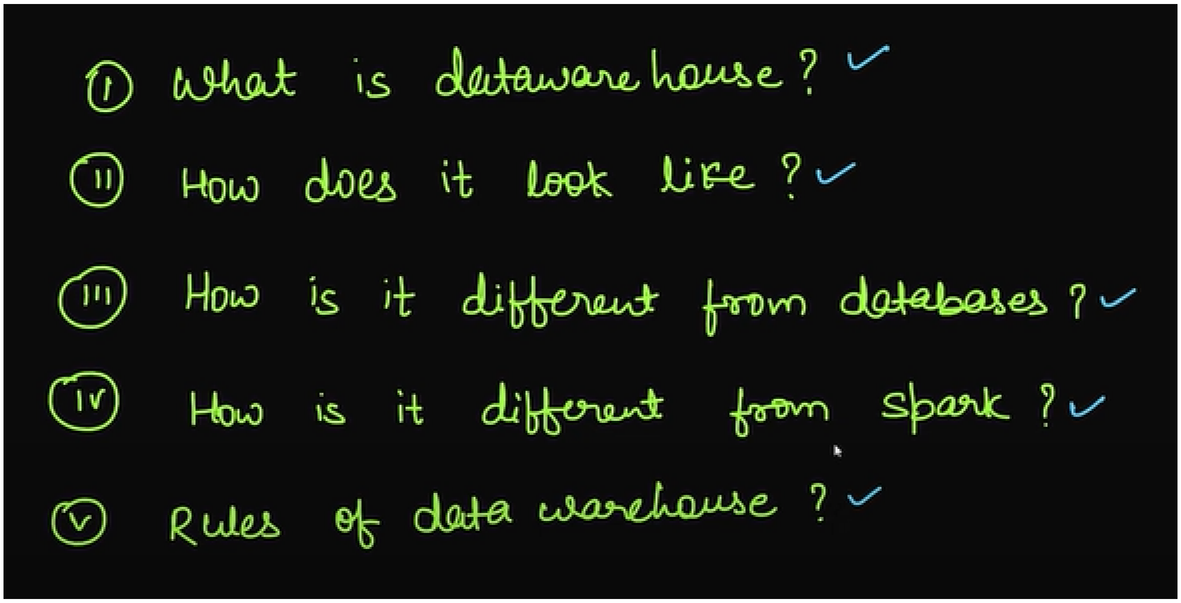
Data Warehouse

1. We’ll learn this topics today



(I) What is datawarehouse? ✓

(II) How does it look like? ✓

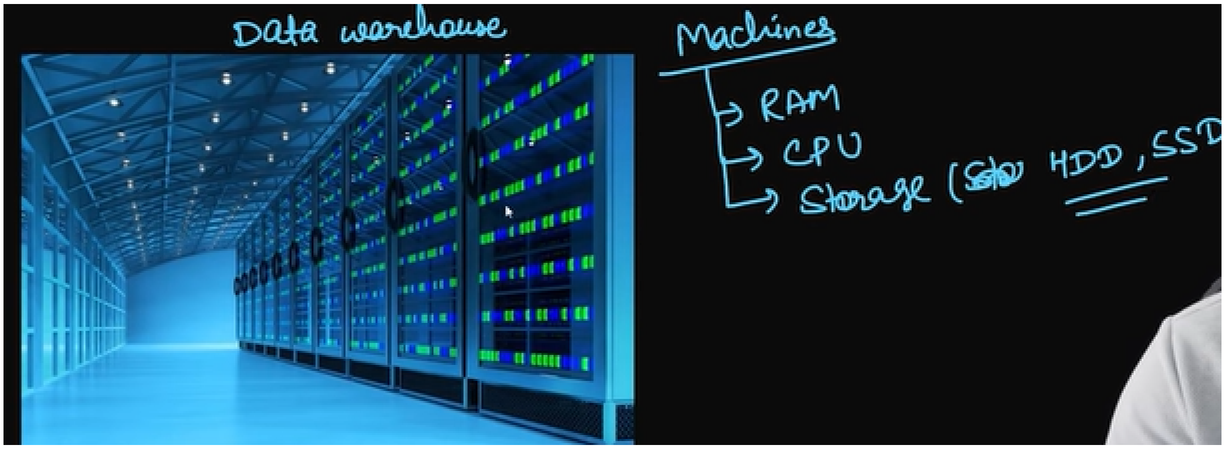
(III) How is it different from databases? ✓

(IV) How is it different from spark? ✓

(V) Rules of data warehouse? ✓

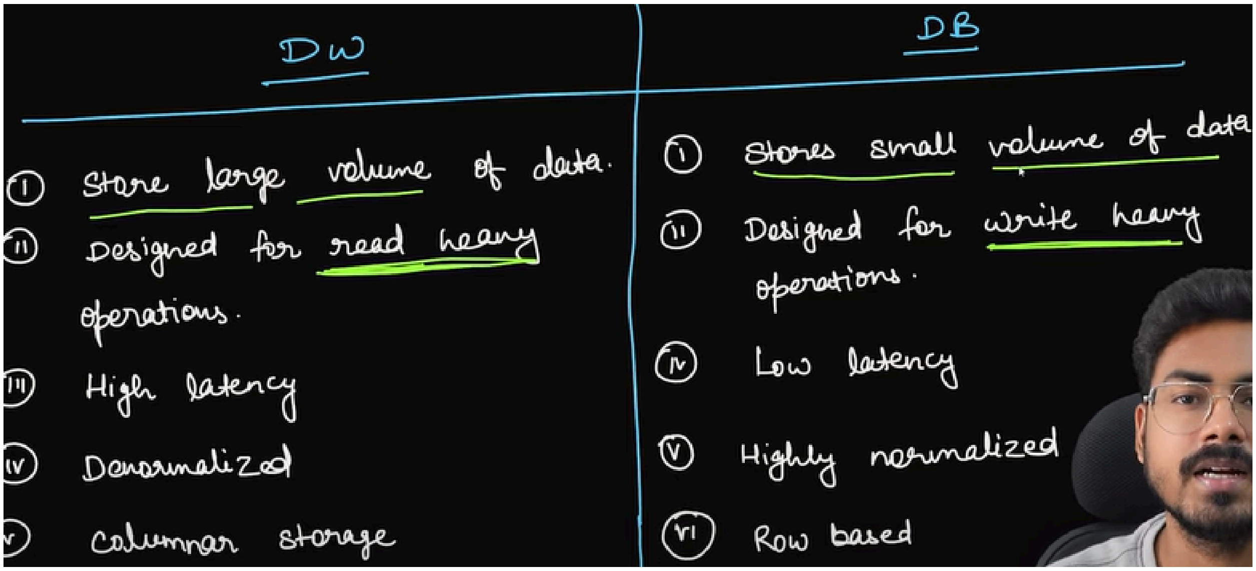
2. Here warehouse is place where we store anything(Goods,books etc)

3. So now in the place of goods..if we store data…then it is a data warehouse



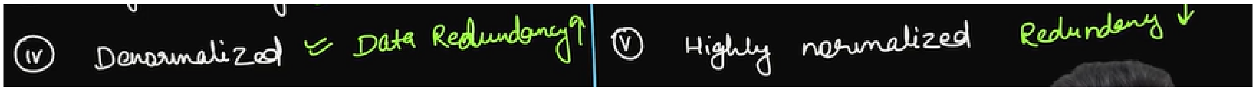
4. Here data will be inside the machines..
5. Each machine may contains number of databases.tables etc

6. Difference bw data warehouse and db



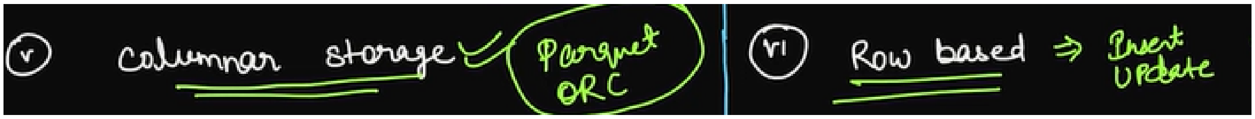
DW	DB
(i) Store large volume of data.	(i) Stores small volume of data
(ii) Designed for read heavy operations.	(ii) Designed for write heavy operations.
(iii) High latency	(iv) Low latency
(iv) Denormalized	(v) Highly normalized
(v) Columnar storage	(vi) Row based

7. DW is designed for heavy read like “select *”, and DB is made for insertions,updates etc which is write heavy



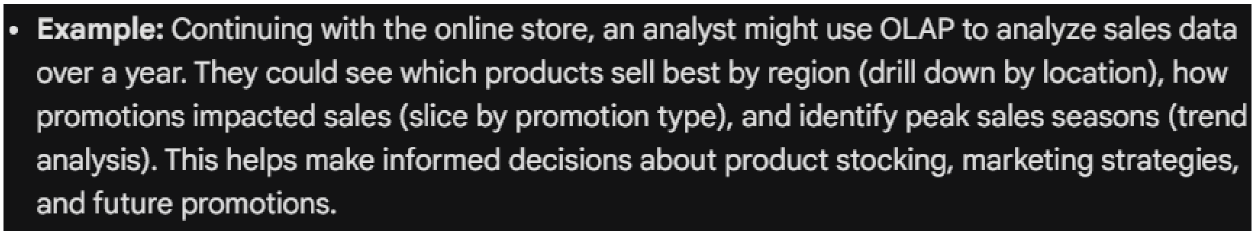
(iv) Denormalized ⇐ Data Redundancy↑ | (v) Highly normalized Redundancy ↓

8. In DW we use columnar storage to read the data quickly..In DB we use row_based which makes insertion and updates faster



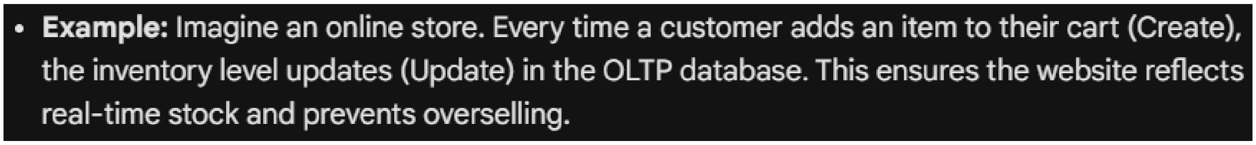
(v) Columnar storage ✓ Parquet ORC | (vi) Row based ⇒ Insert Update

9. DW is for OLAP(Teradata,ExaData,SnowFlake,Redshift) and DB is for OLTP(MySQL,Oracle,PostGres)
10. OLAP example



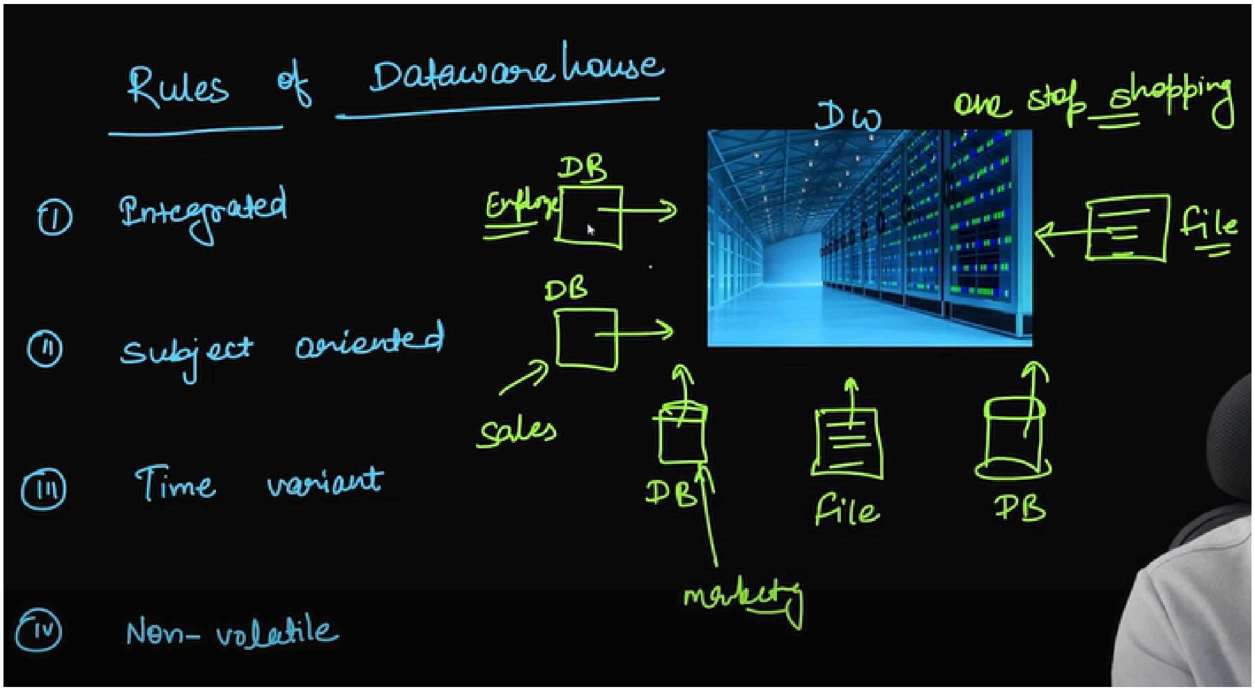
- **Example:** Continuing with the online store, an analyst might use OLAP to analyze sales data over a year. They could see which products sell best by region (drill down by location), how promotions impacted sales (slice by promotion type), and identify peak sales seasons (trend analysis). This helps make informed decisions about product stocking, marketing strategies, and future promotions.

11. OLTP



- **Example:** Imagine an online store. Every time a customer adds an item to their cart (Create), the inventory level updates (Update) in the OLTP database. This ensures the website reflects real-time stock and prevents overselling.

12. Rules of Data Warehouse

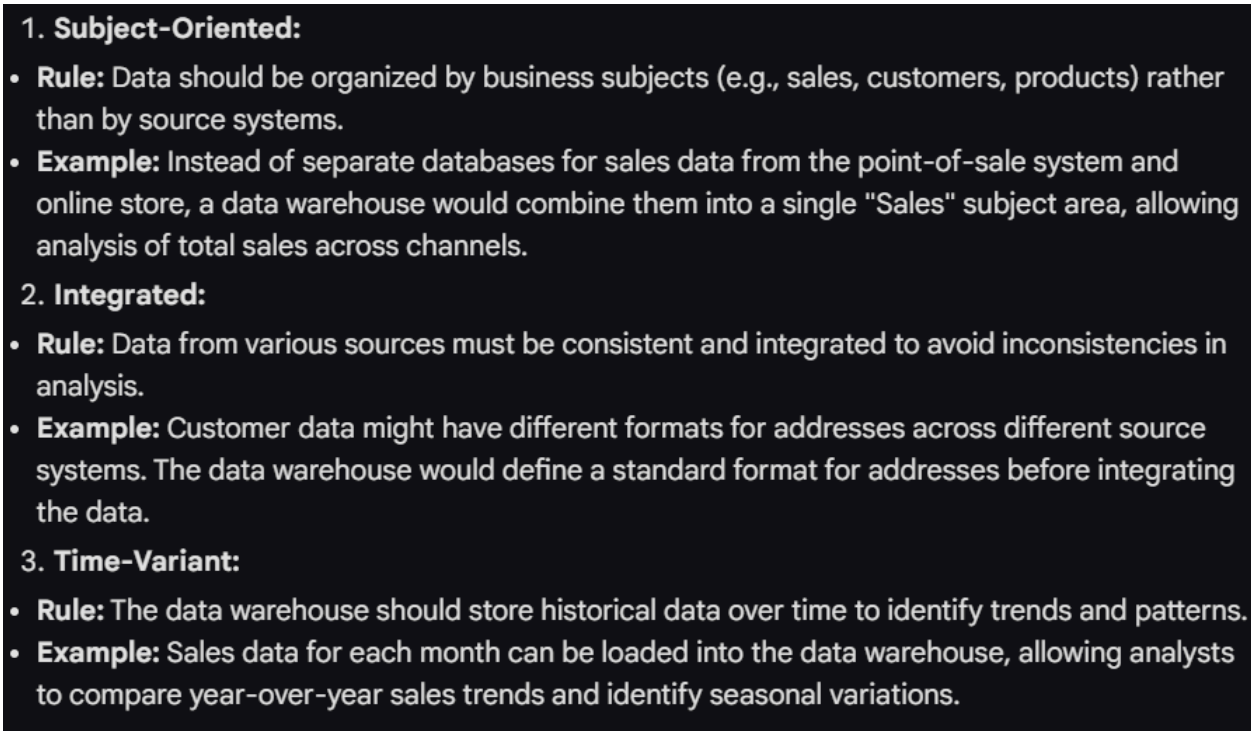


Rules of Datawarehouse

1. Integrated
2. Subject oriented
3. Time variant
4. Non- volatile

13. Here data will be integrated from multiple source(DB’s ,files)..and DW is a one stop shopping for our data
14. Now from our DW..we can analyse the emp_data, sales_data etc …so this how integration works
15. Subject Oriented

16. For example a business_user can work on how much has occured



1. **Subject-Oriented:**

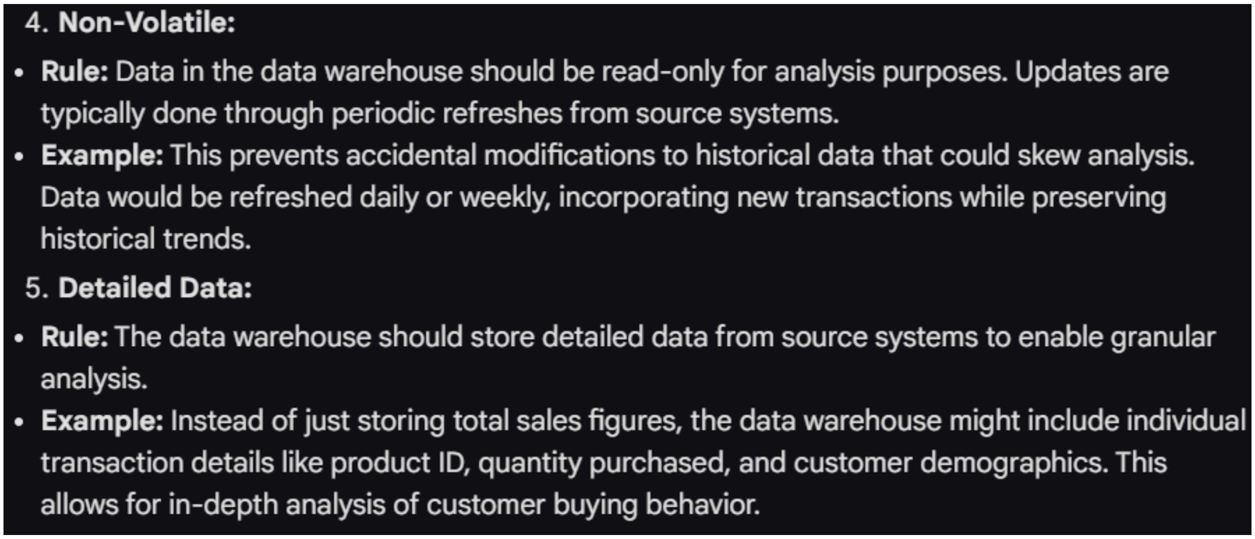
- **Rule:** Data should be organized by business subjects (e.g., sales, customers, products) rather than by source systems.
- **Example:** Instead of separate databases for sales data from the point-of-sale system and online store, a data warehouse would combine them into a single "Sales" subject area, allowing analysis of total sales across channels.

2. **Integrated:**

- **Rule:** Data from various sources must be consistent and integrated to avoid inconsistencies in analysis.
- **Example:** Customer data might have different formats for addresses across different source systems. The data warehouse would define a standard format for addresses before integrating the data.

3. **Time-Variant:**

- **Rule:** The data warehouse should store historical data over time to identify trends and patterns.
- **Example:** Sales data for each month can be loaded into the data warehouse, allowing analysts to compare year-over-year sales trends and identify seasonal variations.



4. **Non-Volatile:**

- **Rule:** Data in the data warehouse should be read-only for analysis purposes. Updates are typically done through periodic refreshes from source systems.
- **Example:** This prevents accidental modifications to historical data that could skew analysis. Data would be refreshed daily or weekly, incorporating new transactions while preserving historical trends.

5. **Detailed Data:**

- **Rule:** The data warehouse should store detailed data from source systems to enable granular analysis.
- **Example:** Instead of just storing total sales figures, the data warehouse might include individual transaction details like product ID, quantity purchased, and customer demographics. This allows for in-depth analysis of customer buying behavior.

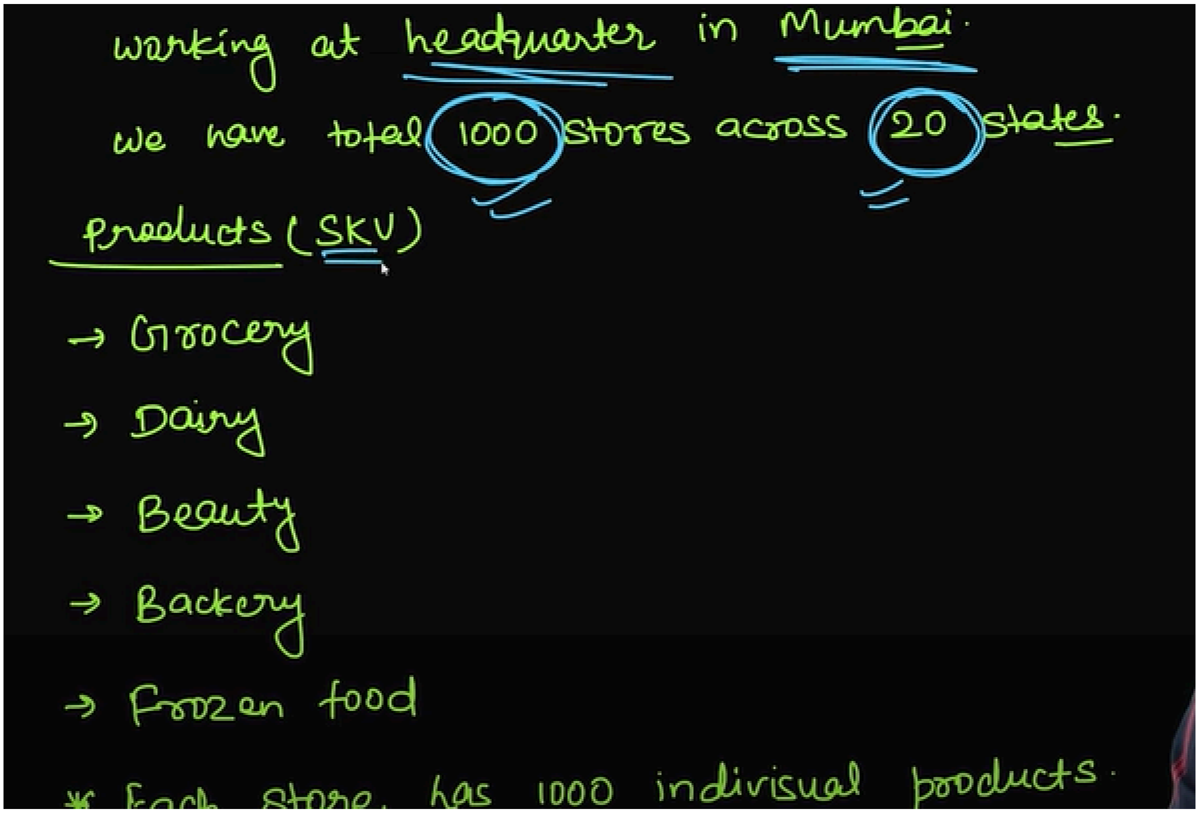
Why do we need a Data Warehouse?

1. Lets take a business use case of retail industry

Business Use case (Retail Industry)

Reliance Mart

2. Here we are working at



working at headquarter in Mumbai.

we have total 1000 stores across 20 states.

products (SKU)

→ Grocery

→ Dairy

→ Beauty

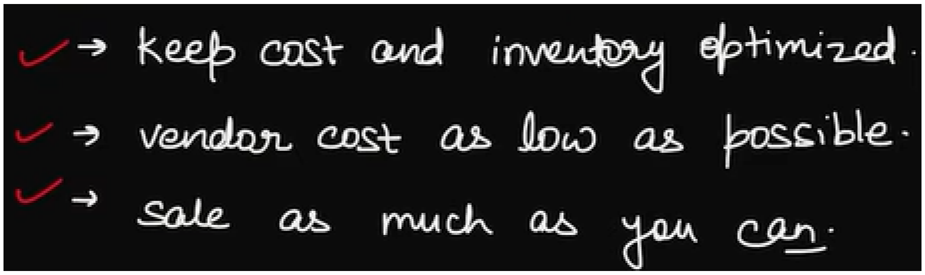
→ Backery

→ Frozen food

* Each store has 1000 indivisual products.

3. Here we want to maximize the profit

4. To increase profit..we have to implement this

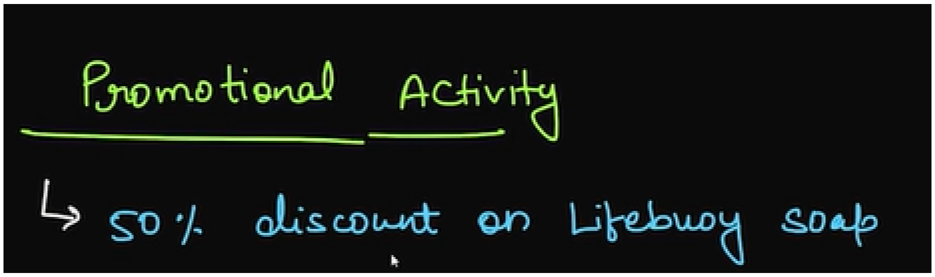


✓ → keep cost and inventory optimized.

✓ → vendor cost as low as possible.

✓ → sale as much as you can.

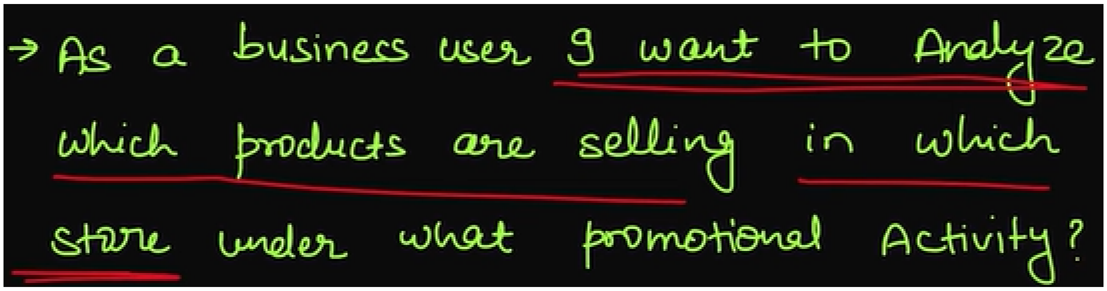
5. To increase sales…we have some promotional activity



Promotional Activity

↳ 50% discount on Lifebuoy soap

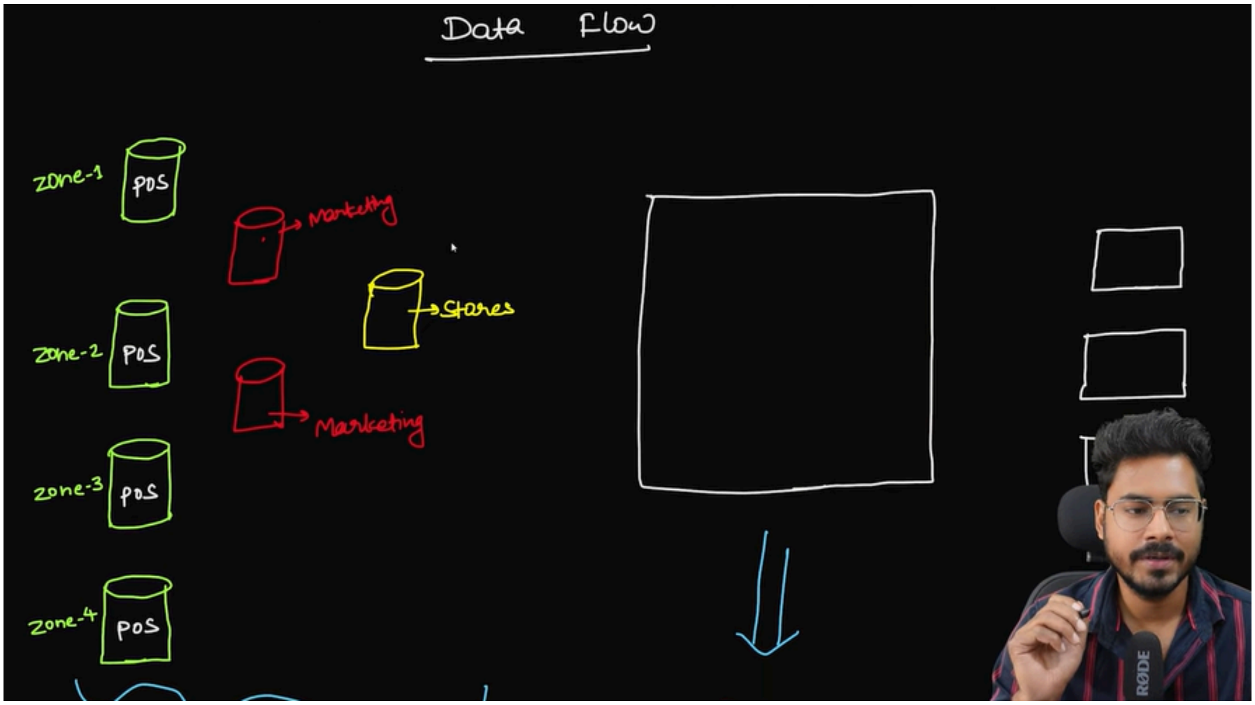
6. Now I need to find solution to this problem



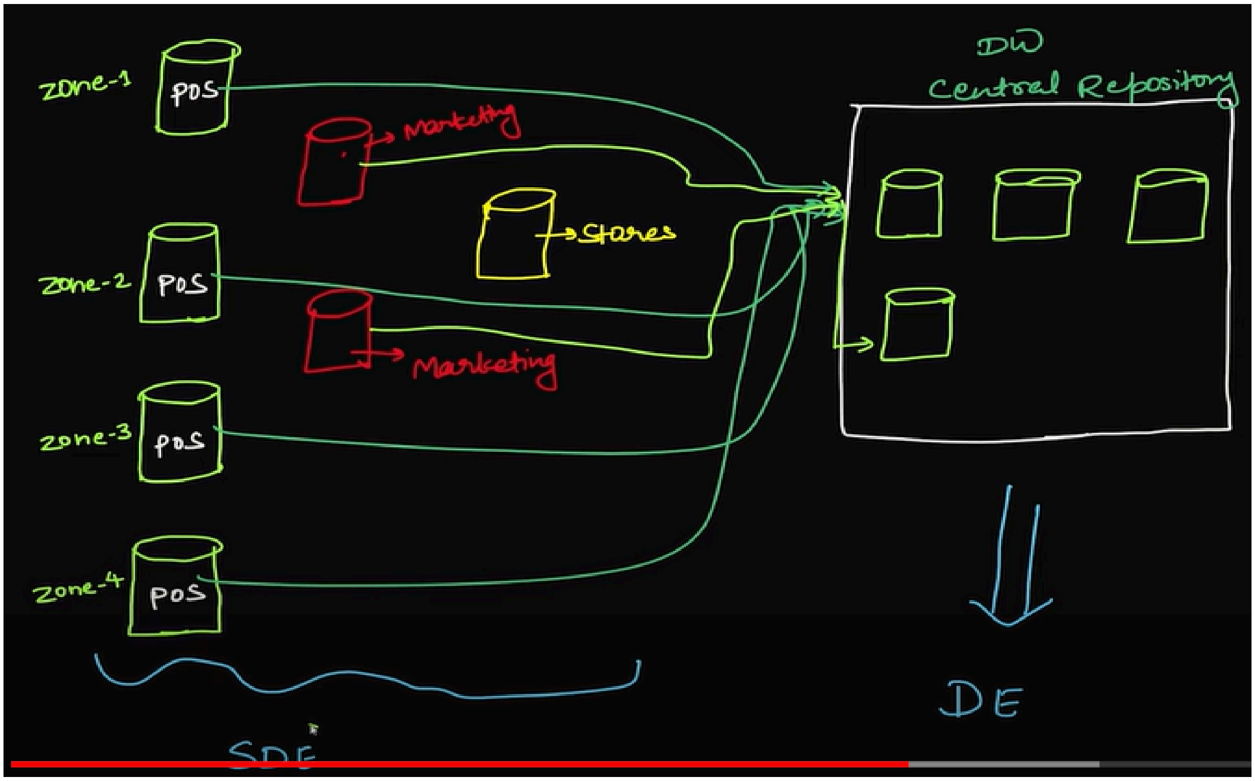
→ As a business user I want to Analyze which products are selling in which store under what promotional Activity?

7. Here we'll solve this in two ways..first we'll solve this in noob way

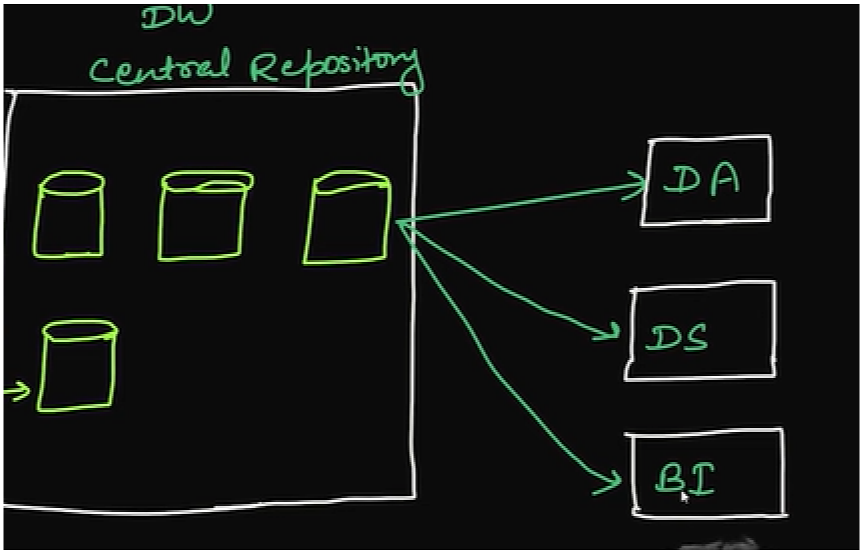
8. Here we have 4 zones..and each has POS(point of sales.. billing)..and we have 2 marketing database…and Stores database



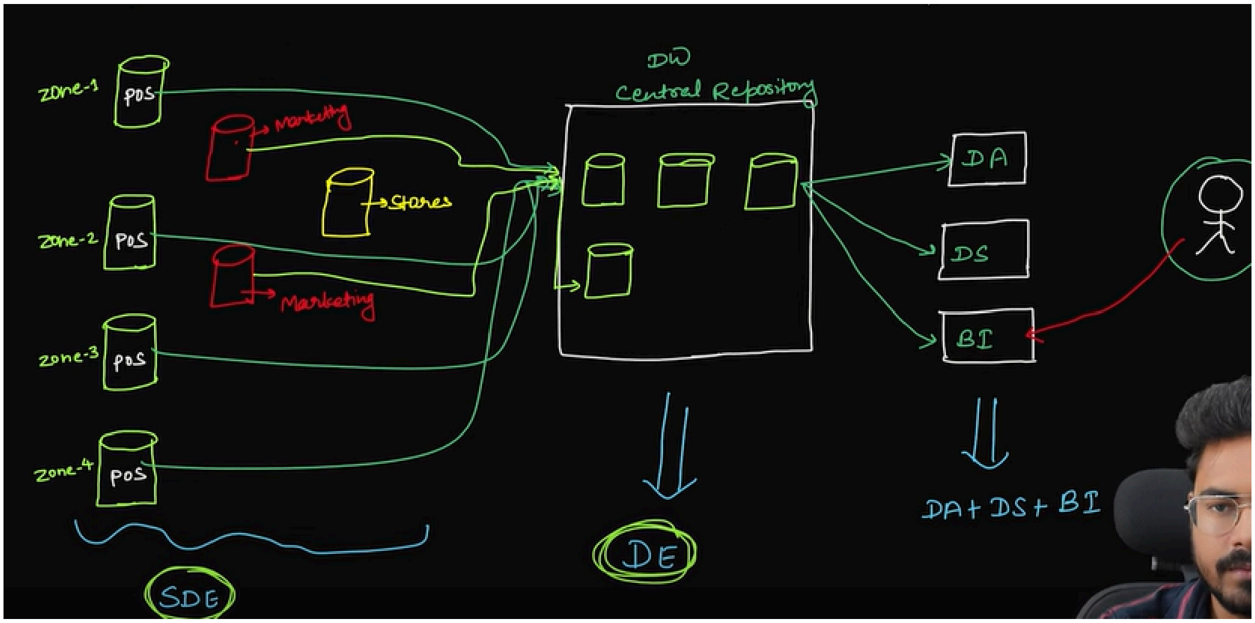
9. SO if there's no DW..then the flow of data will look like this



10. Here we have a central repository..which contains multiple databases..and our data of zones,marketing and stores..has come to central repository
11. A data Engineer will pull the data from this sources and brings to a central repository

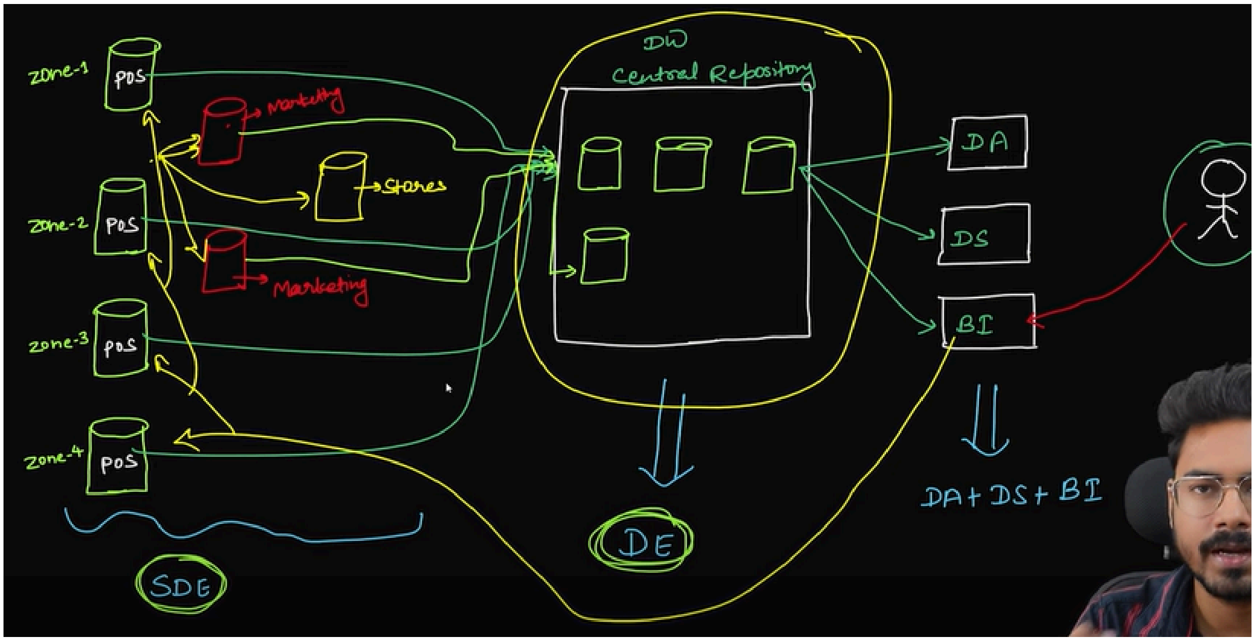
12. 

Now from here DA's,DS's,BI's can pull this data for their own purposes
13. Now we'll focus on our main question of promotional activity

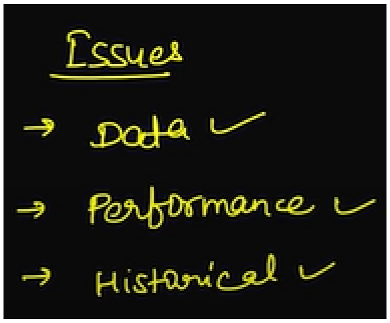


14. Now if there's no DE team..then we'll not be having central repo of data

15. SO if there’s no DW..then BI analyst must go to the database of each zone abd join them with stores to get the required data…so its a lot of work and computation involved



16. Another issues



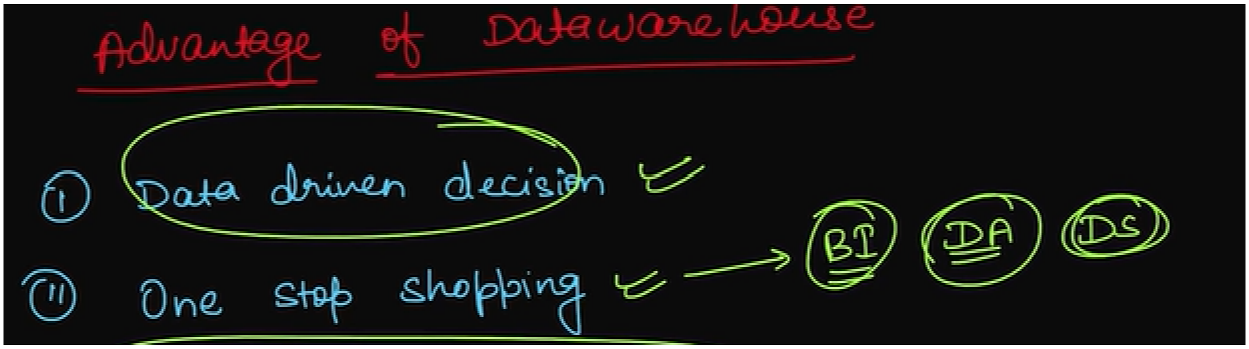
Issues

→ Data ✓

→ Performance ✓

→ Historical ✓

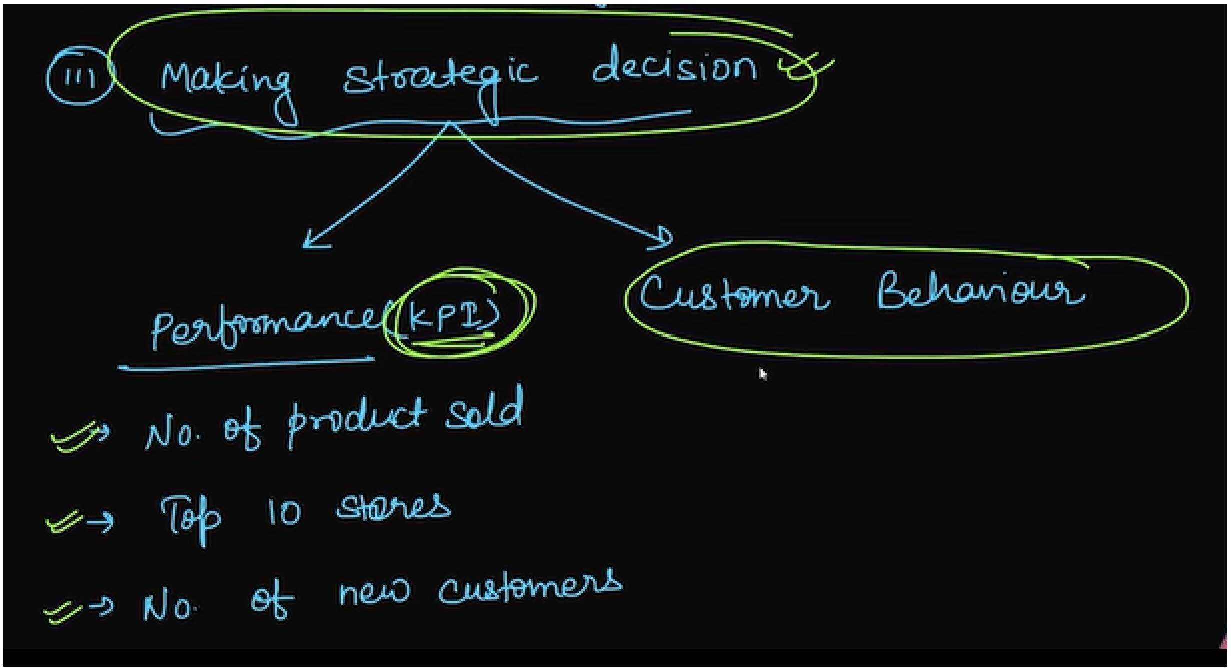
17. Now what if there’s a DE team handling the data warehouse



Advantage of Datawarehouse

① Data driven decision ✓

ⅱ One stop shopping ✓ → BI DA DS



ⅲ Making Strategic decision ✓

Performance (KPI)

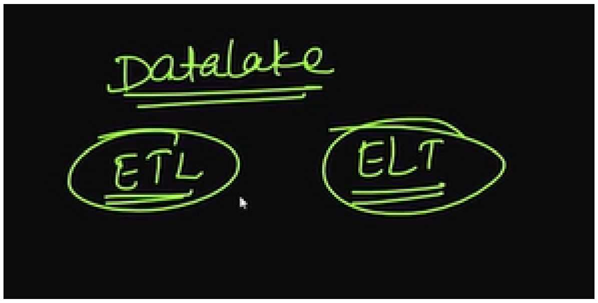
Customer Behaviour

✓→ No. of product sold

✓→ Top 10 stores

✓→ No. of new customers

18. So with the DW ..we can make quick decisions..which helps to stand out in the market

19. Next we’ll learn  Datalake >> DW

Datalake

ETL ELT