upGrad

Data Warehousing and ETL

upGrad

Course: Data Engineering - I

Lecture On: Data Warehousing and ETL

Instructor: **Ganesh Nerale**

Session - 2 | Dimensional Modelling and Data Marts

upGrad

Learning Objectives

Segment 02 — Understanding the use of Factless Tables. Describing various types of attributes in fact and dimension tables.

Segment 03 — Understanding why dimension data may change and describe ways to handle these changes.

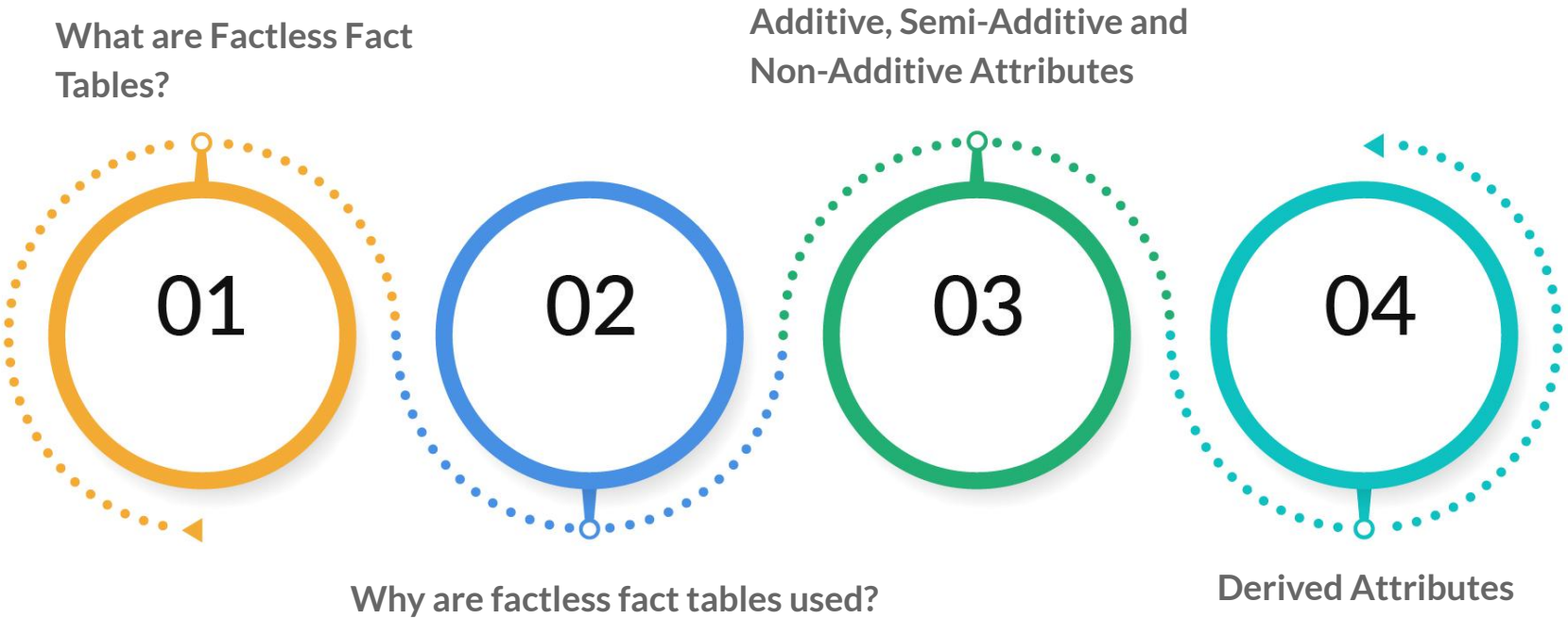
Segment 04 — Understanding the use of snowflake schemas.

Segment 05 — How the companies plan their data warehouse architectures?

upGrad

Segment - 2 | Factless Facts and Different Attributes

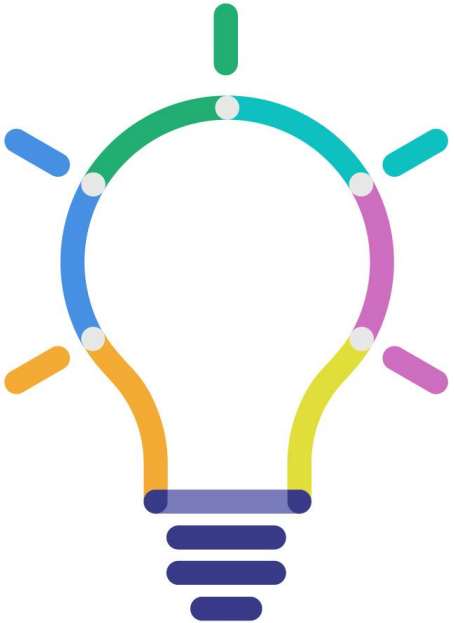
Learning Objectives



upGrad

Factless Fact Table

Such fact tables do not contain any numeric attribute.



Factless Facts and Different Attributes

upGrad

Factless Fact Tables

01

Factless Fact tables provide analysis for business intelligence but do not contain any numeric attribute.

02

You cannot apply any operations of addition or averaging to the attributes of factless fact tables.

03

These tables can be used in both the following cases:
To capture a descriptive event
To analyse a business activity that did not happen

Factless Fact Tables

upGrad

Capturing a descriptive event

- Capturing a meeting that happened between a sales representative and a customer.
- A fact table contains only foreign key columns.

Meetings
Meeting ID
Employee ID
Customer ID
Date ID
Channel ID

Use Case:
Used when a fact table is required only to establish a relation between various dimensions. A new row is added if a transaction happens.

Capturing a business process that did not happen

- Fact tables contain fact attributes, which are not numeric.

Sales
Date ID
Product ID
Store ID
Discount ID

All the sold items

Discount
Product ID
Store ID
Discount Indicator

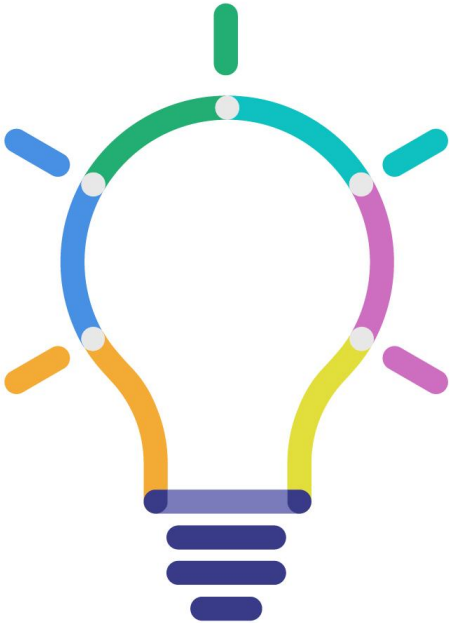
All the items on discount

Use Case:
Items that are not sold is a business process that did not happen. To find the items that were on discount but did not sell, you need to compare the sales table containing data of only sold items with the discount table containing data of all the products that were on discount.

upGrad

Additive Attributes
When such attributes are added, the output is another important metric.

Factless Fact Table
Such fact tables do not contain any numeric attribute.



Factless Facts and Different Attributes

2 Additive Attributes

upGrad

The added output is a useful business measure.

They can be added across all the dimensions.

Date ID
Product ID
Store ID
Customer ID
Sales Total Amount

Additive Across Date Dimension?
The total amount generated in the first week of April

Additive Across Store Dimension?
The total amount generated by a store

Additive Across Product Dimension?
The total amount generated by selling One8 jeans

Additive Across Customer Dimension?
The total amount spent by a customer

Analysis:
The total amount spent by a customer at a particular store in the first week of April

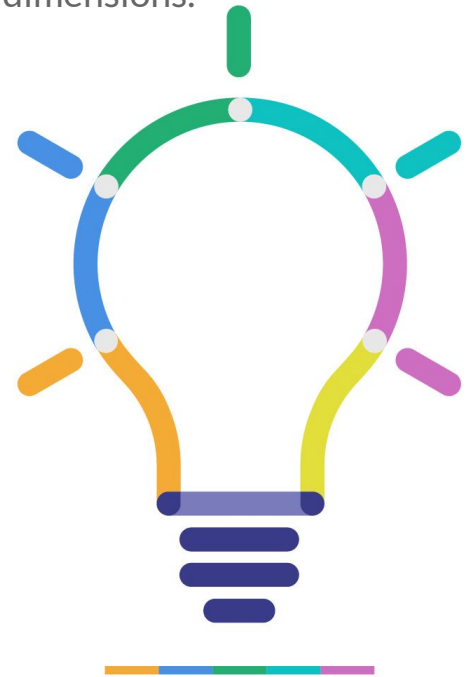
Analysis:
The total amount generated by selling one8 products in the first week of April

upGrad

Semi-Additive Attributes
Such attributes are not additive across all the dimensions.

Additive Attributes
When such attributes are added, the output is another important metric.

Factless Fact Table
Such fact tables do not contain any numeric attribute.



Factless Facts and Different Attributes

3 Semi-Additive Attributes

upGrad

They can be added across some dimensions.

The added output is not a useful business measure for some dimensions.

Date ID
Product ID
Store ID
Quantity remaining

Additive Across Store Dimension
The total quantity remaining at a particular store

Additive Across Date Dimension
The quantity remaining at the end of each day is not additive.

Additive Across Product Dimension?
The total quantity of the remaining one8 products

Analysis:
The total quantity of the remaining one 8 products on Monday

Analysis:
The total quantity of the remaining one8 products at a particular store

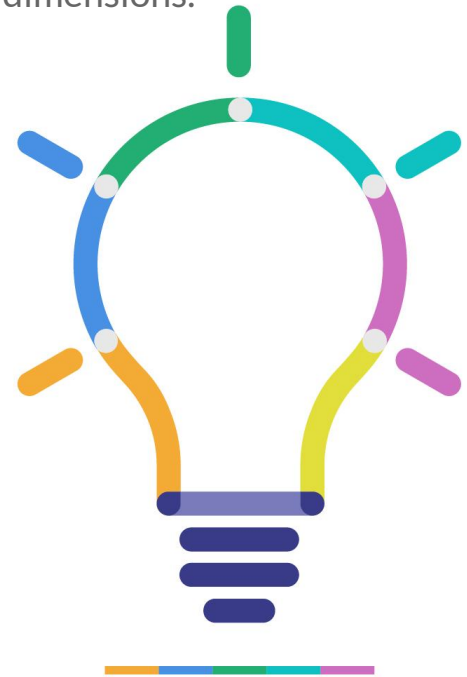
upGrad

Semi-Additive Attributes
Such attributes are not additive across all the dimensions.

Additive Attributes
When such attributes are added, the output is another important metric.

Non-Additive Attributes
When such attributes are added, the output is not a useful quantity.

Factless Fact Table
Such fact tables do not contain any numeric attribute.



Factless Facts and Different Attributes

4 Non-Additive Attributes

upGrad

They cannot be added across any dimension.

The added output is not a useful business measure for any dimension.

Additive Across Store Dimension?
We cannot add the discount percentage attribute across the store dimension.

Date ID
Product ID
Store ID
Discount Percentage

Additive Across Date Dimension?
We cannot add the discount percentage attribute across the date dimension.

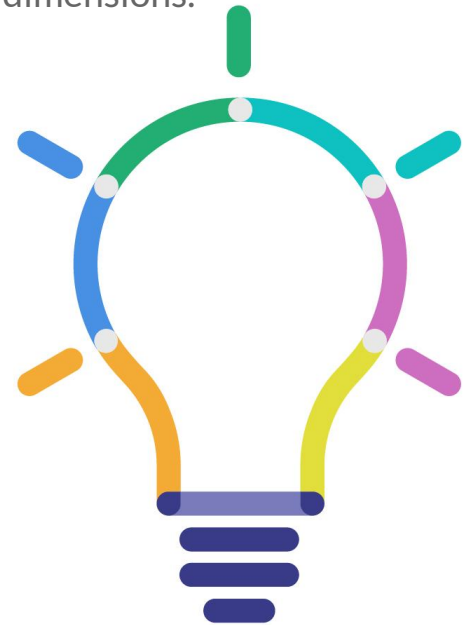
Additive Across Product Dimension?
We cannot add the discount percentage attribute across the product dimension.

- Some non-additive attributes such as percentage and ratio values are calculated by performing operations on additive attributes.
- If such a case, you should also keep the additive attributes in the fact table from which the non-additive attributes are calculated.

upGrad

Semi-Additive Attributes
Such attributes are not additive across all the dimensions.

Additive Attributes
When such attributes are added, the output is another important metric.



Non-Additive Attributes
When such attributes are added, the output is not a useful quantity.

Factless Fact Table
Such fact tables do not contain any numeric attribute.

Derived Attributes
Such fact attributes can be calculated using other fact attributes.

Factless Facts and Different Attributes

5 Derived Attributes

upGrad

Derived attributes are those whose values depend on other attributes.

Their values can be derived by performing additions, averaging or ratio operations on other attributes.

Although they can be calculated using other attributes, if they are important business metrics used frequently, you can include them in the fact table.

Profit is a derived attribute. It is calculated using charges and revenue attributes.

Summary | Factless Tables and Different Attributes

upGrad

01 Factless Fact Tables does not store any numeric attributes.

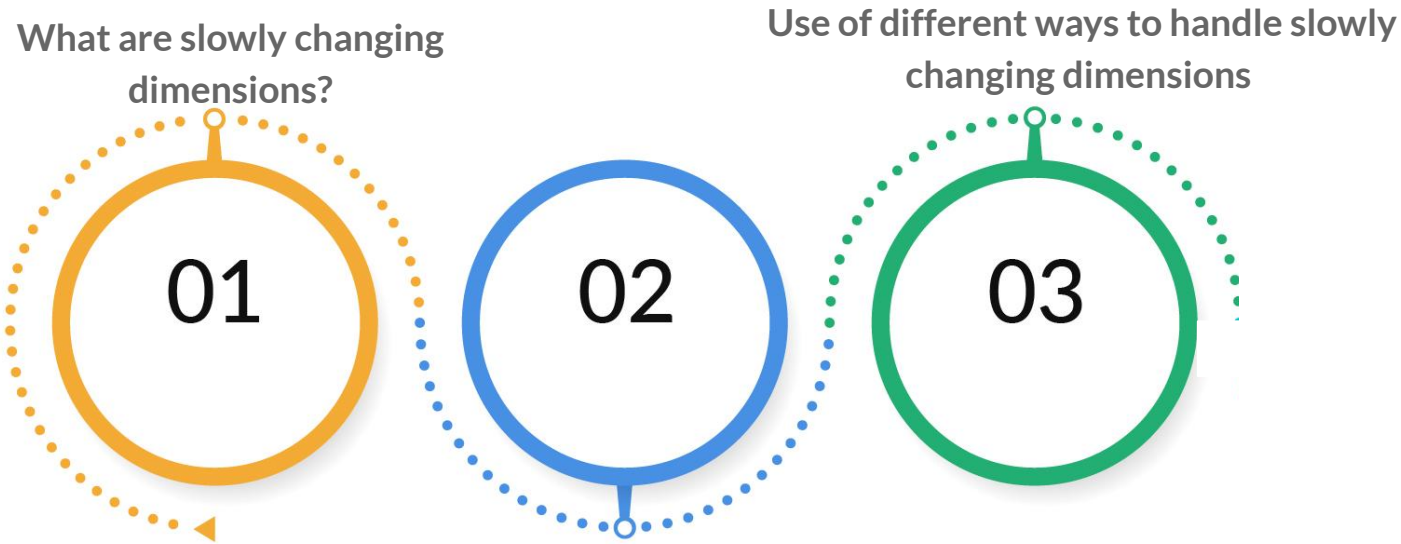
02 Additive Attributes can be added across all dimensions.
Semi - Additive Attributes can be added across some dimensions.
Non - Additive Attributes cannot be added across all dimensions.

03 Derived attributes can be calculated using other attributes.

Segment - 3 | Slowly Changing Dimensions

upGrad

Learning Objectives



Slowly Changing Dimensions

upGrad

01

Slowly Changing Dimension

The values stored in the dimensions are invalid owing to changes in business rules. Such values must be handled in the data warehouse according to the business requirements.

02

Product Dimension

The values in the rows of a product dimension change when the information regarding a particular product changes.

Product ID	Product Name	Product Type	Brand Name	Product Manager
1235	Jeans	Clothes	One8	Virat

The company wants to make Rohit the product manager instead of Virat. There are three ways to do it, which are as follows: Type 1, Type 2 and Type 3.

Type 01 Method

upGrad

By the **Type 1 Method**, we change the value stored in the row.

Product ID	Product Name	Product Type	Brand Name	Product Manager
1235	Jeans	Clothes	One8	Rohit

- It is used when the data warehouse does not want to keep track of the data changes.
- You will not be able to track the performance of the product individually with either Rohit or Virat as its manager.

Type 02 Method

upGrad

By the **Type 2 Method**, we create a new row with all the attributes having the same value except for the attribute whose value has to be updated.

Row Number	Product ID	Product Name	Product Type	Brand Name	Product Manager
101	1235	Jeans	Clothes	One8	Rohit
102	1235	Jeans	Clothes	One8	Virat

- It is used when the data warehouse has to keep track of the data changes.
- If the product manager changes once more, there will be another row.
- You will be able to track the performance of the product individually with either Rohit or Virat as its manager.

Type 03 Method

upGrad

By the **Type 3 Method**, we create a seperate column to store both the values.

Product ID	Product Name	Product Type	Brand Name	Product Manager	Previous Product Manager	Date
1235	Jeans	Clothes	One8	Rohit	Virat	20 January 2020

- It is used when the data warehouse has to keep track of the data changes.
- If the product manager is changed once more, the information about Virat as the manager for that product will not be available.

Summary | Slowly Changing Dimensions

In a Type - 1 method, the change is directly made to the column value.

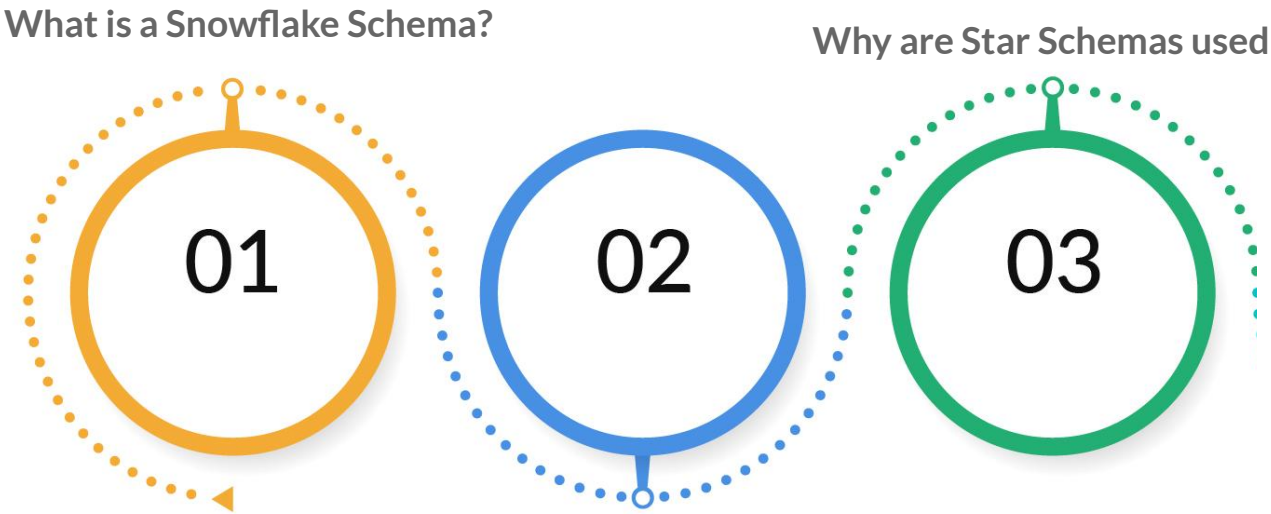
In a Type - 2 method, a new row is made with the required new value for the column.

In a Type - 3 method, a new column is made to track the change in the data value.

Segment - 4 | Snowflake Schema

upGrad

Learning Objectives

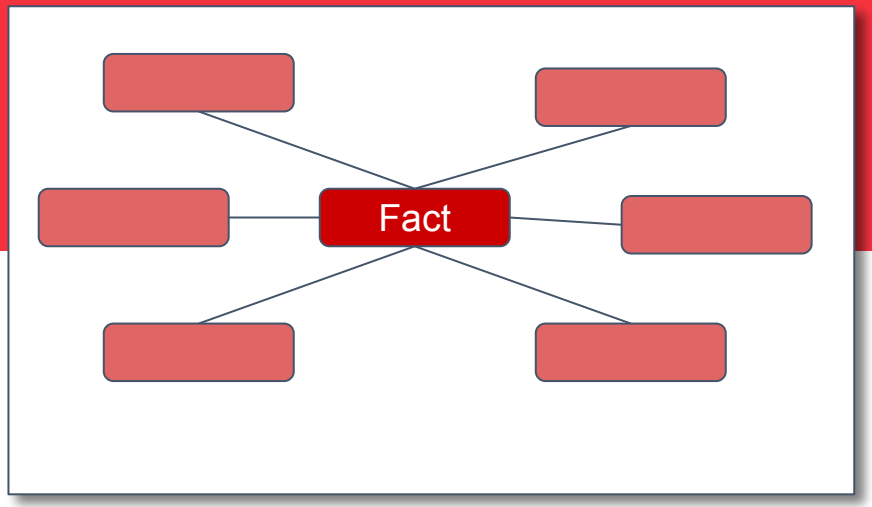


Snowflake Schema

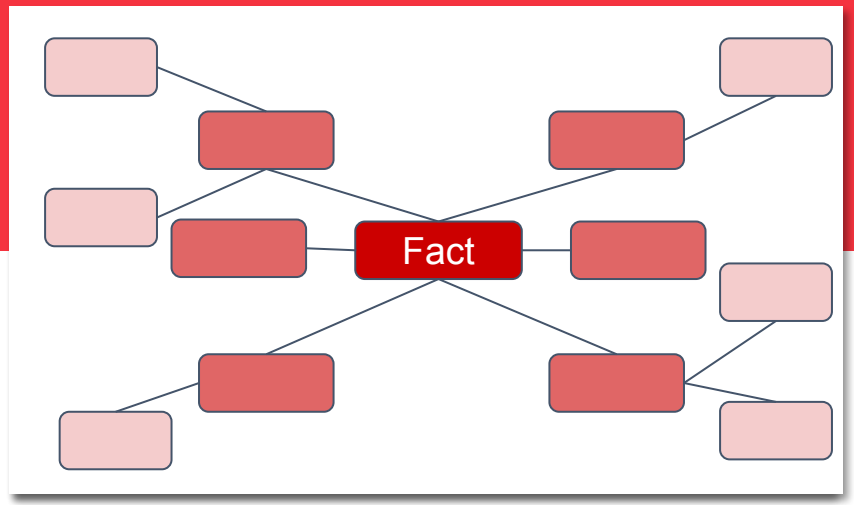
upGrad

In a star schema, the tables are neither in 2NF or 3NF.

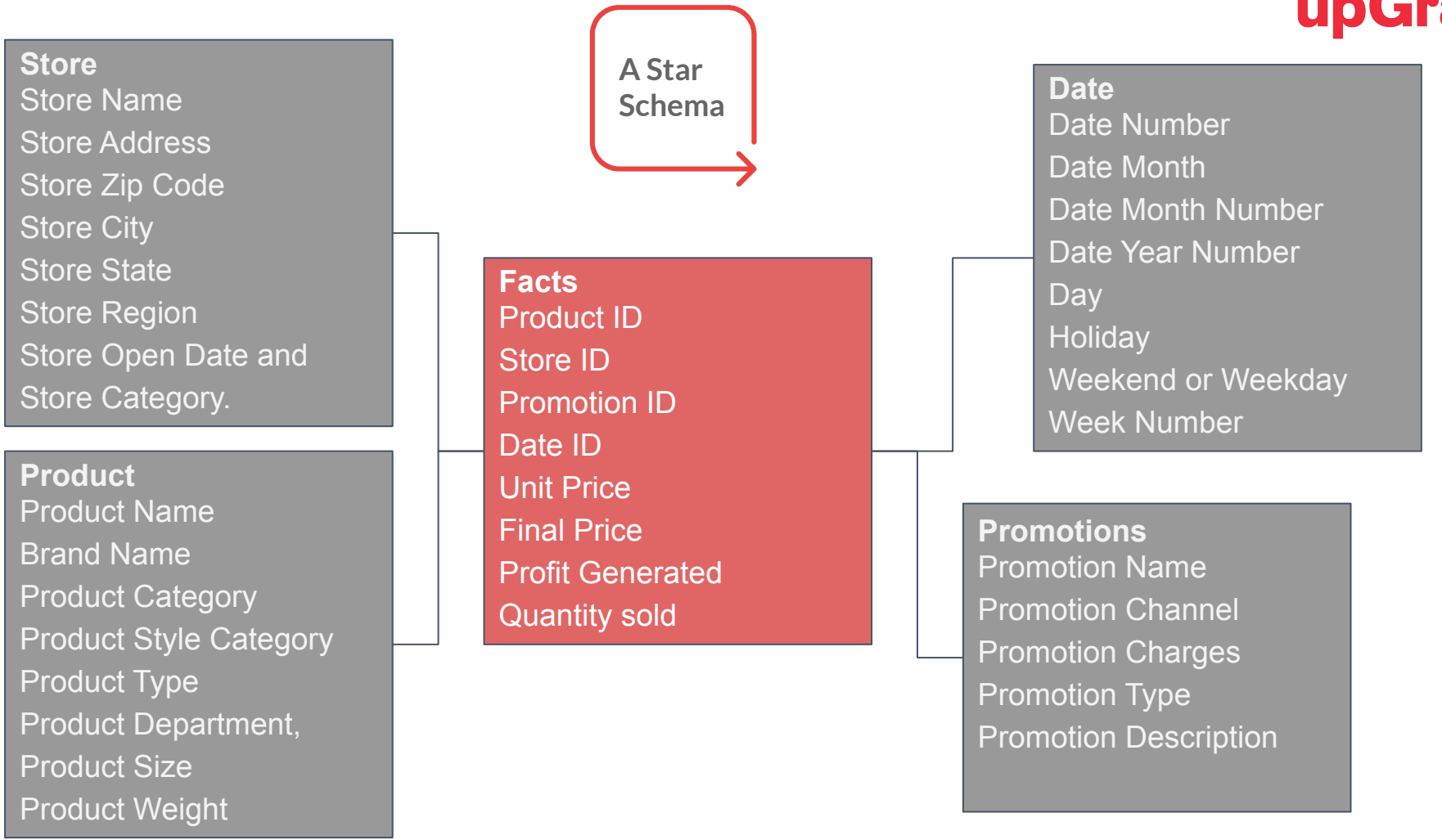
Data related to one dimension is kept together.



In a snowflake schema, the tables are in either in 2NF or 3NF.

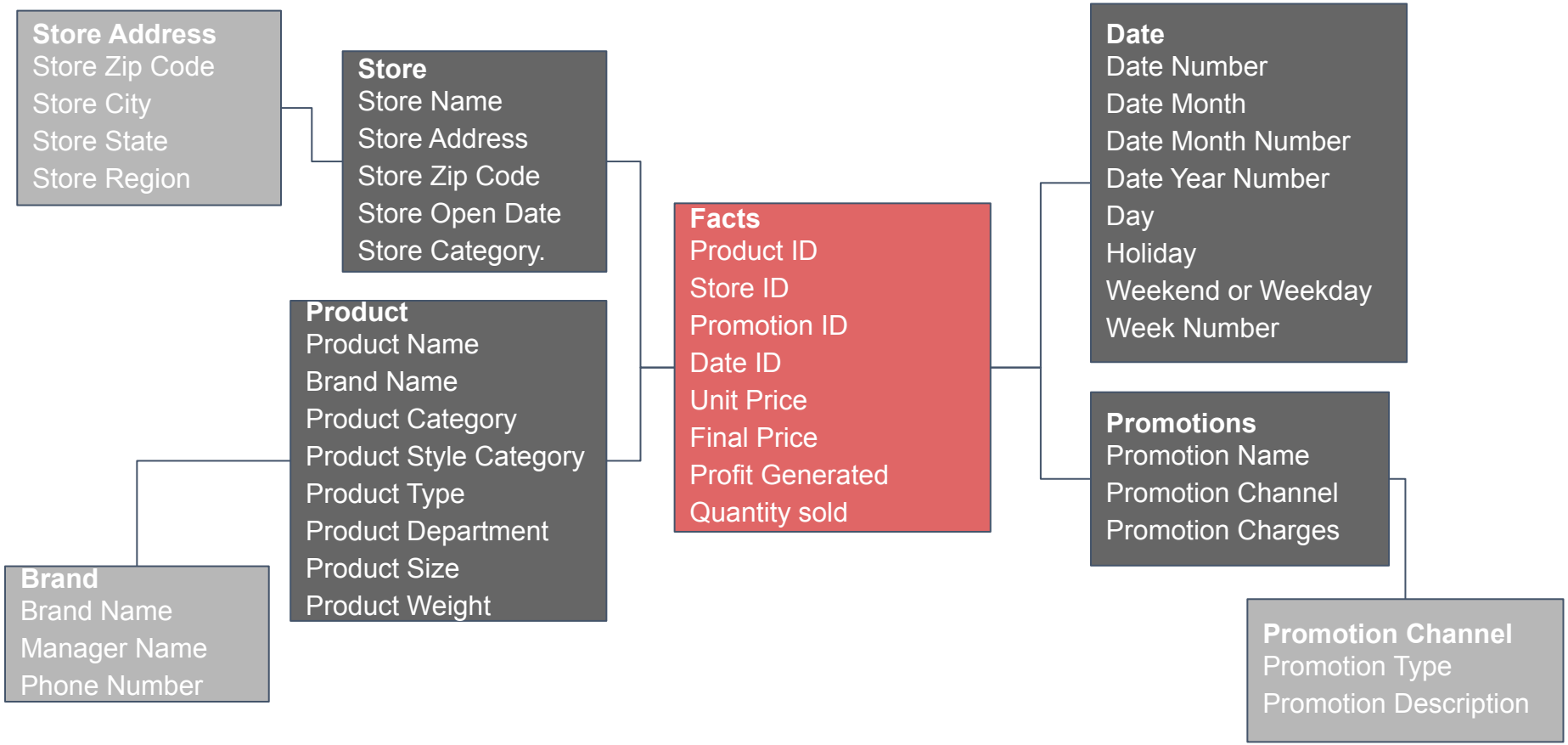


upGrad


A Star Schema
Store
Store Name
Store Address
Store Zip Code
Store City
Store State
Store Region
Store Open Date and
Store Category.
Product
Product Name
Brand Name
Product Category
Product Style Category
Product Type
Product Department,
Product Size
Product Weight
Facts
Product ID
Store ID
Promotion ID
Date ID
Unit Price
Final Price
Profit Generated
Quantity sold
Date
Date Number
Date Month
Date Month Number
Date Year Number
Day
Holiday
Weekend or Weekday
Week Number
Promotions
Promotion Name
Promotion Channel
Promotion Charges
Promotion Type
Promotion Description

Snowflake Schema

upGrad



Star Schema Vs Snowflake Schema

upGrad

Product ID	Product Name	Product Type	Brand Name	Brand Manager
1235	Jeans	Clothes	One8	Virat
1236	Shirt	Clothes	One8	Virat

Product ID	Product Name	Product Type	Brand ID
1235	Jeans	Clothes	12
1236	Shirt	Clothes	12

Brand ID	Brand Name	Brand Manager
12	One8	Virat

01 Faster query results

02 The size of the dimension table is lower than that of fact tables.

03 A complete picture of one dimension in one table

If the dimension table is not used frequently, the snowflake schema can be used.

If the dimension is very wide, the snowflake schema can be used.

Summary | Snowflake Schema

upGrad

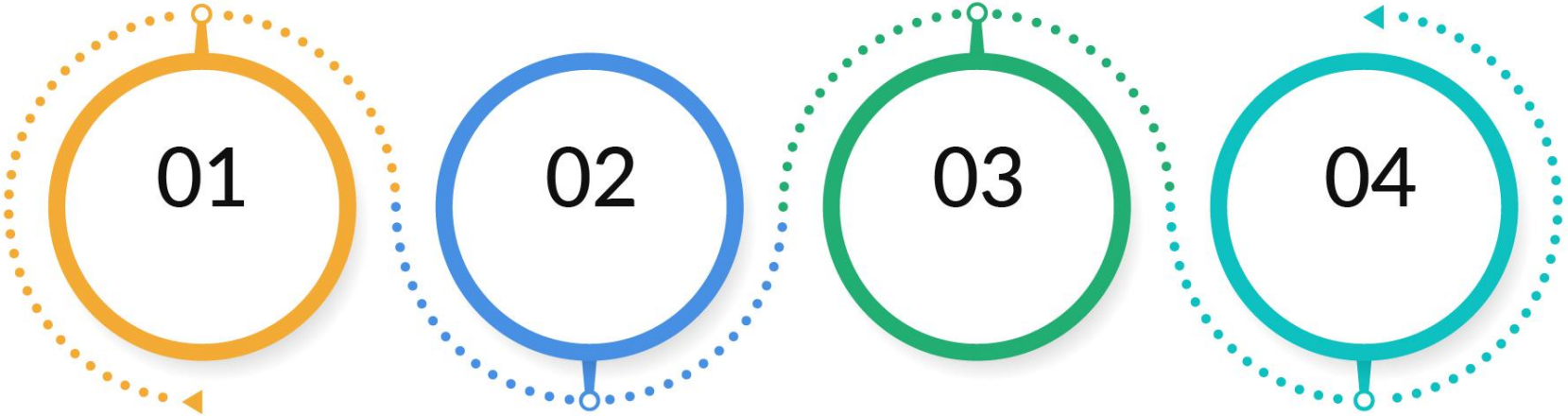
01 The dimension tables in a snowflake are in 2NF or 3NF.

02 A Star Schema is used more because the number of tables are less and the analysis is fast.

03 A snowflake schema can be used where the dimension tables are not frequently used.

Segment - 5 | Data Marts

Learning Objectives



Data Marts

To query a particular set of data that involves a functional activity of a business, every team has to analyse the entire data warehouse.

FROM IDEA

TO RESULT

Data marts store specific data for different functional activities of a company. Query results are obtained quickly.

Data Marts

upGrad

You do not want every team to run analytics on the entire data.

Data is more organised and is readily available. Query results are obtained quickly.

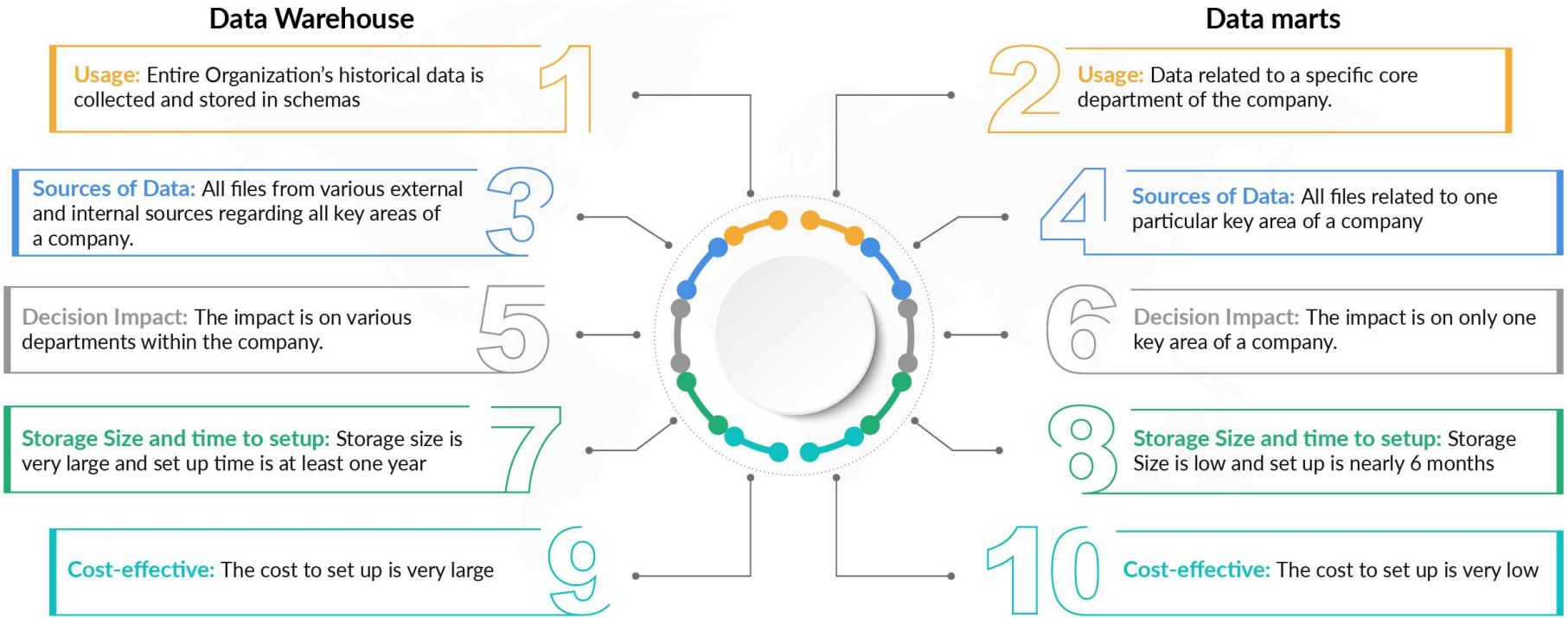
A data mart is a subset of a data warehouse.

A dart mart is a collection of information for a particular key functional area of an organisation.

Sales, marketing, inventory, promotions, product management, finance and resources are all different sections within a company. A data mart stores information specific to every section.

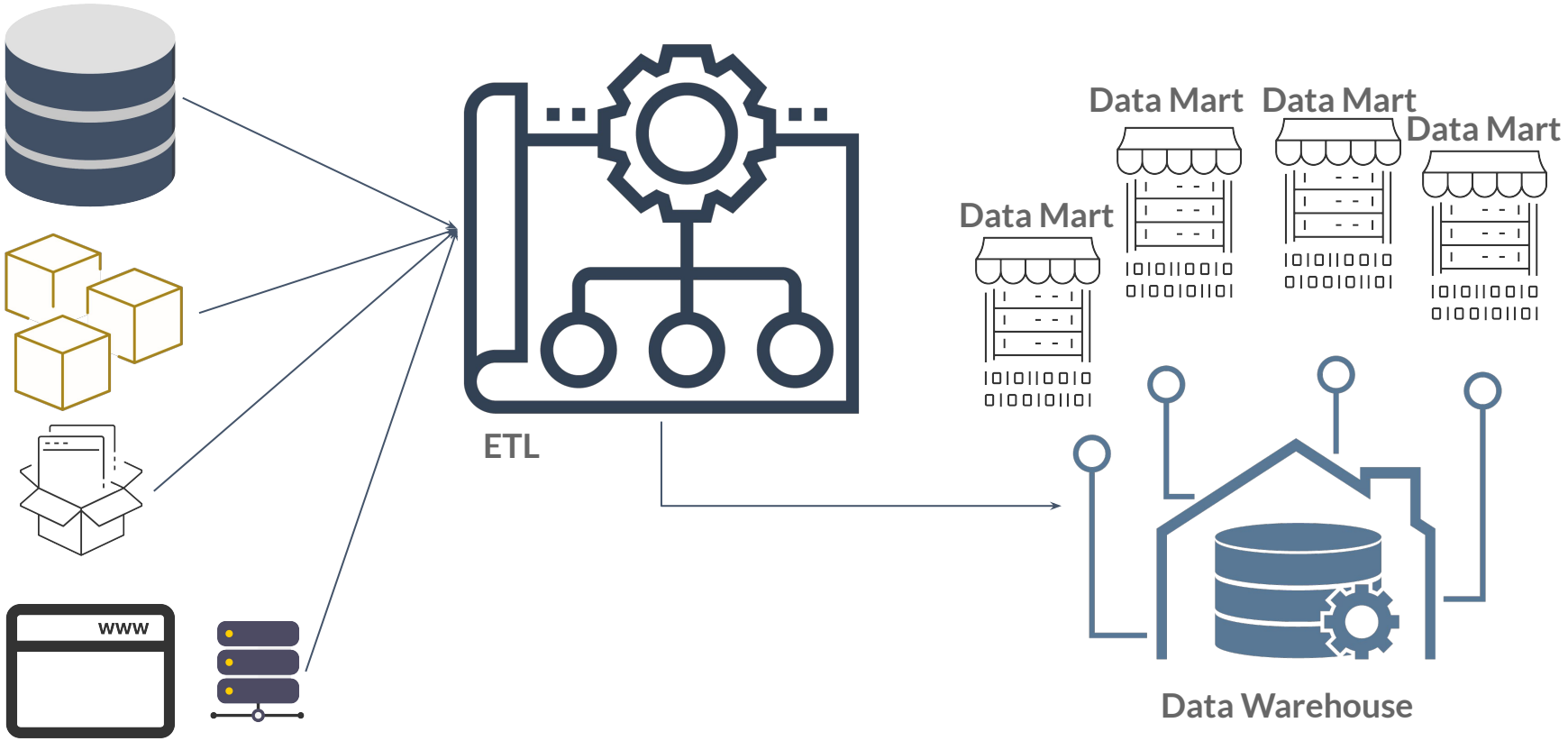
Data Marts vs Data Warehouses

upGrad



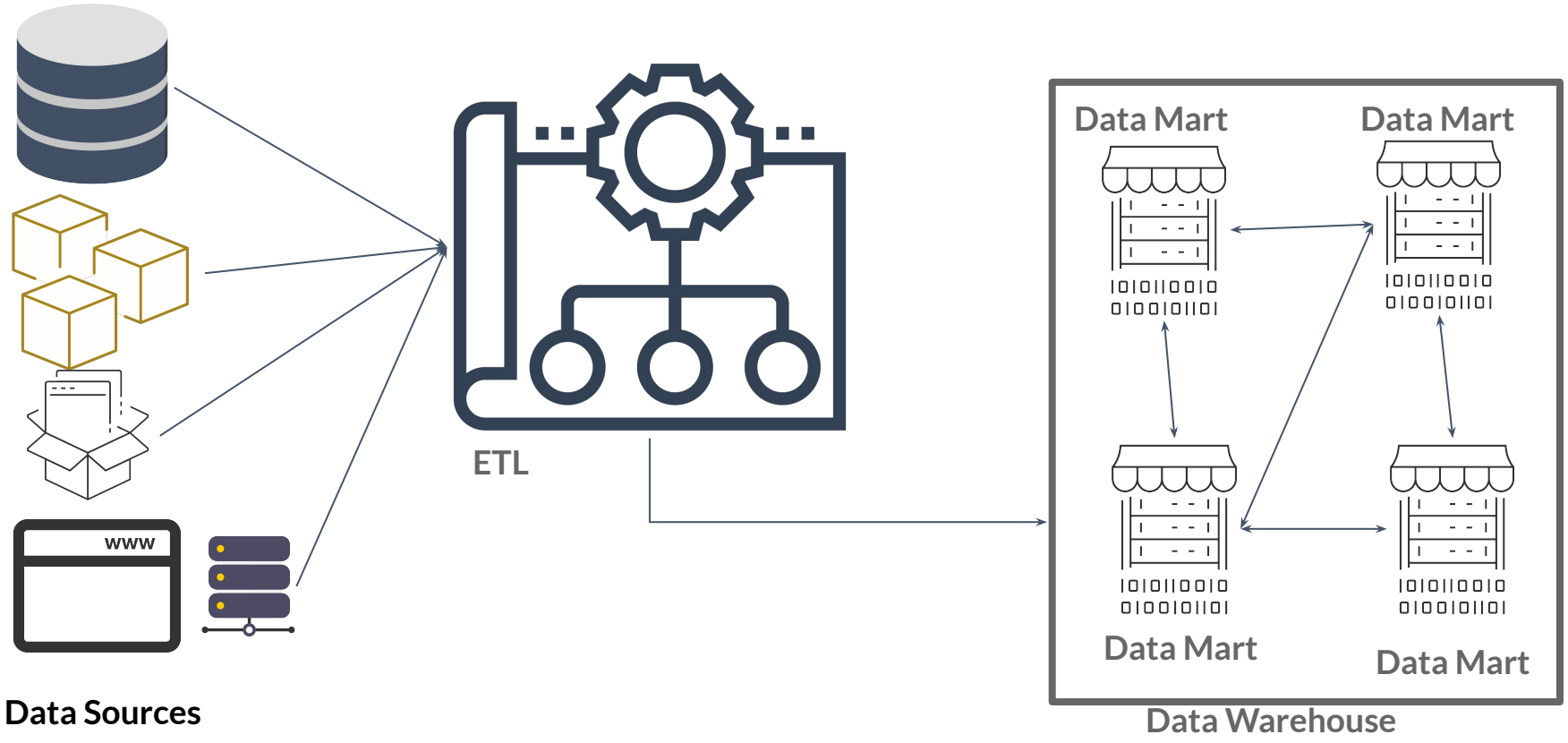
Bill Inmon Architecture of a Data Warehouse

upGrad



Kimball Architecture of a Data Warehouse

upGrad



The Two architectures of Data Warehouses

upGrad

Inmon Architecture

Data is stored using the E-R model in a data warehouse. Data marts are built using dimensional models.

A data warehouse is built first, and then, data marts are built for specific uses.

Inmon Architecture is built keeping the needs of the entire company in mind.

Business users can access the data in a data warehouse only through data marts.

Kimball Architecture

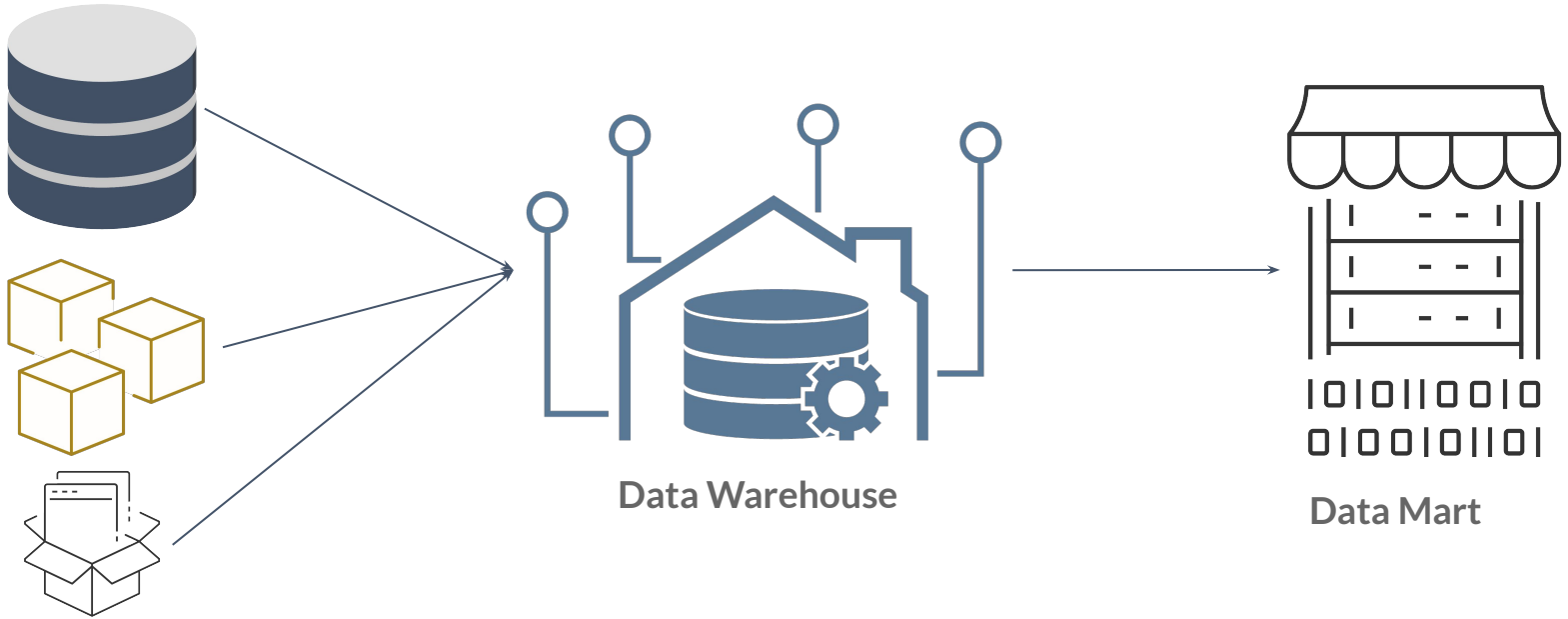
Data is stored using dimensional models in data marts, which build the data warehouse.

Data marts are built first, and then, data warehouses are built using these data marts.

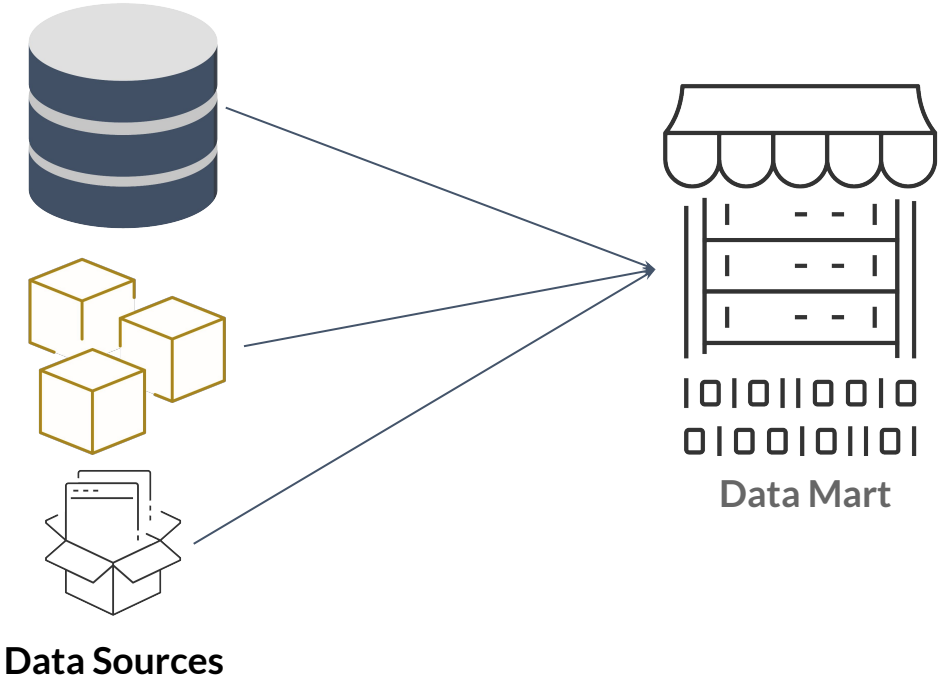
The Kimball architecture is built keeping the needs of particular functional areas in mind.

Business users can directly access the data in data warehouses.

Types of Data Marts: Dependent

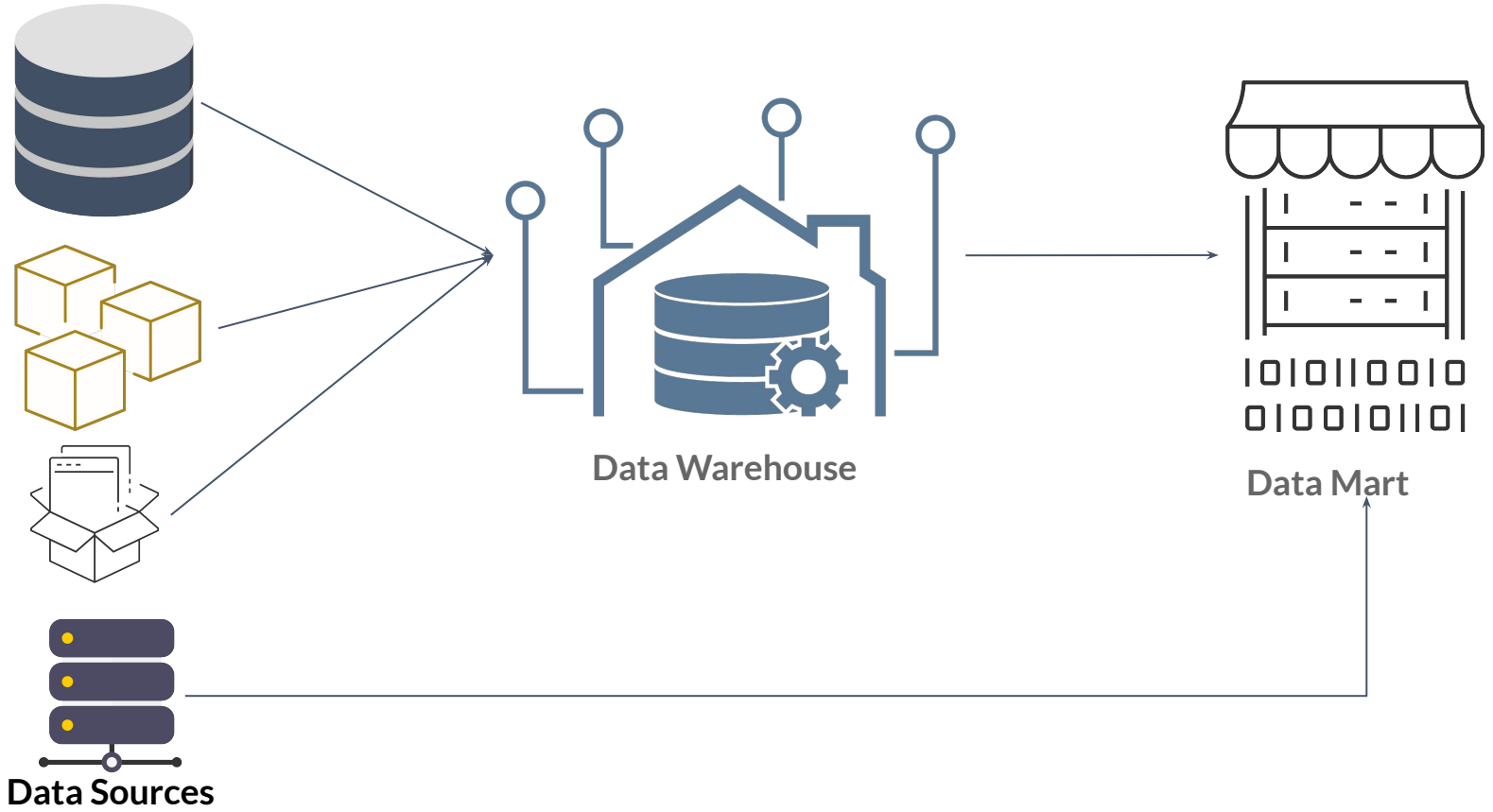


Types of Data Marts: Independent



Types of Data Marts: Hybrid

upGrad



Summary | Data Marts

A data mart is subset of Data Warehouses.

In Bill-Inmon Architecture, Data Marts are build from data warehouse.
In Kimball Architecture, Data Marts build the data warehouse.

For Dependent Data Marts, the source of data is data warehouse.
For Independent Data Marts, the source of data are original data sources.
For Hybrid Data Marts, the source of data are both data warehouses and original data sources.

Session Summary

01 **Factless Fact tables** provide analysis for business intelligence but do not contain any numeric attribute.

02 **Additive attributes** can be added across all dimensions.

03 **Semi-Additive attributes** can be added across some dimensions.

04 **Non-Additive Attributes** cannot be added across any dimension.

05 There are three ways to handle a slowly changing dimension, which are as follows: **Type 1**, **Type 2** and **Type 3**

06 **Snowflake schemas** have dimension tables in 2NF or 3NF. Star schemas are fast for analysis because the number of tables are less.

07 **Data marts** are used for specific data storage and analysis.

08 There are three different types of data marts, which are as follows: **Dependent**, **Independent** and **Hybrid.**

09 The two different architectures for data warehouses are **Bill Inmon's** and **Kimball's.**

10 **Kimball architecture:** Data marts build a data warehouse.
Bill Inmon architecture: Data marts are built from a data warehouse.

upGrad

Thank You