Chapter-1
Data Warehouse Fundamentals

PROF.NEHA SHAH

Agenda of Chapter 1

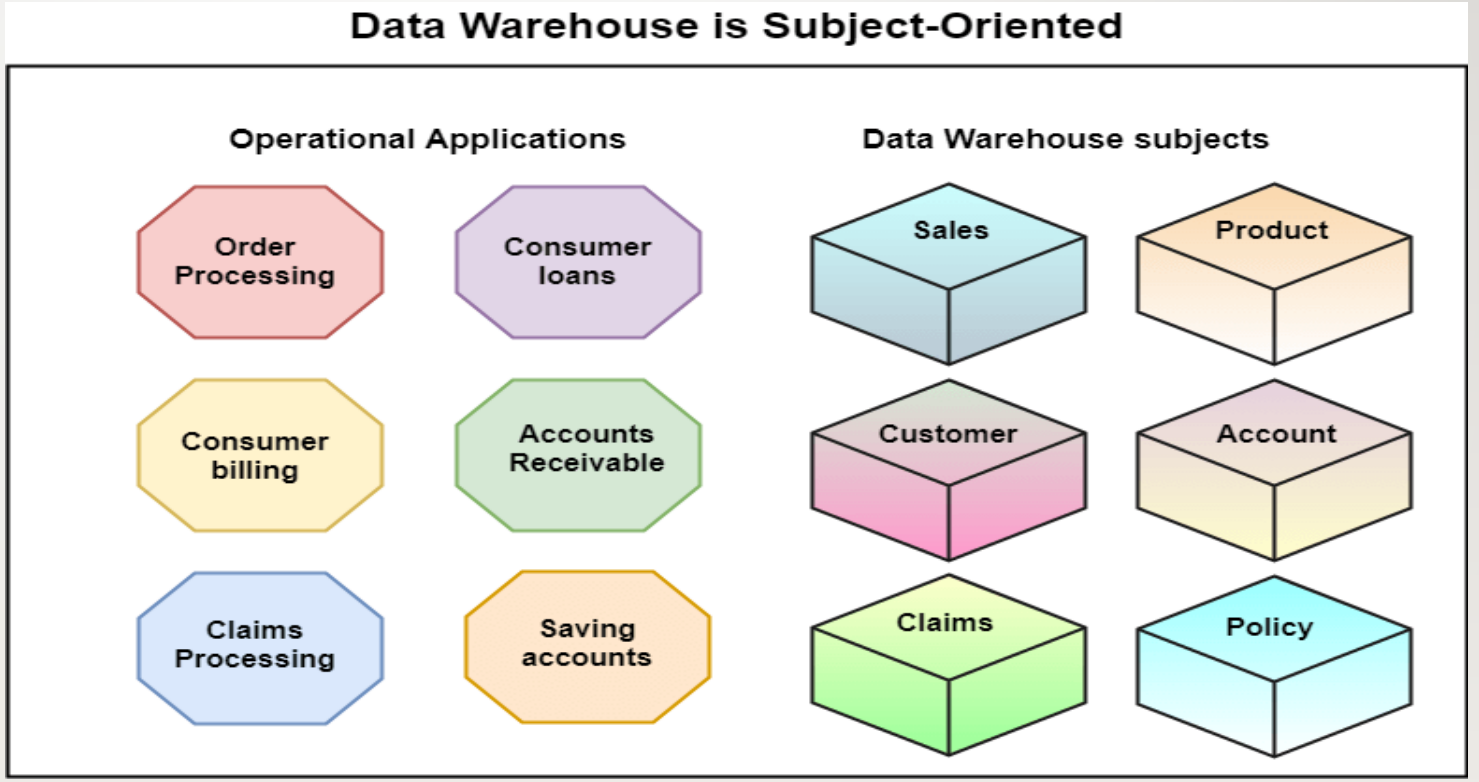
- 1. Data Warehouse Fundamentals
  - 1.1. Introduction to Data Warehouse, OLTP Systems; Differences between OLTP Systems and Data Warehouse:
  - 1.2. Characteristics of Data Warehouse; Functionality of Data Warehouse:
  - 1.3. Advantages and Applications of Data Warehouse; Top- Down and Bottom-Up Development Methodology:
  - 1.4. Tools for Data Warehouse Development: Data Warehouse Types
  - 1.5. Planning and Project Management in constructing Data warehouse: Data Warehouse Project;
  - 1.6. Data Warehouse development Life Cycle, Kimball Lifecycle Diagram

Data warehouse Fundamental

- Data warehouse-a logical collection of information-gathered from many different operational databases- that supports business analysis activities and decision-making tasks.
- The primary purpose of a data warehouse is to aggregate information throughout an organization into a single repository for decision-making purposes

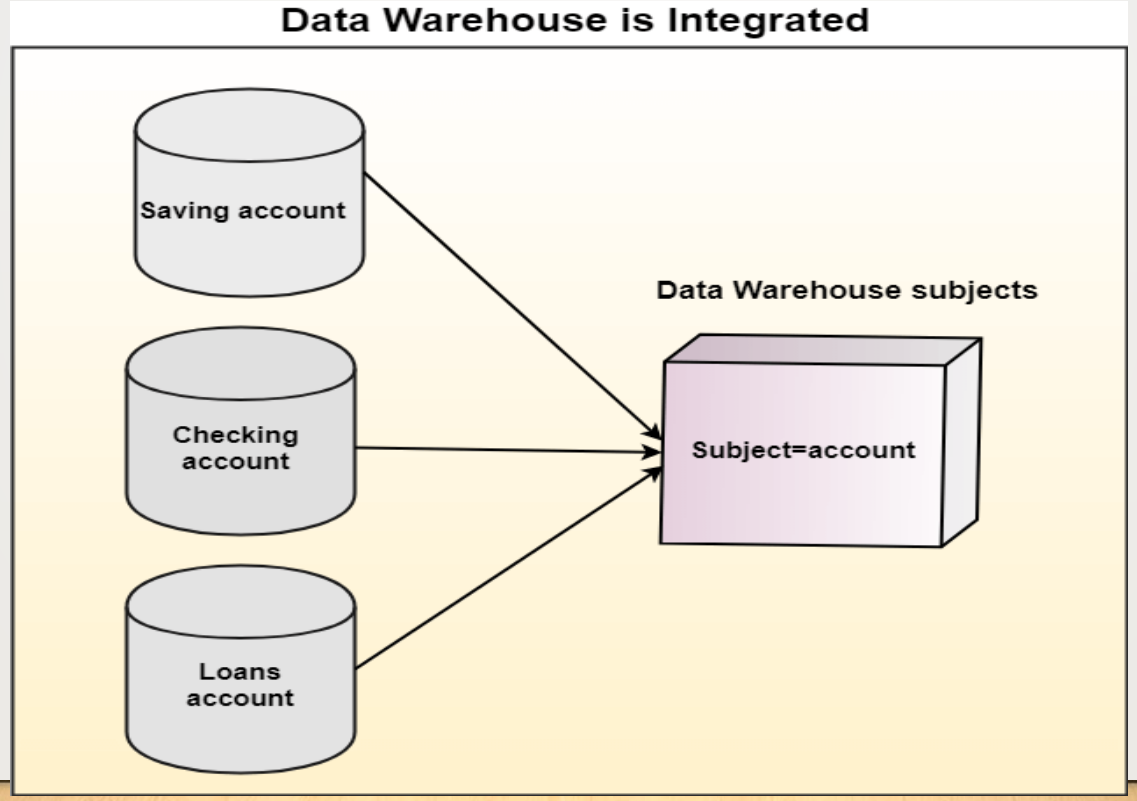
What is a Data Warehouse?

- "Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions.“
- subject-oriented

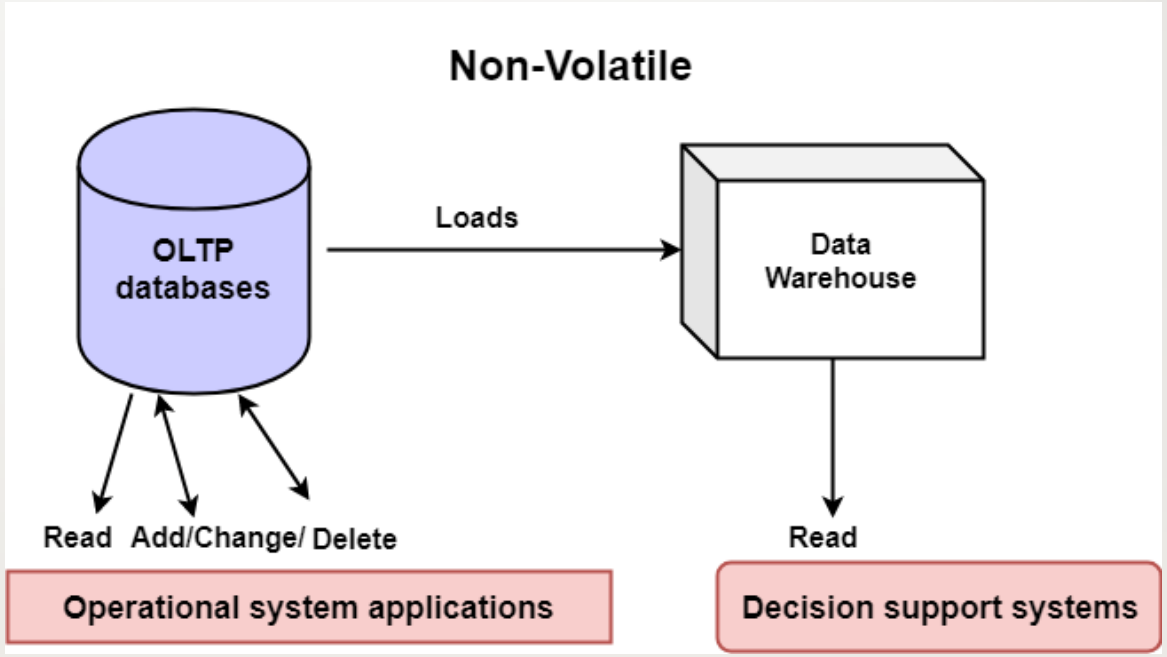


- Integrated

- A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.



- Time-Variant- Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse.
- Non-Volatile
- The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse



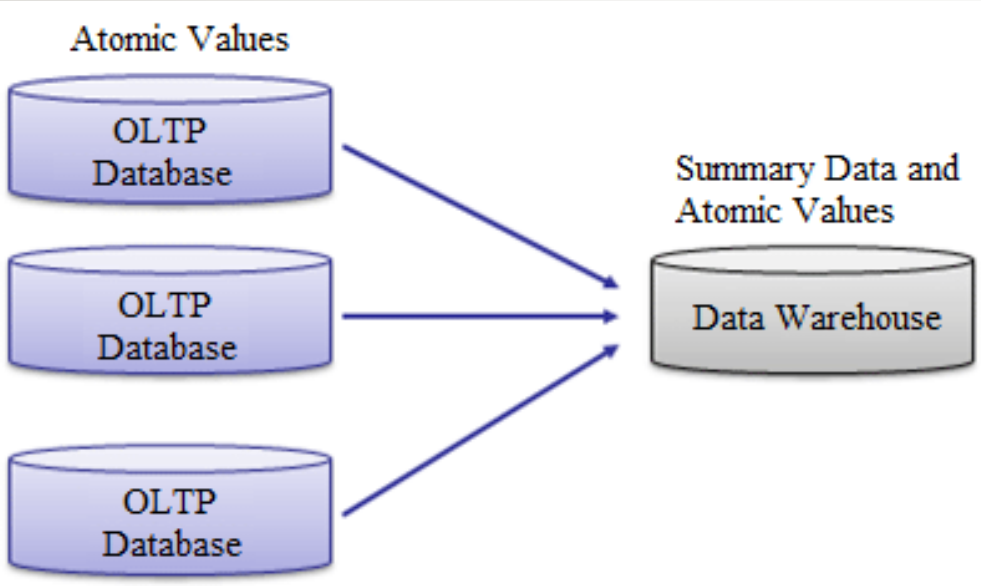
What is OLTP?

- **OLTP** is an operational system that supports transaction-oriented applications in a 3-tier architecture. It administers the day to day transaction of an organization.
- OLTP is basically focused on query processing, maintaining data integrity in multi-access environments
- OLTP system is an online database changing system. Therefore, it supports database query such as insert, update, and delete information from the database.

Characteristics of OLTP

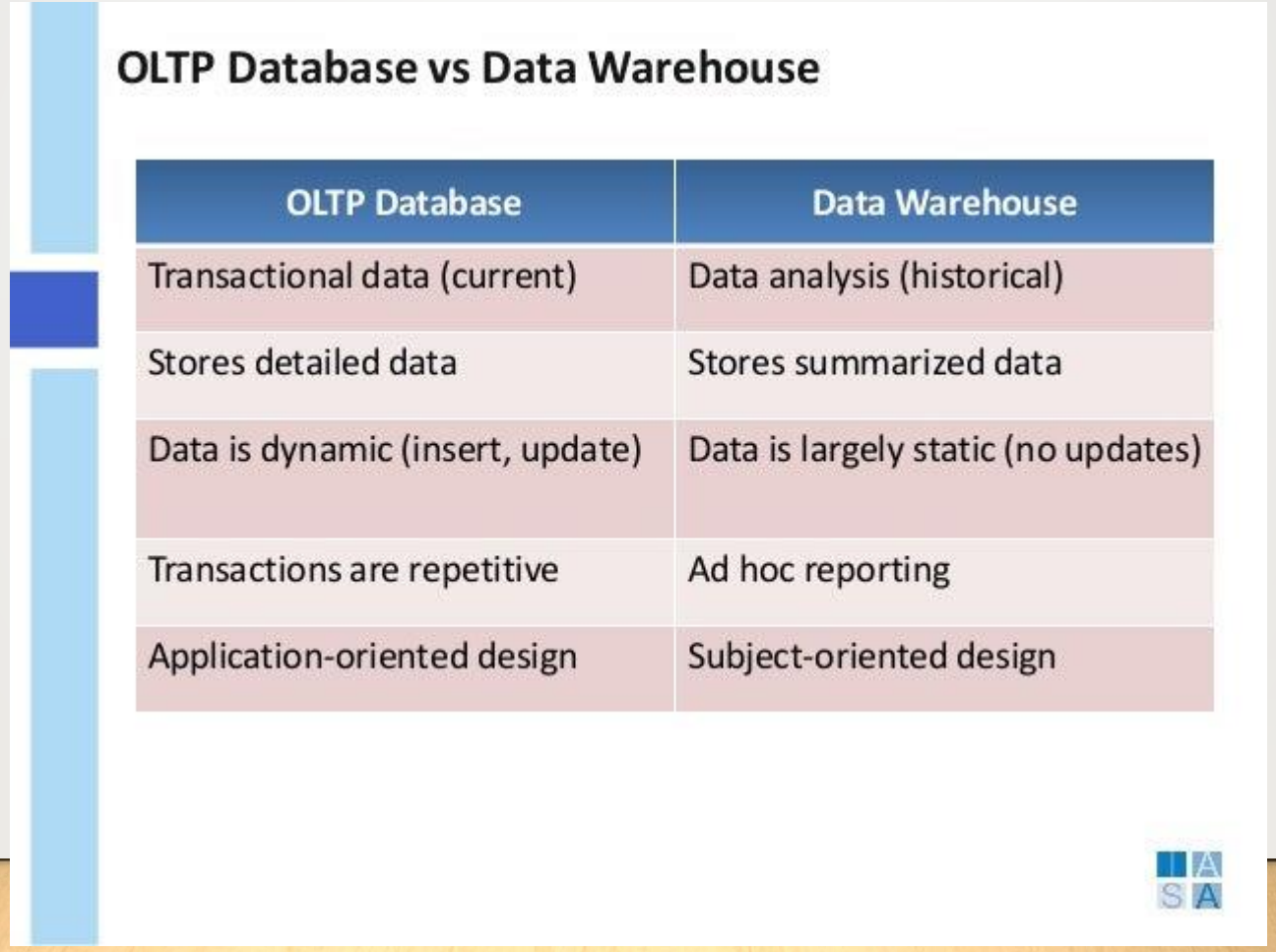
- OLTP uses transactions that include small amounts of data.
- Indexed data in the database can be accessed easily.
- OLTP has a large number of users.
- It has fast response times
- Databases are directly accessible to end-users
- OLTP uses a fully normalized schema for database consistency.
- The response time of OLTP system is short.
- It strictly performs only the predefined operations on a small number of records.
- OLTP stores the records of the last few days or a week.
- It supports complex data models and tables.

OLTP Architecture

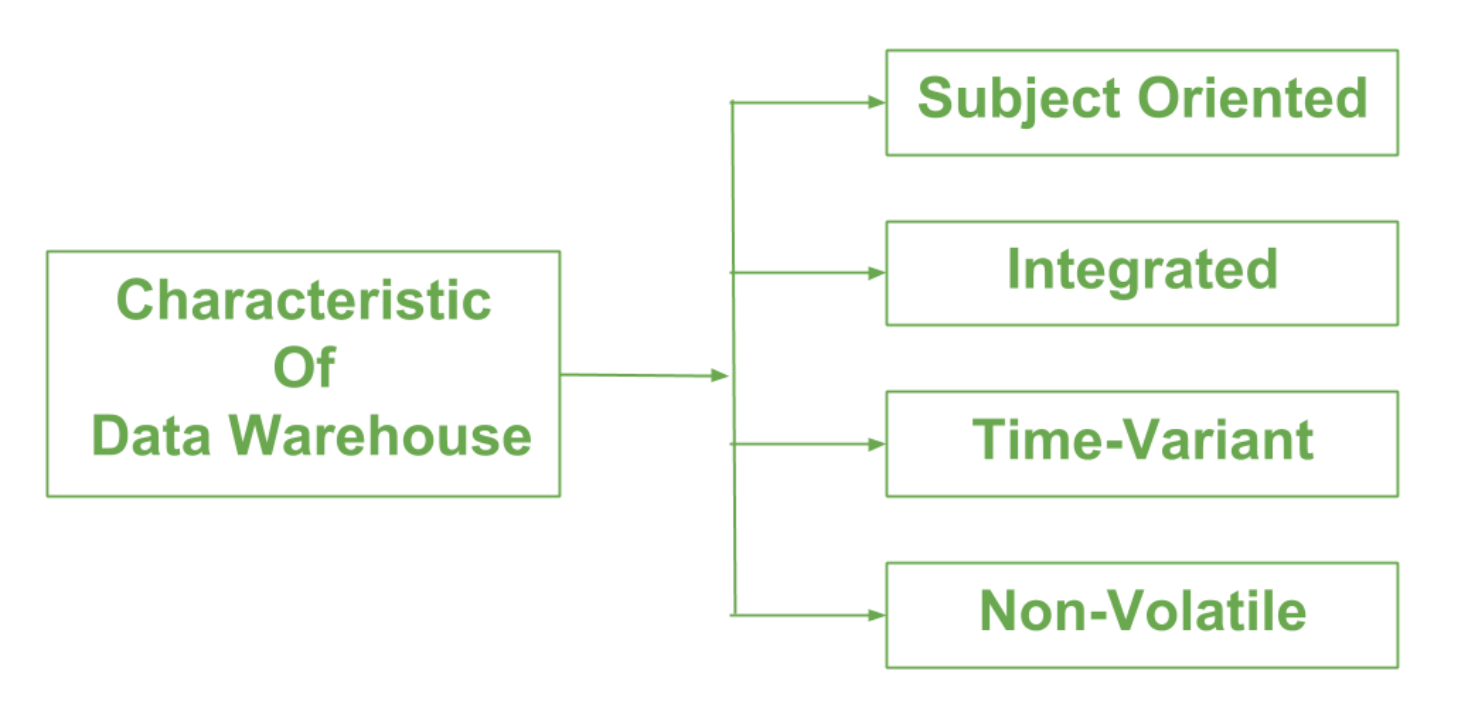


- Online transaction processing is a software program that supports transaction processes conducted online. An OLTP system ensures that businesses and individuals can complete transactions quickly, efficiently and accurately, Businesses may use OLTP for online banking. shopping and point-of-sale terminals.

Differences between OLTP Systems and Data Warehouse:



Characteristics of Data Warehouse



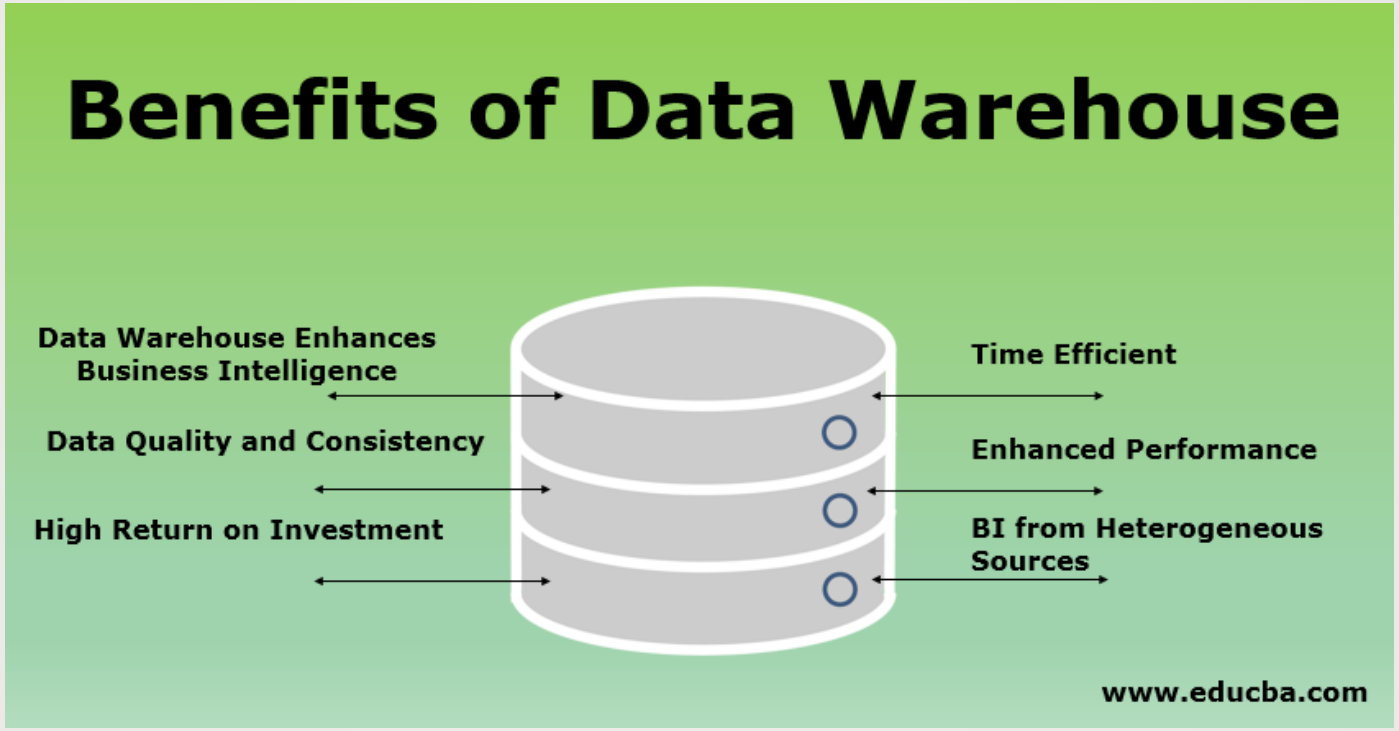
Functionality of Data Warehouse:

- Data Consolidation:
- Data Cleaning:
- Data Integration:
- Data Storage:
- Data Transformation:
- Data Analysis:
- Data Reporting:
- Data Mining:
- Performance Optimization:

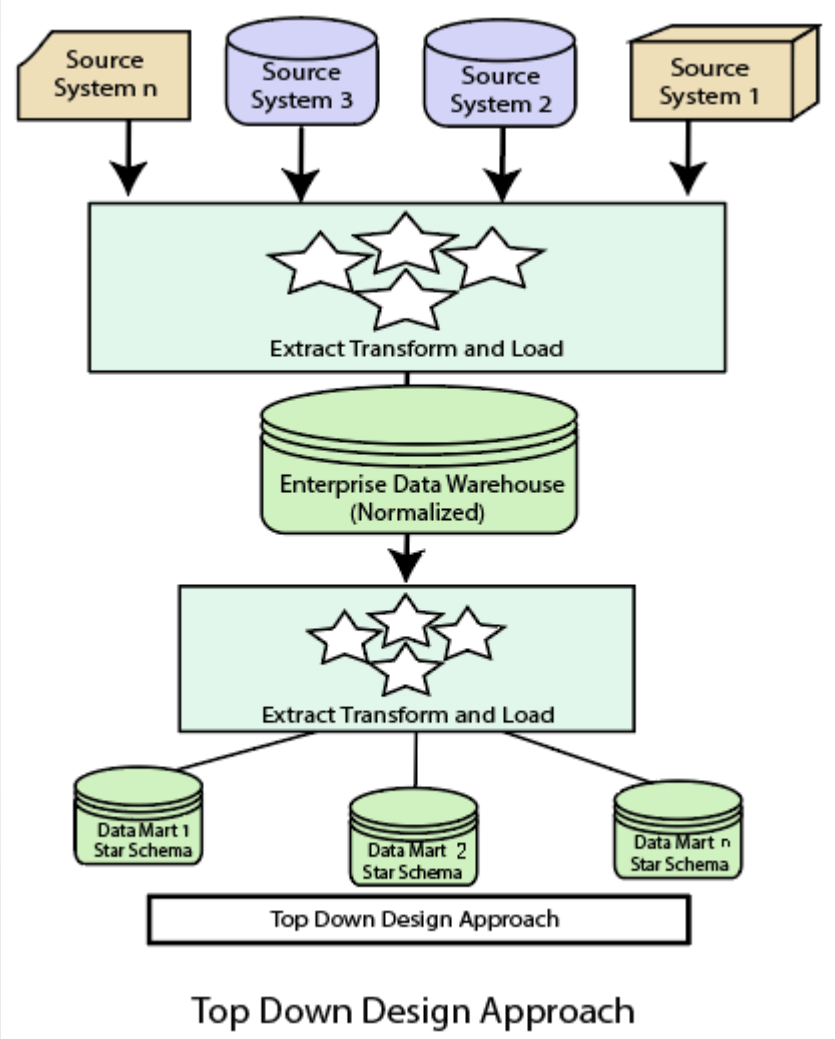
Benefits of Data Warehousing

- Data warehouses enable end-users to **access a wide variety of data**
- Business analysts and decision makers can **analyze the current trends** in the market to predict future trends
- Data warehouse provides **consistent data**
- It helps to **increase productivity** and **decrease computing costs**
- Data warehouses contain data that has been **integrated** from a number of different sources
- The results obtained can be presented in a variety of formats in the form of **reports, graphs**, etc.

Advantages and Applications of Data Warehouse

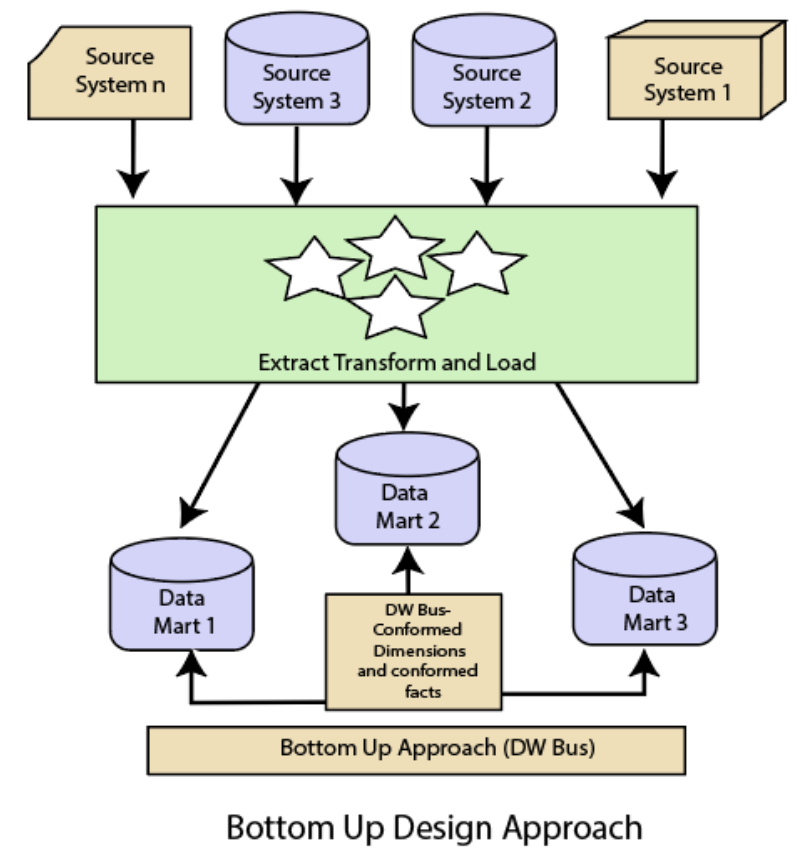


Top- Down Development Methodology:



Top Down Design Approach

Bottom-Up Design Approach



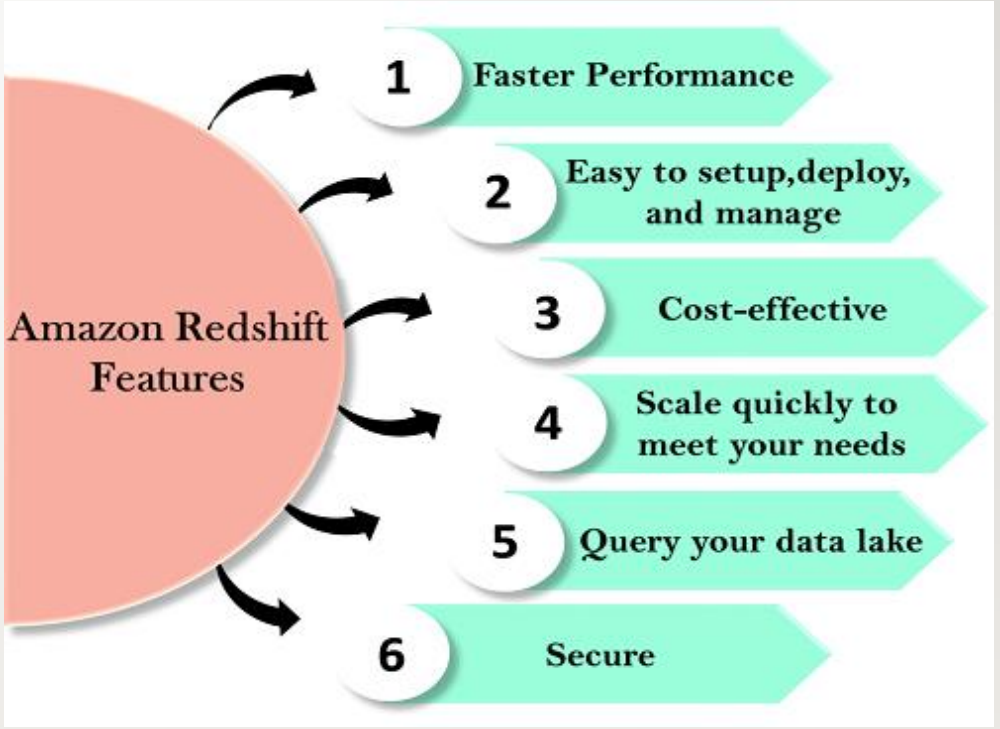
Bottom Up Design Approach

Tools for Data warehouse development:

- organizations had to build lots of infrastructure for data warehousing but today, cloud computing technology has amazingly reduced the efforts as well as the cost of building data warehousing for businesses.

Amazon Redshift features

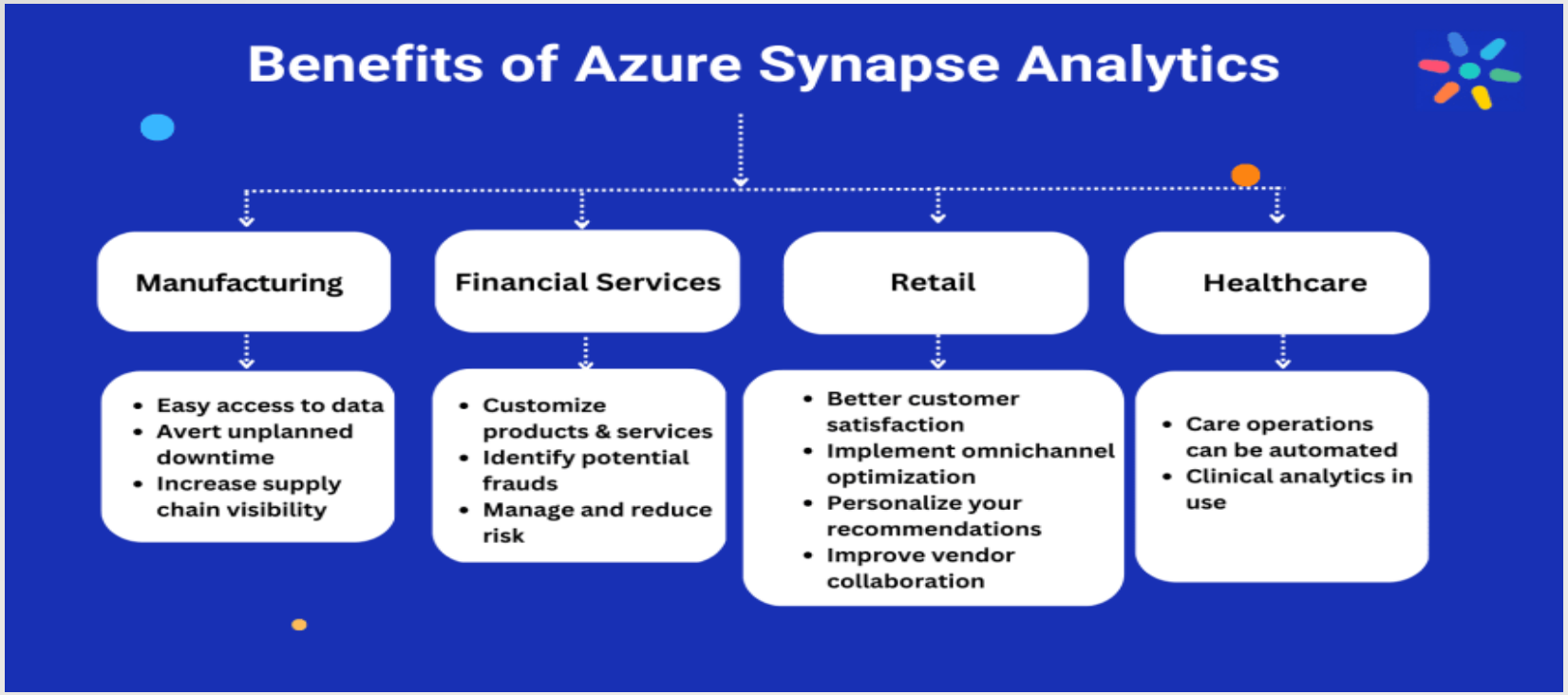
Amazon Redshift is a cloud-based fully managed petabytes-scale data warehouse By the Amazon Company. It starts with just a few hundred gigabytes of data and scales to petabytes or more.
It is a relational database management system (RDBMS) therefore it is compatible with other RDBMS applications.
Amazon Redshift offers quick querying capabilities over structured data by the use of SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections.



Azure Synapse Analytics

- Azure Synapse offers cloud data warehousing, dashboarding, and machine learning analytics in a single workspace.
- It ingests all types of data, including relational and non-relational data, and it lets you explore this data with SQL.
- Azure Synapse uses **massively parallel processing** or MPP database technology, which allows it to manage analytical workloads and also aggregate and process large volumes of data in an efficient manner.
- It gives you the ability to query massive data stores using either an on-demand serverless deployment (which scales automatically as needed to handle any processing or load) or provisioned resources.
- It is compatible with a wide range of scripting languages like Scala, Python, .Net, Java, R, SQL, T-SQL, and Spark SQL.
- It facilitates easy integration with Microsoft and Azure solutions like Azure Data Lake, Azure Blob Storage, and more.
- It includes the latest security and privacy technologies such as real-time data masking, dynamic data masking, always-on encryption, Azure Active Directory authentication, and more.

Azure Synapse Analytics



Tools for Data warehouse development:

- **Amazon Redshift**- Developers can store and examine massive amounts of data with ease with Amazon Redshift, a cloud-based data warehousing tool from AWS. It offers a versatile and scalable framework for managing both organized and semi-structured information.
- **Use Case**
- A retail organization that wants to save and analyze sales data from several channels, including online stores, physical stores, and social media platforms. Insights on client behavior, price adjustments, and streamlined supply chain management are all possible benefits that Redshift can provide.

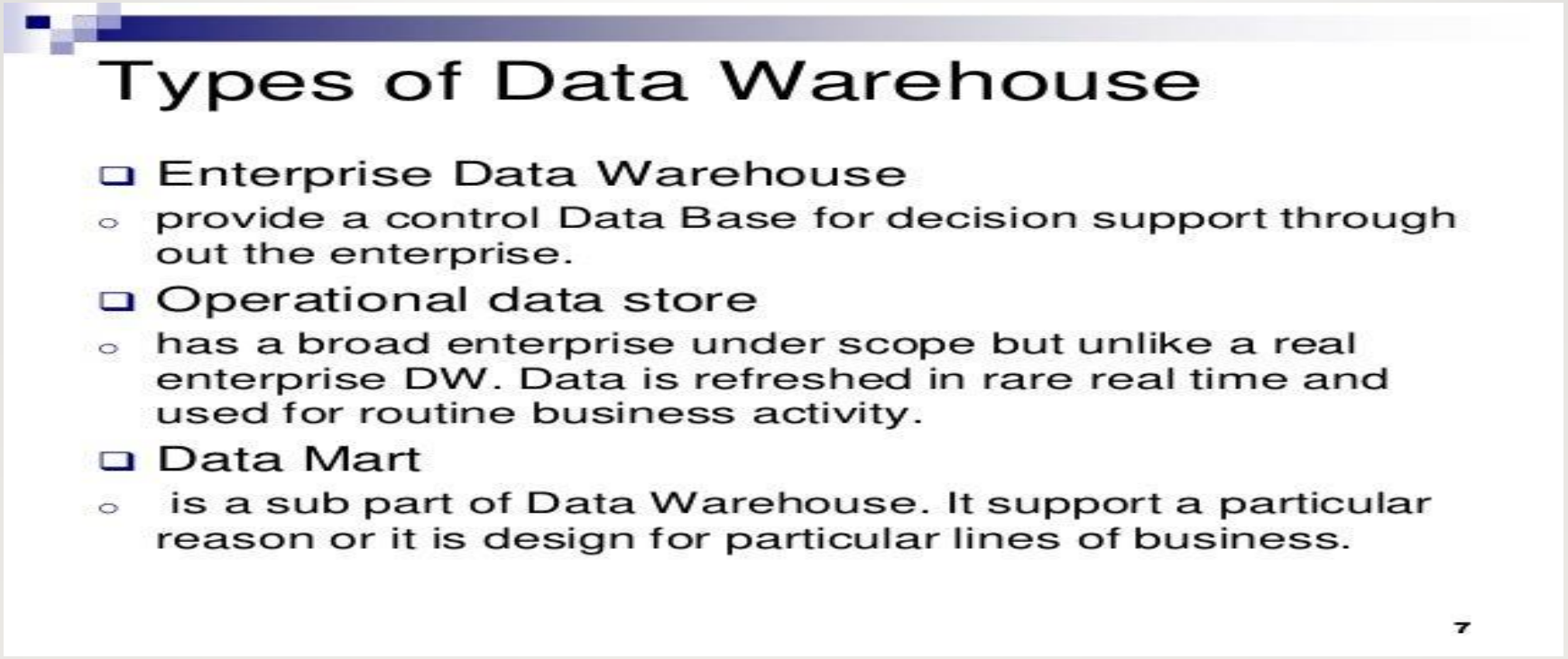
- **Google BigQuery**
- As a cloud-based data warehouse and analytics platform, Google BigQuery delivers on both performance and scalability. Fast and easy analysis of massive datasets is made possible by this tool for developers in the form of SQL-like queries. If your company collects a lot of data from the Internet of Things, social media, or mobile apps, Google BigQuery is the perfect tool for you.
- **Use Case**
- Google BigQuery may be used by a healthcare provider to store and analyze patient data from many sources (e.g., EHRs, RIS/PACS, ambulatory ECGs, and wearables). By seeing patterns, foreseeing potential outcomes, and creating individualized treatment plans, this can help the business provide better care for its patients.

- **Microsoft Azure Synapse Analytics**
- Cloud-based data warehousing and analytics platform Microsoft Azure Synapse Analytics offers scalability and security. It enables developers to merge information from disparate locations and services, including local databases, the cloud, and the Internet of Things. If your company needs a safe, scalable place to save and examine massive volumes of data, Microsoft Azure Synapse Analytics is the way to go.
- **Use Case**
- A financial services organization may useMicrosoft Azure Synapse Analytics to store and analyze data from a wide variety of customer transactions made via ATMs, internet banking, and mobile apps. This has the potential to enhance fraud detection, risk management, and the overall customer service provided by the organization.

- **Oracle Autonomous Data Warehouse**
- With Oracle Autonomous Data Warehouse, you get access to a powerful and scalable data warehousing platform in the cloud. It enables programmers to store and examine massive volumes of data, both organized and unstructured, from a wide range of sources. If your company has to store and analyze massive volumes of data in a short amount of time, Oracle Autonomous Data Warehouse is the way to go.
- **Use Case**
- Oracle Autonomous Data Warehouse is useful for a manufacturing firm to store and analyze production data from a wide variety of sources, including industrial equipment, sensors, and IoT devices. As a result, the company's manufacturing processes may be optimized, quality control can be strengthened, and downtime can be cut.

- **Snowflake**
- Data warehouse solutions that are both scalable and versatile may be found in Snowflake, a cloud-based platform. Structured and semi-structured data may be stored and analyzed with ease. Snowflake is a great option for companies who need to effectively store and analyze massive volumes of data.
- **Use case:** Snowflake may be used by an e-commerce firm to collect and analyze data from a wide variety of channels, including online storefronts, social media, and mobile apps. The corporation may use this to boost client retention, streamline pricing, and better manage stock.

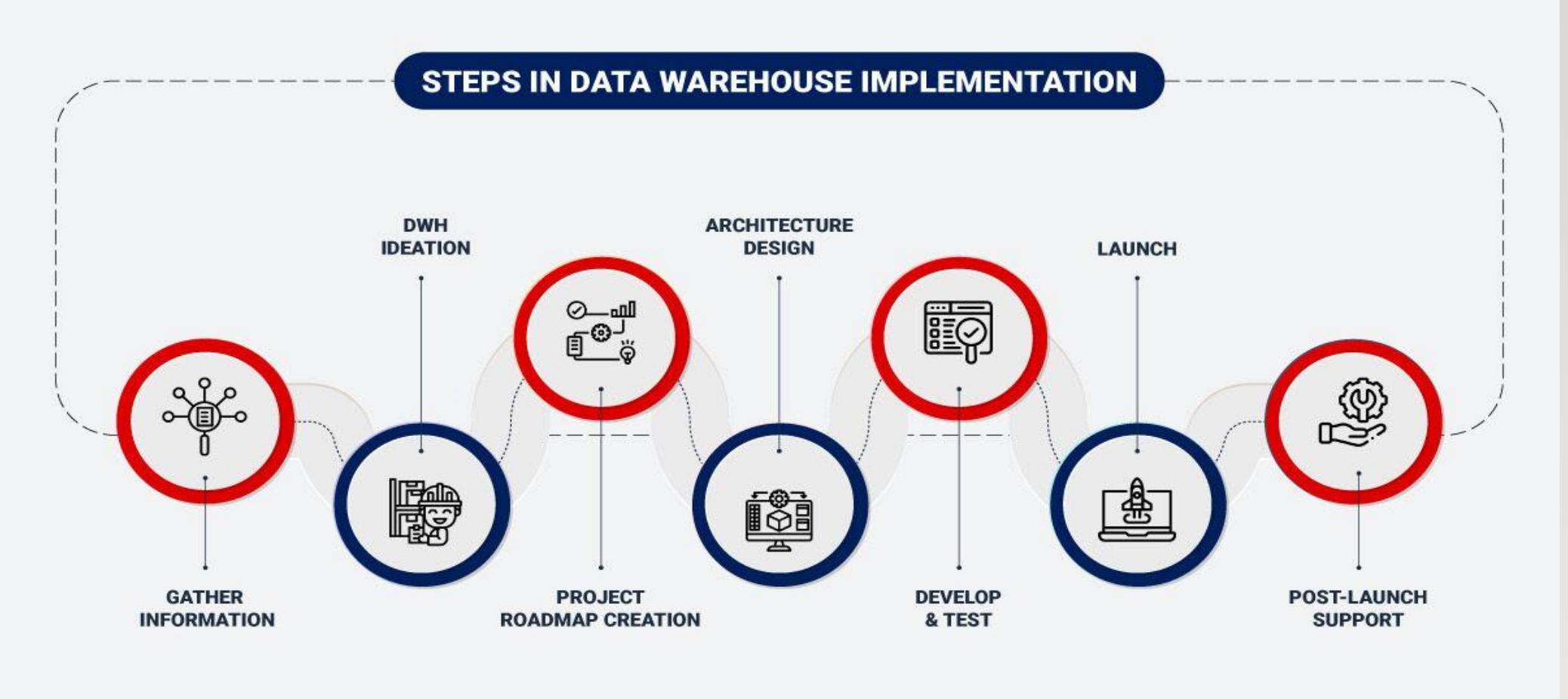
Types of Data Warehouses



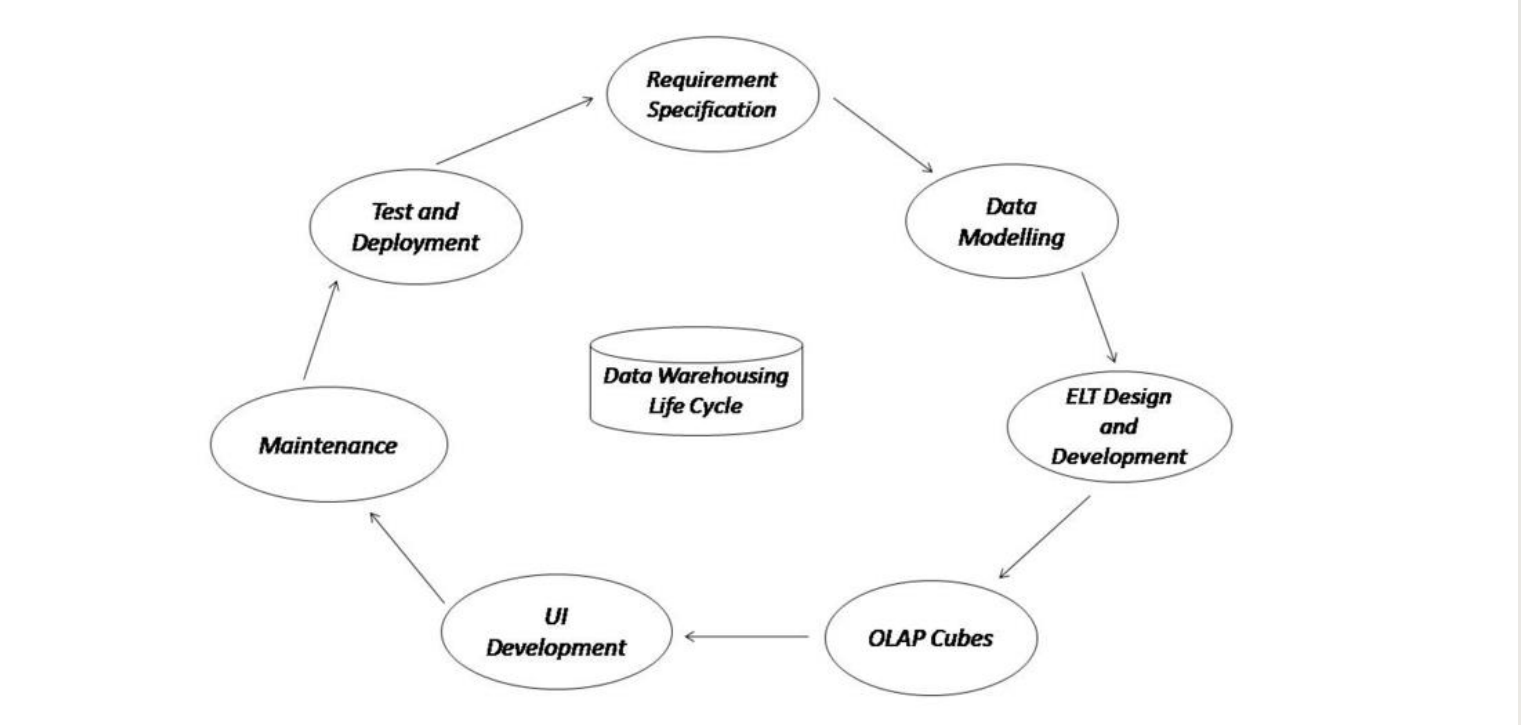
Three main types of Data Warehouses (DWH) are:

Enterprise Data Warehouse (EDW): (Strategic Information) Useful for Decision Making	Operational Data Store(ODS) (Day to day to used data)	Data Mart:
Enterprise Data Warehouse (EDW) is a centralized warehouse. It provides decision support service across the enterprise.	Data warehouse is refreshed in real time.	A data mart is a subset of the data warehouse.
It also provides the ability to classify data according to the subject and give access according to those divisions. Example- Sales, Production, No product selling	it is widely preferred for routine activities like storing records of the Employees. Example- Employee come late so cut salary	It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

Data Warehouse Project



Data Warehouse Development Life cycle Model

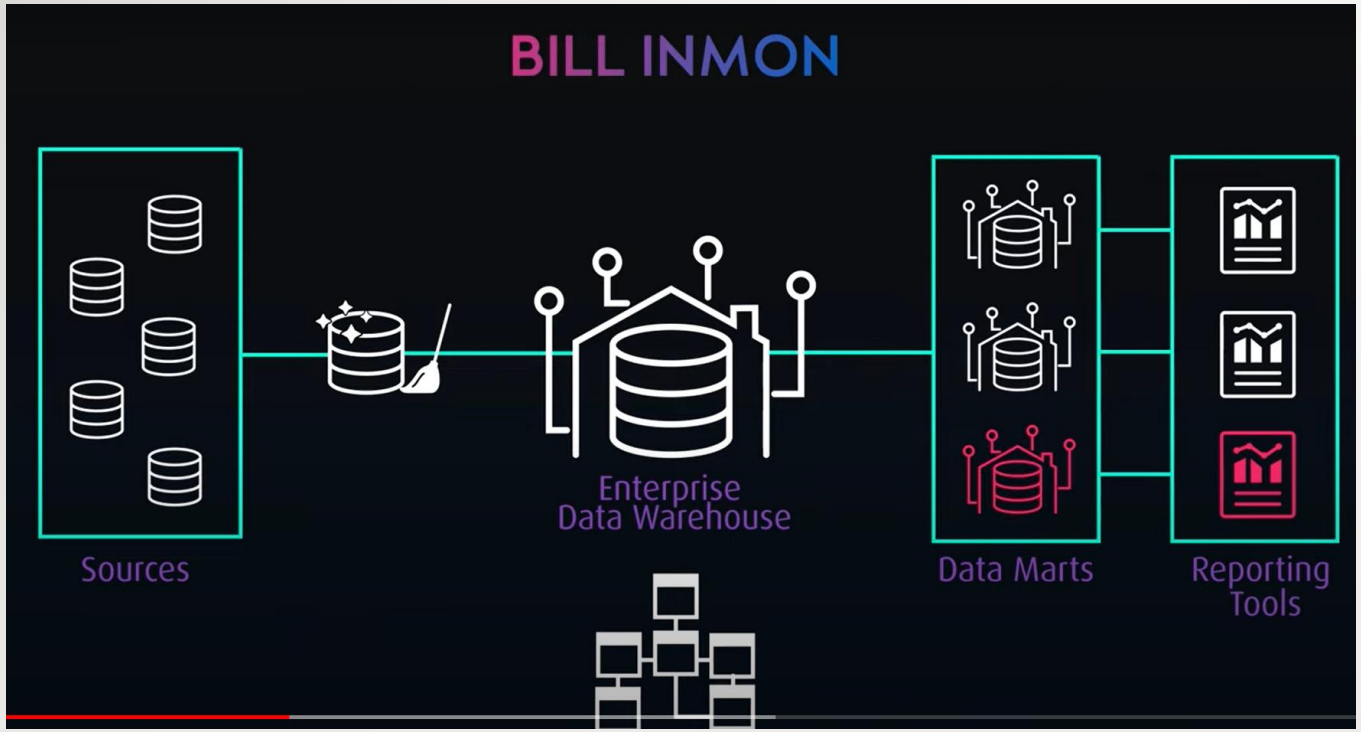


Kimball Lifecycle Diagram

Kimball

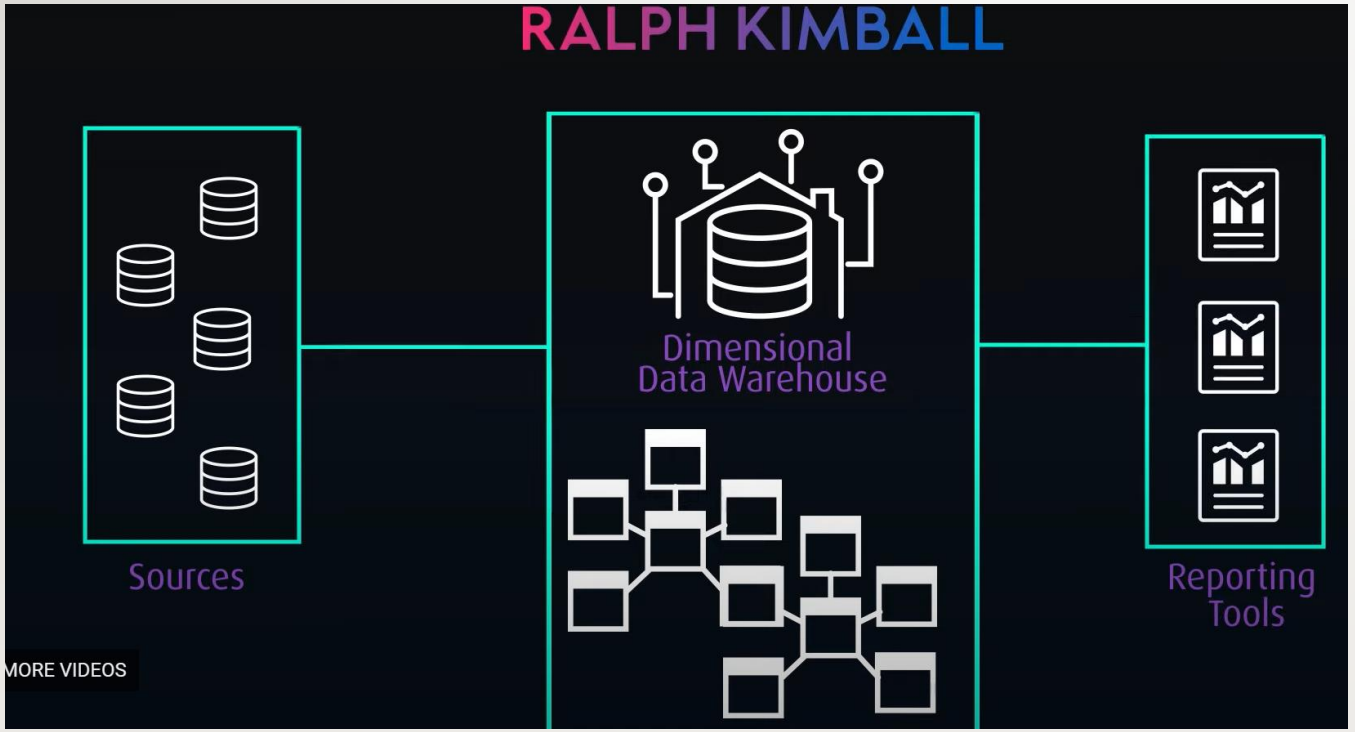
Inmon

Bill Inmon



- Only used fully cleaned data
- Retain Normalized data structure
- Complete organization data
- Dis Advantage
- Slower reporting
- Isolated Data Mart

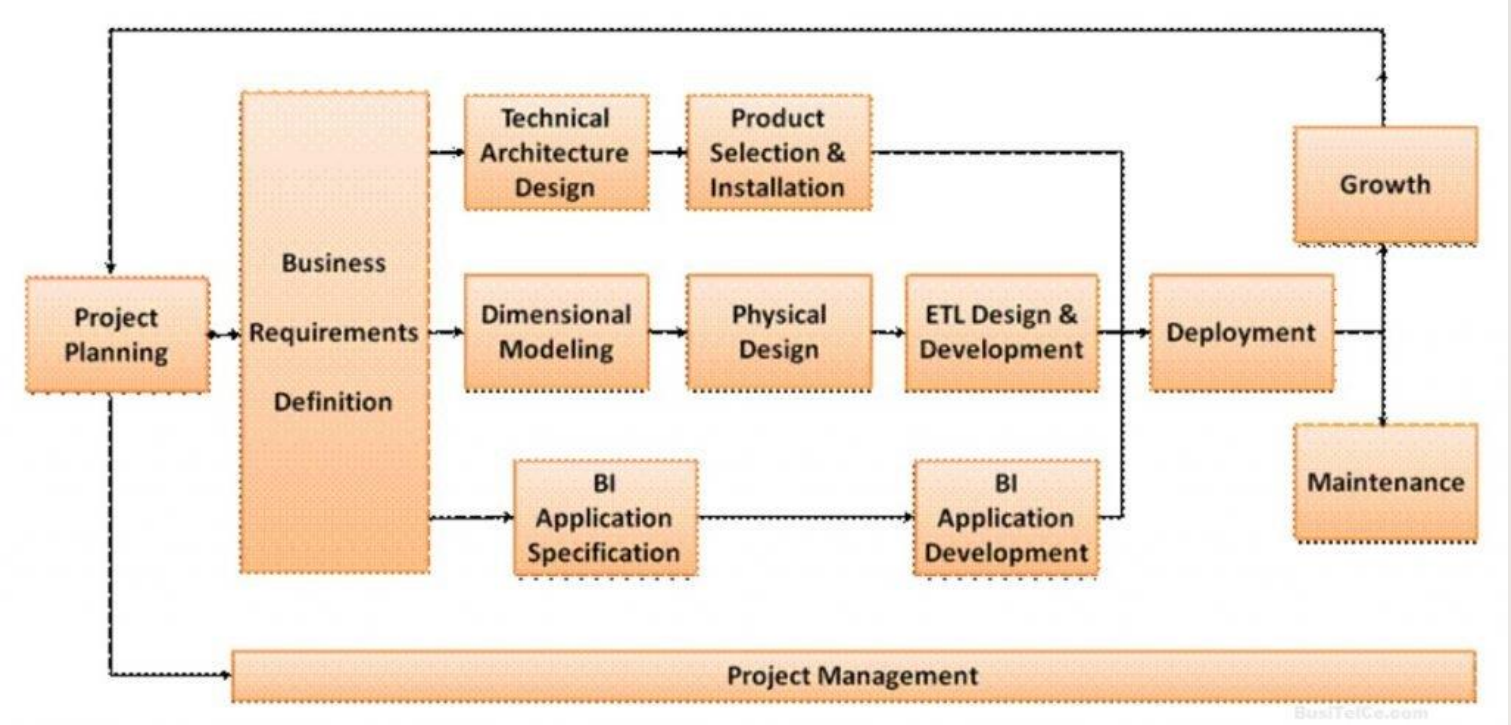
Ralph KimBall



Use Friendly Model
Incremental Build

Complex ETL Process
Multiple data mart
Crated

Kimball Lifecycle Diagram



Summary of Kimball methodology

- Starts with one data mart (ex. sales); later on additional data marts are added (ex. collection, marketing, etc.)
- Data flows from source into data marts, then into the data warehouse
- Kimball approach is faster to implement as it is implemented in stages
- Before we go ahead with details of the methodology, let us take a quick view on some essential definitions of the terms used.

- **Business Requirements Definition**
- Success of the project depends on a solid understanding of the business requirements.
- What follows business requirement definition is three concurrent tracks focusing on
- Technology
- Data
- Business intelligence applications

- **Technical Architecture Design**
- Objective here is to finalize overall architectural framework and vision. We do this based on consider business requirements, current technical environment, and planned strategic technical directions of the organization. Based on the technical architecture
  - **Evaluation and selection of**
  - Products that will deliver needed capabilities
    Hardware platform
    Database management system
    Extract-transformation-load (ETL) tools
    Data access query tools
    Reporting tools must be evaluated
  - Installation of selected products/components/tools
  - Testing of installed products to ensure appropriate end-to-end integration within the data warehouse environment.

Dimensional modeling

- Data track primaraly deals with design of the dimensional model. Dimensional modeling is a vast subject area comprising of many methods, suggestions, and best practices. Here, a detailed data analysis of a single business process is performed to identify the fact table granularity, associated dimensions and attributes, and numeric facts. Primary constructs of a dimensional model are fact tables and dimension tables. Two important design menthods in dimension modeling are – star schema and snowflake schema.

BI Application track

- Star schema contains a central fac table directly connected to dimension table. The snowflake schema is a variation of the star schema used in a data warehouse. The snowflake schema is a more complex schema than the star schema because the tables which describe the dimensions are normalized

Physical design

- We start with a logical data model definition and then proceed to physical design of the model. In physical design, we define the data structures. Some key activities include:
- setting up the database environment
- setting up appropriate security
- preliminary performance tuning strategies, from indexing to partitioning and aggregations.
- if required, OLAP databases are also designed during this process.

BI Application track

- BI applications deliver business value from the DW/BI solution, rather than just delivering the data. The goal is to deliver capabilities that are accepted by the business to support and enhance their decision making.
- **Deployment**
- Deployment should be deferred until all the pieces such as training, documentation, and validated data are not ready for production release. Also, it is critical that deployment be well orchestrated and adequately planned.

Maintenance

- Maintenance begins once the system is deployed into production. Maintenance work ensures ongoing support, education, and communication with business users. Also, technical operational tasks that are necessary to keep the system performing optimally are conducted as needed.
- **Growth process**
- Organization has to reason to be happy if the data warehouse system tends to grow. DW growth is considered as a symbol of success.