Chapter 1

Introduction to Data warehouse and Dimensional Modeling

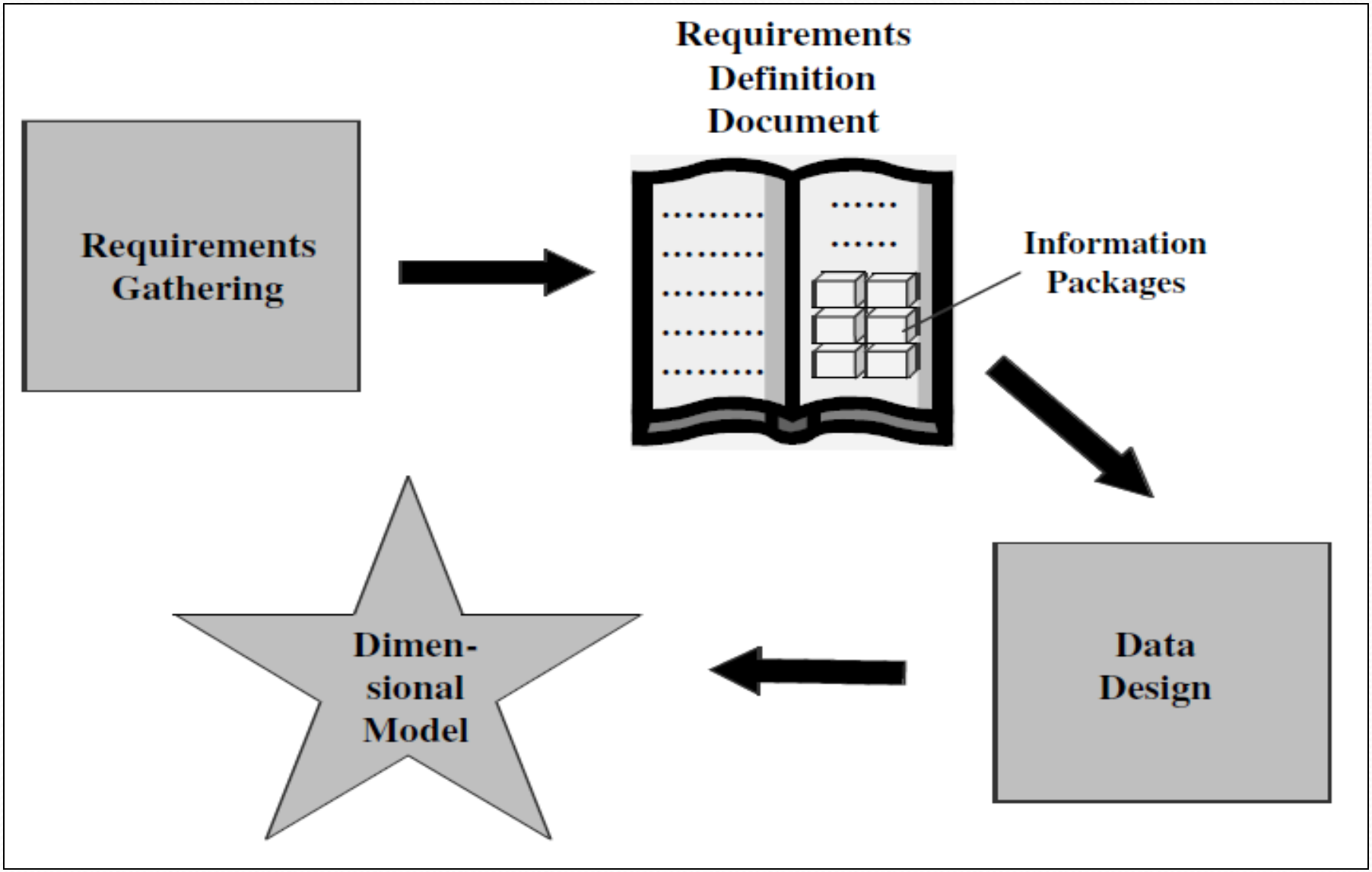
Based on CO1: Identify applications which require data warehouse and select the suitable architecture required for any data ware house applications

By, Safa Hamdare

Outline: Introduction to Dimensional Modeling

- E-R modeling versus Dimensional Modeling
- Information Package Diagram (IPD)
- Contents of fact/dimension tables
- Dimensional Model:
  - 1. Basics of Star Schema + Advantage & Disadvantage
- Advance Concept
  - Slowly changing dimensions
  - Large dimensions
  - Basics of Snowflake Schema + Advantage & Disadvantage
  - Aggregate tables
  - Fact Constellation Schema or Families of Star

Requirements to Design



Design decisions to be taken

1. Choosing the process-deciding subjects
2. Choosing the grain- level of data
3. Identifying and confirming dimensions
4. Choosing the facts
5. Choosing the duration of the database

Dimensional Modeling

1. What is Dimensional Modeling? Explain in detail. (Dec 2011)
2. A Dimension table is wide; the fact table is deep. Explain. (5 m, Dec 2010, Dec 2019)
3. Why is entity-relationship modeling technique not suitable for the data warehouse? How is dimensional modeling different? (10 m, Dec 2019)

Dimensional Modeling: (Dec 2011)

- Dimensional Modeling is the name of a logical design technique used for data warehouse
- The ***objective of dimensional modeling*** is to represent a set of business measurements in a standard framework that is easily understandable by end users.
- A Dimensional Model contains the same information as an ER model but packages the data in a symmetric format whose design goals are:
  - User Understandability
  - Query Performance
  - Resilience to change

ER Model v/s Dimension Model

- ER diagram is a complex diagram, used to represent multiple processes. A single ER diagram can be broken down into several DM diagrams.
- In DM, we prefer keeping the tables de-normalized, whereas in a ER diagram, our main aim is to remove redundancy
- ER model is designed to express microscopic relationships between elements. DM captures the business measures
- DM is designed to answer queries on business process, whereas the ER model is designed to record the business processes via their transactions.

Entity-Relationship vs. Dimensional Models

E-R DIAGRAM	DIMENSIONAL MODEL
• One table per entity • Minimize data redundancy • Optimize update • The Transaction Processing Model	• One fact table for data organization • Maximize understandability • Optimized for retrieval • The data warehousing model

Converting an E-R Diagram into a Data Warehouse

- Determine the purpose of the mart *(Business Scenario)*
- Identify an association table as the central fact table
- Determine facts to be included
- Replace all keys with surrogate keys and put it in fact table
- Promote foreign keys in related tables to the fact table
- Add time dimension
- Refine the dimension tables

Dimensional Model Components:

1. Facts & Dimension
2. Information Package Diagram
3. Dimension Table
4. Fact Table

Facts & Dimensions

- Numeric measurements (values) that represent a specific business aspect or activity.
- Can be computed or derived at run time
- Updated periodically with data from operational databases
- Stored in a fact table (at the center of the star scheme)
- A Fact depends on many factors.

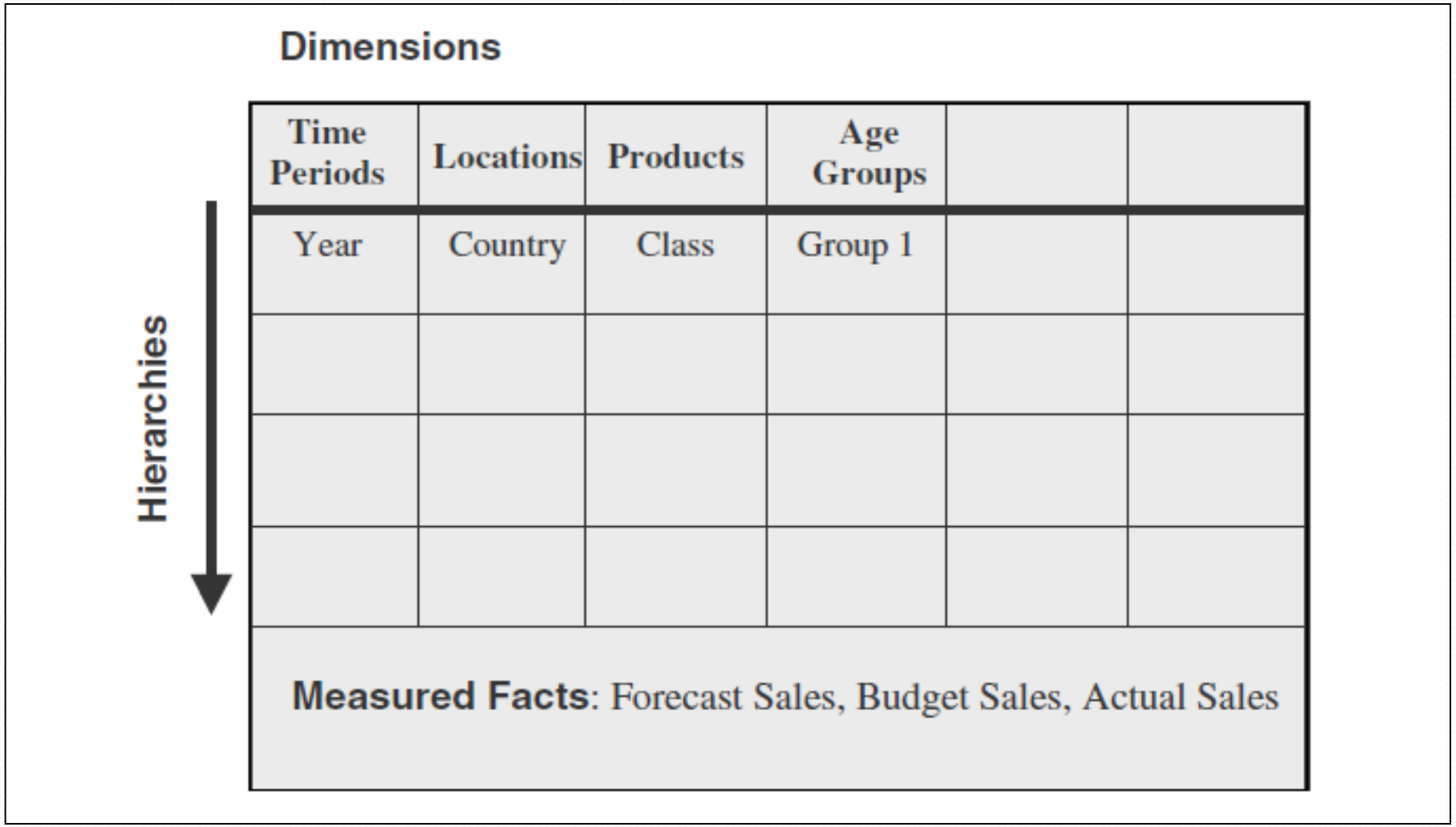
For E.g., ***Fact****- Sales_amount of month, depends on* ***factors****- product, location and time.*

These factors are known as ***dimensions*** (on which a given fact depends).

Information Package Diagram (IPD): **(Dec 2010)**

- Information package diagram is the approach to determine the requirement of data warehouse.
- It gives the metrics which specifies the business units and business dimensions.
- *The information package diagram defines the relationship between the subject or **dimension** matter and key performance measures (**facts**).*
- The information package diagram shows the details that users want so its effective for communication between the user and technical staff.

An information Package Diagram



Dimensional modeling basics: IPD

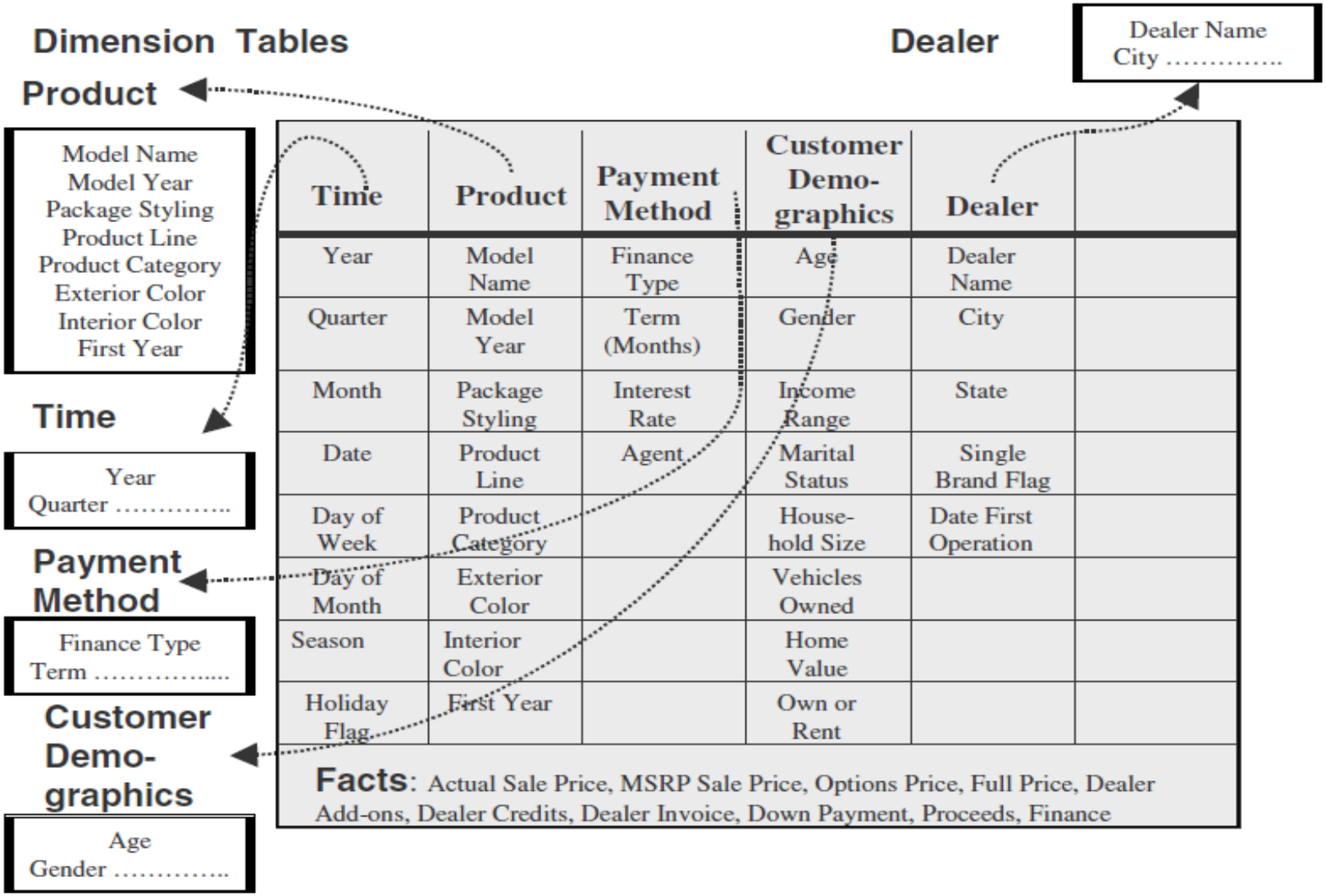
(Automobile maker sales)

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		
Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance					

Dimension table

- ***Dimensions*** are factors on which a given fact depends.
- The *Sales_amount fact* can be thought of as a function of 3 variables *{ f (product, location, time)}*
- Dimension table contains information about a particular dimension, like:
  1. Dimension table key
  2. Table is wide
  3. Textual attributes
  4. Attributes not directly related
  5. Not normalized
  6. Drilling down, rolling up
  7. Multiple hierarchies
  8. Fewer number of records

Formation of the automaker dimension tables



How much sales did the jeep tata mahindra, 2005 model with vxi options, generate in january 2000 at spectra auto dealership for buyers who owned their homes, financed by icici prudential financing?

Fact table

- A ***Fact table*** is the primary table in each dimensional model that is meant to contain measurements of the business.
- Contains primary information of the warehouse
  1. Concatenated key (from all Dimensional Table)
  2. Data grain
  3. Fully additive measures
  4. Semi-additive measures(derived attributes)
  5. Table deep, not wide
  6. Degenerate dimensions(attributes which are neither fact or a dimension)

Formation of the automaker sales fact table

+ Includes Key of each Dimensions

Automaker Sales

Fact Table

Actual Sale Price
MSRP Sale Price
Options Price
Full Price
Dealer Add-ons
Dealer Credits
Dealer Invoice
Down Payment
Proceeds
Finance

Dimensions

Time	Product	Payment Method	Customer Demo-graphics	Dealer	
Year	Model Name	Finance Type	Age	Dealer Name	
Quarter	Model Year	Term (Months)	Gender	City	
Month	Package Styling	Interest Rate	Income Range	State	
Date	Product Line	Agent	Marital Status	Single Brand Flag	
Day of Week	Product Category		House-hold Size	Date First Operation	
Day of Month	Exterior Color		Vehicles Owned		
Season	Interior Color		Home Value		
Holiday Flag	First Year		Own or Rent		

Facts: Actual Sale Price, MSRP Sale Price, Options Price, Full Price, Dealer Add-ons, Dealer Credits, Dealer Invoice, Down Payment, Proceeds, Finance

Exercise: Design Problems on IPD, Fact and dimension table

(a) The college wants to record the grades for the courses completed by students. There are four dimensions :—

(i) Course
(ii) Professor
(iii) Student
(iv) Period.

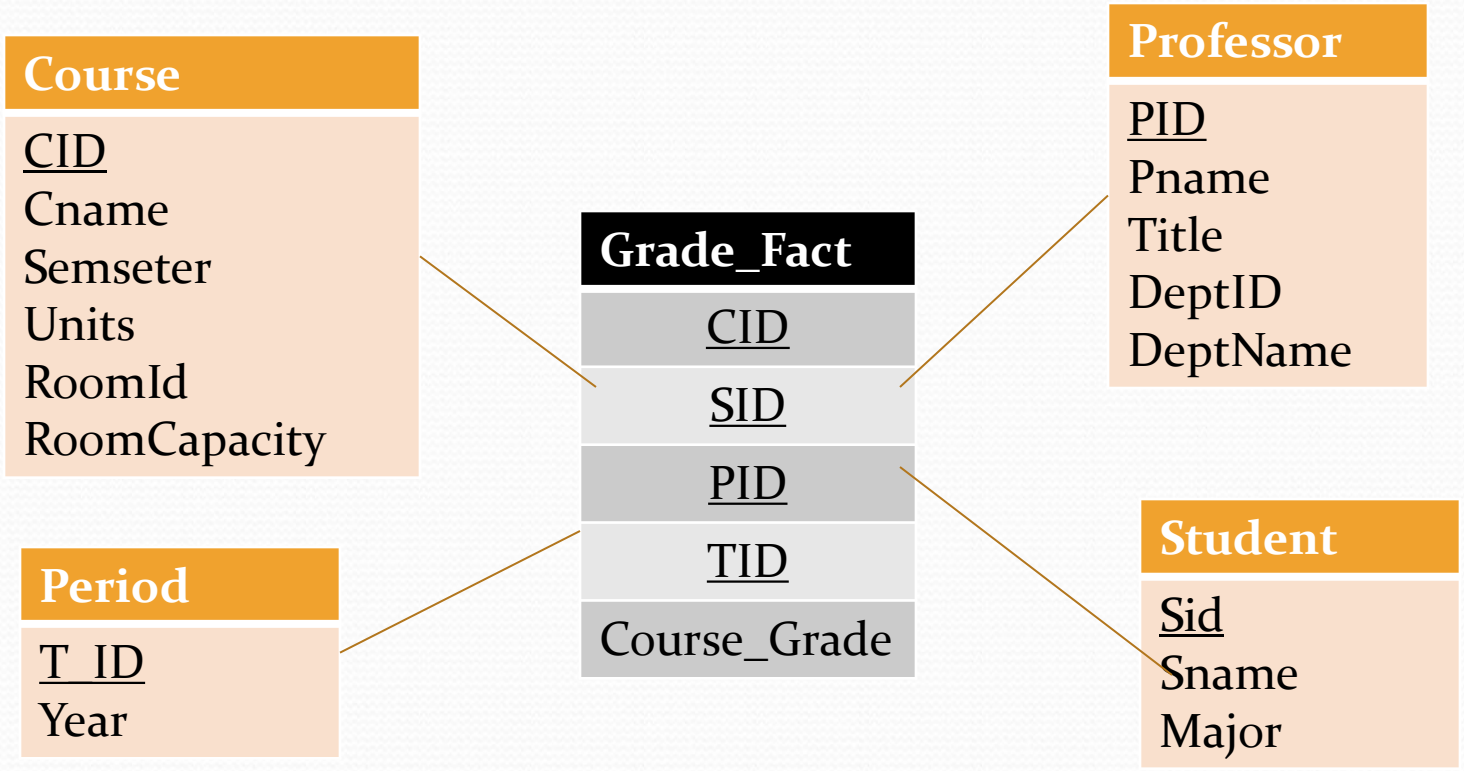
The only fact that is to be recorded in the table is course-grade :–

(i) Design star schema.
(ii) Write DMQL for the above star schema.

Exercise: IPD

Period	Course	Professor	Student
T_ID Year	CID Cname Semseter Units RoomId RoomCapacity	PID Pname Title DeptID DeptName	Sid Sname Major
Fact: Course_grade			

Exercise: Fact table and Dimension table



Dimensional Models (Dec 2011)

- A DE normalized relational model
  - Made up of tables with attributes
  - Relationships defined by keys and foreign keys
- Organized for understandability and ease of reporting rather than update
- Queried and maintained by SQL or special purpose management tools.
- Every Dimensional model is composed of at least one table with a multipart key called the ***fact table*** and a set of smaller tables called ***dimension table***.
- Fact tables are normalized, whereas dimension tables are not.
- Fact tables are very large as compared to dimension tables.

A Dimension table is wide; the fact table is deep. Explain. **(Dec 2010)**

1. Dimension table has got all the detail information's of their respective table
   - For e.g. customer dimension table will contain all the related info about customers
   - whereas fact table contains the main data, which contains the surrogate keys of every dimensions along with other measures.
2. A dimension table contains a higher granular information so have less no of records and it needs to have all the necessary details (more columns) related to the grain of the table.

- On the other side A fact table has the lowest level grain of a subject area. Lower grain cause more number of rows in the fact Table.

Dimensional Model

-Star Schema

Dimensional Model: **Star Schema**

- The most popular schema design for data warehouse is the *Star Schema*.
- Each dimension is stored in a *dimension table* and each entry is given its unique identifier i.e. Primary key.
- The *dimension tables* are related to one or more *fact tables.*
- The fact table contains a composite key made up of the identifiers(primary key) called as surrogate key from the dimension tables.
- Fact table has foreign keys to all dimension tables in a star schema.

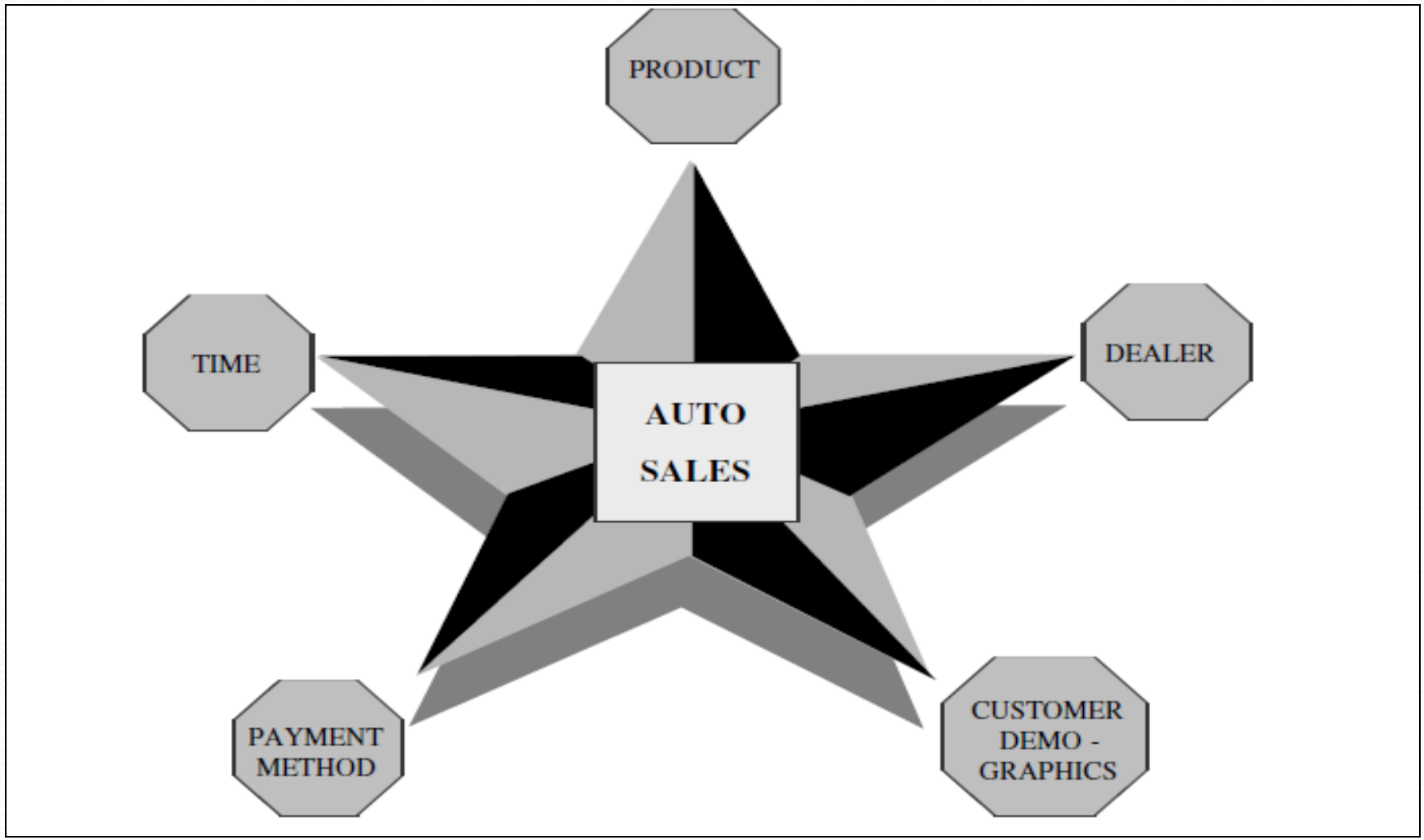
Star Schema characteristics:

1. Star schema is a relational model with one-to-many relationship between the fact table and the dimension tables.
2. De-normalized relational model.
3. Easy to understand. Reflects how users think. This makes it easy for them to query and analyze the data.

Advantages:

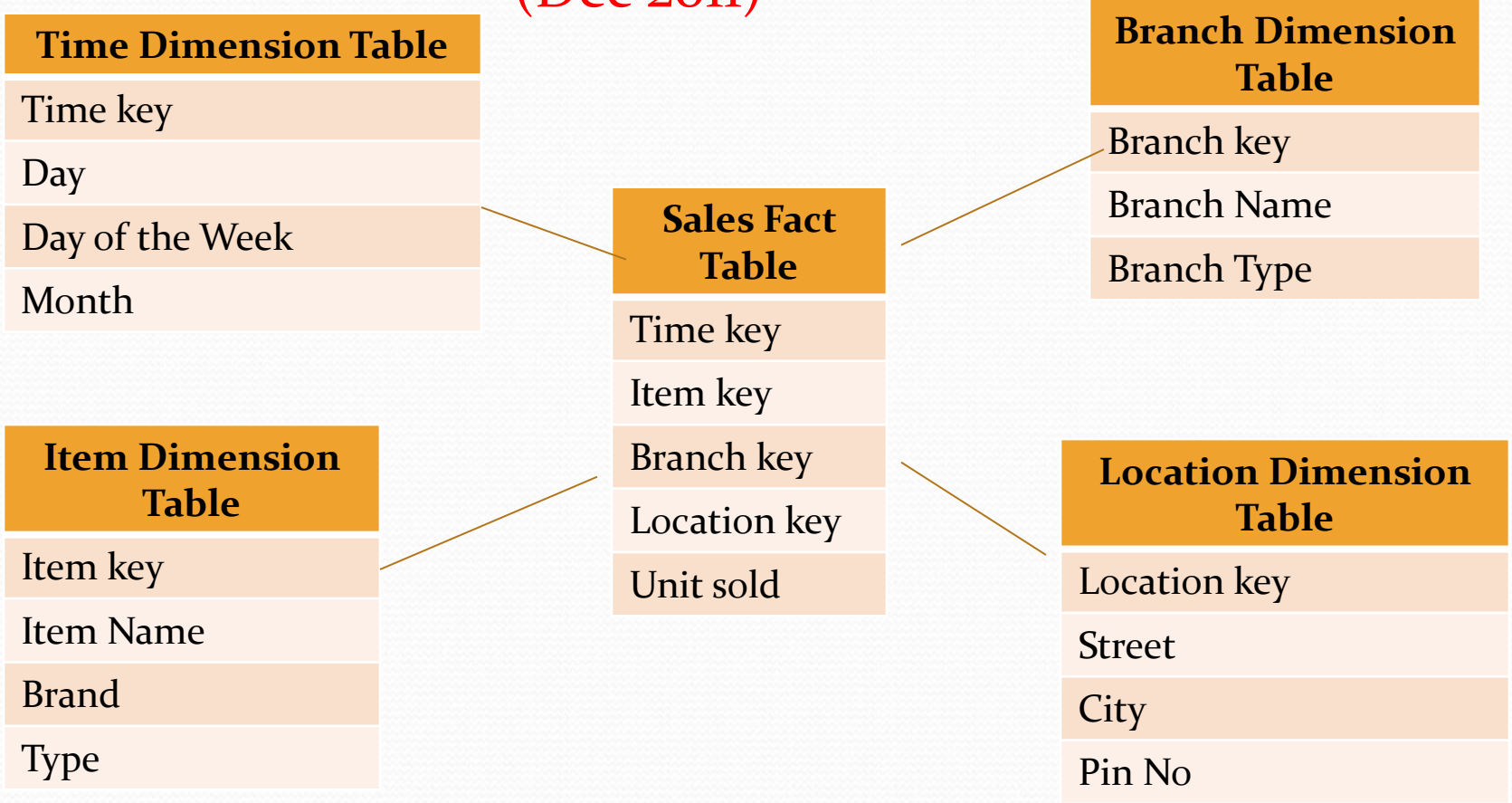
1. Easy for users to understand
2. Optimizes navigation.
3. Enhances query extraction.
4. Most suitable for query processing- Ability to drill down or roll up.

STAR SCHEMA for automaker sales (Dec 2011)



Design Problems on Star Schema

1. Design Star Schema for Auto sales analysis of company. (Dec 2011)



Design Problems on Star Schema

(a) The college wants to record the grades for the courses completed by students. There are four dimensions :—

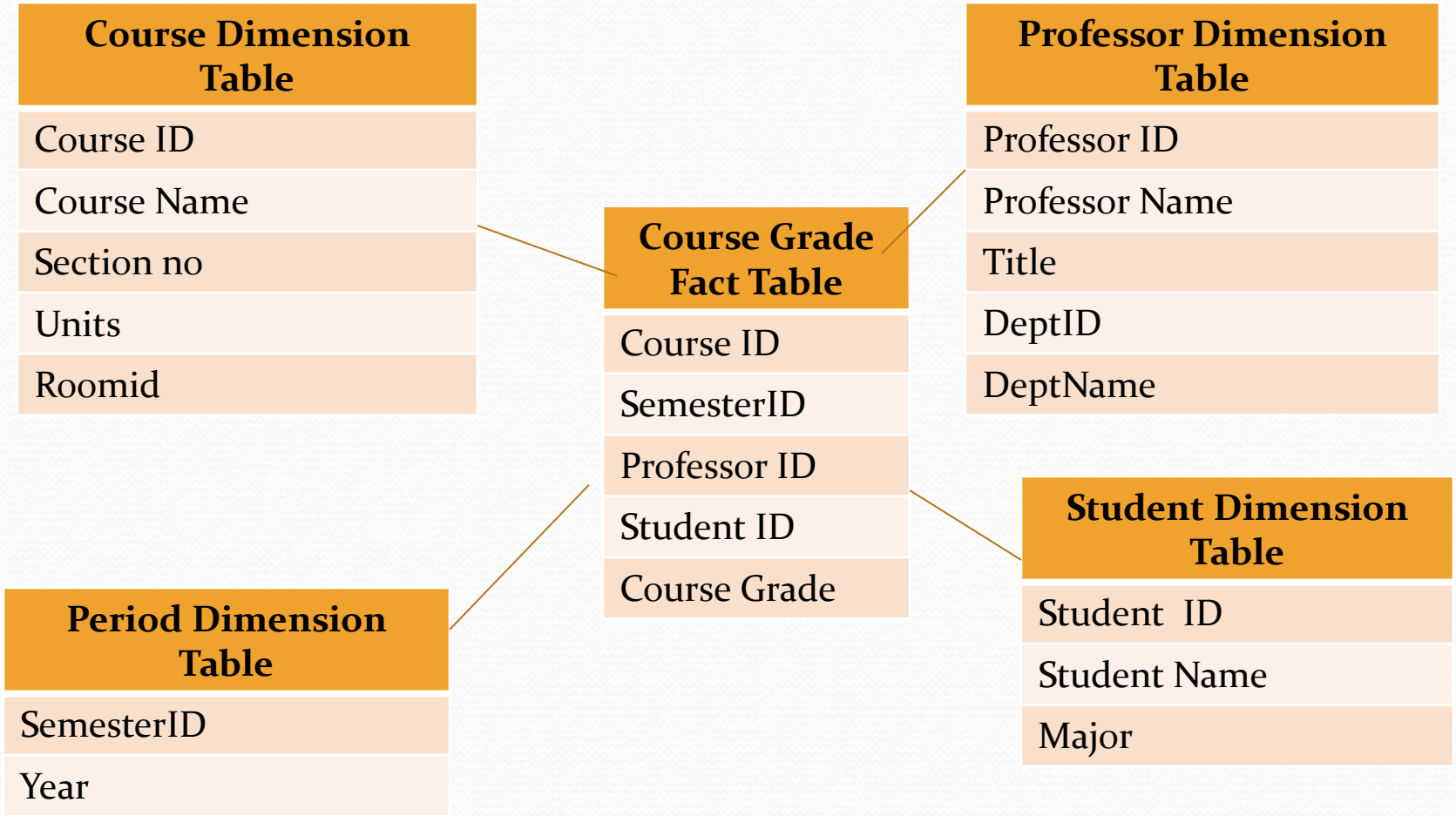
(i) Course
(ii) Professor
(iii) Student
(iv) Period.

The only fact that is to be recorded in the table is course-grade :–

(i) Design star schema.
(ii) Write DMQL for the above star schema.

Solution for college Star Schema

2. Design Star Schema for College. (May 2011)



May 2016

For a Super market chain, consider the following dimensions namely product, store, time and promotion. The schema contains a central fact table for sales.

i. Design star schema for the above application.

ii. Calculate the maximum number of base fact table records for warehouse with the following values given below:

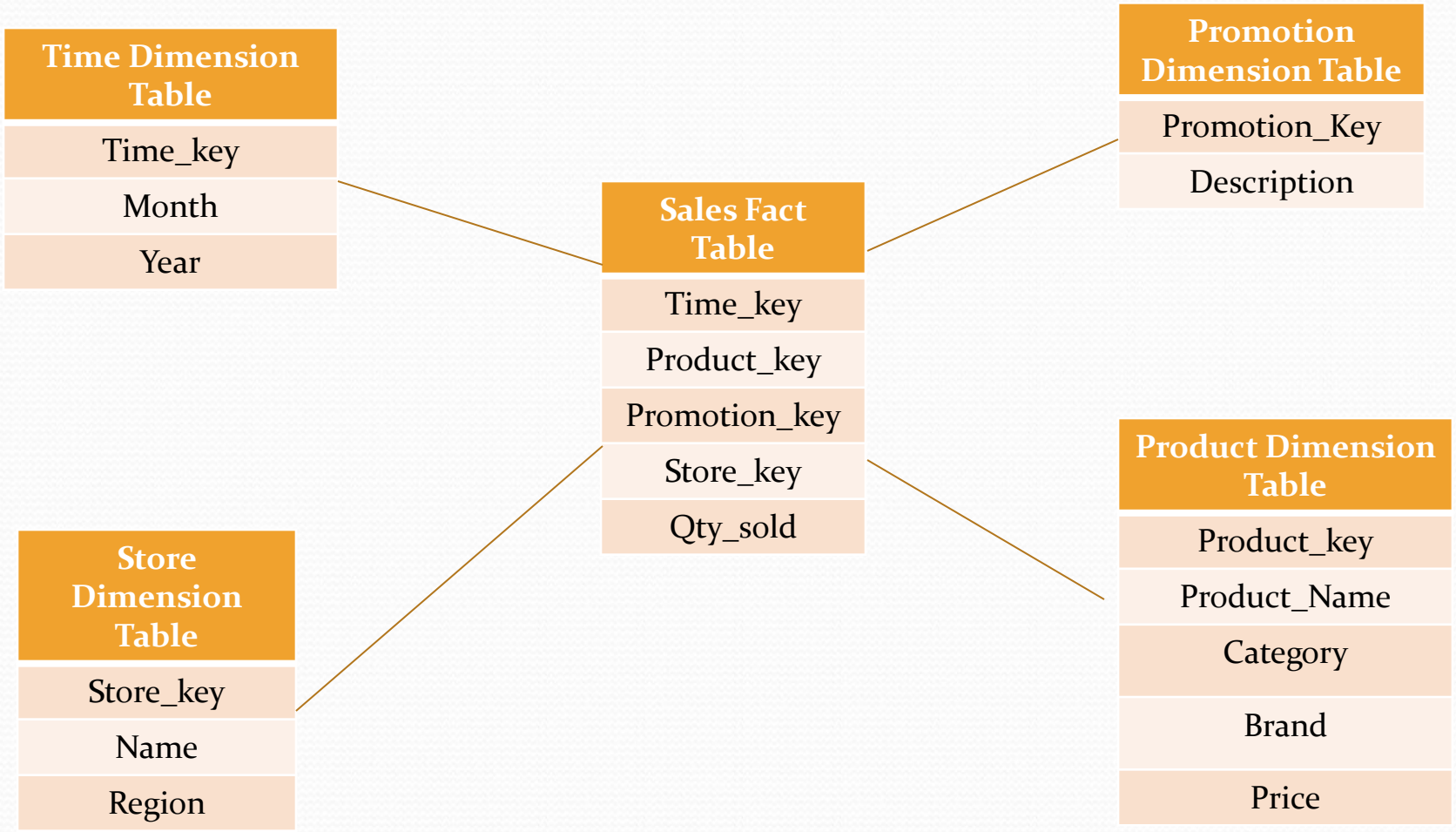
- Time period – 5 years
- Store – 300 stores reporting daily sales
- Product – 40,000 products in each store (about 4000 sell in each store daily)

May- June 2013 (5 Marks)

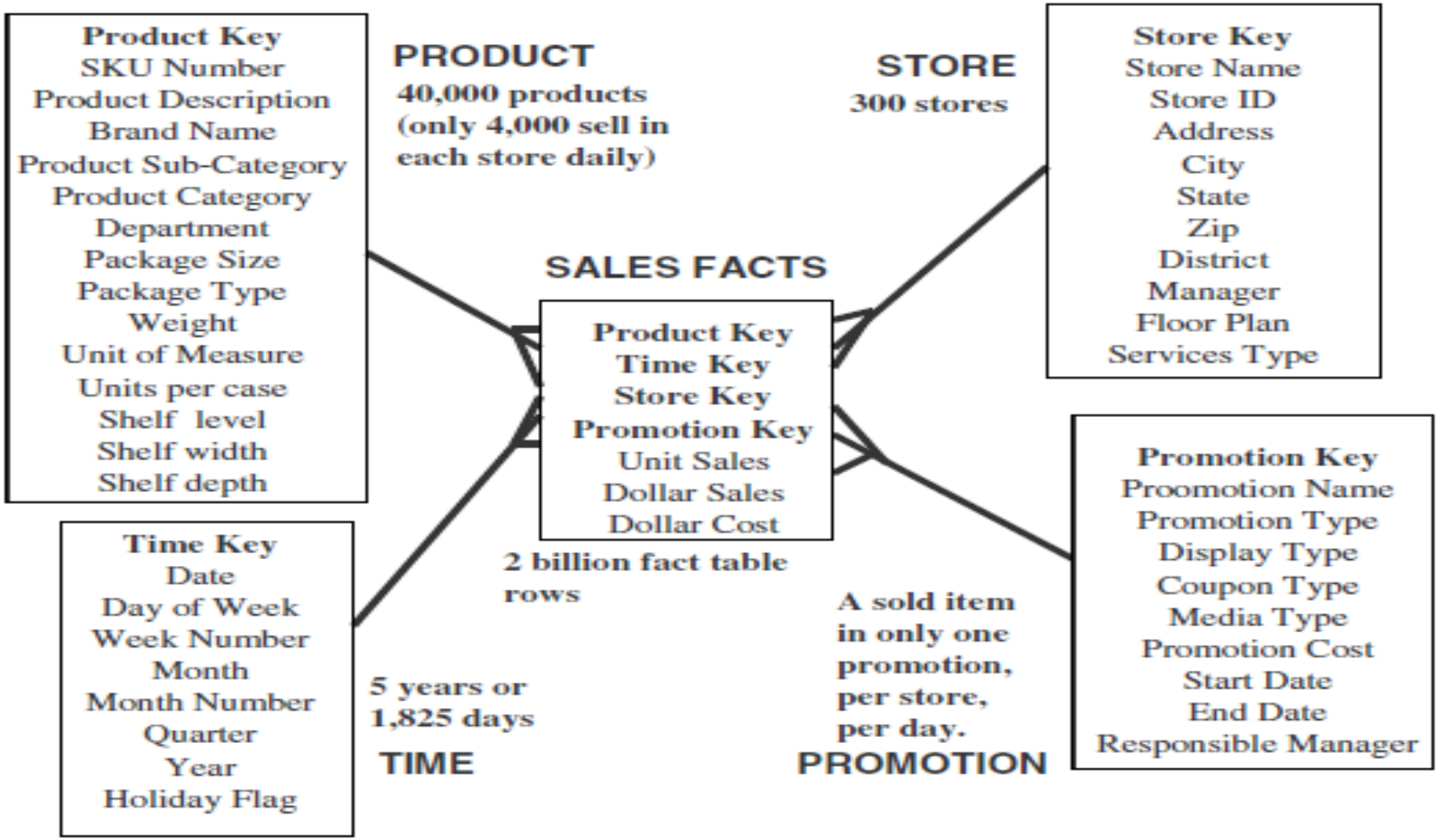
Calculate the maximum number of base fact table records for warehouse with the following values given below :

- Time period: 5 years
- Store: 300 stores reporting daily sales
- Product: 40,000 products in each store (about 4000 sell in each store daily)

2. Star schema for Supermarket Sales



Fact Table Size



Maximum number of base fact table records: 1825 x 300 x 4000 x 1 = 2 billion

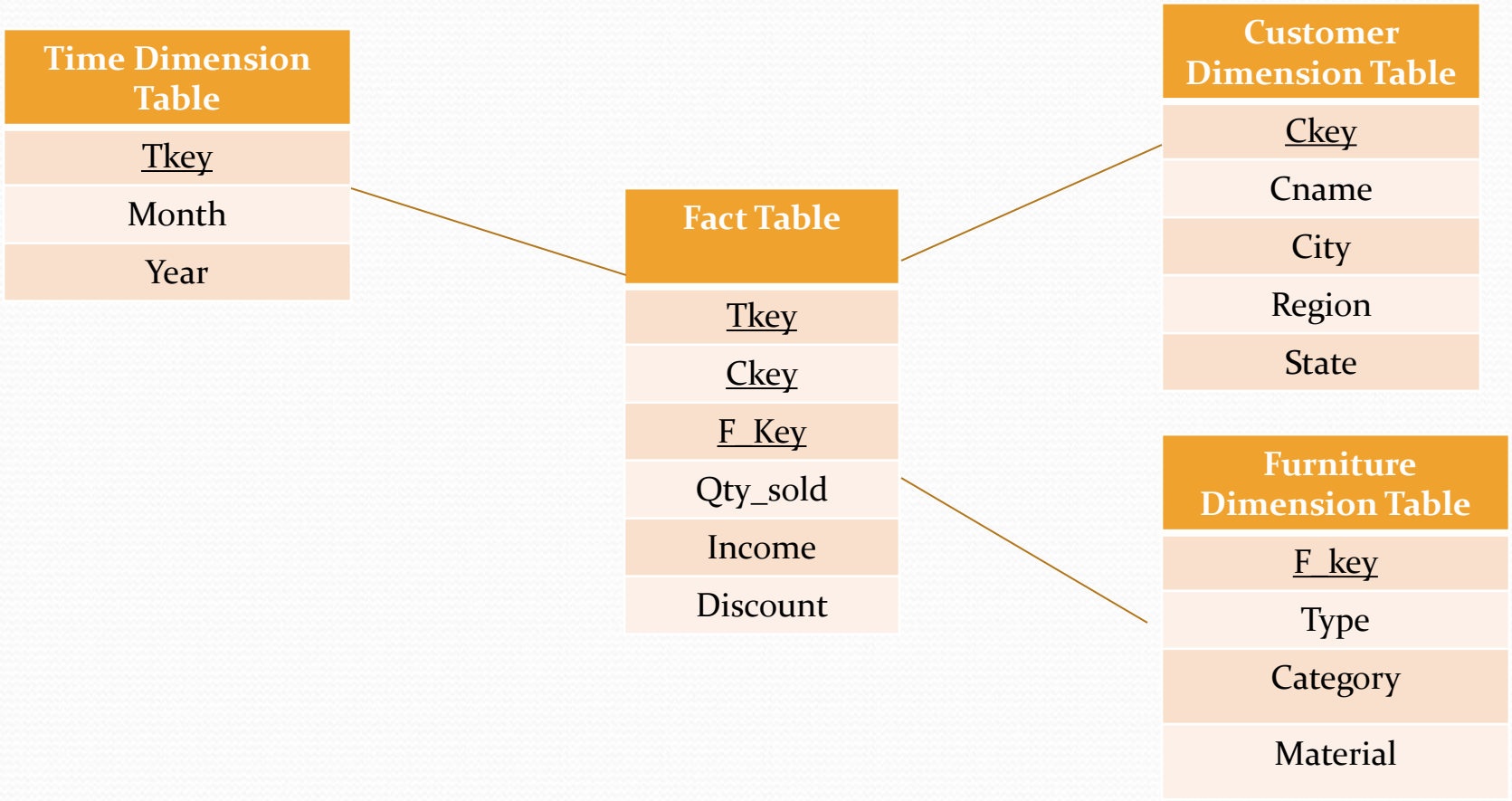
May 2017

What is dimensional modelling? Design the data warehouse for wholesale furniture Company. The data warehouse has to allow analysing the company's situation at least with respect to the Furniture, Customer and Time. More ever, the company needs to analyse: The furniture with respect to its type, category and material. The customers with respect to their spatial location, by considering at least cities, regions and states. The company is interested in learning the quantity, income and discount of its sales.

Dec 2016, Dec 2017

Consider following dimensions for a Hypermarket chain: Product, Store, Time and Promotion. With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures).

Star schema Example Wholesale furniture company



May 2018

i. Design star & snowflake schema for "Hotel Occupancy" considering dimensions like Time, Hotel, Room, etc.

ii. Calculate the maximum number of base fact table records for the values given below:

Time period: 5 years

Hotels: 150

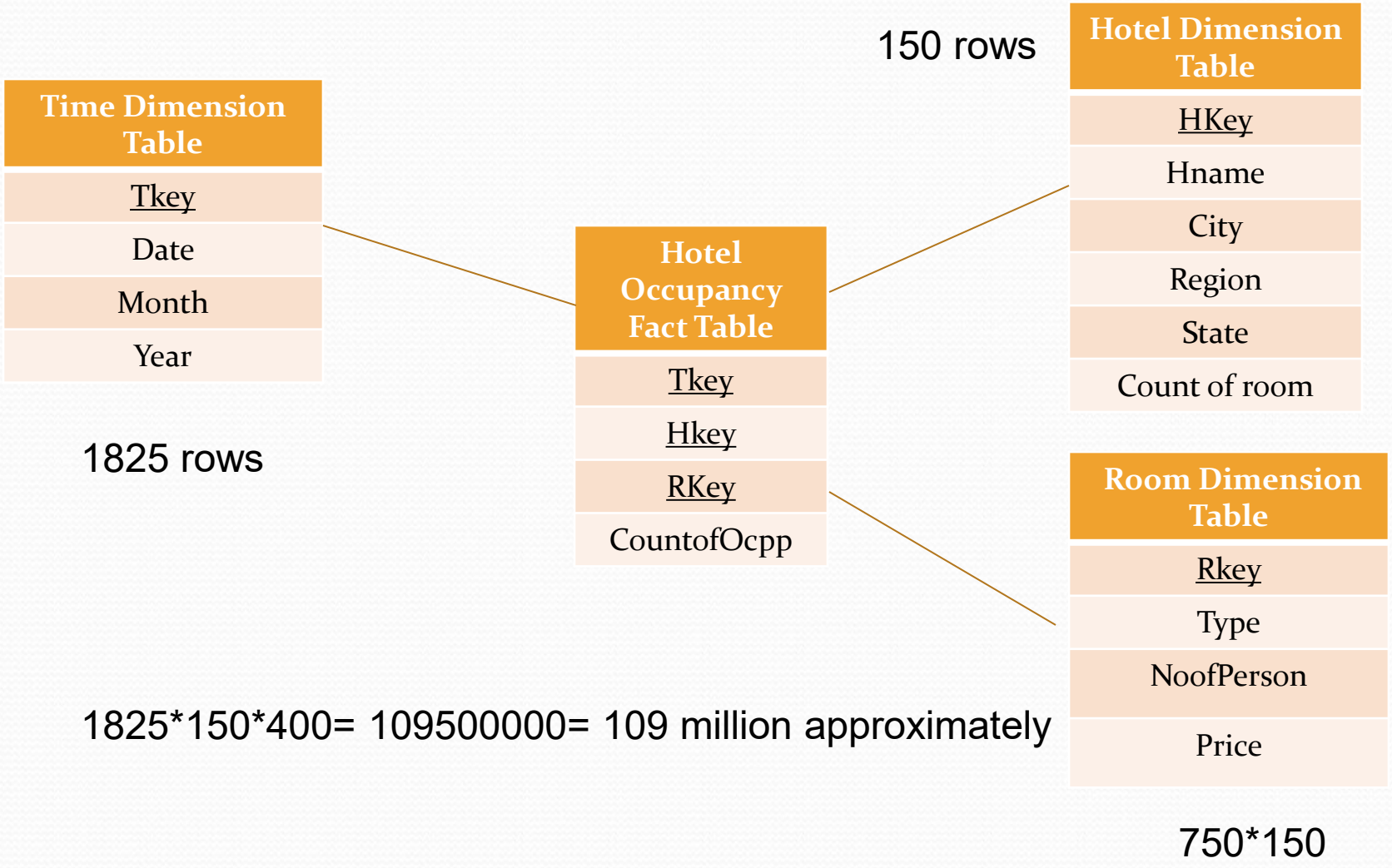
Rooms: 750 rooms in each Hotel (about 400 occupied in each hotel daily).

May 2019

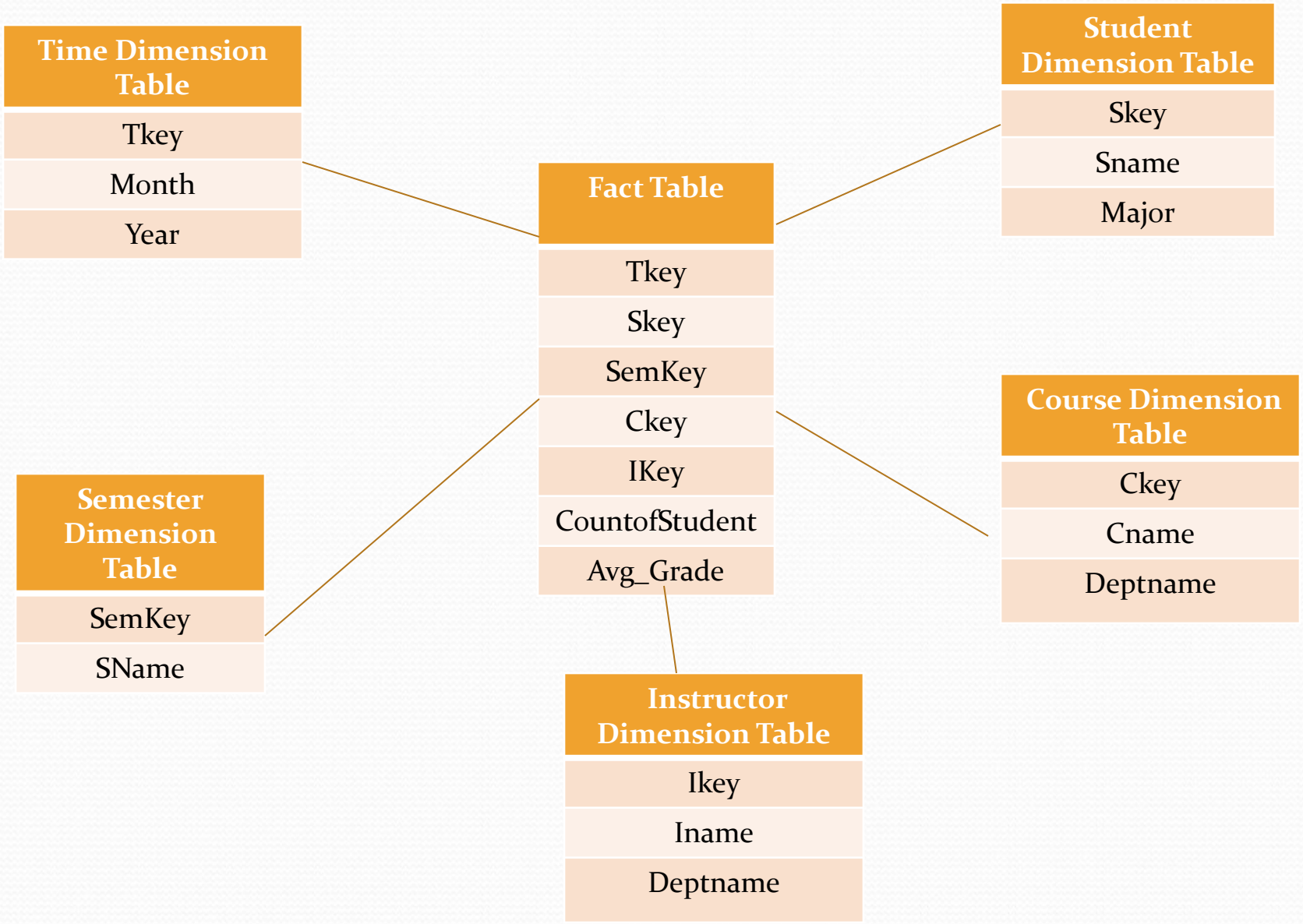
Suppose that a data warehouse for *DB-University* consists of the four dimensions *student, course, semester,* and *instructor*, and two measures count and *avg-grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg-grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg-grade* stores the average grade for the given combination. **[10]**

i. Draw a *snowflake schema* diagram for the data warehouse.

Star schema Example for hotel



Star schema Example for University

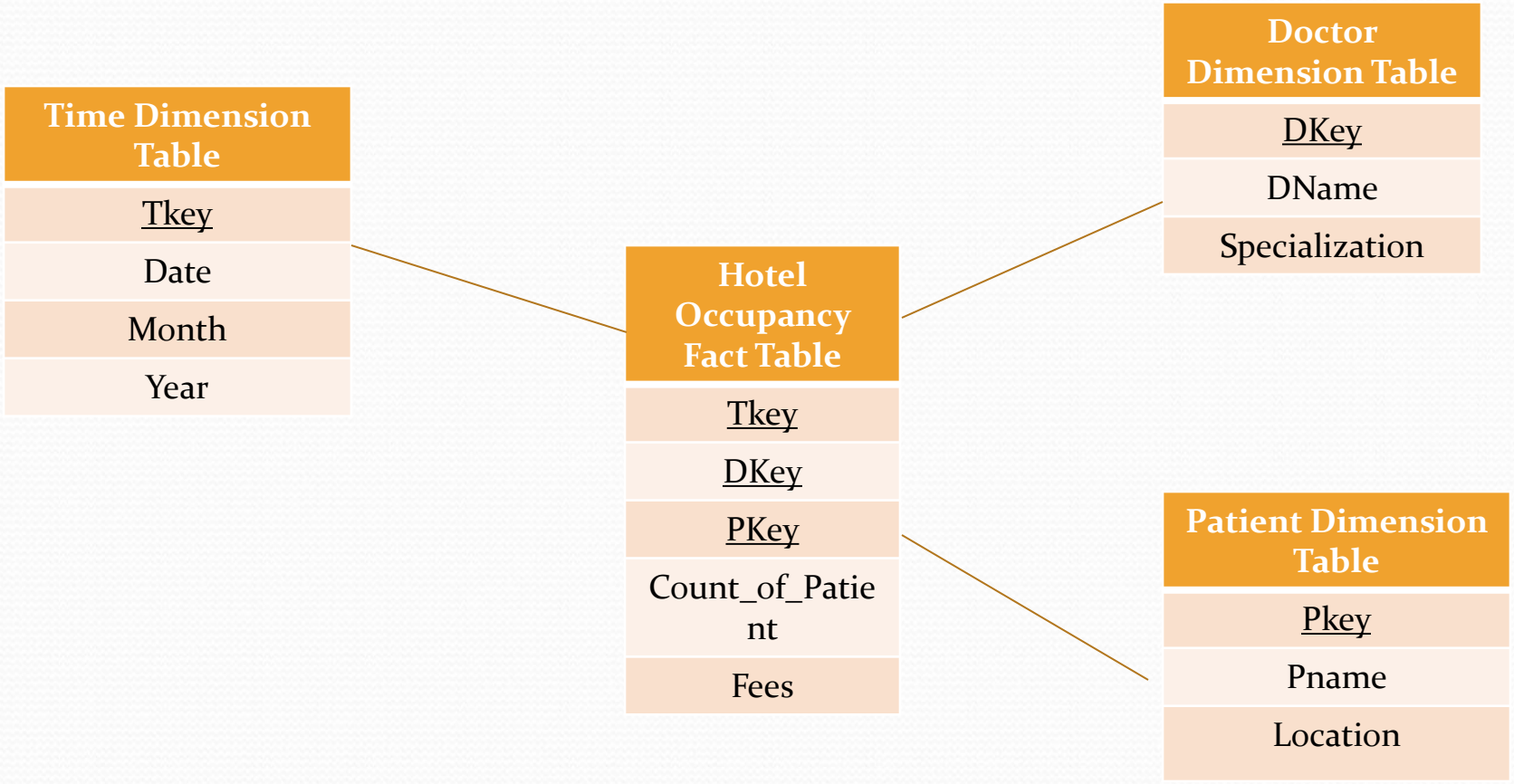


Dec 2019

a) Suppose that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

(i) Draw a star schema diagram for the above data warehouse.

Star schema Example for Doctor



Fact-less Fact Table

1. Define Fact less Fact tables with suitable example. (May 2011,Dec 2012)
2. Write short note on Fact less Fact table. (Dec 2016)

Factless Fact table

- A fact table is said to be empty if it has no measures to be displayed. Fact table represents **events**
- Contains no data, only keys which is used to represent an event which is identified by the combination of conditions referenced in the dimension tables.
- Used only to put relation between the elements of various dimensions.
- Often used to represent ***many to many relationships***.

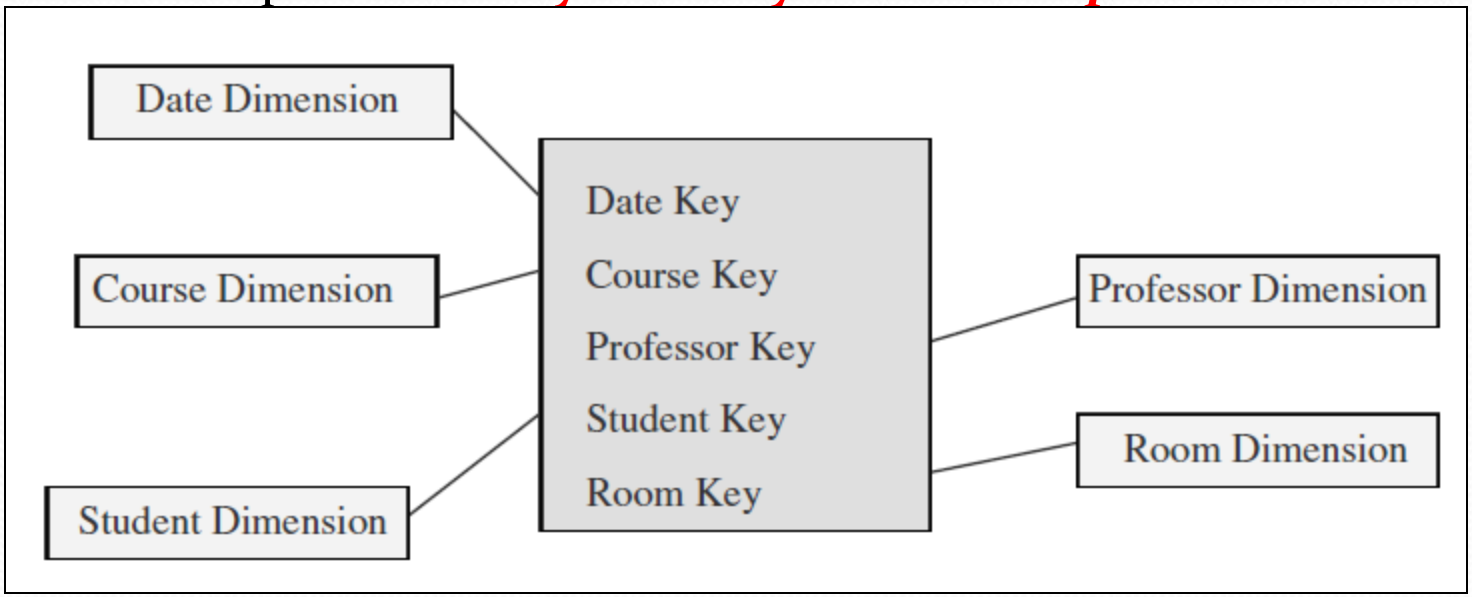


Fig: Example of Factless Fact

Example of Factless Fact Table:

- **Event Tracking Tables:** Use a Factless fact table to track events of interest to the organization .
  - **For Example**, ***attendance at a cultural event*** can be tracked by creating a fact table containing the following Foreign Keys (event p.k, Entertainer p.k, participant p.k, date key) i.e. links to all the dimension tables.This table then can be queried to find out information, such as which cultural events are the most popular among the participants.
  - *Which class has the maximum attendance? , What is average number of attendance of a given course?*
- **Coverage Tables:** It is used to support negative analysis report.
  - **For Example,** a ***store that did not sell a product*** for a given period. To produce such report, we need to have a fact table to capture all the possible combinations. We can then figure out what is missing.
  - *Ex-Visitors to the office, List of people for the web click, tracking student attendance or registration events.*

Dimensional modeling: Advanced Concepts

In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem. (May 2019 5marks)

Objectives

- Updates to dimensional table: (May 2016, May 2017)
  - Slowly changing dimensions & its Types
- Large dimensions
- Snowflake schema
- Aggregate tables
- Fact Constellation Schema or Families of Star. (Dec 2017)

Slowly Changing Dimensions

- Dimension tables are non-volatile and mostly read-only.
- More rows are added to the Dimension tables over time.
- Most dimensions are generally constant over time.
- *Many dimensions change slowly*
- Though the key does not change other description and attributes change slowly over time
- Dimension table attributes are not overwritten

Updating the Dimension table:

(May 2016, May 2017)

- Changes to certain attributes of a row become important at times.
- There are many types of changes that affect the dimension tables based on what information must be preserved.
- **Types of Changes:**
  1. Type 1 Change: Correction of Errors
  2. Type 2 Change: Preservation of History
  3. Type 3 Change: Tentative Soft Revision

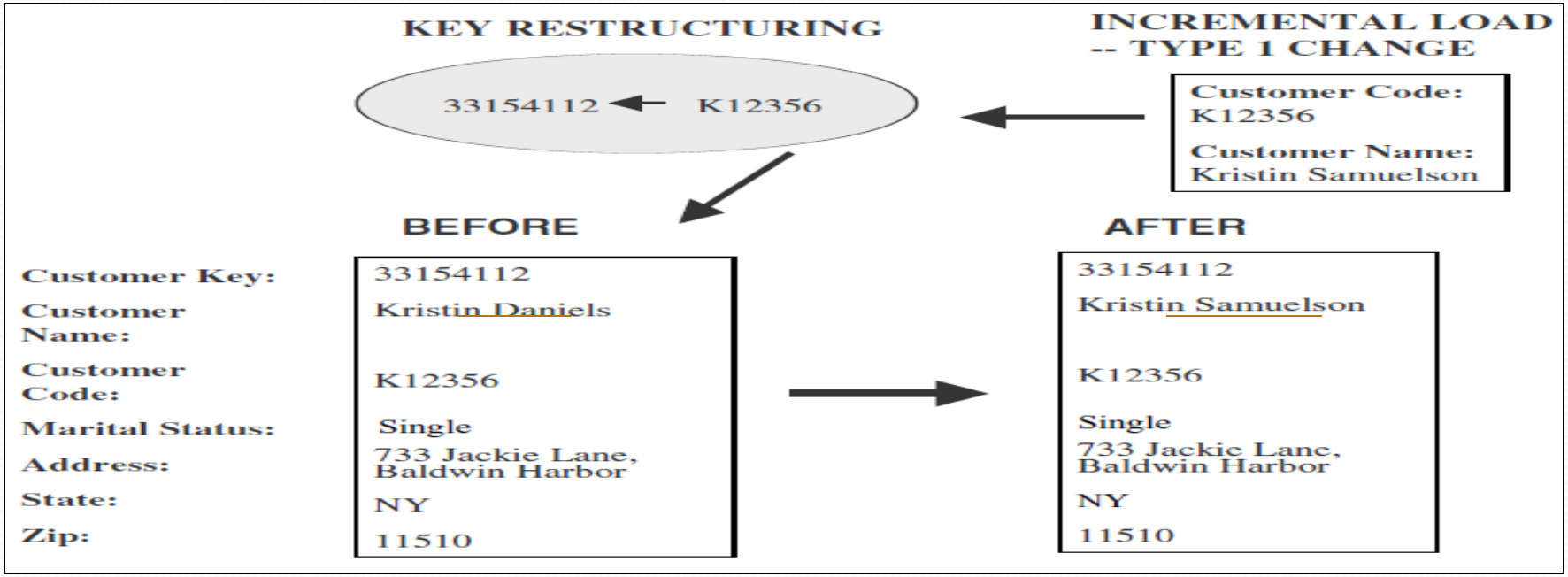
Type 1 Change: Correction of Errors

- Usually relate to correction of errors in the source systems.
- *E.g., spelling error in customer names; change of names of customers;*
- There is no need to preserve the old values here.
- The old value in the source system needs to be discarded.
- The changes made **need not be preserved** or noted.

Values are overwritten here..

Type 1.. Example

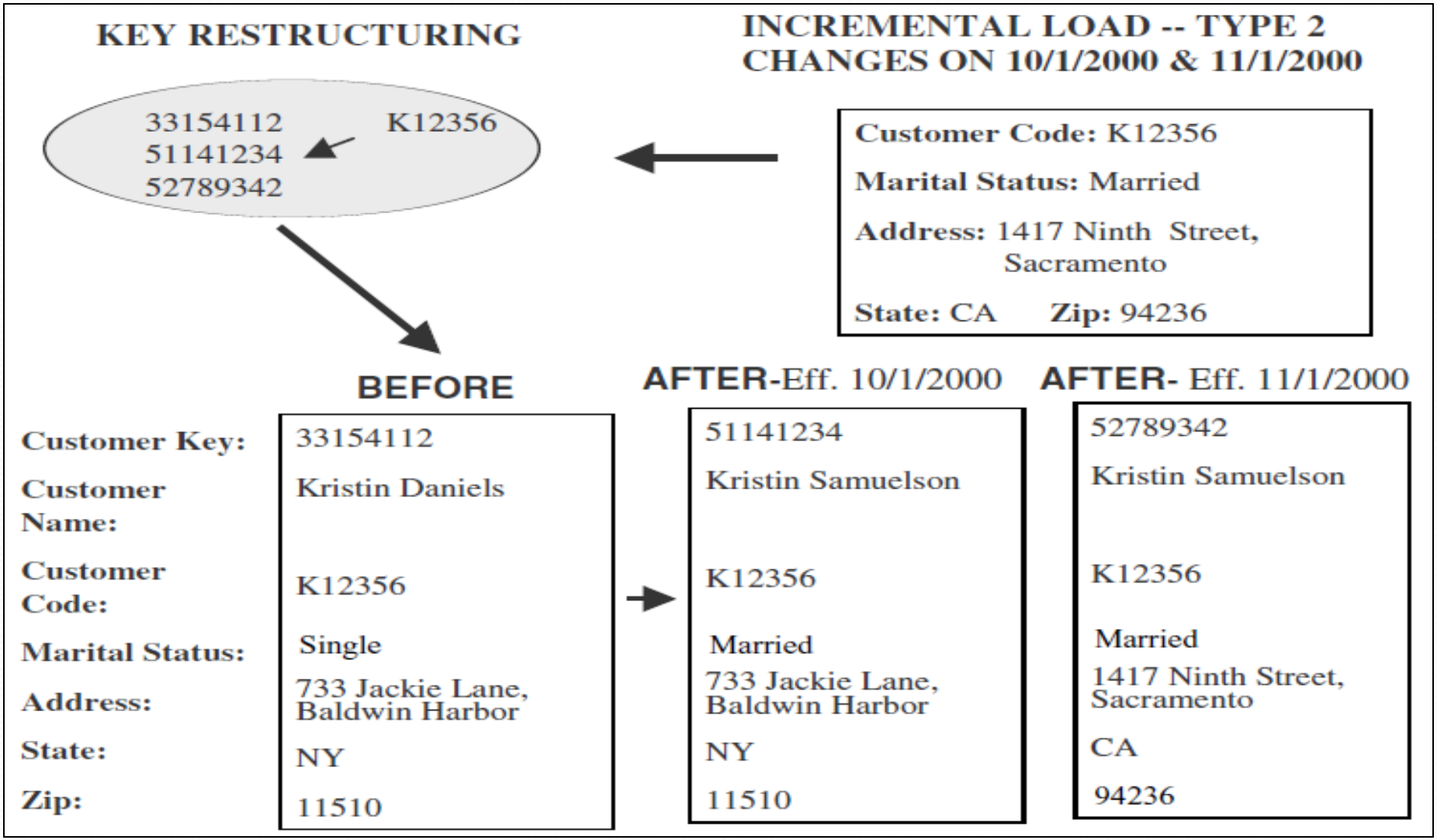
- Overwrite attribute value in the dimension table row with new value.
- No other changes are made to the dimension table row.
- The key is not disturbed
- Easiest type of change to implement.



Type 2 Change: Preservation of history

- True changes in the source systems.
- *E.g., Change of marital status; Change of address*
- There is a **need to preserve history**
- This type of changes partition the warehouse
- Every change for the same attribute must be preserved.
- **Applying these changes:**
  1. Add a new dimension table row with new value of the changed attribute
  2. No changes are made to the existing row.
  3. New rows are inserted with a new surrogate key.

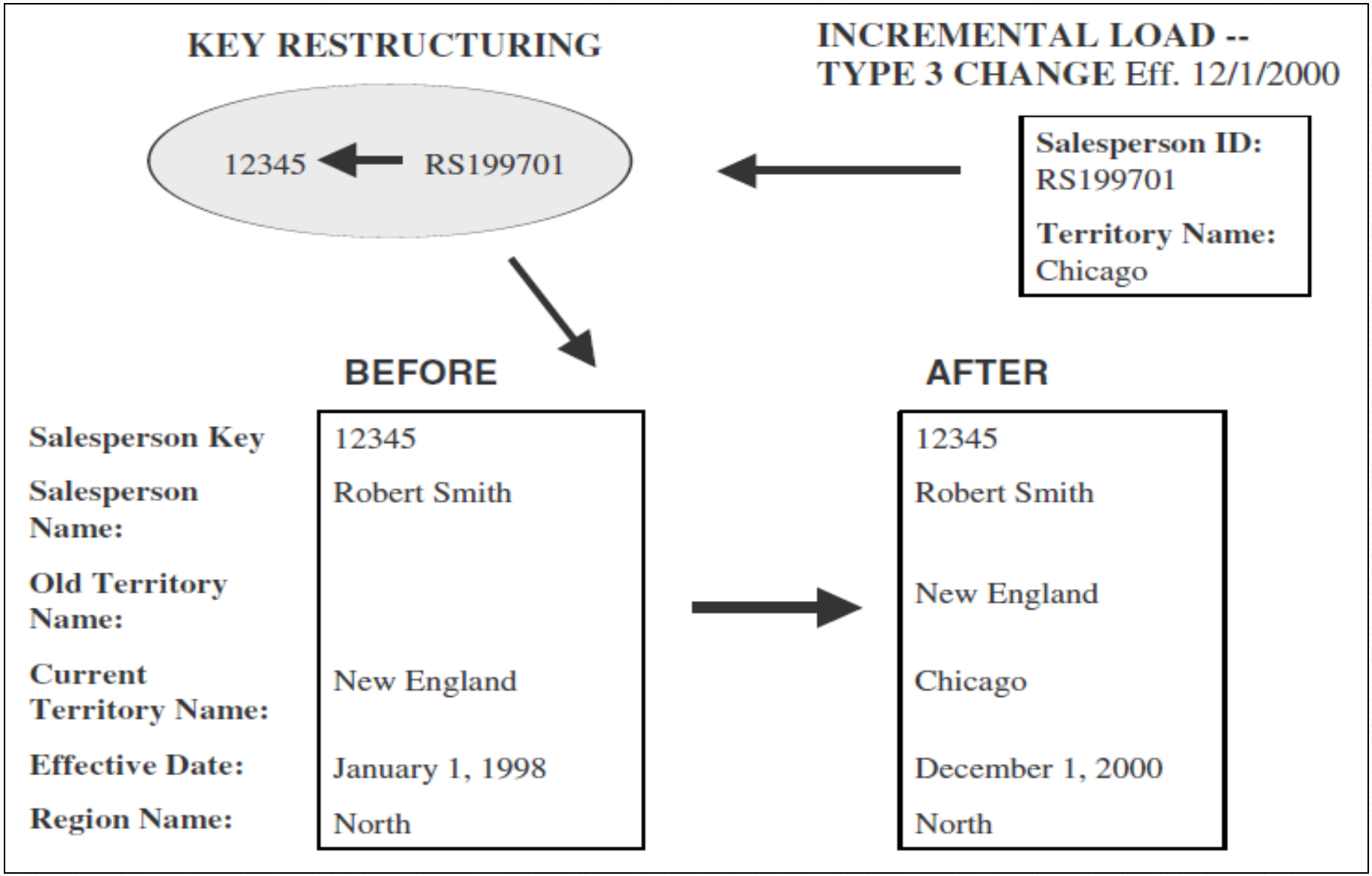
Type 2.. Example



Type 3 Change: Tentative Soft Revision

- Tentative changes in the source system
- *E.g., if an employee will get posted for a short period to a different location*
- Need to keep track of **history with old and new values**
- Used to compare performances across the transition
- **Applying these changes**
  1. An "old" field is added in the dimension table
  2. Push existing value of attribute from "current" to "old"
  3. Update the "current" field with the new value with effective date

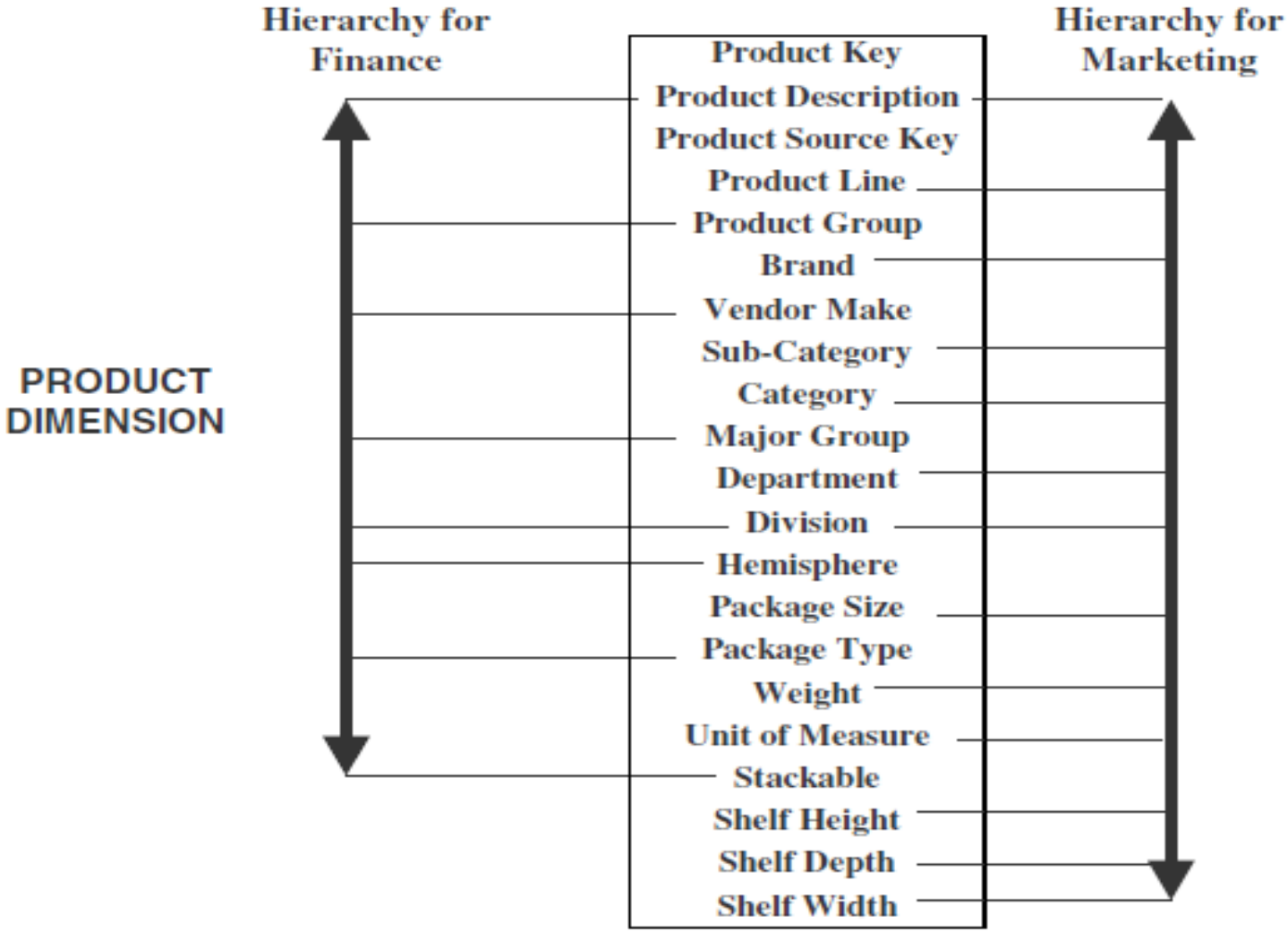
Type 3.. Example



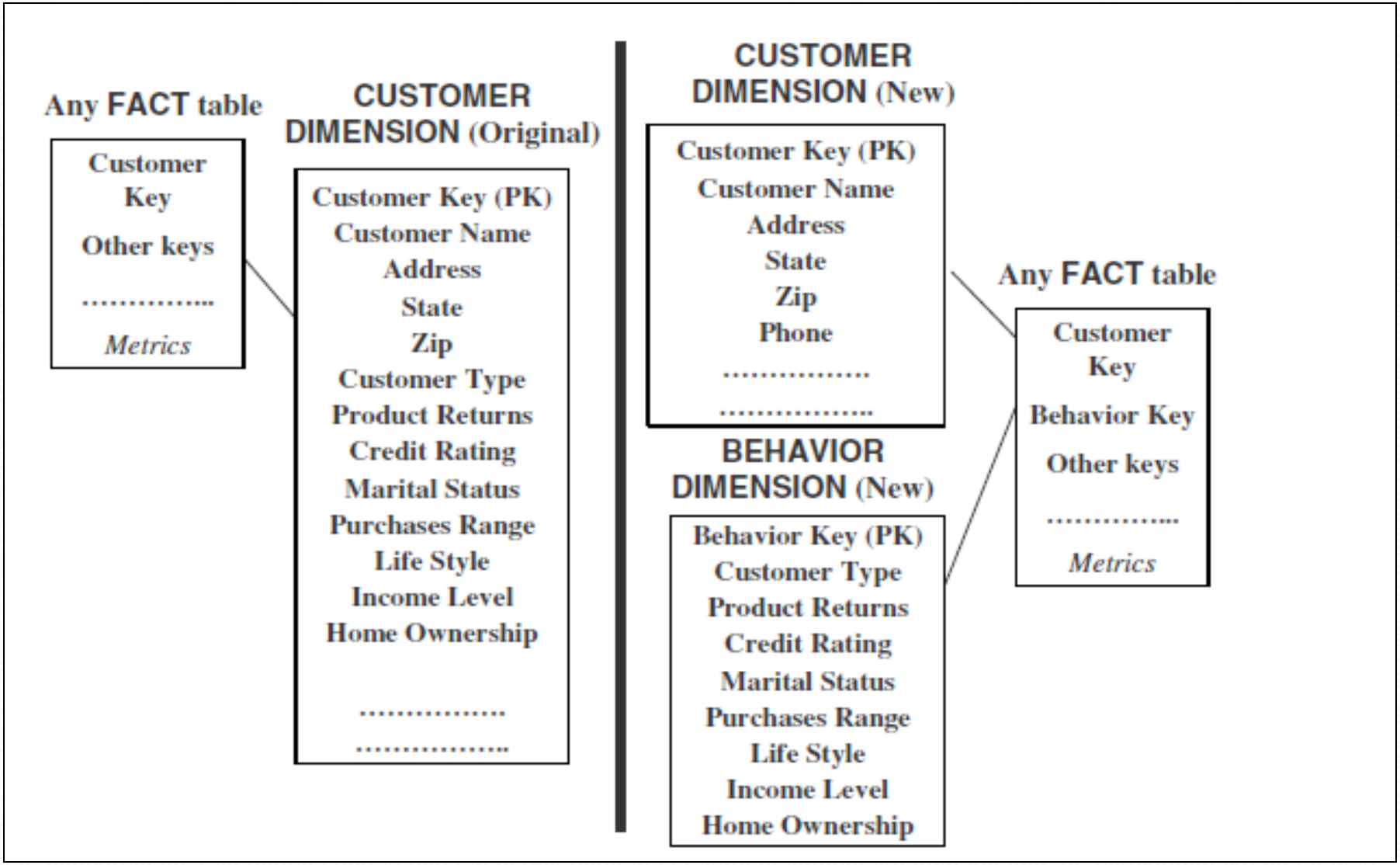
Large Dimensions

- Very deep(large number of rows)
- Very wide(large number of attributes)
- Have multiple hierarchies
- Rapidly changing dimensions

Multiple hierarchies



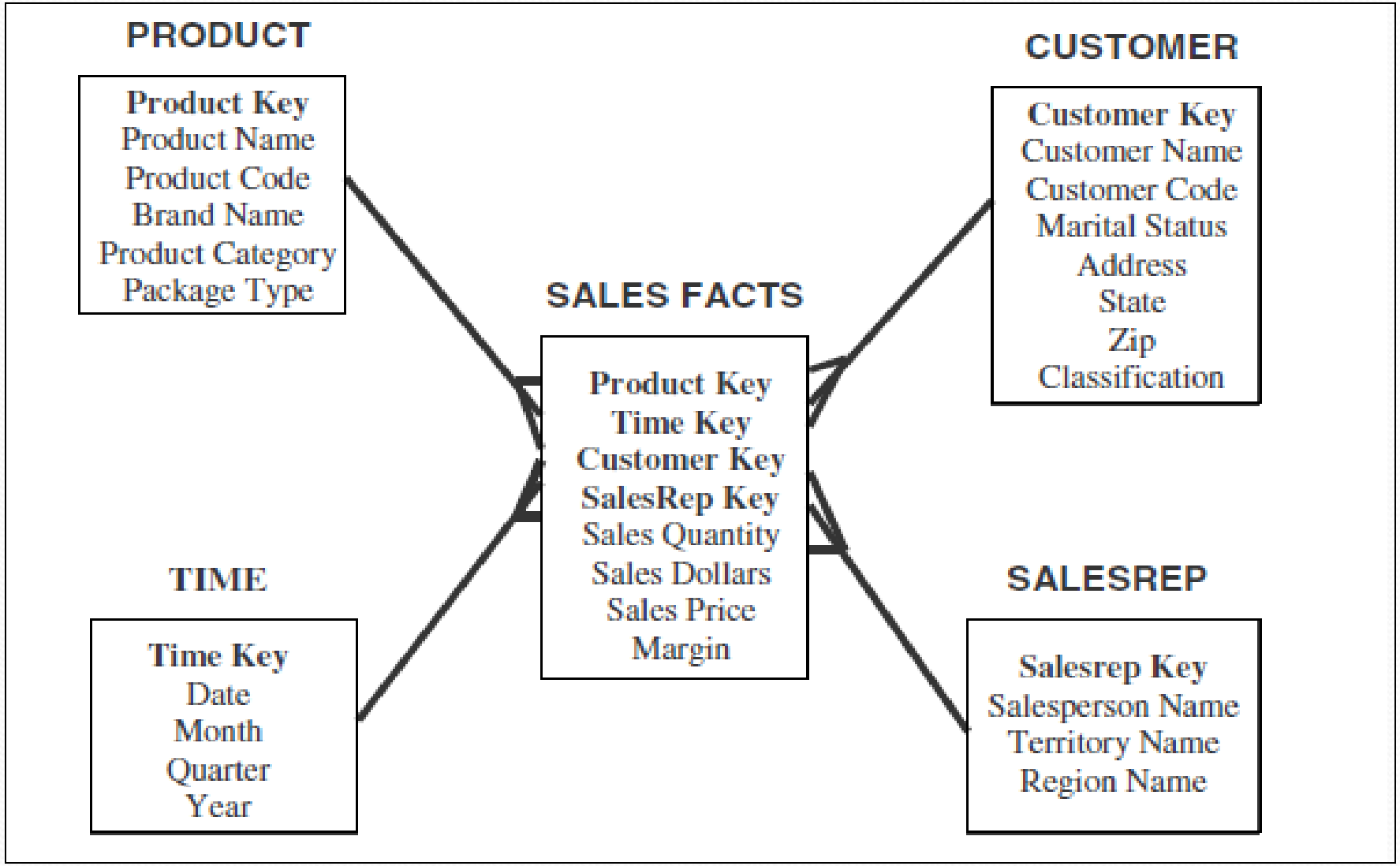
Rapidly Changing dimensions



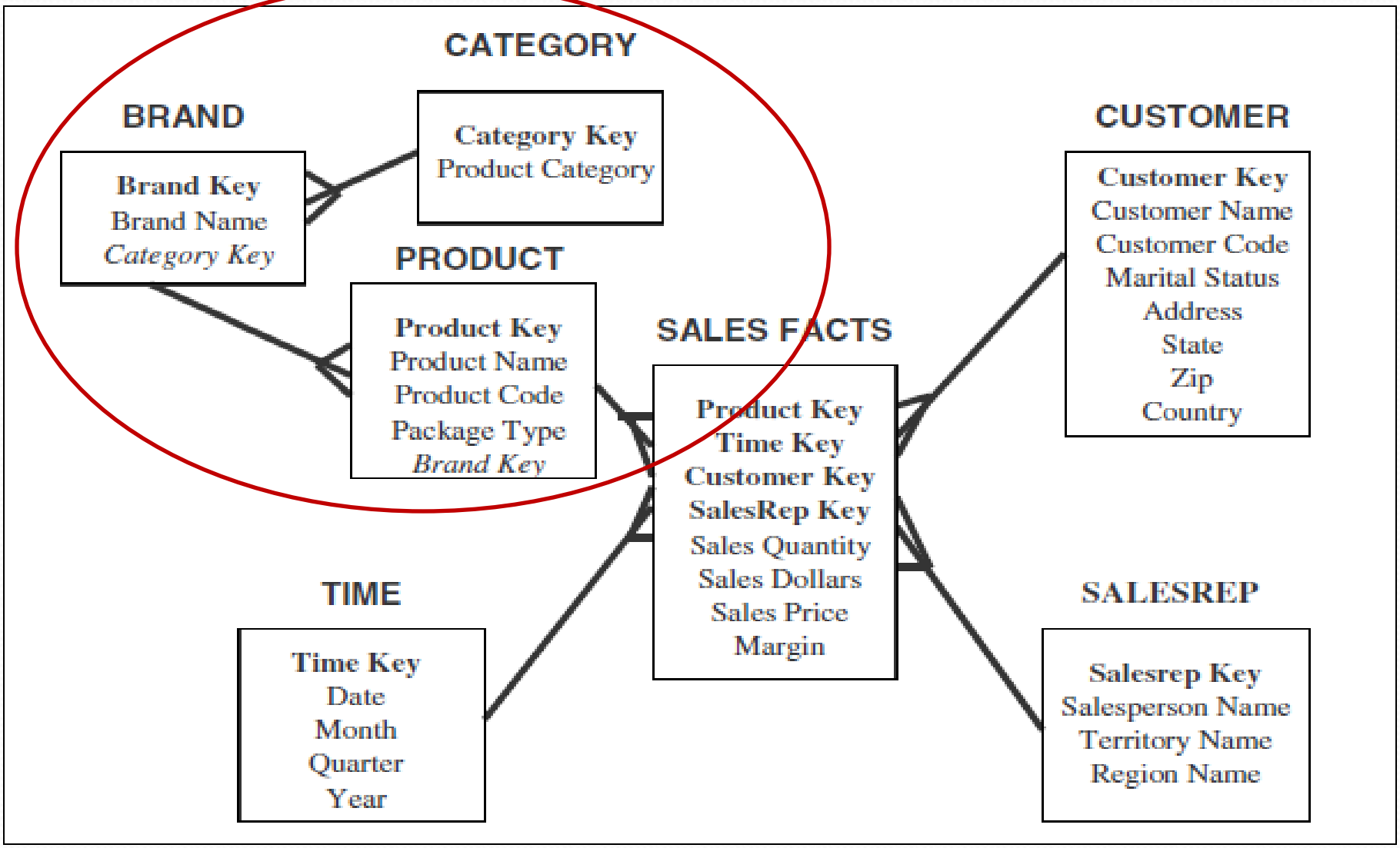
Dimensional Model: **Snowflake Schema**

- ***A variation of the star schema, in which all or some of the dimension tables may be normalized.***
- If all the dimensions in a star schema are normalized, the resulting schema is called a ***Snowflake Schema*** and if only few dimensions are normalized, we call it a ***Starflake Schema*****.**
  1. Eliminates Redundancy
  2. Generally used when a dimension table is wide.
  3. Saves Space
  4. Complex Querying is required.

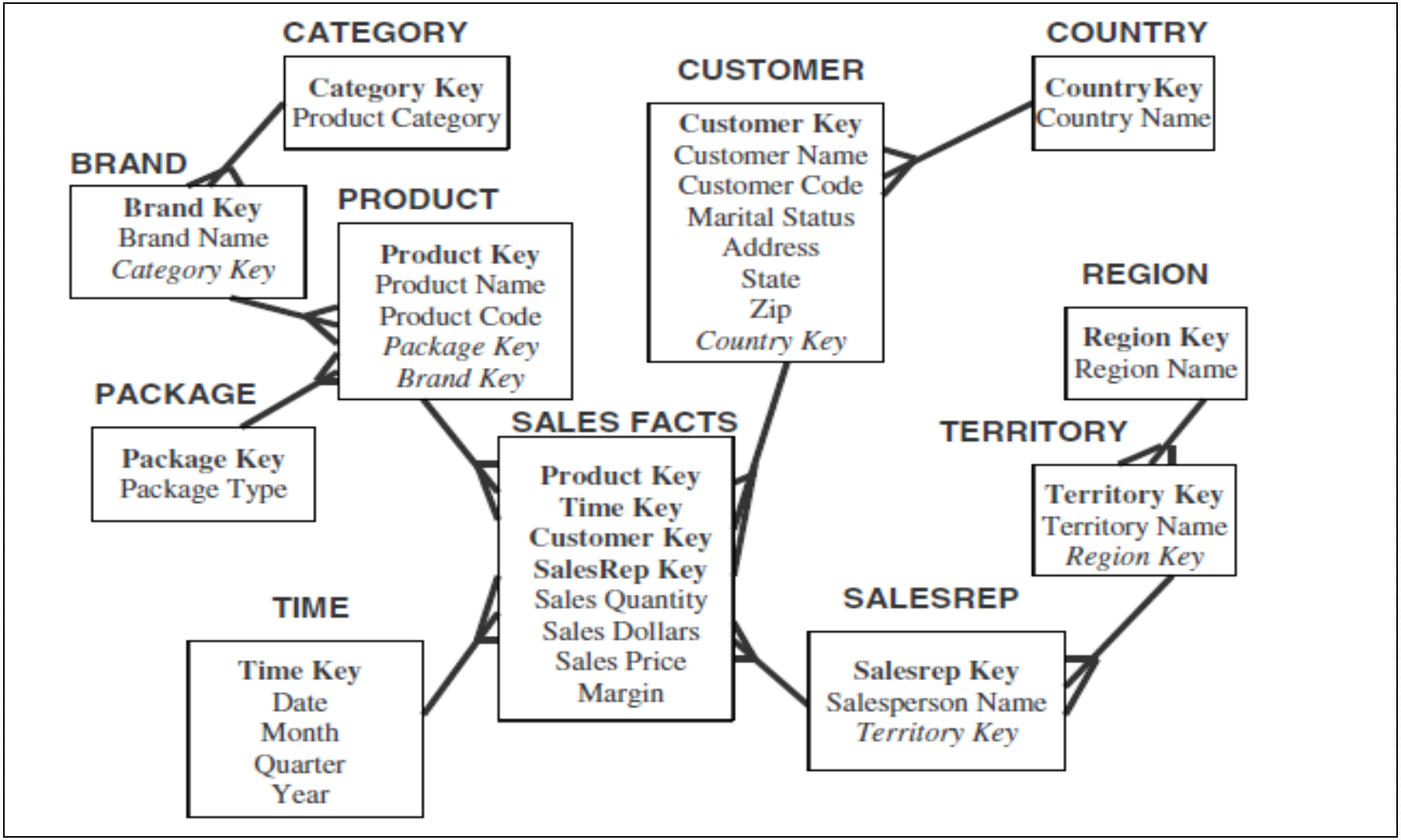
Example: Star Schema for Sales



Normalized Product Dimension



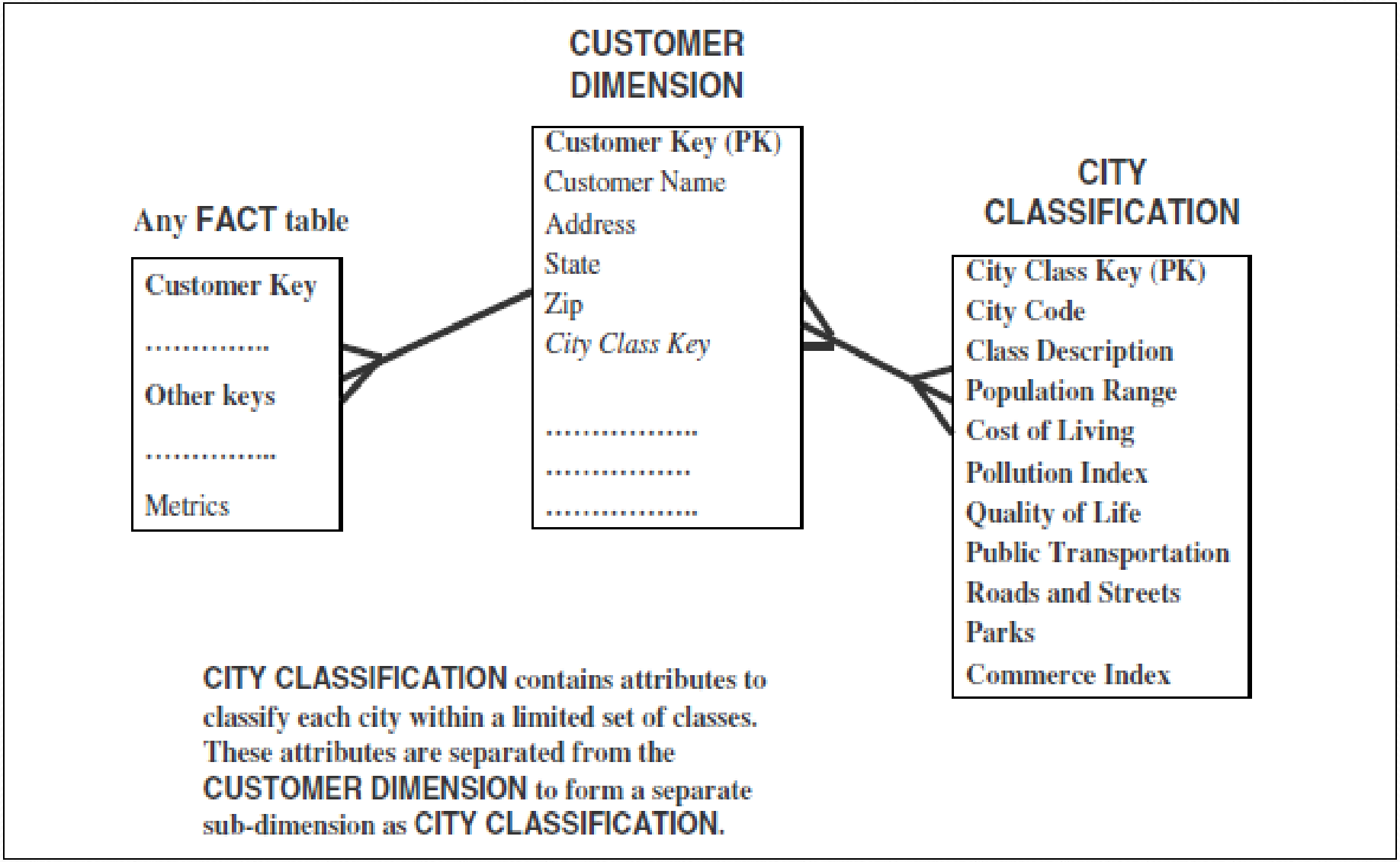
Example: Sales Snowflake Schema



Advantages and disadvantages: **Snow Flake Schema**

- **Advantages**
  - Small savings in storage space
  - Normalized structures are easier to update and maintain
- **Disadvantages**
  - Schema is less intuitive (spontaneous)
  - Browsing becomes difficult
  - Degraded query performance because of additional joins

Important to know: When to Snowflake

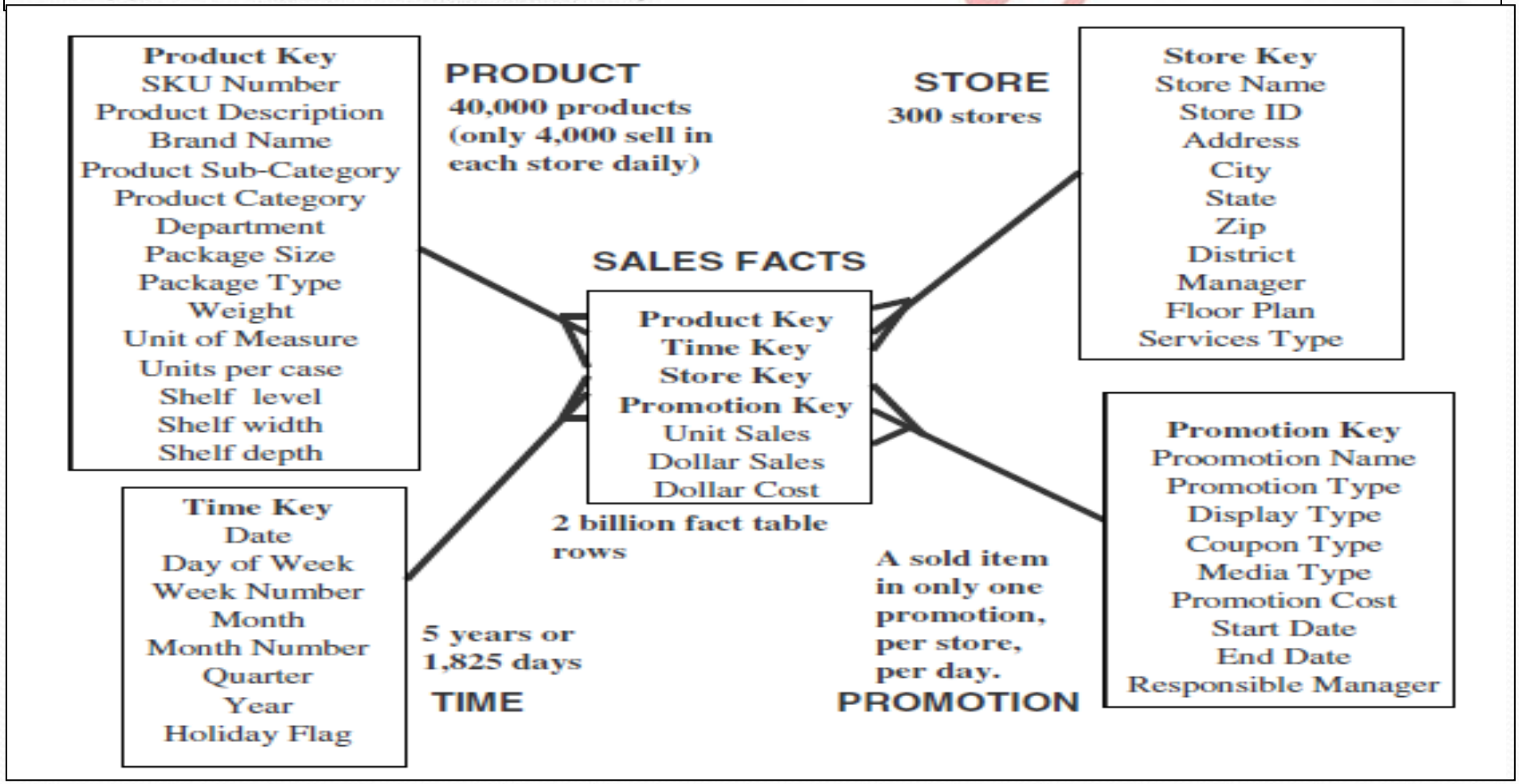


Differentiate between Star Schema & Snowflake Schema

Star Schema	Snowflake Schema
Contains dimension tables mapped around one or more FACT tables	Contains in-depth joins because the tables are split into many pieces
It is a de-normalized table	It is normalized form of Star Schema
No need to use complicated joins	We have to use complicated joins since we have more tables with interconnecting data
Queries results fast	There would be some delay in processing the query
All the primary keys of the dimension table are in the fact table	More dimensions table are there which are linked by primary foreign key relation

Dec 2016, Dec 2017 (Convert into snow flake schema)

Consider following dimensions for a Hypermarket chain: Product, Store, Time and Promotion. With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures).



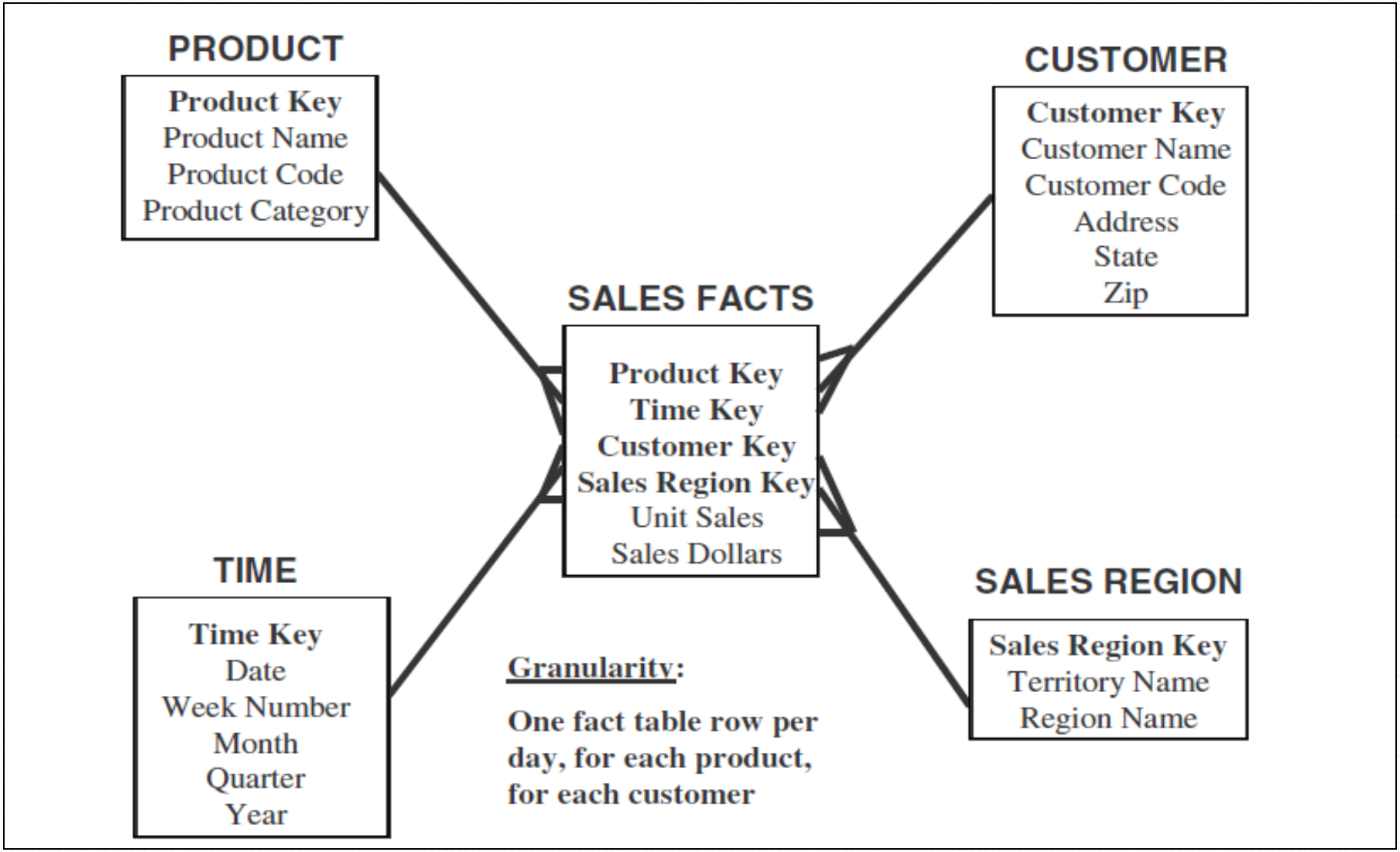
Aggregate Fact Tables

Aggregate fact tables:

- Contain **Pre-calculated Summaries derived** from the most granular (detailed) fact table.
- Created as a specific summarization across any number of dimensions.
- **As the size of fact table is Large:**

1. Reduces runtime processing
2. Speed up query extraction

Star schema with most granular fact table



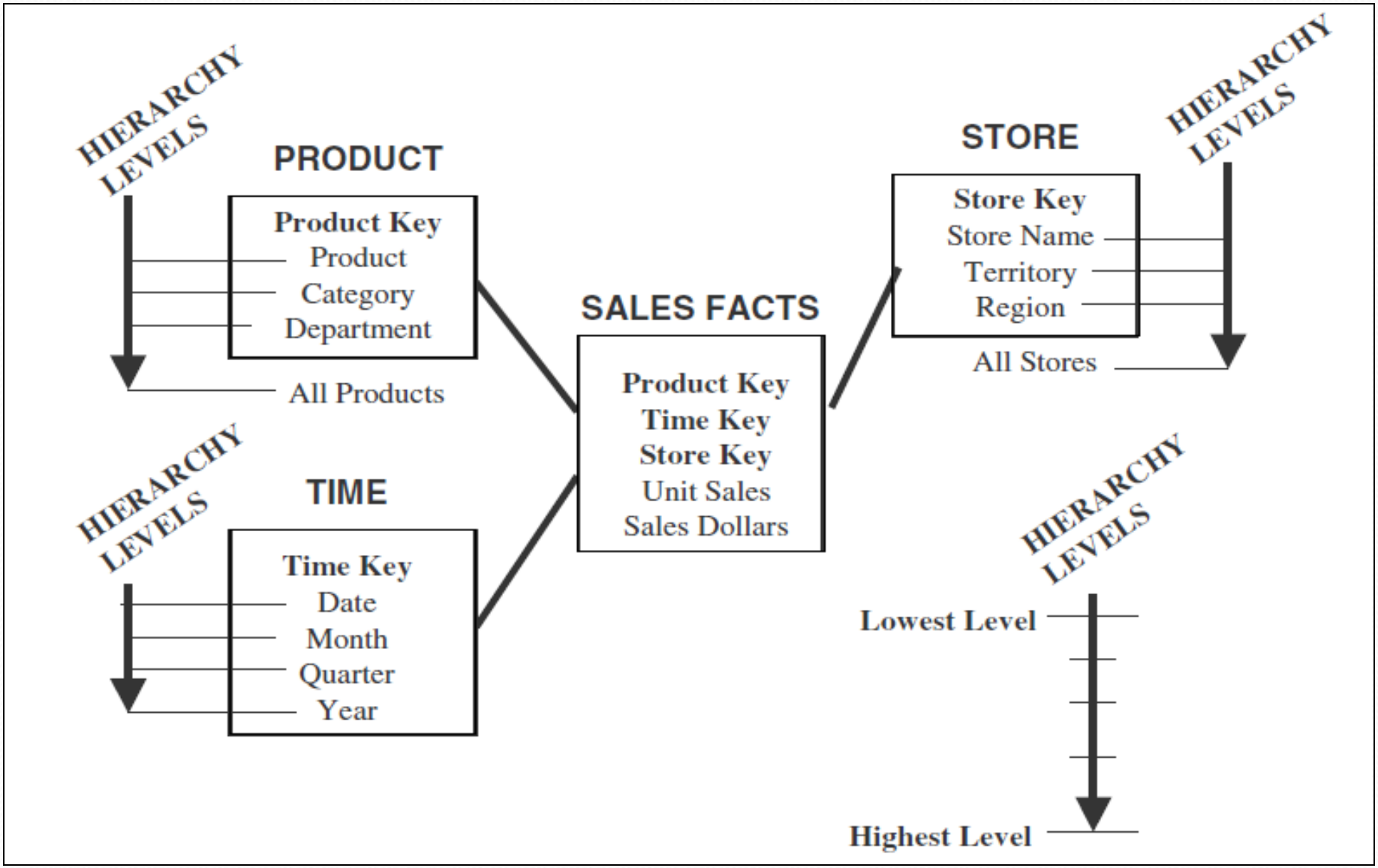
Why need aggregate fact tables?

Query 1: All fact table rows where the customer key relates to customer number 12345678, the product key relates to product Widget-1, and the time key relates to the seven days in the first week of December 2000. Assuming that a customer may make at most one purchase of a single product in a single day, only a maximum of 7 fact table rows participate in the summation.

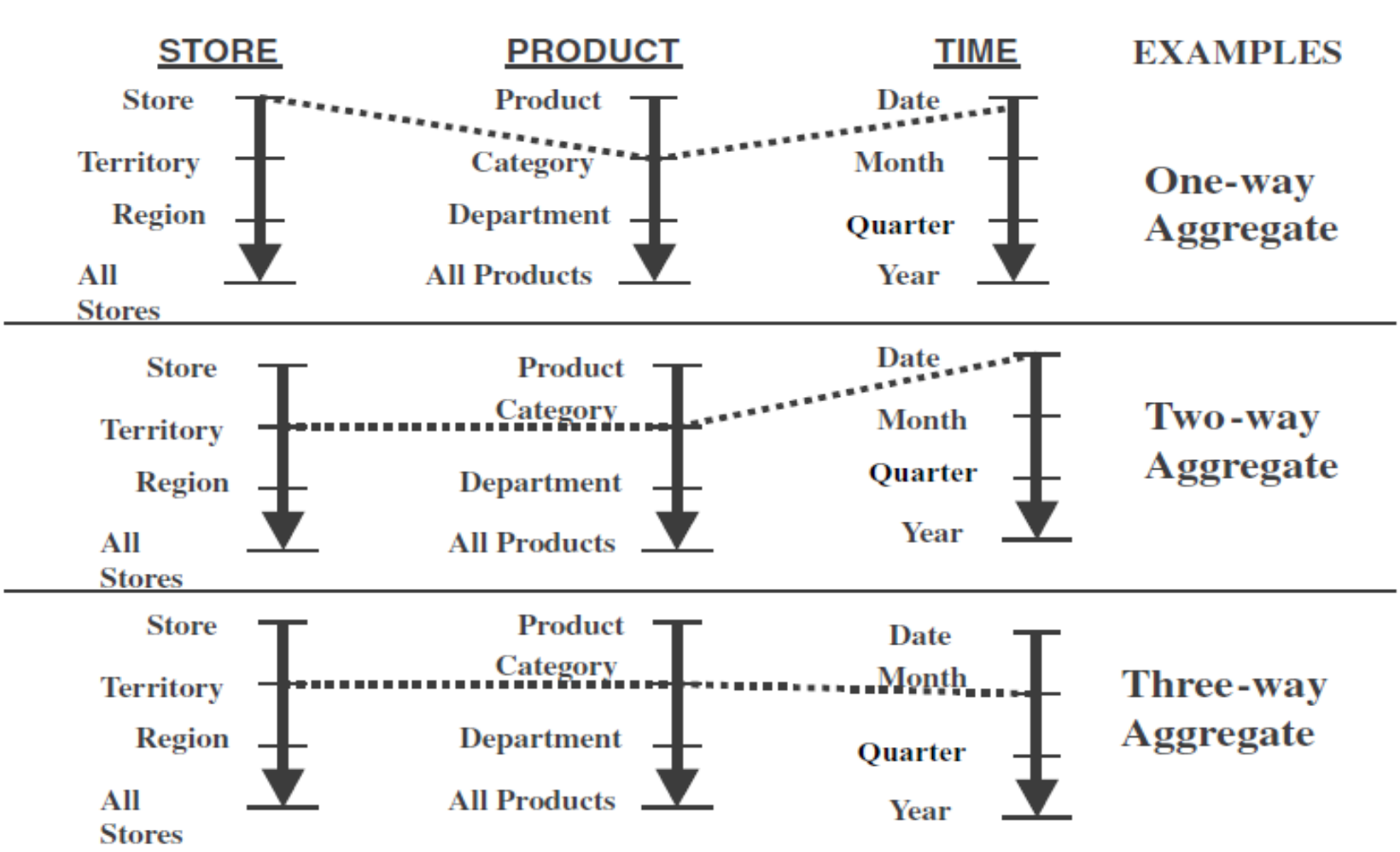
Query 2: All fact table rows where the customer key relates to customer number 12345678, the product key relates to product Widget-1, and the time key relates to about 90 days of the first quarter of 2000. Assuming that a customer may make at most one purchase of a single product in a single day, only about 90 fact table rows or less participate in the summation.

Query 3: All fact table rows where the customer key relates to all customers in the South-Central territory, the product key relates to all products in the product category Bigtools, and the time key relates to about 180 days in the first two quarters of 2000. In this case, clearly a large number of fact table rows participate in the summation.

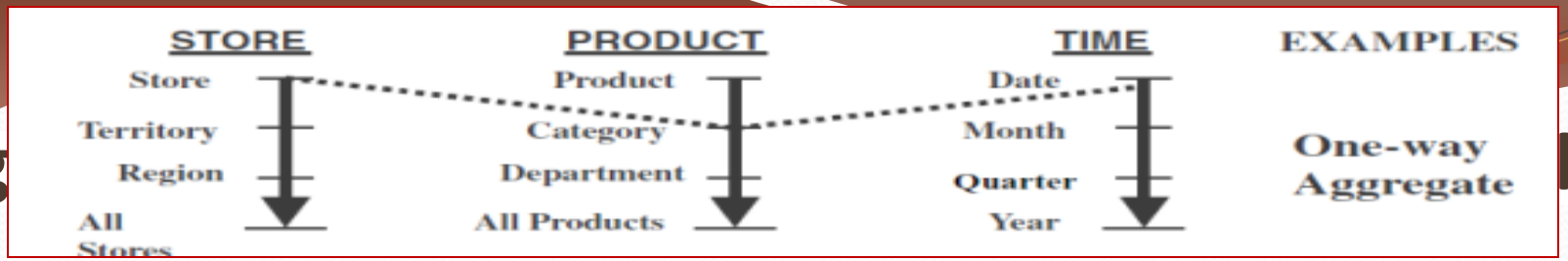
Aggregating Fact Tables:

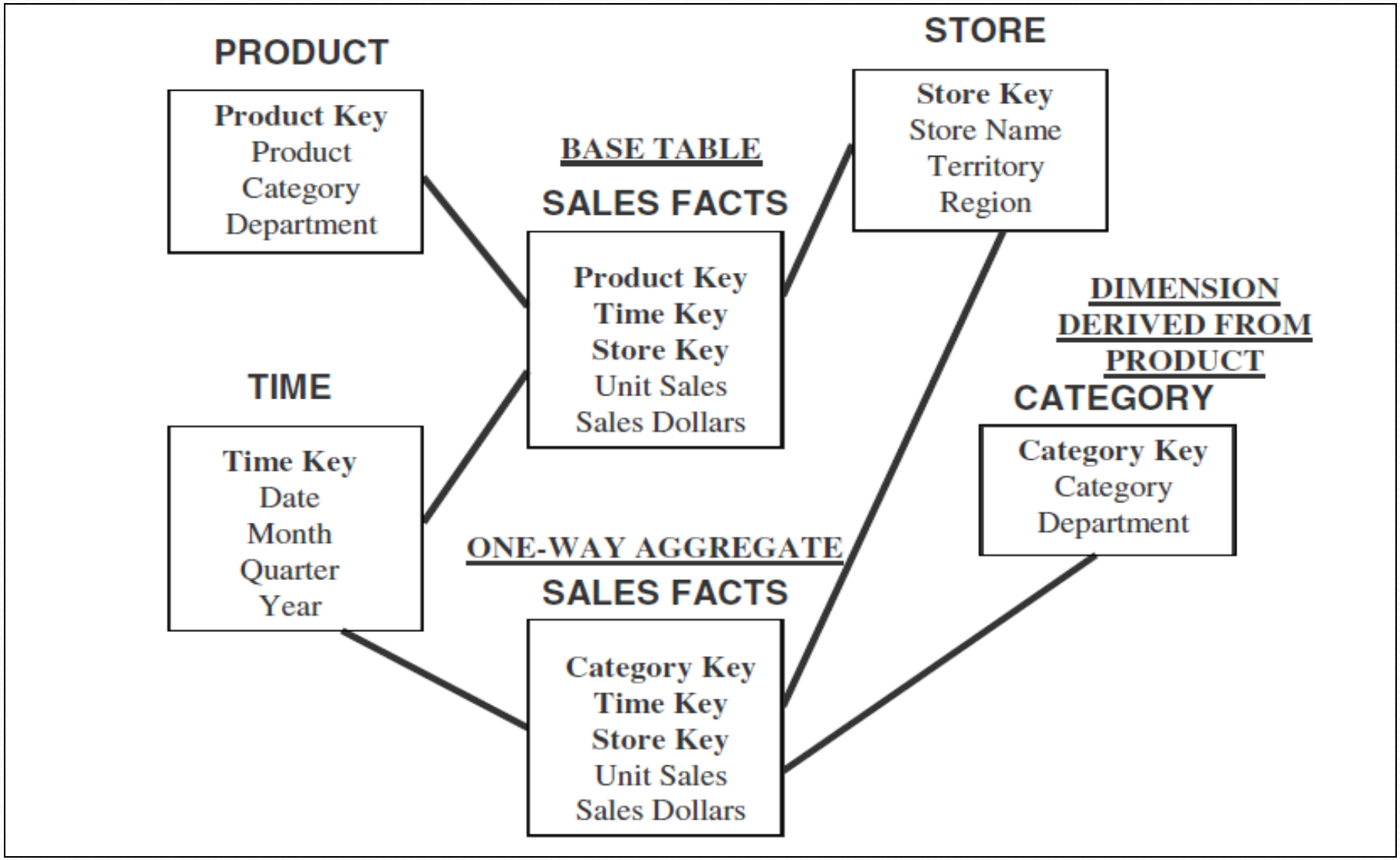


Forming Aggregate fact tables



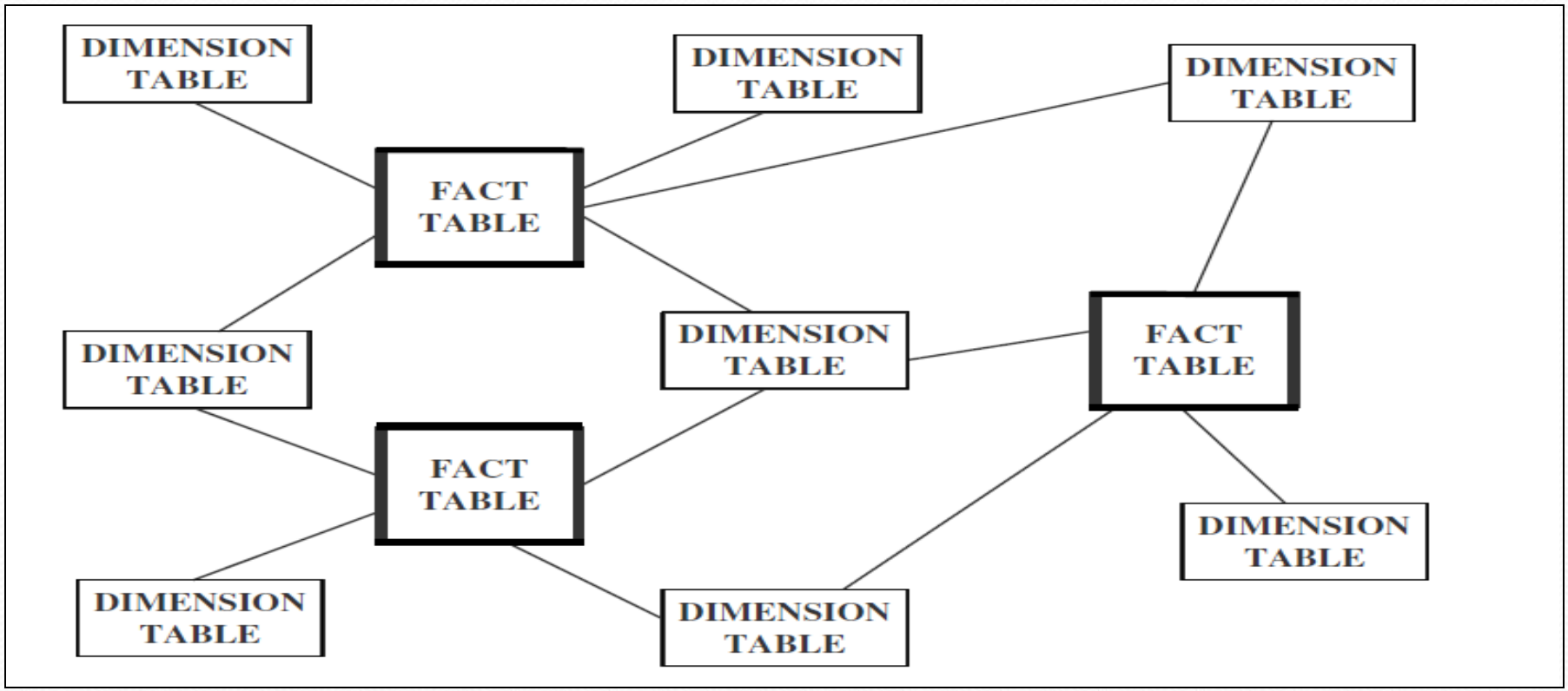
Ag ble


STORE
Store
Territory
Region
All Stores
PRODUCT
Product
Category
Department
All Products
TIME
Date
Month
Quarter
Year
EXAMPLES
One-way Aggregate


PRODUCT
Product Key
Product
Category
Department
STORE
Store Key
Store Name
Territory
Region
BASE TABLE
SALES FACTS
Product Key
Time Key
Store Key
Unit Sales
Sales Dollars
TIME
Time Key
Date
Month
Quarter
Year
DIMENSION DERIVED FROM PRODUCT
CATEGORY
Category Key
Category
Department
ONE-WAY AGGREGATE
SALES FACTS
Category Key
Time Key
Store Key
Unit Sales
Sales Dollars

Fact Constellation: (Family of Stars) (Dec 2017)

- This schema is more complex than star or snowflake varieties, as Multiple fact tables share dimension tables.
- ***This schema is viewed as collection of stars hence called galaxy schema or fact constellation.***



Fact Constellation

- A Schema of this type should only be used for application that need a high level of Sophistication.
- ***For each star schema or snowflake schema it is possible to construct a fact constellation schema.***
- This solution is very flexible, however it may be hard to manage and support.
- **Advantage:**
  1. Useful in cases when some facts are associated with a given dimensional level and other facts with a deeper dimensional level.
- **Disadvantage:**
  1. **Complicated Design:** Many variants of aggregation must be considered

University Questions

1. What is Dimensional Modeling? Explain in detail.(Dec 2011)
2. A Dimension table is wide; the fact table is deep. Explain. (Dec 2010)
3. Define Fact less Fact tables with suitable example. (May 2011,Dec 2012)
4. Write short note on Updates to dimensional table. (May 2016, May 2017)
5. Write short note on Snowflake Schema. (May 2011, Dec 2011,Dec 2012)
6. Write short note Fact Constellation Schema or Families of Star. (Dec 2017)
5. Design Star Schema for auto sales Analysis of Company? (Dec 2011)

University Questions

6. Star Schema (May 2011)

(a) The college wants to record the grades for the courses completed by students. There are four dimensions :—

(i) Course
(ii) Professor
(iii) Student
(iv) Period.

The only fact that is to be recorded in the table is course-grade :–

(i) Design star schema.
(ii) Write DMQL for the above star schema.

7. IPD, Star Schema (Dec 2012)

Consider the following database for a chain of bookstores.

BOOKS (Booknum, Primary_author, Topic, Total_stock, price)
BOOKSTORE (Storenum, City, State. Zip, Inventory_value)
STOCK (Storenum, Booknum, Qty)

With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.

(a) Design an information package diagram. (5)
(b) Design a star schema for the data warehouse clearly identifying the Fact table(s), Dimension table(s), their attributes and measures. (5)

May 2017

What is dimensional modelling? Design the data warehouse for wholesale furniture Company. The data warehouse has to allow analysing the company's situation at least with respect to the Furniture, Customer and Time. More ever, the company needs to analyse: The furniture with respect to its type, category and material. The customers with respect to their spatial location, by considering at least cities, regions and states. The company is interested in learning the quantity, income and discount of its sales.

Dec 2016, Dec 2017

Consider following dimensions for a Hypermarket chain: Product, Store, Time and Promotion. With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures).

May 2018

i. Design star & snowflake schema for "Hotel Occupancy" considering dimensions like Time, Hotel, Room, etc.

ii. Calculate the maximum number of base fact table records for the values given below:

Time period: 5 years

Hotels: 150

Rooms: 750 rooms in each Hotel (about 400 occupied in each hotel daily).

May 2016

For a Super market chain, consider the following dimensions namely product, store, time and promotion. The schema contains a central fact table for sales.

i. Design star schema for the above application.

ii. Calculate the maximum number of base fact table records for warehouse with the following values given below:

- Time period – 5 years
- Store – 300 stores reporting daily sales
- Product – 40,000 products in each store (about 4000 sell in each store daily)

May 2019

Suppose that a data warehouse for *DB-University* consists of the four dimensions *student, course, semester,* and *instructor*, and two measures count and *avg-grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg-grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg-grade* stores the average grade for the given combination. **[10]**

i. Draw a *snowflake schema* diagram for the data warehouse.

Dec 2019

a) Suppose that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.

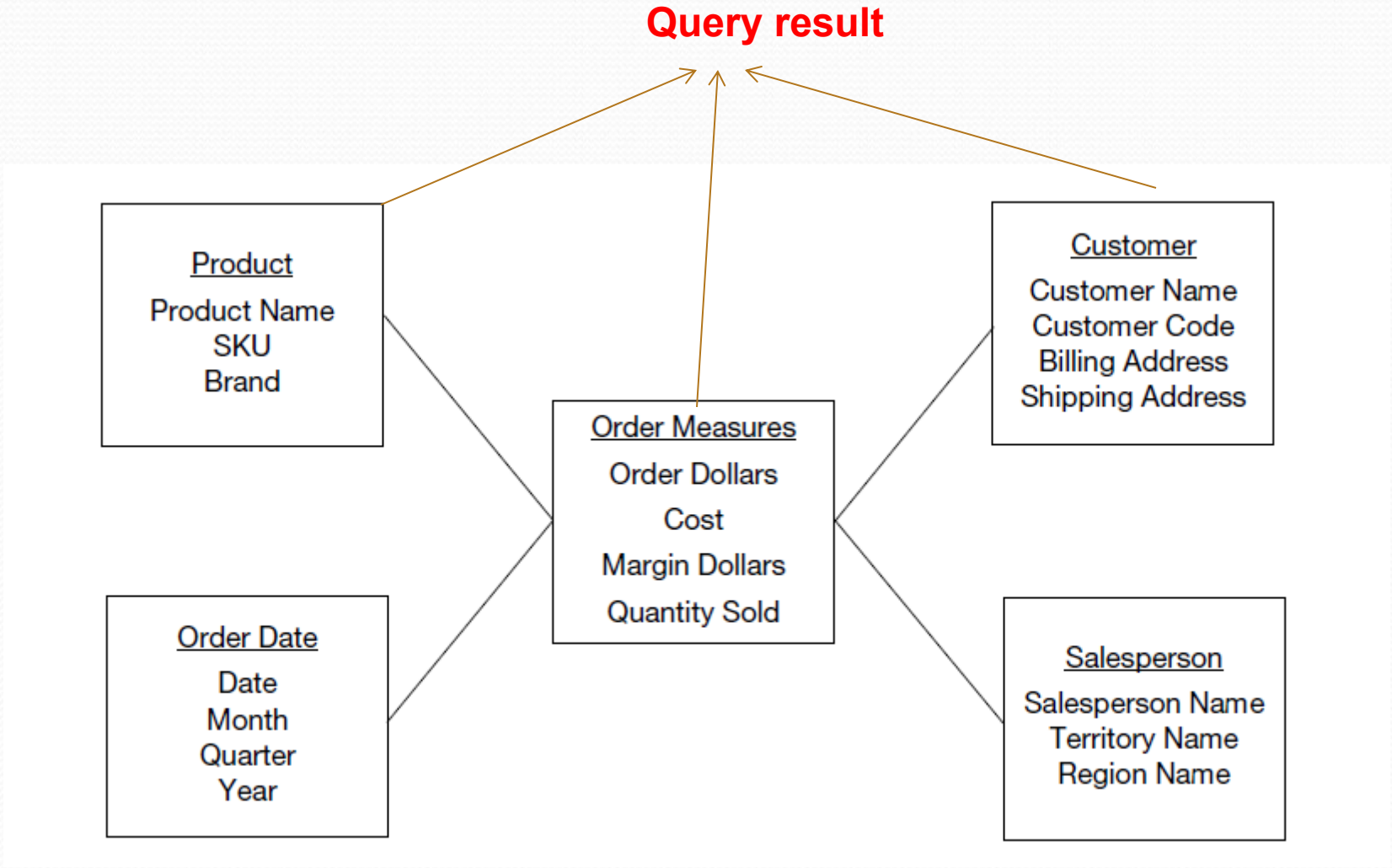
(i) Draw a star schema diagram for the above data warehouse.

Thank You

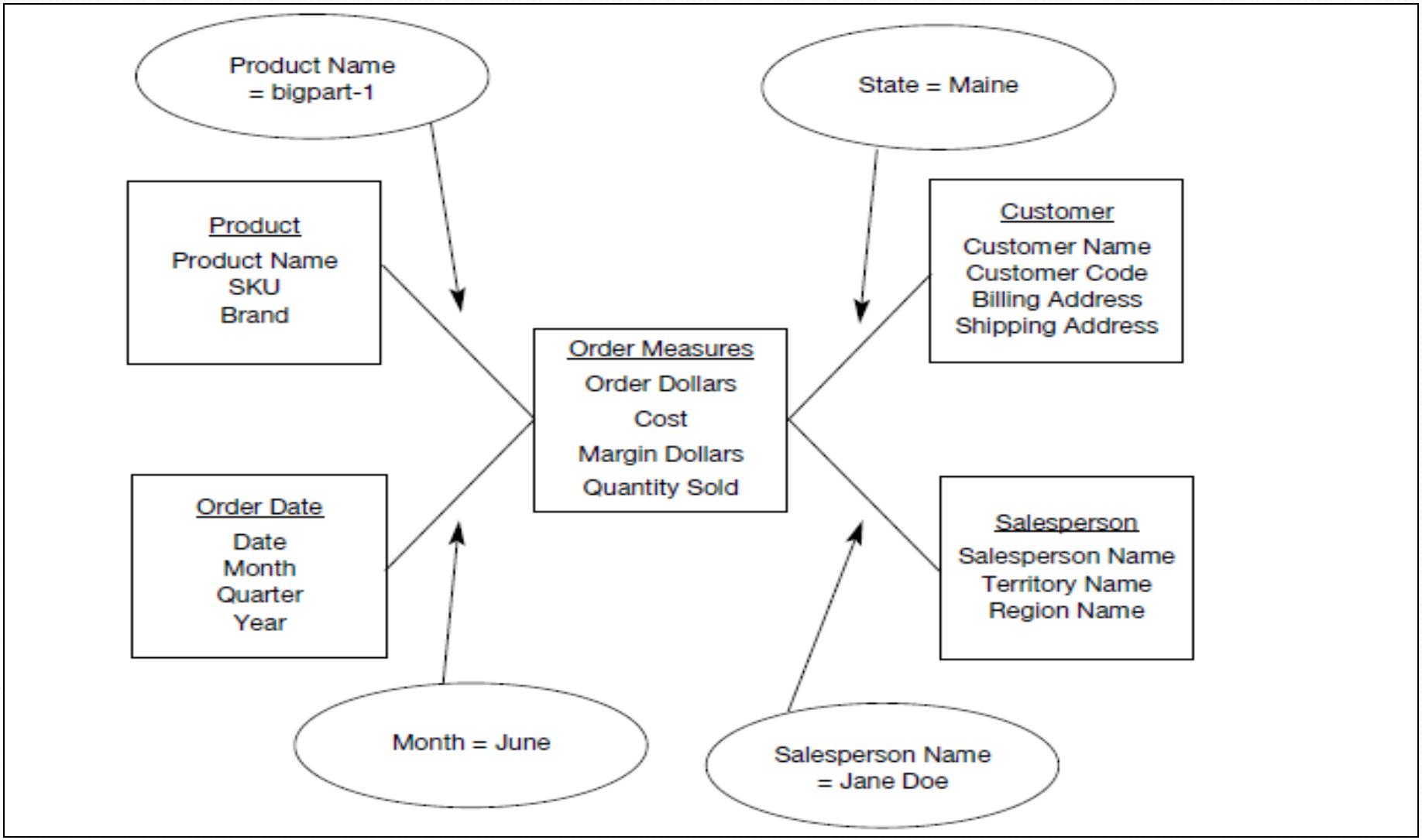
Sample Star Schema Examples

---For Reference Study

1. Star Schema-example of order analysis



Understanding query from the star schema of order analysis



Understanding drill down analysis from the star schema of order analysis