BUAN 6346

Big Data

Project 3

Project 3 includes 3 steps as follows. For each step follow the description and requirements of the step and provide your answers/results in the report.

- The entire report should be prepared in a Jupyter Notebook
- Export the Jupyter Notebook as a PDF and submit the PDF only
- If you cannot export as PDF, please take screen shots of your Jupyter Notebook, put them in a word document, export as PDF and submit the PDF.
- Please do not submit any zip files

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

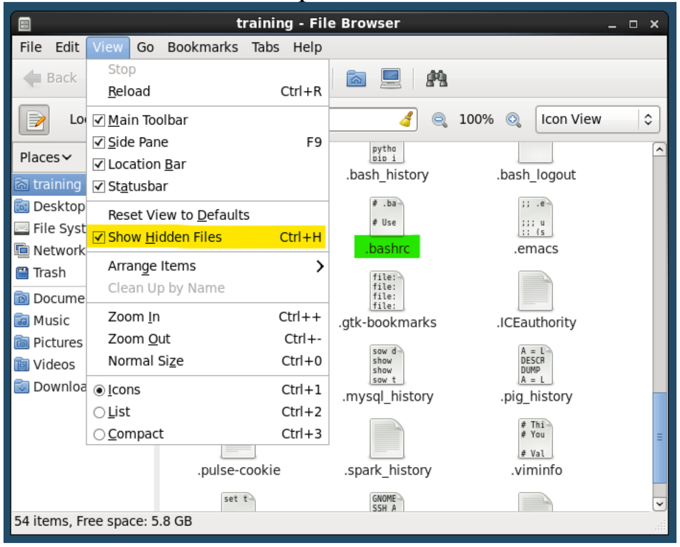
- **Codes**: provide all the codes you used in a clear and readable format
- **Outputs**: provide all the outputs generated in the Jupyter Notebook by execution of the codes
- **Explanations**: provide explanations of what you are doing.

When preparing the report in Jupyter Notebook, apply Headings and text formatting using Markdown for Jupyter Notebooks (Ultimate Markdown Guide for Jupyter Notebook).

Step 0: Jupyter Notebook Setup

In order to be able to perform the required task in this project you need to set up Jupyter Notebook in your VM. Jupyter Notebook is already available in your VM, but the `pyspark` command will open the interactive shell not the Jupyter Notebook. To make `pyspark` run open Jupyter Notebook, follow the steps below: **[No documentation required for this step.]**

1. Find the `.bashrc` file which is located in the following directory in local file system of your VM:
   `/home/training`
   the file is hidden so you need to enable the option 'Show Hidden Files' to see it.



2. Edit the file by uncommenting the following line (remove the # sign before the line) in the file:
   `export PYSPARK_DRIVER_PYTHON_OPTS='notebook --ip 127.0.0.1 --port 3333 --no-mathjax'`
3. Save the file and restart your terminal.
4. In the terminal type `pyspark` and hit Enter. This should now open Jupyter Notebook in your browser.

5. You can now create a new Python notebook from the Jupyter Notebook page and start coding.

Step 2: Data Analysis Using Spark Core

In this step you will use Spark Core to perform analysis on the data you transferred to Hadoop in the previous project. This means that you need to create RDDs from your data and perform operations on them to find the answers to the questions.

A. Blockchain Data Analysis – Part 1

In this step you will use Spark RDDs to analyze and summarize the dataset you stored on HDFS in step **1.A** of the previous project. Use RDDs and their suitable operations to provide the following summaries:

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. Stock Market Data Analysis

In this step you will use Spark RDDs to analyze and summarize the dataset you stored on HDFS in step **1.B** of the previous project. Use RDDs and their suitable operations to provide the following summaries.

1. How many records are there in the table?
2. How many different days are there in the table?
3. How many records per each day are there in the table?
4. What are the symbols in the table?
5. What is the highest price for each symbol?
6. What is the lowest price for each symbol?
7. What is the average price for each symbol?
8. What is the range of price for each symbol?
9. What is the date on which each symbol experienced the highest price?

C. Blockchain Data Analysis – Part 2

In this step you will use Spark RDDs to analyze and summarize the dataset you stored on HDFS in step **1.C** of the previous project. Use RDDs and their suitable operations to provide the following summaries.

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

Step 3: Data Analysis Using Spark SQL

In this step you will use Spark SQL to perform analysis on the data you transferred to Hadoop in the previous project. This means that you should only use Spark Dataframes and their operations to find the answers to the questions.

A. Blockchain Data Analysis – Part 1

In this step you will use Spark SQL to analyze and summarize the dataset you stored on HDFS in step **1.A** of the previous project. Use Dataframes and their suitable operations to provide the following summaries:

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. Stock Market Data Analysis

In this step you will use Spark SQL to analyze and summarize the dataset you stored on HDFS in step **1.B** of the previous project. Use Dataframes and their suitable operations to provide the following summaries.

1. How many records are there in the table?
2. How many different days are there in the table?
3. How many records per each day are there in the table?
4. What are the symbols in the table?
5. What is the highest price for each symbol?
6. What is the lowest price for each symbol?
7. What is the average price for each symbol?
8. What is the range of price for each symbol?
9. What is the date on which each symbol experienced the highest price?

C. Blockchain Data Analysis – Part 2

In this step you will use Spark SQL to analyze and summarize the dataset you stored on MySQL Database in step **1.C** of the previous project. Use Dataframes and their suitable operations to provide the following summaries (Please Note that you need to create Dataframes directly from the data stored on MySQL Database not on HDFS).

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