UNIRAZAK
UNIVERSITI TUN ABDUL RAZAK

# FINAL EXAMINATION
# MARCH 2024

---

|  |  |
|---|---|
| **COURSE TITLE** | **BIG DATA & ALGORITHM** |
| **COURSE CODE** | **RBAN3263** |
| **DATE/DAY** | **29 JUNE 2024 / SATURDAY** |
| **TIME/DURATION** | **05:00 PM - 07:00 PM / 02 Hour(s) 00 Minute(s)** |

---

**INSTRUCTIONS TO CANDIDATES :**

1. Please read the instruction under each section carefully.
2. Candidates are reminded not to bring into examination hall/room any form of written materials or electronic gadget except for stationery that is permitted by the Invigilator.
3. Students who are caught breaching the Examination Rules and Regulation will be charged with an academic dishonesty and if found guilty of the offence, the maximum penalty is expulsion from the University.

(This Question Paper consists of 5 Printed Pages including front page)

*****DO NOT OPEN THE QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO*****

**This question paper consists of TWO (2) sections. Answer ALL questions. [100 MARKS]**

**SECTION A** **(20 Marks)**

**There are TEN (10) questions in this section. Answer ALL questions in the answer booklet provided.**

1. What is the goal of big data preprocessing?

   A. To increase data variability
   B. To reduce data storage requirements
   C. To maximize data volume
   D. To prepare data for analysis and modelling

2. What is a common goal when using machine learning for big data applications?

   A. Reducing the volume of data
   B. Achieving high accuracy and speed
   C. Increasing data skew
   D. Maintaining high redundancy

3. Which of the following big data application scenarios involves continuous data collection and processing to forecast future events?

   A. Transactional data management
   B. Predictive Analytics
   C. Data warehousing
   D. Static data analysis

4. What kind of data is in Log files?

   A. Structured
   B. Unstructured
   C. Semi-structured
   D. Metadata

5. Which of the following are example(s) of real-time big data preprocessing?

   A. Complex Event Processing (CEP) platforms
   B. Stock market data analysis
   C. Bank fraud transactions detection
   D. Both A and C

6. Which of the following is a characteristic of Hadoop's MapReduce model?

   A. Real-time processing
   B. Batch Processing
   C. Iterative Algorithms
   D. In-memory Processing

7. What is a common challenge when working with data set repositories containing large volumes of unstructured data such as text and images?

   A. Data retrieval speeds may be significantly slower
   B. Data is usually stored in a single format, limiting flexibility
   C. Storing large volumes of structured data alongside unstructured data
   D. Lack of standard metadata for unstructured data

8. The classification model is trained on a big data set and performs well on the training set but poorly on the test set. What is the most likely issue with the model?

   A. Underfitting
   B. Overfitting
   C. Lack of data diversity
   D. Model complexity is too low

9. You are working with a large dataset with both numeric and categorical features. Which preprocessing technique would you apply before using a fuzzy model?

   A. Remove categorical features
   B. Standardize numeric features
   C. Convert all features to binary values
   D. Convert all features to numeric values

10. You encounter a dataset with a high level of data skew during classification model training. What approach can help mitigate this issue?

   A. Random forest algorithm
   B. One-hot encoding
   C. Cross-validation
   D. SMOTE (Synthetic Minority Over-sampling Technique)

**SECTION B** **(60 Marks)**

**Answer ALL questions.**

**Question 1** **(30 marks)**

Much has been written about Big Data and the need for advanced analytics within industry, academia, and government. The availability of new data sources and the rise of more complex analytical opportunities have created a need to rethink existing data architectures to enable analytics that take advantage of Big Data.

i. What are the -**THREE (3)**characteristics of Big Data, and what are the main considerations in processing Big Data? (8 marks)

ii. Explain the differences between Business Intelligence and Data Science. (8 marks)

iii. What are the key skill sets and behavioral characteristics of a data scientist? (6 marks)

iv. Define the difference between structured and unstructured data. Support with examples. (8 marks)

**Question 2** **(30 marks)**

When comparing big data algorithms in Hadoop versus Spark, it is important to consider the fundamental differences between the two platforms, as well as how they handle data processing and algorithm implementation.

i. Describe the architectural differences between Hadoop and Spark. (10 marks)

ii. Compare the ease of use and development experience in Hadoop and Spark. (10 marks)

iii. Identify use cases where Hadoop may be preferred over Spark, and vice versa. Provide examples of applications or scenarios where each platform excels. (10 marks)

*** **END OF QUESTION PAPER** ***