STAT 4410/8416 Homework 6

Gerjol, nicholas

Due on Dec 12, 2021

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. Explain what the following commands do.

```
hadoop fs -mkdir wordcount/input
hadoop fs -put myFile.txt myHdfs/test.dat
```

The first command here makes an input directory on the HDFS called input and the second command copies the text file and the dat file into the directory.

b. Explain what the following pig commands will do.

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

The first command loads the file, the second command takes the first 10 data elements and stores them into d, then dump outputs the 10 items in d.

c. Write down two differences between Pig and Hive. Which code will run faster?

Pig is used for data manipulation while hive is used to process structured data. Pig was developed by Yahoo and Hive was developed by facebook and Pig is faster than hive.

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

Use 10 processors to run the process.