IT Business Review

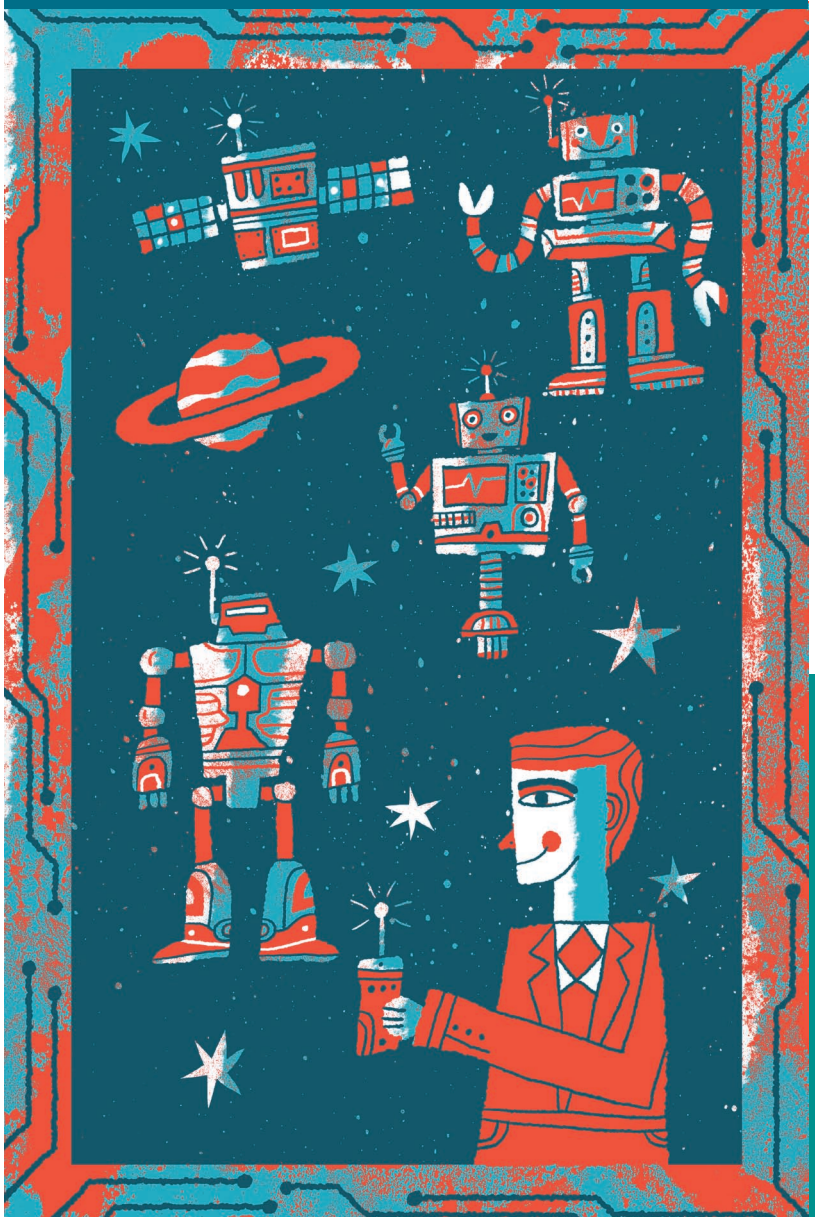
www.itbusinessreview.fr

N° 7 - Juin 2014

N° 7

Comment Europ Assistance surveille l'état de santé de son SI

page 18



MANAGEMENT
- Les modèles d'organisation de la DSI

RELATIONS FOURNISSEURS
- Comment se débarrasser des consultants qui s'incrustent

INFRASTRUCTURES
- De l'APM au maintien en conditions opérationnelles

REPÈRES
- Quand le big data impose plus de sécurité
- La maturité numérique des entreprises françaises
- Dématérialisation : la spirale déflationniste
- Gestion des données : quels bénéfices ?

SERVICES
- Les DSI ont-ils toutes les cartes ?

SÉCURITÉ
- Maîtrise des risques dans les projets

GESTION DE PROJET
- Comment rater son projet big data
- Comment rater ses applications métiers

Business Intelligence : comment intégrer Hadoop en douceur

page 12

Infrastructure décisionnelle : comment intégrer Hadoop en douceur

Par Aurélie Chandèze

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.

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

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 nœuds du *cluster*.

- ***Hadoop YARN*** 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. •

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

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 nœud 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 nœuds 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).

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 : ***http://lc.cx/ITBR7-1***

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, 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.

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. •

Liens utiles

- *A non-geek big data playbook: Hadoop and the Enterprise Data Warehouse*, de Tamara Dull. Lien vers le document : **http://lc.cx/ITBR7-1**
- *Trends in Enterprise Hadoop Deployments*, livre blanc d'IDC pour Red Hat. Lien vers le document : **http://lc.cx/ITBR7-2**
- *Eight considerations for utilizing big data analytics with Hadoop*, par Fern Halper, TDWI Research. Lien pour télécharger le document : **http://lc.cx/ITBR7-3**

Carte de visite

Mouloud Dey
Directeur Solutions et Marchés Émergents
SAS France
Domaine de Grégy - Grégy sur Yerres
77 257 Brie Comte Robert Cedex
e-mail : comsas@fra.sas.com
Tél. : 0820 22 11 11 (0,09 € ttc/min)
www.sas.com/france

Les motivations pour utiliser Hadoop

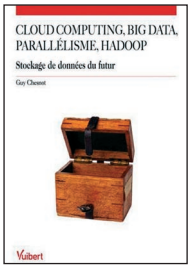
Analyse des données transactionnelles liées aux ventes et aux points de vente	**44 %**
Analyse des comportements des consommateurs en ligne	**38 %**
Développer de nouveaux services	**38 %**
Analyse des données liées aux opérations	**37 %**
Analyse des données issues de terminaux ou de machines	**32 %**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

Source : IDC. Lien : http://lc.cx/ITBR7-2.

Pour en savoir plus sur Hadoop

Par Aurélie Chandèze

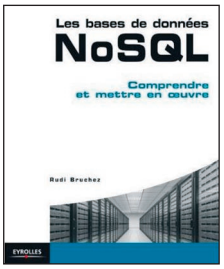
Une affaire de stockage



Cloud Computing, big data, parallélisme, Hadoop - Stockage de données du futur, de Guy Chesnot, préface de Yann Serra, Vuibert, 2012, 224 pages.

Ecrit par un architecte spécialiste du stockage et travaillant chez Silicon Graphics, ce livre a pour sujet central le stockage des données. Il dresse un panorama des différentes solutions et technologies existantes pour le stockage des *big data*, en étudiant notamment les différents systèmes de fichiers. Hadoop et son système de fichiers HDFS sont donc abordés dans un contexte d'usage bien spécifique, celui du stockage, et une vingtaine de pages seulement y sont consacrées. Cet ouvrage s'adresse donc plutôt à des architectes ou décideurs techniques en phase de réflexion sur leurs infrastructures de stockage.

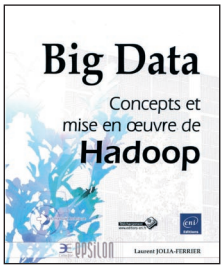
Une affaire de base de données



Les bases de données NoSQL : comprendre et mettre en œuvre, de Rudi Bruchez, Eyrolles, 2013, 279 pages.

Consultant indépendant, expert en bases de données, Rudi Bruchez s'est penché dans ce livre sur les différentes solutions de stockage de données non-relationnelles qui ont émergé au cours de la dernière décennie chez les grands acteurs américains du Web, regroupées aujourd'hui sous le qualificatif « NoSQL ». L'ouvrage comporte notamment un panorama des principales solutions, permettant de situer Hadoop par rapport à d'autres technologies NoSQL comme HBase, CouchDB, MongoDB, Riak, Redis ou Cassandra. L'auteur fournit également des pistes de réflexion pour savoir quand utiliser ce type de solutions et laquelle choisir. A ce titre, il vise un public plutôt technique : experts en bases de données, architectes logiciels ou développeurs. Les chefs de projets pourront aussi y trouver des clés pour aborder les phases de choix techniques.

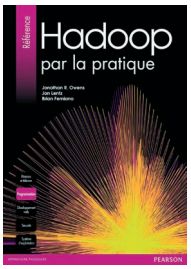
Une affaire de mise en œuvre



Big data - Concepts et mise en œuvre de Hadoop, de Laurent Jolia-Ferrier, ENI Editions, collection Epsilon, 2014, 207 pages.

Centré sur Hadoop, ce livre est écrit par un consultant indépendant, spécialisé dans le *big data* et certifié *Cloudera Hadoop Developer*. Il se présente comme un ouvrage introductif, destiné à un public non familier de Hadoop et des technologies associées visant à démystifier le *big data* et Hadoop. Les chapitres sont principalement consacrés aux premiers pas avec Hadoop : installation sur un poste de travail, aspects matériels, développement de programmes et mise en œuvre d'un *cluster*. Un chapitre est également consacré aux évolutions introduites par la version 2 d'Hadoop, disponible depuis octobre 2013. Pour les décideurs, le livre fournit également quelques éléments sur l'écosystème et les usages possibles de la technologie.

Une affaire de bonnes pratiques



Hadoop par la pratique, de Jonathan R. Owens, Jon Lentz et Brian Femiano, Pearson, 2014, 330 pages.

Le plus récent des quatre ouvrages, ce livre est écrit par trois experts techniques, utilisant régulièrement Hadoop dans leur environnement professionnel, et il se veut centré sur la pratique. De ce fait, alors que les ouvrages précédents sont davantage destinés à être lus en amont des projets, celui-ci intervient plutôt une fois le choix d'Hadoop effectué et les projets lancés. Il cible un public principalement technique : développeurs et administrateurs en phase d'apprentissage de Hadoop. Les différents chapitres traitent chacun d'un problème technique et proposent une solution agrémentée d'exemples de codes : extractions et manipulations de données, analyses d'agrégats, analyse graphique, apprentissage automatique, dépannage ou encore administration. •

IT Business Review