Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2022/2023

COURSE NAME	:	INTRODUCTION TO BIG DATA
COURSE CODE	:	BEE 40903
PROGRAMME CODE	:	BEE / BEJ
EXAMINATION DATE	:	JULY/AUGUST 2023
DURATION	:	3 HOURS

INSTRUCTION

1. ANSWER ALL QUESTIONS
2. THIS FINAL EXAMINATION IS CONDUCTED VIA CLOSED BOOK.
3. STUDENTS ARE PROHIBITED TO CONSULT THEIR OWN MATERIAL OR ANY EXTERNAL RESOURCES DURING THE EXAMINATION CONDUCTED VIA CLOSED BOOK

THIS QUESTION PAPER CONSISTS OF SEVEN (6) PAGES

Part A. Objectives Question

Answer all questions.

Q1. The estimated volume of data that will be processed by Big Data solutions is significant and expected to continue to grow. Choose from the following, Data in which bytes size is called Big Data.

A. Tera
B. Giga
C. Peta
D. Meta

(1 Mark)

Q2. To enables data scientists to extract more value from their data while also enabling the scientists' organizations to become more customer centric as a result of their knowledge, State the V's of Big Data should a data scientists know?

A. 2
B. 3
C. 4
D. 5

(1 Mark)

Q3. In computers, a X is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence. Information can be derived from X in computing if the X provides a symbolic representation of facts or concepts from which some probability can be calculated. While the summarizing of very large X sets might result in smaller X sets that are primarily composed of symbolic X, symbolic X are distinct in their own right on any sized X set, no matter how large or tiny it is. Solve what is X.

A. Data
B. Knowledge
C. Program
D. Algorithm

(1 Mark)

Q4. In Big Data environments, relate in what Variety of data includes

A. Includes multiple formats and types of data
B. Includes structured data in the form of financial transactions,
C. Includes semi-structured data in the form of emails and unstructured data in the form of images
D. All of the mentioned above

(1 Mark)

Q5. Select in Which of the following are Benefits of Big Data Processing?

A. Cost Reduction
B. Time Reductions
C. Smarter Business Decisions
D. All of the mentioned above

(1 Mark)

Q6. Decide the following statement; Structured data conforms to a data model or schema and is often stored in tabular form. Structured data is data that has been organized according to a data model or schema and is frequently kept in tabular format. Due to the fact that it is used to record relationships between distinct things, it is most typically kept in a relational database. Enterprise applications and information systems, such as ERP and CRM systems, are frequently responsible for the generation of structured data.

A. True
B. False

(1 Mark)

Q7. SQL cannot be used to process or query this Data. Point out from the following, Data that does not conform to a data model or data schema is known as

A. Structured data
B. Unstructured data
C. Semi-structured data
D. All of the mentioned above

(1 Mark)

Q8. Amongst which of the following is/are not Big Data Technologies?

A. Apache Hadoop
B. Apache Spark
C. Apache Kafka
D. Apache Pytarch

(1 Mark)

Q9. Infer what is involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

A. Parallel data processing
B. Single channel processing
C. Multi data processing
D. None of the mentioned above

(1 Mark)

Q10. Amongst which of the following can be considered as the main source of unstructured data.

A. Twitter
B. Facebook
C. Webpages
D. All of the mentioned above

(1 Mark)

Q11. Choose amongst which of the following shows an example of unstructured data,

A. Students roll number, age
B. Videos
C. Audio files
D. Both B and C

(1 Mark)

Q12. Big data deals with high-volume, high-velocity and high-variety information assets,

A. True
B. False

(1 Mark)

Q13. Evaluate what reporting and visualization enables.

A. Processing of data
B. User friendly representation
C. Both A and B
D. None of the mentioned above

(1 Mark)

Q14. Infer in which Data interpretation refers to

A. Process of attaching meaning to the data
B. Convert text into insightful information
C. Effective conclusion
D. All of the mentioned above

(1 Mark)

Q15. The significance of metadata is to provide information about a dataset's characteristics and structure.

A. True
B. False

(1 Mark)

Q16. Data throttling refers to the performance of a solution is throttled.

A. True
B. False

(1 Mark)

Q17. The Big data analytics work on the unstructured data, where no specific pattern of the data is defined.

A. True
B. False

(1 Mark)

Q18. Select which is a platform for developing data flows for the extraction, transformation, and loading (ETL) of huge datasets, as well as for data analysis.

A. Spark
B. HBase
C. Hive
D. Pig

(1 Mark)

Q19. Custom extensions built in what programming language are also supported by Hive.

A. Java
B. C#
C. C
D. C++

(1 Mark)

Q20. In the context of big data analytics, Apache Hive is a distributed, fault-tolerant data warehousing system that can handle huge amounts of data. A data warehouse is a centralized repository of information that can be easily evaluated in order to make data-driven decisions that are informed. Hive lets users to read, write, and manage petabytes of data using SQL, which makes it a powerful tool for data scientists. In short in order to analyze all of this Big Data, Hive is a tool that has been developed. Dictate the above statement.

A. True
B. False

(1 Mark)

Part B. Subjective questions.

Answer all questions

Q21 (a) Define big data and explain its significance in today's digital landscape.

(5 marks)

(b) Discuss the three main characteristics that differentiate big data from traditional data analysis and give examples to illustrate each characteristic.

(15 marks)

Q22 Missing and outlier data are the most challenging in performing data analysis.

(a) Differentiate between missing and outlier data

(4 marks)

(b) State two indicators that the datasets contains outlier

(2 marks)

(c) Give example how missing and outlier can affect decision making

(4 marks)

(d) In database, there are two technologies that are widely used. In Big Data, NoSQL has become a well-known database for application. Based on your understanding, discuss TWO (2) benefits of using NoSQL as compared to SQL

(10 marks)

Q23 (a) State any TWO (2) common types of data visualization and their usage

(10 marks)

(b) Describe the following components in package ggplot in R

(i) Data

(2 marks)

(ii) Aesthetic

(3 marks)

(iii) Layers

(3 marks)

-END OF QUESTIONS –