# ITLDM801 - DATA MINING AND DATA WAREHOUSING

NYIRIMANA Jean Marie Vianney

BACHELOR OF TECHNOLOGY (BTech) IN INFORMATION TECHNOLOGY

March 2025

RWANDA POLYTECHNIC

NGOMA COLLEGE

# Contents



RWANDA POLYTECHNIC

NGOMA COLLEGE

# Introduction to Data Warehousing

## What Is A Data Warehouse?

- A data warehouse is a powerful database model that significantly enhances the user's ability to quickly analyze large, multidimensional data sets.
- It cleanses and organizes data to allow users to make business decisions based on facts.
- Hence, the data in the data warehouse must have strong analytical characteristics.
- Creating data to be analytical requires that it be **subject-oriented**, **integrated**, **time-referenced**, and **non-volatile**.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Introduction to Data Warehousing...

**Subject-Oriented Data**

* In a data warehouse environment, information used for analysis is organized around subjects: employees, accounts, sales, products, and so on.
* This subject-specific design helps in reducing the query response time by searching through very few records to get an answer to the user's question.

**Integrated Data**

* Integrated data refers to de-duplicating information and merging it from many sources into one consistent location.
* Much of the transformation and loading work that goes into the data warehouse is centered on integrating data and standardizing it.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Introduction to Data Warehousing...

## Time-Referenced Data

* Time-referenced data essentially refers to its time-valued characteristic. For example, the user may ask:
  *What were the total sales of product 'A' for the past three years on New Year's Day across region 'Y'?*
* Time-referenced data when analyzed can also help in spotting the hidden trends between different associative data elements, which may not be obvious to the naked eye.
* This exploration activity is termed **data mining**.

## Non-Volatile Data

* The non-volatility of data, characteristic of data warehouses, enables users to dig deep into history and arrive at specific business decisions based on facts.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Introduction to Data Warehousing...

## Need for Data Warehousing

1. **Handling Large Volumes of Data:** Traditional databases can only store a limited amount of data (MBs to GBs), whereas a data warehouse is designed to handle much larger datasets (TBs), allowing businesses to store and manage massive amounts of historical data.
2. **Enhanced Analytics:** Transactional databases are not optimized for analytical purposes. A data warehouse is built specifically for data analysis, enabling businesses to perform complex queries and gain insights from historical data.
3. **Centralized Data Storage:** A data warehouse acts as a central repository for all organizational data, helping businesses to integrate data from multiple sources and have a unified view of their operations for better decision-making.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Introduction to Data Warehousing...

1. **Trend Analysis:** By storing historical data, a data warehouse allows businesses to analyze trends over time, enabling them to make strategic decisions based on past performance and predict future outcomes.
2. **Support for Business Intelligence:** Data warehouses support business intelligence tools and reporting systems, providing decision-makers with easy access to critical information, which enhances operational efficiency and supports data-driven strategies.

RP RWANDA POLYTECHNIC

NGOMA COLLEGE

# Warehousing architecture

- Data warehouse architecture is a data storage framework's design of an organization.
- It takes information from raw data sets and stores it in a structured and easily digestible format.
- A data warehouse architecture plays a vital role in the data enterprise.
- As databases assist in storing and processing data, and data warehouses help in analyzing that data.
- Data warehousing is a process of storing a large amount of data by a business or organization.
- The data warehouse is designed to perform large complex analytical queries on large multi-dimensional datasets in a straightforward manner.
- Data warehouses extract data from different resources, which are in different fonts, convert it into a unique form, and place data in Data Warehouse.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Warehousing architecture....

## Types of Data Warehouse Architectures

* Data warehouse architecture defines the arrangement of the data in different databases.
* As the data must be organized and cleansed to be valuable, a modern data warehouse structure identifies the most effective technique of extracting information from raw data.
* Using a dimensional model, the raw data in the staging area is extracted and converted into a simple consumable warehousing structure to deliver valuable business intelligence.
* When designing a data warehouse, there are three different types of models to consider, based on the approach of number of tiers the architecture has.
  (i) **Single-tier data warehouse architecture**
  (ii)**Two-tier data warehouse architecture**
  (iii)**Three-tier data warehouse architecture**

RP RWANDA POLYTECHNIC NGOMA COLLEGE

inframeWarehousing architecture....

**(i) Single-tier data warehouse architecture**

* The single-tier architecture (Figure 1) is not a frequently practiced approach. The main goal of having such architecture is to remove redundancy by minimizing the amount of data stored.
* Its primary disadvantage is that it doesn't have a component that separates analytical and transactional processing.

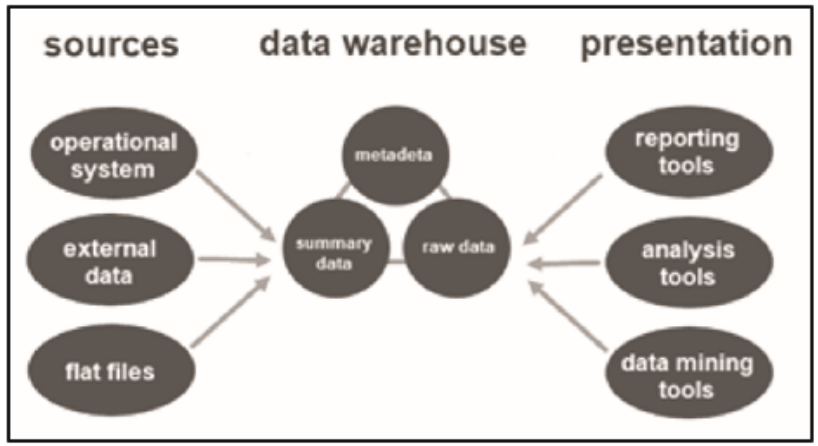


Figure 1: Single Tier Data Warehouse Architecture

# Warehousing architecture....

## (ii) Two-tier data warehouse architecture

* The two-tier architecture (Figure 2) includes a staging area for all data sources, before the data warehouse layer.
* By adding a staging area between the sources and the storage repository, you ensure all data loaded into the warehouse is cleansed and in the appropriate format.

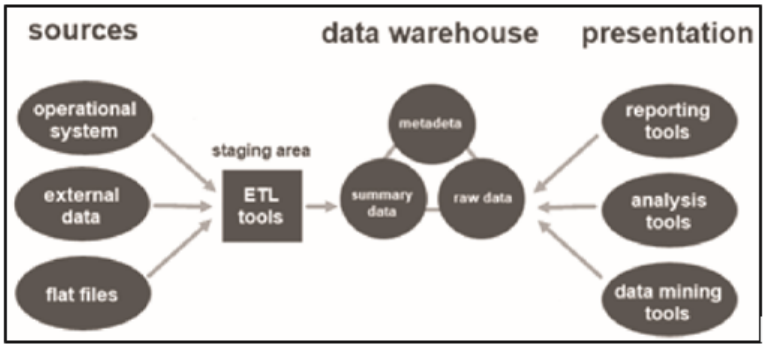


Figure 2: Two-Tier Data Warehouse Architecture

# Warehousing architecture....

**(iii) Three-tier data warehouse architecture**

* The three-tier approach (Figure 3) is the most widely used architecture for data warehouse systems.
  Essentially, it consists of three tiers:
* 1. **The bottom tier** is the database of the warehouse, where the cleansed and transformed data is loaded.
* 2. **The middle tier** is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.
* 3. **The top-tier** is where the user accesses and interacts with the data. It represents the front-end client layer. You can use reporting tools, query, analysis or data mining tools.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# Warehousing architecture....

## (iii) Three-tier data warehouse architecture...

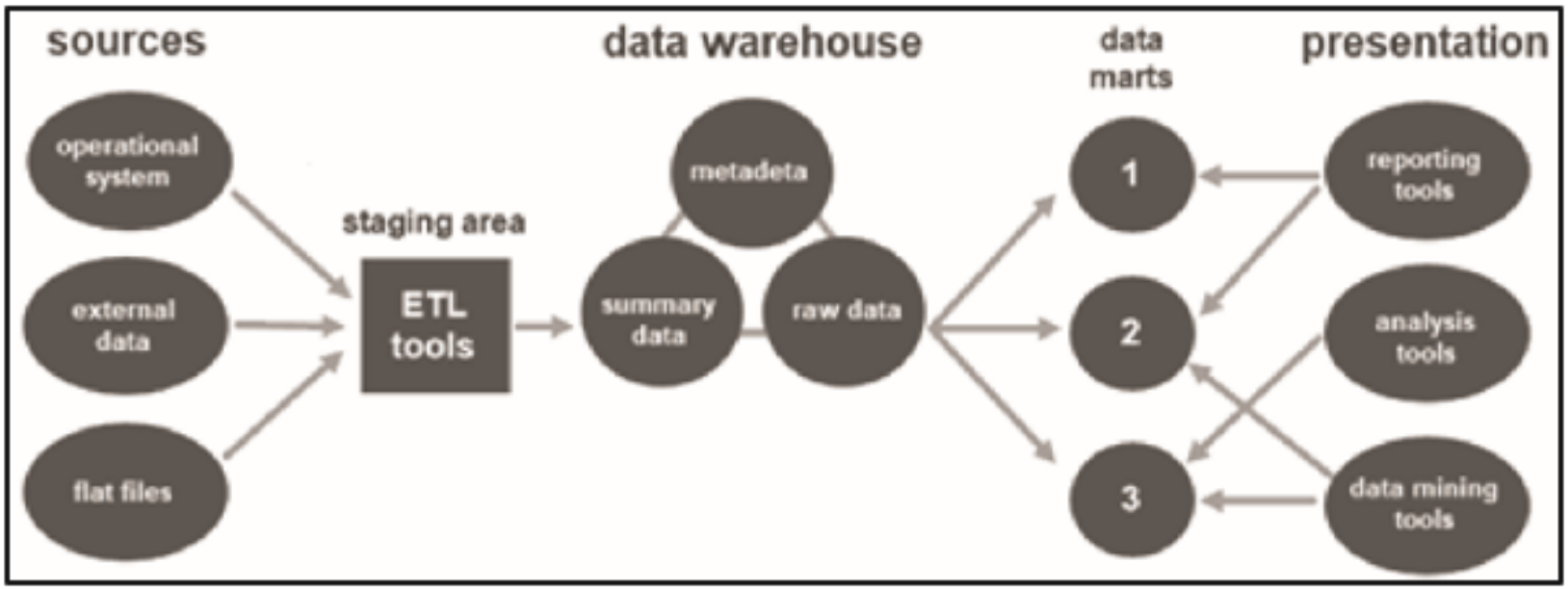


Figure 3: Three-Tier Data Warehouse Architecture

RWANDA POLYTECHNIC NGOMA COLLEGE

# Warehousing architecture....

## (iii) Three-tier data warehouse architecture...

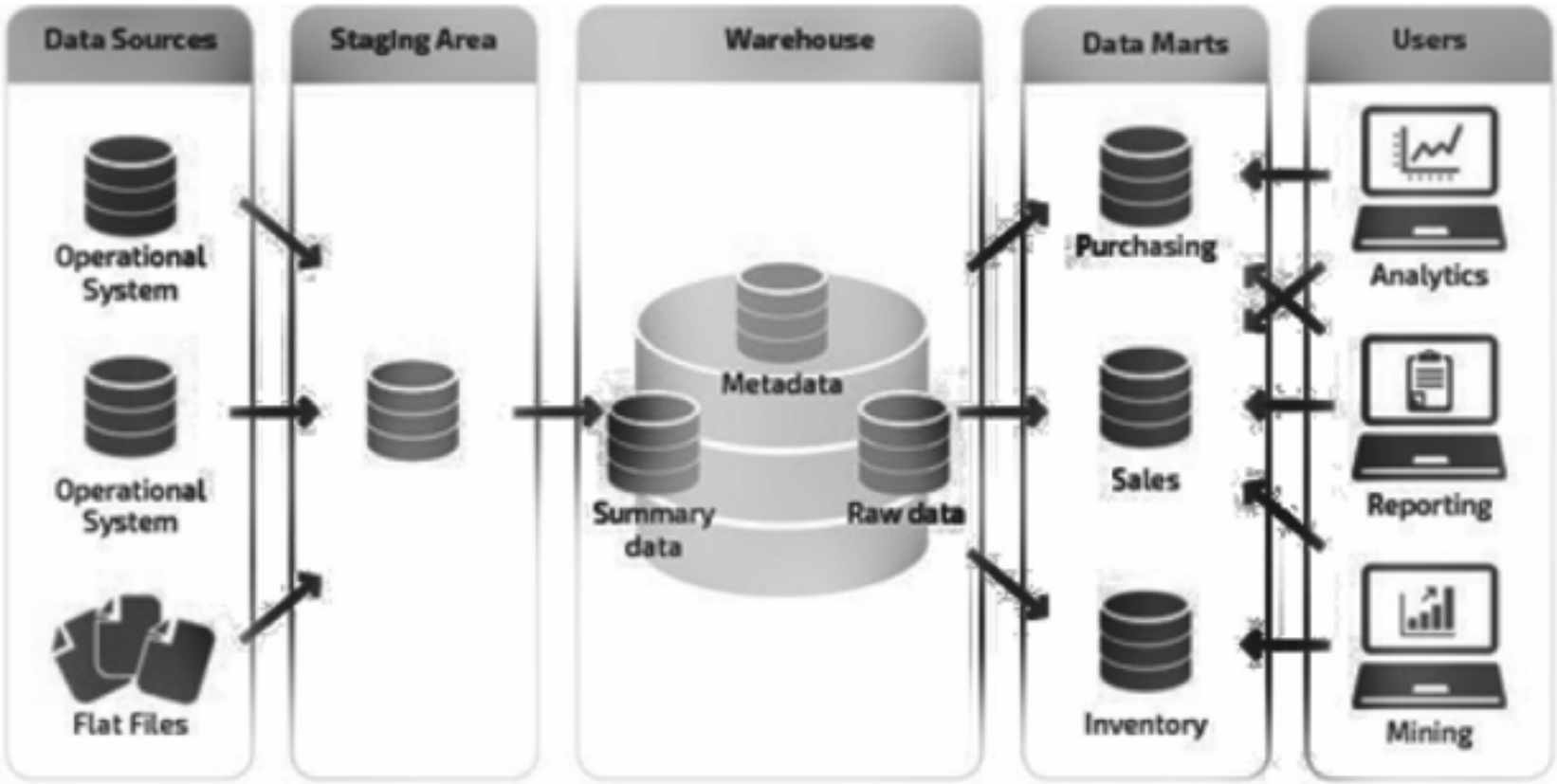


Figure 4: Three-Tier Data Warehouse Architecture

# Data Warehouse Models

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

**1. Enterprise warehouse**

* An enterprise warehouse collects all of the information about subjects spanning the entire organization.
* It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
* It typically contains detailed data as well as summarized data and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
* An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms.
* It requires extensive business modeling and may take years to design and build.

# Data Warehouse Models....

## 2. Datamart

* A data mart contains a subset of corporate-wide data that is of value to a specific group of users.
* The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to a customer, item, and sales.
* The data contained in data marts tend to be summarized.
* Depending on the source of data, data marts can be categorized into the following two classes: **Independent data marts** are sourced from data captured from one or more operational systems or external information providers, or data generated locally within a particular department or geographic area.
  **Dependent data marts** are sourced directly from enterprise data warehouses.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Data Warehouse Models....

**3. Virtual warehouse**

- A virtual warehouse is a set of views over operational databases.
- For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.

# Analytical Processing

## ANALYTICAL PROCESSING

# Data Cube or OLAP approach in Data Mining

## What is OLAP?

- OLAP stands for Online Analytical Processing, which is a technology that enables multi-dimensional analysis of business data.
- It provides interactive access to large amounts of data and supports complex calculations and data aggregation.
- OLAP is used to support business intelligence and decision-making processes.
- Grouping of data in a multidimensional matrix is called data cubes.
- In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes.
- This multidimensional data is represented in the data cube as the cube represents a high-dimensional space.
- The Data cube pictorially shows how different attributes of data are arranged in the data model.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Data Cube or OLAP approach in Data Mining....

Below is the diagram of a general data cube.

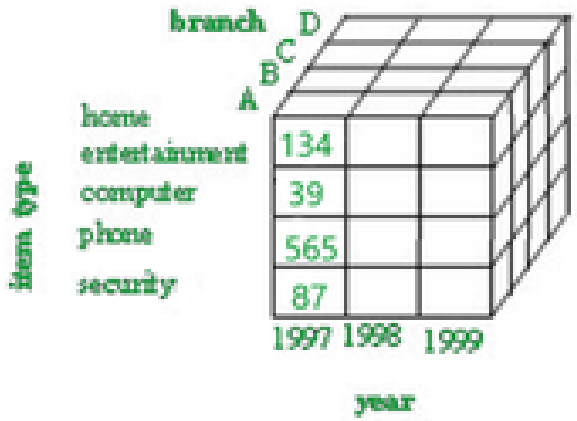


Figure 5: 3D cube

The example above is a 3D cube having attributes like branch(A,B,C,D),item type(home,entertainment,computer,phone,security), year(1997,1998,1999) .

# Data Cube or OLAP approach in Data Mining....

**Data cube classification:**
The data cube can be classified into two categories:

- **Multidimensional data cube:** It basically helps in storing large amounts of data by making use of a multi-dimensional array.
It increases its efficiency by keeping an index of each dimension.
Thus, dimensional is able to retrieve data fast.

- **Relational data cube:** It basically helps in storing large amounts of data by making use of relational tables.
Each relational table displays the dimensions of the data cube.
It is slower compared to a Multidimensional Data Cube.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# Data Cube or OLAP approach in Data Mining....

**Data cube operations:**

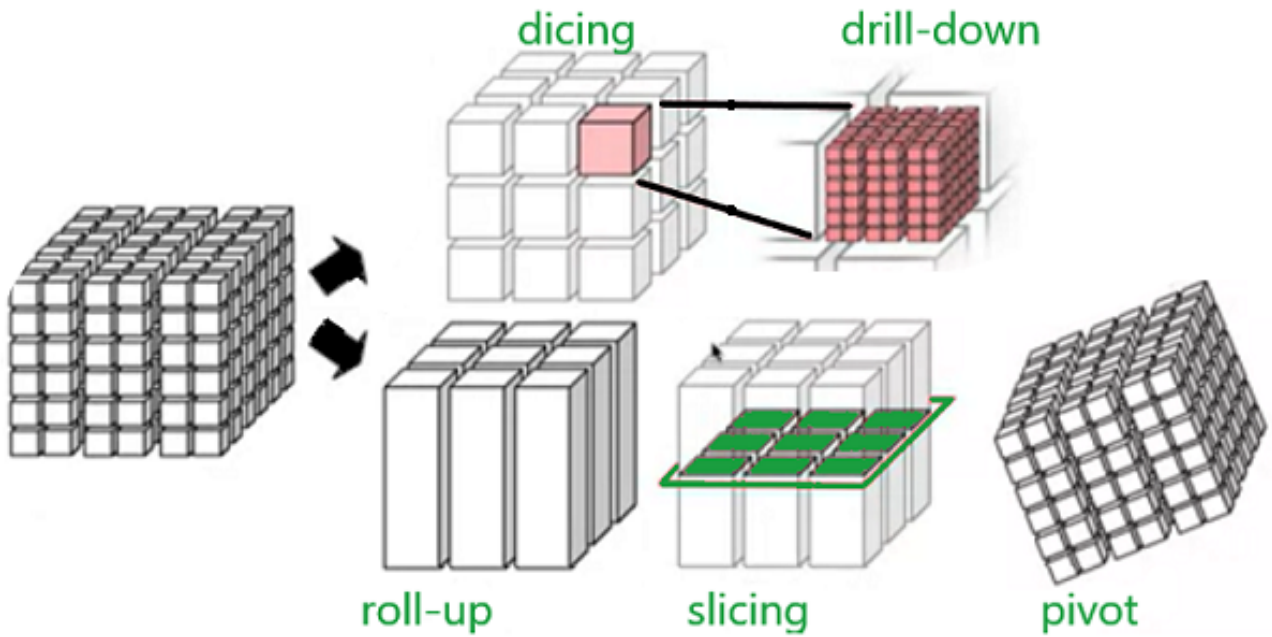


Figure 6: 3D Cube operations

# Data Cube or OLAP approach in Data Mining....

Data cube operations are used to manipulate data to meet the needs of users. These operations help to select particular data for the analysis purpose. There are mainly 5 operations listed below:

* **1. Roll-up:** operation and aggregate certain similar data attributes having the same dimension together.
  For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.
* **2. Drill-down:** this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis.
  It zooms into more detail. For example- if Rwanda is an attribute of a country column and we wish to see villages in Rwanda, then the drill-down operation splits Rwanda into states, districts, towns, cities, villages and then displays the required information.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Data Cube or OLAP approach in Data Mining....

* **3. Slicing:** this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a particular attribute. For example, country="jamaica", this will display only about jamaica and only display other countries present on the country list.
* **4. Dicing:** this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure). For example- the user wants to see the annual salary of Jharkhand state employees.
* **5. Pivot:** this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn't change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Comparison of OLAP and OLTP

| Category | OLAP | OLTP |
|---|---|---|
| Definition | Online database query management | Online database modifying system |
| Data Source | Historical data from various databases | Operational current data only |
| Method Used | Uses a data warehouse | Uses a standard DBMS |
| Application | Data Mining, Analytics, Decision making | Business tasks |
| Normalization | Tables are not normalized | Tables are normalized (3NF) |
| Usage of Data | Planning, problem-solving, decision-making | Day-to-day fundamental operations |
| Task | Multi-dimensional view of business tasks | Snapshot of present business tasks |
| Purpose | Extracts information for analysis | Insert, Update, Delete operations |
| Volume of Data | Large (TB, PB) | Small (MB, GB) |
| Query Speed | Slow (may take hours) | Fast (operates on 5% of data) |
| Update Frequency | Rarely updated | Frequently updated |
| Backup | Infrequent | Regular and rigorous |
| Processing Time | Lengthy (complex queries) | Fast (simple queries) |
| Users | CEO, MD, GM | Clerks, Managers |
| Operations | Mostly read | Read and write |
| Nature of Audience | Customer-focused | Market-focused |
| Productivity | Enhances business analysts' efficiency | Enhances users' productivity |

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

- Schema is a logical description of the entire database.
- It includes the name and description of records of all record types including all associated data-items and aggregates.
- Much like a database, a data warehouse also requires to maintain a schema.
- A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

RWANDA POLYTECHNIC NGOMA COLLEGE

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

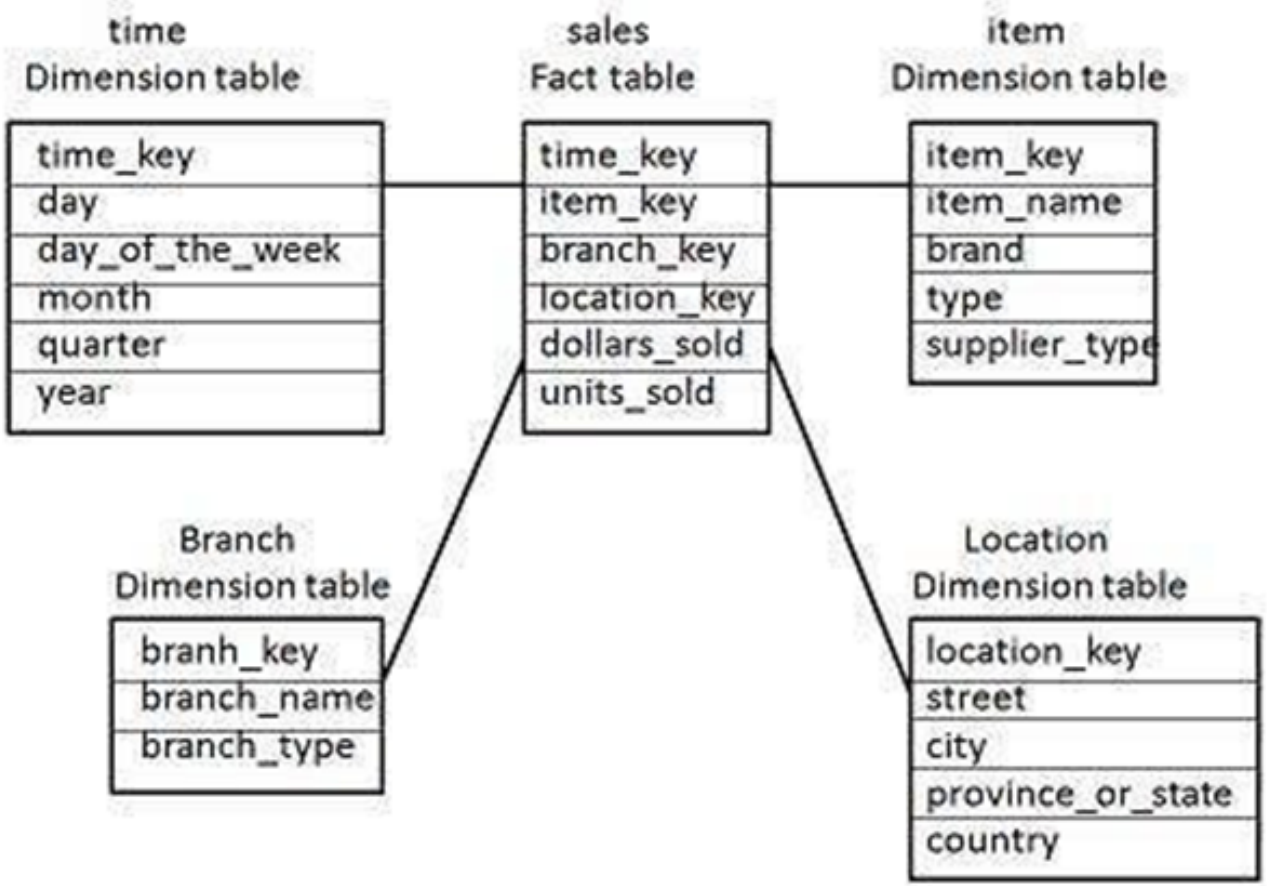


Figure 7: Star Schema

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

## STAR SCHEMA

Each dimension in a star schema is represented with only one-dimension table.

- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.
- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

RWANDA POLYTECHNIC NGOMA COLLEGE

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

## SNOWFLAKE SCHEMA

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized.
- For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.
- • The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

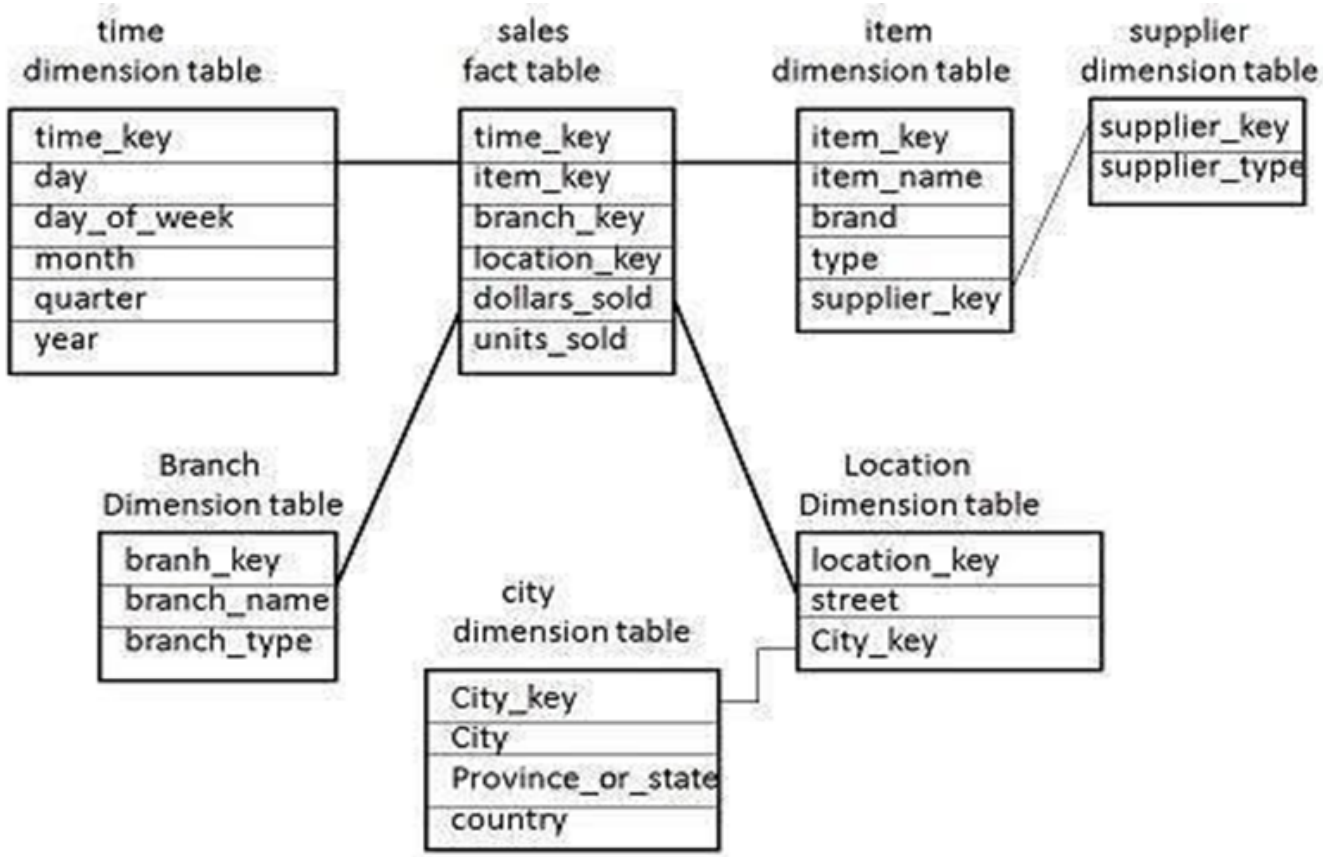


Figure 8: SNOWFLAKE SCHEMA

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

## FACT CONSTELLATION SCHEMA

* A fact constellation has multiple fact tables. It is also known as galaxy schema.
* The following diagram shows two fact tables, namely sales and shipping. The sales fact table is same as that in the star schema.
* The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
* The shipping fact table also contains two measures, namely dollars sold and units sold.
* It is also possible to share dimension tables between fact tables.
* For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# SCHEMAS FOR MULTI-DIMENSIONAL DATA MODEL

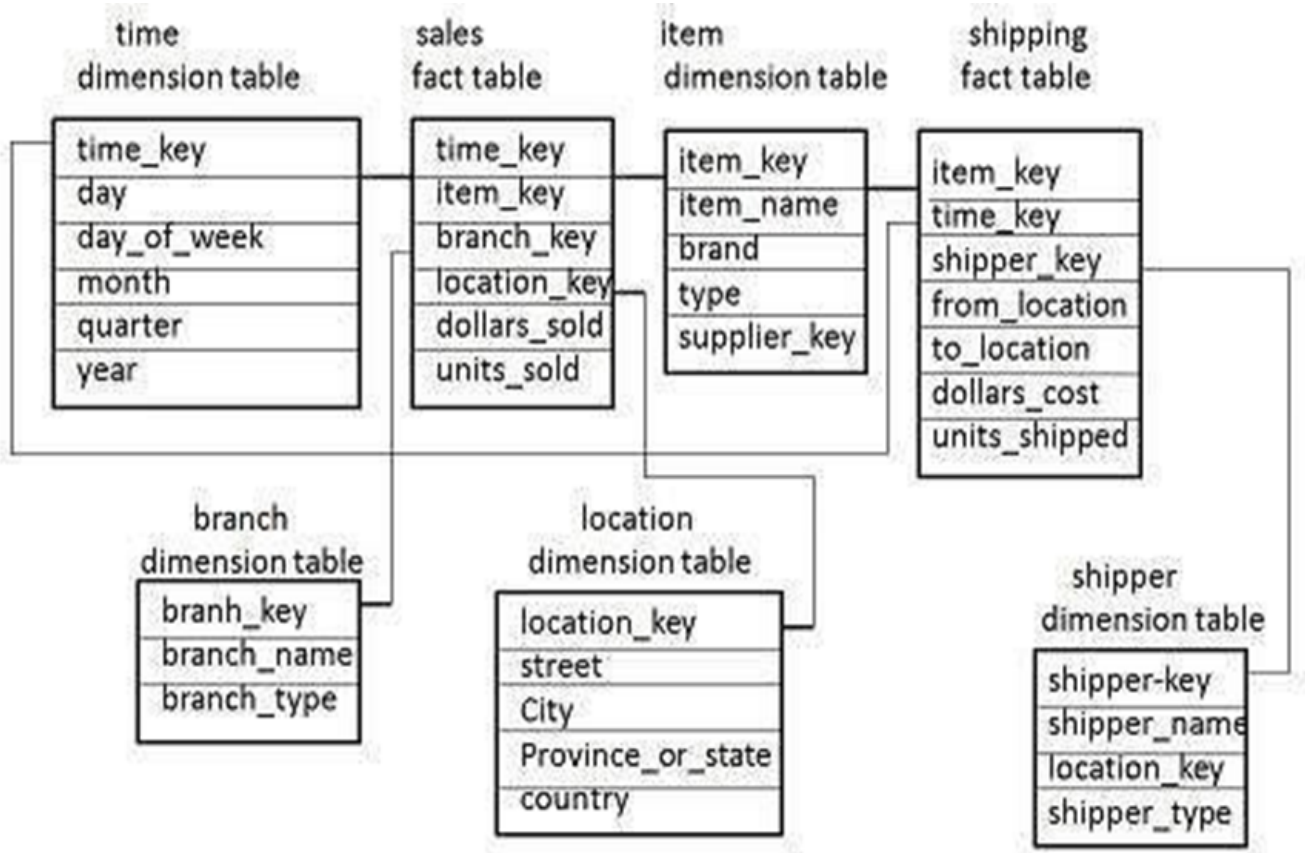


Figure 9: FACT CONSTELLATION SCHEMA

# Data processing overview

- Data preprocessing is the process of transforming raw data into a useful, understandable format.
- Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete.
- Data preprocessing resolves such issues and makes datasets more complete and efficient to perform data analysis.
- It's a crucial process that can affect the success of data mining and machine learning projects.
- It makes knowledge discovery from datasets faster and can ultimately affect the performance of machine learning models.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# Data processing overview

- Data preprocessing is essential to effectively build models with these features.
- Numerous problems can arise while collecting the data.
- You may have to aggregate data from different data sources, leading to mismatching data formats, such as integer and float.
- By preprocessing data, we make it easier to interpret and use.
- This process eliminates inconsistencies or duplicates in data, which can otherwise negatively affect a model's accuracy.
- Data preprocessing also ensures that there aren't any incorrect or missing values due to human error or bugs.
- In short, employing data preprocessing techniques makes the database more complete and accurate.

RP RWANDA POLYTECHNIC

NGOMA COLLEGE

# Purpose of Data Preprocessing

Typical location properties in vast real-world datasets and databases are incomplete, chaotic and unfailing information. Unfinished data can arise for a number of reasons:

* Important attributes cannot always be usable.
* Valid data cannot be recorded because of misunderstanding or equipment failure.
* Data that disputed other recorded data should have been discarded.
* Missing data may be inferred, especially for tuples with missing values for certain attributes.
* Data collection techniques may be unreliable.
* At the moment of data entry, human or computer errors may have existed.

RWANDA POLYTECHNIC

NGOMA COLLEGE

# Purpose of Data Preprocessing

- Data processing failures can also occur.
- There might be technological disadvantages, such as narrow buffer sized for simultaneous data transfer and usage coordination.
- Data Routines for the cleaning of data are used by filling the missing values.
- Relieve noise effects, detect or remove outliers and address inconsistencies.
- Data integration is the hybrid process with many cubes or archives of databases. However, in multiple databases, those characteristics that define a specific may have separate titles, which lead to inconsistencies and redundancies.
- Data transformation is a process approach such as standardizations and consolidation that constitutes additional preprocessing processes that contribute to mining process results.

RWANDA POLYTECHNIC NGOMA COLLEGE

# Purpose of Data Preprocessing

* Data reduction obtains a simpler data collection image, which is significantly smaller in duration, but provides the same analytical efficiency (or nearly the same). A multitude of methods for data reduction are in use.
This includes:
- **Data Aggregation** (e.g., Creating a data cube).
- **Attribute subset selection** (By similarity, eliminating unnecessary attributes)
- **Dimensionality Reduction** (e.g., using encoding systems such as minimal encoding lengths or wavelets).
- **Numerosity Reduction** (e.g., "Replace" details by alternating, smaller representations such as clusters or parametric structures.
- **Generalization** (e.g., Data were minimized with the usage of the definition hierarchy).

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# Factors Contributing to Data Quality

Factors Contributing to Data Quality Before looking at how data is preprocessed, some factors contributing to data quality as are given below:

* **Accuracy:** Accuracy means that the information is correct. Outdated information, typos, and redundancies can affect a dataset's accuracy.
* **Consistency:** The data should have no contradictions. Inconsistent data may give you different answers to the same question.
* **Completeness:** The dataset shouldn't have incomplete fields or lack empty fields. This characteristic allows data scientists to perform accurate analyses as they have access to a complete picture of the situation the data describes.
* **Validity:** A dataset is considered valid if the data samples appear in the correct format, are within a specified range, and are of the right type.
* **Timeliness:** Data should be collected as soon as the event it represents occurs.

# Data preprocessing stages

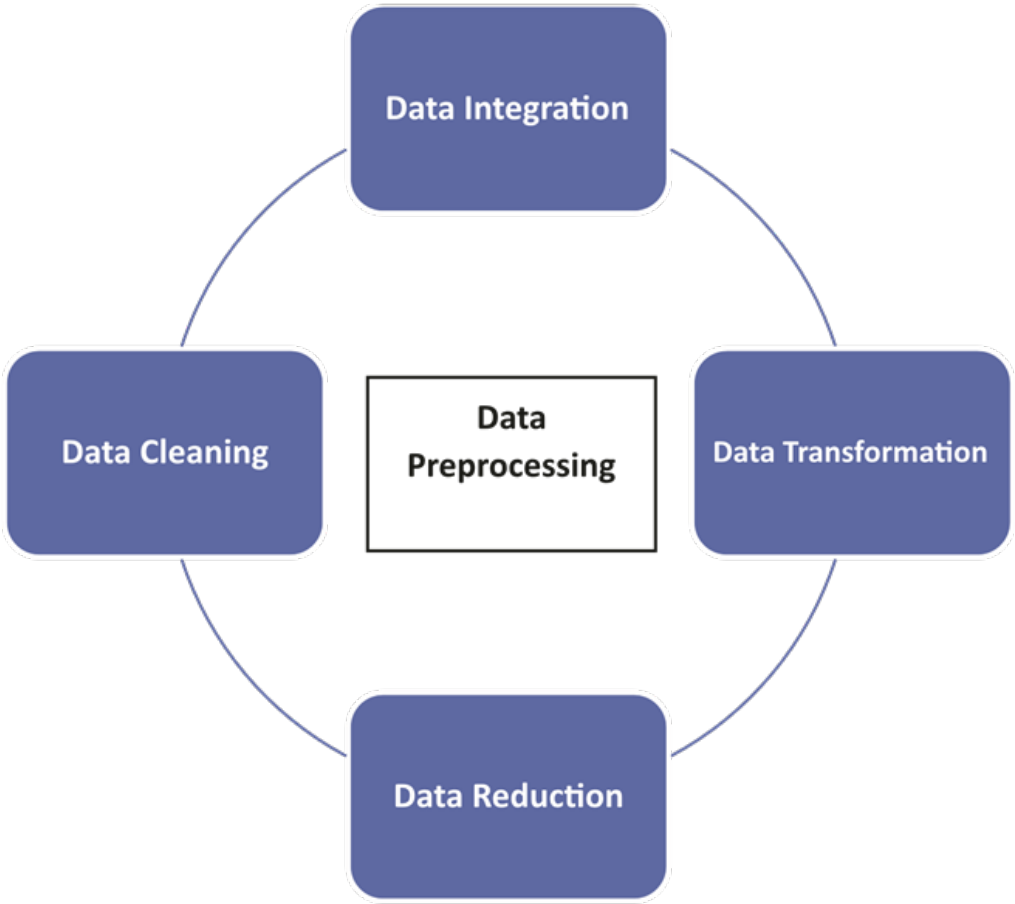


*Figure 1: Stages of Data Preprocessing*

# 1. Data Cleaning

- It is the process of identifying and correcting errors or inconsistencies in the dataset.
- It involves handling missing values, removing duplicates, and correcting incorrect or outlier data to ensure the dataset is accurate and reliable.
- Clean data is essential for effective analysis, as it improves the quality of results and enhances the performance of data models.
- **Missing Values:** This occur when data is absent from a dataset. You can either ignore the rows with missing data or fill the gaps manually, with the attribute mean, or by using the most probable value. This ensures the dataset remains accurate and complete for analysis.

RP RWANDA POLYTECHNIC

NGOMA COLLEGE

# 1. Data Cleaning

* **Noisy Data:** It refers to irrelevant or incorrect data that is difficult for machines to interpret, often caused by errors in data collection or entry. It can be handled in several ways:
-**Binning Method:** The data is sorted into equal segments, and each segment is smoothed by replacing values with the mean or boundary values.
-**Regression:** Data can be smoothed by fitting it to a regression function, either linear or multiple, to predict values.
-**Clustering:** This method groups similar data points together, with outliers either being undetected or falling outside the clusters. These techniques help remove noise and improve data quality.
-**Removing Duplicates:** It involves identifying and eliminating repeated data entries to ensure accuracy and consistency in the dataset. This process prevents errors and ensures reliable analysis by keeping only unique records.

RP RWANDA POLYTECHNIC NGOMA COLLEGE

# 2. Data Integration

- It involves merging data from various sources into a single, unified dataset.
- It can be challenging due to differences in data formats, structures, and meanings.
- Techniques like record linkage and data fusion help in combining data efficiently, ensuring consistency and accuracy.
- **Record Linkage** is the process of identifying and matching records from different datasets that refer to the same entity, even if they are represented differently. It helps in combining data from various sources by finding corresponding records based on common identifiers or attributes.
- **Data Fusion** involves combining data from multiple sources to create a more comprehensive and accurate dataset. It integrates information that may be inconsistent or incomplete from different sources, ensuring a unified and richer dataset for analysis.

RWANDA POLYTECHNIC NGOMA COLLEGE

# 3. Data Transformation

- It involves converting data into a format suitable for analysis.
- Common techniques include normalization, which scales data to a common range; standardization, which adjusts data to have zero mean and unit variance; and discretization, which converts continuous data into discrete categories.
- These techniques help prepare the data for more accurate analysis.
- **Data Normalization:** The process of scaling data to a common range to ensure consistency across variables.
- **Discretization:** Converting continuous data into discrete categories for easier analysis.
- **Data Aggregation:** Combining multiple data points into a summary form, such as averages or totals, to simplify analysis.
- **Concept Hierarchy Generation:** Organizing data into a hierarchy of concepts to provide a higher-level view for better understanding and analysis.

RWANDA POLYTECHNIC NGOMA COLLEGE

# 4. Data Reduction

- It reduces the dataset's size while maintaining key information.
- This can be done through feature selection, which chooses the most relevant features, and feature extraction, which transforms the data into a lower-dimensional space while preserving important details.
- It uses various reduction techniques such as:
  -**Dimensionality Reduction:** A technique that reduces the number of variables in a dataset while retaining its essential information. **(e.g., Principal Component Analysis)**
  -**Numerosity Reduction:** Reducing the number of data points by methods like sampling to simplify the dataset without losing critical patterns.
  -**Data Compression:** Reducing the size of data by encoding it in a more compact form, making it easier to store and process.

RWANDA POLYTECHNIC NGOMA COLLEGE

THANK YOU!!!!

RWANDA POLYTECHNIC

NGOMA COLLEGE