← **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** 75%

**To pass** 80% or higher

**Try again**

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

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

**Expand**

**Correct**
Correct. Keeping data in-memory avoids disk I/O, which speeds up processes.

**2.** One component of Spark architecture is the executor. Which of the following is true? **1 / 1 point**

- ○ Executors control the function of drivers.
- ○ Only one executor can function at a time.
- ○ Executors work on only one worker node.
- ● Executors complete single tasks.

**Expand**

**Correct**
Correct. Executors complete tasks on worker nodes and pass results back to the driver.

**3.** Which of the following is true of datasets? **1 / 1 point**

- ● Datasets are strongly typed and therefore provide compile-time type safety.
- ○ Datasets compute more slowly than RDDs.
- ○ APIs are available in Scala and Java as well as other languages.
- ○ DataFrames are built on top of datasets.

**Expand**

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

**4.** Which of these is one of the four phases of Catalyst query optimization? **0 / 1 point**

- ○ Physical optimization
- ○ Logical planning
- ○ Analysis
- ● Code analysis

**Expand**

**Incorrect**
Incorrect. Please refer to the Catalyst and Tungsten video.

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

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

**Expand**

**Incorrect**
Incorrect. Please review the Using Apache Spark on IBM Cloud video.

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

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

**Expand**

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

**7.** Why does Spark queue tasks and wait for available cores? **1 / 1 point**

- ○ To keep the number of cores in the pool low
- ○ To start as many tasks as possible
- ● To maximize parallel processing
- ○ To use more cores

**Expand**

**Correct**
Correct. Once a task is in queue, Spark assigns it to any available executor.

**8.** If a task fails due to a dependency problem, what is the best way to identify the issue? **1 / 1 point**

- ○ Cataloging the libraries on the system
- ○ Checking required data files for corruption
- ○ Checking APIs
- ● Examining the event log for stack trace errors

**Expand**

**Correct**
Correct. These identify which libraries the application loaded.

**9.** Select the answers that describe the relationship between Big Data and today’s personal assistants including Google, Alexa Siri and others. **0 / 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.
- ☑ 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 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.
- ☐ Assistants base their answers solely on structured data sources.

**Expand**

**Incorrect**
You didn't select all the correct answers

**10.** What is “scaling out”? **1 / 1 point**

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

**Expand**

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

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

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

**Expand**

**Correct**
Yes! This answer is correct!

**12.** What is the current projected yearly growth rate for data? **0 / 1 point**

- ● 90 percent
- ○ 40 percent
- ○ 25 percent
- ○ 75 percent

**Expand**

**Incorrect**
This answer is 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**

- ○ HDFS
- ○ MapReduce
- ○ Hadoop Common
- ● YARN

**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.** What happens when Spark performs a shuffle? Select all that apply. **1 / 1 point**

- ☐ Removes partitions
- ☐ Increases cluster parallelism
- ☑ Boundaries between stages are marked

  ✓ **Correct**
  Correct. a shuffle marks the boundary between stages.
- ☑ Datasets redistributed across cluster

  ✓ **Correct**
  Correct. when Spark performs a shuffle, it redistributes the dataset across the cluster.

**Expand**

**Correct**
Great, you got all the right answers.

**15.** Which configuration method enables you to adjust settings on a per-machine basis? **1 / 1 point**

- ○ Properties
- ○ Manual

● Environment variables

○ Logging

Expand

✓ **Correct**

Correct. Feedback about why this answer is correct

**16.** What are the required additional considerations when deploying Spark applications on top Kubernetes using client mode? Select all that apply.

**0 / 1 point**

☑ The executors must be able to communicate and connect with the driver program.

✓ **Correct**

Correct. the 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.

☑ Use the driver's pod name to set 'spark.kubernetes.driver.pod.name'.

✓ **Correct**

Correct. Use the driver's pod name to set 'spark.kubernetes.driver.pod.name'. This will ensure the executor pods are cleaned up along with the driver pod when the application finishes.

☑ Use the executor name to set 'spark.kubernetes.excutor.pod.name'. This will ensure the executor pods are cleaned up along with the driver pod when the application finishes.

! **This should not be selected**

Incorrect. Please review the Running Spark on Kubernetes video.

Expand

⊗ **Incorrect**

You chose the extra incorrect answers.

**17.** The biggest component of Big Data is _______.

**1 / 1 point**

● Hadoop

○ Kubernetes

○ HDP

○ Apache Spark

Expand

✓ **Correct**

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

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

**1 / 1 point**

○ It can run independently from Hadoop.

● It allows a high level of parallel jobs across nodes.

○ It's specialized for the social media industry.

○ It reduces the data footprint.

Expand

✓ **Correct**

Correct. This saves time and gives flexibility.

**19.** Which of the following characteristics are part of Hive rather than a traditional relational database?

**1 / 1 point**

☑ Designed on the methodology of write once, read many

✓ **Correct**

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

☑ Can handle petabytes of data

✓ **Correct**

Correct! Hive can handle petabytes of data.

☐ Can handle terabytes of data

☐ Used to maintain a Database and uses SQL

Expand

✓ **Correct**

Great, you got all the right answers.

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

**1 / 1 point**

○ 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.

● 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.

Expand

✓ **Correct**

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