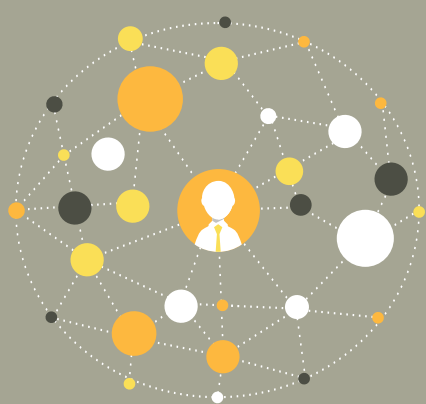


7 BONNES RAISONS DE CHOISIR HADOOP POUR VOTRE PROJET BIG DATA

Hadoop est un framework libre et open source destiné à faciliter la création d'applications distribuées de stockage et traitement de données à grande échelle.

Zoom sur les avantages des clusters Hadoop dans la prise en charge de grands volumes de données, en particulier non structurées.



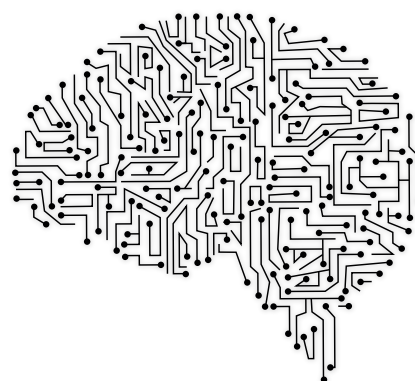
1 SAISIR, STOCKER ET TRAITER DE TRÈS GRANDS VOLUMES DE DONNÉES

À l'ère du Big Data, l'analyse de données nécessite de pouvoir déployer des ressources très importantes afin de pouvoir traiter simultanément plusieurs pétaoctets d'informations. Hadoop fonctionne avec deux modules principaux, MapReduce et HDFS (Hadoop Distributed File System), qui lui permettent de distribuer et paralléliser les opérations sur plusieurs milliers de nœuds selon une logique nœud maître/nœuds esclaves tout en tirant parti de la localité entre données et traitement sur le même nœud.

Hadoop permet donc de dépasser les limitations matérielles de puissance de calcul et d'opérer une gestion globale des données stockées.

2 UN SYSTÈME ÉCHELONNABLE

L'échelonnabilité de l'architecture Hadoop permet un traitement massif mais surtout évolutif des données. Grâce à une gestion intelligente d'un grand nombre d'ordinateurs mis en réseau, la solution peut instantanément accroître sa capacité de calcul en fonction des besoins.



3 LA POSSIBILITÉ DE STOCKER TOUS TYPES DE DONNÉES

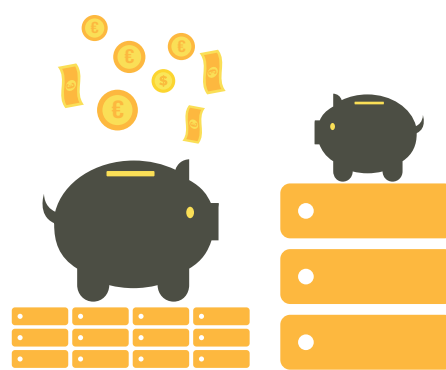


Outre le volume important de données disponibles, l'autre caractéristique du Big Data est la grande variété de ces données générées en permanence par tous types de supports.

Or une des spécificités des solutions NoSQL est de permettre la gestion de tous types d'informations et en particulier de données non structurées qui se prêtent peu aux bases de données traditionnelles : textes, images, fichiers audio...

4 UNE SOLUTION PEU COÛTEUSE

L'utilisation de clusters HDFS présente l'avantage d'offrir une solution de stockage économique par rapport à un système de sauvegarde sur bande. En effet, les machines mises en réseau sont équipées de disques durs banalisés (commodity hardware), le coût par téraoctet est donc beaucoup moins élevé que dans les solutions traditionnelles.



5 UNE GARANTIE CONTRE LES PANNES MATÉRIELLES



Hadoop repose également sur l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence leur gestion doit être intégrée au framework.

Les capacités de haute disponibilité du HDFS permettent de basculer le serveur de métadonnées principal sur une sauvegarde en cas d'échec, le logiciel de regroupement isolant le serveur défaillant.

6 UN FRAMEWORK 100% OPEN SOURCE DISPOSANT D'UNE COMMUNAUTÉ ACTIVE DE DÉVELOPPEURS

Le caractère open source de la solution présente l'avantage de garantir sa pertinence à long terme grâce à une communauté active de développeurs contribuant au code et proposant des corrections au fur et à mesure.



7 UNE TECHNOLOGIE LARGEMENT ADOPTÉE DANS LE SECTEUR DE BIG DATA



À l'heure actuelle, de grandes sociétés du Big Data comme Facebook ou Yahoo! ont un système de gestion et de traitement des données intégrant Hadoop, et la plupart des grandes entreprises ont lancé un projet autour de cette solution.