Cloudera

CDP-3002
CDP Data Engineer- Certification Exam

For More Information – Visit link below:
https://www.examsempire.com/

Product Version

1. **Up to Date products, reliable and verified.**
2. **Questions and Answers in PDF Format.**

https://examsempire.com/

Latest Version: 6.0

Question: 1

What is the primary advantage of using Apache Spark for distributed processing compared to traditional single-node processing?

A. Improved data visualization capabilities
B. Increased storage capacity
C. Faster processing of large datasets
D. Enhanced data security

Answer: C

Question: 2

What is the role of a Spark driver in a distributed processing job?

A. Manages communication between executors and workers
B. Coordinates tasks across the cluster
C. Stores and processes intermediate data
D. Performs computations on individual data partitions

Answer: B

Question: 3

How does Spark achieve fault tolerance during distributed processing?

A. By replicating data across all nodes in the cluster
B. By implementing automatic checkpointing of intermediate results
C. By restarting failed tasks on different nodes
D. Spark employs a combination of data lineage tracking, checkpointing, and task re-execution to ensure job completion even in the event of node failures.

Answer: D

Question: 4

What is the difference between RDDS and DataFrames in Spark?

A. RDDs are mutable, while DataFrames are immutable
B. RDDs are lower-level abstractions, while DataFrames offer higher-level functionalities
C. RDDs represent distributed datasets, while DataFrames work with structured data
D. Both B and C

Answer: D

Question: 5

Which Spark component is responsible for managing the execution of tasks on worker nodes?

A. Spark Driver
B. Spark Executor
C. spark SQL
D. Spark Core

Answer: B

Question: 6

What is the purpose of partitioning data in Spark?

A. To improve data compression efficiency
B. To enable parallel processing across multiple nodes
C. To enforce data access control
D. To optimize data visualization

Answer: B

Question: 7

How does Spark handle data shuffling during distributed processing?

A. By broadcasting all data to each executor
B. By transferring only required data between executors
C. By storing all data on a single node
D. Spark doesn't perform data shuffling

Answer: B

Question: 8

You are working with a large, skewed dataset in Spark. How would you optimize processing to mitigate the impact of skew and improve performance?

A. Use salting on the skewed column during data partitioning.
B. Broadcast the skewed data to all executors.
C. Implement custom partitioners to evenly distribute skewed values.
D. Salting randomizes data distribution within partitions, custom partitioners ensure balanced distribution of skewed values, and broadcasting avoids redundant data transfers on shuffled stages.

Answer: D

Question: 9

How can you leverage Spark Streaming for real-time data processing and analytics?

A. By defining a streaming DataFrame with a window function.
B. By utilizing Structured Streaming with Kafka as the source and sink.
C. By implementing custom logic for data ingestion, transformation, and output.
D. Both A and B.

Answer: D

Question: 10

Explain the concept of lineage tracking in Spark and its benefits for fault tolerance and debugging.

A. Lineage tracks the dependencies between data transformations, enabling efficient re-execution on failures.
B. It creates a log of all operations performed on data, aiding in debugging issues.
C. Lineage allows for caching intermediate results, improving performance.
D. Both A and B.

Answer: D

Thank You for Trying Our Product

Special 16 USD Discount Coupon: NSZUBG3X

Email: support@examsempire.com

Check our Customer Testimonials and ratings available on every product page.

Visit our website.

https://examsempire.com/