UNIT 1 FUNDAMENTALS OF DATA WAREHOUSE

Structure

1.0 INTRODUCTION

A database often contains information or data collection that is generally stored electronically in a computer system. It is easy to access, manage, modify, update, monitor, and organize the data. Data is stored in the tables of the database.

The process of consolidating data and analyzing it to obtain some insights has been around for centuries, but we just recently began referring to this as data warehousing. Any operational or transactional system is only designed with its own functionality and hence, it could handle limited amounts of data for a limited amount of time. The operational systems are not designed or architected for long term data retention as the historical data is little to no importance to them. However, to gain a point-in-time visibility and understand the high-level operational aspects of any business, the historical data plays a vital role. With the emergence of matured Relational Database Management Systems (RDBMS) in 1960s, engineers across various enterprises started

architecting ways to copy the data from the transactional systems over to different databases via manual or automated mechanism and use it for reporting and analysis. As the data in the transactional systems would get purged periodically, it would not be the case in these analytical repositories as their purpose was to store as much data as possible; hence the word "data warehouse" came into existence because these repositories would become a warehouse for the data.

Data Warehousing (DW) as a practice became very prominent during late 80s when the enterprises started building decision support systems that were mainly responsible to support reporting. As there was a rapid advancement in the performance of these relational database during late 1990s and early 2000s, Data Warehousing became a core part of the Information Technology group across large enterprises. In fact, some of the vendors like *Netezza, Teradata* started offering customized hardware to manage data warehouse architectures within state-of-the-art machines. Data Warehousing had evolved to be on top of the list of priorities since mid 2000s. Data supply chain ecosystem has grown exponentially in the current world and so is the way enterprises architect their data warehouses.

A well architected data warehouse serves as an extended vision for the enterprise where multiple departments can gain actionable insights to manage key business decisions that could drive operational excellence or revenue generating opportunities for the enterprise.

This unit covers the basic features of data warehousing, its evolution, characteristics, online transaction processing (OLTP), online analytical processing, popular platforms and applications of data warehouses.

1.1 OBJECTIVES

After going through this unit, you shall be able to:

- understand the evolution of data warehouse;
- describe various characteristics of data warehouse;
- describe the benefits and applications of a data warehouse;
- discuss the significance of metadata in data ware house;
- list and discuss the types of data warehouses, and
- identify the popular data warehouse platforms;

1.2 EVOLUTION OF DATA WAREHOUSE

The relational database revolution in the early 1980s ushered in an era of improved access to the valuable information contained deep within data. It was soon discovered that databases modeled to be efficient at transactional processing were not always optimized for complex reporting or analytical needs.

In fact, the need for systems offering decision support functionality predates the first relational model and SQL. But the practice known today as Data Warehousing really saw its genesis in the late 1980s. An IBM Systems Journal article published in 1988, *An architecture for a business information system* coined the term "business data warehouse," although a future progenitor of the practice, Bill Inmon, used a similar term in the 1970s. Considered by many to be the Father of Data Warehousing, Bill Inmon, an American Computer Scientist is first began to discuss the principles around the Data Warehouse and even coined the term. Throughout the latter 1970s into the 1980s, Inmon worked extensively as a data professional, honing his expertise in all manners of relational Data Modeling. Inmon's work as a Data Warehousing pioneer took off in the early 1990s when he ventured out on his own, forming his first company, Prism Solutions. One of Prism's main products was the Prism Warehouse Manager, one of the first industry tools for creating and managing a Data Warehouse.

In 1992, Inmon published *Building the Data Warehouse*, one of the seminal volumes of the industry. Later in the 1990s, Inmon developed the concept of the Corporate Information Factory, an enterprise level view of an organization's data of which Data Warehousing plays one part. Inmon's approach to Data Warehouse design focuses on a centralized data repository modeled to the third normal form. Inmon's approach is often characterized as a top-down approach. Inmon feels using strong relational modeling leads to enterprise-wide consistency facilitating easier development of individual data marts to better serve the needs of the departments using the actual data. This approach differs in some respects to the "other" father of Data Warehousing, Ralph Kimball.

While Inmon's Building the Data Warehouse provided a robust theoretical background for the concepts surrounding Data Warehousing, it was Ralph Kimball's *The Data Warehouse Toolkit*, first published in 1996, that included a host of industry-honed, practical examples for OLAP-style modeling. Kimball, on the other hand, favors the development of individual data marts at the departmental level that get integrated together using the Information Bus architecture. This bottom up approach fits-in nicely with Kimball's preference for star-schema modeling. Both approaches remain core to Data Warehousing architecture as it stands today. Smaller firms might find Kimball's data mart approach to be easier to implement with a constrained budget. Dimensional modeling in many cases is easier for the end user to understand.

According to Bill Inmon, "*A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process*".

According to Ralph Kimball, "*Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model*".

1.3 DATA WAREHOUSING AND ITS NEED

Data Warehouse is used to collect and manage data from various sources, in order to provide meaningful business insights. A data warehouse is usually used for linking and analyzing heterogeneous sources of business data. The data warehouse is the center of the data collection and reporting framework developed for the BI system. Data warehouse systems are real-time repositories of information, which are likely to be tied to specific applications. Data warehouses gather data from multiple sources (including databases), with an emphasis on storing, filtering, retrieving and in particular, analyzing huge quantities of organized data. The data warehouse operates in information-rich environment that provides an overview of the company, makes the current and historical data of the company available for decisions, enables decision support transactions without obstructing operating systems, makes information consistent for the organization, and presents a flexible and interactive information source.

1.3.1 Need for Data Warehouse

Data warehouses are used extensively in the largest and most complex businesses around the world. In demanding situations, good decision making becomes critical. Significant and relevant data is required to make decisions. This is possible only with the help of a well-designed data warehouse. Following are some of the reasons for the need of Data Warehouses:

*Enhancing the turnaround time for analysis and reporting***:** Data warehouse allows business users to access critical data from a single source enabling them to take quick decisions. They need not waste time retrieving data from multiple sources. The business executives can query the data themselves with minimal or no support from IT which in turn saves money and time.

*Improved Business Intelligence***:** Data warehouse helps in achieving the vision for the managers and business executives. Outcomes that affect the strategy and procedures of an organization will be based on reliable facts and supported with evidence and organizational data.

*Benefit of historical data***:** Transactional data stores data on a day to day basis or for a very short period of duration without the inclusion of historical data. In comparison, a data warehouse stores large amounts of historical data which enables the business to include time-period analysis, trend analysis, and trend forecasts.

*Standardization of data***:** The data from heterogeneous sources are available in a single format in a data warehouse. This simplifies the readability and accessibility of data. For example, gender is denoted as Male/ Female in Source 1 and m/f in Source 2 but in a data warehouse the gender is stored in a format which is common across all the businesses i.e. M/F.

Immense ROI (Return On Investment): Return On Investment refers to the additional revenues or reduces expenses a business will be able to realize from any project.

Now, let us study the benefits.

1.3.2 Benefits of Data Warehouse

Several enterprises adopt data warehousing as it offers many benefits, such as streamlining the business and increasing profits. Following are some of the benefits of having a data warehouse:

Scalability - Businesses today cannot survive for long if they cannot easily expand and scale to match the increase in the volume of daily transactions. DW is easy to scale, making it easier for the business to stride ahead with minimum hassle.

Access to Historical Insights - Though real-time data is important, historical insights cannot be ignored when tracing patterns. Data warehousing allows businesses to access past data with just a few clicks. Data that are months and years old can be stored in the warehouse.

Works On-Premises and on Cloud - Data warehouses can be built on-premises or on cloud platforms. Enterprises can choose either option, depending on their existing business system and the long-term plan. Some businesses rely on both.

Better Efficiency - Data warehousing increases the efficiency of the business by collecting data from multiple sources and processing it to provide reliable and actionable insights. The top management uses these insights to make better and faster decisions, resulting in more productivity and improved performance.

Improved Data Security - Data security is crucial in every enterprise. By collecting data in a centralized warehouse, it becomes easier to set up a multi-level security system to prevent the data from being misused. Provide restricted access to data based on the roles and responsibilities of the employees.

Increase Revenue and Returns - When the management and employees have access to valuable data analytics, their decisions and actions will strengthen the business. This increases the revenue in the long run.

Faster and Accurate Data Analytics - When data is available in the central data warehouse, it takes less time to perform data analysis and generate reports. Since the data is already cleaned and formatted, the results will be more accurate.

Let us study the various approaches in detail in the following section.

1.4 DATA WAREHOUSE DESIGN APPROACHES

Data Warehouse design approaches are very important aspect of building data warehouse. Selection of right data warehouse design could save lot of time and project cost.

There are two different Data Warehouse Design Approaches normally followed when designing a Data Warehouse solution and based on the requirements of your project you can choose which one suits your particular scenario. These methodologies are a result of research from Bill Inmon (Top-Down Approach) and Ralph Kimball(Bottom up Approach).

1.4.1 Top-down Approach

Bill Inmon's design methodology is based on a top-down approach which is illustrated in the Figure 1. In the top-down approach, the data warehouse is designed first and then data mart is built on top of data warehouse.

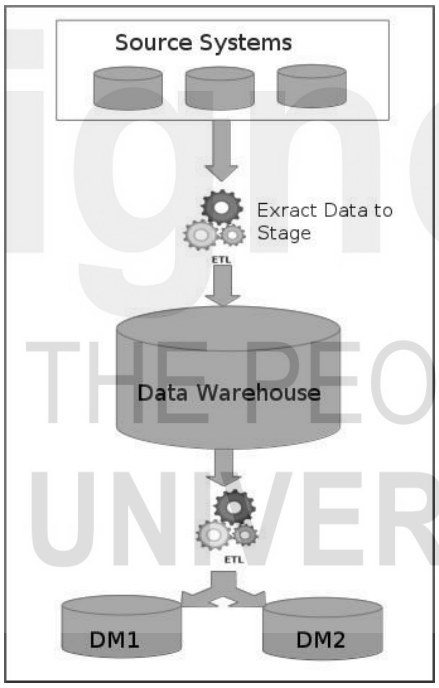

Figure 1: Top-Down DW Design Approach

Below are the steps that are involved in top-down approach:

- Data is extracted from the various source systems. The extracts are loaded and validated in the stage area. Validation is required to make sure the extracted data is accurate and correct. You can use the ETL tools or approach to extract and push to the data warehouse.
- Data is extracted from the data warehouse in regular basis in stage area. At this step, you will apply various aggregation, summarization techniques on extracted data and loaded back to the data warehouse.
- Once the aggregation and summarization is completed, various data marts extract that data and apply the some more transformation to make the data structure as defined by the data marts.

1.4.2 Bottom-up Approach

Ralph Kimball's data warehouse design approach is called dimensional modelling or the Kimball methodology which is illustrated in Figure 2. This methodology follows the bottom-up approach.

As per this method, data marts are first created to provide the reporting and analytics capability for specific business process, later with these data marts enterprise data warehouse is created.

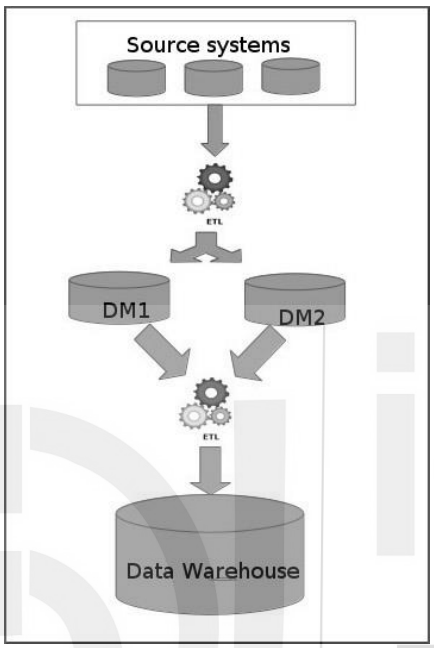

Figure 2: Bottom-Up DW Design Approach

Basically, Kimball model reverses the Inmon model i.e. Data marts are directly loaded with the data from the source systems and then ETL process is used to load in to Data Warehouse. The above image depicts how the top-down approach works.

Below are the steps that are involved in bottom-up approach:

- The data flow in the bottom up approach starts from extraction of data from various source systems into the stage area where it is processed and loaded into the data marts that are handling specific business process.

- After data marts are refreshed the current data is once again extracted in stage area and transformations are applied to create data into the data mart structure. The data is the extracted from Data Mart to the staging area is aggregated, summarized and so on loaded into EDW and then made available for the end user for analysis and enables critical business decisions.

Having discussed the data warehouse design strategies, let us study the characteristics of the DW in the next section.

1.5 CHARACTERISTICS OF A DATA WAREHOUSE

Data warehouses are systems that are concerned with studying, analyzing and presenting enterprise data in a way that enables senior management to make decisions. The data warehouses have four essential characteristics that distinguish them from any other data and these characteristics are as follows:

- **Subject-oriented**

A DW is always a subject-oriented one, as it provides information about a specific theme instead of current organizational operations. On specific themes, it can be done. That means that it is proposed to handle the data warehousing process with a specific theme (subject) that is more defined. Figure 3 shows Sales, Products, Customers and Account are the different themes.

A data warehouse never emphasizes only existing activities. Instead, it focuses on data demonstration and analysis to make different decisions. It also provides an easy and accurate demonstration of specific themes by eliminating information that is not needed to make decisions.

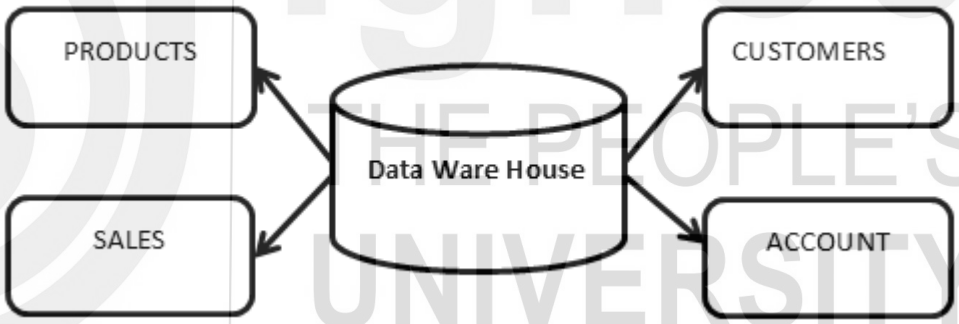

Figure 3: Subject-oriented Characteristic Feature of a DW

- **Integrated**

Integration involves setting up a common system to measure all similar data from multiple systems. Data was to be shared within several database repositories and must be stored in a secured manner to access by the data warehouse. A data warehouse integrates data from various sources and combines it in a relational database. It must be consistent, readable, and coded.. The data warehouse integrates several subject areas as shown in the figure 4.

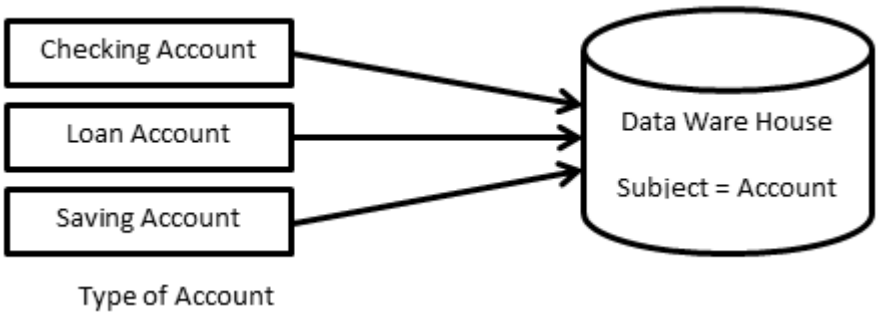

Figure 4: Integrated Characteristic Feature of a DW

- **Time-Variant**

Information may be held in various intervals such as weekly, monthly, and yearly as shown in Figure 5. It provides a series of limited-time, variable rate, online transactions. The data warehouse covers a broader range of data than the operational systems. When the data stored in the data store has a certain amount of time, it can be predictable and provide history. It has aspects of time embedded within it. One other facet of the data warehouse is that the data cannot be changed, modified or updated once it is stored.

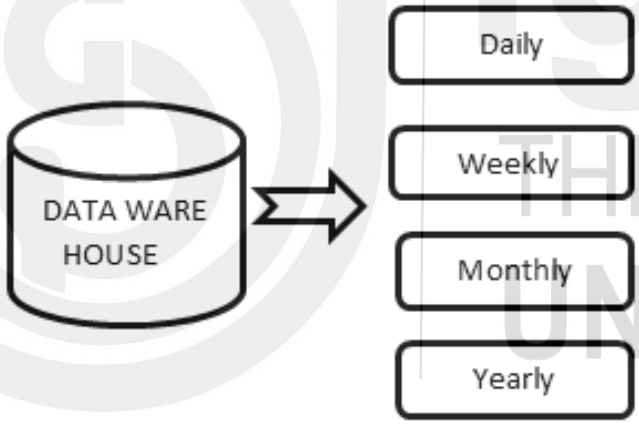

Figure 5: Time- Variant Characteristic Feature of a DW

- **Non-Volatile**

The data residing in the data warehouse is permanent, as the name non - volatile suggests. It also ensures that when new data is added, data is not erased or removed. It requires the mammoth amount of data and analyses the data within the technologies of warehouse. Figure 6 shows the non-volatile data warehouse vs operational database. A data warehouse is kept separate from the operational database and thus the data warehouse does not represent regular changes in the operational database. Data warehouse integration manages different warehouses relevant to the topic.

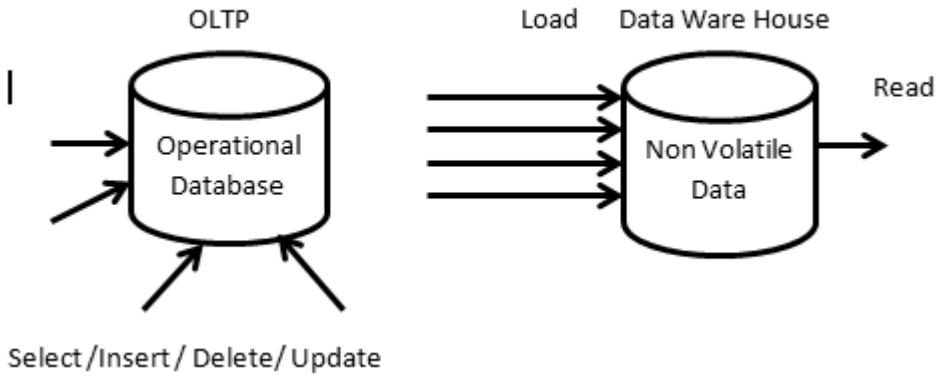

Figure 6: Non –Volatile Characteristic Feature of DW

1.5.1 How Data Warehouse Works?

A data warehouse is a central repository in which one or more sources of information are collected. Data in the data warehouse may be Structured, Semi-structured, or Unstructured. Data are processed, transformed, and accessed by end users for use in business intelligence reporting and decision-making. A data warehouse integrates disparate primary sources into a comprehensive source. Through the integration of all this information, an organization can maintain a more holistic level of customer service. This ensures that all available data is properly considered. Data warehouse enables data mining to find patterns of information that increases profits.

The figure 7 shows the important components of the data warehouse.

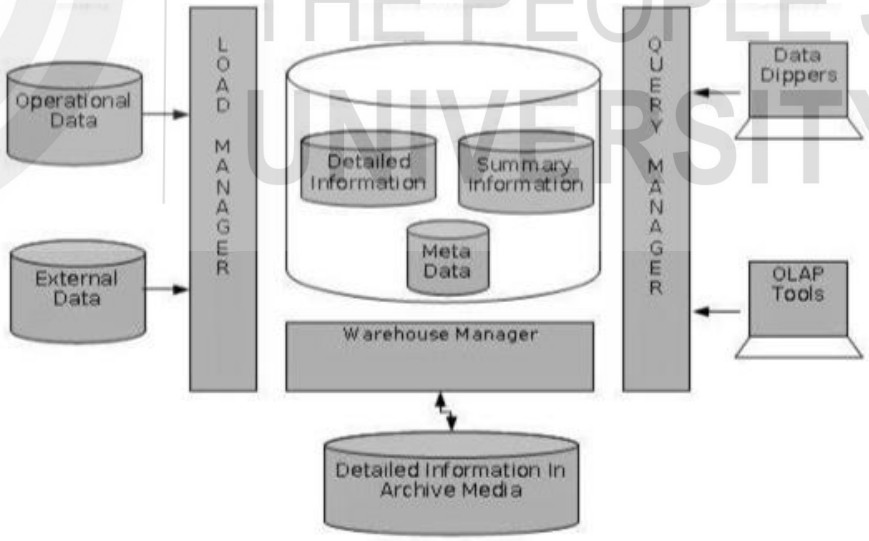

Figure 7: Components of a Data Warehouse

- **Load Manager**

Load Manager Component of data warehouse is responsible for collection of data from operational system and converts them into usable form for the users. This component is responsible for importing and exporting data from operational systems. This component includes all of the programs and

applications interfaces that are responsible for pooling the data out of the operational system, preparing it, loading it into warehouse itself it performs the following tasks such as identification of data, validation of data about the accuracy, extraction of data from original source, cleansing of data by eliminating meaningless values and making it usable, data formatting, data standardization by getting them into a consistent form, data merging by taking data from different sources and consolidating into one place and establishing referential integrity.

- **Warehouse Manager**

The warehouse manager is the center of data-warehousing system and is the data warehouse itself. It is a large, physical database that holds a vast am0unt of information from a wide variety of sources. The data within the data warehouse is organized such that it becomes easy to find, use and update frequently from its sources.

- **Query Manager**

Query Manager Component provides the end-users with access to the stored warehouse information through the use of specialized end-user tools. Data mining access tools have various categories such as query and reporting, on-line analytical processing (OLAP), statistics, data discovery and graphical and geographical information systems.

- **End-user access tools**

This is divided into the following categories, such as:

- Reporting Data
- Query Tools
- Data Dippers
- Tools for EIS
- Tools for OLAP and tools for data mining.

☞ Check Your Progress 1

1) What is a Data Warehouse and why is it important?

……

……

……

……..

2. Mention the characteristics of a Data Warehouse.

……

……

1.6 OLTP AND OLAP

Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) are the two terms which look similar but refer to different kinds of systems. Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

1.6.1 Online Transaction Processing (OLTP)

An OLTP system captures and maintains transaction data in a database. Each transaction involves individual database records made up of multiple fields or columns. Examples include banking and credit card activity or retail checkout scanning.

In OLTP, the emphasis is on fast processing, because OLTP databases are read, written, and updated frequently. If a transaction fails, built-in system logic ensures data integrity.

1.6.2 Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and business intelligence projects. In OLAP, the emphasis is on response time to these complex queries. Each query involves one or more columns of data aggregated from many rows.

Examples include year-over-year financial performance or marketing lead generation trends. OLAP databases and data warehouses give analysts and decision-makers the ability to use custom reporting tools to turn data into information. Query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact the accuracy of business intelligence insights.

OLTP is operational, while OLAP is informational. A glance at the key features of both kinds of processing illustrates their fundamental differences, and how they work together. The table (Table 1) below summarizes differences between OLTP and OLAP.

Table 1: OLTP Vs OLAP

	OLTP	**OLAP**
Characteristics	Handles a large number of small transactions	Handles large volumes of data with complex queries
Query types	Simple standardized queries	Complex queries
Operations	Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting

Response time	Milliseconds	Seconds, minutes, or hours depending on the amount of data to process
Design	Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing
Source	Transactions	Aggregated data from transactions
Purpose	Control and run essential business operations in real time	Plan, solve problems, support decisions, discover hidden insights
Data updates	Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs
Space requirements	Generally small if historical data is archived	Generally large due to aggregating large datasets
Backup and recovery	Regular backups required to ensure business continuity and meet legal and governance requirements	Lost data can be reloaded from OLTP database as needed in lieu of regular backups
Productivity	Increases productivity of end users	Increases productivity of business managers, data analysts, and executives
Data view	Lists day-to-day business transactions	Multi-dimensional view of enterprise data
User examples	Customer-facing personnel, clerks, online shoppers	Knowledge workers such as data analysts, business analysts, and executives
Database design	Normalized databases for efficiency	Denormalized databases for analysis

OLTP provides an immediate record of current business activity, while OLAP generates and validates insights from that data as it's compiled over time. That historical perspective empowers accurate forecasting, but as with all business intelligence, the insights generated with OLAP are only as good as the data pipeline from which they emanate.

☞ Check Your Progress 2

1) Why a data warehouse is separated from Operational Databases?

……

……

……

2) Mention the key differences between a database and a data warehouse.

……

……

1.7 DATA GRANULARITY

Granularity is one of the main elements in the modeling of DW data. Granularity of data refers to detail levels. Multiple levels of detail may be available depending on the requirements. At least two granular levels exist for many data warehouses. The relation between detailing and granularity is important to understand. It means greater detail of the data (less summary) when you speak of less granularity or fine granularity. Greater granularity means fewer details or gross granularity (greater summarization). The operational data is stored at the lowest level of information. The sale units will be stored and stored in the outlet system at the unit level of product per transaction. The amount ordered is collected and stored by the customer at unit level per order in the order entry system. You can add up the individual transactions whenever you need the summary data. When you search for items in a product ordered this month and add them together, you have the total of all product orders entered in that month. Referral data is generally not tracked in an operating system. When a user requests an analysis in the data warehouse, the user has to view summary data first. A user can successfully promote the entire product unit throughout a given area. The user may then wish to examine the region's breakdowns. The next step might be to look at the following levels of the sales units in each store. The analysis often starts at a high level, and is reduced in detail.

Therefore, in a data store, the summary of data at different levels can be maintained effectively. You can provide answers at the simplest level or the most detailed level. The detail level of data is the level of granularity in an existing data warehouse. The more detail in the data, the finer the size of the data. You need to save a lot of permanent data in the data warehouse in order to safely store data. The type of granularity depends on how data will be processed, and the performance expectations. For example, each year details of each month, day, hour, minute, second, and so forth are available.

1.8 METADATA AND DATA WAREHOUSING

In a data warehouse, data is stored using a common schema controlled by a common dictionary. Within the data dictionary, data is kept about the logical data structures, file and address data, index information and others. The metadata should contain the following data warehouse information:

- The data structure based on the programmer's view
- Data structure based on DSS analysts' view
- The DW's data sources
- The data transformation at the moment of its migration to DW
- Model of data
- The connection between the data model and the DW
- Data extraction history

In the DW environment, metadata is a major component. Metadata helps to control reporting accuracy, validates the transformation of data and ensures calculation accuracy. Metadata also complies with the company end-users' definition of business terms. More details on metadata are provided in the next Unit.

1.9 DATA WAREHOUSING APPLICATIONS

In different sectors, there are numerous applications such as e-commerce, telecommunication, transport, marketing, distribution and retail. Given below are some of the applications of data warehouses:

Investment and Insurance: In this sector, data warehousing is used to analyze the customer, market trends and other patterns of data. The two sub-sectors where data warehousing plays an important role are Forex and stock markets.

Healthcare: A data warehousing system is used to forecast outcomes of a treatment generate its reports and share the data with different units. These units can be the research labs, medical units, and insurance providers. Enterprise data warehouses serve as the backbone of healthcare systems as they are updated with recent information which is crucial for saving lives.

Retail: Be it distribution, marketing, examining pricing policies, keeping a track of promotional deals, and finding the pattern in the customer buying trends: data warehousing solves it all. Many retail chains incorporate enterprise data warehousing for business intelligence and forecasting.

Social Media Websites: Social networking sites such as *Facebook, Twitter, LinkedIn* etc. are based on large data sets analyses. These sites collect data on members, groups; locations etc. and store this information in a single central repository. Data warehouse is necessary to implement the same data, because of its high volume of data.

Banking: Most banks are now using warehouses to see account/cardholder spending patterns. They use this to make special offers, deals, etc. available.

Government: In addition to store and analyze taxes used to detect tax theft, government uses the data warehouse.

Airlines: It is used in the airline system for operational purposes such as crew assignments, road profitability analyses, flight frequency programs promotions, etc.

Public sector: Information is collected in the public sector's data warehouse. It helps government agencies and departments manage their data and records.

1.10 TYPES OF DATA WAREHOUSES

There are three different types of traditional Data Warehouse models as listed below:

i. Enterprise
ii. Operational
iii. Data Mart

(i) Enterprise Data Warehouse

An enterprise provides a central repository database for decision support throughout the enterprise. It is a central place where all business information from different sources and applications are made available. Once it is stored, it can be used for analysis and used by all the people across the organization. The enterprise goal is to provide a complete overview of any particular object in the data model.

(ii) Operational Data Warehouse

These features have a sizable enterprise-wide scope, but unlike the substantial enterprise warehouse, data is refreshed in near real-time and used for routine commercial activity. It assists in obtaining data straight from the database, which also helps data transaction processing. The data present in the Operational Data Store can be scrubbed, and the duplication which is present can be reviewed and fixed by examining the corresponding market rules.

(iii) Data Mart

Data Mart may be a subset of knowledge warehouse, and it supports a specific region, business unit, or business function. Data Mart focuses on storing data for a particular functional area, and it contains a subset of data saved in a memory. Data Marts help in enhancing user responses and also reduces the volume of data for data analysis. It makes it more comfortable to go forward with the report. More on Data Marts can be studied in the next Unit.

1.11 POPULAR DATA WAREHOUSE PLATFORMS

A data warehouse is a critical database for supporting data analysis and acts as a conduit between analytical tools and operational data stores. The most popular data warehousing solutions include a range of useful features for data management and consolidation.

You can use them to extract/curate data from a range of environments, transform data and remove duplicates, and ensure consistency in your analytics.

Google BigQuery

BigQuery is a cost-effective data warehousing tool with built-in machine learning capabilities. You can integrate it with *Cloud ML* and *TensorFlow* to create powerful AI models. It can also execute queries on petabytes of data for real-time analytics. This scalable and serverless cloud data warehouse is ideal for companies that want to keep costs low. If you need a quick way to make informed decisions through data analysis, BigQuery is one of the solutions.

AWS Redshift

*Redshif*t is a cloud-based data warehousing tool for enterprises. The platform can process petabytes of data quite fast. That's why it's suitable for high-speed data analytics. It also supports automatic concurrency scaling. The automation increases or decreases query processing resources to match workload demand.

Although tooling provided by Amazon reduces the need to have a database administrator full time, it does not eliminate the need for one. Amazon Redshift is known to have issues with handling storage efficiently in an environment prone to frequent deletes.

Snowflake

Snowflake is a data warehousing solution that offers a variety of options for public cloud technology. With Snowflake, you can make your business more data-driven. You may use Snowflake to set up an enterprise-grade cloud data warehouse. With Snowflake, you can analyze data from various unstructured and structured sources. However, Snowflake is dependent on *Azure, Amazon Web Services (AWS), Google Cloud Services (GCS*). The support can be a problem whenever one of those cloud servers has an independent outage.

Microsoft Azure Synapse

Microsoft Azure is a robust platform for data management, analytics, integration, and more, with solutions spanning AI, blockchain, and more than a dozen unique databases for varying use cases. Among them is Azure Synapse, formerly known as *Azure SQL Data Warehouse*, a platform built for analytics, providing you the ability to query data using either serverless or provisioned resources at scale.

Azure Synapse brings together the two worlds of data warehousing and analytics with a unified experience to ingest, prepare, manage, and serve data for immediate BI and machine learning. The broader Azure platform includes thousands of tools, including others that interface with the various Azure databases.

1.12 SUMMARY

In this unit you have studied about the evolution, characteristics, benefits and applications of data ware house.

Operational database system provides day-to-day information, although strategic decision-making cannot be used easily. Data Warehouse is a concept designed to aid strategic information. Data Warehouse allows people to make decisions and provides flexible, convenient and interactive sources of strategic intelligence. A data warehouse combines several technologies because it collects data from various operational data base systems and external sources such as magazines, newspapers and reports from the same industry, removes contradictions, transforms the data and then stores them in formats suited to easy access for decision-making purposes. The defining characteristics of the data warehouse are: Subject oriented, integrated, time-variant, and non-volatile.

Data warehouses are meant to be used by executives, managers, and other people at higher managerial levels who may not have much technical expertise in handling the databases.

Advantages of data warehouses include better decisions, increased productivity, lower operational costs, enhanced asset and liability management, and better CRM.

1.13 SOLUTIONS/ANSWERS

Check Your Progress 1

1) Data Warehousing (DW) is a process for collecting and managing data from diverse sources to provide meaningful insights into the business. A Data Warehouse is typically used to connect and analyze heterogeneous sources of business data. The data warehouse is the centerpiece of the BI system built for data analysis and reporting.

   It is amalgam of technologies and components which helps to use data strategically. Instead of transaction processing, it is the automated collection of a vast amount of information by a company that is configured for demand and review. It's a process of transforming data into information and making it available for users to make a difference in a timely way.

   The archive of decision support (Data Warehouse) is managed independently from the operating infrastructure of the organization. The data warehouse, however, is not a product but rather an environment. It is an organizational framework of an information

system that provides consumers with knowledge regarding current and historical decision help that is difficult to access or present in the conventional operating data store.

Data storage platforms also sort data on a variety of subjects like customers, products or business.

- Data storage is a tool that companies can use increasingly important for corporate intelligence:
- Make uniformity possible. All research data gathered and shared to decision makers worldwide should be used in a uniform format. Standardization of data from various sources reduces the risk of misinterpretation as well as overall accuracy of interpretation.
- Take better business decisions. Successful entrepreneurs have a thorough understanding of data, and are good at predicting future trends. The data storage system helps users access various data sets at speed and efficiency.
- Data storage platforms allow companies to access their business' past history and evaluate ideas and projects. This gives managers an idea of how they can improve their sales and management practices.

2). Following are the four main characteristics of a data warehouse:

i) Subject oriented

A data warehouse is subject-oriented, as it provides information on a topic rather than the ongoing operations of organizations. Such issues may be inventory, promotion, storage, etc. Never does a data warehouse concentrate on the current processes. Instead, it emphasized modeling and analyzing decision-making data. It also provides a simple and succinct description of the particular subject by excluding details that would not be useful in helping the decision process.

(ii) Integrated

Integration in Data Warehouse means establishing a standard unit of measurement from the different databases for all the similar data. The data must also get stored in a simple and universally acceptable manner within the Data Warehouse. Through combining data from various sources such as a mainframe, relational databases, flat files, etc., a data warehouse is created. It must also keep the naming conventions, format, and coding consistent. Such an application assists in robust data analysis. Consistency must be maintained in naming conventions, measurements of characteristics, specification of encoding, etc.

(iii) Time-variant

Compared to operating systems, the time horizon for the data warehouse is given period and provides historical information. It contains a temporal element, either explicitly or implicitly.One such

location in the record key system where Data Warehouse data shows time variation is. Each primary key contained with the DW should have an element of time either implicitly or explicitly. Just like the day, the month of the week, etc.

(iv) Non-volatile

Also, the data warehouse is non-volatile, meaning that prior data will not be erased when new data are entered into it. Data is read-only, only updated regularly. It also assists in analyzing historical data and in understanding what and when it happened. The transaction process, recovery, and competitiveness control mechanisms are not required. In the Data Warehouse environment, activities such as deleting, updating, and inserting that are performed in an operational application environment are omitted.

Check Your Progress 2

1) Data Warehouse systems are segregated from production databases so that they aren't intermingled and cause conflicts.

- There is a database available for tasks such as searching records, indexing, and digital archiving. Data warehouse queries are often complex due to their varied and complex nature.
- It is possible to manage multiple transactions simultaneously through business databases. Concurrency control and recovery mechanisms are needed to ensure that the database in operational databases is robust and consistent.
- The operational database query allows for reading and modification of operations, whilst the read access to stored information is required for OLAP queries only.
- A database of operations maintains current information. In contrast, historical data is kept in a warehouse.

2) A database stores the current data required to power an application. A data warehouse stores current and historical data from one or more systems in a predefined and fixed schema, which allows business analysts and data scientists to easily analyze the data. The table below summarizes differences between databases, data warehouses:

Table 2: Database Vs Data Warehouse

Characteristic Feature	Database	Data Warehouse
Workloads	Operational and transactional	Analytical
Data Type	Structured or semi-structured	Structured and/or semi-structured

Schema Flexibility	Rigid or flexible schema depending on database type	Pre-defined and fixed schema definition for ingest (schema on write and read)
Data Freshness	Real time	May not be up-to-date based on frequency of ETL processes
Users	Application developers	Business analysts and data scientists
Pros	Fast queries for storing and updating data	The fixed schema makes working with the data easy for business analysts
Cons	May have limited analytics capabilities	Difficult to design and evolve schema Scaling compute may require unnecessary scaling of storage, because they are tightly coupled

1.14 FURTHER READINGS

1. William H. Inmon, Building the Data Warehouse, Wiley, 4th Edition, 2005.
2. Data Warehousing Fundamentals, Paulraj Ponnaiah, Wiley Student Edition
3. Data Warehousing, Reema Thareja, Oxford University Press