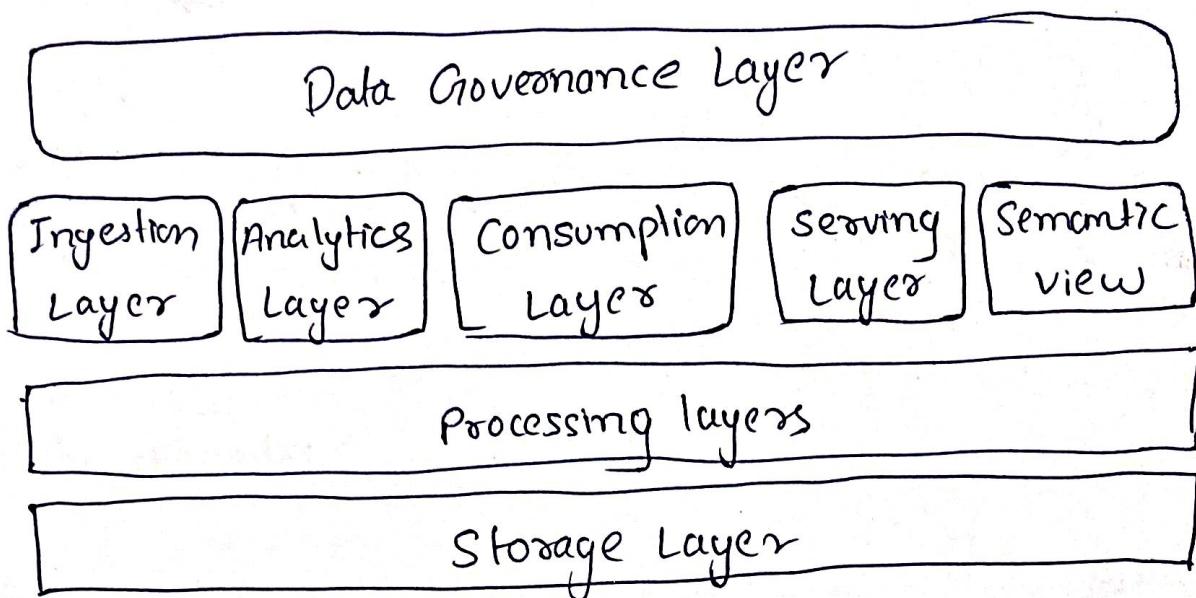


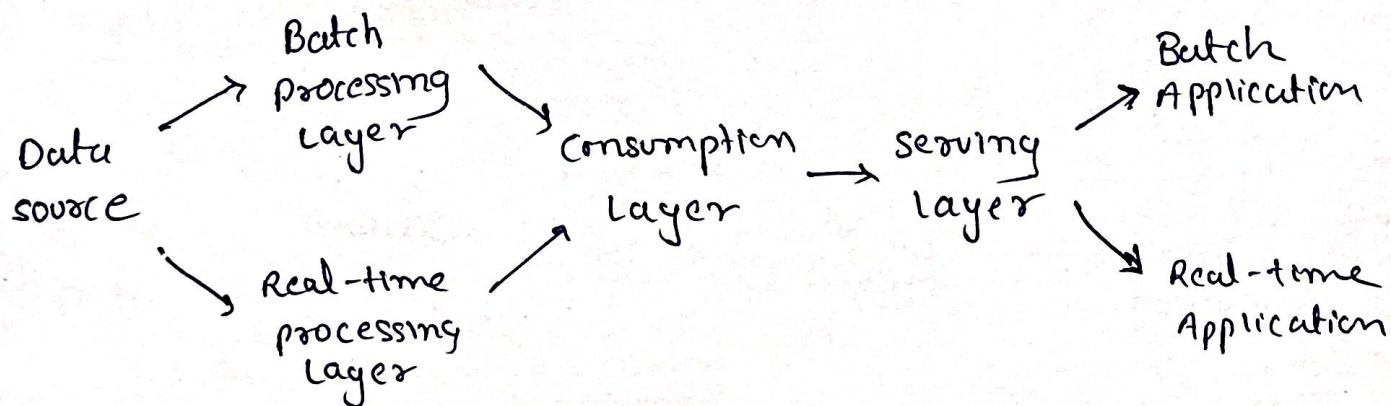
Data Architecture

①

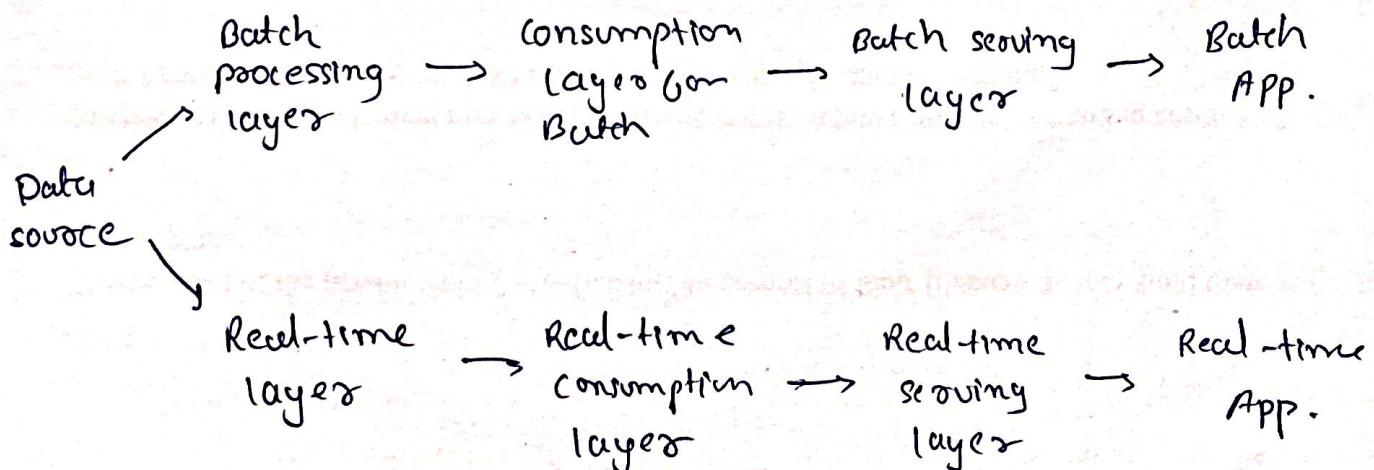


Data platform architecture

② Lambda Architecture :

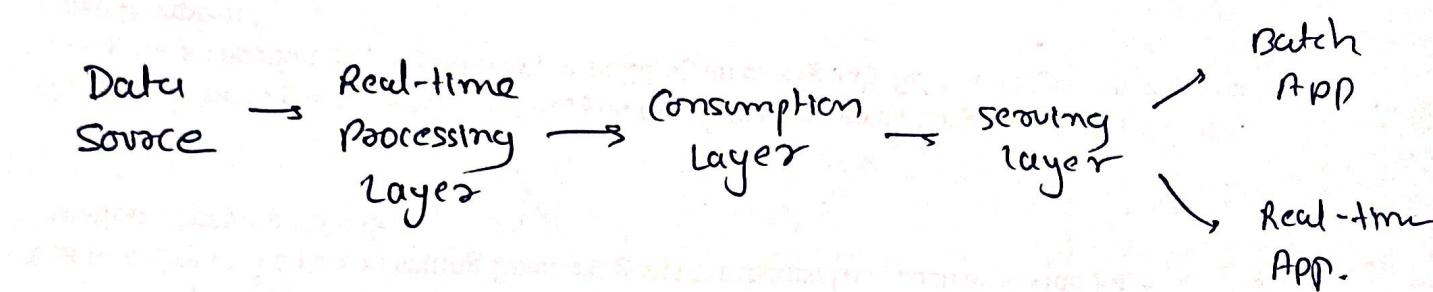


Combined Lambda Arch.



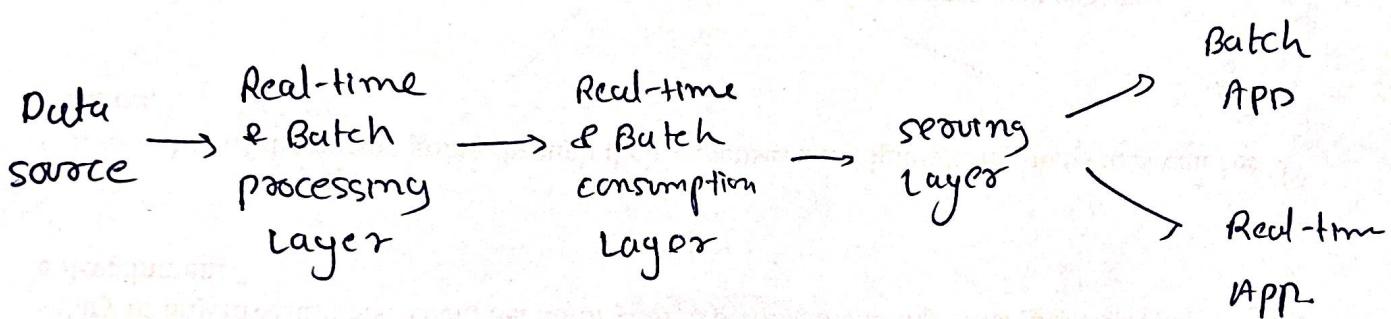
(3)

Kappa Architecture



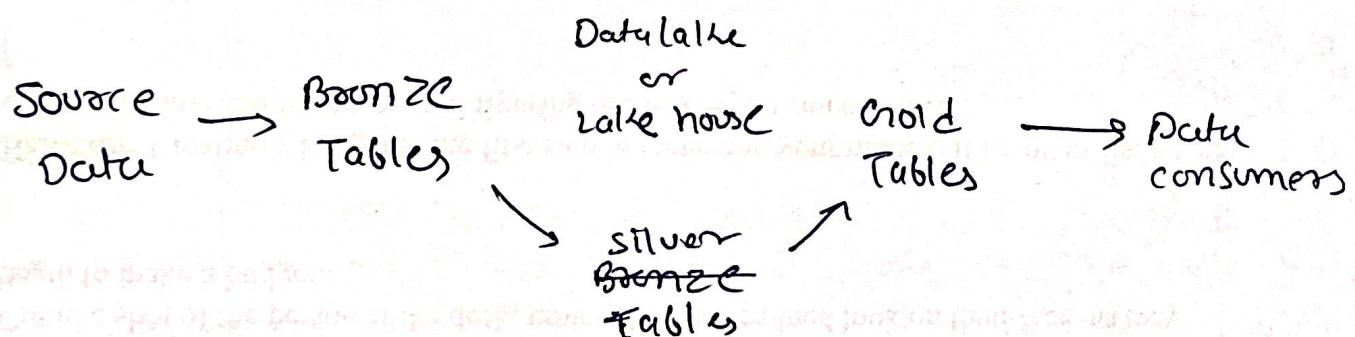
(4)

Delta Architecture

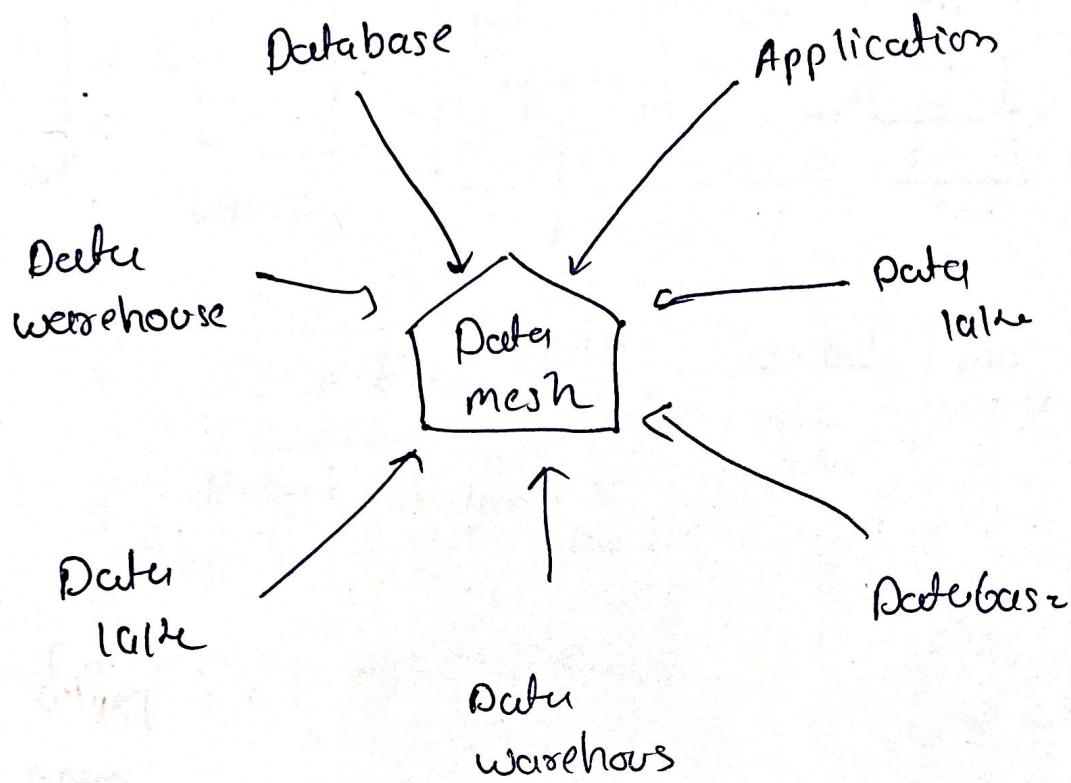


(5)

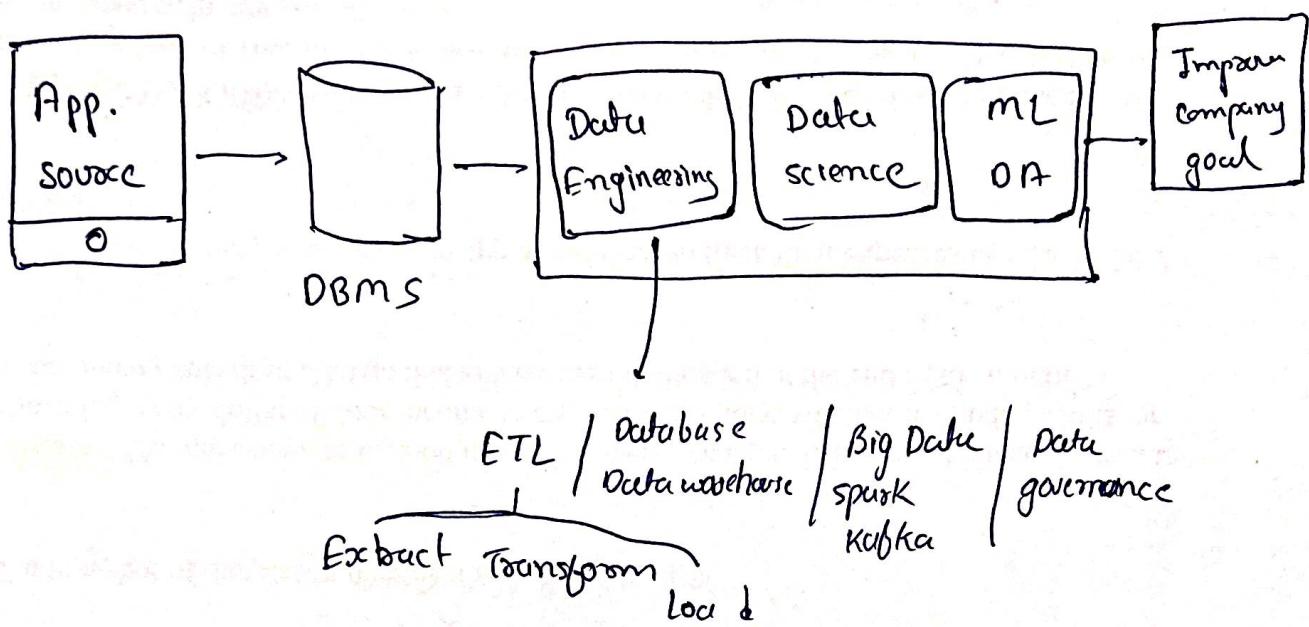
Medallion architecture



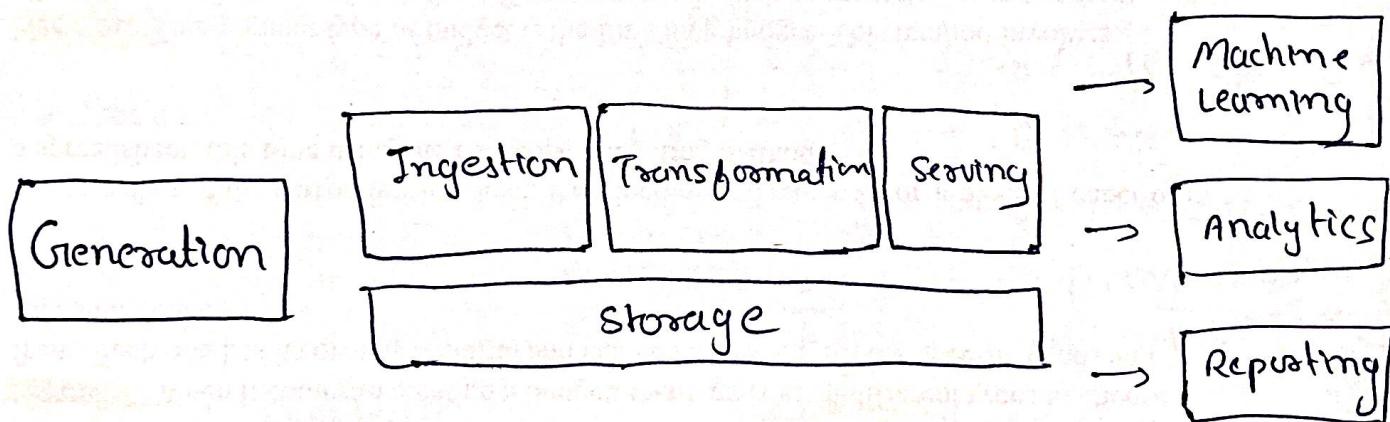
③ Data mesh Architecture



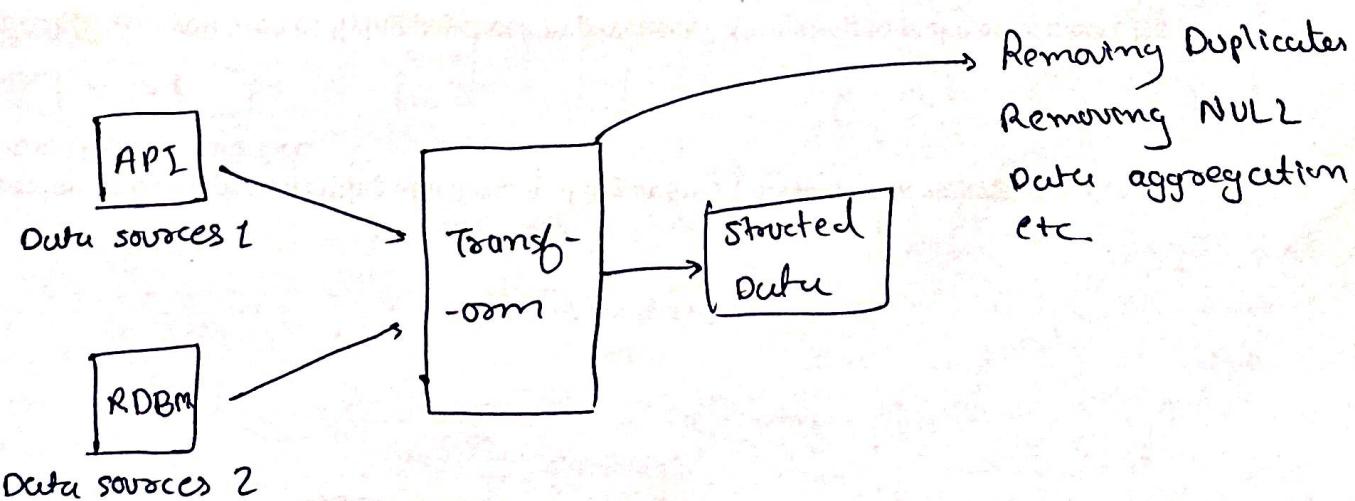
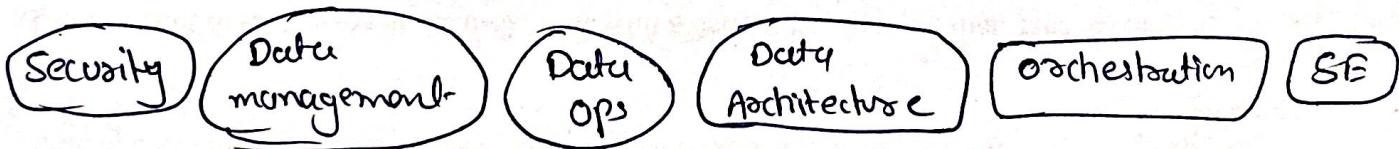
Data Engineering



⇒ Data Engineering Life-cycle



Undercurrents:



=>

Data generation

- Transactional info
- IoT devices
- Social media
- Log & machine
- API / Third party

Data storage

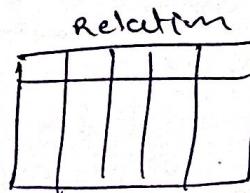
- Relational DB
- NoSQL
- Data warehouse
- Object storage
↳ Data lake, S3

⇒ DataBase management system

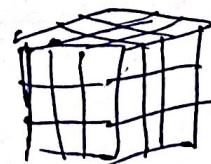
- PostgreSQL
- MySQL
- Oracle



Structured query language (SQL)



relational

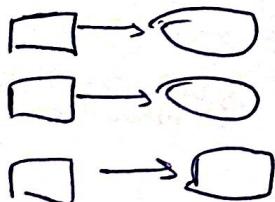


analytical

- row wise operation
- CRUD oper.
- Data modelling
(Defining relation w/ tables)

NoSQL Database

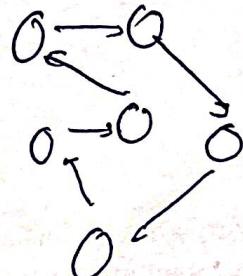
Key-value



Column-family



Graph



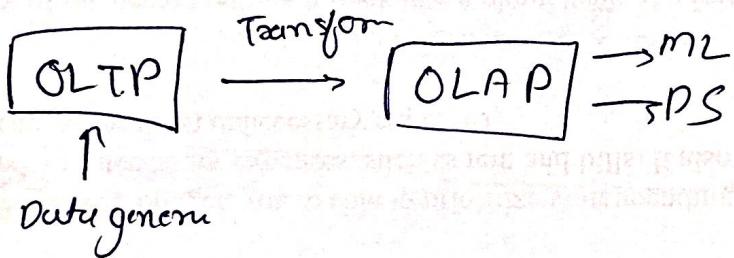
Docu.



⇒ Data Storage Processing

- | | |
|-----------------------------------|--------------------------------|
| OLTP | OLAP |
| - Online transactional processing | - Online Analytical processing |
| - Relational DB | - Data warehouse |
| - Transactional Data | - Analytical workload |
| - Row base | - Column base |
| - Effective for CRUD | - Effective for analysis. |

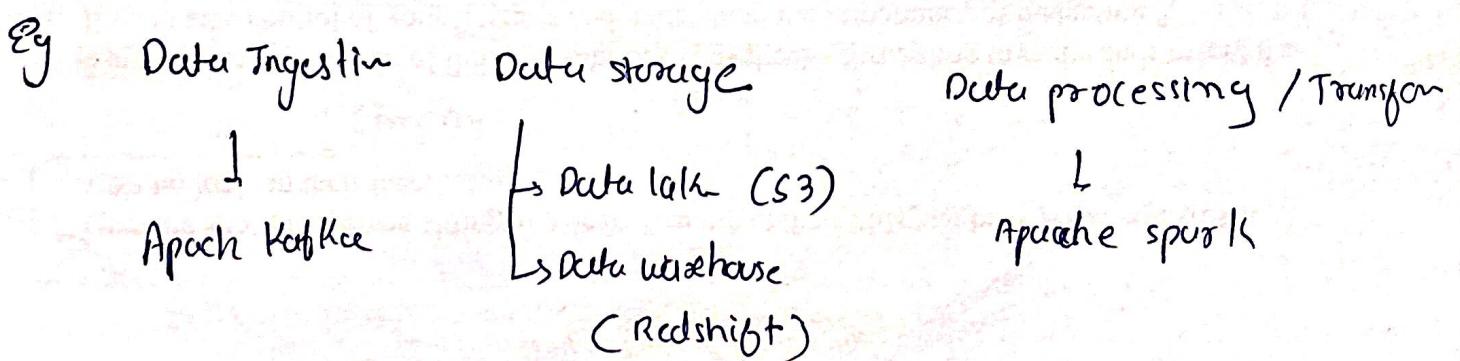
⇒ generally



⇒ Data Architecture

1) Business Needs (Operational Arch) : Answers why ?

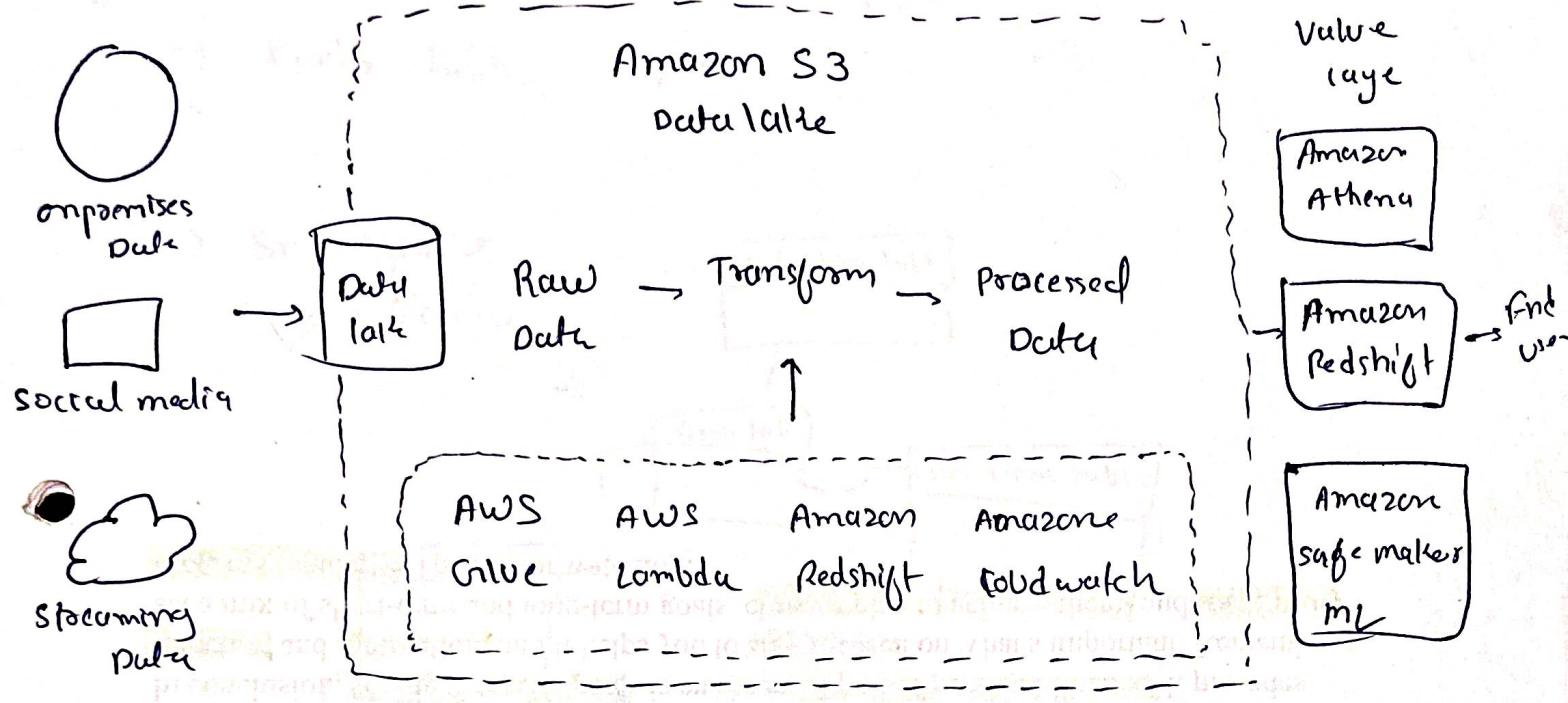
2) Technology Integration (Technical Arch) : How ?



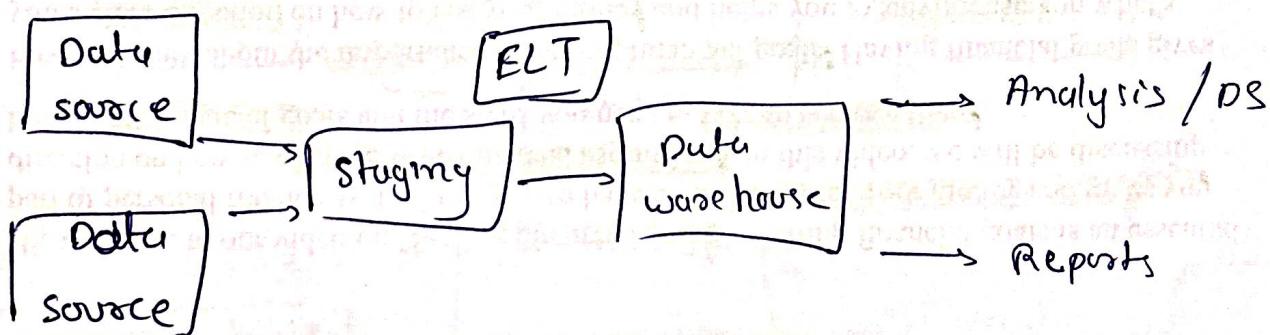
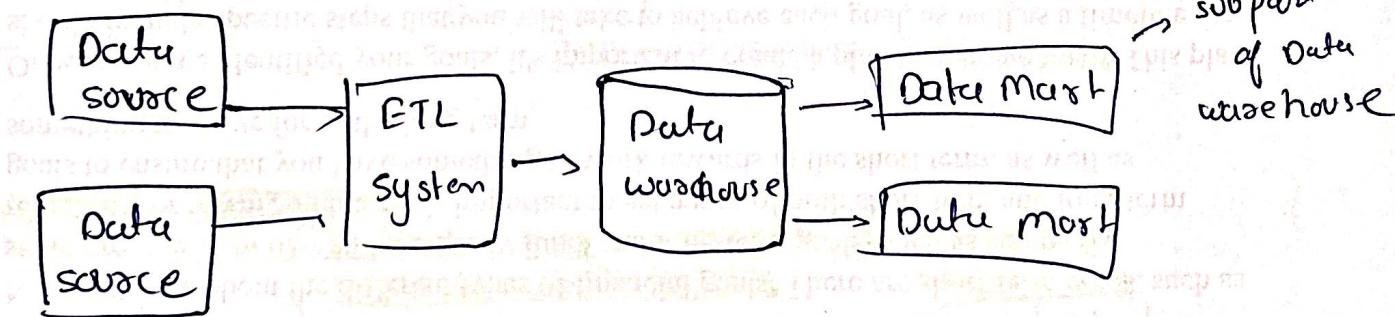
Source Data

Store, Ingest & Backup

Visualize



⇒ Data Warehouse

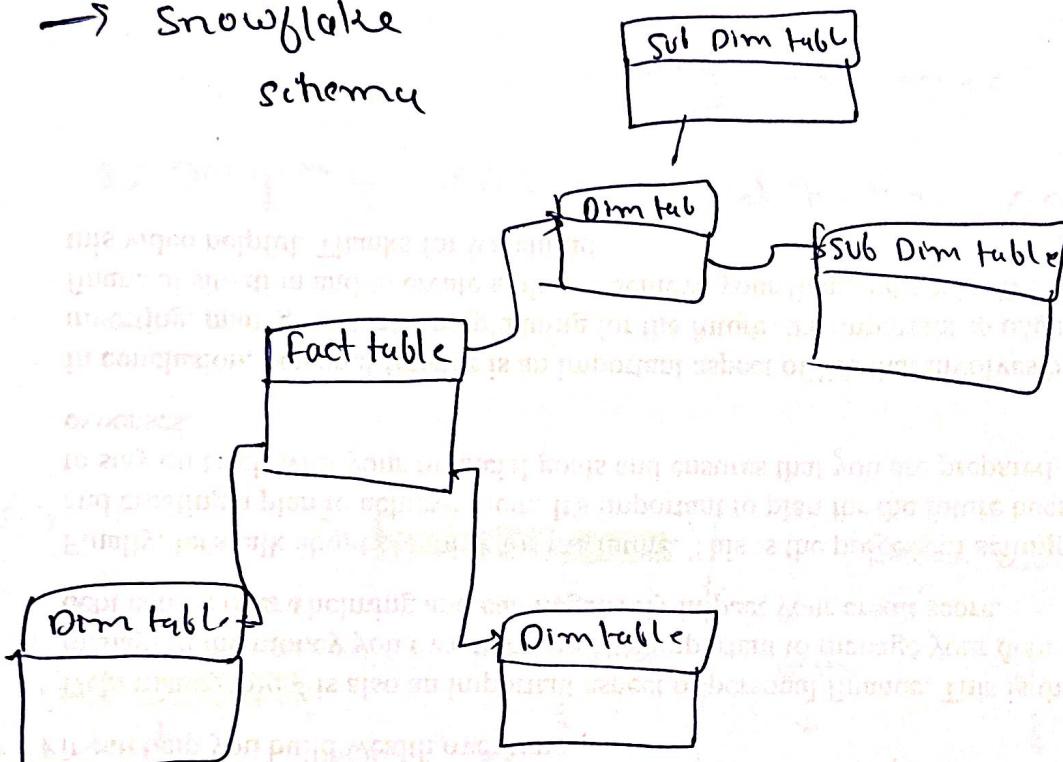


=> Dimension modelling

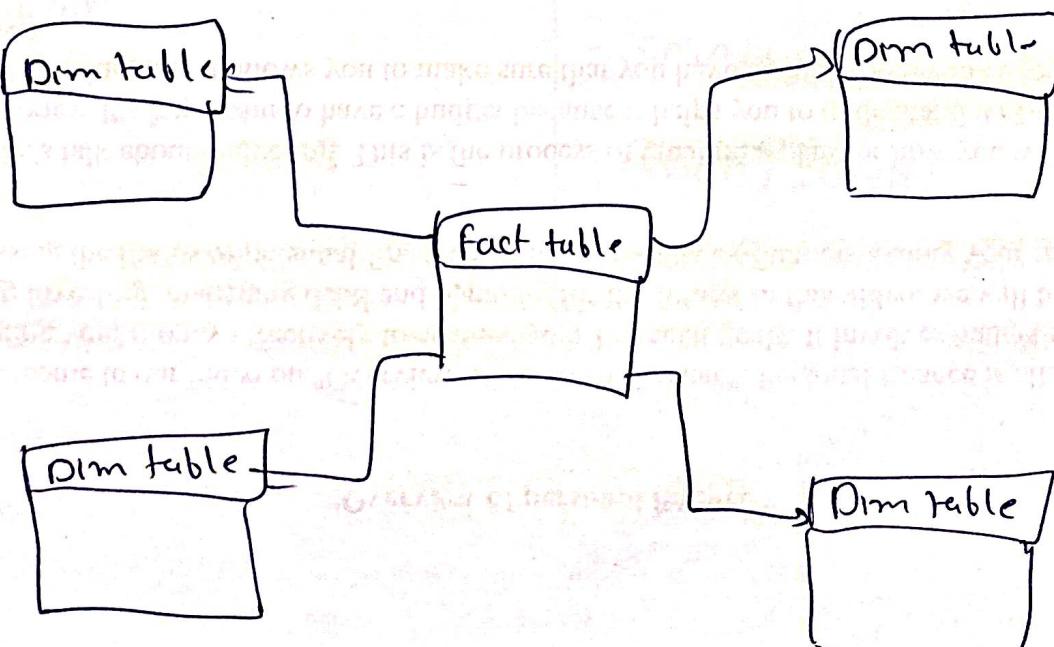
1) Dimensions table (multiple)

2) Facts table -(1)

→ Snowflake
schema



=> Star schema



\Leftrightarrow Slowly changing Dimension

1) Designing SCD1 : overwrite the value & no history

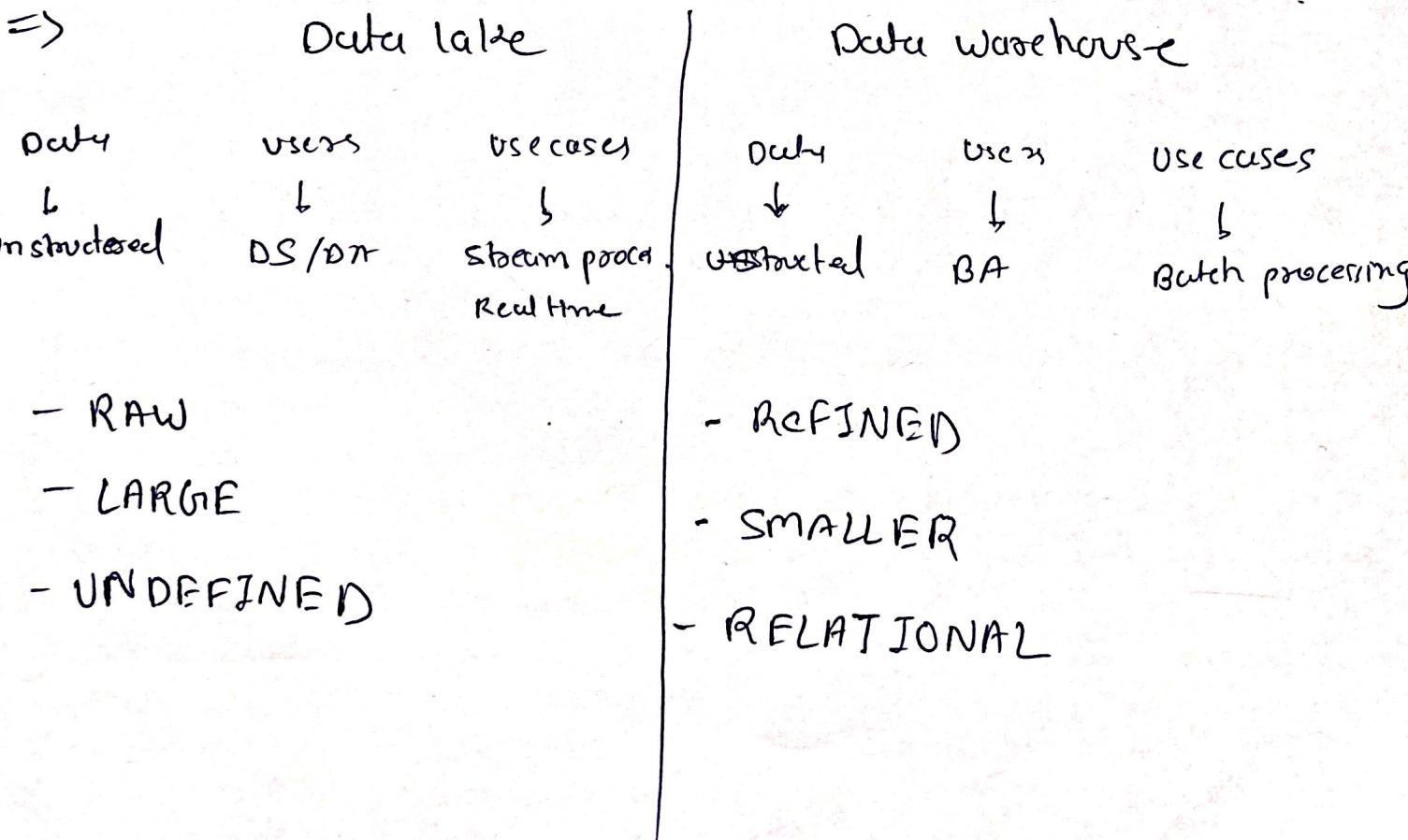
2) Designing SCD2 : Using flag : isActive col

Using version name max(version)

very date range end date == NULL

3) Designing SCD3 : partial history (prev city)

4) Designing SCD6 : combine 1, 2 & 3.



* MUST Learn

- Python
- SQL
- Linux Commands
- Data warehouse [Snowflake , Amazon Redshift , BigQuery]
- Data processing [Apache spark , Apache Kafka]
- Data Orchestration [Airflow]

SQL cheat sheet

1) Level 0

- CREATE DATABASE student ;
- USE student ;
- DELETE FROM student # Delete the rows where marks < 20 ;
- UPDATE table-name # Change the tables data
SET column = new value
where condition
- DROP table-name if exists # Delete the table
- ALTER table-name # changing table
Add column-name datatype constraint schema

2) Level 1

- SELECT statement
- WHERE clause # for condition
- GROUP BY & ORDER BY # for aggregate function.
- HAVING clause
- LIMIT & ALIASING

3) Level 2

- JOINS (INNER, LEFT, RIGHT, SELF)
- UNIONs - window function
- STRING function
- CASE Statement
- Subqueries

ii) level 3

- CTEs
- Temp tables
- Creating Procedures
- Events & Triggers