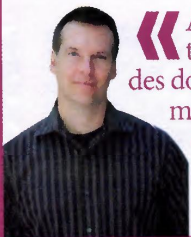




DOSSIER



« Apache Hadoop traitera la moitié des données créées dans le monde dans les cinq prochaines années. »

Eric Baldeschwieler, Hortonworks

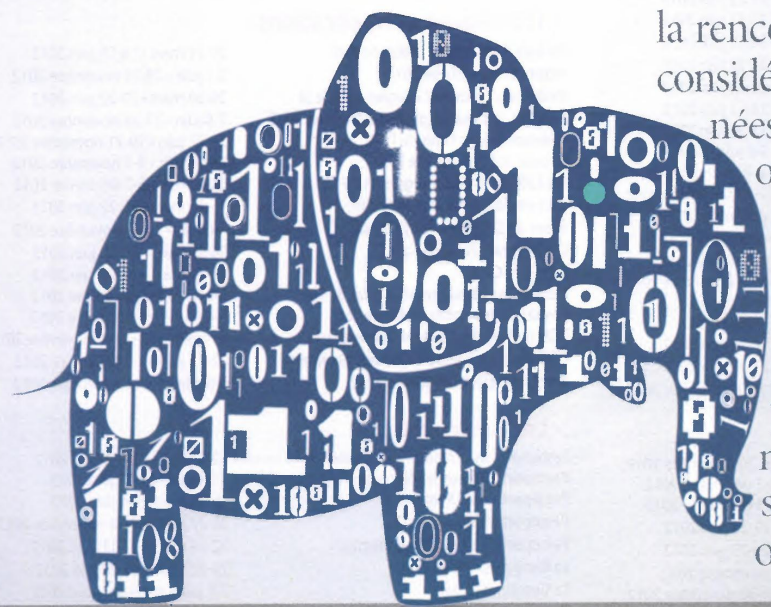
100 Md€ pourraient être économisés par les gouvernements européens s'ils utilisaient le big data pour optimiser leur fonctionnement.

Source: McKinsey Global Institute

À VENIR
La conférence **Strata à Santa Clara (Californie)** du 28 février au 1^{er} mars, puis le Congrès Big Data, les 20 et 21 mars à la Cité universitaire de Paris.

Big data

MIEUX CONNAÎTRE ET MIEUX EXPLOITER LES DONNÉES DE SON ENTREPRISE



Si le concept de big data suscite autant d'intérêt, c'est qu'il traduit la rencontre entre des besoins considérables d'analyse de données non structurées et une offre technologique mature et bon marché. Un nombre croissant de sociétés s'apprête ainsi à mieux analyser les comportements de consommateurs, les interactions sur les réseaux sociaux ou la navigation sur le web.



**DOSSIER RÉALISÉ PAR
VINCENT BERDOT ET ALAIN CLAPAUD**

À SAVOIR
Cloudera, Hortonworks et MapR, trois start up spécialisées dans la pile open source Hadoop, ont levé plus de 150 M\$ en deux ans.

83 % des entreprises interrogées dans le cadre d'une étude réalisée par EMC s'attendent à connaître une pénurie de spécialistes de l'exploration de données (Data Scientists).

« Avec le big data, une médecine s'appuyant sur des tests ADN et les variantes pharmacologiques pourrait être spécifique à chaque personne. »

Shahid N. Shah, expert américain en e-santé



Enjeux DIX ANS POUR EN ARRIVER LÀ...

Au-delà du buzz, le big data traduit un mouvement de fond amorcé depuis cinq à dix ans, tendant à dépasser les limites des bases de données traditionnelles.

Au cœur du système, Hadoop : un socle technologique open source, que les géants du web ont fait sortir de leurs laboratoires.

Les buzz words ont toujours existé et font partie intégrante du cycle de toute innovation. Mais avec celui-ci, on frise l'overdose. Le big data part du postulat que les entreprises sont submergées par des flots de données semi ou non structurées. Parmi les multiples exemples couramment avancés, citons les comportements d'achat sur le web, la nature des interactions sur un réseau social, les relevés de consommations d'énergie sur les compteurs intelligents ou encore les tickets d'ap-

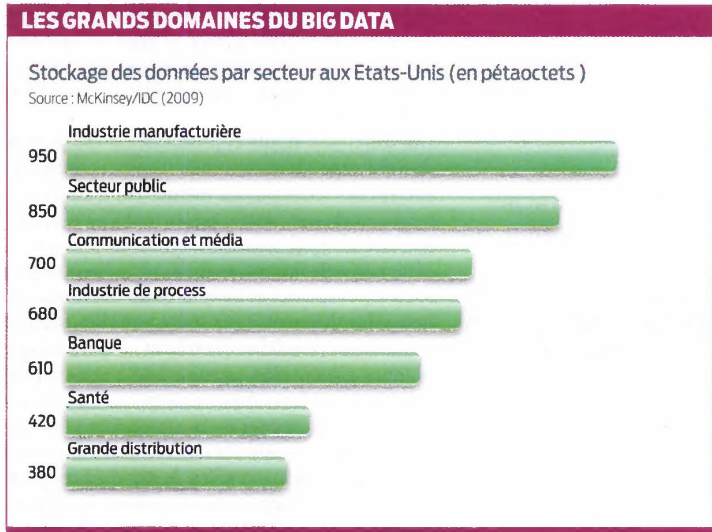
pels des opérateurs téléphoniques. La notion de big data recouvre à la fois la capture de ces données volumineuses et désordonnées, et, surtout, leur traitement. Au final, il promet de donner du sens à la montagne de précieuses informations sur laquelle sont assises les entreprises, et qu'elles délaissent, faute de moyens.

Méfiant, il faut l'être, car le message selon lequel les entreprises seront amenées à crouler sous des déluges de données est ressassé depuis des décennies. Sans que l'apocalypse annoncée ait eu lieu. Le big data ne serait-il alors qu'une simple resucée d'un message savamment entretenu par les fournisseurs de stockage, d'intégration de données, de datawarehouse et de décisionnel ?

Des besoins transactionnels, mais surtout analytiques

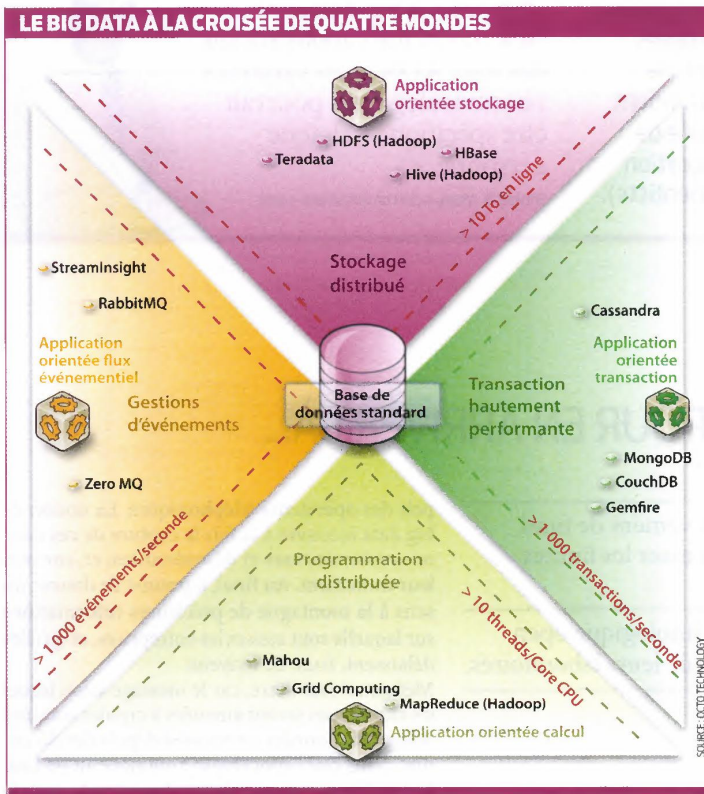
Fort heureusement non, pas uniquement. Au-delà du recyclage marketing, le big data traduit un mouvement de fond amorcé depuis cinq à dix ans : les bases de données traditionnelles, comprenez relationnelles, peinent de plus en plus à répondre à certains besoins transactionnels. « Elles échouent à monter en charge lorsque les connexions vers un site web sont simultanées et très nombreuses, explique Romain Chaumais, cofondateur de l'intégrateur Ysance. D'où l'émergence des technologies estampillées NoSQL. Les bases en mémoire, les bases documentaires ou encore celles orientées colonne ou clé-valeur ont progressivement tué le dogme des bases relationnelles. » Mais ce qui fait tant parler du big data, ce sont surtout les besoins analytiques : indexation, transformation, recherche, calcul, élaboration de modèles, exécution d'algorithmes, etc. Là encore, les bases classiques peinent à réaliser ces fonctions sur des données peu structurées, sans schéma clair, sans grande qualité et stockées en masse (au-delà de 10 téraoctets).

Et c'est là qu'intervient Hadoop, une pile open source précisément conçue pour s'atteler à ces





DOSSIER BIG DATA



open source Lucene, s'empare de ces travaux et réalise le premier prototype d'Hadoop. Devant le succès de Google, et sa capacité à « avaler » si facilement le web, Yahoo, qui a déjà perdu de sa splendeur, cherche à investir cette technologie. Il embauche Doug Cutting en 2006, et crée, l'année dernière, une filiale dédiée à Hadoop: Hortonworks. Depuis, d'autres géants du web, tels que Facebook ou LinkedIn, s'en sont inspirés.

Un système réservé aux grands ?

Reste à savoir pourquoi Hadoop explose aujourd'hui. Plusieurs raisons à cela. D'abord, la pile a gagné en stabilité – la version 1.0 a été lancée en décembre dernier – et en fonctionnalités. « Elle présente l'avantage de couvrir toute la chaîne: stockage, intégration, traitement et requêtage des données. Seule la partie liée à l'alimentation est encore à revoir », explique Julien Cabot, responsable de l'activité finance chez Octo Technology. Certes, certaines entreprises ou centres de recherche n'ont pas attendu Hadoop pour développer ou acquérir des systèmes de calcul haut volume. Mais le plus souvent pour des coûts élevés. Hadoop, lui, et c'est une autre raison de son succès potentiel, promet une équation économique imbattable. Parce qu'il repose sur l'open source, s'appuie sur du matériel banalisé et sur des traitements hautement distribués, il parvient à maintenir un rapport linéaire entre le volume des données à analyser, la valeur de ces analyses et les investissements nécessaires en matériel. « Avec les systèmes analytiques traditionnels, le coût de l'analyse croît de façon exponentielle par rapport à l'augmentation du volume des données », avance Jack Norris, vice-président marketing de MapR. Pour Octo Technology, Hadoop devrait même résoudre un problème de fond: « Sur les trente dernières années, l'informatisation des processus métier était le principal levier pour gagner en productivité. Or, aujourd'hui ces gains plafonnent. Pour dépasser cette limite, il faut revenir aux fondamentaux, en décortiquant ces processus. Et cela ne se fera qu'avec de l'analyse de données et du big data. »

Ces différents atouts et promesses d'Hadoop ne doivent cependant pas nous tromper: toutes les entreprises n'ont pas les mêmes contraintes que les géants du web. Attention surtout à ne pas céder aux sirènes inflationnistes, qui, sous prétexte de cette nouvelle manne analytique, inciteraient les sociétés à capturer et à explorer un maximum de nouvelles données, sans même qu'elles sachent exactement ce qu'elles cherchent... D'autant que nombre d'entre elles peinent déjà à ordonner et à uniformiser leurs données internes. Le big data ne ferait ici qu'ajouter de la complexité à la complexité. Enfin, en sortant de l'escarcelle des bases de données traditionnelles, le big data exige de nouvelles compétences. Notamment en programmation et en exploration de données. ■

enjeux. Elle se résume à deux éléments: un modèle de programmation (MapReduce) et un système de fichiers (HDFS), tous deux hautement distribués. Avec eux, les traitements décrits précédemment sont parallélisés sur différents nœuds.

Des composants au départ développés par Google

En fait, Hadoop incarne quasiment à lui seul ce concept de big data. Il concentre en tout cas l'énergie de toute l'industrie, qu'il s'agisse de start up ou d'acteurs traditionnels. Pourquoi un tel engouement? Cette technologie a fait ses preuves par l'exemple. Pour la petite histoire, le succès de Google lui est en partie imputable. Il y a dix ans, alors qu'il ne pèse encore rien sur le marché des moteurs de recherche, le futur géant du web développe ce qui deviendra les composants phare d'Hadoop. « Pour stocker, traiter et indexer 5 milliards de pages web, il construit MapReduce, Google Big Table, Google File System. Il exploite ces éléments pendant trois ans et en fait la description dans une publication académique », raconte Charles Zedlewski, vice-président produit de Cloudera, l'une des trois start up américaines pionnières dans le big data, avec Hortonworks et MapR. Six mois plus tard, un certain Doug Cutting, l'un des fondateurs du moteur de recherche

En s'appuyant sur une haute distribution des traitements, Hadoop promet une équation économique imbattable