CM2606 Data Engineering

Data Warehousing 01

Week 05 | Piumi Nanayakkara

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

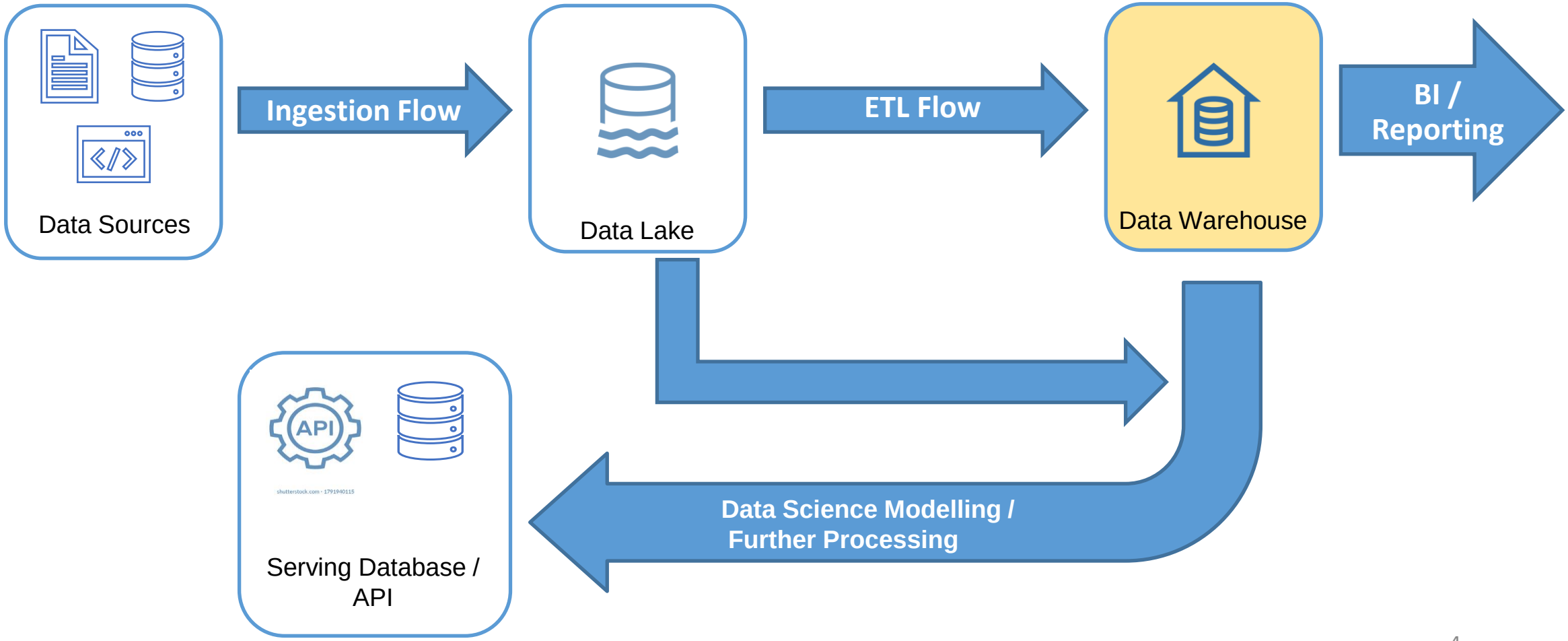
Learning Outcomes

- Covers LO1 and LO2 for Module
- On completion of this lecture, students are expected to be able to:
  - Explain the concept of a data warehouse and a data mart
  - Identify and describe dimensional modelling and different schemas available

CONTENT

- Importance of a Datawarehouse
- Data Mart
- Dimensional Modelling
  - Terminology
  - Schemas
  - Dimension and Fact Tables
    - Characteristics & Types

Data Pipeline: Common Usage



Data Warehouse

- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process. – Bill Inmon, Father of Data Warehousing
- **Subject-Oriented**: used to analyze a particular subject area. E.g., Sales
- **Integrated:** Integrates data from multiple data sources.
- **Time-Variant:** Historical data is kept
- **Non-volatile:** Once data is in the data warehouse, it will not change.

Why is it needed?

- Need of an analytical database for analysis and reporting – OLAP (Online Analytical Processing)
  - Production databases are optimized for writes – OLTP (Online Transactional Processing)
- Avoid disrupting the production databases
- Consolidate data from multiple sources
- Need for single source of truth
- To break the data silos

Data Silos

- Named after the structures farmers use to store different types of grain
- A collection of data held by one department
  - Not fully or easily accessible for other departments
- Disadvantages:
  - Create a data barrier
  - Inconsistencies in data
  - Hard for leaders to get a holistic 360^0 view

Additional Benefits

- Make information accessible easily
  - Centralized Storage
  - Supports analytical reporting (OLAP), querying and decision making
- Present information timely and consistently
  - Daily/hourly or near real time
  - Credible, quality assured data
- Provides Security
  - Access controls
  - Data Masking and encryption
  - Separate from operational database

Data Mart

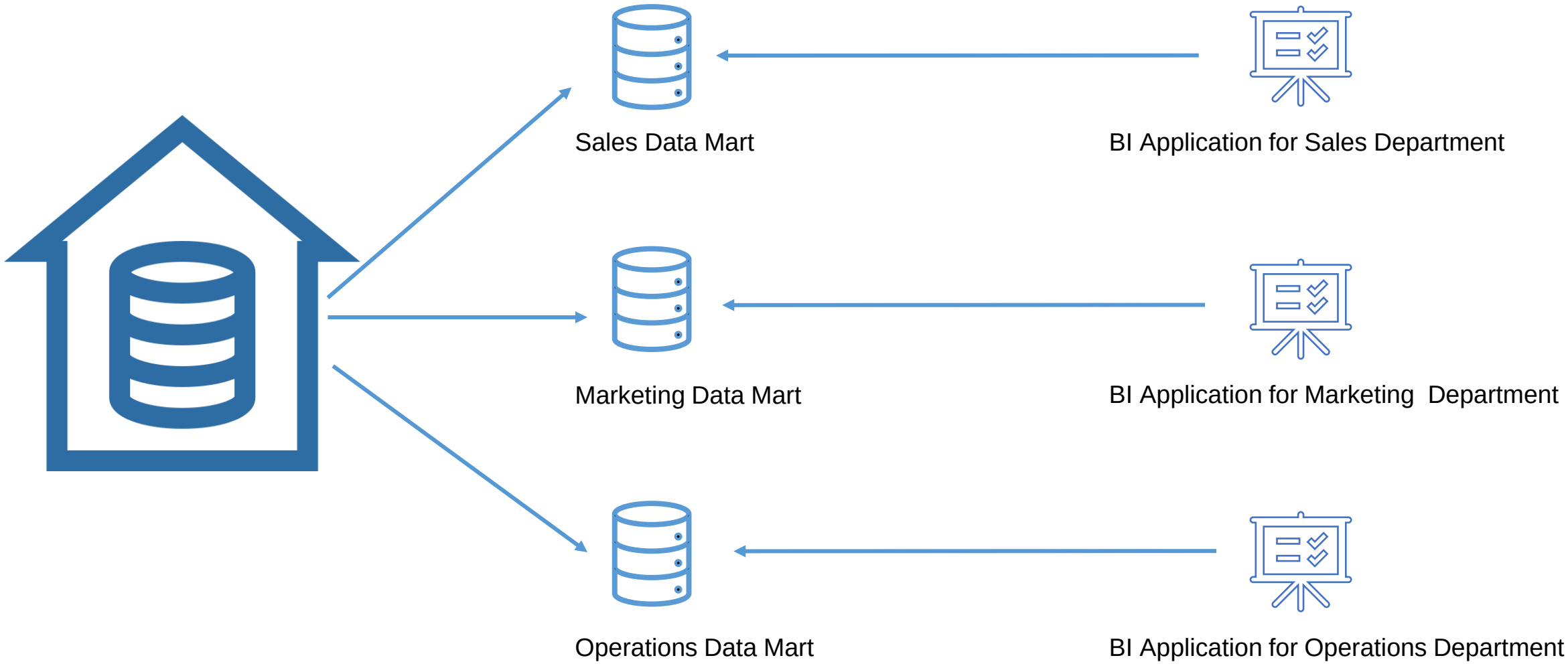
- Small scale implementation when your organization does not need a fully fledged Enterprise Data Warehouse
- Contains a subset of data that is stored in a data warehouse focusing on a particular subject area
- Draw on fewer, more specialized data sources.
- Easy to implement and cost-effective
- Single subject matter expert can define its structure and configuration.

Data Mart Creation

- Bottom-Up Approach
  - ETL loads the data into the Data Marts
  - Data warehouse could be built as the aggregate of all data marts.

- Top-Down Approach
  - ETL loads information to the Data Warehouse directly.
  - Data Marts are created based on the data loaded to warehouse
  - Provides a consistent view of information flow

Top-Down Approach



Types of Data Mart

- Dependent Data Mart
  - Top-Down Approach
  - Data Marts are always created based on central data warehouse
- Independent Data Mart
  - Data Mart Directly source data from row data sources
  - May extend to create a central data warehouse (bottom-up approach)
- Hybrid Data Mart
  - Data is fed from both row data sources as well as central data warehouse
  - Useful when a user wants an ad hoc integration

Dimensional Modelling

- Used to represent 3NF data in a relational database differently in a data warehouse
- Main goal is to improve data retrieval whereas in Normal Form modelling focus is to remove redundancies.
- Dimensional Modelling happens in logical layer which can be mapped to any database in physical layer
  - Relational or Multi-Dimensional databases
- A dimensional model includes fact tables and lookup tables.
  - Fact tables connect to one or more lookup tables
  - Dimensions and hierarchies are represented by lookup tables.
  - Attributes are the non-key columns in the lookup tables.

Dimensional Modelling: Terminology

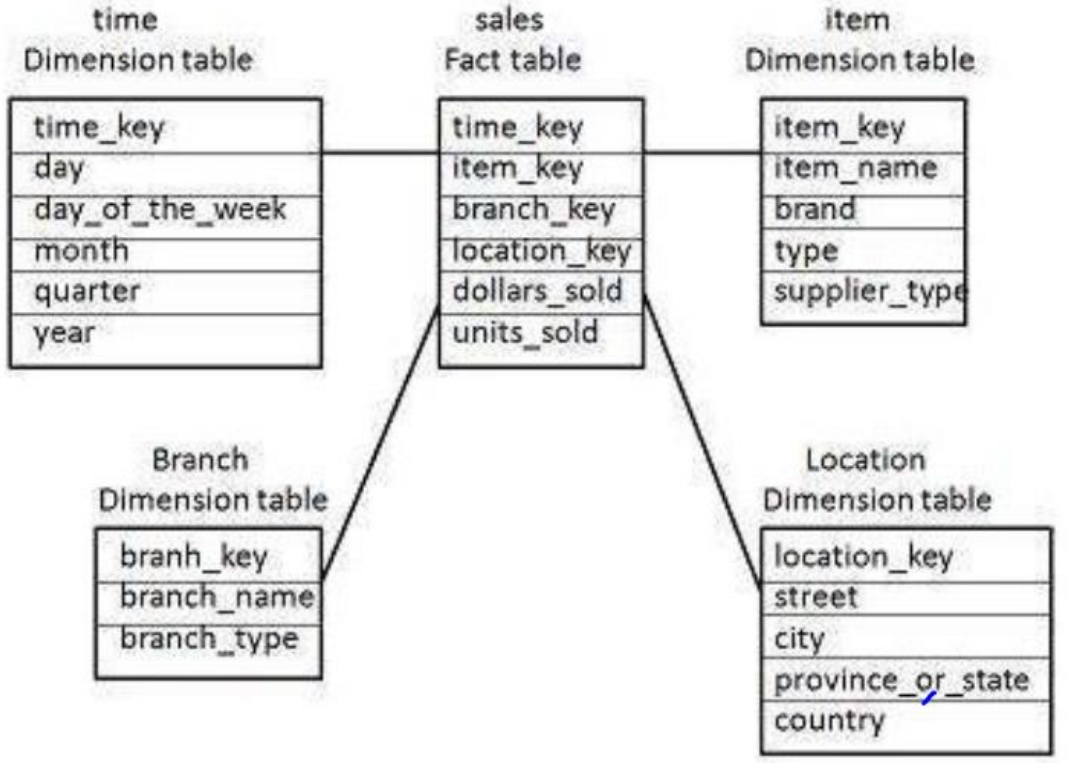
- Fact Table: Table that contains the measures of interest with the appropriate granularity.
  - E.g., Sales amount by store by day, Sales amount by product per month
- Dimensions: Different aspects of the fact in consideration
  - E.g., Time Dimension, Store Dimension, Product Dimension
- Hierarchy: Relationship between different attributes within a dimension
  - E.g., Possible hierarchy for time dimension: Year → Quarter → Month → Day
- Dimension / Lookup Table: Table that contains attributes of a dimension
  - E.g., Store Dimension table containing Store ID, Store Name, Store Size

Dimensional Modeling: Schemas

- A database schema defines how the data is organized and how the relations among them are associated.
- When performing dimension modelling the atomic data is loaded into dimensional structures. Then the dimensional models / schemas are generated or build around the business processes.
- Dimensional modeling schemas provide techniques to join facts and dimension
- Dimensional models are scalable and can easily accommodate unexpected new data

Star Schema

- Bunch of relational database tables whose relationships form a star
- A fact table in the middle connected to a set of dimension tables
- The fact tables are in 3NF form, and the dimension tables are in denormalized form.



Snowflake Schema

- An extension of star schema where some dimensional hierarchies are normalized into a set of smaller dimension tables.
- This forms a shape of a snowflake
- No redundancy, thus less storage. The tables are easy to manage and maintain
- However, due to this, more joins would be required when querying.

Galaxy / Fact Constellation Schema

- A group of different fact tables that have few similar dimensional tables

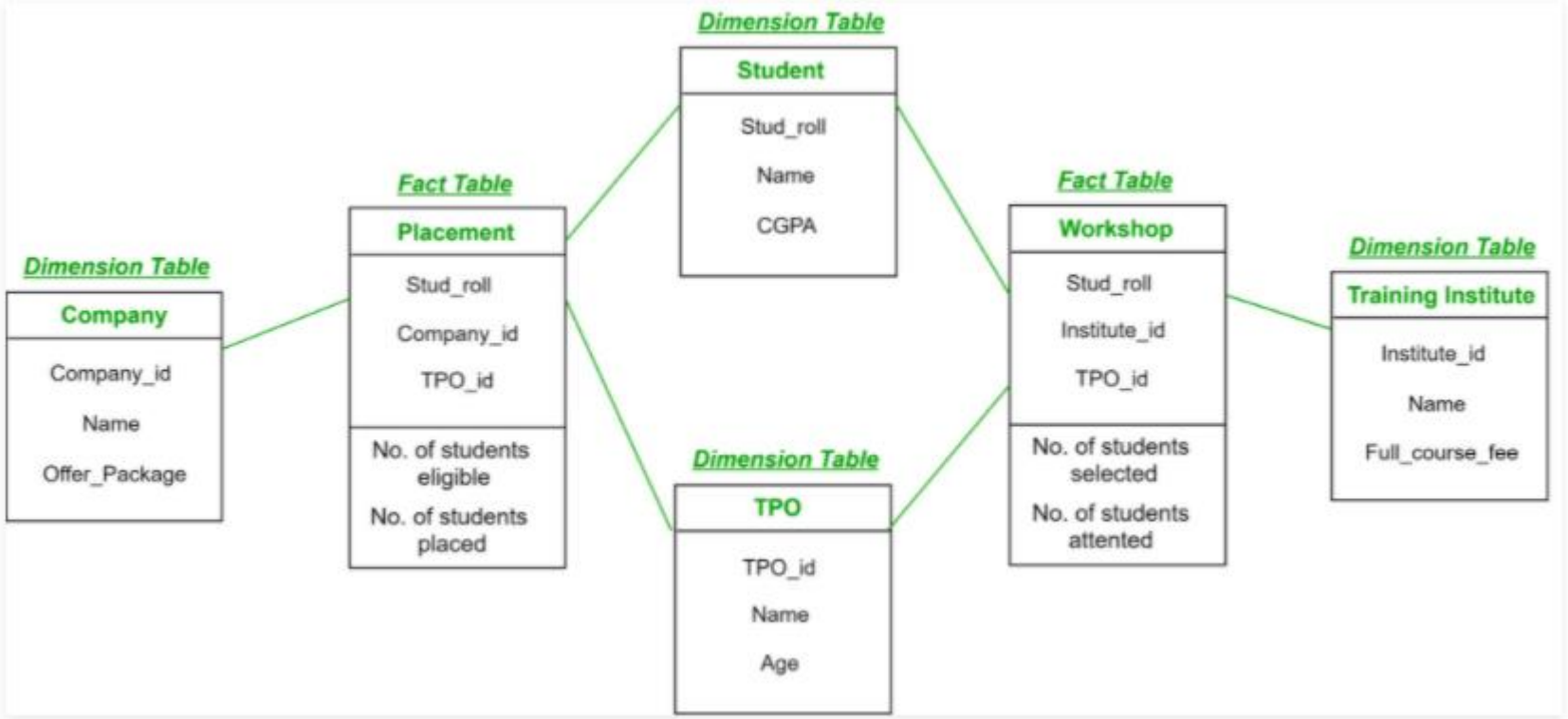


Image Source

Fact Tables: Characteristics

- Each row represents direct facts and connected to associated dimension tables via foreign keys
- Primary key is a composite key made up of all or a subset of foreign keys
  - A surrogate key can also be created to work as a primary key.
- Usually, the fact table is in third-normal form (3NF), while dimensional tables are denormalized.
- Grain of a fact table indicates level at which information is measured.
  - E.g., One row for per store / per product / per Day

Types of Facts

- **Additive:** Can be summed up through all the dimensions in the fact table.
  - E.g., Table containing sales transactions per product in each store :
    Date, Store Id, Product Id, **Sales Amount**

- **Semi-Additive:** Can be summed up for some of the dimensions in the fact table, but not the others.
  - E.g., Table recording the current balance for each account at the end of each day:
    Date, Account no, **Current Balance**

- **Non-Additive:** Cannot be summed up for any of the dimensions present in the fact table.
  - E.g., Table containing sales transactions per product in each store :
    - Date, Store Id, Product Id, **Unit Price**, Sales Quantity

Types of Fact Tables

- **Transaction Fact Table:**
  - Represent an event that occurs at any instantaneous point in time.
  - Capture lowest grain data / most detailed level
  - Data is generally in additive nature
  - E.g., Record Transaction in a PoS (Point of Sales) System:
    Timestamp, Customer Id, Product Id, Qty, Store Id, Price

- **Snapshot (Periodic) Fact Table:**
  - Describes the state of things in a particular instance of time
  - The 'grain' or 'level of resolution' is the period, not the individual transaction
  - E.g., Fact table recording the current balance for each account at the end of each day:
    Date, Account no, Current Balance
  - Provide an overview of the trend lines in the key performance indicators
  - A transaction fact table could be used as a source for this

Types of Fact Tables

- **Accumulated / Cumulative Fact Table**:
  - Describes what has happened over a period, in a given process with definite start and end
  - Will be filled when an order goes through the cycle
  - Grain: One row per entire lifetime of an event
  - E.g., Fact Table containing all the facts related to order processing

    Order Date, Invoice Date, Shipment Date, Return Date, Delivery Date, Store Id, Invoiced Qty, Ordered Qty, Shipped Qty, Returned Qty

- **Fact less Fact Table**:
  - Transaction fact tables which contain no measures.
  - E.g., Student taking a class of a certain lecturer:

    Student Id, Lecturer Id, Module Code, Semester ID

Types of Dimensions

- **Conformed Dimension:**
  - Dimension that is shared across multiple data marts or subject areas.
  - Possible to use the same dimension table across different projects without making any changes (Conforms to all Fact tables)
    - E.g., Time Dimension Table, definition of a year should be same for both HR and Finance
  - Guarantees consistent reporting across organization

- **Junk Dimension:**
  - Grouping of typically low cardinality attributes,
  - It contains different or various attributes which are unrelated to any other attribute.
    - E.g., Payment Modes (Cash or Credit Card) and Store Types (Super Market or Hyper Market) in a Sales Transaction
    - Create a JUNK dimension tables containing rows for possible combinations of above two fields. Add a surrogate key and use it in FACT table

Types of Dimensions

- **Degenerated Dimension:**
  - A dimension that is derived from fact table and does not have its own dimension table
    - E.g., Order No. in a Sales fact table

- **Role Playing Dimension:**
  - Dimensions which are often used for multiple purposes within the same database
    - E.g., Using date dimension for “order date”, “invoice date”, and “shipment date”.

      Using Customer Address as “Billing Address” and “Shipping Address”
  - If we use single dimension table called “Date”, all above dates would have the same value
  - Solution: Create multiple views from the dimension table and link them to the fact table

Frequency of change in Dimensions

- Unchanging / Static Dimensions (UCD)
  - Dimensions values are static and will not change.
    - E.g., Birthdate of a customer
- Slowly Changing Dimension (SCD)
  - Attribute values changes slowly over time
    - E.g., Address and phone number of a customer.
- Rapidly Changing Dimension (RCD)
  - Attribute values changes rapidly leading to performance implications
    - E.g., Weight and BMI of a patient

SCD Types

- Organizations need to keep track of changes in these dimension values
- There are few approaches to handle this known as SCD types:
- Type 0: Always retains original
- Type 1 : Keeps latest data, old data is overwritten
  - There's no historical data, easy maintenance, reduce size

SCD Types

- Type 2 : Keeps the history of old data by adding new row
  - Not recommended where a new attribute could be added in future

Customer ID	Name	Mobile No	Effective From	Effective Till	Flag
123	A. Perera	+94123456	2022-01-01	2022-01-31	0
123	A. Perera	+94987654	2022-01-31	Null	1

- Type 3 : Adds new attribute to store changed value
  - Keeps limited history about changed data

Customer ID	Name	Previous Mobile No	Current Mobile No
123	A. Perera	+94123456	+94987654

SCD Types

- Type 4 : Uses separate history table

Original Table:

Customer ID	Name	Mobile No
123	A. Perera	+94987654

History Table:

Customer ID	Name	Mobile No	Created Date
123	A. Perera	+94123456	2022-01-01
123	A. Perera	+94987654	2022-01-31

SCD Types

- Type 6 : Combination of type 1, 2 and 3

Customer ID	Name	Current Mobile No	Previous Mobile No	Effective From	Effective To	Flag
123	A. Perera	+94123456	+94000111	2022-01-01	2022-01-31	0
123	A. Perera	+94987654	+94123456	2022-01-31	Null	1

Handling Rapidly Changing Dimensions

- Separate RCDs into separate dimension table and connect to main dimension via a mini dimension

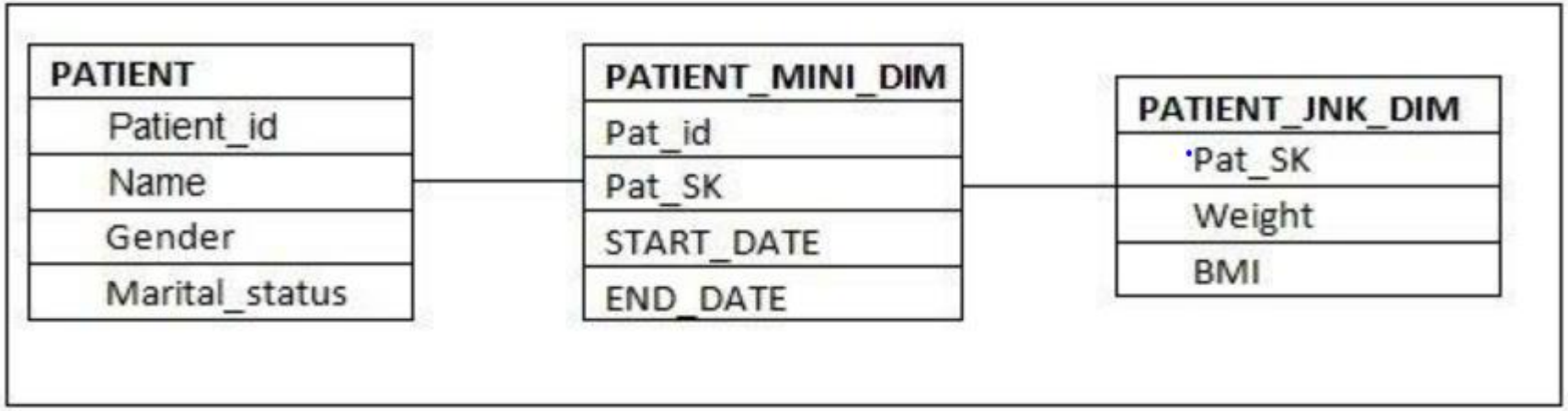


Image Source

Surrogate Key for Dimension tables

- Anonymous integer primary key
- Generated as a sequence and not driven by application data
- When data is accumulated from multiple sources values of the same column (primary Key) may be of different formats
- When adding duplicate rows (e.g., SCD 2), primary key would be repeated
- Due to small size very effective when joining with fact table

Further Reading

- The Data Warehouse Toolkit, : The Definitive Guide to Dimensional Modeling, 3rd Edition by Ralph Kimball (Author), Margy Ross (Author)