Rob Bearden, Hortonworks : « le Big Data ne se résume pas au datalake »

Croisé à l'occasion de l'étape parisienne de son roadshow Future of Data, le Pdg de Hortonworks revient sur l'évolution du Big Data dans les entreprises. Pour l'éditeur de la distribution Hadoop, les enjeux se déplacent du stockage de tous types de données dans un datalake, vers les traitements analytiques temps réel proches des points de capture de l'information.

Hortonworks, une société californienne cotée en bourse depuis décembre 2014, a réalisé un chiffre d'affaires de 122 M$ en 2015 (et vise 260 M$ en 2016). Entretien avec son Pdg, qui affiche sa volonté d'investir en France et en Allemagne. Une filiale hexagonale doit d'ailleurs voir le jour début octobre.

Silicon.fr : Comment envisagez-vous l'extension du champ du Big Data ?

Rob Bearden : Nous pensons que traiter les données en mouvement crée beaucoup de valeur pour l'entreprise. Ensuite, quand ces données sont amenées au repos, dans le datalake, comprendre le contexte dans lequel elles ont été créées devient plus compliqué. La valeur du datalake réside dans le fait de rassembler toutes les données de l'entreprise, celles sur les clients, les produits, l'historique des ventes…

En termes de technologies, comment cette évolution se traduit-elle ?

Quand Hadoop a démarré, c'était une plateforme fantastique pour gérer de gros volumes de données. Mais elle était limitée, dans son architecture, à la gestion d'un unique dataset et à des traitements en mode batch. Nous avons amené des évolutions que nous pensons très importantes à cette architecture, permettant de réunir différents datasets dans une architecture centralisée et d'y greffer à la fois des processus en mode batch et des applications interactives et temps réel. Pour y parvenir, nous avons créé le gestionnaire de ressources Yarn, un outil sophistiqué permettant de réunir toutes les données dans un environnement centralisé et de gérer tous types d'applications. Ce qui nous permet de greffer des moteurs de traitement comme Spark. Un outil sur lequel nous avons d'ailleurs beaucoup travaillé pour l'adapter aux attentes des entreprises.

Nombre de grandes entreprises françaises ont déployé un datalake. Les nouveaux projets vont-ils au-delà de cette ambition consistant à réunir toutes les données dans un réceptacle unique et à décharger le datawarehouse ?

Une fois le datalake en place, les entreprises sont confrontées au défi de son exploitation. Il s'agit de bâtir des analyses prédictives relatives aux clients et aux produits. Les deux prochaines étapes pour les entreprises consistent à agir sur les données dès le point de collecte, par exemple un capteur. Pensez à Fitbits (qui conçoit des traqueurs d'activité, NDLR) qui pourrait, en cas d'anomalie dans les données, avertir un docteur en temps réel. Dans ce scénario, des traitements analytiques tournant pendant la nuit, en mode batch, ne sont d'aucune utilité. L'autre axe d'investissement consiste à disposer d'une vue à 360° de son activité. Par exemple, une entreprise qui aurait déjà injecté les informations sur les interactions en ligne avec ses clients, va y ajouter l'historique des

ventes, la durée des contacts avec le centre d'appel, l'historique des paiements… Des données qui résident habituellement dans des systèmes transactionnels différents.

Plusieurs éditeurs, comme Cloudera ou MapR, commercialisent aussi des distributions Hadoop. En quoi Hortonworks est-il différent ?

Le premier point de différenciation réside dans notre architecture. Alors que MapR et Cloudera ont basé leur plateforme sur l'architecture originelle d'Hadoop, Hortonworks l'a étendue avec Yarn, ce qui donne la possibilité de réunir toutes les données dans un environnement centralisé, plutôt que de les disperser dans toute une série de silos. Les plateformes de nos concurrents ont certes intégré Yarn, mais cela se limite à un dataset unique. Sous Yarn, ces distributions restent très axées sur des silos de données. En faisant tourner des services natifs Yarn, nous pouvons faire tourner simultanément des traitements en batch et des applications temps réel. Ce qui ouvre le champ des possibles en matière d'usages… Là où nos concurrents sont avant tout limités aux données au repos, déjà chargées dans Hadoop.

Par ailleurs, notre technologie est 100 % Open Source, ce qui accélère l'adoption par l'écosystème. Nous avons trois fois plus de développeurs investis dans les projets Apache que n'importe lequel de nos concurrents.

A lire aussi :

Big Data : MapR surfe sur la vague Spark pour consolider Hadoop

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

Romain Chaumais, Ysance : « le Big Data en temps réel n'est pas une exigence, c'est une libération »