UNIVERSITI TEKNOLOGI PETRONAS

FINAL EXAMINATION
MAY 2024 SEMESTER

COURSE   :   **TEB3113 - BIG DATA ANALYTICS**

DATE   :   **12 AUGUST 2024 (MONDAY)**

TIME   :   **9:00 AM - 12:00 NOON (3 HOURS)**

INSTRUCTIONS TO CANDIDATES

1.   Answer **ALL** questions in the Answer Booklet.

2.   Begin **EACH** answer on a new page in the Answer Booklet.

3.   Indicate clearly answers that are cancelled, if any.

4.   Where applicable, show clearly steps taken in arriving at the solutions and indicate **ALL** assumptions, if any.

5.   **DO NOT** open this Question Booklet until instructed.

Note :

    i.   There are **SIX (6)** pages in this Question Booklet including the cover page

    ii. **DOUBLE-SIDED** Question Booklet.

1. a. Describe the following terminology:

    i. governance

    [2 marks]

    ii. velocity

    [2 marks]

    iii. provenance

    [2 marks]

    b. Explain the following data types of data with **ONE (1)** example each.

    i. Human generated data

    [4 marks]

    ii. Machine generated data

    [4 marks]

    c. Differentiate between conceptual data models versus logical data models for UTP Library system.

    [6 marks]

2. a. Analyse the importance of high veracity in predictive analytics system development with **ONE (1)** example.

[5 marks]

b. Salif EM Corp is a petroleum company that is currently exploring predictive analytics potential for optimal mitigating actions on pipeline corrosion issues. As Data Scientist (DS), you believed Cross-Industry Standard Process model for Data Mining (CRISP-DM) is the most suitable process model to be used.

Recommend **TWO (2)** activities for each of the below CRISP-DM phases based on pipeline predictive analytics:

i. Data Preparation

[5 marks]

ii. Modelling

[5 marks]

iii. Testing and evaluation

[5 marks]

3. RichArc Automation, a global leader in industrial automation and digital transformation solutions, collaborates with oil and gas companies to monitor and maintain expensive exploration assets. These assets, critical to operations, include drilling rigs, pipelines, and offshore platforms. Predictive maintenance is crucial in this sector to prevent costly downtime and ensure operational efficiency. Oil and gas exploration assets operate in harsh environments, facing challenges such as corrosion, mechanical wear, and unpredictable weather conditions. Failure of these assets can lead to significant financial losses due to downtime, maintenance costs, and potential environmental risks. Traditional maintenance approaches based on fixed schedules or reactive responses were inadequate for the complex and dynamic conditions these assets operate under. By leveraging IoT, analytics, and machine learning, companies in the oil and gas sector can proactively manage asset health, reduce costs, and maintain competitive advantage in the global market.

a. Examine **FIVE (5)** main advantages for RichArc Automation from deploying predictive maintenance. Justify your answer.

[10 marks]

b. Recommend a solution on how RichArc Automation utilizes IoT technology to enhance predictive maintenance in oil and gas exploration assets.

[5 marks]

c. Analyse suitable data visualization to integrate predictive maintenance insights into the overall maintenance management system. Justify your answer.

[5 marks]

4. Abu Bakar Corp, an e-commerce company, plans to harness the capabilities of Hadoop and MapReduce to efficiently handle and extract valuable insights from its vast data volumes. The company aims to distribute the workload across a Hadoop cluster and leverage the parallel processing power of MapReduce. Illustrate using a diagram the functions of each MapReduce stages:

   i.   Map

[5 marks]

   ii. Shuffle

[5 marks]

   iii. Reduce

[5 marks]

   iv. Combine

[5 marks]

5.   HealthCare Co., a leading provider of healthcare services, relies heavily on data to drive its operations, from patient care to administrative functions and strategic decision-making. As the company expanded, it faced significant challenges related to data quality. These issues were affecting patient outcomes, operational efficiency, and regulatory compliance. HealthCare Co. was encountering discrepancies in patient records, incomplete data entries, and inaccurate information due to manual data entry processes and disparate data sources. This lack of data integrity led to delays in patient care, billing errors, and difficulties in complying with healthcare regulations.

a.   Discuss the following data quality assessment dimensions for HealthCare Co., with ONE (1) example each.

   i. Accuracy

[5 marks]

   ii. Completeness

[5 marks]

b.   Discuss **FIVE (5)** ways how HealthCare Co. can implement prescriptive analytics.

[10 marks]

-END OF PAPER-