← Back **Final Assessment**
Graded Assignment • 1h 40m

English ⌵ **Due** Jul 12, 11:59 PM WIB

Your grade: 85%

Your latest: **85%** • 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 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.

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

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

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 a Hadoop file system into different nodes for distributed tasks and then to the APIs.

● Data flows from the Hadoop file system into the compute interface and then into different nodes to perform distributed/parallel tasks.

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. Your team is responsible for analyzing the customers' behavior and preferences. To do this task, you guide your team to create datasets for performing complex data transformations. There are three ways to create datasets. Which of the following answers help to create datasets? Select all that apply.

0 / 1 point

☑ A JSON file and custom classes

Correct
Correct! Datasets can be created using a JSON file and custom classes.

□ A text file by using an explicit schema declaration and the "String" data type

☑ DataFrames combined within a dataset

This should not be selected
Incorrect. Please refer to the Data-Frames and Datasets video.

☑ The toDS function in Scala

Correct
Correct! The toDS function in Scala can help create datasets.

4. Which of the following features belongs to Tungsten?

1 / 1 point

○ Prohibits Loop unrolling

● Places intermediate data in CPU registers

○ Relies on the JVM object model

○ Generates virtual function dispatches

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

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. 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**

● uber-JAR

○ Dependency file

○ Spark bin directory

○ Driver file

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

7. By default, how much memory does Spark use? **1 / 1 point**

● All available memory minus 1 GB and all available cores

○ All available memory minus 1 GB

○ All available memory and all available cores

○ All available memory minus five available cores

✓ **Correct**
Correct! Spark uses all available memory minus 1 GB and all available cores.

8. Consider that you are a senior data engineer at your organization, and your team currently faces an application dependency issue. Due to the recent adaptation of Apache Spark, you and your team evaluate the best possible way to identify the issue. Which of the following options will help you? **1 / 1 point**

● Examine the event log for stack trace errors

○ Catalogue the libraries on the system

○ Check APIs

○ Check the required data files for corruption

✓ **Correct**
Correct! This identifies which libraries the application loaded.

9. Which of the following is a common application of big data? **1 / 1 point**

○ Write new video games

○ Run automotive assembly lines

● Optimize recommendation engines on websites like Amazon and Google

○ Optimize streaming video services

✓ **Correct**
Correct! Optimizing recommendation engines on websites like Amazon and Google is a common application of big data.

Correct! Optimizing recommendation engines on websites like Amazon and Google is a common application of big data.

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 can be stored on private servers.

○ 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

○ Distributing work among the nodes differently to balance the load

○ Adding larger single nodes to increase capacity

○ Changing the software that runs the nodes to increase efficiency

● Adding nodes to increase capacity

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

12. Which of the following options is associated with semi-structured data?

1 / 1 point

● Includes some metadata that identifies certain characteristics

○ Includes databases and spreadsheets

○ Includes sensor data from Internet of Things devices

○ Includes prestructured data model

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

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 set of open-source programs and procedures that make up an ecosystem.

○ A collection of common utilities and libraries.

○ A proprietary data-processing platform.

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 executors and cores increase?

1 / 1 point

○ Jobs divide into tasks

○ Shuffle requirement arises

○ Data partition transforms

● Cluster parallelism increases

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

15. Which configuration method enables the adjustment of settings on a per-machine basis?

1 / 1 point

● Environment variables

○ Properties

○ Manual

○ Logging

✓ **Correct**
Correct! Environment variables enable the adjustment of settings on a per-machine basis.

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

1 / 1 point

○ For better security

○ For low costs

○ As a limitation to the scope of information used

● As a development environment

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

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

1 / 1 point

○ It can only be changed by a designated organization.

● It is free to use.

○ It is not efficient for large and complex projects.

○ It allows limited users to propose changes to the project

✓ **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

○ Runs independently from Hadoop

● Allows a high level of parallel jobs across nodes

○ Reduces the data footprint

✓ **Correct**
Correct! This saves time and gives flexibility.

19. You are a data engineer in a tech startup. Your team uses HIVE as a data warehouse software within Hadoop as it can read, write, and manage tabular-type datasets and even perform data analysis. What are the three components of Hive architecture that help in achieving effective data analysis?

1 / 1 point

○ Services, Metastore, Database

○ Storage, Computing, Command Line Interface

● Clients, Services, Storage and Computing

○ Clients, Services, Execution

✓ **Correct**
Correct! These three components each have multiple parts as well.

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

0 / 1 point

● Jobs created by the SparkSQL in the executor.

○ Jobs held over as incomplete from a previous stage

○ Jobs completed in the cluster manager

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

⊗ **Incorrect**

Incorrect. Review the Monitoring Application Progress video.