Subject Name: DATA WAREHOUSING AND MINING

Unit No:1

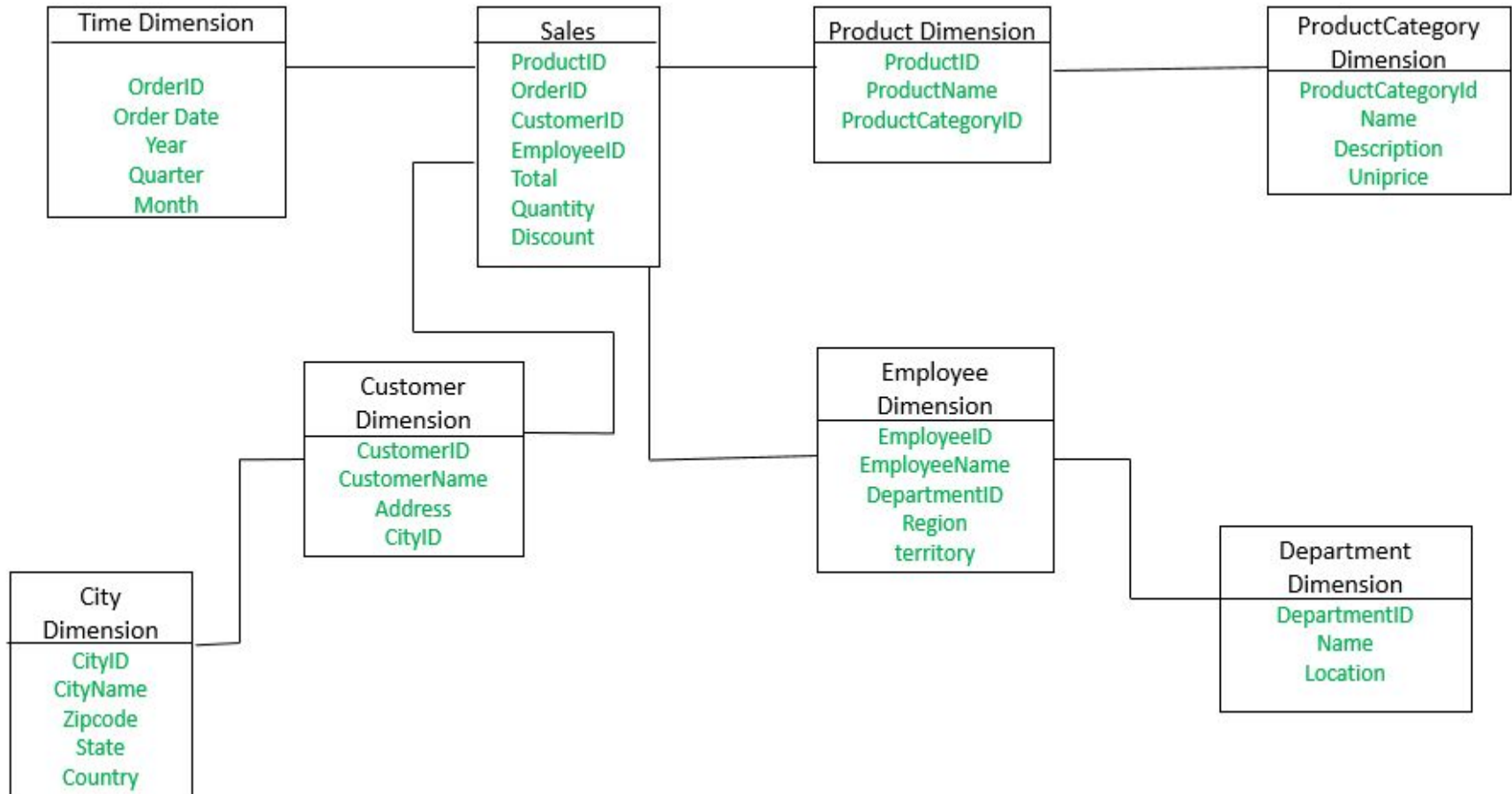**Unit Name: Data Warehousing Fundamentals**

Index

Snowflake Schema in Data Warehouse Model

- The snowflake schema is a **variant of the star schema**.

- Here, the centralized fact table is connected to multiple dimensions.

- In the snowflake schema, **dimension are present in a normalized from in multiple related tables.**

- The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent table.

- The **snowflake effect affects only the dimension tables and does not affect the fact tables**

Snowflake Schema in Data Warehouse Model

- The snowflake schema is a **variant of the star schema**.

- Here, the centralized fact table is connected to multiple dimensions.

- In the snowflake schema, **dimension are present in a normalized from in multiple related tables.**

- The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent table.

- The **snowflake effect affects only the dimension tables and does not affect the fact tables**

Snowflake Schema in Data Warehouse Model

Time Dimension
OrderID
Order Date
Year
Quarter
Month

Sales
ProductID
OrderID
CustomerID
EmployeeID
Total
Quantity
Discount

Product Dimension
ProductID
ProductName
ProductCategoryID

ProductCategory Dimension
ProductCategoryId
Name
Description
Uniprice

Customer Dimension
CustomerID
CustomerName
Address
CityID

Employee Dimension
EmployeeID
EmployeeName
DepartmentID
Region
territory

City Dimension
CityID
CityName
Zipcode
State
Country

Department Dimension
DepartmentID
Name
Location

Lecture 4: Snowflake Schema ,Factless Fact Table

Snowflake Schema in Data Warehouse Model

- The **Employee dimension table** now contains the attributes: EmployeeID, EmployeeName, DepartmentID, Region, Territory.

- The **DepartmentID** attribute links with Employee table with the Department dimension table.

- The **Department dimension** is used to provide detail about each department, such as Name and Location of the department.

- The **Customer dimension table** now contains the attributes: CustomerID, CustomerName, Address, CityID.

- The **CityID** attributes links the Customer dimension table with the City dimension table.

- The **City dimension table** has details about each city such as CityName, Zipcode, State and Country.

Lecture 4: Snowflake Schema ,Factless Fact Table

Snowflake Schema in Data Warehouse Model

- The **main difference between star schema and snowflake schema** is that the **dimension table of the snowflake schema are maintained in normalized form to reduce redundancy.**

- The advantage here is that such table(normalized) are easy to maintain and save storage space.

- However, it also means that more joins will be needed to execute query. This will adversely impact system performance.

Lecture 4: Snowflake Schema ,Factless Fact Table

Snowflake Schema in Data Warehouse Model

Advantages: There are two main advantages of snowflake schema given below:

- It provides structured data which **reduces the problem of data integrity**.
- It uses **small disk space because data are highly structured**.

Disadvantages:

- Snowflaking **reduces space consumed** by dimension tables, **but compared with the entire data warehouse the saving is usually insignificant.**
- **Avoid snowflaking** or normalization of a dimension table, **unless required** and appropriate.
- **Do not snowflake hierarchies of one dimension table into separate tables.** Hierarchies should belong to the dimension table only and should never be snowfalked.
- Multiple hierarchies can belong to the same dimension has been designed at the lowest possible detail.

Factless Fact Table

- A fact table without any measures is known as factless fact table. It's basically an intersection of dimension.

- The concept of factless fact table does not make sense & seems to be of not much use because of fact table. Essentially is all about facts, & there are no facts in a factless fact table

- However there are circumstances where using a factless fact table makes sense in data ware housing.

- Factless fact table provide flexibility in data warehouse design.

- It contains many-many relationships between dimensions.

- These table do not contain numeric textual facts

Lecture 4: Snowflake Schema ,Factless Fact Table

Commonly used Examples

- In tables such as keeping the attendance record students.

- Identifying product promotion events

- Tracking attendance of students or registration events.

- Tracking insurance related accident events.

Lecture 4: Snowflake Schema ,Factless Fact Table

Example

- Think about a record of student attendance in classes.

- In this case, the fact table would consist of 3 dimensions: the student dimension, the time dimension, and the class dimension.

- This factless fact table would look like the following:

FACT_ATTENDANCE

STUDENT_ID
CLASS_ID
TIME_ID

FIGURE 2.2 STUDENT ATTENDANCE

For example, one can easily answer the following questions with this factless fact table:

• How many students attended a particular class on a particular day?

• How many classes on average does a student attend on a given day?

Without using a factless fact table, we will need two separate fact tables to answer the above two questions.

Types of Fact Table

There are two types:-

- Factless fact table for events.

- Factless fact table for conditions.

Factless fact table for events.

- Factless fact table for events is a table that records events.

- In dimensional data warehouse, numerous event-tracking tables appear to be factless sometimes.

- There may be a situation where no fact seems to be related to an important business process & you may have events that you want to track, but you cannot find any measurements. In such situations, create a typical transact-grained fact table that comprises no facts.

Lecture 4: Snowflake Schema ,Factless Fact Table

Factless fact table for conditions:-

- If there are no clear transactions, factless fact tables are used to design the conditions or other important relationships among the different dimensions.

- A factless fact table helps in creating analysis reports that comprises negative aspects of a business. For e.g. Book store that did not sell a single book for a given period.

Lecture 4: Snowflake Schema ,Factless Fact Table

Fact Constellation in Data Warehouse modelling

- Fact Constellation is a schema for **representing multidimensional model**.

- It is a **collection of multiple fact tables having some common dimension tables.**

- It **can be viewed as a collection of several star schemas** and hence, also known as **Galaxy schema.**

- It is **one of the widely used schema for Data warehouse designing** and it is much more complex than star and snowflake schema.

- **For complex systems, we require fact constellations.**

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

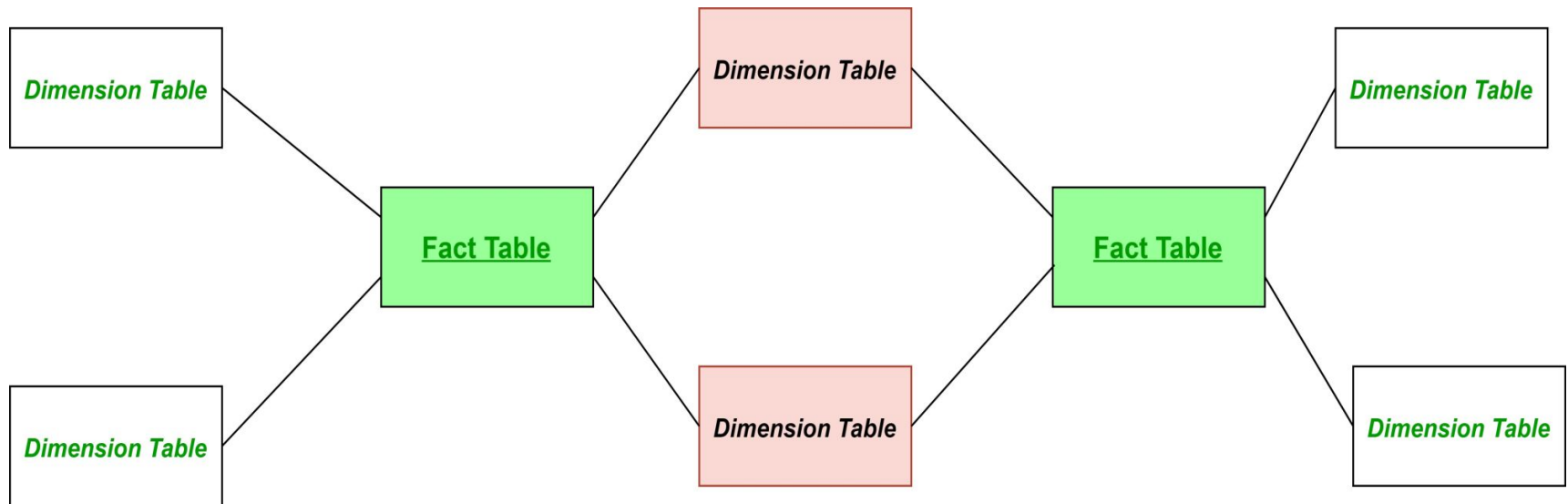# Fact Constellation in Data Warehouse modelling

Figure – General structure of Fact Constellation
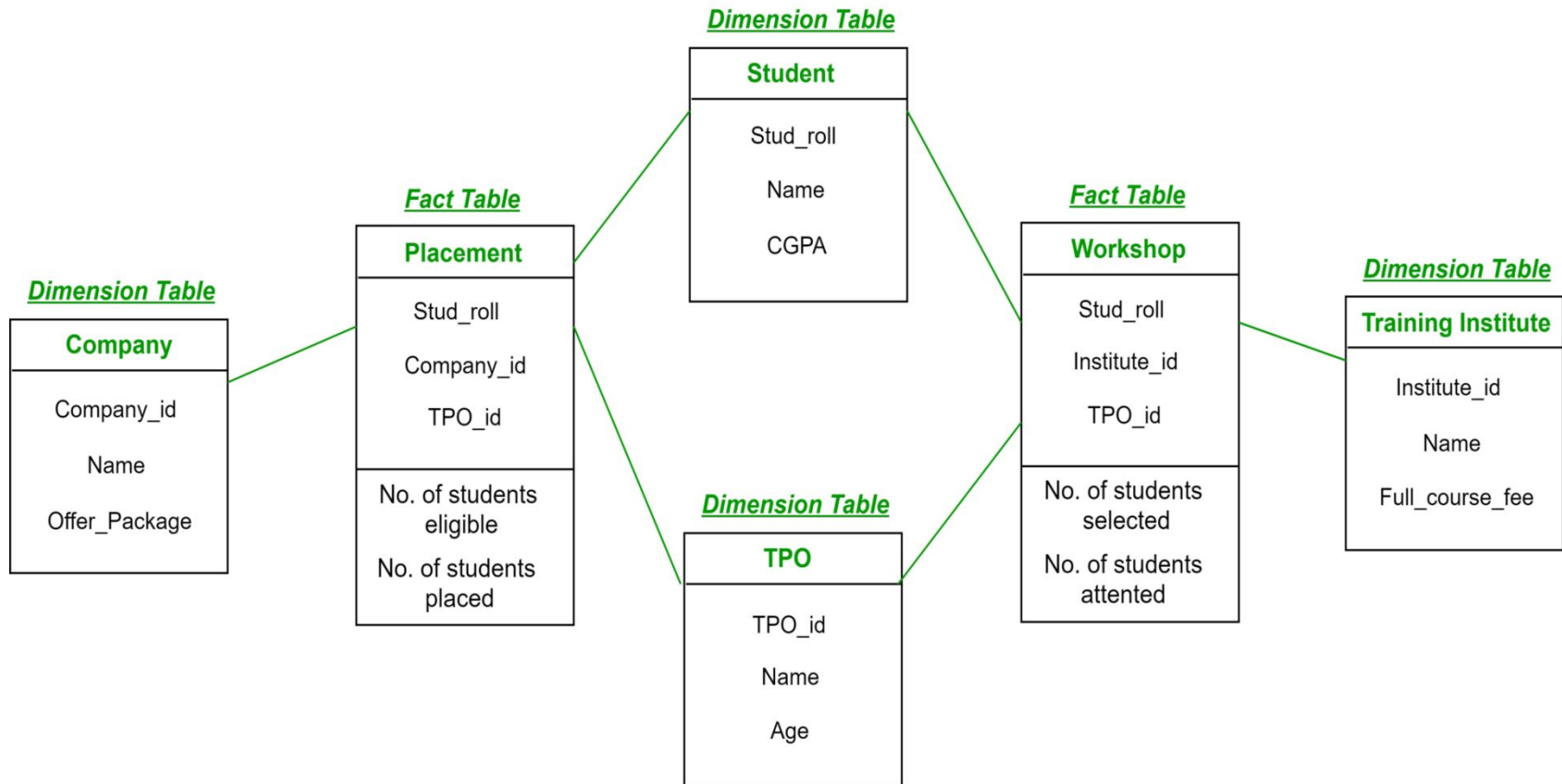
- Here, the pink colored Dimension tables are the common ones among both the star scheme's

- Green colored fact tables are the fact tables of their respective star scheme's

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Fact Constellation in Data Warehouse modelling

Example

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Fact Constellation in Data Warehouse modelling

In above demonstration:
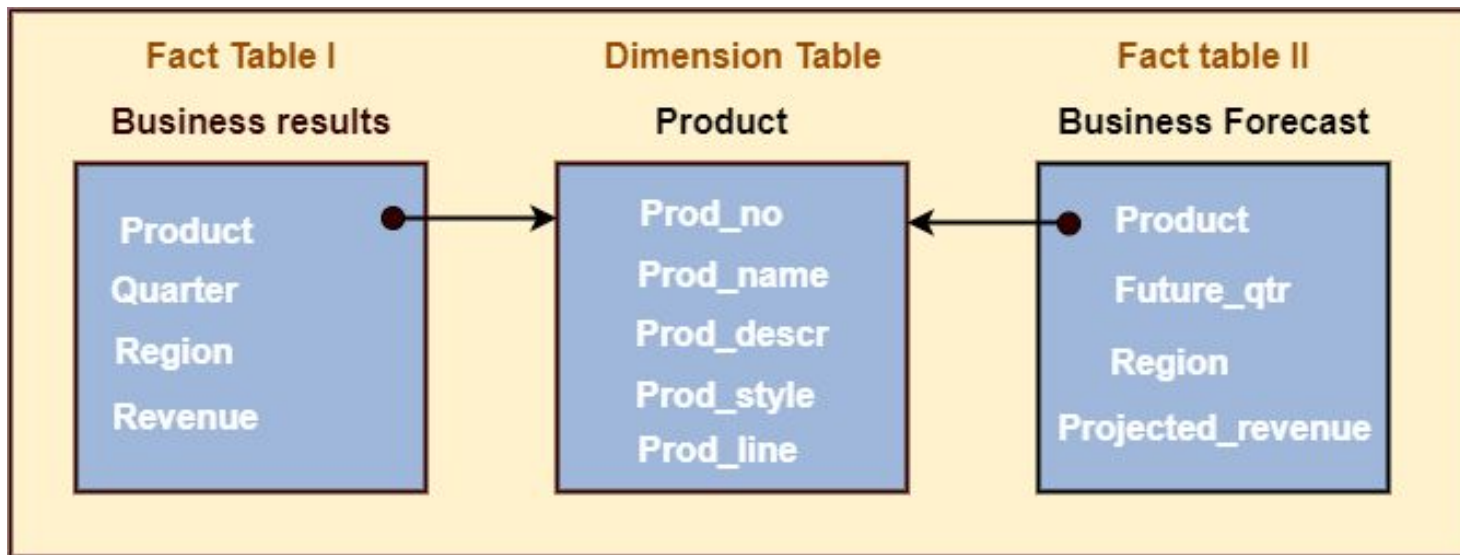
- **Placement** is a **fact table** having attributes: (Stud_roll, Company_id, TPO_id) with facts: (Number of students eligible, Number of students placed).

- **Workshop** is a **fact table** having attributes: (Stud_roll, Institute_id, TPO_id) with facts: (Number of students selected, Number of students attended the workshop).

- **Company is a dimension table** having attributes: (Company_id, Name, Offer_package).

- **Student is a dimension table** having attributes: (Student_roll, Name, CGPA).

- **TPO is a dimension table** having attributes: (TPO_id, Name, Age).

- **Training Institute is a dimension table** having attributes: (Institute_id, Name, Full_course_fee).

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Fact Constellation in Data Warehouse modelling

- So, there are **two fact tables** namely, **Placement** and **Workshop** which are part of two different star schemas having dimension tables – Company, Student and TPO in Star schema with fact table Placement and dimension tables – Training Institute, Student and TPO in Star schema with fact table Workshop.

- Both the **star schema have two dimension tables common** and hence, **forming a fact constellation or galaxy schema.**

- **Advantage**: Provides a **flexible schema**.

- **Disadvantage**: It is much more complex and hence, **hard to implement and maintain**.

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Example 2

- Fact Constellation Schema describes a logical structure of data warehouse or data mart.
- Fact Constellation Schema can design with a collection of de-normalized FACT, Shared, and Conformed Dimension tables.

FACT Constellation Schema

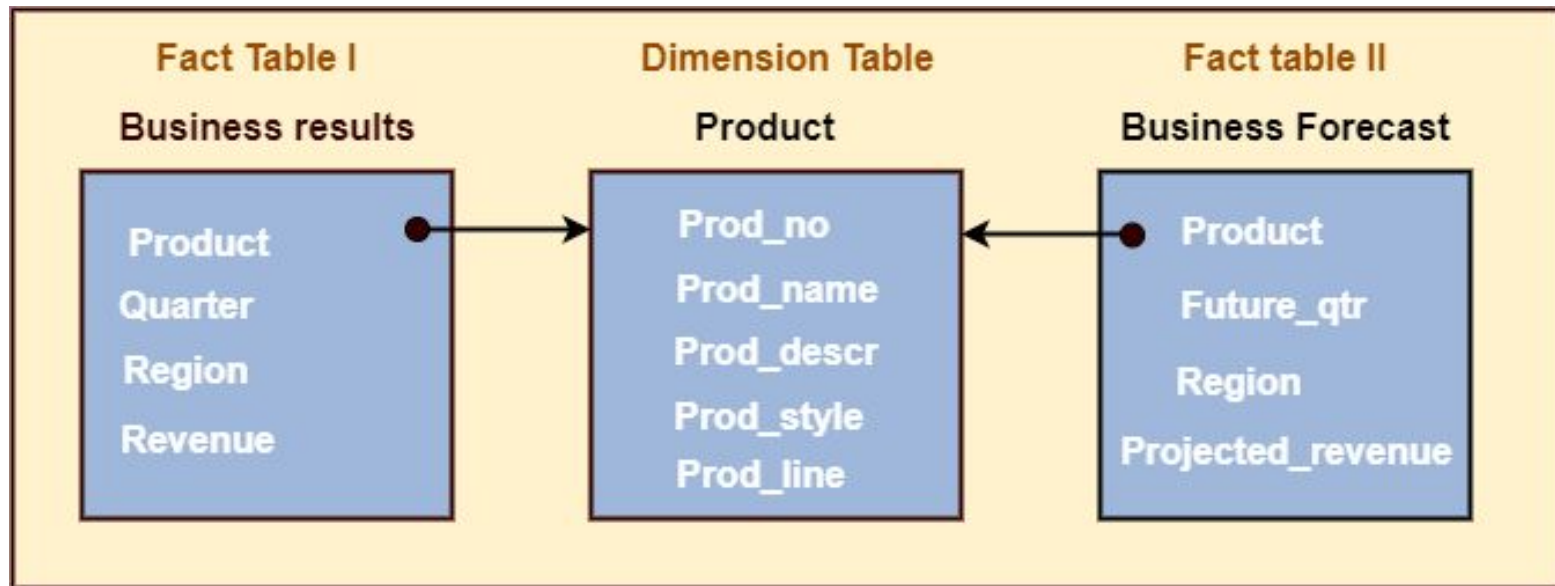Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Example 2

- Fact Constellation Schema describes a logical structure of data warehouse or data mart.

- Fact Constellation Schema can design with a collection of de-normalized FACT, Shared, and Conformed Dimension tables.

FACT Constellation Schema

Lecture 5: Fact Constellation Schema. Update to the dimension tables.

Exmaple 2

Time dimension table

time_key
day
day-of-week
month
quarter
year

Sales fact table

time_key
item key
branch_key
location_key
Rupee_sold
units_sold

Item dimension table

item_key
item name
brand
type
Supplier_type

Shipping fact table

item_key
time_key
shipper_key
from_location
to location
Rupee_cost
units_shpipped

Shipper dimension table

shipper_key
shipper_name
location_key
shipper_type

Location dimension table

location key
street
city
province_or_state
Country

Branch dimension table

branch_key
branch_name
branch_type

24 Lecture 5: Fact Constellation Schema. Update to the dimension tables.
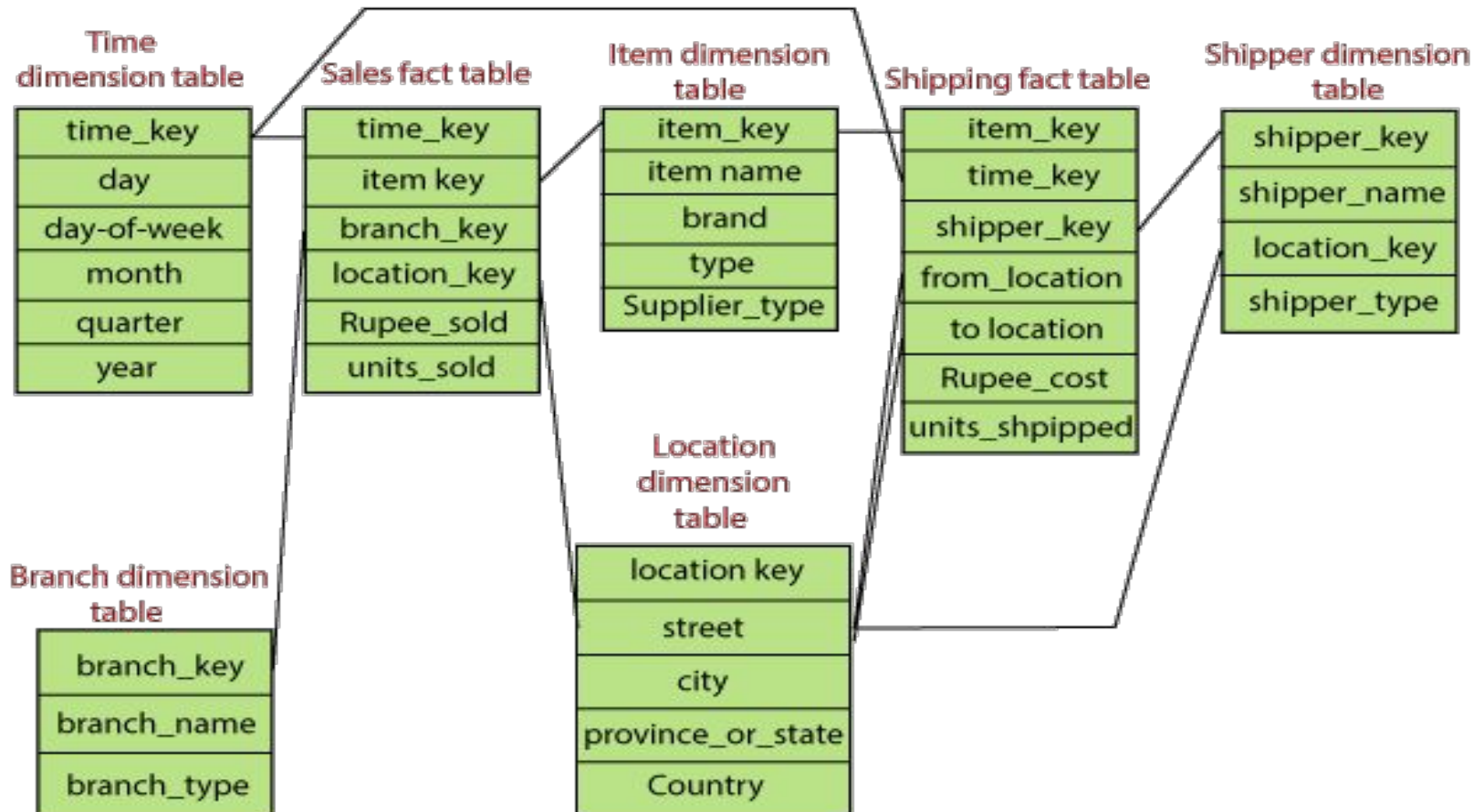
Cont.

- This schema defines two **fact tables, sales, and shipping.**

- Sales are treated along four dimensions, namely, time, item, branch, and location.

- The schema contains a fact table for sales that includes **keys to each of the four dimensions**, along with two measures: Rupee_sold and units_sold.

- The shipping table has **five dimensions, or keys:** item_key, time_key, shipper_key, from_location, and to_location, and

- two measures: Rupee_cost and units_shipped.

Problem Statements to design star and snowflake schema

Types Of Schemas:- Star Schema

> Each dimension in a star schema is represented with a **one-dimension table** which contains a set of attributes.

> **Fact table** is at the center. which contains keys to every dimension table & attributes like: *units sold* and *revenue*.

Dimension Table

Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT

Dimension Table

Date Dim
Date_ID
Year
Month
Quarter
Date

Fact Table

Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue

Dimension Table

Branch Dim
Branch _ID
Name
Address
Country

Dimension Table

Product
Product_ID
Product_Name
Model_ID
Variant_ID

Lecture 5: Fact Constellation Schema. Update to the dimension tables.
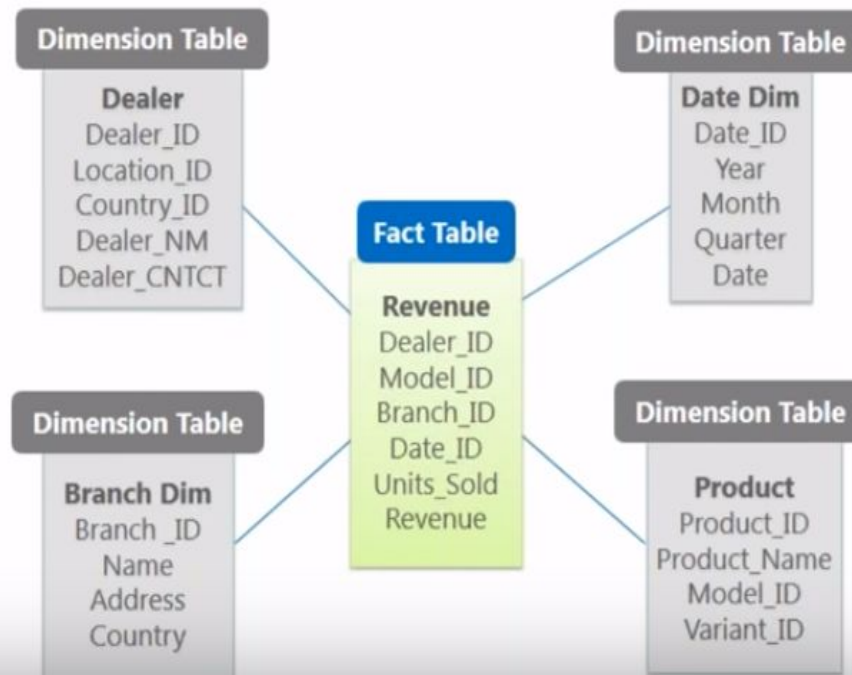
Problem Statement 1

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.

Lecture 8 - Problem statement given to student to design star scheme, snowflake schema
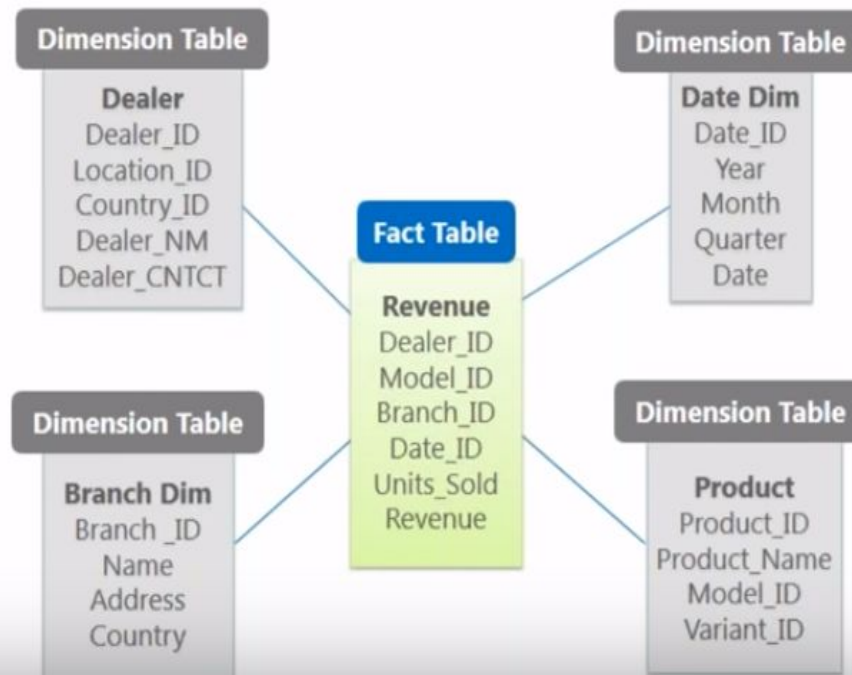
Problem Statement 2

- All Electronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions Date, Dealer , branch, and Product. These dimensions allow the store to keep track of things like sales of items and items were sold

Lecture 8 - Problem statement given to student to design star scheme, snowflake schema

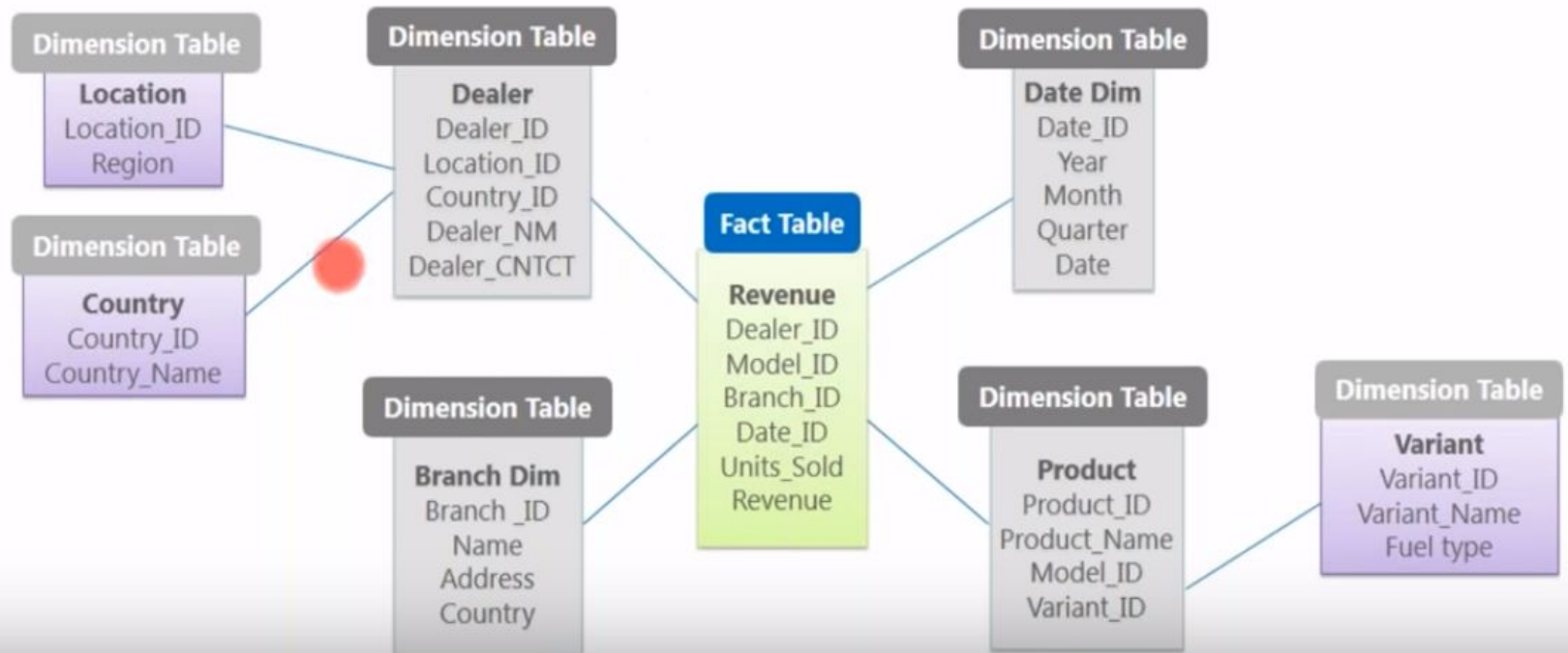## One More Example of Star Schema

Types Of Schemas:- Star Schema

> Each dimension in a star schema is represented with a **one-dimension table** which contains a set of attributes.
> **Fact table** is at the center. which contains keys to every dimension table & attributes like: *units sold* and *revenue*.

Dimension Table

Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT

Dimension Table

Date Dim
Date_ID
Year
Month
Quarter
Date

Fact Table

Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue

Dimension Table

Branch Dim
Branch _ID
Name
Address
Country

Dimension Table

Product
Product_ID
Product_Name
Model_ID
Variant_ID

Lecture 8 - Problem statement given to student to design star scheme, snowflake schema

One More example of Snowflake Schema
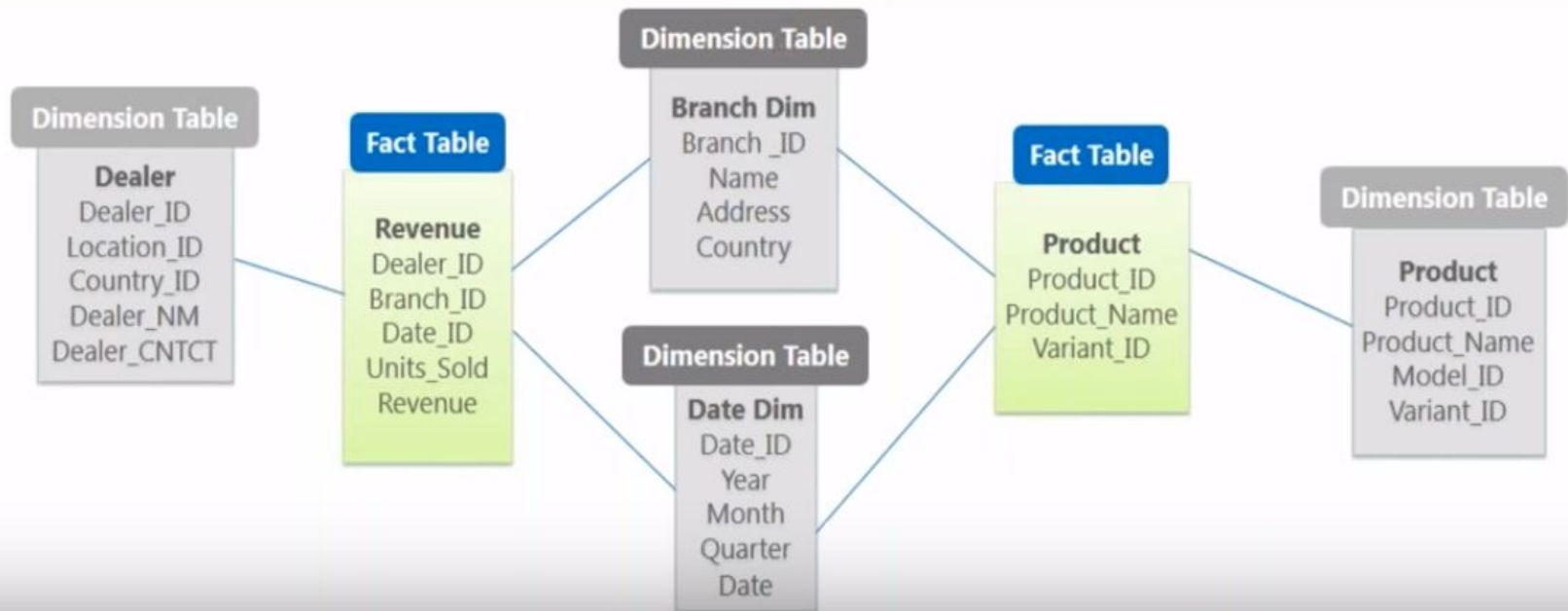

Types Of Schemas:- Snowflake Schema

➤ Dimension tables in the **Snowflake schema** are **normalized**. (Split into additional tables).

➤ **Dealer** dimension table is split into **Location** & **Country**. Product dimension table is split into **Product** & **Variant**.

Dimension Table
Location
Location_ID
Region

Dimension Table
Country
Country_ID
Country_Name

Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT

Dimension Table
Branch Dim
Branch _ID
Name
Address
Country

Fact Table
Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue

Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date

Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID

Dimension Table
Variant
Variant_ID
Variant_Name
Fuel type

Lecture 8 - Problem statement given to student to design star scheme, snowflake schema

One More example of Galaxy Schema

Types Of Schemas:- Galaxy Schema

➤ Also known as **Fact Constellation** schema. Contains more than **1 Fact table**.

➤ Below, there are two fact tables: **Revenue** and **Product**.

➤ Dimensions which are shared are called **Conformed Dimensions**.

Dimension Table

Dimension Table

Fact Table

Branch Dim
Branch _ID
Name
Address
Country

Fact Table

Dimension Table

Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT

Revenue
Dealer_ID
Branch_ID
Date_ID
Units_Sold
Revenue

Product
Product_ID
Product_Name
Variant_ID

Product
Product_ID
Product_Name
Model_ID
Variant_ID

Dimension Table

Date Dim
Date_ID
Year
Month
Quarter
Date

31   Lecture 8 - Problem statement given to student to design star scheme, snowflake schema

Subject Name: <u>DATA WAREHOUSING AND MINING</u>

Unit No:2

Unit Name: Online Analytical Processing(OLAP) & ETL

Faculty Name :Mr.Tushar Ghorpade

Ms. Saguna Ingle

Ms. Dhanashri A. Bhosale

What Is OLAP and OLTP

What is OLAP?

- Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time.
- **The primary objective is data analysis and not data processing**.

What is OLTP?

- Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization.
- **The primary objective is data processing and not data analysis**

Example of OLAP System

Example of OLAP

Any Datawarehouse system is an OLAP system. Uses of OLAP are as follows:

• A company might compare their mobile phone sales in September with sales in October, then compare those results with another location which may be stored in a sperate database.

• Amazon analyzes purchases by its customers to come up with a personalized homepage with products which likely interest to their customer.

Example of OLTP System

Example of OLTP system

- An example of OLTP system is ATM center.

- Assume that a couple has a joint account with a bank. One day both simultaneously reach different ATM centers at precisely the same time and want to withdraw total amount present in their bank account.

- However, the person that completes authentication process first will be able to get money.

- In this case, OLTP system makes sure that withdrawn amount will be never more than the amount present in the bank.

- The key to note here is that OLTP systems are optimized for **transactional superiority instead data analysis.**

OLTP Applications

Other examples of OLTP applications are:

- Online banking
- Online airline ticket booking
- Sending a text message
- Order entry
- Add a book to shopping cart

Lecture no 7: OLTP Vs OLAP ,OLAP Operations

Thank You