← Back

Final Assessment
Graded Assignment • 1h 40m

English

Due Jul 12, 11:59 PM WIB

Your grade: 85%

Your latest: **80%** • Your highest: **85%** • To pass you need at least 70%. We keep your highest score.

Next item →

1. You are a data engineer for an online retail company that has decided to introduce various discount schemes for its customers. Due to high demand, many customers are browsing through your website simultaneously. To manage the traffic, your team has decided to employ distributed computing.

   1 / 1 point

   Which of the following best explains the use of distributed computing in such a scenario?

   ● Distributed computing is a group of computers working together to share the same memory

   ○ Distributed computing is the same as parallel computing.

   ○ Distributed computing requires all participating computers and fails if any is disabled.

   ○ Distributed computing is unscalable with no modular growth.

   ✓ **Correct**
   Correct! Distributed computing is a group of computers working together with a shared memory.

2. As you initiate the journey to explore Apache Spark, you focus on understanding the specific path through which the data flows. What are the three major components of Spark architecture that frame that path?

   0 / 1 point

   ○ Data storage, cluster management framework, and APIs

   ● Hadoop, data storage, and executors

   ○ Cluster management framework, Spark Core, and task schedulers

   ○ Hadoop, APIs, and Spark Core

   ⊗ **Incorrect**
   Incorrect. Review the Scale out and Data Parallelism in Apache Spark video.

3. Which of the following are characteristics of datasets?

   0 / 1 point

   ○ Strongly typed; use APIs in Java, Scala, Python, and R; built on top of RDDs; are the latest data abstraction added to Spark

   ○ Strongly typed; use unified Java and Scala APIs; built on top of DataFrames; are the latest data abstraction added to Spark

   ○ Strongly typed; use APIs in Java, Scala, Python, and R; built on top of DataFrames; added in earlier Spark versions

   ● Strongly typed; use APIs in Java, Scala, Python, and R; built on top of DataFrames; are the latest data abstraction added to Spark

   ⊗ **Incorrect**
   Incorrect. Review the Datasets and DataFrames in Spark video.

4. Which of the following is one of the four phases of Catalyst query optimization?

   1 / 1 point

   ○ Logical planning

   ○ Code analysis

   ○ Physical optimization

   ● Analysis

   ✓ **Correct**
   Correct! The other three stages are logical optimization, physical planning, and code generation.

5. Which of the following options is used by the AIOps tools to facilitate IT operations?

   0 / 1 point

   ● Cluster managers

   ○ Big data

○ Machine learning

○ Apache Spark

⊗ **Incorrect**
Incorrect. Review the Using Apache Spark on IBM Cloud video.

6. Spark dependencies require driver and cluster executor processes to be able to access the application project. With which of the following options do Java and Scala applications provide this access?

1 / 1 point

○ Driver file

○ Spark bin directory

○ Dependency file

● uber-JAR

✓ **Correct**
Correct! This is a single JAR file containing all dependencies. Hence, the application is portable through the cluster.

7. As a data engineer, you need to run a command to specify the number of executor cores for a Spark standalone cluster for the application.Which of the following commands will help you?

1 / 1 point

● Use the command ‘--total-executor-cores’ followed by the number of cores.

○ Use the command ‘–app--total--cores’ followed by the number of cores.

○ Use the command ‘–app--total-executor-cores’ followed by the number of cores.

○ Use the command ‘--app--executor-cores’ followed by the number of cores.

✓ **Correct**
Correct! The command ‘--total-executor-cores’ followed by the number of cores specifies the number of executor cores for a Spark standalone cluster *for the application.*

8. You are a data engineer working for a startup. Your team has recently adapted the use of Spark. However, due to an error, the critical phase of the data processing gets stuck. Your team receives an error message for the same, but due to the new adaptation of Spark, your team needs clarification. Which of the following actions does Apache Spark perform if a task fails due to an error?

1 / 1 point

○ Attempts to locate a missing dependency

○ Terminates the application and reports an error to the driver

● Attempts to rerun the task for a set number of retries

○ Continues with related executor tasks

✓ **Correct**
Correct! The cause of an application failure can usually be found in the driver event log.

9. Which option will describe the relationship between big data and today’s personal assistants, including Google, Alexa, Siri, and others? Select all that apply.

1 / 1 point

☑ Personal assistants also rely on unstructured data sources, including personal data in the form of photos, videos, and text that people send to each other as the bulk of data collected by consumer goods companies.

✓ **Correct**
Correct! Personal assistants use unstructured data sources, including personal data in the form of photos, videos, and texts that people send each other as the bulk of data collected by consumer goods companies.

□ Assistants base their answers solely on structured data sources.

☑ Personal assistants use data sources, including location tracking and historical shopping data, to help provide predictive answers based on personal preferences.

✓ **Correct**
Correct! Assistants combine data from a multitude of sources and apply algorithms and AI to provide users with what the user will deem to be a correct answer.

☑ Assistants take questions and provide answers via some of the most advanced neural networks that exist.

> **Correct**
> Correct! Advanced neural networks process the user’s words and even voice tone when creating responses to questions and requests.

10. Which of the following best describes big data? **1 / 1 point**

● It is complex and requires specialized software to interpret and make it available for human interpretation.

○ It is only generated by certain specialized sensors and devices.

○ It can be stored on private servers.

○ It refers to just large volumes of data.

> **Correct**
> Correct! Big data arrives at a massive volume and with little or no structure.

11. Which of the following statements relates to parallel processing? **1 / 1 point**

○ It’s not particularly flexible.

● It’s the best technique for processing Big Data.

○ It can be inefficient and time-consuming.

○ It isn’t easy to scale.

> **Correct**
> Correct! Parallel processing works well for Big Data because of its speed and flexibility, among other reasons.

12. Which of the following options is associated with semi-structured data? **1 / 1 point**

● Includes some metadata that identifies certain characteristics

○ Includes sensor data from Internet of Things devices

○ Includes databases and spreadsheets

○ Includes prestructured data model

> **Correct**
> Correct! Semi-structured data combines unstructured and structured data.

13. As a data engineer, you know that Hadoop is a set of open-source programs and procedures and it can handle parallel jobs as well for efficient processing of the data. Which of the following best describes the use case for Hadoop? **1 / 1 point**

○ For processing many small files

○ For processing data with dependencies

● For processing enormous data sets

○ For processing transactions

> **Correct**
> Correct! Hadoop is a good solution for considerable data work.

14. Which of the following options explains the process of a driver program? **1 / 1 point**

○ Act in parallel to do work

○ Has similar processes as others in the application

● Create work and send it to the cluster

○ Run multiple threads

✓ **Correct**

Correct! The driver process can be run on a cluster node or another machine.

15. Which of the following is the correct precedence order for Spark property configuration?

0 / 1 point

● Spark-submit configuration, programmatically, spark-defaults.conf file

○ Spark-defaults.conf file, spark-submit configuration, programmatically

○ Programmatically, spark-defaults.conf file, spark-submit configuration

○ Programmatically, spark-submit configuration, spark-defaults.conf file

✗ **Incorrect**

Incorrect. Review the Setting Apache Spark Configuration video.

16. You are a data engineer whose work mainly revolves around big data, data management, etc. You require a platform that runs containerized applications on a cluster in a more resilient and flexible way. To fulfill this purpose, you opt for Kubernetes as an option. Among the following options, which statement is associated with one of the characteristics of Kubernetes?

1 / 1 point

○ It only runs in the cloud.

○ It cannot be run on a single machine.

○ It cannot be deployed automatically.

● It is portable.

✓ **Correct**

Correct! Kubernetes can be run in the cloud or on-premises.

17. Which of the following is true of open-source software?

1 / 1 point

○ It allows limited users to propose changes to the project

○ It can only be changed by a designated organization.

● It is free to use.

○ It is not efficient for large and complex projects.

✓ **Correct**

Correct! Open-source software is free, and the source code is open for review, to use, or re-use as needed in other projects.

18. Which of the following is a key advantage of MapReduce?

1 / 1 point

○ Focuses on the social media industry

○ Reduces the data footprint

● Allows a high level of parallel jobs across nodes

○ Runs independently from Hadoop

✓ **Correct**

Correct! This saves time and gives flexibility.

19. As a data engineer, you encourage using HIVE in your team because of the advantages it offers. Among the following, which statement can appropriately describe the difference between HIVE and a traditional RDBMS?

1 / 1 point

○ Hive is designed to read and write as many times as it needs, whereas RDBMS is based on the methodology of write once and read many.

● The maximum size Hive can handle is petabytes, whereas the maximum size that RDBMS can handle is terabytes.

○ Hive is suited for real-time data analysis, whereas RDBMS is for static data analysis.

○ Hive does not support partitioning, whereas RDBMS supports partitioning.

Correct
Correct! This makes it well suited to working with big data.

20. Which of the following is included in the Spark workflow?

1 / 1 point

○ Jobs completed in the cluster manager

● Jobs transferring results back to the driver or writing to disk

○ Jobs created by the SparkSQL in the executor.

○ Jobs held over as incomplete from a previous stage

Correct
Correct! Workflows can include completing jobs, transferring results back to the driver, or writing to disk.