# CSP 554 – Big Data Technologies

## Fall 2023 – All Sections

### Final Exam

**Part I** – Short Answer (Show Points/Results) – 10 points each, 60 points total

1. A distributed database management system (DBMS) allows for a client to perform writes of records in the case of a network Partition. Identify whether this is a CP, AP, or CA system in terms of the CAP Theorem. How would this designation change if the system is reconfigured to only allow for reads of records during a network Partition?
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 given a tuple of (UserID, Account ID, Amount). In order to speed-up processing, it is suggested to partition this topic by UserID. Will this result in incorrect balances/state for for a given AccountID? Why or why not? **Note**: Multiple users *may* perform deposits/withdrawals for a single account.
3. Provide a brief description of the computed result from the below lineage of RDDs in a Spark Context. Which operations are Transformations and which are Actions? Identify which Transformations are Narrow and which ones are Wide.

‘c *= file.flatMap(lambda r: r.split(‘ ‘))*

*.map(lambda w: (w,1))*

*.reduceByKey(lambda x,y: x+y)*

*.map(lambda: (k,v): (v,k))*

*.sortByKey(False)*

*.take(10)*’

1. A tunable distributed DBMS allows for specification of Read Quorum (R) and Write Quorum (W). If the cluster consists of 10 nodes with a replication factor of N=3, what are the minimum values needed for R and W in order to prioritize Availability? In terms of the PCELC theorem, what trade-off is being made during normal operations?
2. Given a Kafka topic for stock prices with N=500 partitions (by StockID), we wish to increase throughput within a consumer group of G=500 workers. Should we decrease or increase the number of workers? If we decrease the number of partitions to have multiple StockIDs, will latency increase or decrease per StockID? Why?
3. A log file contains a series of entries with the below timestamps and content. Given a Spark Structured Stream with fixed interval micro-batches configured to 1 second, what would be the resulting sequence of DataFrames at 1 second intervals? If we configured our trigger to 1 minute, what would be the resulting sequence of DataFrames at 1 minute intervals?

2023-11-30 | 11:00:30 | INFO | mem=4G

2023-11-30 | 11:00:30 | WARN | mem=8G

2023-11-30 | 11:00:35 | ERROR | mem=16G

2023-11-30 | 11:05:50 | INFO | mem=4G

**Part II** – Long Answer (Show Reasoning/Calculations) – 20 points each, 40 points total

1. Given a JSON data object representing a stock market company 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:

‘{

symbol: INTC,

region: North America,

industry: Semiconductors,

competitors: {

symbol: TSMC,

region: Asia

}

}’

1. Given two Spark DataFrames with the below columns and sample rows, provide an overview/sketch of how one can obtaining an average of Amount spent per AppId 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 AppId, UserId, Amount

345, 3100 S State, Chicago, IL, USA 1507, 345, 471.99

346, 188 Regent St, London, W1B 5BT, UK 1508, 221, 225.00

347, 400 9th Ave, New York, NY, USA 1509, 247, 504.50