Your grade: 90%

Your latest: **90%** • Your highest: **90%** • To pass you need at least 70%. We keep your highest score.

Next item

1. How does Apache Spark solve read/write problems encountered by other tools? **1 / 1 point**

- ○ By only using certain processors in the distributed group
- ◉ By keeping much of the required data in-memory
- ○ By leveraging redundancy
- ○ By using special proprietary APIs

Correct
Correct! Keeping data in-memory avoids disk I/O, which speeds up the process.

2. You are a newly recruited data engineer at your organization that uses Apache Spark for efficient data processing. Being curious, you start learning about the intriguing details of the data flow process. You learn that there are three Apache Spark components: data storage, compute interface, and cluster management framework. In which order does your organization's data flow through these components? **1 / 1 point**

- ○ Data flows from API into different nodes for parallel tasks and then into a Hadoop file system.
- ○ Data flows from the compute interface to various nodes for distributed tasks and then goes to the Hadoop file system.
- ◉ Data flows from the Hadoop file system into the compute interface and then into different nodes to perform distributed/parallel tasks.
- ○ Data flows from a Hadoop file system into different nodes for distributed tasks and then to the APIs.

Correct
Correct! The data from a Hadoop file system flows into the compute interface or API, which then flows into different nodes to perform distributed/parallel tasks.

3. Which of the following best describes datasets? **1 / 1 point**

- ○ Datasets act as a base for DataFrames.
- ◉ Datasets are strongly typed and provide compile-time type safety.
- ○ Datasets compute more slowly than RDDs.
- ○ Datasets are primarily used for real-time stream processing.

Correct
Correct! Compile-time type safety means Spark can detect syntax and semantic errors in production applications before deployment.

4. Which of the following best describes Tungsten? **1 / 1 point**

- ◉ Does not generate virtual function dispatches
- ○ Does not support on-demand JVM byte code generation
- ○ Does not enable computation of algorithms using STRIDE-based memory access
- ○ Relies on the JVM object model

Correct
Correct! This reduces multiple CPU calls.

5. How does IBM Spectrum Conductor help in avoiding downtime when running Spark? **0 / 1 point**

- ○ By sharing cluster resources

- By deploying multiple versions
- By dividing cluster resources dynamically
- By automating troubleshooting

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

6. Being a data engineer, you understand the importance of using Apache Spark to analyze big data. You know that Spark Shell is a command-line tool that makes it easy to run Spark codes and create and test Spark applications. Other than these advantages, how else does Spark Shell simplify working with data? 1 / 1 point

- By running in driver deploy mode
- By creating virtual environments so that applications can run separately
- By creating an uber-JAR
- By automatically initializing the SparkContent and SparkSession variables

Correct
Correct! This process enables you to start working with data immediately.

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--executor-cores' followed by the number of cores.
- Use the command '–app--total-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 currently facing an issue with the Spark application, which is disrupting the efficient processing of your organization's data. To debug the same, the first step will be to recognize the area of the issue. Which of the following options helps you to identify the common areas where Spark application issues can occur? 1 / 1 point

- User code, Configuration, Application Dependencies, Resource allocation, Network Communication
- User code, Configuration, Application Dependencies, Resource allocation, External logins
- User code, Configuration, Application Dependencies, Resource allocation, Network security measures
- User code, Configuration, Application Dependencies, Cloud provider choice

Correct
Correct! User code, Configuration, Application Dependencies, Resource allocation, and Network Communication are common areas where Spark application issues can happen.

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

- [ ] Assistants base their answers solely on structured data sources.
- [x] 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.

- [x] 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.

- [x] 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. Select the answer that identifies the main components that describe the dimensions of big data. **1 / 1 point**

- (•) Velocity, Volume, Variety, and Veracity
- ( ) Volume, Variety, Volatility, and Visibility
- ( ) Velocity, Volume, Visibility, and Volatility
- ( ) Velocity, Volume, Variety, and Validity

Correct
Correct! Four main components that describe the dimensions of Big Data are velocity, volume, variety, and veracity.

11. What is data scaling? **0 / 1 point**

- (•) Data scaling divides workloads to run in parallel.
- ( ) Data scaling is the process of transforming data values for end use.
- ( ) Data scaling is only applicable within cloud environments.
- ( ) Data scaling is a technique to manage, store, and process the overflow of data.

Incorrect
Incorrect. Review the Parallel Processing and Scalability video.

12. What has contributed significantly to the launch of the Big Data era? **1 / 1 point**

- ( ) The usage of social media
- ( ) The rise of streaming media
- ( ) The decrement in the production of data
- (•) The emergence of cloud computing

Correct
Correct! Cloud computing offers flexibility, cost savings, and high capacity, making big data workable.

13. You are a data analyst who provides data analytics solutions to clients. Your client needs a solution to process the customer data generated during the season of peak sales. They need an insightful solution that can help them manage terabytes of data. To address their concern, you introduce the concept of Hadoop. Which of the following best explains Hadoop? **1 / 1 point**

- ( ) A powerful database system.
- ( ) A proprietary data-processing platform.
- (•) A set of open-source programs and procedures that make up an ecosystem.
- ( ) A collection of common utilities and libraries.

Correct

Correct! Hadoop is a set of open-source programs and procedures that make up an ecosystem. It has many components that work together.

14. What happens when Spark performs a shuffle? 1 / 1 point

- ○ Divides jobs into tasks
- ○ Removes partitions
- ○ Increases cluster parallelism
- ◉ Redistributes datasets across the cluster

Correct
Correct! When Spark performs a shuffle, it redistributes the dataset across the cluster.

15. Which of the following is the correct precedence order for Spark property configuration? 1 / 1 point

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

Correct
Correct! This is the precedence order that Spark uses to apply configuration settings.

16. What could be the possible reasons to host Kubernetes on a local machine? 1 / 1 point

- ◉ As a development environment
- ○ As a limitation to the scope of information used
- ○ For low costs
- ○ For better security

Correct
Correct! Using Kubernetes locally can help you determine the best way to deploy it.

17. What is the biggest component of big data? 1 / 1 point

- ○ HDP
- ○ Apache Spark
- ○ Kubernetes
- ◉ Hadoop

Correct
Correct! Hadoop and its components, plus the tools that work with it, comprise the biggest part of Big Data software by far.

18. How does MapReduce keep track of its tasks? 1 / 1 point

- ◉ Using unique keys
- ○ Using variables
- ○ Using tags
- ○ Using cluster managers

Correct

Correct! MapReduce tracks its tasks using a unique key.

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 suited for real-time data analysis, whereas RDBMS is for static data analysis.
- (selected) The maximum size Hive can handle is petabytes, whereas the maximum size that RDBMS can handle is terabytes.
- 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.
- Hive does not support partitioning, whereas RDBMS supports partitioning.

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

20. You are a data engineer working for a growing startup. Your team requires an interface that gives you the running application information by showing jobs, stages, and tasks. 1 / 1 point

How does Spark Application UI benefit you for monitoring applications?

- Pinpoints and removes corrupted data.
- Limits usage if the system is overloaded.
- (selected) Quickly identifies failed jobs and locates the root cause of failure.
- Installs new applications when necessary.

Correct
Correct! This keeps nodes from staying out of use and increases efficiency.