Reportage dans la Silicon Valley : un outil de recherche Cloudera pour le Big Data Hadoop

Reportage réalisé dans le cadre du IT Press Tour 2013 (juin 2013)

Retour chez Cloudera, l'une des trois premières distributions historiques du big data Hadoop, que nous retrouvons une nouvelle fois dans ses locaux, au centre de San Francisco. Une start-up, qui a su séduire les investisseurs – Cloudera a levé 65 millions de dollars ! – en imposant sa distribution Hadoop basée sur le cœur open source du projet de la fondation Apache – auquel elle contribue largement -, un cœur étendu via des développements en partie propriétaires.

Lire « Cloudera, l'état de l'art d'Hadoop et du Big Data analytique ».

Peter Cooper-Ellis – un vétéran avec ses 30 ans d'industrie du logiciel, qui a rejoint l'éditeur au poste de vice-président Engineering après être passé par la case VMware – nous a présenté Cloudera comme « *la plateforme du big data pour le stockage de volumes massifs de données, en partenariat avec des sociétés qui font de l'analytique. Nous sommes la première distribution packagée de Hadoop open source. Et en 2012, nous avons annoncé Cloudera 4, la première plateforme big data analytique d'entreprise mature.* »

Il est vrai que les chiffres alignés par la start-up sont éloquents : au bout de 5 ans, Cloudera affiche 400 employés, 600 partenaires, des dizaines de milliers de nodes, 20 milliards d'évènements enregistrés par jour, 250 millions de tweets sur Twitter. 70 % des smartphones aux US aboutissent sur l'Hadoop Cloudera, qui participe également à la standardisation des institutions financières, et compte parmi ses clients Box, eBay, Experian, Expedia, Monsanto, ou encore Nokia.

Cloudera Impala

Fin 2012, Cloudera annonçait Impala, un moteur de requêtes SQL interactif pour Hadoop. L'outil comprend également un moteur de requêtes MPP natif, un runtime séparé de MapReduce, des requêtes low latency. Le tout en open source.

Les avantages d'Impala ? « *Porter l'expérience SQL sur le Big Data. Nous apportons la capacité de poser de nouvelles questions de type business intelligence et analytique sur plus de données,* explique **Justin Erickson**, Director Product Management de Cloudera*. Notre plateforme réduit les délais de migration des données et les temps de latence des applications analytiques, tout en conservant la fidélité des requêtes* ». En revanche, si Impala propose un modèle pour développer des analytiques, il ne dispose pas de modèles de requêtes packagés.

Cloudera Search

La dernière nouveauté annoncée par Cloudera, et qui nous a été présentée lors de notre visite dans la Silicon Valley, s'appelle Cloudera Search. Il s'agit d'un moteur de recherche pour données big data, toujours en open source. Cet outil est intégré à Cloudera CDH (le cœur Apache Hadoop open source augmenté de 9 projets open source provenant de l'écosystème Hadoop) et repose sur le projet Apache Solr, une plateforme logicielle de recherche s'appuyant sur le moteur de recherche open source Lucene.

« *L'objectif est d'étendre le ROI du big data avec un outil analytique d'exploration, simple d'emploi pour tous, et qui ne soit pas séparé du stockage* », nous a indiqué Justin Erickson. « *Basé sur Lucene, Search est intégré à Cloudera Manager et CDH. Et il profite de HDFS comme base d'index.* »

Cloudera voit « *Hadoop comme le point central de la donnée* ». Et demain ? « *Nos réflexions tournent autour de l'accès Hadoop via Windows…* ». En attendant, l'éditeur a signé un partenariat avec SAS pour faire tournerles outils de ce dernier sur la distribution Cloudera (lire Mouloud Dey (SAS): « Jouer le rôle de chef d'orchestre Big Data au cœur d'une cohabitation hétérogène »).

Nouvelle rencontre avec Doug Cutting

Que le monde est petit ! Doug Cutting est à l'origine de Hadoop (lire « Cloudera : une brève histoire d'Hadoop, de son créateur, et d'une révolution »), mais également de Lucene qu'il a développé en 1999. Nous retrouvons une nouvelle fois cette sympathique figure de la Silicon Valley, grand amateur de la France. L'occasion de l'interroger sur le devenir d'Hadoop.

« *Nous assistons à la transition du processing local vers le web processing. C'est une transition fondamentale dans notre façon de traiter les données. L'usage des PC partout et l'adoption des technologies génèrent de plus en plus de données. L'open source a largement participé au succès de Hadoop, qui est une plateforme majeure pour ouvrir de nouvelles voies dans le traitement de la donnée. C'est le leader du big data. En rendant Solr efficace, nous avons transformé Hadoop en OS du big data-as-a-service. Ma vision du futur ? C'est le processing sur le streaming et le learning.* »

Voir aussi

Silicon.fr étend son site dédié à l'emploi IT
Silicon.fr en direct sur les smartphones et tablettes