One hql script to run end to end job, if any stage fails in between how to restart from that particular stage again and not to reprocess entire thing

Map and mappartitions - map partitions applies map once per partition

Spark partition and hive partitions

Spark partitions are done for performance optimization and data is segregated not on basis of some common value but rather more or less equally distributed. Hive distributes it either randomly using bucketing for map side queries and faster latency or for distributing against a particular value

Hive optimization techniques

Vectorization - Vectorized query execution streamlines operations by processing a block of 1024 rows at a time

Bucketing -

Partitioning -

Mapside joins - like broadcasting

Smb joins - sort merge bucketing joins, used for 2 large tables,

Parquet vs orc - parquet goe well with spark for storage and data aggregation as its columnar and orc goes well with hive for bucketing, dynamic partitioning , vectorizations etc

Repartition is on key or data - its on data, can be on key if mentioned column name

Driver program - The spark driver is **the program that declares the transformations and actions on RDDs of data and submits such requests to the master**. In practical terms, the driver is the program that creates the SparkContext, connecting to a given Spark Master.

Classes vs objects - main method lies in object and it has features like singleton object, class is like a blueprint, it stores variables and methods which can be extended within any object

Out of memory exceptions - during broad transformations, using improper persist, wrong infrastructure

How to schedule java and spark job in oozie

Clientmode and cluster mode - when client machine is on cluster itself we can use client mode to submit jobs but if its away it can lead to job failures if node crashed

How spark access data

How to filter null data in df

Df vs rdd - df is for structured data with schema and run time safety used fr aggregations, joins etc, rdd is for structured/unstructed data where affaining schema is not necessary and performance is not a primary concern

Data skewness - repartitioning

How to handle different delimiters in dq

Hive udfs

Mapside joins - bucketing

How to use spark standalone

Safe mode in spark

How to retrieve corrupted blocks in hdfs