Samenvatting Databanken III - Hadoop

TIN 2 - HoGent

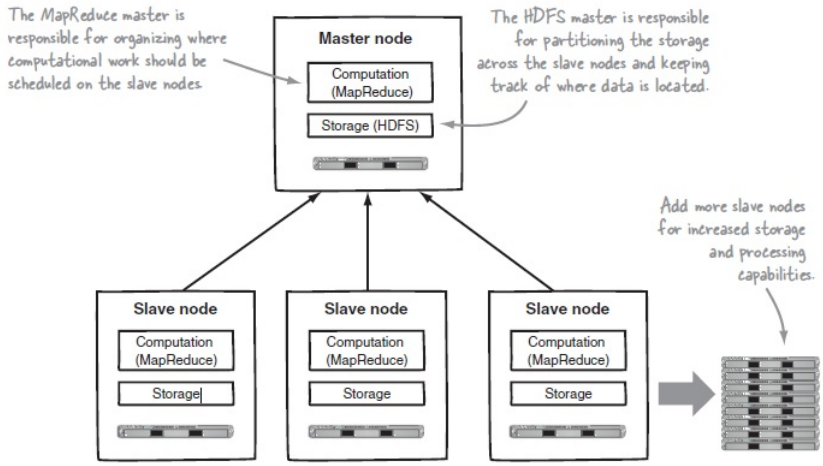
Lorenz Verschingel

24 oktober 2015

1 Intro to Hadoop

1.1 Core Hadoop components

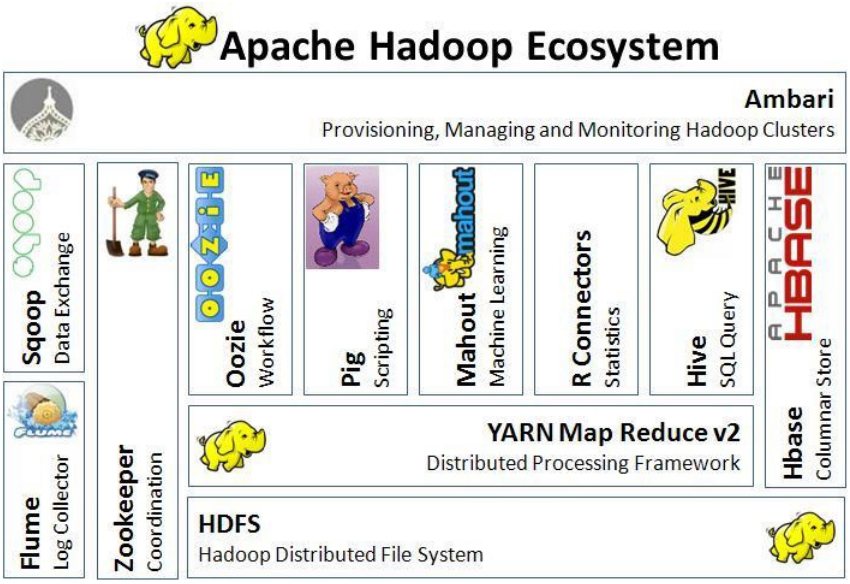
Hadoop is een project dat zich bezighoudt met het opslaan van data. Deze manier van opslaan gebeurd in het Hadoop Distributed File System (HDFS). Een manier om de data te verwerken is MapReduce. Het hoofdconcept van Hadoop is dat de data opgesplitst wordt over verschillende machines, die een cluster genoemd worden.



Figuur 1: Cluster

De verwerking van data gebeurd in waar de data opgeslagen is.

1.2 Hadoop ecosystem



Figuur 2: Hadoop Ecosystem

- **HDFS**: Storage layer
- **MapReduce**: Software layer die de verwerking doet
- **Hive**: Zorgt ervoor dat HiveQL (gelijkend aan sql) wordt omgezet naar MapReduce
- **HBase**: Zorgt voor random acces
- **Mahout**: Data-mining
- **Sqoop**: Exporteren en importeren naar relationele database
- **Zookeeper**: Coördinatie

1.3 Limitations

1. **Availability**: Opgelost in Hadoop 2.x. Voordien single-master model → single-point of failure.
2. **Security**: default is het security model disabled, hierdoor is de enige beveiliging de ownership en permissions. Kerberos kan geïntegreerd worden.
3. **HDFS**: inefficiënt voor kleine files en geen transparante compressie.
4. **MapReduce**: batch-based → kan niet gebruikt worden voor real-time toepassingen
5. **Ecosystem versus compatibilities**: version-dependency challenges to running Hadoop