

DOSSIER BIG DATA

Architecture technique FORCES ET LIMITES DU SOCLE TECHNOLOGIQUE HADOOP

Principal atout de la pile open source : la finesse des modèles prédictifs qu'elle génère est directement fonction du volume de données qu'elle est capable de brasser. Considérable.

Mais la maîtrise d'Hadoop exige de nouvelles compétences qui restent rares. Notamment en programmation parallèle.

Depuis un an, l'engouement du marché se concentre sur une dimension particulière du big data : l'analyse de données. Cet enthousiasme résulte clairement de la montée en puissance d'Hadoop, une technologie qui, grâce à son modèle de programmation et à son système de fichiers, tous deux hautement distribués, est capable de gérer d'immenses quantités de données non structurées. Seulement, le socle Hadoop a beau être très prometteur, il n'est pas adapté à tous les usages. Tour d'horizon de ses forces et faiblesses.

Selon la start up californienne Cloudera, Hadoop répond à deux besoins spécifiques : « D'une part, explique Charles Zedlewski, vice-président produit, celui d'un traitement massif des données n'ayant pas de schéma clair, et de leur transformation vers un format plus structuré. Par exemple, la construction d'un index de pages web. L'autre besoin concerne l'analytique avancée, c'est-à-dire l'élaboration de modèles prédictifs - lutte contre la fraude, type de publicité à proposer en ligne... »

L'AVIS DE L'EXPERT

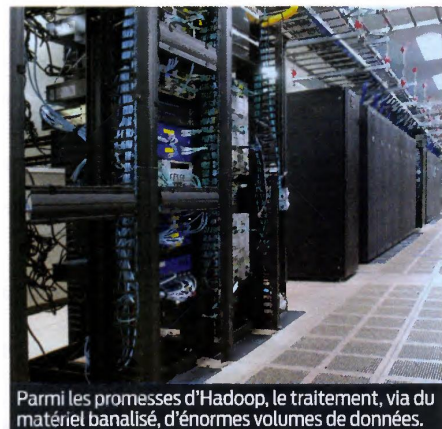


Julien Cabot, responsable de l'activité finance chez Octo Technology

« Les compétences en programmation parallèle sont encore rares sur le marché »

Hadoop est très prometteur, car il permet des analyses jusque-là impossibles ou trop coûteuses à réaliser. Il n'en exige pas moins son lot de compétences spécifiques, en particulier en programmation parallèle. De très bons développeurs spécialisés, par exemple en Java ou dans la conception de requêtes décisionnelle, ne maîtriseront pas le modèle de programmation de MapReduce, au cœur d'Hadoop.

Ces compétences en programmation parallèle relèvent aujourd'hui des ingénieurs spécialisés dans les grilles de calcul. Généralement dans le milieu bancaire. Ils sont actuellement peu nombreux sur le marché. Par ailleurs, rappelons qu'Hadoop exige une maîtrise fine de l'architecture réseau et serveur. Ce dernier point a tendance à être sous-estimé sous prétexte qu'Hadoop repose sur du matériel banalisé.

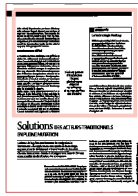


Parmi les promesses d'Hadoop, le traitement, via du matériel banalisé, d'énormes volumes de données.

dans des environnements changeants. » Pourquoi changeants ? Historiquement, les entreprises ont été forcées d'agréger un minimum leurs données de détail pour être en mesure de conserver, par exemple, un historique des ventes le plus ancien possible. Seulement, si le contexte dans lequel les transactions ont été opérées est chamboulé - nouvelle structure de l'entreprise ou nouvelle classification des produits -, les analyses ne peuvent plus être rejouées. « Le fait d'avoir agrégé des données interdit tout retour en arrière. Hadoop, lui, permet de conserver toutes les données de détail et de simuler des scénarios », indique pour sa part Eric Baldeschwieler, PDG d'Hortonworks, une filiale de Yahoo également positionnée sur Hadoop.

Des calculs sur la totalité des données

Plus généralement, le stockage de données de bas niveau rend les utilisateurs libres d'explorer des axes auxquels ils n'auraient pas pensé initialement. Sur ce point, Hadoop se rapproche des offres de décisionnel « en mémoire » qui, elles non plus, n'exigent pas de configurer « dans le dur » les axes d'analyse. Autre atout indéniable du socle : l'élaboration des algorithmes, qui repose sur la totalité des données stockées. Une minirévolution pour Jack Norris, vice-président marketing de MapR, une autre start up de la Silicon Valley spécialisée dans Hadoop : « Dans les approches traditionnelles qui utilisent des bases SQL, les algorithmes sont construits avec des échantillons de données. Et plus cet échantillon est important, plus le coût de l'analyse est élevé. Il croît même de manière exponen-



tielle. » Bref, l'équation des partisans d'Hadoop est la suivante: l'argent dépensé pour nettoyer et organiser les données (avant l'élaboration du modèle prédictif), enrichir sans cesse les algorithmes et embaucher pour cela des compétences très pointues, peut être économisé par un stockage massif des données dans la pile, dont le volume est garant de la précision du résultat.

Un traitement en différé

A l'inverse, ne vous attendez pas à ce qu'Hadoop réalise à la milliseconde des transactions financières. Ce socle technologique reste profondément associé aux traitements différés. Par ailleurs, son processus d'alimentation de données, bien que récemment amélioré, manque de souplesse, car bien trop séquentiel. Ne lui demandez pas non plus d'effectuer des opérations relevant de la business intelligence classique. « *L'analyse opérationnelle des ventes de la semaine par zone et par produit reste plus adaptée au sein d'un datawarehouse classique* », reconnaît Charles Zedlewski qui envisage Hadoop comme un socle d'archivage pour les entrepôts de données.

Enfin, même si ses origines remontent à une décennie, la pile Hadoop est à peine stabilisée. Et ses différents composants (Hive, H Catalog, Pig...) ne demandent qu'à être enrichis. Preuve de la jeunesse de l'écosystème: selon Cloudera, lorsqu'il est déployé en entreprise, principalement dans les banques et chez les opérateurs téléphoniques, il ne touche en moyenne que 20 à 30 utilisateurs. Charles Zedlewski tient cependant à relativiser

Hadoop permet
d'économiser
l'argent
généralement
dépensé
en nettoyage
et en organisation
des données

DATES CLÉS

La technologie Hadoop

2001: Google crée Big Table (base de données compressées) et l'algorithme MapReduce.
De 2004 à 2006: suite à la publication de ces éléments par Google, Doug Cutting lance un prototype Hadoop, puis rejoint Yahoo qui stabilise Hadoop.
Mars 2009: création de Cloudera qui embauche Doug Cutting en fin d'année, puis de MapR, une start up qui enrichit Hadoop d'une gestion de stockage propriétaire.
Juin 2011: Yahoo fonde Hortonworks, filiale dédiée au big data. Avec Cloudera, elle dispute le titre de plus gros contributeur d'Hadoop.
Décembre 2011: lancement de la v.1.0.0 d'Hadoop, qui succède à la v.0.22.0.

cette faible audience. Selon lui, elle tient aussi au fait « *qu'Hadoop exige des compétences particulières pour travailler sur des données qui n'ont pas de schéma* ». C'est d'ailleurs cet argument que font valoir les défenseurs des bases historiques, comme François Guerin, responsable avant-vente de Sybase: « *Le langage SQL peut paraître démodé à certains, mais il intègre des règles de structuration qui font défaut à Hadoop et, plus généralement, aux bases NoSQL. Notamment le contrôle d'intégrité.* » ■