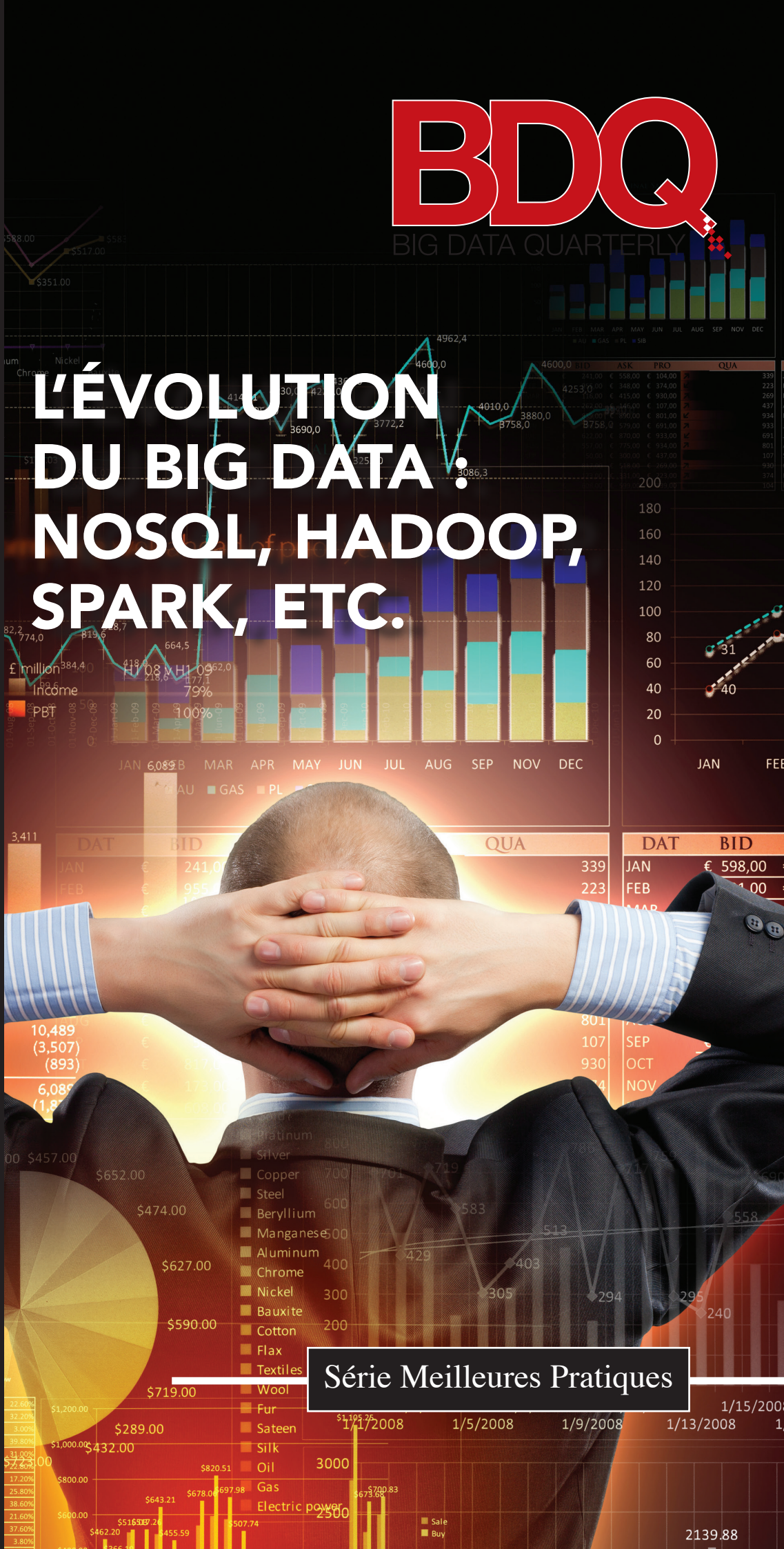
BDQ
BIG DATA QUARTERLY

L'ÉVOLUTION DU BIG DATA : NOSQL, HADOOP, SPARK, ETC.



Série Meilleures Pratiques

NOSQL, HADOOP ET SPARK ENRICHISSENT LES INITIATIVES BIG DATA

Série Meilleures Pratiques

Tandis que l'écosystème Big Data poursuit son expansion, de nouvelles technologies viennent répondre aux exigences de gestion, de traitement, d'analyse et de stockage des données. Elles permettent ainsi aux entreprises de tirer parti des nombreuses sources d'informations qui affluent en permanence.

Bases de données NoSQL, projets open source tels que Spark, Hive, Drill, Kafka, Arrow et Storm ou produits commerciaux locaux et cloud, l'avenir du Big Data se façonne autour de nouvelles approches novatrices sur l'ensemble du cycle de gestion des données. Les domaines les plus pressants sont le traitement des données en temps réel, l'analyse interactive, l'intégration et la gouvernance des données, ainsi que la sécurité.

La révolution Big Data gomme les strictes délimitations entre les types de données, tous étant considérés de la même manière. Il s'agit dès lors d'une opportunité unique d'utiliser le potentiel de toutes ces données pour fournir des informations aux décideurs.

À NOUVEAUX PROBLÈMES, NOUVELLES SOLUTIONS

En répondant aux besoins de stockage et de gestion des données peu adaptées aux lignes et aux colonnes, les technologies NoSQL occupent le devant de la scène : elles représentent un monde plus vaste qui se connecte à Internet dans son ensemble, à l'Internet des objets et aux clouds.

Les bases de données NoSQL peuvent fonctionner avec des équipements standard, prendre en charge les données non structurées, non relationnelles qui affluent dans les entreprises et maîtriser la prolifération de nouvelles sources. Elles sont disponibles dans un grand nombre de structures qui ouvrent des possibilités avec de nouveaux types de sources de données, ce qui permet d'utiliser les connaissances institutionnelles enfermées dans les PC et les silos de chaque département. Par exemple, la technologie émergente de chaînes de blocs est conçue pour stocker des données, de type grand livre, selon une approche hautement distribuée sur tout l'Internet.

Les quatre principaux types de bases de données dans la catégorie NoSQL sont les suivants : les magasins clé-valeur, qui permettent de stocker les données sans schéma, avec clé et valeurs réelles, les bases de données organisées en familles de colonnes, qui stockent les données dans des colonnes, les bases de données graphe, dont les structures de stockage utilisent nœuds, arêtes et propriétés, et les bases de données orientées documents, qui permettent le simple stockage d'agrégations de documents.

Et, malgré l'implicite du nom, les fournisseurs de bases de données NoSQL répondent de plus en plus aux besoins de leurs clients qui souhaitent utiliser SQL comme langage de requête principal. Le paysage NoSQL est encore relativement nouveau, mais il évolue rapidement et propose de nouvelles fonctionnalités permettant d'améliorer accessibilité, interopérabilité, sécurité et gouvernance et montre les prémisses de son futur potentiel pour l'entreprise.

HADOOP, L'ÉCOSYSTÈME QUI MONTE

Hadoop, figure centrale des technologies Big Data, a fêté cette année son 10ème anniversaire : son expansion va bien au-delà d'une plateforme de stockage et de traitement en batch de grands volumes de données sur des serveurs standard. Le framework Apache Hadoop, composé d'Hadoop Common, du système HDFS (Hadoop Distributed File System), d'Hadoop YARN et d'Hadoop MapReduce, est un composant clé de la plupart des projets Big Data et de la création de lacs de données. En outre, on compte aujourd'hui plus d'une centaine de projets open source dans toute la pile Hadoop, ce qui renforce la sécurité, la flexibilité et l'accessibilité d'Hadoop grâce à des fonctionnalités d'entreprise supplémentaires.

Avant Hadoop, la capture et l'analyse de données de tous types impliquaient l'utilisation d'outils propriétaires, autrement dit un engagement coûteux et gourmand en ressources. Hadoop, tout comme l'écosystème open source dans lequel il s'inscrit, offre une option d'analyse Big Data plus rentable, à la portée d'un plus grand nombre d'utilisateurs. Pour les gestionnaires de données, le défi consistera toutefois à acquérir les compétences nécessaires pour bâtir ces environnements open source.

L'ÉMERGENCE DE SPARK

Spark, framework d'analyse de données, est l'une des technologies les plus récentes de l'écosystème Hadoop. Il s'agit d'un framework ancré dans l'univers Apache Hadoop. Certains définissent même Spark comme un "système d'exploitation pour l'analyse," ce qui laisse entendre qu'il peut constituer la base d'une vaste palette de fonctions et d'applications d'analyse de données pouvant s'y superposer.

Hadoop et Spark appartiennent à la vague open source qui continue d'offrir des fonctionnalités utiles aux entreprises de toutes tailles, autrement dit un moyen rentable et hautement évolutif de stocker et de déployer des ensembles de données volumineux et variés.

Certains adeptes de Spark soulignent que le framework open source reprend les rênes là où Hadoop s'essouffle. Pour les débutants, Spark prend naturellement en charge les exigences en temps réel et offre un traitement plus rapide et un moteur d'analyse plus robuste, grâce à des fonctionnalités de traitement parallèle in-memory. Entre autres avantages, le framework inclut des bibliothèques résidentes qui permettent un développement plus rapide des applications qui ciblent et traitent des ensembles de données structurés, semi-structurés et non structurés. Le framework Spark permet de gérer de nombreuses tâches, streaming, ETL classique, lacs de données et applications de streaming en temps réel les plus récentes.

Les principaux composants de Spark sont SparkSQL, langage d'accès aux sources de données par requêtes SQL, Spark MLib, qui fournit des fonctions d'analyse prédictive, SparkR, qui permet l'accès aux données Spark, et GraphX, une API conçue pour les calculs de graphe avec bibliothèque intégrée d'algorithmes communs.

LE COMPLÉMENT D'HADOOP

Spark n'a pas pour vocation de remplacer Hadoop mais de le compléter. Spark repose sur le système HDFS (Hadoop Distributed File System) et s'adapte bien aux environnements qui intègrent déjà des compétences ou des outils Hadoop. La principale différence réside dans le fait que Spark est uniquement un moteur d'analyse, tandis qu'Hadoop est un système de gestion et de stockage des données. Spark n'a toutefois pas besoin d'Hadoop pour la gestion et le stockage des données. Les gestionnaires de données peuvent ainsi conserver leurs configurations actuelles, notamment Hadoop ou autres environnements de données, sans migration.

Les entreprises se trouvent souvent aux prises avec des silos et des frameworks de données. Même les projets Hadoop ont tendance à finir dans leurs propres silos. Extrêmement polyvalent, le framework Spark peut être déployé de nombreuses manières, pour de nombreuses fonctions d'analyse. Les fonctionnalités Spark peuvent également être intégrées aux applications existantes, par exemple celles qui s'appuient sur Java. Une autre fonctionnalité clé de Spark est la prise en charge des ensembles de données distribués résilients, ou RDD (Resilient Distributed Datasets), ce qui permet de gérer et stocker les objets partout dans l'infrastructure, aussi bien sur disque qu'en mémoire. Cela garantit également une très haute disponibilité.

L'analyse de données a longtemps été réservée à des équipes spécialisées d'analystes dont la plupart avaient tendance à travailler à distance des autres équipes de l'entreprise. Spark peut pourtant constituer l'étape la plus importante vers le saint graal de l'analyse : un accès omniprésent dans l'entreprise, par tous les employés, tous niveaux confondus. Spark est accessible à de nombreux utilisateurs, en particulier ceux qui sont concernés par la gestion et l'analyse des données et qui en font le fer de lance de la stratégie de l'entreprise. Spark est également un framework idéal pour les data scientists et les analystes, car de nombreuses fonctionnalités complexes sont déjà intégrées dans son "système d'exploitation pour l'analyse." Le framework prend en charge les langages de programmation les plus répandus et les applications peuvent être rapidement créées à partir des fonctionnalités Spark de base, notamment apprentissage machine, traitement en temps réel et traitement graphe.

Spark est également une plateforme technologique en maturation, autrement dit certains aspects du framework restent à découvrir par les gestionnaires de données et à affiner par la communauté open source et les fournisseurs. Certains utilisateurs, par exemple, signalent un manque de convivialité dans certains aspects de la solution qui nécessitent quelques détours. Il s'agit d'un phénomène normal pour toutes les technologies émergentes. Mais au fur et à mesure que les développeurs et les fournisseurs de Spark en améliorent les fonctionnalités, il occupera une place plus centrale dans l'entreprise.

LA PRESSION CONCURRENTIELLE EN MATIÈRE D'ANALYSE

Aujourd'hui, les entreprises se doivent d'être compétitives en matière d'analyse, dans une économie mondiale hyper concurrentielle. Nombre d'entre elles créent des magasins Big Data, mais manquent de moyens efficaces pour convertir ces données en informations qu'elles peuvent exploiter, à l'endroit et au moment où elles en ont besoin.

"En associant le traitement Big Data d'Hadoop et plus largement son écosystème, notamment l'analyse accélérée par Spark et la richesse des bases de données NoSQL, on obtient un potentiel tourné vers les problématiques de l'entreprise et ses opportunités."

— Joe McKendrick

cloudera

L'ingénierie de données avec Apache Hadoop

QU'EST-CE QUE L'INGÉNIERIE DE DONNÉES ?

L'ingénierie de données est le processus de création d'une infrastructure d'analyse ou de produits internes afin de permettre la collecte, le nettoyage, le stockage et le traitement (en batch ou en temps réel) des données, dans le but de répondre aux questions de l'entreprise. Cette tâche est habituellement menée par un data scientist, un statisticien ou une personne exerçant une fonction connexe.

Quelques exemples d'ingénierie de données:

- Création de pipelines cumulant les données de sources multiples
- Productionisation, à grande échelle, de modèles d'apprentissage machine
- Création d'outils prédéfinis utilisés pour le processus d'interrogation (par ex. UDF)

Les ingénieurs de données s'appuient sur l'écosystème Apache Hadoop, notamment ses composants comme Apache Spark, Apache Kafka et Apache Flume, pour constituer la base de cette infrastructure. Quels que soient les cas d'usage et les composants concernés, cette infrastructure doit offrir la conformité requise en termes de sécurité, de lignage des données et de gestion des métadonnées.

Cet e-book sur l'ingénierie de données présente les concepts techniques liés à la création et à la maintenance d'une infrastructure sur un hub de données d'entreprise Hadoop. Nous présentons ici certains de ces concepts de manière approfondie, comme vous le verrez dans les différentes sections de l'e-book.

SCHÉMAS D'ARCHITECTURE

Pour un traitement des données en temps quasi réel

Pour pouvoir réussir un déploiement en production, il est impératif d'évaluer le schéma de l'architecture de streaming qui sera le mieux adapté à votre cas d'usage. Dans cet e-book, nous présentons quatre grands schémas de streaming et leur mode de mise en œuvre sur le plan architectural.

DÉTECTION DES FRAUDES

Pour concevoir une architecture efficace pour la détection des fraudes, il suffit de s'inspirer du cerveau humain (et de s'entourer de Spark Streaming et de Kafka). La détection des fraudes s'intéresse essentiellement à la détection des anomalies et aux réactions à ces anomalies.

Une architecture efficace de détection des fraudes nécessite que trois sous-systèmes travaillent en cohésion afin de détecter les anomalies dans les flux d'événements : opérationnalisation des systèmes en temps réel, de traitement des flux et de traitement hors ligne.

SESSIONISATION EN TEMPS QUASI RÉEL

Avec Spark et Hadoop

Dans cette section, nous décrivons les fonctionnalités courantes et avancées de Spark Streaming via le cas d'usage de sessionisation en temps quasi réel d'événements de sites Web, puis de chargement de statistiques sur cette activité dans Apache HBase et enfin d'ajout de graphes dans votre outil décisionnel privilégié pour l'analyse.

APACHE KAFKA POUR LES DÉBUTANTS

Apache Kafka fait beaucoup parler d'elle ces temps-ci. Si LinkedIn, lieu de création de Kafka, est l'utilisateur le plus connu, il existe aujourd'hui de nombreuses sociétés qui déploient cette technologie avec succès. Maintenant que l'information circule, tout le monde veut savoir : que fait-elle ? Pourquoi tout le monde veut-il l'utiliser ? En quoi est-elle mieux que les solutions existantes ? Les avantages justifient-ils de remplacer les systèmes et l'infrastructure existants ? Dans cette section, nous allons entre autres répondre à ces questions.

CONVERSION DE MAPREDUCE VERS SPARK

À l'origine, Hadoop a été conçu pour le traitement des journaux à grande échelle et les opérations ETL orientées batch.

Au vu de son expansion, des architectures alternatives comme Impala et Spark ont été créées pour répondre aux nouveaux besoins. Spark s'est tellement développé qu'il est prêt à succéder à MapReduce en tant que paradigme de calcul Hadoop universel.

Cette section de l'e-book explique justement qu'il est parfaitement possible de reprendre dans Spark des calculs de type MapReduce.

ADAPTER LES T CHES APACHE SPARK

Lorsqu'on écrit du code Apache Spark et qu'on examine les API publiques, on entend souvent des termes tels que "transformation, action et RDD."

De même, un nouveau vocabulaire fait son apparition lorsque les choses ne se déroulent pas comme prévu ou que l'application est anormalement lente : on entend des mots comme job, phase et tâche. À nouveau, il est essentiel à ce stade de comprendre Spark pour pouvoir exécuter de bons programmes Spark, autrement dit, des programmes RAPIDES!

Cette section présente les concepts de base relatifs aux modalités d'exécution des programmes Spark sur un cluster et propose des recommandations pratiques sur la capacité du modèle d'exécution Spark à écrire des programmes efficaces.

RÉSUMÉ

Si les stratégies modernes d'ingénierie de données vous intéressent, téléchargez cet e-book et n'hésitez pas à contacter Cloudera pour toute question.

À PROPOS DE CLOUDERA

Cloudera fournit la plateforme de gestion de données et d'analyse la plus rapide, la plus simple et la plus sûre au monde ; elle repose sur Apache Hadoop et les technologies open source les plus récentes.

CLOUDERA
www.cloudera.com

DATASTAX
The power behind the moment.

NoSQL, la simplicité

Pour toutes les entreprises à la recherche d'un avantage concurrentiel dans l'économie numérique actuelle, il ne s'agit pas de savoir si mais plutôt comment elles doivent faire évoluer leur activité pour pouvoir survivre. Aujourd'hui, les clients ne vous accordent pas de deuxième chance lorsqu'il s'agit d'expérience numérique. Si vous ne répondez pas à leurs attentes, ils s'en vont et ne reviennent pas.

Il est fondamental de proposer une expérience client extraordinaire.
Pour atteindre cet objectif avec vos applications cloud qui créent de la valeur en temps réel pour vos clients, vous devez utiliser une plateforme de base de données distribuée, hautement réactive et intelligente.

POURQUOI PRÉFÉRER NOSQL AUX SGBDR ?

Selon The Forrester Wave™ : Le rapport Big Data NoSQL, Q3 2016, une étude menée auprès de plus de 2000 décideurs du secteur des technologies de données et de l'analyse a révélé que plus de 60% des entreprises avaient déjà mis en œuvre, prévoyaient de mettre en œuvre ou développaient/amélioraient leurs solutions NoSQL au cours des 12 prochains mois. "NoSQL est incontournable, il est devenu impératif de prendre en charge les applications nouvelle génération," explique Noel Yuhanna, analyste principal chez Forrester Research.

DE QUOI ONT BESOIN LES APPLICATIONS CLOUD?

Une application cloud est une application comportant de nombreux terminaux, notamment navigateurs, appareils mobiles et/ou machines, qui sont géographiquement dispersés, avec une grande quantité de transactions, toujours disponibles et instantanément réactifs, quel que soit le nombre d'utilisateurs ou de machines qui utilisent cette application. Les clients attendent que leur soient fournies des informations personnalisées et la possibilité d'agir quand et où ils le décident.

Chaque application cloud est certes unique, mais la base de données doit respecter plusieurs exigences fondamentales pour la compétitivité de l'entreprise sur le marché.

- Distribuée pour garantir 100 % de temps d'activité
- Réactive pour réduire la latence et permettre l'augmentation ou la réduction des capacités au gré des besoins de l'entreprise
- Intelligente pour s'adapter aux différents types de modèles de données et de charges de travail, tout en gérant les problèmes qui apparaissent

Et tout ceci doit s'effectuer sur une plateforme unique, sécurisée et adaptée à l'entreprise, avec une gestion intelligente et une surveillance simplifiée.

Les technologies SGBDR ne parvenant pas à répondre à ces attentes, les entreprises se tournent vers les bases de données NoSQL.

MAÎTRISER LA VÉRITABLE PUISSANCE DE NOSQL AVEC DATASTAX ENTERPRISE (DSE)

Seule la solution DataStax Enterprise vous permet de maîtriser la véritable puissance d'une base de données multi-modèle NoSQL et d'offrir une expérience client haute disponibilité, des performances accélérées et de puissantes recommandations contextuelles. Tout ceci en parallèle avec des méthodes DevOps agiles et des outils de gestion permettant à l'équipe Ops de surveiller et d'exploiter plus facilement vos environnements de production.

Il est possible d'atteindre une haute disponibilité via les procédures de basculement des SGBDR existants, mais la disponibilité réellement continue d'une application cloud nécessite une architecture conçue pour ne jamais tolérer aucun point de défaillance. DataStax Enterprise s'appuie sur l'architecture de base d'Apache Cassandra™, dont la conception totalement distribuée permet à chaque nœud d'un cluster de fonctionner indépendamment des opérations de la base de données. On atteint ainsi 100% de temps d'activité. Non pas 99,99%, mais 100%.

DataStax Enterprise – résoudre les défis les plus complexes de l'expérience client numérique

"DataStax assure notre avenir," explique Christos Kalantzis, responsable Ingénierie de la base de données cloud, Netflix. Il se rappelle du cauchemar de la panne Netflix qui a duré plus de 48 heures suite à une défaillance de sa base de données Oracle. "Nous ne pouvions pas risquer que cela se reproduise. Oracle n'a pas été créé pour le cloud et ne fonctionne pas dans le cloud au niveau que nous attendons."

Jeunes entreprises numériques ou entreprises vieilles d'une centaine d'années, toutes commencent déjà à créer de la valeur grâce aux données connectées non collectées. Cette capacité à comprendre les interactions d'un client sur différents silos et canaux numériques et à personnaliser et offrir des expériences contextuelles très pertinentes se généralise rapidement.

Les clients ne donneront pas de deuxième chance à votre entreprise si elle ne fournit pas une excellente expérience client numérique. Avec DataStax, tout cela ira de soi. Créez et gérez des applications cloud qui vont bien au-delà des attentes de vos clients en termes de rapidité, d'accès et d'expérience pertinente, en maîtrisant la véritable puissance d'une base de données multi-modèle NoSQL pour créer de la valeur en temps réel.

DATASTAX www.datastax.com

CASK

Vos solutions Big Data à l'épreuve du futur

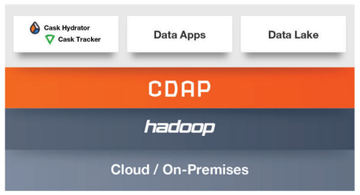
Le Big Data fête cette année son 10ème anniversaire et beaucoup de choses ont changé depuis l'invention d'Apache Hadoop. Au tout début, de nouveaux projets se sont rapidement développés dans l'écosystème et ont couvert les principales zones fonctionnelles nécessaires pour traiter avec une incomparable efficacité le volume, la vélocité et la variété des données. Plus tard, d'autres projets ont été développés afin de compléter les principales fonctionnalités requises pour les divers cas d'usage et les possibilités d'accès des développeurs ; plus récemment aussi avec la croissance d'Apache Spark.

DÉVELOPPEURS ET ENTREPRISES NE VEULENT PAS CRÉER DES INTÉGRATIONS MAIS DES SOLUTIONS

Lorsque l'adoption d'Hadoop a pris de l'ampleur, les éditeurs ont commencé à fournir des outils d'intégration spécifiques, pour l'ETL, l'intégration de données, la sécurité, la gouvernance, etc., afin de simplifier le développement des solutions Big Data. Dans la plupart des cas actuels, toutefois, la création d'une application Big Data nécessite l'intégration d'un nombre important et croissant de ces outils spécifiques, ce qui oblige les développeurs Big Data à porter leur attention non plus sur la logique des applications et des données (IP et valeur) mais sur la logique d'intégration. Par conséquent, développeurs et entreprises finissent par consacrer un temps précieux aux tâches d'infrastructure et d'intégration tout en gérant une multitude de fournisseurs et la complexité de leurs outils, plutôt que de mettre leur énergie au service des applications et des informations exploitables.

PASSER DU PROTOTYPE À LA PRODUCTION: LE DÉFI

Tout comme les applications classiques à trois niveaux, les applications de données distribuées n'apparaissent pas comme par magie ; elles sont créées, testées, optimisées et planifiées pour passer du stade de prototype à l'environnement de production. Mais il est tellement différent de développer du code sur un ordinateur portable ou une station de travail et dans un environnement de production multi-nœud hautement distribué qu'il est souvent nécessaire de beaucoup recoder et reconfigurer. Cela augmente considérablement le délai de mise en œuvre. Certains des défis ne concernent peut-être même pas les compétences des développeurs initiaux, en particulier lorsque des exigences spécifiques de production (packages, versions, etc.) entrent en jeu. Au fil de l'évolution des solutions Big Data, les entreprises doivent surmonter cet obstacle, opérationnaliser efficacement les applications de données du prototype à la production et offrir un environnement en libre-service aux utilisateurs métier avant qu'ils ne commencent à extraire la valeur de leurs données.



Cask Data Application Platform (CDAP): une couche d'intégration sur Hadoop

LA SOLUTION EST L'INTÉGRATION UNIFIÉE POUR BIG DATA AVEC CDAP

La plateforme CDAP (Cask Data Application Platform) a été conçue pour résoudre les problèmes qui apparaissent entre la phase de prototype et la phase de production, localement ou dans le cloud, sur un ordinateur portable ou sur un serveur de 1000 nœuds. Il s'agit d'une plateforme d'intégration réellement unifiée qui combine les fonctionnalités de gestion des applications et d'intégration des données avec un environnement en libre-service sans code et une gouvernance adaptée à l'entreprise. } Elle garantit la cohérence des données et des processus entre les applications et les technologies d'infrastructure sous-jacentes sur de multiples environnements et entre les différentes parties du système informatique.

Pour que les applications Hadoop créées sur CDAP résistent à l'épreuve du futur, cette plateforme 100% open source et extensible agit comme une couche d'abstraction et sépare la logique d'intégration de la logique des applications et des données. Les futurs changements apportés aux données ou à la logique métier deviennent bien plus simples et les opérations en cours pour les solutions Big Data sont rationalisées.

La plateforme CDAP inclut également Cask Hydrator, une extension glisser-déposer qui permet aux utilisateurs de créer rapidement et facilement des pipelines de données sans code pour ingérer et transformer des données dans Hadoop, ce qui accélère le processus de création des lacs de données de l'entreprise.

Une deuxième extension CDAP appelée Cask Tracker permet de découvrir les données, de vérifier l'accès aux données et d'en suivre le lignage, toutes ces tâches étant nécessaires pour la gouvernance de l'entreprise.

La plateforme CDAP permet le développement et le déploiement rapides d'applications Big Data, notamment la possibilité de les migrer aisément vers la phase de production, tout en répondant aux strictes conditions de gouvernance. Avec CDAP, vous n'aurez besoin de rien d'autre pour que votre environnement Hadoop emprunte le chemin de l'efficacité et de la sécurité vers la production, ainsi que pour aisément évoluer avec l'écosystème tout en assurant le futur des solutions Big Data que vous créez aujourd'hui. Pour en savoir plus sur ces produits, consultez le site Web de Cask et suivez l'actualité des produits et de la société sur Twitter @caskdata.

CASK
www.cask.com

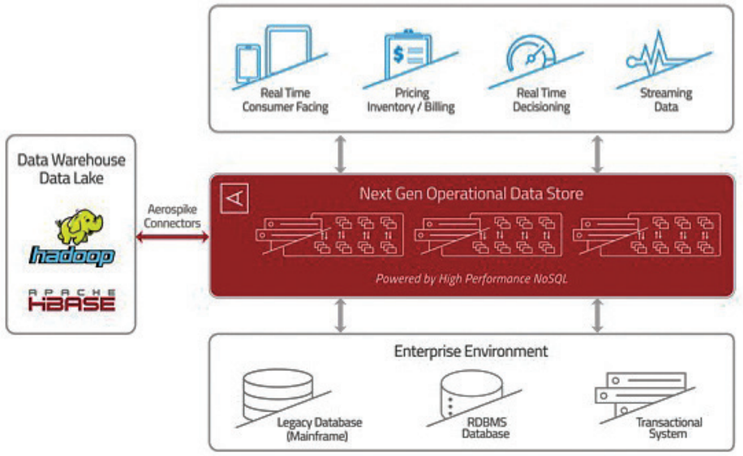
AEROSPIKE

Une base de données NoSQL haute performance pour l'entreprise

Dans un monde strictement relationnel, la base de données était souvent un élément freinant l'innovation. Aujourd'hui, la mise en œuvre de la base de données adéquate est devenue un élément favorisant l'innovation.

La base de données NoSQL haute performance pour l'entreprise Aerospike est un magasin clé-valeur avec stockage flash optimisé. Notre technologie open source inclut des outils et des packages conçus pour aider les développeurs à créer des applications nouvelle génération sans se soucier de la programmation de bas niveau. Aerospike est la première base de données derrière certains des plus grands noms des secteurs de l'AdTech, des télécommunications et des services financiers. Les problématiques que nous gérons peuvent être des charges de travail stratégiques nécessitant une performance prévisible à grande échelle, un temps d'activité élevé et haute disponibilité, le tout avec un coût total de possession minimal.

Grâce à Aerospike, connectez-vous à Hadoop et Spark pour simplifier l'intégration de vos systèmes. Hadoop est excellent pour analyser de vastes quantités de données, mais il ne convient pas pour une latence de quelques millisecondes. La majorité des utilisateurs d'Aerospike, aujourd'hui, intègrent leurs clusters Hadoop avec Aerospike pour offrir le meilleur des deux univers.



- Simplicity
- Maintainability
- Durability
- Consistency
- Scalability
- Low Cost
- Reduced Data Lag

AEROSPIKE AFFICHE DES PERFORMANCES RECORD AVEC RAM OU SSD

Les applications actuelles ont besoin d'atteindre des niveaux de performance inégalés. Aerospike Database exploite au mieux le matériel, que vous choisissiez la RAM ou les SSD pour le stockage des données. Des tests d'évaluation menés par Intel révèlent une incroyable performance quelle que soit l'option utilisée.

2,5 Mt/s avec la RAM sur un serveur unique

Pour ceux qui attendent une vitesse optimale, la RAM fournit le meilleur débit avec la latence la plus faible.

1 Mt/s avec les SSD sur un serveur unique

Pour ceux qui souhaitent stocker de gros volumes de données par serveur, les SSD offrent quasiment la même latence à un prix bien moindre.

AEROSPIKE AFFICHE D'EXCELLENTES PERFORMANCES SUR GOOGLE COMPUTE ENGINE

Aerospike a été testé en haut débit sur GCE (Google Compute Engine), avec un cluster de 10 nœuds, avec RAM et SSD. La taille de chaque enregistrement était de 1500 octets.

- La performance en écriture à 100 % était comparable pour la RAM et les SSD
- La performance en lecture était nettement plus élevée avec la RAM mais les coûts mensuels de stockage étaient de 0,56 USD/Go pour les SSD contre 8,46 USD pour la RAM. Pour de nombreuses applications nécessitant de grandes capacités de stockage, les SSD constituent le choix privilégié en termes d'économies

Tous les résultats des tests et la méthodologie sont disponibles ici:

"Nous atteignons 2,5 millions d'impressions par seconde au maximum de l'activité, alors que nous pourrions aller bien au-delà, et nous comptabilisons 90 milliards d'impressions par jour ; ce temps d'activité est attendu 24h/24 et 7j/7 et Aerospike nous le garantit à 100 %."

— Geir Magnusson, directeur technique, AppNexus

AEROSPIKE
www.aerospike.com