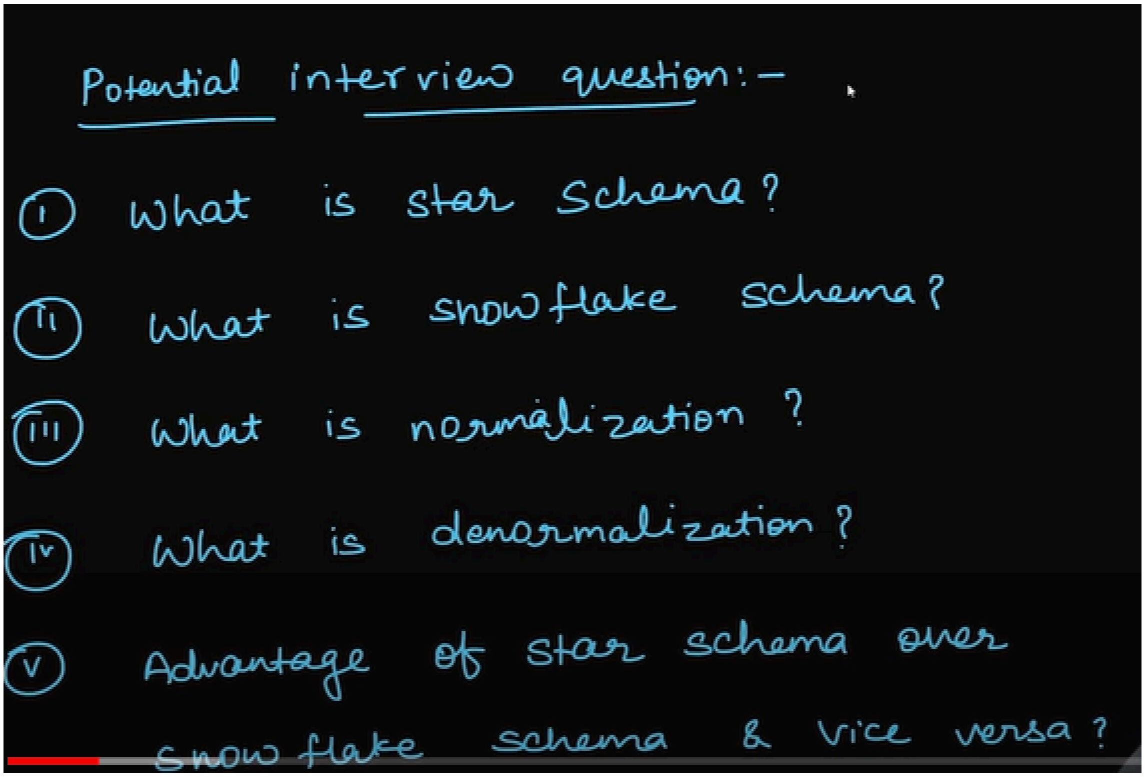
Star and Snowflake Schema

1. Potential Interview Questions
2. 

   Potential interview question:-

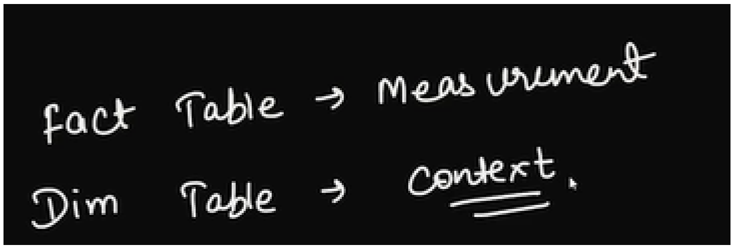
   (i) What is star Schema?

   (ii) What is snowflake schema?

   (iii) What is normalization ?

   (iv) What is denormalization ?

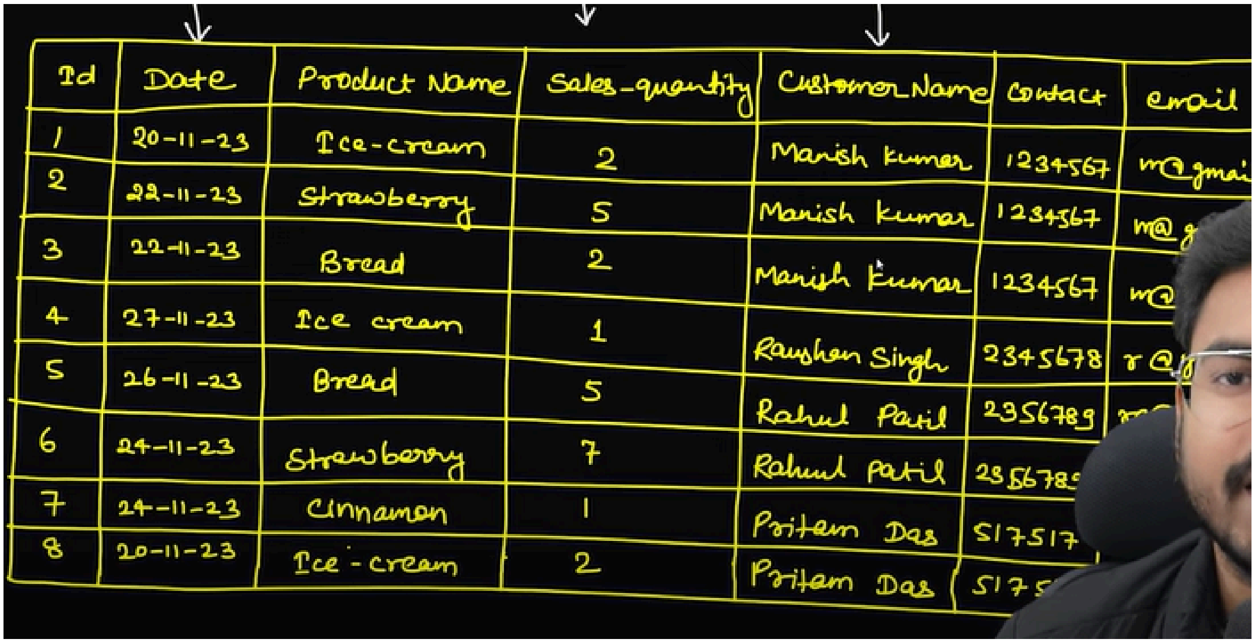
   (v) Advantage of star schema over snowflake schema & vice versa ?

3. Here we know that 

   Fact Table → Measurement

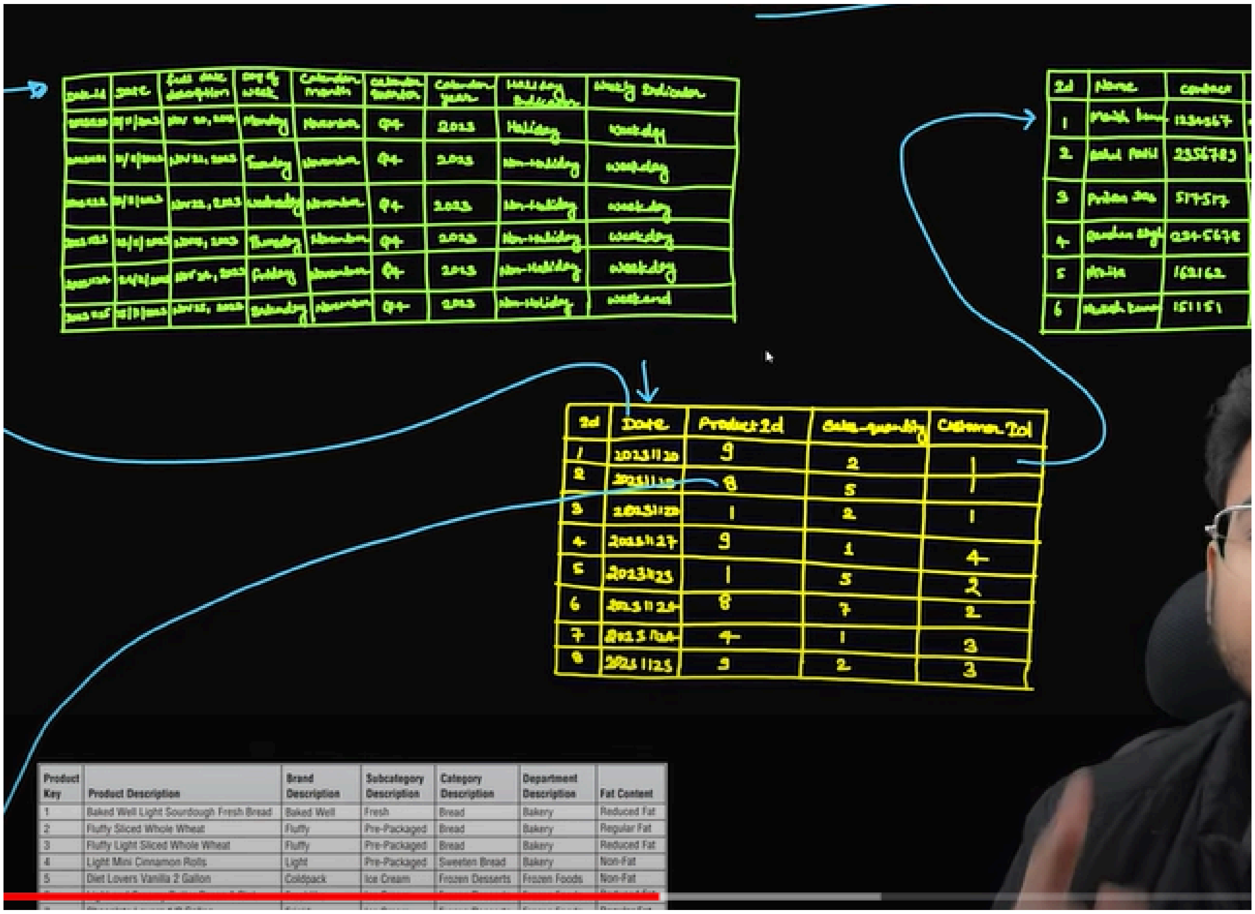
   Dim Table → Context.

4. Consider this table..here we have both dimension table and fact table column



Id	Date	Product Name	Sales-quantity	Customer_Name	Contact	email
1	20-11-23	Ice-cream	2	Manish kumar	1234567	m@gmai
2	22-11-23	Strawberry	5	Manish kumar	1234567	m@g
3	22-11-23	Bread	2	Manish Kumar	1234567	m@
4	27-11-23	Ice cream	1	Raushan Singh	2345678	r@
5	26-11-23	Bread	5	Rahul Patil	2356789	[illegible]
6	24-11-23	Strawberry	7	Rahul patil	235678	
7	24-11-23	Cinnamon	1	Pritam Das	517517	
8	20-11-23	Ice-cream	2	Pritam Das	517	

5. Date,Product_name and customer_name can act as dim table
6. Now let us know why our transactional table..must be divided into facts and dims
7. Here we can see our data in customer_name and product name is redundant..so we'll store this data in another table
8. So here we have date dim table,customer dim and product_dim table
9. Now in Star Schema our fact table…will be joined with dimensional tables

10. 

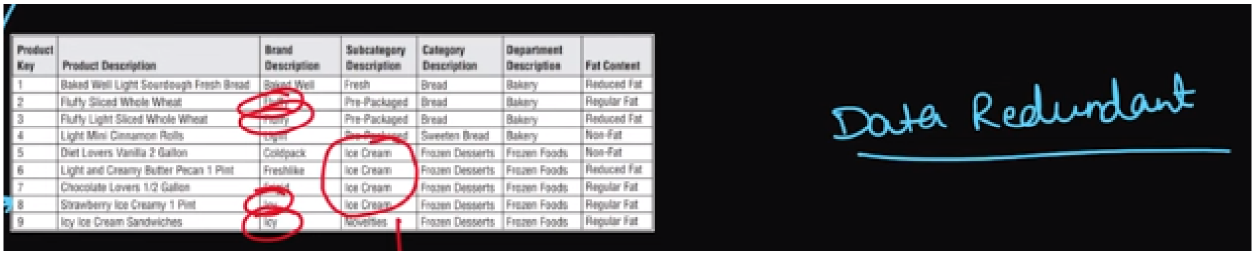
DateId	Date	Full date description	Day of week	Calendar month	Calendar quarter	Calendar year	Holiday Indicator	Weekly Indicator
20231120	20/11/2023	Nov 20, 2023	Monday	November	Q4	2023	Holiday	weekday
20231121	21/11/2023	Nov 21, 2023	Tuesday	November	Q4	2023	Non-Holiday	weekday
20231122	22/11/2023	Nov 22, 2023	Wednesday	November	Q4	2023	Non-Holiday	weekday
20231123	23/11/2023	Nov 23, 2023	Thursday	November	Q4	2023	Non-Holiday	weekday
20231124	24/11/2023	Nov 24, 2023	Friday	November	Q4	2023	Non-Holiday	weekday
20231125	25/11/2023	Nov 25, 2023	Saturday	November	Q4	2023	Non-Holiday	weekend

Id	Name	Contact
1	Manish Kumar	1234567
2	Rahul [illegible]	2356789
3	[illegible]	517517
4	[illegible] Singh	0234-5678
5	[illegible]	162162
6	Manish Kumar	151151

Id	Date	Product Id	Sale quantity	Customer Id
1	20231120	9	2	1
2	20231120	8	5	1
3	20231120	1	2	1
4	20231127	9	1	4
5	20231123	1	5	2
6	20231124	8	7	2
7	20231124	4	1	3
8	20231125	9	2	3

Product Key	Product Description	Brand Description	Subcategory Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Fresh	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Reduced Fat
4	Light Mini Cinnamon Rolls	Light	Pre-Packaged	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Ice Cream	Frozen Desserts	Frozen Foods	Non-Fat

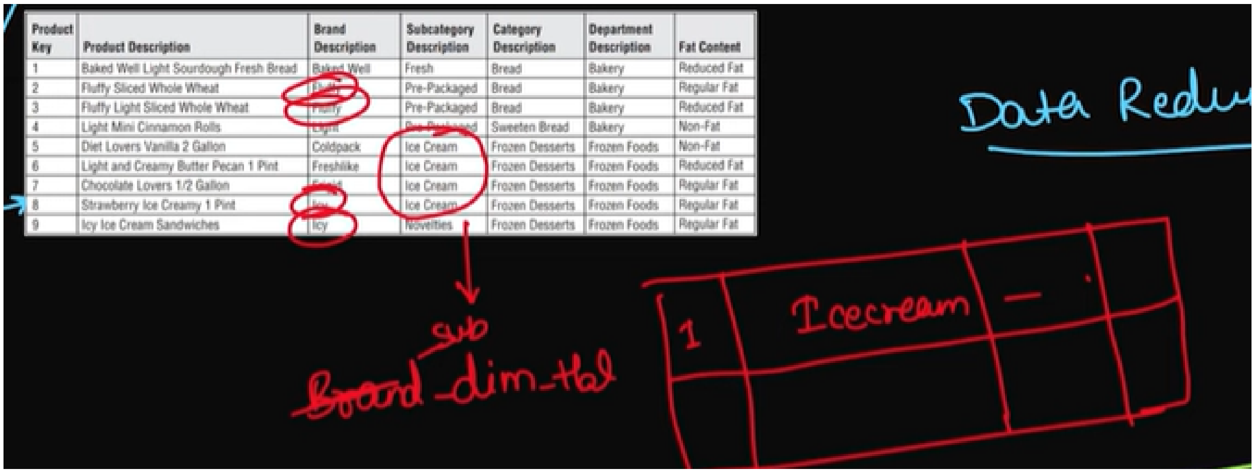
11. And here the main problem data redundancy



Product Key	Product Description	Brand Description	Subcategory Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Fresh	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Reduced Fat
4	Light Mini Cinnamon Rolls	Light	Pre-Packaged	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Ice Cream	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Ice Cream	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Novelties	Frozen Desserts	Frozen Foods	Regular Fat

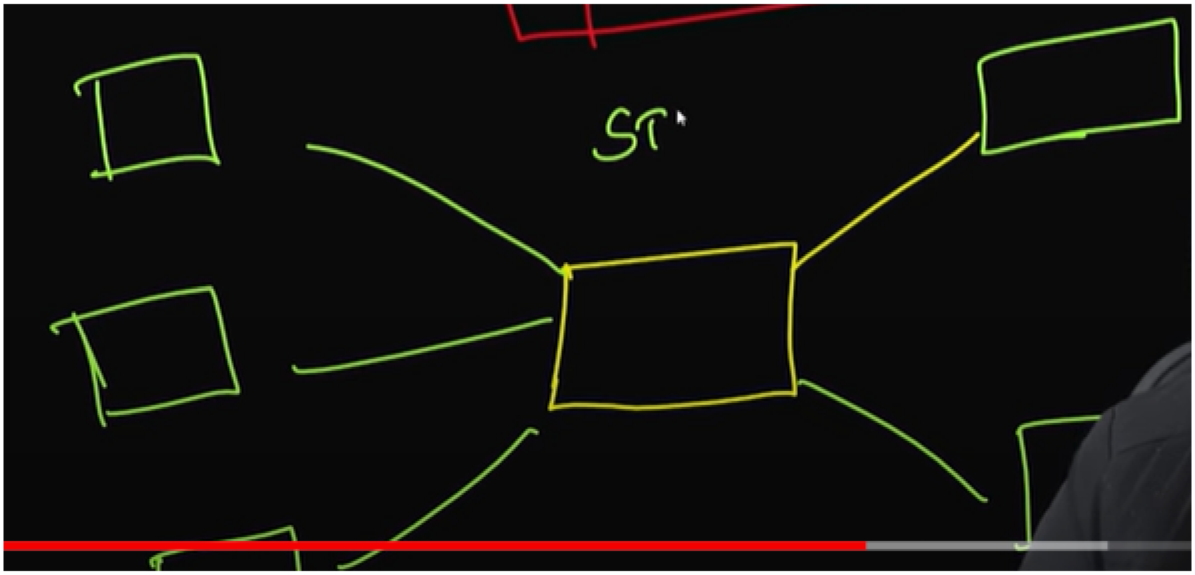
Data Redundant

12. So we create another table for as shown pic to avoid data redundancy in a table

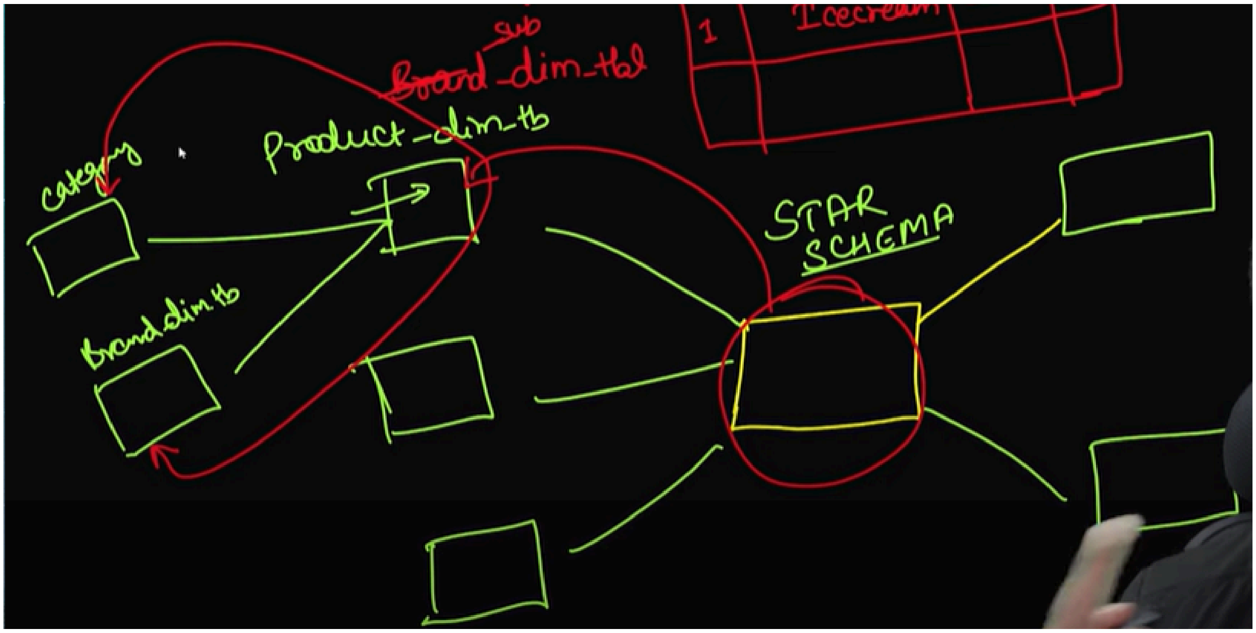


Product Key	Product Description	Brand Description	Subcategory Description	Category Description	Department Description	Fat Content
1	Baked Well Light Sourdough Fresh Bread	Baked Well	Fresh	Bread	Bakery	Reduced Fat
2	Fluffy Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Regular Fat
3	Fluffy Light Sliced Whole Wheat	Fluffy	Pre-Packaged	Bread	Bakery	Reduced Fat
4	Light Mini Cinnamon Rolls	Light	Pre-Packaged	Sweeten Bread	Bakery	Non-Fat
5	Diet Lovers Vanilla 2 Gallon	Coldpack	Ice Cream	Frozen Desserts	Frozen Foods	Non-Fat
6	Light and Creamy Butter Pecan 1 Pint	Freshlike	Ice Cream	Frozen Desserts	Frozen Foods	Reduced Fat
7	Chocolate Lovers 1/2 Gallon	Frigid	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
8	Strawberry Ice Creamy 1 Pint	Icy	Ice Cream	Frozen Desserts	Frozen Foods	Regular Fat
9	Icy Ice Cream Sandwiches	Icy	Novelties	Frozen Desserts	Frozen Foods	Regular Fat

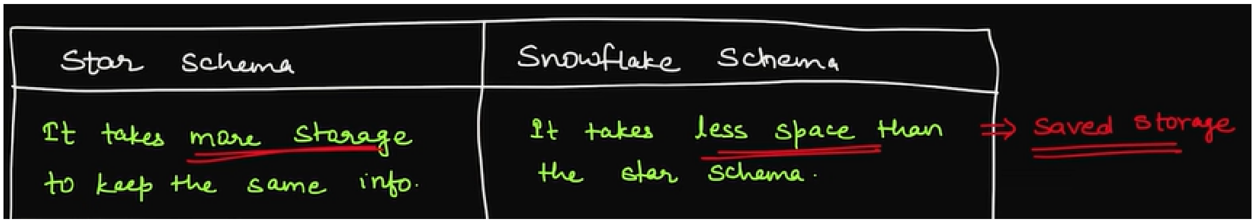
13. So if we are going just one level to get the data ..then it is called star schema



14. SO if we are dealing with more than 1 table to get the data..then it is snowflake schema

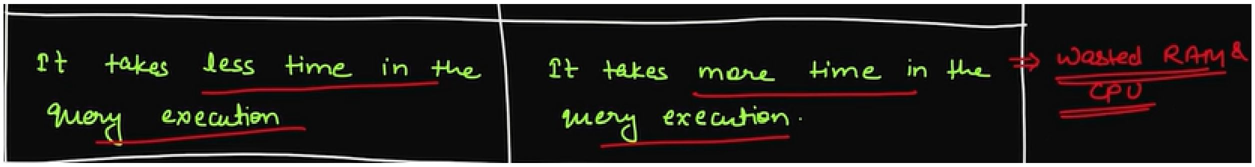


15. Here in the above pic..product_dim has 2 more tables attached
16. Diff bw star schema and snowflake schema



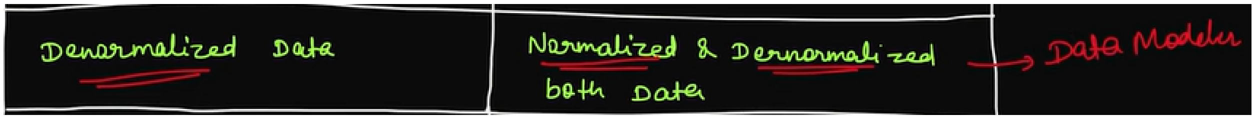
Star Schema	Snowflake Schema	
It takes more storage to keep the same info.	It takes less space than the star schema.	⇒ Saved storage

Snowflake schema removes data redundancy by creating separate tables



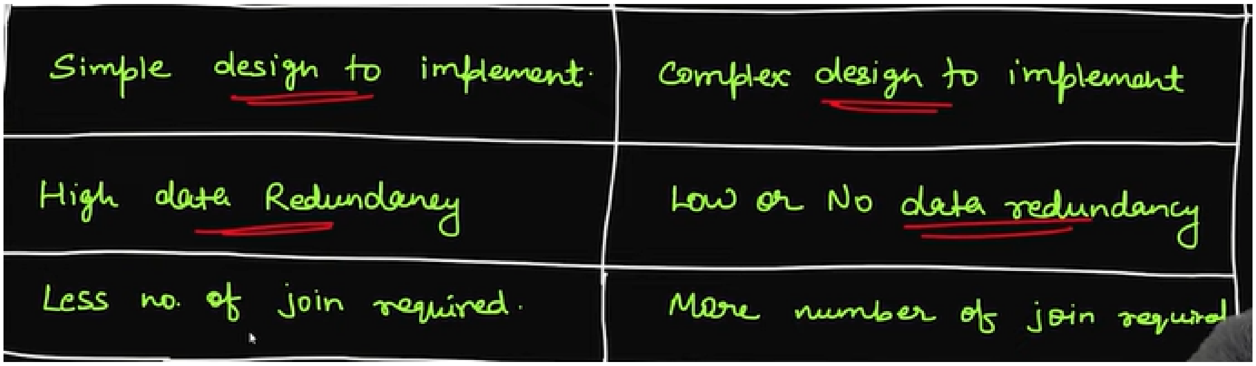
It takes less time in the query execution	It takes more time in the query execution.	⇒ Wasted RAM & CPU

In snowflake we need to use joins as there are diff tables..so it wastes RAM and CPU

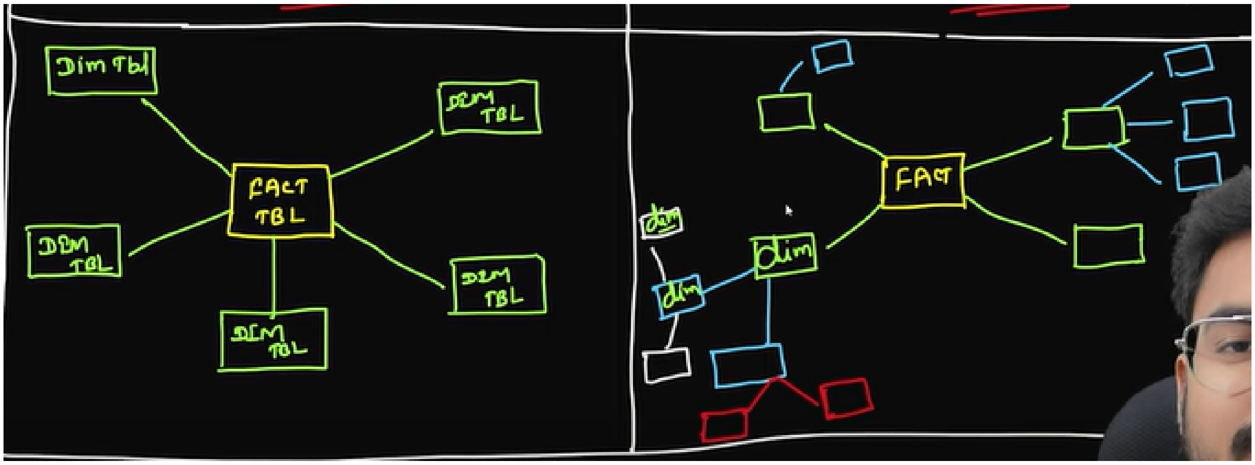


Denormalized Data	Normalized & Dernormalized both Data	→ Data Modeler

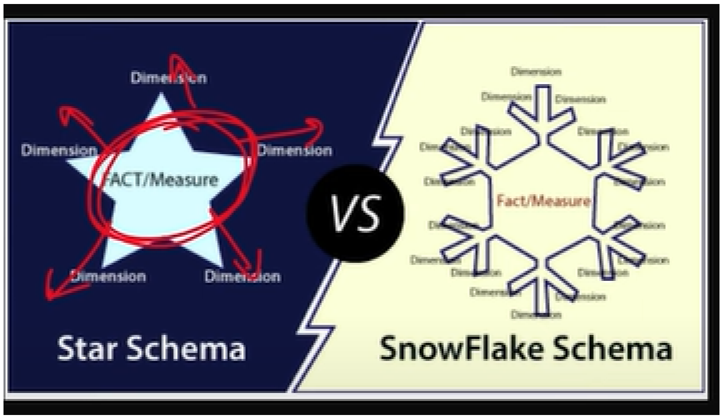
in snowflake it can be both normalized and denormalized ..depends on data modeler



Simple design to implement.	Complex design to implement
High data Redundancy	Low or No data redundancy
Less no. of join required.	More number of join required



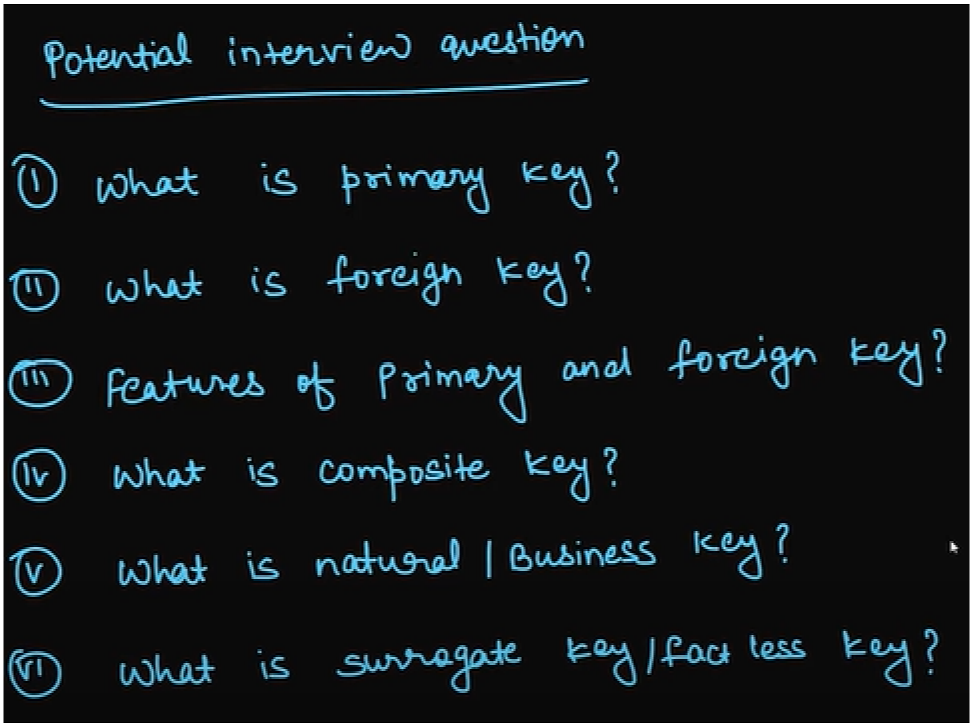
Structure of starschema and snowflake schema



17. Normalization and denormalization explained : https://g.co/gemini/share/e1bc9101c008

Primary Key and Foreign Key

1. Potential Interview Questions



Potential interview question

(i) What is primary key?

(ii) what is foreign key?

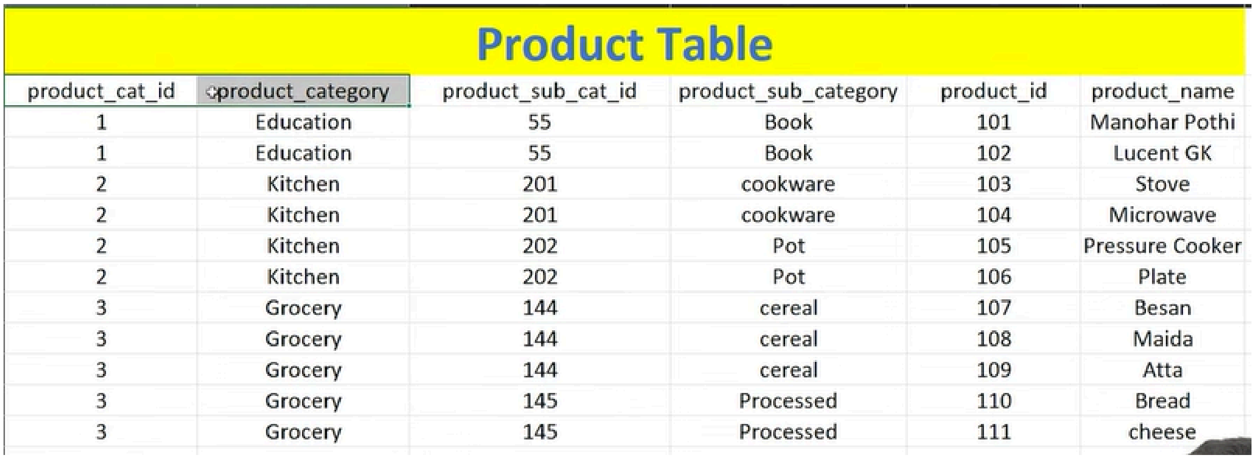
(iii) Features of Primary and foreign key?

(iv) What is composite key?

(v) What is natural | Business key?

(vi) What is surrogate key/fact less key?

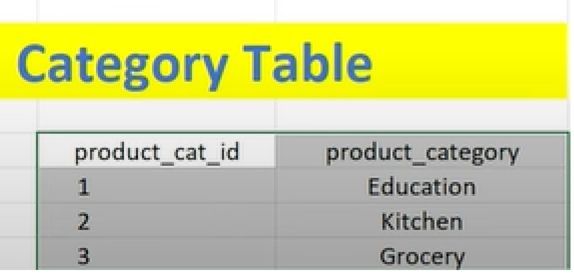
2. Lets consider this sample table



Product Table

product_cat_id	product_category	product_sub_cat_id	product_sub_category	product_id	product_name
1	Education	55	Book	101	Manohar Pothi
1	Education	55	Book	102	Lucent GK
2	Kitchen	201	cookware	103	Stove
2	Kitchen	201	cookware	104	Microwave
2	Kitchen	202	Pot	105	Pressure Cooker
2	Kitchen	202	Pot	106	Plate
3	Grocery	144	cereal	107	Besan
3	Grocery	144	cereal	108	Maida
3	Grocery	144	cereal	109	Atta
3	Grocery	145	Processed	110	Bread
3	Grocery	145	Processed	111	cheese

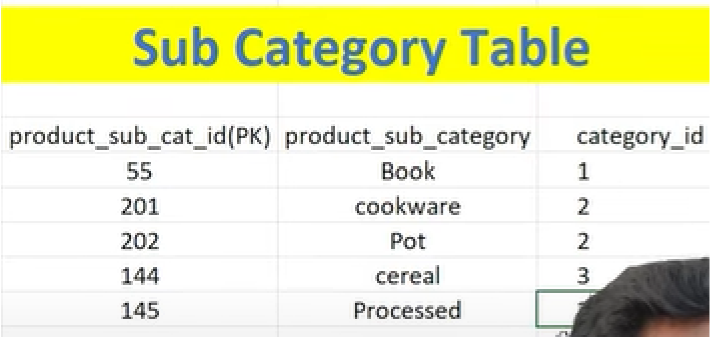
3. we have product_id…which uniquely identifies each row and there wont be any duplicates in the table it is called primary key
4. Here for product_category…there are categories which are repeating multiple times



Category Table

product_cat_id	product_category
1	Education
2	Kitchen
3	Grocery

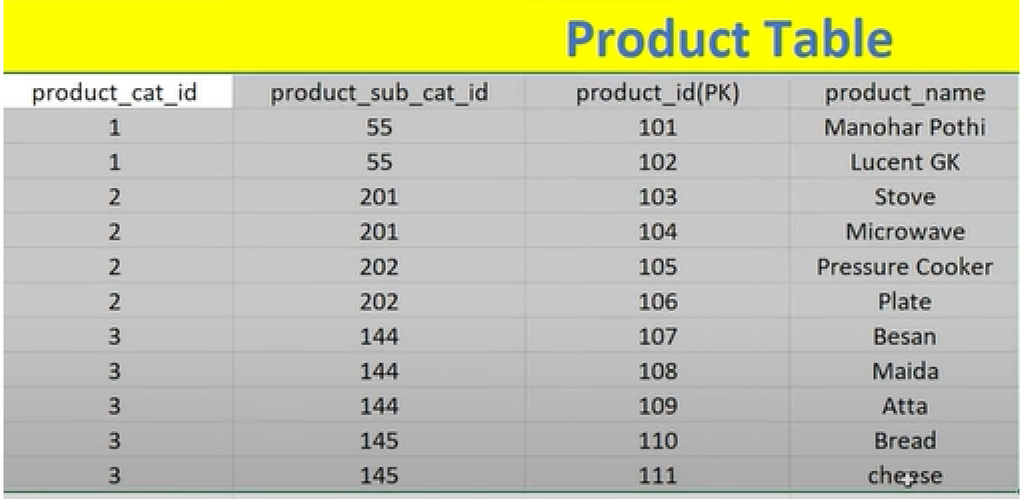
5. So we'll just store this data in another table
6. And we did same for the sub categories too



Sub Category Table

product_sub_cat_id(PK)	product_sub_category	category_id
55	Book	1
201	cookware	2
202	Pot	2
144	cereal	3
145	Processed	[illegible]

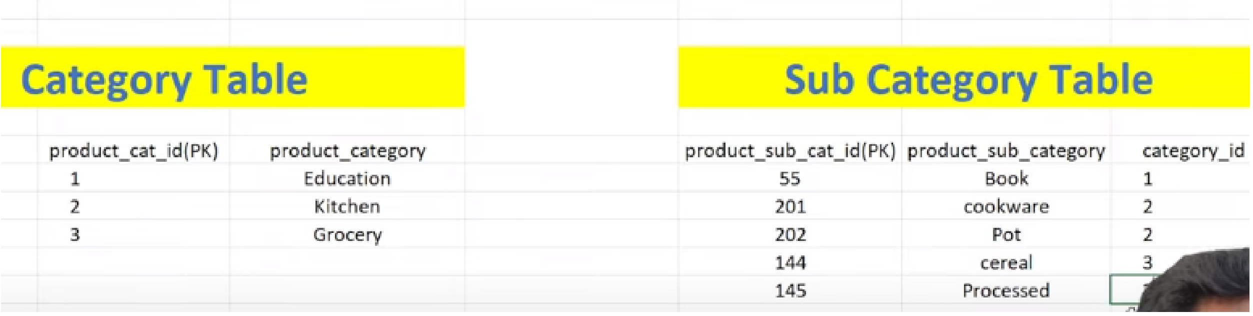
7. After removing this columns ..our original table will look like this…and if we need categories or sub cate…we can join the tables on id's



Product Table

product_cat_id	product_sub_cat_id	product_id(PK)	product_name
1	55	101	Manohar Pothi
1	55	102	Lucent GK
2	201	103	Stove
2	201	104	Microwave
2	202	105	Pressure Cooker
2	202	106	Plate
3	144	107	Besan
3	144	108	Maida
3	144	109	Atta
3	145	110	Bread
3	145	111	cheese

8. Here id are the primary columns of their table



Category Table

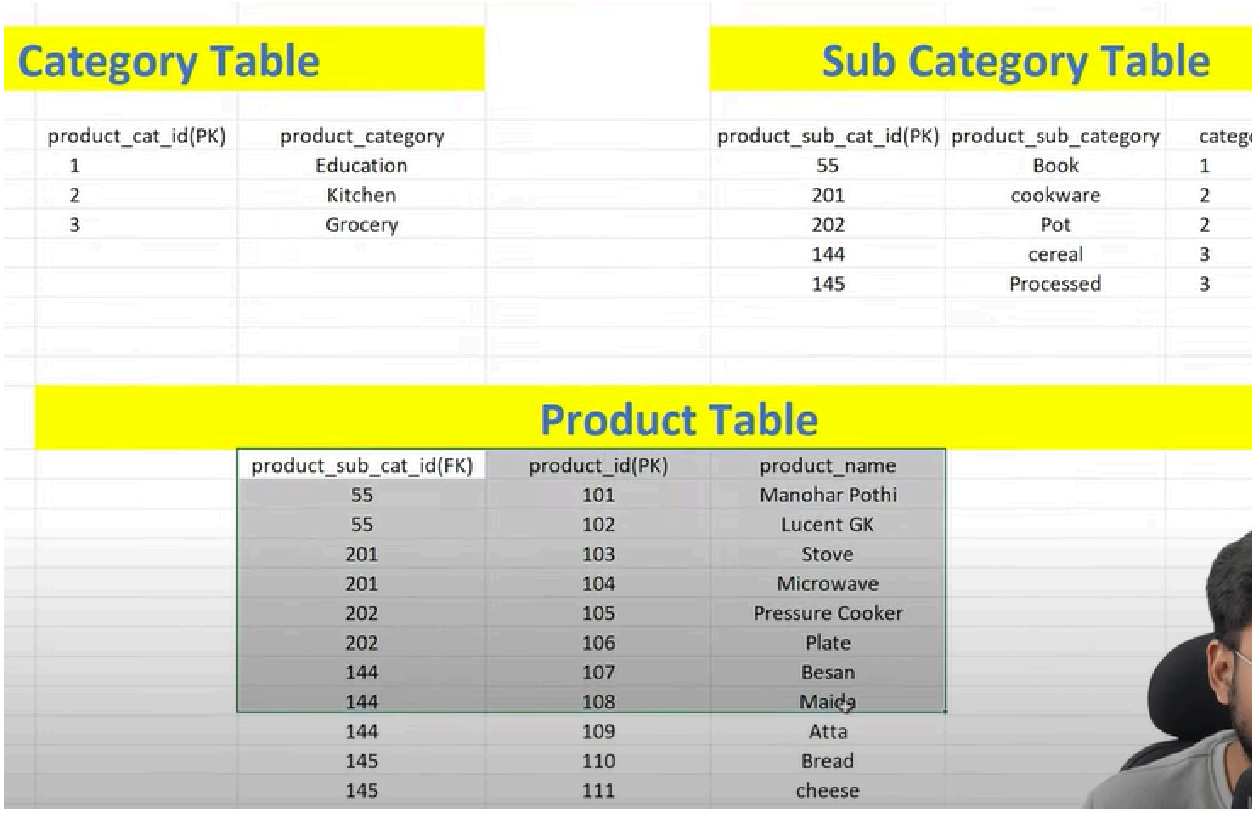
product_cat_id(PK)	product_category
1	Education
2	Kitchen
3	Grocery

Sub Category Table

product_sub_cat_id(PK)	product_sub_category	category_id
55	Book	1
201	cookware	2
202	Pot	2
144	cereal	3
145	Processed	[illegible]

9. Now in our original table…if we there are primary keys of other tables..then they are foreign keys

10. And Also the foreign key column may have the duplicates



Category Table

product_cat_id(PK)	product_category
1	Education
2	Kitchen
3	Grocery

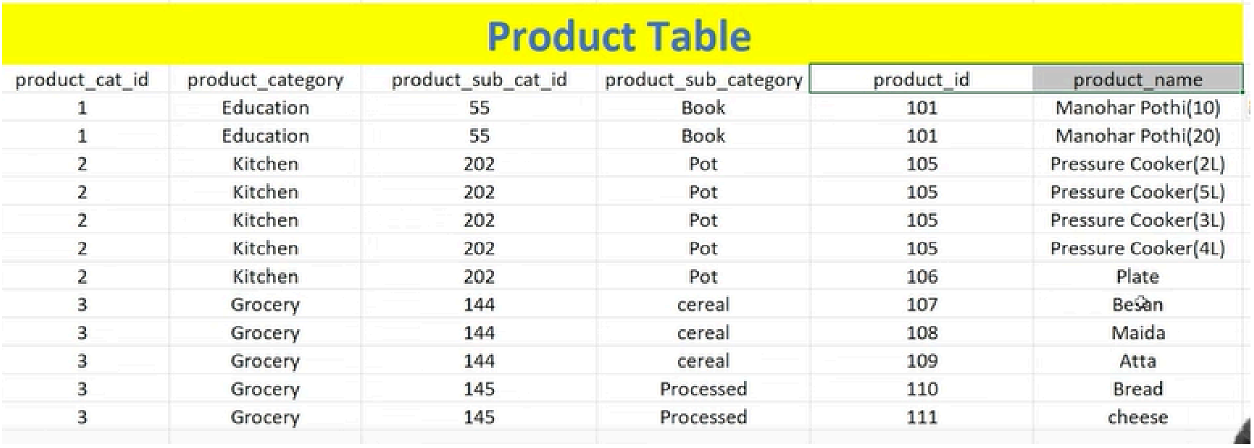
Sub Category Table

product_sub_cat_id(PK)	product_sub_category	categ
55	Book	1
201	cookware	2
202	Pot	2
144	cereal	3
145	Processed	3

Product Table

product_sub_cat_id(FK)	product_id(PK)	product_name
55	101	Manohar Pothi
55	102	Lucent GK
201	103	Stove
201	104	Microwave
202	105	Pressure Cooker
202	106	Plate
144	107	Besan
144	108	Maida
144	109	Atta
145	110	Bread
145	111	cheese

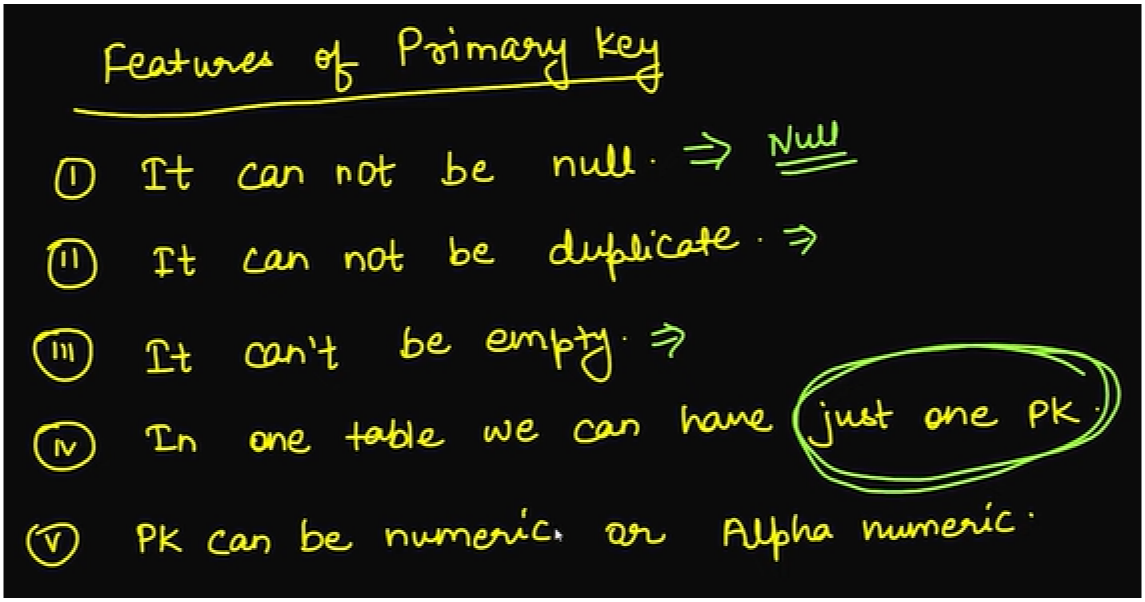
11. Now we can use joins..and get all the data required
12. Consider another table



Product Table

product_cat_id	product_category	product_sub_cat_id	product_sub_category	product_id	product_name
1	Education	55	Book	101	Manohar Pothi(10)
1	Education	55	Book	101	Manohar Pothi(20)
2	Kitchen	202	Pot	105	Pressure Cooker(2L)
2	Kitchen	202	Pot	105	Pressure Cooker(5L)
2	Kitchen	202	Pot	105	Pressure Cooker(3L)
2	Kitchen	202	Pot	105	Pressure Cooker(4L)
2	Kitchen	202	Pot	106	Plate
3	Grocery	144	cereal	107	Besan
3	Grocery	144	cereal	108	Maida
3	Grocery	144	cereal	109	Atta
3	Grocery	145	Processed	110	Bread
3	Grocery	145	Processed	111	cheese

13. From the table above..if we clearly observe..we dont have any single column thats a primary key..
14. Here product_id is repeated multiple times…so its not a primary key
15. So here we will use both product_id and product_name as a key which identifies each row uniquely …so this type of key is called composite key

16. 

Features of Primary key

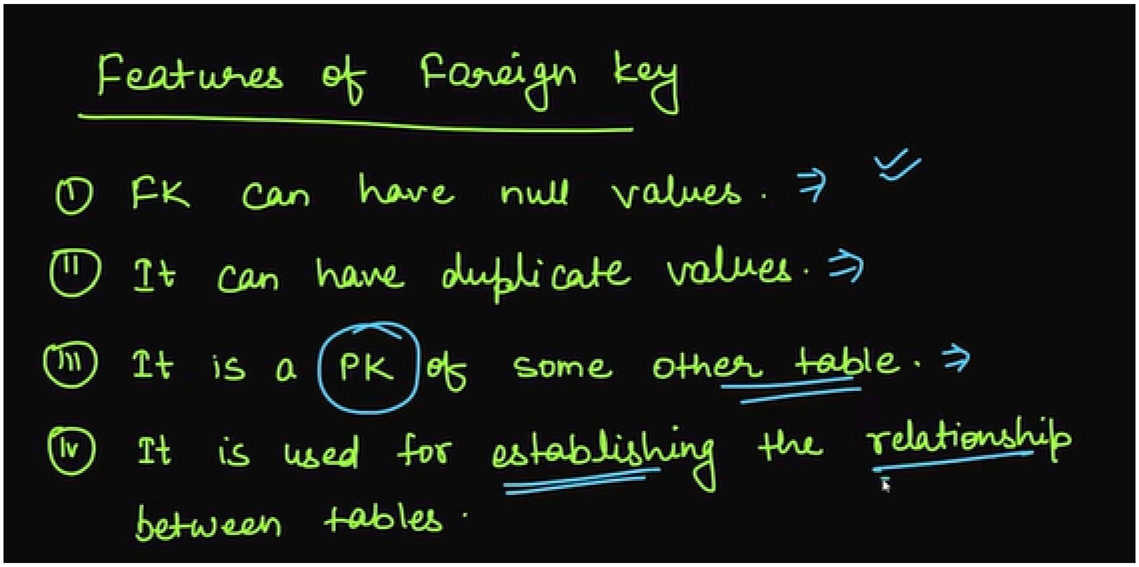
(I) It can not be null. ⇒ Null

(II) It can not be duplicate. ⇒

(III) It can't be empty. ⇒

(IV) In one table we can have just one PK.

(V) PK can be numeric or Alpha numeric.

17. 

Features of Foreign key

(I) FK can have null values. ⇒ ✓

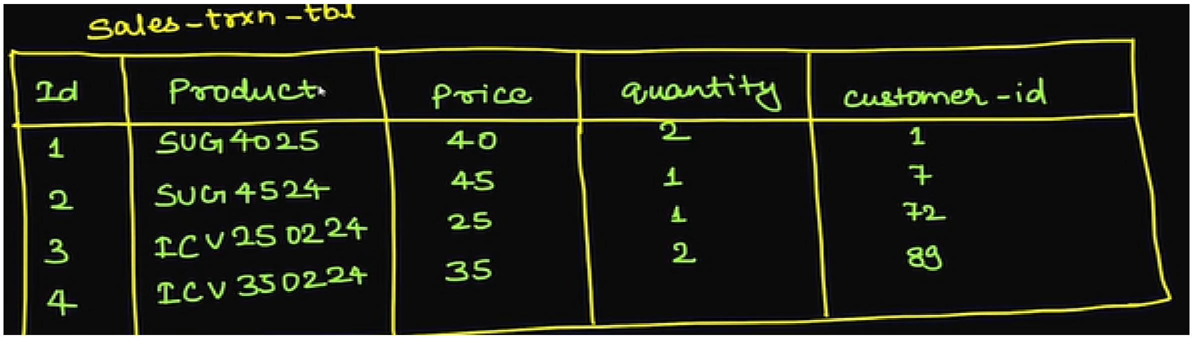
(II) It can have duplicate values. ⇒

(III) It is a PK of some other table. ⇒

(IV) It is used for establishing the relationship between tables.

Surrogate key and Natural key

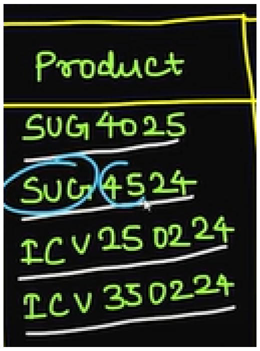
1. Lets consider a sales transaction table



sales-trxn-tbl

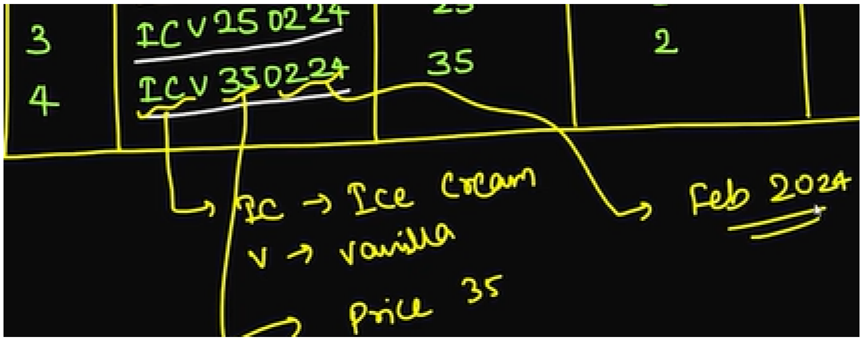
Id	Product	Price	quantity	customer-id
1	SUG4025	40	2	1
2	SUG4524	45	1	7
3	ICV250224	25	1	72
4	ICV350224	35	2	89

2. Product name SUG4025 - (SUGAR's price is 40/kg and its expiry is 25(2025))..so this kind of key is called business keys



Product
SUG4025
SUG4524
ICV250224
ICV350224

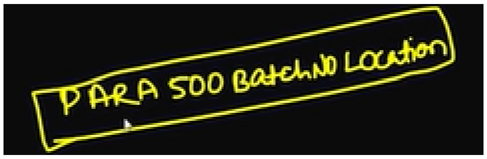
3. So likewise we can derive for every product its price and expiry date
4. Similarly for ICV250224



3	ICV250224	25	2
4	ICV350224	35	

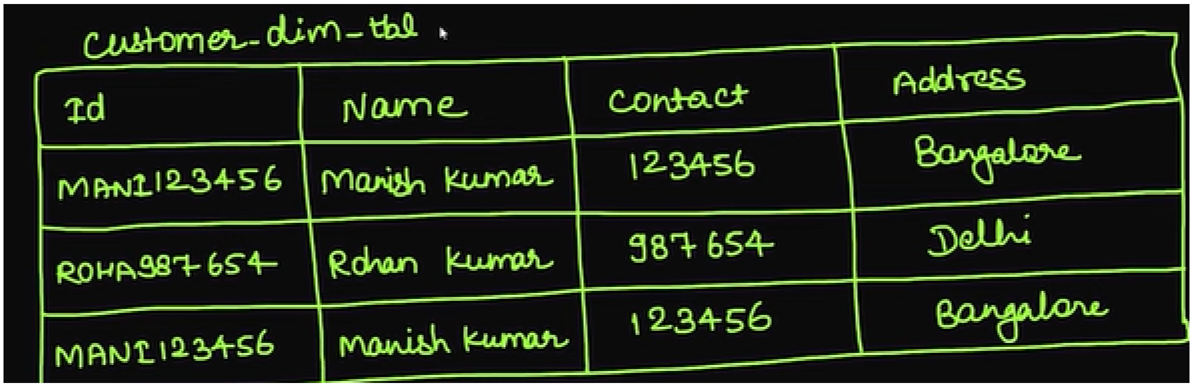
IC → Ice Cream
V → vanilla
Price 35
Feb 2024

5. Similarly for pharmacy table..we can design product name as



PARA 500 BatchNO Location

6. So here natural key can be uniquely identified or not
7. Now lets consider these table

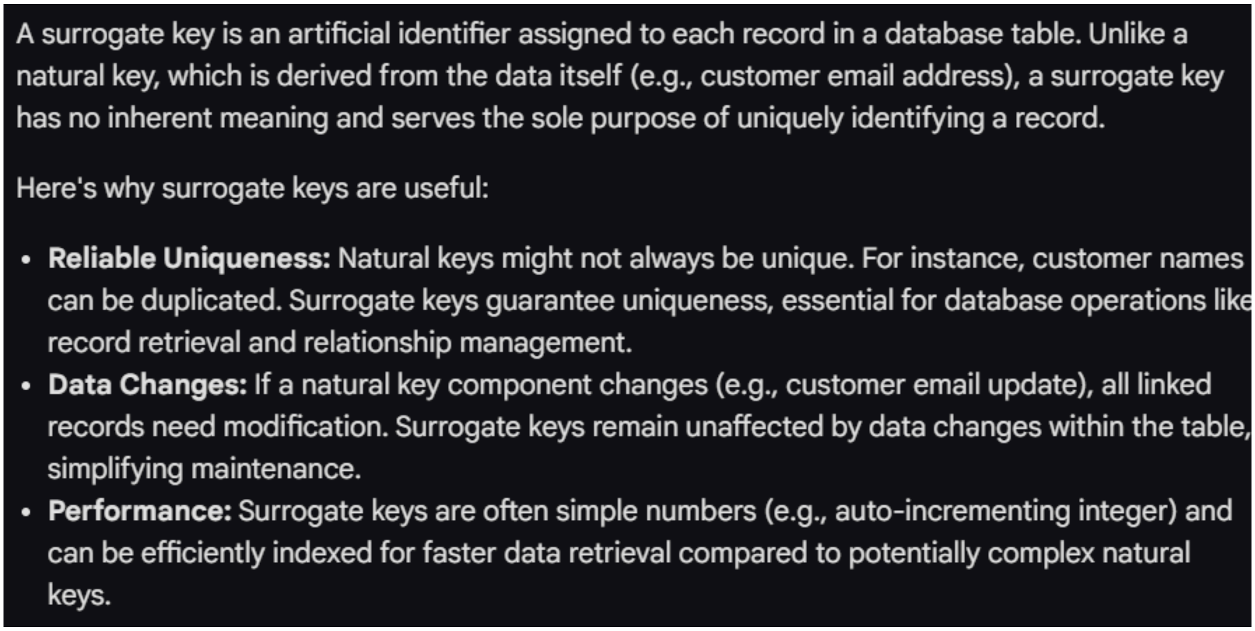


customer_dim_tbl

Id	Name	Contact	Address
MANI123456	Manish Kumar	123456	Bangalore
ROHA987654	Rohan Kumar	987654	Delhi
MANI123456	Manish Kumar	123456	Bangalore

8. For surrogate key ..refer vide0

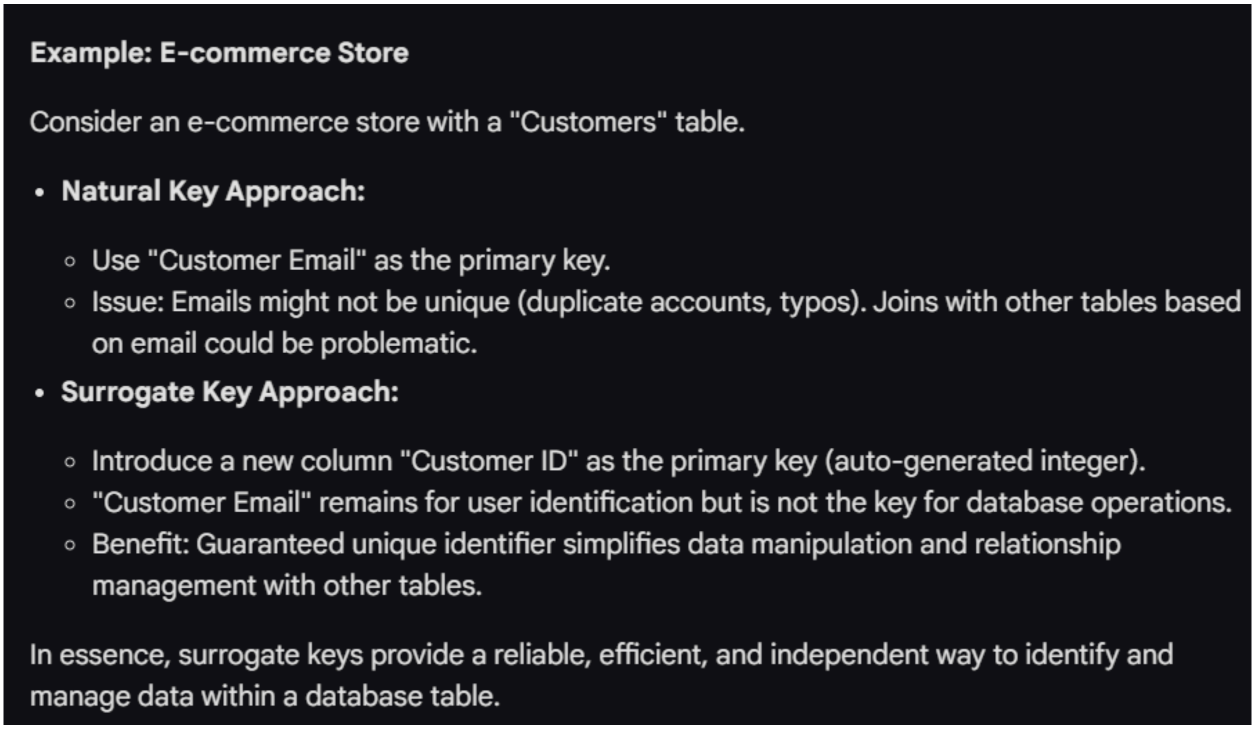
9.



A surrogate key is an artificial identifier assigned to each record in a database table. Unlike a natural key, which is derived from the data itself (e.g., customer email address), a surrogate key has no inherent meaning and serves the sole purpose of uniquely identifying a record.

Here's why surrogate keys are useful:

- **Reliable Uniqueness:** Natural keys might not always be unique. For instance, customer names can be duplicated. Surrogate keys guarantee uniqueness, essential for database operations like record retrieval and relationship management.
- **Data Changes:** If a natural key component changes (e.g., customer email update), all linked records need modification. Surrogate keys remain unaffected by data changes within the table, simplifying maintenance.
- **Performance:** Surrogate keys are often simple numbers (e.g., auto-incrementing integer) and can be efficiently indexed for faster data retrieval compared to potentially complex natural keys.

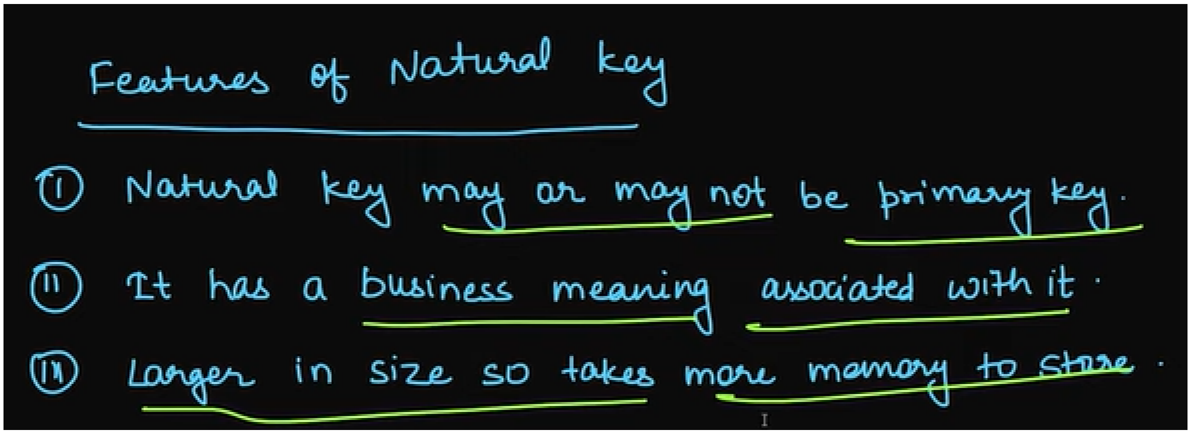


Example: E-commerce Store

Consider an e-commerce store with a "Customers" table.

- **Natural Key Approach:**
  - Use "Customer Email" as the primary key.
  - Issue: Emails might not be unique (duplicate accounts, typos). Joins with other tables based on email could be problematic.
- **Surrogate Key Approach:**
  - Introduce a new column "Customer ID" as the primary key (auto-generated integer).
  - "Customer Email" remains for user identification but is not the key for database operations.
  - Benefit: Guaranteed unique identifier simplifies data manipulation and relationship management with other tables.

In essence, surrogate keys provide a reliable, efficient, and independent way to identify and manage data within a database table.

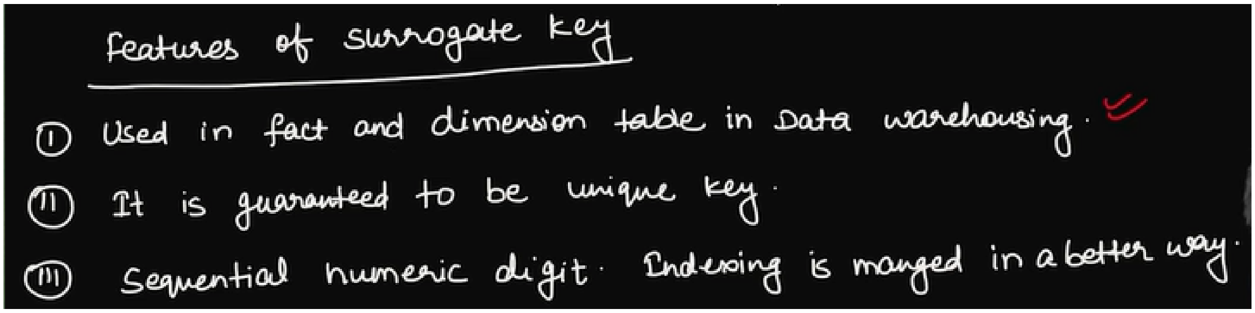
10. 

Features of Natural key

(I) Natural key may or may not be primary key.

(II) It has a business meaning associated with it.

(III) Larger in size so takes more memory to store.



Features of surrogate key

(I) Used in fact and dimension table in Data warehousing. ✓

(II) It is guaranteed to be unique key.

(III) Sequential numeric digit. Indexing is manged in a better way.