LeMagIT
L'informatique pour et par les pros
TechTarget
E-handbook
Contenu PRO+

Hadoop à l'ère du multicloud et de l'analytique en temps réel

Dans ce guide

Introduction.

Le lac de données a connu une forte vague de popularité. Il a permis de développer les cas d'usage Big Data et analytique, puis de machine learning. Ces installations sur site ont petit à petit migré dans le cloud. Même si les frameworks comme Hadoop n'ont pas perdu leurs qualités, les éditeurs et les entreprises ont trouvé de nouveaux moyens de traiter leurs jeux de données, par exemple avec Apache Spark, et de les stocker, avec Amazon S3. Dès lors, la différence entre Data Lake et Data Warehouse est de plus en plus ténue.

Ce passage dans le cloud n'a pas réussi à Cloudera, HortonWorks, et MapR. Le premier a racheté le deuxième, et le troisième appartient aujourd'hui à HPE.

Cela ne veut pas pour autant dire qu'Hadoop est mort. Ces éditeurs prennent différentes voies. HPE défend une approche propriétaire. Cloudera voit sa rédemption dans le mutlicloud et dans l'agrégation de technologies, à l'instar de Microsoft, Google, et AWS.

Dans ce guide

À travers ce guide, comprenez les bénéfices d'une architecture Hadoop. Puis découvrez les critères qui vous aideront à bien choisir la distribution répondant à vos besoins en matière de Big Data et d'analytique. Enfin, restez informé sur les perspectives de ce marché en pleine transformation.

Dans ce guide

Cloudera ouvre les voies du multicloud à ses clients

Craig Stedman et Gaétan Raoul, journalistes

Cloudera a lancé à la fin du mois de septembre sa plateforme Big Data combinant ses technologies et celles d'Hortonworks dans le cloud AWS. Le support du multicloud est d'ores et déjà annoncé.

Cloudera, le pionner du Big Data et d'Hadoop a officiellement lancé la première itération de sa nouvelle Cloudera Data Platform (CDP). La promesse du support du multicloud lui donnera un attrait renouvelé auprès des utilisateurs qui se sont tournés vers AWS et d'autres leaders du cloud pour gérer leurs systèmes Big Data.

Précisons que le multicloud n'est pas encore une réalité chez Cloudera. Pour l'instant, trois services dédiés au data warehouse, au machine learning et l'analytique sont disponibles uniquement sur AWS. L'éditeur prévoit de connecter ses outils à Microsoft Azure plus tard cette année, et en 2020 sur Google Cloud Platform.

Dans ce guide

Cloudera est le dernier éditeur indépendant sur le marché des plateformes Big Data après sa fusion avec son ancien rival Hortonworks en janvier dernier et le rachat de MapR par HPE en août. Malgré son nom, la société tire toujours 90 % de ses revenus des installations sur site. De fait, elle a été durablement touchée par l'accélération des déploiements Big Data dans le cloud, avec pour conséquence le départ du PDG Tom Reilly et du co-fondateur et stratège en chef Mike Olson. Les faibles résultats du premier trimestre, après l'intégration d'Hortonworks, les ont poussés vers la sortie.

Rivaliser avec les géants du cloud

Selon les analystes ayant participé à un briefing de deux jours à New York la semaine dernière, Cloudera Data Platform devrait donner de meilleurs arguments pour rivaliser avec AWS, Microsoft et Google dans le cloud.

« Ils avancent dans la bonne direction », selon Doug Henschen, analyste chez Constellation Research. En plus des nouveaux services cloud, il relève que CDP prend en charge le stockage objet dans le cloud et qu'il apporte des fonctions administratives et une UX communes au cloud et sur site. Par ailleurs, CDP est capable de répliquer les données, les politiques de gouvernance, et la cartographie (Data Lineage) entre les systèmes, peu importe où ils sont exécutés.

Dans ce guide

« Cloudera offre une flexibilité et un choix qu'aucun service cloud ne propose actuellement », déclare Doug Henschen.

Le temps révolu d'Hadoop sur CDP

L'éditeur a également réinventé la manière dont il stocke et traite les données, selon William McKnight, président de McKnight Consulting Group. Les clients peuvent toujours utiliser Hadoop Distributed Files System au sein de CDP, mais Cloudera espère que la majorité d'entre eux choisira les options de stockage cloud natif. L'objectif : séparer les ressources de calcul des espaces des entrepôts de données.

Avec ce changement d'orientation, « Pour Cloudera, l'approche de la gestion de données uniquement basée sur HDFS relève du passé », a affirmé William McKnight. L'expert considère que la portabilité de CDP dans le cloud en fera une meilleure alternative pour des utilisateurs cherchant à déployer des applications de machine learning et des services d'analytique avancés à l'échelle d'une entreprise.

L'autre élément clé de cette nouvelle plateforme est le support de l'orchestration de conteneurs Kubernetes. L'engagement de Cloudera envers Kubernetes et les installations hybrides de systèmes cloud et sur site devrait lui permettre de prendre en charge une variété d'options d'infrastructure de cloud computing, selon Lynne Baer, analyste chez Amalgam Insights. CDP offre également aux

Dans ce guide

utilisateurs de Cloudera une approche pragmatique pour développer leurs entrepôts de données et leurs systèmes d'analyse basés sur Hadoop, a-t-elle ajouté.

Cloudera veut éviter la fuite de ses clients

Toujours selon Lynne Baer, la priorité de l'éditeur est de retenir les utilisateurs existants dans son giron. L'analyste assure que Cloudera peut compter sur plus de 900 clients qui génèrent au moins 100 000 dollars en revenus récurrents par an. Ce n'est pas suffisant selon elle : « le véritable défi est d'augmenter le panier moyen au fil du temps plutôt que de gagner de nouveaux clients ». Dans un tweet, Judith Hurwitz, présidente et CEO d'Hurwitz & Associates fait le même constat.

Tony Baer (sans lien avec Lynne Baer), directeur chez dbInsight et actif sur Tweeter, écrit en effet que logiquement les clients de CDP seront ceux « ayant des besoins et des ressources sophistiqués ». Sans surprise, il compare la situation de Cloudera avec celle que vit actuellement Teradata.

Les services Cloudera Data Warehouse, Cloudera Machine Learning et Cloudera Data Hub basés sur CDP sont facturés à l'heure d'utilisation dans le cloud AWS ; le taux horaire varie en fonction de la configuration des instances supportées qu'un client déploie.

Dans ce guide

La version sur site nommée CDP Data Center est disponible en préversion pour des utilisateurs sélectionnés par Cloudera. Elle sera lancée plus tard cette année à partir de 10 000 dollars par nœud et par an.

Dans ce guide

Cloudera vs AWS EMR : quelle distribution Hadoop choisir pour vos projets Big Data

George Lawton, journaliste

Les entreprises qui ont besoin d'une plateforme Big Data doivent généralement sonder eux-mêmes le marché pour choisir un fournisseur. La complémentarité des services AWS est indéniable, mais la solution de Cloudera est-elle un meilleur choix ?

MPLS Quand il a fusionné avec Hortonworks en janvier 2019, Cloudera a complété son offre Hadoop pour mieux concurrencer les fournisseurs cloud, AWS en premier lieu.

La Cloudera Data Platform (CDP), la solution Big Data issue de la fusion des deux éditeurs, est une offre open source hébergée sur le cloud. Elle est conçue pour défier Amazon Elastic MapReduce (EMR), un autre service cloud basé sur Hadoop. CDP est disponible depuis le début du mois d'octobre depuis AWS. Elle sera accessible depuis Microsoft Azure et Google Cloud platform d'ici à la fin de l'année 2020.

Dans ce guide

L'influence d'IBM sur l'offre de Cloudera

En juin 2019, Cloudera a entamé un partenariat avec IBM afin de proposer des solutions Big Data et IA, revendues par l'un et l'autre. Elles se nomment Cloudera Enterprise Data Hub, Data Flow, IBM Watson Studio et Big SQL.

Dans cet article, nous cherchons à savoir ce que ce partenariat entre Cloudera et IBM pourrait signifier pour les utilisateurs de workloads Big Data. Comment cela modifie-t-il le rapport de force entre Cloudera et Amazon EMR ?

Selon Dave Mariani, fondateur et CSO d'AtScale, un spécialiste de la virtualisation de data warehouse, ce partenariat est une reconduction de l'accord entre Big Blue et Hortonworks réalisé avant la fusion.

Auparavant, Cloudera et Hortonworks se concentraient sur la distribution du système basé sur HDFS et d'outils pour gérer d'importants lacs de données. Grâce à ces capacités, les entreprises pouvaient sauvegarder toutes leurs données à un seul endroit et les réutiliser à diverses fins analytiques.

En pratique, ces sociétés ont souffert de problèmes de performance liés au déploiement d'Hadoop sur site. En conséquence, elles ont choisi de se tourner vers les fournisseurs cloud pour structurer leur gestion de données.

Après la fusion, l'association entre IBM et Cloudera pourrait aider les clients à résoudre les problèmes de performances grâce aux services complets de

support et d'assistance fournis par IBM. De son côté, AWS offre un ensemble d'outils pour automatiser de nombreux aspects de déploiements Big Data. Amazon EMR constitue un choix intéressant pour les entreprises dotées de compétences liées aux technologies d'AWS.

Cloudera vs Amazon EMR

Ainsi, la plateforme CDP semble attrayante pour des acteurs qui posent les premières pierres d'une stratégie Big Data. Ces derniers doivent souvent coordonner leurs données et applications réparties sur site et dans le cloud. En revanche, ce partenariat ne risque pas de changer la donne pour les entreprises ayant déjà adopté les services AWS.

Tout comme IBM, Cloudera soutient une approche hybride et multicloud. CDP devrait être la mieux adaptée pour les entreprises prêtes à adopter cette stratégie, selon Dave Mariani. Il considère que cela empêche l'enfermement auprès d'un seul éditeur.

L'approche d'IBM en matière de développement d'applications consiste à utiliser Kubernetes et des containers pour que les charges de travail puissent être exécutées n'importe où : sur site, dans le cloud privé ou dans le cloud public. AWS, quant à lui, exécute les workloads liés à ces services depuis son infrastructure uniquement.

Dans ce guide

Bien que le multicloud semble une approche viable, Dave Mariani ne s'attend pas à ce que beaucoup d'entreprises empruntent cette route prochainement. Il a discuté avec plusieurs utilisateurs du cloud. Ceux-ci ont choisi un seul fournisseur et ont possiblement adopté les services d'un autre acteur pour la partie backup. Selon lui, le principal avantage pour les clients d'un fournisseur unique repose sur la complémentarité des services et des outils permettant à l'IT de ne plus s'occuper de l'intégration système.

Par exemple, Amazon EMR repose sur le stockage objet S3 connecté au data catalog AWS Glue et à la base de données Redshift. Les points forts d'AWS proviennent de l'intégration des APIs, de la disponibilité et du déploiement dans les différentes régions géographiques, ainsi que l'interopérabilité de l'ensemble de sa gamme de services.

Ces connexions « natives » désavantagent les solutions tierces telles que CDP par rapport à EMR. Surtout si les acheteurs de plateformes Big Data sont entraînés et certifiés par la filiale d'Amazon.

En revanche, Cloudera l'emporte sur AWS quand les entreprises cherchent des services, un support, une implémentation et une conformité haut de gamme pour leur plateforme de données, selon Marty Puranik, président d'Atlantic.net, un fournisseur d'hébergement.

La sécurité, la gouvernance et les métadonnées de Cloudera Data Platform sont normalement intégrées dans la couche d'échange entre les sources de

données et les workloads analytiques. Pour automatiser ce processus, Cloudera a mis au point un catalogue de données partagées nommé SDX. Afin de maintenir un haut niveau de sécurité dans Amazon EMR, les développeurs doivent gérer eux-mêmes le chiffrement entre les différentes applications.

Cependant, CDP n'est pas nativement compatible avec les notebooks basés sur Jupyter. Ceux-ci fonctionnent avec l'ensemble des produits AWS tels que S3, DynamoDB et Redshift. Déployer CDP implique plus de travail pour se connecter à ces environnements Jupyter. Ils sont utiles pour la visualisation, le nettoyage, la création de modèles de données. Les documents à partager peuvent contenir du code, des équations, des éléments visuels et du texte.

Déploiement et coût

Les différences de coût entre Cloudera et AWS se mesurent en termes de déploiement, de conformité aux réglementations, de sécurité et de performances. AWS s'adresse aux entreprises ayant une expertise interne et des centres d'excellence en matière de cloud computing, tandis que Cloudera et IBM offrent davantage de conseils par le biais de services professionnels.

« Le prix d'Amazon EMR affiché sur l'étiquette est plus bas, mais peut largement grimper si vous ne maîtrisez pas les outils », affirme Marty Puranik.

Dans ce guide

Par exemple, les entreprises peuvent payer des frais supplémentaires importants s'ils envoient plus de données dans le cloud que n'en nécessite un traitement analytique. Un autre gros problème peut venir d'une mauvaise configuration, comme le fait de laisser les buckets S3 ouverts. Capital One en a subi les frais : l'erreur a potentiellement touché 100 millions de clients.

Si les utilisateurs n'ont pas souscrit aux offres d'un fournisseur cloud ou s'ils ne sont pas sûrs de l'infrastructure dont ils ont besoin, ils devraient étudier l'offre de Cloudera et d'IBM avec attention, même si le coût de la solution est plus élevé que celui d'Amazon EMR. Pour bien faire son choix, rien de mieux que de lancer les workloads d'essais sur une machine virtuelle.

« Commencez par de plus petits projets, si possible, et déterminez celui qui convient le mieux à votre organisation », assure Marty Puranik.

Dans ce guide

7 étapes pour créer son data lake

David Loshin, PDG Knowledge Integrity Inc.

Peupler un cluster Hadoop de données qui ne sont ni organisées ni gérées correctement risque de nuire à vos projets analytiques. Voici 7 étapes clé qui vous permettrons de mieux utiliser les données de votre data lake.

Le concept du data lake (lac de données) a vu le jour avec l'émergence du Big Data et de l'intérêt des entreprises pour Hadoop comme plate-forme de stockage et de gestion. Cependant, le fait de plonger aveuglément dans le déploiement d'un lac de données Hadoop ne portera pas nécessairement votre entreprise sur les terres du Big Data -- du moins, ce ne sera pas une réussite.

C'est particulièrement vrai lorsque ces données en volume sont placées dans un environnement Hadoop de manière désordonnée. Cette approche pose plusieurs problèmes qui peuvent sérieusement entraver l'utilisation d'un lac de données – et avec l'analytique.

Par exemple, il est difficile de documenter et d'identifier les objets stockés ou leurs sources et leur provenance. Les data scientists et autres analystes ont donc du mal à trouver des données pertinentes dans un cluster Hadoop.

Dans ce guide

Difficile également pour les gestionnaires de données de savoir qui accède aux jeux de données et de déterminer le niveau d'accès adéquat.

Sans un processus bien géré, il sera également compliquer d'organiser les données et de regrouper des objets similaires pour faciliter l'accès et l'analyse.

Pourtant, aucun de ces problèmes n'a trait à l'architecture physique du data lake ou à l'environnement Hadoop sous-jacent. Les plus gros freins sont bien liés à un manque de planification ou à une mauvaise gestion des données.

Comment procéder étape par étape

La bonne nouvelle est que ces difficultés peuvent être facilement surmontées. Voici sept étapes qu'il convient de considérer :

1 - **Créer une taxonomie pour classer les données.** L'organisation des objets de données dans un lac de données repose sur leur classification. Identifiez alors chaque aspect clé des données comme le type de données, le contenu, les scénarios d'utilisation, les groupes d'utilisateurs possibles et la criticité des données. Cette dernière a trait à la protection des données personnelles et de l'entreprise, comme celles sur les clients ou celles sur la propriété intellectuelle.

2 - **Concevoir une architecture de données adéquate.** Appliquez la classification pour organiser les données dans votre environnement Hadoop. Le résultat doit comprendre par exemple la hiérarchie des fichiers pour le stockage

Dans ce guide

des données, les conventions de nommage des fichiers et des dossiers, les méthodes et les contrôles d'accès pour différents ensembles de données, et les mécanismes pour la distribution des données.

3 - **Utiliser des outils de profilage des données.** Dans bien des cas, l'absence de connaissances sur les données peut être minimisée en partie par l'analyse de leur contenu. Les outils de profilage des données peuvent être utiles pour recueillir de l'information dans les objets de données ; ce qui permet de les classifier. Le profilage des données dans le cadre d'un data lake permet également d'identifier les problèmes de qualité des données. Ceux-ci doivent être mesurés pour les corriger et s'assurer que les analystes travaillent avec les bonnes informations.

4 - **Normaliser l'accès aux données.** La multiplication des méthodes d'accès aux données par différentes équipes d'analyse - dont un grand nombre ne sont pas documentées – est aussi l'un des freins à un usage efficace d'un lac de données. La mise en place d'une API simple et commune peut simplifier l'accès aux données.

5 - **Développer un catalogue de données.** Un autre obstacle porte sur le fait que les utilisateurs potentiels ne savent pas ce qu'il y a dans un lac de données et où se trouvent les jeux de données dans Hadoop (ni leur qualité ni leur source, par exemple). Un catalogue de données collaboratif permet de documenter ces détails - parmi tant d'autres - pour chaque donnée. Il permet par exemple de capturer des métadonnées structurelles et sémantiques, la

provenance et la source, et des informations sur les privilèges d'accès. Un catalogue de données fournit également un forum permettant aux groupes d'utilisateurs de partager des expériences et des conseils.

6 - **Mettre en place des protections suffisantes des données.** Outre les aspects classiques de la sécurité IT, il convient également d'utiliser d'autres méthodes pour empêcher l'exposition des informations sensibles. Cela porte par exemple sur le chiffrement et le masquage des données ou sur une surveillance automatisée – cela permet de générer des alertes en cas d'accès ou de transferts de données non autorisés.

7 - **Evangélisation interne.** Enfin, assurez-vous que les utilisateurs de votre lac de données sont conscients qu'une gestion dynamique des données est nécessaire. Formez-les à trouver les ensembles de données dans les catalogues et à configurer les applications analytiques.

Pour que votre data lake donne tout son potentiel, il est crucial d'avoir un plan pour traiter les données avant de les migrer dans Hadoop. En appliquant ce qui est décrit dans cet article, vous contribuerez à rationaliser le déploiement d'un lac de données. Plus important encore, la bonne combinaison planification - organisation - gouvernance vous aidera à optimiser vos investissements et à réduire le risque d'échec.

Dans ce guide

Les principales distributions Hadoop sur le marché

Linda Rosencrance, journaliste

Voici les principales distributions Hadoop sur le marché et un ensemble d'éléments pour choisir celle qui convient le mieux aux pratiques analytiques en entreprise.

Pour répondre aux besoins des entreprises qui déploient Hadoop, les éditeurs et les fournisseurs cloud ont lancé des distributions commerciales et des technologies open source associées. Voici les solutions les plus répandues avant mai 2019.

Note de l'éditeur : Grâce à des recherches approfondies concernant le marché Hadoop, les rédacteurs de TechTarget se sont concentrés sur les éditeurs qui dominent le marché, en plus de ceux qui offrent les fonctionnalités traditionnelles et avancées. Notre recherche repose sur des données provenant de sondages TechTarget, ainsi que des rapports de cabinets de conseil, dont Gartner et Forrester.

Dans ce guide

Alibaba Cloud E-MapReduce

L'essor Alibaba Cloud Elastic MapReduce, aussi connue sous le nom E-MapReduce ou EMR, est une distribution Hadoop hébergée spécialisée dans le traitement massif et l'analyse de données. Construit sur les instances Alibaba Cloud Elastic Service, EMR est basée sur Hadoop et Apache Spark.

La solution permet aux entreprises de gérer leurs données dans un large éventail de scénarios comme l'analyse des tendances, le data warehousing et le traitement des données en ligne et hors ligne. Pour cela, EMR permet d'utiliser les composants Apache Hive, Kafka, Flink, Druid et TensorFlow.

Cette solution est censée simplifier l'import et l'export de données en provenance d'autres systèmes de stockage cloud ou de SGBD, à l'aide d'Alibaba Cloud Object Storage Service et Distributed Relational Database Service.

La plupart des retours clients sur Gartner Peer Insights semblent aimer le produit pour sa facilité de déploiement. Il leur permet aussi « d'ingérer, de structurer et d'analyser les informations », selon le site d'Alibaba, tout comme de gérer les clusters. Cependant, un des utilisateurs considère la plateforme comme trop compliquée et non fonctionnelle.

Les fonctionnalités d'Alibaba Cloud EMR sont les suivantes :

- Déploiement et expansion automatisés des clusters : les clients peuvent déployer et étendre les clusters depuis une interface web sans avoir besoin de gérer les équipements et le logiciel. Ceux-ci peuvent être liés à Hadoop, Kafka, Druid et Zookeeper. Les clients peuvent également ajouter, configurer et maintenir des composants en fonction des besoins de l'entreprise, et ajouter tout type de nœuds aux clusters existants.
- Planification des workflows : cette fonctionnalité facilite l'orchestration et la planification des tâches. Elle prend en charge la gestion graphique de ces dernières ainsi que leurs dépendances pour permettre aux entreprises de les exécuter et de les orchestrer. Ces flux sont produits sous forme de graphes orientés acycliques (DAG).
- Composants multiples : EMR comprend Hadoop, Spark, Hive, Kafka et Storm
- Support complet de l'écosystème Alibaba : l'outil prend en charge la lecture et l'écriture des données provenant des services de messagerie Alibaba Cloud, y compris les services Message Queue et Message Service, et supporte l'intégration SDK.
- Intégration des données : Elastic MapReduce s'intègre à des outils open source, hors ligne, en temps réel et avec ceux d'Alibaba.

Amazon EMR

Amazon EMR est un outil dédié au traitement big data et à l'analytique. Il offre un service extensible. AWS le présente comme une alternative aux clusters déployés en interne.

Les clusters Amazon EMR ont vu le jour en même temps que les frameworks Hadoop ou Spark. Ils sont habituellement couplés avec des utilitaires open source comme Hive ou Apache Pig.

Une fois combinés, ces frameworks peuvent traiter, analyser et transformer de vastes quantités de données. Ils interagissent également avec des bases de données ou des espaces de stockage objets comme DynamoDB ou S3 (Simple Storage Service). L'intégration avec les outils d'AWS permet, en principe, aux équipes de tirer des indicateurs des données analysées.

Sur le papier, les entreprises peuvent instantanément provisionner les capacités de calcul et de stockage nécessaires pour effectuer des tâches comme de l'indexation Web, l'analyse de logs, de l'apprentissage machine, du data mining, de l'analyse financière, de la recherche scientifique ou de la recherche bioinformatique. Par ailleurs, le service dispose d'une option pour faire évoluer automatiquement ou manuellement les capacités à la volée, suivant les besoins.

Enfin, EMR Notebooks fournit un environnement managé basé sur l'application Jupyter qui permet aux analystes, aux développeurs et aux data scientists de préparer, visualiser les données, bâtir des applications, collaborer entre eux et effectuer des analyses interactives en utilisant les clusters EMR.

Toutefois certains utilisateurs ont déclaré sur TrustRadius que si les fonctionnalités de machine learning d'EMR basées sur Hadoop et Spark sont de bonnes factures, elles ne sont pas aussi faciles à utiliser que celles de certains concurrents.

Azure HDInsight

Microsoft Azure HDInsight est lui aussi un service managé dans le cloud. Il repose sur des composants proposés dans la distribution Hortonworks Data Platform (HDP). HDInsight est vendu comme un moyen de déployer Hadoop et les autres solutions d'analyse de données Apache de manière plus économique.

Les clients peuvent utiliser les frameworks open source les plus populaires comme Hadoop, Spark, Hive, LLAP, Kafka, Storm, MapReduce et d'autres. Les scénarios envisageables sont nombreux : ETL, Data Warehousing, machine learning et internet des objets. Par ailleurs, Microsoft y adjoint ses propres services comme SQL Data Warehouse, Azure CosmosDB, Data Lake Storage,

Blob Storage, Event Hubs et Data Factory. Le but est de fournir une panoplie d'outils afin de construire des pipelines analytiques.

Azure HDInsight se connecte également à Azure Log Analytics, ce qui permet en principe de suivre les clusters depuis une seule interface. Le service est compatible avec un ensemble d'environnements de développement dont Visual Studio, Eclipse, IntelliJ, Jupyter et Zeppelin. Les développeurs peuvent utiliser des langages de programmation courants tels que Scala, Python, R, Javascript et .NET.

Tout comme le service d'AWS, il réclame des connaissances approfondies pour le maîtriser. « En général, cela demande tellement de temps pour apprendre aux clients à l'utiliser qu'il est plus facile de simplement le contrôler pour eux », déclare un utilisateur sur le site de notation Web G2.

Cloudera CDH

Cloudera Distribution Hadoop plus communément nommé CDH était le produit phare de Cloudera avant la fusion avec Hortonworks. Il inclut encore une fois Hadoop, Spark, Kafka et plus d'une douzaine de projets open source, tous étroitement intégrés au sein de la solution. CDH, offre les fonctionnalités clé d'Hadoop, c'est-à-dire un stockage évolutif, du calcul distribué, ainsi qu'une interface Web. La plateforme open source sous licence Apache comprend une

solution de traitement unifié par lots, des outils de recherche dont l'un basé sur des requêtes SQL, tout comme un système de contrôle d'accès par rôle.

Cette plateforme permet de stocker, de traiter, de découvrir et de réaliser des modèles associés à un grand volume de données. Elle dispose des fonctionnalités suivantes :

- Stockage des données structurées et non structurées
- Plusieurs types d'analyses des données partagées dont le machine learning, les traitements par batch ou en streaming et les fonctions analytiques SQL.
- Une seule plateforme disponible dans des environnements hybrides et multicloud.

Le framework Impala disponible depuis CDH permet d'effectuer des requêtes SQL directement sur les données stockées dans HDFS, Apache Hbase ou S3. Impala s'appuie sur de nombreuses technologies Hive dont le langage de requête HiveQL, les connecteurs ODBC (Open Data Base Connectevity) et Query UI.

Partie intégrante de CDH et disponible depuis Cloudera Enterprise, Impala est un moteur de traitement massivement parallèle (MPP) open source et analytique d'Hadoop.

Les retours concernant CDH sont bons sur le site Web G2. Les clients affirment qu'elle est facile à utiliser et remplit son rôle pour maintenir et stocker les données dans le cloud. Précisons que le support de la plateforme sera maintenu jusqu'en 2022 par Cloudera.

Google Cloud Dataproc

Google Cloud Dataproc est un service cloud managé pour lancer des clusters Spark et Hadoop. Le fournisseur assure que ce dernier accélère les traitements. Ceux qui duraient plusieurs heures prennent normalement quelques minutes.

DataProc est connecté avec d'autres services GCP (Google Cloud Platform), ce qui permet de disposer, selon le géant du cloud, d'une plateforme complète pour le traitement des données, l'analytique et le machine learning.

Cloud Dataproc propose les fonctionnalités suivantes :

- Gestion automatisée des clusters : cela permet la gestion des déploiements, le monitoring et le logging.
- Clusters redimensionnables : les clients peuvent choisir comment créer et gérer la taille de leurs clusters avec des options concernant le type de machines virtuelles, l'espace de stockage, le nombre de nœuds et la bande passante attribuée.

- Intégration : Cloud DataProc est nativement intégrée avec cloud storage, BigQuery, Bigtable, Stackdriver Logging et Stackdriver Monitoring
- Gestion des versions : un utilisateur peut permuter entre plusieurs versions d'images d'Hadoop, de Spark et autres.
- Haute disponibilité : les équipes exécutent des clusters avec plusieurs nœuds principaux et paramètrent les tâches pour qu'elles redémarrent en cas d'échec.
- Outils pour les développeurs : le service propose plusieurs outils pour gérer un cluster. Web UI, le SDK Google Cloud, des APIs RESTful et des accès SSH.
- Actions d'initialisation : permet d'installer ou de personnaliser les paramètres et les bibliothèques nécessaires au fonctionnement des clusters.
- Configuration manuelle ou automatique : gère le matériel et les logiciels suivant les besoins de l'entreprise.

Les avis disponibles sur le site web G2 sont globalement bons bien que certains utilisateurs pointent quelques problèmes d'interface.

Hortonworks Data Platform

Après la fusion d'Hortonworks et de Cloudera en janvier 2019, l'éditeur a lancé sa plateforme unifiée Cloudera Data Platform. Pourtant, Cloudera maintient le support de CDH et de HDP jusqu'en janvier 2022.

La Hortonworks Data Platform dispose peu ou prou des mêmes fonctionnalités que CDH en se basant uniquement sur des composants sous licence Apache. Cette distribution repose sur le système de stockage HDFS et Hadoop YARN.

YARN, un élément essentiel du projet Hadoop, est un gestionnaire centralisé pour la planification et la gestion de ressources du système. Il surveille également les opérations de traitement effectuées sur chaque nœud d'un cluster. Surtout, il permet de prendre en charge un plus grand nombre de traitements analytiques différents.

La version 3.1.0 de HDP ajoute de nouvelles fonctionnalités censées faciliter le travail des analystes. Le déploiement des applications est plus agile. La plateforme supporte davantage de workloads de machine learning et de deep learning ; elle permet de faire du data warehousing en temps réel et doit améliorer la sécurité et la gouvernance. L'éditeur assure qu'elle donne la possibilité d'exploiter leurs données plus rapidement dans des environnements hybrides.

L'architecture modernisée permet de stocker les données dans le cloud dans leur format d'origine sur Azure Data Lake Storage, Azure Blob, Amazon S3 et Google Cloud Storage. Elle prend également en charge les données en transit et au repos sur site et dans le cloud.

Sur Gartner Peer Insight, les clients déclarent que le produit rencontre de nombreux petits bugs que l'équipe de développement doit encore réparer. D'autres affirment que les clusters HDP sont difficiles à mettre en place dans de grands groupes.

MapR

MapR est une distribution d'Hadoop conçue pour les entreprises. Cette plateforme permet le stockage et le traitement d'importants volumes de données à l'aide de technologies open source sous licence Apache et quelques outils maisons. Selon l'éditeur racheté par HPE, ces composants propriétaires permettent une meilleure gestion tout en améliorant la résilience et la qualité des données présentes dans les clusters Hadoop.

MapR mise sur MapR XD Distributed File and Object Store, un système de fichier distribué, auparavant nommé MapR-FS, qui remplace HDFS. MapR Database prend la place de la base de données Hbase et MapR Control System constitue l'interface utilisateur de la plateforme.

Dans ce guide

Elle est compatible avec toutes les APIs et les outils de traitement de données de l'écosystème Hadoop. Les clients peuvent facilement migrer les datas vers d'autres distributions et vice-versa.

MapR Snapshots est conçu pour améliorer la protection des données. L'utilisateur sauvegarde des instantanés des fichiers et des tables à la demande ou à intervalles réguliers. Par ailleurs, l'éditeur fournit des services prêts à l'emploi de continuité d'activité et de reprise après sinistre.

La distribution comporte enfin un environnement de test hébergé sur une machine virtuelle qui inclut des tutoriels et des démonstrations d'applications pour les débutants.

Les retours clients disponibles depuis Gartner Peer Insight décrivent un produit efficace. Cependant, certains utilisateurs pointent du doigt les tarifs pratiqués et un support trop peu soutenu de Spark.

Qubole

Qubole Data Service (QDS) offre un déploiement automatisé et optimisé d'Apache Hadoop.

QDS est une plateforme cloud native vendue par son éditeur comme une solution complète pour l'analytique en profondeur, l'intelligence artificielle et le machine learning à partir d'une architecture Big Data. Elle dispose d'outils de

Dans ce guide

recherche SQL, de notebooks et des tableaux de bord basés sur des moteurs open source.

L'infrastructure partagée permet de gérer depuis un seul espace de travail les flux ETL, les workloads analytiques, d'IA et de machine learning à l'aide d'outils comme Spark, Presto, TensorFlow, Hadoop ou encore Hive.

Qubole se veut agnostique et propose à ses clients d'accéder, de configurer et de gérer leurs clusters Big Data depuis n'importe quel cloud et leur permet d'accéder en libre-service aux données à l'aide de l'interface de leur choix.

Ils peuvent requêter les données depuis une console Web dans le langage de programmation de leur choix, créer des applications intégrées à l'aide de l'API REST, d'utiliser le SDK pour ce faire, et se connecter à des outils métiers via ODBC ou JDBC.

Selon les témoignages clients « Qubole simplifie la gestion des clusters et des jobs Spark qu'ils soient planifiés ou non ».

« C'est un choix judicieux si vous voulez les outils de données les plus populaires et que vous ne voulez pas passer du temps à les maintenir vous-même », écrit un utilisateur sur le site Web G2.

Dans ce guide

Accéder à plus de contenu exclusif PRO+

Vous avez accès à cet e-Handbook en tant que membre via notre offre PRO+ : une collection de publications gratuites et offres spéciales rassemblées pour vous par nos partenaires et sur tout notre réseau de sites internet.

L'offre PRO+ est gratuite et réservée aux membres du réseau de sites internet TechTarget.

Profitez de tous les avantages liés à votre abonnement sur: http://www.lemagit.fr/eproducts

Images : Fotolia

Dans ce guide

David Castaneira | *Editeur*
TechTarget
22 rue Léon Jouhaux, 75010 Paris
www.techtarget.com