Hadoop Basic Concepts and HDFS

Basic Concepts

• Hadoop is based on Google's map reduce and Google file system (GFS).

• Hadoop is first created by Doug Cutting in 2005 at Yahoo.

Main Components:

– HDFS: Distributed File System

– MapReduce: Programming Paradigm

Other tools:

– Hbase: Hadoop column database; supports batch and random reads and limited queries.

– ZooKeeper: It's a coordination service that gives you the tools you need to write correct distributed applications.

– Pig: Data processing language and execution environemnt

– Hive: Data warehouse language with SQL interface

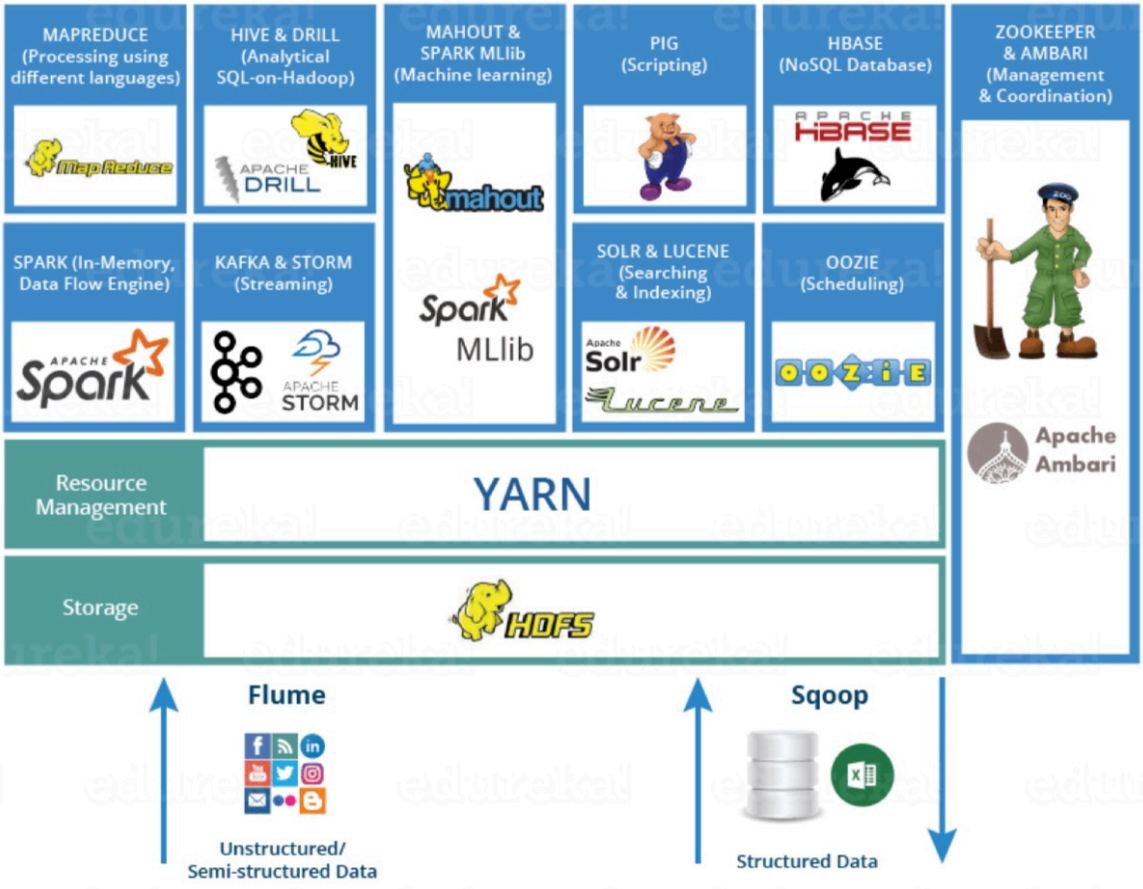
– Oozie: Workflow scheduler system for MR jobs

– Hue: Web application for interacting with Apache Hadoop.

– Impala: Low-latency data warehouse language with SQL interface

– Spark: Spark is a framework for writing fast, distributed programs.

Spark solves similar problems as Hadoop MapReduce does but with a fast in-memory approach and a clean functional style API.

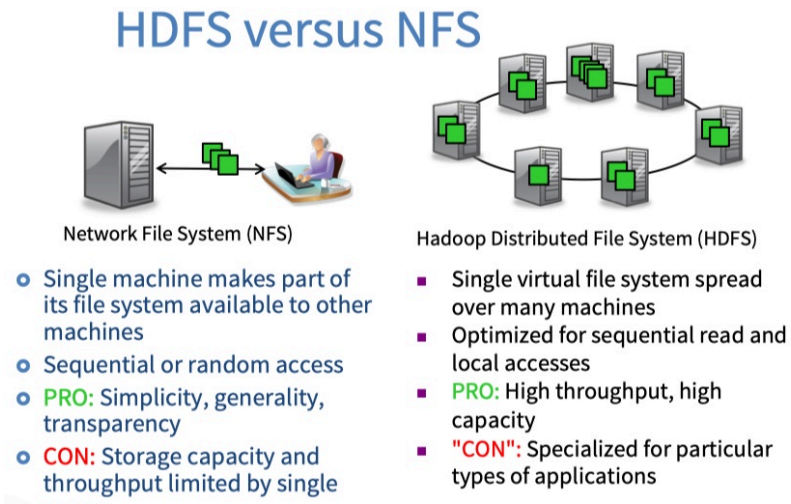


HDFS

• HDFS is the abbreviation for Hadoop Distributed File System.

• HDFS is a distributed file system that is fault tolerant, scalable and extremely easy to expand.

• HDFS is the primary distributed storage for Hadoop applications.



By default, HDFS makes 3 copies of each file.

• HDFS is optimized to support high-streaming read performance. This means that if an application is reading from HDFS, it should avoid (or at least minimize) the number of seeks.

• HDFS supports only a limited set of operations on files — writes, deletes, appends, and reads, but not updates.

• HDFS does not provide a mechanism for local caching of data. Data should simply be re-read from the source.

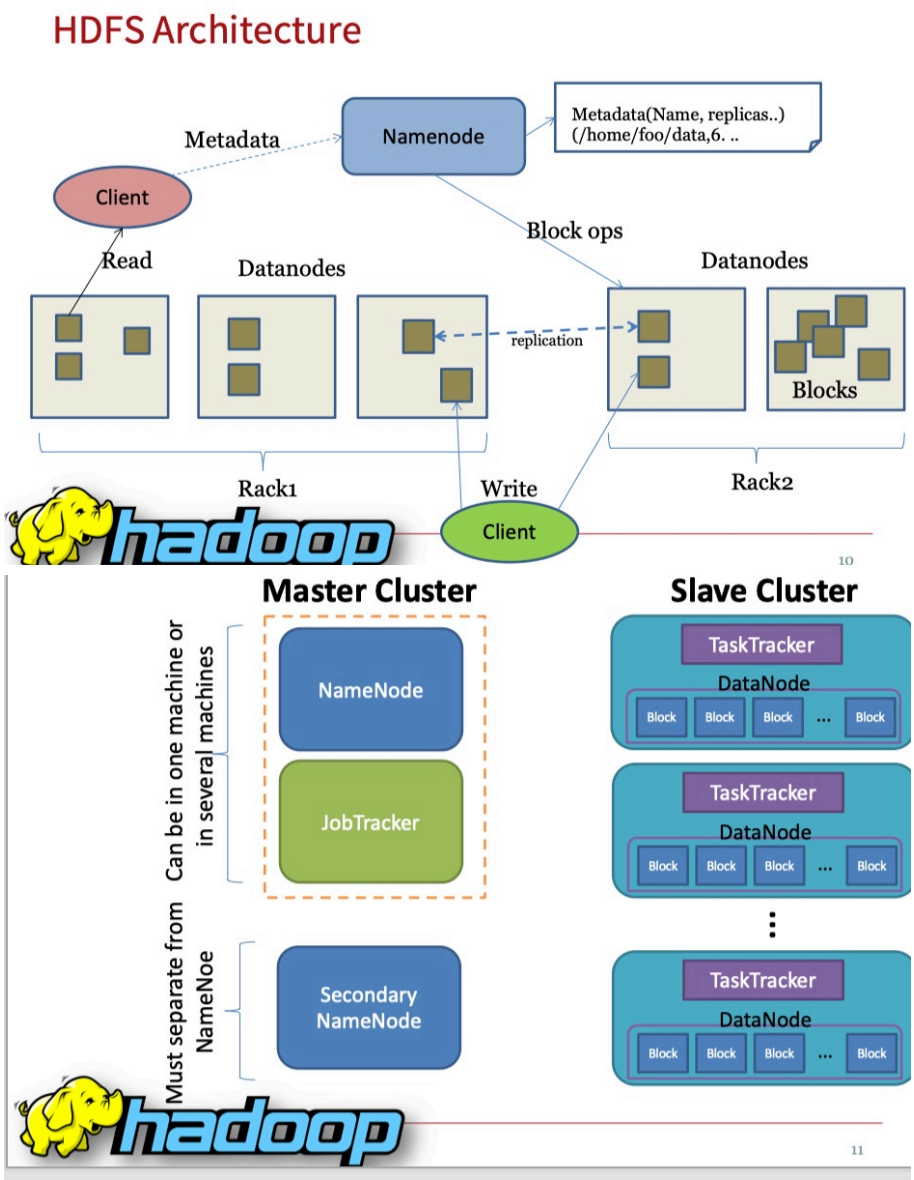
HDFS Architecture

Master/slave architecture HDFS cluster consists of a single Namenode, a master server that

manages the file system namespace and regulates access to files by clients.

There are a number of DataNodes usually one per node in a cluster. The DataNodes manage storage attached to the nodes that they run on.

DataNodes: serves read, write requests, performs block creation, deletion, and replication upon instruction from Namenode.



– List all files:

• hdfs dfs -ls

– Copy a file from local machine to HDFS:

• hdfs dfs –put <localsrc> <dest>

– Copy a file from HDFS to local machine:

• hdfs dfs –get <src> <localdest>

– Remove file from HDFS:

• hdfs dfs –rm <file>

– Remove a directory from HDFS:

• hdfs dfs –rm -r <folder>