Big data: architectures and data analytics

HDFS and Hadoop

HDFS – command line

HDFS

- The content of a HDFS file can be accessed by means of
  - Command line commands
  - A basic web interface provided by Apache Hadoop
    - The HDFS content can only be browsed and its files downloaded from HDFS to the local file system
    - Uploading functionalities are not available
  - Vendor-specific web interfaces providing a full set of functionalities (upload, download, rename, delete, …)
    - E.g., the HUE web application of Cloudera

3

HDFS – user folder

- Each user of the Hadoop cluster has a personal folder in the HDFS file system
  - The default folder of a user is

    /user/*username*

4

HDFS – command line

- The **hdfs** command can be executed in a Linux shell to read/write/modify/delete the content of the distributed file system
- The parameters/arguments of hdfs command are used to specify the operation to execute

5

HDFS – command line

- List the content of a folder of the HDFS file system

  hdfs dfs -ls *folder*

- Example

  hdfs dfs -ls /user/garza

- shows the content (list of files and folders) of the /user/garza folder

6

HDFS – command line

- Example

  hdfs dfs -ls .
- shows the content of the home of the current user
  - i.e., the content of /user/*current_username*
  - . = user home
- The mapping between the local linux user and the user of the cluster is based on
  - A Kerberos ticket if Kerberos is active
  - Otherwise the local linux user is considered

7

HDFS – command line

- Show the content of a file of the HDFS file system

  hdfs dfs -cat *file*
- Example

  hdfs dfs -cat /user/garza/document.txt
- Shows the content of the /user/garza/document.txt file stored in HDFS

8

HDFS – command line

- Copy a file from the local file system to the HDFS file system

  hdfs dfs -put *local_file HDFS_path*
- Example

  hdfs dfs -put /data/document.txt /user/garza/
- Copy the local file /data/document.txt in the folder /user/garza of HDFS

9

HDFS – command line

- Copy a file from the HDFS file system to the local file system

  hdfs dfs -get *HDFS_path local_file*
- Example

  hdfs dfs -get /user/garza/document.txt /data/
- Copy the HDFS file /user/garza/document.txt in the local file system folder /data/

10

HDFS – command line

- Delete a file from the HDFS file system
  hdfs dfs -rm *HDFS_path*
- Example
  hdfs dfs -rm /user/garza/document.txt
- Delete from HDFS the file /user/garza/document.txt

11

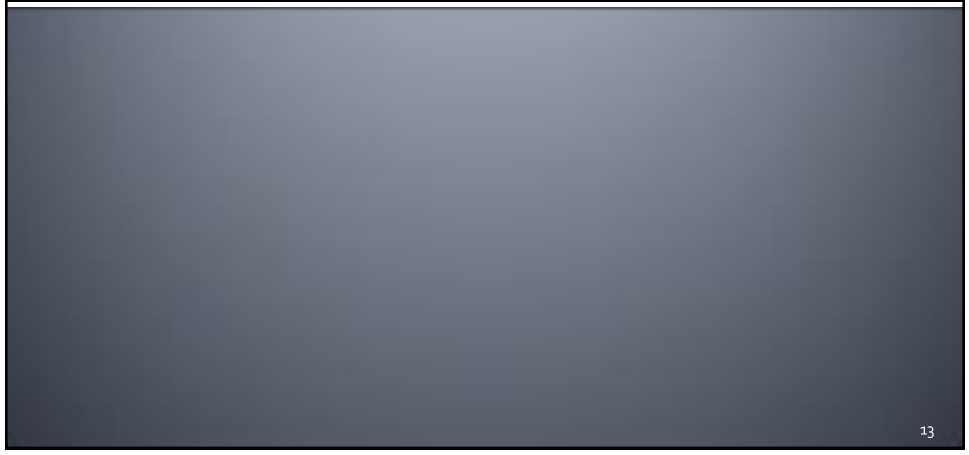
HDFS – command line

- There are many other linux-like commands
  - rmdir
  - du
  - tail
  - …
- Useful link
  - https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html

12

HDFS and Hadoop

Hadoop – command line



Hadoop – command line

- The Hadoop programs are executed (submitted to the cluster) by using the hadoop command
  - It is a command line program
  - Hadoop is characterized by a set of parameters
    - E.g., the name of the jar file containing all the classes of the MapReduce application we want to execute
    - The name of the Driver class
    - The parameters/arguments of the MapReduce application

Hadoop – command line example (1)

- The following command executes/submits a MapReduce application

hadoop jar *MyApplication.jar it.polito.bigdata.hadoop.DriverMyApplication 1 inputdatafolder/ outputdatafolder/*

- It executes/submits the application contained in MyApplication.jar

15

Hadoop – command line example (2)

- The Driver Class is

  it.polito.bigdata.hadoop.DriverMyApplication

- The application has three arguments
  - Number of reducers (args[0])
  - Input data folder (args[1])
  - Output data folder (args[2])

16