Hadoop

Plan :

Introduction

1. Les indispensables d’Hadoop : HDFS et MapReduce
   1. NameNode
   2. Secondary Node, Backup Node et Checkpoint Node
   3. EdgeNode
   4. MapReduce
   5. File / Compression
2. Ecosysteme Hadoop
   1. Outils de gestion et configuration de l’Application
   2. Outils d’accès aux données
   3. Outils

Les bases de données relationnelles (SQL) atteignent actuellement leurs limites de performances et de stockage mais un nouveau modèle, le NoSQL (Not Only SQL) permet désormais de franchir ces limites.

Ces bases sont beaucoup moins couteuses car elles peuvent fonctionner sur du matériel peu cher. Elles s’adaptent également facilement. Cela vient du fait qu’il n’y a pas de limites sur le nombre de disques dur de la base de données, et que le système voit le système de stockage comme une seule entité. Les disques sont donc faciles à ajouter ou à enlever, et n’importe quel disque peut être ajouté, notamment des disques peu couteux.

Nous nous intéresserons dans ce document au framework Hadoop, qui est l’éléphant du BigData et est utilisé par la majorité des utilisateur de BigData.

Hadoop est un Framework Java libre permettant de développer des applications distribués et échelonables (scalables). Il est utilisé pour les bases de données et regroupe de nombreuses librairies.

Hadoop est basé sur le HDFS (Hadoop Distributed File System) qui est un modèle de stockage et sur l’algorithme MapReduce qui permet d’effectuer des calculs parallèles et d’avoir des données fiables.