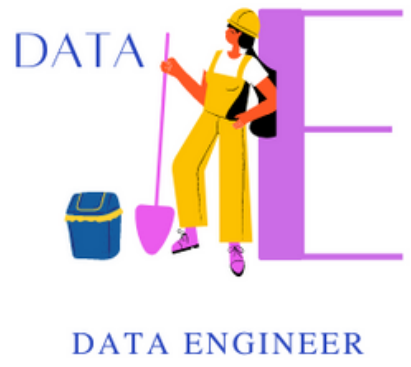
Hive Optimization



1. Table Structure Level Optimization
    i. **Partitioning**
    ii. **Bucketing**
2. Optimize The Query
3. Query Expression By Window Function

Partitioning

- Split the data based on Column like Country/City/State. into multiple machines
- When the cardinality of the data is low, then we use partitioning
- When we partition the data then it creates the directories after the process in the warehouse

If I partition the data based upon the city column ;
Then location of the directory will be like :

/user/hive/warehouse/db/city = Delhi
/user/hive/warehouse/db/city = Hyderabad
/user/hive/warehouse/db/city = Bangalore

- **Types of Partitioning Technique**
  - Static (Manual Process)
  - Dynamic (Automated Process)

Static Partitioning
vs.
Dynamic Partitioning

Static

when we have idea about the data then go for Static partitioning

It's manual process

It's not scalable as it's manual

Dynamic

when we don't have idea about the data then go for dynamic partitioning

It's automated by Hive

It's scalable cause it is automated

Bucketing

1. Bucketing is a process of splitting the data while cardinality is high.
   (if we do partitioning when cardinality is high then it will create alot of directories which is not a recommended process)

2. When we apply bucketing technique , It creates Files not Directories
3. It optimizes the Join Operation