Options d'analyse du Big Data sur AWS

Janvier 2016

© 2016, Amazon Web Services, Inc. ou ses sociétés apparentées. Tous droits réservés.

Mentions légales

Ce document est fourni à titre informatif uniquement. Il présente l'offre de produits et les pratiques actuelles d'AWS à la date de publication de ce document, des informations qui sont susceptibles d'être modifiées sans préavis. Il incombe aux clients de procéder à leur propre évaluation indépendante des informations contenues dans ce document. Chaque client est responsable de son utilisation des produits ou services AWS, chacun étant fourni « en l'état », sans garantie d'aucune sorte, qu'elle soit explicite ou implicite. Ce document n'offre pas de garantie, représentation, engagement contractuel, condition ou assurance de la part d'AWS, de ses sociétés apparentées, fournisseurs ou concédants de licence. Les responsabilités et obligations d'AWS vis-à-vis de ses clients sont régies par les contrats AWS. Le présent document ne fait partie d'aucun contrat et ne modifie aucun contrat entre AWS et ses clients.

Table des matières

Résumé

Ce livre blanc aide les architectes, les spécialistes des données et les développeurs à mieux comprendre les options d'analyse du Big Data disponibles dans le cloud AWS en fournissant une vue d'ensemble des services, avec les informations suivantes :

- Modèles d'utilisation préconisés
- Coût du modèle
- Performances
- Durabilité et disponibilité
- Évolutivité et souplesse
- Interfaces
- Anti-modèles

Ce livre blanc s'achève sur des scénarios illustrant l'usage des fonctions d'analyse. On y retrouve aussi des ressources supplémentaires pour se lancer dans l'analyse du Big Data sur AWS.

Introduction

À mesure que notre société se numérise, la quantité de données créées et collectées augmente et s'accélère considérablement. L'analyse de ces données en perpétuelle croissance devient un défi avec les outils d'analyse traditionnels. Nous avons besoin d'outils innovants pour relier les données générées et les données qui peuvent être analysées efficacement.

Les outils et les technologies Big Data offrent des opportunités et des défis pour pouvoir analyser efficacement les données afin de mieux comprendre les préférences des clients, d'acquérir un avantage concurrentiel sur le marché et de développer votre activité. Les architectures de gestion des données ont évolué du modèle traditionnel d'entreposage de données à des architectures plus complexes qui répondent à plus d'exigences, telles que le traitement en temps réel et par lots, les données structurées et non structurées, les transactions à vitesse élevée, etc.

Amazon Web Services (AWS) fournit une large plate-forme de services gérés pour vous aider à créer, sécuriser et mettre à l'échelle rapidement et sans difficulté des applications Big Data de bout en bout. Que vos applications requièrent un traitement de données par lot ou un streaming en temps réel, AWS fournit l'infrastructure et les outils nécessaires pour mener à bien votre prochain projet Big Data. Pas de matériel à acquérir, pas d'infrastructure à maintenir et à mettre à l'échelle : uniquement ce dont vous avez besoin pour collecter, stocker, traiter et analyser le Big Data. AWS dispose d'un écosystème de solutions analytiques spécialement conçu pour gérer cette quantité croissante de données et fournir un aperçu de votre activité.

L'avantage d'AWS dans l'analyse du Big Data

L'analyse de grands ensembles de données nécessite une capacité de calcul importante dont la taille peut varier en fonction de la quantité de données d'entrée et du type d'analyse. Cette caractéristique des charges de travail du Big Data est parfaitement adaptée au modèle de cloud computing par répartition, dans lequel les applications peuvent facilement évoluer en fonction de la demande. Au fur et à mesure que les exigences évoluent, vous pouvez facilement redimensionner votre environnement (horizontalement ou verticalement) sur AWS afin de répondre à vos besoins, sans avoir à attendre de matériel supplémentaire ou à investir trop pour fournir une capacité suffisante.

Quant aux applications critiques sur une infrastructure plus traditionnelle, les concepteurs de systèmes n'ont pas d'autre choix que de fournir un volume de ressources supérieur, car une augmentation des données supplémentaires due à une hausse des besoins de l'entreprise doit être quelque chose que le système est en mesure de gérer. En revanche, sur AWS, vous pouvez provisionner davantage de capacité et de calcul en quelques minutes, ce qui signifie que vos applications Big Data augmentent et diminuent en fonction de la demande, et que votre système fonctionne le mieux possible.

En outre, vous bénéficiez d'une informatique flexible sur une infrastructure globale avec un accès aux nombreuses régions géographiques[1] que propose AWS, mais aussi de la possibilité d'utiliser d'autres services évolutifs qui permettent de développer des applications Big Data sophistiquées. Ces autres services incluent

Amazon Simple Storage Service (Amazon S3)[2] afin de stocker des données et AWS Data Pipeline[3] pour orchestrer les tâches à transférer et transformer facilement ces données. AWS IoT[4] qui permet aux appareils connectés d'interagir avec les applications cloud et d'autres appareils connectés.

En outre, AWS dispose de nombreuses options afin de vous aider à obtenir des données dans le cloud, y compris les appareils sécurisés tels qu'AWS Import/Export Snowball[5] pour accélérer les transferts de données à l'échelle du pétaoctet et Amazon Kinesis Firehose[6] pour charger les données de streaming et des connexions privées scalables via AWS Direct Connect.[7] Étant donné que l'utilisation des mobiles ne cesse d'évoluer, vous pouvez utiliser la suite de services AWS Mobile Hub[8] pour collecter et mesurer l'utilisation et les données ou exporter ces données vers un autre service afin d'obtenir des analyses personnalisées.

Ces fonctionnalités de la plate-forme AWS en font une solution idéale pour résoudre les difficultés inhérentes au Big Data, et de nombreux clients sont parvenus à mettre en œuvre des charges de travail d'analyse du Big Data sur AWS. Pour plus d'informations sur les cas d'utilisation, consultez la page Big Data et HPC. Alimenté par le cloud AWS.[9]

Les services suivants sont décrits dans le but de collecter, traiter, stocker et analyser le Big Data :

- Amazon Kinesis Streams
- AWS Lambda
- Amazon Elastic MapReduce
- Amazon Machine Learning
- Amazon DynamoDB
- Amazon Redshift
- Amazon Elasticsearch Service
- Amazon QuickSight

En outre, les instances Amazon EC2 sont disponibles pour les applications de Big Data autogérées.

Amazon Kinesis Streams

Amazon Kinesis Streams[10] vous permet de créer des applications personnalisées qui traitent ou analysent les données de streaming en temps réel. Amazon Kinesis Streams peut capturer et stocker en continu des téraoctets de données par heure, provenant de centaines de milliers de sources diverses comme les flux de clics sur des sites web, les transactions financières, les flux de données sur les réseaux sociaux, les journaux informatiques et les événements de suivi des emplacements.

Avec Amazon Kinesis Client Library (KCL), vous pouvez créer des applications Amazon Kinesis et utiliser les données de streaming afin d'alimenter les tableaux de bord en temps réel, de générer des alertes, de mettre en place une tarification et une publicité dynamique, et bien plus encore. Vous pouvez également émettre des données depuis Amazon Kinesis Streams vers d'autres services AWS comme Amazon S3, Amazon Redshift, Amazon EMR et AWS Lambda.

Mettez en service le niveau d'entrée et de sortie requis pour votre flux de données, par blocs de 1 méga-octet par seconde (Mo/s), en utilisant l'AWS Management Console, l'API[11] ou des kits SDK.[12] Vous pouvez à tout moment augmenter ou réduire votre flux de données, sans avoir besoin de le redémarrer, et sans aucune répercussion sur les sources transmettant ces données vers le flux. En quelques secondes, les données mises dans un flux sont disponibles pour analyse.

Les données de flux sont stockées dans plusieurs zones de disponibilité au sein d'une région durant 24 heures. Pendant cet intervalle, les données peuvent être lues, relues, utilisées en remplacement et analysées, ou transférées vers un système de stockage à long terme, tel qu'Amazon S3 ou Amazon Redshift. Le KCL permet aux développeurs de se consacrer à la création d'applications métier en éliminant les charges lourdes liées à l'équilibrage des charges des données de flux, à la coordination des services distribués et au traitement des données insensibles aux défaillances.

Modèles d'utilisation préconisés

Amazon Kinesis Streams est utile partout où il est nécessaire de déplacer rapidement les données des producteurs (sources de données) et de les traiter en continu. Ce traitement peut consister en la transformation des données avant

une émission vers un autre système de stockage de données, la génération des métriques et des analyses en temps réel, la dérivation et l'agrégation de plusieurs flux en flux plus complexes, ou un traitement en aval. Voici quelques scénarios d'utilisation classiques d'Amazon Kinesis Streams pour les analyses.

- **Analyse de données en temps réel** : Amazon Kinesis Streams permet l'analyse de données en temps réel sur des données en continu, telles que l'analyse des données de navigation sur le site web et l'analyse de l'engagement des clients.
- **Apport et traitement des flux de données et des journaux** : Avec Amazon Kinesis Streams, vous pouvez demander aux producteurs de transférer les données directement dans un flux Amazon Kinesis. Par exemple, vous pouvez envoyer des journaux du système et de l'application à Amazon Kinesis Streams et accéder au flux en quelques secondes. Cela évite la perte des données du journal en cas de défaillance du serveur frontal ou du serveur d'applications, et réduit le stockage des journaux locaux à la source. Amazon Kinesis Streams fournit une saisie de données accélérée car vous ne compressez pas les données sur les serveurs avant de les soumettre pour approbation.
- **Métriques et rapports en temps réel** : Vous pouvez utiliser les données intégrées dans Amazon Kinesis Streams afin d'extraire des métriques et générer des indicateurs clés de performance pour alimenter les rapports et les tableaux de bord en temps réel. Cela permet à la logique applicative de traitement des données de fonctionner sur les données en continu, plutôt que d'attendre l'arrivée des lots de données.

Coût du modèle

Amazon Kinesis Streams a un prix par répartition simple, sans frais initiaux ni frais minimum, et vous ne payez que pour les ressources que vous consommez. Un flux Amazon Kinesis se compose d'une ou de plusieurs partitions. Chacune vous donne une capacité de 5 transactions de lecture par seconde, jusqu'à un maximum de 2 Mo de données lues par seconde. Chaque partition peut prendre en charge jusqu'à 1 000 transactions d'écriture par seconde et un maximum de 1 Mo de données écrites par seconde.

La capacité de données de votre flux se fait en fonction du nombre de partitions que vous spécifiez pour le flux. La capacité totale du flux est la somme de la capacité de

chaque partition. Il n'existe que deux composantes de tarification, une charge horaire par partition et une charge pour chaque million de transactions PUT. Pour plus d'informations, consultez la page Tarifs d'Amazon Kinesis Streams.[13] Les applications qui s'exécutent sur Amazon EC2 et traitent les flux Amazon Kinesis entraînent également des coûts Amazon EC2 standard.

Performances

Amazon Kinesis Streams vous permet de choisir la capacité de débit dont vous avez besoin en termes de partitions. Pour chaque partition dans un flux Amazon Kinesis, vous pouvez capturer jusqu'à 1 méga-octet de données par seconde, à un rythme de 1 000 transactions d'écriture par seconde. Vos applications Amazon Kinesis lisent les données de chaque partition à un débit pouvant atteindre 2 Mo/s. Vous pouvez mettre en service autant de partitions que nécessaire afin d'obtenir la capacité de débit souhaitée ; par exemple, un flux de données de 1 Go par seconde nécessiterait 1 024 partitions.

Durabilité et disponibilité

Amazon Kinesis Streams réplique les données de manière synchronisée sur trois zones de disponibilité au sein d'une région AWS, vous permettant de bénéficier d'une haute disponibilité des données et d'optimiser leur durabilité. De plus, vous pouvez stocker un curseur dans DynamoDB pour suivre sur la durée ce qui a été lu à partir d'un flux Amazon Kinesis. Dans le cas où votre application échoue au milieu de la lecture des données du flux, vous pouvez redémarrer votre application et utiliser le curseur afin de retrouver l'endroit exact où l'application a échoué.

Évolutivité et souplesse

Vous pouvez augmenter ou diminuer la capacité du flux à tout moment en fonction de vos besoins opérationnels ou commerciaux, sans interruption du traitement des flux en cours. En utilisant des appels d'API ou des outils de développement, vous pouvez automatiser la mise à l'échelle de votre environnement Amazon Kinesis Streams afin de répondre à la demande et vous assurer de ne payer que ce dont vous avez besoin.

Interfaces

Il existe deux interfaces pour Amazon Kinesis Streams : une entrée utilisée par les applications producteur pour mettre des données dans Amazon Kinesis

Streams, et une sortie pour traiter et analyser les données entrantes. Les producteurs peuvent écrire des données à l'aide de l'API Amazon Kinesis PUT, d'une abstraction du kit de développement logiciel AWS (SDK) ou d'un toolkit,[14] d'Amazon Kinesis Producer Library (KPL),[15] ou d'Amazon Kinesis Agent.[16]

Pour le traitement des données qui ont déjà été placées dans un flux Amazon Kinesis, des bibliothèques client sont fournies pour développer et exploiter des applications de traitement de données en continu en temps réel. Le KCL[17] joue le rôle d'intermédiaire entre Amazon Kinesis Streams et vos applications d'entreprise qui contiennent la logique de traitement spécifique. Il existe également une intégration à lire à partir d'un flux Amazon Kinesis dans Apache Storm via l'Amazon Kinesis Storm Spout.[18]

Anti-modèles

Amazon Kinesis Streams possède les anti-modèles suivants :

- **Débit constant à petite échelle** : bien qu'Amazon Kinesis Streams fonctionne pour le streaming de données à 200 Ko / sec ou moins, il est conçu et optimisé pour des débits de données plus importants.
- **Stockage et analyse de données à long terme** : Amazon Kinesis Streams n'est pas adapté au stockage de données à long terme. Par défaut, les données sont conservées pendant 24 heures et vous pouvez prolonger la période de rétention de sept jours maximum. Vous pouvez déplacer toutes les données qui doivent être stockées pendant plus de 7 jours vers un autre service de stockage durable tel qu'Amazon S3, Amazon Glacier, Amazon Redshift, ou DynamoDB.

AWS Lambda

AWS Lambda[19] vous permet d'exécuter le code sans devoir mettre en service ou gérer des serveurs. Vous payez uniquement le temps de calcul utilisé et ne déboursez rien lorsque votre code n'est pas en cours d'exécution. Avec Lambda, vous pouvez exécuter le code pour quasiment n'importe quel type d'application ou service principal, sans avoir à vous préoccuper de leur administration. Il vous suffit de télécharger votre code et Lambda s'occupe de tout ce qui est nécessaire à son exécution et à son évolution tout en garantissant une haute disponibilité. Vous pouvez configurer le code de manière à le déclencher automatiquement depuis d'autres services AWS ou l'appeler directement à partir de n'importe quelle application web ou mobile.

Modèles d'utilisation préconisés

Lambda vous permet d'exécuter du code en réponse à certains déclencheurs, tels que la modification de données, un changement d'état au niveau du système, ou encore une action effectuée par l'utilisateur. Lambda peut être directement déclenché par certains services AWS, tels que Amazon S3, DynamoDB, Amazon Kinesis Streams, Amazon Simple Notification Service (Amazon SNS) et Amazon CloudWatch, ce qui vous permet de créer différents systèmes de traitement des données en temps réel.

- **Traitement de fichiers en temps réel** : vous pouvez déclencher Lambda afin d'appeler un processus dans lequel un fichier a été modifié ou téléchargé sur Amazon S3. Par exemple, pour passer une image en couleur vers les niveaux de gris une fois celle-ci téléchargée sur Amazon S3.
- **Traitement des flux en temps réel** : vous pouvez utiliser Amazon Kinesis Streams et Lambda pour traiter les données en continu afin d'analyser les flux de navigation, de filtrer les journaux et d'analyser les réseaux sociaux.
- **ETL** : vous pouvez utiliser Lambda pour exécuter des travaux qui transforment les données et les chargent depuis un référentiel de données vers un autre.
- **Remplacer cron** : utilisez des expressions de planification pour exécuter une fonction Lambda à intervalles réguliers. C'est une solution moins coûteuse et plus accessible que l'exécution de cron sur une instance EC2.
- **Traiter les événements AWS** : de nombreux autres services, tels qu'AWS CloudTrail, peuvent agir en tant que sources d'événements simplement en se connectant à Amazon S3 et en utilisant des notifications de compartiment S3 afin de déclencher des fonctions Lambda.

Coût du modèle

Avec Lambda, vous payez à hauteur de votre consommation. Vous êtes facturé en fonction du nombre de requêtes pour vos fonctions et de l'heure d'exécution de votre code. L'offre Lambda gratuite comprend un million de requêtes offertes, ainsi que 400 000 Go/secondes de temps de calcul par mois. Des frais de 0,20 USD par million de requêtes sont ensuite facturés (0,0000002 USD par requête). De plus, la durée de l'exécution de votre code est calculée en fonction de la quantité de mémoire allouée. Vous devez payer 0,00001667 USD par Go/seconde utilisé. Pour plus d'informations, consultez la section Tarifs d'AWS Lambda.

Performances

Après avoir déployé votre code dans Lambda pour la première fois, vos fonctions sont généralement prêtes à être invoquées dans les secondes qui suivent le chargement. Lambda est conçu pour traiter les événements en quelques millisecondes. La latence sera plus importante suite à la création ou la mise à jour d'une fonction Lambda, ou si cette dernière n'a pas été utilisée récemment.

Durabilité et disponibilité

Lambda est conçu pour fournir, via la réplication et la redondance, une haute disponibilité à la fois pour le service lui-même, mais aussi pour les fonctions qu'il exploite. Il n'y a pas de fenêtres de maintenance ou d'arrêts planifiés pour l'un ou l'autre. En cas d'échec, les fonctions Lambda invoquées de façon synchrone répondront avec une exception. Les fonctions Lambda invoquées de manière asynchrone sont relancées au moins trois fois. Au-delà de cette limite, il est possible que l'événement soit rejeté.

Évolutivité et souplesse

Le nombre de fonctions Lambda que vous pouvez exécuter ne souffre aucune limite. Toutefois, Lambda comporte par défaut une limite de sécurité de 100 exécutions simultanées par compte et par région. Un membre de l'équipe du support AWS peut augmenter cette limite.

Lambda est conçu pour effectuer des mises à l'échelle automatiques en votre nom. Il n'y a pas de limites fondamentales à la mise à l'échelle d'une fonction. Lambda alloue de la capacité de façon dynamique pour correspondre aux événements entrants.

Interfaces

Les fonctions Lambda peuvent être gérées de diverses manières. Vous pouvez facilement référencer, supprimer, mettre à jour et surveiller vos fonctions Lambda à l'aide du tableau de bord de la console Lambda. Vous pouvez également utiliser l'interface de ligne de commande et le SDK AWS pour gérer vos fonctions Lambda.

Vous pouvez déclencher une fonction Lambda à partir d'un événement AWS, comme les notifications de compartiments Amazon S3, DynamoDB Streams, CloudWatch Logs, Amazon SES, Amazon Kinesis Streams, Amazon SNS, Amazon Cognito, et bien plus encore. Tout appel d'API effectué dans un service

prenant en charge AWS CloudTrail peut être traité en tant qu'événement dans Lambda en répondant aux journaux d'audit CloudTrail. Pour plus d'informations sur les sources d'événements, consultez Composants de base : Fonction et sources d'événements AWS Lambda.[20]

Lambda prend en charge les langages de programmation tels que Java, Node.js et Python. Votre code peut inclure des bibliothèques existantes, et même natives. Les fonctions Lambda peuvent lancer facilement des processus via des langages pris en charge par l'AMI Amazon Linux,[21] dont Bash, Go et Ruby. Pour de plus amples informations, veuillez consulter la documentation pour Node.js,[22] Python[23] et Java.[24]

Anti-modèles

Lambda possède les anti-modèles suivants :

- **Applications de longue durée** : chaque fonction Lambda doit être terminée dans les 300 secondes. Pour les applications de longue durée qui peuvent nécessiter des travaux de plus de cinq minutes, Amazon EC2 est recommandé. Vous pouvez aussi créer une chaîne de fonctions Lambda où la fonction 1 appelle la fonction 2, qui appelle la fonction 3, et ainsi de suite jusqu'à la fin du processus.
- **Sites web dynamiques** : bien qu'il soit possible d'exécuter un site web statique avec Lambda, l'exécution d'un site web très dynamique et volumineux peut entraîner des performances prohibitives. L'utilisation d'Amazon EC2 et d'Amazon CloudFront est un cas d'utilisation recommandé.
- **État des applications** : le code Lambda doit être écrit dans un style dit « sans état » ; c'est-à-dire qu'il doit supposer qu'il n'existe pas d'affinité avec l'infrastructure de calcul sous-jacente. Il est possible que l'accès au système de fichiers local, les processus enfants et les artefacts similaires ne s'étendent pas au-delà de la durée de la requête et tout état persistant doit être stocké dans Amazon S3, DynamoDB ou un autre service de stockage disponible sur Internet.

Amazon EMR

Amazon EMR[25] est une infrastructure de calcul hautement distribuée qui permet de traiter et de stocker rapidement les données de manière rentable. Amazon EMR utilise Apache Hadoop, une infrastructure open source, pour

répartir vos données et leur traitement sur un cluster redimensionnable regroupant des instances EC2 et vous permet d'utiliser tous les outils Hadoop courants, tels que Hive, Pig, Spark, etc. Hadoop fournit une infrastructure afin d'exécuter le traitement et l'analyse du Big Data. Amazon EMR prend en charge toutes les tâches de provisionnement, de gestion et de maintenance de l'infrastructure et des logiciels d'un cluster Hadoop.

Modèles d'utilisation préconisés

Le framework flexible d'Amazon EMR réduit les problèmes de traitement et les ensembles de données en tâches plus petites et les répartit entre de nombreux nœuds de calcul dans un cluster Hadoop. Cette capacité se prête à de nombreux modèles d'utilisation impliquant des analyses du Big Data. Voici quelques exemples :

- Traitement et analyse des fichiers journaux
- Mouvement de données ETL massif
- Modélisation des risques et analyse des menaces
- Ciblage publicitaire et analyse du parcours de navigation
- Génomique
- Analyses prédictives
- Analyse et exploration de données ad hoc

Pour plus d'informations, reportez-vous à la section Bonnes pratiques pour Amazon EMR.[26]

Coût du modèle

Avec Amazon EMR, vous pouvez lancer un cluster persistant indéfiniment ou un cluster temporaire qui prend fin une fois l'analyse terminée. Dans les deux cas, vous ne payez que pour les heures d'activité du cluster.

Amazon EMR prend en charge divers types d'instances EC2 (standard, processeur élevé, mémoire élevée, E/S élevées, etc.) et d'options tarifaires Amazon EC2 (à la demande, réservée et ponctuelle). Lorsque vous lancez un cluster Amazon EMR (également appelé « flux de travail »), vous choisissez le nombre et le type d'instances Amazon EC2 à approvisionner. Les frais Amazon EMR sont en sus du prix Amazon EC2. Pour plus d'informations, consultez la page Tarifs d'Amazon EMR.[27]

Performances

Les performances d'Amazon EMR dépendent du type d'instances EC2 sur lesquelles vous choisissez d'exécuter votre cluster et du nombre d'instances que vous avez choisi pour exécuter vos analyses. Vous devez choisir un type d'instance adapté à vos besoins de traitement, avec suffisamment de mémoire, de stockage et de puissance de traitement. Pour plus d'informations sur les spécifications d'instances EC2, consultez la section Types d'instances Amazon EC2.[28]

Durabilité et disponibilité

Par défaut, Amazon EMR tolère les défaillances pour les échecs du nœud principal et continue l'exécution des travaux si un nœud esclave tombe en panne. Actuellement, Amazon EMR ne met pas automatiquement en service un autre nœud afin de convertir des esclaves défaillants. Toutefois, les clients peuvent surveiller l'état des nœuds et remplacer les défaillants avec CloudWatch.

Pour vous aider à gérer une défaillance du nœud principal, peu probable, nous vous recommandons de sauvegarder vos données sur un système de stockage persistant tel qu'Amazon S3. Le cas échéant, vous pouvez choisir d'exécuter Amazon EMR avec la distribution MapR,[29] qui fournit une architecture no-NameNode tolérant plusieurs pannes simultanées avec système automatique de basculement et de reprise. Les métadonnées sont distribuées et répliquées, tout comme les données. Sur une architecture no-NameNode, le nombre de fichiers pouvant être stockés n'est pas restreint et il n'y a aucune dépendance à un NAS externe.

Évolutivité et souplesse

Avec Amazon EMR, vous pouvez facilement redimensionner un cluster en cours d'exécution.[30] Vous pouvez ajouter des nœuds principaux qui contiennent le système de fichiers distribué Hadoop (HDFS) à tout moment afin d'augmenter la puissance de traitement et la capacité de stockage HDFS (ainsi que le débit). De plus, vous pouvez utiliser Amazon S3 de manière native ou via EMFS, avec ou à la place de HDFS local, qui vous permet de découpler vos capacités de mémoire et de calcul à partir de votre stockage en fournissant une plus grande flexibilité et rentabilité.

Vous pouvez également ajouter et supprimer des nœuds de tâches à tout moment, capables de traiter des travaux Hadoop, mais pas de gérer HDFS. Certains clients ajoutent des centaines d'instances à leurs clusters au moment du

traitement par lots, puis suppriment les instances excédentaires une fois le traitement terminé. Vous ne pouvez pas, par exemple, connaître la quantité de données que vos clusters géreront dans 6 mois. Vous pourriez aussi avoir des besoins de traitement irréguliers. Avec Amazon EMR, inutile de connaître à l'avance vos futurs besoins ou de prévoir des pics de demande, puisque vous pouvez facilement ajouter ou supprimer de la capacité à tout moment.

En outre, vous pouvez ajouter tous les nouveaux clusters de différentes tailles et les supprimer à tout moment en quelques clics dans la console ou via un appel d'API programmatique.[31]

Interfaces

Amazon EMR prend en charge de nombreux outils qui fonctionnent avec Hadoop et qui peuvent être utilisés pour analyser le Big Data. Chacun possède sa propre interface. Voici un bref résumé des options les plus populaires :

Hive

Hive est un progiciel d'analyse et de stockage de données open source qui fonctionne sur Hadoop. Hive fonctionne avec Hive QL, un langage basé sur SQL qui permet aux utilisateurs de structurer, synthétiser et interroger des données. Hive QL s'étend au-delà de la norme SQL standard, en ajoutant une assistance de première classe pour les fonctions de traitement (« map ») et d'agrégation (« reduce ») et les données complexes extensibles définies par l'utilisateur, telles que JSON et Thrift. Cette fonctionnalité permet le traitement de sources de données complexes et non structurées telles que des documents texte et des fichiers journaux.

Hive permet d'utiliser des extensions utilisateur, grâce aux fonctions définies par l'utilisateur et écrites en langage Java. Amazon EMR a apporté de nombreuses améliorations à Hive, notamment l'intégration directe avec Amazon DynamoDB et Amazon S3. Avec Amazon EMR, vous pouvez, par exemple, charger des partitions de table automatiquement depuis Amazon S3. Vous pouvez aussi écrire des données dans des tables au sein d'Amazon S3 sans utiliser de fichiers temporaires et accéder à des ressources telles que des scripts pour des opérations de map et/ou reduce personnalisées, ainsi que des bibliothèques supplémentaires. Pour plus d'informations, consultez Apache Hive[32] dans *Guide de version Amazon EMR*.

Pig

Pig est un package analytique open source qui fonctionne sur Hadoop. Pig fonctionne avec Pig Latin, un langage de type SQL qui permet aux utilisateurs de structurer, synthétiser et interroger des données. En plus des opérations de type SQL, Pig Latin ajoute également une assistance de première catégorie en ce qui concerne les fonctions de traitement (« map ») et d'agrégation (« reduce ») et les données complexes extensibles définies par l'utilisateur. Cette fonctionnalité permet le traitement de sources de données complexes et non structurées telles que des documents texte et des fichiers journaux.

Pig permet d'utiliser des extensions utilisateur, grâce aux fonctions définies par l'utilisateur et écrites en langage Java. Amazon EMR a apporté de nombreuses améliorations à Pig, notamment la possibilité d'utiliser plusieurs systèmes de fichiers (normalement, Pig ne peut accéder qu'à un seul système de fichiers distant), de charger des fichiers JAR et des scripts Amazon S3 (comme « REGISTER s3://my-bucket/piggybank.jar ») et des fonctionnalités supplémentaires pour le traitement de chaînes et de la date/heure. Pour plus d'informations, consultez Apache Pig[33] dans *Guide de version Amazon EMR*.

Spark

Spark est un moteur d'analyse de données open source basé sur Hadoop et reprenant les principes de base de MapReduce en mémoire. Spark fournit une vitesse supplémentaire pour certaines analyses et est à la base de la puissance d'autres outils tels que Shark (système de stockage de données piloté par SQL), Spark Streaming (applications de streaming), GraphX (systèmes graphiques) et MLlib (apprentissage machine). Pour plus d'informations, consultez l'article du blog Installation d'Apache Spark sur un cluster Amazon EMR.[34]

HBase

HBase est une base de données open source distribuée et non relationnelle, conçue sur le modèle de BigTable de Google. Elle a été développée dans le cadre du projet Hadoop d'Apache Software Foundation et fonctionne avec le système de fichiers distribué Hadoop (HDFS) afin de lui fournir des capacités comparables à celles de BigTable. HBase vous fournit un moyen efficace et tolérant aux pannes de stocker de grandes quantités de données dispersées à l'aide de la compression et du stockage basés sur des colonnes. En outre, HBase permet une recherche rapide des données, car celles-ci sont stockées en mémoire et non sur le disque.

HBase est optimisé pour les opérations d'écriture séquentielle et très efficace pour l'insertion, la mise à jour et la suppression de lots. HBase fonctionne de manière fluide avec Hadoop en partageant son système de fichiers et en servant d'entrée et de sortie directe pour les tâches dans Hadoop. Il intègre également Apache Hive, ce qui permet les requêtes de type SQL sur les tables HBase, se joint aux tables Hive et permet la prise en charge de Java Database Connectivity (JDBC). Avec Amazon EMR, vous pouvez sauvegarder HBase dans Amazon S3 (de façon totale ou incrémentielle, manuelle ou automatisée) et effectuer des restaurations à partir d'une sauvegarde créée précédemment. Pour plus d'informations, consultez HBase et EMR[35] dans le *Guide du développeur Amazon EMR*.

Impala

Impala est un outil open source dans l'écosystème Hadoop. Il est utilisé pour effectuer des requêtes interactives et ad hoc à l'aide de la syntaxe SQL. Au lieu d'utiliser MapReduce, Impala exploite un moteur de traitement massivement parallèle (MPP) similaire à celui des systèmes de gestion de bases de données relationnelles traditionnels (SGBDR). Grâce à cette architecture, vous pouvez interroger très rapidement vos données dans des tables HDFS ou HBase et tirer parti de la capacité de Hadoop à traiter divers types de données et à fournir des schémas lors de l'exécution. Cela fait d'Impala un excellent outil pour effectuer des analyses interactives à faible latence.

Impala dispose également de fonctions définies par l'utilisateur en Java et en C++, et peut se connecter aux outils BI via les pilotes ODBC et JDBC. Impala utilise le métastore Hive pour stocker des informations sur les données d'entrée, y compris les noms de partition et les types de données. Pour plus d'informations, consultez Impala and EMR[36] dans le *Guide du développeur Amazon EMR*.

Hunk

Hunk a été développé par Splunk pour rendre les données de la machine accessibles, utilisables et utiles pour tous. Avec Hunk, vous pouvez explorer, analyser et visualiser de manière interactive les données stockées dans Amazon EMR et Amazon S3, en exploitant les analyses Splunk sur Hadoop. Pour plus d'informations, consultez la section Amazon EMR avec Hunk : Splunk Analytics pour Hadoop et NoSQL.[37]

Presto

Presto est un moteur de requête SQL distribué open source optimisé pour l'analyse ad hoc de données à faible latence. Il prend en charge la norme ANSI SQL, y compris les requêtes complexes, les agrégations, les jonctions et les fonctions de fenêtrage. Presto peut traiter des données provenant de plusieurs sources de données, notamment le système de fichiers distribué Hadoop (HDFS) et Amazon S3.

Autres outils tiers

Amazon EMR prend également en charge une variété d'applications et outils courants dans l'écosystème Hadoop, tels que R (statistiques), Mahout (apprentissage-machine), Ganglia (surveillance), Accumulo (base de données NoSQL sécurisée), Hue (interface utilisateur pour analyser les données Hadoop), Sqoop (connecteur de base de données relationnelle), HCatalog (gestion de tables et de stockage), etc.

De plus, vous pouvez installer votre propre logiciel sur Amazon EMR pour vous aider à répondre aux besoins de votre entreprise. AWS de déplacer rapidement de grandes quantités de données d'Amazon S3 vers HDFS, de HDFS vers Amazon S3 et entre les compartiments Amazon S3 en utilisant S3DistCp[38] d'Amazon EMR, une extension de l'outil open source DistCp qui utilise MapReduce pour déplacer efficacement de grandes quantités de données.

Vous pouvez éventuellement utiliser le système de fichiers EMR (EMRFS), une implémentation de HDFS qui permet aux clusters Amazon EMR de stocker des données sur Amazon S3. Vous pouvez activer Amazon S3 à l'aide du chiffrement côté serveur et côté client Amazon S3, ainsi qu'une vue constante d'EMRFS. Lorsque vous utilisez EMRFS, un magasin de métadonnées est intégré de manière transparente dans DynamoDB pour vous aider à gérer les interactions avec Amazon S3 et vous permet d'avoir plusieurs clusters EMR utilisant facilement les mêmes métadonnées et le même stockage EMRFS sur Amazon S3.

Anti-modèles

Amazon EMR possède les anti-modèles suivants :

- **Petits ensembles de données** : Amazon EMR est conçu pour le traitement en parallèle massif. Si votre ensemble de données est assez petit pour s'exécuter rapidement sur une seule machine, en un seul thread, la surcharge ajoutée à des tâches de mappage et de réduction peut ne pas

valoir la peine pour de petits ensembles de données pouvant facilement être traités en mémoire sur un seul système.

- **Exigences de transaction ACID** : bien qu'il existe des moyens d'obtenir des propriétés ACID (atomicité, cohérence, isolation et durabilité) pleines ou limitées sur Hadoop, une autre base de données, telle qu'Amazon RDS ou une base de données relationnelle s'exécutant sur Amazon EC2, est sans doute la meilleure option pour les charges de travail exigeantes.

Amazon Machine Learning

Amazon ML[39] est un service qui facilite l'utilisation de l'analyse prédictive et de la technologie d'apprentissage machine. Amazon ML fournit des outils de visualisation et des assistants pour vous guider dans le processus de création de modèles d'apprentissage machine (ML) sans avoir à apprendre les algorithmes et la technologie ML complexes. Une fois vos modèles prêts, Amazon ML facilite la génération de prédictions pour votre application à l'aide de simples API, sans devoir implémenter un code de génération de prévision personnalisé ou gérer une infrastructure.

Amazon ML peut créer des modèles ML basés sur des données stockées dans Amazon S3, Amazon Redshift ou Amazon RDS. Des assistants intégrés vous guident à travers les étapes de l'exploration interactive de vos données, de la formation du modèle ML, de l'évaluation de la qualité du modèle et de l'ajustement des résultats afin de les aligner sur les objectifs commerciaux. Une fois qu'un modèle est prêt, vous pouvez demander des prédictions soit par lot, soit en utilisant l'API en temps réel à faible latence.

Modèles d'utilisation préconisés

Amazon ML est idéal pour découvrir des modèles dans vos données et utiliser ces derniers afin de créer des modèles ML qui peuvent générer des prédictions sur de nouveaux points de données invisibles. Par exemple, vous pouvez :

- **Activer les applications pour signaler les transactions douteuses** : créer un modèle d'apprentissage-machine qui prédit si une nouvelle transaction est légitime ou frauduleuse.
- **Prévoir la demande en produits** : saisissez les informations des commandes historiques pour prévoir les quantités de commandes futures.
- **Personnaliser le contenu de l'application** : prédisez les éléments qui intéressent le plus l'utilisateur et récupérez ces prédictions depuis votre application en temps réel.

- **Prévoir l'activité des utilisateurs** : analysez le comportement des utilisateurs pour personnaliser votre site web et offrir une meilleure expérience utilisateur.
- **Ecouter des réseaux sociaux** : intégrez et analysez les réseaux sociaux susceptibles d'avoir une incidence sur les décisions commerciales.

Coût du modèle

Avec Amazon ML, vous ne payez que pour ce que vous utilisez. Pas de frais minimum, pas d'engagement initial. Amazon ML facture un taux horaire pour le temps de calcul utilisé afin de créer des modèles prédictifs, puis vous payez pour le nombre de prédictions générées pour votre application. En ce qui concerne les prédictions en temps réel, vous payez également un tarif de capacité horaire basé sur la quantité de mémoire requise pour exécuter votre modèle.

La facturation pour l'analyse des données, la formation des modèles et l'évaluation est basée sur le nombre d'heures de calcul nécessaires à leur réalisation et dépend de la taille des données d'entrée, du nombre d'attributs, ainsi que du nombre et des types de transformations appliquées. L'analyse des données et les frais de création de modèle coûtent 0,42 USD l'heure. Les frais de prévision sont classés par lots et en temps réel. Les prédictions de lots sont de 0,10 USD toutes les 1 000 prédictions, arrondies aux 1 000 suivantes, tandis que les prédictions en temps réel sont de 0,0001 USD par prédiction, arrondies au centime près. Pour les prédictions en temps réel, des frais de capacité de 0,001 USD par heure sont également réservés pour chaque tranche de 10 Mo de mémoire allouée à votre modèle.

Lors de la création du modèle, vous spécifiez la taille de mémoire maximale de chaque modèle pour gérer les coûts et contrôler les performances prédictives. Vous ne payez les frais de capacité réservés que lorsque votre modèle est activé pour les prévisions en temps réel. Les frais applicables aux données stockées dans Amazon S3, Amazon RDS ou Amazon Redshift sont facturés séparément. Pour plus d'informations, consultez la page Tarifs d'Amazon Machine Learning.[40]

Performances

Le temps nécessaire pour créer des modèles ou pour demander des prédictions par lots à partir de ces modèles dépend du nombre d'enregistrements de données d'entrée, des types et de la distribution des attributs dans ces enregistrements et de la complexité de la « recette » de traitement que vous spécifiez.

La plupart des requêtes de prédictions en temps réel renvoient une réponse en moins de 100 ms, ce qui les rend assez rapides pour les applications web, mobiles ou de bureau interactives. La durée exacte nécessaire à l'API en temps réel pour générer une prédiction varie en fonction de la taille de l'enregistrement des données d'entrée, et de la complexité de la « recette » [41] du traitement des données associée au modèle d'apprentissage-machine qui génère les prédictions. Chaque modèle d'apprentissage-machine activé pour les prédictions en temps réel peut être utilisé pour demander jusqu'à 200 transactions par seconde par défaut, et ce nombre peut être augmenté en contactant le service client. Vous pouvez surveiller le nombre de prédictions demandées par vos modèles d'apprentissage-machine à l'aide des métriques CloudWatch.

Durabilité et disponibilité

Amazon ML est conçu pour une haute disponibilité. Il n'y a ni fenêtres de maintenance ni arrêts programmés. Le service s'exécute dans les centres de données fiables et hautement disponibles d'Amazon. En outre, cette API offre une réplication de la pile de services configurée sur trois sites dans chaque région AWS afin d'assurer la tolérance aux pannes en cas de défaillance du serveur ou de panne de zone de disponibilité.

Évolutivité et souplesse

Vous pouvez traiter des ensembles de données allant jusqu'à 100 Go pour créer des modèles ML ou pour demander des prédictions par lots. Pour les grands volumes de prédictions par lots, vous pouvez diviser vos enregistrements de données d'entrée en blocs distincts afin de permettre le traitement d'un volume de données de prédiction plus important.

Par défaut, vous pouvez exécuter jusqu'à cinq tâches simultanées. En contactant le service clientèle, vous pouvez augmenter cette limite. Étant donné qu'Amazon ML est un service géré, il n'y a pas de serveurs à provisionner et, par conséquent, vous pouvez évoluer en fonction de la croissance de votre application sans devoir provisionner ou payer pour des ressources non utilisées.

Interfaces

La création d'une source de données est aussi simple que l'ajout de vos données à Amazon S3. Vous pouvez aussi extraire des données directement à partir des bases de données Amazon Redshift ou MySQL gérées par Amazon RDS. Une fois

votre source de données définie, vous pouvez interagir avec Amazon ML à l'aide de la console. L'accès par programme à Amazon ML est activé par les SDK AWS et Amazon ML.[42] Vous pouvez également créer et gérer des entités Amazon Machine Learning à l'aide de l'interface de ligne de commande AWS disponible sur les systèmes Windows, Mac et Linux / UNIX.

Anti-modèles

Amazon ML possède les anti-modèles suivants :

- **Très grands ensembles de données** : même si Amazon ML peut prendre en charge jusqu'à 100 Go de données, l'ingestion de données à l'échelle du téraoctet n'est pas prise en charge à l'heure actuelle. Amazon EMR est un outil tout à fait indiqué pour exécuter Machine Learning Library (MLlib) de Spark.
- **Tâches d'apprentissage non prises en charge** : Amazon ML peut être utilisé pour créer des modèles ML qui effectuent une classification binaire (choisir l'un des deux choix, et fournir une mesure de confiance), une classification multiclasse (étendre les choix à plus de deux options) ou une régression numérique (prédire un nombre directement). Les tâches ML non prises en charge telles que la prédiction de séquence ou la mise en cluster non supervisée peuvent être abordées en utilisant Amazon EMR pour exécuter Spark et MLlib.

Amazon DynamoDB

Amazon DynamoDB[43] est un service de base de données NoSQL rapide et entièrement géré qui simplifie et rentabilise le stockage et la récupération de toute quantité de données, ainsi que le traitement du trafic des requêtes. DynamoDB contribue à alléger la charge administrative liée à l'exploitation et à la mise à l'échelle d'un cluster de bases de données distribuées hautement disponible. Cette solution de stockage répond aux exigences de latence et de débit des applications très exigeantes en fournissant une latence de quelques millisecondes seulement et des performances prévisibles, avec un débit fluide et une évolutivité du stockage.

DynamoDB stocke les données structurées dans des tables, indexées par clé primaire, et permet un accès en lecture et en écriture à faible latence aux éléments allant de 1 octet à 400 Ko. DynamoDB prend en charge trois types de

données (nombre, chaîne et binaire), dans des ensembles scalaires et à valeurs multiples. Il prend en charge les magasins de documents tels que JSON, XML ou HTML dans ces types de données. Les tables n'ont pas de schéma fixe, de sorte que chaque élément de données peut avoir un nombre différent d'attributs. La clé primaire peut être une clé de hachage composite ou à un seul attribut.

DynamoDB propose à la fois des index secondaires globaux et locaux offrant une flexibilité supplémentaire pour l'interrogation d'attributs autres que la clé primaire. DynamoDB fournit à la fois des lectures constantes (par défaut) et des lectures très constantes (facultatif), ainsi que des transactions implicites au niveau des éléments pour leur mise en place, les mises à jour, les suppressions, les opérations conditionnelles et l'incrémentation/décrémentation.

DynamoDB est intégré à d'autres services, tels qu'Amazon EMR, Amazon Redshift, AWS Data Pipeline et Amazon S3, pour l'analyse, l'entrepôt de données, l'importation/exportation de données, la sauvegarde et l'archivage.

Modèles d'utilisation préconisés

DynamoDB est idéal pour les applications existantes ou nouvelles qui ont besoin d'une base de données NoSQL flexible avec de faibles latences de lecture et d'écriture et la possibilité d'augmenter ou de réduire le débit et le stockage sans modification de code ni de temps d'arrêt.

Les cas d'utilisation courants incluent les exemples suivants :

- Applications mobiles
- Jeux
- Diffusion de publicités numériques
- Vote en direct
- Interaction de l'audience pour les événements en direct
- Réseaux de capteurs
- Ingestion de fichiers journaux
- Contrôle d'accès pour le contenu web
- Stockage de métadonnées pour les objets Amazon S3
- Panier pour les achats en ligne
- Gestion de session web

La plupart de ces cas d'utilisation nécessitent une base de données hautement disponible et évolutive, car les temps d'arrêt ou la dégradation des performances ont un impact négatif immédiat sur les activités d'une entreprise.

Coût du modèle

Avec DynamoDB, vous ne payez que pour ce que vous utilisez et il n'y a pas de frais minimum. DynamoDB dispose de trois composantes de tarification : capacité de débit provisionnée (par heure), stockage de données indexées (par Go par mois), transfert de données interne ou externe (par Go et par mois). Les nouveaux clients peuvent commencer à utiliser DynamoDB gratuitement dans le cadre de l'utilisation gratuite d'AWS.[44] Pour plus d'informations, consultez la section Tarifs d'Amazon DynamoDB.[45]

Performances

Les disques SSD et l'indexation limitative sur les attributs fournissent un haut débit et une faible latence[46] et réduisent considérablement le coût des opérations de lecture et d'écriture. Au fur et à mesure de la croissance des données, des performances prévisibles sont requises afin de maintenir une faible latence pour les charges de travail. Vous pouvez obtenir ces performances prévisibles en définissant la capacité de débit provisionnée requise pour une table donnée.

En coulisse, le service gère l'allocation des ressources pour atteindre le débit demandé. Vous n'avez pas à vous soucier des instances, du matériel, de la mémoire ou d'autres facteurs susceptibles d'affecter le débit d'une application. Vous pouvez réserver les capacités de débit allouées librement. Elles peuvent être augmentées ou diminuées à la demande.

Durabilité et disponibilité

DynamoDB dispose d'une tolérance de panne intégrée qui réplique automatiquement et de manière synchrone les données de trois centres de données dans une région pour plus de disponibilité, mais aussi pour protéger les données contre les pannes individuelles de machines, voire d'installations. DynamoDB Streams[47] capture l'activité des données de votre table et permet de configurer la réplication régionale d'une région géographique à l'autre pour plus de disponibilité.

Évolutivité et souplesse

DynamoDB est à la fois hautement évolutif et souple. Il n'y a pas de limite à la quantité de données que vous pouvez stocker dans une table DynamoDB et le service alloue automatiquement plus d'espace lorsque vous stockez davantage de

données à l'aide des API d'écriture DynamoDB. Les données sont partitionnées automatiquement et en fonction de vos besoins, tandis que l'utilisation des disques SSD fournit des temps de réponse à faible latence à n'importe quelle échelle. De plus, le service est assez souple, dans la mesure où il suffit d'« augmenter »[48] ou de« réduire »[49] la capacité de lecture et d'écriture d'une table en fonction de l'évolution de vos besoins.

Interfaces

DynamoDB fournit une API REST de bas niveau, ainsi que des SDK de niveau supérieur pour Java, .NET et PHP qui enveloppent l'API REST de bas niveau et fournissent des fonctions de correspondance objet-relationnel (ORM). Ces API fournissent une interface de gestion et de données pour DynamoDB. L'API offre actuellement des opérations qui permettent la gestion de tables (création, listage, suppression et obtention de métadonnées) et l'utilisation d'attributs (obtention, écriture et suppression d'attributs, interrogation à l'aide d'un index et analyse complète).

Bien que le SQL standard ne soit pas disponible, vous pouvez utiliser l'opération de sélection DynamoDB pour créer des requêtes de type SQL qui extraient un ensemble d'attributs en fonction des critères que vous fournissez. Vous pouvez également utiliser DynamoDB à l'aide de la console.

Anti-modèles

DynamoDB possède les anti-modèles suivants :

- **Application pré-écrite liée à une base de données relationnelle classique** : si vous tentez de déposer une application existante sur le cloud AWS et que vous devez continuer d'utiliser une base de données relationnelle, vous avez la possibilité d'utiliser Amazon RDS (Amazon Aurora, MySQL, PostgreSQL, Oracle ou SQL Server) ou l'une des nombreuses bases de données AMI Amazon EC2 préconfigurées. Vous pouvez également installer le logiciel de base de données de votre choix sur une instance EC2 que vous gérez.
- **Jonctions ou transactions complexes** : tandis que de nombreuses solutions sont en mesure de tirer parti de DynamoDB pour prendre en charge leurs utilisateurs, il est possible que votre application nécessite des jonctions, des transactions complexes et d'autres infrastructures

relationnelles fournies par les plates-formes de base de données traditionnelles. Si tel est le cas, vous pouvez étudier Amazon Redshift, Amazon RDS ou Amazon EC2 avec une base de données autogérée.

- **Données de grands objets binaires (BLOB)** : si vous envisagez de stocker des données BLOB volumineuses (supérieures à 400 Ko), telles que des vidéos, des images ou de la musique, envisagez d'utiliser Amazon S3. Toutefois, DynamoDB a encore un rôle à jouer dans ce scénario, en assurant le suivi des métadonnées (par ex. nom de l'élément, taille, date de création, propriétaire, emplacement, etc.) concernant vos objets binaires.
- **Données volumineuses avec un faible taux d'E/S** : DynamoDB utilise des disques SSD et est optimisé pour les charges de travail avec un taux d'E/S élevé par Go stocké. Si vous envisagez de stocker de très grandes quantités de données auxquelles vous accédez rarement, il serait peut-être plus sage d'opter pour d'autres options de stockage telles qu'Amazon S3.

Amazon Redshift

Amazon Redshift[50] est un service d'entrepôt de données rapide, entièrement géré et d'une capacité de plusieurs pétaoctets qui permet d'analyser de manière simple et rentable toutes vos données à l'aide de vos outils d'aide à la décision existants. Il est optimisé pour les ensembles de données allant de quelques centaines de gigaoctets à un ou plusieurs pétaoctets, et est conçu pour coûter moins d'un dixième du coût de la plupart des solutions traditionnelles d'entreposage de données.

Amazon Redshift fournit des performances de requête et d'E/S rapides pour pratiquement n'importe quel ensemble de données, quelle que soit sa taille, en faisant appel à la technologie de stockage en colonnes, tout en parallélisant et en distribuant des requêtes sur plusieurs nœuds. Il automatise la plupart des tâches administratives courantes associées à l'allocation, à la configuration, à la surveillance, à la sauvegarde et à la sécurisation d'un entrepôt de données, ce qui le rend facile et peu coûteux à gérer et à entretenir. Cette automatisation vous permet de créer des entrepôts de données à l'échelle du pétaoctet en quelques minutes au lieu des semaines ou mois nécessaires pour les implémentations traditionnelles sur site.

Modèles d'utilisation préconisés

Amazon Redshift est idéal pour le traitement analytique en ligne (OLAP) en utilisant vos outils de veille économique existants. Les entreprises utilisent Amazon Redshift pour effectuer les opérations suivantes :

- Analyser les données de ventes globales pour plusieurs produits
- Stocker les données boursières historiques
- Analyser les impressions publicitaires et les clics
- Agréger les données de jeu
- Analyser les tendances sociales
- Mesurer la qualité clinique, l'efficacité des opérations et la performance financière dans les soins de santé

Coût du modèle

Un cluster d'entrepôt de données Amazon Redshift ne nécessite aucun engagement à long terme ou coûts initiaux. De cette manière, vous n'avez pas à subir les dépenses d'investissement et la complexité qui accompagnent souvent la planification et l'achat de capacités en amont. Les frais sont calculés en fonction de la taille et du nombre de nœuds de votre cluster.

Il n'y a pas de frais supplémentaires si votre stockage de sauvegarde est inférieur ou égal à 100 % du stockage mis en service. Par exemple, si vous avez un cluster actif avec 2 nœuds XL, pour un total de 4 To de stockage, AWS fournit jusqu'à 4 To de stockage de sauvegarde sur Amazon S3, sans frais supplémentaires. Le stockage de sauvegarde au-delà de la taille de stockage provisionnée et les sauvegardes stockées une fois votre cluster fermé seront facturés aux Tarifs d'Amazon S3[51] standard. Aucuns frais de transfert de données ne sont appliqués pour la communication entre Amazon S3 et Amazon Redshift. Pour plus d'informations, consultez la page de présentation des Tarifs d'Amazon Redshift.[52]

Performances

Amazon Redshift repose sur différentes innovations afin d'offrir des performances très élevées en matière d'interrogation sur des ensembles de données dont la taille peut aller d'une centaine de gigaoctets à plusieurs pétaoctets ou plus. Il utilise un stockage en colonnes, la compression des données et les cartes de zones pour réduire le volume d'E/S nécessaires à l'exécution des requêtes.

Amazon Redshift dispose d'une architecture d'entrepôt de données à traitement massivement parallèle (MPP, Massively Parallel Processing), qui lui permet de traiter en parallèle les opérations SQL et de les distribuer afin d'exploiter toutes les ressources disponibles. Le matériel sous-jacent a été conçu pour un traitement des données extrêmement performant. Il utilise le stockage connecté local afin d'optimiser le débit entre les processeurs et les lecteurs, ainsi qu'un réseau maillé 10 GigE dans le but d'optimiser le débit entre les nœuds. Les performances peuvent être ajustées en fonction de vos besoins d'entreposage de données : AWS propose des calculs denses (DC) avec des disques SSD, ainsi que des options de stockage denses (DS).

Durabilité et disponibilité

Amazon Redshift détecte et remplace automatiquement un nœud défaillant dans votre cluster d'entrepôt de données. Le cluster d'entrepôt de données est en lecture seule jusqu'à ce qu'un nœud de remplacement soit provisionné et ajouté à la base de données, ce qui ne prend généralement que quelques minutes. Amazon Redshift met immédiatement à disposition votre nœud de remplacement et diffuse d'abord les données les plus fréquemment consultées depuis Amazon S3 pour vous permettre de reprendre l'interrogation de vos données le plus rapidement possible.

De plus, votre cluster d'entrepôt de données reste disponible en cas de panne d'un lecteur. Comme Amazon Redshift met en miroir vos données sur le cluster, il utilise les données à partir d'un autre nœud pour recréer les disques défaillants. Les clusters Amazon Redshift résident dans une zone de disponibilité,[53] mais si vous souhaitez disposer d'une configuration multi-AZ pour Amazon Redshift, vous pouvez configurer un miroir, puis gérer automatiquement la réplication et le basculement.

Évolutivité et souplesse

En quelques clics dans la console ou via un appel d'API,[54] vous pouvez facilement modifier le nombre ou le type de nœuds dans votre entrepôt de données à mesure que vos performances ou vos besoins en capacités évoluent. Amazon Redshift vous permet de commencer avec un seul nœud de 160 Go et de passer à un pétaoctet ou plus de données utilisateur compressées en utilisant plusieurs nœuds. Pour plus d'informations, consultez la section À propos des clusters et des nœuds,[55] dans la rubrique des clusters Amazon Redshift dans le *manuel de gestion d'Amazon Redshift*.

Lors du redimensionnement, Amazon Redshift place votre cluster existant en lecture seule, fournit un nouveau cluster de la taille choisie, puis copie les données de votre ancien cluster vers votre nouveau en parallèle. Au cours de ce processus, vous ne payez que pour le cluster Amazon Redshift actif. Pendant la mise en service du nouveau cluster, vous pouvez continuer à interroger l'ancien. Une fois vos données copiées dans votre nouveau cluster, Amazon Redshift redirige automatiquement les requêtes vers votre nouveau cluster et supprime l'ancien.

Interfaces

Amazon Redshift dispose de pilotes JDBC et ODBC personnalisés que vous pouvez télécharger depuis l'onglet Connect Client de la console, ce qui vous permet d'utiliser un large éventail de clients SQL familiers. Vous pouvez également utiliser les pilotes PostgreSQL JDBC et ODBC standard. Pour plus d'informations sur les pilotes Amazon Redshift, consultez la section Amazon Redshift et PostgreSQL.[56]

Il existe de nombreux exemples d'intégrations validées auprès de vendeurs de BI et d'ETL populaires.[57] Des chargements et déchargements sont tentés en parallèle dans chaque nœud de traitement afin de maximiser la vitesse à laquelle vous pouvez ingérer des données dans votre cluster d'entrepôt de données, ainsi qu'en provenance et à destination d'Amazon S3 et de DynamoDB. Vous pouvez facilement charger des données en continu dans Amazon Redshift à l'aide d'Amazon Kinesis Firehose, ce qui vous permet d'effectuer des analyses en temps quasi réel avec les outils d'analyse décisionnelle et les tableaux de bord existants que vous utilisez déjà. Les métriques relatives à l'utilisation des capacités de calcul, de mémoire et de stockage, ainsi que le trafic en lecture / écriture de votre cluster d'entrepôts de données Amazon Redshift sont disponibles gratuitement via la console ou les opérations de l'API CloudWatch.

Anti-modèles

Amazon Redshift possède les anti-modèles suivants :

- **Petits ensembles de données** : Amazon Redshift est conçu pour un traitement parallèle sur un cluster. Si votre ensemble de données est inférieur à une centaine de gigaoctets, vous ne pourrez pas bénéficier de tous les avantages d'Amazon Redshift. Dans ce cas, il vous est conseillé de vous tourner vers Amazon RDS.

- **Traitement des transactions en ligne (OLTP)** : Amazon Redshift est conçu pour les charges de travail d'entrepôt de données produisant des capacités analytiques extrêmement rapides et peu coûteuses. Si vous avez besoin d'un système transactionnel rapide, vous pouvez choisir un système de base de données relationnelle traditionnel conçu sur Amazon RDS ou une offre de base de données NoSQL, telle que DynamoDB.
- **Données non structurées** : les données d'Amazon Redshift doivent être structurées par un schéma défini plutôt que de prendre en charge une structure de schéma arbitraire pour chaque ligne. Si vos données ne sont pas structurées, vous pouvez extraire, transformer et charger (ETL) sur Amazon EMR afin que les données soient prêtes à être chargées dans Amazon Redshift.
- **Données BLOB** : si vous envisagez de stocker de gros fichiers binaires (tels que des vidéos, des images ou de la musique), vous pouvez envisager de stocker les données dans Amazon S3 et de référencer leur emplacement dans Amazon Redshift. Dans ce scénario, Amazon Redshift effectue le suivi des métadonnées (telles que le nom, la taille, la date de création, le propriétaire, l'emplacement, etc.) de vos objets binaires, mais les objets volumineux eux-mêmes sont stockés dans Amazon S3.

Amazon Elasticsearch Service

Amazon ES[58] est un service géré qui facilite le déploiement, le fonctionnement et l'évolutivité d'Elasticsearch dans le cloud AWS. Elasticsearch est un moteur de recherche et d'analyse distribué en temps réel. Il vous permet d'explorer vos données à une vitesse et à une échelle jamais égalée jusqu'ici. Il est utilisé pour la recherche en texte intégral, la recherche structurée et l'analyse. Il est même possible de faire les 3 en même temps.

Vous pouvez installer et configurer votre cluster Amazon ES en quelques minutes à l'aide de la console. Amazon ES gère les tâches inhérentes à la configuration d'un domaine ; de la mise à disposition de la capacité d'infrastructure demandée à l'installation du logiciel Elasticsearch.

Une fois votre domaine actif, Amazon ES automatise les tâches d'administration courantes, telles que l'exécution de sauvegardes, la surveillance des instances et l'application de correctifs au logiciel qui alimente votre instance Amazon ES. Il détecte et remplace automatiquement les nœuds Elasticsearch défaillants, réduisant ainsi les frais généraux associés aux

infrastructures autogérées et au logiciel Elasticsearch. Le service vous permet d'adapter facilement votre cluster via un seul appel d'API ou quelques clics dans la console.

Avec Amazon ES, vous bénéficiez d'un accès direct à l'API open source Elasticsearch afin que le code et les applications que vous utilisez avec vos environnements Elasticsearch existants fonctionnent sans difficulté. Il prend en charge l'intégration avec Logstash, un pipeline de données open source qui vous aide à traiter les journaux et autres données d'événements. Il inclut également une prise en charge intégrée de Kibana, une plate-forme d'analyse et de visualisation open source qui vous aide à mieux comprendre vos données.

Modèles d'utilisation préconisés

Amazon ES est idéal pour interroger et rechercher de grandes quantités de données.

Les entreprises peuvent utiliser Amazon ES pour effectuer les opérations suivantes :

- Analyser les journaux d'activité, tels que les journaux pour les applications ou les sites web destinés aux clients
- Analyser les journaux CloudWatch avec Elasticsearch
- Analyser les données d'utilisation de produit provenant de divers services et systèmes
- Analyser les ressentis des réseaux sociaux et les données de CRM, et trouver des tendances pour les marques et les produits
- Analyser les mises à jour de flux de données provenant d'autres services AWS, tels que Amazon Kinesis Streams et DynamoDB
- Fournir aux clients une expérience de recherche et de navigation riche
- Surveiller l'utilisation des applications mobiles

Coût du modèle

Avec Amazon ES, vous ne payez que pour les ressources de calcul et de stockage que vous utilisez. Pas de frais minimum, pas d'engagement initial. Vous êtes facturé pour les heures d'instance Amazon ES, le stockage Amazon EBS (si vous avez sélectionné cette option) et les frais standard de transfert de données.[59]

Si vous utilisez des volumes EBS pour le stockage, Amazon ES vous permet de choisir le type de volume. Si vous choisissez le stockage SSD (Provisioned IOPS),[60]

vous êtes facturé pour le stockage, ainsi que pour le débit que vous provisionnez. Toutefois, vous n'êtes pas facturé pour les E/S que vous consommez. Vous avez également la possibilité de payer un stockage supplémentaire en fonction de la taille cumulée des volumes EBS attachés aux nœuds de données de votre domaine.

Pour chaque domaine Amazon ES, vous obtenez un espace de stockage gratuit pour les instantanés automatiques. Les instantanés manuels sont facturés en fonction des taux de stockage Amazon S3. Pour plus d'informations, consultez Tarifs du Amazon Elasticsearch Service.[61]

Performances

Les performances d'Amazon ES dépendent de plusieurs facteurs, notamment le type d'instance, la charge de travail, l'index, le nombre de partitions utilisées, les réplicas en lecture et les configurations de stockage (stockage d'instance ou stockage EBS, par exemple les disques SSD). Les index sont constitués de partitions de données pouvant être distribuées sur différentes instances dans plusieurs zones de disponibilité.

Les répliques en lecture des partitions sont conservées par Amazon ES dans une zone de disponibilité différente si la détection de zone est activée. Amazon ES peut utiliser le stockage d'instance SSD rapide afin de stocker des index ou plusieurs volumes EBS. Un moteur de recherche permet une utilisation intensive des périphériques de stockage et rend les disques plus rapides, ce qui accélère les requêtes et les performances de recherche.

Durabilité et disponibilité

Vous pouvez configurer la haute disponibilité de vos domaines Amazon ES en activant l'option « Zone Awareness » (Prise en compte des zones) au moment de la création du domaine ou lors de la modification d'un domaine actif. Lorsque Zone Awareness est activée, Amazon ES distribue les instances prenant en charge le domaine dans deux zones de disponibilité différentes. Ensuite, si vous activez des réplicas dans Elasticsearch, les instances sont automatiquement distribuées de manière à fournir une réplication sur plusieurs zones.

Vous pouvez créer une durabilité des données pour votre domaine Amazon ES grâce à des instantanés automatiques et manuels. Vos instantanés pourront vous permettre de restaurer votre domaine ou d'en créer un nouveau à partir de données préchargées. Les instantanés sont stockés dans Amazon S3, qui est un

espace de stockage d'objets sécurisé, durable et hautement évolutif. Par défaut, Amazon ES crée automatiquement des instantanés quotidiens de chaque domaine. En outre, vous pouvez utiliser les API d'instantané Amazon ES pour créer des instantanés manuels supplémentaires. Les instantanés manuels sont stockés dans Amazon S3. Les instantanés manuels peuvent être utilisés pour la reprise après sinistre entre régions et pour fournir une durabilité supplémentaire.

Évolutivité et souplesse

Vous pouvez ajouter ou supprimer des instances et modifier facilement les volumes Amazon EBS pour prendre en charge la croissance des données. Vous pouvez écrire quelques lignes de code pour surveiller l'état de votre domaine via les métriques d'Amazon CloudWatch et faire appel à l'API Amazon ES pour augmenter ou diminuer la capacité de votre domaine en fonction des seuils que vous définissez. Le service exécute la mise à l'échelle sans aucun temps d'arrêt.

Amazon ES prend en charge un volume EBS (taille maximale de 512 Go) par instance associée à un cluster. Avec un maximum de 10 instances autorisées par cluster Amazon ES, les clients peuvent allouer environ 5 To de stockage à un seul domaine Amazon ES.

Interfaces

Amazon ES prend en charge l'API Elasticsearch,[62] de sorte que le code, les applications et les outils courants que vous utilisez avec les environnements Elasticsearch existants fonctionnent sans difficulté. Les kits SDK AWS prennent en charge toutes les opérations d'API Amazon ES, ce qui facilite la gestion et l'interaction avec vos domaines par le biais de votre technologie préférée. L'interface de ligne de commande AWS ou la console peut être utilisé pour la création et la gestion de vos domaines.

Amazon ES prend en charge l'intégration à plusieurs services AWS, y compris les données diffusées en continu à partir d'Amazon S3, Amazon Kinesis Streams et DynamoDB Streams. Les intégrations utilisent une fonction Lambda en tant que gestionnaire d'événements dans le cloud, qui répond aux nouvelles données en les traitant et les diffusant dans votre domaine Amazon ES. Amazon ES s'intègre également à CloudWatch pour surveiller les métriques de domaine Amazon ES, et à CloudTrail pour auditer la configuration des appels d'API vers des domaines Amazon ES.

Amazon ES inclut l'intégration avec Kibana, une plate-forme d'analyse et de visualisation open source, et prend en charge l'intégration à Logstash, un pipeline de données open source qui vous aide à traiter les journaux et autres données d'événements. Vous pouvez configurer votre domaine Amazon ES en tant que magasin principal pour tous les journaux issus de votre implémentation Logstash afin d'intégrer facilement des données structurées et non structurées provenant de diverses sources.

Anti-modèles

Amazon ES possède les anti-modèles suivants :

- **Traitement des transactions en ligne (OLTP)** : Amazon ES est un moteur de recherche et d'analyse distribué en temps réel. Il ne prend pas en charge les transactions ou les manipulations de données. Si vous avez besoin d'un système transactionnel rapide, alors vous devriez opter pour un système de base de données relationnelle traditionnel basé sur Amazon RDS, ou une base de données NoSQL offrant des fonctionnalités telles que DynamoDB.
- **Stockage de plusieurs péta-octets** : avec un maximum de 10 instances autorisées par cluster Amazon ES, vous pouvez allouer environ 5 To de stockage à un seul domaine Amazon ES. Pour les charges de travail plus importantes, envisagez d'utiliser Elasticsearch autogéré sur Amazon EC2.

Amazon QuickSight

En octobre 2015, AWS a présenté l'aperçu d'Amazon QuickSight, un service de Business Intelligence (BI) rapide et basé sur le cloud qui vous permet de créer facilement des visualisations, d'effectuer des analyses ad hoc et d'obtenir rapidement des informations commerciales à partir de vos données.

QuickSight utilise un nouveau moteur de calcul en mémoire (SPICE) parallèle, super rapide, pour effectuer des calculs avancés et afficher des visualisations rapidement. QuickSight s'intègre automatiquement avec les services de données AWS, permet aux entreprises d'atteindre des centaines de milliers d'utilisateurs et offre des performances de requête rapides et réactives via le moteur de recherche de SPICE. Au dixième du coût des solutions traditionnelles, QuickSight vous permet d'offrir des fonctionnalités BI à tous les membres de votre entreprise. Pour en savoir plus et vous inscrire à cette version d'évaluation, consultez QuickSight.[63]

Amazon EC2

Amazon EC2,[64] doté d'instances agissant comme des machines virtuelles AWS, constitue une plate-forme idéale pour l'exploitation de vos propres applications d'analyse du Big Data autogérées sur l'infrastructure AWS. Presque tous les logiciels que vous pouvez installer sur des environnements virtualisés Linux ou Windows peuvent être exécutés sur Amazon EC2 et vous pouvez utiliser le modèle de tarification par répartition. Ce que vous n'obtenez pas, ce sont les services gérés au niveau de l'application qui sont fournis avec les autres services mentionnés dans ce livre blanc. Il existe de nombreuses options pour l'analyse du Big Data auto-gérée. Voici quelques exemples :

- Une offre NoSQL, telle que MongoDB
- Un entrepôt de données ou un stockage en colonnes comme Vertica
- Un cluster Hadoop
- Un cluster Apache Storm
- Un environnement Apache Kafka

Modèles d'utilisation préconisés

- **Environnement spécialisé** : lors de l'exécution d'une application personnalisée, d'une variante d'un ensemble Hadoop standard ou d'une application non couverte par l'une de nos autres offres, Amazon EC2 offre la flexibilité et la scalabilité nécessaires pour répondre à vos besoins informatiques.
- **Exigences de conformité** : certaines exigences de conformité peuvent nécessiter que vous exécutiez des applications vous-même sur Amazon EC2, au lieu d'une offre de service géré.

Coût du modèle

Amazon EC2 propose une variété de types d'instances dans un certain nombre de familles d'instances (standard, processeur élevé, mémoire élevée, E/S élevées, etc.) et différentes options tarifaires (à la demande, réservée et ponctuelle). En fonction des besoins de votre application, il peut être judicieux d'utiliser des services supplémentaires avec Amazon EC2, tels qu'Amazon Elastic Block Store (Amazon EBS) pour le stockage persistant directement rattaché, ou Amazon S3 comme système de stockage d'objets durable. Chaque modèle a un prix différent. Si vous exécutez votre application Big Data sur Amazon EC2, vous êtes responsable des frais de licence, comme s'il s'agissait de

votre propre centre de données. AWS Marketplace[65] propose de nombreux progiciels Big Data tiers, préconfigurés pour être lancés en un clic.

Performances

Les performances dans Amazon EC2 dépendent du type d'instance que vous choisissez pour votre plate-forme Big Data. Chaque type d'instance dispose d'une quantité différente de CPU, de RAM, de stockage, d'IOP et de capacité de mise en réseau, ce qui vous permet de choisir le bon niveau de performance en fonction de vos besoins.

Durabilité et disponibilité

Les applications critiques doivent être exécutées dans un cluster sur plusieurs zones de disponibilité au sein d'une région AWS afin que toute défaillance d'instance ou de centre de données n'affecte pas les utilisateurs de l'application. Pour les applications critiques non actives, vous pouvez sauvegarder votre application sur Amazon S3 et la restaurer dans n'importe quelle zone de disponibilité de la région en cas de défaillance d'une instance ou d'une zone. D'autres options existent, selon l'application que vous utilisez et les exigences, telles que la mise en miroir de votre application.

Évolutivité et souplesse

Auto Scaling[66] est un service qui vous permet d'augmenter ou de réduire automatiquement la capacité de votre Amazon EC2 en fonction des conditions que vous définissez. Avec Auto Scaling, vous pouvez veiller à ce que le nombre d'instances EC2 augmente en toute transparence pendant les pics de la demande pour maintenir les performances et diminue automatiquement lors des creux de la demande afin de réduire les coûts. Auto Scaling convient particulièrement aux applications utilisées de façon variable sur le plan horaire, quotidien ou hebdomadaire. Auto Scaling est activé par CloudWatch et disponible sans frais supplémentaires à ceux de CloudWatch.

Interfaces

Amazon EC2 peut être interfacé par programmation via une API, des SDK, ou la console. Les métriques relatives à l'utilisation des capacités de calcul, de mémoire, de stockage et de consommation réseau, ainsi que le trafic en lecture / écriture de vos instances sont disponibles gratuitement via la console ou les opérations de l'API CloudWatch.

Les interfaces de votre logiciel d'analyse de données volumineuses que vous utilisez sur Amazon EC2 varient en fonction des caractéristiques du logiciel que vous choisissez.

Anti-modèles

Amazon EC2 possède les anti-modèles suivants :

- **Service géré** : si vous recherchez une offre de service géré où vous analysez la couche d'infrastructure et l'administration à partir de l'analyse du Big Data, ce modèle de gestion de votre propre logiciel d'analyse sur Amazon EC2 n'est peut-être pas le bon choix.
- **Manque d'expertise ou de ressources** : si votre entreprise ne dispose pas ou ne souhaite pas dépenser les ressources ou l'expertise nécessaires pour installer et gérer une installation à haute disponibilité pour le système en question, vous devriez envisager d'utiliser les services AWS équivalents tels que Amazon EMR, DynamoDB, Amazon Kinesis Streams ou Amazon Redshift.

Résolution des problèmes du Big Data sur AWS

Dans ce livre blanc, nous avons examiné certains outils à votre disposition sur AWS pour l'analyse du Big Data. C'est un bon point de référence lorsque vous commencez à concevoir vos applications Big Data. Cependant, il y a des aspects supplémentaires que vous devez prendre en compte lorsque vous choisissez les bons outils pour votre cas d'utilisation spécifique. En général, chaque charge de travail analytique aura certaines caractéristiques et exigences, qui détermineront les outils à utiliser. Par exemple :

- En combien de temps avez-vous besoin des résultats d'analyse ? En temps réel, en quelques secondes ou en une heure ?
- Quelle valeur ces analyses fourniront-elles à votre entreprise et quelles sont les contraintes budgétaires existantes ?
- Quelle est la taille des données et quel est son taux de croissance ?
- Comment les données sont-elles structurées ?
- Quelles sont les capacités d'intégration des producteurs et des consommateurs ?

- Quelle est la latence acceptable entre les producteurs et les consommateurs ?
- Quel est le coût des temps d'arrêt ou dans quelle mesure la solution doit-elle être disponible et durable ?
- La charge de travail analytique est-elle constante ou souple ?

Chacune de ces caractéristiques ou exigences vous aide à vous orienter dans la bonne direction pour utiliser l'outil. Dans certains cas, vous pouvez facilement faire correspondre votre charge de travail d'analyse de données volumineuses à l'un des services en fonction d'un ensemble d'exigences. Cependant, dans la plupart des charges de travail d'analyse de données volumineuses du monde réel, il existe de nombreuses caractéristiques et exigences, parfois conflictuelles, pour le même ensemble de données.

Par exemple, certains ensembles de résultats peuvent avoir des exigences en temps réel lorsqu'un utilisateur interagit avec un système, alors que d'autres analyses peuvent être groupées et exécutées quotidiennement. Ces différentes exigences pour un même ensemble de données doivent être découplées et résolues en utilisant plus d'un outil. Si vous essayez de résoudre ces deux exemples dans le même ensemble d'outils, soit vous finissez par sur-allouer, et par conséquent surpayer pour un temps de réponse inutile, soit votre solution n'est pas suffisante pour répondre à vos utilisateurs en temps réel. Si vous faites correspondre l'outil le mieux adapté à chaque ensemble de problèmes analytiques, les résultats reflètent l'utilisation la plus rentable de vos ressources de calcul et de stockage.

Le Big Data n'engendre pas forcément des coûts onéreux. Par conséquent, lorsque vous concevez vos applications, il est important de vérifier leur rentabilité. Si ce n'est pas rentable par rapport aux différentes solutions, c'est que la conception n'est pas optimale. Il existe une autre idée reçue : le fait de disposer de plusieurs ensembles d'outils pour résoudre un gros problème de données est plus onéreux ou plus difficile à gérer que le fait de disposer d'un seul outil conséquent. Si vous reprenez l'exemple des deux exigences différentes pour un même ensemble de données, la demande en temps réel peut être faible sur le processeur mais élevée sur les E/S, alors que les demandes qui exigent un traitement plus lent peuvent être très intensives. Le découplage peut s'avérer beaucoup moins onéreux et plus facile à gérer, car vous pouvez concevoir chaque outil selon des spécifications exactes et non de façon excessive. Avec l'AWS

prépayé et uniquement lorsque vous utilisez l'infrastructure en tant que modèle de service, l'offre est intéressante car vous pouvez exécuter l'analyse par lots en une heure seulement et ne payer que les ressources de calcul pour cette heure. Vous pouvez trouver cette approche plus facile à gérer, comparée à un système unique qui tente de répondre à toutes les exigences. Résoudre différentes exigences avec un seul outil, c'est comme essayer d'ajuster une cheville carrée (demandes en temps réel) dans un trou rond (grand entrepôt de données).

La plate-forme AWS facilite le découplage de votre architecture en permettant à différents outils d'analyser le même ensemble de données. Les services AWS sont intégrés, de sorte que le déplacement d'un sous-ensemble de données d'un outil à un autre peut être effectué très facilement et rapidement à l'aide de la parallélisation. Nous allons mettre cela en pratique en étudiant un certain nombre de scénarios de problèmes réels liés à l'analyse du Big Data et en parcourant une solution architecturale AWS.

Exemple 1 : Entrepôt de données d'entreprise

Un fabricant de vêtements multinational regroupe plus d'un millier de magasins de vente au détail, vend certaines lignes dans des grandes surfaces et des magasins de rabais, et dispose d'un site Internet. Ces trois chaînes semblent agir indépendamment l'une de l'autre sur le plan technique. Elles disposent de différents services de gestion, de points de vente et de comptabilité. Il n'existe aucun système qui fusionne tous ces ensembles de données pour permettre au PDG d'avoir un aperçu de l'entreprise. Le PDG souhaite disposer d'une image de ses chaînes à l'échelle de l'entreprise et être en mesure d'effectuer des analyses ad hoc en cas de besoin. Voici des exemples d'analyses désirées par l'entreprise :

- Quelles tendances existent parmi les différentes chaînes ?
- Quelles sont les régions géographiques les plus sollicitées parmi toutes les chaînes ?
- Quelle est l'efficacité de leurs publicités et coupons ?
- Quelles tendances existent parmi chaque ligne de vêtements ?
- Quelles forces externes peuvent avoir un impact sur les ventes, par exemple, le taux de chômage, ou la météo ?
- Comment les effets de magasin affectent-ils les ventes (par exemple les employés/la direction, un centre commercial linéaire par rapport à

un centre commercial fermé, l'emplacement de la marchandise dans le magasin, la promotion, les têtes de gondole, les circulaires, etc.) ?

Un entrepôt de données d'entreprise est un excellent moyen de résoudre ce problème. L'entrepôt de données doit collecter des données provenant des différents systèmes des trois chaînes et des registres publics pour les données météorologiques et économiques. Chaque source de données envoie les données de consommation sur une base quotidienne à l'entrepôt de données. Chaque source de données pouvant être structurée différemment, un processus ETL est effectué afin de reformater les données en une structure commune. Ensuite, l'analyse peut être effectuée simultanément sur les données de toutes les sources. Pour ce faire, nous utilisons l'architecture de flux de données suivante :

Entrepôt de données d'entreprise

1. La première étape de ce processus consiste à obtenir les données provenant de nombreuses sources différentes sur Amazon S3. Amazon S3 a été choisi parce qu'il s'agit d'une plate-forme de stockage hautement durable, peu coûteuse et évolutive qui peut être écrite en parallèle à partir de nombreuses sources différentes à un coût très faible.
2. Amazon EMR est utilisé pour transformer et nettoyer les données du format source vers la destination et un format. Amazon EMR est intégré à Amazon S3 afin d'autoriser les threads parallèles de débit de chaque nœud de votre cluster vers et depuis Amazon S3. Généralement, les entrepôts de données obtiennent de nouvelles données chaque soir à partir de ses nombreuses sources. Étant donné que ces analyses ne sont pas nécessaires au beau milieu de la nuit, il n'y a qu'une seule exigence. Le processus de transformation doit être terminé le matin, lorsque le PDG et les autres utilisateurs professionnels ont besoin des résultats. Grâce à cette exigence,

vous pouvez utiliser le marché Amazon EC2 Spot[67] pour réduire davantage le coût de la transformation. Une bonne stratégie Spot serait de commencer à enchérir à un prix très bas à minuit, et d'augmenter continuellement le prix au fil du temps, jusqu'à ce que la capacité soit accordée. À mesure que vous vous rapprochez de la date limite, si les enchères Spot n'ont pas été fructueuses, vous pouvez revenir à la tarification à la demande afin de vous assurer que vous répondez toujours aux exigences en termes de délai d'exécution. Chaque source peut avoir un processus de transformation différent sur Amazon EMR, mais avec le modèle AWS prépayé, vous pouvez créer un cluster Amazon EMR distinct pour chaque transformation et l'ajuster afin qu'il soit à la puissance nécessaire pour effectuer tous les travaux de transformation de données au prix le plus bas possible, sans rivaliser avec les ressources des autres tâches.

3. Chaque tâche de transformation transfère ensuite les données formatées et nettoyées vers Amazon S3. Amazon S3 est à nouveau utilisé ici, car Amazon Redshift peut consommer ces données sur plusieurs threads en parallèle depuis chaque nœud. Cet emplacement sur Amazon S3 sert également d'enregistrement historique et constitue une source fiable formatée entre les systèmes. Les données d'Amazon S3 peuvent être utilisées par d'autres outils d'analyse si des exigences supplémentaires sont introduites au fil du temps.
4. Amazon Redshift charge, trie, distribue et compresse les données dans ses tables afin que les requêtes analytiques puissent s'exécuter efficacement et en parallèle. Amazon Redshift est conçu pour les charges de travail de l'entrepôt de données et peut facilement être développé en ajoutant un autre nœud à mesure que la taille des données augmente avec le temps et que l'entreprise se développe.
5. Pour visualiser les analyses, vous pouvez utiliser Amazon QuickSight ou l'une des nombreuses plates-formes de visualisation partenaires via la connexion OBDC / JDBC à Amazon Redshift. C'est là que les rapports et les graphiques peuvent être consultés par le chef de la direction et son personnel. Ces données peuvent désormais être utilisées par les dirigeants pour prendre de meilleures décisions concernant les ressources de l'entreprise, ce qui peut finalement augmenter les bénéfices et la valeur pour les actionnaires.

Cette architecture est très flexible et peut facilement être étendue si l'entreprise se développe, si d'autres sources de données sont importées, si de nouveaux canaux sont ouverts ou si une application mobile est lancée avec des données spécifiques du client. À tout moment, des outils supplémentaires peuvent être intégrés et l'entrepôt peut être mis à l'échelle en quelques clics en augmentant le nombre de nœuds dans le cluster Amazon Redshift.

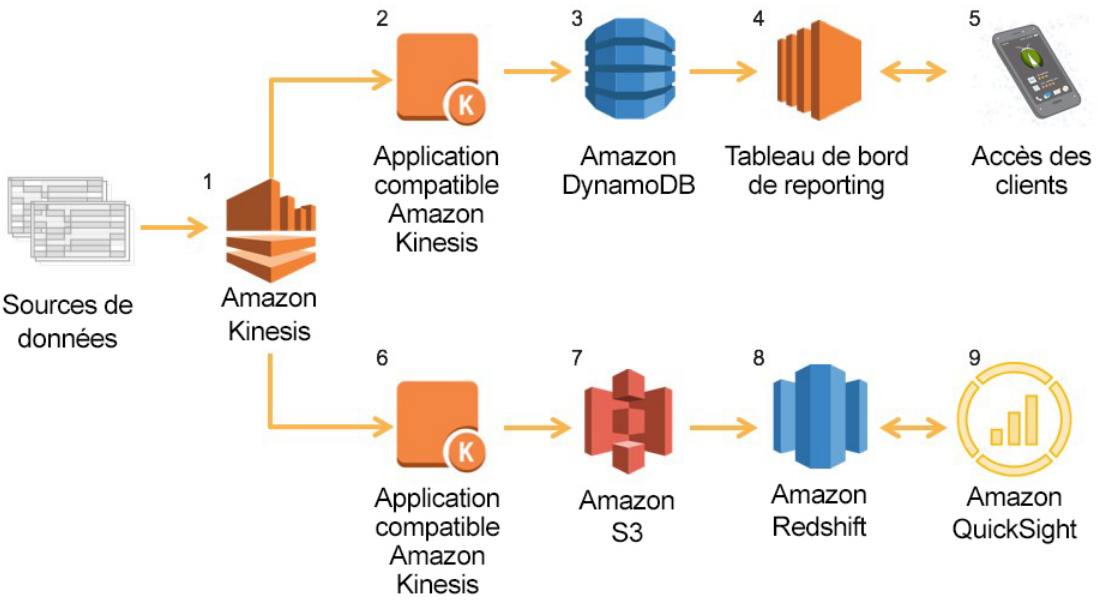
Exemple 2 : Capture et analyse des données de capteur

Un fabricant de climatiseurs international dispose de beaucoup de grands climatiseurs qu'il vend à diverses entreprises commerciales et industrielles. Non seulement ils vendent les climatiseurs mais, pour mieux se positionner par rapport à leurs concurrents, ils offrent également des services complémentaires où vous pouvez voir des tableaux de bord en temps réel dans une application mobile ou un navigateur web. Chaque unité envoie ses informations de capteur pour traitement et analyse. Ces données sont utilisées par le fabricant et ses clients. Par le biais de cette capacité, le fabricant peut visualiser l'ensemble de données et repérer les tendances.

Actuellement, avec cette capacité, ils ont quelques milliers d'unités pré-achetées. Ils prévoient de les livrer aux clients dans les deux prochains mois et espèrent que, à terme, des milliers d'unités à travers le monde utiliseront cette plate-forme. En cas de réussite, ils aimeraient étendre cette offre à leur gamme de produits grand public, avec un volume beaucoup plus important et une plus grande part de marché. La solution doit être capable de gérer des quantités massives de données et d'évoluer à mesure qu'elles se développent sans interruption. Comment devriez-vous concevoir un tel système ? D'abord, divisez-le en deux flux de travail, tous deux provenant des mêmes données :

- Les informations actuelles de l'unité A/C avec des exigences en temps quasi réel et un grand nombre de clients qui utilisent cette information.
- Toutes les informations historiques sur les unités de climatisation pour exécuter les tendances et les analyses à usage interne.

Voici l'architecture de flux de données pour résoudre ce problème de Big Data :



Capture et analyse des données de capteur

1. Le processus commence avec chaque unité de conditionnement d'air fournissant un flux de données constant à Amazon Kinesis Streams. Ceci fournit une interface souple et durable à laquelle les unités peuvent parler et qui peut être mise à l'échelle sans problème à mesure qu'augmente le nombre d'unités de climatisation vendues et mises en ligne.
2. En utilisant les outils fournis par Amazon Kinesis Streams tels que Kinesis Client Library ou SDK, une simple application est créée sur Amazon EC2 afin de lire les données dans Amazon Kinesis Streams, les analyser et déterminer si elles justifient une mise à jour du tableau de bord en temps réel. L'application recherche les changements dans le fonctionnement du système, les fluctuations de température et les erreurs éventuelles rencontrées par les unités.
3. Ce flux de données doit se produire en temps quasi réel afin que les clients et les équipes de maintenance puissent être alertés le plus rapidement possible en cas de problème avec l'unité. Les données du tableau de bord contiennent des informations agrégées sur les tendances, mais il s'agit principalement de l'état actuel, ainsi que des erreurs système. Ainsi, les données nécessaires pour remplir le tableau de bord sont plutôt petites. De plus, les accès potentiels à ces données provenant des sources suivantes seront nombreux :

- Les clients qui procèdent à des vérifications sur leur système via un appareil mobile ou un navigateur
- Les équipes de maintenance contrôlant l'état de leur flotte
- Les données, les algorithmes d'intelligence et les analyses de la plate-forme de reporting détectent les tendances qui peuvent ensuite être envoyées sous forme d'alertes, par exemple si le ventilateur du climatiseur fonctionne de manière inhabituellement longue et que la température du bâtiment ne baisse pas.

DynamoDB a été choisi pour stocker cet ensemble de données en temps quasi réel car il est à la fois hautement disponible et évolutif. Le débit de ces données peut être facilement augmenté ou réduit pour répondre aux besoins des consommateurs à mesure que la plate-forme est adoptée et que son utilisation augmente.

4. Le tableau de bord de création de rapports est une application web personnalisée qui fonctionne avec cet ensemble de données et s'exécute sur Amazon EC2. Il fournit un contenu basé sur l'état et les tendances du système, et alerte les clients et les équipes de maintenance de tout problème pouvant survenir avec l'unité.
5. Le client accède aux données depuis un appareil mobile ou un navigateur web afin d'obtenir l'état actuel du système et visualiser les tendances historiques.

Le flux de données (étapes 2 à 5) qui vient d'être décrit est conçu pour fournir des informations en temps quasi réel aux consommateurs. Il est développé et conçu pour une faible latence et peut évoluer très rapidement afin de répondre à la demande. Le flux de données (étapes 6 à 9) qui est représenté dans la partie inférieure du diagramme n'a pas des exigences de vitesse et de latence aussi strictes. Cela permet à l'architecte de concevoir une pile de solutions différente qui peut contenir de plus grandes quantités de données à un coût par octet d'information beaucoup plus faible et choisir des ressources de calcul et de stockage moins coûteuses.

6. Pour lire le flux Amazon Kinesis, il existe une application Amazon Kinesis distincte qui s'exécute en principe sur une instance EC2 plus petite, à évolution plus lente. Tandis que cette application va analyser le même ensemble de données que le flux de données supérieur, l'objectif ultime de ces données consiste à les stocker à long terme et à héberger l'ensemble de

données dans un entrepôt de données. Cet ensemble de données finit par regrouper toutes les données envoyées par les systèmes et autorise l'exécution d'un plus vaste ensemble d'analyses, sans les exigences en temps quasi réel.

7. Les données sont transformées par l'application Amazon Kinesis dans un format qui convient au stockage à long terme, au chargement dans son entrepôt de données et au stockage sur Amazon S3. Les données sur Amazon S3 servent non seulement de point d'ingestion parallèle à Amazon Redshift, mais constituent aussi un système de stockage durable qui regroupera toutes les données qui circuleront au sein de ce système. Il peut s'agir d'une source unique fiable. Cela peut être utilisé pour charger d'autres outils d'analyse en cas d'exigences supplémentaires. Amazon S3 inclut une intégration native avec Amazon Glacier si des données doivent être stockées dans un stockage à froid à long terme et à faible coût.
8. Amazon Redshift est de nouveau utilisé comme entrepôt de données pour le plus grand ensemble de données. Il peut évoluer facilement lorsque l'ensemble de données augmente en ajoutant un autre nœud au cluster.
9. Pour visualiser les analyses, l'une des nombreuses plates-formes de visualisation partenaires peut être utilisée via la connexion OBDC / JDBC à Amazon Redshift. C'est là que les rapports, les graphiques et les analyses ad hoc peuvent être effectués sur l'ensemble de données afin de trouver certaines variables et tendances pouvant entraîner une sous-performance ou une rupture des unités de climatisation.

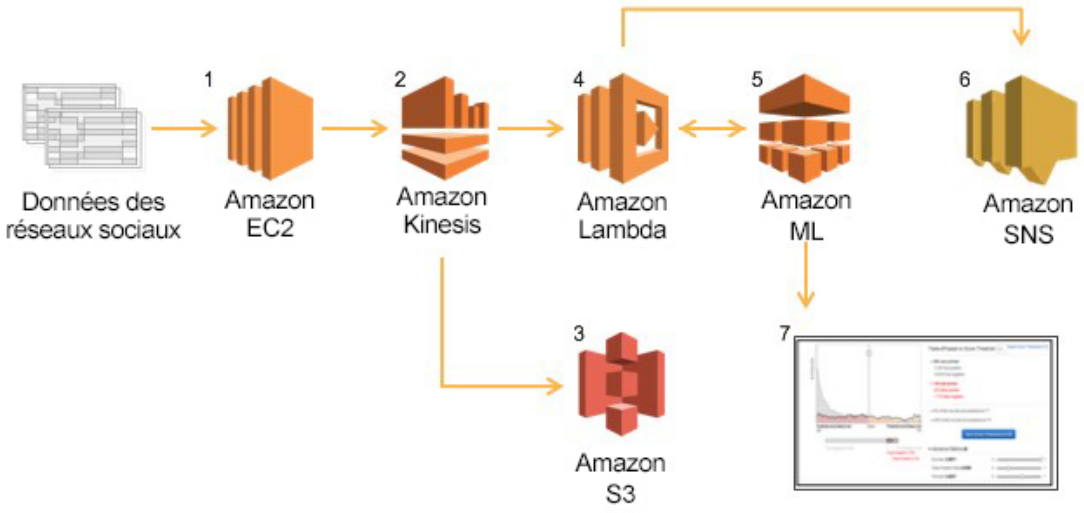
Cette architecture peut débuter modestement, puis se développer en fonction des besoins. De plus, si vous découplez les deux flux de travail les uns des autres, ils peuvent se développer à leur propre rythme en fonction des besoins sans engagement préalable, permettant au fabricant d'évaluer la réussite ou l'échec de cette nouvelle offre sans trop investir. Vous pouvez facilement imaginer d'autres ajouts, tels qu'Amazon ML, pour pouvoir prédire exactement la durée d'une unité de climatisation et d'envoyer en prévention des équipes de maintenance sur la base de ses algorithmes de prédiction afin d'offrir aux clients le service et l'expérience les plus parfaits possible. Ce niveau de service serait un facteur de différenciation de la concurrence et conduirait à une augmentation des ventes futures.

Exemple 3 : Analyse des ressentis sur les réseaux sociaux

Un grand fabricant de jouets s'est développé très rapidement et a élargi sa gamme de produits. Après chaque nouvelle sortie de jouet, l'entreprise veut comprendre dans quelle mesure les consommateurs apprécient et utilisent leurs produits. De plus, l'entreprise veut s'assurer que ses consommateurs ont une bonne expérience avec leurs produits. Au fur et à mesure que l'écosystème des jouets se développe, l'entreprise veut s'assurer que ses produits sont toujours pertinents pour ses clients et qu'ils peuvent planifier les futurs projets en fonction des commentaires émis. L'entreprise veut récupérer les éléments suivants à partir de réseaux sociaux :

- Comprendre comment les consommateurs utilisent leurs produits
- Assurer la satisfaction du client
- Planifier les futures feuilles de route

Capturer les données à partir de divers réseaux sociaux est relativement facile, mais le défi est de développer l'intelligence via la programmation. Une fois les données ingérées, l'entreprise souhaite être en mesure d'analyser et de classer les données de manière rentable et programmatique. Pour ce faire, l'architecture suivante peut être utilisée :



Analyse des ressentis sur les réseaux sociaux

1. La première chose à faire, c'est décider des réseaux sociaux qu'il faut consulter. Créez ensuite une application qui interroge ces réseaux via leurs API correspondantes et lancez-la sur Amazon EC2.
2. Ensuite, un flux Amazon Kinesis est créé, car nous pouvons avoir plusieurs sources de données : Twitter, Tumblr, et ainsi de suite. De cette façon, un nouveau flux peut être créé pour chaque nouvelle source de données ajoutée, et vous pouvez utiliser le code et l'architecture de l'application existante. En outre, dans cet exemple, un nouveau flux Amazon Kinesis est créé pour copier les données brutes dans Amazon S3.
3. Pour l'archivage, l'analyse à long terme et la référence historique, les données brutes sont stockées dans Amazon S3. Des modèles de lots Amazon ML supplémentaires peuvent être exécutés à partir de données situées dans Amazon S3 pour effectuer une analyse prédictive et suivre les tendances d'achat des consommateurs.
4. Comme indiqué dans le schéma d'architecture, Lambda est utilisé pour le traitement et la normalisation des données, ainsi que la demande de prédictions à partir d'Amazon ML. Une fois la prédiction Amazon ML renvoyée, la fonction Lambda peut prendre des mesures en fonction de la prédiction (par exemple, acheminer le message d'un réseau social vers l'équipe du service client pour révision ultérieure).
5. Amazon ML est utilisé pour effectuer des prédictions sur les données d'entrée. Par exemple, un modèle ML peut être conçu pour analyser un commentaire issu d'un réseau social afin de déterminer si le client a exprimé un sentiment négatif au sujet d'un produit. Pour obtenir des prédictions précises avec Amazon ML, commencez par les données d'entraînement et assurez-vous que vos modèles ML fonctionnent correctement. Si vous créez des modèles d'apprentissage-machine pour la première fois, consultez Didacticiel : Utilisation d'Amazon ML pour prédire les réponses à une offre marketing.[68] Comme mentionné précédemment, si plusieurs sources de données de réseaux sociaux sont utilisées, un modèle ML différent est proposé pour chacune d'entre elles afin d'assurer la précision de la prédiction.
6. Enfin, les données exploitables sont envoyées à Amazon SNS à l'aide de Lambda et envoyées aux ressources appropriées par SMS ou par e-mail pour une analyse plus approfondie.

7. Dans le cadre de l'analyse des ressentis, la création d'un modèle Amazon ML mis à jour régulièrement est impérative afin d'obtenir des résultats précis. Des métriques supplémentaires pour un modèle spécifique peuvent être affichées graphiquement via la console, telles que l'exactitude, le taux de faux positifs, la précision et le rappel. Pour plus d'informations, consultez Étape 4 : Passer en revue les performances prédictives du modèle d'apprentissage-machine et définir un seuil.[69]

En combinant Amazon Kinesis Streams, Lambda, Amazon ML et Amazon SES, nous avons créé une plateforme d'écoute sociale évolutive et facilement personnalisable. Il est important de noter que cette image ne représente pas la création d'un modèle ML. Cette action sera effectuée au moins une fois ; mais généralement, elle sera effectuée sur une base régulière pour tenir le modèle à jour. La fréquence de création d'un nouveau modèle dépend de la charge de travail et n'est réellement utilisée que pour rendre le modèle plus précis si les choses évoluent.

Conclusion

À mesure que de plus en plus de données sont générées et collectées, l'analyse des données nécessite des outils évolutifs, flexibles et performants afin de fournir des informations en temps opportun. Cependant, les entreprises font face à un écosystème grandissant de données volumineuses où de nouveaux outils émergent et « meurent » très rapidement. Par conséquent, il peut être très difficile de suivre le rythme et de choisir les bons outils.

Ce livre blanc propose une première étape pour vous aider à relever ce défi. Avec un large éventail de services gérés pour collecter, traiter et analyser le Big Data, la plate-forme AWS facilite la création, le déploiement et la mise à l'échelle des applications Big Data, ce qui vous permet de vous concentrer sur les problèmes métier.

AWS fournit de nombreuses solutions pour répondre aux besoins d'analyse du Big Data. La plupart des solutions d'architecture Big Data utilisent plusieurs outils AWS afin de créer une solution complète : cela permet de répondre aux exigences métier les plus strictes de la manière la plus rentable, performante et souple possible. Nous obtenons une architecture Big Data flexible, capable de s'adapter à votre activité sur l'infrastructure globale AWS.

Participants

Les personnes et entreprises suivantes ont participé à l'élaboration de ce document :

- Erik Swensson, directeur, architecture des solutions, Amazon Web Services
- Erick Dame, architecte de solutions, Amazon Web Services
- Shree Kenghe, architecte de solutions, Amazon Web Services

Suggestions de lecture

Les ressources suivantes peuvent vous aider à faire vos premiers pas dans l'exécution d'analyses du Big Data sur AWS :

- Visitez aws.amazon.com/big-data[70]

  Consultez la gamme complète de services Big Data, ainsi que des liens vers d'autres ressources telles que les partenaires Big Data AWS, les didacticiels, les articles et les offres AWS Marketplace sur les solutions Big Data. Contactez-nous si vous avez besoin d'aide.

- Consultez le blog AWS consacré au Big Data[71]

  Le blog propose des exemples concrets et des idées mises à jour régulièrement pour vous aider à collecter, stocker, nettoyer, traiter et visualiser le Big Data.

- Laissez-vous tenter par le test du Big Data[72]

  Explorez l'écosystème riche de produits conçus pour relever les défis liés au Big Data à l'aide d'AWS. Les tests, développés par des partenaires technologiques et consultants du réseau de partenaires AWS (APN), sont fournis gratuitement à des fins de démonstration, de formation et d'évaluation.

- Suivez un cours de formation AWS sur le Big Data[73]

  La formation Big Data sur AWS présente des solutions pour le Big Data basées sur le cloud et Amazon EMR. Vous découvrirez comment utiliser Amazon EMR afin de traiter des données grâce au vaste écosystème

d'outils Hadoop tels que Pig et Hive. Vous apprendrez également à créer des environnements Big Data, à travailler avec DynamoDB et Amazon Redshift, à comprendre les avantages d'Amazon Kinesis Streams et à tirer parti des meilleures pratiques afin de concevoir des environnements Big Data sécurisés et économiques.

- Affichez l'étude de cas client sur le Big Data[74]

  Apprenez de l'expérience d'autres clients qui ont conçu des plates-formes Big Data puissantes et performantes sur le cloud AWS.

Révisions du document

Janvier 2016	Ajout d'informations sur Amazon Machine Learning, AWS Lambda, Amazon Elasticsearch Service ; mise à jour générale
Décembre 2014	Première publication

Remarques

1 http://aws.amazon.com/about-aws/globalinfrastructure/

2 http://aws.amazon.com/s3/

3 http://aws.amazon.com/datapipeline/

4 https://aws.amazon.com/iot/

5 https://aws.amazon.com/importexport/

6 http://aws.amazon.com/kinesis/firehose

7 https://aws.amazon.com/directconnect/

8 https://aws.amazon.com/mobile/

9 http://aws.amazon.com/solutions/case-studies/big-data/

10 https://aws.amazon.com/kinesis/streams

11 http://docs.aws.amazon.com/kinesis/latest/APIReference/Welcome.html

12 http://docs.aws.amazon.com/aws-sdk-php/v2/guide/service-kinesis.html

13 http://aws.amazon.com/kinesis/pricing/

14 http://aws.amazon.com/tools/

15 http://docs.aws.amazon.com/kinesis/latest/dev/developing-producers-with-kpl.html

16 http://docs.aws.amazon.com/kinesis/latest/dev/writing-with-agents.html

17 https://github.com/awslabs/amazon-kinesis-client

18 https://github.com/awslabs/kinesis-storm-spout

19 https://aws.amazon.com/lambda/

20 http://docs.aws.amazon.com/lambda/latest/dg/intro-core-components.html

21 https://aws.amazon.com/amazon-linux-ami/

22 http://docs.aws.amazon.com/lambda/latest/dg/nodejs-create-deployment-pkg.html

23 http://docs.aws.amazon.com/lambda/latest/dg/lambda-python-how-to-create-deployment-package.html

24 http://docs.aws.amazon.com/lambda/latest/dg/lambda-java-how-to-create-deployment-package.html

25 http://aws.amazon.com/elasticmapreduce/

26 https://media.amazonwebservices.com/AWS_Amazon_EMR_Best_Practices. pdf

27 http://aws.amazon.com/elasticmapreduce/pricing/

28 http://aws.amazon.com/ec2/instance-types/

29 http://aws.amazon.com/elasticmapreduce/mapr/

30 http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-manage-resize.html

31 http://docs.aws.amazon.com/ElasticMapReduce/latest/API/Welcome.html

32 http://docs.aws.amazon.com/ElasticMapReduce/latest/ReleaseGuide/emr-hive.html

33 http://docs.aws.amazon.com/ElasticMapReduce/latest/ReleaseGuide/emr-pig.html

34 http://blogs.aws.amazon.com/bigdata/post/Tx15AY5C50K70RV/Installing-Apache-Spark-on-an-Amazon-EMR-Cluster

35 http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-hbase.html

36 http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-impala.html

37 http://aws.amazon.com/elasticmapreduce/hunk/

38 http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/UsingEMR_s3distcp.html

39 https://aws.amazon.com/machine-learning/

40 https://aws.amazon.com/machine-learning/pricing/

41 http://docs.aws.amazon.com/machine-learning/latest/dg/suggested-recipes.html

42 http://docs.aws.amazon.com/machine-learning/latest/APIReference/Welcome.html

43 https://aws.amazon.com/dynamodb

44 http://aws.amazon.com/free/

45 http://aws.amazon.com/dynamodb/pricing/

46 Nombre standard de millisecondes à un chiffre pour les temps de réponse moyens côté serveur

47 http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html

48 DynamoDB vous permet d'augmenter à 100 % votre niveau de débit alloué à l'aide d'une opération d'appel de l'API UpdateTable unique. Pour augmenter votre débit de plus de 100 %, appelez à nouveau UpdateTable.

49 Vous pouvez augmenter votre débit provisionné aussi souvent que vous le souhaitez. Cependant, la limite est fixée à deux diminutions par jour.

50 https://aws.amazon.com/redshift/

51 http://aws.amazon.com/s3/pricing/

52 http://aws.amazon.com/redshift/pricing/

53 http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html

54 http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html

55 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes

56 http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html

57 http://aws.amazon.com/redshift/partners/

58 https://aws.amazon.com/elasticsearch-service/

59 https://aws.amazon.com/ec2/pricing/

60 https://aws.amazon.com/ebs/details/

61 https://aws.amazon.com/elasticsearch-service/pricing/

62 https://aws.amazon.com/elasticsearch-service/faqs/

63 https://aws.amazon.com/quicksight

64 https://aws.amazon.com/ec2/

65 https://aws.amazon.com/marketplace

66 http://aws.amazon.com/autoscaling/

67 http://aws.amazon.com/ec2/spot/

68 http://docs.aws.amazon.com/machine-learning/latest/dg/tutorial.html

69 http://docs.aws.amazon.com/machine-learning/latest/dg/step-4-review-the-ml-model-predictive-performance-and-set-a-cut-off.html

70 http://aws.amazon.com/big-data

71 http://blogs.aws.amazon.com/bigdata/

72 https://aws.amazon.com/testdrive/bigdata/

73 http://aws.amazon.com/training/course-descriptions/bigdata/

74 http://aws.amazon.com/solutions/case-studies/big-data/