UNIT 2 DATA WAREHOUSE ARCHITECTURE

Structure

2.0 INTRODUCTION

In the previous unit we had studied about the data warehousing and related topics. Despite numerous advancements over the last five years in the arena of Big Data, cloud computing, predictive analysis, and information technologies, data warehouses have only gained more significance. For the success of any data warehouse, its architecture plays an important role. Since three decades, the data warehouse architecture has been the pillar of the corporate data ecosystems.

This unit present various topics including the basic concept of data warehouse architecture, its types, significant components and layers of data ware house architecture, data marts and their designing.

2.1 OBJECTIVES

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

- Understand the purpose of data warehouse architecture;
- Describe the process of storing the data in a data warehouse;
- List and discuss the various types of data warehouse architectures;
- Discuss various components and layers of data warehouse architecture;
- To summarize the functionality of data marts, their benefits and various types, and
- To know the ways of structuring and designing the data marts.

2.2 DATA WAREHOUSE ARCHITECTURE AND ITS TYPES

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.

2.2.1 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

The details of each of the architecture are given below:

(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.

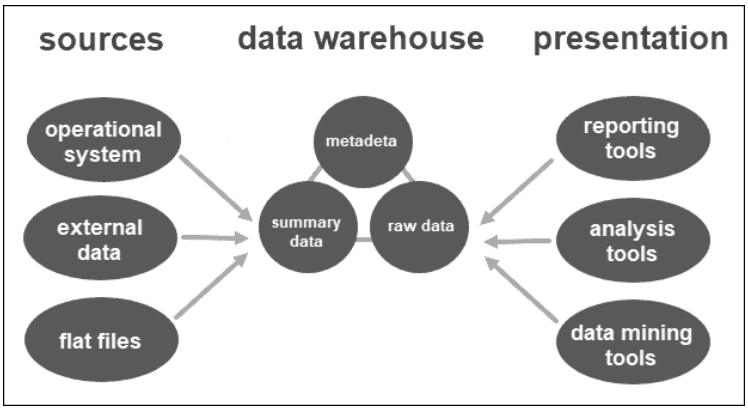

Figure 1: Single Tier Data Warehouse Architecture

(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.

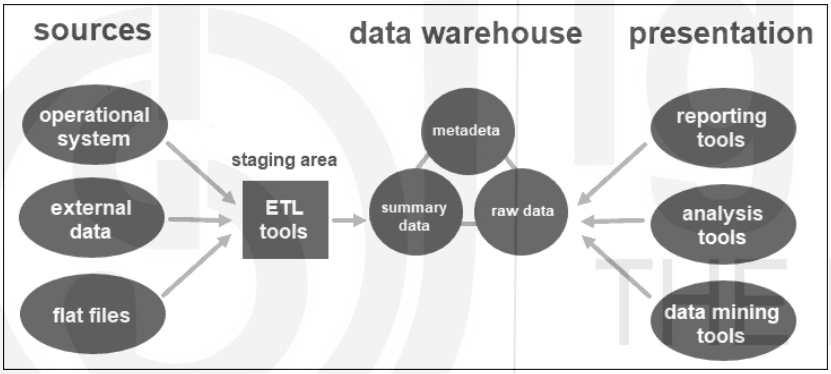

Figure 2: Two- Tier Data Warehouse Architecture

(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.

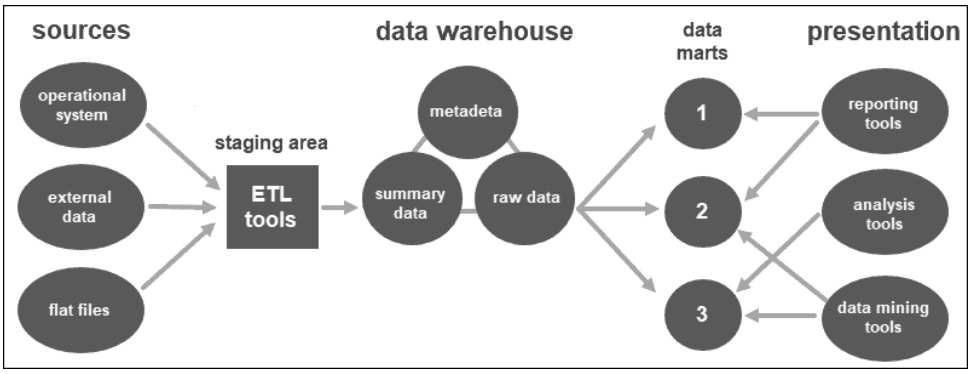

Figure 3: Three- Tier Data Warehouse Architecture

Figure 4 illustrates the complete data warehouse architecture with the three tiers:

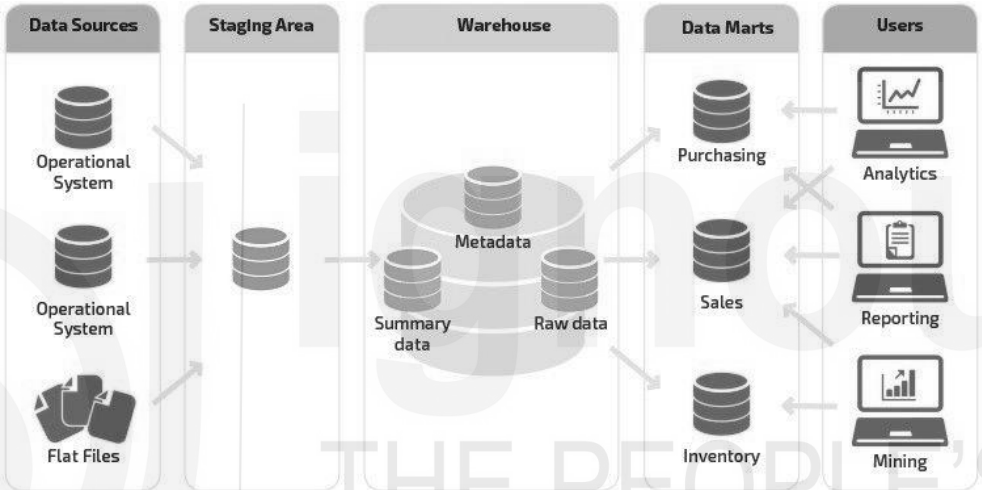

Figure 4: 3-Tiers of Data Warehouse

2.2.2 Cloud-based Data Warehouse Architecture

Cloud-based data warehouse architecture is relatively new when compared to legacy options. This data warehouse architecture means that the actual data warehouses are accessed through the cloud. There are several cloud based data warehouses options, each of which has different architectures for the same benefits of integrating, analyzing, and acting on data from different sources. The difference between a cloud-based data warehouse approach compared to that of a traditional approach include:

- **Up-front costs:** The different components required for traditional, on-premises data warehouses mandate pricey up-front expenses. Since the components of cloud architecture are accessed through the cloud, these expenses don't apply.
- **Ongoing costs:** While businesses with on-prem data warehouses must deal with upgrade and maintenance costs, the cloud offers a low, pay-as-you-go model.

- **Speed:** Cloud-based data warehouse architecture is substantially speedier than on-premises options, partly due to the use of ELT — which is an uncommon process for on-premises counterparts.
- **Flexibility:** Cloud data warehouses are designed to account for the variety of formats and structures found in big data. Traditional relational options are designed simply to integrate similarly structured data.
- **Scale:** The elastic resources of the cloud make it ideal for the scale required of big datasets. Additionally, cloud-based data warehousing options can also scale down as needed, which is difficult to do with other approaches.

Cloud-based platforms make it possible to create, share, and store massive data sets with ease, paving the way for more efficient and effective data access and analysis. Cloud systems are built for sustainable business growth, with many modern Software-as-a Service (SaaS) providers separating data storage from computing to improve scalability when querying data.

Some of the more notable cloud data warehouses in the market include Amazon Redshift, Google BigQuery, Snowflake, and Microsoft Azure SQL Data Warehouse.

Now, let's learn about the major components of a data warehouse and how they help build and scale a data warehouse in the next section.

2.3 COMPONENTS OF DATA WAREHOUSE ARCHITECTURE

A data warehouse design consists of six main components:

- Data Warehouse Database
- ETL
- Metadata
- Data Warehouse Access Tools
- Data Warehouse Bus
- Data Warehouse Reporting Layer

The details of all the components are given below.

2.3.1 Data Warehouse Database

The central component of a DW architecture is a data warehouse database that stocks all enterprise data and makes it manageable for reporting. Obviously, this means you need to choose which kind of database you'll use to store data in your warehouse.

The following are the four database types that you can use:

- Typical relational databases are the row-centered databases you perhaps use on an everyday basis —for example, Microsoft SQL Server, SAP, Oracle, and IBM DB2.
- Analytics databases are precisely developed for data storage to sustain and manage analytics, such as Teradata and Greenplum.
- Data warehouse applications aren't exactly a kind of storage database, but several dealers now offer applications that offer software for data management as well as hardware for storing data. For example, SAP Hana, Oracle Exadata, and IBM Netezza.
- Cloud-based databases can be hosted and retrieved on the cloud so that you don't have to procure any hardware to set up your data warehouse—for example, Amazon Redshift, Google BigQuery, and Microsoft Azure SQL.

2.3.2 Extraction, Transformation, and Loading Tools (ETL)

ETL tools are central components of enterprise data warehouse architecture. These tools help extract data from different sources, transform it into a suitable arrangement, and load it into a data warehouse.

The ETL tool you choose will determine:

- The time expended in data extraction
- Approaches to extracting data
- Kind of transformations applied and the simplicity to do so
- Business rule definition for data validation and cleansing to improve end-product analytics
- Filling mislaid data
- Outlining information distribution from the fundamental depository to your BI applications

2.3.3 Metadata

Before we delve into the different types of metadata in data mining, we first need to understand what metadata is. In the data warehouse architecture, metadata describes the data warehouse database and offers a framework for data. It helps in constructing, preserving, handling, and making use of the data warehouse.

There are two types of metadata in data mining:

- **Technical Metadata** comprises information that can be used by developers and managers when executing warehouse development and administration tasks.
- **Business Metadata** comprises information that offers an easily understandable standpoint of the data stored in the warehouse.

Metadata plays an important role for businesses and the technical teams to understand the data present in the warehouse and convert it into information.

2.3.4 Data Warehouse Access Tools

A data warehouse uses a database or group of databases as a foundation. Data warehouse corporations generally cannot work with databases without the use of tools unless they have database administrators available. However, that is not the case with all business units. This is why they use the assistance of several no-code data warehousing tools, such as:

- **Query and reporting tools** help users produce corporate reports for analysis that can be in the form of spreadsheets, calculations, or interactive visuals.
- **Application development tools** help create tailored reports and present them in interpretations intended for reporting purposes.
- **Data mining tools for data warehousing** systematize the procedure of identifying arrays and links in huge quantities of data using cutting-edge statistical modeling methods.
- **OLAP tools** help construct a multi-dimensional data warehouse and allow the analysis of enterprise data from numerous viewpoints.

2.3.5 Data Warehouse Bus

It defines the data flow within a data warehousing bus architecture and includes a data mart. A data mart is an access level that allows users to transfer data. It is also used for partitioning data that is produced for a particular user group.

2.3.6 Data Warehouse Reporting Layer

The reporting layer in the data warehouse allows the end-users to access the BI interface or BI database architecture. The purpose of the reporting layer in the data warehouse is to act as a dashboard for data visualization, create reports, and take out any required information.

Constructing a data warehouse is primarily dependent on a particular business. And so, every data warehouse architecture has four layers. Let's study them in following section.

2.4 LAYERS OF DATA WAREHOUSE ARCHITECTURE

In general, the data warehouse architecture can be divided into four layers. They are:

i. Data Source Layer
ii. Data Staging Layer
iii. Data Storage Layer
iv. Data Presentation Layer

Let us study the various layers and their functionality.

(i) Data source layer

The data source layer is the place where unique information, gathered from an assortment of inner and outside sources, resides in the social database. Following are the examples of the data source layer:

- ***Operational Data*** — Product information, stock information, marketing information, or HR information
- ***Social Media Data*** — Website hits, content fame, contact page completion
- ***Outsider Data*** — Demographic information, study information, statistics information

While most data warehouses manage organized data, thought ought to be given to the future utilization of unstructured data sources, for example, voice accounts, scanned pictures, and unstructured text. These floods of data are significant storehouses of information and ought to be viewed when building up your warehouse.

(ii) Data Staging Layer

This layer dwells between information sources and the data warehouse. In this layer, information is separated from various inside and outer data sources. Since source data comes in various organizations, the data extraction layer will use numerous technologies and devices to extricate the necessary information. Once the extracted data has been stacked, it will be exposed to high-level quality checks. The conclusive outcome will be perfect and organized data that you will stack into your data warehouse. The staging layer contains the given parts:

- **Landing Database and Staging Area**

The landing database stores the information recovered from the data source. Before the data goes to the warehouse, the staging process does stringent quality checks on it. Arranging is a basic step in architecture. Poor information will add up to inadequate data, and the result is poor business dynamic. The arranging layer is where you need to make changes in accordance with the business process to deal with unstructured information sources.

- **Data Integration Tool**

Extract, Transform and Load tools (ETL) are the data tools used to extricate information from source frameworks, change, and prepare information and load it into the warehouse.

(iii) Data Storage Layer

This layer is the place where the data that was washed down in the arranging zone is put away as a solitary central archive. Contingent upon your business and your warehouse architecture necessities, your data storage might be a data

warehouse center, data mart (data warehouse somewhat recreated for particular departments), or an Operational Data Store (ODS).

(iv) Data Presentation Layer

This is where the users communicate with the scrubbed and sorted out data. This layer of the data architecture gives users the capacity to query the data for item or service insights, break down the data to conduct theoretical business situations, and create computerized or specially appointed reports.

You may utilize an OLAP or reporting instrument with an easy to understand Graphical User Interface (GUI) to assist users with building their queries, perform analysis, or plan their reports.

2.4.1 Best Practices for Data Warehouse Architecture

Designing the data warehouse with the designated architecture is an art. Some of the best practices are shown below:

- Create data warehouse models that are optimized for information retrieval in both dimensional, de-normalized, or hybrid approaches.
- Select a single approach for data warehouse designs such as the top-down or the bottom-up approach and stick with it.
- Always cleanse and transform data using an ETL tool before loading the data to the data warehouse.
- Create an automated data cleansing process where all data is uniformly cleaned before loading.
- Allow sharing of metadata between different components of the data warehouse for a smooth retrieval process.
- Always make sure that data is properly integrated and not just consolidated when moving it from the data stores to the data warehouse. This would require the 3NF normalization of data models.
- Monitor the performance and security. The information in the data warehouse is valuable, though it must be readily accessible to provide value to the organization. Monitor system usage carefully to ensure that performance levels are high.
- Maintain the data quality standards, metadata, structure, and governance. New sources of valuable data are becoming available routinely, but they require consistent management as part of a data warehouse. Follow procedures for data cleaning, defining metadata, and meeting governance standards.
- Provide an agile architecture. As the corporate and business unit usage increases, they will discover a wide range of data mart and warehouse needs. A flexible platform will support them far better than a limited, restrictive product.
- Automate the processes such as maintenance. In addition to adding value to business intelligence, machine learning can automate data warehouse technical management functions to maintain speed and reduce operating costs.
- Use the cloud strategically. Business units and departments have different deployment needs. Use on-premise systems when required,

and capitalize on cloud data warehouses for scalability, reduced cost, and phone and tablet access.

2.5 DATA MARTS

A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse. For example, many companies may have a data mart that aligns with a specific department in the business, such as finance, sales, or marketing.

2.5.1 Data Mart Vs Data Warehouse

Data marts and data warehouses are both highly structured repositories where data is stored and managed until it is needed. However, they differ in the scope of data stored: data warehouses are built to serve as the central store of data for the entire business, whereas a data mart fulfills the request of a specific division or business function. Because a data warehouse contains data for the entire company, it is best practice to have strictly control who can access it. Additionally, querying the data you need in a data warehouse is an incredibly difficult task for the business. Thus, the primary purpose of a data mart is to isolate—or partition—a smaller set of data from a whole to provide easier data access for the end consumers.

A data mart can be created from an existing data warehouse—the top-down approach—or from other sources, such as internal operational systems or external data. Similar to a data warehouse, it is a relational database that stores transactional data (time value, numerical order, reference to one or more object) in columns and rows making it easy to organize and access.

On the other hand, separate business units may create their own data marts based on their own data requirements. If business needs dictate, multiple data marts can be merged together to create a single, data warehouse. This is the bottom-up development approach.

In a nut-shell, following are the differences:

- Data mart is for a specific company department and normally a subset of an enterprise-wide data warehouse.
- Data marts improve query speed with a smaller, more specialized set of data.
- Data warehouses help make enterprise-wide strategic decisions, data marts are for department level, tactical decisions.
- Data warehouse includes many data sets and takes time to update, data marts handle smaller, faster-changing data sets.

- Data warehouse implementation can take many years, data marts are much smaller in scope and can be implemented in months.

2.6 BENEFITS OF DATA MARTS

Data marts are designed to meet the needs of specific groups by having a comparatively narrow subject of data. And while a data mart can still contain millions of records, its objective is to provide business users with the most relevant data in the shortest amount of time.

With its smaller, focused design, a data mart has several benefits to the end user, including the following:

- **Cost-efficiency:** There are many factors to consider when setting up a data mart, such as the scope, integrations, and the process to extract, transform, and load (ETL). However, a data mart typically only incurs a fraction of the cost of a data warehouse.
- **Simplified data access:** Data marts only hold a small subset of data, so users can quickly retrieve the data they need with less work than they could when working with a broader data set from a data warehouse.
- **Quicker access to insights:** Intuition gained from a data warehouse supports strategic decision-making at the enterprise level, which impacts the entire business. A data mart fuels business intelligence and analytics that guide decisions at the department level. Teams can leverage focused data insights with their specific goals in mind. As teams identify and extract valuable data in a shorter space of time, the enterprise benefits from accelerated business processes and higher productivity.
- **Simpler data maintenance:** A data warehouse holds a wealth of business information, with scope for multiple lines of business. Data marts focus on a single line, housing under 100GB, which leads to less clutter and easier maintenance.
- **Easier and faster implementation:** A data warehouse involves significant implementation time, especially in a large enterprise, as it collects data from a host of internal and external sources. On the other hand, you only need a small subset of data when setting up a data mart, so implementation tends to be more efficient and include less set-up time.

2.7 TYPES OF DATA MARTS

There are three types of data marts that differ based on their relationship to the data warehouse and the respective data sources of each system.

- **Dependent data marts** are partitioned segments within an enterprise data warehouse. This top-down approach begins with the storage of all business data in one central location. The newly created data marts extract a defined subset of the primary data whenever required for analysis.
- **Independent data marts** act as a standalone system that doesn't rely on a data warehouse. Analysts can extract data on a particular subject or business process from internal or external data sources, process it, and then store it in a data mart repository until the team needs it.
- **Hybrid data marts** combine data from existing data warehouses and other operational sources. This unified approach leverages the speed and user-friendly interface of a top-down approach and also offers the enterprise-level integration of the independent method.

2.8 STRUCTURE OF A DATA MART

A data mart is a subject-oriented relational database that stores transactional data in rows and columns, which makes it easy to access, organize, and understand. As it contains historical data, this structure makes it easier for an analyst to determine data trends. Typical data fields include numerical order, time value, and references to one or more objects.

Companies organize data marts in a multidimensional schema as a blueprint to address the needs of the people using the databases for analytical tasks. The three main types of schema:

Star

Star schema is a logical formation of tables in a multidimensional database that resembles a star shape. In this blueprint, one fact table—a metric set that relates to a specific business event or process—resides at the center of the star, surrounded by several associated dimension tables.

There is no dependency between dimension tables, so a star schema requires fewer joins when writing queries. This structure makes querying easier, so star schemas are highly efficient for analysts who want to access and navigate large data sets.

Snowflake

A snowflake schema is a logical extension of a star schema, building out the blueprint with additional dimension tables. The dimension tables are normalized to protect data integrity and minimize data redundancy.

While this method requires less space to store dimension tables, it is a complex structure that can be difficult to maintain. The main benefit of using snowflake schema is the low demand for disk space, but the caveat is a negative impact on performance due to the additional tables.

Vault

Data vault is a modern database modeling technique that enables IT professionals to design agile enterprise data warehouses. This approach enforces a layered structure and has been developed specifically to combat issues with agility, flexibility, and scalability that arise when using the other schema models.

Data vault eliminates star schema's need for cleansing and streamlines the addition of new data sources without any disruption to existing schema.

2.9 DESIGNING THE DATA MARTS

Data marts guide important business decisions at a departmental level. For example, a marketing team may use data marts to analyze consumer behaviors, while sales staff could use data marts to compile quarterly sales reports. As these tasks happen within their respective departments, the teams don't need access to all enterprise data.
Typically, a data mart is created and managed by the specific business department that intends to use it. The process for designing a data mart usually comprises the following steps:

(i) Essential Requirements Gathering

The first step is to create a robust design. Some critical processes involved in this phase include collecting the corporate and technical requirements, identifying data sources, choosing a suitable data subset, and designing the logical layout (database schema) and physical structure.

(ii) Build/Construct

The next step is to construct it. This includes creating the physical database and the logical structures. In this phase, you'll build the tables, fields, indexes, and access controls.

(iii) Populate/Data Transfer

The next step is to populate the mart, which means transferring data into it. In this phase, you can also set the frequency of data transfer, such as daily or weekly. This usually involves extracting source information, cleaning and transforming the data, and loading it into the departmental repository.

(iv) Data Access

In this step, the data loaded into the data mart is used in querying, generating reports, graphs, and publishing. The main task involved in this phase is setting up a meta-layer and translating database structures and item names into corporate expressions so that non-technical operators can easily use the data mart. If necessary, you can also set up API and interfaces to simplify data access.

(v) Manage

The last step involves management and observation, which includes:

- Controlling ongoing user access.
- Optimization and refinement of the target system for improved performance.
- Addition and management of new data into the repository.
- Configuring recovery settings and ensuring system availability in the event of failure.

2.10 LIMITATIONS WITH DATA MARTS

Prospective builders of data warehouses are frequently advised to "start small" with a data mart and use that kernel to expand gradually into a full blown data warehouse. This approach to warehousing generally leads to failed projects for several reasons.
Sometimes the new data mart is so successful that the configuration is overrun by user demands. The databases grow too large too fast, response times become unacceptably long, and user frustration leads to searching for other ways to get the answers.

The more common reason for failure is that the data mart is immediately unsuccessful because it is designed in such a way that users are unable to retrieve the sort of information they want and need to extract from the data. Databases are highly denormalized to respond to a small set of canned queries; summaries, rather than detail data, comprise the database so that fine-grained exploratory data analysis is not possible; and support for ad hoc queries is either absent or so poor as to discourage users from bothering with them.

The very factors that frequently defeat data mart projects are also the most commonly recommended approaches to designing data marts and data warehouses in the popular data warehousing literature:

- Denormalization (dimensional modeling)
- Storing aggregates at the expense of detail data
- Skewing performance toward a small, preselected set of queries at the expense of all other exploratory analyses

Check your Progress 1

1. Define data warehouse architecture.

……
……
……
……
……

2. What is the correct flow of the data warehouse architecture?

……
……
……
……
……

3. Mention some Data Mart Use Cases.

……
……
……
……
……

2.11 SUMMARY

Data warehouse architecture is the design and building blocks of the modern data warehouse. In this unit we have studied the basic building blocks of the data warehouse, data warehouse architecture, its types, architecture models, data marts, designing of data marts and limitations.

In this next unit we will study about Dimensional Modeling.

2.12 SOLUTIONS / ANSWERS

Check Your Progress 1:

1. The method for defining the entire architecture of data communication processing as well as the presentation that exists for end-clients is the data warehouse architecture. Every data warehouse is different, and each of them

is characterized based on the standard vital components.

In simple words, a data warehouse is an information system that consists of commutative and historical data from single or multiple sources. The process of reporting and analysis of data in the organizations is simplified with the help of different data warehousing concepts. There are different approaches to constructing a data warehouse architecture. Any approach is used based on the requirements of the organizations.

2. On every operational database, there are a certain fixed number of operations that have to be applied. There are different well-defined techniques for delivering suitable solutions. Data warehousing is found to be more effective when the correct flow of the data warehouse architecture is completely followed.

   The four different processes that contribute to a data warehouse are extracting and loading the data, cleaning and transforming the data, backing up and archiving the data, and carrying out the query management process by directing them to the appropriate data sources.

3. Data marts are used to solve specific organizational problems, especially those that are unique to one department. Typical use cases for a data mart include:

   Focused Analytics
   Analytics is perhaps the most common application of data marts. The data in these repositories is entirely relevant to the requirements of the business department, with no extraneous information, resulting in faster and more accurate analysis. For example, financial analysts will find it easier to work with a financial data mart, rather than working with an entire data warehouse.
   Fast Turnaround
   Data marts are generally faster to develop than a data warehouse, as the developers are working with fewer sources and a limited schema. Data marts are ideal for data projects operating under challenging time constraints.
   Permission Management
   Data marts can be a risk-free way to grant limited data access without exposing the entire data warehouse. For example, dependent data mart contains a segment of warehouse data, and users are only able to view the contents of the mart. This prevents unauthorized access and accidental writes.
   Better Resource Management
   Data marts are sometimes used where there is a disparity in resource usage between different departments. For example, the logistics department might perform a high volume of daily database actions, which causes the marketing team’s analytics tools to run slow. By

providing each department with its own data mart, it's easier to allocate resources according to their needs.

2.13 FURTHER READINGS

1. William H. Inmon, Building the Data Warehouse, Wiley, 4th Edition, 2005.
2. Data Warehousing Fundamentals, Paulraj Ponnaiah, Wiley Student Edition, 2001.
3. Data Warehousing, Reema Thareja, Oxford University Press, 2009.