AfterWork

Apache Spark DataFrames Project

Project Deliverable

You will be required to submit:

- A GitHub repository with your project written in Pyspark.

Instructions

As a Data professional, you need to perform an analysis by answering questions about some stock market data on Safaricom from the years 2012-2017.

You will need to perform the following:

Data Importation and Exploration

- Start a spark session and load the stock file while inferring the data types.
- Determine the column names
- Make observations about the schema.
- Show the first 5 rows
- Use the describe method to learn about the data frame

Data Preparation

- Format all the data to 2 decimal places i.e. format_number()
- Create a new data frame with a column called HV Ratio that is the ratio of the High Price versus volume of stock traded for a day

Data Analysis

- What day had the Peak High in Price?
- What is the mean of the Close column?
- What is the max and min of the Volume column?
- How many days was the Close lower than 60 dollars?
- What percentage of the time was the High greater than 80 dollars?
- What is the Pearson correlation between High and Volume?
- What is the max High per year?
- What is the average Close for each Calendar Month?

Data description

- Dataset URL (CSV File): https://bit.ly/3pmchka