Neha Maddali

## CPRE/SE 419 Software Tools for Large Scale Data Analytics

## Spring 2023

## Take-home Final Exam

## Due: Wednesday, May 10, at 11:59PM

### Part I.A – Provide brief/concise answer (with justification):

1. **(5 pts.) What are the long running daemons in YARN?**
   Running daemons in YARN include ResourceManager (RM), NodeManager (NM), and ApplicationMaster (AM). RM is responsible for managing the allocation of resources and scheduling of tasks across the cluster. NM runs on each node in the cluster and is responsible for managing the resources allocated to it by the RM. AM is responsible for managing the execution of a single application on the cluster. RM and NM are YARN's long running daemons.

2. **(5 pts.) In the context of CAP theorem for distributed data systems, give an example of a database that belongs to the "CP" part of the spectrum.**
   An example of a database that belongs to the CP part of the spectrum is MongoDB which is a NoSQL.

3. **(5 pts.) Explain briefly the difference between NoSQL and NewSQL.**
   NoSQL is non-relational and highly scalable, but NewSQL is relational and designed for large-scale transactional workloads since it has SQL as the primary mechanism for the application interaction. NewSQL has ACID support and NoSQL provides CAP.

4. **(5 pts.) What are the main features of EDA (Event Driven Architectures)**
   EDA is an approach to software design that emphasizes the production, detection and consumption of events in order to create robust and flexible software systems. The main features of EDA are events, asynchronous communication, event driven workflows, loose coupling, and fault tolerance. EDA provides a flexible approach to software design that can enable the create of complex, scalable and reliable systems.

5. **(5 pts.) Explain briefly the turnstile models of data streams.**
   The turnstile model is a type of data stream model that is used to track changes in the aggregate value of a dataset over time. It allows processing of streams with positive and negative updates, allowing it to estimate properties of an underlying vector based on a sequence of updates. The turnstile model is helpful for real-time analytics since it can account for changing values and provide accurate insights.

Neha Maddali

**Part 1.B – Explain the following (please try not to be overly-verbose (i.e. sufficiently detailed, but concise discussions)):**

1. **(9 pts.) Explain the concept of a partitioner in MapReduce (i.e., what is its use), and give an example of types of Partitioners readily available for MR jobs in HDFS.**
   Partitioners in MR are responsible for distributing the intermediate key-value pairs among reducers so that it can make sure that there is an even workload distribution and minimize network congestion. An example of a readily available partitioner for MR jobs in HDFS is the HashPartitioner which uses a hash function to determine the reducer assignment.

2. **(9 pts.) In the context of distributed databases, explain the Bully Algorithm for coordinator selection.**
   The Bully Algorithm is a coordinator selection method in distributed systems. It is where the process with the highest identifier is elected as the coordinator. Processes with a lower identifier can initiate an election, however if a process with a higher identifier detects the election, it takes over and becomes the coordinator.

3. **(9 pts.) Define the concept of RDD in Spark. What are its main benefits?**
   RDD is an immutable, fault-tolerant distributed collection of objects that can be processed in parallel across a cluster of machines. RDDs can be made from data stored in HDFS, from data stored in other distributed storage systems, or from data generated in memory or on disk. The main benefits include distributed processing, fault tolerance, in-memory processing, lazy evaluation, and transformation and action operations.

4. **(9 pts.) What is the difference between tumbling window and sliding window?**
   Tumbling windows are useful when we want to analyze non-overlapping segments of the data stream, while sliding windows are useful when we want to analyze overlapping segments of the data stream.

**Part III – Algorithmic questions**

1. **(14 pts.) Consider the following input stream: A, B, C, E, A, A, A, D, F, E, F, F, F, B, B, C, C, D, F, F, F, What is the outcome of executing Misra-Gries algorithms for selecting the (approximate) heavy hitters in the stream, when the capacity of the map is limited to c = 3. Justify/explain your answer.**
   Initialize a frequency map.
   For each item coming from the stream:
   - If current element is in frequency map, increase the frequency counter by 1
   - If the current element is not in the frequency map and the map is not full, where the number of elements is less than the capacity, add the current element to the map with a frequency counter of 1
   - If current element is not in the frequency map and the map is full, where the number of elements is equal to the capacity, decrease the frequency counter of

all elements in the map. If a counter reaches 0, remove the corresponding element from the map. The current element will not be added to the map.

This algorithm can be implemented in python like so:

```
def function(stream, cap):
    freq_map = {}
    for el in stream:
        if el in freq_map:
            freq_map[el] += 1
        elif len(freq_map) < cap:
            freq_map[el] = 1
        else:
            for x in list(freq_map.keys()):
                freq_map[x] -= 1
                if freq_map[x] ==0:
                    del freq_map[x]
    return freq_map
data_stream = [‘A’, ‘B’, ‘C’, ‘E’, ‘A’, ‘A’, ‘A’, ‘D’, ‘F’, ‘E’, ‘F’, ‘F’, ‘F’, ‘B’, ‘B’, ‘C’,
‘C’, ‘D’, ‘F’, ‘F’, ‘F’]
cap = 3
```

The outcome of the frequency map is {‘A’: 4, ‘F’: 7, ‘B’: 3}

2. **(13 pts.) Consider a scenario in which a distinct sampling is required, for a sample of size k=3. Assume that following are the values for the hash function h(A) = 0.7; h(B) = 0.8; h(C) = 0.4; h(D) = 0.9; h(E) = 0.3. Show the content of the sample for each arrival of an element in the following stream: A, D, D, A, B, A, A, A, C, C, E, A, B, A, D.**

| Arriving element | Sample 1 | Sample 2 | Sample 3 |
|---|---|---|---|
| A | A(0.7) | | |
| D | A(0.7) | D(0.9) | |
| D | A(0.7) | D(0.9) | |
| A | A(0.7) | D(0.9) | |
| B | A(0.7) | D(0.9) | B(0.8) |
| A | A(0.7) | D(0.9) | B(0.8) |
| A | A(0.7) | D(0.9) | B(0.8) |
| A | A(0.7) | D(0.9) | B(0.8) |
| C | A(0.7) | C(0.4) | B(0.8) |
| C | A(0.7) | C(0.4) | B(0.8) |
| E | A(0.7) | C(0.4) | E(0.3) |
| A | A(0.7) | C(0.4) | E(0.3) |
| B | B(0.8) | C(0.4) | E(0.3) |
| A | A(0.7) | C(0.4) | E(0.3) |
| D | D(0.9) | C(0.4) | E(0.3) |

3. **(14 pts.) Consider the composite event C ≡ A; B (i.e., C is detected whenever an occurrence of the primitive event "A" is followed by an occurrence of the primitive event "B"). Consider the following stream of events – where the first parameter is the primitive event, and the 2nd parameter is the time-instant of its detection (for example: e(B,4) denotes that the primitive event B was detected at time-stamp 4): e(A,1), e(A, 2), e(A,3), e(B,4), e(A,5), e(B,6) How many detections of "C" will occur under the recent context and how many under chronicle context for selecting the participating primitive events (illustrate/justify your answer)?**

   Under recent context, we only consider the most recent occurrence of each primitive event. For primitive event A, the most recent occurrence is 5. For primitive event B the most recent occurrence is 6. Since we have a recent occurrence of A at time 5, followed by a recent occurrence of B at time 6, there is a detection of C. Therefore, there is 1 detection of C under the recent context.

   Under chronicle context, we consider all occurrences of each primitive event in the stream. For primitive event A, occurrence times were 1, 2, 3, 5. For primitive event B, occurrence times were 4, 6. Look at each occurrence of A and check if there is a subsequent occurrence of B. If there is, then there is a detection of C. We can identify two detections of C: occurrence of A at time 1 followed by occurrence of B at time 4, and occurrence of A at time 5 followed by occurrence of B at time 6. So, there are 2 detections of C under the chronicle context.

4. **(5 pts.) For this problem, you will need to have each member of a team work on it separately, and you will need to provide the individual answers! The purpose is to combine your skills in file-searching/scanning and user's preferences matching. Specifically, you are to execute the following request:**
   a. **Go through the "Contacts" in your mobile phone;**
   b. **Select your favorite person for executing the activity in Step III below;**
   c. **Upon completing all the other problems from this assignment, call and tell your favorite person1 from Step II above in a loud and clear manner: "I am done with 419!!!"**
   d. **Write the answer that you received.**
   Response from my friend Saachi Dalvi: "YAY! Congrats, you're finally done with finals!"