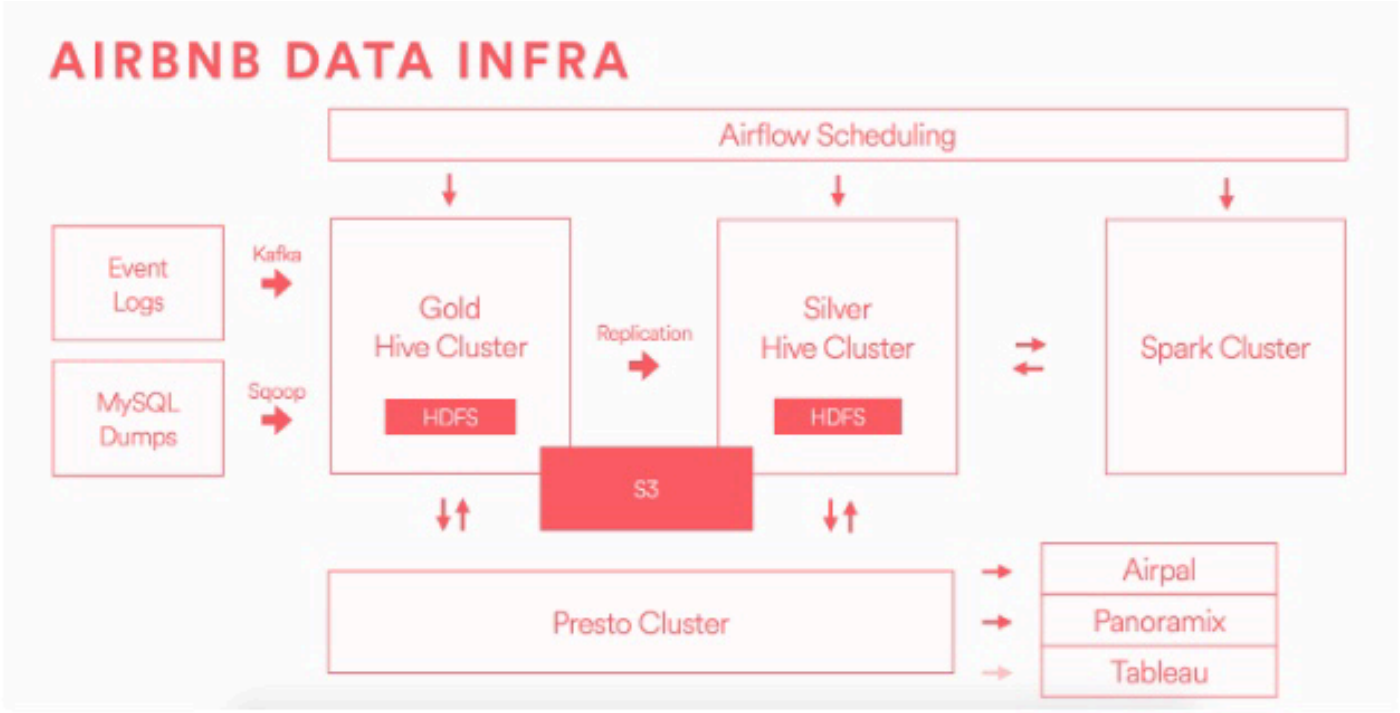
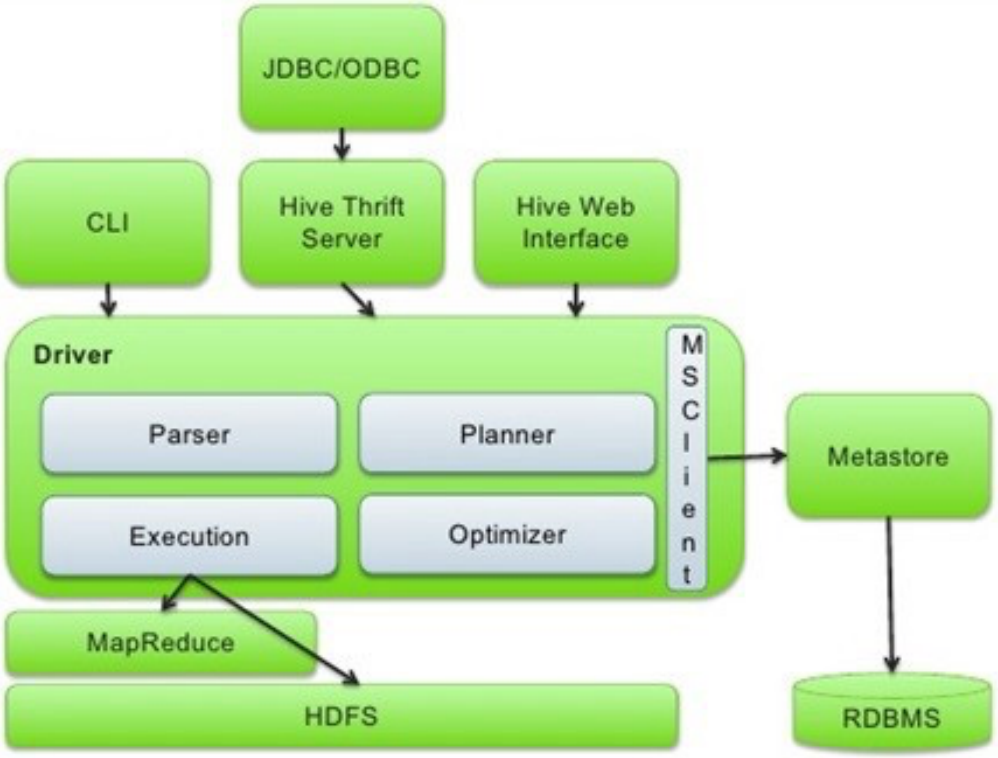
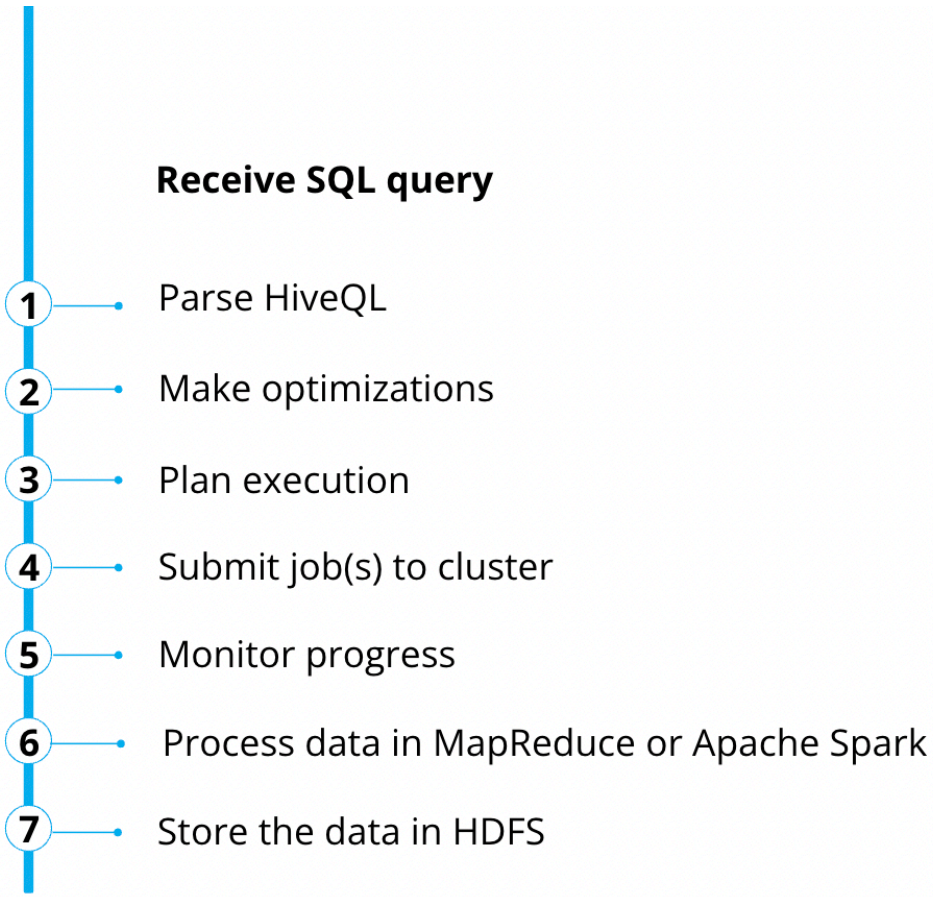
Agenda

What are data warehouses and what is OLAP?
How does Apache hive act as a data warehouse?
What makes hive as Hive? (Architecture)
How to interact with Hive?
What are different HQL commands that are used commonly?
What are the optimization techniques in Hive ?
Use cases of hive/ where to use hive/why and why not

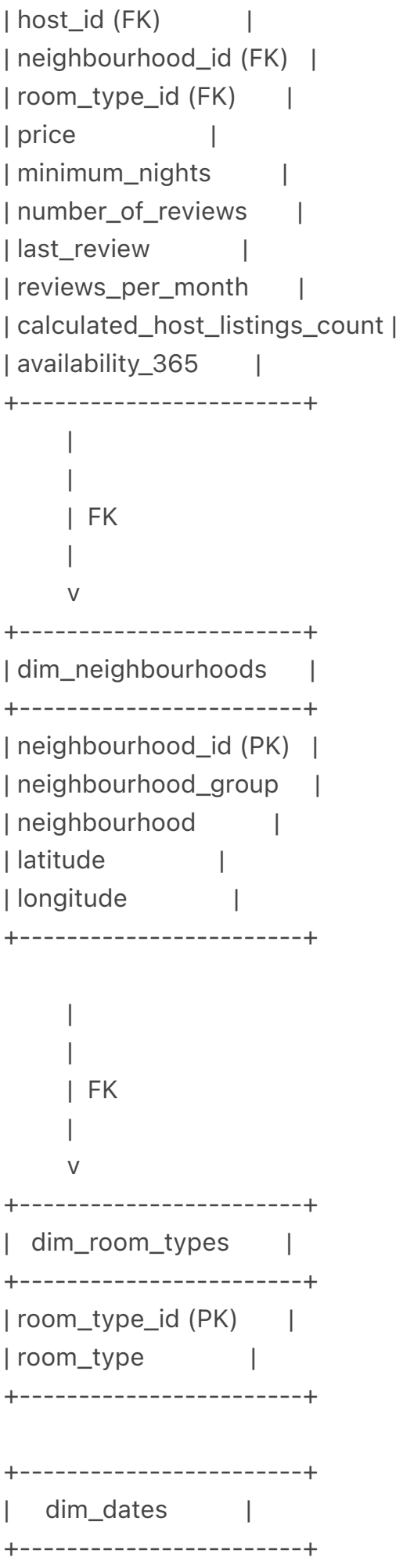


Apache Hive Architecture






```
+-----------------------+
|    dim_hosts          |
+-----------------------+
| host_id (PK)          |
| host_name             |
| host_since            |
| host_location         |
| host_response_time    |
| host_response_rate    |
+-----------------------+
       |
       |
       |  FK
       |
       v
+-----------------------+
|   fact_listings       |
+-----------------------+
| listing_id (PK)       |
```


| host_id (FK) |
| neighbourhood_id (FK) |
| room_type_id (FK) |
| price |
| minimum_nights |
| number_of_reviews |
| last_review |
| reviews_per_month |
| calculated_host_listings_count |
| availability_365 |
FK
dim_neighbourhoods
neighbourhood_id (PK)
neighbourhood_group
neighbourhood
latitude
longitude
FK
dim_room_types
room_type_id (PK)
room_type
dim_dates

| date_id (PK) |
| date |
| year |
| quarter |
| month |
| day |
| day_of_week |
+----------+

Hive

↳ Datawarehouse developed by FB in 2002

↳ open Source Apache

↳ High abstraction layer on top of MR.

↳ use HQL

↳ Suitable for structured Data

Hive Architecture

(a) Hadoop Core Components (Hive-site.xml)

(a) HDFS → Storage

default ↳ /user/hive/warehouse

(b) Map Reduce

SQL → (MR) Jobs

(b) [Metastore] ... all metadata

(b) ↳ Crucial part for ... information related to table, Columns, partitions, ... present in it. (Col N + Col DT)

↳ Derby | Mysql | oracle | Postgres | Sqlite.

(c) Hive Server :-
↳ enables client to execute query against Hive.
↳ Designed to provide better support for open API like (JDBC/ODBC)

(d) Compiler
↳ Parses the query
↳ Does semantic analysis
↳ Query expressions
↳ generates an execution plan (table MD)

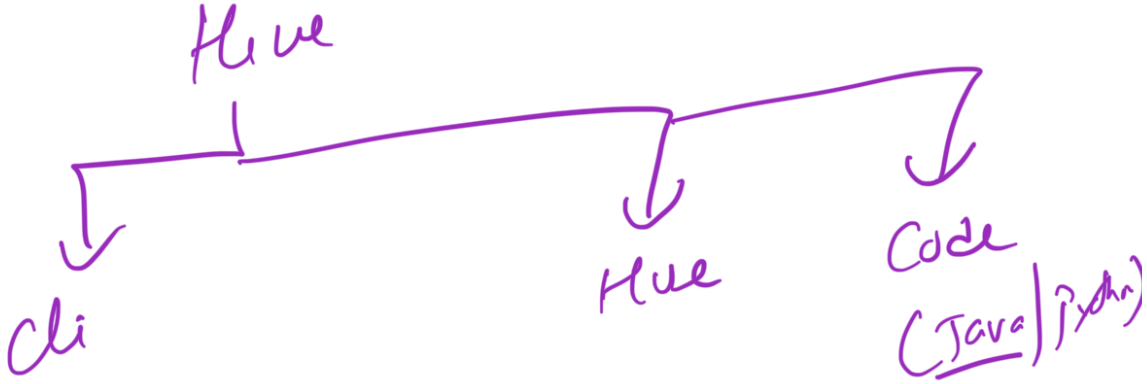
(e) Planner

(f) Execution Engine :
↳ Planner presents the execution plan to Engine.

(g) ... form of

↳ plan is ... stages.

↳ Sql/HQL will be converted to Map Reduce Jobs. (Jar)



Hive (*)	Beeline
→ Native cli tool	→ JDBC based CLI
→ NO Concurrency support	→ Concurrency support
→ Fat client (Jar)	→ thin client → only relevant
→ Higher resource usage as it connects directly to MS.	→ lower resource usage as it runs on HiveServer2
→ NO support to advanced	→ Supports all advanc features like ACID, CLAP.

Question : Get the details of all travelers about their choice of stays in New York city i.e whether they book Entire home/apt or Private Room or they preferred Shared room

Ans.1 (a) Create database ;

(b) Create tables

(c) Insert Data

(d) Query

(e) Sumedh → Mgr → 202 → folder → .db → HDFS

(1) Database → 27aug 2024

(2) Table

Optimizations

(1) Apache Tez (2013-15)

Set hive.execution.engine = tez;

(2) Apache Spark