CCU

Big Data Analytics - Fall 2024

Exercise 2: Spark Walmart Data Analysis

Due Date: 2025/11/17 23:59 pm

Target

In this exercise we will get some practices with the usage of Spark DataFrame , you need use Spark to find the correct answer of asked questions about some stock market data, in this case Walmart Stock from the years 2012-2017.

Abstract

The goals of this assignment are to

(1) become familiar with Spark, PySpark and Spark DataFrame, and have the ability to set up the environment for development Spark applications

In this exercise, you need to first follow the steps in the course slides to properly set up the PySpark development environment. Then, follow the instructions below to download the practice dataset from E-course or GitHub and answer the final question..

Exercises 1

1. ***Download the dataset(a or b)***
    a. From E-course: https://ecourse2.ccu.edu.tw/pluginfile.php/1454707/mod_resource/content/0/walmart_stock.csv
    b. From Github: https://github.com/pratikbarjatya/spark-walmart-data-analysis-exercise/blob/master/walmart_stock.csv

2. ***Write A PySpark program to find the answers for following***

questions

a. Please finish all the questions(9 questions) during **pages 79-82** in **week6_PySpark.pptx (60%)**
b. **Please answer the following two additional questions:**
    i. What is the max price spread ***(max High – min Low in one year)*** per year **(20%)**
    ii. What is the Pearson correlation between Volume and HV Ratio? ***(20%)***

Homework Submissions

Exercises 1:

For all questions in exercise 1, kindly ensure to capture and include copies or screenshots of your coding process and the resulting outputs in your comprehensive report, which can be prepared using either Microsoft Word or PowerPoint. Once your report is finalized and ready, please proceed to upload it directly to the E-course platform for submission and further evaluation. Your attention to detail and adherence to these instructions are highly appreciated.