Congratulations! You passed!
TO PASS 80% or higher

Keep Learning

GRADE
100%

Module 3 Quiz

LATEST SUBMISSION GRADE
100%

1. Decoupling storage and compute means storing data in one location and processing it using a separate resource. What are the benefits of this design principle? (Select all that apply.) — 1 / 1 point

- [x] It allows for elastic resources so larger storage or compute resources are used only when needed

Correct
Decoupled resources that aren't utilized can easily be shut down.

- [x] It makes updates to new software versions easier

Correct
New database and computation versions can be installed on new hardware due to the ephemeral nature of the underlying data.

- [ ] It results in copies of the data in case of a data center outage
- [x] Resources are isolated and therefore more manageable and debuggable

Correct
With each component of the architecture responsible for specific tasks, debugging is significantly easier.

2. You want to run a report entailing summary statistics on a large dataset sitting in a database. What is the main resource limitation of this task? — 1 / 1 point

- ( ) CPU: the transfer of data is more demanding than the computation
- (•) IO: the transfer of data is more demanding than the computation
- ( ) IO: computation is more demanding that the data transfer
- ( ) CPU: computation is more demanding than the data transfer

Correct
The main bottleneck here is the transfer of data across the network.

3. Processing virtual shopping cart orders in real time is an example of... — 1 / 1 point

- (•) Online Transaction Processing (OLTP)
- ( ) Online Analytical Processing (OLAP)

Correct
Processing real time information involves transactional processing.

4. When are BLOB stores an appropriate place to store data? (Select all that apply.) — 1 / 1 point

- [x] For storing large files

Correct
BLOB stores scale effectively infinitely.

- [x] For a "data lake" of largely unstructured data

Correct
BLOB stores are the backbone of most data lakes.

- [x] For cheap storage

Correct
BLOB stores are significantly cheaper than databases.

- [ ] For online transaction processing on a website

5. JDBC is the standard protocol for interacting with databases in the Java environment. How do parallel connections work between Spark and a database using JDBC? **1 / 1 point**

- (•) Specify a column, number of partitions, and the column's minimum and maximum values. Spark then divides that range of values between parallel connections.
- ( ) Specify the numPartitions configuration setting. Spark then creates one parallel connection for each partition.
- ( ) Specify the number of partitions using REPARTITION. Spark then creates one parallel connection for each partition.
- ( ) Specify the number of partitions using COALESCE. Spark then creates one parallel connection for each partition.

✓ **Correct**
Spark uses the max and min of a range of values to know which connection should receive which data.

6. What are some of the advantages of the file format Parquet over CSV? (Select all that apply.) **1 / 1 point**

- [x] Compression

✓ **Correct**
Parquet is compressed by default and has many additional compression options.

- [x] Parallelism

✓ **Correct**
Parquet easily parallelized so one file is written per Spark connection.

- [x] Columnar

✓ **Correct**
Parquet is a column-based rather than a row-based format.

- [ ] Corruptible

7. SQL is normally used to query tabular (or "structured") data. Semi-structured data like JSON is common in big data environments. Why? (Select all that apply.) **1 / 1 point**

- [x] It allows for complex data types

✓ **Correct**
Complex types like arrays are allowed in JSON.

- [x] It allows for data change over time

✓ **Correct**
JSON allows for schema evolution over time.

- [ ] It allows for easy joins between relational JSON tables
- [x] It allows for missing data

✓ **Correct**
JSON does not require all keys to appear in a dataset.

- [x] It does not need a formal structure

✓ **Correct**
No formal structure is needed to be declared in advance like with relational tables.

8. Data writes in Spark can happen in serial or in parallel. What controls this parallelism? **1 / 1 point**

- (•) The number of data partitions in a DataFrame
- ( ) The number of jobs in a write operation
- ( ) The number of stages in a write operation
- ( ) The numPartitions setting in the Spark configuration

✓ **Correct**

Controlling the data partitions controls the parallelism of data writes.

9. Fill in the blanks with the appropriate response below: 1 / 1 point

A _____ table manages _____and a DROP TABLE command will result in data loss.

- (●) Managed, both the data and metadata such as the schema and data location
- ( ) Managed, only the metadata such as the schema and data location
- ( ) Unmanaged, only the metadata such as the schema and data location
- ( ) Unmanaged, both the data and metadata such as the schema and data location

✓ **Correct**
When dropping a managed table, the underlying data will be deleted too.