Deree
School of Graduate and Professional Education

Big Data Architectures
Winter 2024

Final Exam

April 2, 2024, 18:30 – 21:30.

Total 100 points. Answer all questions.

Question 1 (20 points)

Consider the producers-consumers problem that was presented in class. Write a Python program in file *producers_consumers.py* that uses semaphores and extends the one presented in class by spawning N producers and $2N$ consumers that use a shared buffer *buf*. Each producer iterates $2N$ times and produces items to *buf* and each consumer iterates N times and consumes items from *buf*. Producers and consumers must synchronize their access and actions to *buf* and must print their actions in each iteration ("produce item *x* to *buf*" or "consume item *x* from *buf*", respectively).

Question 2 (20 points)

A number is *perfect* if the sum of its divisors excluding itself is equal to the number itself. For example, 6 is perfect since 1+2+3=6.

Write a Python program in file *perfect_numbers.py* that uses the *ThreadPoolExecutor* class to find the number of perfect numbers that are < *max_num* and report on its timing requirements. Summarize your results, i.e., the size of the problem ($max\_num = 1000, 10000, 100000, 1000000$) vs. the number of treads $(4, 8, 16)$.

Write the same program in file *perfect_numbers_t.py* without use of *ThreadPoolExecutor* by using plain Python threads.

Question 3 (20 points)

Consider the file BTC.csv that contains historical data on the Bitcoin prices in US dollars since its inception. Write a Map-Reduce application in file *btc.py* to output:

1. The average closing price of BTC in 2020.
2. The number of dates for which the closing price was over 60000 USD.
3. The number of days for which the closing price was less than the average of the highest and lowest price.
4. The ten highest closing prices and the corresponding dates.
5. The year and month for which the average closing price was between 20000 and 22000 USD.

The application should use a single mapper and a single reducer.

Question 4 (20 points)

Consider the file CoffeeShop.csv that contains historical sales data of a coffee retailer with several shops in New York City. Write a plain Spark application (no use of Spark Data Frames or Spark SQL) in file *coffee_Shop.py* to display

1. The number of transactions per month.
2. The sum of the total bills per month.
3. The number of transactions in June 2023 which contained Coffee in product_category.
4. The average bill amount for the Astoria store.
5. The number of transactions for which Brazilian-Organic (product_detail attribute) Coffee Beans were sold in the Lower Manhattan store.

No use of built-in methods and functions for calculating max, min, average, standard deviation is allowed.

Question 5 (20 points)

Answer the following questions.

1. (4 points) Assume two sensors that monitor the health of a press machine in a factory assembly line send their data to a Kafka topic *PressData* for predictive maintenance analysis. The order of data of each sensor and across the two sensors is important to draw valid conclusions. Two applications A1 and A2 perform independent analytics processing of the sensor data. How many partitions should topic *PressData* have and how many consumer groups should be created for this scenario?
2. (4 points) Consider a distributed banking application that is required to be available 24/7/365. What property must be sacrificed under this assumption and why?
3. (4points) Explain what is Eventual Consistency. Give an example of an application for which eventual consistency is acceptable.
4. (4 points) What is the best speedup that can be achieved for a program for which 25% of its execution is parallelizable when it runs on a machine with 3 processors? Can this speedup be achieved in practice? Explain your answer.
5. (4 points) Synchronization: Consider the dining philosophers problem that was presented in class. Each philosopher spends her life as

```
Forever
    Wakeup
    Grab left chopstick
    Grab right chopstick
    Eat
```

Put down right chopstick
Put down left chopstick
Sleep

Give an execution scenario that may result to an undesirable state assuming three philosophers. Explain what is the reason of reaching this state and give a solution to overcome the problem.

Put the answers to Question 5 in file Question5.pdf. Use the template Question5.doc.

Submission

Put all .py files and Question5.pdf into a rar or zip compressed file with name <Lastname>_<Firstname>.rar or .zip (e.g., Efremidis_Sofoklis.zip) and submit it to Blackboard. Only submissions to Blackboard will be graded.

Penalties

Violation of any naming convention(s) will result in 20 points reduced from your grade.

Submitting anything but the .py files and Question5.pdf will result in 20 points reduced from your grade.