

Your grade: **75%**

Your latest: **75%** • Your highest: **75%** • 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 unscalable with no modular growth.
- ☐ Distributed computing requires all participating computers and fails if any is disabled.
- ☐ Distributed computing is the same as parallel computing.

✓ 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

- ☐ Hadoop, data storage, and executors
- ☐ Data storage, cluster management framework, and APIs
- ☐ Hadoop, APIs, and Spark Core
- ☒ Cluster management framework, Spark Core, and task schedulers

✗ Incorrect

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

3. Which of the following best describes datasets?

1 / 1 point

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

✓ Correct

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

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

1 / 1 point

- ☐ Logical planning
- ☒ Analysis
- ☐ Physical optimization
- ☐ Code analysis

✓ Correct

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

5. You are a data analyst working with a fast-growing startup. Your organization keeps adapting new technology to keep pace with clients' requirements. You are currently evaluating the tools that can create production-ready environments for AI and machine learning. Which of the following will assist you to create production-ready environments for AI and machine learning?

1 / 1 point

- ☐ Apache Hadoop Cluster

- ☒ IBM Watson
- ☐ Spark
- ☐ IBM Analytics Engine



Correct! IBM Watson provides services, support, and holistic workflows.

6. What is the name of the Spark unified interface?

1 / 1 point

- ☒ spark-submit
- ☐ spark-default
- ☐ YARN
- ☐ SUI



Correct! The spark-submit script is found in the bin/' directory.

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 '--app--executor-cores' followed by the number of cores.
- ☐ Use the command '--app--total--cores' followed by the number of cores.
- ☒ Use the command '--total-executor-cores' followed by the number of cores.
- ☐ Use the command '--app--total-executor-cores' followed by the number of cores.



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



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

9. Which of the following is a major component of the Internet of Things (IoT) ecosystem?

1 / 1 point

- ☐ End-users and device owners
- ☐ Solar power generation
- ☐ Social media platforms
- ☒ Smart devices and sensors



Correct! Smart devices and sensors play a crucial role in the IoT ecosystem by collecting data and enabling connectivity.

10. Which of the following best describes big data?

1 / 1 point

- ☐ It can be stored on private servers.
- ☒ 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 refers to just large volumes of data.

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

11. What does “scaling out” mean?

1 / 1 point

- ☒ Adding nodes to increase capacity
- ☐ Changing the software that runs the nodes to increase efficiency
- ☐ Distributing work among the nodes differently to balance the load
- ☐ Adding larger single nodes to increase capacity

✓ **Correct**
Correct! This is a sustainable solution to growing infrastructure needs.

12. What is the current projected yearly growth rate for data?

0 / 1 point

- ☐ 75 percent
- ☐ 40 percent
- ☐ 90 percent
- ☒ 25 percent

✗ **Incorrect**
Incorrect. Review the Beyond the Hype video.

13. Which of the following Hadoop core components prepares the RAM and CPU for Hadoop to run data in batch, stream, interactive, and graph processing?

1 / 1 point

- ☐ MapReduce
- ☒ YARN
- ☐ Hadoop Common
- ☐ HDFS

✓ **Correct**
Correct! YARN is short for "yet another resource negotiator" and is one of the most important components because it prepares the RAM and CPU for Hadoop to run data in such kinds of processing.

14. Which of the following options explains the process of a driver program?

1 / 1 point

- ☐ Has similar processes as others in the application
- ☒ Create work and send it to the cluster
- ☐ Act in parallel to do work
- ☐ 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-defaults.conf file, spark-submit configuration, programmatically

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

✗ **Incorrect**

Incorrect. Review the Setting Apache Spark Configuration video.

16. What are the required additional considerations when deploying Spark applications on top Kubernetes using client mode?

0 / 1 point

- ☐ Driver's pod name must be set to "spark.name."
- ☐ Executors must be able to communicate and connect with the driver program.
- ☐ Drivers and executors are not required to connect and communicate with each other.
- ☒ Executor's name must be set to "spark.kubernetes.executor.pod.name."

✗ **Incorrect**

Incorrect. Review the Running Spark on Kubernetes video.

17. What is the biggest component of big data?

1 / 1 point

- ☐ Kubernetes
- ☒ Hadoop
- ☐ Apache Spark
- ☐ HDP

✓ **Correct**

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

18. You are a data analyst in a tech startup. You use MapReduce as a programming model in Hadoop to process big data. In the process, the first stage is when you input a file. The next stage is the "split," then "map" task, then shuffling, followed by the "reduce" task and output. Which of the following happens in the "map" task of MapReduce?

1 / 1 point

- ☐ Aggregate a set of results
- ☐ Produce a final report
- ☐ Give consistent names to pieces of data
- ☒ Process data into key value pairs

✓ **Correct**

Correct! The reduce task can then shuffle and collate this information into a report.

19. Your team is switching to using Hive from traditional relational databases to effectively manage and optimize the data infrastructure. Which of the following characteristics is a part of Hive rather than a traditional RDBMS that makes it a more efficient software?

1 / 1 point

- ☐ Can handle terabytes of data
- ☐ Suited for real-time data analysis
- ☐ Maintains a Database and uses SQL
- ☒ Designed on the methodology of write once, read many

✓ **Correct**


Correct! Hive is designed on the methodology of write once, read many.

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

0 / 1 point

- ☒ Jobs created by the SparkSQL in the executor.
- ☐ Jobs transferring results back to the driver or writing to disk

- ☐ Jobs held over as incomplete from a previous stage
- ☐ Jobs completed in the cluster manager

 **Incorrect**

Incorrect. Review the Monitoring Application Progress video.