UNIT-1

Introduction to Data Warehouse

What Is A Data Warehouse?

- It is a central repository where information is coming from one or more data sources.it helps to make decision easily.

- 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.
- It is a multi dimensional model.

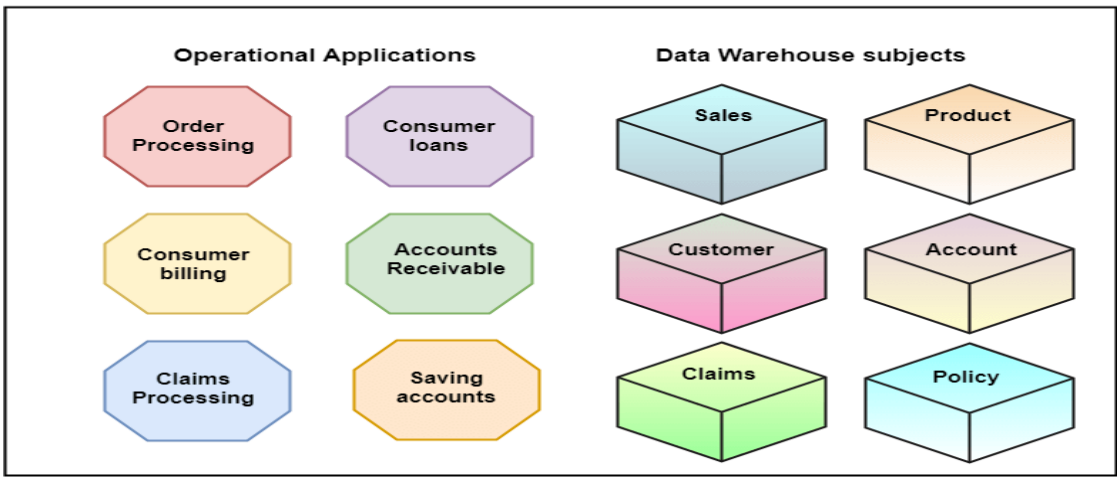
Characteristics of Data Warehouse:

1.Subject-Oriented Data
2.Integrated Data
3.Time-Referenced Data
4.Non-Volatile Data
5.Granularity

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.

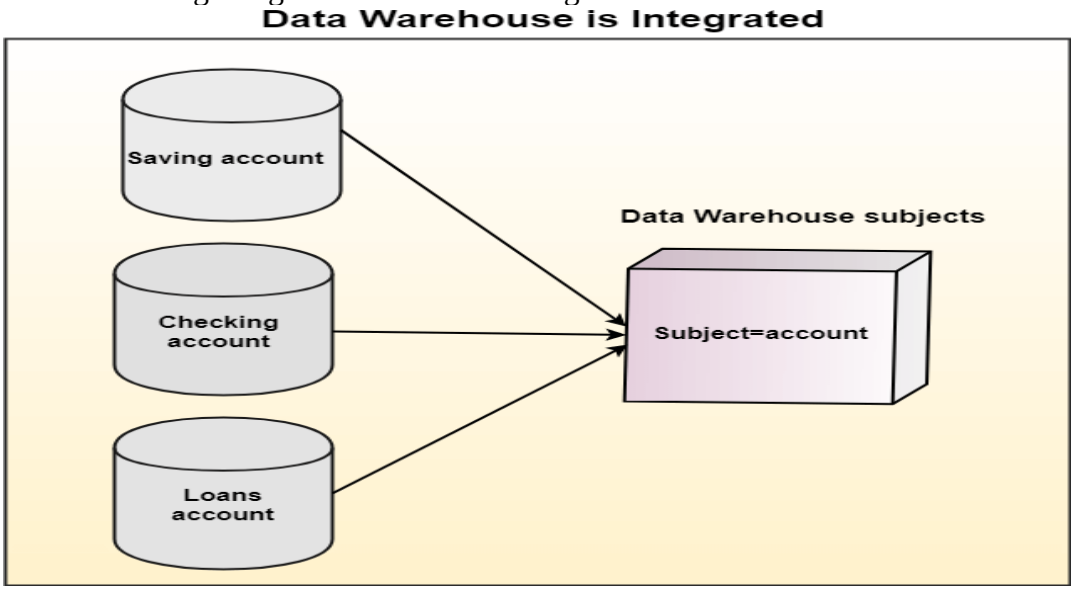
Data Warehouse is Subject-Oriented



Integrated Data:

- Integrated data refers to de-duplicating information and merging it from many sources into one consistent location.

- When short listing your top 20 customers, you must know that "HAL" and "Hindustan Aeronautics Limited" are one and the same.
- Much of the transformation and loading work that goes into the data warehouse is centred on integrating data and standardizing it.

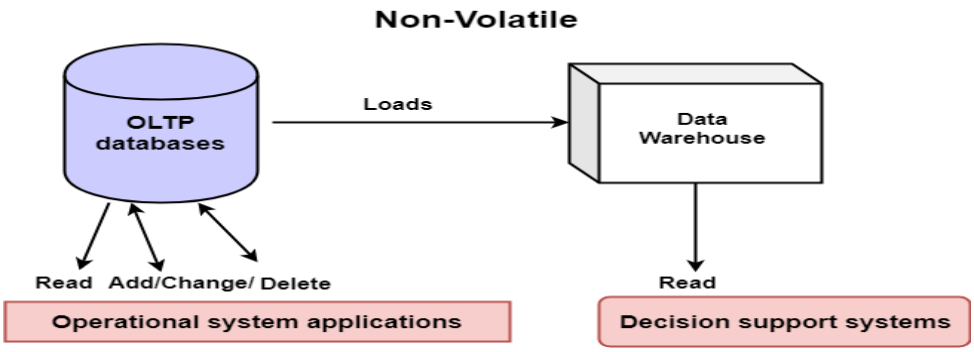


Time-Referenced Data:

- 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 Years 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 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, i.e., update, insert, and delete operations are not performed.
- The non-volatility of data, characteristic of data warehouse, enables users to dig deep into history and arrive at specific business decisions based on facts.



Data Granularity:in data warehouse data to keep data at different levels

Why A Data Warehouse?

The Data Access Crisis

- If there is a single key to survival in the 1990s and beyond, it is being able to analyze, plan, and react to changing business conditions in a much more rapid fashion.
- In order to do this, top managers, analysts, and knowledge workers in our enterprises, need more and better information.
- Every day, organizations large and small, create billions of bytes of data about all aspects of their business; millions of individual facts about their customers, products, operations and people.
- But for the most part, this is locked up in a maze of computer systems and is exceedingly difficult to get at. This phenomenon has been described as "data in jail".

Data Warehousing:

- The idea of data warehousing came to the late 1980's when IBM researchers Barry Devlin and Paul Murphy established the "Business Data Warehouse."
- Data warehousing is a field that has grown from the integration of a number of different technologies and experiences over the past two decades.
- These experiences have allowed the IT industry to identify the key problems that need to be solved.
- It is the process of developing,managing and securing the electronic storage of data by a business or organization in a digital data warehouse .

Operational vs. Informational Systems

- **Operational systems**, as their name implies, are the systems that help the every day operation of the enterprise.
- These are the backbone systems of any enterprise, and include order entry, inventory, manufacturing, payroll and accounting.
- Due to their importance to the organization, operational systems were almost always the first parts of the enterprise to be computerized.

- **Informational systems** deal with analyzing data and making decisions, often major, about how the enterprise will operate now, and in the future.
- Not only do informational systems have a different focus from operational ones, they often have a different scope.
- Where operational data needs are normally focused upon a single area, informational data needs often span a number of different areas and need large amounts of related operational data.

DATA WAREHOUSE ADVANTAGES & DISADVANTAGES:

DATA WAREHOUSE ADVANTAGES:

1. DW make access to a wide variety of data easier for end users.2.provide key i/f for business decision making
3. Improves the quality of decisions made
4. Especially useful for the medium & large term
5. It provides a great power of information processing
6. Facilities decision making in business
7. Companies get an increase in productivity
8. It allows you to plan more effectively
9. Reduce response times &Operating costs
10. improve relationships with suppliers & customers

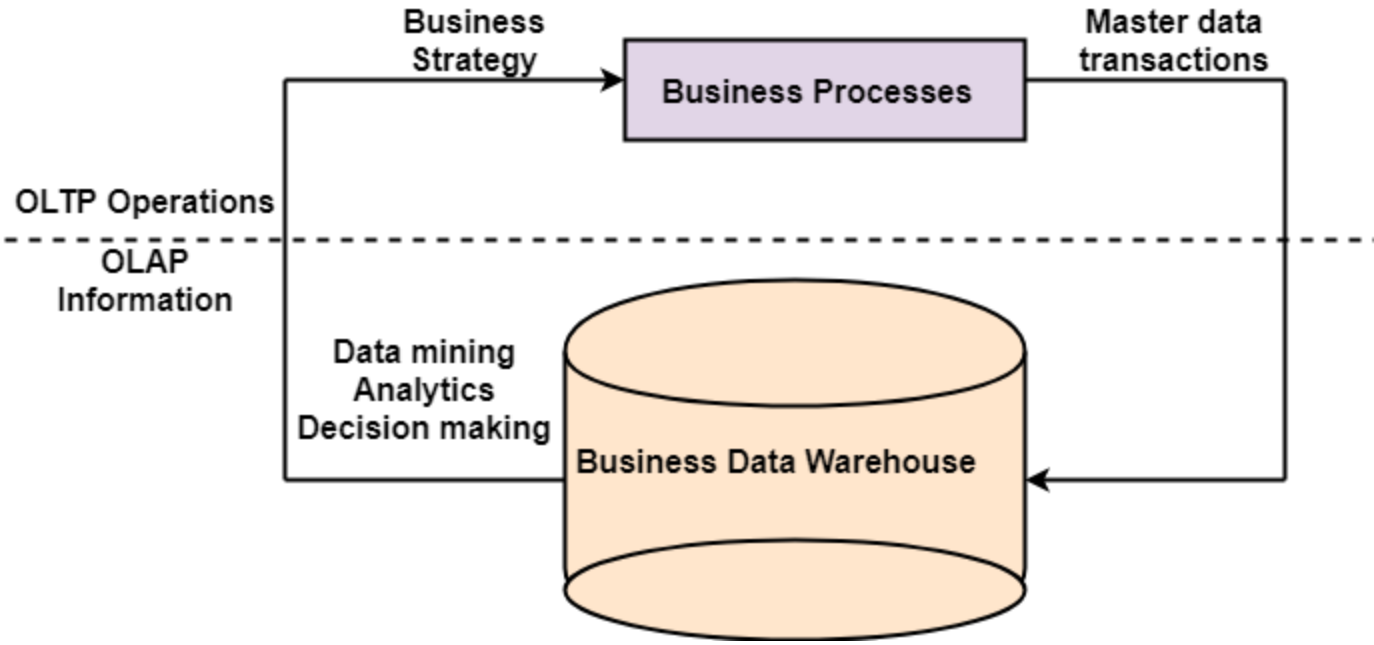
DATA WAREHOUSE DIS ADVANTAGES:

1. DW can suppose high costs, maintenance costs are high
2. DW may become obsolete relatively soon
3. It require continuous cleaning, transformation ,data integration, maintenance
4. Once implemented it can be difficult to add new data sources
5. they have a complex & multidisciplinary design
6. they require a restructuring of the OS
7. They require a review of the data model, objects, transactions, additions to storage
8. In an implementation process difficulties may be encountered in relation to the different objectives that an organization intends

Difference between OLTP and OLAP:

OLTP (On-Line Transaction Processing) is featured by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). The primary significance of OLTP operations is put on very rapid query processing, maintaining record integrity in multi-access environments, and effectiveness consistent by the number of transactions per second. In the OLTP database, there is an accurate and current record, and schema used to save transactional database is the entity model (usually 3NF).

OLAP (On-line Analytical Processing) is represented by a relatively low volume of transactions. Queries are very difficult and involve aggregations. For OLAP operations, response time is an effectiveness measure. OLAP applications are generally used by Data Mining techniques. In OLAP database there is aggregated, historical information, stored in multi-dimensional schemas (generally star schema).



Data Warehouse (OLAP)	**Operational Database(OLTP)**
It involves historical processing of information	It involves day-to-day processing.
OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
It is used to analyze the business.	It is used to run the business.
It focuses on Information out.	It focuses on Data in.
It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
It is subject oriented.	It is application oriented.
It contains historical data.	It contains current data.
It provides summarized and consolidated data.	It provides primitive and highly detailed data.
It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
The number of users is in hundreds.	The number of users is in thousands.
The number of records accessed is in millions.	The number of records accessed is in tens.
These are highly flexible.	It provides high performance.

What are the components in Kimball's DW/BI Architecture?

Kimball's DW/BI Architecture:

There are four separate and distinct components to consider in the DW/BI environment: operational source systems, ETL system, data presentation area, and business intelligence application.

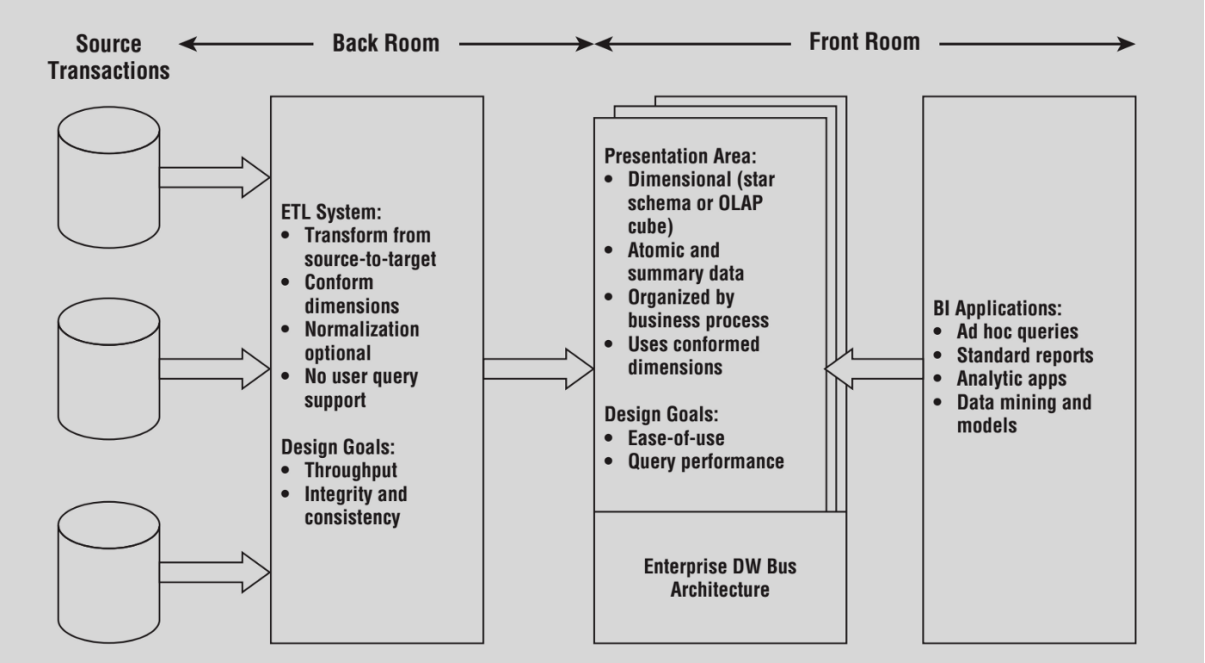


Fig:Core elements of the Kimball DW/BI architecture.

Operational Source Systems :

- These are the operational systems of record that capture the business's transactions.
    - Source systems as outside the data warehouse because presumably you have little or no control over the content and format of the data in these operational systems.
    - The main priorities of the source systems are processing performance and availability.
    - Operational queries against source systems are narrow, one-record-at-a-time queries that are part of the normal transaction flow and severely restricted in their demands on the operational system.
    - In many cases, the source systems are special purpose applications without any commitment to sharing common data such as product, customer, geography, or calendar with other operational systems in the organization. Of course, a broadly adopted cross-application enterprise resource planning (ERP) system or operational master data management system could help address these shortcomings.

Extract, Transformation, and Load System :

- The extract, transformation, and load (ETL) system of the DW/BI environment consists of a work area, instantiated data structures, and a set of processes.

- . Extraction is the first step in the process of getting data into the data warehouse environment.
- Extracting means reading and understanding the source data and copying the data needed into the ETL system for further manipulation.
- After the data is extracted to the ETL system, there are numerous potential transformations, such as cleansing the data (correcting misspellings, resolving domain conflicts, dealing with missing elements, or parsing into standard formats), combining data from multiple sources, and de-duplicating data.
- The ETL system adds value to the data with these cleansing and conforming tasks by changing the data and enhancing it.
- The final step of the ETL process is the physical structuring and loading of data into the presentation area's target dimensional models.
- When the dimension and fact tables in a dimensional model have been updated, indexed, supplied with appropriate aggregates, and further quality assured, the business community is notified that the new data has been published.
- The ETL system is typically dominated by the simple activities of sorting and sequential processing
- The creation of both normalized structures for the ETL and dimensional structures for presentation means that the data is potentially extracted, transformed, and loaded twice—once into the normalized database and then again when you load the dimensional model.
- Although enterprise-wide data consistency is a fundamental goal of the DW/BI environment, there may be effective and less costly approaches than physically creating normalized tables in the ETL system, if these structures don't already exist.

Presentation Area to Support Business Intelligence:

- The DW/BI presentation area is where data is organized, stored, and made available for direct querying by users, report writers, and other analytical BI applications.
- Dimensional modeling is the most viable technique for delivering data to DW/BI users.
- The presentation area is that it must contain detailed, atomic data.
- Atomic data is required to withstand assaults from unpredictable adhoc user queries.
- The presentation data area should be structured around business process measurement events.
- All the dimensional structures must be built using common, conformed dimensions.
- Data in the queryable presentation area of the DW/BI system must be dimensional, atomic (complemented by performance-enhancing aggregates), business process-centric, and adhere to the enterprise data warehouse bus architecture. The data must not be structured according to individual departments' interpretation of the data.

Business Intelligence Applications :

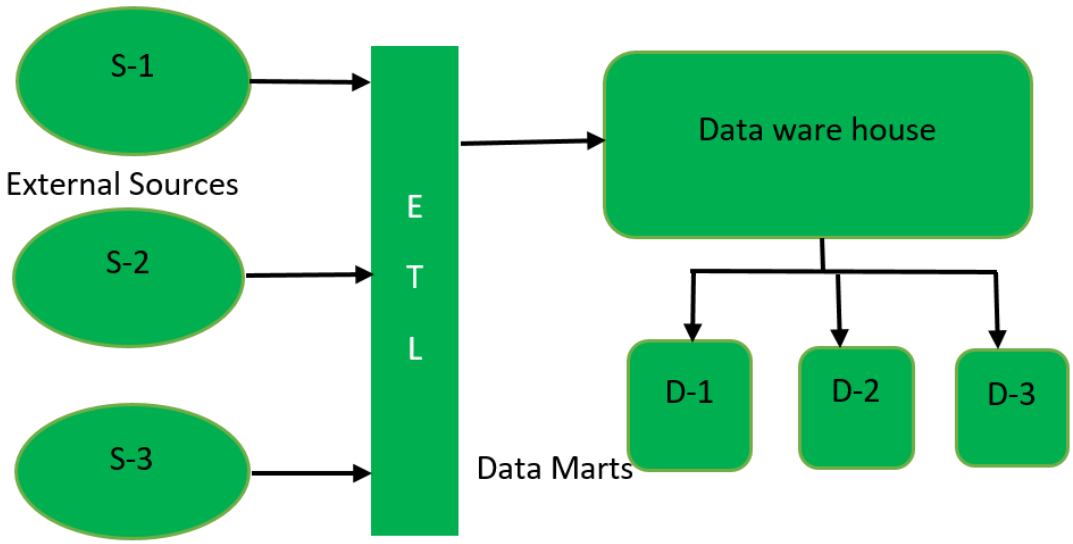
- The final major component of the Kimball DW/BI architecture is the business intelligence (BI) application.
- The term BI application loosely refers to the range of capabilities provided to business users to leverage the presentation area for analytic decision making.
- A BI application can be as simple as an ad hoc query tool or as complex as a sophisticated data mining or modeling application.

Alternative DW/BI Architectures:

- **Data mart** is such a storage component which is concerned on a specific department of an organization. It is a subset of the data stored in the datawarehouse. Data mart is focused only on particular function of an organization and it is maintained by single authority only, e.g.m finance, Marketing. Data Marts are small in size and are flexible.
- **Types of Data Mart:**
  There are three types of data marts:

Dependent Data Mart :

- Dependent Data Mart is created by extracting the data from central repository, Datawarehouse. First data warehouse is created by extracting data (through ETL tool) from external sources and then data mart is created from data warehouse. Dependent data mart is created in top-down approach of datawarehouse architecture. This model of data mart is used by big organizations.



Independent Data Mart Architecture:

- Independent Data Mart is created directly from external sources instead of data warehouse. First data mart is created by extracting data from external sources and then datawarehouse is created from the data present in data mart. Independent data mart is designed in bottom-up approach of datawarehouse architecture. This model of data mart is used by small organizations and is cost effective comparatively.

- Independent data marts are not difficult to design and develop. They are beneficial to achieve short-term goals but may become cumbersome to manage—each with its own ETL tool and logic—as business needs expand and become more complex.
    - An advantage to this model is that individual business units can run the data mart that suits them best.

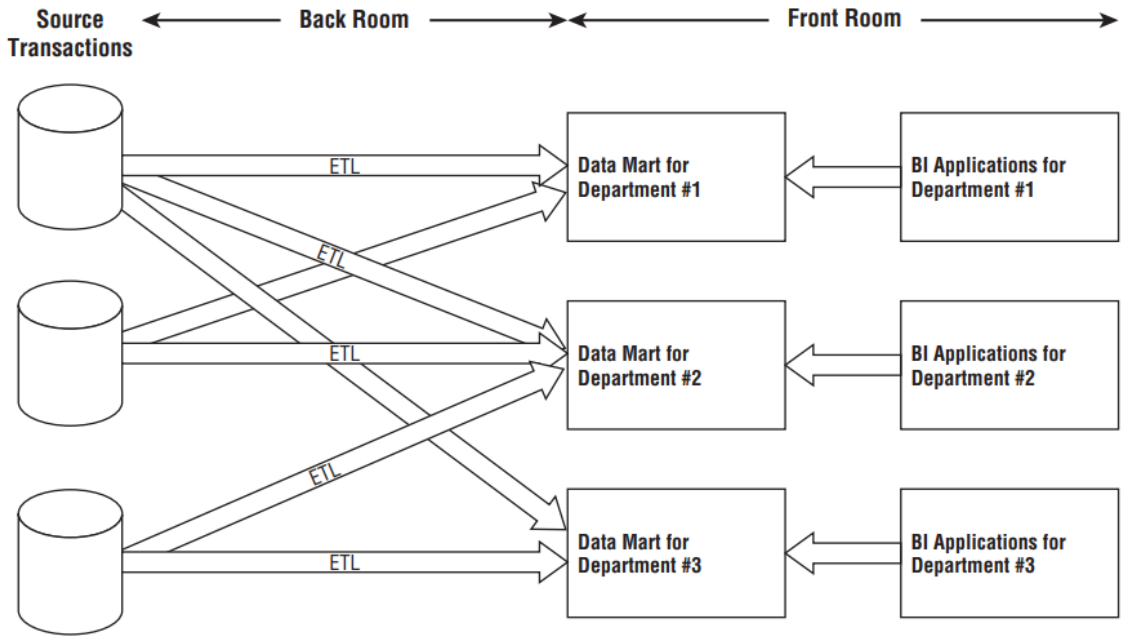
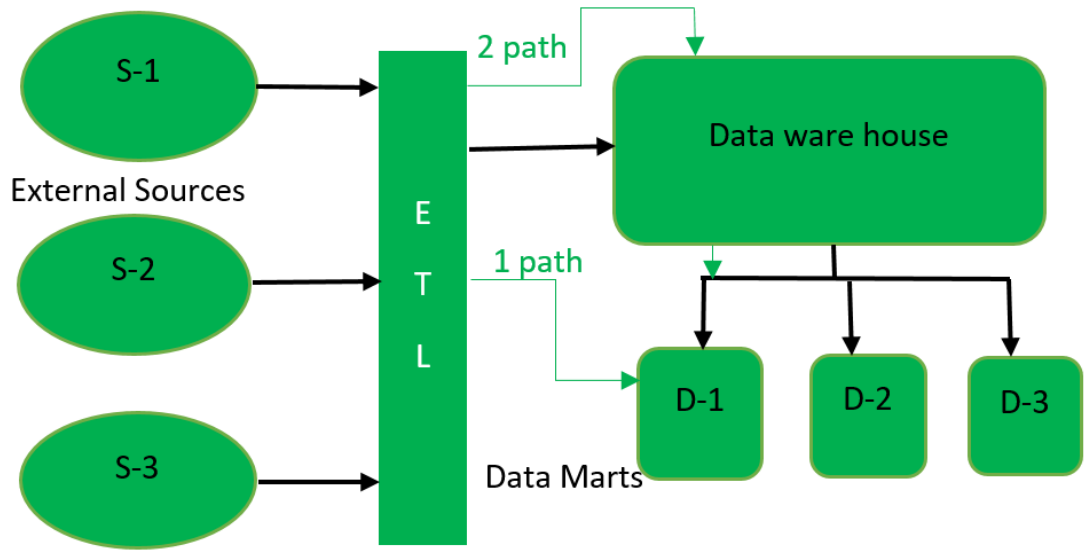


Figure 1-8: Simplified illustration of the independent data mart "architecture."

Hybrid Data Mart –



This type of Data Mart is created by extracting data from operational source or from data warehouse. 1Path reflects accessing data directly from external sources and 2Path reflects dependent data model of data mart.