Repartition vs coalesce

Repartition -> used to increase or decrease the number of partition to improve parallelism

->induces full shuffle -> creates new stage

->evenly distribution of data in partition

Coalesce -> used to decrease the number of partition

-> no shuffling

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

External tables are created on top of the existing data

When you drop managed table , it deletes both metadata and data

Where as in external table , it deletes only metadata but data remains same. No impact on the data

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

Dataframe writer :

Mode : overwrite

Ignore

Append

patientdf.write.format("orc").mode("overwrite").save("file:///home/hduser/target/patient\_orc")

patientdf.write.format("orc").mode("ignore").save("file:///home/hduser/target/patient\_orc")

patientdf.write.format("orc").mode("append").save("file:///home/hduser/target/patient\_orc")

Storage optimization 👍

1. File format ⇒ parquet or orc --to optimize the storage space
2. You can create partitions based on the columns ,where the distinct column values are less (20-200)
3. You can create bucket to fragment the data where distinct column values are more (1000--10000)

Hash value (key) mod no of bucket

Hash value (store ) mod 4

Hash value(“abc store) mod 4 ⇒1

78645 mod 4 ⇒

Xyz store -> 556780 mod 4⇒ 0

Pqrs store --. 11111321 mod 4 ⇒ 1

B1 B2 B3 b4

0 1 2 3

Xyz store abc store ,pqrs

Sampling →