UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2021/2022

COURSE NAME	:	INTRODUCTION TO BIG DATA
COURSE CODE	:	BEE 40903
PROGRAMME CODE	:	BEE / BEJ
EXAMINATION DATE	:	JULY 2022
DURATION	:	3HOURS

INSTRUCTION

1. ANSWER ALL QUESTIONS
2. THIS FINAL EXAMINATION IS AN **ONLINE** ASSESSMENT AND 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 **TWELVE (12)** PAGES

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. Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media.

A. True
B. False

(1 Mark)

Q4. 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)

Q5. In Big Data environments, relate in what Velocity refers to

A. Data can arrive at fast speed
B. Enormous datasets can accumulate within very short periods of time
C. Velocity of data translates into the amount of time it takes for the data to be processed
D. All of the mentioned above

(1 Mark)

Q6. 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)

Q7. In Big Data environment, relate in what Veracity of data refers to

A. Quality or fidelity of data
B. Large size of the data that cannot be process
C. Small size of the data that can easily process
D. All of the mentioned above

(1 Mark)

Q8. 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)

Q9. 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)

Q10. 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)

Q11. 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)

Q12. 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)

Q13. 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)

Q14. 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)

Q15. Scalability, elasticity, resource pooling, self-service, low cost and fault tolerance are the features one of the following. It also refer to refers to the technology that remotely receives this data and conducts any actions that have been specified on that information.

A. Cloud computing
B. Power BI
C. System development
D. None of the mentioned above

(1 Mark)

Q16. Amongst which of the following is/are the cloud deployment models,

A. Public Cloud
B. Private Cloud
C. Hybrid Cloud
D. All of the mentioned above

(1 Mark)

Q17. Virtualization separates resources and services from the underlying physical delivery environment that they are delivered in. Big data virtualization is a method that focuses on the creation of virtual structures for large-scale data storage and processing environments. The usage of big data virtualization can be beneficial to businesses and other organizations because it helps them to make use of all the data assets, they have collected in order to achieve a variety of goals and objectives. Identify the statement as;

A. True
B. False

(1 Mark)

Q18. What is a Virtual Machine (VM)?

A. Virtual representation of a physical computer
B. Virtual representation of a logical computer
C. Virtual System Integration
D. All of the mentioned above

(1 Mark)

Q19. In the given Virtual Architecture figure Q19, infer the missing layer,

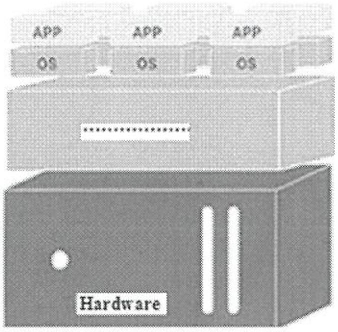

Figure Q19

A. Virtualization layer
B. Storage layer
C. Abstract layer
D. None of the mentioned above

(1 Mark)

Q20. Big data deals with high-volume, high-velocity and high-variety information assets,

A. True
B. False

(1 Mark)

Q21. Distinguish which type hypervisor runs directly on the underlying host system. It is also known as "Native Hypervisor" or "Bare metal hypervisor".

A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor
C. Both A and B
D. None of the mentioned above

(1 Mark)

Q22. Distinguish which type hypervisor is also known as "Hosted Hypervisor".

A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor
C. Both A and B
D. None of the mentioned above

(1 Mark)

Q23. Select In which the layered architecture of Big Data Stack, Interfaces and feeds,

A. Internally managed data
B. Data feeds from external sources.
C. It provides access to each and every layer & components of big data stack
D. All of the mentioned above

(1 Mark)

Q24. Select what is the supporting physical infrastructure is fundamental to the operation and scalability of big data architecture.

A. Redundant physical infrastructure
B. Integrated System
C. Integrated Database
D. All of the mentioned above

(1 Mark)

Q25. The physical infrastructure of a big data is based on a distributed computing model. Justify the above statement

A. True
B. False

(1 Mark)

Q26. Security infrastructure refers the data about your constituents needs to be protected to which of the following.

A. Meet compliance requirements
B. Protect the privacy
C. Both A and B
D. None of the mentioned above

(1 Mark)

Q27. 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)

Q28. 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)

Q29. The significance of metadata is to provide information about a dataset's characteristics and structure. Justify the statement.

A. True
B. False

(1 Mark)

Q30. Justify the following statement: Data throttling refers to the performance of a solution is throttled,

A. True
B. False

(1 Mark)

Q31. Justify the following statement: The Big data analytics work on the unstructured data, where no specific pattern of the data is defined.

A. True
B. False

(1 Mark)

Q32. Select Amongst which of the following represents the Use of Hadoop,

A. Robust and Scalable
B. Affordable and Cost Effective
C. Adaptive and Flexible
D. All of the mentioned above

(1 Mark)

Q33. 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)

Q34. In contrast to relational databases, Hive is a query engine that supports the elements of SQL that are specifically designed for querying data.

A. True
B. False

(1 Mark)

Q35. Custom extensions built in what programming language are also supported by Hive.

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

(1 Mark)

Q36. 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)

Q37. Hadoop is a framework that can be used in conjunction with a number of related products. Identify among the most common cohorts are ______.

A. MapReduce, Hive and HBase
B. Hive, Spark and HBase
C. Spark, Hive and ZooKeeper
D. Spark, HBase and Hive

(1 Mark)

Q38. _____ is a shell utility that can be used to run Hive queries in either interactive or batch mode, depending on the situation.

A. $HIVE_HOME/bin/hive
B. $HIVE/bin/
C. $HIVE_HOME/hive
D. All of the mentioned above

(1 Mark)

Q39. Amongst which of the following is/are true with reference to User-defined Functions of Hive.

A. function that fetches one or more columns from a row as arguments
B. It returns a single value
C. Both A and B
D. None of the mentioned above

(1 Mark)

Q40. Choose in which manner HDFS operates

A. Master-slave architecture
B. Master-worker architecture
C. Worker-slave architecture
D. All of the mentioned above

(1 Mark)

Q41. Compare amongst which of the following is not aligns as a characteristic of HDFS?

A. HDFS file system is well suited for storing data associated with applications that require low latency data access.
B. HDFS is well-suited for storing data connected to applications that require low-latency data access to be performed.
C. HDFS is not suited for instances in which multiple/simultaneous writes to the same file are required.
D. None of the mentioned above

(1 Mark)

Q42. Descriptive analytics is a statistical method that is used to search and summarize X in order to identify patterns or meaning. In this study, we conduct a reflective analysis of learner data with the goal of providing insight into past patterns of behaviour and performance in online learning environments. Propose what is X

A. Account data
B. Historical data
C. Financial data
D. None of the mentioned above

(1 Mark)

Q43. From the following, select two techniques used in descriptive analytics to discover historical data.

A. Data ingestion and data mining
B. Data warehouse and data storage
C. Data aggregation and data mining
D. All of the mentioned above

(1 Mark)

Q44. Select amongst which of the following is / are true with reference to hypothesis?

A. A statement that the researcher wishes to put to the test using the information gathered during a study.
B. A research question that will be answered as a result of the findings.
C. A theory that serves as the foundation for the research.
D. the application of statistics to determine the extent to which the outcomes could have been caused by chance

(1 Mark)

Q45. Prescriptive analytics makes the use of machine learning to help Business organizations to decide a course of action based on a computer program's predictions. Justify with example supporting this.

(2 Mark)

Q46. The following are the goodness of prescriptive analytics,

- Exhausting valuable resources on housing data that does not inform business decisions
- Spending time sifting through unutilized data sets
- Missing out on unique revenue streams and insights

Conclude your opinion on above statement

(2 Mark)

Q47. Large Volume of data is considered as big data and Velocity is the speed in which data is process and becomes accessible, with example, discuss the statement.

(2 Mark)

-END OF QUESTIONS –

FINAL EXAMINATION

SEMESTER / SESSION : SEM II 2021/2022 PROGRAMMECODE : BEE/BEJ
COURSE NAME: INTRODUCION TO BIG DATA COURSE CODE : BEE 40903