← Back **Final Exam**
Graded Quiz • 1h 40m

Due Oct 22, 11:59 PM +07

Try again once you are ready

Grade received 75%
Latest Submission Grade 65%
To pass 80% or higher

Retake the assignment in **23h 2m**

1. Which of the following Apache Spark benefits helps manage big data processing? **1 / 1 point**

○ In-memory processing
● Unified framework
○ Python
○ Scala

Expand

Correct
Correct. Apache Spark creates a comprehensive, unified framework to manage big data processing.

2. The three Apache Spark components are data storage, compute interface, and cluster management framework. Which order does the data flow through these components? **0 / 1 point**

○ Data flows from Hadoop file system, into compute interface and then into different nodes for distributed.
● Data flows from API into different nodes for parallel tasks, and then into a Hadoop file system.
○ Data flows from a Hadoop file system into different nodes for distributed task, but then flows to the APIs.
○ Data flows from compute interface to various nodes for distributed tasks and then goes to the Hadoop file system.

Expand

Incorrect
Incorrect. Please review the Scale out Data Parallelism in Apache Spark video.

3. Which of the following is NOT a way to create a dataset? **0 / 1 point**

○ From a text file using an explicit schema declaration and the “String” data type
○ Using the toDS function in Scala
○ Combine DataFrames within a dataset.
● From a JSON file and custom classes

Expand

Incorrect
Incorrect. Please refer to the Data-Frames and Datasets video.

4. Which of the following features belong to Tungsten? **1 / 1 point**

☐ Prohibits Loop unrolling.
☑ Manages memory explicitly and does not rely on the JVM object model or garbage collection.

✓ **Correct**
This answer is correct. Tungsten manages memory explicitly and does not rely on the JVM object model or garbage collection.

☐ Generates virtual function dispatches
☑ Places intermediate data in CPU registers

✓ **Correct**
Yes! Tungsten places intermediate data in CPU registers.

Expand

Correct
Great, you got all the right answers.

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

○ Deploy multiple versions
○ Automatic troubleshooting
● Cluster resources divided dynamically
○ Shares cluster resources

Expand

Correct
Correct. this avoids downtime.

6. What is the name of the Spark unified interface? **1 / 1 point**

○ spark-default
○ SUI
● spark-submit
○ YARN

Expand

Correct
Correct. The spark-submit script is found in the ‘bin/’ directory.

7. Which command specifies the number of executor cores for a Spark standalone cluster for the application? **1 / 1 point**

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

Expand

Correct
This answer is correct!

8. If a task fails due to an error, Spark ______________. **0 / 1 point**

○ continues with related executor tasks.
○ can attempt to rerun the task for a set number of retries.
○ attempts to locate a missing dependency.
● terminates the application and reports an error to the driver.

Expand

Incorrect
Incorrect. Please refer to the Debugging Apache Spark Application Issues video.

9. Select the answers that describe the relationship between Big Data and today’s personal assistants including Google, Alexa Siri and others. **1 / 1 point**

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

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

☐ Assistants base their answers solely on structured data sources.
☑ 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**
Yes! Personal assistants use unstructured data sources including personal data in the form of photos, videos, and texts that people send to each other as the bulk of data collected by consumer goods companies.

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

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

Expand

Correct
Great, you got all the right answers.

10. Select the answer that identifies the main components that describe the dimensions of Big Data. **1 / 1 point**

○ Volume, Variety, Volatility, and Visibility
○ Velocity, Volume, Visibility, and Volatility
○ Velocity, Volume, Variety, and Validity
● Velocity, Volume, Variety, and Veracity

Expand

Correct
Yes! Four main components describe the dimensions of Big Data.

11. What is Data Scaling? **1 / 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.

Expand

Correct
Yes! This answer is correct!

12. Semi-structured data __________. **0 / 1 point**

○ Has a pre-structured data model.
● Includes sensor data from Internet of Things devices.
○ Includes databases and spreadsheets.
○ Includes some metadata that identifies certain characteristics.

Expand

Incorrect
Incorrect. Please refer to 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**

○ Hadoop Common
○ HDFS
● YARN
○ MapReduce

Expand

Correct
Correct. Yarn is short for "yet another resource negotiator" and it's one of the most important components because it prepares the RAM and CPU for Hadoop to run data in these types of processing.

14. Increasing Executors and cores __________. **1 / 1 point**

○ requires a shuffle
● increases cluster parallelism
○ transforms data partitions
○ necessitates dividing jobs into tasks

Expand

Correct
Correct. Tasks run in separate threads until all cores are used.

15. What is the Spark property configuration that follows a precedence order, with the highest being configuration set programmatically, then spark-submit configuration and lastly configuration set in the spark-defaults.conf file?

0 / 1 point

○ Setting how many cores are used
○ Application name
● Properties related to the application
○ Application version

Expand

Incorrect
Incorrect. Please review the Setting Apache Spark Configuration video.

16. Why might you want to host Kubernetes on a local machine?

1 / 1 point

○ To keep costs low
○ For better security
○ To limit the scope of information used
● As a test development environment

Expand

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

17. Which of the following is NOT true of open-source software?

0 / 1 point

○ It works very well for large, complex projects.
● It's the industry standard for servers worldwide.
○ It can only be changed by a designated organization.
○ It is profitable to use.

Expand

Incorrect
Incorrect. Please refer to the Open Source and Big Data video.

18. What is a key advantage of MapReduce that was mentioned in the course?

1 / 1 point

● It allows a high level of parallel jobs across nodes.
○ It reduces the data footprint.
○ It's specialized for the social media industry.
○ It can run independently from Hadoop.

Expand

Correct
Correct. This saves time and gives flexibility.

19. What is one way that Hive is different from traditional RDBMS?

0 / 1 point

○ Designed to read and write as many times as it needs.
● Suited for real-time data analysis.
○ Does not support partitioning.
○ Maximum size it can handle is petabytes.

Expand

Incorrect
Incorrect. Please refer to the Hive video.

20. Select the option that most closely matches the steps associated with the Spark Application Workflow.

1 / 1 point

● The application creates a job. Spark divides the job into one or more stages. The first Stage starts tasks. The tasks run and as one stage completes, the next stage starts. When tasks and stages complete the next job can begin.
○ The application creates a job. As one job completes another job begins. As the jobs complete, the application is restarted.
○ The application creates a task. Spark divides the task into one or more jobs. The first job starts a stage. The stages run and as one task completes, the next job starts. When tasks and stages complete the next job can begin.
○ The job opens the application and tasks run that are divided in stages. As a stage completes, the next task runs.

Expand

Correct
Correct! You selected the answer that matches the steps associated with the Spark Application Workflow.