Deree
School of Graduate
and Professional
Education

Big Data Architectures
Winter 2023

Assignment #3

Due: Sunday March 19 2023, 23:59.

Exercise 1 (25 points)

Consider the file BGF-WTF-A2-EUR_fund1.csv, which contains historical data on an investment fund (source: https://www.blackrock.com/uk/individual/products/230009/blackrock-world-technology-a2-eur-fund). Each row of the file contains the fund's data since its inception date as follows:

1. Date
2. Closing NAV (Net Asset Value) per share
3. NAV change from previous day
4. Percentage NAV change from previous day

For example, if the closing NAV on April 1, 2021 was 66,21 and on March 31, 2021 was 64,28, the NAV change is 1,93 and the % NAV change is 1,93/64,28 * 100 % = 3,00 (%).

Write a pyspark application in file BGF-WTF-A2-EUR_fund.py that uses plain Spark methods (i.e., no Spark dataframes or Spark SQL) to read the file BGF-WTF-A2-EUR_fund1.csv and calculate the following.

1. The maximum and minimum daily change and percentage change for the full history of the Fund.
2. The average NAV and the standard deviation of NAV for the full history of the Fund.
3. The average NAV and standard deviation for 2020
4. The average NAV per month for the full history of the Fund, with the most recent month listed first.

Exercise 2 (25 points)

Consider the file ontime_flights.csv that contains historical data on US flights for April 2014. Write a Spark application (no use of Spark Dataframes or Spark SQL) in file ontime_flights.py to display

1. The airline with the maximum total (departure and arrival) delay
2. The average departure delays out of JFK
3. The total delays (departure and arrival) per day
4. The average total delays of American Airlines (AA)
5. The number of flights flown from JFK to LAX per day

Exercise 3 (25 points)

Consider the file earthquakes.csv that contains historical data on earthquakes that struck Greece since 1/7/1965. Write a Spark application in file earthquakes.py using Spark Dataframes to display

1. The date, time, and magnitude of the ten most powerful earthquakes in Greece.
2. The number of earthquakes per year that struck after 2000.
3. The minimum, maximum and average magnitude of earthquakes per year between 2010 and 2020 inclusive.
4. The five most powerful earthquakes that struck the Athens area (latitude between 37.5 and 39.0 and longitude between 23.35 and 23.55)

Exercise 4 (25 points)

Consider the file all_stocks_5yr.csv that contains S&P historical data (2013 – 2018). Write a Spark application using Spark Data Frames in file *all_stocks_5yr.py* to display

1. The dates with the highest volume of transactions per ticker (column Name)
2. The average volume of transactions per month
3. The average volume of transactions per ticker
4. The dates and the highest spread (high – low) per ticker
5. The average stock price (close) per ticker for 2016
6. All ticker symbols in the dataset

Submission

Put all .py files into a rar or zip compressed file with name <Lastname>_<Firstname>.rar or .zip (e.g., Nikolaou_Maria.zip) and submit it to Blackboard.

Penalties

Violation of any naming conventions will result into 25 points reduced from your grade.