AWS ホワイトペーパー

人工知能、機械学習、および生成 AI の AWS クラウド 導入フレームワーク

人工知能、機械学習、および生成 AI の AWS クラウド 導入フレームワーク: AWS ホワイトペーパー

Table of Contents

人工知能、機械学習、および生成 AI の AWS クラウド 導入フレームワーク

クラウドを活用した AI トランスフォーメーションを加速

出版日:2024 年 2 月 13 日 (ドキュメント履歴)

このドキュメントでは、人工知能 (AI)、機械学習 (ML)、および生成 AI の AWS クラウド 導入フレームワークについて説明します。これは、AI からビジネス価値を生み出すことを目指す組織のメンタルモデルを記述したフレームワークです。このフレームワークでは、AI と ML に関する組織の能力が成熟するにつれて、お客様がたどる AI ジャーニーについて説明します。組織が AI に成熟するのに役立つ基礎的能力を掘り下げることで、このジャーニーを説明します。最後に、これらの基礎的能力の目標状態の概要を示し、それらを段階的に進化させながらその過程でビジネス価値を生み出す方法を述べ、規範的ガイダンスを提供します。

AI の概要

人工知能 (AI) は、従来はヒューマンインテリジェンスを必要とするタスクを実行可能なインテリジェントマシンを作成、または少なくとも模倣することを目的とした幅広い分野です。これらのタスクには、自然言語の理解、視覚的な認識、意思決定や問題解決にいたるまで、あらゆるものが含まれます。多くの AI システム間の共通点の 1 つは、確率的成果の追求です。つまり、人間の判断の複雑さを反映して、高い確実性で予測や決定を生成することです。その後、このようなシステムを使用して、知的作業を自動化または強化できます。

現在、ほとんどの AI は機械学習 (ML) 上に構築されています。ML は、コンピュータがデータに基づいて学習し、意思決定を行う手法の開発に焦点を当てた AI の分野です。機械学習モデルは、明示的なプログラミングに頼るのではなく、例から一般化し、無数のアプリケーションに対して高い汎用性を実現します。機械学習内のさまざまな手法の 1 つに深層学習があります。深層学習は、複数のレイヤーを持つニューラルネットワークを使用して、データの複雑な要因を分析する特殊なサブセットです。深層学習は、画像やテキストなどの非構造化データの処理に特に優れており、画像や音声認識など、多くの複雑なタスクで問題を解決しています。

深層学習における新たな機能は生成 AI です。これにより、AI は新しく、潜在的にオリジナルなコンテンツを生成または作成できます。この革新的な分野は、人間のような考え方や推論機能の側面を模倣した出力を生成する能力として認識され始めています。コンピューティング能力、データ可用性、アルゴリズムイノベーションの進歩により、生成 AI が可能になり、スポーツやアートから科学研究まで、幅広いアプリケーションへの道が開かれました。

これらの分野と手法は、人工知能の階層化され相互に関連する様を表し、それぞれが、ますます幅広いタスク群を自律的に実行するシステムの開発に寄与します。AI のアプリケーションと能力は急速に拡大し続ける可能性が高いため、日々の活動に不可欠なものであり、複雑な問題を解決するための重要なツールでもあります。

「生成 AI は、数少ないイノベーションが提供する形で、人々の想像を駆り立てています。研究者や開発者の領域を超えて進化するにつれて、コンシューマーエクスペリエンスの向上から複雑な企業課題の解決まで、幅広いアプリケーションを持つことが証明されています。人間のようなテキストを生成したり、AI 主導のコードスニペットでコード作成者を支援したり、インテリジェントなチャットボットによる顧客インタラクションを自動化したりする場合でも、可能性は無限にあるかのように見えます。これらのアプリケーション以外にも、生成 AI は、スケーラビリティ、カスタマイズ、インテリジェンスをバランスよく備え、テクノロジーが人間の能力を高め、実現可能性を高める方法を再検討するのに役立ちます。大規模導入を前にして、このテクノロジーがもたらすとされていることは、タスクをより効率的に達成するだけでなく、さまざまな分野において何ができるかを根本的に再定義することででもあります。」 – Andy Jassy、Amazon CEO

Note

さらに、人工知能 (AI) という用語は、そのさまざまなサブ分野を含む総称として使用されます。AI の専門分野のいくつかは、生成 AI や機械学習など、一般的な AI とは別の呼称で呼ばれます。

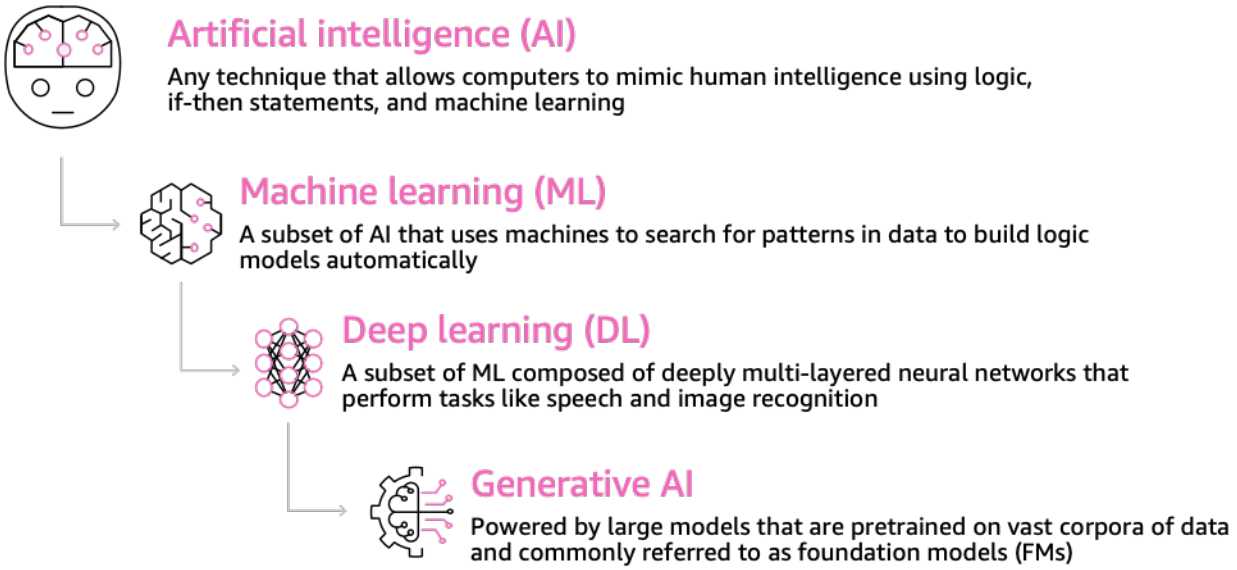

図 1: 人工知能、機械学習、深層学習、生成 AI の分類を示す図

AWS CAF-AI の概要

人工知能、機械学習、および生成 AI の AWS クラウド導入フレームワーク (CAF-AI) は、機械学習と AI ジャーニーの出発点であり、ガイドでもあります。このフレームワークは、これらの専門分野における中間計画と戦略を策定し、情報を提供することを目的としています。これは、チーム内だけでなく、同僚や AWS パートナーと協力して AI 戦略について議論するためのリソースとして使用することができます。

ジャーニーのどの段階にいるかにもよりますが、特定のセクションに集中して、そこでスキルを磨くことも、ドキュメント全体を使って成熟度を判断し、短期的な改善点を導くのに役立てることもできます。CAF-AI は、企業レベルで AI を採用する際に考慮すべきすべての事項を要約し、継続的に成長および更新し続けており、単一の概念実証 (POC) の先に進むのに役立ちます。目標は、AI の導入に成功するように、AWS クラウド導入フレームワーク (AWS CAF) についてお客様が理解し期待しているのと同じ規範的ガイダンスを提供することです。CAF は、一連の基本的な組織能力に支えられ、世界中の数千の組織がクラウドトランスフォーメーションのジャーニーを加速するためにうまく活用してきた規範的ガイダンスを提供します。

AWS CAF-AI でも、これらの基礎的能力に依存していますが、AI が求める変化に対応できるよう、多くの機能を強化しています。さらに、組織が AI ジャーニーの一部として考慮すべき新しい基礎的能力を特定して追加します。

AWS CAF: クラウド導入フレームワーク

AWS では、過去 10 年間にわたり、お客様のクラウド導入戦略の基礎として、AWS クラウド導入フレームワーク (AWS CAF) を構築してきました。このフレームワークを進化させる一方で、そのインサイトとメンタルモデルがさまざまなお客様の多くに使用できるように、クラウド以外の特定のテクノロジーにほとんど縛られないようにしました。ただし、人工知能はまったく新しい種類のテクノロジーであり、すべての業種とほとんどのお客様に大きな影響を与えています。クラウドテクノロジーを通じて加速された AI の導入のジャーニーにあるお客様を支援するために、CAF-AI を構築しました。

Well-Architected の実現状況の確認

AWS Well-Architected フレームワークは、クラウド内でのシステム構築に伴う意思決定の長所と短所を理解するのに役立ちます。このフレームワークの 6 つの柱により、信頼性、安全性、効率、費用対効果、持続可能性の高いシステムを設計および運用するための、アーキテクチャのベストプラクティスを確認できます。AWS Management Console で無料で提供されている AWS Well-Architected Tool を使用すると、柱ごとに一連の質問に答えることで、これらのベストプラクティスに照らしてワークロードを評価できます。

Machine Learning Lens では、AWS クラウド で機械学習ワークロードを設計、デプロイ、構築する方法に焦点を当てています。このレンズは、Well-Architected Framework で説明されているベストプラクティスを発展させます。

クラウドアーキテクチャに関する専門的なガイダンスやベストプラクティス (リファレンスアーキテクチャのデプロイ、図、ホワイトペーパー) については、AWS アーキテクチャセンターを参照してください。

人工知能クラウドトランスフォーメーションのバリューチェーン

人工知能は、特殊なテクノロジーから強力で広く利用可能なビジネス機能に進化しました。機械学習は、今やイノベーションの二ューウェーブを促進しています。つまり、データは発明の起源であり、ML は組織にとって過去を説明するだけでなく、未来を予測し、有意義な行動を規定するためのまったく新しい機能でもあります。この機能がすべての市場とビジネスに与える影響から、あらゆる業界の組織が AI への投資を増やしています。この投資は、顧客に関するインサイトの向上、従業員の効率の向上、イノベーションの加速を通じて競争上の優位性を生み出すことができます。これは、縦型と横型の両方のユースケースにまたがる広大な問題領域に AI を適用できることが原動力となっています。

特に、AI を適用できるビジネス上の問題領域は、単一の機能や領域ではなく、AI が経済的に大きな違いをもたらす市場での競争条件を再設定する機会を持つ、ビジネスのあらゆる機能やあらゆる業界領域にわたって大きな可能性を秘めているということです。AI は、何十年もの間経済的に解決できなかった問題や、単に技術的に AI なしでは取り組むことが不可能だった問題に対するソリューションやソリューションへの道筋が得られるため、結果としてもたらされるビジネス上の成果は計り知れません。

一例として、追加データをほとんど加えずにドメイン固有の機能を実行する大規模な AI モデルが新たに登場したことで、組織に大きな変化を引き起こし、ビジネスの差別化に役立っています。これらは主に生成 AI の分野であり、幅広い関心を集め、人々の想像を駆り立てています。ただし、このようなモデルの開発、適用、調整は複雑になる可能性があります。

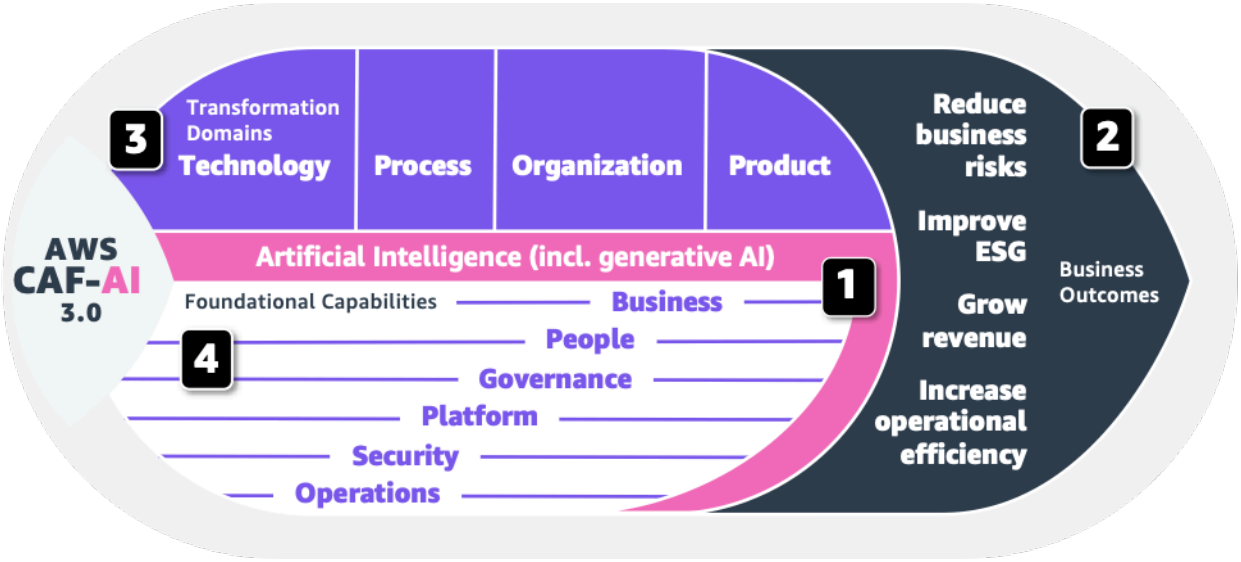

図 2:AWS CAF-AI トランスフォーメーションのバリューチェーン (ピンクとマゼンタの部分はすべて、ここで構築した元の CAF の分野です)。

上の図は、変化する市場環境と急速に加速する分野に直面する中で、AI の導入についてどのように考えるべきかという方向を示しています。

1. AI は組織に新しい能力を提供します。
2. これらの新しい能力により、ユーザーとユーザーの組織は具体的なビジネス成果を生み出すよう努めます。これらのビジネス成果には、ビジネスリスクの軽減 (生産チェーン内の壊れた部品や欠陥部品の検出など)、環境、社会、ガバナンス (ESG) パフォーマンスの向上 (環境保護コンプライアンスレポートの自動要約とフラグ付けなど)、新規および既存の収益の増加 (顧客への製品およびサービスの推奨事項のパーソナライズなど)、または業務効率の向上 (出張領収書の分類と内部予約コードへのマッピングなど) などが含まれます。ただし、このようなビジネス成果を生み出すには、AI を導入する能力が鍵となります。
3. AI を導入するには、組織には少なくとも 4 つのドメインにおいてトランスフォーメーションを実行する必要があります。
    a. テクノロジー: 技術力を確立し、AI の活用と導入を可能にすることに重点を置いたドメイン。
    b. プロセス: AI の能力を通じて事業運営のデジタル化、自動化、最適化、革新に重点を置いたドメイン。
    c. 組織: ビジネスチームとテクノロジーチームが、AI を駆使して、顧客価値の創造と戦略的意図の達成に向けた取り組みをどのように調整しているか。
    d. 製品: AI の機能を活用した新しい価値提案 (製品、サービス) と収益モデルを作成することで、ビジネスモデルを再考します。
4. これらのドメインを変革し、AI を活用できるかどうかは、ビジネス、人員、ガバナンス、プラットフォーム、セキュリティ、およびオペレーションにおける基礎的な能力に依存します。

AI の導入を成功させるには、ジャーニーを次のように計画してください。

- AI によって何が可能になるかを理解したうえで、逆算して考えます。
- 長期的にどのようなビジネス成果が期待できるかを定義します。
- ビジネスがたどらなければならないトランスフォーメーションを確立します。
- このジャーニーを支える基礎的能力を開発します。

AI トランスフォーメーションジャーニー

すべての大規模な技術導入のアジェンダは、特に AI のような急速に進化している技術を導入する場合、長い道のりです。トランスフォーメーションと導入のジャーニーは組織によって大きく異なりますが、私たちは AI の導入が成功するパターンを学んできました。そこで、このカスタマージャーニーのリスクを軽減するために、AWS CAF-AI では、何千ものお客様から学習した以下の点をベストプラクティスとして提供しています。それでも、AI ジャーニーがそれぞれの組織で異なることには変わりありません。

AI トランスフォーメーションジャーニーを開始したり、進めたりする際には、次の図 3 にも示されている 4 つの重要な要素を検討してください。

1. ジャーニーの目的、つまり、達成を目指すビジネス成果から逆算して、取り組みを進めます。
2. ジャーニーの原動力としての AI フライホイール。AI フライホイールとは、初期の高品質データ (タイムリーで、関連性があり、価値があり、有効なデータ) を使用して AI システムをトレーニングまたは調整し、その後、予測を提供するという好循環のことです。このような予測はビジネスの成果にプラスの影響を与え、その結果、顧客関係の増加またはより深い顧客関係の構築につながり、作成されるデータの増加またはデータの品質向上を促します (ネットワークとフライホイール効果)。
3. データおよびデータ戦略は、AI フライホイールを推し進めます。
4. 基礎的な能力は、何よりも AI の導入の成否を左右します。

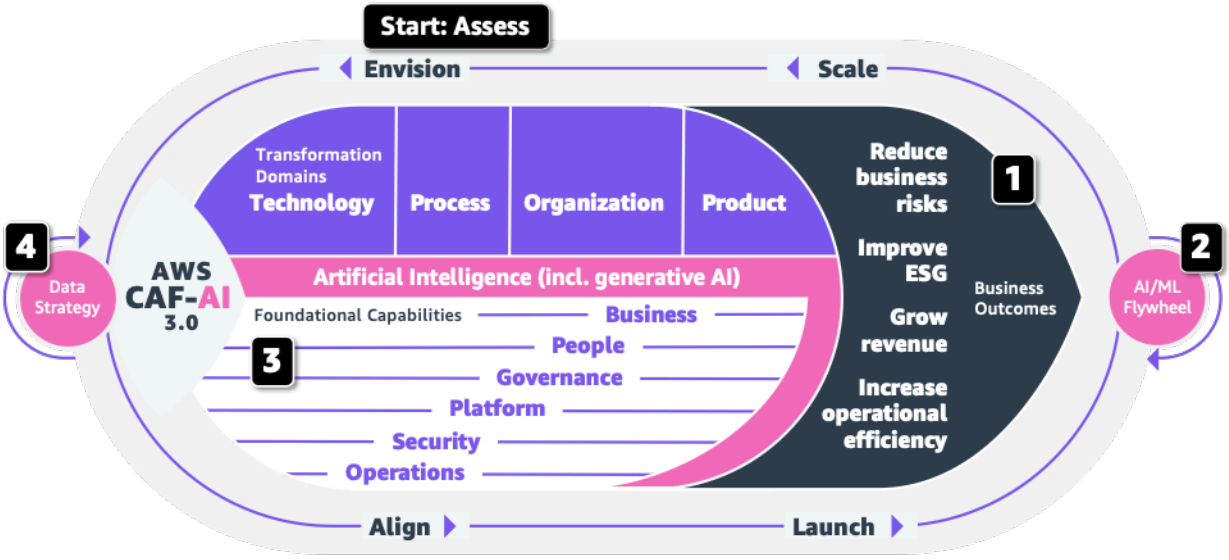

図 3: AWS CAF-AI クラウドのトランスフォーメーションジャーニー

このジャーニーへの取り組みは、反復的かつ段階的な改善を土台とします。また、AWS 担当者 (担当アカウントチームなど) に連絡して、AWS ML ストラテジスト、エンタープライズストラテジスト、ML アドバイザーから支援を受けることをお勧めします。初期評価の後、次の 4 つのステージに基づいて導入サイクルが始まります。

- 構想: この第 1 フェーズは、AI がビジネス成果を加速するのにどのように役立つかを構想することに重点を置いています。つまり、ビジネス目標に沿ってトランスフォーメーションの機会を特定し、優先順位を付けます。トランスフォーメーションのイニシアチブを、主要なステークホルダー (つまり、変化に影響を与えて推進できる上級の関係者) と測定可能なビジネス成果に関連付けます。また、この初期段階で、これらのイニシアチブや機会がどのデータ資産やソースに依存しているかを必ず特定してください。機会を起点にしてデータ要件を考えます。
- 調整: この第 2 フェーズでは、基礎的能力に焦点を当てます。組織間の依存関係を特定し、ステークホルダーの懸念や課題を明らかにします。AI の導入は、他のテクノロジーよりもはるかに部門を超えた取り組みです。構想段階で設定した目標を内部で調整することが重要です。そうすることで、クラウドおよび AI の準備を全体的に改善するための戦略を立て、ステークホルダーの足並みを揃えて将来の賛同を得て、関連する組織的な変更管理活動を促進できます。
- 着手: このフェーズでは、初期の概念実証から本番稼働までのパイロットイニシアチブの実施に注力し、ビジネス価値の向上を実証します。パイロットは、組織とビジネスに大きな影響を与えるだけでなく、AI を適用することで有意義なメリットが得られる必要があります。パイロットイニシアチブが成功するかどうかにかかわらず、将来の方向性に影響を与えることができます。ここから学ぶことで、本番環境に移行する前に取り組み方を調整できます。
- スケーリング: このフェーズでは、幅広く持続的な価値を実現するために、本番環境のパイロットをスケーリングすることに重点を置いています。ここでのスケーリングには、ソリューションやイニシアチブの技術的能力だけでなく、ビジネスや顧客を対象とするリーチも含まれます。このアクティビティにより、あなたの活動が顧客価値に変換されます。

これらのサイクルを繰り返しながら、1 つのサイクルで達成できることの限界を認識します。野心を持って高い目標を掲げることは大切ですが、同じサイクル内ですべてを実行しようとすると、組織を落胆させる可能性があります。このため、より大きな全体像と多くの実用的で実行可能なステップ、およびこれらの小さなステップに関する測定可能な KPI を組み合わせることが重要です。これにより、すべてのステップで組織が目標に近づきます。一度ですべてを達成しようとしないでください。AI のトランスフォーメーションジャーニーを進めていく中で、基礎的能力を進化させ、AI の準備状況を改善してください。

基礎的な AI 能力

AI のトランスフォーメーションジャーニーを繰り返し進めるには、ビジネス、人材、ガバナンス、プラットフォーム、セキュリティ、オペレーションにおいて、AI を導入するための基礎的能力が必要です。能力とは、プロセスを使用してリソース (人材、テクノロジー、その他の有形または無形の資産など) を投入し、成果を達成する組織的能力です。次の図は、クラウドと AI の導入のための基礎的能力の一覧を示しています (ピンク部分)。グレーで表示されている項目は、AI の導入では変更されることのない既存の CAF 機能です。

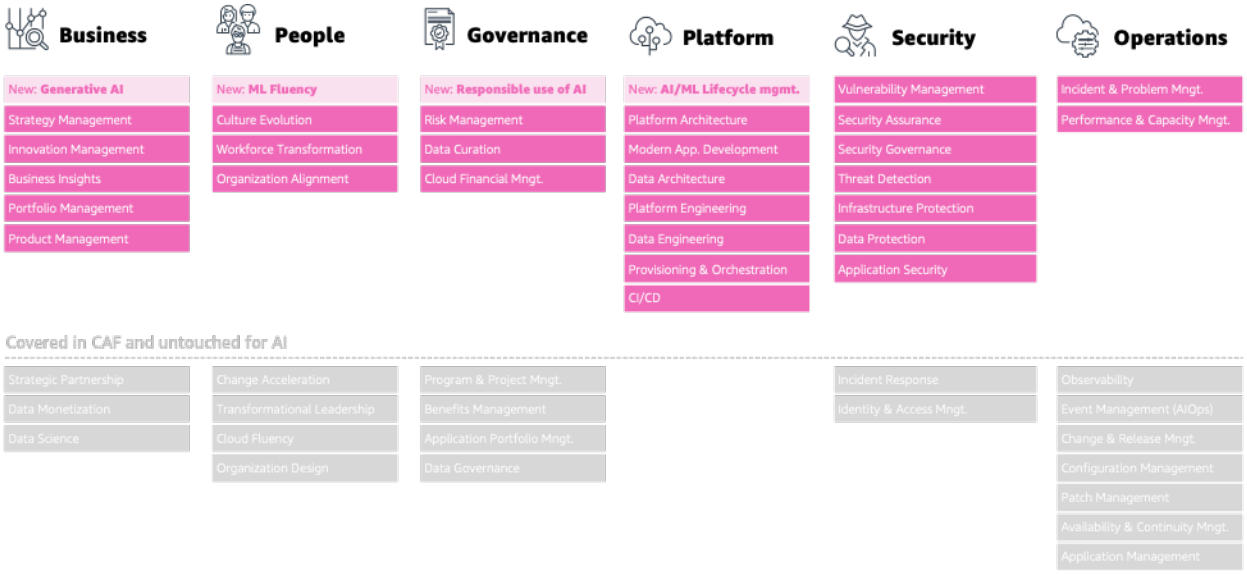

図 4: AWS CAF-AI の基礎的能力

ビジネスのパースペクティブのセクションにおける製品管理能力を例として考えてみましょう。製品管理は、クラウドベースの製品開発を成功させるために既に必須となっている能力ですが、クラウドでの AI サービスでは、その実装は大きく異なります。このドキュメントの残りの部分では、AI の導入に関する差異と特定のニーズについて説明します。その他の能力については、「AWS クラウド導入フレームワーク」ドキュメントで説明されています。どの経営幹部レベルの関係者がこれらの能力のどれを所有しているかは、組織によって異なります。多くの場合、複数の関係者が 1 つ以上の能力に共通の関心を持っています。このドキュメントを読み進めやすくするために、ある視点に関心を持つ代表的な関係者をリストアップしました。

- ビジネスのパースペクティブ: このパースペクティブは、AI への投資によってデジタルトランスフォーメーションと AI トランスフォーメーションという目標とビジネス成果を確実に加速する上で役立ちます。特に、このパースペクティブでは、能力の多くを充実させ、AI を主役に据え、リスクを軽減し、お客様のアウトプットや成果を高め、AI 戦略の効果的な策定を可能にする方法に

ついて説明します。一般的な関係者には、最高経営責任者 (CEO)、最高財務責任者 (CFO)、最高執行責任者 (COO)、最高情報責任者 (CIO)、最高技術責任者 (CTO) が含まれます。

- 人員のパースペクティブ: このパースペクティブは、AI テクノロジーとビジネスをつなぐ架け橋となり、継続的な成長と学習の文化を進化させ、変化が日常になることを目指しています。AI の時代における将来の競争優位性に最も影響を与える能力、つまり適切な人材、その言語、それを育む文化に焦点を当てて、AWS CAF を拡張します。一般的な関係者には、最高人事責任者 (CHRO)、CIO、COO、CTO、クラウドディレクター、および一般的にはその他の部門横断的な企業全体のリーダーが含まれます。
- ガバナンスのパースペクティブ: このパースペクティブは、組織のメリットを最大化し、トランスフォーメーションに関するリスクを最小限に抑えながら、AI イニシアチブを調整するのに役立ちます。私たちは、AI の開発とスケーリングの両方に関連するリスク、つまりコストの性質の変化に特に注意を払っています。さらに、新しい CAF-AI の論点、つまり AI の責任ある使用をこのパースペクティブに追加しています。一般的な関係者には、最高トランスフォーメーション責任者、CIO、CTO、CFO、最高データ責任者 (CDO)、最高リスク責任者 (CRO) が含まれます。
- プラットフォームのパースペクティブ: このパースペクティブは、AI 対応または組み込み型のサービスと製品の両方を運用できるだけでなく、新しいカスタム AI ソリューションを開発する能力も得られる、エンタープライズグレードのスケーラブルなクラウドプラットフォームの構築に役立ちます。AI の開発が通常の開発タスクとどのように異なるか、また実務者がその変化にどのように適応できるかを明らかにするために、能力を充実させています。一般的な関係者には、CTO、テクノロジーリーダー、ML 運用エンジニア、データサイエンティストが含まれます。
- セキュリティのパースペクティブ: このパースペクティブは、データとクラウドワークロードの機密性、整合性、可用性の達成に役立ちます。ここでは AWS CAF からのベストプラクティスに大きく依存しますが、AI システムに影響を与える可能性のある攻撃ベクトルについて推論する方法と、クラウドを通じてそれらに対処する方法について詳しく説明します。一般的な関係者には、最高情報セキュリティ責任者 (CISO)、最高コンプライアンス責任者 (CCO)、内部監査リーダー、セキュリティアーキテクトおよびエンジニアが含まれます。
- オペレーションのパースペクティブ: このパースペクティブは、クラウドサービス、特に AI ワークロードをビジネスニーズを満たすレベルで確実に構築するのに役立ちます。AI ワークロードを管理する方法、運用を維持する方法、信頼性の高い価値創造を確保する方法に関するガイダンスを提供します。一般的な関係者には、インフラストラクチャおよび運用リーダー、ML 運用エンジニア、サイト信頼性エンジニア、情報技術サービスマネージャーが含まれます。

これらの各パースペクティブには、能力に対処する、または能力を改善する、自然な、または論理的な順序があり、それにより AI トランスフォーメーションジャーニーのためのアクション領域の順序が決まります。次の図は、AI 戦略の経験豊富な実装者と一緒にこの順序と評価の例を示したもので

す。これは、これらの能力のどれが組織に存在しているか、およびそれがどの程度成熟しているかを確認するのに適しています。

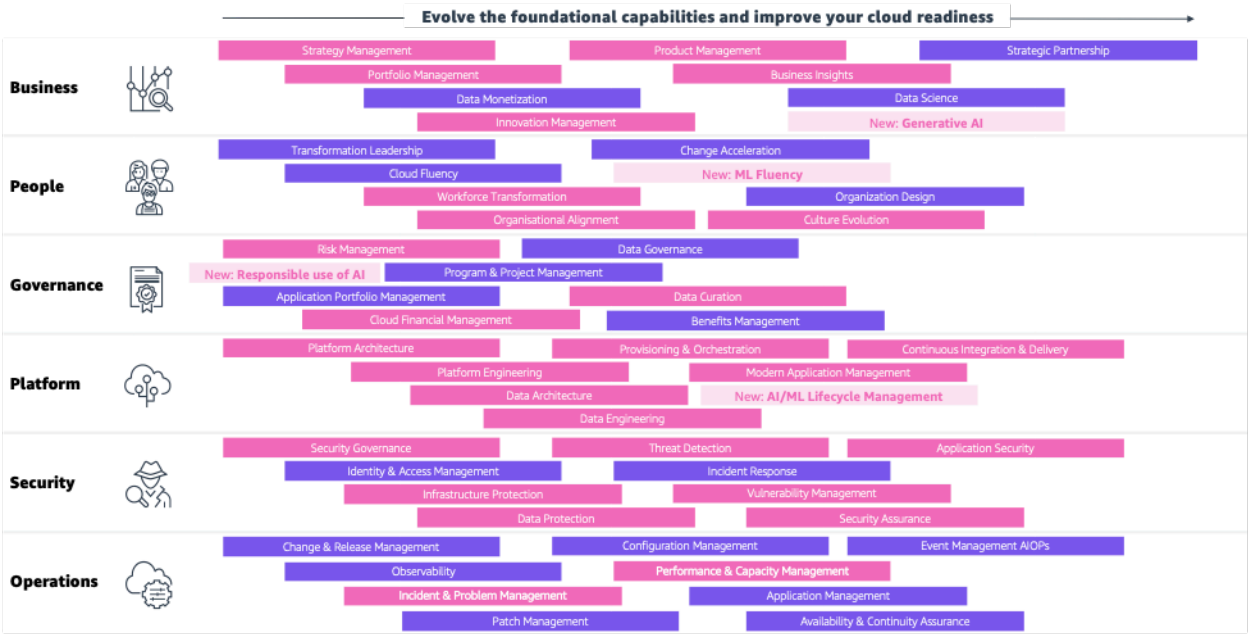

図 5: 成熟度と進化によって順序付けられる AWS CAF-AI の基礎的能力

ビジネスのパースペクティブ: AI の時代における AI 戦略

クラウドは組織のイノベーションを加速させますが、ML や AI などの新しい技術パラダイムは、まったく新しい組織的能力、製品、およびサービスを可能にします。何十年もの間、意思決定プロセスが複雑だったり、それを伝えるデータが構造化されていなかったり、意思決定の環境が絶えず変化したりするビジネス上の問題は、コンピューターサイエンスの方法では解決できないことが証明されてきました。

ML の最近の進歩によって状況は変わり、突然、機械が言語を見たり理解したり、過去のデータから学習して結果を予測したりする必要がある問題に対処できるようになりました。新しく、すぐに利用できる ML 機能により、運転支援や自動化を敬遠する自動車会社など、確立された組織の長年の市場仮説に疑問が投げかけられています。したがって、この視点は、企業がこれらのユースケースを最大限に活用できるようにする能力を対象としています。

基礎的能力	説明
戦略管理	人工知能と機械学習を通じて新しいビジネス価値を引き出します。

基礎的能力	説明
製品管理	データ主導型の製品や AI が組み込まれた製品または対応製品を管理します。
ビジネスインサイト	あいまいな質問に答えたり、過去のデータから予測したりする AI の能力。
ポートフォリオ管理	実現可能な価値の高い AI 製品とイニシアチブを特定し、優先順位を付けます。
イノベーション管理	長年の市場仮説に疑問を投げかけ、現在のビジネスを変革します。
新規: 生成 AI	大規模な AI モデルの汎用機能を活用します。
データ収益化	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
戦略的パートナーシップ	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
データサイエンス	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。

戦略管理

人工知能と機械学習を通じて新しいビジネス価値を引き出します。

機械学習は新しい価値提案を可能にし、ひいてはビジネスリスクの軽減、収益の増加、運用効率、ESG の改善など、ビジネス成果の向上につながります。そのため、まず、ビジネスと顧客を中心に据えた AI 導入の目標を定義し、AI テクノロジーの導入に向けて段階的に進む実行可能な戦略でそれを支えます。導入戦略が、これらの新しい能力を活用した、具体的な (短期的かつ測定可能な) または少なくとも野心的な (長期的で測定が困難な) ビジネスインパクトに基づくことを確認します。AI の導入による短期的影響と長期的影響の両方を考慮します。

既存のビジネスの問題や顧客の問題と、AI がそれらに与える影響から逆算して考えます。AI の機会の優先順位付けに近づいたら、どのようなデータが、どのようにシステム機能を強化するかについて検討します。ML 製品やサービスのデータフライホイールの自己強化特性、つまり、新しいデータが

システムの改善につながり、顧客基盤が拡大し、ビジネスが恩恵を受けるデータ量が増えることを最初から考慮します。

このようなフライホイールを構築する際には、取得したデータが価値提案 (まれでコストがかかることがある) の周囲の防御堀となるかどうかを考慮します。幅広い影響力のある AI テクノロジーが既に市場を取り巻いている状況に基づいて、近い将来、製品やサービスの機能に対する顧客の期待は高まることが考えられ、AI 機能もその期待の一部であることを考慮します。

各機会について、既存の AI システムを構築、調整、または導入する必要があるかどうかを考慮します。例えば、基盤モデルの幅広い創発能力を使用する予定があるが、ゼロから作成する能力がないため、特定のニーズに合わせてカスタマイズすることに焦点を置きます。ビジネスを推進するドメイン固有の汎用システムを作成することを目標としている場合は、データ基盤に投資します。

製品管理

データ主導型の製品や AI が組み込まれた製品または対応製品を管理します。

AI システムの開発とライフサイクルは従来のソフトウェアやクラウド製品とは異なるため、AI ベースの製品の構築と管理は大きな課題となる可能性があります。AI ベースの製品の開発、運用、結果の継続的な作成 (直接予測など) には、特定の緩和戦略を必要とする潜在的にコストのかかる不確実性が伴います。

AI を構築したり、製品に組み込んだりするときには、顧客やユーザーが期待する価値向上から逆算して、測定可能なビジネスプロキシを AI システムがサポート、強化、自動化できる個別の意思決定ポイントにマッピングします。これらのそれぞれについて、ML ソリューションドメインで考えられるメトリクス (金融セクターでの不正取引を検出することで得られる価値が、期待される金銭的利益と、ML 対応のトランザクション分類器の相関精度またはリコール率にどのように変換されるかなど) 、および対応する ML 問題 (分類問題、意図抽出問題、生成 AI など) を定義します。これらの定式化された ML 問題とそれぞれのソリューションが一緒になって、製品に対して ML がもたらす価値の向上を形成します。

重要なこととして、これらの ML ソリューションは、ユーザーと製品に特定のデータ要件を課すため、それぞれについてデータの 4 V を調査する必要があります。この知識をボトムアップで構築する際には、必ず、ビジネス、データ、経営幹部、ML 関係者をソリューションの評価に参加させます。ML 製品は、データ、ドメイン、テクノロジーを 1 つの予測システム、場合によっては規範的なシステムに融合するため、これらすべてが必要です。適切なライフサイクル管理を通じて、AI ベースの製品を進化させる道を切り開き、ユーザーが AI システムからの確率ベースの出力にどのように反応するか (システムの信頼性が低いときに潔く失敗するなど) を考慮し、ソリューションを採用した場合にどのような影響があるかを検討して、責任を持って AI を使用するようにします。

製品の ML 機能の範囲を適切に設定するには、どの質問が重要かについての理解を深めて、AI 製品管理能力を向上させます。つまり、例えば、ML コンポーネントのリスクを軽減するために、実験的な、多くの場合時間制限のあるアプローチを採用し、これらの実験から学んだことを本番環境レベルのシステムにどのように変換するかを最初から検討する必要があります。同様に、フィードバックループの設計をシステムの情報の流れに組み込む (または明示的に防止する) 必要があります。時間が経つにつれて、データメッシュ (データゾーンも参照) やデータレイクアーキテクチャなどのテクノロジーを通じて他の ML システムの出力に基づいて、また、チームと製品グループ間の適切な知識伝達を確立すること (SageMaker モデルカードなどによる実装) によって、より広範な組織が新しい AI 製品を構築できるようになります。

ビジネスインサイト

あいまいな質問に答えたり、過去のデータから予測したりする AI の能力。

ビジネスインテリジェンス (BI) は、多くの場合、記述的分析や診断的分析を含み、企業が AI の使用準備のジャーニーを始めるときによく使用されます。ただし、記述的分析や診断的分析の域を超えて、ML は予測機能、さらに規範的機能を可能にし、これらが組み合わさって AI ジャーニーを形成します。分析ユニットと BI ユニットの範囲は、AI 主導のユニットから組織的に予想される範囲とは異なっていることを認識することが重要です。

今日、多くの企業は、内容領域専門家 (SME) にインサイトをふるいにかけ、データ内の特定の観察結果の原因 (why) を抽出することを求めています。しかし、AI の手法を使用して、BI はこれらの SME を補足し始めており、why と what if を特定することにより、思考プロセスに組み込むべき新しいインサイトを与え始めています。これにより、データと AI は、突然、予測的意思決定の原動力になります。

BI プラクティスを AI 対応のプラクティスに移行し、一般的により高いレベルのアナリティクスに移行する準備をする際、限界を超える優れた方法は、アルゴリズムを診断分析に使用し、問題記述に影響を及ぼす主な変数または根本原因の理解に役立てることです。分析における組織の成熟度が組織の各部門でサイロ化されないようにし、AI ジャーニーを加速させるために、成熟度の高い組織と成熟度の低い組織を相互受粉させる方法を検討します。

トランスフォーメーションの初期段階では、効果的な方法とは、クラウドイニシアチブに密接に結びついた分析 (必ずしも AI ではない) のためのセンターオブエクセレンスを構築することです。このようなセンターオブエクセレンス (COE) は、データ主導型の予測と分析への民主的なアクセスを通じて、即値を提供し、より大きな目標を推進します。最も重要なのは、AI を使用して主要なビジネス上の意思決定を知らせるリズムを作ることです。そうすれば、真のビジネス成果に対する AI の価値の認識を促すことができます。

ポートフォリオ管理

実現可能な価値の高い AI 製品とイニシアチブを特定し、優先順位を付けます。

ML イニシアチブの課題は、長期的な価値を犠牲にすることなく短期的な成果を示さなければならないということです。最悪の場合、短期的な考え方が技術的な AI の概念実証 (POC) につながり、無関係なビジネス上の技術が重視されるために、その技術段階を超えることはできません。ML プロジェクトや製品を特定、優先順位付け、実行する際の最初の目標は、目に見えるビジネス成果を実現することでなければなりません。

どこかから始めることが非常に重要であり、小さな成果を上げることで組織への信頼を高めることができます。なぜなら、ビジネスの他の部分で AI を使用できる場所に人々がつながるのに役立つからです。同時に、複数の AI プロジェクトや製品を通じて解決しようとしている顧客やビジネス上の大きな問題を検討し、それらを階層的なポートフォリオにまとめ、そのポートフォリオの下位層が上位層に対応できるようにします。特定の AI 機能を 1 回で構築することはできません。むしろ、それらは互いに支え合っています。例えば、金融業界では、顧客に新商品を勧める前に、現在何が重要かを分類できなければなりません。そのため、取引の分類が事前の提案より優先されます。ポートフォリオの各レイヤーは、組織全体に付加価値をもたらすはずです。

図 6: データ戦略による自己補強フライホイールの作成

次に、このポートフォリオに AI フライホイールのデザインを埋め込みます。ポートフォリオがもたらす価値がビジネス上の成果を後押しし、ひいてはポートフォリオにメリットをもたらす追加のデータを生み出します (図 6 を参照)。このフライホイールは、単一製品レベルである必要はなく、ポートフォリオを通じて達成できます。 ポートフォリオが進化し拡大するにつれて、何を構築するかよりも何を購入するかを優先することが重要になります。Not Invented Here 症候群を押し返します。

どのユースケースとどのソリューションが市場に既に存在し、どの程度の成熟度かの調査は、後回しにすべきではありません。また、どのソリューションがカスタムモデリングを必要とするかも調査し、適切な AI 製品とクラウド環境を選択することで、AI ワークフォースの効率を高めます。ポートフォリオを技術的に管理するだけでも非常に複雑であることを認識してください。希少な AI ワークフォースの効率を維持するには、決断力を持って大胆に取り組み、分析麻痺から脱却しましょう。

最後に、ポートフォリオが拡大し、組織のより多くの部分が AI を使い始めるようになったら、ビジネスユニット、チーム、および信頼できる AWS パートナー間の効率的なコラボレーションを可能にします (AWS DataZones、AWS Redshift、および AWS CleanRoom を参照)。

イノベーション管理

長年の市場仮説に疑問を投げかけ、現在のビジネスを変革します。

この視点の序文で述べたように、ML がビジネスに提供する新しい機能は、既存のビジネスやバリューチェーンを混乱させる可能性があり、多くの場合破壊的です。この汎用テクノロジーの力は、さまざまな分野で見られ、感じられ、事実上、例外はありません。AI 研究の長期的な目標は、知能を複製するか、少なくとも模倣することであるためです。知恵を働かせ、複雑な情報を処理し、推論して洞察を導き出し、行動を起こすという歴史的な人間の能力は、今や、高度な基盤モデルと生成 AI の挑戦を受けようとしています。イノベーションロードマップとイノベーション管理の実践において、これらの AI 研究の中長期的目標に向けて、短期的で現実的に適用可能な価値提案で橋渡しをします。

そのためには、まず、内部と外部の両方の視点から、変化し続ける顧客の期待とニーズを調査します。CAF-AI が提案するビジネス成果は、これらのニーズと期待を特定する上で指針となります。ML 対応製品または搭載製品のバリューチェーンを考えてみましょう。そして、コスト削減のためのイノベーション (プロセスの改善など)、収益と利益の増加のためのイノベーション (製品の改善など)、またはまったく新しい収入チャネルのためのイノベーション (新しい製品やサービスなど) を区別します。

ML を社内外の関係者や顧客に対する独自の差別化要因として活用し、位置づけます。ML を統合して、新しい機能を活用し、既存の機能を強化して、自動化によって労力を削減します。アクセスするデータに含まれるドメイン固有の知識を十分に活用し、強化します。AI システムの健全なデータバリューチェーンを設計して、長期的な価値創出を可能にします。ML ベースの製品の中には、時間の経過とともにしか成長しないものや、イノベーションサイクルが一部の企業が慣れ親しんでいるものよりも長くなる可能性があることに落胆しないでください。ML 対応製品の単一ラインを構築する一方で、価値創造プロセスの第一級市民にデータを集め、消費用の内部データ製品を作成することで、組織全体のイノベーションへの道を開きます。

さらに、イノベーション管理に対するこのトップダウンアプローチに加えて、社内の AI チャンピオンを通じて草の根運動を巻き起こします。これらのチャンピオンには、ビジネスオーナー、プロダクトマネージャー、技術専門家、経営幹部などがあります。大胆な目標と達成可能な目標のバランスを常に保ってください。一般的なソフトウェアシステムと環境は、ユーザー数の増加とともに価値が高まりますが、ML システムの価値は、主にそれをより効果的にするデータによって決まります。したがって、AI イノベーションを管理するということは、過去のデータをアーカイブするだけでなく、データ戦略を実現することも意味します。組織の境界を越えて管理されアクセス可能な、高品質で価値のあるデータが増えれば、AI のアイデアやプロジェクトに重点が置かれるようになります。

新規: 生成 AI

大規模な AI モデルの汎用機能を利用します。

AI の全体的な目標は、一般的な品質で、追加費用をほとんどまたはまったくかけずに多くの複雑な問題領域に適用できるシステムを作成することです。この研究の特に強力な流れの 1 つが生成 AI です。これは、会話、ストーリー、画像、動画、音楽など、新しいコンテンツやアイデアを生み出すことができる AI の一種です。生成 AI は、膨大な量のデータであらかじめトレーニングされ、一般に基盤モデル (FM) と呼ばれる非常に大規模なモデルによって支えられています。このような FM の可能性が、ドメインとタスクをまたいで一般化する能力を支えています。このような基盤モデルは、ナレッジワークのコストを大幅に削減できるため、何らかの形で組織やビジネスに影響を与えます。この強力な AI の分野の導入を計画する際には、3 つの考慮事項があります。そのような FM を構築する必要があるか

1. ゼロから、ビジネスに合わせて独自にカスタマイズするか
2. 事前にトレーニングされたモデルを微調整して、既に学習した能力を活用するか
3. 追加のチューニングを行わずに、サプライヤーの既存の FM を使用するか

この 3 つの中から選択することが不可欠であり、正しい選択はビジネスケースによって異なります。多くの場合、これらの大規模モデルの真の価値を引き出すには、ドメイン固有のデータでコンテキスト化し (ケース 2)、さまざまなタスクに適用する必要があります。これは、事前にトレーニングされた大規模なモデルには、ゼロから作成する (ケース 1) にはコストがかかる新しい機能 (推論など) が既に備わっているためです。したがって、基盤モデルと生成 AI を使用する際には、ほとんどデータがない状態に適応して学習するトレーニング済みモデルの能力を利用します。

多くの企業にとって、このアプローチは、ビジネス上の問題に適した基盤モデルを選択し、(例えば、インストラクションチューニングや少量データ学習を通じて) カスタマイズし、ドメインまたは顧客固有のデータを使用して微調整することを意味します。生成 AI と基盤モデルの有効性と差別化能力は、他の AI システムと同様に、データ戦略とデータフライホイールに大きく依存します。データは本番環境でのモデルの動作に影響し、生成 AI システムに関するガードレールを確立することは非常に難しいため、どちらの方法を選択する場合でも、使用するデータに満足していることを確認してください。

人員のパースペクティブ: AI ファーストに向けた文化と変化

AI を導入し、確実かつ反復的に価値を創造することは、単なる技術的な課題ではありません。どの AI イニシアチブも、それを守り、推進する人々に大きく依存しています。汎用テクノロジーとして

の AI は複数のセクターに影響を与えますが、従業員がその能力を活用している組織は成功します。優れた AI システムがいかに実現されるか、つまり、関係者、ビジネスユニット、プラクティス間のコラボレーションを通じて実現されることを考えると、なおさらです。

AI が人間の労働を自動化する可能性についてよく言われますが、実際には人間の労働を豊かにしたり、補ったり、さらに増強したりします。一部のドメインでは自動化が可能ですが、今日の AI は主に、人間にとって特に複雑だと認識されているタスクを支援することを目的としています。AI ファーストの組織は、運用コストを削減し、収益を増やして、やりがいのある有意義な仕事を従業員に提供していることがわかります。このパースペクティブの焦点は、組織を結集し、適切な人材を育成して、ビジネス上の価値ある問題を探す際に同じ言語を話すことです。AI の導入においては、なおさら文化が重要です。このパースペクティブには、次の表に示す 7 つの能力が含まれます。一般的な関係者には、CIO、COO、CTO、クラウドディレクター、および部門横断的な企業全体のリーダーが含まれます。

基礎的能力	説明
新規: ML フルーエンシー	共有言語とメンタルモデルの構築。
労働力のトランスフォーメーション	ユーザーからビルダーまで、AI 人材の誘致、活用、管理。
組織の連携	組織間のコラボレーションの強化と依存
文化の進化	AI を導入する場合はなおさら、文化は王様です。
トランスフォーメーションのリーダーシップ	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
クラウドフルーエンシー	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
組織デザイン	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。

新規: ML フルーエンシー

共有言語とメンタルモデルの構築。

人工知能と機械学習の境界と意味的範囲は十分に特定されていません。どちらの用語にも、さまざまなメンタルモデルや感情的な解釈が多数含まれているため、関係者が何を意味しているのかについて、社内で意見を一致させることが重要です。これらの言葉の意味についてほぼ一致した視点を広め、その言葉に興味を持つ関係者を将来の社内 AI チャンピオンとして特定します。

最初の解釈レイヤーが組織全体に広がったら、2 つ目の、より技術的なレイヤーに取り組みます。AI プロジェクトと要件は、用語や重要度が異なります。製品管理の実務からエンジニアリングやデータサイエンスの実践まで、効果的に機能するためにどのような共同理解が必要かについて意見を一致させてください。効果的な方法は、異なる実務間のインターフェースワードを定義することです。例えば、ML で成功を測定する方法と、ビジネスドメインで成功を測定する方法などです。

これらの調整は、組織全体の賛同を得るのに役立つので、ML フルー遠視–と ML 文化のトレーニングを通じて実施してください。この理解は、ビジネスオーナーが ML ユースケース特有の側面に適応し、顧客に対する期待を設定するうえで非常に重要になるでしょう。

最後に、組織内と顧客の両方に AI のアウトプットを最もよく伝える方法を検討します。顧客のメンタルモデルや用語はそれぞれ異なることを考慮してください。そのため、例えば、AI システムを潔く失敗させて信頼を維持することは困難です。適切な言語と流暢さがあれば、効率が向上するだけでなく、顧客の関心に合わないシステムを構築するリスクも軽減されます。

労働力のトランスフォーメーション

ユーザーからビルダーまで、AI 人材の誘致、活用、管理。

AI 戦略を推し進めることができる人材を引き付け、定着させ、再トレーニングすることは、AI の成功にとって最も重要な側面の 1 つです。AI の成功に必要な役割は数多くあり、その中には外部委託できるものもあれば、社内の労働力のみが影響力を持つものもあります。最初のステップとして、AI 戦略のリーダーは、ビジネスと緊密に連携して、内部から価値を引き出す必要があります。この役割を外注先が担うことはめったにありません。

AI の導入を成功させるために必要なさまざまな役割を雇用または育成して、リーダーの能力を高めましょう。

- 技術的な人材 (データサイエンティスト、応用科学者、ディープラーニングアーキテクト、ML エンジニアなど)。
- ロードマップを管理し、ニーズを特定する、技術者以外の製品に関する人材 (ML プロダクトマネージャー、ML ストラテジスト、ML エバンジェリストなど)。

雇用戦略を全体的な AI 戦略および目標と緊密に連携させます。

- 科学的に野心的な大規模なイニシアチブには、長年の経験を持つ博士号取得者が適しているかもしれませんが、ML ストラテジストなど、ビジネスに近い人物で補完するのが最善策です。
- 既存の人材の一部を AI の役割に移行させることは、組織全体での導入にとって有益です。
- 確立されたソリューション、基盤モデル、または組織の手の届かない AI 作業に基づいて AI 機能を構築する予定がある場合は、ML エンジニアと深層学習アーキテクトを雇用するのが最も合理的です。

この社内労働力に加えて、早い段階で適切な AWS パートナーに連絡して、AI アジェンダが立ち消えに終わらないようにします。人材がいない場合は、AI のビジョンを社外に広め、成果を生み出し、新しい人材を鼓舞する取り組みを始めます。歴史的に供給が需要に追いついていないため、AI の人材を維持することは難しいことを最初から認識してください。もう 1 つの要因は、現実世界の AI が多くの人材を AI に引き込む学術研究とは大きく異なる点です。AI の専門家が協力したり、会議でプレゼンテーションを行ったり、ホワイトペーパーを書いたりといった機会を設けることで、この要因に対処できます。

ただし、人員の自然減少は避けられません。柔軟に対応し、適切なタイミングで人材を雇用するプロセスを確立して、人員の減少が発生したときに補充できるリソースを確保しておきます。CAF-AI の他の部分で参照しているプロセスは、企業が人員減少に対して堅牢であるために不可欠です。AI の分野で活躍するために必要な新しいスキルを学ぶ継続的な再トレーニングの機会を通じて、AI の人材を育成します。このアプローチには、プロジェクトを実行できるだけでなく、ビジネスに関する深い知識を持つ人材を確保できるという利点もあります。最後に、AI における人材と価値の比率は他の分野よりも低いことを認識してください。通常、有能なプラクティショナーで構成された小規模なチームは、知的というよりも機械的な作業が少ないため、大規模なチームよりも優れた成果を上げます。

組織の連携

組織間のコラボレーションの強化と依存

AI が組織の最重要課題になると、その価値と知識を組織全体に広めるために、カプセル化され権限を与えられた個別のユニットを提供することが、典型的な第一歩です。AI センターオブエクセレンス (COE) は、AI に焦点を当てたチームを雇用し、進化させるというこの役割を果たすことができるユニットです。この組織内の命令系統が、組織内の AI 戦略の所有権を持つ関係者と一致していることを確認し、経営幹部への近道を確保してください。そうすることで、必要なときに意思決定や変更を迅速に行えるようになり、新しいチームが自分のリズムを取り戻せるようになります。同時に、このような COE のインセンティブを自社の戦略、ビジネス、そして最も重要なことに顧客に合わせる

ことが重要です。よくある間違いは、ビジネス価値をもたらさない AI ユニットを進化させることです。

時間が経つにつれて、労働力のトランスフォーメーションにより、より広い組織や他のビルダーが COE と既存の AI サービスを効果的に利用し、効果的にコラボレーションできるようになるはずです。Not Invented Here 症候群を防止してください。組織がビジネス要件を満たしていれば、クラウドですぐに利用できるものを再構築することはありません。COE と人材がエンジニアリング精神を養い、異なるシステムを維持するコストを認識し、DevOps 精神を文化にもたらす MLOps のベストプラクティスを確立するようにしてください。このようなユニット、他の社内ビルダー、および AI の人材が進化する間に、データ主導型の製品メンタリティを確立することで、データフライホイールを実現します。組織全体のビジネスがデータを共有および管理できるだけでなく、データ製品の活気あるエコシステムを確立できるようにします。ただし、そのようなデータ製品を自分のために構築しないでください。

文化の進化

AI を導入する場合はなおさら、文化は王様です。

AI ファーストの文化を育むには、古いメンタルモデルを壊さなければならないことが多いため、長くやりがいのあるプロセスです。一般的なクラウドやソフトウェア開発では、ビルダーが複雑なルールやシステムを体系化できるようにすることに文化の重点が置かれています。AI は、必要なアウトプットを生み出す適切なインプットを求める文化に大きく依存しています。テクノロジーを中心とする文化を回避するには、ビルダー、ビジネス、その他の関係者がビジネスチャンスや顧客のニーズから、あらゆる AI の課題に至るまで、逆算して取り組むという考え方を取り入れましょう。

逆算とは、ビジネス環境の変化によって期待される結果を事前に定式化し、その変化を実現するために何が必要かを検討することです。ある意味、これが AI システムの構築方法です。つまり、期待される出力を定義し、その出力を可能にする信号を含む入力を探すということです。

このような価値主導の考え方が整ったところで、AI ファーストの文化の基礎を詳しく見てみましょう。

- 実験的な考え方とアジャイルエンジニアリングの実践の組み合わせ
- チームやビジネスユニット間のコラボレーションと依存
- ボトムアップおよびトップダウンの AI 機会発見
- 顧客価値に基づく幅広く包括的な AI 導入ソリューション設計

次のことを行って、AI ファーストの文化を広げ始めましょう。

- ビルダーが AI システムを試せるようにしましょう。実験のためではなく、AI システムを構築するには、どのソリューションパスが機能し、どれが行き止まりかを調査する必要があるからです。経路がわかっている既存の AI サービスの導入に伴うリスクの軽減を検討すると役に立ちます。

  実験は許可しますが、アジャイルの考え方を AI の不確実性に合わせて調整します。ビジネス価値の高い複雑な AI 問題の多くはまだ解決されていないため、複雑なプロジェクトの時間と労力の見積もりを確実に定義することはできないことを認識してください。その場合、期待される顧客価値が最も高いものを強化します。
- データがチーム間のインターフェースとなり、価値が互いに連携して生み出される文化を取り入れます。ビジネスから離れたデータサイエンスチームを構築するのではなく、コラボレーションのフライホイールを生み出す文化を構築するように注意します。
- 組織のあらゆるレベルで価値を見出し、認識して、活用する文化を強化します。これには、リーダーシップにインセンティブを与えて昇進させ、現状に挑戦することが含まれます。
- AI の影響や使用が懸念される環境の構築は、意見を聞くだけでなく、意思決定プロセスに影響を与えます。

ガバナンスのパースペクティブ: AI 主導の組織の管理

組織の AI イニシアチブの管理、最適化、スケーリングは、ガバナンスのパースペクティブの中核をなします。AI ガバナンスを組織の AI 戦略に組み込むことは、信頼を構築し、AI テクノロジーを大規模にデプロイし、課題を克服してビジネスの変革と成長を促進する上で役立ちます。一貫性を高めることで、AI ガバナンスは、組織の目標に沿って AI テクノロジーが倫理的に使用され、効果的に管理されるようにします。そのために、AI ガバナンスフレームワークは、組織的なリスク、倫理的なデプロイ、データ品質と使用状況、さらには規制コンプライアンスに対処し、AI ワークロードのさまざまなコストパターンを管理するための一貫したプラクティスを組織内に作成します。AI のデプロイのためのスケーラブルなプロセスと標準の作成により、組織はビジネスユニット全体に取り組みを拡張して、長期的なビジネス価値を生み出すことができます。

AI ガバナンスのプラクティスの構築は、組織の AI 戦略に密接に沿って行う必要があります。最初のステップは、すべての主要なステークホルダーを特定し、複数のビジネスユニットの代表者でチームをまとめることです。このチームは次の作業を担当します。

- コンプライアンスや倫理目標などのガバナンス目標の定義、潜在的なリスク領域の特定。
- データ、透明性、責任ある AI、コンプライアンスを含むポリシーとガイドラインの開発。
- AI システム、パフォーマンス、コンプライアンス、バイアスのモニタリング、および定義済みのしきい値に基づいてアクションを決定するメカニズムの定義。

- ビジネス目標の達成や AI の安全性の確保のための、結果と既存のポリシーの継続的な修正。

このパースペクティブでは、ガバナンスの課題に対するいくつかのソリューションに触れ、新しい能力、つまり AI の責任ある使用について説明します。これは、AI 分野における将来的な競争上の優位性を決定する要素です。

基礎的能力	説明
クラウド財務管理 (CFM)	クラウドでの AI のコストを計画、測定、最適化します。
データキュレーション	データカタログと製品から価値を創造します。
リスク管理	クラウドを活用して、AI に内在するリスクを軽減し、管理します。
AI の責任ある使用	責任ある使用を通じて、継続的な AI イノベーションを促進します。
プログラムとプロジェクトの管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
データガバナンス	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
福利厚生管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
アプリケーションポートフォリオ管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。

クラウド財務管理 (CFM)

クラウドでの AI のコストを計画、測定、最適化します。

クラウドでの AI プロジェクトの管理では、トレーニングと推論のコスト構造を計画する必要があります。これは、個々のプロジェクトの予算編成や AI イニシアチブの全体的な資金編成の際に事前

に検討することが重要です。AI ライフサイクルにおけるこのようなコスト構造の例として、ジグザグコストや、低/高/低/高コストのフェーズがあります。

- ソリューションの構築に必要なデータの品質を確立または向上させるため、初期コストが高くなることがあります。ただし、データの準備が整っている場合、この初期コストは非常に低く抑えられる可能性があります。その後、概念実証フェーズが続きますが、このフェーズは不安定なものになることがあります。
- コンピューティングの側面では、ほとんどの AI 概念実証 (POC) イニシアチブは比較的低コストに抑えることもできますが、大規模なモデルのトレーニング (生成 AI のコンテキスト) やドメインに特化した ML モデルの継続的な再トレーニングなど、コスト増に直結しやすいいくつかの技術的な側面があります。このような場合、AWS Trainium を使用した Amazon Elastic Compute Cloud (Amazon EC2) Trn1、または AWS Inferentia2 を使用した Amazon EC2 Inf2 インスタンスなど、専用の AI ハードウェアを活用して、コストを低く抑えることができます。適切な人材、AI サービス、および AWS パートナーにアクセスできる場合は、知見を活用してユースケースのさまざまなフェーズと AI 戦略全体に必要なリソースを見積もってもらいます。可能であれば、ML メトリクスを少しずつ改善するために何が必要かを判断して、投資を最適化する方法を決定します。
- 最初のシステムを構築した後は、システム機能を一般化したり、ユーザーへのシステムの導入に不可欠なエッジケースやロングテールデータを取得したりするために、次の実用最小限の製品 (MVP) フェーズのコストが比較的高くなる可能性があります。生成 AI 機能を必要とするユースケースでは、基盤モデルの使用や微調整を行うことで、大幅なコスト削減が可能な場合があります。これは、初期トレーニングのコストがサプライヤーやベンダーによって (Amazon BedrockTitan Foundation Model など) によって吸収されるためです。
- AI モデルがデプロイされた後、推論自体はリクエストの量に大きく依存し、多くの場合、推論コスト自体も比較的低くなります。そうでない場合は、専用の AWS Inferentia アーキテクチャを活用します。このステージでは、モデルメトリクスをモニタリングし、ドリフトにフラグを付けることで、変更やアルゴリズムの再トレーニングが必要になる可能性に気づくことができます。クラウドでのスケーリングの低コスト性を活用できます。AI のライフサイクルを通じて、コストを追跡し、すべてのリソースと ML ワークロードにタグを付けることが重要です。

コストの可視化対策ができたら、データ、トレーニング、および時間の経過に伴う推論コストを分析することが重要です。テキスト、予測、文書処理など大量の問題の種類があり、初期段階ではそれほどコストがかからない場合でも、規模に比例して増加します。音声と音声データに依存する他の AI の問題は、初期費用がはるかに高く、POC 段階でも大幅な請求が発生しないように、明確な目標を設定する必要があります。AI ビジョンをビジネス目標に合わせることで、どのように作業範囲を絞るかが決まり、モデルコストとモデルパフォーマンスのトレードオフを計算するメカニズムを確立す

ることは、ROI を維持する上で重要です。さらに、データ収集のコストは、組織がデータプロセスを中心に確立するメカニズムの影響を強く受けます。新しいデータやマスターデータを取得する標準的なプロセスは、(コピー/読み取り/コピーや ETL の必要性を減らして) AI に使用できる形式にデータを維持するのと同じくらい、コストを抑える上で重要です。クラウドは、管理されたデータサービスとゼロ ETL パターンを通じて、これらすべての課題を支援します。

これ以外にも、基盤となるビジネス目標に AI イニシアチブを常に結び付けてください。新しい収益源に関連する場合は、どのくらいの収益がどの成功基準に関連しているのかを想定し、ビジネス価値を AI のメトリクスに変換します。AI の責任ある使用の必要性を認識しないことによる過小評価されがちなコストも考慮に入れてください。その重要性から、AI の責任ある使用をこのパースペクティブでの新しい能力として追加しました。

データキュレーション

データカタログと製品から価値を創造します。

データの取得、ラベル付け、クリーニング、処理、操作ができると、処理速度が向上し、価値実現までの時間が短縮され、モデルのパフォーマンス (精度など) が向上します。モデルの精度が落ちた場合は、アルゴリズムに入力しているデータに戻って、データの充実、増加、または改善を検討します。そうすることは、多くの場合、モデリングだけで再構築したり、パフォーマンスの次のパーセントを引き出したりするよりもはるかに簡単です。

ML を念頭に置いたデータ収集は、AI ロードマップの達成に不可欠であり、自分自身や他のリーダーたちに「データを民主化することで AI のイノベーションを実現できるのか」、「組織はデータを製品とみなしているのか」、「私のデータは組織全体で発見できるか」と問いかける必要があります。これらの質問への答えは「はい」と「いいえ」の間にあることが多いですが、覚えておくべき重要なことは、データを現代の発明の起源として認識することです。取り組みの早期からデータをコードとして扱い、ビジネスにおける最も重要な要素としてデータを考えることが重要です。

データ品質評価とガバナンスに関するルールは、データの利用を加速するか、すべての進展を止める可能性があります。この 2 つのバランスを取り、適切なツールを使用して、組織全体がイノベーションを起こせるようにします。データセットの直接的な所有者やデータスチュワードは、堅牢なデータエコシステムの構築に役立ちます。最初は小さく始め、継続的にデータメッシュに追加して、データフライホイールを回転し続けます。ユーザーの種類に応じて、さまざまな方法でデータにアクセスしたり検索したりできるようにします。このアプローチにより、環境内で行われている作業をより詳細に把握でき、シャドー DataOps を回避できます。

人間が読める使いやすいデータリポジトリと辞書は、組織のデータセットに関する一元的かつ組織的なデータとメタデータのリポジトリとなります。これにより、あらゆるスキルレベルのチームが

データを探し、活用し、コラボレーションを行い、データを活用したビジネス価値の創造を始めることができるようになります。そのため、他のケースで必要となる追加投資費用の決定が格段にスピードアップします。データの可能性を高めるには、次のようなさまざまな方法があります。外部データソースの購入、ML アルゴリズムによるデータの拡張または作成、社内データにラベルを付けるチームをクラウドソーシングする、ビジネスプラクティスを変更してデータの生成とキャプチャを自動化する、といったことも可能です。これらの各リソースをいつ使用するかを決めるためのプラクティスを戦略的に策定します。

リスク管理

クラウドを活用して、AI に内在するリスクを軽減し、管理します。

すべての新しいテクノロジーには新しいリスクが伴いますが、AI モデルの設計と開発のプロセス、AI のデプロイ、および AI の長期的な運用と適用の両方に関連するリスクの管理は、AI モデルが持つ不確定性により、困難なものになります。また、これらのリスクには財務上のものも含まれます。AI 開発イニシアチブの成果を前もって保証するのは難しいため (アウトプットに合わせてシステムを最適化するよりも、具体的にシステムを構築するという性質)、開発プロセスの埋没コストのリスクを考慮することから始めましょう。モデルカードや敵対的なインプットなどの確固たるプラクティスとメカニズム (POC、Minimum Loveable Products、実用最小限の製品 (MVP) など) を確立して、リスクを軽減し制御します。

さらに、法的および倫理的性質を持つリスクもあります。これには、例えば、欧州連合など、地域の議会によって分類されたリスクと、隠れたフィードバックループやキャリブレーションされていないアウトプットによる誤解、さまざまな人々に成果が予期せず与える可能性のあるマイナスの影響など、AI に固有のものが含まれます。また、業務上、組織上、さらに社会的な使用や影響 (エコーチェンバーや顧客行動への長期的な影響など) も考慮してください。詳細については、「AI の責任ある使用」を参照してください。

安全性が重要な環境だけでなく、必要に応じてシステムを制約する保護手段やアーキテクチャを開発して採用することが重要です。ダウンストリームの AI システムにサブシステムの障害が伝播したり悪化させたりしないことを確認します。説明可能性、透明性、解釈可能性など、どのテーマが重要かを考慮します。これらのリスクを管理するには、AI の影響を受ける単一の意思決定やアクションだけでなく、対象となるプロセスや大規模なシステム全体にわたって管理してください。世界中のデータや概念のばらつきがシステムにもたらす長期的な課題を把握し、悪意のある攻撃者への対策に投資してください (「セキュリティのパースペクティブ: AI/ML システムのコンプライアンスと保証」を参照)。最後に、特定のドメインで人間と同等のレベルに到達するまでの複雑さを取り繕わないでください。

AI の責任ある使用

責任ある AI 実践を通じて、継続的な AI イノベーションを促進します。

最近まで、この強力な新技術の責任ある使用は、組織が AI ソリューションの技術的な側面のみにフォーカスし、特定のビジネス目標を達成しようする中で、後回しにされることがよくありました。ただし、AI システムは膨大な量のデータから学習し、その内容は必ずしも意図したとおりであるとは限らないことがわかってきたため、AI の責任ある使用の重要度が高まっています。責任ある AI の実践は、継続的な AI イノベーションを促進するための鍵であり、AI ソリューションの開発、デプロイ、使用が、倫理的に、透明性を持って、偏見なく行われるようにすることが重要です。アプリケーションの用途や影響範囲が広ければ広いほど、その重要性が高まります。したがって、AI ジャーニーの早期とライフサイクル全体を通じて、AI の責任ある使用 (RAI) を考慮し、これに対処してください。

AI ソリューションの安全性を確保し、従業員、顧客、および社会全体に悪影響を及ぼさないように、AI リーダーシップチームの一部として機能したり、緊密に連携したりするために、複数のビジネスユニット (リサーチ、人事、多様性と包含、法務、政府および規制関連、調達、コミュニケーションなど) の代表者による AI ガバナンスボードを設置します。このボードは、AI テクノロジーの開発、デプロイ、使用に関する倫理と責任のガイダンスを統括し、業界の規制および AI に焦点を当てたコンプライアンスへの準拠を推進する責務を担います。時間の経過とともに、責任ある AI が設計、開発、運用に与える影響をスケールします。システムが個人、各ビジネスユニット、ユーザー、顧客、そして社会にどのように影響するかを考慮してください。クラウドでの AI の拡張速度を考えると、説明可能性、公平性、ガバナンス、プライバシー、セキュリティ、堅牢性、透明性など、責任ある AI の主要な側面がどのように組み込まれているか、またさまざまな文化や人口統計がテクノロジーによってどのように影響を受けるかを考慮する必要があります。AI の責任ある使用や、それがイニシアチブにどのように影響するかについて、よく考え抜かれた原則や信条を含め、AI ビジョンの重要な一部にしましょう。特に、アルゴリズムの公平性、多様で包括的な表現、偏見の検出などを含めてください。

可能な場合は、設計による説明可能性を AI のライフサイクルに組み込み、意図した偏見と意図しない偏見の両方を認識して発見するためのプラクティスを確立してください。適切なツールを使用して、現状の監視と、リスクの通知に役立てることを考慮してください。責任を持って AI を使用する文化を可能にするベストプラクティスを使用し、チームがこれらの要素を調査できるシステムを構築または使用します。このコストは、アルゴリズムが本番稼働状態になる前に累積されますが、被害を軽減することで中期的には報われるでしょう。特に、基盤モデルの構築、調整、使用を計画している場合は、ハルシネーション、著作権侵害、モデルデータの漏洩、モデルのジェイルブレークなど、新たな懸念事項について把握しておいてください。元のベンダーまたはサプライヤーが開発に RAI ア

プローチを採用しているかどうか、どのように採用しているかを尋ねます。これはビジネスケースに直接波及します。

> **Note**
>
> AWS AI の責任ある使用チームは、この件について、実用的で充実したホワイトペーパーを作成しました。

プラットフォームのパースペクティブ: AI のインフラストラクチャと応用

AI アルゴリズムと ML アルゴリズム、およびその使用が高度化する中で、これらのアルゴリズムを実行するシステムやプロセスが古びてしまうことがあります。他の効率的な製造プロセスと同様に、統一され、一貫性のある製品を提供する AI 開発用のシステムとプラットフォームが必要です。この製品は、ビジネス価値を促進するアルゴリズム的な結果です。基礎的能力に沿ったプラットフォームを開発することで、競争上の優位性を引き出し、イノベーションを加速できます。リスク軽減プラットフォームは信頼性が高く、拡張可能で、このホワイトペーパーの他のパースペクティブに沿った、長期的なビジネス価値を可能にする基本的な能力の成果に基づいています。

AI 対応のプラットフォームは、コンポーネントを目的と意図に沿ったものにする一連の設計原則に基づいて、ML ライフサイクルのすべての側面を経時的にカバーする必要があります。その中心となるのは、分散データと管理対象データの管理とアクセスです。このデータは、個々のユーザーの特定のニーズを満たす方法で準備され、提供されます。さらに、包括的なエンドツーエンドの開発エクスペリエンスを通じて、新しい AI システムの開発をサポートする必要があります。また、既存の AI 機能や基盤モデルを活用することも不可欠です。これらのモデルをトレーニングしたら、オーケストレーション、モニタリング、共有を行い、ダウンストリームのコンシューマーとアプリケーション、システム、またはプロセスを統合することができます。これらのアクティビティは、一貫した改善を目的として、提供されたフィードバックに継続的に対応するプラットフォームイネーブルメントチームによって管理されます。

基礎的能力	説明
プラットフォームアーキテクチャ	繰り返し可能な AI の価値を実現するための原則、パターン、ベストプラクティス

基礎的能力	説明
最新のアプリケーションの開発	適切に設計された AI ファーストのアプリケーションを構築します。
AI ライフサイクルの管理と MLOp	機械学習ワークロードのライフサイクルを管理します。
データ アーキテクチャ	目的に合った AI データアーキテクチャを設計します。
プラットフォームエンジニアリング	強化された機能を備えた AI の環境を構築します。
データエンジニアリング	AI 開発のデータフローを自動化します。
プロビジョニングとオーケストレーション	承認された AI 製品を作成、管理、配布します。
継続的インテグレーションと継続的デリバリー	AI の進化を加速します。

プラットフォーム アーキテクチャ

繰り返し可能な AI の価値を実現するための原則、パターン、ベストプラクティス。

機械学習が研究主導のテクノロジーからエンジニアリングの実践へと成熟するにつれ、その応用から確実かつ反復的に価値を生み出す必要性がますます重要になっています。プラットフォームアーキテクチャの目的は、さまざまな CAF の視点から入力を検討し、ビジネス目標に沿った基盤を設計して、AI ライフサイクルの導入とイネーブルメントを保証することです。まず、プラットフォームのステークホルダーの成熟度と能力と、ML スタックから何が必要かを理解します。例えば、構築済みの既製の AI サービス、ローコードおよび AutoML 機能の使用を有効化し、専門家でないユーザーが AI を使用できるようにするかについて考えます。または、インフラストラクチャに直接アクセスできる ML フレームワークの使用やカスタマイズなど、AI 開発ライフサイクル全体を通じた専門家による使用をサポートすることを目的とするかについて考えます。特に生成 AI の分野では、このような質問はプラットフォームアーキテクチャに大きな違いをもたらします。次の 3 つのレイヤーで、AI に関する具体的な要件を検討します。

1. コンピューティングレイヤー: AI では、トレーニングと推論に大量のコンピューティングリソース (基盤モデル用) が必要になる場合があります。使用のガードレールに加えて、コストとパフォー

マンスの比較は、組織の標準を設定する上で重要な要素の 1 つです。従来の CPU や GPU よりも優れた価格パフォーマンスでコストを削減できる専用のハードウェアの使用を検討してください。

2. ML および AI サービスレイヤー: プラットフォームが ML および AI サービスの開発、デプロイ、イテレーションをどのようにサポートするかを設計します。ML サービスは、専門家のステークホルダーがカスタムモデル (基盤モデル など) をトレーニングまたはチューニングできるようなものである必要がありますが、AI サービスはモデルと機能 (生成 AI ドメイン内の大規模でコストのかかる基盤モデルなど) の使用が可能なものである必要があります。この分離は必ずしも容易なことではなく、要件は異なります。
3. 使用レイヤー: AI 機能のダウンストリームコンシューマーは、このレイヤーでオペレーションを行います。これは、ダッシュボードのようなシンプルなものでも、プロンプトエンジニアリングによる基盤モデルの拡張、または検索拡張生成 (RAG) アプリケーションなどの特定の生成 AI アーキテクチャのような複雑なものでもかまいません。

プラットフォームを構築する際は、データ、モデル開発プロセス、デプロイ (データの必要なセグメンテーションなど) に影響する業界固有の法的要件を分析し、それに応じてガードレールを適用します。データのプライバシーやデータガバナンスなど、ダウンストリームチームが使用する際の標準を特定し、公開します。次に、準拠した環境とインフラストラクチャのプロビジョニングを合理化し、新しい AI ユースケースの開発とデプロイを高速化します。チームが AI ワークフローの重要なチェックポイントとしてどのようにヒューマンインザループ (HITL) およびヒューマンオンザループ (HOTL) 機能を使用するかを理解して、フィードバックループをプラットフォームに統合する計画を立てます。最後に、バイアス検出、説明可能性、レビュー担当者など、モデルの動作が変化した際の ML 固有のモニタリングの必要性を特定します。

AI サプライチェーンにモジュール設計を採用することは、独立したスケーリングと更新を可能にするために重要です。このモジュールアプローチは、より迅速なデータラベリングに役立ち、さまざまなコンポーネントの所有権と説明責任を明確にします。どのクラウドネイティブソリューションを標準化するかを決定する際には、コスト、信頼性、耐障害性、パフォーマンスなどの要素を考慮する必要があります。これらのベストプラクティスはすべて、設計ガイドラインや標準とともに、組織内のすべての実務者がアクセスできる中央リポジトリに公開する必要があります。プラットフォームの導入を測定するフィードバックメカニズムとメトリクスを実装すると、AI イニシアチブを継続的に把握できるため、情報に基づいた意思決定に役立ちます。

最新のアプリケーションの開発

適切に設計された AI ファーストのアプリケーションを構築します。

> **Note**
>
> AWS Well Architected Framework – 機械学習レンズ は、ワークロードとアーキテクチャの設計パターンとベストプラクティスの最終的なソースです。

AI テクノロジーが成熟するにつれ、これはアプリケーション開発のあらゆる側面に関連します。

1. AI で強化されたアプリケーション開発: AI を活用してソフトウェア開発ライフサイクル (SDLC) を強化します。AI サービスとツールを使用して、生成およびオートコンプリート機能を持つアプリケーションを推進したり、潜在的なコードの問題を特定するレビュープロセスや、効率的でエラーのない開発を確保して、パフォーマンスとテストを自動化することによって、合理化を行います。概念化からソフトウェアのメンテナンスまで、AI 機能を使用して SDLC を見直します。
2. 差別化要因としての AI: AI をソフトウェアに統合すると、ユーザーエクスペリエンスを向上させたり、価値提案の中核に据えるたりすることができます。AI はソフトウェアの機能を向上させ、ユーザーのニーズや期待と密接に一致し、最終的にユーザーが使用しやすい製品を提供することができます。このようなアプリケーションを開発する場合、データがシステム内をどのように移動するか、それによって AI システムがどのように変化するか、データによって生成される出力、これらの出力がコンシューマーや顧客によってどのように解釈されるか、ユーザーが使用する新しいデータにそれらの出力がどのようにつながるかを考慮します。AI で既に確立されている設計原則に基づいて、アーキテクチャ上の決定を行います。
3. AI モデルの開発: AI をソフトウェアに統合する際は、既存のモデルを適応させるか、オープンソースのオプションを利用するか、カスタムソリューションを作成するかを検討します。モダンアプリケーション開発が進化するにつれて、AI をマスターするには日々の活動が必要です。特定のデータを使用し、ニーズに合わせてモデルを微調整して、ユースケースに合わせたさらなるカスタマイズが必要な場合があります。

3 つの側面すべてについて、アプリケーションと開発プロセスを、小規模で管理しやすいサイズに分割する方法を検討してください。マイクロサービスまたはマルチモデルアプローチをアジャイルプラクティスと連携させることで、柔軟性を高め、完成までの期間を短縮し、変化により適切に対応できます。このアプローチは、反復テスト、実験、改良の必要性が高い AI 開発で特に有効です。AI システムは、顧客とユーザーによって実際には異なって認識されており、多くのユーザーにはシステムの使用に役立つメンタルモデルが欠落していることを、開発チームに明確に理解してもらいます。つまり、顧客やユーザーが使用するすべての AI ベースのアプリケーションは、ユーザーエクスペリエンス (UX) を新たに見直すことで直接的なメリットを得られます。

AI のライフサイクル管理

AI のライフサイクル管理は、組織の能力とともに成熟するアーキテクチャとエンジニアリングのパースペクティブに分かれます。

アーキテクチャのパースペクティブは、AI のライフサイクル管理の設計、計画、概念的な側面に焦点を当てます。機械学習ワークロードのライフサイクルの管理は、包括的なアプローチを必要とする複雑なタスクです。このライフサイクルは、次の主な 3 つの要素から構成されています。

1. ビジネス成果と顧客価値の特定、管理、提供
2. AI ソリューションの技術的コンポーネントの構築と進化
3. AI システムの経時的な運用は、機械学習オペレーション (MLOps)、または大規模モデルの場合、基盤モデルオペレーション (FMOps) とも呼ばれます。

これらの各コンポーネントは複雑なため、Well Architected Framework: ML レンズで詳細なガイダンスを提供しています。異なる AI 戦略においては、これら 3 つの要素に対する視点も異なります。例えば、全体的な目標がカスタムモデルから新製品を開発することである場合、一般公開されているサービスを通じて社内の運用効率の向上に努める場合とは異なる観点からライフサイクル管理を構築する必要があります。どのアプローチを採用する場合でも、一元化されたリポジトリとバージョン管理を使用して AI アーティファクトを保存し、モデル系統とデータ系統を追跡します。

エンジニアリングのパースペクティブは、AI のライフサイクル管理の実装と運用に重点を置いています。このプロセスを効率化するには、AI モデルのデプロイとモニタリングを自動化する MLOps プラクティスを実装して、工数の削減、信頼性の向上、デプロイまでの時間の短縮、オブザーバビリティの向上を行うことが重要です。概念からデプロイ、さらにはモニタリングまで、AI のライフサイクルを管理するための定義済みプロセスに沿ったものにします。このプロセスには、パフォーマンスのモニタリングを含む、データの収集と保存、モデルのトレーニングとデプロイ、モデルのモニタリングと評価 (CAF-AI の運用セクション) の手順を含める必要があります。これにより、早期に問題を検知し、モデルの継続的な進化をサポートできます。最後に、パフォーマンスが低下したり、新しいデータを受け取ったりした場合などに、AI モデルを再トレーニングするための自動化フレームワークを確立します。

業界のベストプラクティスに関連して現在の状況をよりよく理解するには、AWS パートナーまたは AWS の MLOps の成熟度を評価し、MLOps とライフサイクルフレームワークに基づいて意思決定を行います。これらのプロセスと標準は、組織の知識だけに頼るシステムに対する最善の自衛策となり、AI の技術的な負担を軽減します。データチームが、厳格な ML メトリクスがビジネスメトリクスにどのように影響を与えるかではなく、厳格な ML メトリクスのみに過度に焦点を合わせること

がよくありますが、これはライフサイクル管理においては失敗です。どの過程においても、MLOps 用に確立したプロセスと標準を再現可能なものにしてください。このような MLOps のベストプラクティスは、サイエンスチームがモデリングに疲労せずに、実験の大規模な並列化に気を取られることなく、結果に集中するのにも役立ちます。

データアーキテクチャ

目的に合った AI データアーキテクチャを設計します。

データは AI の鍵であり、データ型とデータ量は急増するため、従来のデータアーキテクチャを進化させる必要があります。特に、AI はビジネス意思決定の中心になりつつあるため、AI においては AI の複雑性に適した保存、管理、分析の新しいアプローチを必要とします。AI ワークロードは、大量のデータだけでなく、モデルのトレーニングと検証に多様で高品質のデータも必要とすることに注意してください。このようなデータは、多くの場合、さまざまな形式や構造で複数のソースから取得されるため、データの移動やタイプに制限があり、従来のデータアーキテクチャでは、この種の多様性やボリュームを効率的に管理することはできません。そのため、進化し続ける最新のデータアーキテクチャの分野についての知識を深めてください。これにより、データレイク、データウェアハウス、その他の専用データストアを一元化でき、ガバナンスの複雑さが軽減されると同時に、AI にとって不可欠なデータの移動が可能になります。

今日の組織では、構造化されスループット最適化済みのストアであるデータウェアハウス、さまざまなサイロからデータを集約し、中央データリポジトリとして機能するデータレイク、NoSQL データベース、検索サービスなどのビジネスアプリケーション専用ストアの 3 つが主要なアーキテクチャであり、それぞれが異なるユースケースをサポートしています。ただし、これらのストア間でデータを移動することは難しく、コストがかかる可能性があります。したがって、AI システムにとってデータ移動の重要性が増すにつれて、アーキテクチャのデータ移動要件は次のようになります。

- 内部から外部へ: データは、まず、さまざまなソースからデータレイクに集約されます。データベースや適切に構造化されたスプレッドシートのように構造化されているものも、メディアやテキストのように構造化されていないものもあります。サブセットは、検索分析やナレッジグラフの構築など、専用の分析のために専用のストアに移動されます。
- 外部から内部へ: データは、最初に、特定のアプリケーションに適した専用ストアに格納されています。例えば、クラウドで実行されているゲームをサポートするために、アプリケーションは特定のストアを使用してゲームの状態とリーダーボードを維持する場合があります。その後、このデータはデータレイクに移動され、より包括的な分析を実施してゲーム体験を向上させることができます。
- ペリメータの周辺: これは、リレーショナルデータベースから NoSQL データベースなど、特殊なデータストア間でデータを移動し、ダッシュボードレポートなど、特定のニーズに対応します。

AI チームの迅速なスピードを維持するには、このようなデータ移動をシームレスに行う必要があります。AI が急速に進化するにつれて、この柔軟性を持つことが鍵となります。AI においてはデータの重要性が高く、データをマシンコードのように処理できるため、AI とデータアーキテクチャの間の境界線は明確ではありません。最新のデータアーキテクチャにより、組織はデータ自体を製品 と見なすことができます。最新のデータアーキテクチャは静的なコンストラクトではなく、新しいデータ型やテクノロジーの出現に柔軟に対応できるよう設計されています。したがって、最新のデータアーキテクチャ、分散データメッシュ、データマートなど、さまざまな新しいアーキタイプのデータアーキテクチャを検討し、すべてのタイプのデータに対して統一されたプラットフォームまたはエコシステムをプロビジョニングします。最後に、現在のアーキテクチャを定期的に見直し、アクセスパターンとニーズを事前に検討し、目的に合ったアーキテクチャを選択します。データセットが見つけやすいものになっており、十分に文書化され、理解しやすいように計画してください。メタデータの原則またはデータドキュメントを作成して、データの意味、他のデータとの関係、オリジン、使用状況、形式など、データに関する情報を記載します。

プラットフォームエンジニアリング

強化された機能を備え、準拠した AI の環境を構築します。

クラウドにより、組織が最先端の AI インフラストラクチャとサービスにアクセスする方法は大きく変化しました。AI へのアクセスを民主化することで、組織は AI ワークフローを簡素化し、規模の経済が持つ莫大なメリットを活用できます。そのため、理にかなった AI プラットフォームにより、AI チームはより低いコストで、より多くのことを実行できます。プラットフォームを適切に構築し、さまざまなステークホルダー (開発者、データチーム、オペレーションなど) に簡素化と抽象化を提供し、作業方法を改善する能力を高めながら、その負荷を軽減します。

- AI サービス: 構築済みのモデルや特定のユースケースを考慮して、構築するプラットフォームと既製の AI サービス間の接続を簡素化することで、チームが最新のデータアーキテクチャを直接使用できるようにします。
- ML サービス: クラウドでは、開発者は AI アプリケーションの開発とデプロイ用に設計された専用の環境を使用できます。AI モデルのトレーニングとデプロイにおいて、このようなマネージド機械学習サービスは不可欠なものです。これらによって、ML システムのエンジニアリングに固有の複雑なプロセスや長期間のプロセスを効果的に処理できます。これらのサービスを採用することで、AI チームは貴重な時間をより戦略的イニシアチブに費やすことができるようになります。
- ML インフラストラクチャ: 高度に専門化された基盤 AI インフラストラクチャを管理することで、プラットフォームでの高負荷をチームから取り除くことができます。インフラストラクチャを所有することは、多くの場合、AI チームを強化することなく逆に妨げとなり、ビジネス価値の棚上げすることになる点に注意してください。

クラウドの主な利点の 1 つは、日常的なタスクを自動化する能力です。ML プラットフォームタスクによって、プロセスの高速化、人為的ミスの軽減、一貫性の確保を行えるため、このタスクを可能な限り自動化します。AI ソリューションが複雑になればなるほど、専用の MLOps プラクティスが重要になります。AI に特化したモニタリングツールを導入開始からプラットフォームに埋め込みます。このようなツールは AI ワークロードのパフォーマンスを追跡し、運用に関する貴重なインサイトを提供するため、問題を早期に特定するのに役立ちます。フィードバックメカニズムは、モデルの微調整とハイパーパラメータの設定に影響します。ワークロードをリアルタイムでモニタリングすることで、組織は AI アプリケーションのパフォーマンスを最も適切に向上させ、発生した問題に迅速に対処できるようになります。

クラウドには幅広い柔軟性がありますが、ガードレールを実装することが不可欠です。リスクの低減と責任のある使用を保証するための定義されたベストプラクティスとセキュリティパラメータの範囲内で開発者が作業できるようにするために、ガイドラインまたは制限を通じてこのようなガードレールを採用します。セーフティネットを提供し、イノベーションを推奨しつつ、組織のセキュリティ、コンプライアンス、またはパフォーマンス基準を侵害しないようにします。

データエンジニアリング

AI 開発のデータフローを自動化します。

データは AI 戦略と開発プロセスの基盤であるため、データエンジニアリングは最初に考慮すべきより重要なものとなりますが、組織やチーム内で簡単に利用できる能力です。データは AI システムの動作を能動的に形成するために使用されるため、データを適切に設計することが非常に重要です。データ準備ツールは、開発プロセスの不可欠な要素です。プラクティス自体は根本的に変化していませんが、その重要性と継続的な進化の必要性は高まっています。合理的かつシームレスな前処理を通じて、データパイプラインとプラクティスを AI 開発プロセスとモデルトレーニングに直接統合することを検討してください。従来の抽出、変換、ロード (ETL) プロセスからゼロ ETL アプローチへの移行を検討してください。データエンジニアリングへのこのようなアプローチにより、データプラクティスと AI プラクティスの間の摩擦を軽減できます。AI チームが複数のソースからのデータをセルフサービス機能として 1 つの統合ビューにまとめることを可能にし、権限を与えます。これを、AI やデータチームがデータを視覚的に探索して理解するのに役立つ視覚化ツールや手法と組み合わせます。

可能な限り、正確性、完全性、信頼性の高いデータに焦点を当てます。効率的なデータの取り扱いと処理を促進するため、(正規化され、一貫性があり、適切に文書化された) 機械学習専用のワークフローの一部として、データモデルまたは変換を設計します。これにより、AI アプリケーションのパフォーマンスが大幅に向上し、開発プロセスの摩擦が減少します。

プロビジョニングとオーケストレーション

承認された AI 製品を作成、管理、配布します。

AI システムのインフラストラクチャ要件は、さまざまな開発段階とデプロイ段階にわたって大幅に変化するため、プロビジョニングとオーケストレーションには既存のクラウド戦略に 2 つ目の視点が必要です。AI トランスフォーメーションジャーニーのどのステージにいるか、そしてそれが MLOps の成熟度にどのように関係しているかを理解します。使用者、データエンジニア、データサイエンティスト、開発者、ビジネスアナリストが各自の業務を行う際は、それぞれニーズと要件が異なることを考慮してください。さまざまなユーザー、特に技術的な専門知識に乏しいユーザーに対して、AI 環境のセルフサービスプロビジョニングを提供する方法を特定します。そのためには、プラットフォームアーキテクチャで承認されたカタログ、ポートフォリオ、製品を作成します。カタログはエンドユーザーに配布でき、ユーザーはカタログ内の製品を使用できます。製品は、Infrastructure as Code (IaC) として定義し、プラットフォームチームが管理する組織ポリシーに従って、パーソナライズされたポータル、または CI/CD パイプラインを介してデプロイできます。一般的なユースケースとしては、データチームが新しいビジネス課題に取り組むために、事前定義されたノートブックとコンピューティングを提供するパーソナライズされたポータルを提供し、プラットフォームチームがリソースをプロビジョニングするのを待つことなく、迅速に実行できるようにすることが挙げられます。一連のツールを必要とするデータサイエンティストなどのより高度な役割に対しては、基盤モデルのアクセラレーターへのアクセス許可の付与を含む AI 環境全体をデプロイするカタログを設定できます。

AI モデルのトレーニングまたはチューニングステップでは、高性能なコンピューティングが必要で、予算とガバナンスの制約に合った事前承認済みのサービスを使用してプロビジョニングを自動化することを検討してください。可能な限り、API レベルおよびフレームワークレベルのオートメーションとオーケストレーションを使用します。AI ワークロードのデプロイを管理し、基盤となるインフラストラクチャの作成を効率化するメカニズムを設計します。

継続的インテグレーションと継続的デリバリー (CI/CD)

AI の進化を加速します。

AI における継続的インテグレーションおよびデリバリーには、根本的に異なる 2 つの視点があります。1 つ目は、カスタムモデルの開発など、モデルの開発とデプロイのプロセスをできる限り自動化して強化することです。2 つ目は、DevOps エクスペリエンスの一部として AI を使用し、それを通じて CI/CD を容易にすることです。

1 つ目のステージでは、組織は AI モデルのデプロイとテストを自動化し、チームはクラウドがもたらすスピードでイノベーションを起こすことができます。カスタムモデルの開発では、データ処理、

モデルトレーニング、モデル評価、後処理、モデル登録、モデルのデプロイなど、複雑なワークフローの管理と AI ワークロードのデプロイと管理を自動化することが目的です。AI 開発プロセスを自動化する際は、ML パイプライン専用のツールを、従来のアプリケーション開発で一般的な方法やツールとともに使用します。適切なアーキテクチャとブループリントにより、データサイエンティストはさまざまなモデルを試し、本番環境に導入する前にモデルを徹底的にテストできます。この機能の構築が組織に適しているかどうかを、時間をかけて考慮します。そのためには、ML モデルの作成速度、更新の必要性、ユースケースの重要度と影響を理解する必要があります。モデルドリフトは時間の経過とともに発生する可能性があるため、再トレーニングのしきい値を定義するなど、検証プロセスをどの程度自動化できるかを検討してください。自動検証は、モデルのパフォーマンスを事前定義された基準に照らしてチェックし、モデルのパフォーマンスが許容可能なしきい値を超えてドリフトすると、自動再トレーニングまたは以前のバージョンへのロールバックをトリガーします。最後に、人によるフィードバックを統合し、モデルの検証、テスト、再トレーニングなどのタスクを自動化することによる反復性で AI ワークロードの信頼性が向上し、データサイエンティストやエンジニアはより戦略的なタスクに集中するために貴重な時間を費やすことができるようになります。これらの側面を統合すると、組織はデータや要件が進化する中で、モデルとカプセル化する AI システムの関連性と有効性を維持しながら、費用対効果の高い方法で AI モデルを進化させることができます。

2 つ目の視点では、AI や AI 以外の関連 DevOps アクティビティに AI を使用し、開発プロセスを強化して、適切な場合は生成 AI を活用します。AI がもたらす最も重要なビジネス価値のいくつかは、AI が開発プロセス自体に適用されていることから生まれます。そのため、ステークホルダーがそれを技術的なワークフローでどのように受け入れることができるかを検討します。つまり、AI を使用してワークロードの異常を分析したり、AI を使用してコードレベルのパフォーマンスを最適化したり、開発者のプロンプトに基づいてコードを生成したりすることができます。組織における DevOps 向けの AI の使用が、セキュリティを念頭に置いたものとなっていることを常に確認してください。

セキュリティのパースペクティブ: AI システムのコンプライアンスと保証

セキュリティは AWS の最優先事項であり、規模にかかわらず、AWS の安全なインフラストラクチャと新しいサービスへの継続的な投資からメリットを得られます。AI AWS ワークロードを開発しているお客様にとって、セキュリティは AWS ソリューション全体の不可欠な部分です。生成 AI は、ビジネス成果を実現するための基盤モデルをスケーリングするための重要なイネーブラーで、生成 AI ワークロードを作成する方法は複数あります。AI のあらゆる側面でセキュリティとプライバシーを統合することは、ビジネス成果の全体的な成功に不可欠です。AI を使用する基本的なビジネスケースには、日常的な生産性向上タスクのシンプルな自動化から、機密データを含む複雑な医療や財務上の意思決定まで、さまざまなビジネス課題の解決が含まれます。リスク管理手法を適用し、こ

のパースペクティブで定義されたセキュリティおよびプライバシー機能を実装して、ビジネスニーズを満たします。

基礎的能力	説明
Vulnerability Management	AI の脆弱性を継続的に特定、分類、修復、軽減する
セキュリティガバナンス	AI ワークロードに関連する役割と責任とともに、セキュリティポリシー、標準、ガイドラインを確立する
セキュリティ保証	AI ワークロードの規制およびコンプライアンス要件に対するセキュリティおよびプライバシーの対策を適用、評価、検証する
脅威検知	AI ワークロードにおける潜在的な AI 関連のセキュリティの脅威や予期しない動作を検出して軽減する
インフラストラクチャの保護	AI ワークロードの運用に使用されるシステムとサービスを保護する
データ保護	AI の開発と使用におけるデータの可視性、安全なアクセス、制御を維持する
アプリケーションセキュリティ	AI ワークロードのソフトウェア開発ライフサイクルプロセス中に脆弱性を検知して軽減する
Identity and Access Management (IAM)	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
インシデントへの対応	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。

脆弱性管理

AI の脆弱性を継続的に特定、分類、修復、軽減します。

AI システムには、プロンプトインジェクション、データポイズニング、モデルインバージョン脆弱性など、注意すべきテクノロジー固有の脆弱性が存在する場合があります。AI システムの 3 つの重要なコンポーネントは、入力、モデル、出力です。これらのコンポーネントは、以下のベストプラクティスで保護でき、ワークロードの潜在的な脆弱性を軽減できます。

- 入力の脆弱性は、モデルへのエントリーポイントを含むすべてのデータに関連します。この入力は、ターゲットモデルや分布のドリフトの目標となります。攻撃者は、時間の経過とともに決定に影響を与えようとしたり、特定のデータに隠れたバイアスや機密性を意図的に挿入したりします。データ品質の自動化と継続的なモニタリングにより、このような入力を強化します。モデルの悪用は AI ソリューションへのプロンプトインジェクションに起因する脆弱性の一例です。データや命令のクラウドは相互に組み合わされているため、急速に進化する基盤モデルの分野では特に注意が必要です。大規模言語モデル (LLM) へのアクセスを特定のユーザーに制限することで、入力検証を実行してデータを命令から分離し、最小特権の原則を使用します。システムコマンド、実行ファイル、および広範な運用上の影響を与えるログアクションへのアクセスは避けてください。
- モデルの脆弱性は、現実世界の虚偽表示とモデル内に表示されるデータの悪用に関するものです。脅威モデリングを使用して、既知の文書化された脅威を軽減することで、モデルを強化します。商用の生成 AI モデルを使用する場合は、そのデータソース、モデルの微調整に使用する条件、およびモデル自体またはサードパーティーライブラリの使用によって影響を受ける可能性のある脆弱性を確認してください。モデルの目標とその結果がモニタリングされ、時間の経過とともに一貫性が保たれていることを検証し、モデルのドリフトを防ぎます。
- 出力の脆弱性は、長期間にわたるシステムとの対話に関連しており、これにより、モデルの入力とプロパティに関する重要な情報を推測できる可能性があります (これは多くの場合、データ漏えいと呼ばれます)。生成 AI では、クロスサイトの脆弱性とリモート実行を軽減するために、その出力がサニタイズされ、直接使用されていないことを確認します。これらは、ワークロードで考慮が必要な脆弱性のほんの一部です。すべての AI システムがこれらの脆弱性を持っているわけではありませんが、実際のワークロードに適用されるリスクには細心の注意を払ってください。プレイブックで規定されている修復を検証するために、定期的なテスト、ゲームデー、机上演習を行います。

セキュリティガバナンス

AI ワークロードに関連する役割と責任とともに、セキュリティポリシー、標準、ガイドラインを確立します。

内部および外部でホストされている商用モデル、またはオープンソースモデルを使用するためのポリシーが明確に定義されていることを確認します。同様に、商用の生成 AI モデルを使用する場合、組織の機密データが商用モデルのプラットフォームに漏洩するリスクを考慮してください (「データ保

護機能」を参照)。セキュリティの取り組みの優先順位付けに役立つ、業界や組織に適用される AI に関連する資産、セキュリティリスク、コンプライアンス要件を理解します。特定されたロールに十分なセキュリティリソースを割り当て、可視性を提供します。

AI に関連するリスクは、プライバシーの侵害、データの操作、不正使用、漏洩した意思決定など、広範囲にわたる影響を与える可能性があります。AI 環境の完全性と機密性を保護するために、堅牢な暗号化、多要素認証、継続的なモニタリング、リスク許容やフレームワーク (NIST AI RMF など) との整合性を実装することが重要です。

以下に、ワークロードの 3 つの重要なコンポーネントについての継続的な指示とアドバイスを示します。

- 入力 - データソースと AI の使用を承認できるユーザーを明確にします。データ分類や機密性、データセット内の規制対象データの存在、データの来歴、データの古さ、データを処理する権利など、承認プロセスのデータの側面を考慮してください。リスクを管理するには、ソースの評価、受信方法、保存または保護方法などの要因を考慮して、入力データの調達に使用するメカニズムを評価します。公開されている AI ソリューションでは機密データを処理できないなど、ソースデータのデータ分類がソリューションの分類と一致していることを確認します。
- モデル - モデルの作成とトレーニングの役割と責任を明確にします。モデルリリースの作成者、承認者、パブリッシャーのアプローチに関連する役割を確立します。リスクを管理するには、関連するツールや個人を含むモデルトレーニングメカニズムを評価して、意図的な、または意図しない脆弱性の導入を防ぎます。出力に影響する脆弱性がないかについて、モデルのアーキテクチャを評価します。すべてのモデルの障害モードが、閉鎖、またはセキュアな状態で失敗するようにして、データが公開されないようにします。
- 出力 - 作成された出力のライフサイクル管理を確立します。分類基準を確立し、潜在的に異なるデータセットやデータの分類の結果に細心の注意を払ってください。リスクを管理するには、適切な保護と保持の制御を定め、個人を特定できる情報 (PII) などの重要性と機密性に基づいてデータを分類し、適切なアクセスコントロールを定義します。データ保護コントロールとライフサイクル管理ポリシーを定義します。プライバシー規制やその他のコンプライアンスへの準拠を含む、堅牢なデータ共有プロトコルを確立します。

セキュリティ保証

AI ワークロードの規制およびコンプライアンス要件に対するセキュリティおよびプライバシーの対策を適用、評価、検証します。

組織、およびサービスや製品を提供する顧客には、実装したコントロールに対する信頼と革新が必要です。顧客やユーザーの AI 関連のリスクや潜在的な悪用に対する認識と機密性が高まるにつれ、高

いセキュリティ基準が満たされることを期待するようになります。サイバーセキュリティを優先し、規制要件を満たし、AI に固有のビジネス目標とリスク許容度に沿ったセキュリティリスクを効果的かつ効率的に管理できる方法で、ソリューションの設計、開発、デプロイ、モニタリングを行います。モニタリングを慎重に行い、法律の専門家、コンプライアンスの専門家、データサイエンティスト、情報テクノロジーの専門家の間に透明性とコラボレーションを提供することで、確実性に対する包括的なアプローチの検証に役立ちます。テスト手順と修復プロセスを実装することで、確実性に対する積極的なアプローチが可能になります。ワークロードの 3 つの重要なコンポーネントを継続的にモニタリングして評価します。

- 入力 — モデルではトレーニングと分析に大量のデータが必要になることが多いため、取り込んだデータのタイプがモデルの目標と成果に沿っていることを確認する必要があります。確立された制御フレームワークへの準拠を理解するための監査メカニズムを確立します。
- モデル — 組織のポリシーに従って、ユーザーが AI を許可されている範囲で使用していることを確認します。ポリシーと制御を実装して、組織が AI を使用するのが適切な場所とそうでない場所を理解していることを確認します。監査メカニズムを確立して、モデルがデータをどのように使用しているか、また組織内のどこで AI 機能が使用されているかを特定します。
- 出力 — 出力の許容可能な使用基準を確立し、データが再利用される場所や追加の AI モデルに再導入される可能性がある場所に注意を払ってください。検出または監査メカニズムを確立して出力データを確認し、生成されたデータを機密データまたは規制対象のデータの推測または再作成に使用できないようになっていることを確認します。医療での診断など、信頼度が最優先される出力の信頼性と作成場所を検証するためのメカニズムを作成します。

個々のプライバシーを保持するには、データの不正なアクセス、悪用、開示を防ぐために、法律および法的なガイドラインを厳密に順守する必要があります。AI の可能性のバランスを取り、プライバシー権を尊重することで、社会の信頼を得ることができ、これらの機能の利点を実現できます。保護情報の詳細については、Well-Architected フレームワークの「MLSEC-05: 機密データのプライバシーを保護する」を参照してください。透明性とインフォームドコンセントなどのメカニズムを確立します。データ保持を機能に必要なものだけに制限し、データ共有に関する同意を実装します。考慮すべきワークロードの 3 つの重要なコンポーネントに関連するプライバシー要件は以下のとおりです。

- 入力 — プライバシー関連の規制の対象となるデータ (GDPR、CCPA、COPPA、PDPA など) がどのように使用されるかを理解し、データを処理するための法的根拠が存在するかを検証します。データレジデンシーおよびデータの保存場所または処理場所を考慮します。規制されたデータの使用ごとに、プライバシー影響評価 (PIA) または同様のプロセスを確立します。

- モデル — モデルをトレーニングまたはチューニングする際は、データ処理の法的根拠が存在するかどうか、および対象データの透明性を示すことができるかどうかを考慮します。モデルからの潜在的な漏洩に関連するプライバシー影響評価または同様のプロセスを確立します。
- 出力 — 規制されたデータが追加のモデルのトレーニングに使用されているかどうか、および個人データの二次使用の制限が適用されるかどうかを考慮します。削除権または忘れられる権利タイプのリクエストに応えるための仕組みを確立します。検出または監査メカニズムを確立して出力データを確認し、生成されたデータが匿名化済みのデータの推測または再作成に使用できないようになっていることを確認します。

脅威検知

AI ワークロードにおける潜在的なセキュリティの脅威や予期しない動作を検出して軽減します。

ML または生成 AI システムの 3 つの重要なコンポーネント (入力、モデル、出力) の保護を改善するには、以下のベストプラクティスを使用し、ワークロードへの脅威を検出して軽減します。

- 入力 — AI ソリューションの脅威の検出は、ビジネスに影響を与える可能性のある脆弱性を軽減するために非常に重要です。入力データをサニタイズして、モデルの使用開始時に脅威を検出します。ユーザーセッションの入力データを追跡し続け、可用性や誤用に影響を与える可能性のある脅威を検出して軽減します。
- モデル — AI システムに固有の脅威モデリングと脅威ハンティングの演習を行い、潜在的な脅威を検出して軽減します。脅威モデルとモニタリングを更新して、予期しないユーザー入力を含むトレーニングモデル、コンテンツまたはトレーニングに使用されるデータセットのポイズニング、プライバシー違反、データ改ざんを含む AI の脅威の概念を含めます。入力データとモデルで使用されるデータを相関させ、異常なアクティビティや悪意のあるアクティビティを検出します。
- 出力 — モデルの目標から逸脱した出力異常をモニタリングし、モデル出力内の機密データを検出するためのチェックを有効にします。ワークロードを対象とした特定済みの既知の脅威を含む脅威カタログを構築します。自動テストを作成して、検出機能を検証し、脅威インテリジェンスの統合を検証して有効性を高め、誤検知を低減します。有効性を高め、誤検知を低減する脅威インテリジェンスの使用を検討してください。

インフラストラクチャの保護

AI ワークロードの運用に使用されるシステムとサービスを保護します。

MLOps は AI ワークロードに DevOps プラクティスを使用し、セキュリティは環境全体を構成するインフラストラクチャに適用する必要があります。AI モデルには安全なエンドポイントを使用し、

レート制限モデルアクセスには Amazon API Gateway を使用します。使用するすべての内部および外部 API で API セキュリティのベストプラクティスを使用し、独自の VPC 外のモデルからの API コールの明示的な許可リストを作成します。セキュリティリファレンスアーキテクチャで規定されているセキュリティ機能の使用から始め、環境に基づいてネットワーク、コンピューティング、ストレージのセキュリティ制御を適用します。

モデルは、ネットワークとサーバーにまたがる複数の環境に分散されます。これらの環境間の通信は、転送中の暗号化を使用して保護する必要があります。開発環境と本番環境を一元的に設定し、セキュリティ管理者によって個別に管理される予防ガードレールと検出ガードレールを適用します。モデルトレーニングなどの機密性の高いタスクの開発環境を分離します。エンドユーザーがセッション分離を使用してエクスペリエンスの整合性を保持し、意図しないデータ開示を防止できることを確認します。コンプライアンスおよびトラブルシューティングの目的で、出力レスポンスおよび関連するセッションデータのログを Write Once Read Many (WORM) ストレージデバイスに記録します。セキュリティ問題の原因となる可能性のあるエッジユースケースを発見して軽減するために、モデルの脆弱性報奨金制度を使用することを検討してください。

データ保護

AI の開発と使用におけるデータの可視性、安全なアクセス、制御を維持します。

データ保護は、AI 開発ライフサイクル全体で、またセキュリティガバナンスによって定義されたデータ保護ポリシー (Well-Architected の機械学習レンズに記載されている MLSEC-07: 関連するデータのみ保持するなど) が運用されている場合に、非常に重要なものです。生成 AI 開発に商用モデルを使用する場合は、データをモデルへの入力として直接使用すると、機密情報が公開される可能性があることに注意してください。同様に、独自のモデルや自己ホスト型モデルに保護データへのアクセスを許可すると、データ関連の権限昇格の余地が生じる可能性があります。必要に応じてモデルの使用条件とサービス利用規約を評価します。モデル開発のトレーニング前フェーズと微調整フェーズでモデル開発のために収集されたデータは、転送、保管、使用において保護する必要があります。データトークン化プロセスを使用して、クリーニング、正規化、変換などのデータ前処理フェーズの一環として、機密データを非機密データトークンに置き換えることを検討してください。モデルで使用されるすべてのデータソース、特にモデルのトレーニングに使用される推論データに対して検証可能なメカニズムを作成します。機密データや、機密クラスのエスカレーションにつながる可能性のあるデータのアラートをモニタリングして作成します。データアクティビティのモニタリング手法を採用し、使用や頻度ごとなどでアクセスパターン検出します。機密データを使用してモデルをトレーニングすることは避けてください。これは、モデル出力からデータが意図せず開示される可能性があるためです (推論中のデータ漏洩など)。さまざまな環境でのトレーニングに使用されるデータをタグ付けしてラベル付けし、データタグとラベルをデータ分類ポリシーと標準に沿ったものにします。非本番環境や開発リージョンのデータ系統とデータアクセスが制御され、モデルの脆弱性を引き起こすデー

タ操作が防止されていることを確認します。CI/CD パイプラインを使用して、整合性を維持するために、テスト環境と本番環境にデータを昇格することを検討してください。データアクセスの監査証跡を作成しながら、機密データをログに記録してマスクします。機密データストア、および設計上、指定されたデータクラスのデータ (機密データなど) を保存しないことを想定しているデータストアにデータ損失防止技術を実装し、機密データの意図しない開示をモニタリングします。モデル出力のデータ品質を検証して、信頼性を高め、ハルシネーションを防ぎます。モデルデータ出力の機密レベルをモニタリングし、機密レベルが上昇した場合に、秘匿化または隔離されたレスポンスによる再分類をトリガーします。例えば、新しい入力データセットがモデルによって使用される場合、またはモデルのトレーニングに使用される場合は、出力データが既存の機密性レベルに準拠していることを確認します。

アプリケーションセキュリティ

AI ワークロードのソフトウェア開発ライフサイクルプロセス中に脆弱性を検知して軽減します。

モデル開発者が、プロンプトテストやその他のセキュリティテストケースを自分の環境でローカルに実行し、CI/CD パイプラインでモデルの使用を検証していることを確認します。テストケースライブラリを作成し管理して、カバレッジを検証し、自動化を有効にします。すべての開発、テスト、本番環境のセキュリティスキャンと統合されたデータとモデルパイプラインを活用し、すべてのモデルアーティファクトを安全なリポジトリに保存します。AI モデルのインベントリを管理し、明確に指定した技術所有者とビジネス所有者にモデルインスタンスを割り当てます。既知の優れたトレーニング済みモデルがバックアップされていることを確認します。ポイントインタイムリカバリを保持して、侵害されたモデルが既知の正常な状態に戻ることができるようにします。モデルとデータのバックアップへのアクセスを保護して侵害されていないことを確認し、モデル復旧を定期的にテストして、既知の正常な状態へのフルリカバリを有効にします。出力結果の有効性をサポートするために、パラメータ、メタデータなど、モデルやデータ開発に関連するデータを追跡し、データの来歴を確認します。運用ランブックを作成して使用し、運用上またはセキュリティ上のインシデントが発生した場合に実行できるデータセットとモデルのロールバックメカニズムを個別にテストして、回復性を実現します。

オペレーションのパースペクティブ: AI 環境の正常性と可用性

ML アプリケーションの運用は、多くのお客様にとって初めてのことです。新しい CAF-AI 機能である AI のライフサイクル管理と MLOps では、これに取り組むためのいくつかのパースペクティブとガイダンスを既に紹介しました。すでに説明した内容以外にも、インシデント管理とパフォーマンスについて考慮する必要があります。この CAF-AI の視点をさらに深く掘り下げるには、AWS Well-Architected Framework の「MLOps の成熟度フレームワーク」と「機械学習レンズ」をご覧になる

ことをお勧めします。両方とも、これらの課題に関する豊富なドキュメントとベストプラクティスで構成されています。

基礎的能力	説明
インシデントと問題の管理	予期しない AI の動作を特定して管理します。
パフォーマンスと容量	AI ワークロードのパフォーマンスをモニタリングおよび管理します。
可観測性	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
イベント管理 (AIOps)	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
変更およびリリース管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
設定管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
パッチ管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。
可用性と継続性管理	この能力は AI にとって充実したものではなく、AWS CAF を参照してください。

インシデントと問題の管理

予期しない AI の動作を特定して管理します。

AI システムは、1 人の人物の専門知識だけでは問題を把握または解決するのに十分ではない状況でよく使用されます。AI システムのこのような性質により、システムやエッジケースの一般的な動作を理解することは困難で、時間の経過とともにパフォーマンスが低下する可能性を予測することが難しくなります。このため、実務者はプロキシと簡略化された統計を通じて AI システムを調べます。AI の観察とモニタリングを導入する際、このような AI システムの簡略化されたビューが重要になります。これは開発の初期段階ですでに当てはまりますが、システムを実際の条件下で使用する場合に特に重要です。

AI システムは確認されても検証されないこと、および持続的で継続的な管理とモニタリングが必要であることを認識したプラクティスを確立するようにしてください。その一例が共変量シフトで、ラボで開発された AI システムのパフォーマンスが、本番環境で見られるものとは大きく異なります。必要に応じて、顧客やユーザーが「好ましくない」または「間違っている」と結果にフラグ付けできるようにしてください。顧客やユーザーが直接やり取りしてインシデントを報告するための手段を設けましょう。最初から、ドリフト、共変量シフト、ブラックスワンイベント、観測されていないデータポイントなどを通じて、データ、そしてパフォーマンスの変更に備えます。システムで可能であれば、適切に障害を発生させ、そのようなインシデントを報告して対応し、そこから学ぶ方法を提供します。システムがうまく機能しない顧客やユーザーはデータに表れないことが多いと予想してください。最後に、このようなインシデントは発生するものと想定し、何も報告されない場合は疑わしいと考えてください。この課題は、AI システムの規模と複雑さに比例して大きくなることが予想されます。例えば、基礎モデルは、単純なデシジョンツリーと比較して修正とモニタリングが非常に困難です。

パフォーマンスと容量

AI ワークロードのパフォーマンスをモニタリングおよび管理します。

AI は、従来のソフトウェアとは異なる開発サイクルをたどるため、パフォーマンスとワークロードのプロファイルが異なります。開発の初期段階でデータが調査され、コストとパフォーマンスの点では、非常に多くのさまざまなワークロードへの適応が求められます。このようなワークロードの大半は、多くの場合、強力なマシン、専用のハードウェア、メモリ効率の高いアーキテクチャを必要とする実験やトレーニングのワークロードです。クラウドを使用すると、それぞれがまばらに、開発ライフサイクルの特定の時点でのみ発生するこのようなワークロードのプロファイルに動的に対応する機能が提供されるため、この多数のワークロードを有効にできます。

時間が経つにつれて、トレーニングと合理化された前処理がワークロードプロファイルを引き継いで支配するようになり、一貫性と予測可能性が高まります。イノベーションのスピードは、この新しいプロファイルに適応し、開発と生産の境界を明確にしながら、両者の間を迅速かつ継続的に移行する能力に影響されます。モデルアーティファクトと、このような合理化されたワークロードを促進しているデータが、潜在的なフォールバックに利用できることを確認してください。モデルがデプロイおよび運用段階に移行したら、非機能要件 (レイテンシーやスループットなど) のコストに対して推論が最適化され、パフォーマンスと容量のモニタリングが実施されていることを確認します。AI のライフサイクル管理 機能については、MLOps の成熟度モデルを導入しました。運用に関するより深い洞察を得るには、このリンクを参照してください。時間の経過とともに複数の種類のワークロードプロファイルが混在および混合し、データサイエンティストが立ち上げ前に個別に開発した際のプロファイル (多くの場合、ラボで呼び出されたもの) とはほぼ異なるものになります。Well-Architected-

Framework と、クラウドでそのようなシステムを構築する方法を説明する専用の ML レンズを詳しく見てみましょう。

結論

このドキュメントでは、CAF-AI の概要、顧客が AI ジャーニーを整理して構造化するためのマップ、導入の成功に必要な能力、それらを反復するためのメンタルモデルについて説明しました。このドキュメントで説明した基礎的な能力を、今後の調査、学習、および AI の専門家との会話での参考情報として活用してください。これらの機能はすべて AWS CAF と結びついており、組織がクラウドジャーニーと AI ジャーニーの両方について考えるのに役立ちます。

寄稿者

本ドキュメントの寄稿者は次のとおりです。

- Alexander Wöhlke、上級 ML 戦略担当、生成 AI イノベーションセンター、AWS CAF-AI リード
- Caleb Wilkinson、上級 ML 戦略担当、生成 AI イノベーションセンター、AWS CAF-AI リード
- Payal Vadhani、セキュリティディレクター、プロフェッショナルサービス
- Mayank Jain、シニアマネージャー、プリンシパル、プロフェッショナルサービス
- Michael Sinnwell、上級セキュリティ CDA、プロフェッショナルサービス
- Mark Lieberg、上級セキュリティコンサルタント、プロフェッショナルサービス
- Matias Undurraga、トランスフォーメーションアーキテクト、モダナイゼーションイノベーション トランスフォーメーション
- Tony Santiago、WW パートナーソリューションアーキテクト、CAF プラットフォームパースペクティブリード
- Dr. Saša Baškarada、AWS クラウド 導入フレームワークの世界的リーダー
- Neil Mackin、主任 ML ストラテジスト、機械学習ソリューションラボ
- Shuja Sohrawardy、上級 ML ストラテジスト、生成 AI イノベーションセンター
- Emily Soward、データサイエンティスト、プロフェッショナルサービス
- Margaret Sharp、テクニカルプログラムマネージャー、エンゲージメントセキュリティ、プロフェッショナルサービス
- Ana Echeverri、上級スペシャリスト AI サービス、WW スペシャリストオーガニゼーション、CAF-AI 評価リード
- Phil Le-Brun、企業戦略担当ディレクター

詳細情報

詳細については、次を参照してください。

- AWS クラウド導入フレームワーク (AWS CAF)
- AWS Well-Architected Framework の機械学習レンズ
- AWS Well-Architected
- AWS アーキテクチャセンター
- AWS Prescriptive Guidance
- AWS ホワイトペーパー

ドキュメント履歴

このホワイトペーパーの更新に関する通知を受け取るには、RSS フィードにサブスクライブしてください。

変更	説明	日付
更新	「はじめに」セクションのセキュリティ、プラットフォーム、ガバナンスのパースペクティブの更新および加筆。	2024 年 2 月 13 日
初版発行	ホワイトペーパーの初回発行。	2023 年 5 月 22 日

> **Note**
>
> RSS の更新を購読するには、使用しているブラウザで RSS プラグインを有効にする必要があります。

注意

お客様は、本書に記載されている情報を独自に評価する責任を負うものとします。本書は、(a) 情報提供のみを目的とし、(b) AWS の現行製品と慣行について説明しており、これらは予告なしに変更されることがあり、(c) AWS およびその関連会社、サプライヤー、またはライセンサーからの契約上の義務や保証をもたらすものではありません。AWS の製品やサービスは、明示または黙示を問わず、一切の保証、表明、条件なしに「現状のまま」提供されます。お客様に対する AWS の責任は AWS 契約によって規定されます。本書は、AWS とお客様との間で締結されるいかなる契約の一部でもなく、その内容を修正するものでもありません。

AWS 用語集

AWS の最新の用語については、「AWS の用語集リファレンス」の「AWS 用語集」を参照してください。