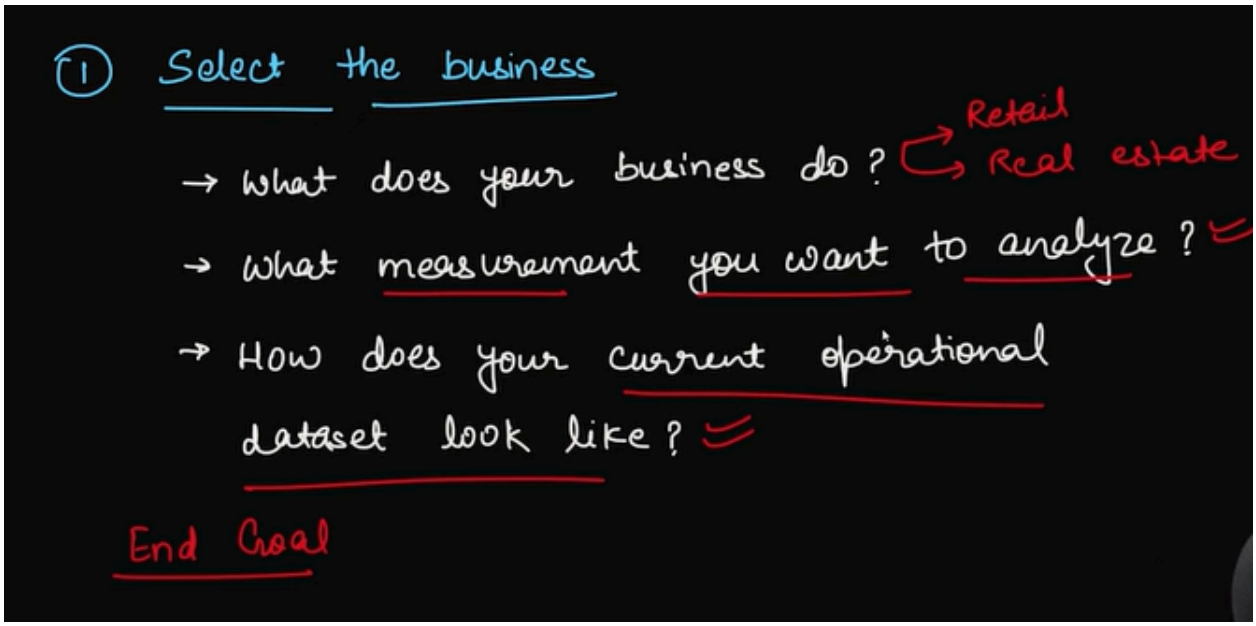
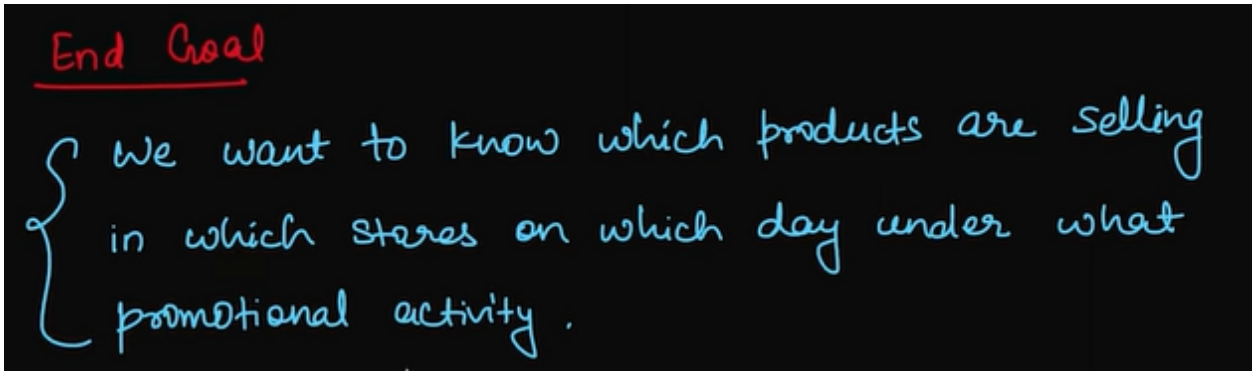
Dimension Modeling fundamentals

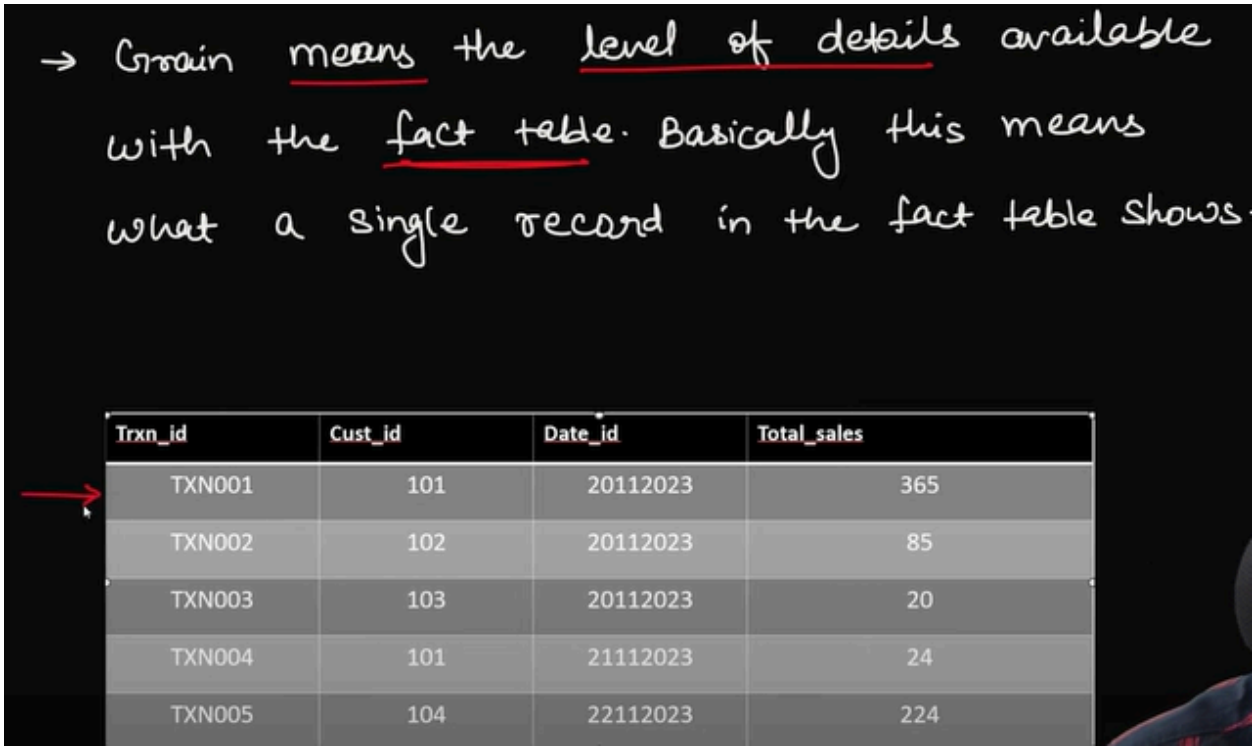
1. How to select fact tables and its columns
2. Lets see 4 steps of dimensional design process
3. First step we have to select the business type



4. So we must know the end goal of our business..based on that we can build table
5. Lets look at our end goal for building dim table



6. Next step is to declare the grain
7. Grain can be defined as levels of details in our data

8. 

Trxn_id	Cust_id	Date_id	Total_sales
TXN001	101	20112023	365
TXN002	102	20112023	85
TXN003	103	20112023	20
TXN004	101	21112023	24
TXN005	104	22112023	224

9. So now here..we can calculate total sales on nov20 - 470
10. But what if we want to calculate the product which sold the most? …now with this data we cannot find the answer
11. Now we'll increase our grain level access

Trxn_id	Cust_id	Product_id	Date_id	Total_sales
TXN001	101	PRD001	20112023	200
TXN002	101	PRD003	20112023	150
TXN003	101	PRD007	20112023	15
TXN004	102	PRD004	20112023	85
TXN005	103	PRD006	20112023	20
TXN006	101	PRD005	21112023	24
TXN007	104	PRD001	22112023	200
TXN008	104	PRD005	22112023	24

12. Here ..we have added product_id..

13. As we increase our grain level's..then the size of the data also increases

5 Record

Trxn_id	Cust_id	Date_id	Total_sales
TXN001	101	20112023	365
TXN002	102	20112023	85
TXN003	103	20112023	20
TXN004	101	21112023	24
TXN005	104	22112023	224

8 Record

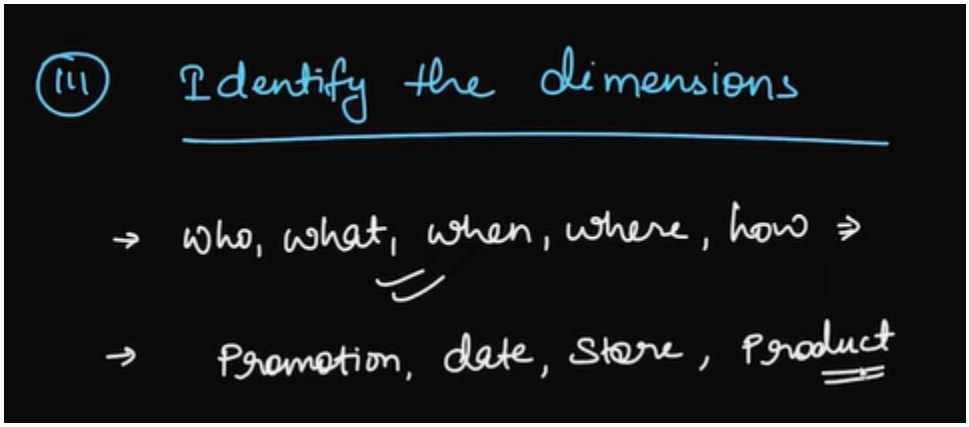
Trxn_id	Cust_id	Product_id	Date_id	Total_sales
TXN001	101	PRD001	20112023	200
TXN002	101	PRD003	20112023	150
TXN003	101	PRD007	20112023	15
TXN004	102	PRD004	20112023	85
TXN005	103	PRD006	20112023	20
TXN006	101	PRD005	21112023	24
TXN007	104	PRD001	22112023	200
TXN008	104	PRD005	22112023	24

14. Here without product_id..the data was in 5 rows..after adding the product_id..the no. of records went to 8
15. So here if you want to know no_of_product solds ..then we need to have another column

Trxn_id	Cust_id	Product_id	Date_id	Total_sales
TXN001	101	PRD001	20112023	200
TXN002	101	PRD003	20112023	150
TXN003	101	PRD007	20112023	15

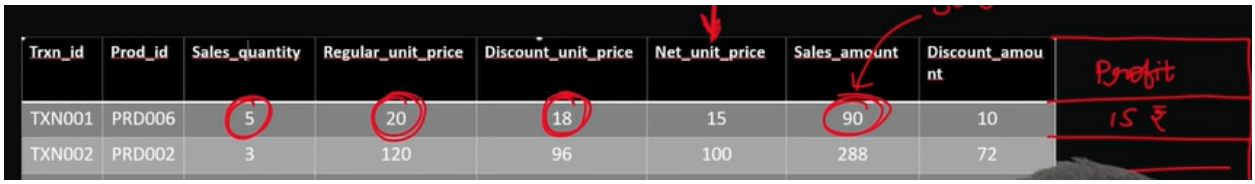
16. Step3 would be to identify the dimensions

17. From the data..we have to what is says



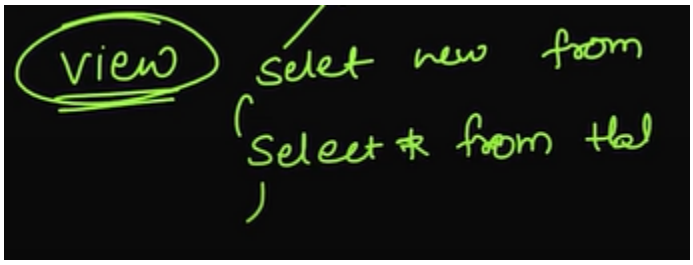
Promotion,date,store,product are the dimensional tables of our prev lectures

18. Step4 is to identify the facts
19. We must know which measurements is related to which dimensional table
20. What is derived fact?n
21. From the above rows…we can calculate our profit using (net_unit_price * 5) - sales_amount



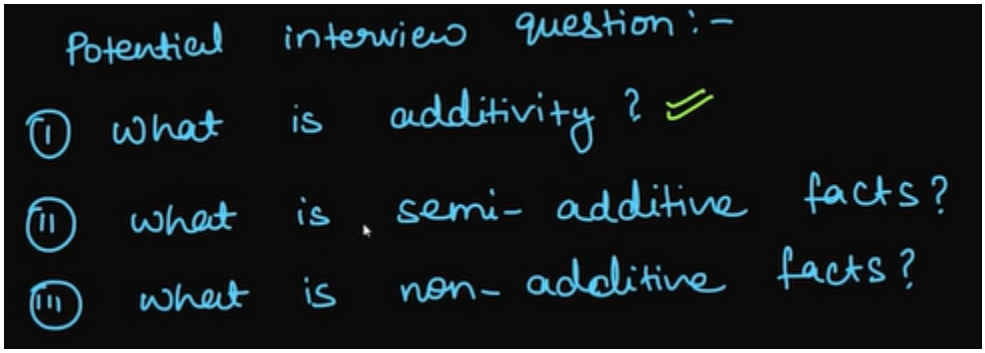
Trxn_id	Prod_id	Sales_quantity	Regular_unit_price	Discount_unit_price	Net_unit_price	Sales_amount	Discount_amount	Profit
TXN001	PRD006	5	20	18	15	90	10	15 ₹
TXN002	PRD002	3	120	96	100	288	72	

22. Now here profit column is the derived fact
23. So by adding derived fact..we are increasing the size of data…so instead we use view..which gives the extra col in the run time

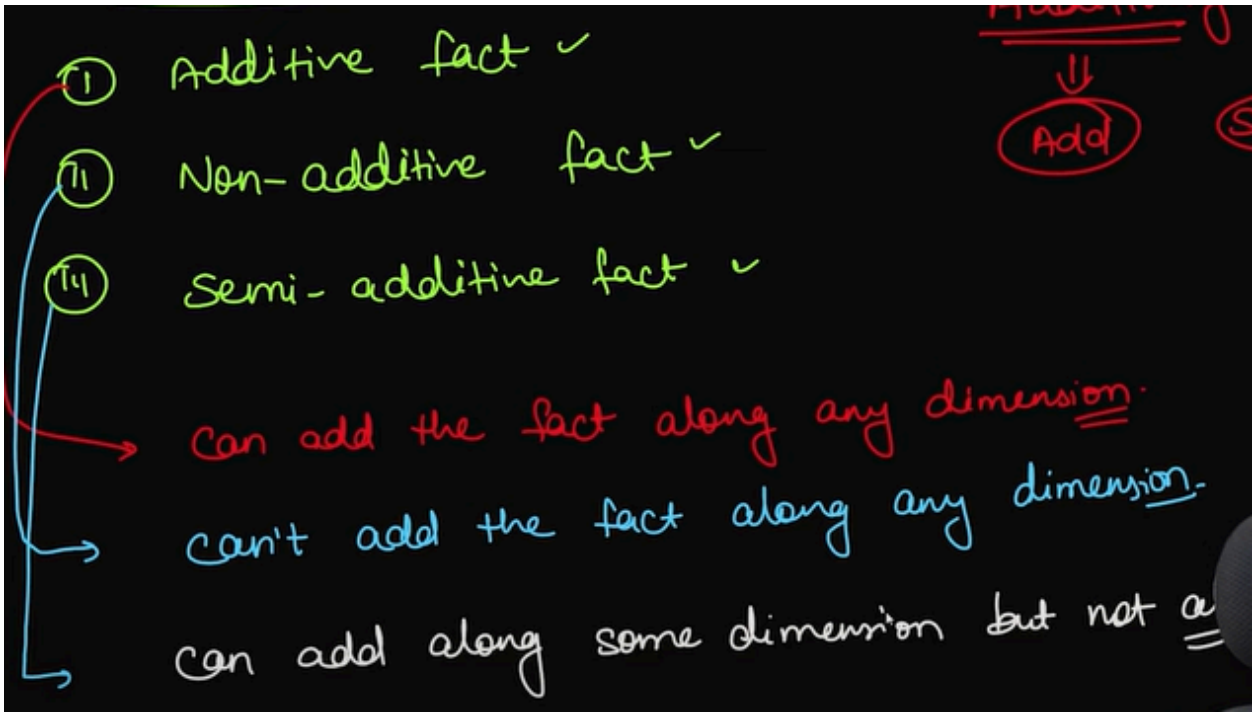


Types of Facts in Fact Table

1. Potential interview ques



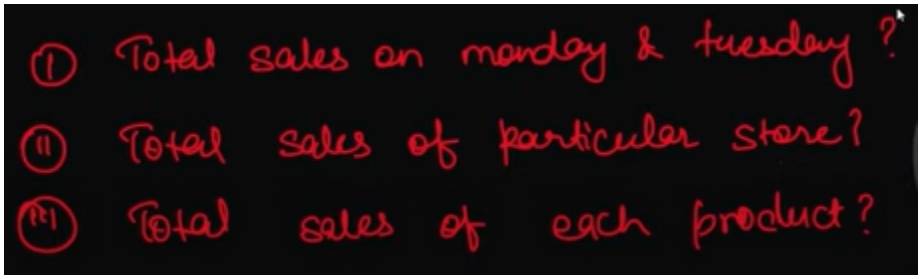
2. We have 3 types of fact tables



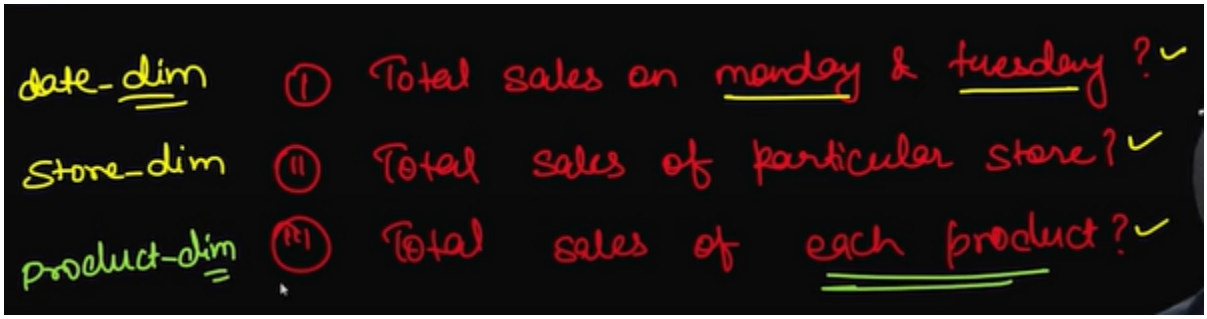
3. Lets understand with examples
4. Lets consider these sample table and our target is sales_amount col

Trxn_id	Prod_id	Sales_quantity	Regular_unit_price	Discount_unit_price	Net_unit_price	Sales_amount	Discount_amount
TXN001	PRD006	5	20	18	15	90	10
TXN002	PRD002	3	120	96	100	288	72
TXN003	PRD004	7	85	68	60	476	119
TXN004	PRD005	1	24	21.6	20	21.6	2.4
TXN005	PRD003	1	150	135	150	135	
TXN006	PRD001	2	200	160	130	320	
TXN007	PRD007	6	5	5	4	30	

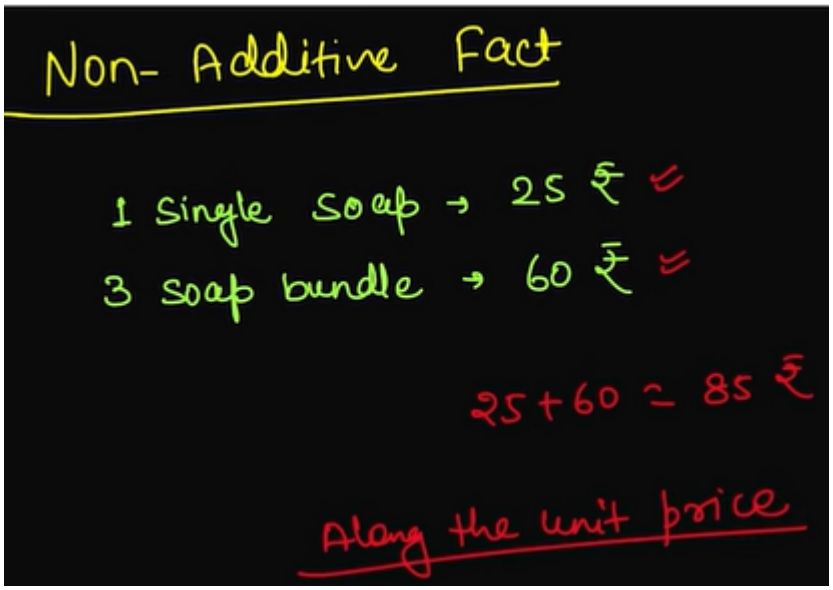
5. Now we need answers for this questions



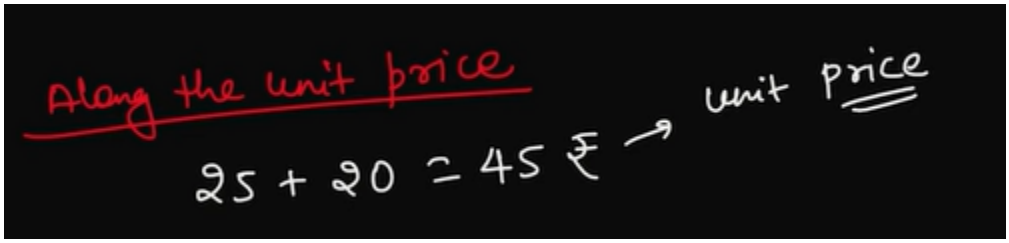
6. Here to get sales on monday & tuesday..we need date_dim, and for sales on particular store we need store_dim..similarly for product sales .. product_dim



7. Here we are adding the fact to any dim..so it is additive fact

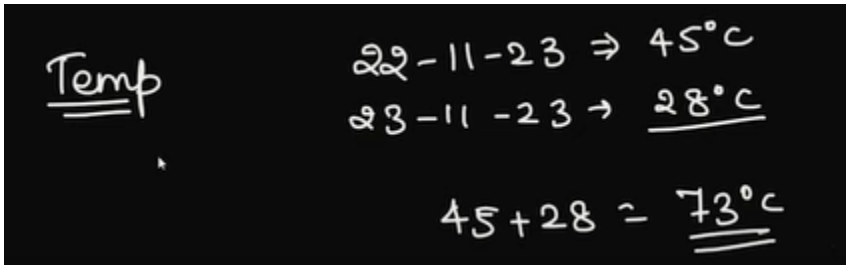


8. Non Additive facts
   Here..we can sum up the total sales of soaps
9. But here…we cannot add up their unit prices

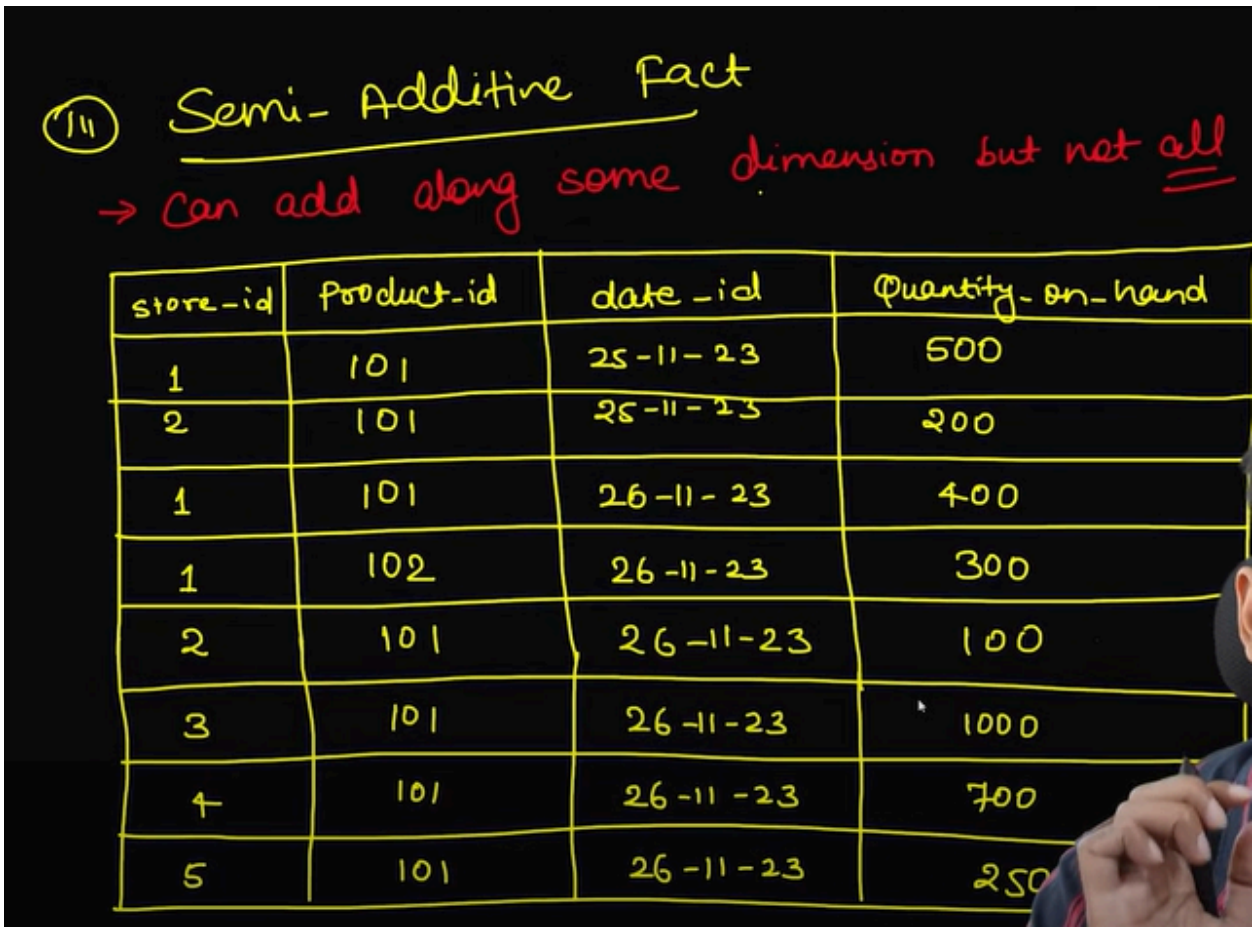


as both unit prices are diff

10. Similarly for temp…we cannot add two temperature



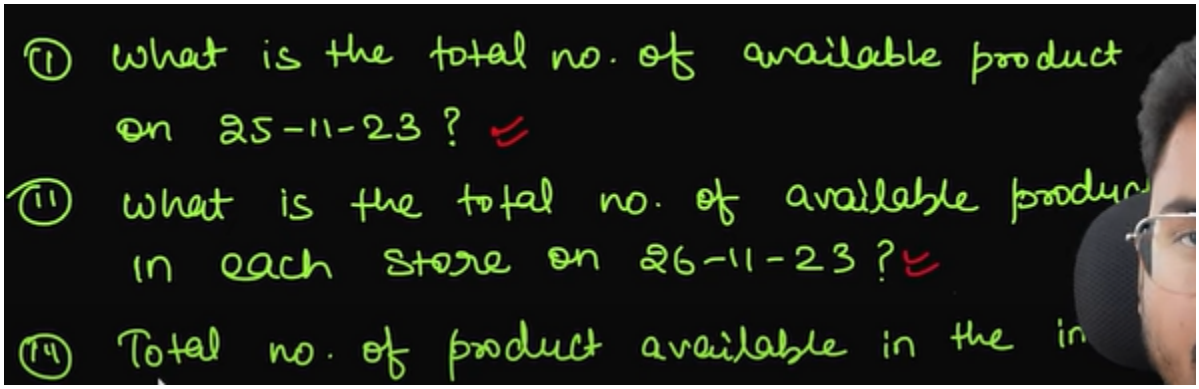
11. Semi Additive Fact



12.

13. First Understand this data..here a store on a particular date is storing a product_id with quantity given

14. Now lets answer this questions

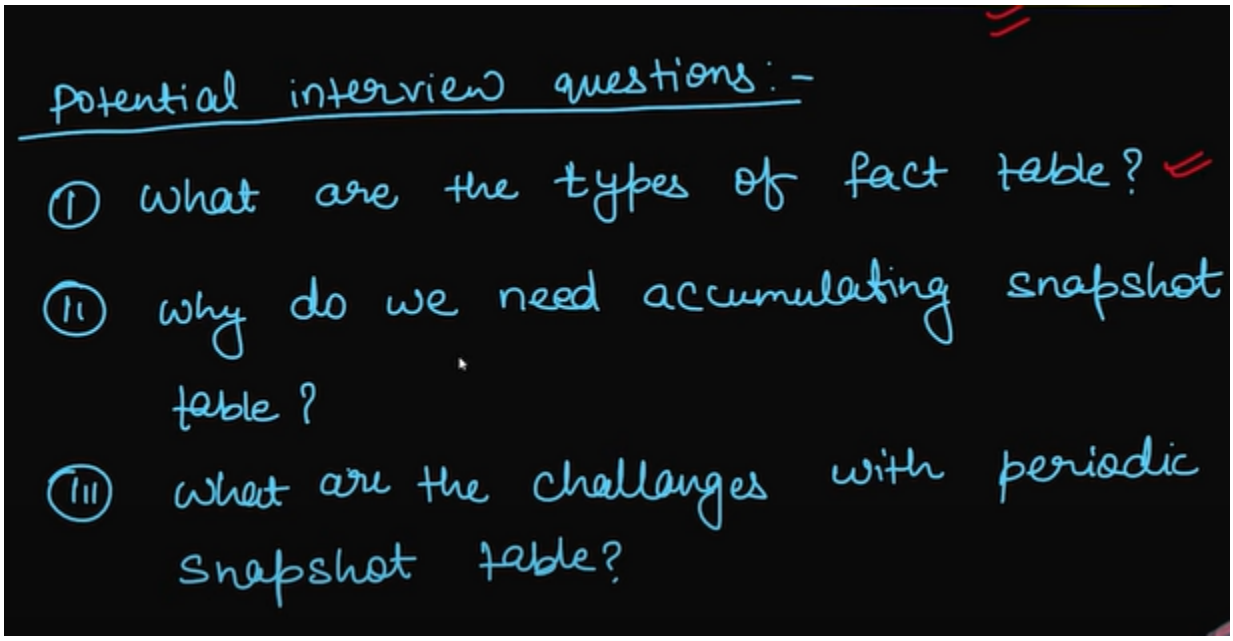


-inventory

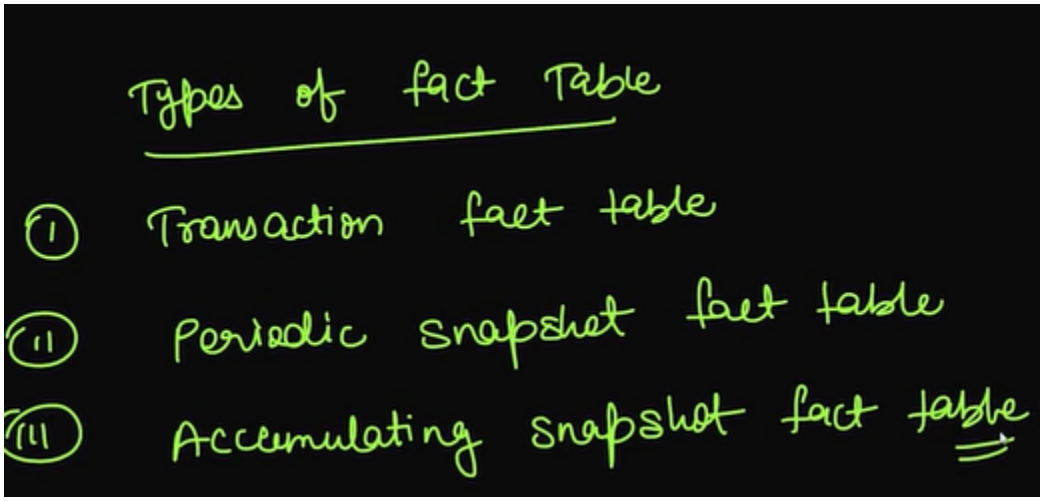
15. Here we can get 1st question and 2nd question
16. But in the 3rd question…as the quantity in store's varies on given date..we cannot calculate the no.of products in the inventory. ..
17. So this is semi additive

Types of Fact Table

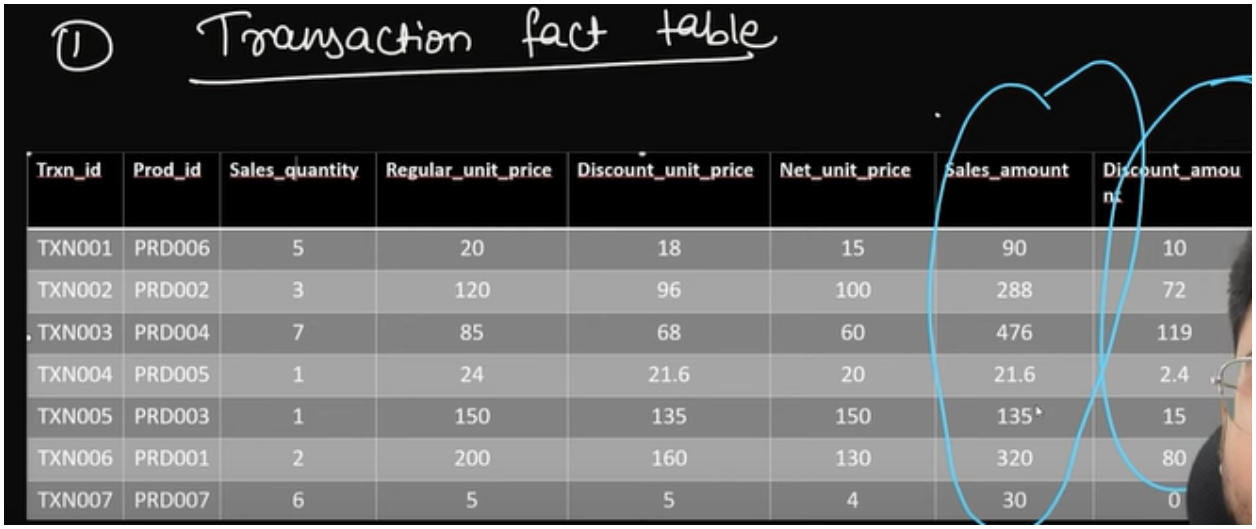
1. Potential Interview Questions



2. We have 3 types of fact table

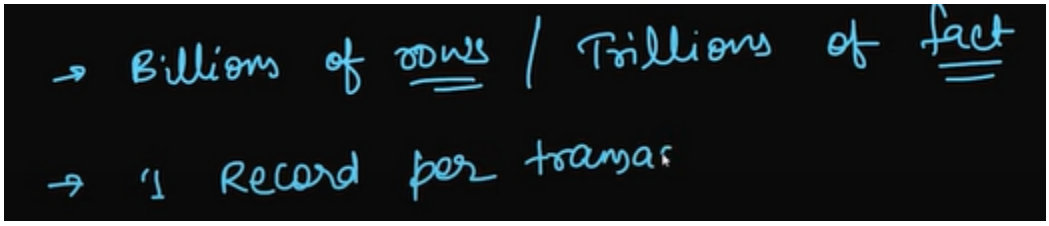


3. Transaction fact…lets consider our table



Trxn_id	Prod_id	Sales_quantity	Regular_unit_price	Discount_unit_price	Net_unit_price	Sales_amount	Discount_amount
TXN001	PRD006	5	20	18	15	90	10
TXN002	PRD002	3	120	96	100	288	72
TXN003	PRD004	7	85	68	60	476	119
TXN004	PRD005	1	24	21.6	20	21.6	2.4
TXN005	PRD003	1	150	135	150	135	15
TXN006	PRD001	2	200	160	130	320	80
TXN007	PRD007	6	5	5	4	30	0

4. Here we might have billions of rows and 1 record per transaction



1. Transaction Fact Table:

- Captures individual transactions or events in detail.
- Like a cash register receipt, it shows each sale with items, price, time, etc.
- Example: A retail store might have a transaction table recording every sale, including customer ID, product ID, quantity, price, and date/time.

5. Periodic snapshot table

2. Periodic Snapshot Fact Table:

- Provides a summary of data at specific periods (day, week, month).
- Like a monthly sales report, it shows totals over a timeframe.
- Example: A website might have a periodic snapshot table summarizing daily website traffic by country, source, and number of visits.

6. Here we are taking snapshot of our quantity each daily

store-id	Product-id	date-id	Quantity-on-hand
1	101	25-11-23	500
2	101	25-11-23	200
1	101	26-11-23	400

7. Now we can calculate the quantities sold on each day
8. Challeges
9. So if we have ..1000 stores and 10000 products then we'll have 10^7 rows daily

25-11-2023 ⇒

1000 Store, 10000 Products ⇒ 1000 × 10000 = 10^7 ⇒

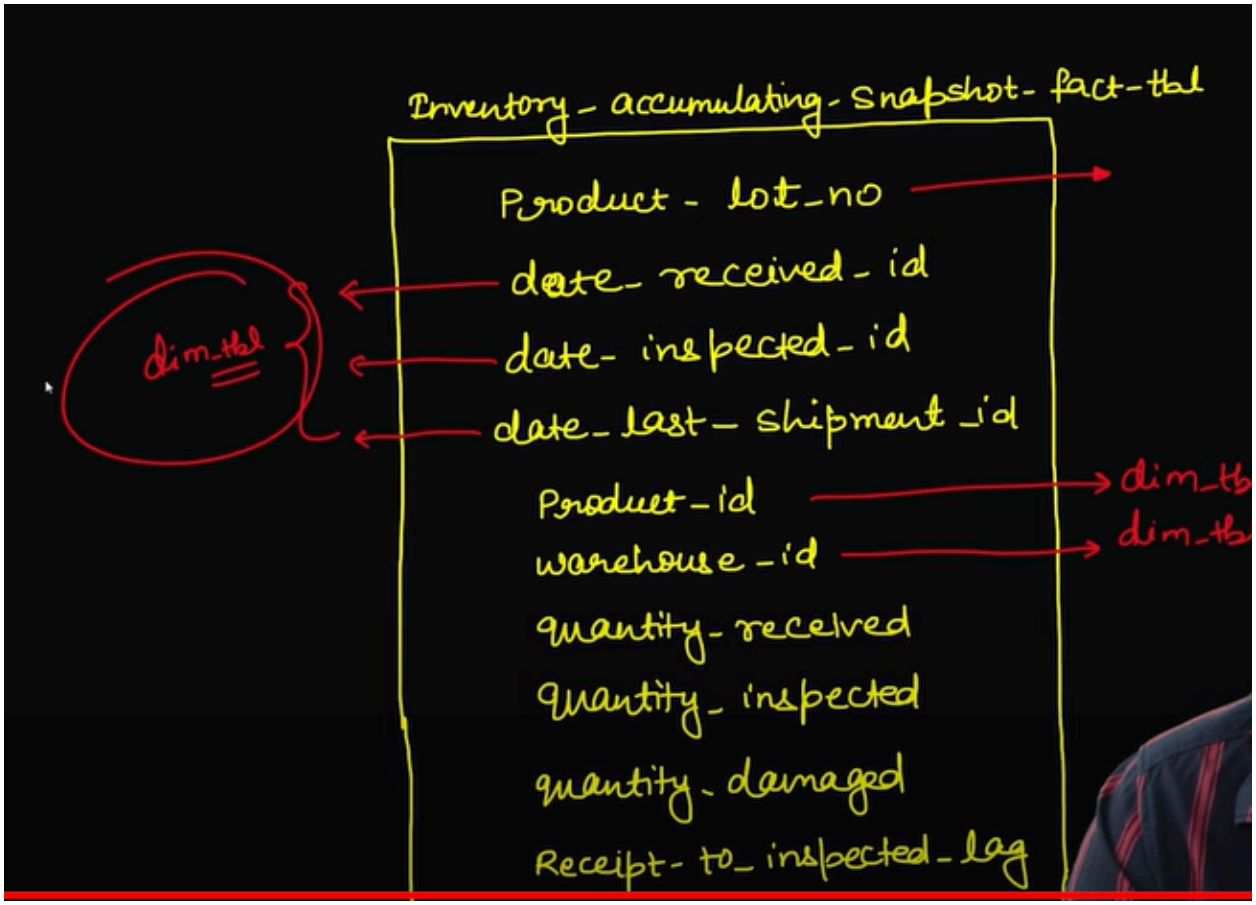
10. And if we take snapshots daily for year 365 × 10M = 3650M Record
11. So Instead of this we'll take snapshots for each month…or we'll use aggregate functions to decrease the no.of rows
12. Accumulating snapshot fact table

→ where we have a fix start and fix end.

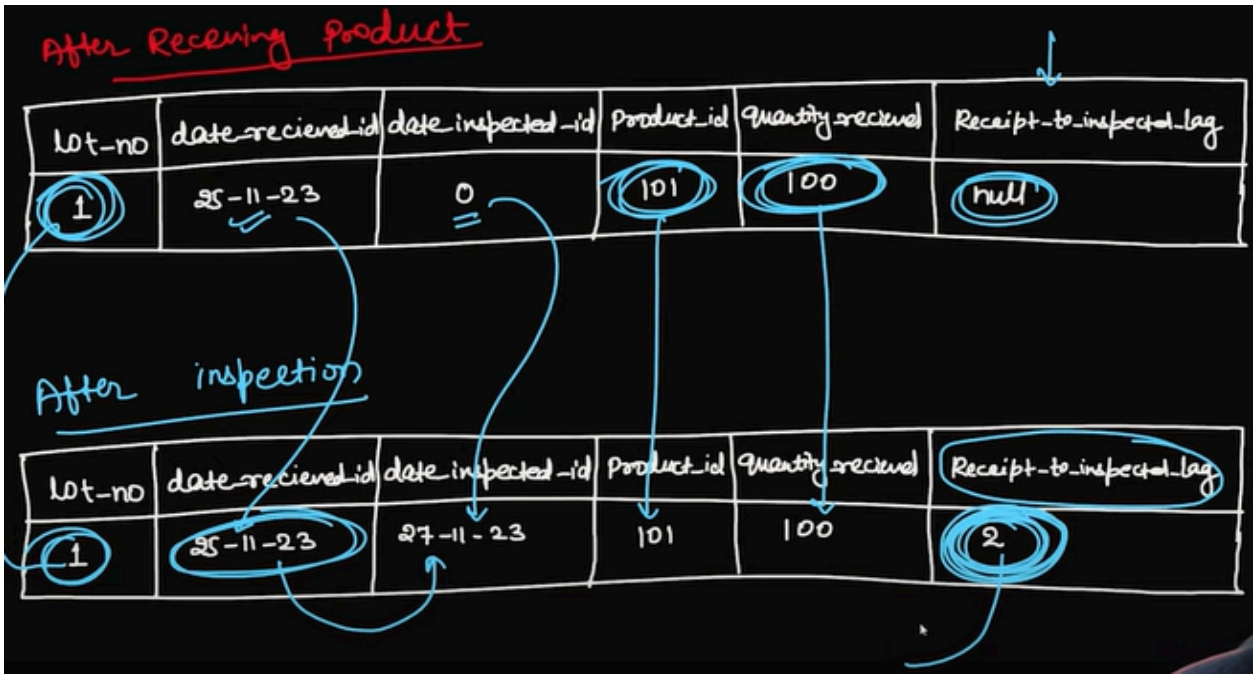
eg. order, warehouse order, inventory order

it has a fixed start and end point

13. Lets take an example table schema



14. Here measurements are quantity_received, quantity inspected,quantity damaged
15. Now after receiving product



16. Here for the same record…we have update date_inspected and receipt-to-inspected lag (2 days)