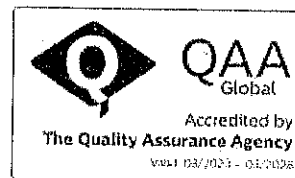




The University of Fiji
(An Entity of Arya Pratinidhi Sabha of Fiji)



SCHOOL OF SCIENCE AND TECHNOLOGY
Semester 2: Year 2023

FINAL EXAMINATION PAPER

ITC333: DISTRIBUTED SYSTEMS

Important Instructions to Students:

1. Read the questions, instructions carefully and answer the questions.
2. You are allowed an **extra 10 minutes reading time** (during this time, you are not allowed to write with pen) and **180 minutes (3 Hours) to complete this paper**.
3. Silent, non-programmable calculators are allowed, but not supplied.
4. Write all your answers in the answer booklet provided. If you need extra sheets, please ask from the supervisor. Write your name, course code and student identification number in the answer booklet and in the attendance sheet.
5. **This Final Examination Paper is worth 100 marks contributes 50% of the total course assessments.** Students must **at least score 40%** in this final exam to pass the course.
6. This paper contains one (1) section A. All questions are compulsory to answer.
7. Write all your answers legibly. The marker has the right to award zero for answers that are not legible.
8. This is a **Closed Book Examination**. The University views plagiarism, cheating and dishonest practice as serious offences and if found, offenders are penalized.

Examination Paper Content (Format):

<u>Section</u>	<u>Question</u>	<u>Content</u>	<u>Marks</u>	<u>Suggested Time</u>
A	1-10	Short Answer Questions	100	180 minutes
		TOTAL	100	180 Minutes

Question 1: The Hadoop Ecosystem**[6 Marks]**

- A. What is Hadoop, and what are the key reasons for implementing a Hadoop system in the context of big data processing and storage? **[2 Marks]**
- B. List and provide explanations for two applications in Hadoop that utilize the MapReduce framework. **[2 Marks]**
- C. List and explain the default NoSQL database included with Hadoop. **[2 Marks]**

Question 2: HDFS and MapReduce**[15 Marks]**

- A. What is HDFS, and how does it store data on a cluster? **[2 Marks]**
- B. Discuss the HDFS architecture and explain the process used by HDFS for reading and writing a file. **[3 Marks]**
- C. List and explain the four strategies used for ensuring Namenode resilience in Hadoop. **[2 Marks]**
- D. What is MapReduce, and how does it process data within a cluster in Hadoop? **[2 Marks]**

Table 1: Sample Movie Lens Data

USER ID	MOVIE ID	RATING	TIMESTAMP
196	242	3	881250949
186	302	3	891717742
196	377	1	878887116
244	51	2	880606923
166	346	1	886397596
186	474	4	884182806
186	265	2	886397596

- E. Describe how MapReduce can be used to determine the number of movies rated by each user using the provided Movie Lens data in Table 1. *Clearly explain the Mapper, Shuffle and Sort, and Reducer steps.* **[6 Marks]**

Question 3: Apache Pig**[8 Marks]**

- A. What is the primary advantage that Apache Pig provides compared to MapReduce? **[2 Marks]**
- B. What delimiter is the following Pig script using? **[2 Marks]**

```
ratings = LOAD '/user/maria_dev/ml-100k/u.data' AS (userID:int, movieID:int, rating:int, ratingTime:int);
```

- C. Rewrite the script in (B) assuming that it uses the “|” delimiter. **[2 Marks]**
- D. Explain the following line of code. **[2 Marks]**

```
fiveStarMovies = FILTER avgRatings BY avgRating > 4.0;
```

Question 4: Apache Spark**[9 Marks]**

- A. What is Apache Spark? Why is Apache Spark so efficient in processing data on a cluster? **[2 Marks]**
- B. List and explain the four (4) main components of Apache Spark. **[2 Marks]**
- C. What is a Spark RDD? List and explain four (4) functions for transforming RDD's. **[3 Marks]**
- D. What are the advantages of using a DataFrame over an RDD object in Apache Spark? **[2 Marks]**

Question 5: Relational datastores**[10 Marks]**

- A. What is Hive? **[1 Mark]**
- B. List and explain six (6) reasons for using Hive in big data processing. **[3 Marks]**
- C. List and explain four (4) reasons for NOT using Hive. **[2 Marks]**
- D. What is the difference between “LOAD DATA” and “LOAD DATA LOCAL” in Hive. **[2 Marks]**
- E. What is Sqoop and how does it handle Big Data? **[2 Marks]**

- A. List and explain four (4) techniques that can be employed on SQL/relational databases to handle Big Data. **[2 Marks]**
- B. List and explain the key components of HBase. **[3 Marks]**
- C. Explain the replicated ring architecture used by Cassandra. **[2 Marks]**
- D. What are the major limitations of CQL in contrast to SQL. **[2 Marks]**
- E. List and explain four (4) advantages of using MongoDB. **[2 Marks]**
- F. What are five (4) considerations to keep in mind while choosing a database for your project? **[4 Marks]**
- G. Consider the following requirements and suggest the best NoSQL database. **[1 Mark]**
- You have a big Spark job that produces movie recommendations for end users nightly.
 - Something needs to vend this data to your web applications.
 - You work for some huge company with massive scale.
 - Downtime is not tolerated.
 - Must be fast.
 - Eventual consistency is acceptable.

- A. What is Apache Drill? **[2 Mark]**
- B. What is Phoenix? **[2 Mark]**
- C. Explain the Phoenix architecture. **[2 Marks]**
- D. List and explain four (4) advantages of using Presto. **[2 Marks]**

- A. What is YARN and how does it work? **[2 Marks]**
- B. Explain how Tez improves the performance of MapReduce jobs. **[2 Marks]**
- C. List and explain the major differences between Mesos and YARN. **[2 Marks]**

- D. How does Myriad help with the integration of YARN and Mesos? **[2 Marks]**
- E. What is ZooKeeper? **[2 Marks]**
- F. Explain the difference between Persistent and ephemeral znodes in Zookeeper. **[2 Marks]**
- G. Explain the steps to set up a workflow in Oozie. **[2 Marks]**

Question 9: Feeding Data into your Cluster

[6 Marks]

- A. Explain the Kafka architecture in detail. **[3 Marks]**
- B. What are the components of an agent in Flume? **[3 Marks]**

Question 10: Feeding Data into your Cluster

[8 Marks]

- A. List and explain the advantages of using Spark Streaming. **[2 Marks]**
- B. Explain the difference between Batch interval, slide interval, and window interval. **[3 Marks]**
- C. Compare and contrast between Spark Streaming, Storm and Flink? **[3 Marks]**

THE END

