Question Paper: Big Data Analytics

Course: MPBA G 517

The duration of the exam is 90 minutes.

Multi-choice Questions

Instructions

- *This section contains 45 multiple-choice questions.*
- *Each question has four options, out of which only one is correct.*
- *Mark the correct option on the answer sheet provided.*
- *Each question carries one mark. There is no negative marking.*

1. What is cloud computing?

A) The use of remote servers on the internet to store, manage, and process data.
B) The use of local servers on the premises to store, manage, and process data.
C) The use of hybrid servers on the edge to store, manage, and process data.
D) The use of virtual servers on the network to store, manage, and process data.

2. What are the benefits of cloud computing?

A) Scalability, elasticity, cost-efficiency, reliability, security.
B) Scalability, rigidity, cost-efficiency, reliability, security.
C) Scalability, elasticity, cost-inefficiency, reliability, security.
D) Scalability, elasticity, cost-efficiency, unreliability, security.

3. What are the three service models of cloud computing?

A) Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS).
B) Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Storage as a Service (SaaS).
C) Infrastructure as a Service (IaaS), Processing as a Service (PaaS), Software as a Service (SaaS).
D) Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Security as a Service (SaaS).

4. What are the three deployment models of cloud computing?

A) Public cloud, private cloud, hybrid cloud.
B) Public cloud, private cloud, community cloud.
C) Public cloud, private cloud, hybrid cloud, community cloud.
D) Public cloud, private cloud, hybrid cloud, edge cloud.

5. What is AWS?

A) Amazon Web Services, a cloud computing platform that offers a variety of services and products.
B) Amazon Web Solutions, a cloud computing platform that offers a variety of solutions and products.
C) Amazon Web Services, a cloud computing platform that offers a variety of services and programs.
D) Amazon Web Solutions, a cloud computing platform that offers a variety of solutions and programs.

6. What are some of the popular services offered by AWS?

A) EC2, S3, Lambda, DynamoDB, RDS, SQS, SNS, Kinesis, EMR, Redshift.
B) EC2, S3, Lambda, DynamoDB, RDS, SQS, SNS, Kinesis, EMR, Spark.
C) EC2, S3, Lambda, DynamoDB, RDS, SQS, SNS, Kinesis, EMR, Hive.
D) EC2, S3, Lambda, DynamoDB, RDS, SQS, SNS, Kinesis, EMR, Hadoop.

7. What is Azure?

A) Microsoft Azure, a cloud computing platform that offers a variety of services and products.
B) Microsoft Azure, a cloud computing platform that offers a variety of solutions and products.
C) Microsoft Azure, a cloud computing platform that offers a variety of services and programs.
D) Microsoft Azure, a cloud computing platform that offers a variety of solutions and programs.

8. What are some of the popular services offered by Azure?

A) Virtual Machines, Blob Storage, Azure Functions, Cosmos DB, SQL Database, Service Bus, Event Hubs, HDInsight, Synapse Analytics.
B) Virtual Machines, Blob Storage, Azure Functions, Cosmos DB, SQL Database, Service Bus, Event Hubs, HDInsight, Spark.
C) Virtual Machines, Blob Storage, Azure Functions, Cosmos DB, SQL Database, Service Bus, Event Hubs, HDInsight, Hive.
D) Virtual Machines, Blob Storage, Azure Functions, Cosmos DB, SQL Database, Service Bus, Event Hubs, HDInsight, Hadoop.

9. What is GCP?

A) Google Cloud Platform, a cloud computing platform that offers a variety of services and products.
B) Google Cloud Platform, a cloud computing platform that offers a variety of solutions and products.
C) Google Cloud Platform, a cloud computing platform that offers a variety of services and programs.

D) Google Cloud Platform, a cloud computing platform that offers a variety of solutions and programs.

10. What are some of the popular services offered by GCP?
A) Compute Engine, Cloud Storage, Cloud Functions, Cloud Firestore, Cloud SQL, Pub/Sub, Dataflow, Dataproc, BigQuery.
B) Compute Engine, Cloud Storage, Cloud Functions, Cloud Firestore, Cloud SQL, Pub/Sub, Dataflow, Dataproc, Spark.
C) Compute Engine, Cloud Storage, Cloud Functions, Cloud Firestore, Cloud SQL, Pub/Sub, Dataflow, Dataproc, Hive.
D) Compute Engine, Cloud Storage, Cloud Functions, Cloud Firestore, Cloud SQL, Pub/Sub, Dataflow, Dataproc, Hadoop.

11. What is Spark?
A) A distributed computing framework that provides fast and general-purpose data processing.
B) A distributed computing framework that provides slow and specific-purpose data processing.
C) A distributed computing framework that provides fast and general-purpose data storage.
D) A distributed computing framework that provides slow and specific-purpose data storage.

12. What are the components of Spark?
A) Spark Core, Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX.
B) Spark Core, Spark SQL, Spark Streaming, Spark MLlib, Spark Graph.
C) Spark Core, Spark SQL, Spark Streaming, Spark ML, Spark GraphX.
D) Spark Core, Spark SQL, Spark Streaming, Spark ML, Spark Graph.

13. What is PySpark?
A) A Python API for Spark that allows users to write Spark applications using Python.
B) A Python API for Spark that allows users to write Spark applications using Java.
C) A Python API for Spark that allows users to write Spark applications using Scala.
D) A Python API for Spark that allows users to write Spark applications using R.

14. What are the benefits of PySpark?
A) It provides a simple and expressive syntax, a large number of libraries and tools, and interoperability with other Python frameworks.
B) It provides a complex and rigid syntax, a large number of libraries and tools, and interoperability with other Python frameworks.
C) It provides a simple and expressive syntax, a small number of libraries and tools, and interoperability with other Python frameworks.
D) It provides a simple and expressive syntax, a large number of libraries and tools, and isolation from other Python frameworks.

15. What are some of the common operations performed on tabular data using PySpark?

A) Reading and writing data from various sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, pivoting and unpivoting data.

B) Reading and writing data from various sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, pivoting and unpivoting data, visualizing data.

C) Reading and writing data from various sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, pivoting and unpivoting data, modeling data.

D) Reading and writing data from various sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, pivoting and unpivoting data, testing data.

16. What are some of the common machine learning tasks that can be performed using PySpark?

A) Classification, regression, clustering, recommendation, dimensionality reduction, feature engineering, model evaluation, model selection, model persistence.

B) Classification, regression, clustering, recommendation, dimensionality reduction, feature engineering, model evaluation, model selection, model deployment.

C) Classification, regression, clustering, recommendation, dimensionality reduction, feature engineering, model evaluation, model selection, model visualization.

D) Classification, regression, clustering, recommendation, dimensionality reduction, feature engineering, model evaluation, model selection, model optimization.

17. What are some of the common real-time analytics tasks that can be performed using PySpark?

A) Reading and writing data from various streaming sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, applying watermarking, outputting data to various sinks.

B) Reading and writing data from various streaming sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, applying watermarking, visualizing data.

C) Reading and writing data from various streaming sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and

merging data, applying user-defined functions, performing window functions, applying watermarking, modeling data.
D) Reading and writing data from various streaming sources, applying schema and data types, selecting and filtering columns and rows, grouping and aggregating data, joining and merging data, applying user-defined functions, performing window functions, applying watermarking, testing data.

18. What is containerization?
A) The process of packaging an application and its dependencies into a standardized unit that can run on any platform.
B) The process of packaging an application and its dependencies into a customized unit that can run on any platform.
C) The process of packaging an application and its dependencies into a standardized unit that can run on a specific platform.
D) The process of packaging an application and its dependencies into a customized unit that can run on a specific platform.

19. What are the benefits of containerization?
A) Portability, isolation, scalability, efficiency, security, compatibility.
B) Portability, isolation, scalability, efficiency, security, flexibility.
C) Portability, isolation, scalability, efficiency, security, simplicity.
D) Portability, isolation, scalability, efficiency, security, reliability.

20. What is Docker?
A) A software platform that enables users to build, run, and share applications using containers.
B) A software platform that enables users to build, run, and share applications using virtual machines.
C) A software platform that enables users to build, run, and share applications using functions.
D) A software platform that enables users to build, run, and share applications using services.

21. What is a Dockerfile?
A) A text file that contains the instructions to build a Docker image.
B) A text file that contains the instructions to run a Docker image.
C) A text file that contains the instructions to share a Docker image.
D) A text file that contains the instructions to delete a Docker image.

22. What is a Docker image?
A) A read-only template that contains the application and its dependencies.

B) A read-only template that contains the application and its configuration.
C) A read-write template that contains the application and its dependencies.
D) A read-write template that contains the application and its configuration.

23. What is a Docker container?
A) A running instance of a Docker image that can be isolated and managed.
B) A running instance of a Docker image that can be shared and distributed.
C) A stopped instance of a Docker image that can be isolated and managed.
D) A stopped instance of a Docker image that can be shared and distributed.

24. What is Docker Hub?
A) A cloud-based service that allows users to store and manage their Docker images.
B) A cloud-based service that allows users to build and run their Docker images.
C) A cloud-based service that allows users to compose and swarm their Docker images.
D) A cloud-based service that allows users to machine and engine their Docker images.

25. What is the main benefit of parallel computing for big data analytics?
a) It reduces the cost of data storage
b) It increases the speed of data processing
c) It improves the quality of data analysis
d) It enhances the security of data transmission

26. Which of the following is an example of real-time monitoring in big data analytics?
a) A dashboard that displays the current traffic conditions of a city
b) A report that summarizes the monthly sales performance of a company
c) A database that stores the historical records of a hospital
d) A spreadsheet that calculates the average temperature of a country

27. Which of the following is a key characteristic of cloud-based architectures for big data analytics?
a) Scalability
b) Portability
c) Reliability
d) All of the above

28. Which of the following cloud components is best suited for storing large volumes of unstructured data?
a) Cloud storage

b) Cloud database
c) Cloud compute
d) Cloud network

29. Which of the following cloud components is best suited for performing complex data analysis and machine learning tasks?
a) Cloud storage
b) Cloud database
c) Cloud compute
d) Cloud network

30. Which of the following is not a challenge of parallel computing?
A) Load balancing
B) Communication overhead
C) Data fragmentation
D) Data consistency

31. Which of the following is an example of data parallelism in Big Data Analytics?
A) Splitting a large file into smaller chunks and processing them simultaneously on different nodes
B) Assigning different tasks to different nodes based on their availability and capability
C) Distributing the execution of a complex algorithm across multiple nodes
D) Replicating the same data on multiple nodes for fault tolerance

32. Which of the following is an advantage of using MapReduce for parallel computing?
A) It abstracts the details of parallelization, distribution, and fault tolerance
B) It supports both structured and unstructured data sources
C) It scales well with the increase of data size and cluster size
D) All of the above

33. Which of the following statements is true about the CAP theorem in the context of distributed systems?
A) It states that it is impossible for a distributed system to simultaneously provide more than two of the following guarantees: Consistency, Availability, and Partition tolerance.
B) It states that it is possible for a distributed system to simultaneously provide all three of the following guarantees: Consistency, Availability, and Partition tolerance.

C) It states that it is impossible for a distributed system to simultaneously provide any of the following guarantees: Consistency, Availability, and Partition tolerance.
D) It states that it is possible for a distributed system to simultaneously provide any one of the following guarantees: Consistency, Availability, and Partition tolerance.

34. Which of the following is a characteristic of parallel computing?
A) It involves multiple processors working on different parts of a problem simultaneously.
B) It involves a single processor working on the same problem sequentially.
C) It involves multiple processors working on the same part of a problem sequentially.
D) It involves a single processor working on different parts of a problem simultaneously.

35. Which of the following is a benefit of distributed computing?
A) It reduces the communication overhead among processors.
B) It increases the reliability and availability of the system.
C) It decreases the scalability and performance of the system.
D) It increases the complexity and cost of the system.

36. Which of the following is a challenge of machine learning using parallel computing?
A) It requires a large amount of data to train the models.
B) It requires a high level of synchronization and coordination among processors.
C) It requires a low level of parallelism and concurrency in the algorithms.
D) It requires a homogeneous and centralized architecture.

37. Which of the following is an example of a cloud-based architecture for big data analytics?
A) MapReduce
B) Hadoop
C) Spark
D) All of the above

38. Which of the following is a common architecture pattern for building a scalable NLP solution?
A) Microservices
B) Lambda
C) Pipeline
D) Monolith

39. Which of the following is a common architecture pattern for building a scalable anomaly detection solution?

A) Stream
B) Batch
C) Hybrid
D) All of the above

40. Which of the following is a key component of a scalable NLP solution that enables distributed processing of large volumes of text data?

A) Data lake
B) Data warehouse
C) Data pipeline
D) Data ingestion

41. Which of the following is a key component of a scalable anomaly detection solution that enables real-time detection and alerting of anomalous events?

A) Data stream
B) Data sink
C) Data source
D) Data monitor

42. What is the name of the process that converts a deep learning model into an executable program that can run on a target device?

A) Model inference
B) Model training
C) Model deployment
D) Model optimization

43. What is the name of the framework that allows developers to create and deploy deep learning models using web technologies such as HTML, CSS, and JavaScript?

A) TensorFlow.js
B) PyTorch
C) Keras
D) Scikit-learn

44. What is the name of the technique that reduces the size and complexity of a deep learning model by removing redundant or less important parameters?

A) Pruning
B) Quantization
C) Compression
D) Regularization

45. What is the name of the platform that provides a cloud-based service for creating, managing, and deploying deep learning models?

A) Google Cloud AI Platform
B) Amazon SageMaker
C) Microsoft Azure Machine Learning
D) All of the above

Detail Questions

Instructions

- *Provide detailed and labelled diagrams where ever needed*
- *Provide pointwise reasoning and explanation/ justification for components selected.*
- *Each question is of 15 marks with no negative marking.*
- *All questions are mandatory to attempt.*
- *Total weightage is 45 marks.*

I. Explain the concept of MapReduce and how it enables parallel processing of large data sets. (15 marks)

II. Describe the steps involved in building a machine learning pipeline for big data using PySpark and create a complete deign flow for social media sentiment analysis use case. (15 marks)

III. Design a big data architecture for a real-time analytics application that monitors the traffic patterns and congestion levels in a city. Use appropriate tools and technologies and justify your choices. Any one cloud based components can be selected. (15 marks)