HDFS Overview

Tushar B. Kute,
http://tusharkute.com

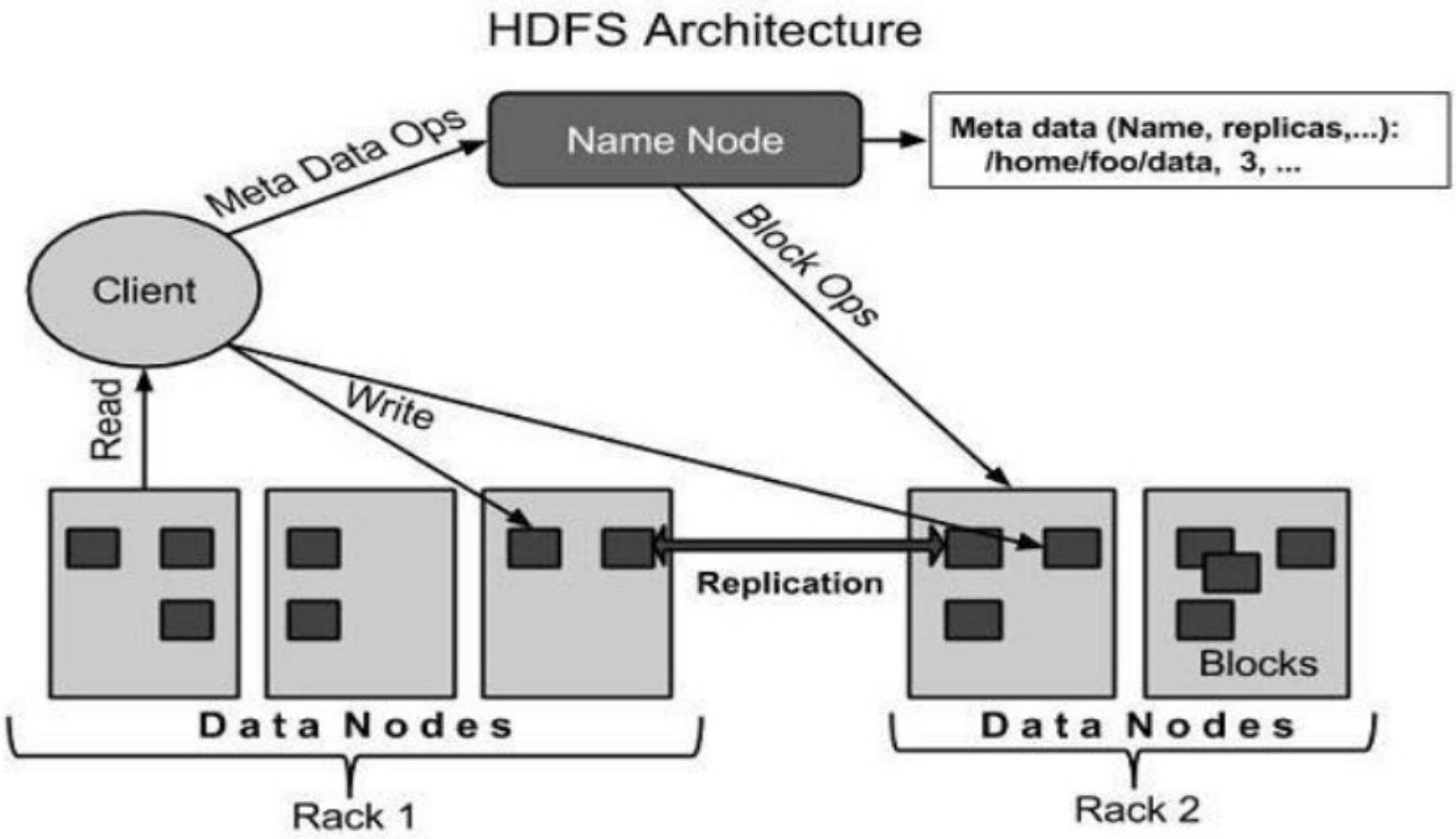
HDFS

- Hadoop File System was developed using distributed file system design.
- It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault-tolerant and designed using low-cost hardware.
- HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines.
- These files are stored in redundant fashion to rescue the system from possible data losses in case of failure.
- HDFS also makes applications available to parallel processing.

Features of HDFS

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication

HDFS Architecture



Namenode

- The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software.
- It is a software that can be run on commodity hardware.
- The system having the namenode acts as the master server and it does the following tasks:
  - Manages the file system namespace.
  - Regulates client's access to files.
  - It also executes file system operations such as renaming, closing, and opening files and directories.

Datanode

- The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/ System) in a cluster, there will be a datanode.
- These nodes manage the data storage of their system.
  - Datanodes perform read-write operations on the file systems, as per client request.
  - They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

Block

- Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes.
- These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block.
- The default block size is 128MB, but it can be increased as per the need to change in HDFS configuration.

Goals of HDFS

- Fault detection and recovery: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and automatic fault detection and recovery.
- Huge datasets: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets.
- Hardware at data: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

Thank you

Web Resources
http://mitu.co.in
http://tusharkute.com

Blogs
http://digitallocha.blogspot.in
http://kyamputar.blogspot.in

tushar@tusharkute.com