Expert Session Video Questions

Started: Nov 19 at 10:06pm

Quiz Instructions

Welcome to the Apache Spark Quiz!

To begin, watch the embedded **expert session video with Andre Zanella** thoroughly.

The quiz questions are based directly on the concepts, examples, and insights shared during the session, so pay close attention to the details discussed.

Once you feel confident, proceed to answer the multiple-choice questions.

Each question has one correct answer, so carefully select the most accurate option based on the video content. You are allowed two attempt(s), and the quiz has a time limit of **20 minutes**.

Ensure you complete and submit your responses before the time expires.

Your final score will reflect your understanding of the material.

Question 1 2 pts

What is the major challenge of working with datasets in telecommunications research, as described by Andre Zanella?

○ Insufficient data for analysis

● Complexity and large size of datasets

○ Lack of efficient tools

○ Privacy issues preventing any data collection

Question 2 2 pts

What was the size of the dataset Andre Zanella described that involved mobile traffic in France for four months?

○ 12 GB

○ 1.7 TB

● 100 MB

○ 5.5 TB

Question 3 2 pts

Why does Spark handle large datasets more efficiently than Pandas, according to the session?

○ Spark processes data in memory without any partitions.

● Spark uses parallelism and fault tolerance while handling chunks of data.

○ Spark avoids using RAM altogether.

○ Spark relies solely on cloud storage.

Question 4 2 pts

What file system does Andre recommend pairing with Spark for better efficiency in cluster environments?

○ NTFS

● HDFS (Hadoop Distributed File System)

○ Ext4

○ FAT32

Question 5 2 pts

What is a key feature of the Parquet file format highlighted in the session?

○ It is editable in notepad.

● It provides better data compression and encoding.

○ It stores data as plain text for simplicity.

○ It cannot be used with Spark.

Question 6 2 pts

According to Andre, what step is crucial when creating a Spark DataFrame?

● Declaring the schema of the DataFrame

○ Saving the DataFrame as a CSV

○ Creating user-defined functions (UDFs)

○ Using machine learning libraries immediately

Question 7 2 pts

What is an example of a task that Spark can handle efficiently using a user-defined function (UDF)?

○ Applying a filter on small datasets

● Aggregating sessions that span multiple time intervals

○ Simple mathematical operations on columns

○ Saving data to Excel files

Question 8 2 pts

What is the significance of lazy evaluation in Spark, as mentioned in the session?

● It avoids immediate execution of operations until explicitly triggered.

○ It prevents memory errors by skipping invalid data.

○ It allows Spark to predict results before executing tasks.

○ It ensures datasets are immediately processed in full.

Question 9 2 pts

Which module in Spark is specifically designed for processing real-time data?

○ Spark SQL

○ MLlib

● Spark Streaming

○ Spark Core

Question 10 2 pts

Why does Andre recommend against starting with RDDs when learning Spark?

○ They are outdated and deprecated.

○ They require extensive setup and cluster management.

● They are too abstract and complex for beginners.

○ They lack documentation and support.

No new data to save. Last checked at 10:20pm Submit Quiz