SAS®
&
HADOOP

Diffuser l'analytique au sein de l'entreprise

sas | THE POWER TO KNOW®

Table des matières

Introduction 3

Section 1 - Big Data Analytics et Hadoop : une révolution technologique... et culturelle 4
Section 2 - Une puissance potentielle inégalée et des coûts de mise en oeuvre modestes 9
Section 3 - Utiliser Hadoop comme catalyseur de vos projets analytiques 13

Chapitre 1 - Data lakes ou lacs de données : comment surnager dans l'abondance des données et comment en tirer profit ?.. 14

Section 1 - Hadoop et les entrepôts de données 15
Section 2 - Intégrer Hadoop dans une infrastructure décisionnelle 17
Les trois modules d'Hadoop 19
L'intégration d'Hadoop aux plateformes BI : l'exemple de SAS 21
Six scénarios pour intégrer Hadoop et les entrepôts de données existants 22
Section 3 - Pourquoi le Data Management est essentiel dans un contexte Hadoop 25
Section 4 - Comment utiliser une plateforme Hadoop sans avoir à jeter ses entrepôts de données ? 29
Section 5 - Des partenariats experts pour une parfaite intégration de vos big data analytics dans Hadoop 35
SAS & Cloudera 35
SAS & Hortonworks 38

Chapitre 2 - Comment mettre en place une stratégie Big Data avec Hadoop 40

Section 1 - Hadoop n'est pas une stratégie big data, c'est un élément du puzzle 41
Section 2 - Fast and Furious: quand les Bid Data Analytics rencontrent Hadoop 42

Chapitre 3 - Maintenant que vous savez nager, plongez ! Quelques applications 43

Des cas d'utilisation variés - Des besoins analytiques nouveaux 44
Industrie électrique : dompter les big data avec Hadoop 45
Hadoop : colonne vertébrale d'American Express 46
Comment Hadoop permet aux Etats-Unis d'accroître leur indépendance énergétique 48
Santé : prendre le virage de l'analytique pour éviter le chaos 50

Conclusion 52

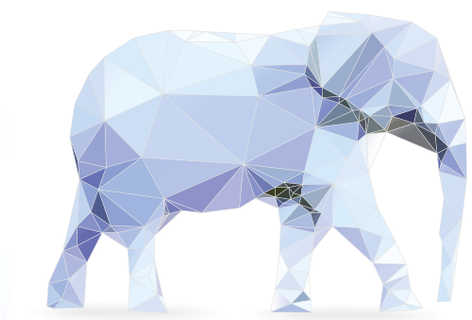
Introduction

Agilité et puissance à l'ère des big data

Si le terme big data n'est désormais plus un simple « buzz word » pour nombre d'entreprises, il n'en reste pas moins encore une problématique. Comment gérer ces masses de données, qui affluent en permanence, de sources multiples et dans les formats les plus hétérogènes ? Comment surnager dans ces véritables lacs de données ? Une des réponses est apparue il y a peu représentée par un logo à joyeux éléphant jaune : Hadoop. Et si la force et le calme de l'éléphant sont bien connus, qui aurait pu imaginer que cet animal était peu coûteux malgré son volume, et parfaitement agile !

C'est ce que nous allons exposer dans cette publication au travers d'une approche qui se veut explicite et illustrée d'exemples concrets.

Big Data Analytics et Hadoop : une révolution technologique... et culturelle



D'ici trois ans, la moitié des données du web pourraient être stockées sur Hadoop. En permettant une baisse historique des coûts de stockage des données, l'écosystème issu de la recherche de Yahoo! donne aujourd'hui tout son sens au big data. Les premières applications sont déjà là, et la révolution commence à prendre forme. Une révolution technologique, mais aussi culturelle : c'est une nouvelle ère de l'analytique qui s'ouvre, dans laquelle SAS s'engage pleinement.

Retour sur la conférence du dernier SAS Forum.

Serge Boulet avait donné le ton : « *La façon traditionnelle d'envisager la donnée a vécu* », lançait le directeur marketing de SAS. « *Nous étions dans un monde où la donnée était rare – et où rien n'était possible à moins d'avoir été prévu à l'avance. Avec le big data s'ouvre* ***une ère d'abondance, et ce qui n'était pas envisageable hier devient possible.*** »

Pour Mouloud Dey, Directeur solutions et marchés émergents chez SAS, deux grands mouvements sont à l'œuvre : d'une part la prise en compte d'un gigantesque volume de données aujourd'hui éparpillées ; d'autre part le développement des solutions en temps réel. Deux mouvements qui sont amenés à converger avec le développement de l'Internet des objets : qui peut nier qu'une révolution est en marche quand on prédit plusieurs dizaines de milliards d'objets connectés d'ici 2020 ?

Sur le plan technologique, cette révolution a son symbole : un éléphant jaune. « ***Hadoop est amené à s'imposer comme un standard de notre industrie****, assure Mouloud Dey. Il nous permet de* ***repenser les questions d'architecture pour nous attaquer au big data, et développer des solutions d'aide à la décision en temps réel****. »*

En réduisant les coûts de stockage, Hadoop s'est imposé comme une urgence IT. Il est aussi devenu une exigence business.

« Arrêtons le blabla, activons la data »

Abed Ajraou, Responsable BI de Solocal Group (ex-PagesJaunes), abonde dans ce sens. Il insiste notamment sur la transformation majeure qu'induit l'intégration de **données non structurées** dans les modèles de business intelligence : *« avec des téraoctets de données, il fallait une nouvelle architecture avec une technologie capable de traiter et d'indexer des bases de données en temps réel »*.

Dans cette logique, il voit aujourd'hui Hadoop *« comme un framework où l'on peut incorporer toutes nos données, structurées ou non, pour mieux les exploiter, un espace où s'ouvrent des possibilités nouvelles – notamment en temps réel »*. Illustration : une entreprise peut aujourd'hui **imaginer un « pré-lancement » de produit** : l'analyse en temps réel des premières réactions de consommateurs et d'internautes viendrait valider ou non (ou ajuster) le lancement de la production.

Au-delà des opportunités technologiques, Abed Ajraou met en avant l'aspect culturel de la transformation en cours. *« Hadoop nous permet de penser l'entreprise différemment : il faut tenter des choses nouvelles,* ***tester les modèles, et laisser parler la donnée****. La décision vient ensuite : si la data dit que ça ne marche pas, eh bien, on tente autre chose. »* Un monde catalysé par la donnée où disparaîtrait la peur de l'échec... Chiche ?

Agilité et gouvernance : deux approches compatibles

Sur le plan technologique aussi, Hadoop engendre de nouveaux paradigmes. Et il était temps ! Car pour Mouloud Dey, le modèle actuel basé sur les entrepôts de données, ne pouvait pas s'emparer du big data : *« Les architectures s'alourdissaient et le time-to-data avait atteint une limite »*.

L'écosystème Hadoop est venu répondre à cette question du stockage des données, et avec lui deux nouveaux modèles ont fait leur apparition. Premier modèle : les **« data lakes »**. Utilisés outre-Atlantique, ils permettent de stocker la donnée sans se poser auparavant la question de l'utilisation qui en sera faite. En Europe, à l'image de Solocal, les nouvelles expériences s'orientent plutôt vers le modèle des **« data labs »**, qui permettent aux utilisateurs métier de se poser de nouvelles questions, en dégageant de nouvelles opportunités plus agiles.

Quel que soit le modèle, l'objectif est le même : il s'agit **pour les métiers de gagner en agilité, et pour les directions IT d'assurer une meilleure gouvernance des données**. Deux approches complémentaires, selon Mouloud Dey, et qui constituent aujourd'hui la priorité de SAS en matière de R&D. *« Nous disposons déjà de solutions qui permettent d'accéder à ces nouvelle formes de données. Nous nous apprêtons maintenant à déployer de nouveaux modèles pour préparer, exploiter et modéliser ces données pour* ***accélérer non seulement le time-to-data, mais surtout le time-to-insight*** *– donc les cycles de décision »*.

« Une approche simple »

Max Brudey, Regional Sales Manager d'Hortonworks, s'est appliqué à démystifier le terme de big data. *« Je ne fais pas de big data,* dit-il. *Chez Hortonworks, nous essayons simplement de* ***créer les plates-formes les plus adaptées pour nos clients, alignées sur leurs outils métiers****, pour leur garantir à la fois la sécurité et la gouvernance de leurs données, et leur permettre de traiter ces données nouvelles [non structurées] qu'ils ne sont pas armés pour gérer aujourd'hui »*.

Pour lui, la question de la structuration des données est secondaire. Le principal blocage à la mise en œuvre de projets analytiques est plutôt lié à l'organisation en silo des entreprises, ou au manque de profondeur historique des données. *« Avec les modèles analytiques classiques, il y a toujours quelque chose qui manque »*, constate-t-il. Mais avec Hadoop, l'approche est différente et vient briser ces silos : ***« on prend la donnée dans son jus, on la stocke de manière massive et on fait tourner les moteurs par-dessus »***.

Une approche simple, insiste-t-il, qui permet des résultats rapides dès lors qu'un premier *use-case* est venu roder le système. Il donne l'exemple de HSBC qui, pour lutter contre la fraude, a appris à gérer des données sur plusieurs continents – malgré des alphabets et des modes de saisie très différents : un premier projet test à Hong-Kong, repris en Angleterre, et le **déploiement mondial** a été rapide.

La clé de cette rapidité a un nom : **YARN**. ***« Avec Yarn, nous avons réellement une notion de donnée unique »***, assure-t-il. Une donnée qui peut être chargée sur un seul cluster et reprise par des experts comme SAS pour faire tourner différents moteurs selon les besoins des métiers.

Le mode d'accès à ces technologies est assez simple, conclut-il, non sans rappeler que le fait de travailler en mode open source, comme Hortonworks sur Hadoop, ne signifie pas que le travail soit gratuit. Tout est toujours une question de valeur ajoutée...

Avec Hadoop, SAS couvre déjà l'ensemble du cycle analytique

Partenaire stratégique des distributions Hadoop les plus importantes du marché, SAS propose déjà des solutions innovantes qui couvrent l'ensemble du cycle de vie des applications analytiques.

Toutes ces actions peuvent être réalisées avec la même interface. A ce jour, c'est unique dans le monde de l'analytique.

Comment SAS peut vous aider

Big Data Analytics

SAS et Hadoop relèvent le défi des Big data. Et ils le remportent.

Que faire avec des volumes massifs de données si vous ne pouvez pas vous en servir ? Ou si leur analyse doit prendre des jours ou des semaines ?
Combiner le pouvoir analytique de SAS aux capacités d'Hadoop, dans un environnement unique et interactif, vous livre les décisions clé en main – pour des résultats rapides et une valeur maximale.

Découvrir les solutions SAS pour Hadoop

Scannez le QR code avec votre smartphone. Nécessite un lecteur de QR code.

sas®
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2014 SAS Institute Inc. All rights reserved. S117789US.0114

Une puissance potentielle inégalée et des coûts de mise en œuvre modestes

Un téraoctet de stockage coûte aujourd'hui moins de 100 $. Pour autant, dans la plupart des entreprises, le coût de traitement d'un téraoctet reste encore très élevé.

A l'heure des **big data** c'est désormais avec un pyjama jaune qu'il faudrait réécrire l'histoire. Hadoop a su imposer un nouveau style «pattes d'eph », mince des genoux et large des pattes, qui se concrétise, au premier abord, par ce clivage étonnant entre une puissance potentielle inégalée et des coûts de mise en œuvre plutôt modestes.

Un téraoctet de **stockage** coûte aujourd'hui moins de 100 $. Pour autant, dans la plupart des entreprises, le coût de traitement d'un téraoctet reste encore très élevé : selon certaines estimations, entre 100 000 et 200 000 $ dans les contextes massivement parallèles des « appliances » de bases de données. Ce même volume peut désormais être traité dans un cluster de **serveurs** banalisés (« commodity hardware ») pour moins de 5 000 $. Hadoop redéfinit ainsi le champ des possibles dès lors qu'on stocke et traite une très grande quantité de données quel qu'en soient le format ou la structure. Un seul cluster **Hadoop** peut désormais comporter 10 000 serveurs, et exécuter plus de 100 000 tâches simultanées.

Telle est la principale proposition de valeur d'Hadoop, un

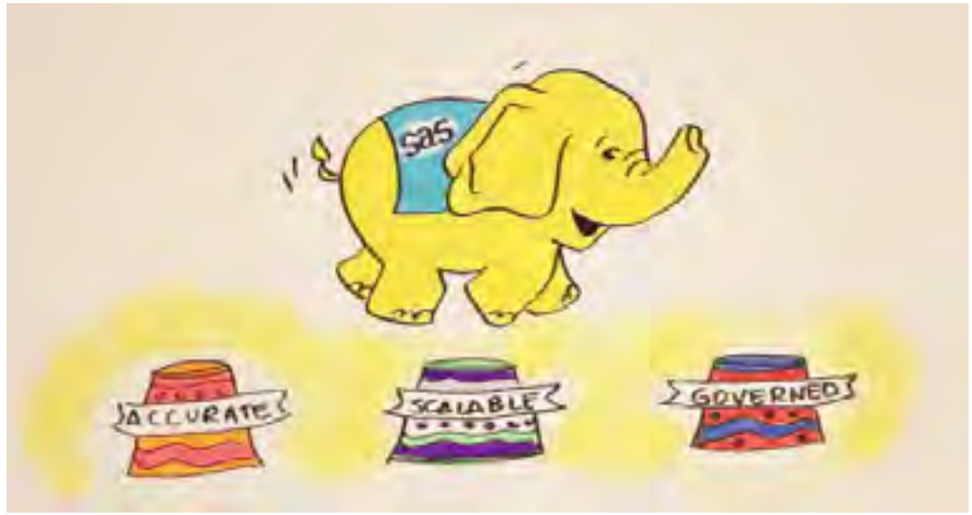
1 - Comprendre l'écosystème Hadoop

projet de l'**Apache** Software Foundation, largement médiatisé par les expériences et les succès très concrets des plus grands acteurs de l'économie numérique tels que Yahoo, **Google** et autres Facebook, qui y ont trouvé une partie des secrets de leur efficacité : stockage et calcul distribué en mode « shared nothing », architecture haute disponibilité tolérante aux pannes, absence de schéma prédéfini pour le stockage de données structurées et non structurées qui, de ce fait, délègue l'organisation effective des données au traitement qui les exploitera.

« Le vieil éléphant sait où trouver de l'eau »

Pour les entreprises plus traditionnelles, le modèle Hadoop peut aussi s'avérer pertinent, notamment pour démultiplier les possibilités des solutions analytiques les plus sophistiquées, et plus particulièrement pour les données «multi-structurées» qui se prêtent mal à l'utilisation de bases de données traditionnelles. Une enquête récente de TDWI Research montre que près de 88 % des entreprises considèrent Hadoop comme une réelle opportunité pour leurs systèmes d'information et la plupart des analystes estiment que d'ici 5 ans plus de la moitié des données mondiales seront traitées par Hadoop.

2 - les piliers SAS



Pour autant, les entreprises hésitent encore à laisser ce puissant animal approcher de leurs datamarts de porcelaine. Les principaux freins sont connus : difficultés de mise en œuvre, nécessité de développements spécifiques, utilisation en batch, besoins en compétences particulières, écosystème en développement semblant manquer de maturité, de pérennité, de support.

Puissant et agile

Pour la plupart d'entre elles, Hadoop n'a pas, au moins pour l'instant, vocation à se substituer aux datawarehouses et datamarts existants mais plutôt à les compléter pour traiter au quotidien, et pour un coût abordable, de vastes ensembles de données nouvelles, le plus souvent de qualité incertaine et qu'il ne serait pas

raisonnable, ni souhaitable, d'intégrer d'emblée dans les systèmes d'information sans investigation ou modélisation préalable. On pense tout naturellement aux promesses des big data : données des sites web, des réseaux et médias sociaux, données de capteurs, données des systèmes embarqués, données de mobilité qui n'ont pas vocation à être stockées dans leur intégralité dans les entrepôts de données et qui forment pourtant le cœur de nouveaux gisements de valeur.

Hadoop and the Enterprise Data Warehouse by Tamara Dull [1]

Dans cette perspective, les premiers usages d'entreprise pourraient bien être ceux d'espaces de transit (« staging area ») et/ou d'analyse permettant à des utilisateurs d'explorer ces données, à l'aide de solutions analytiques et de data visualisation, d'en dériver des informations qui iront éventuellement compléter utilement les systèmes existants ou resteront disponibles pour des analyses ultérieures.

Le coût d'appropriation, relativement modeste, permet ainsi de travailler sans échantillonnage sur l'intégralité et la diversité des données disponibles, structurées ou non, sans a priori sur leurs volumes, leurs contenus, leurs structures, leurs formats. Cette souplesse pourrait aussi être synonyme d'agilité pour les utilisateurs.

Mais la principale difficulté est que Hadoop reste un écosystème faisant appel à de nombreux outils dont la prise en main peut constituer un réel obstacle à une large diffusion dans l'entreprise. Pour dépasser ces limites, il est essentiel que l'ensemble des profils concernés, du « business analyst » au « data scientist », puissent collaborer autour de solutions innovantes de business intelligence appropriées à leurs besoins en termes de data visualisation, de reporting, mais aussi de modélisation, de data et de text mining.

Une convergence entre les environnements décisionnels existants et ce nouvel écosystème pourrait bien être la recette du succès d'Hadoop dans l'entreprise.

Les utilisateurs pourraient ainsi de manière très réactive, analyser de nouvelles formes de données, les modéliser, obtenir très rapidement des résultats, faire plusieurs itérations ou simulations sans dépendre de la lourdeur des projets traditionnels et sans cannibaliser les systèmes d'informations existants.

Les big data imposent clairement un nouveau rythme aux directions des systèmes d'information. Pour celles qui entonneront la « marche des éléphants » il faudra impérativement mettre la valeur ajoutée de l'environnement Hadoop en perspective des environnements existants et donner aux utilisateurs des solutions qui leur donnent une autonomie et une agilité équivalentes dans les deux contextes.

MOULOUD DEY,
Directeur solutions et marchés émergents, SAS Institute

[1] ***Hadoop and the Enterprise Data Warehouse – a non-geek's big data playbook, Tamara Dull***

***Publié sur le Journal du net, voir l'article*.**

3 - Le Processus de décision

Utiliser Hadoop comme catalyseur de vos projets analytiques

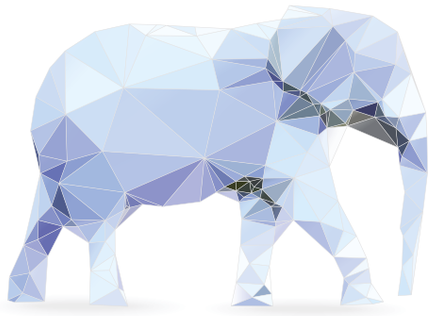
La réputation de Hadoop n'est pas usurpée. Information financière, tweets, transactions commerciales, etc., la célèbre solution de calcul distribué permet de traiter, stocker et analyser à moindre coût tout type de données en temps réel. Pourtant, certaines entreprises considèrent encore Hadoop comme un simple espace de stockage à prix discount. Exemples à l'appui, Paul Kent, vice-président big data chez SAS, démontre au contraire son potentiel pour déployer des modèles d'analyse prédictive sans précédent.

Lorsque vous postez un message sur Facebook, vous interagissez avec un site dont le bon fonctionnement repose sur Hadoop. Yahoo, Netflix, LinkedIn, Amazon ou encore Twitter sont d'autres illustres utilisateurs de ce système de stockage et de traitement des données devenu incontournable. Selon Paul Kent, le modèle économique de ces géants de l'Internet dépend largement du logiciel ouvert qui a pour emblème… un éléphant jaune. Hadoop dispose en effet de nombreux avantages comme son accessibilité, la rapidité de sa mise en œuvre et surtout son faible coût.

Ses capacités en calcul parallèle font de Hadoop la technologie naturelle à la base d'un système d'analyse de données efficace, affirme Paul Kent. Pas seulement pour comprendre ce qui s'est passé, mais aussi pour construire des modèles prédictifs permettant d'anticiper sur ce qui pourrait arriver et prendre les décisions en conséquence. Pour convaincre le lecteur, l'expert liste une série d'exemples concrets : AT&T se sert de Hadoop comme plate-forme analytique pour désamorcer d'éventuels désabonnements ; Cisco en a besoin pour fabriquer 30 000 scores d'appétence par trimestre ; une grande société de services financiers utilise Hadoop pour construire une gigantesque base de données référençant pêle-mêle ventes additionnelles, montées en gamme et activités de prévention.

Pour conclure, Paul Kent souligne que l'adoption de Hadoop ne nécessite aucun changement dans l'infrastructure informatique de l'organisation. Des outils visuels et interactifs existent sur le marché pour simplifier la découverte de tendances dans les données par de multiples utilisateurs.

Lire l'article intégral ici



Data lakes ou lacs de données : comment surnager dans l'abondance des données et comment en tirer profit ?

Hadoop et les entrepôts de données

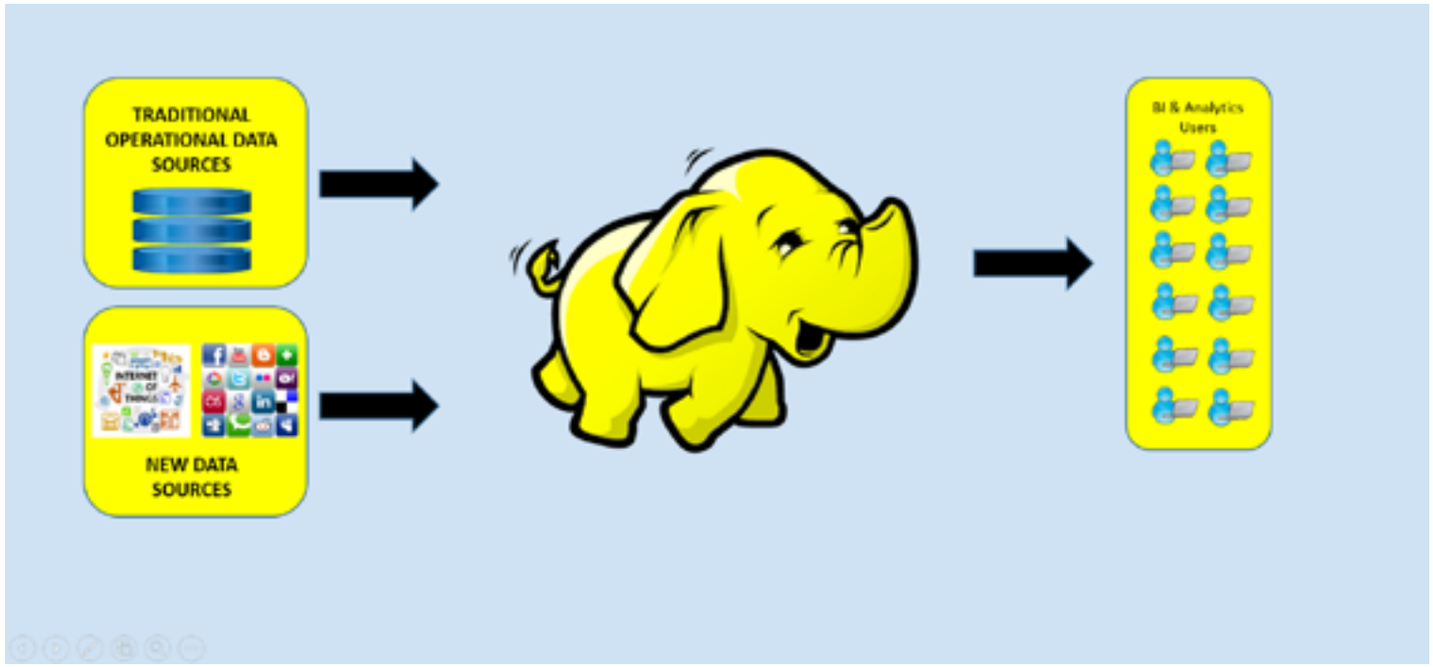
Qu’est-ce que le data lake Hadoop ?

Comme pour tous les nouveaux concepts, le concept de « data lake » est en pleine évolution et il règne sur le marché une grande confusion dans son emploi.

Le concept original

Si l’on s’appuie sur la définition du data lake (wikitionary), le concept consisterait à déployer un unique data lake basé sur Hadoop dans lequel la grande majorité des données d’une organisation serait déplacées et stockées supprimant ainsi le besoin de recourir aux datawares et aux datamarts.

Pour y parvenir, la plupart des entreprises devront mettre en œuvre de telles charges de développement que cet idéal risque de ne pas devenir réalité de sitôt. Sait-on même si, d’ici 3 ou 5 ans, Hadoop pourra supporter l’ensemble des applications que supportent les datawares, tout en satisfaisant le niveau de service escompté ou les exigences réglementaires ?

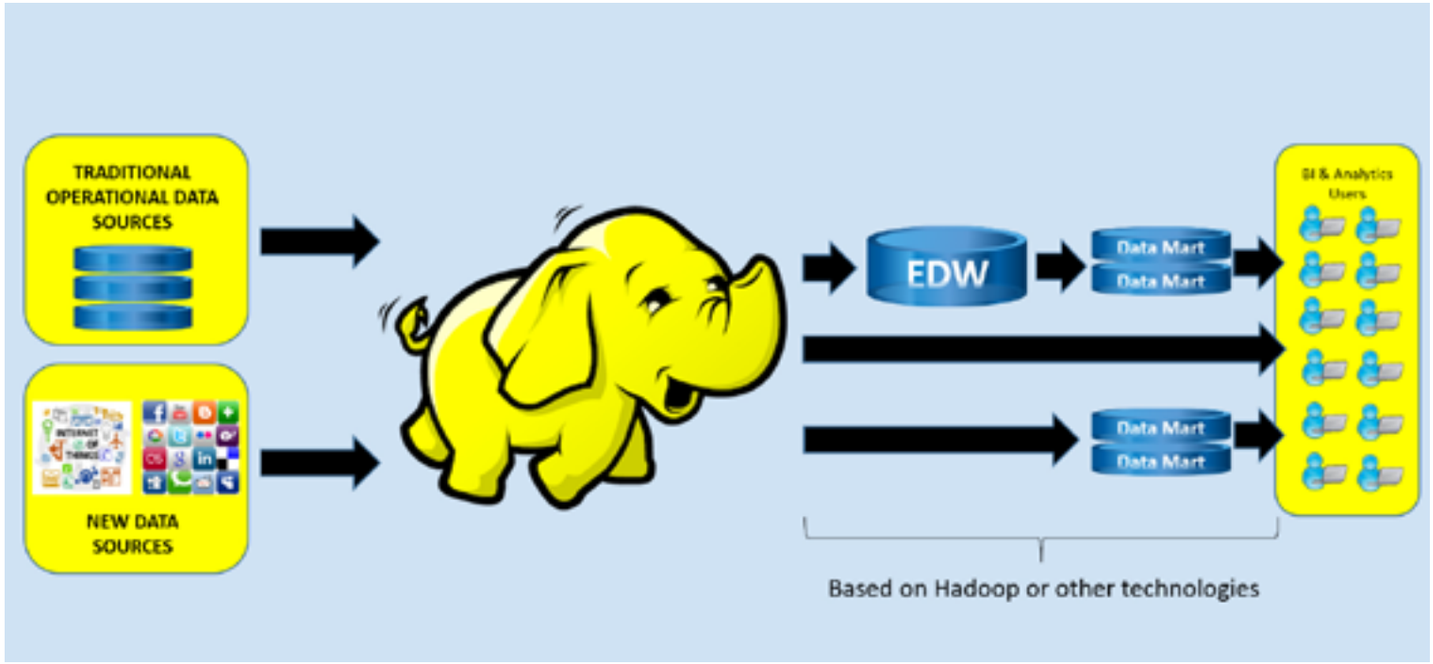


Vision évolutive

Par la force des choses, les entreprises sont en train d'aménager le concept de « data lake » comme un nouvel écosystème de données complémentaire de leur environnement actuel. Dans cette vision évolutive du « data lake », Hadoop devient l'environnement de données principal, recueillant toutes les données brutes en amont et alimentant après transformation, les SGBD et les appliances existants ainsi que les nouveaux entrepôts de données structurés dans Hadoop tel Hive, Cloudera Impala et Pivotal Hawq, qui assurent une structure aux applications en aval.

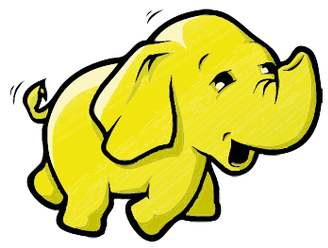
Ce réservoir de données, comme certains l'appellent déjà, constitue une alternative à la vision centralisée du « data lake », auquel elle apporte des principes de structure et de contrôle.

On peut néanmoins distinguer ces deux formes de plateformes centralisées Hadoop en retenant que le réservoir de données est conçu avec une idée de structure et de contrôle contrairement au « data lake ».



Source: SAS Blog, Mark Torr, Senior Director, Analytical Platform

Télécharger l'article complet ici



Retour sommaire

Intégrer Hadoop dans une infrastructure décisionnelle

Infrastructure décisionnelle : comment intégrer Hadoop en douceur

Le *big data* représente un gisement de valeur important pour les entreprises. Les technologies associées, comme Hadoop, ne constituent pas un monde à part dans les systèmes d'information. L'adoption de ces solutions encore jeunes représente un certain nombre d'enjeux, tant pour les DSI que pour les utilisateurs métier. À ce titre, l'intégration d'Hadoop avec les outils de *Business Intelligence* et d'analyse des données peut rendre ce type de projets possible en réduisant les coûts, en facilitant la prise en main et en augmentant les possibilités d'interagir avec les données.

Dans les entreprises, le traitement de la donnée est en train de changer d'échelle. Succès des médias sociaux, développement des objets connectés et des capteurs intelligents, dématérialisation de plus en plus poussée des échanges : tous ces phénomènes multiplient les sources de données potentiellement exploitables, générant, dans certains cas, des données à haute vélocité, c'est-à-dire qui se renouvellent très rapidement.

Face à cette masse de données, bien souvent non-structurées, les architectures analytiques classiques ont du mal à suivre et atteignent vite leurs limites (coûts et performances). Les entreprises qui souhaitent explorer ces big data afin d'y trouver de nouveaux gisements de valeur se sont donc mises en quête d'autres solutions. Parmi celles-ci, le *framework* d'analyse Hadoop est, à l'heure actuelle, l'une des options suscitant le plus d'intérêt chez les DSI. Dans une étude publiée en 2013, le cabinet d'analystes Gartner souligne ainsi que « les organisations sont conscientes de la force que l'analyse basée sur Hadoop peut apporter aux projets big data », en particulier quand ceux-ci concernent *« des données faiblement structurées, du texte, de l'analyse comportementale ou qu'il s'agit de requêtes basées sur des notions de temps. »* Dans ce même rapport, Gartner a également estimé que, à l'horizon 2015, près de 65 % des applications analytiques avancées embarqueraient Hadoop.

Qu'est-ce que Hadoop ?

Hadoop est à l'origine un *framework* open source destiné à l'analyse répartie de grands volumes de données, structurées ou non. Il a été créé par Doug Cutting, puis confié à la fondation Apache en 2009. Son architecture distribuée lui permet de bénéficier d'une vraie capacité à monter en puissance : Hadoop implémente un ensemble d'algorithmes conçus pour répartir les traitements sur des grappes de serveurs, chaque nœud assurant une partie des calculs et du stockage. De ce fait, il se différencie des autres solutions positionnées sur le créneau du big data, dont la plupart reposent sur des architectures matérielles spécifiques de type appliance.

Sur cette base, des acteurs ont choisi de bâtir des offres packagées destinées aux entreprises, ajoutant des outils de gestion, de sécurité et de gouvernance et fournissant des services de support. Parmi ceux-ci figurent Cloudera, Hortonworks, MapR, IBM InfoSphere BigInsights ou encore Pivotal.

Les trois modules d'Hadoop

À l'heure actuelle, le projet Hadoop comporte trois grands modules :

- ***Hadoop Distributed File System (HDFS)*** est un système de fichiers distribué permettant d'accéder rapidement aux données. Il fournit, par ailleurs, des mécanismes de résilience et de résistance aux pannes, grâce à la réplication des données sur différents noeuds du cluster.
- ***Hadoop YAR N*** regroupe des outils de planification des traitements et de gestion des clusters.
- ***Hadoop MapReduce*** implémente des algorithmes de traitement en parallèle des données. Ceux-ci, inspirés d'un modèle mis au point par Google, permettent par exemple de trouver des amis, de calculer le nombre moyen de contacts dans un réseau social ou encore de traiter des données d'accès au Web, pour analyser le trafic et les comportements.

À ceux-ci viennent s'ajouter un certain nombre de modules complémentaires, notamment Hive, qui permet d'interroger les données stockées dans Hadoop à travers HiveQL, un langage de requête proche de SQL, Pig, un langage de scripts pour interagir avec de larges ensembles de données, ou encore Spark et Tez, d'autres systèmes de traitement des données qui offrent une alternative à MapReduce.

Pourquoi Hadoop intéresse les entreprises

L'attractivité majeure d'une technologie comme Hadoop tient au fait que, fonctionnant sur des serveurs standards, elle ne nécessite pas l'achat de matériel spécifique. Elle s'avère donc particulièrement séduisante, face à d'autres solutions au coût d'entrée plus élevé, et devient encore plus compétitive, dès lors qu'il s'agit de monter en puissance, tant au niveau des coûts de déploiement que du support. Par ailleurs, Hadoop ne nécessite pas de connaître la structure des données qu'il stocke, contrairement aux bases de données relationnelles qui imposent de définir tables et relations avant de pouvoir accueillir des données. Cette souplesse rend Hadoop particulièrement adapté au stockage de données non structurées, comme du texte ou des documents multimédia.

Grâce à ces avantages, Hadoop peut répondre à des cas d'usages très divers, auparavant coûteux ou complexe à aborder. Dans le domaine financier, il s'agit, par exemple, de l'analyse de risques et de la détection de fraudes ; dans le domaine médical, du suivi des essais cliniques sur grande échelle ou de la recherche génomique ; dans le marketing, de la mise en place de recommandations d'achats ou de contenus personnalisées, en fonction des données fournies par les consommateurs. Du côté des industriels, il peut

s'agir d'optimiser l'usage des équipements en analysant l'ensemble des données émises par les machines et les capteurs. Enfin, les informaticiens peuvent plus facilement analyser des ensembles de journaux et des données de supervision, afin de déterminer la cause d'une panne complexe.

Malgré ces atouts, Hadoop souffre toutefois de quelques points faibles. Ainsi, l'interrogation des données s'effectue, pour l'essentiel, sous forme de requêtes en mode batch. Même si les temps de traitement sont rapides, cela ne permet pas à l'utilisateur d'explorer les données de manière dynamique, à l'instar de ce que proposent, par exemple, les outils d'analyse visuelle. Par ailleurs, Hadoop ne propose pas, à ce jour, d'environnement de développement intégré pour simplifier l'écriture des requêtes. Il nécessite donc un bon niveau de compétences en programmation pour être pris en main, ce qui exclut la majorité des utilisateurs métier. Enfin, sur un plan plus large, Hadoop est une technologie jeune : son implémentation est à la portée d'une DSI capable d'y consacrer du temps,

Les principales forces et faiblesses d'Hadoop

Les forces	Les faiblesses
• Performances • Coûts (d'entrée et de support) • Système de stockage souple • Écosystème large et en plein essor	• Latence liée à l'exécution des traitements en mode *batch* • Manque d'outils de requête et d'exploration • Manque d'outils de management et de gouvernance des données
Les opportunités	**Les menaces**
• Analyse de données auparavant inexploitées, car peu ou pas structurées • Stockage de données volumineuses	• Rareté des compétences • Solution inadaptée aux utilisateurs métier

Source : IT Business Review.

L'intégration d'Hadoop aux plateformes BI : l'exemple de SAS

Actuellement, la plupart des entreprises hors du secteur informatique sont en phase de découverte d'Hadoop. Dans le même temps, plusieurs fournisseurs IT explorent des modes d'intégration plus poussés, afin de tirer parti des capacités de stockage et de traitement d'Hadoop dans leurs solutions.
Certains acteurs vont même jusqu'à embarquer Hadoop dans leur écosystème, afin de faciliter son adoption par les clients, notamment les fournisseurs d'infrastructures. Après l'intégration d'Hadoop aux environnements de stockage et d'exploitation, la prochaine étape logique serait de transformer et de gérer ces données, puis d'utiliser des outils analytiques pour les explorer rapidement et en extraire de l'information.
Pourtant, les acteurs issus du monde décisionnel et analytique sont encore rares à suivre cette piste. Parmi ceux qui ont choisi de tirer parti de l'essor d'Hadoop figure SAS. Depuis plus de deux ans, l'éditeur travaille à l'intégration de la technologie à ses offres, une démarche qui s'est déjà concrétisée à travers plusieurs solutions. A l'heure actuelle, l'intégration de Hadoop avec la plateforme SAS est possible à plusieurs niveaux, permettant de combiner les possibilités des deux environnements, en fonction des objectifs et du degré d'adoption d'Hadoop souhaités.

1. Hadoop comme source de données

Le premier niveau d'intégration est de considérer Hadoop comme une source de données parmi d'autres, consommable depuis les applications analytiques. Cette intégration est facilitée par des connecteurs adaptés. Il est possible, par exemple, d'interroger des données stockées dans Hadoop en utilisant les interfaces natives de SAS, sans avoir besoin d'être expert MapReduce ou Hive.

2. Hadoop comme système de stockage permanent

Un second niveau d'intégration, plus poussé, consiste à utiliser un cluster Hadoop comme système de stockage permanent pour l'ensemble des applications décisionnelles et analytiques. Les données résidentes, stockées dans HDFS, sont chargées en mémoire dans chaque noeud du cluster. Elles restent donc accessibles en permanence par les applications analytiques. Dévoilé en 2012, le serveur SAS LASR Analytic Server, sur lequel repose l'offre In-Memory de SAS, a été conçu pour permettre ce type d'utilisation. Les données peuvent être manipulées

à la volée, sans avoir besoin de relancer une requête à chaque nouvelle analyse.
Paul Kent, Vice-président Big Data chez SAS, donne quelques chiffres pour illustrer les possibilités de cette architecture en termes de performances : avec celle-ci, il est possible de parcourir un milliard d'enregistrements par seconde, d'en faire une analyse synthétique en 0,2 seconde et d'établir 45 paires de corrélations simultanées en 5 secondes environ.

3. Hadoop comme système de calcul

Dans la forme d'intégration la plus avancée, ce sont à la fois les données et les traitements qui sont répartis sur les noeuds du cluster Hadoop. Les analystes peuvent ainsi bénéficier de la puissance de calcul d'Hadoop directement depuis leurs applications SAS.
La solution SAS Visual Analytics s'appuie notamment sur ce modèle, pour proposer des fonctionnalités d'analyse et d'exploration visuelle sur des données volumineuses et/ou peu ou pas structurées.
Les utilisateurs sont les premiers bénéficiaires d'une telle stratégie. Ils peuvent tirer profit des performances d'Hadoop de manière transparente, à partir d'outils d'analyse qui leur sont familiers. De cette façon, la courbe d'apprentissage associée au déploiement d'une nouvelle technologie se trouve considérablement simplifiée.

En 2014, SAS a décidé d'offrir ce même choix aux statisticiens et data scientists, avec deux nouvelles solutions : la première proposant des fonctionnalités d'analyse prédictive et statistique, à travers une interface visuelle et intuitive, et la seconde un environnement de programmation, permettant de
construire et de déployer des modèles d'analyse sophistiqués sur Hadoop.
À travers cette approche sur plusieurs niveaux, SAS laisse aux entreprises le choix : celles-ci peuvent aborder Hadoop de manière progressive, en fonction de leurs besoins.
L'intégration du framework dans une plateforme décisionnelle permet également de bénéficier du meilleur des deux mondes, en palliant les faiblesses de Hadoop et en accentuant ses forces.

Six scénarios pour intégrer Hadoop et les entrepôts de données existants

Ces scénarios envisagent différentes possibilités pour combiner les environnements décisionnels classiques et les environnements big data. Les systèmes décisionnels et analytiques classiques reposent sur des entrepôts de données (datawarehouses), alimentés à partir des données structurées, issues des bases transactionnelles, et servant de source pour les applications de Business Intelligence (BI).
Les systèmes de big data se caractérisent, quant à eux, par des sources de données non structurées qui alimentent un cluster Hadoop, lui-même servant de source pour les applications de big data.

Scénario 1 : utiliser Hadoop comme une plateforme intermédiaire de collecte et de préparation (staging) des données structurées entre les bases transactionnelles et les entrepôts de données.

Scénario 2 : utiliser Hadoop comme plateforme intermédiaire pour le traitement et la transformation des données structurées, avant leur chargement dans les entrepôts de données, le cluster jouant alors un rôle proche des solutions d'ETL (extraction, transformation et chargement des données).

mais les compétences nécessaires pour le déployer et le mettre en œuvre sur une plus large échelle sont encore rares sur le marché. Une enquête menée, fin 2013, par IDC pour Red Hat auprès d'entreprises envisageant de déployer Hadoop vient corroborer ces constats. Parmi les défis rencontrés dans ce type de projet au niveau applicatif, les décideurs IT interrogés citent en premier lieu la complexité de l'intégration des données et le manque de compétences.

Tirer parti d'Hadoop : quelle stratégie adopter ?

En tant qu'alternative aux entrepôts de données traditionnels, Hadoop est en train de devenir un composant incontournable des projets qui traitent de multiples sources de données. Les décideurs informatiques ont donc tout intérêt à s'y préparer dès à présent, risquant d'y être confrontés à plus ou moins brève échéance.

Adopter une démarche progressive permet de démarrer en douceur avec Hadoop. À travers les étapes proposées ci-dessous, les équipes de la DSI peuvent prendre le temps de se familiariser avec la technologie, tout en accompagnant sa montée en puissance.

Scénario 3 : utiliser Hadoop pour traiter et transformer des données à la fois structurées et non structurées avant de les charger dans les entrepôts de données.

Scénario 4 : utiliser Hadoop comme environnement, pour archiver toutes les données (structurées, non structurées, données des datawarehouses…) et comme source pour les applications big data (et pour les applications décisionnelles et analytiques si celles-ci ont été adaptées).

Scénario 5 : utiliser Hadoop comme une source de données complémentaire, capable de stocker des données structurées et non structurées et alimentant à la fois les entrepôts de données et les applications de BI classiques (si celles-ci ont été adaptées).

Scénario 6 : faire d'Hadoop la base principale, stockant les données structurées et non structurées et alimentant l'entrepôt de données, les applications big data, les applications décisionnelles et d'analyse statistique.

Ces six scénarios sont issus de :
A non-geek big data playbook : Hadoop and the Enterprise Data Warehouse, de Tamara Dull.

Lien vers le document

1. Évaluer la compatibilité de Hadoop avec l'infrastructure

Selon les analystes de Gartner, la première étape est de vérifier que les infrastructures actuelles sont compatibles avec Hadoop, sur le plan matériel, mais aussi applicatif. Il ne faut pas perdre de vue qu'Hadoop n'a pas vocation à fonctionner de manière isolée. Il est donc indispensable d'évaluer, à travers un ou plusieurs projets pilotes, comment Hadoop peut s'intégrer avec les autres bases de données et datawarehouses de l'entreprise.

2. Tester différents scénarios

De multiples usages d'Hadoop sont possibles en fonction du contexte de l'entreprise. Dans un second temps, il convient de déterminer quels sont les scénarios les plus pertinents (voir encadré). Dans une DSI en phase de découverte, Hadoop peut se positionner comme un moteur de calcul adossé à une base de données classique, notamment pour stocker des données non structurées. Dans une entreprise plus avancée dans l'adoption de Hadoop, il est en revanche souhaitable d'examiner des scénarios permettant l'exploration en direct des big data.

3. Rechercher les fonctions analytiques susceptibles d'intéresser les utilisateurs métier

En parallèle du travail de débroussaillage mené par les équipes IT, Gartner préconise, côté métier, de rechercher et d'identifier des projets Hadoop porteurs d'une forte valeur pour l'entreprise. Ce travail est l'une des missions des data scientists, une nouvelle fonction en pleine émergence. À la croisée de l'informatique, de l'analyse statistique et de l'expertise métier, ces experts doivent connaître les différentes sources de données de l'entreprise et mettre à profit leurs compétences pour trouver de nouvelles façons de les exploiter.

4. Considérer l'analyse en mémoire

La multiplication des accès aux disques peut limiter les gains de performance liés à l'adoption d'Hadoop. Pour construire des applications analytiques fluides, qui permettent aux utilisateurs d'expérimenter rapidement plusieurs scénarios, il est intéressant d'enrichir Hadoop avec des capacités d'analyse en mémoire, comme le recommande Fern Halper, directrice de recherche à l'institut TDWI.

5. Industrialiser et outiller

Dès que les premiers projets pertinents sont identifiés et enclenchés, il est souhaitable de réfléchir à l'industrialisation et à l'outillage. En effet, il faut non seulement se conformer aux exigences de fiabilité et de robustesse en vigueur dans l'entreprise, mais également prendre en compte les besoins et les pratiques des utilisateurs métier en matière d'applications décisionnelles, un domaine en pleine évolution. Cela signifie réfléchir au choix de la distribution à adopter, à l'infrastructure matérielle à mettre en place, mais aussi aux modes d'accès et de consommation de l'information que l'on souhaite favoriser. Enfin, il importe de veiller à industrialiser le déploiement des modèles d'analyse sur les big data stockés dans Hadoop, faute de quoi tout projet, si prometteur soit-il, ne dégagera pas de bénéfices sur le plan opérationnel.

Source : IT Business Review N° 7 Juin 2014 par Aurélie Chandèze

Les motivations pour utiliser Hadoop	
Analyse des données transactionnelles liées aux ventes et aux points de vente	44 %
Analyse des comportements des consommateurs en ligne	38 %
Développer de nouveaux services	38 %
Analyse des données liées aux opérations	37 %
Analyse des données issues de terminaux ou de machines	32 %
Modernisation d'une solution de datawarehouse existante	23 %
Analyse de données (sites Web, e-mails)	17 %

Source : IDC.

Pourquoi le Data Management est essentiel dans un contexte Hadoop

- **Hadoop promet l'exploitation de volumes quasi illimités de données de tout type en temps réel. Y a-t-il des freins ou des points d'attention dans la mise en œuvre d'un projet analytique dans une infrastructure Hadoop ?**

Une partie de la réponse est d'ores et déjà dans la question. Hadoop est avant tout une infrastructure technique décuplant les capacités d'exploitation des données. Par conséquent, quelle que soit la finalité du projet, Hadoop n'adresse que cette thématique technique. Ainsi les sujets liés à la gouvernance, la qualité, la sécurité des données, qui sont au demeurant des défis permanents pour toute organisation hors contexte Hadoop, sont définitivement des points de vigilance.

Des modèles de maturité en matière de gestion des données ont déjà été présentés par le passé, mais peu sont ceux qui ont anticipé la démultiplication des capacités de stockage et d'analyse offertes par les technologies Hadoop. Le déploiement d'une infrastructure Hadoop, par analogie, revient à investir dans un immense entrepôt logistique, situation qui crée de facto une crise de « coordination » évidente. L'organisation du stockage, de l'approvisionnement, des ressources, de la sécurité des accès ainsi que la définition de procédures qualité restent à définir ; et sans nul doute que de nombreuses organisations qui éprouvent aujourd'hui des difficultés de gestion des données avec leur « small data » ne feront qu'amplifier ces mêmes difficultés en passant sur Hadoop. La maîtrise et la coordination du patrimoine et de la production de données est un pré requis pour la mise en œuvre d'une stratégie analytique.

- **Comment résoudre les problèmes de variété des données ?**

Est-ce réellement un problème ? La question réside davantage dans le fait de déterminer quelles informations pourraient être exploitées. Auparavant, seules les données dites structurées pouvaient être industrialisées dans des processus analytiques puisque les bases de données ne supportaient majoritairement que ce type de

données. Hadoop offre de nouvelles possibilités, notamment dans le traitement de données non structurées, documents, articles, médias, qui via l'utilisation de capacités analytiques permettent de dégager des tendances, comme par exemple l'analyse de sentiment. Mais là encore, pour citer Aristote « la richesse consiste bien plus dans l'usage qu'on en fait que dans la possession.». Il convient donc de définir quelles données, structurées ou non, pourraient être utilisées dans le cadre de la mise en œuvre d'un projet analytique, et que les opérations d'acquisition et de traitement de ces données soient maîtrisées. Les capacités stockage offertes par les technologies Hadoop ne doivent pas conduire à une accumulation de données inutiles.

• **A quelle étape du projet aborder le Data Management ?**

Avant le projet, au démarrage, pendant, à la fin et encore après ! C'est une activité constante qui suit l'évolution des besoins et des usages. La gouvernance des données, c'est-à-dire la définition des règles d'accès, de traitement, de contrôle, et d'organisation de l'information n'est pas figée dans le temps et doit accompagner l'organisation dans la mise en place de nouveaux projets.

• **Quelle réponse SAS apporte à cette problématique ?**

Hadoop est un axe majeur de développement pour SAS. Des partenariats stratégiques ont été conclus avec les acteurs principaux du marché : Cloudera, Hortonworks, MapR.

En termes de fonctionnalités, c'est la capacité pour SAS de fournir des solutions permettant d'accéder aux environnements Hadoop, mais également d'exécuter des opérations analytiques et de gestion des données directement dans ces environnements.

La plateforme SAS intègre donc par défaut des connecteurs pour Apache Hive et Cloudera Impala permettant d'accéder à des environnements Hadoop. Par ailleurs, en 2015, SAS présente SAS® Data Loader for Hadoop, une nouvelle solution de « Self-Service Data Management », qui a pour objectif de vulgariser l'accès, la manipulation et la mise en qualité des données au sein du monde Hadoop. Les utilisateurs métier pourront désormais effectuer, en toute simplicité, des actions de gestion de données au sein d'environnements Hadoop, sans avoir à maîtriser les aspects techniques de ces plateformes.

En savoir plus sur SAS Data Loader for Hadoop

VINCENT REJANY
SWE Business Solution Manager – Information Management

Ryan Schmiedl, Senior Director of Product Management chez SAS, détaille les apports de la solution Data Management de SAS à l'environnement Hadoop.

Analytique et Hadoop

Tant de données.
En si peu de temps.
C'est si facile.

A quoi bon disposer d'immenses volumes de données si vous ne pouvez pas vous en servir ? Ou si leur analyse doit prendre des jours ou des semaines ?
Combiner le pouvoir analytique de SAS avec Hadoop, dans un environnement unique et interactif, vous apporte les décisions sur un plateau – pour des résultats rapides et une valeur décuplée.

Découvrir les solutions SAS pour Hadoop

Scannez le QR code avec votre smartphone.
Nécessite un lecteur de QR code.

sas
THE POWER TO KNOW®

SAS et tout autre nom de produit et de service de SAS Institute Inc. sont des marques déposées de SAS Institute Inc. pour les USA et tous les autres pays. ® indique un dépôt aux USA. Les autres marques et noms de produits sont des marques déposées de leurs entreprises respectives. Copyright © 2014 SAS Institute Inc. Tous droits réservés.

Comment utiliser une plateforme Hadoop sans avoir à jeter ses entrepôts de données ?

Trois approches pour mettre en place une plateforme Hadoop sans supprimer votre datawarehouse

Attirées par les promesses d'Hadoop, les entreprises cherchent à déterminer comment une plateforme de données Hadoop pourrait cohabiter avec leurs architectures IT voire complètement les remplacer. Ces promesses reposent sur deux arguments clés : les coûts et l'agilité.

Economique à de nombreux égards

L'efficience économique est l'argument le plus souvent avancé en faveur de l'adoption d'une plateforme de données Hadoop. Dans une telle approche, les coûts logiciels et le support, par nœud ou par CPU, et matériels sont beaucoup moins élevés que ceux de n'importe quelle base de données du marché. Autre bénéfice conséquent, il est possible d'accroître l'espace disque, le nombre de cpu ou la mémoire sur un nœud ou un cluster sans impact conséquent sur le coût de licences ou sur les mises à jour incrémentales.

Les entreprises espèrent limiter la croissance des coûts considérables induite par l'expansion exponentielle de leurs bases de données en adoptant une plateforme de données Hadoop ou en supprimant les bases de données qui ne sont plus absolument nécessaires.

Hadoop est aujourd'hui considéré comme la solution la plus économique pour répondre à l'explosion du volume de données et à aux impacts de l'internet des objets sur ce dernier.

Néanmoins, toutes les entreprises ne sont pas focalisées sur l'avenir et considèrent surtout Hadoop comme un levier de réduction de coûts, leur permettant de gérer leurs problématiques de volumétrie plus efficacement grâce à l'agilité qu'apporte Hadoop. Le chargement des données dans Hadoop est indépendant de leur format ou de

leur structure et on peut accéder ou modéliser les données ou des flux pour les utilisateurs instantanément comme à la lecture d'un schéma.

Une plus grande agilité

Deux grands axes ressortent :

1. Hadoop représente l'opportunité d'utiliser les larges quantités de données non-structurées et semi-structurées qui ne l'étaient pas jusqu'à maintenant pour des raisons économiques.

2. Grâce aux schémas en consommation ou en usage plutôt qu'en conception on peut définir les dimensions et modèles de données à l'utilisation, lorsque les problématiques métier surviennent, au lieu de les définir et de les résoudre en amont. Il n'est plus nécessaire de mettre à jour un énorme modèle de données central ni de construire les flux l'alimentant avant que le besoin métier n'émerge. De plus, les métiers ont le plus souvent besoin de modèles spécifiques. Cette fonction permet donc d'ajouter des données et de les exposer aux métiers sans trop d'effort et dans un délai court.

Cette agilité apportée aux métiers constitue comme la première motivation à implémenter Hadoop dans nombre d'entreprises.

Trois modèles de plateformes de données Hadoop

Bien que chaque entreprise soit spécifique, trois scénarios communs de mise en œuvre d'Hadoop se dégagent. La plateforme de données Hadoop est utilisée en parallèle des plateformes traditionnelles qu'elles utilisent aujourd'hui.

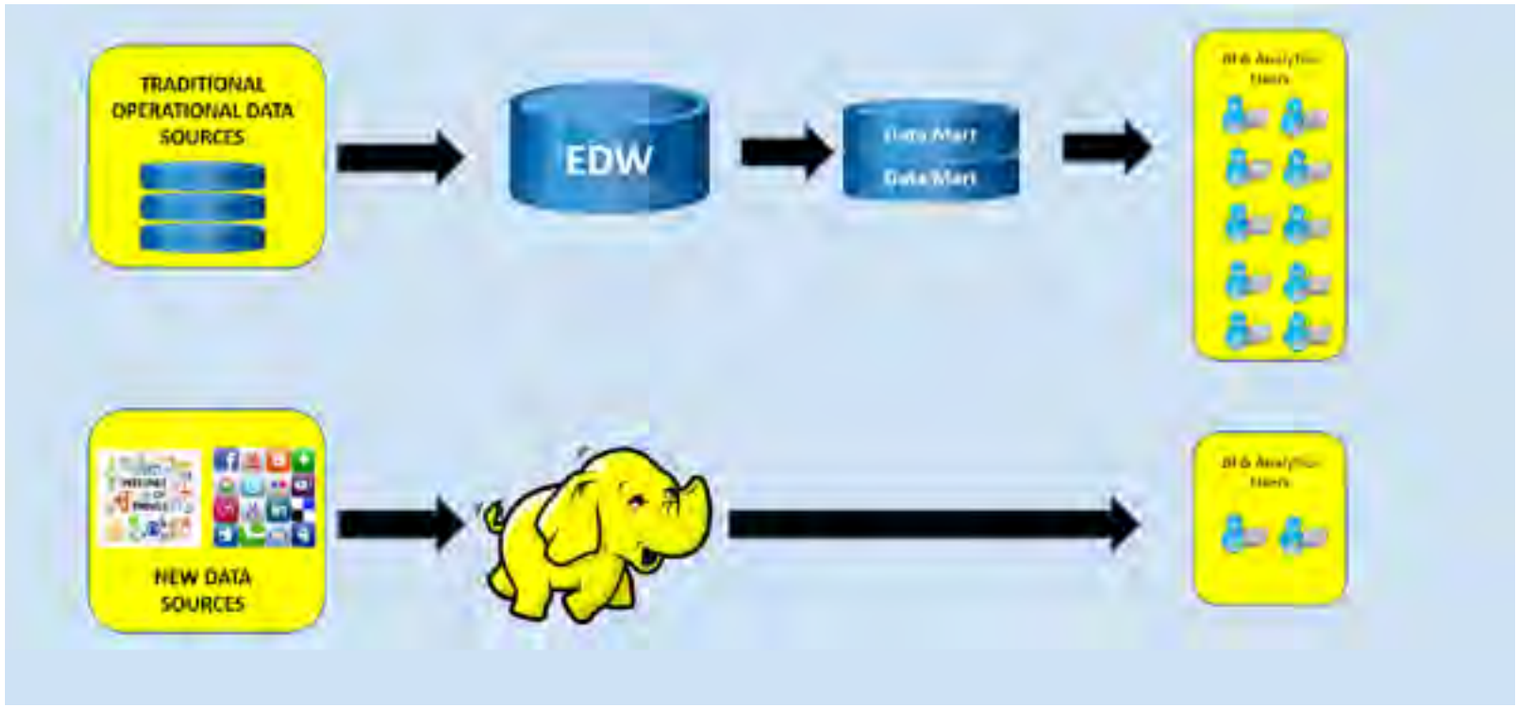
1. Hadoop en tant que nouvel entrepôt de données.

2. Hadoop en tant que plateforme de données alimentant le datawarehouse de l'entreprise.

3. Hadoop en tant que plateforme de données servant de base à la BI et à l'analytique.

Observons chacun de ces cas et comment les équipes IT pourraient évaluer leur pertinence.

Hadoop en tant que nouvel entrepôt de données

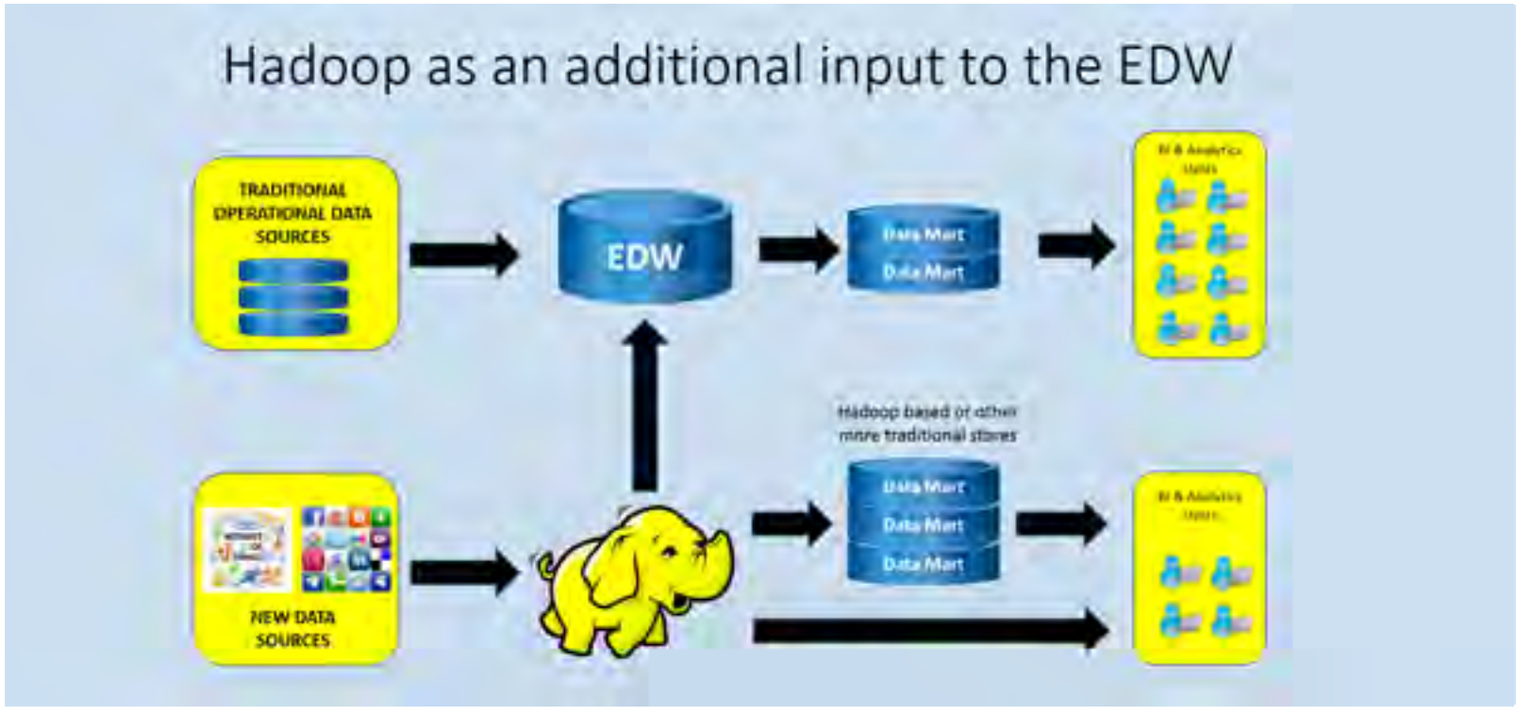


Dans ce scénario, Hadoop est utilisé pour gérer de nouvelles formes de données qui ne sont pas prises en charge par le datawarehouse de l'entreprise aujourd'hui. C'est le cas des données non-structurées et semi-structurées ou des données qui ne sont pas reconnues comme immédiatement utiles. Cette approche permet d'extraire de la valeur de données jusqu'alors inexploitées à faible coût et sans impacter le datawarehouse.

Hadoop est ici un soutien aux stratégies métier innovantes devant faire appel à de nouvelles données ou le moyen le plus économique d'intégrer les données non-structurées et semi-structurées dans un espace déterminé.

Dans ce type d'approche, il y a souvent peu d'utilisateurs Hadoop.

Hadoop en tant que plateforme de données alimentant le datawarehouse de l'entreprise



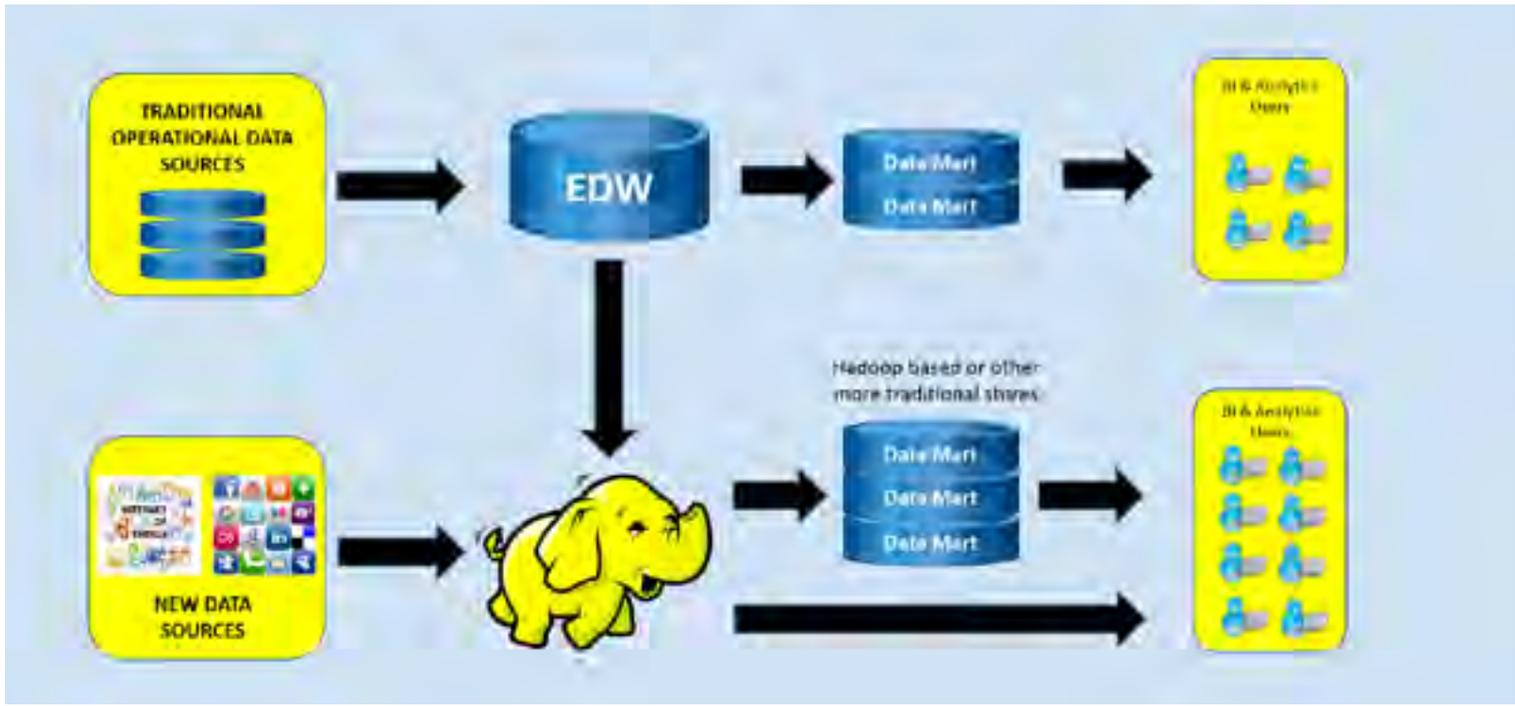
Dans cette approche Hadoop est aussi utilisé pour gérer, à faible coût, de nouvelles formes de données, mais les connaissances qui en sont retirées alimentent le datawarehouse de l'entreprise pour être consommées par de nombreux utilisateurs. Hadoop complète la stratégie de données et le datawarehouse demeure la seule source d'informations pour la majorité des utilisateurs. La conséquence de cette stratégie est que la taille du datawarehouse va croître sensiblement et que cela va induire une augmentation des coûts assez difficile à maîtriser.

Les informations utiles sont souvent découvertes grâce à la Business Intelligence et à l'analytique. De nombreuses entreprises y ont dédié un datamart structuré qui peut être stocké directement dans Hadoop (HIVE, Cloudera Impala, Pivotal HAWQ ou même des appliances/RDBMS/SAS).

Les données stockées dans Hadoop sont utilisées par un plus grand nombre d'utilisateurs que dans le premier scenario mais elles ne sont toujours pas consommées en masse.

L'idée, ici, est de limiter les coûts d'un datawarehouse en pleine expansion en attendant qu'une donnée s'avère véritablement utile pour tous avant de la charger dans le datawarehouse. Ce nouvel environnement permet aussi de concevoir, à moindre coût, les stratégies métier innovantes, qui requièrent souvent de gros volumes et une grande variété de données, et de ne mettre à jour le datawarehouse qu'avec celles qui s'avèrent pertinentes sans perturber ce qui est en place.

Hadoop en tant que plateforme de données servant de base à la BI et à l'analytique.



Dans ce scénario Hadoop est l'entrepôt principal pour tout ce qui touche à la Business intelligence et à l'analytique. Les processus touchant au datawarehouse d'entreprise ne sont quasiment pas touchés mais de nouveaux flux sont ajoutés afin de copier beaucoup d'information contenue dans le dataware vers Hadoop et seront complétées par de nouvelles données qui, elles, ne seront jamais dans le datawarehouse . Cela pourra permettre de réduire la taille du datawarehouse, ou du moins de limiter sa croissance.

Ainsi, le dataware peut continuer à supporter les fonctions importantes, telles que le reporting réglementaire et les systèmes en place s'appuyant sur le datawarehouse pourront continuer à fonctionner sans changement.

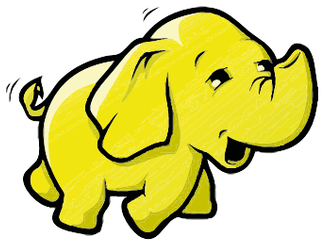
La plupart des tâches relatives à la BI et à l'analytique s'exécuteront directement sur l'entrepôt Hadoop qui contiendra des informations plus complètes et détaillées (ex : données transactionnelles), provenant de multiples systèmes, et un historique bien plus étendu.

Les rapports pourront ainsi s'exécuter sur des volumes de données élargis et les modèles analytiques à un niveau de granularité plus fin et sur un historique étendu en vue d'être mis en œuvre au niveau opérationnel en aval.

Dans cette configuration, Hadoop devient la source centrale pour la plupart des utilisateurs.

Source: SAS Blog, Mark Torr, Senior Director, Analytical Platform

Télécharger l'article complet ici



Pour en savoir plus, télécharger le livre blanc «Bringing the Power of SAS® to Hadoop»

Des partenariats experts pour une parfaite intégration de vos big data analytics dans Hadoop

SAS & Cloudera

SAS and Cloudera enable customers to achieve competitive advantage by gaining value from all of their data, through a proven combination of enterprise-ready storage, processing, analytics and management for Big Data.

Why Use SAS and Cloudera?

By making all of your data usable, not just what's in your databases, SAS and Cloudera let you see relationships that were hidden before and reveal answers that have always been just out of reach. You can start making more decisions based on hard data instead of hunches and look at complete data sets, not just samples.

The partnership of SAS® and Cloudera® offers enterprises compelling solutions to address the challenges of Big Data and enterprise data management and analytics.

With SAS and Cloudera, enterprises can run market leading analytical systems such as SAS® Visual Analytics or SAS® High-Performance Analytics Server natively on their Hadoop cluster, or leverage SAS/Access® to connect their Cloudera-powered Big Data sources to SAS applications.

Personne ne
connaît Hadoop
aussi bien
que Cloudera
cloudera

Plateforme Enterprise Data Hub:

La clé du succès pour les entreprises axées sur l'information

La gestion unifiée des données a longtemps été un objectif partagé par les départements métiers et informatique (IT). Cet objectif était jusqu'alors inaccessible en raison des limites et du coût inhérents aux approches traditionnelles. Pendant des décennies, les entrepôts et magasins de données basés sur la technologie de bases de données relationnelles ont constitué la seule option pour le déploiement de solutions analytiques, et les archives et réseaux de stockage les seules possibilités pour conserver des données à long terme.

Avec Cloudera Enterprise, les entreprises peuvent désormais maîtriser facilement l'augmentation rapide du volume et de la variété de données à laquelle elles sont confrontées, traitant une part croissante des données et des charges de travail au moyen de l'infrastructure existante tout en optimisant l'efficacité de ces systèmes. Basée sur le framework Apache Hadoop, la plateforme Cloudera Enterprise étend les investissements existants à un environnement central à la fois évolutif, flexible et sécurisé, permettant d'exploiter toutes les données sans restriction.

Une plateforme EDH (Enterprise Data Hub) est un nouvel environnement extrêmement puissant. À l'avenir, la majorité des données d'entreprise atterriront d'abord dans une plateforme EDH et elles y resteront de plus en plus. À court terme, les plateformes EDH apportent une flexibilité sans précédent pour analyser et traiter les données de manière approfondie et économique selon de nouvelles approches. Les entreprises qui déploient une plateforme EDH au côté de leur infrastructure existante se doteront des meilleurs atouts pour tirer leur épingle du jeu dans le monde des données modernes.

Cloudera sur le Web

Lire nos blogs : http://www.cloudera.com/blog et http://vision.cloudera.com

Sur Twitter, @clouderafr : http://twitter.com/cloudera

Sur Facebook : http://www.facebook.com/cloudera

SAS & Hortonworks

With the release of Apache Hadoop YARN in October of last year, more and more solution providers are moving from single-application Hadoop clusters to a versatile, integrated Hadoop 2 data platform. This allows them to host multiple applications — eliminating silos, maximizing resources and bringing true multi-workload capabilities to Hadoop.

That is why we're extremely excited to have **Paul Kent**, Vice President of Big Data at SAS, share his insights on the value of Apache Hadoop YARN and the benefits it brings to SAS and its users.

Here's Paul…

As much as it could be a refrain from the Montessori school playground, this theme of "Share your Cluster" echoes across many modern Apache Hadoop deployments.

Data Architects are plotting to assemble all their data in one system – something that is now achievable thanks to the economics of modern Apache Hadoop systems. Once assembled, this collection of data now has sufficient gravity to attract the application processing towards it – folks are becoming intolerant of the idea that we should make another copy (and have to reconcile, secure and govern that copy) to facilitate processing.

Move the Work to the Data

One of the original themes in Hadoop is to move the work to the data. MapReduce is a beautiful expression of this paradigm – the work is expressed in Java classes, and these classes are literally copied and executed across all the data nodes that participate in processing the data.

Apache Hadoop YARN facilitates this idea at the next level. YARN allows for users to move different kinds of computation to the data while sharing both the data and the resources of the cluster – which enables other newer applications to run alongside traditional MapReduce workloads.

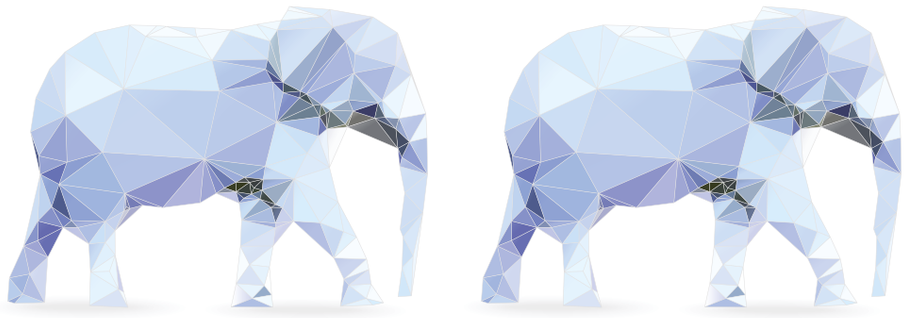
The SAS Approach to Hadoop YARN Integration

The **SAS High-Performance Analytics** (HPA) products and SAS LASR Analytic Server based products have their roots in classical High Performance Computing. At their heart, both began their life as a traditional MPI application. The “mpirun” command launches an instance of the application on each host in the host list. These instances are bound together by infrastructure so that they can communicate easily with one another. SSH is used to create processes on the cluster, and as a practical matter, we use password-free SSH or Kerberos keys to save from having to repeat passwords over and over.

For SAS workloads, the ability for worker tasks to communicate with each other to solve the problem team-style is preferred to the MapReduce model where each worker is assigned a slice of the dataset and must process that slice in relative isolation, handing off the results to a downstream task without learning anything from other workers doing the same processing on their slice.

Adopting YARN allows us to use the YARN infrastructure to set the boundaries for the processes needed to run SAS HPA products and SAS LASR Analytic Server based products. CPU and memory can be capped, facilitating a better sharing model for the cluster. Our early adoption and expanded integration with YARN going forward positions SAS as a good citizen at the center of shared Hadoop clusters. Our engineers have been working with Hortonworks engineers on the optimal way to integrate SAS HPA and SAS LASR Analytic Server with Hadoop YARN and the result of these efforts means that customers will get the maximum benefits from their Hadoop clusters.

This is an exciting development for us. SAS applications bring highly advanced, in-memory analytic processing to the data in Hadoop and enable a rich set of additional use cases with high performance analytic needs. Together, our combined technologies offer customers more flexibility to choose best of breed SAS HPA and LASR analytic applications in conjunction with their trusted Hadoop workloads as they build and deploy bigger Hadoop clusters



Comment mettre en place une stratégie Big Data avec Hadoop

Hadoop n’est pas une stragégie big data, c’est un élément du puzzle

Il y a encore peu, les organisations ne pouvaient stocker que des quantités limitées de données. Les analystes étaient alors contraints à analyser des échantillons d’un sous-ensemble de données dans le but de simuler une plus grande population de données, alors qu’en utilisant l’ensemble des données le résultat aurait été plus précis.

Hadoop permet de surmonter les problèmes de volumétrie et de coordination associés au traitement de milliards de données qui, auparavant, n’auraient pas pu être stockées. SAS permet de réaliser des analyses à la mesure des possibilités offertes par Hadoop.

Sans entrer dans des détails techniques, on peut résumer l’apport de SAS dans un environnement Hadoop en trois points :

- SAS utilise les données provenant de Hadoop comme de n’importe quelle autre source de données. SAS Data Integration supporte Hadoop comme toute autre plateforme de stockage de données. Ses outils graphiques permettent aux utilisateurs d'accéder, de traiter et gérer les données et les processus Hadoop à partir de l'environnement SAS habituel. Ceci est essentiel, compte tenu de la pénurie de compétences familiarisées avec Hadoop.
- La puissance de SAS Analytics associée à Hadoop permet de garantir aux organisations et entreprises utilisatrices une réponse à leurs attentes en termes de précision, de productivité et d’évolutivité.
- SAS apporte la gouvernance nécessaire à l’environnement Hadoop. SAS propose une approche robuste pour le cycle de gestion de l'information sur Hadoop qui inclut la gestion de données et des process analytiques. C’est un énorme avantage comparé à d'autres produits du marché qui se concentrent sur le transfert des données vers et hors d’Hadoop.

Pour en savoir plus, télécharger le livre blanc ici

Retour sommaire

Fast and Furious: quand les Big Data Analytics rencontrent Hadoop

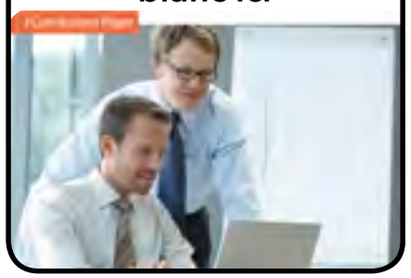
Vous l'aurez compris au gré de votre lecture, SAS et Hadoop sont des compléments naturels. Hadoop offre un stockage distribué et une puissance de traitement. SAS supporte Hadoop comme toute autre source de données, lui apporte ses technologies de data management, d'exploration de données et d'analyses avancées. Leur intégration mutuelle vous fournit des bénéfices majeurs pour extraire un maximum de valeur de vos big data :

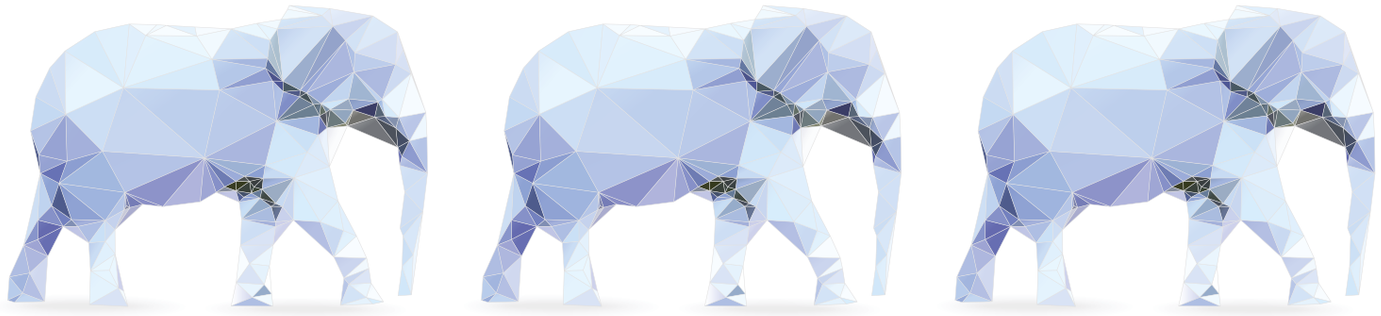
- Précision. Appliquez des algorithmes analytiques avancés et des techniques de machine-learning SAS aux données sous Hadoop et bénéficiez des meilleurs résultats.
- Evolutivité. Obtenez une réponse à vos problématiques métiers, à l'échelle de votre département comme de votre entreprise, avec la plus grande fiabilité et en un temps record réduit à quelques secondes, et non plus en heures ou en jours.
- Productivité. Plusieurs utilisateurs peuvent simultanément, rapidement et de manière interactive analyser les données dans Hadoop en utilisant le langage de programmation analytique en mémoire.

Vous pouvez travailler avec des données Hadoop à partir de l'environnement familier et interactif de SAS Analytics soit avec l'interface graphique de SAS Visual Statistics ou via la programmation simplifiée de SAS in-memory Statistics for Hadoop.

Ainsi, vos processus de traitement tourneront à plein régime, permettant des analyses toujours plus étendues et plus approfondies - et vous libérerez des ressources pour d'autres requêtes sous Hadoop.

Pour en savoir plus, télécharger le livre blanc ici





Maintenant que vous savez nager, plongez ! Quelques applications

Des cas d'utilisation variés - Des besoins analytiques nouveaux

Comme on l'a vu au long de ces pages, Hadoop offre des perspectives nouvelles aux entreprises, quels que soient les secteurs ou les métiers. Ce tableau non exhaustif liste un certain nombre d'applications possibles.

Secteur	Cas d'utilisation	Type de données
Financial Services	*New Account Risk Screens*	*Text, Server Logs*
	Fraud Prevention	*Server Logs*
	Trading Risk	*Server Logs*
	Maximize Deposit Spread	*Text, Server Logs*
	Insurance Underwriting	*Geographic, Sensor, Text*
	Accelerate Loan Processing	*Text*
Telecom	*Call Detail Records (CDRs)*	*Machine, Geographic*
	Infrastructure Investment	*Machine, Server Logs*
	Next Product to Buy (NPTB)	*Clickstream*
	Real-time Bandwidth Allocation	*Server Logs, Text, Sentiment*
	New Product Development	*Machine, Geographic*
Retail	*360° View of the Customer*	*Clickstream, Text*
	Analyze Brand Sentiment	*Sentiment*
	Localized, Personalized Promotions	*Geographic*
	Website Optimization	*Clickstream*
	Optimal Store Layout	*Sensor*
Manufacturing	*Supply Chain and Logistics*	*Sensor*
	Assembly Line Quality Assurance	*Sensor*
	Proactive Maintenance	*Machine*
	Crowdsourced Quality Assurance	*Sentiment*

Industrie électrique : dompter les big data avec Hadoop

Avec le développement des smart grids, les fournisseurs d'électricité se retrouvent noyés sous les données issues des compteurs intelligents. Une solution existe aujourd'hui pour traiter, analyser et tirer de la valeur de cette avalanche d'informations : Hadoop. Kathy Ball, consultante énergie chez SAS, fait le point sur cette technologie dans un article publié sur le site Intelligent Utility.

Avec Hadoop, rien n'est trop grand. Pourtant, en étant avisé que le nombre d'objets connectés à Internet devrait avoisiner les 50 milliards en 2020, on prend assez vite la mesure des enjeux qui attendent l'industrie électrique ces cinq prochaines années. Pour la consultante de SAS, les entreprises du secteur capables de transformer en ressources stratégiques les données générées par ces objets de tous les jours, disposeront d'un avantage compétitif sur plusieurs plans : rentabilité du réseau, amélioration du service client ou encore réputation. C'est à ce niveau qu'intervient Hadoop.

Hadoop est une plate-forme open source qui permet de stocker et de traiter de manière efficace les big data, en répartissant les tâches de calcul sur des machines interconnectées. Les fournisseurs d'électricité, par exemple, peuvent exploiter la masse de données issue des compteurs intelligents installés sur le réseau, en association avec des commentaires laissés par les clients, pour détecter des tendances et faire des recommandations aux usagers. De quoi optimiser l'utilisation du réseau électrique et améliorer la relation client. Pour illustrer son propos, Kathy Ball cite l'exemple d'un électricien qui a utilisé le potentiel analytique de Hadoop pour segmenter sa base client et identifier ceux qui pourraient participer aux programmes dits de réponse à la demande, dont le but est d'adapter la consommation d'électricité pour éviter les pics de production.

Lire l'article

Hadoop : colonne vertébrale d'American Express

American Express est un mastodonte de la finance : 107,2 millions de cartes de crédit actives à travers le monde, pour un total mondial réglé de 952,4 milliards de dollars (cartes privées et d'entreprises) l'année dernière. Pour conserver cette belle situation, l'entreprise a initié une profonde transformation de son infrastructure technique, étayée par la mise en œuvre d'Hadoop qui l'aide à proposer de nouveaux services innovants. Retour sur la transformation numérique d'Amex.

Présente dans 130 pays, l'entreprise compte près de 63 000 employés, a généré 33 milliards de dollars de revenus et 5,4 milliards de dollars de revenu net en 2013. Mais American Express ne fait pas qu'offrir des cartes de paiement et proposer des services aux commerçants, l'entreprise fournit également les réseaux qui hébergent et traitent les transactions entre acheteurs et vendeurs. Sachant qu'elle touche un pourcentage sur chaque transaction, elle propose donc des informations, applications et offres spéciales aux détenteurs de cartes comme aux commerçants pour augmenter le nombre de transactions, et assurer ainsi sa croissance.

L'analyse de données est donc au cœur des activités d'American Express. En plus des informations provenant de son propre réseau, l'entreprise a noué des partenariats avec les opérateurs d'autres systèmes de paiement ou de fidélisation, par exemple avec le groupe Metro en Allemagne, avec Walmart aux Etats-Unis et bientôt avec McDonalds.

Pour collecter et tirer pleinement parti de ces précieuses informations, American Express a créé des applications sociales et mobiles basées sur la technologie Hadoop. Destinées aux particuliers et aux professionnels, elles proposent des réductions en temps réel selon la situation géographique, les préférences et les achats précédents,

des analyses de tendance trimestrielles, ainsi que des analyses comparatives et concurrentielles. Ces applications reposent sur un cluster Hadoop et visent à rassembler l'équivalent de dix ans de données pour, entre autres, assurer la conformité avec la réglementation Bâle III. La R&D dispose d'un second cluster Hadoop et un troisième est utilisé pour l'analytique : développement de modèles et analyses ponctuelles.

L'objectif d'American Express est de fidéliser ces deux populations, en leur offrant une valeur tangible : 100 millions de dollars ont été offerts sous la forme de bons de réduction aux titulaires de cartes, et les services offerts aux commerçants contribuent à fidéliser leurs clients.

Lire l'article

Comment Hadoop permet aux Etats-Unis d'accroître leur indépendance énergétique

Si les big data sont le nouveau pétrole de l'économie mondiale, l'industrie pétrolière elle-même surfe sur la vague pour trouver de nouveaux relais de croissance. L'Agence Internationale de l'Energie annonce qu'en termes de production de pétrole, les Etats-Unis dépasseront l'Arabie Saoudite et la Russie d'ici à 2016. Dans les domaines de la prospection, de l'exploitation, de la maintenance et de la sécurité, les technologies de l'information tiennent une grande place parmi les outils utilisés, notamment Hadoop qui facilite l'analyse de grands volumes de données.

Au cours de ses dernières années, la production de pétrole a beaucoup augmenté aux Etats-Unis. C'est une bonne nouvelle pour l'autonomie énergétique du pays mais aussi au niveau international : la production mondiale a cessé d'augmenter, c'est donc une nouvelle offre qui voit le jour pour répondre à la demande mondiale, en croissance constante. Côté outils, **Hadoop**, qui est une infrastructure de traitement massif de données, est au cœur de l'organisation et de la rationalisation de la production et du raffinage du pétrole et du gaz. En effet, les nombreux capteurs installés sur des équipements à la pointe de la technologie collectent de grands volumes de données qui sont analysées à des fins d'optimisation et de rationalisation.

Hadoop permet d'optimiser la découverte et l'exploration des gisements : les capteurs sismiques génèrent des données dont l'analyse permet de déterminer les emplacements et méthodes optimaux pour la production, tout en améliorant la rentabilité et la sécurité. De même, on peut évaluer le volume de pétrole disponible et mettre ainsi en place des méthodes d'extraction plus durables. Parce que les puits de pétrole évoluent tout au long de l'exploitation, il faut également surveiller l'évolution des réserves pour s'assurer du succès de l'exploitation.

L’infrastructure d’architecture distribuée d’Hadoop permet également d’améliorer la sécurité de l’exploitation. La maintenance des équipements est plus précise grâce à des capteurs dont les données sont analysées pour détecter les erreurs, les pannes mécaniques ou les dangers potentiels. Identifier les problèmes potentiels permet de les anticiper et améliore la sécurité de l’installation dans son ensemble, pour le personnel et pour l’environnement, en réduisant les risques de blessures, d’incendies, d’explosion, d’effondrement, etc. Enfin, au-delà des risques physiques, Hadoop permet de détecter, de contrer voire d’anticiper les vols ou altérations de données, l’implantation de virus ou encore les failles de sécurité.

Lire l’article

Santé : prendre le virage de l’analytique pour éviter le chaos

La conclusion de ce rapport signé Ernst & Young (EY) et consacré aux sciences du vivant est plutôt radicale. Après avoir analysé cinq tendances majeures du secteur, le cabinet de conseil britannique estime que seules les entreprises de santé ayant préparé le déferlement des big data et de son corollaire analytique seront en mesure d’affronter les nouveaux défis et de saisir les opportunités qui placent le patient au centre de leur activité.

Nous sommes en 2015. Comme des millions de personnes, Julie transmet pression sanguine et glycémie via sa montre connectée sur une plate-forme d’analyse spécialisée. Celle-ci lui prédit un risque élevé d’arrêt cardiaque dans les trois mois et planifie d’urgence un rendez-vous chez le cardiologue. Après une série de tests, ce dernier constate la formation d’une plaque dans les artères coronaires de sa patiente et lui prescrit un traitement adapté à ses marqueurs génétiques. En adaptant son mode de vie et son régime alimentaire, Julie a probablement évité le pire.

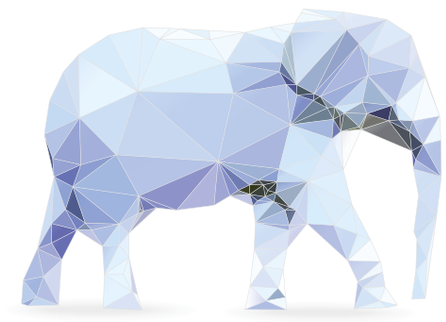
Cette histoire fictive et néanmoins réaliste concentre tous les ingrédients de la révolution en cours dans le domaine des sciences du vivant. Selon EY, l’avènement de cette nouvelle ère est imminent — deux à cinq ans tout au plus : les technologies permettant le prélèvement, la transmission et le traitement des données personnelles de santé arrivent à maturité ; la médecine personnalisée a fortement progressé avec la génomique ; les outils analytiques sont désormais à la portée de tous ; l’usage du cloud computing et de frameworks comme Hadoop s’accélère. Last but not least, les institutions de santé misent sur la prévention pour réduire les coûts de l’assurance maladie.

Pour mettre en œuvre les projets de développement basés sur l'analytique et atteindre un succès durable, EY propose une démarche qui positionne les data au cœur de l'entreprise et conseille aux organisations de conclure des partenariats avec des experts.

Lire l'article

Ces 3 articles sont extraits du site **Business Analytics Info**.

Conclusion



La question n’est plus de savoir si Hadoop trouvera une place dans votre organisation, mais plutôt quand il trouvera sa place, et dans quel rôle.

Hadoop pourra vous aider à économiser sur vos coûts de stockage, mais aussi bien apporter de la valeur comme élément crucial dans votre architecture analytique du futur. Hadoop vous aidera à mieux répondre aux besoins analytiques de vos utilisateurs. Peu importe le cas, Hadoop a prouvé sa valeur et vient de passer le stade du « early adopter ».

Si vous ne savez pas encore trop bien par où commencer, ou si vous avez déjà fait vos premiers pas dans le monde Hadoop, les solutions SAS vous permettront d’accéder à toute la valeur de vos projets. D’une part en baissant le niveau d’entrée dans Hadoop pour rendre son environnement plus accessible au grand public, et d’autre part en offrant une architecture de traitement analytique in-memory complémentaire à ce que peut offrir Hadoop.

MATHIAS COOPMANS
Business Development manager, SAS

SAS CEO Jim Goodnight discusses SAS on Hadoop with Chris Dingle from Rogers Communications at the SAS Global Forum 2014 Opening Session.

THE POWER TO KNOW®

Rejoignez SAS sur les réseaux sociaux et restez informés

SAS Institute s. a. s. – DOMAINE DE GREGY - GREGY-SUR-YERRES - 77257 BRIE COMTE ROBERT - FRANCE

TEL. : +33(0) 1 60 62 11 11 FAX : +33(0) 1 60 62 11 99

www.sas.com/france

SAS et tout autre nom de produit et de service de SAS Institute Inc. sont des marques déposées de SAS Institute Inc. pour les USA et tous les autres pays. ® indique un dépôt aux USA. Les autres marques et noms de produits sont des marques déposées de leurs entreprises respectives. Copyright © 2015 SAS Institute Inc. Tous droits réservés.

EBK01FCE0315

Retour sommaire