MapR : la distribution Hadoop pour le cloud Google

Si avec Compute Engine **Google** se lance dans le cloud (voir Compute Engine : l'IaaS hautes performances de Google), le montage technologique de l'offre IaaS (Infrastructure as a Service) du géant de la recherche Internet nous a interpellés et nous restons prudents quant aux ambitions du projet (Google Compute Engine se coupe de 90 % des clients du cloud). En revanche, s'il est un domaine qui pourrait tirer profit du cloud de Google, c'est l'analytique big data. Un juste retour des choses ?

Du développement de MapReduce à Hadoop sur le cloud Google

Rappelons tout d'abord que Google est en partie à l'origine de **Hadoop**, la plateforme big data open source devenue la référence. Tout du moins c'est Google qui a initié voici quelques années le développement de MapReduce, le module analytique qui accompagne le système de fichiers des grosses volumétries du big data.

MapReduce a ensuite inspiré la communauté de développement de Hadoop, tandis que son auteur **Doug Cutting**, en opposition avec la stratégie de Google (qui s'est révélée plus propriétaire que le discours de l'époque), quittait la firme pour rejoindre Yahoo et continuer de développer Hadoop dans l'esprit de l'open source. Nous avons rencontré Doug Cutting lors d'un voyage sur la Silicon Valley, lire notre article Cloudera : une brève histoire d'Hadoop, de son créateur, et d'une révolution.

C'est la distribution de **MapR Technologies**, une version commerciale de la distribution Apache Hadoop, qui a été retenue par Google pour alimenter l'offre analytique big data de Compute Engine. Google devrait donc rapidement proposer un service d'analyse de la donnée sur son cloud basé sur un large cluster piloté par Hadoop. Une version bêta privée gratuite de MapR sur Google Compute Engine est disponible, mais reste réservée à un petit nombre d'utilisateurs sélectionnés.

Démonstration de Hadoop sur Compute Engine

MapR a réalisé une démonstration d'Hadoop sur Google Compute Engine lors de la conférence **Google I/O** qui s'est tenue fin juin. MapR s'est appuyé sur un cluster de 1256 nœuds, avec 5024 cœurs de processeurs et 1256 disques, pour réaliser une transaction analytique de 1 To, qui a pris 1 minute et 20 secondes. En comparaison, le record pour la même transaction est de 1 minute et 2 secondes, sur un cluster physique privé alignant 200 serveurs supplémentaires, de double de cœurs et quatre fois plus de disques.

Proposé à la demande, l'analytique big data de Google Compute Engine pourrait séduire les organisations qui ne craignent d'affronter ni le cloud, ni certaines pratiques de Google qui peuvent

laisser planer le doute sur l’exploitation par le moteur des données qui lui sont confiées…