Mike Olson, Cloudera : « dans Hadoop, temps réel et batch sur les mêmes données »

Croisé dans les allées du salon Big Data Paris 2016 (que se tient en ce moment au Palais des Congrès de la Porte Maillot), Mike Olson est un des fondateurs de Cloudera, l'un des principaux éditeurs de distribution Hadoop aux côtés de Hortonworks et MapR. Après avoir été le Pdg de la société de sa création à 2013, il est désormais son directeur de la stratégie. Mike Olson revient sur les récentes évolutions de l'environnement Hadoop, avec l'apparition de frameworks, comme Spark ou Storm, permettant de bâtir des applications temps réel.

Silicon.fr : cette édition 2016 de Big Data Paris consacre Hadoop comme plate-forme pour les applications temps réel, alors que la technologie a avant tout été conçue pour des environnements batch. Pourquoi ce virage ?

Mike Olson : Quand la technologie a été inventée au sein de Google, dans les années 2000-2002, elle était pensée pour le stockage d'énormes volumes de données, pas pour le streaming ou le processing de données. Après le lancement du projet Hadoop en 2005-2006, nous avons monté Cloudera ; l'ambition était simple : la technologie était déjà employée par les grands noms du Web et nous voulions la diffuser dans les entreprises traditionnelles. A l'époque, le processing de données passait par le framework MapReduce, conçu pour le mode batch.

Même si ce dernier reste pertinent sur les données historiques, d'autres moteurs ont depuis vu le jour. C'est le cas d'Impala que nous avons développé. Ou encore de Spark, un outil dédié aux analyses de données en temps réel que nous avons adopté en 2013. Citons aussi le projet Apache Solar qui a donné naissance à Cloudera Search. Tous ces frameworks ouvrent aux développeurs d'applications de nouvelles portes leur permettant de concevoir des applications temps réel s'appuyant sur la même plate-forme que celle utilisée pour le mode batch. Ils ont à leur disposition une collection d'outils bien plus puissante qu'à l'origine. Et toutes ces applications peuvent fonctionner sur les mêmes données. En parallèle de ces évolutions, l'écosystème Hadoop s'est aussi enrichi de fonctions de sécurité –gestion des accès, chiffrement… – qu'attendaient des secteurs comme la banque ou la santé.

Quels sont les types d'applications temps réel qui sont aujourd'hui déployés au-dessus de Hadoop ?

M.O. : Les applications améliorant la connaissance client, par exemple. C'est le choix qu'a fait par exemple Mark & Spencer pour son système d'engagement client temps réel. Les systèmes améliorant la conception des produits également. SFR emploie ainsi une application Hadoop pour repenser le design de sa plate-forme de messagerie. On peut aussi citer les applications de conformité ou de gestion de la sécurité. L'objectif est ici de surveiller en temps réel l'activité dans un datacenter ou de détecter en direct les attaques des cybercriminels.

Existe-t-il un marché de remplacement d'applications bâties auparavant sur des bases de

données relationnelles ?

M.O. : D'abord, rappelons que des technologies comme Oracle ou Teradata reviennent à environ 40 000 dollars par To, contre entre 1 000 et 2 000 dollars pour Hadoop. Ce dernier bénéficie par ailleurs d'une architecture moderne, extensible. Malgré ces constats, je ne vois pas les entreprises débrancher leur datawarehouse traditionnel pour le remplacer par Hadoop. Simplement, cette technologie leur offre une option supplémentaire pour stocker beaucoup plus de données et avec beaucoup plus de flexibilité.

A lire aussi :

Big Data : la Silicon Valley a toujours le béguin pour Hadoop

Big Data : Amazon renforce son offre Hadoop / Spark

Comment Criteo transforme Hadoop en moteur de sa rentabilité