CSP 554 – Big Data Technologies

Spring 2023 – All Sections

Final Exam - Sample Questions

Part I – Short Answer (Show Points/Results) – 5 points each, 30 points total

1. Given the following lineage of RDDs in a Spark Context: ‘*data = file.filter(lambda s: s.contains(‘ERROR’)).map(lambda s: s.split(‘|’)[2]).count()*’ which operations are Transformations and which are Actions? When referencing the resulting RDD in another operation, how can we ensure these results are not recomputed each time?

2. A messaging platform based on Kafka will be deployed for consumer banking transactions. A messaging topic is defined for user account operations consisting of: 1) deposits, and 2) withdrawals consisting of the Account ID and an Amount. In order to speed-up processing, it is suggested to partition this topic by Account ID. Will this result in incorrect account balances/state? Why or why not?

3. A distributed database management system (DBMS) allows for a client to continue performing inserts and updates to records even in the case of a network partition. Identify whether this is a CP, AP, or CA system in terms of the CAP Theorem. What corresponding Write Quorum level would this be under the PACELC Theorem?

Part V – Long Answer (Show Reasoning/Calculations) – 10 points each, 20 points total

1. Given a JSON data object representing a patient electronic medical record with the below (partial/incomplete) structure, outline how this record would be represented in each of the following systems: 1) Key-Value Store, 2) Wide-Column Store, and 3) Document Store:

```
‘{
        name: Mario,
        occupation: Plumber,
        diagnosis: {
                        name: Mushroom Poisoning,
                        date: 2023-03-01
                }
}’
```

2. Given two Spark DataFrames with the below columns and sample rows, provide an overview/sketch of how one can obtaining an average of Rating values by Country. Provide two solutions: 1) Using Spark SQL commands and 2) Using Spark Transformation functions. Your code does not have to be syntactically correct - however, you must explain your reasoning.

DF1	DF2
UserId, Address, City, State, Country	MovieId, UserId, Rating
345, 3100 S State, Chicago, IL, USA	1507, 345, 4
346, 188 Regent St, London, W1B 5BT, UK	1508, 221, 2
347, 400 9th Ave, New York, NY, USA	1509, 247, 5