STAT 4410/8416 Homework 6

Mamoundou Dramera

Due on May 8, 2022

1. **Big data tools:** The Hadoop Distributed File System (HDFS) allows us to manipulate massive amount of data using scalable computing power. Please answer the questions below based on HDFS. You don't have to show the results, just explain.

a. [2 pts] Explain what the following commands do.

```
hadoop fs -mkdir wordcount/input : It creates a input directory
hadoop fs -put myFile.txt myHdfs/test.dat :
```

First, It creates a input directory

Second, it put local file to Hadoop file system. It copies the content into test.dat from myFile.txt

b. [3 pts] Explain what the following `pig` commands will do.

```
dat = LOAD 'myHdfs/test.dat';
d = LIMIT dat 10;
DUMP d;
```

First, It loads file into pig variable called dat;

Second, It gives the limit number of tuples which is 10.

Finally, it executes the command and display.

c. [3 pts] Write down two differences between `Pig` and `Hive`. Which code will run faster?

For Pig, It is:

- a PigLatin language;
- structured and semi-structured data
- It loads data faster
- It is developed by Yahoo

For Hive:

- It uses HiveDL languages
- it uses for structured data
- it is used to create reports
- It loads data slowly
- developed by Facebook

I believe pig will run faster than Hive

d. [2 pts] If a data manipulation process takes 10 days to complete, what can you do to finish it in one day?

We can use the parallel computing to perfom the task by using 10 processors to run the data in one day