Agenda →

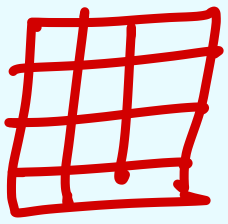
1. what is data Engg ?
2. In AI world → why DE is Important ? ✓ !
3. what does DE actually do ?
4. Skills ?
5. **SQL** → ?

1. what is DE?

→

1. **Structured ?**

↓

Table →  ✓

→ Purchase ?
?

2. Semi structured ← PDF ,

Email , Hashtag

id	doc loco
1	. ‿ ‿

(JSON
event log)

3. unstructured ← Review

Image

→ Img → Text

Text

→ videos → voice

→ Reviews

→ **Extract** ✓

- → **website** ✓
- → **API** → Paytm
  - → GPAY
  - →
- → **files**
  - →

→ **Store**

?

Data warehouse   Database

↗ Data lake ?

→ **Transforming** → clean

→ **Delivering for analysis**

Machine learni

Model GenAI

Data ?

(web data ?)

Joota ?

230

Purchased

(shoe)

Data !

(internaly)

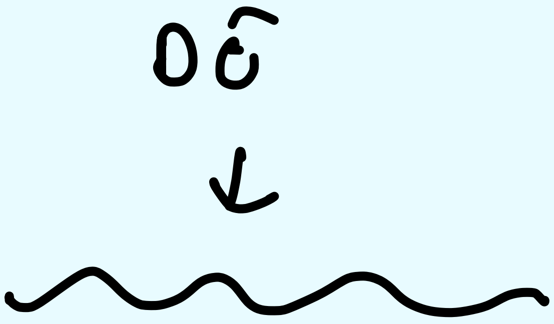
✓ ✓

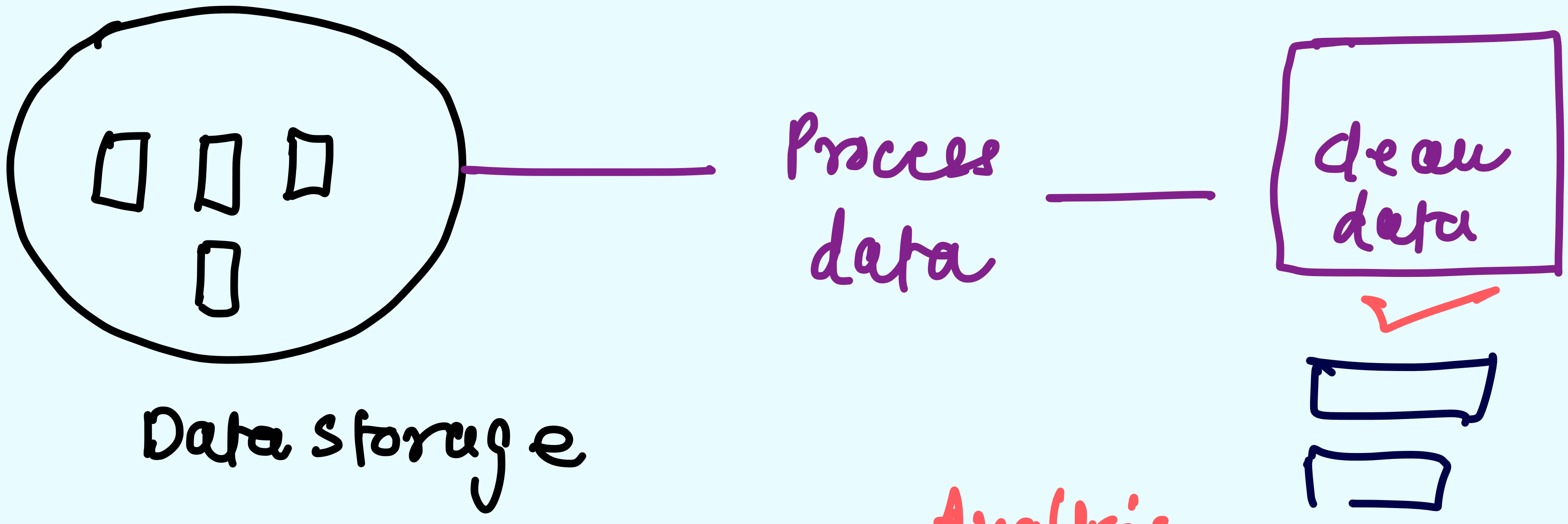
(DÊ)

R-II

↳ Data ?

|| ✓





Data → structured → (Analysis) Data Analyst ✓

→ (structured + unstructured) → DS ✓

Icecream

DS equation

+ = ML

New data

✓

(ML + (Decision Algo) = AI ✓

→ Skills

→ Python - scripting

→ SQL → query (Tables)

→ Java or Scala (Big data)

—x— spark

2) Data tools

→ ETL → Apache Airflow / ADF

(Extract | Transform | Load]

→ Hadoop | spark

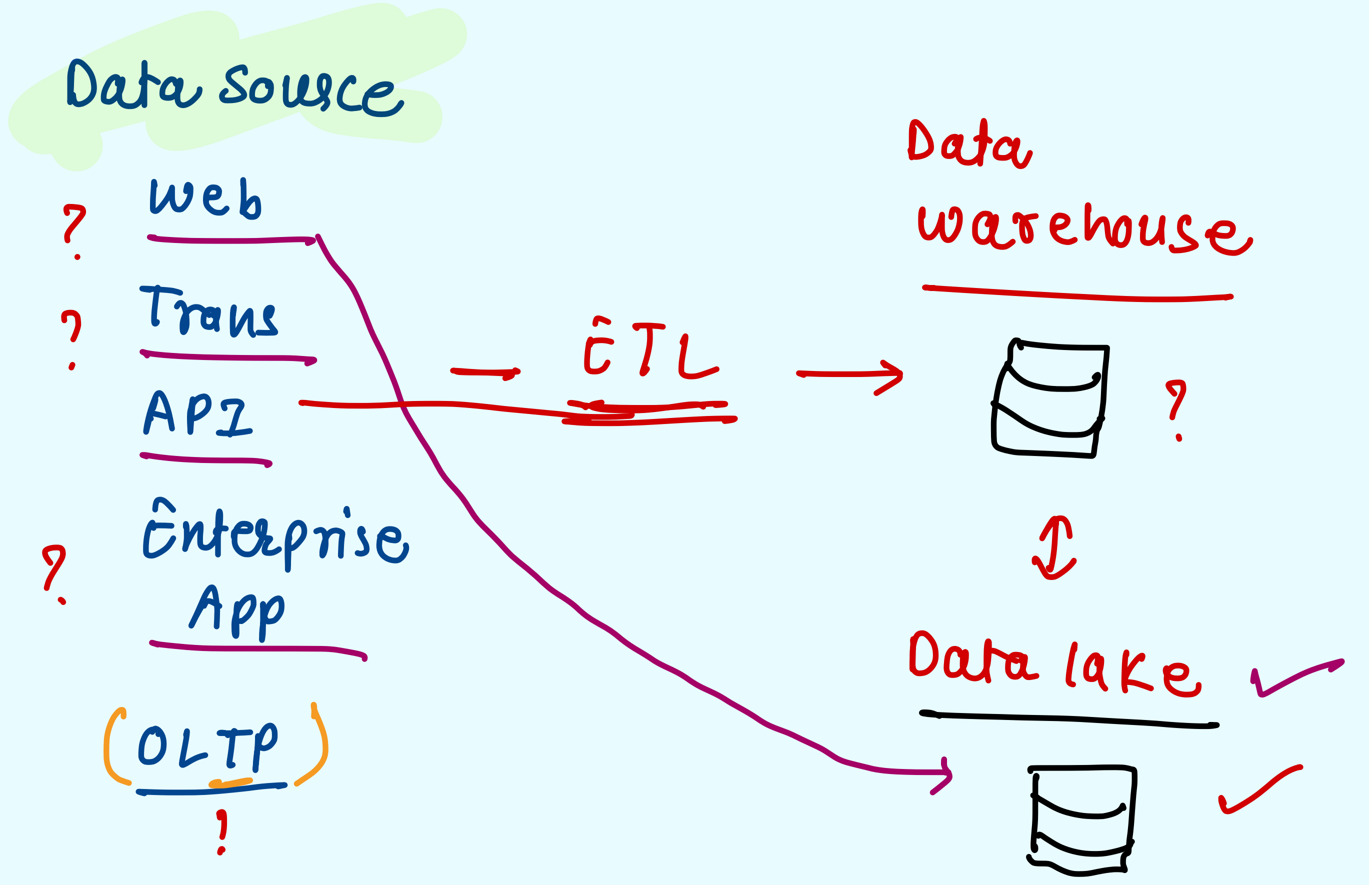
→ AWS / GCP / Azure (Cloud)

Database | storage

- structured — SQL
- semi structure → NoSQL

Myntra

Data Arch



online Transcation processing

OLTP → Database ?

→ (OLTP)

Pos ? →

Row	custid	order	order item	unit	Price
	1	c1	01	2	100

Myntra →

Inv

item-id	in stock	datetime
1	10	dt t1
1	8	dt t2

water

Data lake

water ? ✓

lake → ETL → water tank

Pay

Image

(GCP/Azure)

storage free

Lake → Treatment Plant — w — water tank — w — overhead tank w → Tap

✓

Data lake — ETL — Data warehouse — (Database) (OLAP) — Eng

↑ (r)

✓

online Analytical Process

lake

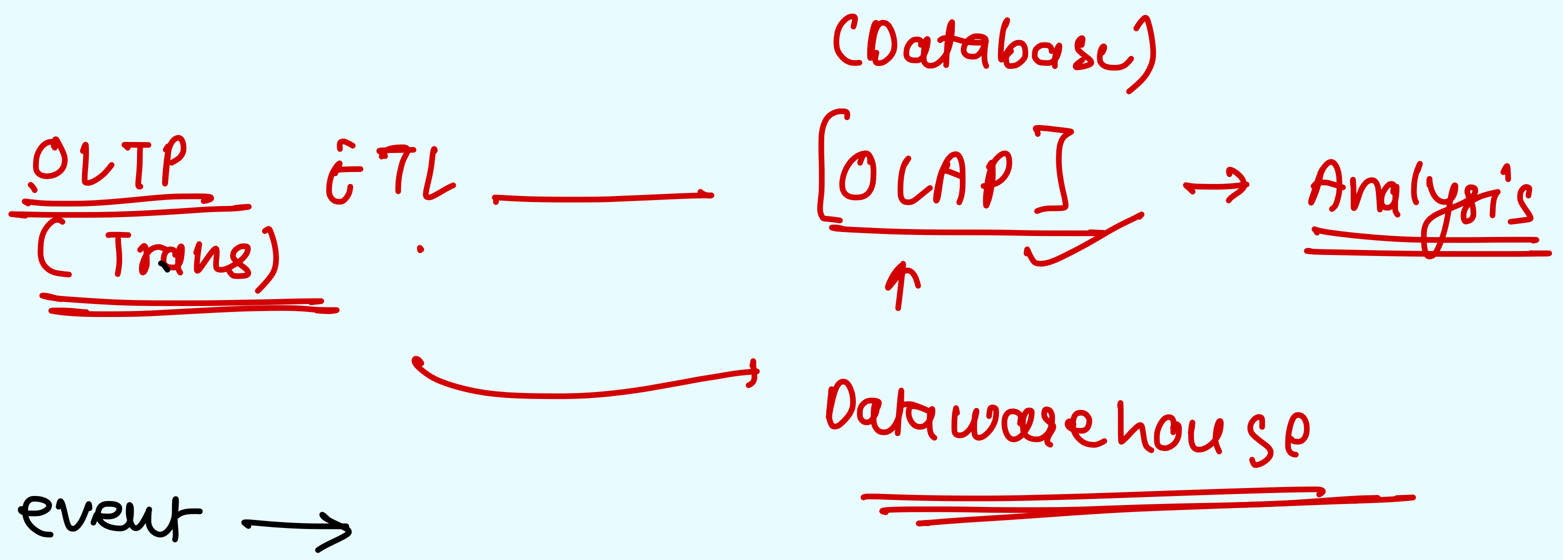
↑

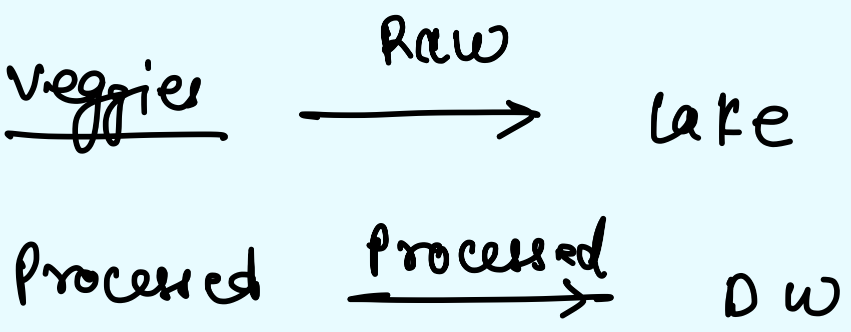
veg

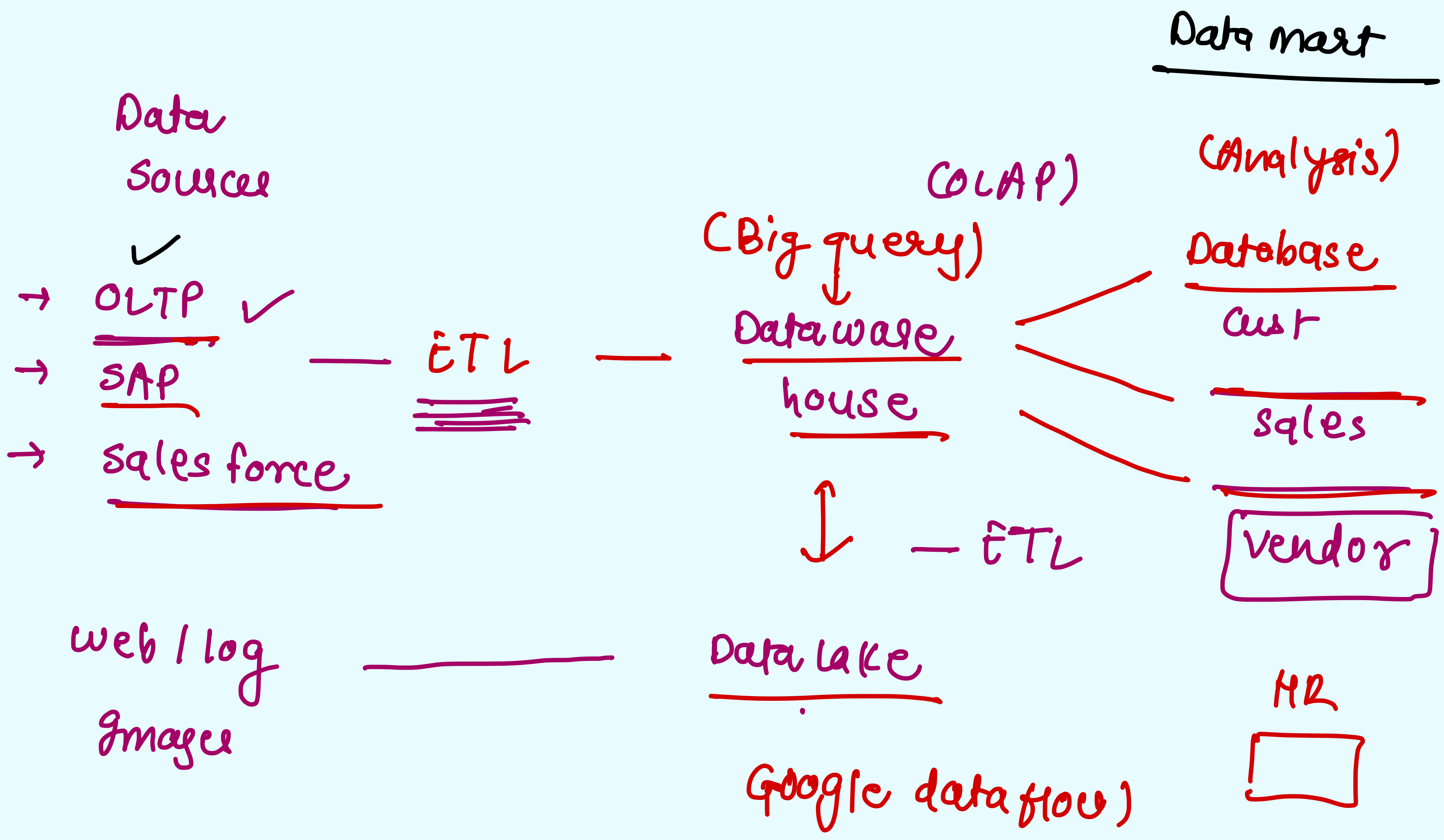
Data warehouse

↑

(Processed)







Database — OLAP (DW) Not Realtime / strategy

— OLTP → Realtime

OLTP (DB) ✓

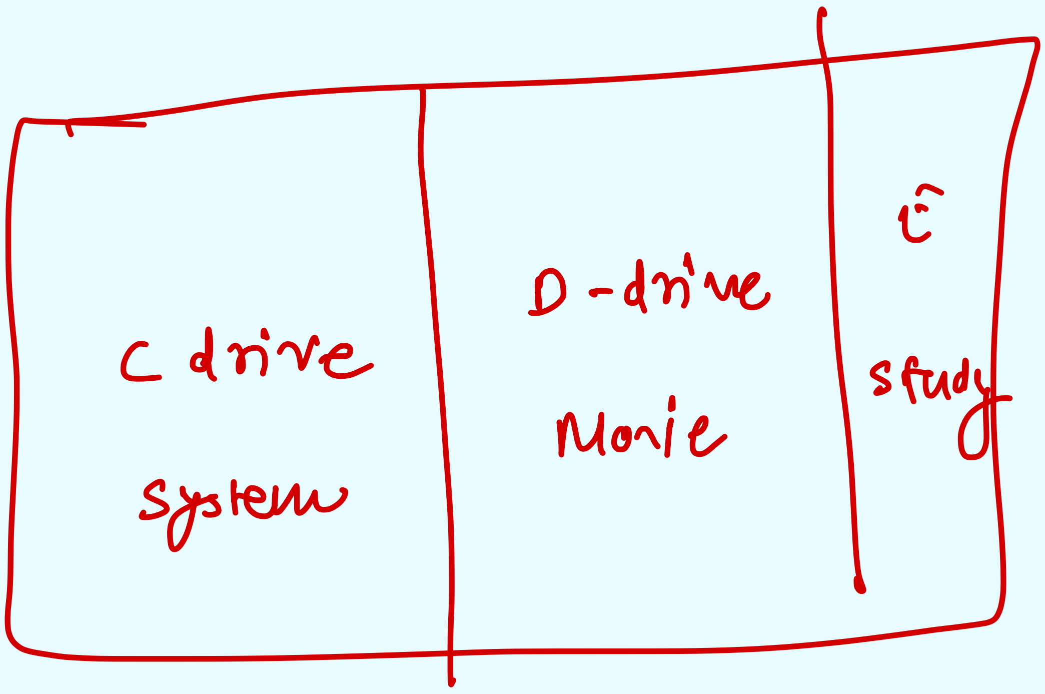
Order system

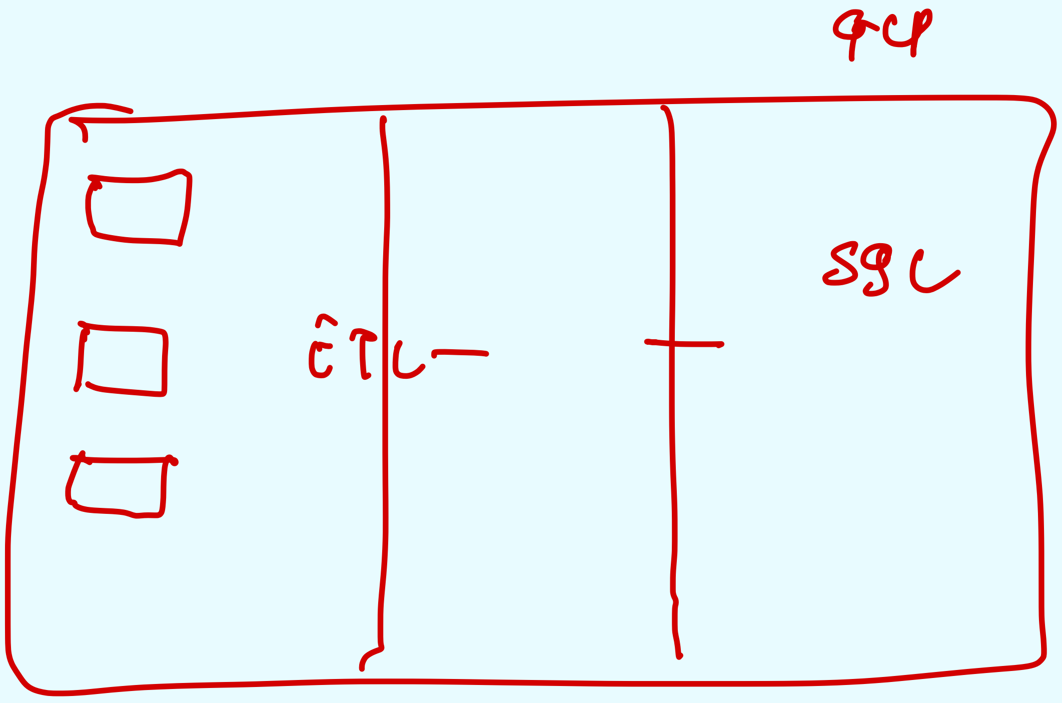
POS

OM

OLAP (DW) ✓ ✓

DB vs D





DB ✓

→ Real time

→ detailed Trans

→ Limited

→ Change freq

DW

Mon order

Trans per day

Many

DC

Raw

Orginal

DW

Processed

Agg

✓

Costlier

?

processed

DW

E

unstructured

id ✓

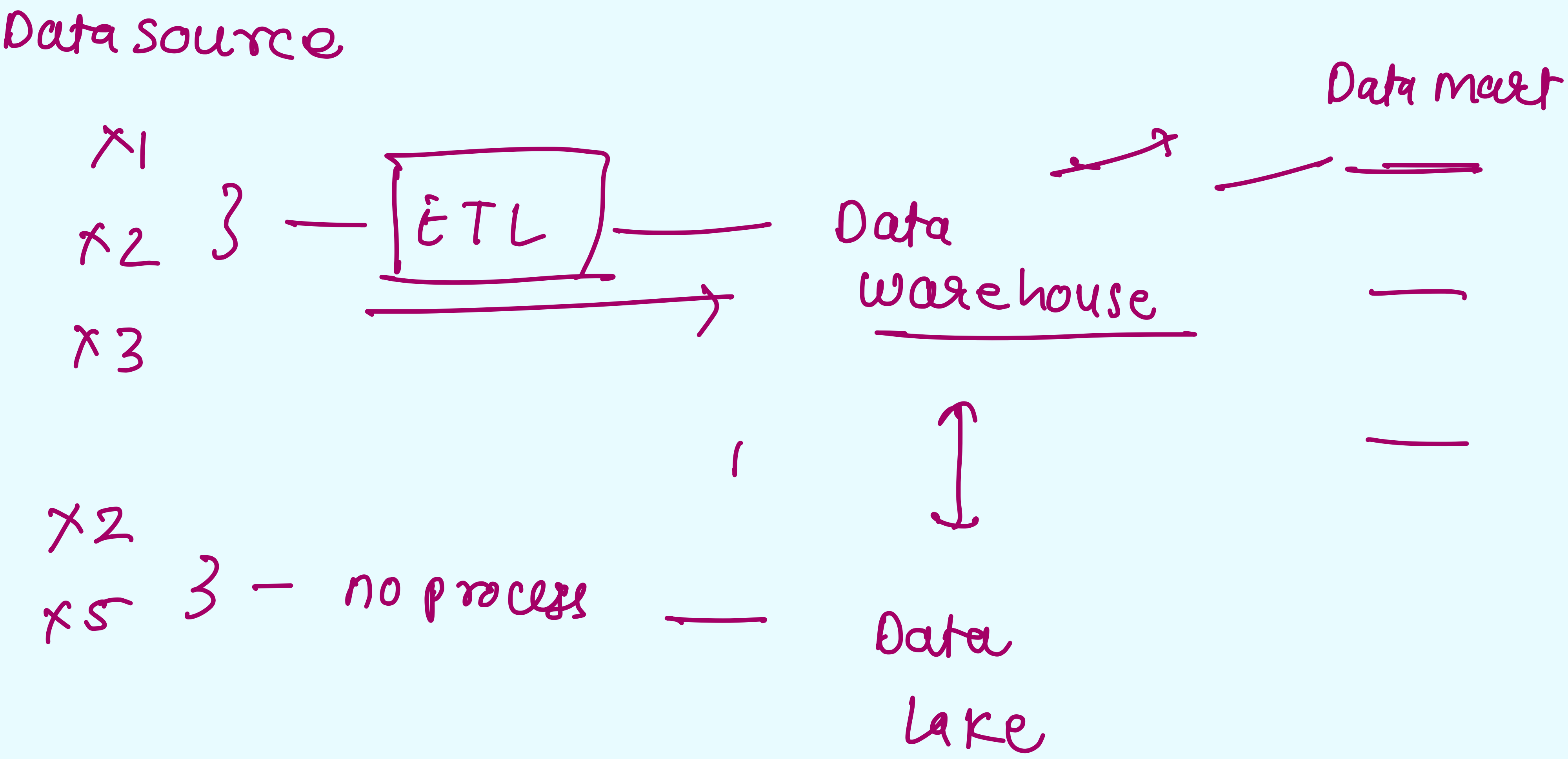
image

Type

location

↑

Not processed → Data lake ✓



Extraction → Hadoop Ecosystem