Amp lab → UCB → spark

Databricks ->

Hadoop ⇒ java

Spark ⇒ scala

**Apache spark vs Databricks Spark**

1. Open source commercial/licensed
2. on-premises/cloud cloud
3. Need to import libraries to get

Advanced features

IAAS PAAS SAAS

Kafka

Pub-sub model ⇒ publisher subscriber

Producer

Consumer

Broker

Zookeeper

Topics ⇒messages

========================================

Hive table vs Delta table

1.Directly wont support ACID Supports ACID

Property

Only insert ,no update/delete DML operations

are supported

Stored in a Disk/hdfs/dbfs Cached

11."Match the following advanced concepts related to Spark's runtime architecture and application deployment with their appropriate descriptions:

Driver's role in task scheduling

spark-submit --packages option

Resolving dependency conflicts with spark.jars.packages

YARN's NodeManager in Spark job execution

Options:

A) The component that is responsible for negotiating resources with the Cluster Manager and plays a crucial role in task distribution to Executors.

B) A spark-submit option that allows the user to specify Maven coordinates of jars to include on the driver and executor classpaths.

C) The YARN component that manages the user job's executors on a node, monitoring their resource usage (CPU, memory, disk, network) and reporting the same to the ResourceManager.

D) A Spark property that can help resolve jar dependency conflicts by specifying the coordinates of the Maven artifacts to be included in the runtime."

12. "For a seasoned Spark data engineer, match the following nuanced Spark configuration parameters with their precise effects on Spark application tuning and performance optimization:

spark.executor.memoryOverhead

spark.memory.fraction

spark.memory.storageFraction

spark.driver.maxResultSize

Options:

A) Adjusts the proportion of JVM memory dedicated for Spark to use, leaving the rest for user data structures and overhead.

B) Specifies the amount of non-heap memory to be allocated per executor, beyond the memory used for caching and storing tasks' data.

C) Controls the size of the maximum result that can be fetched by the driver from the executors at one time, affecting the driver's ability to handle large collected datasets.

D) Dictates the fraction of Spark memory to be reserved for storage of Spark internal data structures, with the remainder available for execution and tasks.