

① Data Engineering -

② Datalake, Datawarehouse, Datamart, DataSwamp, Metadata.

③ Data pipeline

④ ETL

⑤ ELT

⑥ CDC

⑦ ETLT (ELT - subpattern)

→ Data pipeline.

① Data Engineering -

→ On-prem

Early 2000s → Database → Managed DBA

Late 2000s → New Apps → New data. → Scattered

Early mid 2010s → Social media ↑ Data driven ↑

Cloud DBs

Database Inconsistent.  
noise. ↓  
not Central storage  
with. ↑  
parallel. ↑

Mid-late 2010s → Warehouse, pipeline & reporting -

Extract, Transform, Load (ETL) → Extract, Load, Transform (ELT) → Extract, Load, Transform (ELT) More Adaptive.

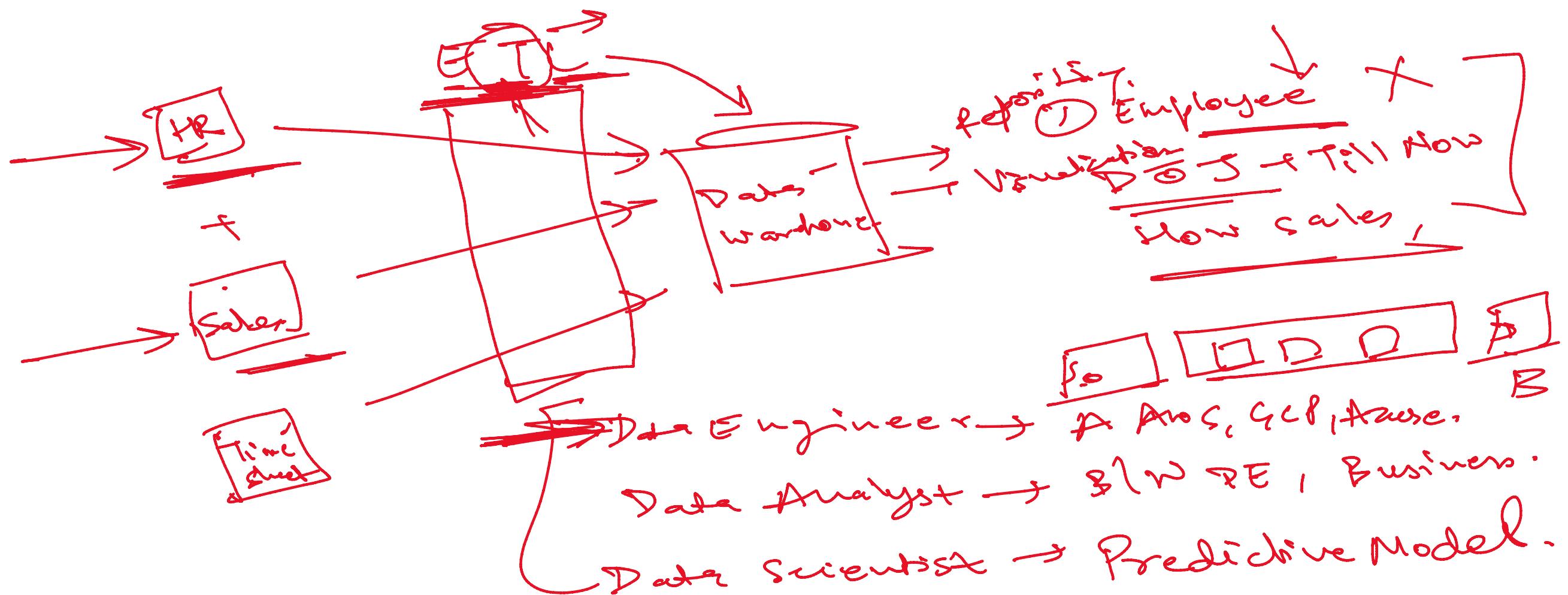
Extract, Transform, Load (ETL) → Extract, Load, Transform (ELT) → Extract, Load, Transform (ELT) More Adaptive.

Cloud migration → Extract, Load, Transform (ELT) More Adaptive.

② Snowflake ① Amazon Redshift

② Snowflake ① Amazon Redshift

Early 2020s → Instagram, YouTube, Twitter, TikTok.



## Dataops →

### ① Data Lake:-

- ① Collection of Huge Amount of data.
- ② Repository
- ③ Collection of Unstructured Data.
- ④ Methodology for transformation of data.

### ② Data Warehouses

- ① Collection of Historical Data + Read more <sup>than transformation</sup>
- ② Structured Data.
- ③ Unstructured Data, Staging Plane.

### ③ Data lakes:-

- ① Central Storehouse all the data.
- ② Any type of data → Structured, Unstructured, Image, video, log etc.

ELT  
↑↑

ETL  
↔↔

video, log etc.

③ Load first, Think later

Metadata — Provide info / structure of Actual Data.

Data Samples

Deteriorated & Unmanaged data later

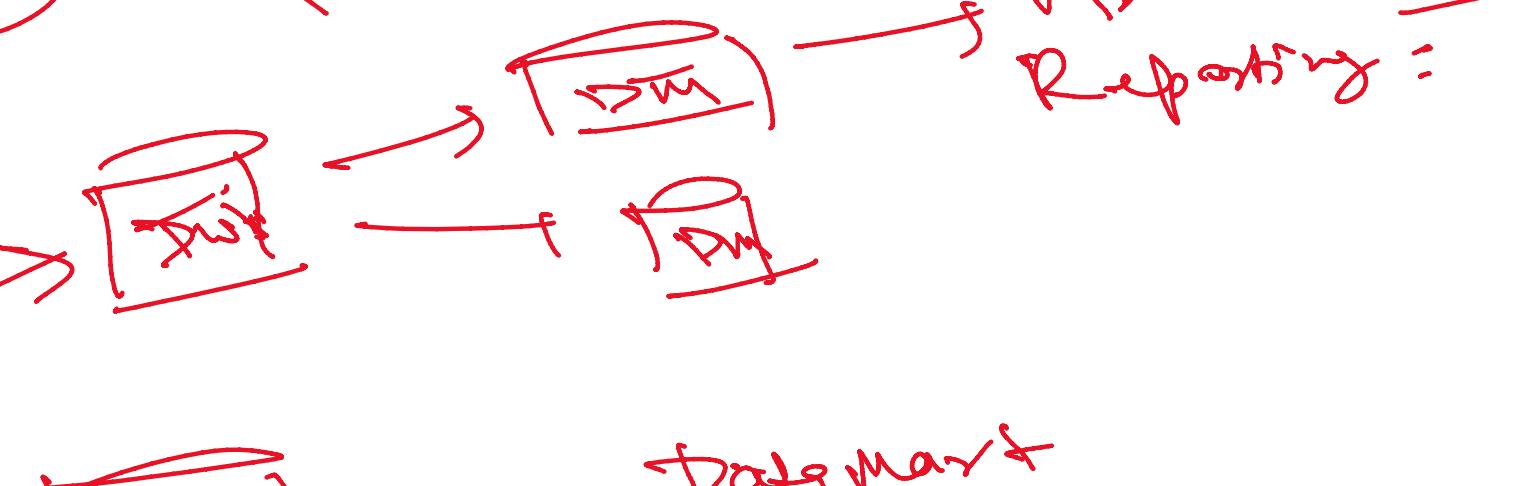
② Data warehouses

↓  
Central  
Warehouse  
in 1980s

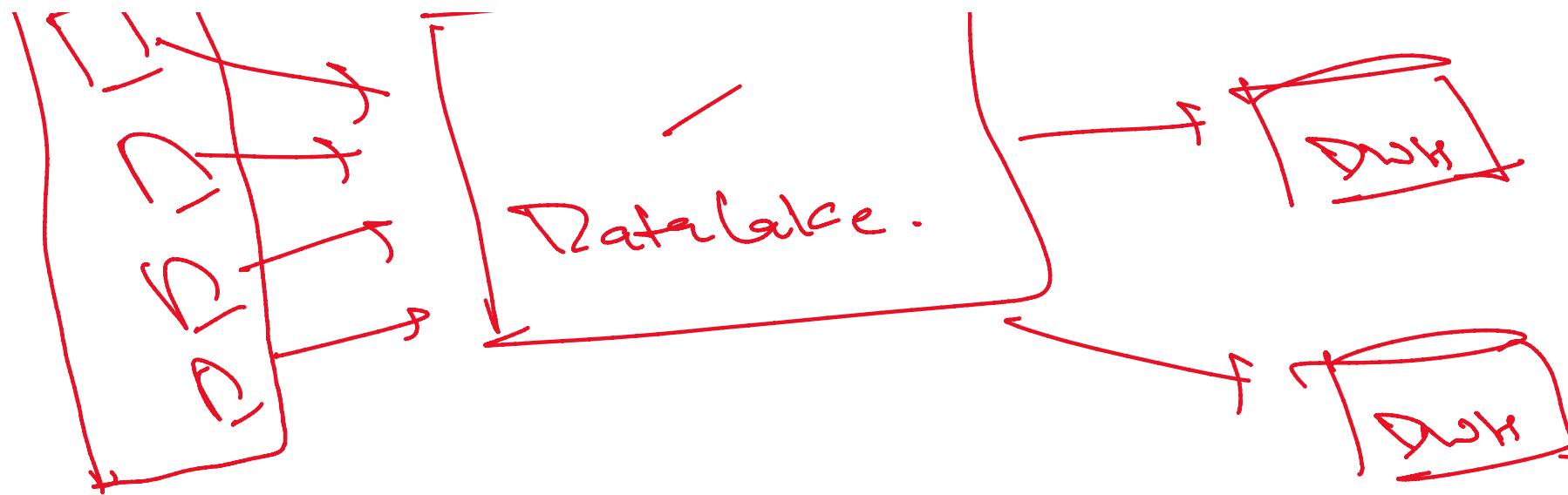
① Think first, load later.  
② Support Structured Data -

③ Business Intelligence -

④ SQL Queries



Data Mart



Data Mart



Subset of

Database

Data lake

1. Unstructured data, semi-structured, structured data.

2. Load first, Think later.

3. Built on low storage unit-

4. Highly Agile.

5. Advanced Analytics.

Data warehouse

1. Structured Data.

2. Think first, load later.

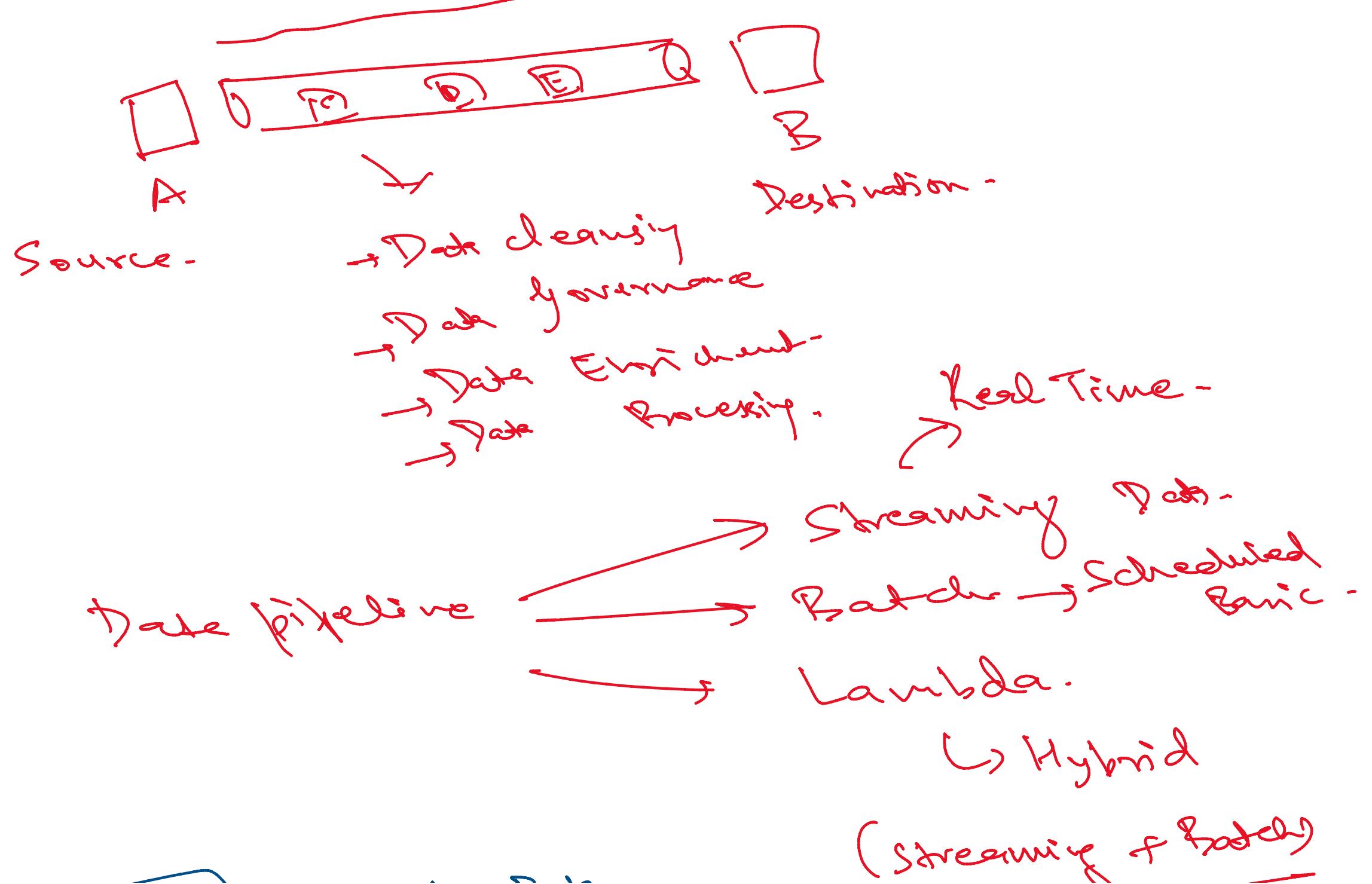
3. Data storage costly -

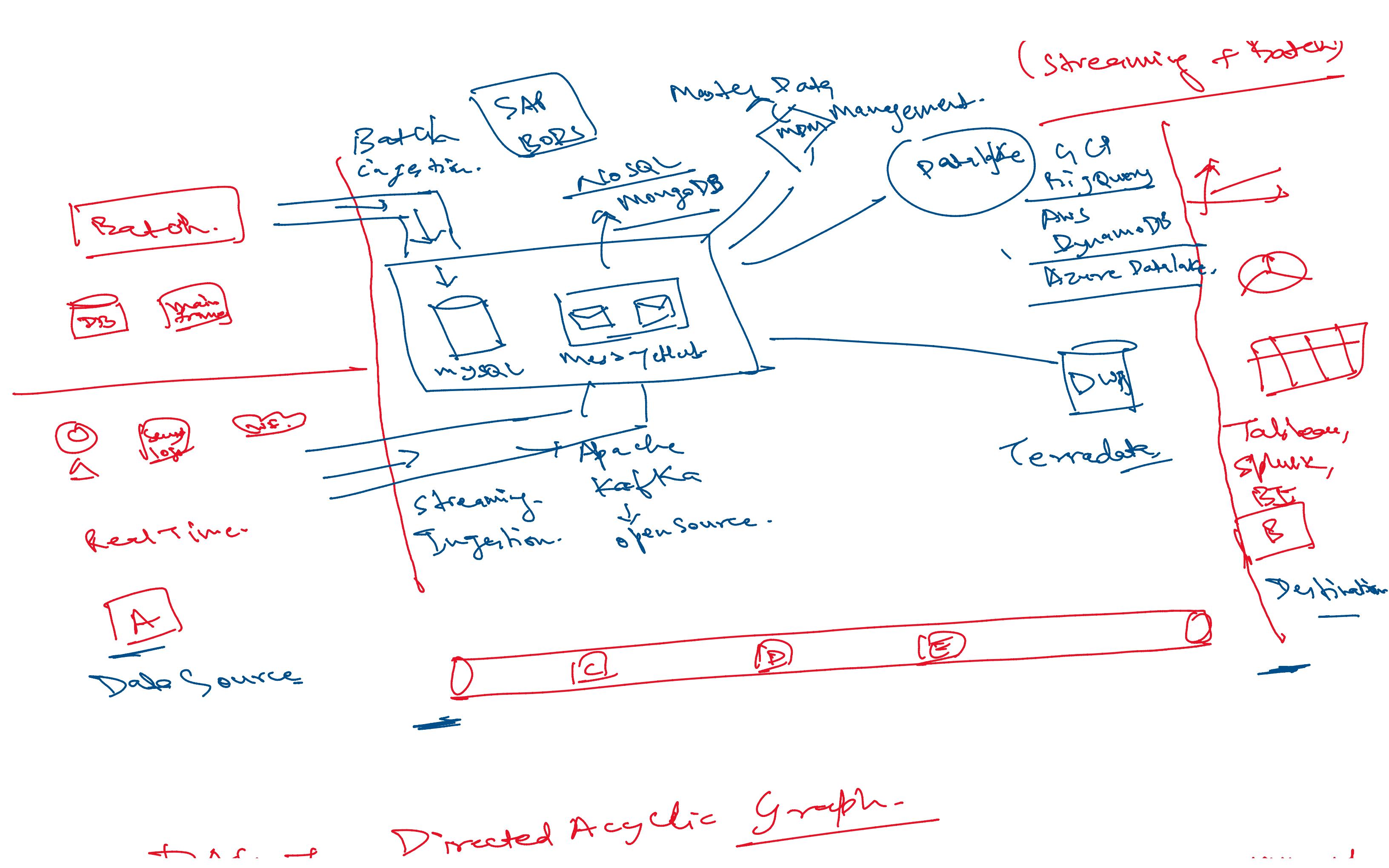
4. less Agile (rigid)

5. Operational Reporting

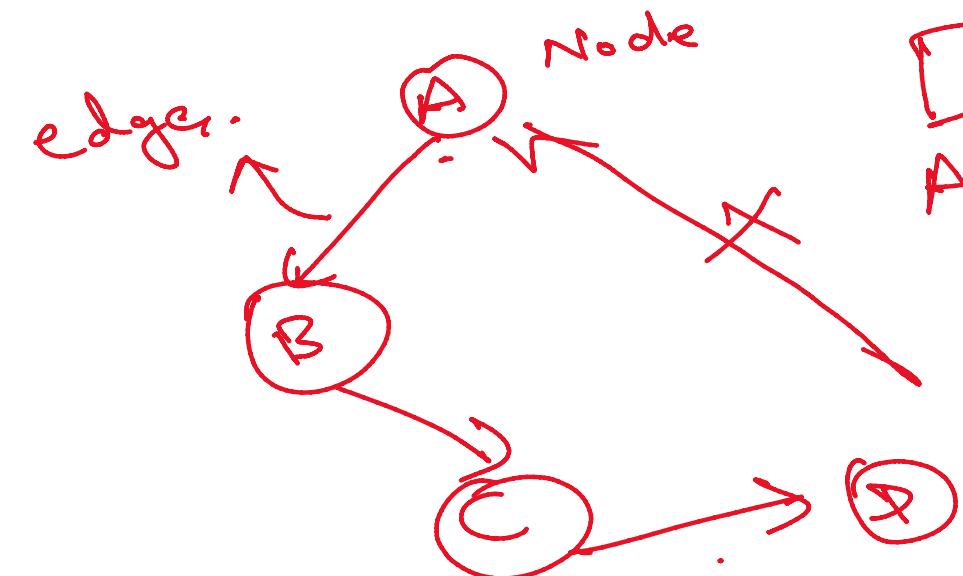
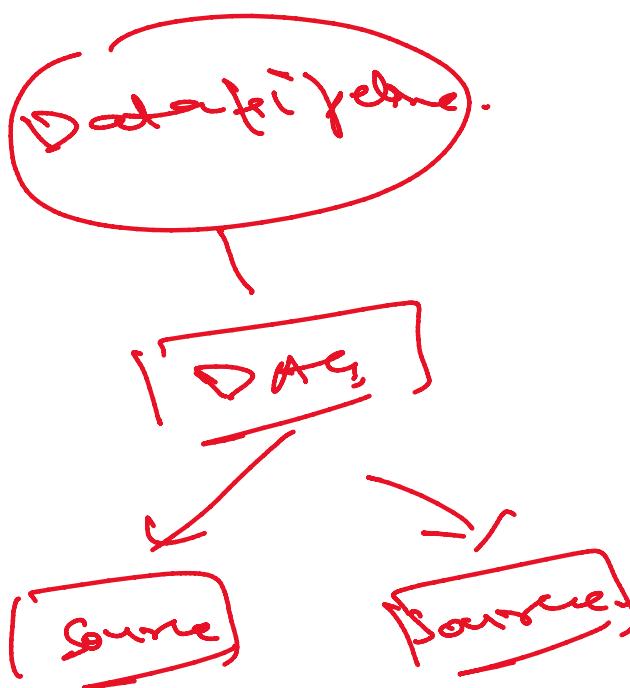
## 5. Advanced Analytics

Data Pipeline :-





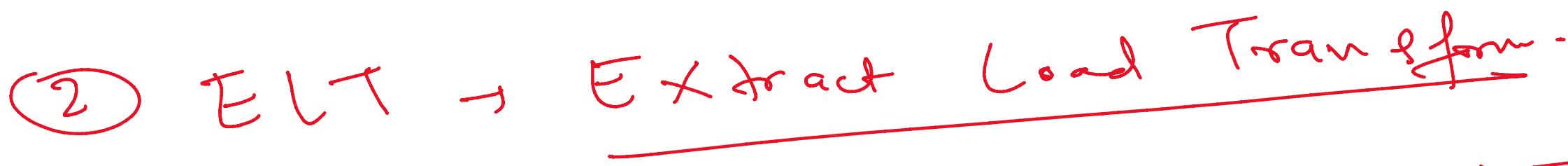
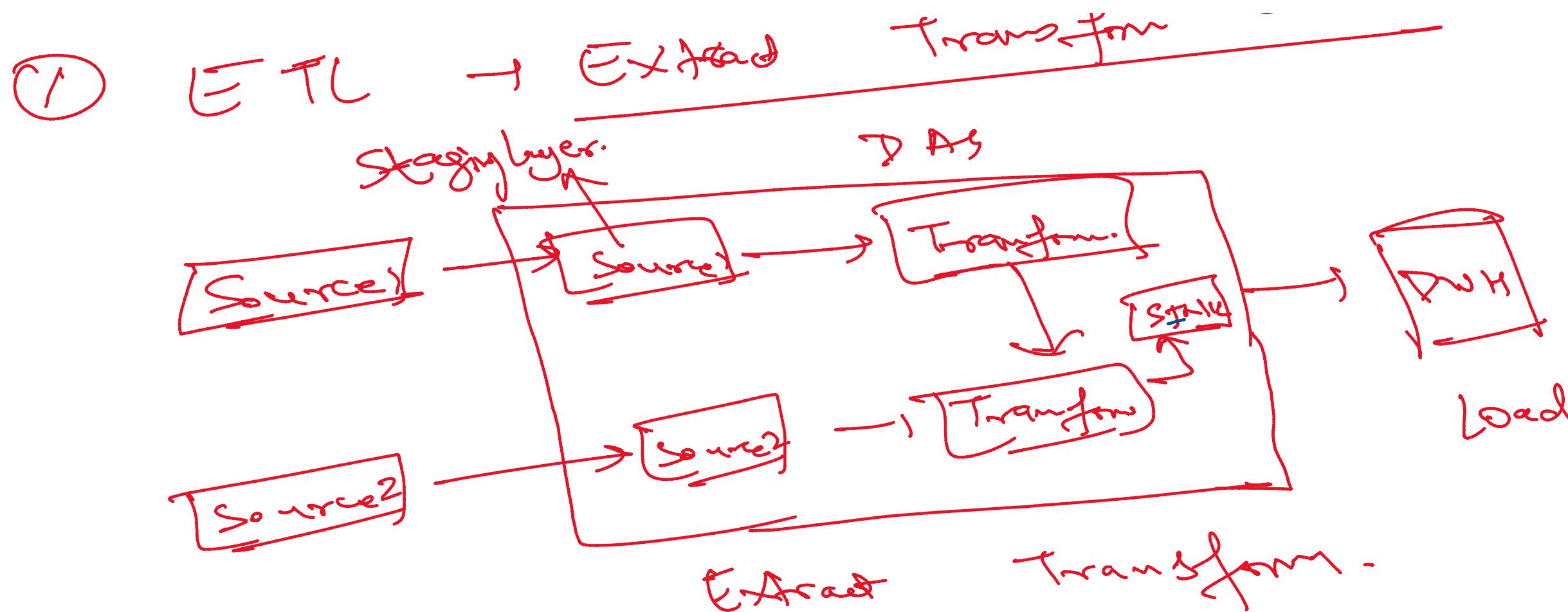
DAG  $\rightarrow$  Directed Acyclic ~~Graph~~  
 Upstream to Downstream Not other way around



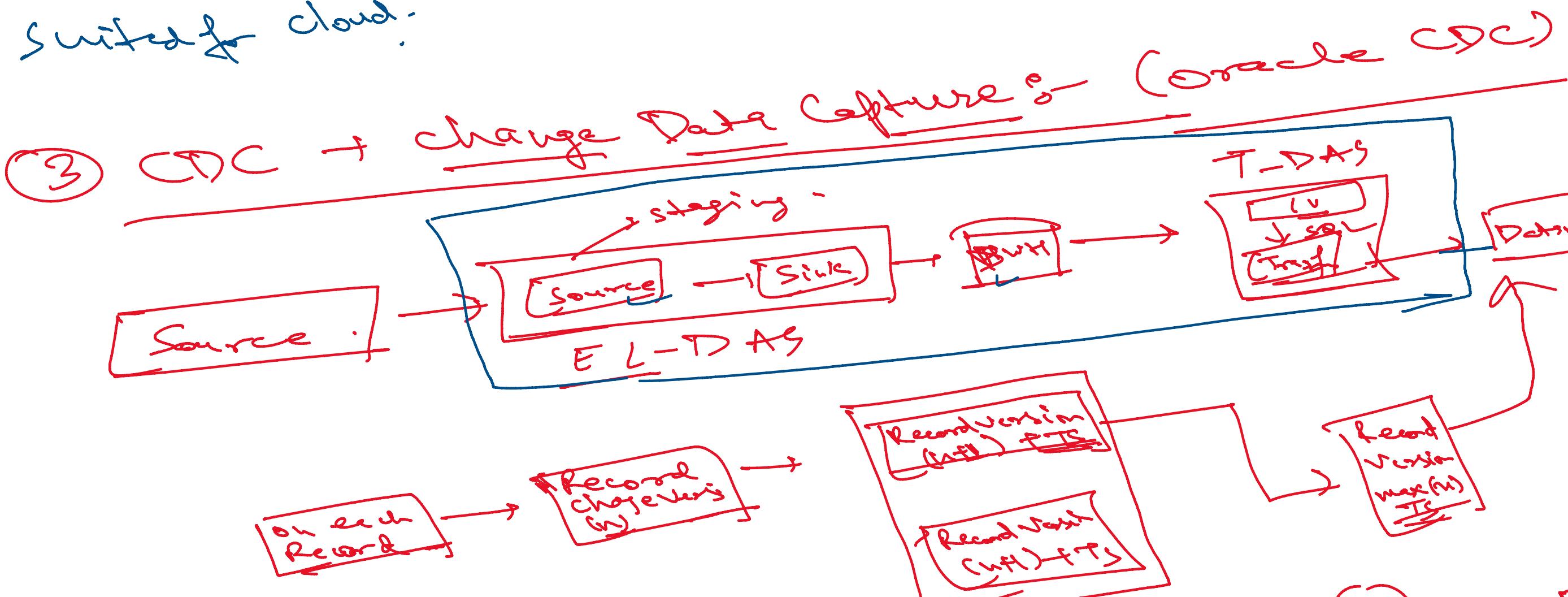
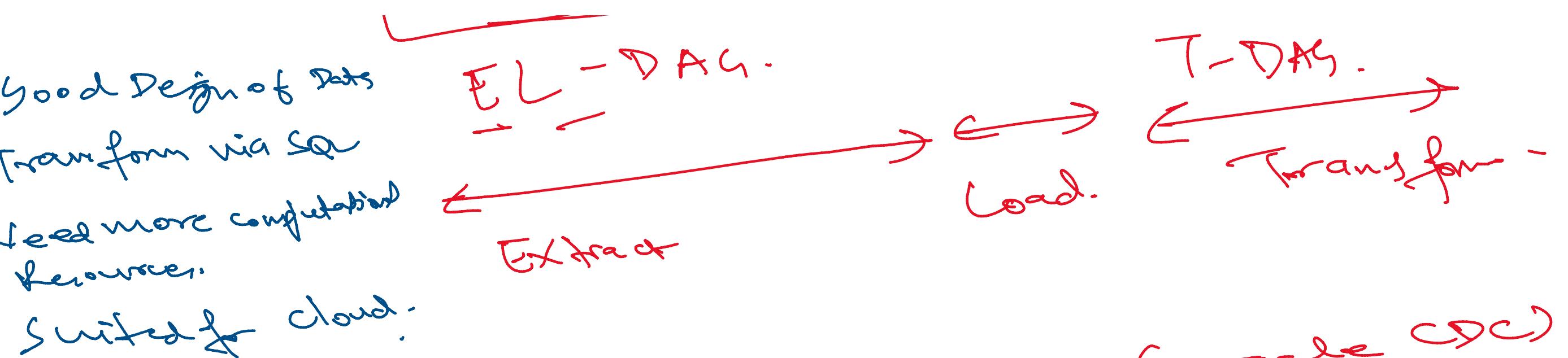
- ① ETL  $\rightarrow$  Extract Transform Load
- ② ECT  $\rightarrow$  Extract Load Transform
- ③ CDC  $\rightarrow$  Change Data Capture
- ④ E+LT  $\rightarrow$  E+LT Sub-pattern

→ → Extract

Transform Load



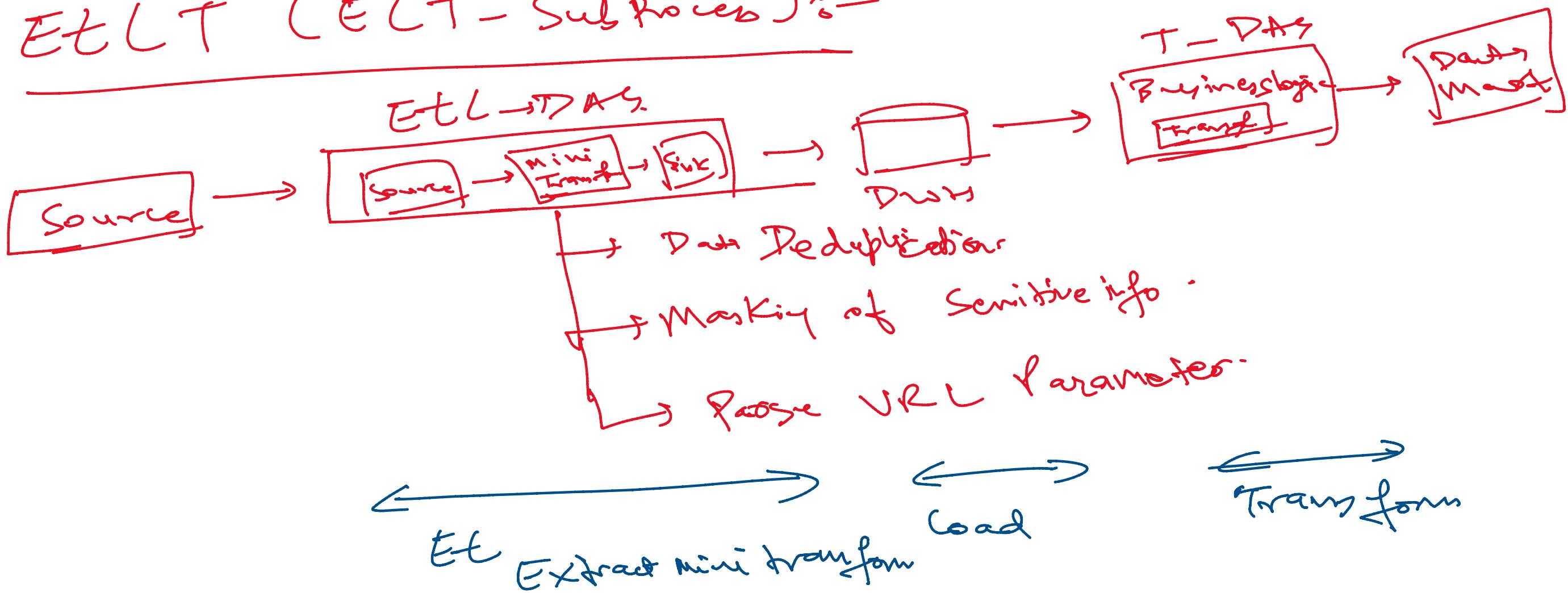
- ① Good Design of Data
  - ② Transform via SQL
  - ③ Need more computational resources.
  - ④ Suited for cloud



- ① Track the changes at the source -
  - ② Get the latest record -

- ② Get the latest record
- ③ Based on ELT Pattern
- ④ Read log from DB logs
- ⑤ Octool → Oracle ~~ORACLE~~

#### 4. ETLT (ELT - Sub Process) :-



① Disconnect from Business logic

② Does "mini transform" at early stage of pipeline

(L-) your migration  
could help in latency.