Designing the Data Warehouse

Dimensional Modeling:

- To create a data warehouse system we must determine how we are going to extract meaningful data and logically group the data.
- Multidimensional modeling is a technique for structuring data around the business concepts.

- Data is modeled and viewed in multiple dimensions.
- These dimensions are the perspectives or entities with respect to which an organization wants to keep records.
- AllElectronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time , product , customer and store.

Information Package Diagram

- First step is to prepare an information package, that allows the data warehouse's designers to layout the requirements for the dimension tables, their hierarchies, and the facts to be modeled.

Information Subject: Sales

Dimensions

Hierarchies/Categories

Time	Product	Customer	Store
Year	Category	License type	State
Quarter	Subcategory	Category	Region
Season	Product name	Size	City
Month		Customer Name	Square footage
Date			Store name
Day of Month			
Day of Week			

Facts: Sales quantity, Item dollar amount, Item cost

Dimension Tables:

- The information package is then used to create the dimension tables.

Product	
PK	**ProductID**
	ProductName Category Subcategory

Store	
PK	**StoreID**
	StoreName StoreRegion StoreState StoreCity StoreFootage

Time	
PK	**TimeID**
	TimeDate DateText DayOfWeek WeekDay DayOfMonth MonthNum MonthText Quarter Season Year

Customer	
PK	**CustomerID**
	CompanyName CustomerCategory LicenseType Size

The Fact Table

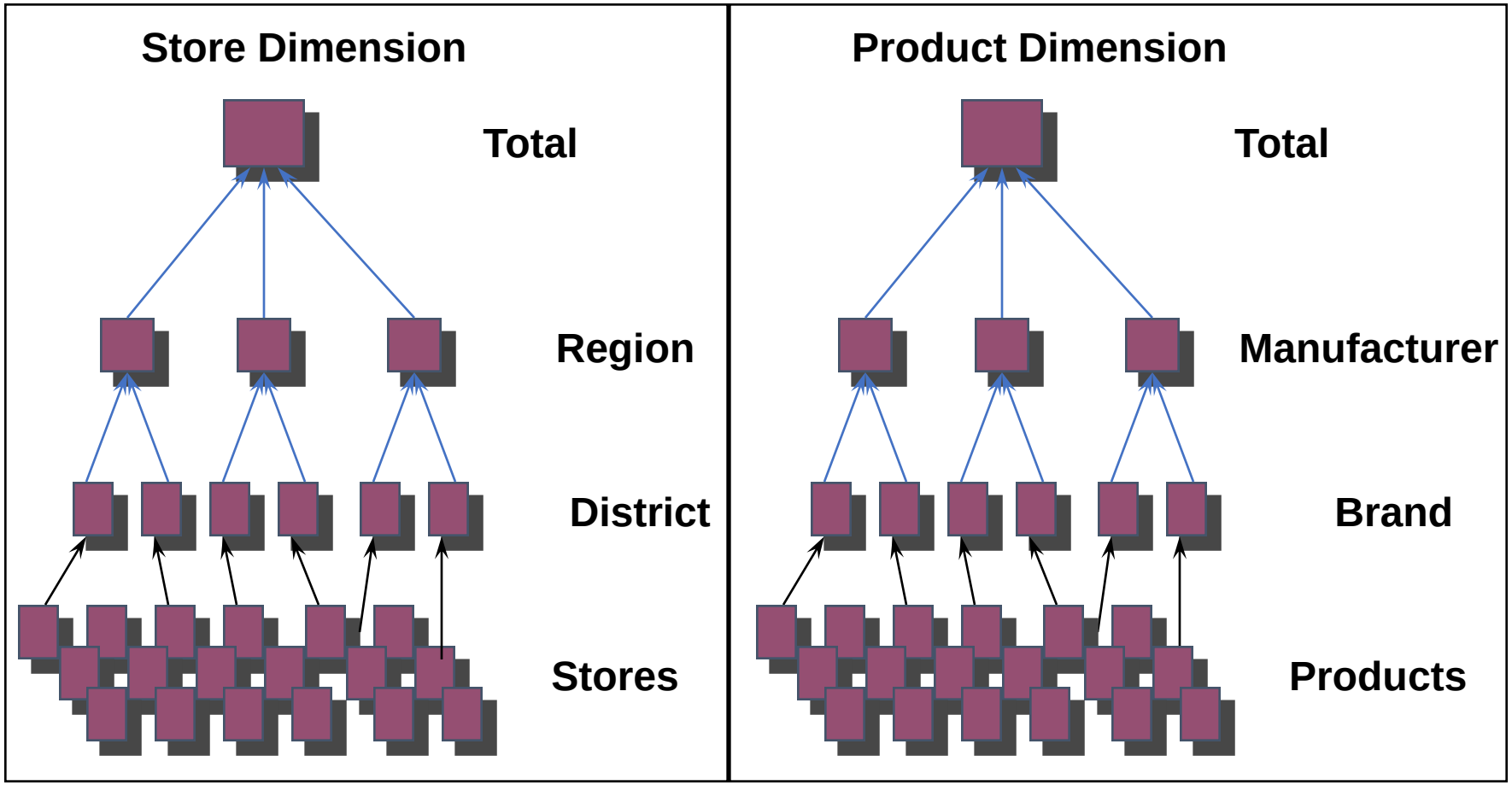
The fact table used for this project was based on sales information.

Sales fact table

ProductID	Fact table keys (PK)
StoreID	
CustomerID	
TimeID	
Sales quantity	Measures
Item dollor amount	
Item cost	

Dimensional Modeling

- All products are grouped into categories and each category if further divided into several subcategories
- This allows related items to be grouped and summarized for high level analysis while retaining the ability to drill down to more specific product detail



Dimensional Modeling

Dimensional Modeling

Customers are organized into three hierarchies: size, license type and category; and each has further hierarchies

The AC sales are high during the summer, but heating sales are high in winter, thus Time dimension is categorized in two categories

Customer			Time	
Size	**License Type**	**Category**	**Calendar**	**Seasonal**
Small (2-3 techs) Medium (4-10 techs) Large (11-20 techs) Corporate (21+ techs)	"A" license (HVAC and refrigeration) "B" license (HVAC only)	HVAC Builder Government Refrigeration Maintenance	Year Quarter Month Day	Year Season Month Day of the week

E-R modeling for OLTP systems

- OLTP systems capture details of events or transactions
- OLTP systems focus on individual events
- An OLTP system is a window into micro-level transactions
- Picture at detail level necessary to run the business
- Suitable only for questions at transaction level
- Data consistency, non-redundancy, and efficient data storage critical

E-R modeling for OLTP systems.

Entity-Relationship Modeling

- Removes data redundancy
- Ensures data consistency
- Expresses microscopic relationships

Dimensional modeling for the data warehouse.

DW meant to answer questions on overall process

- DW focus is on how managers view the business
- DW reveals business trends
- Information is centered around a business process
- Answers show how the business measures the process
- The measures to be studied in many ways along several business dimensions.

Dimensional modeling for the data warehouse.

Dimensional Modeling

- Captures critical measures
- Views along dimensions
- Intuitive to business users

Data Warehouse Schema

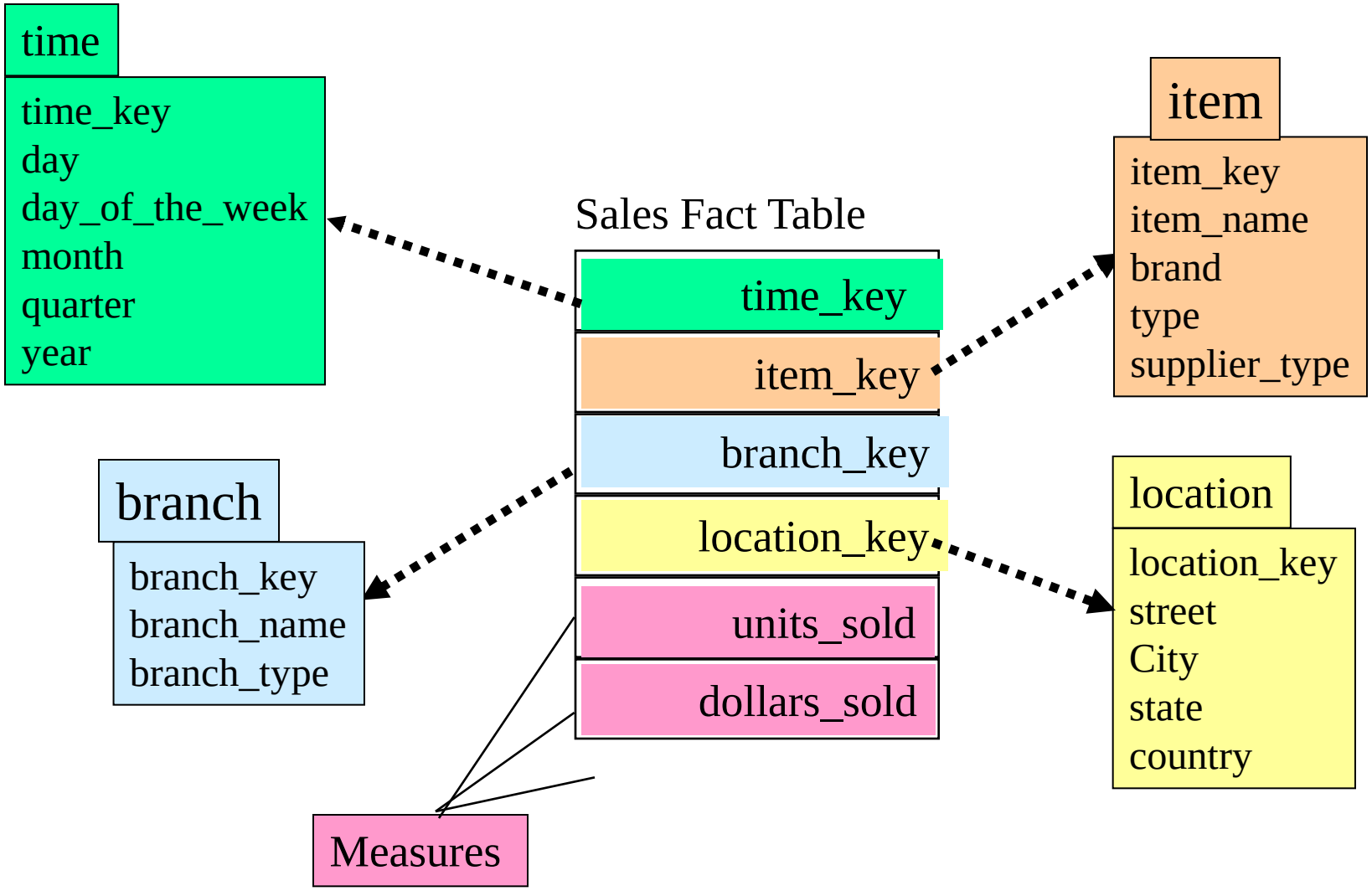
- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

- The major focus will be on the star schema which is commonly used in the design of many data warehouse.

Star Schema

- This is the most common modeling paradigm for designing data warehouse.
- In this model a data warehouse consists of:
  - a large central table (fact table) containing the bulk of the data, with no redundancy,
  - a set of smaller attendant tables (dimension tables), one for each dimension. Dimension tables are not normalized
- The diagram below show an example of star schema

Example of Star Schema



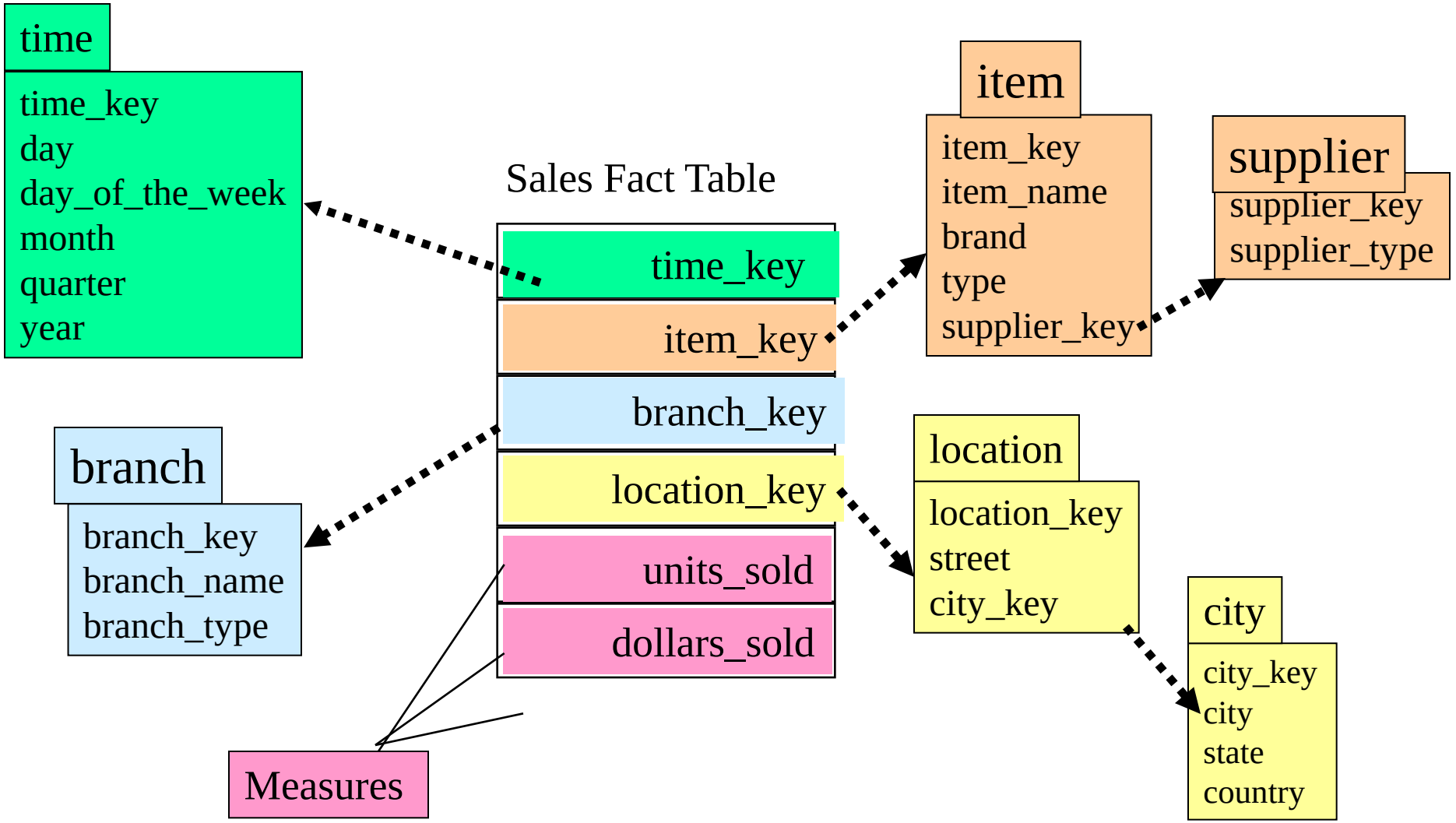
Star Schema

- A star schema for *AllElectronics* sales is shown in Figure in the above slide. Sales are considered along four dimensions namely, *time, item, branch*, and *location*.
- The schema contains a central fact table for *sales* that contains **keys** to each of the four dimensions, along with **two measures**: *dollars sold* and *units sold*.
- To minimize the size of the fact table, dimension identifiers (such as *time key* and *item key*) are system-generated identifiers.

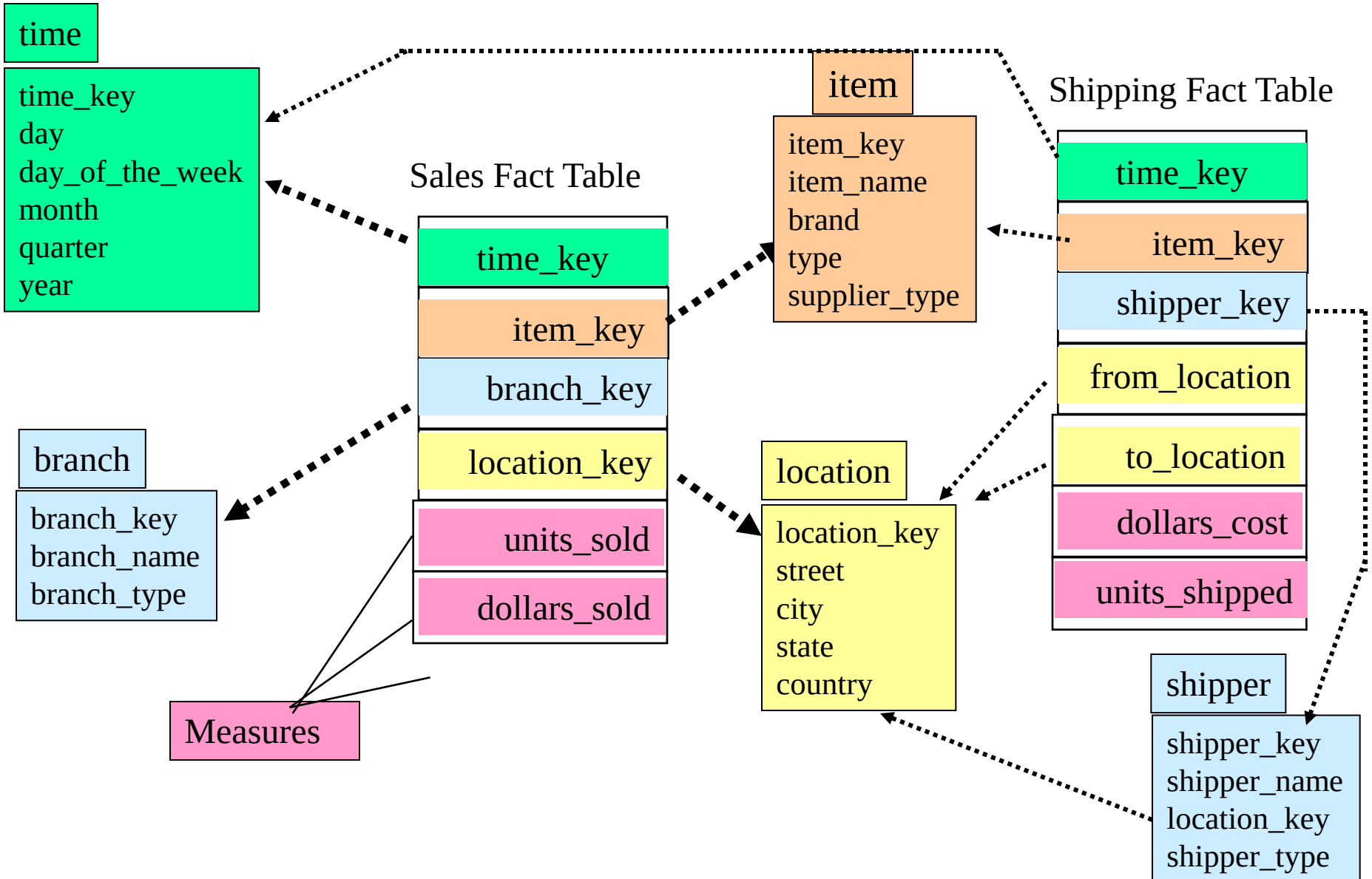
Star Schema

- Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes.
- For example, the *location* dimension table contains the attribute set {*location key, street, city, province or state, country*}

Example of Snowflake Schema



Example of Fact Constellation



Example

Q1. (A) A bank wants to develop a data warehouse for effective decision-making about their loan schemes. The bank provides loans to customers for various purposes like House Building Loan, Car Loan, Educational Loan, Personal Loan, etc. The whole country is categorized into a number of regions, namely, North, South, East and West. Each region consists of a set of states. Loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. The data warehouse should record an entry for each disbursement of loan to customer. With respect to the above business scenario,

(1) Design an information package diagram. Clearly explain all aspects of the diagram (5)

(2) Draw a star schema for the data warehouse clearly identifying the Fact table(s), Dimension table(s), their attributes and measures. (5)

Datalake:

DWH: A data warehouse is a centralized repository of structured and curated data that is extracted, transformed, and loaded (ETL) from various sources.
A data warehouse supports business intelligence (BI) and analytics by providing a consistent and reliable view of the data across different dimensions, such as time, location, product, customer, and so on.
A data warehouse **is typically designed with a predefined schema, or a logical structure, that defines the tables, columns, and relationships of the data**.

A **data lake** is a distributed repository of raw and unstructured data that is ingested from various sources in its original format.
A data lake supports data exploration and discovery by allowing users to store and access any type of data, such as text, images, audio, video, and so on.
A data lake is typically designed with a schema-on-read approach, meaning that the structure and meaning of the data are determined at the time of querying, rather than at the time of loading.

Storage:
One of the key differences between data warehouse and data lake architectures is how they store data.
A data warehouse stores data in a structured and normalized way, using relational databases or columnar formats. This reduces data redundancy and improves data quality, but also requires more processing and storage resources.

A data lake stores data in a flat and flexible way, using object storage or file systems(raw data). This enables data scalability and diversity, but also increases data complexity and governance challenges.

Processing:
Another key difference between data warehouse and data lake architectures is how they process data.
A data warehouse processes data before loading it into the repository, using ETL tools and pipelines. This ensures that the data is clean, consistent, and ready for analysis, but also limits the scope and speed of data ingestion.

A data lake processes data after loading it into the repository(ELT), using various tools and frameworks, such as Hadoop, Spark, or SQL. This enables faster and more diverse data ingestion, but also requires more skills and resources to analyze the data.

use cases

A third key difference between data warehouse and data lake architectures is how they support different use cases. A data warehouse is best suited for use cases that require structured and standardized data for reporting and analysis, such as dashboards, OLAP cubes.

A data warehouse can answer predefined and repeatable questions, such as "What is the monthly revenue by region?" or "How many customers bought product X in the last quarter?".

A data lake is best suited for use cases that require raw and unstructured data for exploration and discovery, such as machine learning, natural language processing, or sentiment analysis.

A data lake can answer ad-hoc and complex questions, such as "What are the main topics of customer reviews?" or "How can we predict customer churn based on behavior patterns?".

Example 4 : For a Supermarket Chain consider the following dimensions, namely Product, store, time , promotion. The schema contains a central fact tables sales facts with three measures unit_sales, dollars_sales and dollar_cost.

Design star schema and calculate the maximum number of base fact table records for the 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)

Promotion : a sold item may be in only one promotion in a store on a given day

(a) **Star schema :**

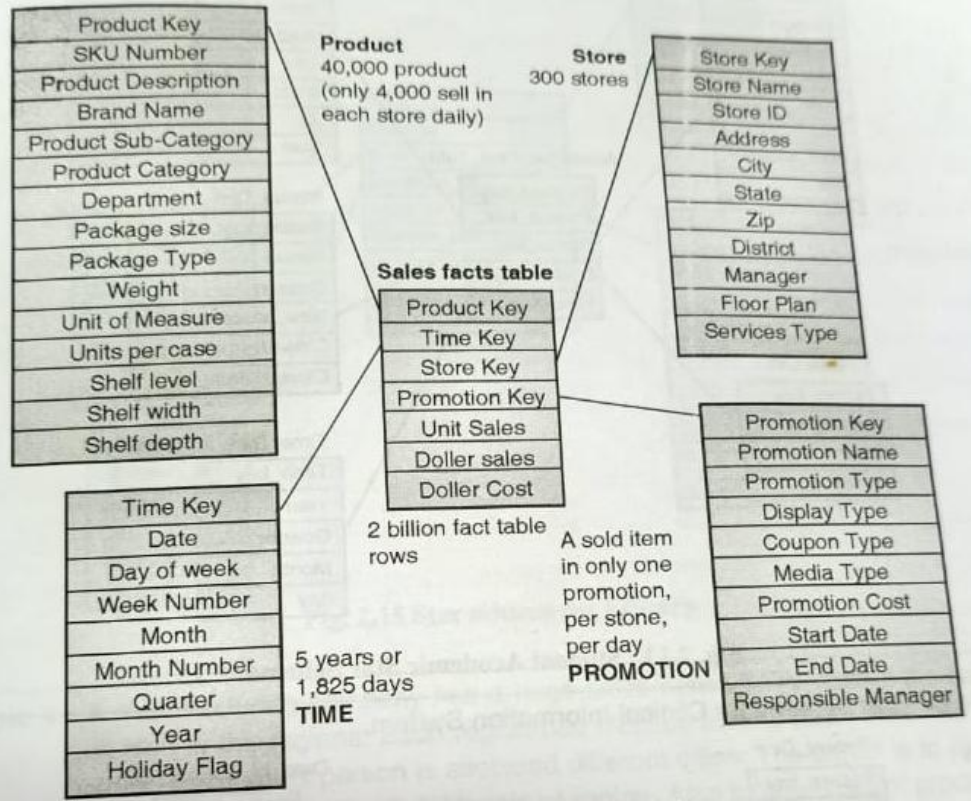


Fig. 2.12 : Sales Promotion Star Schema

(b) Time period = 5 years × 365 days = 1825

There are 300 stores,

Each stores daily sale = 4000

Promotion = 1

Maximum number of fact table records: $1825 \times 300 \times 4000 \times 1 = 2$ billion

2 : The Mumbai university wants you to help design a star schema to record grades for course completed by students. There are four dimensional tables namely course_section, professor, student, period with attributes as follows :

Course_section Attributes: Course_Id, Section_number, Course_name, Units, Room_id, Roomcapacity. During a given semester the college offers an average of 500 course sections

Professor Attributes: Prof_id, Prof_Name, Title, Department_id, department_name

Student Attributes: Student_id, Student_name, Major. Each Course section has an average of 60 students

Period Attributes: Semester_id, Year. The database will contain Data for 30 months periods. The only fact that is to be recorded in the fact table is course Grade

Answer the following Questions

(a) Design the star schema for this problem

(b) Estimate the number of rows in the fact table, using the assumptions stated above and also estimate the total size of the fact table (in bytes) assuming that each field has an average of 5 bytes.

(c) Can you convert this star schema to a snowflake schema ? Justify your answer and design a snowflake schema if it is possible.

Solution :

(a) Star Schema

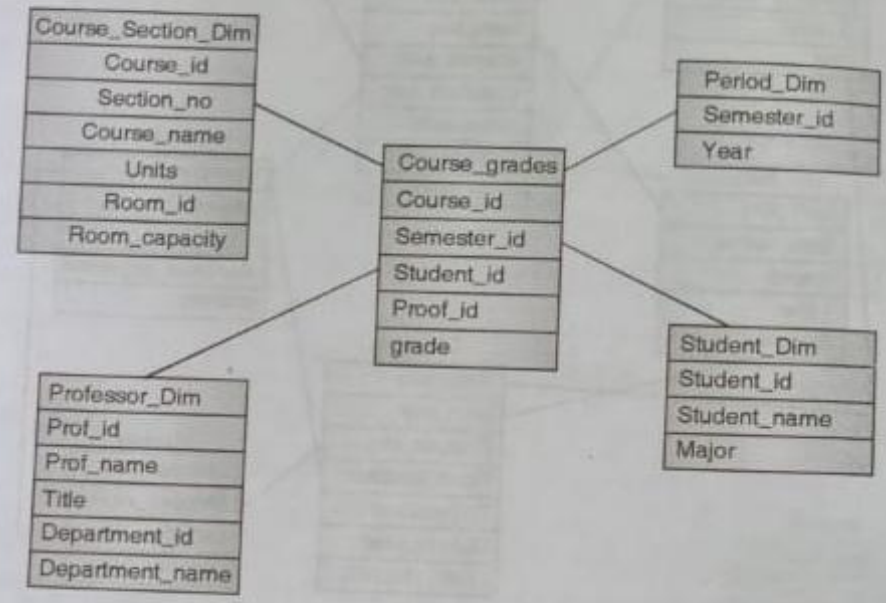


Fig. 2.9 : University Star Schema

(b) Total Courses Conducted by university = 500

Each Course has average students = 60

University stores data for 30 months

Total Student in University for all courses in 30 months = 500*60 = 30000

Time Dimension = 30 months = 5 Semesters (Assume 1 semester = 6 months)

Now, Number of rows of fact table = 30000*5 = 150000 (one student has 5 grades for 5 semesters)

(c) ***Snowflake Schema :*** Yes, the above star schema can be converted to a snowflake schema, considering the following assumptions

- Courses are conducted in different rooms, so course dimension can be further normalized to rooms dimension as shown in the

- Professor belongs to a department, and department dimension is not added in the star schema, so professor dimension can be further normalized to department dimension.
- Similarly students can have different major subjects, so it can also be normalized as shown in the figure below.

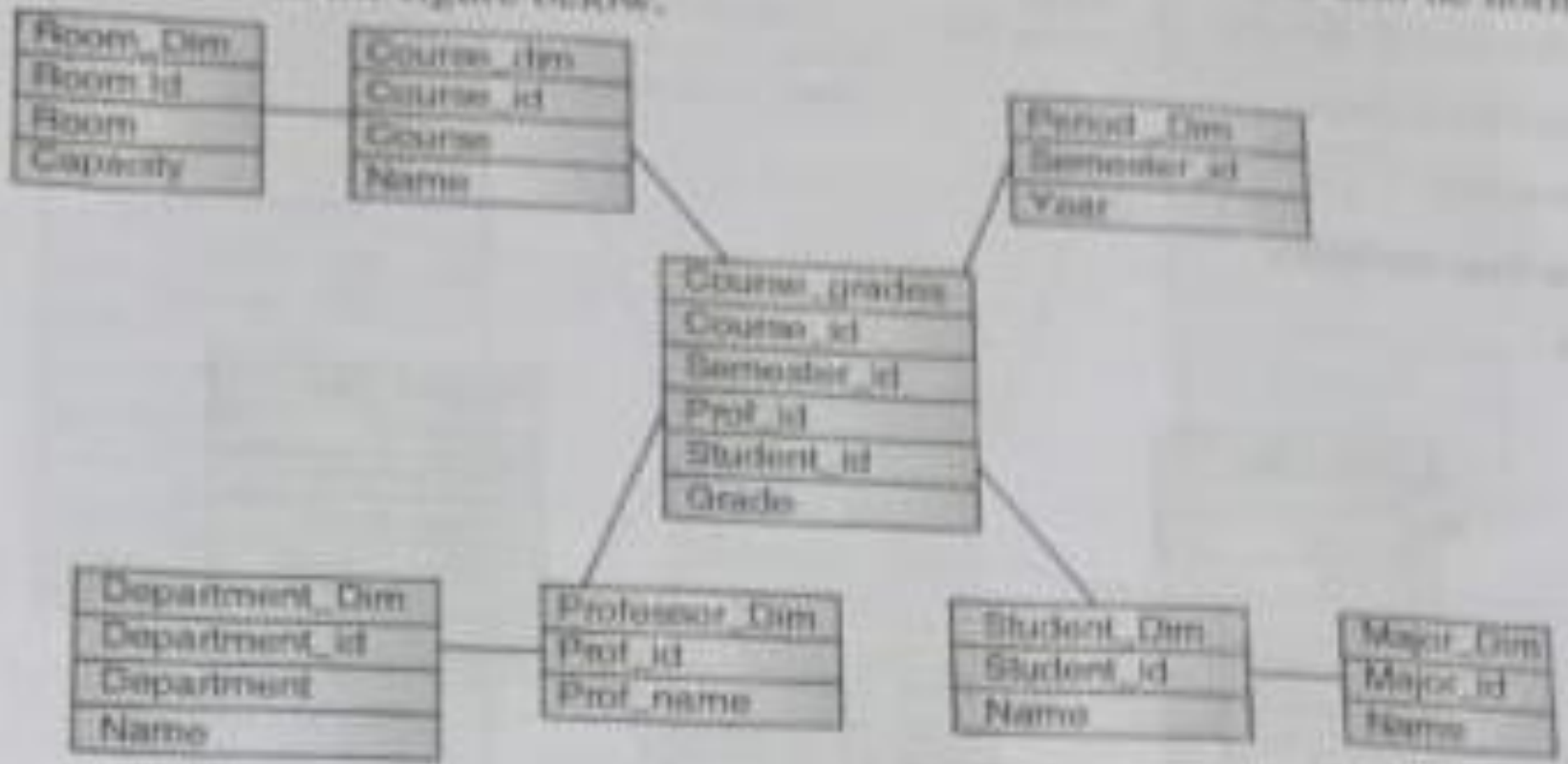


Fig. 2.10 : University Snowflake Schema

Example 3 : Give Information Package for recording information requirements for "Hotel Occupancy" considering dimensions like Time, Hotel etc. Design star schema from the information package.

Table 2.5 : Information Package for Hotel Occupancy

Hotel	Room Type	Time
Hotel Id	Room id	Time id
Branch Name	room type	Year
Branch Code	room size	Quarter
Region	number of beds	Month
Address	type of bed	Date
city/stat/zip	max occupants	day of week
construction year	Suite	day of month,
renovation year		holiday flag

Facts

(a) Occupied Rooms
(b) Vacant Rooms
(c) Unavailable Rooms
(d) No of occupants
(e) Revenue

Draw the Star Schema

Solution :

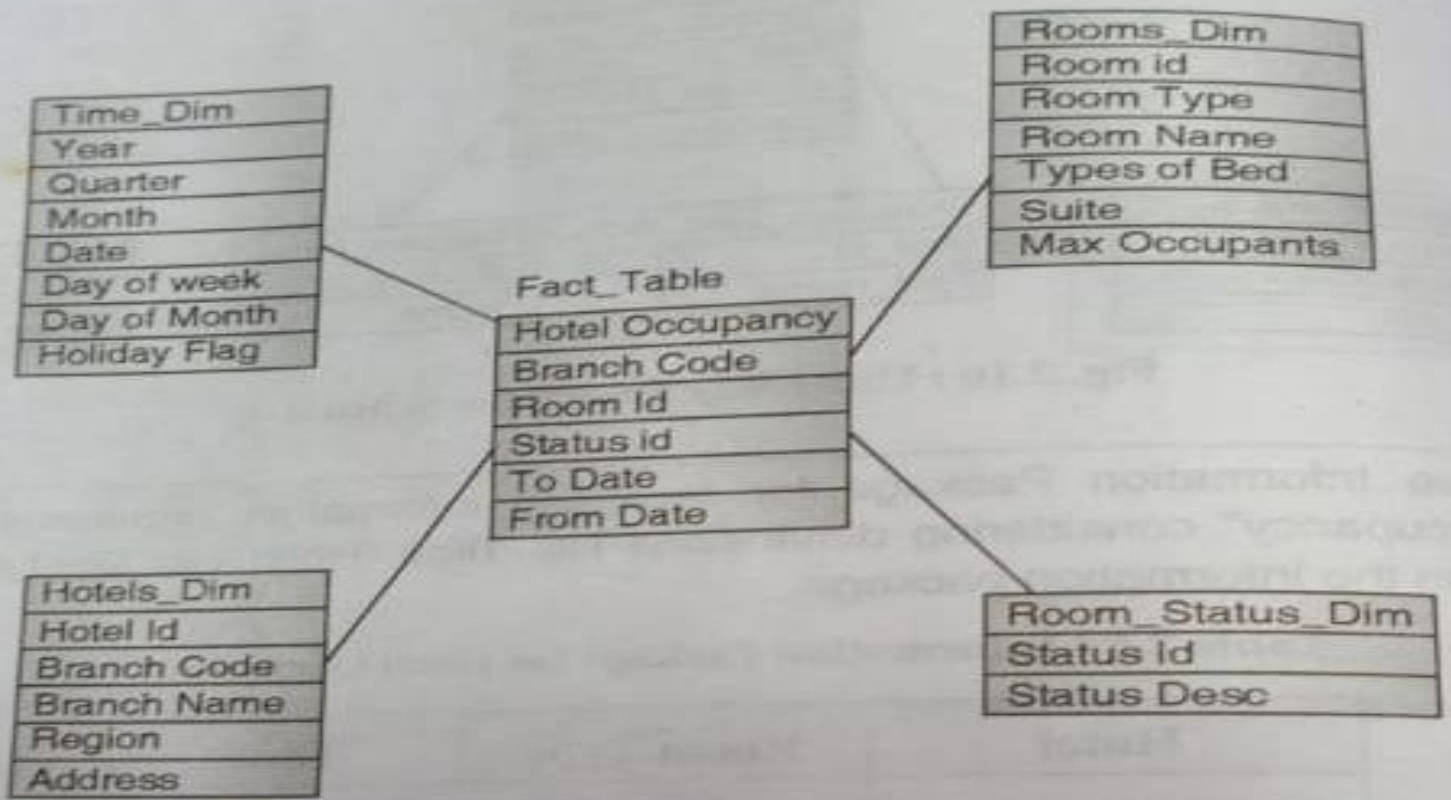


Fig. 2.11 : Hotel Occupancy Star Schema