Assignment-2 Total = 50pts.

Learning Outcomes

Understand the concepts of parallel and distributed databases including the map-reduce parallelism framework, NoSQL, Spark, data replication and distributed query processing.

1. Download Apache Hadoop from

https://hadoop.apache.org/ and test it [5pts + 5pts]. Take screen shots for your test and write how you tested the correct installation. Here, it is not necessary for you to demonstrate multinode cluster.

You may take the help from the book by Hadoop: Definitive Guide: Tom White and the links in the Resource1 in the folder, Week1.

2. Download Apache Spark from

https://spark.apache.org/and test it. [5pts + 5pts]

Tutorial Links.

[Big Data Processing with Apache Spark - Part 2: Spark SQL (infoq.com)](https://www.infoq.com/articles/apache-spark-sql/)

<https://mapr.com/blog/using-apache-spark-dataframes-processing-tabular-data/>

You may test out Apache Spark with excel or a text file. You may create your own data file. Take screen shots for your test and write how you tested the correct installation.

3. What are the differences in the architecture of MapReduce and Apache Spark? [5pts]

4. What is distributed query processing? Give an example. Does MapReduce and Apache Spark allow distributed processing? [ 5pts + 5pts + 5pts]. Read Chapter 22.1 and Chapter 22.10.

5. Read the paper by The Google Filesystem (See the Reading Folder, file= gfs-sosp2003.pdf).

Write a one page summary for the paper (single spaced 11pts). The audience(reader) of the paper should be a student who has the knowledge of undergraduate database and operating system. Given the page limit is one, you have to pick the main idea behind the paper and give example. [10pts]

You will upload word documents with screenshots testing of Apache Hadoop and Apache Spark. Also, the answers for question (2) and question (3) should be in the document. Upload the dataset that you created. The purpose of this homework is to learn the concepts used MapReduce and Spark.