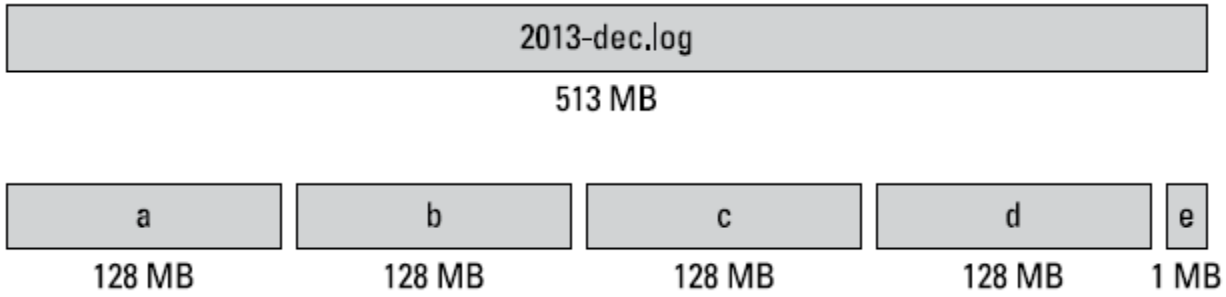
HDFS TUTORIAL

Hands-on Session

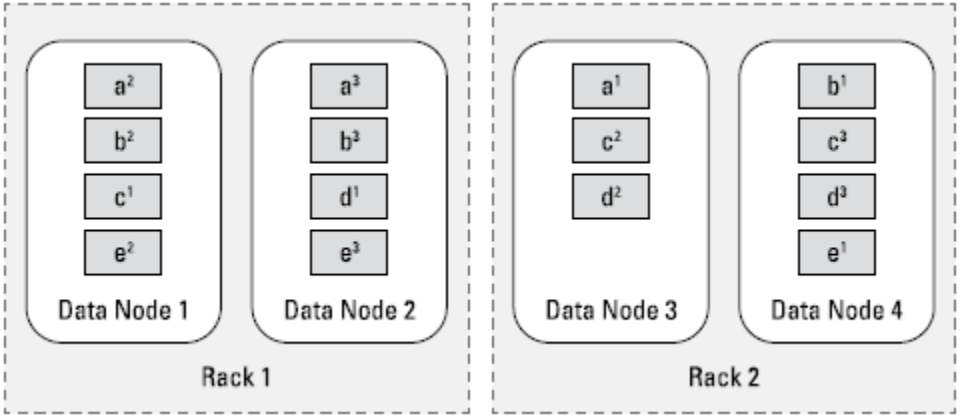
by Suchitra Jayaprakash
suchitra@cmi.ac.in

HDFS

- HDFS is distributed file system.
- It stores data by splitting files into blocks.



- Blocks are distributed across multiple nodes.
- Replicates blocks on multiple nodes.
- It provides fault tolerant storage.



(source: Hadoop for Dummies)

HDFS Architecture

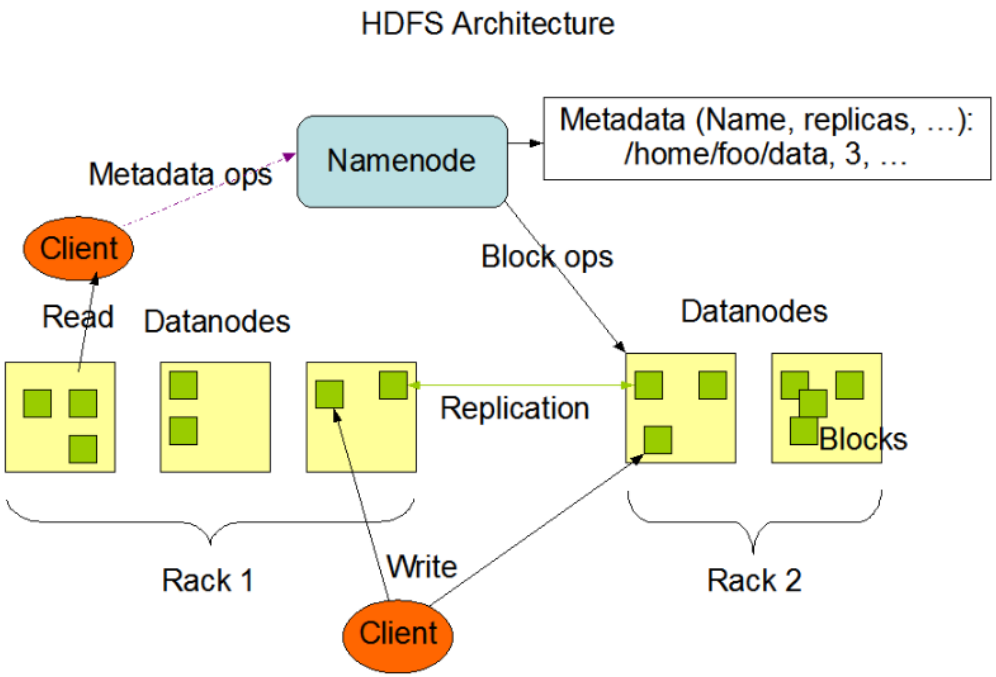
- HDFS follows master/slave architecture.

- Master Node / Name Node

  manages the file system name space (meta data) and regulates access to files by clients.

- Slave Node / Data Node

  Stores data blocks attached to a node. Block size – 64 MB to 128 MB.



- HDFS follows Write once and Read multiple times.

(source: Cloudera website)

START CLOUDERA

- Start cloudera quick start

docker run --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true -p 50070:50070 -p 8088:8088 -p 50075:50075 cloudera/quickstart /usr/bin/docker-quickstart

Port	Purpose
50070	Name node web interface
8088	job tracker :- yarn
50075	Data node

HDFS Shell Command

- mkdir command

  hadoop fs -mkdir DATA

  hadoop fs -mkdir DATA2

  It will create a new directory named DATA under the location /user/root

- LS command

  hadoop fs -ls

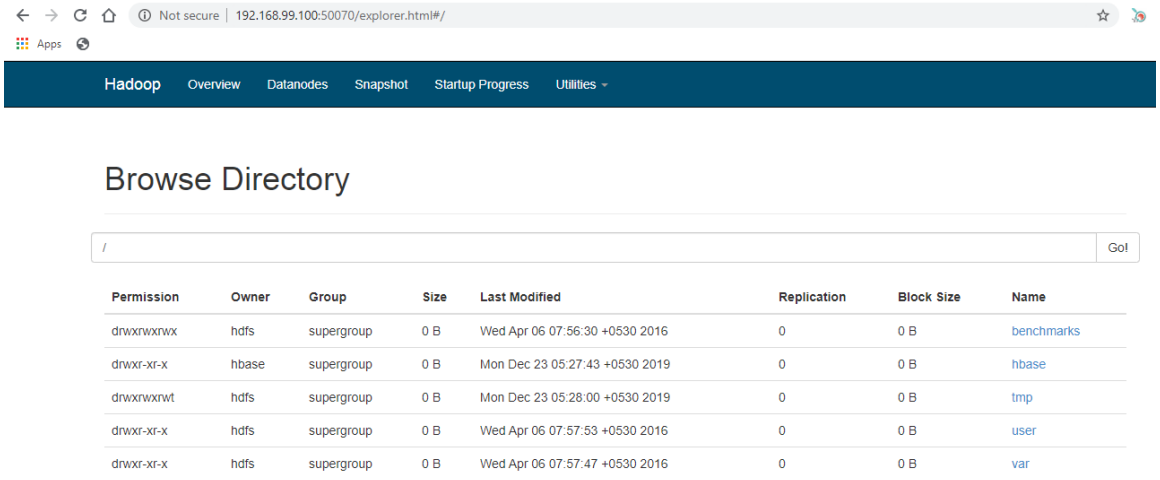
  It will list all the available files and subdirectories under default directory.

- Name node Web UI

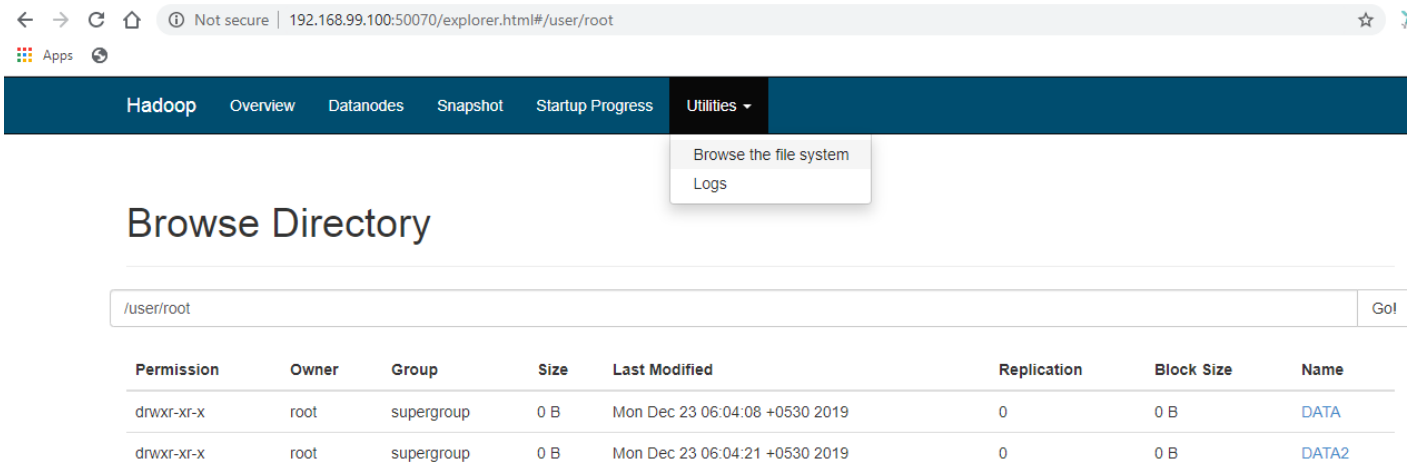
  To see name node web interface , Open http://192.168.99.100:50070/ for docker tool box or http://localhost:50070 in browser.

HDFS Shell Command

- Click on utilities : browse the file system.



- Click on user then root,



HDFS Shell Command

- Copy files from HOST

  docker cp c:/tmp/sample.txt <containerid>:/tmp/sample.txt

  copy files from host machine to docker container. (run the command in command prompt).

- Put command

  hadoop fs -put /tmp/sample.txt DATA

  It uploads files to hadoop distributed file system.

- Get command

  hadoop fs -get DATA/sample.txt

  copies the files to the local filesystem. Run ls –ltr to check file on docker container.

HDFS Shell Command

- copyFromLocal command

  hadoop fs -copyFromLocal /tmp/sample1.txt DATA

  copys file from local file system to HDFS location.

- moveFromLocal command

  hadoop fs -moveFromLocal /tmp/sample2.txt DATA

  copys file from local file system to given destination and source file is deleted.

- Print block count

  hadoop fsck /user/root/DATA -files -blocks

HDFS Shell Command

- Mv command

  hadoop fs -mv DATA/sample2.txt DATA2/sample2.txt

  move file from one directory to other directory.

- touchz command

  hadoop fs -touchz DATA2/sample3.txt

  creates files in given location.

- rm command

  hadoop fs -rm DATA2/sample3.txt

  removes the file or empty directory in the given path.

HDFS Shell Command

- tail command

  hadoop fs -tail DATA/output.txt

  Display trailing Kilobytes of file content.

- Cat command

  hadoop fs -cat DATA/sample1.txt

  Copies file content to *stdout*

- CP command

  hadoop fs -cp DATA/output.txt DATA2/output.txt

  Copy files from source to destination.

HDFS Shell Command

- DU command

  hadoop fs -du DATA

  Displays disk usage, in bytes, for all the files in the current folder.

  Stat command

  hadoop fs -stat DATA

  Prints information about folder.

  chmod command

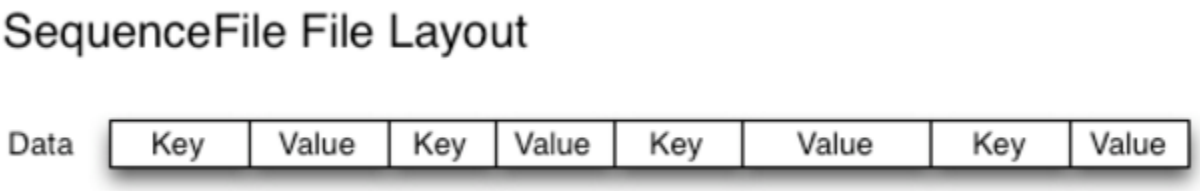
  hadoop fs -chmod -R 777 DATA

  Changes the file permissions

Small files problem

- Issues :
  - Processing too many small files is a problem in Hadoop.
  - Overhead for name node.
  - Map task process a block of input at a time. So more number of maper job.
- Solution : Sequence file format
  - Sequnce file is a flat file consisting of binary key/value pairs.
  - Filename is used as key and file content as value.
  - SequenceFile java API provides SequenceFile.Writer,

SequenceFile.Reader and SequenceFile.Sorter classes for writing, reading and sorting data.



HDFS Shell Command

- How to merge content of two text files?

hadoop fs -cat DATA/sample1.txt DATA2/sample2.txt

hadoop fs -cat DATA/* | hadoop fs -put - DATA2/output1.txt

HDFS Config files

- All HDFS related configuration are done by adding or updating properties in xml files:

  hdfs-site.xml

  core-site.xml

- Type below command to see config file

  ls -l /etc/hadoop/conf/

  less /etc/hadoop/conf/hdfs-site.xml

  less /etc/hadoop/conf/core-site.xml

THANK YOU