Pourquoi Hadoop est en passe de détrôner le datawarehouse traditionnel (avis d'expert)

Nous le pressentions il y a 2 ans. Nous en sommes aujourd'hui intimement convaincus, preuves à l'appui : le Big Data, et, en particulier l'environnement Hadoop, fait désormais **jeu égal avec les entrepôts de données traditionnels**. Prix, avancées technologiques, simplification de la prise en main… Rien ne s'oppose désormais à ce que ce nouveau socle analytique remplace les datawarehouse historiquement déployés dans les entreprises pour des besoins décisionnels. D'ailleurs nombre d'entreprises ont d'ores et déjà franchi le pas, parmi lesquelles certains de nos clients.

Comment expliquer qu'une technologie encore inconnue du grand public il y a quelques années menace des architectures vieilles de plusieurs décennies ? Première raison : Hadoop, avec son écosystème, **respecte aujourd'hui la plupart des exigences d'intégration** avec les systèmes d'information. Les distributions Hadoop proposent aujourd'hui les outils indispensables aux équipes d'administration pour sécuriser l'accès aux données, gérer l'allocation des ressources, automatiser les sauvegardes et monitorer le bon fonctionnement de l'ensemble. Mais si Hadoop gagne aujourd'hui ses galons de plateforme d'entreprise, c'est aussi et surtout pour sa nouvelle réactivité.

Jusque-là, la star des technologies Big Data ne supportait que le mode batch. La plus petite des opérations prenait plusieurs minutes… Impossible, dès lors, de rivaliser avec les datawarehouse qui traitent les requêtes décisionnelles à la volée (calculs de chiffre d'affaires, de marges ou de prévisions). Or, avec des projets tels que Yarn, Impala, Spark, Drill ou Presto, **les plateformes Hadoop s'ouvrent désormais aux requêtes interactives et instantanées**. De même, avec Storm et Kenesis, elles capturent et analysent au fil de l'eau les données transitant dans les flux. Hadoop met ainsi **un pied dans le temps réel**. Enfin, lorsqu'il est associé à une brique Elasticsearch ou Solr**,** il se mue en moteur de recherche à la capacité d'indexation quasi infinie.

Plus de capacités, pour moins cher

Stockage, traitement massif, requête interactive, requête transactionnelle, outil de recherche… Hadoop n'a donc plus rien à envier aux datawarehouse. D'autant que la plateforme se montre parfaitement **compatible avec les outils de transformation et d'intégration de données,** d'une part, et avec les **applications de reporting, d'analyse prédictive et de visualisation,** d'autre part**.**

Autre argument de taille censé faire pencher la balance : le prix. À en juger par les déploiements de nos clients, **un projet Hadoop est en moyenne cinq fois moins cher** qu'un datawarehouse classique. Ce chiffre comprenant le matériel, le logiciel et le déploiement de l'infrastructure. Sans compter qu'une plateforme Big Data stocke environ cinq fois plus d'informations qu'un datawarehouse traditionnel. Aux données de ventes, sont en effet associées toutes les

informations relatives aux comportements des clients en magasin, sur le web ou les réseaux sociaux, etc.

Un accès facilité à la techno

Enfin, si ce n'était pas le cas en 2010, Hadoop est aujourd'hui devenu LA référence du Big Data. De quoi sécuriser au maximum les investissements consentis par les entreprises. Les communautés open source l'ont totalement adopté, notamment la **fondation Apache**. Même les géants de l'infrastructure s'y rangent **(IBM, Microsoft, Oracle**, etc). Et sur le terrain, la plupart des grands groupes le testent et envisagent de le mettre en production prochainement. Ces derniers doivent être rassurés : cette technologie est pérenne. Open source, elle n'est « enfermée » par aucun éditeur et jouit d'un écosystème très riche, très actif et très productif.

Se pose pourtant, diront certains, **la question des compétences**. Peu de profils, en effet, sont à même de déployer ces plateformes et d'investiguer les données qui y sont stockées. C'est vrai, mais là encore, ce frein tend à se résorber. De plus en plus d'outils (en particulier ceux issus des projets évoqués plus haut) n'exigent plus de compétences spécifiques en matière de programmation parallèle. Ils tendent à être accessibles par les personnes qui, dans les entreprises, gèrent déjà le décisionnel et exploitent les bases de données.

Pour autant, la fin des datawarehouse n'est pas pour tout de suite. Culture du changement oblige, les deux socles analytiques devraient **cohabiter encore longtemps.** Mais l'on ne voit pas comment et pourquoi les organisations continueraient à payer le prix fort pour des analyses qui, à terme, seront accessibles à moindre coût via une plateforme incommensurablement plus riche et plus ouverte.

En complément, sur le même sujet :

Pourquoi NoSQL s'impose face aux SGBDR traditionnelles (avis d'expert)