Data Pipeline の使用に役立つリソースを次に示します。

- AWS Data Pipeline 製品情報 – AWS Data Pipeline に関する情報のメインウェブページです。
- AWS Data Pipeline テクニカル FAQ – この製品について開発者からよく寄せられる上位 20 の質問です。
- リリースノート – 現在のリリースの概要を提供します。新機能、解決された問題、既知の問題が具体的に記載されています。
- AWS Data Pipeline ディスカッションフォーラム – Amazon Web Services に関する技術的な質疑応答の場である開発者向けのコミュニティベースのフォーラムです。

- クラスとワークショップ – AWS のスキルを磨き、実践的な経験が得るために役立つセルフペースラボに加えて、ロールベースのコースと特別コースへのリンクです。
- AWS デベロッパーセンター – チュートリアルの検索、ツールのダウンロード、AWS デベロッパーイベントの確認を行います。
- AWS デベロッパーツール - AWS アプリケーションを開発および管理するためのデベロッパーツール、SDK、IDE ツールキット、およびコマンドラインツールへのリンクです。
- ご利用開始のためのリソースセンター — AWS アカウント をセットアップする方法、AWS コミュニティに参加する方法、最初のアプリケーションを起動する方法を説明します。
- ハンズオンチュートリアル - ステップ バイ ステップのチュートリアルに従って、最初のアプリケーションを AWS で起動します。
- AWS ホワイトペーパー – アーキテクチャ、セキュリティ、エコノミクスなどのトピックについて、AWS のソリューションアーキテクトや他の技術エキスパートが記述した AWS の技術ホワイトペーパーの包括的なリストへのリンクです。
- AWS Support Center – AWS Support のケースを作成して管理するためのハブです。フォーラム、技術上のよくある質問、サービスヘルスステータス、AWS Trusted Advisor など、他の役立つリソースへのリンクも含まれています。
- AWS Support – AWS Support に関する情報のメインウェブページです。クラウド内でのアプリケーションの構築および実行を支援するために 1 対 1 での迅速な対応を行うサポートチャネルとして機能します。
- お問い合わせ - AWS の請求、アカウント、イベント、不正使用、その他の問題などに関するお問い合わせの受付窓口です。

- AWS サイトの利用規約 – 当社の著作権、商標、お客様のアカウント、ライセンス、サイトへのアクセス、その他のトピックに関する詳細情報。

ドキュメント履歴

このドキュメントは、 の 2012-10-29 バージョンに関連付けられています AWS Data Pipeline。

変更	説明	リリース日
AWS Data Pipeline は新規顧客には利用できなくなりました	AWS Data Pipeline は、新規顧客には利用できなくなりました。の既存のお客様は、通常どおりサービスを AWS Data Pipeline 引き続き使用できます。詳細はこちら	2025 年 7 月 25 日
を使用して特定の手順を実行するためのドキュメントを追加しました AWS CLI。 AWS Data Pipeline コンソール関連の手順を削除しました。	詳細については、パイプラインのクローン作成、パイプラインのログの表示、およびCLI を使用して Data Pipeline テンプレートからパイプラインを作成するを参照してください。	2023年5月26日
から他の代替サービスに移行 AWS Data Pipeline するためのコンテンツとサンプルを追加しました。	AWS Glue、 AWS Step Functions、または Amazon のいずれか AWS Data Pipeline に移行するためのトピックを更新MWAAし、各代替、サービス間の概念マッピング、サンプルに関する詳細情報を追加しました。詳細については、「からのワークロードの移行 AWS Data Pipeline」を参照してください。	2023年3月31日
AWS Data Pipeline のサポートに関する情報を追加しましたIMDSv2。	AWS Data Pipeline は、Amazon EMR および Amazon EC2リソースIMDSv2の をサポートします。詳細については、AWS Data Pipeline でのデータ保護、EmrCluster、およびEc2Resourceを参照してください。	2022年12月16日
から他の代替サービス AWS Data Pipeline への移行に	現在、お客様により良いデータ統合エクスペリエンスを提供する他の AWS サービスがあります。の一般的なユースケース AWS Data Pipeline は、、 AWS Step Functions AWS Glue、または Amazon に移行できます	2022年12月16日

変更	説明	リリース日
関するトピックを追加しました。	MWAA。詳細については、「からのワークロードの移行 AWS Data Pipeline」を参照してください。	
サポートされているAmazon EC2およびAmazon EMRインスタンスのリストを更新しました。 インスタンスAMIsに使用される HVM (ハードウェア仮想マシン) IDsのリストを更新しました。	サポートされている Amazon EC2および Amazon EMRインスタンスのリストを更新しました。詳細については、「パイプラインの作業アクティビティ用にサポートされているインスタンスタイプ」を参照してください。 インスタンスAMIsに使用される HVM (ハードウェア仮想マシン) IDsのリストを更新しました。詳細については、「構文」を参照し、`imageId`を検索してください。	2018年11月9日

変更	説明	リリース日
Amazon EBSボリュームをクラスターノードにアタッチし、Amazon EMRクラスターをプライベートサブネットに起動するための設定を追加しました。	`EMRcluster` オブジェクトに設定オプションを追加しました。これらのオプションは、Amazon EMRクラスターを使用するパイプラインで使用できます。 `coreEbsConfiguration` 、、および `TaskEbsConfiguration` フィールドを使用して`masterEbsConfiguration` 、Amazon EMRクラスターのコアノード、マスターノード、およびタスクノードへの Amazon EBSボリュームのアタッチメントを設定します。詳細については、「クラスターノードにEBSボリュームをアタッチする」を参照してください。 `emrManagedMasterSecurityGroupId` 、`emrManagedSlaveSecurityGroupId` 、および `ServiceAccessSecurityGroupId` フィールドを使用して、プライベートサブネットにAmazon EMRクラスターを設定します。詳細については、「プライベートサブネットで Amazon EMRクラスターを設定する」を参照してください。 `EMRcluster` 構文の詳細については、「EmrCluster」を参照してください。	2018年4月19日
サポートされているAmazon EC2およびAmazon EMRインスタンスのリストを追加しました。	パイプライン定義でインスタンスタイプを指定しない場合に、 がデフォルトで AWS Data Pipeline 作成するインスタンスのリストを追加しました。サポートされている Amazon EC2および Amazon EMRインスタンスのリストを追加しました。詳細については、「パイプラインの作業アクティビティ用にサポートされているインスタンスタイプ」を参照してください。	2018年3月22日
オンデマンドパイプラインのサポートを追加しました。	• パイプラインを再アクティブ化することで再実行できるオンデマンドパイプラインのサポートが追加されました。	2016年2月22日

変更	説明	リリース日
RDS データベースの追加サポート	• rdsInstanceId 、region、およびjdbcDriverJarUri がRdsDatabaseに追加されました。 • RdsDatabase もサポートするためにSqlActivity内のdatabaseが更新されました。	2015年8月17日
その他のJDBCサポート	• JdbcDatabase もサポートするためにSqlActivity内のdatabaseが更新されました。 • jdbcDriverJarUri がJdbcDatabaseに追加されました。 • initTimeout がEc2ResourceとEmrClusterに追加されました。 • runAsUser がEc2Resourceに追加されました。	2015年7月7日
HadoopActivity、アベイラビリティーゾーン、スポットサポート	• Hadoopクラスターに並列作業を送信できるようになりました。詳細については、「HadoopActivity」を参照してください。 • Ec2ResourceとEmrClusterで使用するスポットインスタンスをリクエストする機能が追加されました。 • 指定したアベイラビリティーゾーンでEmrCluster リソースを起動する機能を追加しました。	2015年6月1日
パイプラインの非アクティブ化	アクティブなパイプラインの非アクティブ化のサポートが追加されました。詳細については、「パイプラインの非アクティブ化」を参照してください。	2015年4月7日
更新されたテンプレートとコンソール	新しいテンプレートを追加しました。テンプレートの開始方法を使用するように開始方法 ShellCommandActivityの章を更新しました。詳細については、「CLI を使用して Data Pipeline テンプレートからパイプラインを作成する」を参照してください。	2014年11月25日

変更	説明	リリース日
VPC サポート	仮想プライベートクラウド () でリソースを起動するためのサポートを追加しましたVPC。	2014年3月12日
リージョンのサポート	複数サービスリージョンのサポートを追加しました。に加えてus-east-1 、 AWS Data Pipeline は eu-west-1 、、ap-northeast-1 、ap-southeast-2 および でサポートされていますus-west-2 。	2014年2月20日
Amazon Redshiftサポート	で Amazon Redshift のサポートが追加されました。これには AWS Data Pipeline、新しいコンソールテンプレート (Redshift にコピー) と、テンプレートをデモンストレーションするためのチュートリアルが含まれます。詳細については、「AWS Data Pipeline を使用した Amazon Redshift へのデータのコピー」、「RedshiftDataNode」、「RedshiftDatabase」、および「RedshiftCopyActivity」を参照してください。	2013年11月6日
PigActivity	Pig のネイティブサポート PigActivityを提供する を追加しました。詳細については、「PigActivity」を参照してください。	2013年10月15日
新しいコンソールテンプレート、アクティビティ、およびデータ形式	新しい HiveCopyActivity と D を含む新しい CrossRegion DynamoDB Copy コンソールテンプレートを追加しましたynamoDBExportDataFormat。	2013年8月21日
カスケードの失敗と再実行	AWS Data Pipeline カスケード障害と再実行動作に関する情報を追加しました。詳細については、「カスケードの失敗と再実行」を参照してください。	2013年8月8日
トラブルシューティングの動画	AWS Data Pipeline 基本的なトラブルシューティングビデオを追加しました。詳細については、「トラブルシューティング」を参照してください。	2013年7月17日

変更	説明	リリース日
アクティブなパイプラインの編集	アクティブなパイプラインの編集およびパイプラインコンポーネントの再実行に関する詳細情報が追加されました。詳細については、「パイプラインの編集」を参照してください。	2013年7月17日
異なるリージョンでのリソースの使用	異なるリージョン内のリソースの使用に関する詳細情報を追加しました。詳細については、「複数のリージョンにあるリソースとパイプラインの使用」を参照してください。	2013年6月17日
WAITING_ON_DEPENDENCIES ステータス	CHECKING_PRECONDITIONS ステータスが WAITING_ON_DEPENDENCIES に変更され、パイプラインオブジェクトの @waitingOn runtime フィールドが追加されました。	2013年5月20日
D ynamoDBData形式	D ynamoDBData形式テンプレートを追加しました。	2013年4月23日
ウェブログの処理に関する動画とスポットインスタンスサポート	「Process Web Logs with AWS Data Pipeline, Amazon EMR, and Hive」ビデオと Amazon EC2スポットインスタンスのサポートを紹介しました。	2013年2月21日
	AWS Data Pipeline デベロッパーガイドの初回リリース。	2012年12月20日