ISOM 671: Managing Big Data (Group Assignment 1)

Name
Email

***There are 3 numbered questions*. Please submit your assignment as a single PDF or Word file by uploading it to course canvas page. *You should provide: all commands, results of any commands, and answers to questions, if any*.**

1. (10 points) Assuming a company is planning to migrate their existing file servers and databases (500TB) to HDFS platform where each datanode if of size 64TB. The company believes they will be utilizing 25% of datanode storage for intermediate tasks.

1.1. Plot a chart showing the number of datanodes needed (Y-axis) based data replication count (X-axis).

1.2. For a selected data replication count, plot a chart for change in datanode need every quarter for next 20 quarters - assuming their data increases by 2% every quarter.

1.3. Based on the AWS cost calculator (https://calculator.aws/#/createCalculator/EMR) and the number of "core/data" nodes estimated in part previous question (1.2), calculate and plot the cost associated with EMR cluster over next 20 quarters. (you can get attached storage space from EMR cluster creation page. Assume that 80% utilization and linear increase in price)

2. (10 points) Assuming you have following 4 data items on a datanode.

NodeID	DocumentID	Data
N1	D1	If life were predictable it would cease to be life, and be without flavor
N1	D2	Challenges are what make life interesting and overcoming them is what makes life meaningful
N1	D3	Life is trying things to see if they work
N1	D4	You will face many defeats in life, but never let yourself be defeated

2.1. Write and submit a mapper code (in Python or pseudocode) that creates a sorted dictionary of each word (key) and frequency (value) in each document.

2.2. Write and submit a reducer code (in Python or pseudocode) that takes the mapped dictionaries and gives the TOTAL frequency of ALL words.

2.3. Write and submit a reducer code (in Python or pseudocode) that takes the mapped dictionaries and gives the AVERAGE frequency of word "life" across all documents.

3. (10 points) Based on a use case published in a paper (http://article.nadiapub.com/IJDTA/vol9_no12/24.pdf), the authors recommend using Hive over Pig Latin. Briefly discuss the use case where Hive performed better than Pig and discuss what industries this use case would be applicable to.