BUAN 6346

Big Data

Project 2

Project 2 includes the following steps. For each step follow the description and requirements of the step and provide your answers/results in the report.

- There is not a page limit for the report.
- Make sure to document every step you take and explain what you have done.

For each step you need to provide the following deliverables in your report **along with explanations**:

- **Codes**: provide all the codes you used in a clear and readable format
- **Logs**: provide all the logs (wherever applicable)
- **Screenshots**: provide screenshots (wherever applicable)

In every step it is clearly specified what type of deliverable is required in the report.

Data Analysis

In this project you will use various tools from the Hadoop ecosystem to analyze and understand the data you transferred to Hadoop in the previous project.

A. Data Analysis using Pig

In this step you will use Pig to analyze and summarize the dataset you stored on HDFS in **Project 1 Step 1.A**. Use Pig Latin scripting language to investigate and understand the data and provide the following summaries: [Deliverable: for each analysis include the Pig Latin code and results]

1. How many total blocks are there in your dataset?
2. What is the largest block height among the blocks in your dataset?
3. What is the date and time for that block?
4. What is the highest number of transactions in your blocks?

B. Data Analysis using Hive – Part 1

In this step you should use Hive/Impala to analyze the dataset you stored on HDFS in **Project 1 Step 1.B**.

1. Create a table in Hive based on the data you stored on HDFS in Project 1 Step 1.B [Deliverable: HiveQL code and result]
2. Use HiveQL to query the table and provide the following information: [Deliverable: for each analysis include the HiveQL code and results]
    a. How many records are there in the table?
    b. How many different days are there in the table?
    c. How many records per each day are there in the table?
    d. What are the symbols in the table?
    e. What is the highest price for each symbol?
    f. What is the lowest price for each symbol?
    g. What is the average price for each symbol?
    h. What is the range of price for each symbol?
    i. What is the date on which each symbol experienced the highest price?

C. Data Analysis using Hive – Part 2

In this step you should use Hive/Impala to analyze the datasets you stored on HDFS in **Project 1 Step 1.C**.

1. Use HiveQL to query the table you created in Project 1 Step 1.C and provide the following information: [Deliverable: for each analysis include the HiveQL code and result]
    a. How many total blocks are there in your blocks table?
    b. What is the largest block height among the blocks in your blocks table?
    c. What is the date and time for that block?
    d. What is the largest number of transactions in your blocks?