Éditeurs et outils logiciels

Nous allons parler de quelque éditeurs et outils logiciels pour le Big Data. Il y a plusieurs sorte d'éditeurs et de logiciels, il y a ceux qui vont comprendre les données et essayer de les structurer si possible et il a aussi ceux qui vont interroger les masse de données pour ensuite pouvoir les analyser. Nous allons parler de deux exemple d'éditeurs et de logiciels dans chaque cas (structurel et traitement). Dans un premier temps, nous allons parler de Solr et NoSQL pour la parti structurel, puis dans un second temps, nous allons parler de Hadoop et Spark. Apache a des beaucoup de modules dans le Big Data, notamment Solr, Hadoop, Spark et etc.

Éditeurs

Solr et NoSQL, on la même utilisation. Ces deux éditeurs vont organiser la masse de donnée pour les comprendre plus facilement. Ils vont indexé les données qui permettrons de faire des recherche plus facile dans la masse de données.

Solr

Fonctionnalités :

- Indexation de documents doc, pdf, ppt, ou xls.
- Indexation d'une base de données.
- Optimisé pour de grandes volumétries de données.
- Une API aux standards ouverts, xml, json et http, permet de d'intégrer facilement Solr à une application tierce.
- Système intégré de réplication et de haute disponibilité.

Avantages :

- Une indexation presque instantanée.
- Flexible et adaptable simplement.
- Une interface d'administration claire.

Inconvénient :

- Achat de Plugins

NoSQL

Fonctionnalité

- Indexez dès l'assimilation et interrogez sans fin vos données.

- S'adapte pour contenir jusqu'à plusieurs pétaoctets de données et des milliards de documents.
- Des résultats ultra-rapides.

Avantages

- Contrairement aux bases de données relationnelles, les bases de données NoSQL sont basées sur des paires clé-valeur.
- Certains types de stockage de bases de données NoSQL incluent différents types de stockages, tels que les stockages de documents, de valeurs de clé, de XML et d'autres modes d'entrepôt de données.
- On pourrait dire que l'implémentation de bases de données NoSQL de source ouverte est rentable. Puisqu'ils n'ont pas besoin de frais de licence et peuvent fonctionner sur du matériel économique.
- Lorsque vous travaillez avec des bases de données NoSQL, qu'elles soient de source ouverte ou qu'elles soient propriétaires, l'extension est plus simple et moins coûteuse que travailler avec des bases de données relationnelles.
- Inconvénient :
- La plupart des bases de données NoSQL ne prennent pas en charge les fonctions de fiabilité.
- Afin de soutenir les fonctionnalités de fiabilité et de cohérence, les développeurs doivent implémenter leur propre code, ce qui ajoute une complexité supplémentaire au système.
- Limite le nombre d'applications sur lesquelles nous pouvons compter pour effectuer des transactions sécurisées et fiables.
- L'incompatibilité avec les requêtes SQL est l'une des complexités trouvées dans la plupart des bases de données NoSQL.

Outils logiciels

Hadoop et Spark, on la même utilisation. Ce sont deux outils logiciel qui vont traiter la masse de donnée pour faire des stats et avoir une meilleure connaissance du monde.

Hadoop

Fonctionnalités

- Composants communs permettant de gérer les systèmes de fichiers distribués. Beaucoup de modules se basent sur ce projet.
- Un système de fichiers distribués conçu pour gérer de grosses volumétries.
- Un framework logiciel qui facilite la réalisation d'applications capables de fonctionner dans un environnement clustérisé.
- Logiciel permettant la requête, analyse des données contenues dans un datawarehouse
- Outil d'analyse, traitement des données par le biais de scripts.
- Compatibilité des cluster Hadoop via une interface web.

Avantages

- Gamme de sources de données

- Rentabilité
- Vitesse de traitement
- Copies multiples

Inconvénient

- Absence de mesure préventives
- Problèmes liés aux petites données
- Fonctionnement risqué

Spark

Fonctionnalité

- Performances rapides
- Intégration simple de plugins, api

Avantages

- Rapidité de traitement
- Dynamique de la nature
- Tolérance aux pannes
- Traitement de flux en temps réel

Inconvénient

- L'absence de support pour le traitement en temps réel
- Problème avec les petits fichiers
- Aucun système de gestion de fichier
- Manque d'algorithmes
- Optimisation manuelle
- Traitement itératif
- Temps de latence

Source

Apache Hadoop: Avantages et Inconvénients

Apache Spark: Avantages et Inconvénients

Hadoop Big Data

Solr

NoSQL