Big Data Technologies

2024-2025

Lab 2

Starting procedure

- Build the container
  `docker build -t hadoop:v1 .`
- Run the container
  `docker run -it -p 8888:8888 -p 9870:9870 -p 8032:8032 hadoop:v1`
- Do the following exercises
- At the end of the lab, just use "exit" to quit the container and stop it

➢ Questions

1. Run the command "jps" to list the running processes. What do you note?
2. What is the directory containing the HDFS configuration files?
3. Where does Namenode store its image on the local file system? What configuration file defines this location?
4. Where on the local file system will Datanode store its blocks? What is the configuration file which contains this location?
5. What is the block replication?
6. Is it possible to have some information about the Namenode via a web browser? If so, what is the "http-address"?
7. It is possible to have some information about the Datanodes via a web browser?
8. The file "core-site.xml" defines the property "fs.defaultFS". What is this property? How can we interpret its value "hdfs://0.0.0.0:9000"?

Hints:

- *To find the file "myfile" in the sandbox, you can use the command:*
  $ find / -name "myfile"
- *To read the file "myfile.txt", you can use the command: # cat myfile.txt | more*

➢ Perform

1. Create a new directory "/lab2" on HDFS.
2. Download the file "alice30.txt" from the website
   http://www.umich.edu/~umfandsf/other/ebooks/alice30.txt
   You can use the command "wget".
   This file contains the plain text of Alice's Adventures in Wonderland by Lewis Carroll.
3. Upload "alice30.txt" to HDFS under /lab2 directory under the name "alice.txt"
4. View the content of the "/lab2/" directory
5. Determine the size of the "alice.txt" file in KB that resides on HDFS (not local directory)
6. Print the first 25 lines to the screen from "alice.txt" on HDFS
7. Copy "alice.txt" to "aliceHdfsCopy.txt"
8. Copy file "alice.txt" back to local file system and name it "aliceCopy.txt"
9. Check the entire filesystem for inconsistencies/problems
10. Delete "alice.txt" from HDFS
11. Delete the "/lab2" directory from HDFS