Assignment-2 (Data Analysis using Apache Spark)

The objective of this assignment is to learn

- Installation and configuration of Apache Spark.
- Installation of HDFS
- Installation and use of different storage engine with Apache Spark

Tasks:

1. Extension of first assignment: Now use Apache Spark to answer all the queries with Postgres as the backend storage engine. You may use any technology for connecting Postgres with Apache Spark. One such method is to use JDBC. You may like to read the blog at (https://zheguang.github.io/blog/systems/2019/02/16/connect-spark-to-postgres.html)
2. MongoDB as a storage engine:
    a. Install MongoDB engine
    b. Import all of your data from the log file to MongoDB engine
    c. Connect Apache Spark to MongoDB and answer all the queries of first assignment.
3. HDFS and Apache Spark
    a. Install HDFS
    b. Copy the log file in HDFS
    c. Use Apache Spark to run queries with data in HDFS.