Le Big Data toujours aussi bouillonnant dans la Silicon Valley

En moins de cinq ans, Hadoop s'est imposé comme une technologie phare de l'industrie informatique. Et ses pionniers sont devenus des références, toujours conscientes de leur nécessité d'expansion, malgré les dizaines ou centaines de millions de dollars amassés pour leur financement. Et déjà de nouvelles venues, très inventives, lorgnent sur le gâteau Big Data.

Cloudera : 740 millions de raisons d'apprécier Intel

Arrivé comme CEO chez Cloudera en juin 2013, **Tom Reilly** a profité du Press Tour européen dans la Silicon Valley pour faire le point sur la société et le rôle majeur d'Intel expliquant son investissement record sur la star du Big Data. Après avoir levé plus de 140 millions de dollars entre 2009 et 2013, Cloudera a en effet reçu 900 millions de dollars en mars 2014, dont 740 d'Intel (voir notre article).

« Aucune entreprise ne nécessite autant d'argent », plaisante Tom Reilly. *« Intel détient près de 50% de part de marché des serveurs dans les datacenters. La société a parié avec succès sur trois technologies majeures : Wintel, Linux via des partenariats avec RedHat – entre autres -, et dans VMWare (notre article). Aujourd'hui, les ERP sont l'application numéro 1 sur les serveurs. Or, Hadoop dépassera l'ERP sur le marché applicatif. Pour son quatrième pari, il est donc logique qu'Intel ait misé sur Cloudera.»* Pour le dirigeant, Hadoop incarne donc logiquement **un levier pour la vente des serveurs** dans les datacenters. CQFD.

« Intel Ventures n'a pas investi avec l'objectif d'obtenir un rapide retour sur investissement, mais bien pour contribuer à étendre le marché potentiel des serveurs. La concentration des processeurs augmente tous les 18 à 24 mois, mais les données doublent tous les 12 mois. Il faut donc combler le décalage en achetant aussi des serveurs. Ainsi, China Mobile a acheté 100 000 serveurs cette année,» assure le dirigeant. *« De notre côté, nous nous efforcerons de concevoir du logiciel optimisé pour la plate-forme Intel.»* Selon lui, on peut déjà imaginer une sorte de « Cloudera on chip », un processeur qui apporte déjà de nombreuses accélérations, comme la sécurité intégrée (chiffrement, etc.).

On se souvient qu'Intel avait lancé sa propre distribution Hadoop en 2013, développement depuis abandonné pour Cloudera à la faveur de son investissement dans la société. *« Intel est un grand client de Cloudera,* renchérit Tom Reilly. *Son arrivée au capital nous a aussi apporté des échanges technologiques conséquents sur la sécurité (stratégique pour Hadoop) et une porte d'entrée vers les très grandes entreprises à travers le monde.»*

Avec 40 employés en Europe, Cloudera annonce que 2015 sera pour elle une grande année, ainsi que pour le Big Data. *« Certes, Cloudera ne représente pas une menace imminente pour Oracle. Toutefois, HDFS et Hadoop représentent certainement un danger pour ce géant et son hégémonie sur les bases de données. Et, dans trois ou quatre ans, Cloudera devrait être trop importante et donc hors de portée pour une acquisition par Oracle »,* conclut le CEO.

Qubole ou BGaaS : Big Data as a Service

Avant la création de Qubole en 2011, **Ashish Thusoo** (cofondateur et CEO) travaillait dans l'équipe Data Infrastructure de Facebook, tandis que **Joydeep Sen Sarma** (cofondateur et dirigeant) était rattaché à l'écosystème chargé du traitement de données sous Hadoop, avait lancé le projet Apache Hive (datawarehouse construit sur Hadoop), et dirigé l'équipe Data Infrastructure.

Rien d'étonnant à retrouver ces deux-là dans Qubole, une start-up prometteuse qu'ils ont créée en 2011. *« Avec Qubole, nous proposons une plate-forme Big Data as a Service combinant les avantages du Cloud (agilité, flexibilité, collaboration, self-service…) et les stouts des moteurs Big Data (Hadoop, Hive, Presto, Spark…). Il devient ainsi possible de passer de 10 à 1 000 nœuds en quelques minutes sans se préoccuper de l'infrastructure. Et même de mixer divers types de nœuds et différentes tâches »,* explique Gil Allouche, responsable marketing chez Qubole. *« Une solution sur site impose de maintenir les clusters et les configurations, de gérer le matériel, et de disposer d'experts pour gérer le tout. »*

Pour déployer des clusters Hadoop sur le Web, Qubole a interfacé son service avec Amazon Web Services, où les instances sont automatiquement créées et supervisées depuis l'interface Qubole. Depuis un an, d'autres plates-formes IaaS sont proposées comme Google Cloud Platform ou Microsoft Azure. Par ailleurs, la solution est compatible Openstack, ce qui élargit encore les possibilités. Le client ne paie que pour la puissance de calcul et s'acquitte du reste auprès du fournisseur de IaaS.

Outre les connecteurs ODBC et JDBC, Qubole propose de nombreux connecteurs applicatifs et continuera à en intégrer ou à en développer selon les besoins.

Une connexion des données ou informations à Qubole permet l'autodétection, sans code ni manipulation. Puis, l'utilisateur peut sélectionner les champs ou les documents qui l'intéressent.
Selon l'éditeur, sa plate-forme serait 4 à 8 fois plus rapide qu'Amazon EMR (Elastic Map Reduce) et deux fois plus rapide qu'Amazon RedShift (datawarehouse). Basé sur Spark, Qubole bénéficie logiquement de meilleures performances que Map Reduce (EMR chez Amazon).

S'appuyant sur des solutions comme Amazon S3 pour le stockage, Qubole bénéficie par ailleurs de la sécurité de ce type de plates-formes : chiffrement, authentification, etc.

Des modèles prédéfinis permettent à des utilisateurs métiers de manipuler des données sans connaissance technique. Outre l'interface graphique, les développeurs peuvent accéder aux fonctions Qubole depuis leurs applications via des API et un SDK (Python).

Parmi plus de 60 clients aux États-Unis, en Inde et en Europe, l'éditeur compte déjà des références reconnues comme Pinterest (réseau social et partage de photos pour des dizaines de millions de membres), qui utilise Qubole pour ses index de recherche de documents. Le plus grand cluster actuellement déployé compte 1 800 nœuds, tandis que le site traite 86 Po de données par mois, et vise les 100 Po par mois en 2015.

Avec plus de 70 employés répartis entre les États-Unis et l'Inde, Qubole se développe et **ouvrira certainement des bureaux en Europe.** La start-up a déjà levé 7 millions de dollars lors d'un premier tour de table auprès de fonds réputés pour leurs investissements technologiques :

Lightspeed Ventures et Charles River. Autant d'atout qui devraient rapidement amener Qubole sur le devant de la scène Big Data.

Platfora embarque le Big Data In-Memory

« Transactions, interactions clients, données machine, tweets, logs… le Big Data résulte d'ensembles de données hétérogènes à corréler : informations diverses, variées et différemment structurées. Platfora propose une solution permettant de gérer tout cela via un accès en quelques secondes à cet ensemble de données », lance **Ben Werther**, fondateur et CEO de Platfora.

Cette start-up a été créée en 2011 par ce dernier, auparavant vice-président Produits chez DataStax, et chef produit chez Greenplum (rachetée par EMC en 2012). Une expérience qui favorise la levée de fonds. Ainsi, Platfora a engrangé 27 millions de dollars en deux tours de tables en deux ans auprès d'Andreessen Horowitz, Sutter Hill Ventures et In-Q-Tel (fonds de la CIA misant sur les technologies liées au renseignement). Puis, un troisième tour (Series C) a apporté 38 millions de dollars supplémentaires en mars 2014.

Le logiciel Platfora s'installe au-dessus d'une solution Hadoop (Cloudera, MapR, Hortonworks, Pivotal, Amazon Web Services et Cisco). Il peut se déployer aussi bien dans un datacenter que sur le Cloud.

« Nous proposons une approche architecturale différente pour remédier à l'impossible balayage ligne par ligne… et à la rigidité d'une structure préparée », explique Ben Werther. *« Platfora masque la complexité technologique en proposant des interfaces visuelles et intuitives aux utilisateurs, qu'ils soient analystes métier, informaticiens, dirigeants ou data scientists.»*
La solution s'installe au-dessus d'un cluster Hadoop et consiste à récupérer les informations en drag&drop afin de les enrichir de liens entre elles dans le Data Catalog. Ensuite, la plateforme puise dans le Data Lake Hadoop et ramène les résultats en mémoire, qui deviennent alors manipulables. Des résultats qui peuvent être mis à jour manuellement ou automatiquement, avec possibilité de préserver des snapshots pour comparer ou conserver des historiques. En plus de Hadoop, d'autres sources peuvent également servir de base au mix des données.

Le logiciel ne propose pas encore l'auto-découverte des schémas de données, de logs de schémas et autres informations. Mais cette évolution est à l'étude.

La plate-forme intègre une gestion évoluée des utilisateurs : privilèges, accès, temps d'utilisation, volumes de données autorisés, etc.

Avec sa version 4.0, Platfora intègre mieux Spark et bénéficie pleinement de ses performances et de la possibilité d'automatiser des tâches (workflow) de bout en bout pour les non-informaticiens. Spark ouvre aussi la solution à divers langages de programmation comme Python, entre autres. De très nombreuses visualisations sont déjà disponibles, et d'autres devraient venir compléter le dispositif.

Parmi les nombreux clients (Disney, JP Morgan, Groupon , Unisys, Paypal, Paytronics, Auto Trader…), la société compte aussi le tour opérateur allemand Tui. Platfora affiche d'ailleurs de grandes ambitions en Europe, où de nombreux projets pilotes sont en cours. La start-up cherche

également des partenaires pour l'aider à les réaliser.

A lire aussi :

Silicon Valley Tour – Le stockage objet au centre de toutes les attentions

Crédit photo : Ai825 / Shutterstock