DATA WAREHOUSING AND MINING

T.E. CSE-DS, Sem V
Academic Year: 2023-24

Data Warehousing Fundamentals: Snowflake Schema, Aggregate Fact Table, Constellation Schema
Lecture 6

The Snowflake Schema

1. The snowflake schema is a variation of the star schema
2. The dimension tables are grouped into multiple tables instead of one large table (splitting the data into additional tables)
3. It normalizes its dimension tables to eliminate the redundancy.
4. Snowflake schemas are generally used when a dimension table becomes very big and star schema can not handle it.
5. You can fully or partially normalize (all/some) of the dimension tables
6. The attributes with low cardinality in each dimension table are removed to form separate tables. These new tables are linked back to the original dimension table through artificial keys

Snowflake Schema

Assume that there are 1,00,000 product dimension rows. These product fall under 100 brands and these brands fall under 10 products categories.

If the user runs a query constraining just on product category and if the product dimension table is not indexed on product category, the query will have to search through 100000 rows.

How does a query execute in Snowflake schema

The query to be executed is “ Fetch the data of quantity ordered of blue shirts that belongs to John Players brand in a sales transaction made on 2 Jan, 2006, by a salesperson who is from Chennai to Indian Customers”

Star Schema

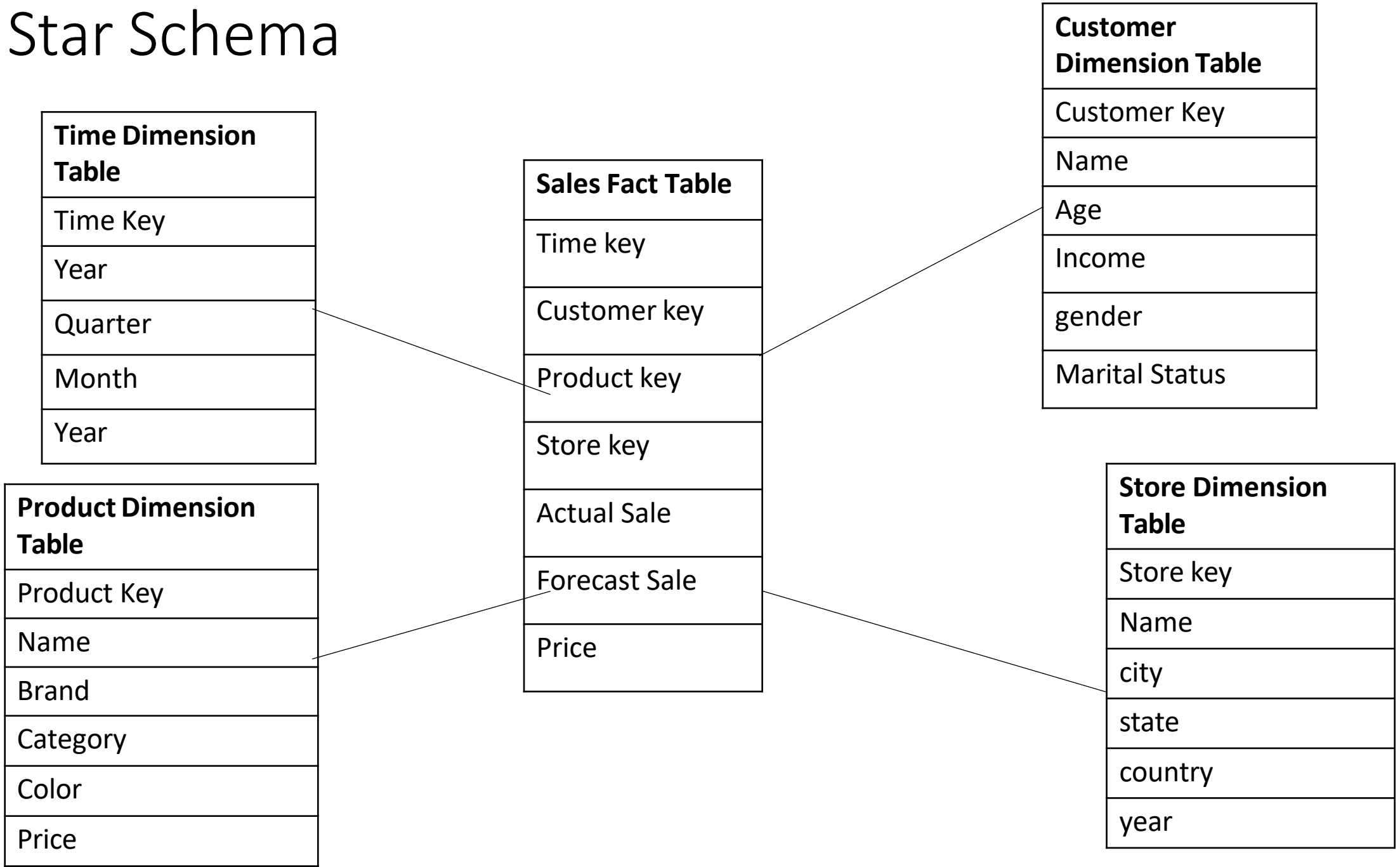
Time Dimension Table
Time Key
Year
Quarter
Month
Year
Product Dimension Table
Product Key
Name
Brand
Category
Color
Price
Sales Fact Table
Time key
Customer key
Product key
Store key
Actual Sale
Forecast Sale
Price
Customer Dimension Table
Customer Key
Name
Age
Income
gender
Marital Status
Store Dimension Table
Store key
Name
city
state
country
year

Table 6.2(a)
Brand dimension

Brand Key	Brand Name	Category Key
B01	John Players	CK01
B02	Peter England	CK02
B03	Allen Solly	CK01

Table 6.2(b)
Colour dimension

Colour Key	Colour Name
C001	Blue
C002	Black
C003	White
C004	Gray

Table 6.2(c)
Category dimension

Category Key	Category Name
CK01	Men's Formal
Ck02	Men's Casual

Table 6.2(d)
Product dimension

Product Key	Product Name Key	Brand Key	Colour Key
P001	Shirts	B01	C001
P002	Trousers	B01	C002
P003	T-shirts	B02	C003
P004	Shirts	B03	C00

Table 6.2(e)
Time dimension

Time Key	Year	Quarter	Month	Week	Date
T0001	2006	First	January	First	01
T0002	2006	First	January	First	02
T0003	2006	First	January	First	03
T0004	2006	First	January	First	04

How does a query execute in Snowflake schema

Now to understand query execution, break it into individual components

i)Blue shirt is the product

ii)John Players is the brand

iii)2nd Jan is the time

iv)Salesperson from Chennai

v)Indian customer

If we have all the primary keys of the dimension table we can fetch the desired rows from the fact table.

Table 6.2(a)
Brand dimension

Brand Key	Brand Name	Category Key
B01	John Players	CK01
B02	Peter England	CK02
B03	Allen Solly	CK01

Table 6.2(b)
Colour dimension

Colour Key	Colour Name
C001	Blue
C002	Black
C003	White
C004	Gray

Table 6.2(c)
Category dimension

Category Key	Category Name
CK01	Men's Formal
Ck02	Men's Casual

Table 6.2(d)
Product dimension

Product Key	Product Name Key	Brand Key	Colour Key
P001	Shirts	B01	C001
P002	Trousers	B01	C002
P003	T-shirts	B02	C003
P004	Shirts	B03	C00

Table 6.2(e)
Time dimension

Time Key	Year	Quarter	Month	Week	Date
T0001	2006	First	January	First	01
T0002	2006	First	January	First	02
T0003	2006	First	January	First	03
T0004	2006	First	January	First	04

Table 6.2(f)
Customer dimension

Customer Key	Name	Age	City	State	Country Key	Marital Status
C0001	Mary	25	XXXXXX	Delhi	CT01	Single
C0002	Joe	32	YYYYYY	Texas	CT02	Married
C0003	Ken	30	ZZZZZZ	Washington	CT02	Single
C0004	Jenny	43	AAAAAA	Punjab	CT01	Married

Table 6.2(g)
Country dimension

Country Key	Country Name
CT01	India
CT02	USA

Table 6.2(h)
Salesperson dimension

Sales Person Key	Salesperson Name	City Key
SP001	Neelam	CI01
SP002	Ram	CI02
SP003	Ishaan	CI03

Table 6.2(i)
City dimension

City Key	City Name
CI01	Chandigarh
CI02	Mumbai
CI03	Chennai

Sales fact table

Time Key	Product Key	Customer Key	Salesperson Key	Sales Amount	Sales Qunatity	Profit
T0001	P001	C0001	S0002	200	20	30
T0002	P003	C0002	S0003	180	17	35
T0003	P002	C0003	SP004	150	15	27
T0002	P001	C0004	SP001	220	18	20
T0002	P001	C0001	SP003	300	10	50

How does a query execute in Snowflake schema

Now we have:
Time key=T0002
Product key=P0001
Customer key=C001 and C0002
Salesperson key=SP003

Resultant row from the fact table

Sales fact table

Time Key	Product Key	Customer Key	Salesperson Key	Sales Amount	Sales Qunatity	Profit
T0001	P001	C0001	S0002	200	20	30
T0002	P003	C0002	S0003	180	17	35
T0003	P002	C0003	SP004	150	15	27
T0002	P001	C0004	SP001	220	18	20
T0002	P001	C0001	SP003	300	10	50

Pros of Snowflake Schema

1. If a dimension has a very large list of attributes and attributes having the hierarchy's the snowflake model is appropriate
2. Detailed data is available
3. Reduce data redundancy

Cons of Snowflake Schema

1. Reduce the effectiveness of browsing since more joins will be needed to execute a query
2. Not that much easy to understand for end user because of its complex structure
3. Navigation in complex structure is difficult
4. Query performance degrades

(Generally snowflake schema is not preferred to use in DW environment)