AWS Well-Architected Framework

Pilier Efficacité des performances

Pilier Efficacité des performances: AWS Well-Architected Framework

Table of Contents

Pilier Efficacité des performances - AWS Well-Architected Framework

Date de publication : 27 juin 2024 (Révisions du document)

Résumé

Ce livre blanc porte sur le pilier Efficacité des performances du cadre AWS Well-Architected Framework. Le but de ce document est de fournir des conseils qui aident les clients à utiliser efficacement les ressources du cloud pour répondre aux besoins de leur entreprise, et à maintenir cette efficacité à mesure que la demande et les technologies évoluent.

Introduction

La AWS Well-Architected Framework vous aide à comprendre les avantages et les inconvénients des décisions que vous prenez lors de la création de charges de travail sur AWS. En utilisant ce cadre, vous apprenez les bonnes pratiques architecturales en matière de conception et d'exploitation de charges de travail fiables, sécurisées, efficaces, économiques et durables dans le cloud. Il vous permet d'évaluer systématiquement vos architectures par rapport aux bonnes pratiques et d'identifier les domaines à améliorer. Nous pensons que le fait d'avoir des charges de travail bien structurées augmente considérablement les chances de réussite métier.

Le cadre repose sur six piliers :

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité des performances
- Optimisation des coûts
- Durabilité

Ce livre blanc porte sur l'application des principes du pilier Efficacité des performances à vos charges de travail. Dans les environnements sur site traditionnels, il est difficile de bénéficier de performances élevées et durables. En appliquant les principes de ce livre blanc, vous pourrez créer

des architectures sur AWS qui fournissent avec efficacité des performances soutenues sur le long terme. Les conseils et les bonnes pratiques présentés dans ce document sont répartis dans cinq domaines clés qui servent de principes directeurs pour la création de solutions cloud performantes sur AWS. Ces domaines d'intérêt sont les suivants :

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Mise en réseau et diffusion de contenu
- Processus et culture

Le présent document est conçu pour ceux et celles qui sont dépositaires de rôles technologiques, comme les directeurs de la technologie, les architectes, les développeurs et les membres de l'équipe d'exploitation. Après avoir lu ce document, vous allez vous familiariser avec les bonnes pratiques et les stratégies d'AWS à utiliser lors de la conception d'architectures cloud performantes.

Efficacité en matière de performance

Le pilier Efficacité des performances se concentre sur l'utilisation efficace des ressources de calcul pour répondre aux exigences et sur la façon de maintenir cette efficacité à mesure que la demande change et que les technologies évoluent.

Rubriques

Principes de conception

Les principes de conception suivants peuvent vous aider à créer des charges de travail efficaces dans le cloud, tout en veillant à ce qu'elles le restent dans la durée.

- Démocratiser les technologies avancées : Facilitez la mise en œuvre de technologies avancées pour votre équipe en déléguant les tâches complexes à votre fournisseur de cloud. Plutôt que de demander à votre équipe informatique de s'informer sur l'hébergement et l'exploitation de nouvelles technologies, envisagez de consommer les technologie en tant que service. Par exemple, les bases de données NoSQL, le transcodage multimédia et le machine learning sont trois technologies qui requièrent des compétences spécialisées. Dans le cloud, ces technologies deviennent des services que votre équipe peut consommer, ce qui lui permet de se dédier au développement de produits plutôt qu'à l'allocation et à la gestion des ressources.
- Envergure mondiale en quelques minutes : En déployant votre charge de travail dans plusieurs régions AWS à travers le monde, vous pouvez offrir une latence plus faible et une meilleure expérience à vos clients à un coût minimal.
- Utiliser des architectures sans serveur : Grâce aux architectures sans serveur, vous n'avez plus besoin de faire fonctionner et de gérer des serveurs physiques pour les activités de calcul traditionnelles. Par exemple, les services de stockage sans serveur peuvent agir comme des sites Web statiques (éliminant le besoin de serveurs Web), et les services d'événements peuvent héberger du code. Ainsi, vous supprimez la charge opérationnelle de gestion des serveurs physiques et réduisez les coûts des transactions, car les services gérés fonctionnent à l'échelle du cloud.

- Expérimenter plus fréquemment : Avec des ressources virtuelles et automatisables, vous pouvez rapidement exécuter des tests comparatifs à l'aide de différents types d'instances, de stockages ou de configurations.
- Envisager la « sympathie mécanique » : utilisez l'approche technologique qui correspond le mieux à vos objectifs. Par exemple, tenez compte des modèles d'accès aux données lorsque vous sélectionnez les approches de stockage ou de base de données de votre charge de travail.

Définition

Concentrez-vous sur les domaines suivants pour assurer l'efficacité des performances dans le cloud :

- Choix d'architecture
- Informatique et matériel
- Gestion des données
- Mise en réseau et diffusion de contenu
- Processus et culture

Adoptez une approche axée sur les données pour créer une architecture performante. Collectez des données sur tous les aspects de l'architecture, depuis la conception générale jusqu'à la sélection et la configuration des types de ressources.

En réexaminant vos choix régulièrement, vous faites en sorte de tirer parti de l'évolution constante du cloud AWS. La surveillance vous offre la garantie d'être informé de tout écart par rapport aux performances attendues. Effectuer des compromis dans votre architecture pour améliorer les performances, comme l'utilisation de la compression, la mise en cache ou l'abaissement des exigences de cohérence.

Choix d'architecture

La solution optimale pour une charge de travail peut varier, et les solutions combinent souvent plusieurs approches. Les charges de travail bien architecturées utilisent plusieurs solutions et permettent d'exploiter différentes fonctionnalités pour améliorer les performances.

De nombreux types et configurations de ressources AWS sont proposés. Il est ainsi plus facile de trouver l'approche qui correspond le mieux à vos besoins. Vous pouvez également rechercher des options qui ne sont pas facilement accessibles avec une infrastructure sur site. Par exemple, un service géré tel que Amazon DynamoDB fournit une base de données NoSQL entièrement gérée avec une latence de moins de dix millisecondes, quelle que soit l'échelle.

Ce domaine d'intérêt partage des conseils et des bonnes pratiques sur la manière de sélectionner des ressources cloud et des modèles d'architecture efficaces et performants.

Bonnes pratiques

- PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles
- PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques
- PERF01-BP03 Tenir compte des coûts dans vos décisions architecturales
- PERF01-BP04 Évaluer l’impact des compromis sur les clients et l’efficacité de l’architecture
- PERF01-BP05 Utiliser des stratégies et des architectures de référence
- PERF01-BP06 Utiliser le benchmarking pour éclairer vos décisions architecturales
- PERF01-BP07 Utiliser une approche orientée données pour les choix architecturaux

PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles

Découvrez en continu les services et configurations disponibles qui vous aident à prendre de meilleures décisions architecturales et à améliorer l’efficacité des performances de votre architecture de charge de travail.

Anti-modèles courants :

- Vous utilisez le cloud comme centre de données hébergé.

- Vous ne modernisez pas votre application après la migration vers le cloud.
- Vous n'utilisez qu'un seul type de stockage pour tout ce que vous devez conserver.
- Vous utilisez les types d'instances qui correspondent le plus à vos standards actuels. Elles peuvent être de plus grande taille au besoin.
- Vous déployez et gérez les technologies disponibles en tant que services gérés.

Avantages liés au respect de cette bonne pratique : En envisageant de nouveaux services et de nouvelles configurations, vous pourriez être en mesure d'améliorer considérablement vos performances, de réduire les coûts et d'optimiser les efforts requis pour maintenir votre charge de travail. Elle peut également vous aider à accélérer le délai de valorisation des produits compatibles avec le cloud.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS publie en permanence de nouveaux services et fonctionnalités susceptibles d'améliorer les performances et de réduire le coût des charges de travail dans le cloud. Il est essentiel de se tenir informé de ces nouveaux services et fonctionnalités pour maintenir l'efficacité des performances dans le cloud. La modernisation de votre architecture de charge de travail vous permet également d'accélérer la productivité, de stimuler l'innovation et de générer de nouvelles opportunités de croissance.

Étapes d'implémentation

- Faites l'inventaire de vos charges de travail logicielles et de l'architecture des services connexes. Déterminez la catégorie de produits sur laquelle vous souhaitez en savoir plus.
- Explorez les offres AWS pour identifier et découvrir les services et les options de configuration pertinents qui peuvent vous aider à améliorer les performances et à réduire les coûts et la complexité opérationnelle.
  - Amazon Web Services Cloud
  - AWS Academy
  - Nouveautés avec AWS
  - Blog AWS
  - AWS Skill Builder
  - Événements et webinaires AWS

  - AWS Training et certifications
  - Chaîne YouTube AWS
  - Ateliers AWS
  - Communautés AWS
- Utilisez des environnements de test (hors production) pour découvrir et tester de nouveaux services sans frais supplémentaires.
- Découvrez en permanence les nouveaux services et fonctionnalités du cloud.

Ressources

Documents connexes :

- Vue d’ensemble d’Amazon Web Services
- Fonctionnalités d'Amazon EC2
- Apprenez étape par étape grâce à un Plan de formation pour les partenaires AWS
- AWS Training and Certification
- Mon parcours d’apprentissage pour devenir architecte de solutions AWS
- Centre d’architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Créer des applications modernes sur AWS

Vidéos connexes :

- AWS re:Invent 2023 - What’s new with Amazon EC2
- AWS re:Invent 2022 - Reduce your operational and infrastructure costs with Amazon ECS
- AWS re:Invent 2023 - Build with the efficiency, agility & innovation of the cloud with AWS
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- This is my Architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques

Utilisez les ressources cloud de l'entreprise, telles que la documentation, les architectes de solutions, les services professionnels ou les partenaires appropriés pour éclairer vos décisions architecturales. Ces ressources vous aident à vérifier et à améliorer votre architecture pour obtenir des performances optimales.

Anti-modèles courants :

- Vous utilisez AWS en tant que fournisseur de cloud ordinaire.
- Vous utilisez les services AWS de manière non conforme à leur utilisation prévue.
- Vous suivez toutes les recommandations sans tenir compte du contexte de votre entreprise.

Avantages liés au respect de cette bonne pratique : En suivant les recommandations d'un fournisseur de cloud ou d'un partenaire approprié, vous pouvez faire les bons choix architecturaux pour votre charge de travail et vous avez confiance dans vos décisions.

Niveau de risque exposé si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

AWS propose un large éventail de recommandations, documentations et ressources qui peuvent vous aider à générer et à gérer des charges de travail cloud efficaces. La documentation AWS fournit des exemples de code, des tutoriels et des explications détaillées sur les services. Outre la documentation, AWS propose des programmes de formation et de certification, des architectes de solutions et des services professionnels qui peuvent aider les clients à explorer différents aspects des services cloud et à mettre en œuvre une architecture cloud efficace sur AWS.

Tirez parti de ces ressources pour obtenir des informations précieuses et des bonnes pratiques, gagner du temps et obtenir de meilleurs résultats dans le AWS Cloud.

Étapes d'implémentation

- Consultez la documentation et les recommandations AWS et suivez les bonnes pratiques. Ces ressources peuvent vous aider à choisir et à configurer efficacement les services, ainsi qu'à améliorer les performances.
  - documentation AWS (comme les guides d'utilisation et les livres blancs)
  - Blog AWS
  - AWS Training et certifications
  - Chaîne YouTube AWS
- Participez à des événements partenaires AWS (tels que les sommets mondiaux AWS, les groupes d'utilisateurs, re:Invent AWS et les ateliers) pour découvrir les bonnes pratiques d'utilisation des services AWS auprès des experts AWS.
  - Événements et webinaires AWS
  - Ateliers AWS
  - Communautés AWS
- Contactez AWS pour obtenir de l'aide lorsque vous avez besoin de conseils ou d'informations supplémentaires sur le produit. Les architectes de solutions AWS et les services professionnels AWS fournissent des conseils pour la mise en œuvre de solutions. les partenaires AWS apportent une expertise AWS pour vous aider à gagner en agilité et favoriser l'innovation au sein de votre entreprise.
- Utilisez AWS Support si vous avez besoin d'une assistance technique pour utiliser un service de manière efficace. Nos plans de support sont conçus pour vous fournir la bonne combinaison d'outils et l'accès à une expertise afin que vous puissiez réussir avec AWS tout en optimisant les performances, en gérant les risques et en maîtrisant les coûts.

Ressources

Documents connexes :

- Centre d'architecture AWS
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Support aux entreprises AWS

Vidéos connexes :

- Voici mon architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP03 Tenir compte des coûts dans vos décisions architecturales

Tenez compte des coûts dans vos décisions architecturales afin d'améliorer l'utilisation des ressources et l'efficacité des performances de votre charge de travail cloud. Lorsque vous êtes conscient des implications financières de votre charge de travail cloud, vous êtes plus susceptible de tirer parti de ressources efficaces et de réduire les pratiques inutiles.

Anti-modèles courants :

- Vous n'utilisez qu'une seule famille d'instances.
- Vous n'évaluez pas les solutions sous licence par rapport aux solutions open source.
- Vous ne définissez pas de stratégies de cycle de vie pour le stockage.
- Vous ne passez pas en revue les nouveaux services et les nouvelles fonctionnalités du AWS Cloud.
- Vous utilisez uniquement le stockage par blocs.

Avantages liés au respect de cette bonne pratique : En tenant compte des coûts dans vos prises de décision, vous pouvez utiliser des ressources plus efficaces et explorer d'autres investissements.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

L'optimisation des charges de travail en termes de coûts peut améliorer l'utilisation des ressources et éviter le gaspillage dans une charge de travail cloud. La prise en compte des coûts dans les décisions architecturales implique généralement de dimensionner correctement les composants de

la charge de travail et de renforcer l'élasticité, ce qui se traduit par une amélioration de l'efficacité des performances de la charge de travail cloud.

Étapes d'implémentation

- Fixez des objectifs de coûts tels que des limites budgétaires pour votre charge de travail cloud.
- Identifiez les composants clés (tels que les instances et le stockage) qui augmentent le coût de votre charge de travail. Vous pouvez utiliser AWS Pricing Calculator et AWS Cost Explorer pour identifier les principaux facteurs de coûts dans votre charge de travail.
- Comprenez les modèles de tarification dans le cloud, par exemple les instances à la demande, réservées, Savings Plans et Spot.
- Utilisez les Bonnes pratiques d'optimisation des coûts Well-Architected afin d'optimiser ces composants clés en termes de coûts.
- Surveillez et analysez en permanence les coûts afin d'identifier les opportunités d'optimisation des coûts dans votre charge de travail.
  - Utilisez Budgets AWS pour recevoir des alertes en cas de coûts inadmissibles.
  - Utilisez AWS Compute Optimizer ou AWS Trusted Advisor pour obtenir des recommandations en matière d'optimisation des coûts.
  - Utilisez Cost Anomaly Detection AWS pour détecter automatiquement les anomalies de coûts et analyser les causes racines.

Ressources

Documents connexes :

- Qu'est-ce que la solution Facturation et gestion des coûts AWS ?
- Optimisation des coûts avec AWS
- Choix d'une stratégie de gestion des coûts AWS
- Guide de gestion des coûts AWS pour les débutants
- Présentation détaillée du tableau de bord Cost Intelligence Dashboard
- Centre d'architecture AWS
- Bibliothèque de solutions AWS
- Centre de connaissances AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2023 - What's new with AWS cost optimization
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2023 - Optimize costs in your multi-account environments

Exemples connexes :

- Code de démonstration AWS Compute Optimizer
- Atelier d'optimisation des coûts
- Playbooks de mise en œuvre technique de la gestion financière dans le cloud
- Optimisation du démarrage : ajustement des performances des applications pour une efficacité maximale
- Atelier d'optimisation sans serveur (performances et coûts)
- Mise à l'échelle d'architectures rentables

PERF01-BP04 Évaluer l'impact des compromis sur les clients et l'efficacité de l'architecture

Lors de l'évaluation des améliorations liées à la performance, identifiez les choix qui affectent vos clients et l'efficacité de la charge de travail. Par exemple, si l'utilisation d'un magasin de données clé-valeur augmente les performances du système, il est important d'évaluer l'impact de la nature constante de cette modification à terme sur les clients.

Anti-modèles courants :

- Vous supposez que tous les gains de performances doivent être mis en œuvre, même s'il existe des compromis en termes d'implémentation.
- Vous n'évaluez les modifications apportées aux charges de travail que lorsqu'un problème de performances a atteint un point critique.

Avantages liés au respect de cette bonne pratique : Lorsque vous évaluez les améliorations potentielles liées aux performances, vous devez décider si les compromis concernant les modifications sont compatibles avec les exigences de charge de travail. Dans certains cas, vous devrez peut-être mettre en place des contrôles supplémentaires pour compenser les compromis.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Identifiez les domaines critiques de votre architecture en termes de performances et d'impact sur les clients. Déterminez la façon dont vous pouvez apporter des améliorations ainsi que les compromis que ces améliorations entraînent et la façon dont ils affectent le système et l'expérience de l'utilisateur. Par exemple, la mise en œuvre de la mise en cache des données permet d'améliorer de manière significative les performances, mais nécessite une stratégie précise concernant la manière et le moment où mettre à jour ou invalider les données mises en cache pour empêcher un comportement incorrect du système.

Étapes d'implémentation

- Comprenez vos exigences en matière de charge de travail et vos SLA.
- Définissez clairement les facteurs d'évaluation. Les facteurs peuvent être liés au coût, à la fiabilité, à la sécurité et aux performances de votre charge de travail.
- Sélectionnez l'architecture et les services qui répondent à vos besoins.
- Menez des expériences et des démonstrations de faisabilité (POC) afin d'évaluer les facteurs de compromis et l'impact sur les clients et l'efficacité de l'architecture. En général, les charges de travail hautement disponibles, performantes et sécurisées consomment davantage de ressources cloud tout en offrant une meilleure expérience client. Comprenez les compromis entre la complexité, les performances et les coûts de votre charge de travail. Généralement, la priorisation de deux des facteurs se fait au détriment du troisième.

Ressources

Documents connexes :

- Bibliothèque Amazon Builders' Library
- KPI Amazon QuickSight
- Amazon CloudWatch RUM

- Documentation X-Ray
- Comprenez les modèles de résilience et les compromis pour concevoir une architecture efficace dans le cloud

Vidéos connexes :

- Optimize applications through Amazon CloudWatch RUM
- AWS re:Invent 2023 - Capacity, availability, cost efficiency: Pick three
- AWS re:Invent 2023 - Advanced integration patterns & trade-offs for loosely coupled systems

Exemples connexes :

- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics
- Client web Amazon CloudWatch RUM

PERF01-BP05 Utiliser des stratégies et des architectures de référence

Utilisez les stratégies internes et les architectures de référence existantes lors de la sélection des services et des configurations en vue d'augmenter votre efficacité lorsque vous concevez et mettez en œuvre votre charge de travail.

Anti-modèles courants :

- Vous autorisez un large éventail de technologies qui peuvent avoir un impact sur les frais généraux de gestion de votre entreprise.

Avantages liés au respect de cette bonne pratique : L'établissement d'une stratégie pour les choix d'architecture, de technologie et de fournisseur permet de prendre des décisions rapidement.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Le fait de disposer de politiques internes en matière de sélection des ressources et de l'architecture fournit des normes et des directives à suivre lors des choix architecturaux. Ces directives simplifient

le processus de prise de décision lors du choix du bon service cloud et peuvent contribuer à améliorer l'efficacité des performances. Déployez votre charge de travail à l'aide de stratégies ou d'architectures de référence. Intégrez les services à votre déploiement dans le cloud. Utilisez ensuite vos tests de performance pour vérifier que vous pouvez continuer à répondre à vos exigences de performance.

Étapes d'implémentation

- Comprenez clairement les exigences de votre charge de travail cloud.
- Passez en revue les stratégies internes et externes pour identifier les plus pertinentes.
- Utilisez les architectures de référence appropriées fournies par AWS ou les bonnes pratiques de votre secteur.
- Créez un continuum composé de stratégies, de normes, d'architectures de référence et de directives normatives pour les situations courantes. Vos équipes pourront ainsi agir plus rapidement. Adaptez les ressources à votre secteur d'activité, le cas échéant.
- Validez ces stratégies et architectures de référence pour votre charge de travail dans les environnements de test.
- Restez informé des normes du secteur et des mises à jour AWS pour garantir que vos stratégies et architectures de référence contribuent à optimiser votre charge de travail cloud.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- AWS Blog Architecture

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2022 - Accelerate value for your business with SAP & AWS reference architecture

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

PERF01-BP06 Utiliser le benchmarking pour éclairer vos décisions architecturales

Définissez des points de référence pour les performances d'une charge de travail existante afin de comprendre ses performances sur le cloud et prendre des décisions architecturales sur la base de ces données.

Anti-modèles courants :

- Vous comptez sur des points de référence courants qui ne reflètent pas les caractéristiques de votre charge de travail.
- Vous utilisez les commentaires et la perception des clients comme seule référence.

Avantages liés au respect de cette bonne pratique : le benchmarking de votre implémentation actuelle vous permet de mesurer l'amélioration des performances.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Utilisez la définition de points de référence avec des tests synthétiques pour évaluer les performances des composants de votre charge de travail. La définition de points de référence est généralement plus rapide à configurer que les tests de charge. Elle est utilisée pour évaluer la technologie pour un composant en particulier. La définition de points de référence est souvent utilisée au début d'un nouveau projet, lorsque vous n'avez pas de solution complète pour le test de charge.

Vous pouvez créer vos propres tests d'évaluation personnalisés ou utiliser un test standard, tel que TPC-DS, pour évaluer vos charges de travail. Les points de référence du secteur sont utiles lorsque vous comparez différents environnements. Les points de référence personnalisés sont utiles pour cibler certains types d'opérations que vous souhaitez effectuer dans votre architecture.

Avec le benchmarking, il est important de préparer votre environnement de test pour obtenir des résultats valides. Exécutez plusieurs fois le même point de référence pour vous assurer d'avoir capturé toute variabilité au fil du temps.

Étant donné que les points de référence sont généralement plus rapides à exécuter que les tests de charge, ils peuvent être utilisés plus tôt dans le pipeline de déploiement et fournir un retour rapide sur les écarts de performances. Lorsque vous évaluez un changement important dans un composant ou un service, un point de référence peut être un moyen rapide pour voir si la modification a un intérêt. L'utilisation de la définition de points de référence avec un test de charge est essentielle, car un test de charge vous indique comment votre charge de travail se comporte dans un environnement de production.

Étapes d'implémentation

- Planifiez et définissez :
  - Définissez les objectifs, la base de référence, les scénarios de test, les métriques (telles que l'utilisation du CPU, la latence ou le débit) et les indicateurs de performance clés de votre test d'évaluation.
  - Concentrez-vous sur les exigences des utilisateurs en termes d'expérience utilisateur et sur des facteurs tels que le temps de réponse et l'accessibilité.
  - Identifiez un outil de benchmarking adapté à votre charge de travail. Vous pouvez utiliser des services AWS, tels qu'Amazon CloudWatch, ou un outil tiers compatible avec votre charge de travail.
- Configurez et instrumentez :
  - Configurez votre environnement et configurez vos ressources.
  - Mettez en œuvre la surveillance et la journalisation pour capturer les résultats des tests.
- Comparez et surveillez :
  - Effectuez vos tests comparatifs et surveillez les métriques pendant le test.
- Analysez et documentez :
  - Documentez votre processus de benchmarking et vos résultats.
  - Analysez les résultats pour identifier les goulots d'étranglement, les tendances et les domaines d'amélioration.
  - Utilisez les résultats des tests pour prendre des décisions architecturales et ajuster votre charge de travail. Cet ajustement peut impliquer la modification des services ou l'adoption de nouvelles fonctionnalités.

- Optimisez et répétez :
  - Ajustez les configurations et les allocations des ressources en fonction de vos critères de référence.
  - Testez à nouveau votre charge de travail après ajustement pour valider vos améliorations.
  - Documentez vos conclusions et répétez le processus pour identifier d'autres domaines d'amélioration.

Ressources

Documents connexes :

- Centre d'architecture AWS
- AWS Partner Network
- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Genomics workflows, Part 5: automated benchmarking
- Benchmark and optimize endpoint deployment in Amazon SageMaker JumpStart

Vidéos connexes :

- AWS re:Invent 2023 - Benchmarking AWS Lambda cold starts
- Benchmarking stateful services in the cloud
- Voici mon architecture
- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Exemples connexes :

- Exemples AWS (langue française non garantie)
- Exemples de kits SDK AWS (langue française non garantie)
- Tests de charge distribuée

- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics (langue française non garantie)
- Client Web Amazon CloudWatch RUM (langue française non garantie)

PERF01-BP07 Utiliser une approche orientée données pour les choix architecturaux

Définissez une approche orientée données claire pour les choix architecturaux afin de vérifier que les services et configurations cloud appropriés sont utilisés pour répondre aux besoins spécifiques de votre entreprise.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne devrait pas être mise à jour au fil du temps.
- Vos choix architecturaux sont basés sur des suppositions et des hypothèses.
- Vous introduisez des modifications d'architecture au fil du temps sans justification.

Avantages liés au respect de cette bonne pratique : En adoptant une approche bien définie pour les choix architecturaux, vous utilisez les données pour influencer la conception de votre charge de travail et prendre des décisions éclairées au fil du temps.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Mobilisez l'expérience et l'expertise des ressources cloud internes ou faites appel à des ressources externes, comme des cas d'utilisation publiés ou des livres blancs pour définir un processus de sélection des ressources et services dans votre architecture. Vous devriez disposer d'un processus bien défini qui encourage l'expérimentation et le benchmarking avec les services qui pourraient être utilisés dans votre charge de travail.

Les backlogs relatifs aux charges de travail critiques doivent non seulement comprendre des témoignages d'utilisateurs proposant des fonctionnalités pertinentes pour les entreprises et les utilisateurs, mais également des récits techniques qui constituent une piste architecturale pour la charge de travail. Cette piste s'inspire des nouvelles avancées technologiques et des nouveaux

services et les adopte sur la base de données et de justifications appropriées. Cela permet de vérifier que l'architecture reste pérenne et ne stagne pas.

Étapes d'implémentation

- Collaborez avec les principales parties prenantes pour définir les exigences en matière de charge de travail, y compris les considérations relatives aux performances, à la disponibilité et aux coûts. Tenez compte de facteurs tels que le nombre d'utilisateurs et le modèle d'utilisation de votre charge de travail.
- Créez une piste architecturale ou un backlog technologique qui est axé en priorité sur le backlog fonctionnel.
- Évaluez les différents services cloud (pour en savoir plus, consultez PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles).
- Explorez les différents modèles architecturaux, tels que les microservices ou le modèle sans serveur, qui répondent à vos exigences en termes de performances (pour en savoir plus, consultez PERF01-BP02 Utiliser les recommandations de votre fournisseur de cloud ou d'un partenaire approprié pour en savoir plus sur les modèles d'architecture et les bonnes pratiques).

- Consultez d'autres équipes, des diagrammes d'architecture et des ressources, tels que les architectes de solutions AWS, Centre d'architecture AWSet AWS Partner Networkpour vous aider à choisir l'architecture adaptée à votre charge de travail.

- Définissez des métriques de performances telles que le débit et le temps de réponse qui peuvent vous aider à évaluer les performances de votre charge de travail.
- Testez et utilisez des métriques définies pour valider les performances de l'architecture sélectionnée.
- Surveillez en continu les performances et effectuez les ajustements nécessaires pour maintenir un niveau optimal de performance pour votre architecture.
- Documentez l'architecture que vous avez sélectionnée et les décisions que vous avez prises comme référence pour les futures mises à jour et les futurs apprentissages.
- Vérifiez en permanence l'approche de sélection de l'architecture et mettez-la à jour en fonction des apprentissages, des nouvelles technologies et des métriques indiquant un changement nécessaire ou un problème dans l'approche actuelle.

Ressources

Documents connexes :

- Bibliothèque de solutions AWS
- Centre de connaissances AWS
- Architectural Patterns to Build End-to-End Data Driven Applications on AWS

Vidéos connexes :

- This is my Architecture
- AWS re:Invent 2021 - Data-driven enterprise: Going from vision to value
- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results

Exemples connexes :

- Exemples AWS
- Exemples de kits SDK AWS

Informatique et matériel

Le choix d'une solution de calcul optimale pour une charge de travail particulière peut varier selon la conception de l'application, les modèles d'utilisation et les paramètres de configuration. Les architectures peuvent utiliser différentes solutions de calcul pour divers composants et permettent différentes fonctionnalités pour améliorer les performances. Choisir une solution de calcul inadaptée pour une architecture peut nuire à ses performances.

Ce domaine d'intérêt partage des conseils et des bonnes pratiques sur la manière d'identifier et d'optimiser les options de calcul pour l'efficacité des performances dans le cloud.

Bonnes pratiques

- PERF02-BP01 Sélectionner les meilleures options de calcul pour votre charge de travail
- PERF02-BP02 Comprendre les configurations et les fonctionnalités de calcul disponibles
- PERF02-BP03 Collecter les métriques liées au calcul
- PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul
- PERF02-BP05 Mettre à l’échelle vos ressources de calcul de manière dynamique
- PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

PERF02-BP01 Sélectionner les meilleures options de calcul pour votre charge de travail

La sélection de l’option de calcul la mieux adaptée à votre charge de travail vous permet d’améliorer les performances, de réduire les coûts d’infrastructure inutiles et de diminuer les efforts opérationnels nécessaires pour maintenir votre charge de travail.

Anti-modèles courants :

- Vous utilisez la même option de calcul que celle utilisée sur site.
- Vous manquez de connaissances sur les options, les fonctionnalités et les solutions de calcul cloud et sur la manière dont elles pourraient améliorer vos performances de calcul.
- Vous surprovisionnez une option de calcul existante pour répondre aux exigences de mise à l’échelle ou de performances, alors qu’une autre option de calcul s’alignerait plus précisément sur les caractéristiques de votre charge de travail.

Avantages liés au respect de cette bonne pratique : en identifiant les exigences de calcul et en les comparant aux options disponibles, vous pouvez optimiser votre charge de travail en termes de ressources.

Niveau de risque exposé si cette bonne pratique n'est pas établie: élevé

Directives d'implémentation

Pour optimiser vos charges de travail cloud afin d'améliorer l'efficacité des performances, il est important de sélectionner les options de calcul les mieux adaptées à votre cas d'utilisation et à vos exigences de performances. AWS fournit une variété d'options de calcul qui sont adaptées aux différentes charges de travail dans le cloud. Par exemple, vous pouvez utiliser Amazon EC2 pour lancer et gérer des serveurs virtuels, AWS Lambda pour exécuter du code sans avoir à provisionner ou à gérer de serveurs, Amazon ECS ou Amazon EKS pour exécuter et gérer des conteneurs ou encore AWS Batch pour traiter d'importants volumes de données en parallèle. En fonction de vos besoins en termes de mise à l'échelle et de calcul, vous devez choisir et configurer la solution de calcul optimale pour votre situation. Vous pouvez également envisager d'utiliser plusieurs types de solutions de calcul dans une seule charge de travail, car chacune présente ses avantages et ses inconvénients.

Les étapes suivantes vous guident dans la sélection des options de calcul adaptées aux caractéristiques de votre charge de travail et à vos exigences de performances.

Étapes d'implémentation

- Comprenez les exigences de calcul de votre charge de travail. Les exigences clés à prendre en compte incluent les besoins de traitement, les modèles de trafic, les modèles d'accès aux données, les besoins de mise à l'échelle et les exigences de latence.
- Découvrez les différentes options de calcul disponibles pour votre charge de travail sur AWS (comme décrit dans PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles). Voici quelques options de calcul AWS clés, leurs caractéristiques et leurs cas d'utilisation courants :

AWS service	Key characteristics	Common use cases
Amazon Elastic Compute Cloud (Amazon EC2)	Has dedicated option for hardware, license requirements, large selection of different	Lift and shift migrations, monolithic application, hybrid

AWS service	Key characteristics	Common use cases
	instance families, processor types and compute accelerators	environments, enterprise applications
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Easy deployment, consistent environments, scalable	Microservices, hybrid environments
AWS Lambda	Calcul sans serveur service that runs code in response to events and automatically manages the underlying compute resources.	Microservices, event-driven applications
AWS Batch	Efficiently and dynamically provisions and scales Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), and AWS Fargate compute resources, with an option to use On-Demand or Spot Instances based on your job requirements	HPC, train ML models
Amazon Lightsail	Preconfigured Linux and Windows application for running small workloads	Simple web applications, custom website

- Évaluez les coûts (tels que le tarif horaire ou le transfert de données) et les frais de gestion (tels que l'application de correctifs et la mise à l'échelle) associés à chaque option de calcul.
- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier l'option de calcul la mieux adaptée à vos exigences en termes de charge de travail.

- Après avoir testé et identifié votre nouvelle solution de calcul, planifiez votre migration et validez vos métriques de performance.
- Utilisez les outils de surveillance AWS tels qu'Amazon CloudWatch et les services d'optimisation tels qu'AWS Compute Optimizer pour optimiser en continu vos ressources de calcul en fonction de modèles d'utilisation réels.

Ressources

Documents connexes :

- Calcul sur le cloud avec AWS
- Types d'instances Amazon EC2
- Conteneurs Amazon EKS : composants master Amazon EKS
- Conteneurs Amazon ECS : instances de conteneur Amazon ECS
- Fonctions : configuration des fonctions Lambda
- Conseils prescriptifs pour les conteneurs
- Conseils prescriptifs pour les modèles sans serveur

Vidéos connexes :

- AWS re:Invent 2023 - AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 - New Amazon Elastic Compute Cloud generative AI capabilities in AMS
- AWS re:Invent 2023 - What's new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Powering next-gen Amazon Elastic Compute Cloud: Deep dive on the Nitro System
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud foundations
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Déployez des modèles de ML à des fins d'inférence avec des performances élevées et à faible coût

Exemples connexes :

- Migrer l'application Web vers des conteneurs (langue française non garantie)
- Exécuter un modèle Hello World sans serveur
- Atelier Amazon EKS
- Atelier Amazon EC2
- Efficient and Resilient Workloads with Amazon Elastic Compute Cloud Auto Scaling
- Migrating to AWS Graviton with Container Services

PERF02-BP02 Comprendre les configurations et les fonctionnalités de calcul disponibles

Découvrez les options et les fonctionnalités de configuration disponibles pour votre service de calcul qui vous aideront à allouer la quantité de ressources appropriée et à améliorer l'efficacité des performances.

Anti-modèles courants :

- Vous ne comparez pas les options de calcul ni les familles d'instances disponibles avec les caractéristiques de la charge de travail.
- Vous surprovisionnez les ressources de calcul pour répondre aux pics de demande.

Avantages liés au respect de cette bonne pratique : Familiarisez-vous avec les fonctionnalités et les configurations de calcul d'AWS pour pouvoir utiliser une solution de calcul optimisée qui répond aux caractéristiques et aux besoins de votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Chaque solution de calcul dispose de configurations et de fonctionnalités uniques pour prendre en charge différentes caractéristiques et exigences de charge de travail. Découvrez comment ces options soutiennent votre charge de travail et déterminez celles qui sont optimales pour votre système. Parmi ces options, citons, par exemple la famille d'instances, les tailles, les fonctionnalités (GPU, E/S), la capacité de débordement (bursting), les délais d'attente, les tailles de fonction, les instances de conteneur et la simultanéité. Si votre charge de travail utilise la même option de calcul

depuis plus de quatre semaines et que vous anticipez que les caractéristiques resteront les mêmes à l'avenir, vous pouvez utiliser AWS Compute Optimizer pour déterminer si votre option de calcul actuelle est adaptée aux charges de travail du point de vue du processeur et de la mémoire.

Étapes d'implémentation

1. Comprenez les exigences de la charge de travail (comme les besoins en UC, la mémoire et la latence).
2. Consultez la documentation AWS et les bonnes pratiques pour en savoir plus sur les options de configuration recommandées qui peuvent vous aider à améliorer vos performances de calcul. Voici quelques options de configuration clés à prendre en compte :

Option de configuration	Exemples
Type d'instance	• Les instances optimisées pour le calcul sont idéales pour les charges de travail qui exigent un ratio vCPU/mémoire plus élevé. • Les instances optimisées pour la mémoire offrent de grandes quantités de mémoire pour soutenir les charges de travail gourmandes en mémoire. • Les instances optimisées pour le stockage sont conçues pour les charges de travail nécessitant un accès séquentiel élevé en lecture et en écriture (IOPS) au stockage local.
Modèle de tarification	• Instances à la demande Les instances à la demande vous permettent d'utiliser la capacité de calcul à l'heure ou à la seconde sans engagement à long terme. Ces instances sont idéales pour dépasser les besoins de base en matière de performances. • Les Savings Plans permettent de réaliser des économies importantes par rapport aux

Option de configuration	Exemples
	instances à la demande, en échange d'un engagement à utiliser une quantité spécifiqu e de puissance de calcul pour une période d'un ou de trois ans. • Les instances Spot vous permettent de tirer parti de la capacité d'instance inutilisée à un prix réduit pour vos charges de travail sans état et tolérantes aux pannes.
Auto Scaling	Utilisez Auto Scaling pour faire correspon dre les ressources de calcul aux modèles de trafic.
Dimensionnement	• Utilisez Compute Optimizer pour recevoir des recommandations optimisées par le machine learning sur la configuration de calcul qui correspond le mieux à vos caractéristiques de calcul. • Utilisez AWS Lambda Power Tuning pour sélectionner la meilleure configuration pour votre fonction Lambda.
Accélérateurs de calcul matériels	• Les instances de calcul accéléré exécutent des fonctions telles que le traitement graphique ou la mise en correspondance de modèles de données de manière plus efficace que les alternatives basées sur le CPU. • Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, par exemple AWS Trainium, AWS Inferenti aet Amazon EC2 DL1

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'instances Amazon EC2
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)
- Conteneurs Amazon EKS : composants master Amazon EKS
- Amazon ECS Containers: Amazon ECS Container Instances
- Functions: Lambda Function Configuration

Vidéos connexes :

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations
- AWS re:Invent 2022 – https://www.youtube.com/watch?v=5B4-s_ivn1o

Exemples connexes :

- Code de démonstration Compute Optimizer
- Atelier sur les instances Spot Amazon EC2
- Charges de travail efficaces et résilientes avec Amazon EC2 AWS Auto Scaling
- Atelier pour développeurs Graviton
- Journée d'immersion AWS pour les charges de travail Microsoft
- Journée d'immersion AWS pour les charges de travail Linux
- Code de démonstration AWS Compute Optimizer
- Atelier Amazon EKS

PERF02-BP03 Collecter les métriques liées au calcul

Enregistrez et suivez les métriques liées au calcul pour mieux comprendre comment fonctionnent vos ressources de calcul et améliorer leurs performances et leur utilisation.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous n'utilisez que les métriques par défaut enregistrées par votre logiciel de surveillance.
- Vous n'examinez les métriques qu'en cas de problème.

Avantages liés au respect de cette bonne pratique : En collectant des métriques liées aux performances, vous pouvez aligner les performances des applications sur les exigences de l'entreprise afin de garantir que vous répondez à vos besoins en matière de charge de travail. Cela peut également vous aider à améliorer en continu les performances et l'utilisation des ressources de votre charge de travail.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Les charges de travail cloud peuvent générer de gros volumes de données telles que des métriques, des journaux et des événements. Dans AWS Cloud, la collecte de métriques est une étape cruciale qui permet d'améliorer la sécurité, la rentabilité, les performances et la durabilité. AWS fournit un large éventail de métriques liées aux performances à l'aide de services de surveillance, tels que Amazon CloudWatch pour vous fournir des informations précieuses. Les métriques telles que l'utilisation de l'UC, l'utilisation de la mémoire, les E/S de disque et les métriques entrantes et sortantes du réseau peuvent fournir des informations sur les niveaux d'utilisation ou les goulots d'étranglement au niveau des performances. Utilisez ces métriques dans le cadre d'une approche fondée sur les données pour ajuster activement et optimiser les ressources de votre charge de travail. Dans un scénario idéal, vous devriez collecter toutes les métriques relatives à vos ressources de calcul sur une plateforme unique, avec des stratégies de conservation mises en œuvre pour atteindre les objectifs financiers et opérationnels.

Étapes d'implémentation

1. Identifiez les métriques liées aux performances qui sont pertinentes pour votre charge de travail. Vous devriez collecter des métriques relatives à l'utilisation des ressources et au fonctionnement de votre charge de travail cloud (comme le temps de réponse et le débit).
   a. Métriques par défaut Amazon EC2
   b. Métriques Amazon ECS par défaut
   c. Métriques par défaut Amazon EKS
   d. Métriques Lambda par défaut
   e. Métriques de mémoire et de disque Amazon EC2
2. Choisissez et configurez la solution de journalisation et de surveillance adaptée à votre charge de travail.
   a. Observabilité native AWS
   b. AWS Distro for OpenTelemetry
   c. Amazon Managed Service for Prometheus
3. Définissez le filtre et l'agrégation requis pour les métriques en fonction de vos exigences en matière de charge de travail.
   a. Quantifier les métriques d'application personnalisées avec Amazon CloudWatch Logs et les filtres de métrique
   b. Collecter des métriques personnalisées avec le balisage stratégique Amazon CloudWatch (langue française non garantie)
4. Configurez des stratégies de conservation des données pour vos métriques afin qu'elles correspondent à vos objectifs sécuritaires et opérationnels.
   a. Métriques de conservation des données pour CloudWatch
   b. Conservation des données pour CloudWatch Logs
5. Si nécessaire, créez des alarmes et des notifications pour vos métriques afin de vous aider à résoudre de manière proactive les problèmes liés aux performances.
   a. Créer des alarmes pour les métriques personnalisées à l'aide de la détection d'anomalies Amazon CloudWatch (langue française non garantie)
   b. Créer des métriques et des alarmes pour certaines pages Web avec RUM Amazon CloudWatch (langue française non garantie)
6. Utilisez l'automatisation pour déployer vos agents d'agrégation de métriques et de journaux.

a. Automation AWS Systems Manager
b. Collecteur OpenTelemetry

Ressources

Documents connexes :

- Monitoring and observability
- Best practices: implementing observability with AWS
- Documentation Amazon CloudWatch
- Collecte des métriques et des journaux des instances Amazon EC2 et serveurs sur site avec l'agent CloudWatch
- Accès à Amazon CloudWatch Logs pour AWS Lambda
- Utiliser CloudWatch Logs avec des instances de conteneur
- Publier des métriques personnalisées
- AWS Answers : journalisation centralisée
- Services AWS publiant des métriques CloudWatch
- Surveillance d'Amazon EKS sur AWS Fargate

Vidéos connexes :

- AWS re:Invent 2023 – [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 – Implementing application observability
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS re:Invent 2023 – Seamless observability with AWS Distro for OpenTelemetry
- Application Performance Management on AWS

Exemples connexes :

- Journée d'immersion AWS pour les charges de travail Linux – Amazon CloudWatch
- Surveillance des clusters et des conteneurs Amazon ECS
- Surveillance avec les tableaux de bord Amazon CloudWatch

- Atelier Amazon EKS

PERF02-BP04 Configurer et dimensionner correctement les ressources de calcul

Configurez et dimensionnez correctement les ressources de calcul en fonction des exigences de performance de votre charge de travail et évitez de sous-utiliser ou de surexploiter les ressources.

Anti-modèles courants :

- Vous ignorez les exigences de performance de votre charge de travail, ce qui entraîne un surprovisionnement ou un sous-provisionnement des ressources de calcul.
- Vous ne choisissez que la plus grande ou la plus petite instance disponible pour toutes les charges de travail.
- Vous n'utilisez qu'une seule famille d'instances pour faciliter la gestion.
- Vous ignorez les recommandations d'AWS Cost Explorer ou de Compute Optimizer concernant le redimensionnement.
- Vous ne réévaluez pas la charge de travail pour voir si de nouveaux types d'instances pourraient convenir.
- Vous ne certifiez qu'un petit nombre de configurations d'instance pour votre organisation.

Avantages liés au respect de cette bonne pratique : Dimensionner correctement les ressources de calcul garantit le fonctionnement optimal dans le cloud en évitant le surprovisionnement et le sous-provisionnement des ressources. Le dimensionnement correct des ressources de calcul se traduit généralement par des performances renforcées, une meilleure expérience client et une baisse des coûts.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Le dimensionnement correct permet aux organisations d'exploiter leur infrastructure cloud de manière efficace et rentable tout en répondant aux besoins de l'entreprise. Le surprovisionnement des ressources cloud peut entraîner des coûts supplémentaires, tandis que le sous-provisionnement peut générer de faibles performances et une expérience client négative. AWS fournit des outils tels que

AWS Compute Optimizer et AWS Trusted Advisor qui utilisent des données historiques pour fournir des recommandations de redimensionnement de vos ressources de calcul.

Étapes d'implémentation

- Choisissez le type d'instance qui correspond le mieux à vos besoins :
  - Comment choisir le type d'instance Amazon EC2 approprié pour ma charge de travail ?
  - Sélection de type d'instance basée sur des attributs pour la flotte Amazon EC2
  - Créez un groupe Auto Scaling en utilisant la sélection du type d'instance basée sur des attributs (langue française non garantie)
  - Optimisation de vos coûts de calcul Kubernetes avec la consolidation Karpenter
- Analysez les différentes caractéristiques de performances de votre charge de travail et la façon dont ces caractéristiques se rapportent à la mémoire, au réseau et à l'utilisation du processeur. Utilisez ces données pour choisir les ressources qui correspondent le mieux aux objectifs de votre charge de travail en termes de profil et de performance.
- Surveillez l'utilisation de vos ressources à l'aide des outils de surveillance d'AWS tels qu'Amazon CloudWatch.
- Sélectionnez la configuration adaptée à vos ressources de calcul.
  - Pour les charges de travail éphémères, évaluez les métriques d'instance Amazon CloudWatch telles que `CPUUtilization` pour identifier si l'instance est sous-utilisée ou surexploitée.
  - Pour les charges de travail stables, vérifiez les outils de redimensionnement AWS tels qu'AWS Compute Optimizer et AWS Trusted Advisor à intervalles réguliers pour identifier les opportunités d'optimisation et de redimensionnement des ressources de calcul.
- Testez les changements de configuration dans un environnement hors production avant de les implémenter dans un environnement réel.
- Réévaluez en permanence les nouvelles offres de calcul et comparez-les aux besoins de votre charge de travail.

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'instances Amazon EC2

- Amazon ECS Containers: Amazon ECS Container Instances
- Conteneurs Amazon EKS : composants master Amazon EKS
- Functions: Lambda Function Configuration
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)

Vidéos connexes :

- Amazon EC2 foundations
- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Exemples connexes :

- Code de démonstration AWS Compute Optimizer
- Atelier Amazon EKS
- Recommandations en matière de redimensionnement

PERF02-BP05 Mettre à l'échelle vos ressources de calcul de manière dynamique

Utilisez l'élasticité du cloud pour mettre à l'échelle vos ressources de calcul de manière dynamique afin de répondre à vos besoins et d'éviter de surprovisionner ou de sous-provisionner la capacité de votre charge de travail.

Anti-modèles courants :

- Vous réagissez aux alertes en augmentant manuellement la capacité.
- Vous utilisez les mêmes recommandations de dimensionnement (généralement, infrastructure statique) que sur site.
- Vous conservez une capacité accrue après un événement de mise à l'échelle au lieu de la réduire.

Avantages liés au respect de cette bonne pratique : En configurant et en testant l'élasticité des ressources de calcul, vous pouvez économiser de l'argent, maintenir les points de référence des performances et améliorer la fiabilité en fonction de l'évolution du trafic.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS apporte la flexibilité nécessaire pour mettre à l'échelle vos ressources de manière dynamique grâce à divers mécanismes de mise à l'échelle afin de répondre à l'évolution de la demande. Combinée aux métriques liées au calcul, la mise à l'échelle dynamique permet aux charges de travail de réagir automatiquement aux changements et d'utiliser l'ensemble optimal de ressources de calcul pour atteindre son objectif.

Vous pouvez utiliser plusieurs approches pour adapter l'offre de ressources à la demande.

- Approche visant à suivre les cibles: surveillez votre métrique de capacité de mise à l'échelle et augmentez ou réduisez automatiquement votre capacité selon vos besoins.
- Mise à l'échelle prédictive: mettez à l'échelle en prévision des tendances quotidiennes et hebdomadaires.
- Approche basée sur un calendrier: planifiez votre propre calendrier de mise à l'échelle en fonction de changements de charge prévisibles.
- Mise à l'échelle des services: choisissez des services (sans serveur, par exemple) conçus pour se mettre à l'échelle automatiquement.

Vous devez vous assurer que les déploiements de charge de travail peuvent gérer les événements de mise à l'échelle ascendante et descendante.

Étapes d'implémentation

- Les instances de calcul, les conteneurs et les fonctions fournissent des mécanismes d'élasticité, soit en combinaison avec l'autoscaling, soit en tant que fonctionnalité du service. Voici des exemples de mécanismes d'autoscaling :

Mécanisme d'autoscaling	Où utiliser
Amazon EC2 Auto Scaling	Pour vous assurer que vous disposez du nombre adéquat d'instances Amazon EC2

Mécanisme d'autoscaling	Où utiliser
	disponibles pour gérer la charge utilisateur de votre application.
Application Auto Scaling	Pour mettre à l'échelle automatiquement les ressources pour les services AWS individuels au-delà d'Amazon EC2, tels que les fonctions AWS Lambda ou les services Amazon Elastic Container Service (Amazon ECS) .
Kubernetes Cluster Autoscaler/Karpenter	Pour mettre à l'échelle automatiquement les clusters Kubernetes.

- La mise à l'échelle est souvent abordée pour les services de calcul, tels que les instances Amazon EC2 ou les fonctions AWS Lambda. Assurez-vous également de prendre en compte la configuration des services non liés au calcul tels que AWS Glue afin de répondre à la demande.
- Vérifiez que les métriques de mise à l'échelle correspondent aux caractéristiques de la charge de travail en cours de déploiement. Si vous déployez une application de transcodage vidéo, une utilisation de 100 % du processeur est attendue. N'en faites pas votre métrique principale. Utilisez plutôt la profondeur de la file d'attente des tâches de transcodage. Vous pouvez utiliser une métrique personnalisée pour votre politique de mise à l'échelle si nécessaire. Pour choisir les bonnes métriques, tenez compte des conseils suivants pour Amazon EC2 :
  - La métrique doit être une métrique d'utilisation valide et décrire à quel point l'instance est occupée.
  - La valeur de la métrique doit augmenter ou diminuer proportionnellement au nombre d'instances dans le groupe Auto Scaling.
- Assurez-vous d'utiliser la mise à l'échelle dynamique plutôt que la mise à l'échelle manuelle pour votre groupe Auto Scaling. Nous vous recommandons également d'utiliser des politiques de mise à l'échelle du suivi des cibles dans votre mise à l'échelle dynamique.
- Vérifiez que les déploiements de charges de travail peuvent gérer les deux événements de mise à l'échelle (augmentation et diminution des charges de travail). À titre d'exemple, vous pouvez utiliser l'historique d'activité pour vérifier une activité de mise à l'échelle pour un groupe Auto Scaling.
- Évaluez votre charge de travail pour les modèles prédictifs et mettez-la à l'échelle de manière proactive pour anticiper les changements prévisibles et prévus de la demande. Avec la mise à l'échelle prédictive, vous pouvez supprimer le besoin de surprovisionner de la capacité. Pour en savoir plus, consultez Mise à l'échelle prédictive avec Amazon EC2 Auto Scaling.

Ressources

Documents connexes :

- Cloud Compute with AWS
- Types d'instances Amazon EC2
- Amazon ECS Containers: Amazon ECS Container Instances
- Conteneurs Amazon EKS : composants master Amazon EKS
- Functions: Lambda Function Configuration
- Contrôle de l'état du processeur pour votre instance Amazon EC2 (langue française non garantie)
- En savoir plus sur la Auto Scaling d'un cluster Amazon ECS
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler

Vidéos connexes :

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Exemples connexes :

- Exemples de groupes Amazon EC2 Auto Scaling
- Atelier Amazon EKS
- Mettez à l'échelle vos charges de travail Amazon EKS en les exécutant sur IPv6

PERF02-BP06 Utiliser des accélérateurs de calcul matériels optimisés

Utilisez des accélérateurs matériels pour exécuter certaines fonctions de manière plus efficace que les alternatives basées sur l'UC.

Anti-modèles courants :

- En ce qui concerne votre charge de travail, vous n'avez pas comparé une instance à usage général à une instance dédiée capable de fournir de meilleures performances à moindre coût.
- Vous utilisez des accélérateurs de calcul matériels pour les tâches qui peuvent être plus efficaces en utilisant des alternatives basées sur l'UC.
- Vous ne surveillez pas l'utilisation du GPU.

Avantages liés au respect de cette bonne pratique : en utilisant des accélérateurs matériels, tels que des unités de traitement graphique (GPU) et des circuits logiques programmables (FPGA), vous pouvez exécuter certaines fonctions de traitement de manière plus efficace.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Les instances de calcul accéléré donnent accès à des accélérateurs de calcul matériels tels que les GPU et les FPGA. Ces accélérateurs matériels exécutent certaines fonctions comme le traitement graphique ou la correspondance de modèles de données plus efficacement que les alternatives basées sur l'UC. De nombreuses charges de travail accélérées, telles que le rendu, le transcodage et le machine learning, sont très variables en termes d'utilisation des ressources. Exécutez ce matériel uniquement pendant le temps nécessaire et mettez-le hors service grâce à l'automatisation lorsque vous n'en avez plus besoin afin d'améliorer l'efficacité globale des performances.

Étapes d'implémentation

- Identifiez quelles instances de calcul accéléré peuvent répondre à vos exigences.
- Pour les charges de travail de machine learning, tirez parti d'un matériel conçu spécialement pour votre charge de travail, comme AWS Trainium, AWS Inferentia et Amazon EC2 DL1. Les instances Inferentia AWS telles que les instances Inf2 offrent des performances/watt jusqu'à 50 % supérieures à celles des instances Amazon EC2 comparables.
- Collectez des métriques d'utilisation pour vos instances de calcul accéléré. Par exemple, vous pouvez utiliser l'agent CloudWatch pour collecter des métriques comme `utilization_gpu` et `utilization_memory` pour vos GPU, comme illustré dans Collecter les métriques des GPU NVIDIA avec Amazon CloudWatch.
- Optimisez le code, le fonctionnement du réseau et les paramètres des accélérateurs matériels pour veiller à ce que le matériel sous-jacent soit pleinement utilisé.

  - Optimiser les paramètres GPU
  - GPU Monitoring and Optimization in the Deep Learning AMI (Surveillance et optimisation des GPI dans l'AMI Deep Learning)
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker
- Utilisez les dernières bibliothèques performantes et les pilotes GPU.
- Utilisez l'automatisation pour libérer les instances GPU lorsqu'elles ne sont pas utilisées.

Ressources

Documents connexes :

- Utilisation des GPU sur Amazon Elastic Container Service
- Instances GPU (langue française non garantie)
- Instances avec AWS Trainium (langue française non garantie)
- Instances avec AWS Inferentia (langue française non garantie)
- Passons à l'architecture ! Architecture avec des puces personnalisées et des accélérateurs (langue française non garantie)

- Accelerated Computing (Calcul accéléré)
- Amazon EC2 VT1 Instances (Instances VT1 EC2)
- Comment choisir le type d'instance Amazon EC2 approprié pour ma charge de travail ?
- Choisissez le meilleur accélérateur d'IA et la meilleure compilation de modèles pour l'inférence de vision par ordinateur avec Amazon SageMaker (langue française non garantie)

Vidéos connexes :

- AWS re:Invent 2021 - How to select Amazon Elastic Compute Cloud GPU instances for deep learning
- AWS re:Invent 2022 - [NEW LAUNCH!] Introducing AWS Inferentia2-based Amazon EC2 Inf2 instances
- AWS re:Invent 2022 - Accelerate deep learning and innovate faster with AWS Trainium
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Exemples connexes :

- Amazon SageMaker and NVIDIA GPU Cloud (NGC)
- Use SageMaker with Trainium and Inferentia for optimized deep learning training and inferencing workloads
- Optimizing NLP models with Amazon Elastic Compute Cloud Inf1 instances in Amazon SageMaker

Gestion des données

La solution optimale de gestion des données pour un système particulier varie en fonction du type de données (bloc, fichier ou objet), des modèles d'accès (aléatoire ou séquentiel), du débit requis, de la fréquence d'accès (en ligne, hors ligne, archivage), de la fréquence de mise à jour (WORM, dynamique), ainsi que des contraintes de disponibilité et de durabilité. Les charges de travail bien architecturées utilisent des magasins de données sur mesure qui intègrent différentes fonctionnalités pour améliorer les performances.

Ce domaine d'intérêt partage des conseils et des bonnes pratiques pour optimiser le stockage des données, les modèles de déplacement et d'accès, ainsi que l'efficacité des performances des magasins de données.

Bonnes pratiques

- PERF03-BP01 Utiliser un magasin de données dédié le mieux adapté à vos besoins en matière de stockage des données et d’accès aux données
- PERF03-BP02 Évaluer les options de configuration disponibles pour un magasin de données
- PERF03-BP03 Collecter et archiver les métriques de performance du magasin de données
- PERF03-BP04 Mettre en œuvre des stratégies pour améliorer les performances des requêtes dans un magasin de données
- PERF03-BP05 Mise en œuvre de modèles d’accès aux données utilisant la mise en cache

PERF03-BP01 Utiliser un magasin de données dédié le mieux adapté à vos besoins en matière de stockage des données et d’accès aux données

Comprenez les caractéristiques des données (telles que la possibilité de partage, la taille, la taille du cache, les modèles d’accès, la latence, le débit et la persistance des données) afin de sélectionner les magasins de données dédiés (stockage ou base de données) adaptés à votre charge de travail.

Anti-modèles courants :

- Vous vous en tenez à un magasin de données, car l’équipe interne sait comment tirer parti de ce type de solution en particulier.

- Vous partez du principe que toutes les charges de travail ont des exigences similaires en termes de stockage de données et d'accès aux données.
- Vous n'avez pas implémentée de catalogue de données pour inventorier vos ressources de données.

Avantages liés au respect de cette bonne pratique : comprendre les caractéristiques des données et les exigences vous permet de déterminer la technologie de stockage la plus efficace et la plus performante pour répondre à vos besoins en matière de charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas établie : élevé

Directives d'implémentation

Lors de la sélection et de la mise en œuvre du stockage des données, assurez-vous que les caractéristiques d'interrogation, de mise à l'échelle et de stockage répondent aux exigences en matière de données de charge de travail. AWS fournit de nombreuses technologies de stockage de données et de base de données, notamment le stockage par blocs, le stockage d'objets, le stockage en continu, le système de fichiers et les bases de données relationnelles, clé-valeur, document, en mémoire, orientées graphe, de séries chronologiques et de registre. Chaque solution de gestion de données propose des options et des configurations pour prendre en charge vos cas d'utilisation et vos modèles de données. En comprenant les caractéristiques et les exigences des données, vous pouvez vous affranchir de la technologie de stockage monolithique et des approches restrictives et universelles pour vous concentrer sur la gestion appropriée des données.

Étapes d'implémentation

- Procédez à l'inventaire des différents types de données qui existent dans votre charge de travail.
- Comprenez et documentez les caractéristiques et les exigences des données, notamment :
  - Type de données (non structurées, semi-structurées, relationnelles)
  - Volume et croissance des données
  - Durabilité des données : persistantes, éphémères, temporaires
  - Exigences ACID (atomicité, cohérence, isolement, durabilité)
  - Modèles d'accès aux données (à lecture intensive ou à écriture intensive)
  - Latence
  - débit
  - IOPS (opérations d'entrée/sortie par seconde)

  - Durée de conservation des données
- Découvrez les différents magasins de données disponibles (services de stockage et de base de données) disponibles pour votre charge de travail AWS qui peuvent répondre aux caractéristiques de vos données, telles qu'elles sont décrites dans PERF01-BP01 Découvrir et se familiariser avec les services et fonctionnalités cloud disponibles. Voici quelques exemples de technologies de stockage AWS et leurs principales caractéristiques :

Type	Services AWS	Principales caractéristiques
Object storage	Amazon S3	Unlimited scalability, high availability, and multiple options for accessibility. Transferring and accessing objects in and out of Amazon S3 can use a service, such as Transfer Acceleration or Points d'accès, to support your location, security needs, and access patterns.
Archiving storage	Amazon S3 Glacier	Built for data archiving.
Streaming storage	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Efficient ingestion and storage of streaming data.
Shared file system	Amazon Elastic File System (Amazon EFS)	Système de fichiers montable auquel plusieurs types de solutions informatiques peuvent accéder.
Shared file system	Amazon FSx	Built on the latest AWS compute solutions to support four commonly used file systems: NetApp ONTAP, OpenZFS, Windows File

Type	Services AWS	Principales caractéristiques
		Server, and Lustre. Amazon FSx La latence, le débit et les IOPS vary per file system and should be considered when selecting the right file system for your workload needs.
Block storage	Amazon Elastic Block Store (Amazon EBS)	Scalable, high-performance block-storage service designed for Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS includes SSD-backed storage for transactional, IOPS-intensive workloads and HDD-backed storage for throughput-intensive workloads.
Relational database	Amazon Aurora, Amazon RDS, Amazon Redshift.	Designed to support ACID (atomicity, consistency, isolation, durability) transactions, and maintain referential integrity and strong data consistency. Many traditional applications, enterprise resource planning (ERP), customer relationship management (CRM), and ecommerce use relational databases to store their data.

Type	Services AWS	Principales caractéristiques
Key-value database	Amazon DynamoDB	Optimized for common access patterns, typically to store and retrieve large volumes of data. High-traffic web apps, ecommerce systems, and gaming applications are typical use-cases for key-value databases.
Document database	Amazon DocumentDB	Designed to store semi-structured data as JSON-like documents. These databases help developers build and update applications such as content management, catalogs, and user profiles quickly.
In-memory database	Amazon ElastiCache , Amazon MemoryDB for Redis	Used for applications that require real-time access to data, lowest latency and highest throughput. You may use in-memory databases for application caching, session management, gaming leaderboards, low latency ML feature store, microservices messaging system, and a high-throughput streaming mechanism

Type	Services AWS	Principales caractéristiques
Graph database	Amazon Neptune	Used for applications that must navigate and query millions of relationships between highly connected graph datasets with milliseco nd latency at large scale. Many companies use graph databases for fraud detection , social networking, and recommendation engines.
Time Series database	Amazon Timestream	Used to efficiently collect, synthesize, and derive insights from data that changes over time. IoT applications, DevOps, and industrial telemetry can utilize time-series databases.
Wide column	Amazon Keyspaces (pour Apache Cassandra)	Uses tables, rows, and columns, but unlike a relational database, the names and format of the columns can vary from row to row in the same table. You typically see a wide column store in high scale industrial apps for equipment maintenance, fleet management, and route optimization.

Type	Services AWS	Principales caractéristiques
Ledger	Amazon Quantum Ledger Datab (Amazon QLDB)	Provides a centralized and trusted authority to maintain a scalable, immutable, and cryptographically verifiable record of transactions for every application. We see ledger databases used for systems of record, supply chain, registrations, and even banking transactions.

- Si vous créez une plateforme de données, tirez parti de l'architecture de données moderne sur AWS pour intégrer votre lac de données, votre entrepôt des données et vos magasins de données dédiés.
- Les principales questions que vous devez vous poser lors du choix d'un magasin de données pour votre charge de travail sont les suivantes :

Question	Things to consider
How is the data structured?	• Si les données ne sont pas structurées, envisagez d'utiliser un magasin d'objets tel que Amazon S3 ou une base de données NoSQL telle que Amazon DocumentDB • Pour les données clé-valeur, envisagez d'utiliser DynamoDB, Amazon ElastiCache for Redis ou Amazon MemoryDB for Redis
What level of referential integrity is required?	• Pour les contraintes de clé étrangère, les bases de données relationnelles telles que Amazon RDS et Aurora peuvent fournir ce niveau d'intégrité. • En règle générale, dans un modèle de données NoSQL, vous dénormalisez les données en un seul document ou en une

Question	Things to consider
	collection de documents à récupérer en une seule requête au lieu de joindre des documents ou des tables.
Is ACID (atomicity, consistency, isolation, durability) compliance required?	• Si les propriétés ACID associées aux bases de données relationnelles sont requises, envisagez d'utiliser une base de données relationnelle telle que Amazon RDS et Aurora. • Si une forte cohérence est requise pour la base de données NoSQL, vous pouvez utiliser des lectures fortement cohérentes avec DynamoDB.
How will the storage requirements change over time? How does this impact scalability?	• Les bases de données sans serveur telles que DynamoDB et Amazon Quantum Ledger Database (Amazon QLDB) seront mises à l'échelle de manière dynamique. • Les bases de données relationnelles ont des limites supérieures sur le stockage alloué et doivent souvent être partitionnées horizontalement à l'aide de mécanismes tels que le partitionnement une fois qu'elles atteignent ces limites.
What is the proportion of read queries in relation to write queries? Would caching be likely to improve performance?	• Les charges de travail à lecture intensive peuvent bénéficier d'une couche de mise en cache, comme ElastiCache ou DAX si la base de données est DynamoDB. • Les lectures peuvent également être déchargées pour lire des réplicas avec des bases de données relationnelles comme Amazon RDS.

Question	Things to consider
Does storage and modification (OLTP - Online Transaction Processing) or retrieval and reporting (OLAP - Online Analytical Processin g) have a higher priority?	• Pour un traitement transactionnel des lectures en l'état à haut débit, envisagez d'utiliser une base de données NoSQL comme DynamoDB. • Pour des modèles de lecture complexes à haut débit (tels que la jointure) avec cohérence, utilisez Amazon RDS. • Pour les requêtes analytiques, envisagez d'utiliser une base de données orientée colonnes, telle que Amazon Redshift, ou d'exporter les données vers Amazon S3 et d'effectuer une analyse à l'aide d'Athena ou d'Amazon QuickSight.
What level of durability does the data require?	• Aurora réplique automatiquement vos données sur trois zones de disponibilité au sein d'une région. Autrement dit, vos données sont très durables avec moins de risque de perte de données. • DynamoDB est automatiquement répliqué sur plusieurs zones de disponibilité, assurant ainsi la haute disponibilité et la durabilité des données. • Amazon S3 offre une durabilité de 99,999999999 %. De nombreux services de base de données, tels que Amazon RDS et DynamoDB, prennent en charge l'exporta tion des données vers Amazon S3 pour une conservation et un archivage à long terme.

Question	Things to consider
Is there a desire to move away from commercial database engines or licensing costs?	• Envisagez d'utiliser des moteurs open source tels que PostgreSQL et MySQL sur Amazon RDS ou Aurora. • Tirez parti d'AWS Database Migration Service et d'AWS Schema Conversion Tool pour migrer des moteurs de bases de données commerciaux vers des moteurs open source.
What is the operational expectation for the database? Is moving to managed services a primary concern?	• L'utilisation d'Amazon RDS au lieu d'Amazon EC2 et de DynamoDB ou d'Amazon DocumentDB au lieu de l'auto-hébergement d'une base de données NoSQL contribue à réduire les frais généraux opérationnels.
How is the database currently accessed? Is it only application access, or are there business intelligence (BI) users and other connected off-the-shelf applications?	• Si vous dépendez d'outils externes, vous devrez peut-être préserver la compatibilité avec les bases de données qu'ils prennent en charge. Amazon RDS est entièrement compatible avec les différentes versions de moteur qu'il prend en charge, notamment Microsoft SQL Server, Oracle, MySQL et PostgreSQL.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier le magasin de données qui peut répondre à vos exigences en termes de charge de travail.

Ressources

Documents connexes :

- Types de volume Amazon EBS
- Stockage Amazon EC2

- Amazon EFS : performances d'Amazon EFS (langue française non garantie)
- Performances d'Amazon FSx for Lustre (langue française non garantie)
- Performances d'Amazon FSx for Windows File Server (langue française non garantie)
- Amazon S3 Glacier : documentation S3 Glacier
- Schémas de conception des bonnes pratiques : optimisation des performances Amazon S3
- Stockage dans le cloud sur AWS
- Caractéristiques d'E/S d'Amazon EBS (langue française non garantie)
- Bases de données dans le cloud AWS
- Mise en cache de bases de données AWS (langue française non garantie)
- DynamoDB Accelerator
- Bonnes pratiques avec Amazon Aurora
- Performances d'Amazon Redshift (langue française non garantie)
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena (langue française non garantie)
- Bonnes pratiques Amazon Redshift Spectrum (langue française non garantie)
- Bonnes pratiques Amazon DynamoDB (langue française non garantie)
- Choisir entre Amazon EC2 et Amazon RDS (langue française non garantie)
- Qu'est-ce qu'Amazon ElastiCache ?

Vidéos connexes :

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimizing storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2022: Building modern data architectures on AWS
- AWS re:Invent 2022: Building data mesh architectures on AWS
- AWS re:Invent 2023: Deep dive into Amazon Aurora and its innovations
- AWS re:Invent 2023: Advanced data modeling with Amazon DynamoDB
- AWS re:Invent 2022: Modernize apps with purpose-built databases
- Amazon DynamoDB deep dive: Advanced design patterns

Exemples connexes :

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Build a Data Mesh on AWS
- Exemples Amazon S3 (langue française non garantie)
- Optimiser le modèle de données à l'aide du partage de données Amazon Redshift (langue française non garantie)
- Migrations des bases de données
- MS SQL Server - Démonstration de réplication AWS Database Migration Service (AWS DMS) (langue française non garantie)
- Atelier pratique sur la modernisation des bases de données
- Exemples Amazon Neptune (langue française non garantie)

PERF03-BP02 Évaluer les options de configuration disponibles pour un magasin de données

Comprenez et évaluez les différentes fonctionnalités et options de configuration disponibles pour vos magasins de données afin d'optimiser l'espace de stockage et les performances de votre charge de travail.

Anti-modèles courants :

- Vous n'utilisez qu'un seul type de stockage (comme Amazon EBS) pour toutes les charges de travail.
- Vous utilisez les IOPS provisionnés pour toutes les charges de travail sans effectuer de test en situation réelle sur tous les niveaux de stockage.
- Vous ne connaissez pas les options de configuration de la solution de gestion de données que vous avez choisie.
- Vous vous concentrez uniquement sur l'augmentation de la taille de l'instance sans examiner les autres options de configuration disponibles.
- Vous ne testez pas les caractéristiques de mise à l'échelle de votre magasin de données.

Avantages liés au respect de cette bonne pratique : en explorant et en expérimentant les configurations de magasin de données, vous pourriez réduire le coût de l'infrastructure, améliorer les performances et réduire l'effort requis pour maintenir vos charges de travail.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Une charge de travail peut comporter un ou plusieurs magasins de données utilisés en fonction des exigences de stockage des données et d'accès aux données. Pour optimiser l'efficacité et le coût de vos performances, vous devez évaluer les modèles d'accès aux données afin de déterminer les configurations de magasin de données appropriées. Pendant que vous explorez les options de magasin de données, tenez compte de divers aspects tels que les options de stockage, la mémoire, le calcul, le réplica en lecture, les exigences de cohérence, le regroupement de connexions et les options de mise en cache. Testez ces différentes options de configuration pour améliorer les métriques d'efficacité des performances.

Étapes d'implémentation

- Comprenez les configurations actuelles (comme le type d'instance, la taille de stockage ou la version du moteur de base de données) de votre magasin de données.
- Consultez la documentation AWS et les bonnes pratiques pour en savoir plus sur les options de configuration recommandées qui peuvent vous aider à améliorer les performances de votre magasin de données. Les principales options de magasin de données à prendre en compte sont les suivantes :

Configuration option	Examples
Offloading reads (like read replicas and caching)	• Pour les tables DynamoDB, vous pouvez décharger les lectures à l'aide de DAX pour la mise en cache. • Vous pouvez créer un cluster Amazon ElastiCache for Redis pour Redis et configurer votre application pour qu'elle lise d'abord les données à partir du cache, en revenant à la base de données si l'élément demandé n'est pas présent.

Configuration option	Examples
	• Les bases de données relationnelles comme Amazon RDS et Aurora et les bases de données NoSQL allouées telles que Neptune et Amazon DocumentDB prennent toutes en charge l'ajout de réplicas en lecture pour décharger les parties lues de la charge de travail. • Les bases de données sans serveur comme DynamoDB se mettent à l'échelle automatiquement. Assurez-vous que vous disposez de suffisamment d'unités de capacité de lecture (RCU) allouées pour gérer la charge de travail.

Configuration option	Examples
Scaling writes (like partition key sharding or introducing a queue)	• Pour les bases de données relationnelles, vous pouvez augmenter la taille de l'instanc e pour qu'elle s'adapte à une charge de travail accrue ou augmenter les IOPS provisionnés pour permettre un débit accru vers le stockage sous-jacent. • Vous pouvez également ajouter une file d'attente devant votre base de données plutôt que d'écrire directement dans la base de données. Ce modèle vous permet de dissocier l'ingestion de la base de données et de contrôler le débit afin que la base de données ne soit pas submergée. • Regrouper vos demandes d'écriture plutôt que de créer de nombreuses transactions de courte durée contribue à améliorer le débit dans les bases de données relationn elles à volume d'écriture élevé. • Les bases de données sans serveur comme DynamoDB peuvent mettre à l'échelle le débit d'écriture automatiquement ou en ajustant les unités de capacité d'écriture allouées (WCU) en fonction du mode de capacité. • Vous pouvez toujours rencontrer des problèmes avec les partitions à chaud lorsque vous atteignez les limites de débit pour une clé de partition donnée. Pour pallier ce problème, choisissez une clé de partition distribuée plus uniformément ou partitionnez en écriture la clé de partition.

Configuration option	Examples
Policies to manage the lifecycle of your datasets	• Vous pouvez utiliser Amazon S3 Lifecycle afin de gérer vos objets au cours de leur cycle de vie. Si vos modèles d'accès sont inconnus, changeants ou imprévisibles, vous pouvez utiliser Amazon S3 Intelligent-Tiering, qui surveille les modèles d'accès et déplace automatiquement les objets qui n'ont pas été consultés aux niveaux d'accès à moindre coût. Vous pouvez utiliser les métriques d'Amazon S3 Storage Lens afin d'identifier les possibilités d'optimisation et les écarts dans la gestion du cycle de vie. • La gestion du cycle de vie Amazon EFS gère automatiquement le stockage des fichiers pour vos systèmes de fichiers.
Connection management and pooling	• Amazon RDS Proxy peut être utilisé avec Amazon RDS et Aurora pour gérer les connexions à la base de données. • Les bases de données sans serveur comme DynamoDB n'ont pas de connexion s associées, mais tenez compte de la capacité allouée et des politiques de mise à l'échelle automatique pour faire face aux pics de charge.

- Réalisez des tests et procédez au benchmarking dans un environnement hors production afin d'identifier l'option de configuration qui répond à vos exigences en termes de charge de travail.
- Après avoir réalisé vos tests, planifiez votre migration et validez vos métriques de performance.
- Utilisez les outils de surveillance AWS (tels qu'Amazon CloudWatch) et d'optimisation (tels qu'Amazon S3 Storage Lens) pour optimiser en continu votre magasin de données à l'aide d'un modèle d'utilisation réel.

Ressources

Documents connexes :

- Stockage dans le cloud sur AWS
- Types de volume Amazon EBS
- Stockage Amazon EC2
- Amazon EFS : performances d'Amazon EFS (langue française non garantie)
- Performances d'Amazon FSx for Lustre (langue française non garantie)
- Performances d'Amazon FSx for Windows File Server (langue française non garantie)
- Amazon S3 Glacier : documentation S3 Glacier
- Schémas de conception des bonnes pratiques : optimisation des performances Amazon S3
- Caractéristiques d'E/S d'Amazon EBS (langue française non garantie)
- Bases de données dans le cloud AWS
- Mise en cache de bases de données AWS (langue française non garantie)
- DynamoDB Accelerator
- Bonnes pratiques avec Amazon Aurora
- Performances d'Amazon Redshift (langue française non garantie)
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena (langue française non garantie)
- Bonnes pratiques Amazon Redshift Spectrum (langue française non garantie)
- Bonnes pratiques Amazon DynamoDB (langue française non garantie)

Vidéos connexes :

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimize storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2023: What's new with AWS file storage
- AWS re:Invent 2023: Dive deep into Amazon DynamoDB

Exemples connexes :

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Amazon EBS Autoscale
- Exemples Amazon S3 (langue française non garantie)
- Exemples Amazon DynamoDB (langue française non garantie)
- Exemples de migration de base de données AWS (langue française non garantie)
- Atelier sur la modernisation des bases de données
- Utilisation des paramètres de votre instance de base de données Amazon RDS for Postgress (langue française non garantie)

PERF03-BP03 Collecter et archiver les métriques de performance du magasin de données

Suivez et archivez les métriques de performance pertinentes pour votre magasin de données afin de comprendre comment fonctionnent vos solutions de gestion des données. Ces métriques peuvent vous aider à optimiser votre magasin de données, à vérifier que les exigences de votre charge de travail sont satisfaites et à fournir une vue d'ensemble claire sur le fonctionnement de la charge de travail.

Anti-modèles courants :

- Vous utilisez uniquement la recherche manuelle des fichiers journaux pour les métriques.
- Vous publiez uniquement des métriques sur les outils internes utilisés par votre équipe et vous n'avez pas une visibilité complète de votre charge de travail.
- Vous n'utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n'examinez les métriques qu'en cas de problème.
- Vous ne surveillez que les métriques au niveau du système et vous ne capturez pas les métriques d'accès aux données ou d'utilisation des données.

Avantages liés au respect de cette bonne pratique : La définition de points de référence pour les performances vous permet de mieux comprendre le comportement normal et les exigences des

charges de travail. Les modèles anormaux peuvent être identifiés et débogués plus rapidement, ce qui améliore les performances et la fiabilité du magasin de données.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

L'enregistrement de plusieurs métriques de performance sur une période donnée est nécessaire pour la surveillance des performances de vos magasins de données. Cette surveillance vous permet non seulement de détecter les anomalies, mais aussi d'évaluer les performances par rapport aux métriques métier afin de vérifier que vous répondez aux besoins de votre charge de travail.

Ces métriques doivent inclure à la fois le système sous-jacent qui prend en charge le magasin de données et les métriques de la base de données. Les métriques système sous-jacentes peuvent inclure l'utilisation du processeur, la mémoire, le stockage sur disque disponible, les E/S de disque, le taux d'accès au cache et les métriques entrantes et sortantes du réseau, tandis que les métriques du magasin de données peuvent inclure les transactions par seconde, les principales requêtes, les taux de requêtes moyens, les temps de réponse, l'utilisation de l'index, les verrouillages de table, les délais d'expiration des requêtes et le nombre de connexions ouvertes. Ces données sont essentielles pour comprendre comment fonctionne la charge de travail et comment la solution de gestion des données est utilisée. Utilisez ces métriques dans le cadre d'une approche fondée sur les données pour ajuster et optimiser les ressources de votre charge de travail.

Utilisez des outils, des bibliothèques et des systèmes qui enregistrent des mesures de performances liées aux performances de la base de données.

Étapes d'implémentation

1. Identifiez les métriques de performances clés que votre magasin de données doit suivre.
   a. Amazon S3 Métriques et dimensions
   b. Surveillance des métriques pour une instance Amazon RDS
   c. Surveillance de la charge de base de données avec Performance Insights sur Amazon RDS
   d. Présentation de la surveillance améliorée
   e. DynamoDB Métriques et dimensions
   f. Surveillance de DynamoDB Accelerator
   g. Surveillance de Amazon MemoryDB for Redis avec Amazon CloudWatch
   h. Quelles métriques dois-je surveiller ?

   i. Surveillance des performances du cluster Amazon Redshift
   j. Timestream Métriques et dimensions
   k. Métriques Amazon CloudWatch pour Amazon Aurora
   l. Journalisation et surveillance dans Amazon Keyspaces (for Apache Cassandra)
   m. Surveillance des ressources Amazon Neptune
2. Utilisez une solution de journalisation et de surveillance approuvée pour collecter ces métriques. Amazon CloudWatch peut récupérer des métriques à partir des ressources de votre architecture. Vous pouvez également récupérer et publier des métriques personnalisées pour faire apparaître des métriques d'entreprise ou des métriques dérivées. Utilisez CloudWatch ou des solutions tierces pour définir des alarmes qui indiquent les dépassements de seuils.
3. Vérifiez si la surveillance du magasin de données peut bénéficier d'une solution de machine learning qui détecte les anomalies de performance.
   a. Amazon DevOps Guru pour Amazon RDS assure la visibilité des problèmes de performances et suggère des actions correctives.
4. Configurez la conservation des données dans votre solution de surveillance et de journalisation en fonction de vos objectifs sécuritaires et opérationnels.
   a. Métriques de conservation des données pour CloudWatch
   b. Conservation des données pour CloudWatch Logs

Ressources

Documents connexes :

- Mise en cache de bases de données AWS
- Les 10 meilleures techniques pour améliorer les performances d'Amazon Athena
- Bonnes pratiques Amazon Aurora
- DynamoDB Accelerator
- Bonnes pratiques Amazon DynamoDB
- Bonnes pratiques Amazon Redshift Spectrum
- Performances Amazon Redshift
- Bases de données cloud avec AWS
- Amazon RDS Performance Insights

Vidéos connexes :

- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Database Performance Monitoring and Tuning with Amazon DevOps Guru for Amazon RDS
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- AWS re:Invent 2023 - Building and optimizing a data lake on Amazon S3
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- Bonnes pratiques pour la surveillance des charges de travail Redis sur Amazon ElastiCache

Exemples connexes :

- Cadre de collecte de métriques pour l'ingestion des jeux de données AWS
- Atelier sur la surveillance Amazon RDS
- Atelier sur les bases de données sur mesure AWS

PERF03-BP04 Mettre en œuvre des stratégies pour améliorer les performances des requêtes dans un magasin de données

Mettez en œuvre des stratégies pour optimiser les données et améliorer les requêtes sur les données afin de renforcer la capacité de mise à l'échelle et l'efficacité des performances pour votre charge de travail.

Anti-modèles courants :

- Vous ne partitionnez pas les données dans votre magasin de données.
- Vous ne stockez les données que dans un seul format de fichier dans votre magasin de données.
- Vous n'utilisez pas d'index dans votre magasin de données.

Avantages liés au respect de cette bonne pratique : En optimisant les performances des données et des requêtes, vous augmentez leur efficacité, vous réduisez les coûts et vous améliorez l'expérience utilisateur.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d’implémentation

L’optimisation des données et des requêtes sont des aspects essentiels de l’efficacité des performances d’un magasin de données, car ils ont un impact sur les performances et la réactivité de l’ensemble de la charge de travail dans le cloud. Les données non optimisées peuvent augmenter l’utilisation des ressources et les goulots d’étranglement, ce qui réduit l’efficacité globale d’un magasin de données.

L’optimisation des données inclut plusieurs techniques pour garantir un stockage des données et un accès aux données efficaces. Cela permet également d’améliorer les performances des requêtes dans un magasin de données. Les principales stratégies incluent le partitionnement des données, la compression des données et la dénormalisation des données, qui permettent d’optimiser les données à la fois pour le stockage et l’accès.

Étapes d’implémentation

- Comprenez et analysez les requêtes essentielles sur les données effectuées dans votre magasin de données.
- Identifiez les requêtes lentes dans votre magasin de données et utilisez des plans de requêtes pour comprendre leur état actuel.
  - Analyse du plan de requête dans Amazon Redshift
  - Utilisation d’EXPLAIN et EXPLAIN ANALYZE dans Athena (langue française non garantie)
- Mettez en œuvre des stratégies pour améliorer les performances des requêtes. Les stratégies clés incluent :
  - L’utilisation d’un format de fichier en colonnes (comme Parquet ou ORC).
  - La compression des données dans le magasin de données pour réduire l’espace de stockage et les opérations d’E/S.
  - Le partitionnement des données pour diviser les données en parties plus petites et réduire le temps d’analyse des données.
    - Partitionnement des données dans Athena (langue française non garantie)
    - Les partitions et la distribution de données
  - L’indexation des données sur les colonnes communes de la requête.
  - Utilisez des vues matérialisées pour les requêtes fréquentes.
    - Compréhension des vues matérialisées
    - Création de vues matérialisées dans Amazon Redshift

  - Choisissez l’opération de jointure appropriée pour la requête. Lorsque vous joignez deux tables, spécifiez la table la plus grande sur le côté gauche de la jointure et la plus petite sur le côté droit de la jointure.
  - La solution de mise en cache distribuée pour améliorer la latence et réduire le nombre d’opérations d’E/S dans la base de données.
  - La maintenance régulière, comme l’exécution de statistiques.
- Expérimentez et testez les stratégies dans un environnement hors production.

Ressources

Documents connexes :

- Amazon Aurora best practices
- Amazon Redshift performance
- Amazon Athena top 10 performance tips
- AWS Database Caching
- Best Practices for Implementing Amazon ElastiCache
- Partitionnement des données dans Athena (langue française non garantie)

Vidéos connexes :

- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Optimize Amazon Athena Queries with New Query Analysis Tools

Exemples connexes :

- Amazon S3 Sélectionner – Interroger des données sans serveurs ni bases de données
- Atelier sur les bases de données sur mesure AWS

PERF03-BP05 Mise en œuvre de modèles d’accès aux données utilisant la mise en cache

Mettez en œuvre des modèles d’accès qui peuvent tirer parti de la mise en cache des données pour une récupération rapide des données fréquemment consultées.

Anti-modèles courants :

- Vous mettez en cache des données qui changent fréquemment.
- Vous utilisez les données mises en cache comme si elles étaient stockées de manière durable et toujours disponibles.
- Vous ne tenez pas compte de la cohérence de vos données mises en cache.
- Vous ne surveillez pas l’efficacité de la mise en œuvre de la mise en cache.

Avantages liés au respect de cette bonne pratique : Le stockage des données dans un cache contribue à améliorer la latence et le débit de lecture, l’expérience utilisateur et l’efficacité globale, tout en réduisant les coûts.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : moyen

Directives d’implémentation

Un cache est un composant logiciel ou matériel destiné à stocker des données afin que les requêtes futures portant sur les mêmes données puissent être traitées plus rapidement ou plus efficacement. Les données stockées dans un cache peuvent être reconstruites en cas de perte en répétant un calcul antérieur ou en les récupérant dans un autre magasin de données.

La mise en cache des données peut être l’une des stratégies les plus efficaces pour améliorer les performances globales de votre application et réduire la charge qui pèse sur vos sources de données principales sous-jacentes. Les données peuvent être mises en cache à plusieurs niveaux dans l’application, par exemple dans l’application effectuant des appels à distance et également connue sous le nom de mise en cache côté client, ou en utilisant un service secondaire rapide pour stocker les données, ce que l’on appelle aussi mise en cache à distance.

Mise en cache côté client

Grâce à la mise en cache côté client, chaque client (une application ou un service qui interroge le magasin de données backend) peut stocker les résultats de ses requêtes uniques localement

pendant une durée spécifiée. Cela permet de réduire le nombre de requêtes adressées à un magasin de données sur le réseau en vérifiant d'abord le cache du client local. En l'absence de résultats, l'application peut alors interroger le magasin de données et stocker ces résultats localement. Ce modèle permet à chaque client de stocker les données dans l'emplacement le plus proche possible (le client lui-même), ce qui se traduit par la latence la plus faible possible. Les clients peuvent également continuer à répondre à certaines requêtes lorsque le magasin de données backend n'est pas disponible, ce qui augmente la disponibilité de l'ensemble du système.

L'un des inconvénients de cette approche est que lorsque plusieurs clients sont impliqués, ils peuvent stocker les mêmes données mises en cache localement. Cela entraîne à la fois une double utilisation du stockage et une incohérence des données entre ces clients. Un client peut mettre en cache les résultats d'une requête et, une minute plus tard, un autre client peut exécuter la même requête et obtenir un résultat différent.

Mise en cache à distance

Pour résoudre le problème de duplication des données entre clients, un service externe rapide, ou cache distant, peut être utilisé pour stocker les données interrogées. Au lieu de vérifier un magasin de données local, chaque client vérifie le cache distant avant d'interroger le magasin de données backend. Cette stratégie permet d'obtenir des réponses plus cohérentes entre les clients, d'améliorer l'efficacité des données stockées et d'augmenter le volume de données mises en cache, car l'espace de stockage évolue indépendamment des clients.

L'inconvénient d'un cache distant est que l'ensemble du système peut connaître une latence plus élevée, car un saut de réseau à réseau supplémentaire est nécessaire pour vérifier le cache distant. La mise en cache côté client peut être utilisée parallèlement à la mise en cache à distance pour une mise en cache à plusieurs niveaux afin d'améliorer la latence.

Étapes d'implémentation

1. Identifiez les bases de données, les API et les services réseau susceptibles de bénéficier de la mise en cache. Les services dont la charge de travail de lecture est importante, qui ont un ratio lecture/écriture élevé ou qui sont coûteux à adapter conviennent à la mise en cache.
   - Mise en cache de bases de données
   - Activez la mise en cache des API pour améliorer la réactivité.
2. Identifiez le type de stratégie de mise en cache le mieux adapté à votre modèle d'accès.
   - Stratégies de mise en cache
   - Solutions de mise en cache AWS

3. Suivez les bonnes pratiques en matière de mise en cache pour votre magasin de données.
4. Configurez une stratégie d'invalidation du cache, telle qu'une durée de vie (TTL), pour toutes les données afin d'équilibrer la fraîcheur des données et de réduire la pression qui pèse sur le magasin de données backend.
5. Activez des fonctionnalités telles que les nouvelles tentatives de connexion automatiques, le backoff exponentiel, les délais d'attente côté client et le regroupement des connexions dans le client, le cas échéant, car elles peuvent améliorer les performances et la fiabilité.
   - Bonnes pratiques : clients de Redis et Amazon ElastiCache for Redis
6. Surveillez le taux d'accès au cache en visant un objectif de 80 % ou plus. Des valeurs inférieures peuvent indiquer une taille de cache insuffisante ou un modèle d'accès qui ne bénéficie pas de la mise en cache.
   - Quelles métriques dois-je surveiller ?
   - Bonnes pratiques pour la surveillance des charges de travail Redis sur Amazon ElastiCache
   - Bonnes pratiques de surveillance avec Amazon ElastiCache for Redis via Amazon CloudWatch
7. Implémentez la réplication des données pour transférer les lectures vers plusieurs instances et améliorer les performances et la disponibilité de lecture des données.

Ressources

Documents connexes :

- Utilisation d'Amazon ElastiCache Well-Architected Lens
- Bonnes pratiques de surveillance avec Amazon ElastiCache for Redis via Amazon CloudWatch
- Quelles métriques dois-je surveiller ?
- Livre blanc « Performances à grande échelle avec Amazon ElastiCache »
- Défis et stratégies en matière de mise en cache

Vidéos connexes :

- Amazon ElastiCache Learning Path
- Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2020 - Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2023 - [LAUNCH] Introducing Amazon ElastiCache Serverless

- AWS re:Invent 2022 - 5 great ways to reimagine your data layer with Redis
- AWS re:Invent 2021 - Deep dive on Amazon ElastiCache for Redis

Exemples connexes :

- Améliorer les performances des bases de données MySQL avec Amazon ElastiCache for Redis

Mise en réseau et diffusion de contenu

La solution de mise en réseau optimale pour une charge de travail varie en fonction de la latence, des exigences de débit, de l'instabilité et de la bande passante. Le choix des options d'emplacement est tributaire des contraintes physiques telles que les ressources pour utilisateur ou sur site. Ces contraintes peuvent être compensées avec les emplacements périphériques ou le placement des ressources.

Sur AWS, la mise en réseau est virtualisée et disponible dans plusieurs types et configurations. Il est ainsi plus facile d'adapter vos méthodes de mise en réseau à vos besoins. AWS propose des fonctionnalités de produit (par exemple, la mise en réseau améliorée, les instances optimisées pour Amazon EC2, Amazon S3 Transfer Acceleration et Amazon CloudFront dynamique) pour optimiser le trafic réseau. AWS propose également des fonctionnalités de mise en réseau (par exemple, le routage de latence Amazon Route 53, les points de terminaison du Amazon VPC, AWS Direct Connect et AWS Global Accelerator) pour réduire l'instabilité ou la distance du réseau.

Ce domaine d'intérêt partage des conseils et des bonnes pratiques pour concevoir, configurer et exploiter des solutions de mise en réseau et de diffusion de contenu efficaces dans le cloud.

Bonnes pratiques

- PERF04-BP01 Compréhension de l’impact de la mise en réseau sur les performances
- PERF04-BP02 Évaluation des fonctionnalités de mise en réseau disponibles
- PERF04-BP03 Choix d’une connectivité dédiée ou d’un VPN approprié pour votre charge de travail
- PERF04-BP04 Utilisation de l’équilibrage de charge pour répartir le trafic entre plusieurs ressources
- PERF04-BP05 Choix de protocoles réseau afin d’améliorer les performances
- PERF04-BP06 Choix du placement de votre charge de travail en fonction des exigences réseau
- PERF04-BP07 Optimisation de la configuration réseau en fonction de métriques

PERF04-BP01 Compréhension de l’impact de la mise en réseau sur les performances

Analysez et comprenez l’impact des décisions liées au réseau sur votre charge de travail afin de fournir des performances efficaces et une meilleure expérience utilisateur.

Anti-modèles courants :

- Tout le trafic passe par vos centres de données existants.
- Vous acheminez l'ensemble du trafic via des pare-feux centralisés au lieu d'utiliser des outils de sécurité réseau natifs cloud.
- Vous configurez des connexions AWS Direct Connect sans connaître les exigences d'utilisation réelles.
- Vous ne tenez pas compte des caractéristiques de la charge de travail et de la surcharge de chiffrement lors de la définition de vos solutions de mise en réseau.
- Vous utilisez des concepts et des stratégies sur site pour les solutions de mise en réseau dans le cloud.

Avantages liés au respect de cette bonne pratique : Comprendre comment la mise en réseau affecte les performances de la charge de travail vous aidera à identifier les goulots d'étranglement potentiels, à améliorer l'expérience utilisateur, à accroître la fiabilité et à réduire la maintenance opérationnelle à mesure que la charge de travail évolue.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

Le réseau est responsable de la connectivité entre les composants d'application, les services cloud, les réseaux périphériques et les données sur site et, par conséquent, il peut avoir un impact majeur sur les performances de la charge de travail. Outre les performances de la charge de travail, l'expérience utilisateur peut également être affectée par la latence du réseau, la bande passante, les protocoles, l'emplacement, la congestion du réseau, l'instabilité, le débit et les règles de routage.

Avoir une liste documentée des exigences de mise en réseau de la charge de travail, y compris la latence, la taille des paquets, les règles de routage, les protocoles et les modèles de trafic pris en charge. Passez en revue les solutions de mise en réseau disponibles et identifiez le service qui répond aux caractéristiques de mise en réseau de votre charge de travail. Les réseaux basés sur le cloud peuvent être rapidement recréés. L'évolution de votre architecture réseau au fil du temps est donc nécessaire pour améliorer l'efficacité des performances.

Étapes d'implémentation :

1. Définissez et documentez les exigences de performance réseau, y compris les métriques tels que la latence du réseau, la bande passante, les protocoles, les emplacements, les modèles de trafic (pics et fréquence), le débit, le chiffrement, l'inspection et les règles de routage.
2. Découvrez les principaux services AWS de mise en réseau tels que les VPC, AWS Direct Connect, Elastic Load Balancing (ELB)et Amazon Route 53.
3. Capturez les principales caractéristiques réseau suivantes :

Caractéristiques	Outils et métriques
Caractéristiques de mise en réseau fondamentales	• Journaux de flux VPC • Journaux de flux AWS Transit Gateway • Métriques AWS Transit Gateway • Métriques AWS PrivateLink
Caractéristiques de mise en réseau des applications	• Elastic Fabric Adapter • Métriques AWS App Mesh • Métriques Amazon API Gateway
Caractéristiques de mise en réseau à la périphérie	• Métriques Amazon CloudFront • Métriques Amazon Route 53 • Métriques AWS Global Accelerator
Caractéristiques de mise en réseau hybride	• Métriques AWS Direct Connect • Métriques AWS Site-to-Site VPN • Métriques AWS Client VPN • Métriques AWS Cloud WAN
Caractéristiques de mise en réseau de la sécurité	• Métriques AWS Shield, AWS WAF et AWS Network Firewall
Caractéristiques de traçage	• AWS X-Ray • VPC Reachability Analyzer • Network Access Analyzer

Caractéristiques	Outils et métriques
	• Amazon Inspector • Amazon CloudWatch RUM

4. Définir des points de référence et et tester les performances du réseau :
   a. Évaluez le débit du réseau, car certains facteurs peuvent affecter les performances du réseau Amazon EC2 lorsque les instances se trouvent dans le même VPC. Mesurez la bande passante du réseau entre les instances Amazon EC2 Linux dans le même VPC.
   b. Effectuez des tests de charge pour expérimenter des solutions et des options de mise en réseau.

Ressources

Documents connexes :

- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Mise en réseau de produits avec AWS
- Transit Gateway
- Transition vers le routage basé sur la latence dans Amazon Route 53
- Points de terminaison d'un VPC

Vidéos connexes :

- AWS re:Invent 2023 - AWS networking foundations
- AWS re:Invent 2023 - What can networking do for your application?
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities

- AWS re:Invent 2023 - A developer’s guide to cloud networking
- AWS re:Invent 2019 - Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2019 - Optimizing Network Performance for Amazon EC2 Instances
- AWS Summit Online - Improve Global Network Performance for Applications
- AWS re:Invent 2020 - Networking best practices and tips with the Well-Architected Framework
- AWS re:Invent 2020 - AWS networking best practices in large-scale migrations

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS
- Atelier pratique sur le pare-feu réseau
- Observation et diagnostic de votre réseau sur AWS
- Détection et résolution des erreurs de configuration du réseau sur AWS

PERF04-BP02 Évaluation des fonctionnalités de mise en réseau disponibles

Évaluez les fonctions de mise en réseau dans le cloud qui peuvent améliorer les performances. Mesurez l’impact de ces fonctions au moyen de tests, de métriques et de l’analyse. Par exemple, tirez parti des fonctionnalités au niveau du réseau qui sont disponibles pour réduire la latence, la distance réseau ou l’instabilité.

Anti-modèles courants :

- Vous restez au sein d’une même région, car c’est là que votre siège social se trouve physiquement.
- Vous utilisez des pare-feux plutôt que des groupes de sécurité pour filtrer le trafic.
- Vous enfreignez le protocole TLS pour inspecter le trafic plutôt que de vous fier aux groupes de sécurité, aux politiques relatives aux points de terminaison et à d’autres fonctionnalités natives cloud.
- Vous utilisez uniquement la segmentation basée sur un sous-réseau au lieu des groupes de sécurité.

Avantages liés au respect de cette bonne pratique : L'évaluation de toutes les options et fonctionnalités de service peut augmenter les performances de vos charges de travail, baisser le coût d'infrastructure, réduire les efforts nécessaires à la maintenance de vos charges de travail et améliorer votre posture générale en matière de sécurité. Vous pouvez utiliser la couverture mondiale d'AWS pour fournir à vos clients une expérience de mise en réseau optimale.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Élevé

Directives d'implémentation

AWS propose des services comme AWS Global Accelerator et Amazon CloudFront qui contribuent à améliorer les performances du réseau, alors que la plupart des services AWS comportent des fonctionnalités de produit (telles que Amazon S3 Transfer Acceleration) pour optimiser le trafic réseau.

Examinez les options de configuration liées au réseau disponibles et leur impact potentiel sur votre charge de travail. L'optimisation des performances dépend de la compréhension de la manière dont ces options interagissent avec votre architecture et de l'impact qu'elles auront à la fois sur les performances mesurées et sur l'expérience utilisateur.

Étapes d'implémentation

- Créer une liste des composants de la charge de travail.
  - Envisagez d'utiliser AWS Cloud WAN pour créer, gérer et surveiller le réseau de votre organisation lors de la création d'un réseau mondial unifié.
  - Surveillez vos réseaux mondiaux et principaux avec les métriques Amazon CloudWatch Logs. Exploitez Amazon CloudWatch RUM, qui fournit des informations permettant d'identifier, de comprendre et d'améliorer l'expérience numérique des utilisateurs.
  - Visualisez la latence réseau globale entre les Régions AWS et les zones de disponibilité, et au sein de chaque zone de disponibilité, avec AWS Network Manager pour mieux comprendre comment les performances de votre application sont liées aux performances du réseau AWS sous-jacent.
  - Utilisez un outil ou un service de base de données de gestion de la configuration (CMDB) comme AWS Config pour créer un inventaire de votre charge de travail et de sa configuration.
- Identifier et documenter le test comparatif pour vos métriques de performances s'il s'agit d'une charge de travail existante, en vous concentrant sur les goulots d'étranglement et les zones à améliorer. Les métriques de mise en réseau liées aux performances diffèrent par charge de travail en fonction des exigences métier et des caractéristiques de charge de travail. Pour commencer,

il pourrait être important d'examiner ces métriques pour votre charge de travail : bande passante, latence, perte de paquets, instabilité et retransmissions.

- S'il s'agit d'une nouvelle charge de travail, réaliser des tests de charge pour identifier les goulots d'étranglement au niveau des performances.
- Concernant l'identification des goulots d'étranglement au niveau des performances, examiner les options de configuration pour les solutions afin d'identifier les opportunités d'amélioration des performances. Découvrez les principales options et fonctionnalités de mise en réseau suivantes :

Opportunité d'amélioration	Solution
Chemin ou itinéraires réseau	Utilisez Network Access Analyzer pour identifier des chemins ou des itinéraires.
Protocoles réseau	Voir PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances
Topologie du réseau	Évaluez vos compromis de performances et opérationnels entre Appairage des VPC et AWS Transit Gateway lors de la connexion de plusieurs comptes. AWS Transit Gateway simplifie la façon dont vous interconnectez tous vos VPC, qui peuvent s'étendre sur des milliers de Comptes AWS et sur les réseaux sur site. Partagez votre AWS Transit Gateway entre plusieurs comptes à l'aide de AWS Resource Access Manager. Voir PERF04-BP03 Choix d'une connectivité dédiée ou d'un VPN approprié pour votre charge de travail
Services de réseau	AWS Global Accelerator est un service qui améliore de 60 % les performances du trafic réseau de vos utilisateurs grâce à l'infrastructure réseau mondiale AWS. Amazon CloudFront contribue à améliorer les performances de votre charge de travail, de

Opportunité d’amélioration	Solution
	diffusion de contenu et de latence à l’échelle mondiale. Utilisez Lambda@edge pour exécuter des fonctions qui personnalisent le contenu diffusé par CloudFront au plus près des utilisateurs, réduire la latence et améliorer les performances. Amazon Route 53 offre des options de routage basé sur la latence, routage de géolocalisation, routage de proximité géographiqueet routage basé sur IP pour vous permettre d’améliorer les performances de votre charge de travail pour satisfaire un public international. Identifiez l’option de routage qui optimiserait les performances de votre charge de travail en examinant le trafic de votre charge de travail et la localisation des utilisateurs lorsque votre charge de travail est distribuée dans le monde entier.

Opportunité d’amélioration	Solution
Fonctionnalités des ressources de stockage	Amazon S3 Transfer Acceleration est une fonction qui permet aux utilisateurs externes de bénéficier des optimisations de mise en réseau de CloudFront pour charger des données dans Amazon S3. Cela améliore le transfert d’importants volumes de données à partir d’emplacements distants qui n’ont pas de connectivité dédiée au AWS Cloud. Les points d’accès multi-régions dans Amazon S3 répliquent le contenu vers plusieurs régions et simplifient la charge de travail en fournissant un point d’accès. Lorsqu’un point d’accès multi-région est utilisé, vous pouvez demander ou écrire des données à Amazon S3 tandis que le service identifie le compartiment à la latence la plus faible.

Opportunité d'amélioration	Solution
Fonctionnalités des ressources informatiques	Les interfaces réseau Elastic (ENI) utilisées par des instances Amazon EC2, des conteneurs et des fonctions Lambda sont limitées par flux. Examinez vos groupes de placement pour optimiser votre débit de mise en réseau EC2. Pour éviter un goulot d'étranglement par flux, créez votre applicati on pour qu'elle utilise plusieurs flux. Pour surveiller et disposer d'une visibilité sur vos métriques de mise en réseau liée au calcul, utilisez les métriques CloudWatch et ethtool. La commande `ethtool` est incluse dans le pilote ENA et expose des métriques liées au réseau supplémentaires qui peuvent être publiées en tant que métriques personnalisées sur CloudWatch. Les adaptateurs réseau élastiques (ENA) fournissent une optimisation supérieure en offrant un meilleur débit pour vos instances dans un groupe de placement du cluster. Un Elastic Fabric Adapter (EFA) est une interface réseau pour les instances Amazon EC2 qui vous permet d'exécuter des charges de travail nécessitant des niveaux élevés de communication entre les nœuds à grande échelle sur AWS. Les instances optimisées Amazon EBS utilisent une pile de configuration optimisée et offrent une capacité dédiée supplémentaire pour les E/S Amazon EBS.

Ressources

Documents connexes :

- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Networking Products with AWS
- Transition vers le routage basé sur la latence dans Amazon Route 53
- VPC Endpoints
- Journaux de flux VPC

Vidéos connexes :

- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2018 – Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS
- Observation et diagnostic de votre réseau
- Détection et résolution des erreurs de configuration du réseau sur AWS

PERF04-BP03 Choix d’une connectivité dédiée ou d’un VPN approprié pour votre charge de travail

Lorsque la connectivité hybride est requise pour connecter des ressources sur site et dans le cloud, allouez une bande passante adéquate pour répondre à vos exigences de performance. Estimez les exigences en matière de bande passante et de latence pour votre charge de travail hybride. Ces chiffres détermineront vos exigences en matière de dimensionnement.

Anti-modèles courants :

- Vous n’évaluez les solutions VPN que pour les exigences de chiffrement de votre réseau.
- Vous n’évaluez pas les options de sauvegarde ni de connectivité redondante.
- Vous n’identifiez pas toutes les exigences de la charge de travail (chiffrement, protocole, bande passante et trafic requis).

Avantages liés au respect de cette bonne pratique : La sélection et la configuration de solutions de connectivité appropriées renforcent la fiabilité de votre charge de travail et optimisent les performances. En identifiant les exigences de la charge de travail, en effectuant une planification appropriée et en évaluant les solutions hybrides, vous pouvez minimiser les modifications coûteuses du réseau physique et les frais généraux opérationnels tout en accélérant le délai de rentabilisation.

Niveau d’exposition au risque si cette bonne pratique n’est pas respectée : Élevé

Directives d’implémentation

Développez une architecture de mise en réseau hybride basée sur vos besoins en bande passante. AWS Direct Connect vous permet de connecter votre réseau sur site en privé à AWS. Cette solution convient lorsque vous avez besoin d’une bande passante élevée et d’une faible latence tout en conservant des performances constantes. Une connexion VPN établit une connexion sécurisée sur Internet. Elle sert uniquement lorsque seule une connexion temporaire est requise, lorsque le coût est un facteur, ou en cas d’urgence en attendant qu’une connectivité réseau physique résiliente soit établie lors de l’utilisation d’AWS Direct Connect.

Si vos besoins en bande passante sont élevés, vous pouvez envisager divers services AWS Direct Connect ou VPN. Le trafic peut être équilibré entre les services, mais nous ne recommandons pas l’équilibrage de charge entre AWS Direct Connect et le VPN en raison des différences de latence et de bande passante.

Étapes d'implémentation

1. Évaluez les besoins en bande passante et en latence de vos applications existantes.
   a. Pour les charges de travail existantes qui migrent vers AWS, exploitez les données de vos systèmes internes de surveillance du réseau.
   b. Pour les nouvelles charges de travail ou pour les charges de travail existantes pour lesquelles vous ne disposez pas de données de suivi, contactez les propriétaires du produit pour obtenir des métriques de performance adéquates et offrir une bonne expérience utilisateur.
2. Sélectionnez une connexion dédiée ou un VPN comme option de connectivité. En fonction de toutes les exigences de la charge de travail (chiffrement, bande passante et trafic requis), vous pouvez choisir AWS Direct Connect ou AWS VPN (ou les deux). Le schéma suivant peut vous aider à choisir le type de connexion approprié.
   a. AWS Direct Connect fournit une connectivité dédiée à l'environnement AWS, de 50 Mbit/s à 100 Gbit/s, en utilisant des connexions dédiées ou des connexions hébergées. Cela vous permet de gérer et de contrôler la latence et de profiter d'une bande passante provisionnée. Ainsi, vos charges de travail peuvent se connecter efficacement à d'autres environnements. Grâce aux partenaires AWS Direct Connect, vous bénéficiez d'une connectivité de bout en bout à partir de plusieurs environnements, ce qui vous permet de disposer d'un réseau étendu aux performances constantes. AWS offre une bande passante de connexion directe évolutive en utilisant soit le débit 100 Gbit/s natif, soit le protocole LAG (Link Aggregation Group), soit le protocole BGP ECMP (Equal-cost multipath).
   b. AWS Site-to-Site VPN fournit un service VPN géré prenant en charge le protocole de sécurité du protocole Internet (IPsec). Lorsqu'une connexion VPN est créée, chaque connexion VPN comprend deux tunnels pour une haute disponibilité.
3. Consultez la documentation AWS pour choisir l'option de connectivité appropriée :
   a. Si vous décidez d'utiliser AWS Direct Connect, sélectionnez la bande passante adaptée à votre connectivité.
   b. Si vous utilisez un AWS Site-to-Site VPN sur plusieurs emplacements pour vous connecter à une Région AWS, utilisez une connexion Site-to-Site VPN accélérée pour pouvoir améliorer les performances du réseau.
   c. Si la conception de votre réseau consiste en une connexion VPN IPsec via AWS Direct Connect, pensez à utiliser un VPN IP privé pour améliorer la sécurité et réaliser une segmentation. Un VPN IP privé AWS Site-to-Site est déployé au-dessus de l'interface virtuelle de transport (VIF).

   d. AWS Direct Connect SiteLink permet de créer des connexions redondantes et à faible latence entre vos centres de données à travers le monde en envoyant les données sur le chemin le plus rapide entre les emplacements AWS Direct Connect, en contournant les Régions AWS.
4. Validez votre configuration de connectivité avant le déploiement en production. Effectuez des tests de sécurité et de performance pour vous assurer qu'elle répond à vos exigences en matière de bande passante, de fiabilité, de latence et de conformité.
5. Surveillez régulièrement les performances et l'utilisation de votre connectivité et optimisez-les si nécessaire.

Organigramme des performances déterministes

Ressources

Documents connexes :

- Networking Products with AWS
- AWS Transit Gateway
- VPC Endpoints
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure
- Client VPN

Vidéos connexes :

- AWS re:Invent 2023 – Building hybrid network connectivity with AWS
- AWS re:Invent 2023 – Secure remote connectivity to AWS
- AWS re:Invent 2022 – Optimizing performance with Amazon CloudFront
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2020 – AWS Transit Gateway Connect

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP04 Utilisation de l’équilibrage de charge pour répartir le trafic entre plusieurs ressources

Répartissez le trafic sur plusieurs ressources ou services pour permettre à votre charge de travail de tirer parti de l’élasticité fournie par le cloud. Vous pouvez également utiliser l’équilibrage de charge afin de décharger la terminaison du chiffrement en vue d’améliorer les performances, d’assurer la fiabilité et de gérer et acheminer efficacement le trafic.

Anti-modèles courants :

- Vous ne tenez pas compte des exigences de votre charge de travail lorsque vous choisissez le type d’équilibreur de charge.
- Vous ne tirez pas parti des fonctions d’équilibrage de charge pour optimiser les performances.
- La charge de travail est exposée directement à Internet sans équilibreur de charge.
- Vous acheminez tout le trafic Internet via des équilibreurs de charge existants.
- Vous utilisez l’équilibrage de charge TCP générique et faites en sorte que chaque nœud de calcul gère le chiffrement SSL.

Avantages liés au respect de cette bonne pratique : un équilibreur de charge gère la charge variable du trafic de vos applications dans une zone de disponibilité unique ou entre plusieurs zones de disponibilité et permet une haute disponibilité, une mise à l’échelle automatique et une meilleure utilisation de votre charge de travail.

Niveau de risque exposé si cette bonne pratique n’est pas établie : élevé

Directives d’implémentation

Les équilibreurs de charge constituent le point d’entrée de votre charge de travail, à partir duquel ils distribuent le trafic vers vos cibles principales, telles que les instances de calcul ou les conteneurs, afin d’améliorer l’utilisation.

Le choix du bon type d’équilibreur de charge est la première étape de l’optimisation de votre architecture. Commencez par énumérer les caractéristiques de votre charge de travail, telles que le protocole (TCP, HTTP, TLS ou WebSockets), le type de cible (instances, conteneurs ou sans serveur), les exigences de l’application (connexions de longue durée, authentification de l’utilisateur ou permanence) et le placement (région, zone locale, Outpost ou isolement de zone).

AWS fournit plusieurs modèles permettant à vos applications d'utiliser l'équilibrage de charge. Application Load Balancer convient parfaitement pour l'équilibrage de charge du trafic HTTP et HTTPS, et fournit un routage avancé des demandes, axé sur la diffusion d'architectures d'application modernes, notamment de microservices et de conteneurs.

Network Load Balancer convient parfaitement pour l'équilibrage de charge du trafic TCP, qui nécessite des performances extrêmes. Il est capable de traiter des millions de requêtes par seconde tout en maintenant de très faibles latences. Il est optimisé pour gérer les tendances soudaines et instables du trafic.

Elastic Load Balancing assure la gestion intégrée des certificats et le déchiffrement SSL/TLS, ce qui vous permet de gérer de façon centralisée les paramètres SSL de l'équilibreur de charge et de décharger les tâches gourmandes en CPU de votre charge de travail.

Après avoir choisi le bon équilibreur de charge, vous pouvez commencer à tirer parti de ses fonctionnalités pour réduire les efforts que votre système backend doit fournir pour servir le trafic.

Par exemple, en utilisant à la fois Application Load Balancer (ALB) et Network Load Balancer (NLB), vous pouvez effectuer un déchargement du chiffrement SSL/TLS. Cela permet d'éviter que la liaison TLS, très gourmande en ressources CPU, ne soit effectuée par vos cibles, et permet également d'améliorer la gestion des certificats.

Lorsque vous configurez le déchargement SSL/TLS dans votre équilibreur de charge, celui-ci se charge du chiffrement du trafic en provenance et à destination des clients, tout en acheminant le trafic non chiffré vers vos systèmes backend. Cela libère vos ressources backend et améliore le temps de réponse pour les clients.

Application Load Balancer peut également servir le trafic HTTP/2 sans avoir besoin de le prendre en charge sur vos cibles. Cette simple décision peut améliorer le temps de réponse de votre application, car HTTP/2 utilise plus efficacement les connexions TCP.

Les exigences de latence de votre charge de travail doivent être prises en compte lors de la définition de l'architecture. Par exemple, si vous avez une application sensible à la latence, vous pouvez décider d'utiliser Network Load Balancer, qui offre des latences extrêmement faibles. Vous pouvez également décider de rapprocher votre charge de travail de vos clients en tirant parti d'Application Load Balancer dans les zones locales AWS ou même AWS Outposts.

L'équilibrage de charge entre zones est un autre élément à prendre en compte pour les charges de travail sensibles à la latence. Avec l'équilibrage de charge inter-zone, chaque nœud d'équilibreur de charge distribue le trafic sur les cibles enregistrées dans toutes les zones de disponibilité activées.

Intégrez Auto Scaling à votre équilibreur de charge. L'un des aspects essentiels d'un système performant est le dimensionnement adéquat de vos ressources backend. Pour ce faire, vous pouvez tirer parti des intégrations d'équilibreurs de charge pour les ressources cibles du système backend. Grâce à l'intégration de l'équilibreur de charge avec les groupes Auto Scaling, les cibles seront ajoutées ou retirées de l'équilibreur de charge selon les besoins en fonction du trafic entrant. Les équilibreurs de charge peuvent également s'intégrer à Amazon ECS et Amazon EKS pour les charges de travail conteneurisées.

- Amazon ECS : équilibrage de charge des services
- Équilibrage de charge d'application sur Amazon EKS
- Équilibrage de la charge réseau sur Amazon EKS

Étapes d'implémentation

- Définissez vos exigences en matière d'équilibrage de charge, notamment en termes de volume de trafic, de disponibilité et de capacité de mise à l'échelle des applications.
- Choisissez le type d'équilibreur de charge adapté à votre application.
  - Utilisez Application Load Balancer pour les charges de travail HTTP/HTTPS.
  - Utilisez Network Load Balancer pour les charges de travail non HTTP qui fonctionnent sur TCP ou UDP.
  - Utilisez une combinaison des deux (ALB comme cible de NLB) si vous souhaitez tirer parti des fonctionnalités des deux produits. Par exemple, vous pouvez le faire si vous voulez utiliser les adresses IP statiques de NLB avec le routage basé sur l'en-tête HTTP d'ALB, ou si vous voulez exposer votre charge de travail HTTP à AWS PrivateLink.
  - Pour une comparaison complète des équilibreurs de charge, consultez Comparaison des produits ELB.
- Utilisez le déchargement SSL/TLS si possible.
  - Configurez des écouteurs HTTPS/TLS avec Application Load Balancer et Network Load Balancer, tous deux intégrés à AWS Certificate Manager.
  - Notez que certaines charges de travail peuvent nécessiter un chiffrement de bout en bout pour des raisons de conformité. Dans ce cas, il est nécessaire de permettre le chiffrement au niveau des cibles.
  - Pour découvrir les bonnes pratiques de sécurité, consultez SEC09-BP02 Application du chiffrement en transit.

- Sélectionnez le bon algorithme de routage (ALB uniquement).
  - L'algorithme de routage peut faire une réelle différence dans la manière d'utiliser vos cibles backend et donc dans leur impact sur les performances. Par exemple, ALB propose deux options pour les algorithmes de routage :
  - Demandes les moins en attente : permet de mieux répartir la charge sur vos cibles backend lorsque les demandes de votre application varient en complexité ou que vos cibles ont des capacités de traitement différentes.
  - Tourniquet : utilisez cette méthode lorsque les demandes et les cibles sont similaires, ou si vous devez distribuer les demandes de manière égale entre les cibles.
- Envisagez un isolement inter-zone ou par zone.
  - Utilisez les zones croisées désactivées (isolement par zone) pour améliorer la latence et les domaines de panne par zone. L'option est désactivée par défaut dans NLB et, dans ALB, vous pouvez la désactiver par groupe cible.
  - Utilisez les zones croisées activées pour une disponibilité et une flexibilité accrues. Par défaut, les zones croisées sont activées pour ALB et, dans NLB, vous pouvez les activer par groupe cible.
- Activez l'option de persistance HTTP pour vos charges de travail HTTP (ALB uniquement). Grâce à cette fonction, l'équilibreur de charge peut réutiliser les connexions backend jusqu'à l'expiration du délai de persistance, ce qui améliore les temps de demande et de réponse HTTP et réduit également l'utilisation des ressources sur vos cibles backend. Pour plus de détails sur la façon de procéder pour Apache et Nginx, consultez Quels sont les paramètres optimaux pour utiliser Apache ou NGINX en tant que serveur principal pour ELB ?
- Activez la surveillance pour votre équilibreur de charge.
  - Activez les journaux d'accès pour vos Application Load Balancer et Network Load Balancer.
  - Les principaux champs à prendre en compte pour ALB sont `request_processing_time`, `request_processing_time` et `response_processing_time`.
  - Les principaux champs à prendre en compte pour NLB sont `connection_time` et `tls_handshake_time`.
  - Soyez prêt à interroger les journaux lorsque vous en aurez besoin. Vous pouvez utiliser Amazon Athena pour interroger les journaux ALB et les journaux NLB.
  - Créez des alarmes pour les métriques liées aux performances, telles que `TargetResponseTime` pour ALB.

Ressources

Documents connexes :

- Comparaison des produits ELB
- Infrastructure mondiale AWS
- Improving Performance and Reducing Cost Using Availability Zone Affinity
- Procédure détaillée de l'analyse des journaux avec Amazon Athena
- Querying Application Load Balancer logs
- Monitor your Application Load Balancers
- Monitor your Network Load Balancer
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group

Vidéos connexes :

- AWS re:Invent 2023: What can networking do for your application?
- AWS re:Inforce 20: How to use Elastic Load Balancing to enhance your security posture at scale
- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Exemples connexes :

- Gateway Load Balancer
- Exemples de CDK et AWS CloudFormation pour l'analyse des journaux avec Amazon Athena

PERF04-BP05 Choix de protocoles réseau afin d'améliorer les performances

Prenez des décisions concernant les protocoles de communication entre les systèmes et les réseaux en fonction de l'impact sur les performances de la charge de travail.

Il existe une relation entre la latence et la bande passante pour atteindre le débit. Si votre transfert de fichiers utilise le protocole de contrôle de transmission (TCP), des latences plus élevées réduiront

très probablement le débit global. Il existe des approches pour résoudre ce problème avec le réglage du protocole TCP et les protocoles de transfert optimisés. Le protocole UDP (User Datagram Protocol) est une solution possible.

Anti-modèles courants :

- Vous utilisez TCP pour toutes les charges de travail, quelles que soient les exigences de performance.

Avantages liés au respect de cette bonne pratique : Vérifiez que vous utilisez un protocole approprié pour la communication entre les utilisateurs et les composants de la charge de travail, afin d'améliorer l'expérience globale des utilisateurs de vos applications. Par exemple, le protocole UDP sans connexion permet d'obtenir une vitesse élevée, mais sans retransmission ni fiabilité élevée. Quoique complet, le protocole TCP nécessite une surcharge plus importante pour le traitement des paquets.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Moyen

Directives d'implémentation

Si vous avez la possibilité de choisir différents protocoles pour votre application et que vous possédez l'expertise nécessaire dans ce domaine, optimisez votre application et l'expérience de l'utilisateur final en utilisant un autre protocole. Notez que cette approche présente des difficultés importantes et ne doit être tentée que si vous avez d'abord optimisé votre application à d'autres égards.

Pour améliorer les performances de votre charge de travail, il est essentiel de comprendre les exigences en matière de latence et de débit, puis de choisir des protocoles réseau qui optimisent les performances.

Quand envisager l'utilisation du protocole TCP

Le protocole TCP assure une livraison fiable des données et peut être utilisé pour la communication entre les composants de la charge de travail où la fiabilité et la livraison garantie des données sont importantes. De nombreuses applications web reposent sur des protocoles basés sur le protocole TCP, tels que HTTP et HTTPS, pour ouvrir des sockets TCP pour la communication entre les composants de l'application. Les e-mails et le transfert de données de fichiers sont des applications courantes qui utilisent également le protocole TCP, car il s'agit d'un mécanisme de transfert simple et fiable entre les composants de l'application. L'utilisation de TLS avec TCP peut ajouter une certaine

surcharge à la communication, ce qui peut entraîner une augmentation de la latence et une réduction du débit, mais elle présente l'avantage de la sécurité. La surcharge provient principalement de la charge supplémentaire du processus de liaison, qui peut prendre plusieurs allers-retours pour se terminer. Une fois la liaison établie, la charge de chiffrement et de déchiffrement des données devient relativement faible.

Quand envisager l'utilisation du protocole UDP

UDP est un protocole orienté sans connexion et convient donc aux applications qui nécessitent une transmission rapide et efficace, comme les données de journal, de surveillance et de VoIP. En outre, envisagez d'utiliser UDP si vous avez des composants de charge de travail qui répondent à de petites requêtes provenant d'un grand nombre de clients, afin de garantir des performances optimales de la charge de travail. Le protocole DTLS (Datagram Transport Layer Security) est l'équivalent UDP du protocole TLS (Transport Layer Security). Lors de l'utilisation de DTLS avec UDP, la charge provient du chiffrement et du déchiffrement des données, car le processus de liaison est simplifié. DTLS ajoute également une petite quantité de charge aux paquets UDP, car il inclut des champs supplémentaires pour indiquer les paramètres de sécurité et pour détecter la falsification.

Quand envisager l'utilisation du protocole SRD

Le protocole SRD (Scalable reliable datagram) est un protocole de transport en réseau optimisé pour les charges de travail à haut débit en raison de sa capacité à répartir le trafic sur plusieurs chemins et à se rétablir rapidement en cas de perte de paquets ou de défaillance d'un lien. Le SRD est donc le mieux adapté aux charges de travail du calcul haute performance (HPC) qui nécessitent un débit élevé et une communication à faible latence entre les nœuds de calcul. Il peut s'agir de tâches de traitement parallèle telles que la simulation, la modélisation et l'analyse de données qui impliquent un grand nombre de transferts de données entre les nœuds.

Étapes d'implémentation

1. Utilisez les services AWS Global Accelerator et AWS Transfer Family pour améliorer le débit de vos applications de transfert de fichiers en ligne. Le service AWS Global Accelerator vous aide à réduire la latence entre vos appareils clients et votre charge de travail sur AWS. Avec AWS Transfer Family, vous pouvez utiliser des protocoles basés sur TCP tels que le protocole de transfert de fichiers Secure Shell (SFTP) et le protocole de transfert de fichiers sur SSL (FTPS) pour dimensionner et gérer en toute sécurité vos transferts de fichiers vers des services de stockage AWS.
2. Utilisez la latence du réseau pour déterminer si le protocole TCP est adapté à la communication entre les composants de la charge de travail. Si la latence du réseau entre votre application

client et le serveur est élevée, la liaison tripartite TCP peut prendre un certain temps, ce qui a un impact sur la réactivité de votre application. Des métriques telles que le délai jusqu'au premier octet (TTFB) et le temps de propagation aller-retour (RTT) peuvent être utilisées pour mesurer la latence du réseau. Si votre charge de travail fournit du contenu dynamique aux utilisateurs, pensez à utiliser Amazon CloudFront, qui établit une connexion persistante avec chaque origine pour le contenu dynamique afin d'éliminer le temps d'établissement de la connexion qui, autrement, ralentirait chaque demande du client.

3. L'utilisation de TLS avec TCP ou UDP peut entraîner une augmentation de la latence et une réduction du débit de votre charge de travail en raison de l'impact du chiffrement et du déchiffrement. Pour de telles charges de travail, envisagez d'activer le déchargement SSL/TLS sur Elastic Load Balancing pour améliorer les performances de la charge de travail en permettant à l'équilibreur de charge de gérer les processus de chiffrement et de déchiffrement SSL/TLS au lieu de laisser les instances backend s'en charger. Cela peut contribuer à réduire l'utilisation du CPU sur les instances backend, ce qui peut améliorer les performances et augmenter la capacité.
4. Utilisez le Network Load Balancer (NLB) pour déployer des services qui reposent sur le protocole UDP, tels que l'authentification et l'autorisation, la journalisation, le DNS, l'IoT et le streaming média, afin d'améliorer les performances et la fiabilité de votre charge de travail. Le NLB distribue le trafic UDP entrant sur plusieurs cibles, ce qui vous permet de faire évoluer votre charge de travail horizontalement, d'augmenter la capacité et de réduire les frais généraux associés à une seule cible.
5. Pour vos charges de travail liées au calcul haute performance (HPC), pensez à utiliser la fonctionnalité Elastic Network Adapter (ENA) Express qui utilise le protocole SRD pour améliorer les performances du réseau en fournissant une bande passante à flux unique plus élevée (25 Gbit/s) et une latence de queue plus faible (99,9 centile) pour le trafic réseau entre les instances EC2.
6. Utilisez l' Application Load Balancer (ALB) pour router et répartir votre trafic gRPC (Remote Procedure Calls) entre les composants de la charge de travail ou entre les clients et les services gRPC. gRPC utilise le protocole HTTP/2 basé sur TCP pour le transport et offre des avantages en termes de performances, tels qu'une empreinte réseau plus légère, la compression, une sérialisation binaire efficace, la prise en charge de nombreux langages et le streaming bidirectionnel.

Ressources

Documents connexes :

- How to route UDP traffic into Kubernetes

- Application Load Balancer
- Mise en réseau améliorée d'EC2 sous Linux
- Capacité réseau améliorée d'EC2 sous Windows
- Groupes de placement EC2
- Activation de la mise en réseau améliorée avec un adaptateur réseau élastique (ENA) sur les instances de Linux
- Network Load Balancer
- Networking Products with AWS
- Transitioning to Latency-Based Routing in Amazon Route 53
- VPC Endpoints

Vidéos connexes :

- AWS re:Invent 2022 – Scaling network performance on next-gen Amazon Elastic Compute Cloud instances
- AWS re:Invent 2022 – Application networking foundations

Exemples connexes :

- AWS Transit Gateway et solutions de sécurité évolutives
- Ateliers sur la mise en réseau AWS

PERF04-BP06 Choix du placement de votre charge de travail en fonction des exigences réseau

Évaluez les options de placement des ressources afin de réduire la latence du réseau et d'améliorer le débit, offrant ainsi une expérience utilisateur optimale en réduisant les temps de chargement des pages et de transfert des données.

Anti-modèles courants :

- Vous regroupez toutes les ressources de charge de travail dans un seul emplacement géographique.

- Vous avez choisi la région la plus proche de votre emplacement, pas celle de l'utilisateur final de la charge de travail.

Avantages liés au respect de cette bonne pratique : l'expérience utilisateur est fortement affectée par la latence entre l'utilisateur et votre application. En utilisant les Régions AWS appropriées et le réseau mondial AWS privé, vous pouvez réduire le temps de latence et offrir une meilleure expérience aux utilisateurs distants.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Les ressources, telles que les instances Amazon EC2, sont placées dans des zones de disponibilité au sein des Régions AWS, des zones locales AWS, de AWS Outposts ou des zones AWS Wavelength. Le choix de cet emplacement influence la latence et le débit du réseau à partir d'un emplacement donné de l'utilisateur. Les services périphériques tels que Amazon CloudFront et AWS Global Accelerator peuvent également améliorer les performances réseau, soit en mettant en cache du contenu aux emplacements périphériques, soit en fournissant aux utilisateurs un chemin optimal vers la charge de travail via le réseau mondial AWS.

Amazon EC2 fournit des groupes de placement pour la mise en réseau. Un groupe de placement est un regroupement logique d'instances permettant de réduire la latence. L'utilisation de groupes de placement avec des types d'instance pris en charge et un adaptateur réseau élastique (ENA) permet aux charges de travail de participer à un réseau 25 Gbit/s à faible latence avec une instabilité réduite. Les groupes de placement sont recommandés pour les charges de travail nécessitant une latence réseau faible, un débit réseau élevé, ou les deux.

Les services sensibles à la latence sont fournis sur des emplacements périphériques via le réseau mondial AWS, comme Amazon CloudFront. Ces emplacements périphériques fournissent généralement des services tels que les réseaux de diffusion de contenu (CDN) et les systèmes de noms de domaine (DNS). Placer ces services en périphérie permet aux charges de travail de répondre avec une faible latence aux requêtes de contenu ou de résolution DNS. Ces services fournissent également des services géographiques tels que le ciblage géographique du contenu (qui fournit des contenus différents en fonction de l'emplacement des utilisateurs finaux) ou le routage en fonction de la latence pour diriger les utilisateurs finaux vers la région plus proche (latence minimum).

Utilisez des services en périphérie pour réduire la latence et permettre la mise en cache de contenu. Configurez correctement le contrôle du cache pour les services DNS et HTTP/HTTPS afin de tirer le plus grand bénéfice de ces approches.

Étapes d'implémentation

- Capturez des informations sur le trafic IP entrant et sortant des interfaces réseau.
  - Enregistrement du trafic IP à l'aide des journaux de flux VPC
  - Comment l'adresse IP du client est préservée dans AWS Global Accelerator
- Analysez les modèles d'accès au réseau dans votre charge de travail afin d'identifier comment les utilisateurs utilisent votre application.
  - Utilisez des outils de surveillance, tels que Amazon CloudWatch et AWS CloudTrail, pour recueillir des données sur les activités du réseau.
  - Analysez les données pour identifier le modèle d'accès au réseau.
- Choisissez les régions pour le déploiement de votre charge de travail en fonction des éléments clés suivants :
  - Emplacement de vos données : pour les applications utilisant de grandes quantités de données (telles que le big data et le machine learning), le code des applications doit s'exécuter aussi près que possible des données.
  - Emplacement de vos utilisateurs : pour les applications orientées utilisateur, choisissez une ou plusieurs régions proches des utilisateurs de votre charge de travail.
  - Autres contraintes : tenez compte de contraintes telles que le coût et la conformité, comme expliqué dans Éléments à prendre en compte lors de la sélection d'une région pour vos charges de travail.
- Utilisez les zones locales AWS pour exécuter des charges de travail telles que le rendu vidéo. Ces zones locales vous permettent de profiter des avantages liés à la présence de ressources de calcul et de stockage plus proches des utilisateurs finaux.
- Utilisez AWS Outposts pour les charges de travail qui doivent rester sur site et pour lesquelles vous souhaitez qu'elles fonctionnent de manière transparente avec le reste de vos charges de travail dans AWS.
- Les applications telles que le streaming vidéo en direct à haute résolution, l'audio haute fidélité et la réalité augmentée ou virtuelle (RA/RV) exigent une latence ultra-faible pour les appareils 5G. Pour ces applications, pensez à AWS Wavelength. AWS Wavelength intègre les services de calcul et de stockage d'AWS dans les réseaux 5G, fournissant ainsi une infrastructure informatique de périphérie mobile pour le développement, le déploiement et la mise à l'échelle d'applications à très faible latence.

- Utilisez la mise en cache locale ou les solutions de mise en cache AWS pour les ressources fréquemment utilisées afin d'améliorer les performances, de réduire les déplacements de données et de diminuer l'impact environnemental.

Service	When to use
Amazon CloudFront	Permet de mettre en cache du contenu statique comme des images, des scripts et des vidéos, ainsi que du contenu dynamique comme des réponses API ou des applications Web.
Amazon ElastiCache	Permet de mettre en cache du contenu pour les applications Web.
DynamoDB Accelerator	Permet d'ajouter une accélération en mémoire à vos tables DynamoDB.

- Utilisez des services capables de vous aider à exécuter le code plus près des utilisateurs de votre charge de travail, tels que les suivants :

Service	When to use
Lambda@edge	Destiné aux opérations exigeantes en puissance de calcul qui sont lancées lorsque des objets ne sont pas dans le cache.
Amazon CloudFront Functions	Destiné aux cas d'utilisation simples comme les requêtes HTTP(S) ou les manipulations de réponse pouvant être lancées par des fonctions de courte durée.
AWS IoT Greengrass	Permet d'exécuter du calcul local, une messagerie et une mise en cache de données pour les appareils connectés.

- Certaines applications nécessitent des points d'entrée fixes ou des performances plus élevées en réduisant la latence et l'instabilité du premier octet et en augmentant le débit. Ces applications

peuvent bénéficier de services de mise en réseau qui fournissent des adresses IP statiques anycast et une terminaison TCP aux emplacements périphériques. AWS Global Accelerator peut améliorer les performances de vos applications jusqu'à 60 % et assurer un basculement rapide pour les architectures multirégionales. AWS Global Accelerator vous fournit des adresses IP statiques anycast qui servent de point d'entrée fixe pour vos applications hébergées dans une ou plusieurs Régions AWS. Ces adresses IP permettent au trafic d'entrer sur le réseau global AWS aussi près que possible de vos utilisateurs. AWS Global Accelerator réduit le temps d'établissement de la connexion initiale en établissant une connexion TCP entre le client et l'emplacement périphérique AWS le plus proche du client. Examinez l'utilisation de AWS Global Accelerator pour améliorer les performances de vos charges de travail TCP/UDP et fournir un basculement rapide pour les architectures multirégionales.

Ressources

Bonnes pratiques associées :

- COST07-BP02 Mettre en œuvre des régions en fonction des coûts
- COST08-BP03 Mettre en œuvre des services pour réduire les coûts de transfert de données
- REL10-BP01 Déployer la charge de travail sur plusieurs emplacements
- REL10-BP02 Sélectionner les emplacements appropriés pour votre déploiement multisite
- SUS01-BP01 Choix d'une région en fonction des exigences et des objectifs de durabilité de l'entreprise
- SUS02-BP04 Optimisation du placement géographique des charges de travail en fonction de leurs exigences réseau
- SUS04-BP07 Réduire le mouvement des données entre les réseaux

Documents connexes :

- Infrastructure mondiale AWS
- AWS Local Zones and AWS Outposts, choosing the right technology for your edge workload (Zones locales AWS et Outpost AWS, choisir la bonne technologie pour votre charge de travail périphérique)
- Groupes de placement
- Zones locales AWS
- AWS Outposts

- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Vidéos connexes :

- Vidéo explicative sur les zones locales AWS
- AWS Outposts : Overview and How it Works
- AWS re:Invent 2023 : A migration strategy for edge and on-premises workloads
- AWS re:Invent 2021 - AWS Outposts: Bringing the AWS experience on premises
- AWS re:Invent 2020: AWS Wavelength: Run apps with ultra-low latency at 5G edge
- AWS re:Invent 2022 - AWS Local Zones: Building applications for a distributed edge
- AWS re:Invent 2021 - Building low-latency websites with Amazon CloudFront
- AWS re:Invent 2022 - Improve performance and availability with AWS Global Accelerator
- AWS re:Invent 2022 - Build your global wide area network using AWS
- AWS re:Invent 2020: Global traffic management with Amazon Route 53

Exemples connexes :

- AWS Global Accelerator Custom Routing Workshop
- Handling Rewrites and Redirects using Edge Functions (Gestion des réécritures et des redirections à l'aide des fonctions de périphérie)

PERF04-BP07 Optimisation de la configuration réseau en fonction de métriques

Utilisez les données collectées et analysées pour prendre des décisions avisées concernant l'optimisation de votre configuration réseau.

Anti-modèles courants :

- Vous supposez que tous les problèmes liés aux performances sont liés à l'application.
- Vous testez uniquement les performances de votre réseau à partir d'un emplacement proche de l'endroit où vous avez déployé la charge de travail.
- Vous utilisez des configurations par défaut pour tous les services du réseau.
- Vous surdimensionnez la ressource réseau afin de fournir une capacité suffisante.

Avantages liés au respect de cette bonne pratique : La collecte des métriques nécessaires de votre réseau AWS et la mise en œuvre d'outils de surveillance du réseau vous permettent de comprendre les performances du réseau et d'optimiser les configurations du réseau.

Niveau d'exposition au risque si cette bonne pratique n'est pas respectée : Faible

Directives d'implémentation

La surveillance du trafic en provenance et à destination des VPC, des sous-réseaux ou des interfaces réseau est essentielle pour comprendre comment utiliser les ressources réseau AWS et comment optimiser les configurations réseau. Les outils de mise en réseau AWS suivants vous permettent d'obtenir des informations supplémentaires sur l'utilisation du trafic, l'accès au réseau et les journaux.

Étapes d'implémentation

- Identifiez les indicateurs clés de performance tels que la latence ou la perte de paquets à collecter. AWS fournit plusieurs outils qui peuvent vous aider à collecter ces métriques. Les outils suivants vous permettent d'obtenir des informations supplémentaires sur l'utilisation du trafic, l'accès au réseau et les journaux.

Outil AWS	Où utiliser
Amazon VPC IP Address Manager.	Utilisez IPAM pour planifier, suivre et surveille r les adresses IP pour vos charges de travail AWS et sur site. Il s'agit d'une bonne pratique pour optimiser l'utilisation et l'allocation des adresses IP.
Journaux de flux VPC	Utilisez les journaux de flux VPC pour capturer des informations détaillées sur le trafic en provenance et à destination des interface

Outil AWS	Où utiliser
	s réseau de vos VPC. Grâce aux journaux de flux VPC, vous pouvez diagnostiquer les règles de groupes de sécurité trop restrictives ou trop permissives et déterminer la direction du trafic vers et depuis les interfaces réseau.
Journaux de flux AWS Transit Gateway	Utilisez les journaux de flux AWS Transit Gateway pour capturer des informations sur le trafic IP à destination et en provenance de vos passerelles de transit.
Journalisation des requêtes DNS	Enregistrez les informations relatives aux requêtes DNS publiques ou privées reçues par Route 53. Grâce aux journaux DNS, vous pouvez optimiser les configurations DNS en comprenant le domaine ou le sous-doma ine qui a été demandé ou les emplacements périphériques Route 53 qui ont répondu aux requêtes DNS.
Reachability Analyzer	Utilisez Reachability Analyzer pour analyser et déboguer l’accessibilité du réseau. Reachabil ity Analyzer est un outil d’analyse de la configuration qui vous permet d’effectuer des tests de connectivité entre une ressource source et une ressource de destination dans vos VPC. Cet outil vous aide à vérifier que votre configuration réseau correspond à la connectivité souhaitée.

Outil AWS	Où utiliser
Network Access Analyzer	Network Access Analyzer vous aide à comprendre l'accès réseau à vos ressources. Vous pouvez utiliser Network Access Analyzer pour spécifier vos exigences en matière d'accès au réseau et identifier les chemins d'accès potentiels qui ne répondent pas à vos exigences spécifiées. En optimisant la configuration de votre réseau correspondant, vous pouvez comprendre et vérifier l'état de votre réseau et démontrer si votre réseau sur AWS répond à vos exigences de conformité.
Amazon CloudWatch	Utilisez Amazon CloudWatch et activez les métriques appropriées pour les options réseau. Veillez à choisir la métrique de réseau adaptée à votre charge de travail. Par exemple, vous pouvez activer des métriques pour l'utilisation d'adresses réseau VPC, la passerelle VPC NAT, AWS Transit Gateway, le tunnel VPN, AWS Network Firewall, Elastic Load Balancing et AWS Direct Connect. La surveillance continue des métriques est une bonne pratique pour observer et comprendre l'état et l'utilisation de votre réseau. Elle vous aide à optimiser la configuration du réseau en fonction de vos observations.

Outil AWS	Où utiliser
AWS Network Manager	AWS Network Manager permet de surveiller les performances historiques et en temps réel du réseau mondial AWS à des fins opérationnelles et de planification. Network Manager fournit une latence réseau globale entre les Régions AWS et les zones de disponibilité et au sein de chaque zone de disponibilité, ce qui vous permet de mieux comprendre le lien entre les performances de votre application et les performances du réseau AWS sous-jacent.
Amazon CloudWatch RUM	Utilisez Amazon CloudWatch RUM pour collecter les métriques fournissant les informations qui vous aideront à identifier, à comprendre et à améliorer l'expérience utilisateur.

- Identifiez les principaux intervenants et les modèles de trafic des applications à l'aide des journaux de flux VPC et AWS Transit Gateway.
- Évaluez et optimisez votre architecture réseau actuelle, y compris les VPC, les sous-réseaux et le routage. À titre d'exemple, vous pouvez évaluer l'impact de l'appairage de VPC ou d'AWS Transit Gateway sur l'amélioration de la mise en réseau de votre architecture.
- Évaluez les chemins de routage de votre réseau pour vérifier que le chemin le plus court entre les destinations est toujours utilisé. Network Access Analyzer peut vous aider à le faire.

Ressources

Documents connexes :

- Journalisation des requêtes DNS publiques
- Qu'est-ce qu'IPAM ?
- Qu'est-ce qu'Reachability Analyzer ?
- Qu'est-ce qu'Network Access Analyzer ?
- Métriques CloudWatch pour vos VPC

- Optimiser les performances et réduire les coûts analytiques des réseaux grâce aux journaux de flux VPC au format Apache Parquet
- Surveillance de vos réseaux mondiaux et principaux avec les métriques Amazon CloudWatch
- Surveiller en permanence le trafic et les ressources du réseau

Vidéos connexes :

- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2020 – Networking best practices and tips with the AWS Well-Architected Framework
- AWS re:Invent 2020 – Monitoring and troubleshooting network traffic

Exemples connexes :

- Ateliers sur la mise en réseau AWS
- Surveillance réseau AWS
- Observation et diagnostic de votre réseau sur AWS
- Détection et résolution des erreurs de configuration du réseau sur AWS

Processus et culture

Lors de la création de l'architecture des charges de travail, vous pouvez adopter certains principes et certaines pratiques pour optimiser l'exécution de charges de travail cloud efficaces et performantes. Ce domaine d'intérêt propose les bonnes pratiques pour l'adoption d'une culture qui favorise l'efficacité des performances des charges de travail dans le cloud.

Tenez compte de ces principes clés pour développer cette culture :

- Infrastructure en tant que code : définissez votre infrastructure en tant que code à l'aide de méthodes telles que les modèles AWS CloudFormation. L'utilisation de modèles vous permet de placer votre infrastructure en mode de contrôle de code source parallèlement au code et aux configurations de votre application. Ceci vous permet d'appliquer les pratiques utilisées pour développer des logiciels à votre infrastructure et ainsi d'itérer rapidement.
- Pipeline de déploiement : utilisez un pipeline de déploiement d'intégration continue (CI) et de livraison continue (CD) (par exemple, référentiel de code source, systèmes de génération, déploiement et automatisation des tests) pour déployer votre infrastructure. Cela vous permet de déployer de manière reproductible et cohérente, le tout à un faible coût, à mesure que vous itérez.
- Métriques bien définies : configurez et surveillez vos métriques pour capturer des indicateurs clés de performances (KPI). Nous vous recommandons d'utiliser des métriques techniques, mais aussi des métriques commerciales. Pour les sites web ou les applications mobiles, les indicateurs clés capturent le temps jusqu'au premier octet ou le rendu. Les autres métriques applicables de manière générale comprennent le nombre de threads, la vitesse de nettoyage de la mémoire et les états d'attente. Les métriques commerciales, telles que les coûts cumulés agrégés par demande, peuvent vous permettre d'identifier des solutions pour réduire vos coûts. Réfléchissez bien à la façon dont vous prévoyez d'interpréter les métriques. Par exemple, vous pouvez choisir le maximum ou le 99e centile plutôt que la moyenne.
- Tests de performances automatiques : dans le cadre de votre processus de déploiement, des tests de performances peuvent se lancer automatiquement une fois que les tests d'exécution plus rapide ont abouti. L'automatisation doit créer un environnement, configurer des conditions initiales (comme des données de test), puis exécuter une série d'analyses comparatives et de tests de charge. Les résultats de ces tests doivent être rattachés à la version de génération afin que vous puissiez suivre l'évolution des performances dans le temps. Pour les tests de longue durée, vous pouvez rendre cette partie du pipeline asynchrone par rapport au reste de la compilation. Sinon, vous pouvez exécuter des tests de performances pendant la nuit en utilisant les instances Spot Amazon EC2.

- Génération de charge : vous devez créer une série de scripts qui reproduisent des parcours utilisateur synthétiques ou préenregistrés. Ces scripts doivent être idempotents et non couplés. Il se peut que vous deviez aussi inclure des scripts « de préparation » pour obtenir des résultats valides. Dans la mesure du possible, vos scripts de test doivent pouvoir répliquer le comportement d'utilisation en production. Vous pouvez utiliser un logiciel ou des solutions de logiciel en tant que service (SaaS) pour générer la charge. Envisagez d'utiliser des solutions AWS Marketplace et des instances Spot. Elles peuvent constituer des solutions rentables pour générer la charge.
- Visibilité des performances : les métriques clés doivent être visibles pour votre équipe, en particulier pour chaque version. Cela vous permet d'identifier les tendances positives ou négatives significatives au fil du temps. Vous devez également afficher les métriques sur le nombre d'erreurs ou d'exceptions pour vous assurer que vous testez un système fonctionnel.
- Visualisation : utilisez des techniques de visualisation qui permettent d'identifier clairement l'origine des problèmes de performances, les points chauds, les états d'attente ou les taux d'utilisation faibles. Superposez les métriques de performance sur les schémas d'architecture, des graphiques ou codes d'appel qui peuvent vous aider à identifier rapidement les problèmes.
- Processus d'évaluation régulier : les architectures qui présentent des performances médiocres sont généralement le résultat d'un processus d'évaluation des performances inexistant ou interrompu. Si votre architecture est peu performante, la mise en œuvre d'un processus d'évaluation des performances vous permet d'apporter des améliorations itératives.
- Optimisation continue : adoptez une culture permettant d'optimiser en permanence l'efficacité des performances de votre charge de travail dans le cloud.

Bonnes pratiques

- PERF05-BP01 Définir des indicateurs clés de performance (KPI) pour mesurer l’état et les performances de la charge de travail
- PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d’une importance critique
- PERF05-BP03 Définir un processus pour améliorer les performances des charges de travail
- PERF05-BP04 Effectuer un test de charge de votre charge de travail
- PERF05-BP05 Utiliser l’automatisation pour résoudre de manière proactive les problèmes liés aux performances
- PERF05-BP06 Maintenir votre charge de travail et vos services à jour
- PERF05-BP07 Vérifier les métriques à intervalles réguliers

PERF05-BP01 Définir des indicateurs clés de performance (KPI) pour mesurer l'état et les performances de la charge de travail

Identifiez les KPI qui mesurent les performances de la charge de travail de manière quantitative et qualitative. Les KPI vous aident à mesurer l'état et les performances d'une charge de travail par rapport à un objectif métier.

Anti-modèles courants :

- Vous surveillez uniquement les métriques au niveau du système pour avoir un aperçu de votre charge de travail et ne comprenez pas les impacts commerciaux possibles.
- Vous supposez que vos KPI sont déjà publiés et partagés en tant que données de métriques standard.
- Vous ne définissez pas de KPI quantitatif et mesurable.
- Vous ne tenez pas compte des objectifs ni des stratégies de l'entreprise pour définir vos KPI.

Avantages liés au respect de cette bonne pratique : en identifiant les KPI spécifiques qui représentent l'état et les performances de la charge de travail, vous pouvez aligner les équipes sur leurs priorités et définir des résultats commerciaux atteignables. Le partage de ces métriques avec tous les départements offre une visibilité et un alignement sur les seuils, les attentes et l'impact commercial.

Niveau de risque exposé si cette bonne pratique n'est pas établie : élevé

Directives d'implémentation

Les KPI permettent aux équipes commerciales et d'ingénierie de s'aligner sur la mesure des objectifs et des stratégies et sur la façon dont ces facteurs se combinent pour générer des résultats commerciaux. Par exemple, une charge de travail de site Web peut utiliser le temps de chargement de la page comme indication des performances globales. Cette métrique serait l'un des éléments de données pris en compte qui mesure l'expérience d'un utilisateur. En plus d'identifier les temps limites de chargement des pages, vous devez documenter le résultat attendu ou le risque commercial si les performances idéales ne sont pas atteintes. Un temps de chargement long des pages affecte directement vos utilisateurs finaux, nuit à leur expérience utilisateur et peut entraîner une perte de clients. Lorsque vous définissez vos seuils de KPI, combinez à la fois les points de référence en vigueur dans votre secteur et les attentes de vos utilisateurs finaux. Par exemple, si le point de référence actuel établi par votre secteur d'activité pour le chargement d'une page Web est un délai de

deux secondes, mais que vos utilisateurs finaux s'attendent à ce qu'une page Web se charge dans un délai d'une seconde, vous devez prendre en compte ces deux éléments de données lors de la définition des KPI.

Votre équipe doit évaluer les KPI de votre charge de travail à l'aide de données précises en temps réel et de données historiques à titre de référence et créer des tableaux de bord qui effectuent des calculs de métriques par rapport à vos données de KPI pour générer des informations opérationnelles et d'utilisation. Les KPI doivent être documentés et inclure les seuils qui soutiennent les objectifs et les stratégies de l'entreprise et doivent être mappés aux métriques surveillées. Les KPI doivent être revus lorsque les objectifs commerciaux, les stratégies ou les exigences des utilisateurs finaux changent.

Étapes d'implémentation

- Identifiez les parties prenantes : identifiez et documentez les principales parties prenantes de l'entreprise, y compris les équipes de développement et d'exploitation.
- Définissez les objectifs : collaborez avec ces parties prenantes pour définir et documenter les objectifs de votre charge de travail. Tenez compte des aspects critiques des performances de vos charges de travail, tels que le débit, le temps de réponse et le coût, ainsi que des objectifs métier, tels que la satisfaction des utilisateurs.
- Passez en revue les bonnes pratiques du secteur : passez en revue les bonnes pratiques du secteur pour identifier les KPI pertinents qui correspondent à vos objectifs en matière de charge de travail.
- Identifiez les métriques : identifiez les métriques qui correspondent à vos objectifs en matière de charge de travail et qui peuvent vous aider à mesurer les performances et les objectifs métier. Établissez des KPI sur la base de ces métriques. Les mesures telles que le temps de réponse moyen ou le nombre d'utilisateurs simultanés sont des exemples de métriques.
- Définissez et documentez des KPI : utilisez les bonnes pratiques du secteur et vos objectifs de charge de travail pour définir des cibles pour votre KPI de charge de travail. Utilisez ces informations pour définir les seuils de KPI pour les niveaux de gravité ou d'alarme. Identifiez et documentez le risque et l'impact du non-respect d'un KPI.
- Mettez en œuvre la surveillance : utilisez des outils de surveillance tels que Amazon CloudWatch ou AWS Config pour collecter les métriques et mesurer les KPI.
- Communiquez visuellement les KPI : utilisez des outils de tableau de bord tels que Amazon QuickSight pour visualiser et communiquer les KPI aux parties prenantes.

- Analysez et optimisez : passez régulièrement en revue et analysez les indicateurs de performance clés pour identifier les domaines de votre charge de travail qui doivent être améliorés. Collaborez avec les parties prenantes pour mettre en œuvre ces améliorations.
- Revoyez et affinez : passez régulièrement en revue les métriques et les KPI pour évaluer leur efficacité, en particulier lorsque les objectifs métier ou les performances de la charge de travail changent.

Ressources

Documents connexes :

- Documentation CloudWatch
- Surveillance, journalisation et performances - Partenaires AWS Partner
- Outils d'observabilité d'AWS
- The Importance of Key Performance Indicators (KPIs) for Large-Scale Cloud Migrations
- How to track your cost optimization KPIs with the KPI Dashboard
- Documentation X-Ray
- Utilisation des tableaux de bord Amazon CloudWatch (langue française non garantie)
- Indicateurs clés de performance Amazon QuickSight (langue française non garantie)

Vidéos connexes :

- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - Manage resource lifecycle events at scale with AWS Health
- AWS re:Invent 2023 - Performance & efficiency at Pinterest: Optimizing the latest instances
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2023 : Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Scaling on AWS for the first 10 million users
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Creating an Effective Metrics Strategy for Your Business | AWS Events

Exemples connexes :

- Création d’un tableau de bord avec Amazon QuickSight (langue française non garantie)

PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d’une importance critique

Comprenez et identifiez les domaines où l’augmentation des performances de votre charge de travail aura un impact positif sur l’efficacité ou l’expérience client. Par exemple, un site web qui comporte un grand nombre d’interactions clients pourrait gagner à utiliser des services de périphérie pour rapprocher la diffusion de contenus des clients.

Anti-modèles courants :

- Vous supposez que les métriques de calcul standard telles que l’utilisation du processeur ou la pression de mémoire, suffisent pour détecter les problèmes de performances.
- Vous n’utilisez que les métriques par défaut enregistrées par le logiciel de surveillance que vous avez sélectionné.
- Vous n’examinez les métriques qu’en cas de problème.

Avantages liés au respect de cette bonne pratique : en comprenant les domaines critiques de performances, les propriétaires des charges de travail peuvent surveiller les KPI et prioriser les améliorations à impact élevé.

Niveau de risque exposé si cette bonne pratique n’est pas établie : élevé

Directives d’implémentation

Mettez en place un suivi de bout en bout afin d’identifier les tendances du trafic, la latence et les domaines de performances critiques. Surveillez vos modèles d’accès aux données afin d’identifier les requêtes lentes ou les données mal fragmentées et partitionnées. Identifiez les zones de charge de travail limitées à l’aide de tests ou de surveillance des charges.

améliorer l’efficacité des performances en comprenant votre architecture, vos modèles de trafic et d’accès aux données, et identifier vos temps de latence et de traitement. Identifier les goulots d’étranglement potentiels qui pourraient avoir une incidence sur l’expérience client à mesure que la charge de travail augmente. Après avoir enquêté sur ces domaines, déterminez quelle solution vous pouvez déployer afin de surmonter ces problèmes de performances.

Étapes d'implémentation

- Mettez en place une surveillance de bout en bout pour capturer tous les composants et métriques de la charge de travail. Voici des exemples de solutions de surveillance sur AWS.

Service	Where to use
Amazon CloudWatch Real-User Monitoring (RUM)	To capture application performance metrics from real user client-side and frontend sessions.
AWS X-Ray	To trace traffic through the application layers and identify latency between components and dependencies. Use X-Ray service maps to see relationships and latency between workload components.
Amazon Relational Database Service Performance Insights	To view database performance metrics and identify performance improvements.
Amazon RDS Enhanced Monitoring	To view database OS performance metrics.
Amazon DevOps Guru	To detect abnormal operating patterns so you can identify operational issues before they impact your customers.

- Effectuez des tests afin de générer des métriques, d'identifier les tendances de trafic, les goulots d'étranglement et les domaines de performance critiques. Voici quelques exemples de méthodes de test :
  - Configurez CloudWatch Synthetic Canaries pour imiter par programmation les activités des utilisateurs basées sur le navigateur à l'aide de tâches cron Linux ou d'expressions de taux afin de générer des métriques cohérentes au fil du temps.
  - Utilisez la solution AWS Distributed Load Testing afin de générer un trafic de pointe ou de tester la charge de travail au taux de croissance attendu.
- Évaluez les métriques et la télémétrie pour identifier vos domaines de performances critiques. Examinez ces domaines avec votre équipe afin de discuter de la surveillance et des solutions pour éviter les goulots d'étranglement.

- Expérimentez des améliorations des performances et mesurez ces changements avec des données. Par exemple, vous pouvez utiliser CloudWatch Evidently pour tester les nouvelles améliorations et les impacts sur les performances de votre charge de travail.

Ressources

Documents connexes :

- What's new in AWS Observability at re:Invent 2023
- Bibliothèque Amazon Builders' Library
- Documentation X-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Vidéos connexes :

- AWS re:Invent 2023 : [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 : Implementing application observability
- AWS re:Invent 2023 : Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - The Amazon Builders' Library: 25 years of Amazon operational excellence
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Exemples connexes :

- Mesurer le temps de chargement des pages avec Amazon CloudWatch Synthetics (langue française non garantie)
- Client Web Amazon CloudWatch RUM (langue française non garantie)
- Kit SDK X-Ray pour Python (langue française non garantie)
- Tests de charges distribuées sur AWS

PERF05-BP03 Définir un processus pour améliorer les performances des charges de travail

Définissez un processus d'évaluation de nouveaux services, les modèles de conception, les types de ressources et les configurations au fur et à mesure qu'elles deviennent disponibles. Par exemple, exécutez des tests de performances existants sur de nouvelles offres d'instances afin de déterminer leur potentiel d'amélioration de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous introduisez des modifications d'architecture au fil du temps sans justification basée sur les métriques.

Avantages liés au respect de cette bonne pratique : en définissant votre processus permettant d'effectuer des modifications d'architecture, vous pouvez utiliser les données collectées pour influencer la conception de votre charge de travail au fil du temps.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Les performances de votre charge de travail présentent quelques contraintes clés. Documentez-les pour connaître les types d'innovations qui pourraient améliorer les performances de votre charge de travail. Utilisez ces informations lors de l'apprentissage de nouveaux services ou la technologie au fur et à mesure de leur disponibilité afin d'identifier les moyens d'atténuer des contraintes ou des goulets d'étranglement.

Identifiez les principales contraintes de performance pour votre charge de travail. Documentez les contraintes environnementales de votre charge de travail pour connaître les types d'innovations qui pourraient améliorer les performances de celle-ci.

Étapes d'implémentation

- Identifiez les indicateurs clés de performance (KPI) : identifiez les KPI des performances de votre charge de travail comme indiqué dans PERF05-BP01 Définir des indicateurs clés de performance

(KPI) pour mesurer l'état et les performances de la charge de travail pour établir une base de référence pour votre charge de travail.

- Mettez en œuvre la surveillance : utilisez les outils d'observabilité AWS pour collecter des métriques de performance et mesurer les KPI.
- Effectuez une analyse : réalisez une analyse approfondie pour identifier les domaines (tels que la configuration et le code d'application) de votre charge de travail qui ne sont pas performants, comme indiqué dans PERF05-BP02 Utiliser des solutions de surveillance pour comprendre les domaines où les performances sont d'une importance critique. Utilisez vos outils d'analyse et de performance pour identifier les stratégies d'amélioration des performances.
- Validez les améliorations : utilisez des environnements de test (sandbox) ou de pré-production pour valider l'efficacité des stratégies d'amélioration.
- Mettez en œuvre les modifications : mettez en œuvre les modifications en production et surveillez en permanence les performances de la charge de travail. Documentez les améliorations et communiquez-les aux parties prenantes.
- Revoyez et affinez : passez régulièrement en revue votre processus d'amélioration des performances afin d'identifier les domaines à améliorer.

Ressources

Documents connexes :

- Blog AWS
- Nouveautés AWS
- AWS Skill Builder

Vidéos connexes :

- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - Optimize your AWS workloads with best-practice guidance

Exemples connexes :

- AWS sur Github

PERF05-BP04 Effectuer un test de charge de votre charge de travail

Effectuez un test de charge de votre charge de travail pour vérifier qu'elle peut supporter la charge de production et identifier les éventuels goulots d'étranglement en termes de performances.

Anti-modèles courants :

- Vous testez les différentes parties et non la totalité de votre charge de travail.
- Vous testez la charge sur une infrastructure qui n'est pas la même que votre environnement de production.
- Vous n'effectuez le test de charge que pour la charge prévue sans aller au-delà, avec pour but de prévoir où vous pourriez rencontrer des problèmes à l'avenir.
- Vous effectuez des tests de charge sans consulter la politique de test Amazon EC2 et sans soumettre de formulaire de soumission d'événements simulés. Cela entraîne l'échec de votre test, car cela ressemble à un événement de déni de service.

Avantages liés au respect de cette bonne pratique : en mesurant vos performances dans le cadre d'un test de charge, vous saurez où vous serez affecté à mesure que la charge augmente. Cela peut vous permettre d'anticiper les changements nécessaires avant qu'ils n'affectent votre charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas établie: faible

Directives d'implémentation

Les tests de charge dans le cloud sont un processus visant à mesurer les performances de la charge de travail cloud dans des conditions réalistes avec la charge utilisateur attendue. Ce processus implique la mise en service d'un environnement cloud de type production, l'utilisation d'outils de test de charge pour générer la charge et l'analyse de métriques pour évaluer la capacité de votre charge de travail à gérer une charge réaliste. Pour effectuer un test de charge, vous devez exécuter des versions de données de production factices ou légèrement altérées (supprimez les données sensibles ou les informations d'identification). Effectuez automatiquement des tests de charge dans le cadre de votre pipeline de livraison et comparez les résultats aux indicateurs de performance clés et aux seuils prédéfinis. Ce processus vous permet de continuer à atteindre les performances requises.

Étapes d'implémentation

- Définissez vos objectifs de test : identifiez les aspects de performance de votre charge de travail que vous souhaitez évaluer, tels que le débit et le temps de réponse.
- Sélectionnez un outil de test : choisissez et configurez l'outil de test de charge adapté à votre charge de travail.
- Configurez votre environnement : configurez l'environnement de test en fonction de votre environnement de production. Vous pouvez utiliser les services AWS pour exécuter des environnements à l'échelle de la production afin de tester votre architecture.
- Mettez en œuvre la surveillance : utilisez des outils de surveillance, tels qu'Amazon CloudWatch pour collecter des métriques sur les ressources figurant dans votre architecture. Vous pouvez également collecter et publier des métriques personnalisées.
- Définissez les scénarios : définissez les scénarios et les paramètres de test de charge (tels que la durée du test et le nombre d'utilisateurs).
- Réalisez des tests de charge : exécutez des scénarios de test à grande échelle. Utilisez le AWS Cloud pour tester votre charge de travail et découvrir où elle ne parvient pas à se dimensionner ou si elle évolue de manière non linéaire. Par exemple, utilisez les instances Spot pour générer des charges à faible coût et découvrir les goulots d'étranglement avant de les rencontrer en production.
- Analysez les résultats de test : analysez les résultats pour identifier les goulots d'étranglement en matière de performances et les domaines à améliorer.
- Documentez et partagez les résultats : documentez et établissez des rapports sur les résultats et les recommandations. Partagez ces informations avec les parties prenantes pour les aider à prendre des décisions éclairées concernant les stratégies d'optimisation des performances.
- Effectuez une itération continue : les tests de charge doivent être effectués à une cadence régulière, en particulier après une modification ou une mise à jour du système.

Ressources

Documents connexes :

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Tests de charges distribuées sur AWS

Vidéos connexes :

- AWS Summit ANZ 2023: Accelerate with confidence through AWS Distributed Load Testing
- AWS re:Invent 2022 - Scaling on AWS for your first 10 million users
- Solving with AWS Solutions: Distributed Load Testing

- AWS re:Invent 2021 - Optimize applications through end user insights with Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Exemples connexes :

- Tests de charges distribuées sur AWS

PERF05-BP05 Utiliser l’automatisation pour résoudre de manière proactive les problèmes liés aux performances

Utilisez les KPI en combinaison avec des systèmes de surveillance et d’alarme pour traiter de manière proactive les problèmes liés aux performances.

Anti-modèles courants :

- Vous autorisez uniquement le personnel des opérations à apporter des modifications opérationnelles à la charge de travail.
- Vous confiez toutes les activités de filtre des alarmes à l’équipe des opérations sans correction proactive.

Avantages liés au respect de cette bonne pratique : en corrigeant de manière proactive les actions d’alarme, le personnel d’assistance peut se concentrer sur les éléments qui ne sont pas exploitables automatiquement. Cela permet au personnel des opérations de gérer toutes les alarmes sans être submergé et de se concentrer uniquement sur les alarmes critiques.

Niveau de risque exposé si cette bonne pratique n’est pas établie : faible

Directives d’implémentation

Utilisez des alarmes pour déclencher des actions automatisées afin de corriger les problèmes dans la mesure du possible. Faites remonter l’alarme aux personnes qui peuvent répondre si une réponse automatique n’est pas possible. Par exemple, vous pourriez disposer d’un système capable de prédire les valeurs attendues de KPI et qui déclenche une alarme lorsqu’elles dépassent certains seuils. Vous pouvez aussi disposer d’un outil capable d’arrêter ou de restaurer automatiquement des déploiements si les valeurs des KPI dépassent celles attendues.

Mettez en place des processus qui rendent visibles les performances pendant que votre charge de travail est en cours d’exécution. Créez des tableaux de bord de surveillance et établissez des normes de référence pour les attentes en matière de performances pour déterminer si les performances de la charge de travail sont optimales.

Étapes d’implémentation

- Identifiez le flux de travail de remédiation : identifiez et comprenez le problème de performance qui peut être résolu automatiquement. Utilisez les solutions de surveillance d’AWS telles que Amazon CloudWatch ou AWS X-Ray pour vous aider à mieux comprendre la cause profonde du problème.
- Définissez le processus d’automatisation : créez un processus de résolution étape par étape qui peut être utilisé pour résoudre automatiquement le problème.
- Configurez l’événement de lancement : configurez l’événement pour lancer automatiquement le processus de résolution. Par exemple, vous pouvez définir un déclencheur pour redémarrer automatiquement une instance lorsqu’elle atteint un certain seuil d’utilisation de l’UC.
- Automatisez la résolution : utilisez des services et des technologies AWS pour automatiser le processus de résolution. Par exemple, AWS Systems Manager Automation fournit une solution sécurisée et évolutive d’automatisation du processus de résolution. Veillez à utiliser une logique d’auto-réparation pour annuler les modifications si elles ne permettent pas de résoudre le problème.
- Testez le flux de travail : testez le processus de résolution automatisée dans un environnement de pré-production.
- Mettez en œuvre le flux de travail : mettez en œuvre la résolution automatisée dans l’environnement de production.
- Développez un manuel :développez et documentez un manuel qui décrit les étapes du plan de résolution, y compris les événements de lancement, la logique de résolution et les actions entreprises. Veillez à former les parties prenantes pour les aider à répondre efficacement aux événements de résolution automatisée.

- Révisez et affinez : évaluez régulièrement l'efficacité du flux de travail de résolution automatisée. Ajustez les événements de lancement et la logique de résolution, si nécessaire.

Ressources

Documents connexes :

- Documentation CloudWatch
- Surveillance, journalisation et performances - Partenaires AWS Partner Network
- Documentation X-Ray
- Utilisation des alarmes et des actions d'alarme dans CloudWatch (langue française non garantie)
- Build a Cloud Automation Practice for Operational Excellence: Best Practices from AWS Managed Services
- Automate your Amazon Redshift performance tuning with automatic table optimization

Vidéos connexes :

- AWS re:Invent 2023 - Strategies for automated scaling, remediation, and smart self-healing
- AWS re:Invent 2023 : [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 : Implementing application observability
- AWS re:Invent 2021 - Intelligently automating cloud operations
- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - Automating patch management and compliance using AWS
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS
- AWS re:Invent 2021 - {New Launch} Automatically detect and resolve issues with Amazon DevOps Guru
- AWS re:Invent 2023 - Centralize your operations

Exemples connexes :

- Personnalisation des alarmes CloudWatch Logs (langue française non garantie)

PERF05-BP06 Maintenir votre charge de travail et vos services à jour

Restez informé des nouveaux services et des nouvelles fonctionnalités cloud pour adopter des fonctionnalités efficaces, résoudre les problèmes et améliorer l'efficacité globale des performances de votre charge de travail.

Anti-modèles courants :

- Vous supposez que votre architecture actuelle est statique et ne sera pas mise à jour au fil du temps.
- Vous ne disposez pas de systèmes ou de rythme régulier pour évaluer la compatibilité des packages et des logiciels mis à jour avec votre charge de travail.

Avantages liés au respect de cette bonne pratique : en mettant en place un processus permettant de rester informé des nouveaux services et des nouvelles offres, vous pouvez adopter de nouvelles fonctionnalités et capacités, résoudre les problèmes et améliorer les performances de la charge de travail.

Niveau de risque exposé si cette bonne pratique n'est pas établie : faible

Directives d'implémentation

Évaluez les méthodes d'amélioration des performances au fur et à mesure que de nouveaux services, modèles de conception et fonctionnalités de produits entrent en scène. Identifiez celles de ces méthodes qui sont susceptibles d'améliorer les performances ou d'accroître l'efficacité de la charge de travail via l'évaluation, la discussion interne ou l'analyse externe. Mettez en place un processus permettant d'évaluer les mises à jour, les nouvelles fonctions et les services pertinents pour votre charge de travail. Par exemple, la création d'une démonstration de faisabilité qui utilise les nouvelles technologies ou la consultation d'un groupe interne. Lorsque vous essayez de nouvelles idées ou services, exécutez des tests de performances pour mesurer leur impact sur les performances de la charge de travail.

Étapes d'implémentation

- Dressez l'inventaire de votre charge de travail : faites l'inventaire des logiciels et de l'architecture de votre charge de travail, et identifiez les composants qui doivent être mis à jour.

- Identifiez les sources de mise à jour : identifiez les sources d'information et de mise à jour relatives aux éléments de votre charge de travail. Par exemple, vous pouvez vous abonner au blog Nouveautés AWS pour découvrir les produits correspondant à votre composant de charge de travail. Vous pouvez vous abonner au flux RSS ou gérer vos abonnements par e-mail.
- Définissez un calendrier de mise à jour : définissez un calendrier pour évaluer les nouveaux services et les nouvelles fonctionnalités pour votre charge de travail.
  - Utilisez AWS Systems Manager Inventory pour récupérer les métadonnées des systèmes d'exploitation, des applications et des instances issues de vos instances Amazon EC2 et rapidement connaître les instances exécutant le logiciel, les configurations requises par votre politique de logiciel et les instances devant être mises à jour.
- Évaluez la nouvelle mise à jour : découvrez comment mettre à jour les composants de votre charge de travail. Profitez de l'agilité du cloud pour tester rapidement la façon dont les nouvelles fonctionnalités peuvent améliorer votre charge de travail afin de gagner en efficacité.
- Utilisez l'automatisation : utilisez l'automatisation pour le processus de mise à jour afin de réduire l'effort de déploiement des nouvelles fonctionnalités et de limiter les erreurs causées par les processus manuels.
  - Vous pouvez utiliser CI/CD pour mettre automatiquement à jour les AMI, les images de conteneurs et d'autres artefacts liés à votre application cloud.
  - Vous pouvez utiliser des outils tels que AWS Systems Manager Patch Manager pour automatiser le processus de mise à jour du système et programmer l'activité à l'aide d'AWS Systems Manager Maintenance Windows.
- Documentez le processus : documentez votre processus d'évaluation des mises à jour et des nouveaux services. Donnez aux propriétaires le temps et l'espace nécessaires pour rechercher, tester, expérimenter et valider les mises à jour et les nouveaux services. Reportez-vous aux exigences opérationnelles documentées et aux KPI pour établir l'ordre de priorité des mises à jour qui auront un impact positif sur les activités.

Ressources

Documents connexes :

- Blog AWS
- Nouveautés AWS
- Implementing up-to-date images with automated EC2 Image Builder pipelines

Vidéos connexes :

- AWS re:Inforce 2022 - Automating patch management and compliance using AWS
- All Things Patch: AWS Systems Manager | AWS Events

Exemples connexes :

- Inventory and Patch Management
- One Observability Workshop

PERF05-BP07 Vérifier les métriques à intervalles réguliers

Vérifiez les métriques qui sont collectées dans le cadre de la maintenance de routine ou en réponse à des événements ou des incidents. Utilisez ces vérifications pour identifier d'une part les métriques qui ont été essentielles pour traiter les problèmes, et d'autre part les métriques supplémentaires, si elles ont été suivies, qui pourraient aider à identifier, traiter ou empêcher les problèmes.

Anti-modèles courants :

- Vous autorisez les métriques à rester dans un état d'alarme pendant longtemps.
- Vous créez des alarmes qui ne sont pas exploitables par un système d'automatisation.

Avantages liés au respect de cette bonne pratique : examinez en permanence les métriques collectées pour vérifier qu'elles identifient, résolvent ou préviennent correctement les problèmes. Les métriques peuvent également devenir caduques si vous les laissez dans un état d'alarme pendant longtemps.

Niveau de risque exposé si cette bonne pratique n'est pas établie : moyen

Directives d'implémentation

Améliorez constamment la surveillance et la collecte des métriques. Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies actuellement. Utilisez cette méthode pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Lorsque vous répondez aux incidents ou aux événements, évaluez les métriques qui ont été utiles dans la gestion du problème et les métriques qui auraient pu aider mais ne sont pas suivies actuellement. Utilisez ce processus pour améliorer la qualité des métriques que vous collectez afin de pouvoir prévenir ou résoudre plus rapidement les incidents futurs.

Étapes d'implémentation

- Définissez des métriques : définissez des métriques de performance critiques à surveiller, alignées sur votre objectif de charge de travail, notamment des métriques telles que le temps de réponse et l'utilisation des ressources.
- Établissez des bases de référence : définissez une base de référence et une valeur souhaitable pour chaque métrique. La base de référence doit fournir des points de référence pour identifier les écarts ou les anomalies.
- Définissez une cadence : définissez une cadence (hebdomadaire ou mensuelle, par exemple) pour examiner les métriques critiques.
- Identifiez les problèmes de performance : au cours de chaque examen, évaluez les tendances et les écarts par rapport aux valeurs de référence. Recherchez les goulots d'étranglement ou les anomalies au niveau des performances. Pour les problèmes identifiés, effectuez une analyse détaillée des causes profondes afin de comprendre la raison principale du problème.
- Identifiez les actions correctives : utilisez votre analyse pour identifier les actions correctives. Cela peut inclure le réglage des paramètres, la correction de bogues et la mise à l'échelle des ressources.
- Documentez les résultats : documentez vos résultats, y compris les problèmes identifiés, les causes racines et les actions correctives.
- Itérez et améliorez : évaluez et améliorez en permanence le processus de révision des métriques. Utilisez les enseignements tirés de la révision précédente pour améliorer le processus au fil du temps.

Ressources

Documents connexes :

- Documentation CloudWatch
- Collecter des métriques et des journaux auprès d'instances Amazon EC2 et de serveurs sur site avec l'agent CloudWatch (langue française non garantie)
- Interrogation de vos métriques avec CloudWatch Metrics Insights

- Surveillance, journalisation et performances - Partenaires AWS Partner Network
- Documentation X-Ray

Vidéos connexes :

- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 : Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS

Exemples connexes :

- Création d’un tableau de bord avec Amazon QuickSight (langue française non garantie)
- CloudWatch Dashboards

Conclusion

Pour atteindre et maintenir l'efficacité des performances, il est nécessaire d'avoir une approche axée sur les données. Vous devriez sérieusement envisager des modèles d'accès et des compromis qui vous permettront d'optimiser les performances. L'utilisation d'un processus d'évaluation basé sur des points de référence et des tests de charge vous permet de sélectionner les configurations et types de ressources appropriés. En traitant votre infrastructure comme du code, vous pouvez faire évoluer votre architecture rapidement et en toute sécurité tout en utilisant les données pour prendre des décisions basées sur des faits en ce qui concerne votre architecture. La mise en place d'une surveillance à la fois active et passive permet de s'assurer que les performances de votre architecture ne se dégradent pas au fil du temps.

AWS s'efforce de vous aider à concevoir des architectures garantissant l'efficacité des performances tout en offrant une valeur commerciale. Utilisez les outils et techniques présentés dans ce document pour garantir votre réussite.

Participants

Les personnes et organisations suivantes ont participé à la préparation du présent document :

- Sam Mokhtari, architecte principal de solutions en matière d'efficacité, Amazon Web Services
- Josh Hart, architecte de solutions, Amazon Web Services
- Richard Trabing, architecte de solutions, Amazon Web Services
- Brett Looney, architecte principal de solutions, Amazon Web Services
- Nina Vogl, architecte principale de solutions, Amazon Web Services
- Eric Pullen, architecte de solutions, Amazon Web Services
- Julien Lépine, responsable des architectes de solutions spécialisés, Amazon Web Services
- Ronnen Slasky, architecte de solutions, Amazon Web Services

Autres lectures

Pour obtenir de l'aide, consultez les ressources suivantes :

- AWS Well-Architected Framework
- Centre d'architecture AWS

Révisions du document

Pour être informé des mises à jour de ce livre blanc, abonnez-vous au flux RSS.

Modification	Description	Date
Livre blanc mis à jour	Les bonnes pratiques ont été mises à jour avec de nouvelles recommandations en matière d'implémentation.	June 27, 2024
Mise à jour et restructuration majeures	Ce pilier a été restructuré pour inclure cinq domaines de bonnes pratiques (contre huit auparavant). Le contenu a été regroupé dans ces cinq domaines et a été mis à jour. Les nouveaux domaines de bonnes pratiques sont Choix d'architecture, Informatique et matériel, Gestion des données, Mise en réseau et diffusion de contenuet Processus et culture.	October 3, 2023
Mise à jour mineure	Suppression du langage non inclusif.	April 13, 2023
Mises à jour du nouveau cadre	Les bonnes pratiques ont été mises à jour avec des recommandations et de nouvelles bonnes pratiques.	April 10, 2023
Livre blanc mis à jour	Les bonnes pratiques ont été mises à jour avec de nouvelles	December 15, 2022

	recommandations en matière d'implémentation.	
Livre blanc mis à jour	Développement des bonnes pratiques et ajout de plans d'amélioration.	October 20, 2022
Mise à jour mineure	Suppression du langage non inclusif.	April 22, 2022
Mise à jour mineure	Ajout du pilier Durabilité dans l'introduction.	December 2, 2021
Mises à jour mineures	Mise à jour des liens.	March 10, 2021
Mises à jour mineures	Délai d'expiration AWS Lambda remplacé par 900 secondes et nom de Amazon Keyspaces (for Apache Cassandra) corrigé.	October 5, 2020
Mise à jour mineure	Correction d'un lien rompu.	July 15, 2020
Mises à jour pour le nouveau cadre	Révision et mise à jour majeures du contenu	July 8, 2020
Livre blanc mis à jour	Mise à jour mineure pour les problèmes grammaticaux	July 1, 2018
Livre blanc mis à jour	Actualisation du livre blanc pour refléter les modifications apportées à AWS	November 1, 2017
Publication initiale	Pilier Efficacité des performances - AWS Well-Architected Framework publié.	November 1, 2016

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the Glossaire AWS Reference.