Congratulations! You passed!

Grade received 100%

To pass 80% or higher

Go to next item

Module 2 Quiz

Latest Submission Grade 100%

1. What are the different units of parallelism? (Select all that apply.) **1 / 1 point**

- [x] Core

  Correct
  A processor has many cores.

- [x] Task

  Correct
  A job can be divided into many tasks.

- [x] Executor

  Correct
  An executor is one worker node in a cluster.

- [x] Partition

  Correct
  A partition is a subset of data.

2. What is a partition? **1 / 1 point**

- ○ A synonym with "task"
- ○ The result of data filtered by a WHERE clause
- ○ A division of computation that executes a query
- ◉ A portion of a large distributed set of data

Correct
Data distributed across the cluster is divided into different partitions.

3. What is the difference between in-memory computing and other technologies? (Select all that apply.) **1 / 1 point**

- ☑ In-memory operates from RAM while other technologies operate from disk

Correct
In-memory operation works using RAM.

- ☑ In-memory operations were not realistic in older technologies when memory was more expensive

Correct
The price of memory has come down drastically enabling Spark to rely on in-memory calculations.

- ☐ In-memory computing is slower than other types of computing
- ☑ Computation not done in-memory (such as Hadoop) reads and writes from disk in between each step

Correct
Hadoop (the precursor to Spark) was much slower because it had to read from and write to disk between every step.

4. Why is caching important? **1 / 1 point**

- ○ It always stores data in-memory to improve performance
- ◉ It stores data on the cluster to improve query performance
- ○ It improves queries against data read one or more times
- ○ It reformats data already stored in RAM for faster access

Correct
By storing data we know we'll see again, caching improves query performance.

5. Which of the following is a wide transformation? (Select all that apply.) **1 / 1 point**

- ☑ ORDER BY

Correct
An ORDER BY transfers data across the network and is therefore a wide transformation.

- ☐ WHERE
- ☑ GROUP BY

Correct
A GROUP BY transfers data across the network and is therefore a wide transformation.

- ☐ SELECT

6. Broadcast joins... **1 / 1 point**

- ○ Transfer the smaller of two tables to the larger, increasing data transfer requirements
- ○ Shuffle both of the tables, minimizing data transfer by transferring data in parallel
- ◉ Transfer the smaller of two tables to the larger, minimizing data transfer
- ○ Shuffle both of the tables, minimizing computational resources

✓ **Correct**

7. Adaptive Query Execution uses runtime statistics to: **1 / 1 point**

- ☑ Dynamically coalesce shuffle partitions

✓ **Correct**
Please revisit the lesson: **Adaptive Query Execution**.

- ☑ Dynamically switch join strategies

✓ **Correct**
Please revisit the lesson: **Adaptive Query Execution**.

- ☑ Dynamically optimize skew joins

✓ **Correct**
Please revisit the lesson: **Adaptive Query Execution**.

- ☐ Dynamically cache data

8. Which of the following are bottlenecks you can detect with the Spark UI? (Select all that apply.) **1 / 1 point**

- [ ] Incompatible data formats
- [x] Shuffle writes

Correct
The Spark UI can show shuffles triggered by Spark actions.

- [x] Shuffle reads

Correct
The Spark UI can show shuffles triggered by Spark actions.

- [x] Data Skew

Correct
Data skew is when partitions are not of similar sizes and can be detected by the Spark UI.

9. What is a stage boundary? **1 / 1 point**

- (•) When all of the slots or available units of processing have to sync with one another
- ( ) Any transition between Spark tasks
- ( ) An action caused by a SQL query is predicate
- ( ) A narrow transformation

Correct
A stage boundary is when all Spark tasks must come together to exchange a result.

10. What happens when Spark code is executed in local mode? **1 / 1 point**

- ○ A cluster of virtual machines is used rather than physical machines
- ○ The code is executed against a local cluster
- ○ The code is executed in the cloud
- ◉ The executor and driver are on the same machine

Correct
Local mode refers to when the executor and driver are the same machine, such as when prototyping Spark code on your laptop.