#ASLI ENGINEERING

Discord's Data Platform

SWIPE

BY

ARPIT BHAYANI

Discard's Data Platform

Why companies need a data platform?

- business decisions and strategies
- power data science - train ML models
- deep product insights

Data doesn't lie

Can they not work with databases directly?

- each service has a separate DB
- databases can be of different kinds

  eg: profile service → MongoDB
  payment service → MySQL

  } Making sense from both the data requires them at the same place (DB)

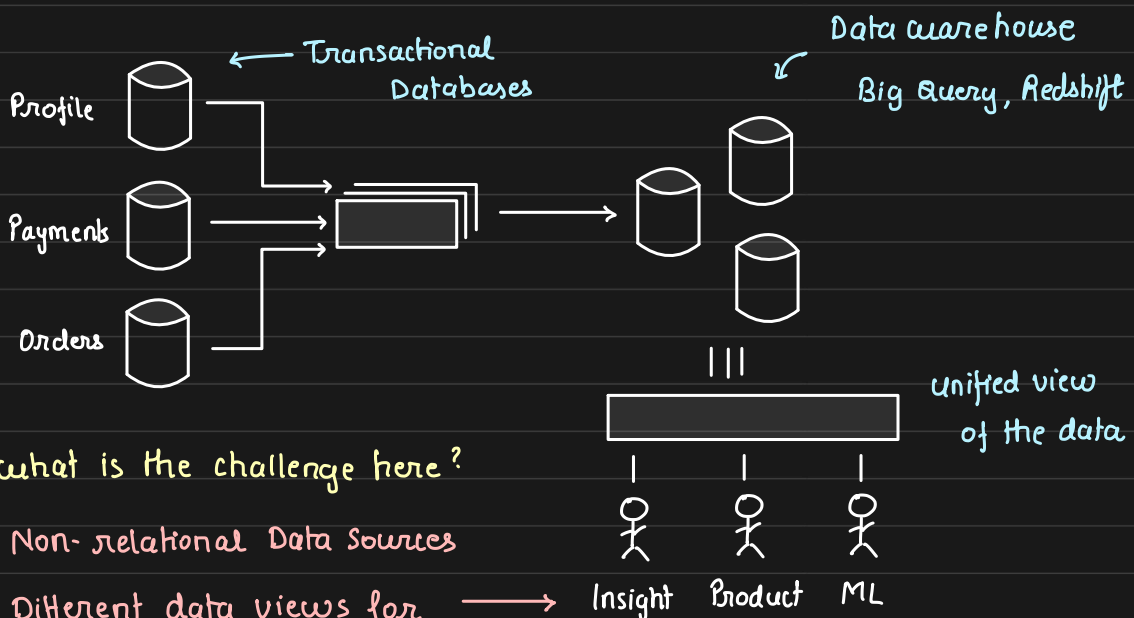
- cannot make cross DB "joins" and "queries"
  ↳ requires a ton of time and computation
- analytics queries will choke transactional DB throughput

Hence all the data is duplicated and stored in a centralized location - Data Warehouse / lake

↳ unified view of data / information

↳ crunching stat / insights is easy

Discord's Data Platform - Derived



So, what is the challenge here?

1. Non-relational Data Sources
2. Different data views for → different end usecases

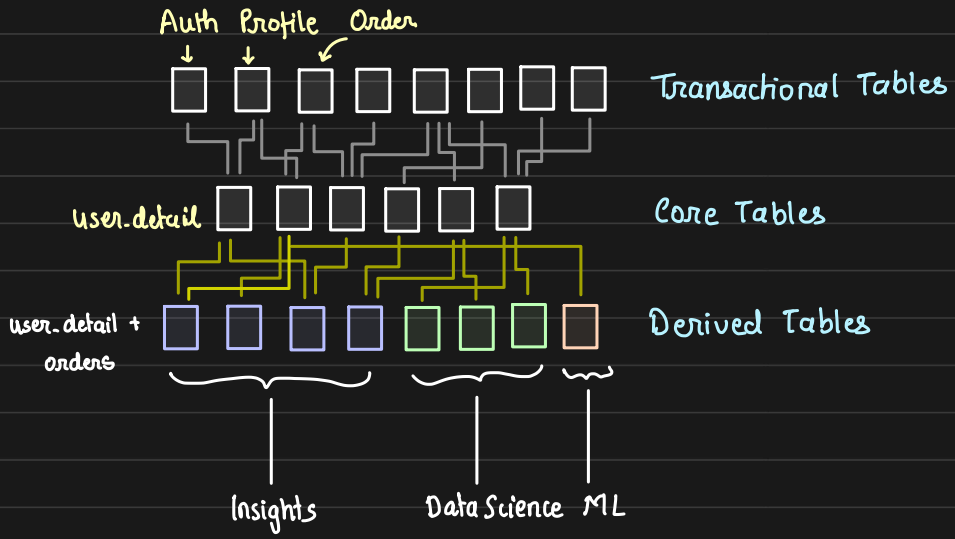
eg: Table 1 + Table 2 + Table 3 → Recommendation

Table 1 + Table 5 → Retention analysis

Table 2 + Table 4 → Payment insights

Note: different tables need to be joined / merged and kept to efficiently power usecases

* End usecase should need to do minimal joins

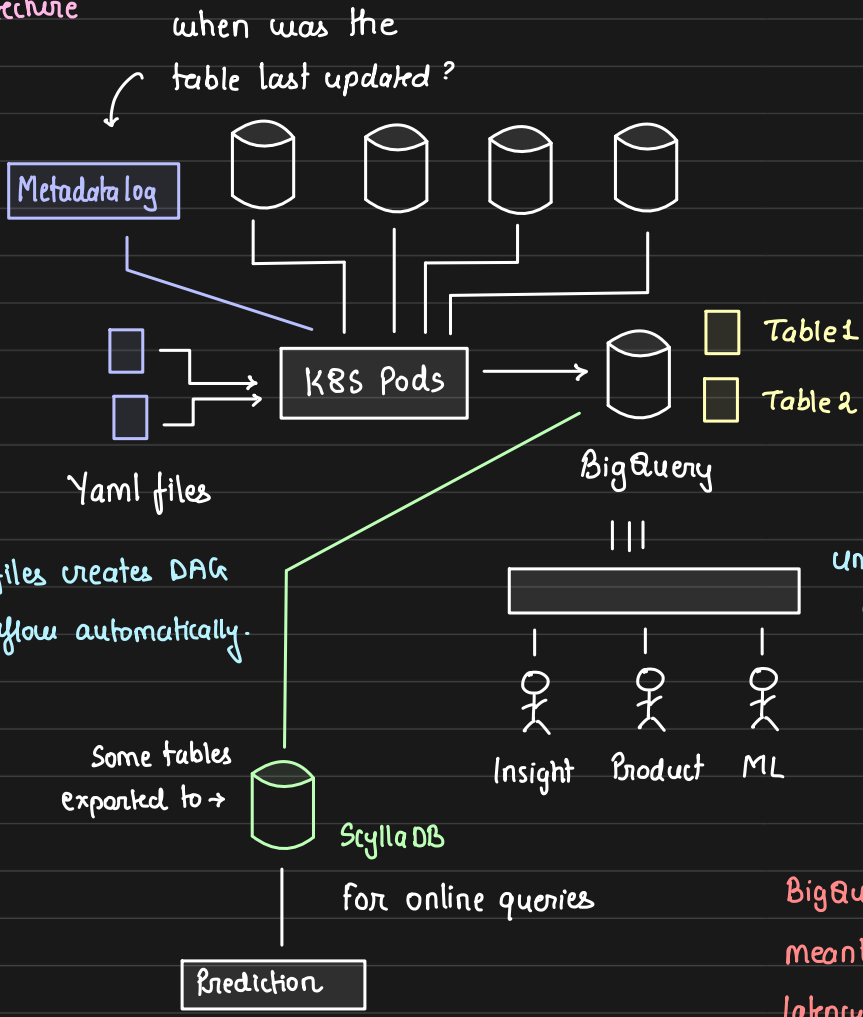


Creating a new derived table

New table is configured in YAML along with details such as

1. columns of new table
2. strategy: merge, append, replace
   ↳ while writing the new values what should we do
3. schedule
4. window
5. partition_by
6. cluster_by → sort within partition
7. dataset
8. sql → select query that is executed and output is put in the derived table

Architecture



BigQuery is not meant for millisecond latency usecase.