upGrad

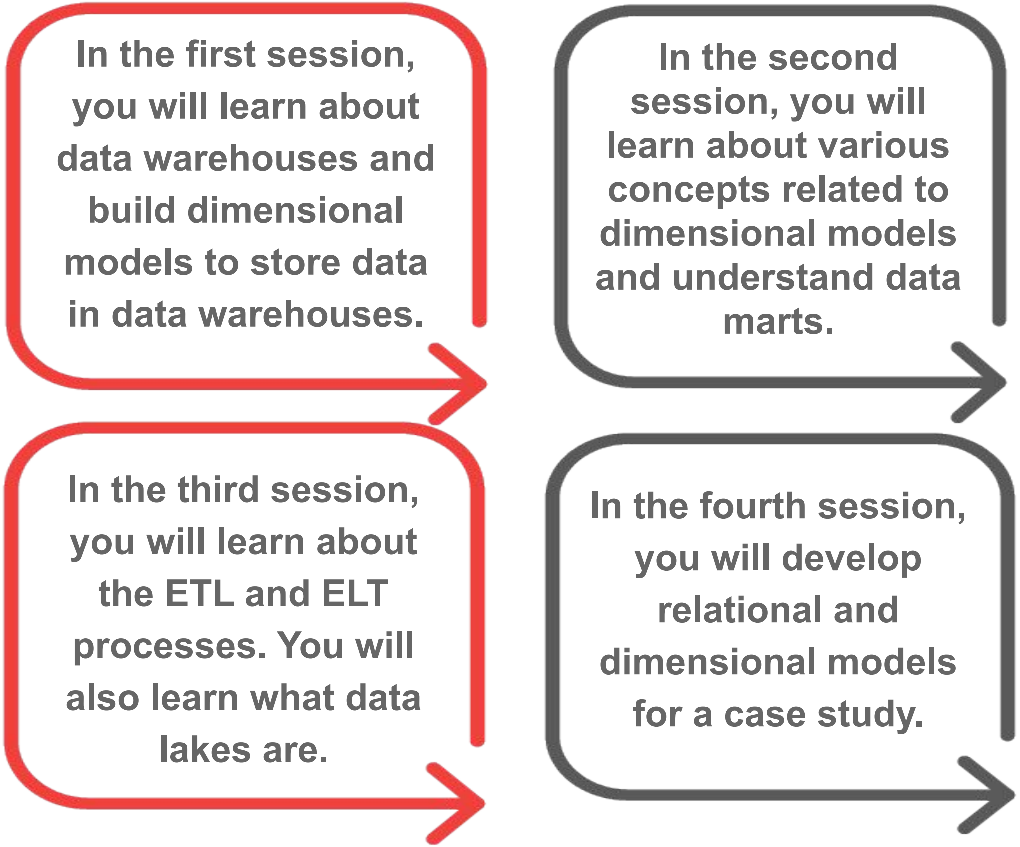
Data Warehousing and ETL

Course: Data Engineering - I

Lecture On: Data Warehousing and ETL

Instructor: Ganesh Nerale

Overview of the Module



In the first session, you will learn about data warehouses and build dimensional models to store data in data warehouses.

In the second session, you will learn about various concepts related to dimensional models and understand data marts.

In the third session, you will learn about the ETL and ELT processes. You will also learn what data lakes are.

In the fourth session, you will develop relational and dimensional models for a case study.

Session 1 | Introduction to Dimensional Modelling

Learning Objectives of the Session

Segment 03 Describing the OLTP systems and their use. Discussion on why relational databases cannot be used for analysis.

Segment 04 Need for a central repository in a company

Segment 05 Discussion on the steps used to build a dimensional model for a particular business process.

Segment 06 Choose the grain at which the data is stored in a particular dimensional model.

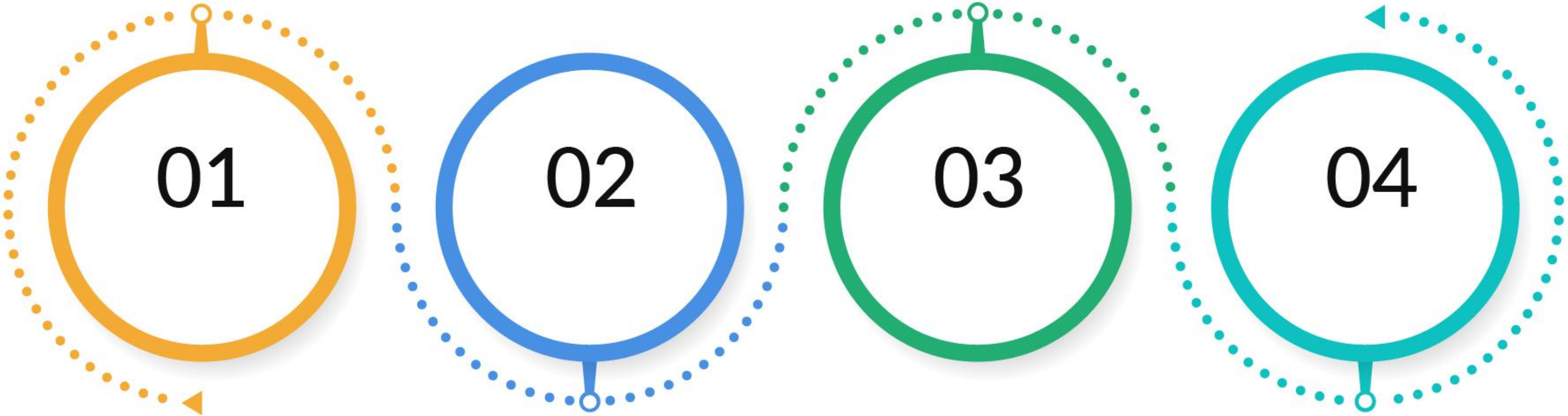
Segment 07 Understanding facts and dimensions in a dimensional model

Segment 3 | OLTP

Learning Objectives

What does OLTP mean?

Why can relational databases not be used for analysis?



What are the uses of relational databases?

Practical analysis of Relational vs Dimensional models

upGrad

Online Transaction Processing

What are Transactional Databases?

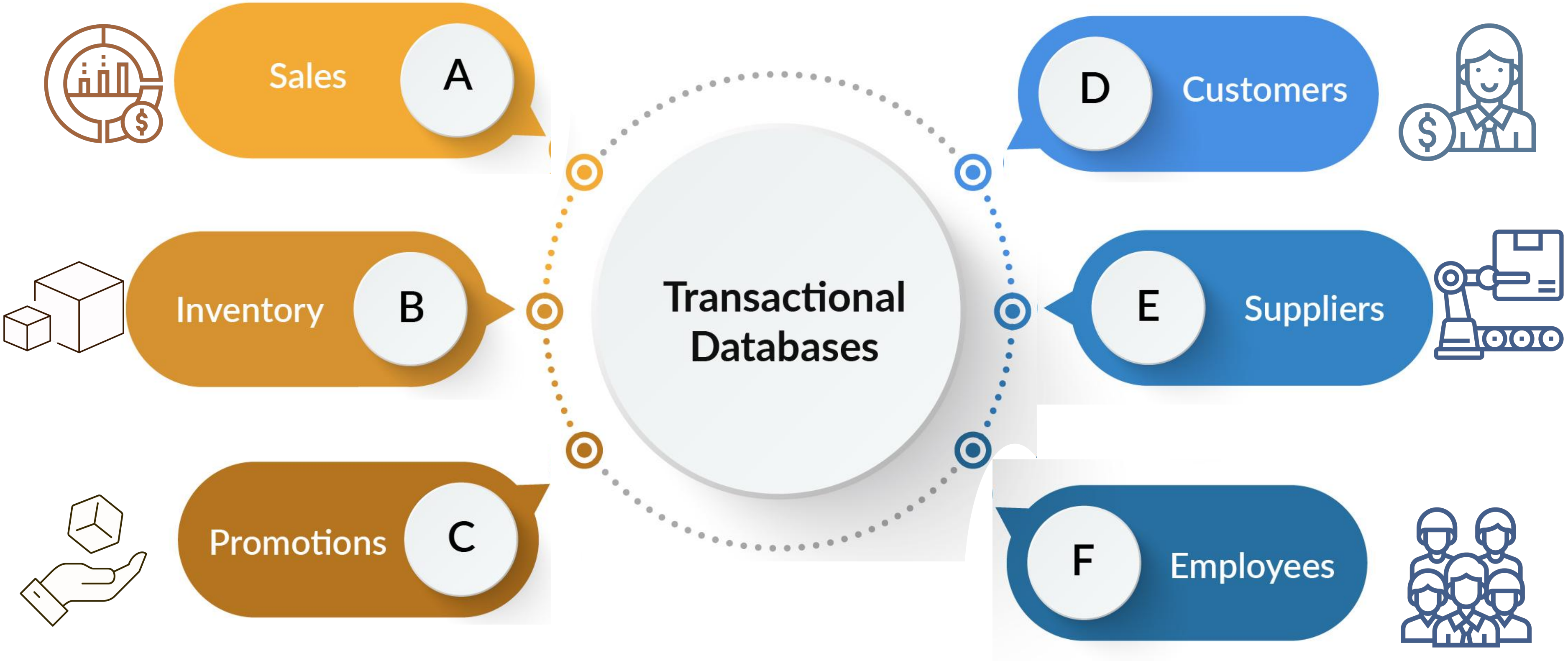
Where are they used?

1 Highly Available, Concurrent and fast response time

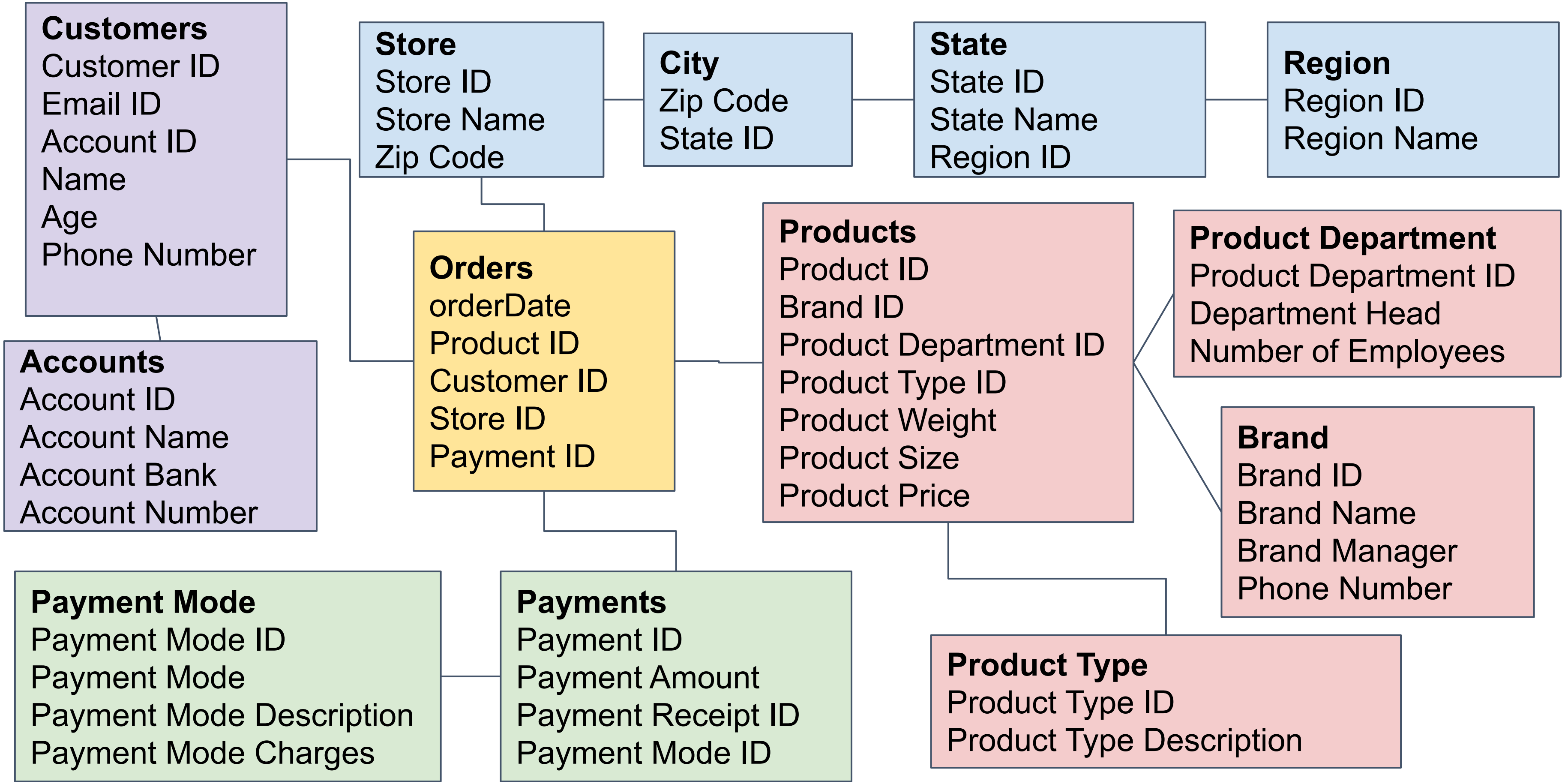
2 Normalised relational databases

3 Optimised for Search, Update, Deletion and Insertion operations

4 Not optimised for analysing and storing historical data


Sales
A
Inventory
B
Promotions
C
Transactional Databases
D
Customers
E
Suppliers
F
Employees

A Relational Database: For a Retail Store



Customers
Customer ID
Email ID
Account ID
Name
Age
Phone Number

Accounts
Account ID
Account Name
Account Bank
Account Number

Store
Store ID
Store Name
Zip Code

City
Zip Code
State ID

State
State ID
State Name
Region ID

Region
Region ID
Region Name

Orders
orderDate
Product ID
Customer ID
Store ID
Payment ID

Products
Product ID
Brand ID
Product Department ID
Product Type ID
Product Weight
Product Size
Product Price

Product Department
Product Department ID
Department Head
Number of Employees

Brand
Brand ID
Brand Name
Brand Manager
Phone Number

Payment Mode
Payment Mode ID
Payment Mode
Payment Mode Description
Payment Mode Charges

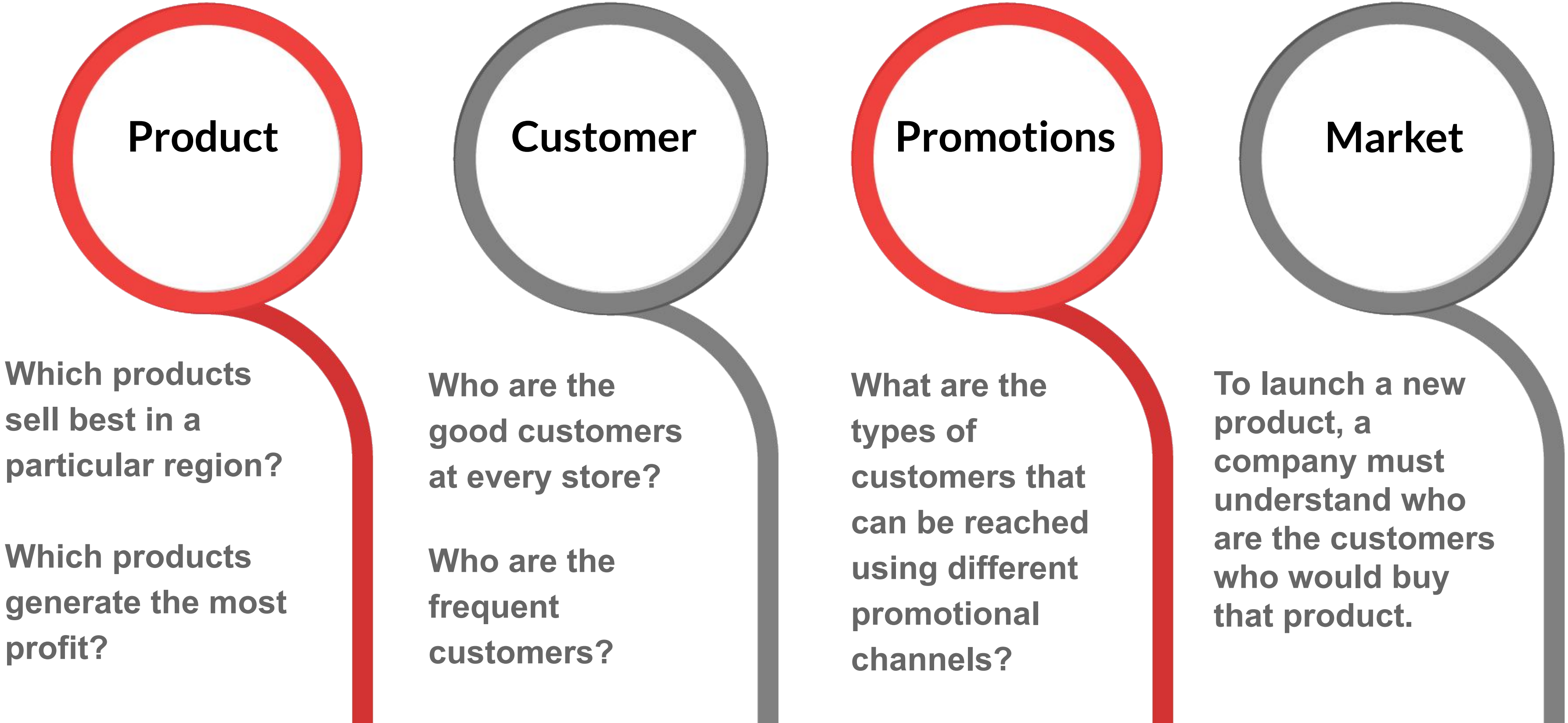
Payments
Payment ID
Payment Amount
Payment Receipt ID
Payment Mode ID

Product Type
Product Type ID
Product Type Description

Purposes of a Relational Database

1. Update every transaction at every store in real time
2. Maintain and manage orders
3. Update which customer has ordered which products?
4. Update which product belongs to which brand?
5. Update the product inventory as soon as order is complete.

Consider these Questions



Product

Which products sell best in a particular region?

Which products generate the most profit?

Customer

Who are the good customers at every store?

Who are the frequent customers?

Promotions

What are the types of customers that can be reached using different promotional channels?

Market

To launch a new product, a company must understand who are the customers who would buy that product.

Why can Relational Databases not be Used for Analysis?

Analysis has to be performed on historical data

Query results are not fast

The tables in relational database are in 3 NF. To store such a large amount of historical data in these many tables consumes more memory.

The query results are not fast, as many tables have to be joined to retrieve the data, and this would consume a significant amount of system resources.

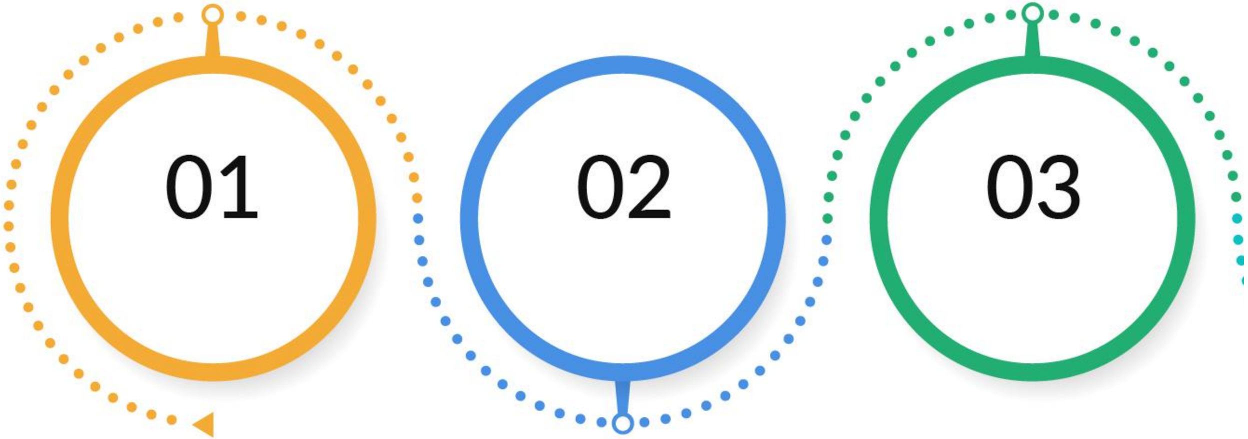
Summary | OLTP

01 An OLTP system has relational database to store the real-time transactions. The tables are in 2NF or 3NF.

02 Since the number of tables are more and the querying process is not fast, relational databases are not used for analyzing data.

Segment 4 | Data Warehouses

Learning Objectives



Need for data warehouses

Features of a data warehouse

OLAP at a glance

The data is not present in a central repository

The company wants to analyse only what is subject

The company wants to know the correct metrics that are to be analysed

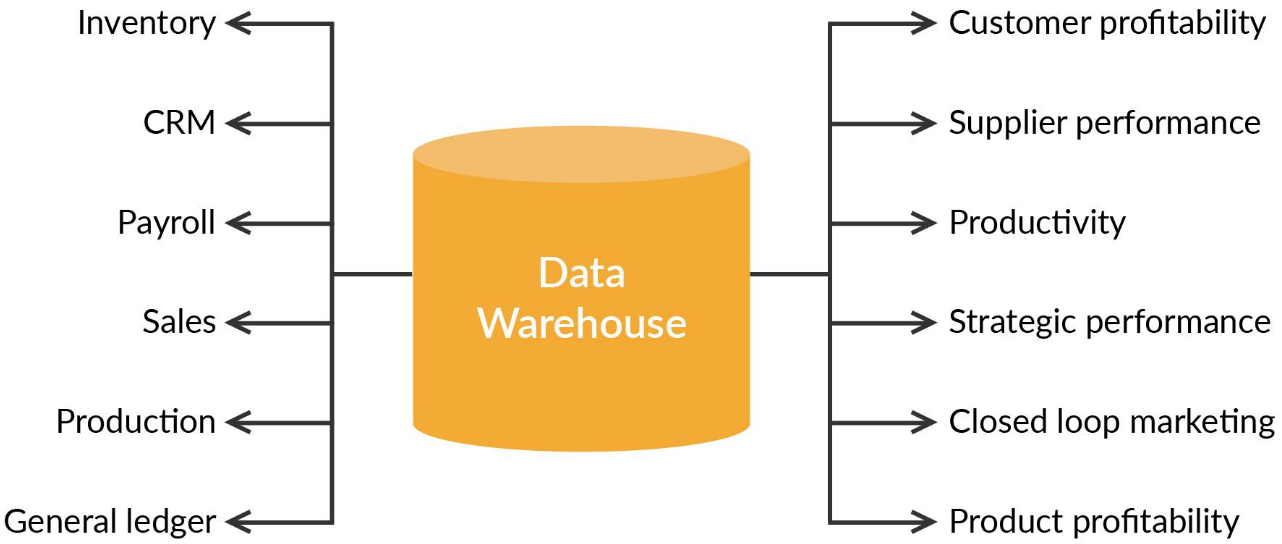
The company has not designed processes to access the data storages

The company wants to analyse the data in every possible way

Need for Data Warehouse

Why Data Warehouses?

A data warehouse integrates the data from one or more databases, so that analysis can be performed to obtain results.



Inventory

CRM

Payroll

Sales

Production

General ledger

Data Warehouse

Customer profitability

Supplier performance

Productivity

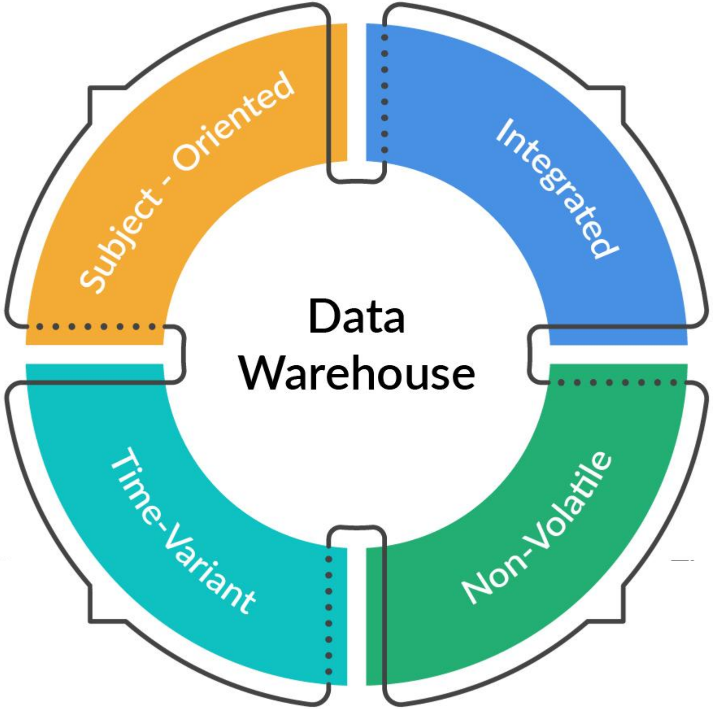
Strategic performance

Closed loop marketing

Product profitability

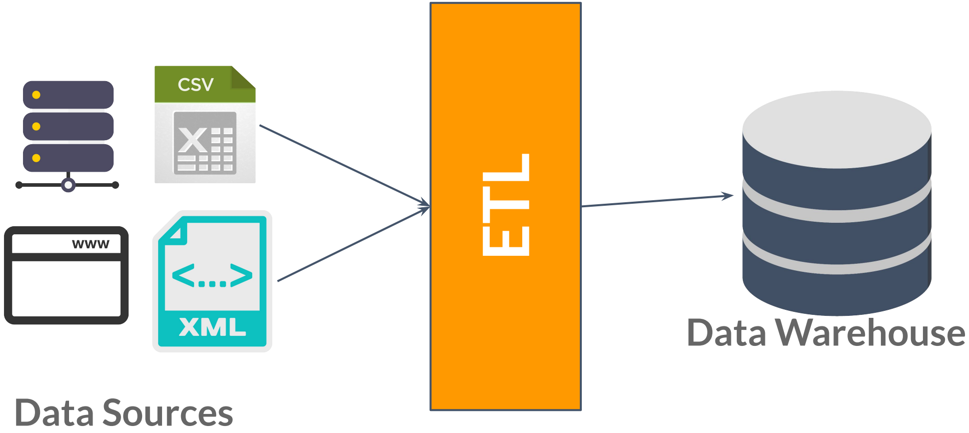
A Central Data Repository

Features of Data Warehouse



OLAP - OnLine Analytical Processing

A data warehouse aggregates data across an organisation, from multiple sources, and then selects, organises and aggregates it for efficient comparison and analysis.



- **Data Extraction:** Involves gathering data from multiple, heterogeneous sources
- **Data Cleaning:** Involves finding and correcting errors in the data
- **Data Transformation:** Involves converting the data from legacy format to warehouse format
- **Data Loading:** Involves sorting, summarising, consolidating and checking the integrity of data, and building indices and partitions
- **Refreshing:** Involves updating from the data sources to the warehouse

Summary | Data Warehouse

01 A Data Warehouse is a central repository that stores data.

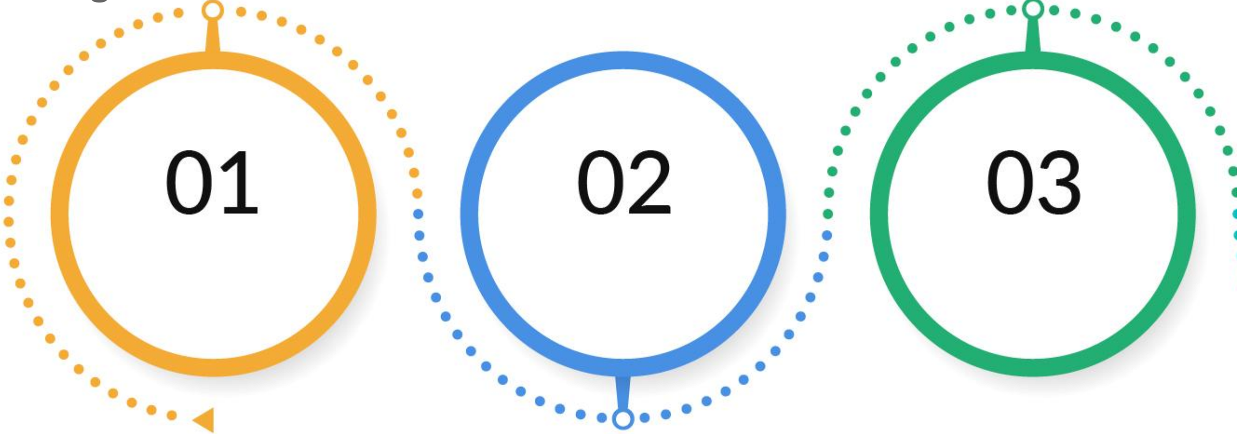
02 Data Warehouses integrate historical data based on one subject.

03 OLAP systems are used for data analysis. The data is extracted, transformed and loaded into data warehouses.

Segment 5 | Introduction to Dimensional Modelling

upGrad

Learning Objectives



Defining the four steps of building a dimensional model

Defining an example to understand the four steps

Understanding the first step

An Example Case Study

upGrad

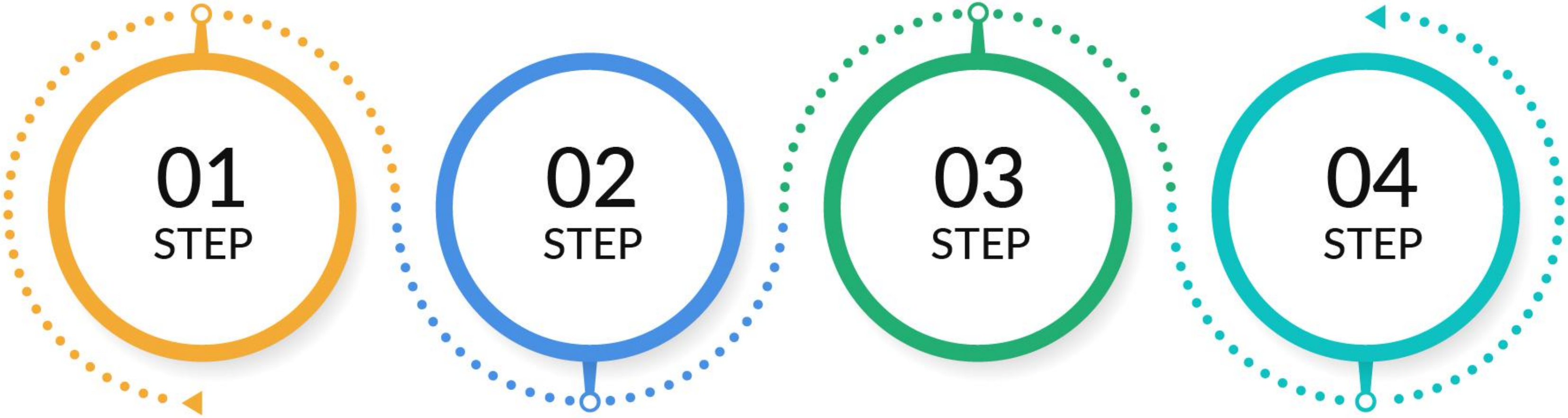
01 There is a hypothetical company named upGrad Fashions, which represents many clothing brands together. It has many local stores across the country.

02 To make data-driven decisions, they are looking to build a data warehouse to get answers to various business questions.

03 upGrad Fashions wants to analyse the sales at their local stores in order to maximise their profits.

Four Steps to a Dimensional Model

Dimensional Model is a database design to store data in a format that supports data analysis. Data is stored in data warehouse using facts and dimension tables of a dimensional model



Define the Business Process

01 STEP

02 STEP

Define the Grain

Define the Dimensions

03 STEP

04 STEP

Define the Facts

upGrad

Define the Business Process

1 The purpose of your dimensional model. What are you building it for?

2 A key performance activity that a company wants to track

3 Data engineers discuss with business users to identify the business process

4 Find out why this process can form the basis of your dimensional model

Define the Business Process

upGrad Fashions wants to keep track of its Inventory.

How many products of each brand are available at a store?

upGrad Fashions wants to keep track of Product Promotions.

Promoted but not sold.
Not Promoted but sold.

upGrad Fashions wants to keep track of Sales.

What are the sales that are occurring at the different stores?

Summary | Introduction to Dimensional Modelling

01 To build a dimensional model, a business process is defined, a grain is chosen and facts and dimensions tables are built.

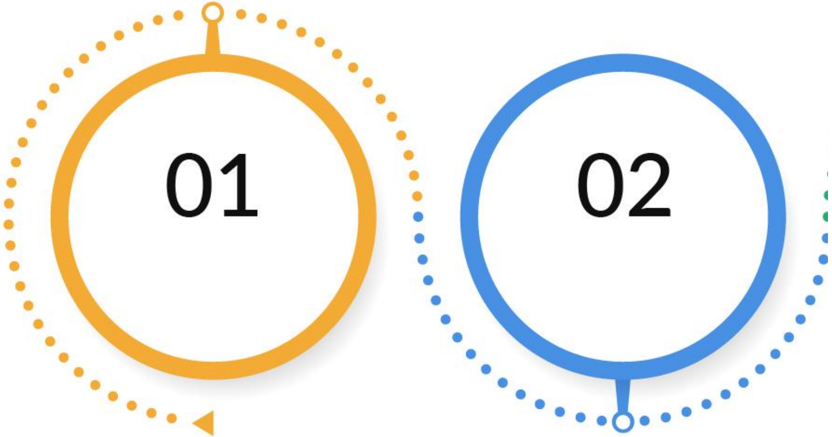
02 A business process is an activity that a business wants to track and analyse.

03 A business process could be analyzing the data of sales, inventory, promotions, discounts or customers.

Segment 6 | Define the Grain

Learning Objectives

Understanding the second step of building dimensional models

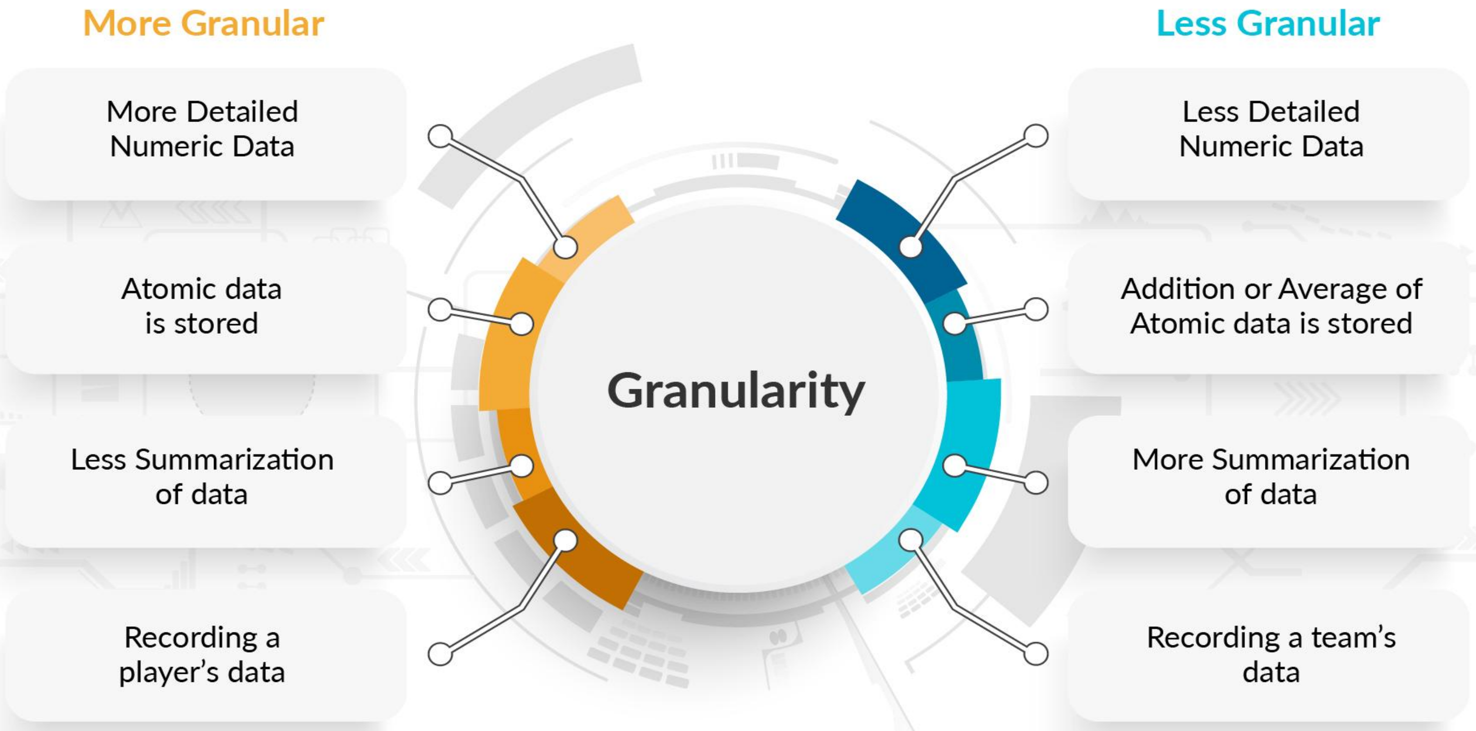


What do more granularity and less granularity mean? How do they affect the dimensional model?

upGrad

2 Define the Grain

How much do you want to drill down into each and every detail?



More Granular

- More Detailed Numeric Data
- Atomic data is stored
- Less Summarization of data
- Recording a player's data

Granularity

Less Granular

- Less Detailed Numeric Data
- Addition or Average of Atomic data is stored
- More Summarization of data
- Recording a team's data

2 Define the Grain

How much do you want to drill down into each and every detail?

upGrad

Analysis of Dimensional Model	Team	Player
Average of the runs scored by a team in all the matches	YES	YES
Which team scored the most runs against a particular team?	YES	YES
Total runs scored by a particular player	NO	YES
Which player scored the most runs against a particular team?	NO	YES

2 Define the Grain

How much do you want to drill down into each and every detail?

upGrad

Monthly

They can track sales for a month. This means there will be one row for every month in the main table of the dimensional model

Daily

Store the information about sales on a daily basis. There will be one row for every day

Every Transaction

Storing each and every transaction at each of the stores of Upgrad fashions. There will be one row for each transaction

Every Product

Storing each and every product being sold in every transaction

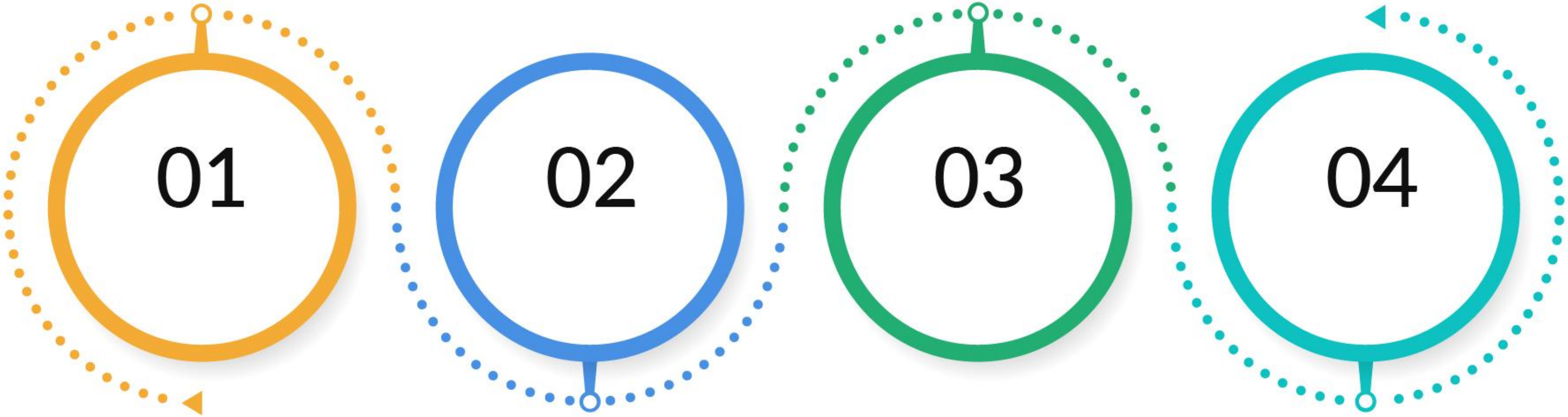
Summary | Define the Grain

01 A more granular data is more detailed. A less granular data is more summarization

02 Storing atomic data related to a business process is useful as it can be used to get summarized data.

Segment 7 | Facts and Dimensions

Learning Objectives



What are Facts?

What are Dimensions?

What are Star Schemas?

Dimensional model for an example

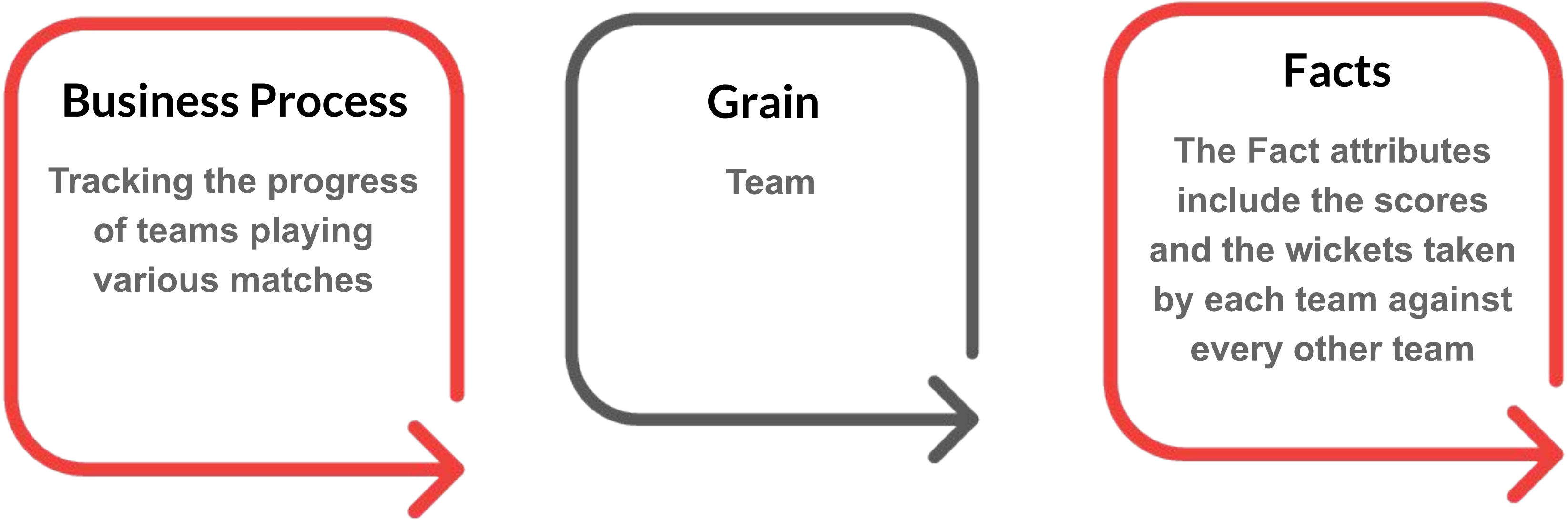
3 Define the Facts and Dimensions

1 A fact table is the main table that contains measurable data. The attributes of the fact table are the measurable metrics of a business process

2 A fact attribute is something that your business process wants to measure

3 The total amount paid by a customer in a transaction is a measurable quantity

3 Define the Facts and Dimensions



Business Process

Tracking the progress of teams playing various matches

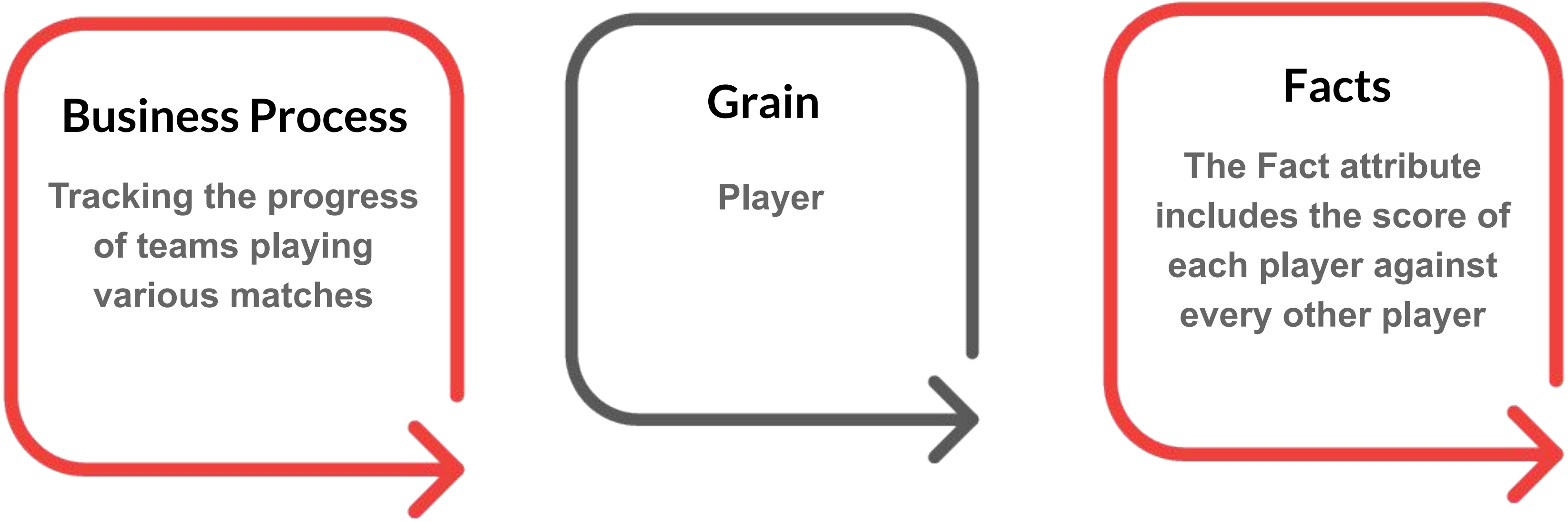
Grain

Team

Facts

The Fact attributes include the scores and the wickets taken by each team against every other team

3 Define the Facts and Dimensions



Business Process

Tracking the progress of teams playing various matches

Grain

Player

Facts

The Fact attribute includes the score of each player against every other player

upGrad

3 Define the Facts and Dimensions

Business Process

The business process is to track sales at every store of upGrad Fashions

Grain

Every product sold at every store

Facts

The Fact attributes include the final prices of products, the discount indicator, the quantity sold, the profit generated

3 Define the Facts and Dimensions

Business Process

The business process is to track sales at every store of upGrad Fashions

Grain

Daily amount of products sold at every store

Facts

The Fact attributes include the total amount generated, the total quantity of each product sold, the profit generated in a day

3 Define the Facts and Dimensions

Business Process

The business process is to track inventory at every store of upGrad Fashions

Grain

Daily inventory of each product daily

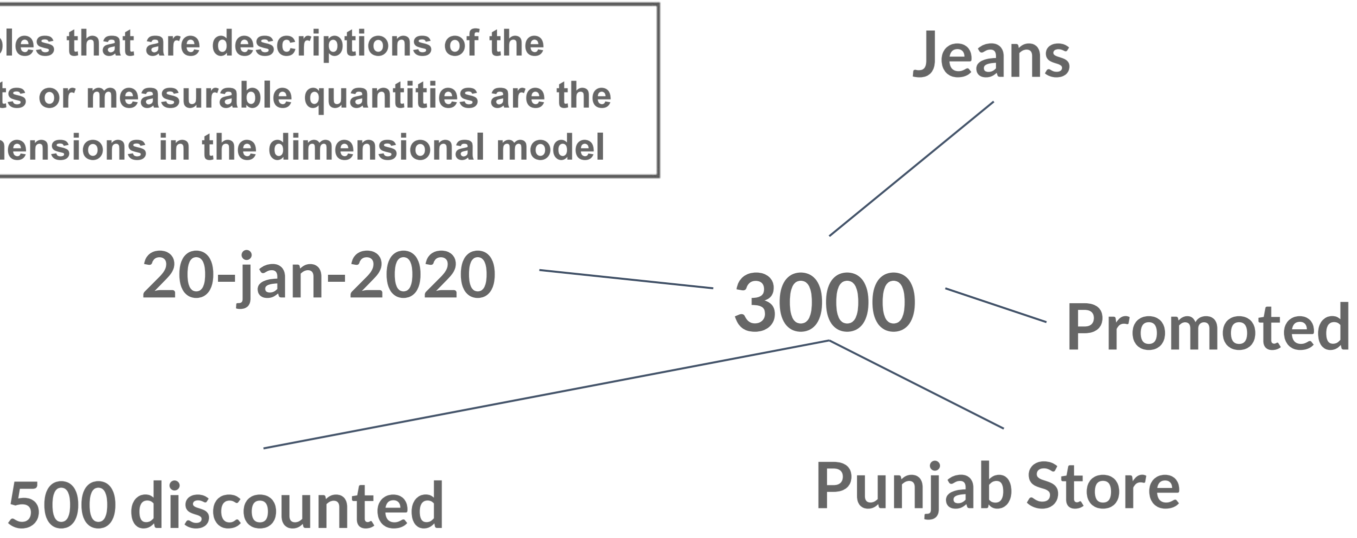
Facts

The Fact attributes include the quantity available.

3 Define the Facts and Dimensions

1 A dimension is detailed information about a fact attribute

2 Tables that are descriptions of the facts or measurable quantities are the dimensions in the dimensional model



Jeans

20-jan-2020

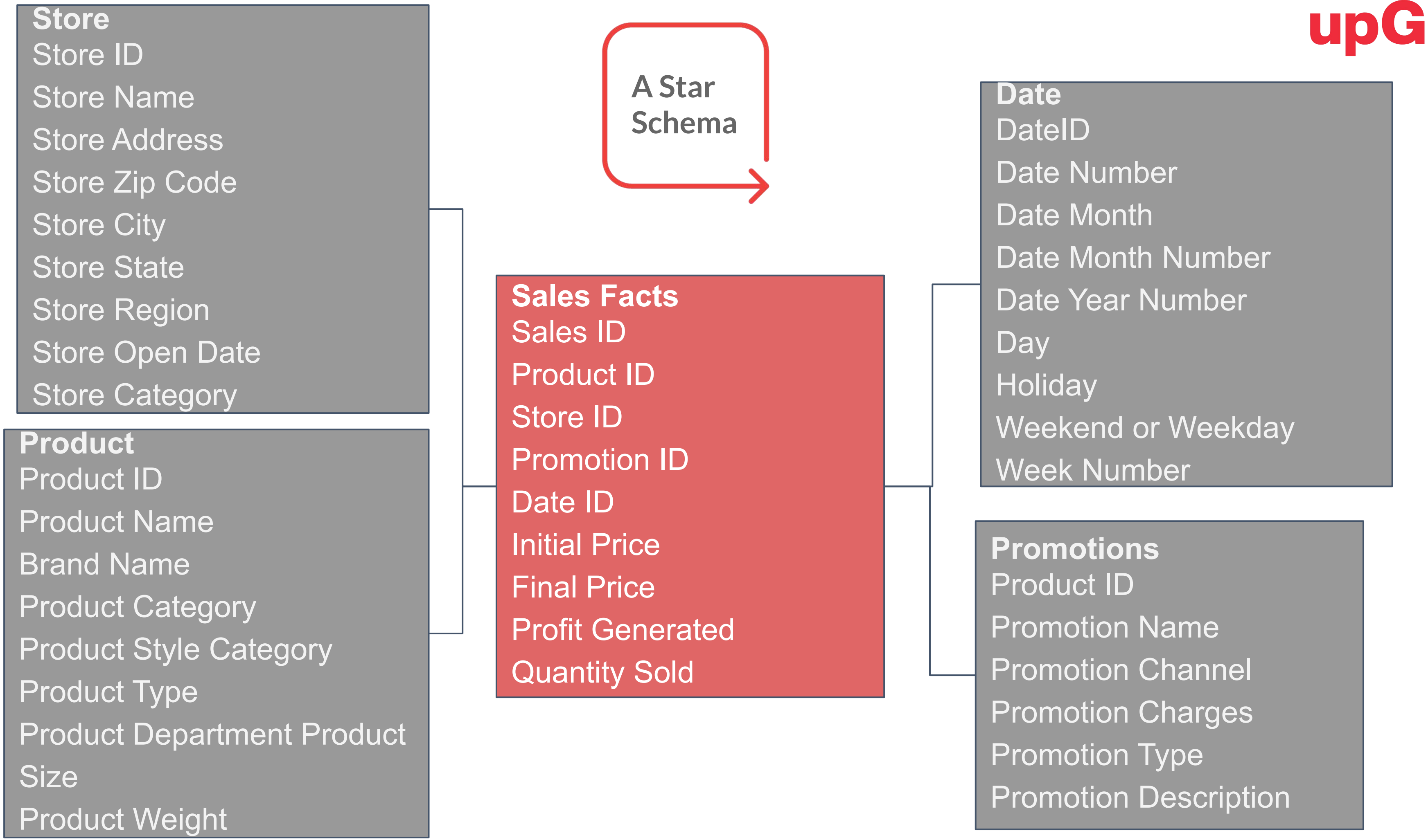
3000

Promoted

500 discounted

Punjab Store

A Star Schema



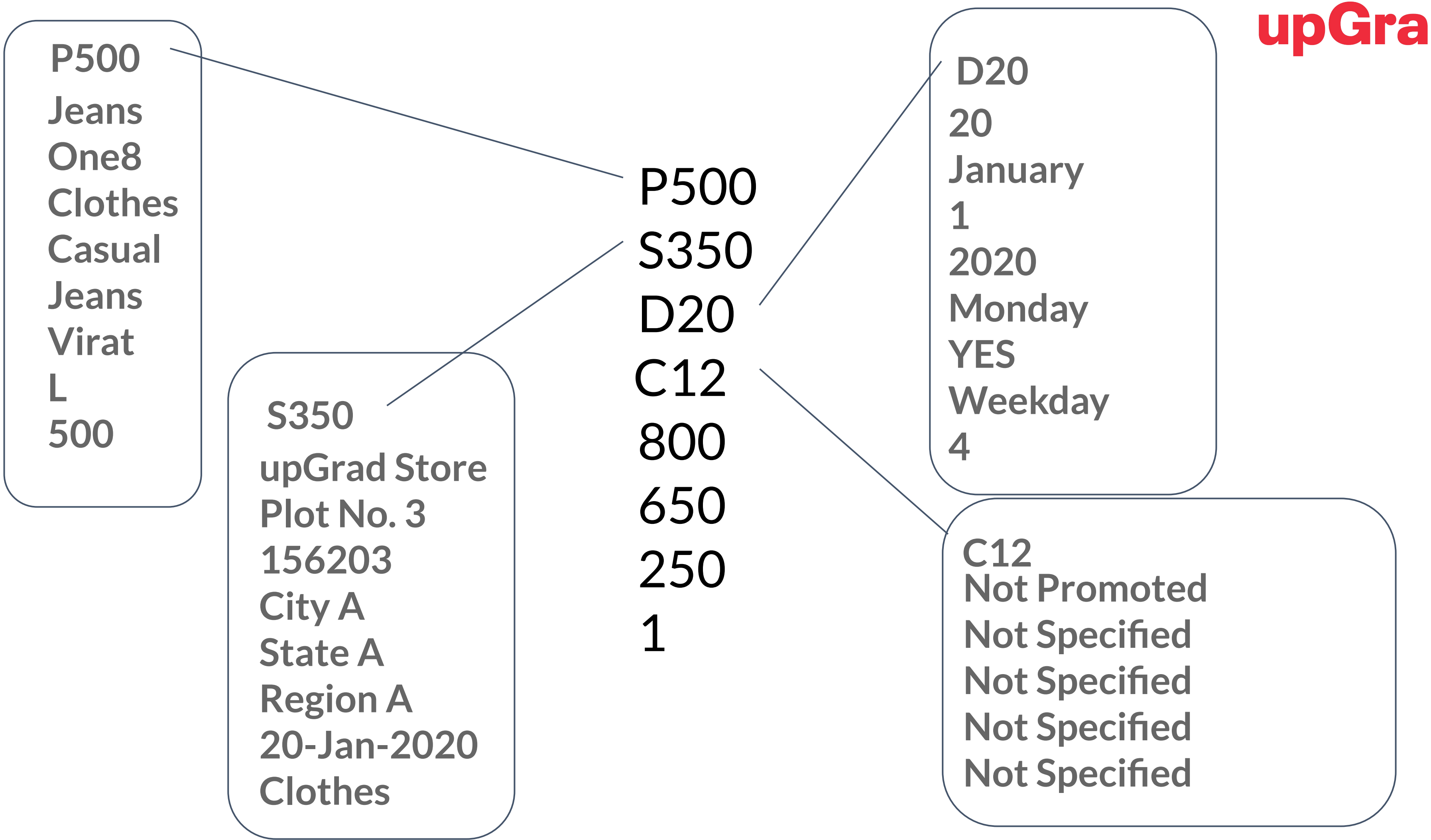
Store
- Store ID
- Store Name
- Store Address
- Store Zip Code
- Store City
- Store State
- Store Region
- Store Open Date
- Store Category

Product
- Product ID
- Product Name
- Brand Name
- Product Category
- Product Style Category
- Product Type
- Product Department Product Size
- Product Weight

Sales Facts
- Sales ID
- Product ID
- Store ID
- Promotion ID
- Date ID
- Initial Price
- Final Price
- Profit Generated
- Quantity Sold

Date
- DateID
- Date Number
- Date Month
- Date Month Number
- Date Year Number
- Day
- Holiday
- Weekend or Weekday
- Week Number

Promotions
- Product ID
- Promotion Name
- Promotion Channel
- Promotion Charges
- Promotion Type
- Promotion Description



P500
Jeans
One8
Clothes
Casual
Jeans
Virat
L
500

P500
S350
D20
C12
800
650
250
1

S350
upGrad Store
Plot No. 3
156203
City A
State A
Region A
20-Jan-2020
Clothes

D20
20
January
1
2020
Monday
YES
Weekday
4

C12
Not Promoted
Not Specified
Not Specified
Not Specified
Not Specified

Summary | Facts and Dimensions

01 Facts are collection of numeric metrics of the business process that company wants to analyse.

02 Dimensions are descriptive details of facts.

03 In a star schema, the tables are not in 2NF or 3NF.

Session Summary

01 An **OLTP** system is used to record real-time data.

02 An **OLAP** system is used to record and analyse historical data.

03 **Data warehouses** are needed to integrate the dat that a company records.

04 **Data warehouses** are **subject-specific**, **non-volatile**, **time-variant** and **integrated** data storages.

05 The number of tables is greater in relational modelling. **Dimensional modelling** is used for analysing and storing historical data.

06 **There are four steps to building a dimensional model:**

a. Define the Process
b. Define the Grain
c. Define the Dimensions
d. Define the Facts

07 A process is a **business activity** that a company has to track and analyse.

08 **More granular** means more detailed and atomic data. **Less granular** means more summarisation.

09 **Facts** are collection of attributes, which are numeric metrics.

10 **Dimensions** are the descriptive attributes related to a particular business concept.

Thank You