UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Big Data Management and Analytics
Session: Distributed Storage - Hadoop I

Lecturers: Petar Jovanovic and Josep Berbegal

1 Tasks To Do Before The Session

It is important that you: (1) carefully read the **instruction sheet** for this lab session, (2) introduce yourself to the **lab's main objectives**, (3) understand the **theoretical background**, and (4) get familiar with the **tools being used**.

2 Part A: Objectives & Questions (15min)

In the first 15 minutes, we will first clarify the main objectives of this lab. We will then learn about the role of HDFS in the Big Data stack and introduce its main components. We will then overview how HDFS takes care of partitioning of input data and ensures the fault-tolerance in the cluster.

3 Part B: In-class Practice (2h 45min)

3.1 Exercise 1 (1h): Setting-up of Hadoop

Follow the enclosed guide to set up and start up a Hadoop/HDFS cluster to work on later on. Let the lecturer know when this is complete.

3.2 Exercise 2 (45min): Data generation

1. Compile the Java project you have been given. You can follow the **Eclipse Setup** instructions attached.
2. Export a runnable JAR (this JAR file is called *labo1.jar* in the examples below).
3. Upload it to your master node using the SCP client.
4. Generate, for instance, a file with 10 million rows and load it into HDFS. You are free to generate larger volumes of data, but note that will take more time on moving data from one node to another without

adding much to your knowledge. Play a little bit and check on what information is displayed in the web interface[1].

```
java -jar labo1.jar write -plainText 10000000 wines.txt
hadoop-2.7.4/bin/hdfs dfs -put wines.txt
```

3.2.1 Exercise 3 (1h): Block splitting and replication

1. With replication factor of 1:

   (a) Load again the data file you previously generated into HDFS (first, remove the previous one).

   ```
   hadoop-2.7.4/bin/hdfs dfs -rm wines.txt
   time hadoop-2.7.4/bin/hdfs dfs -D dfs.replication=1 -put wines.txt
   ```

   How long did it take?

   Answer:
   real 0m43.839s
   user 0m13.497s
   sys 0m4.010s

   (b) Now try to explore a little bit more on what has been going on.

   ```
   hadoop-2.7.4/bin/hdfs fsck /user/bdma**/wines.txt
   ```

   What is the size of the file in HDFS? How many blocks have been stored? What is the average block size? Discuss if such results make sense to you.

   Answer:
   Total size: 1314950750 B
   Total files: 1
   Total blocks (validated): 10 (avg. block size

   (c) Now log into both DataNodes and try to explore the directory where you configured Hadoop to store the data.

   ```
   du -shx data/
   ```

   How much of the file is stored on each node? Does such distribution of data benefits the parallelism of applications executing in the cluster?

   Answer:
   slave1: 748M
   slave2: 517M
   Yes, one distribution such as seen in this case, broken

[1] After starting your HDFS cluster, the HDFS web interface should be available at port 50070. Thus, you should check your email to see what public address opens to your master:50070.

2. With replication factor of 2:

   (a) Remove the previous file and repeat the same steps as with replication factor of 1.

   ```
   hadoop-2.7.4/bin/hdfs dfs -rm wines.txt
   time hadoop-2.7.4/bin/hdfs dfs -D dfs.replication=2 -put wines.txt
   ```

   Are there any differences in the results you obtain? If so, what are these differences?

   Answer:

   real 0m59.459s
   user 0m15.119s
   sys 0m4.448s

3. With replication factor of 3:

   (a) Now we start understanding how replication works, try to anticipate what is going to happen with replication factor of 3. **What volume of data you expect to be physically stored?**

   Answer: By using replication factor of 3, one of the slaves is expected to have more volume of data stored.

   (b) Repeat the same steps as with replication factor of 1 and 2. Remove the previous file.

   ```
   hadoop-2.7.4/bin/hdfs dfs -rm wines.txt
   time hadoop-2.7.4/bin/hdfs dfs -D dfs.replication=3 -put wines.txt
   ```

   (c) **Discuss the results obtained. What "under-replicated blocks" message means? Why is it showing up now?**

   Answer: Under-replicated blocks, under this circumstance, means that it was not possible to replicate the blocks to a third DataNode, since there are only two DataNodes

---

Additional comments: