# DATA WAREHOUSE

Ishika Jain
19111027, 6th Semester
Department of Biomedical Engineering
National Institute of Technology , Raipur

Guided by :
Prof. Saurabh Gupta Sir

**ABSTRACT**

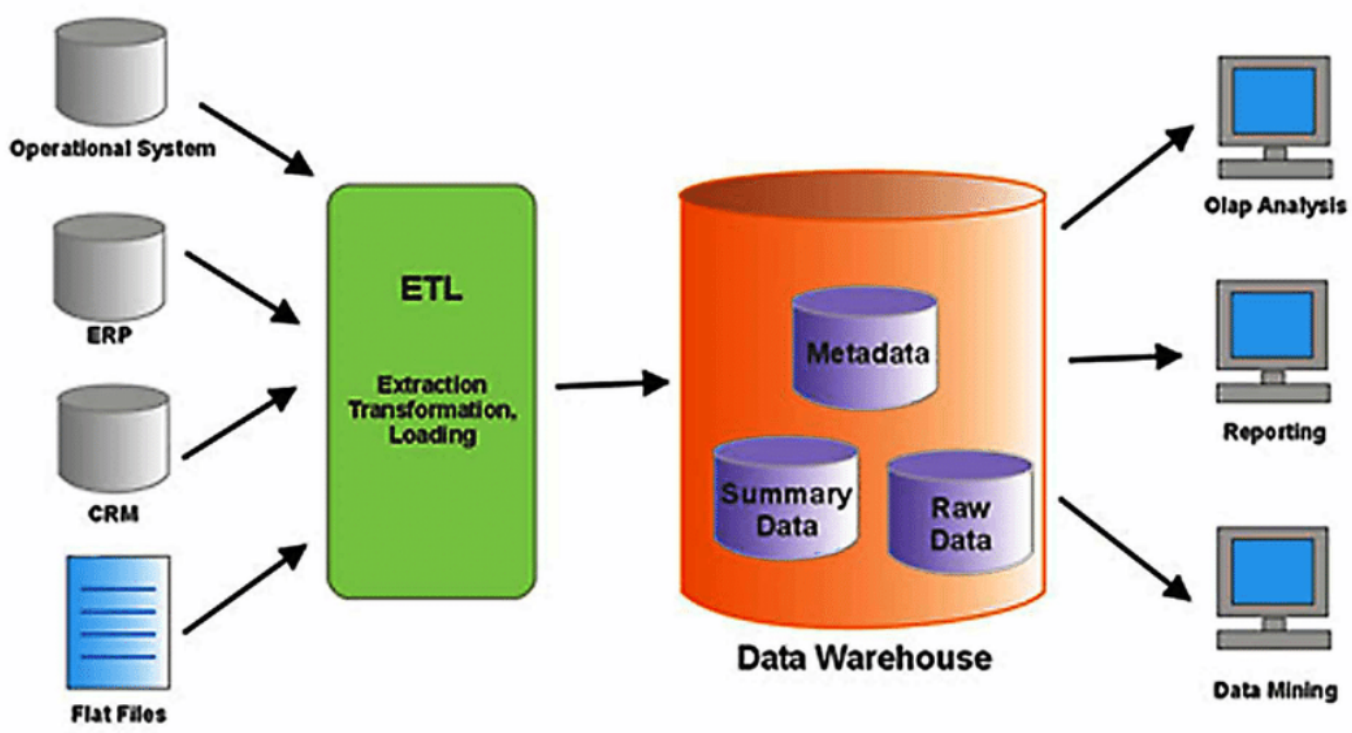
A data warehouse is a centralized repository for the storage of data from one or more aggregated sources, updated immediately with real-time data yet retaining prior data for a more comprehensive dataset. The data in the data warehouse can be different types of data in various formats for disparate sources such as electronic health records and other clinical data and operational and administrative records – all can be in other formats and come from multiple sources of technologies and people. A Healthcare data warehouse often depends on integration tools to support extraction, transformation and loading (ETL) from proprietary healthcare systems such as EPIC, Cerner, and many others. This term paper deals with the fundamentals of a data ware house along with its importance in healthcare and its implementation in SQL.

**TABLE OF CONTENTS**

## 1 WHAT IS A DATA WAREHOUSE?

A data warehouse (DW) is a digital storage system that connects and harmonizes large amounts of data from many different sources. Its purpose is to feed business intelligence (BI), reporting, and analytics, and support regulatory requirements – so companies can turn their data into insight and make smart, data-driven decisions. Data warehouses store current and historical data in one place and act as the single source of truth for an organization.

- This is an efficient method essential for the decision making process and the forecasting process of an organization.
- It is a form of a Relational Database Management System (RDBMS Technology), consisting of an entire set of relational databases, which gives an idea about strategies that need to be planned for the future of a company.
- The Data warehouse is used for research and collection of data from different sources, such that there are minimal number of clients yet interactions are wide.
- There are various functions of a data warehouse. It is not widely used in the transaction process, but has the ability to act as an analytical tool with few tables.
- Data Warehouse environment contains an extraction, transportation, and loading (ETL) solution, an online analytical processing (OLAP) engine, customer analysis tools, and other applications that handle the process of gathering information and delivering it to business users.




Data warehouse model

## 2 DATABASE vs DATA WAREHOUSE

Database consists of a repository of real-time information which is specific to a particular application. This is a transactional system as it is used in the transaction process. Data Warehouse consists of a repository of a huge volume of structured data, which compiles relational databases and information centers from a wide variety of sources which is of huge help to a company for decision making. This is not used in transactions, but is beneficial from a historical perspective.

For example, Database will give information regarding a person's current phone number while data warehouse will contain information about all the phone numbers used by a person in a predefined period of time in the past. This is because it will compile all the data from all the databases.

## 3 NEED FOR A DATA WAREHOUSE

There are five reasons behind the need of Data warehousing:

- Non-technical People- Business users are the non technical people who need to gather information in a summarized, elementary fashion, this function is fulfilled by Data Warehousing.

- Storing the historical data- The time variable related data from the past needs to be stored for future use.
- Strategic decision making- Data warehousing helps in making strategic decisions based on the data given in the warehouse.
- Data consistency and quality of data- Data warehousing helps in maintaining the consistency and uniformity of the data, even though it has been derived from heterogeneous sources.
- Response time is fast- Data warehousing provides a significant degree of flexibility and faster response time that helps it to deal with a lot of load and queries.

## 4 CHARACTERISTICS

The four characteristics of a data warehouse, also called features of a data warehouse, include:

1. Subject Oriented: Analysis of the data for the decision makers of a business can be done easily by constricting to a particular subject area of the Data warehouse. This makes understanding and analysis of the data concise and straightforward by excluding the unwanted information on some subject that is not needed for decision-making. This means that the ongoing operations of an organization are not taken into consideration.

2. Integrated

   Data warehouses consist of data from different variable sources integrated under one platform. This data obtained is extracted and transformed maintaining uniformity without depending on the source it was obtained from, this feature is known as Integrated. Standards are established which are universally acceptable for the data present in the warehouse.

3. Time Variant One of the important properties of the data warehouse is the historical perspective it holds. It keeps the huge volume of data from all databases stored in accordance with the elements of time. It consists of a temporal element and extensive time horizon. Inability to change the element of time is an essential aspect of time variance. Record key is used to display time variance.

4. Non-Volatile

   Data is updated by uploading data in the data warehouse to protect data from momentary changes. This means that once a data is fed, there can be no alteration or changes made. The inability to be erased is called the non-volatile character of the data warehouse environment. data is read only and allows only two functions to be performed: Access and Loading.

## 5 COMPONENTS OF ITS ARCHITECTURE

1. Operational Source – An operational Source is a data source consists of Operational Data and External Data. Data can come from Relational DBMS like Informix, Oracle.
2. Load Manager – The Load Manager performs all operations associated with the extraction of loading data in the data warehouse. These tasks include the simple transformation of data to prepare data for entry into the warehouse.
3. Warehouse Manage – The warehouse manager is responsible for the warehouse management process. The operations performed by the warehouse manager are the analysis, aggregation, backup and collection of data, de-normalization of the data.
4. Query Manager – Query Manager performs all the tasks associated with the management of user queries. The complexity of the query manager is determined by the end-user access operations tool and the features provided by the database.
5. Detailed Data – It is used to store all the detailed data in the database schema. Detailed data is loaded into the data warehouse to complement the data collected.
6. Summarized Data – Summarized Data is a part of the data warehouse that stores predefined aggregations These aggregations are generated by the warehouse manager.
7. Archive and Backup Data – The Detailed and Summarized Data are stored for the purpose of archiving and backup. The data is relocated to storage archives such as magnetic tapes or optical disks.
8. Metadata – Metadata is basically data stored above data. It is used for extraction and loading process, warehouse, management process, and query management process.
9. End User Access Tools – End-User Access Tools consist of Analysis, Reporting, and mining. By using end-user access tools users can link with the warehouse.

## 6 ADVANTAGES OF DATA WAREHOUSING

- Analysis: It helps in effective analysis and better decision-making process.
- Data Access: It makes accessing data fast and simple.
- Data Quality: It helps in maintaining consistency in quality of the data.

- Trends: It helps in understanding trends better and forecast decisions easily, leading to an increase in productivity.
- Data Volume: It includes repositories consisting of huge volumes of data and helps in managing such data.
- Demands: It is an efficient method to handle demands of different users.
- Queries: It is designed to be able to easily handle complex queries.
- Data Maintenance: It helps in maintaining the data with respect to history of the information provided.
- Storage: It is valuable for merger organizations since it allows storage of heterogeneous data.
- Extra Functions: Coding, Descriptions, Flagging, Fixations are all possible due to data warehousing.
- Restructuring Data: Data can be restructured, though it is not erased, based on the operations desired by the business user.
- Operational Business Applications: It adds value to operational business applications like the customer relationship management (CRM) systems.
- Data Model: It can merge data to form a common data model.
- Less Time Consuming Cheap: It takes less time to provide output and is cheaper than other programs.
- Use of Analytics: It uses better BI analytics (Business Intelligence tools) called enterprise data warehouse (EDW), a centralized data repository to analyze and generate reports.

## 7 DISADVANTAGES OF DATA WAREHOUSING

- Reduced Flexibility: Homogenization of data makes working a little less flexible and also leads to loss of data. This problem can be solved by monitoring the data cleaning process.
- Copyright Issues: Since data is added to a centralized warehouse, copyright issues may occur that makes data insecure.
- Increased Reports: Large volume of data means increased amount of reports generation and use of resources. Categorization of data can prevent this.
- Maintenance Problems: Problems in this system remain hidden and hence, need proper maintenance.

## 8 IMPLEMENTATION USING SQL

Mentioned below is the process of getting data from the Lake into the Warehouse.We are using SQL to perform all transformations. It's the standard language for relational database management systems (which is what a Data Warehouse should be) and it's the environment that we use for the Data Lake. Working in a SQL-based model is ideal because a variety of tools and platforms already exist to write and execute queries. Also, data engineers, analysts, and some business users already understand how to use it.Views allow us to quickly reformat what the data looks like without needing to build a new Data Warehouse or incurring costs from storing any additional data.
So let's look at a messy table with all of the hard to understand/query fields.

| 2 id columns | | Nulls and inconsistent naming | | | Column name and values not descriptive | JSON would need to be parsed | Deprecated data |
|---|---|---|---|---|---|---|---|
| **Id** | **External_Id** | **Name** | **Display Name** | **Location** | **Type** | **Info** | **is_deleted** |
| 21590 | 68791 | Doug Gonzalez | D Gonzalez | Texas | 1 | { groups: ["Admin", "R&D"] title: "Director of R&D", status: "active" } | False |
| 13107 | 32699 | | Sales | USA | 3 | { groups: "Sales" title: "", status: "active" } | True |
| 29448 | 28175 | Josh | Josh Redman | US | 2 | { groups: ["Marketing", "HR"] title: "CMO", status: "inactive" } | False |
| 32641 | 19873 | Hannah To | | San Paulo | 1 | { groups: ["Sales", "Editor"] title: "Account Executive", status: "active" } | False |

We then want to make all of the following changes:

| Drop unused column | | Add consistent column | | Standardize | Make column name and values descriptive | Parse relevant fields, drop original column | |
|---|---|---|---|---|---|---|---|
| **Id** | **External_Id** | **Name** | **Display Name** | **Email** | **Location** | **Access Level** | **Info** | **Status** |
| 21590 | 68791 | Doug Gonzalez | D Gonzalez | dgonzalez@gmail.com | USA | Can view | { groups: ["Admin", "R&D"] title: "Director of R&D", status: "active" } | active |
| 13107 | 32699 | | Sales | lisaf@yahoo.com | USA | Can admin | { groups: "Sales" title: "", status: "active" } | active |
| 29448 | 28175 | Josh | Josh Redman | josh@gmail.com | USA | Can edit | { groups: ["Marketing", "HR"] title: "CMO", status: "inactive" } | inactive |
| 32641 | 19873 | Hannah To | | hannah@aol.com | Brazil | Can view | { groups: ["Sales", "Editor"] title: "Account Executive", status: "active" } | active |

Filter row that was deprecated

We can create this as a series of SQL statements in a dbt file of common table expressions with a final CREATE VIEW query at the bottom:

```
-- drop unused column External_id
WITH t1 AS (
      SELECT Id, Name, Display Name, Email, Location, Type, Info, Status
      FROM dl_table
),

-- Add consistent column Email
t2 AS (
      SELECT Id, Name, Display Name, Email, Location, Type, Info, Status, is_deleted
      FROM t1
      JOIN dl_email
      ON t1.Id = dl_email.Id
),

--Standardize Location column
t3 AS (
      SELECT Id, Name, Display Name, Email,
      CASE WHEN Location = "US" THEN "USA"
            WHEN Location = "Texas" THEN "USA"
            WHEN Location = "Sao Paulo" THEN "Brazil"
            ELSE Location
            END AS "Location",
        Type, Info, Status, is_deleted
      FROM t2
)
```

```
--Make column names and values descriptive for Type
t4 as (
      SELECT Id, Name, Display Name, Email, Location,
      CASE WHEN Type = "1" THEN "Can view"
            WHEN Type = "2" THEN "Can edit"
            WHEN Type = "3" THEN "Can admin"
            END AS "Access Level",
        Info, Status, is_deleted
FROM t3
)

--Parse relevant fields, drop original column for Info
t5 as (
      SELECT Id, Name, Display Name, Email, Location, Access Level,
            CASE WHEN Info = "%active" THEN "active"
                  WHEN Info = "%inactive" THEN "inactive"
                  END AS "Status",
            is_deleted
FROM t4
)

-- filter row that was deprecated from is_deleted, and drop column
t6 as (
      SELECT Id, Name, Display Name, Email, Location, Access Level, Status
      FROM t5
      WHERE is_deleted != True
)

-- create view for Data Warehouse
CREATE VIEW dw_table AS
      SELECT *
      FROM t6
```

We now have a clean view of the original data by creating Views for the Data Warehouse and lightly cleaning and denormalizing the data so that it is easier to query.

| Id | Name | Display Name | Email | Location | Access Level | Status |
|---|---|---|---|---|---|---|
| 21590 | Doug Gonzalez | D Gonzalez | dgonzalez@gmail.com | USA | Can view | active |
| 29448 | Josh | Josh Redman | josh@gmail.com | USA | Can edit | inactive |
| 32641 | Hannah To | | hannah@aol.com | Brazil | Can view | active |

## 9 HEALTHCARE DATA WAREHOUSE

Healthcare data can be used for analytics often categorized into three significant areas which are descriptive, predictive, and prescriptive analytics. This data can be used by many different experts in the healthcare field, ranging from clinicians to healthcare provider administrators and those on the Payer side (claims adjusters, underwriters, provider network managers, and so forth).

Examples of sources of healthcare data:

- Medical records, inpatient, outpatient
- Vital records such as claims
- Financial records such as reimbursements
- Disease and cause of mortality registries and other
- Population Health records such as HEDIS
- Administrative records
- Prescriptions
- Laboratory test
- Monitoring
- Social determinants of health (SDOH)

Within a healthcare data warehouse, descriptive analytics or trend analysis can be done for a patient to determine what has happened to them over a period of time. This data then can be anonymized and aggregated over larger populations and then used for predictive and prescriptive analysis to prescribe solutions for

the individual patient or to understand how medical solutions support people in general. Payers can use healthcare data to determine what rates to set for group policies and individuals as customers and what reimbursement schedules to set for in-network and out-of-network providers. Statistical data from various populations of people or individuals can lead to research advancements, cures, improved preventive measures, and the overall health of the world's population. Payers and providers can use data in an enterprise warehouse to deliver better-valued care while at the same time reducing cost and improving the economic value of the service offered to all.

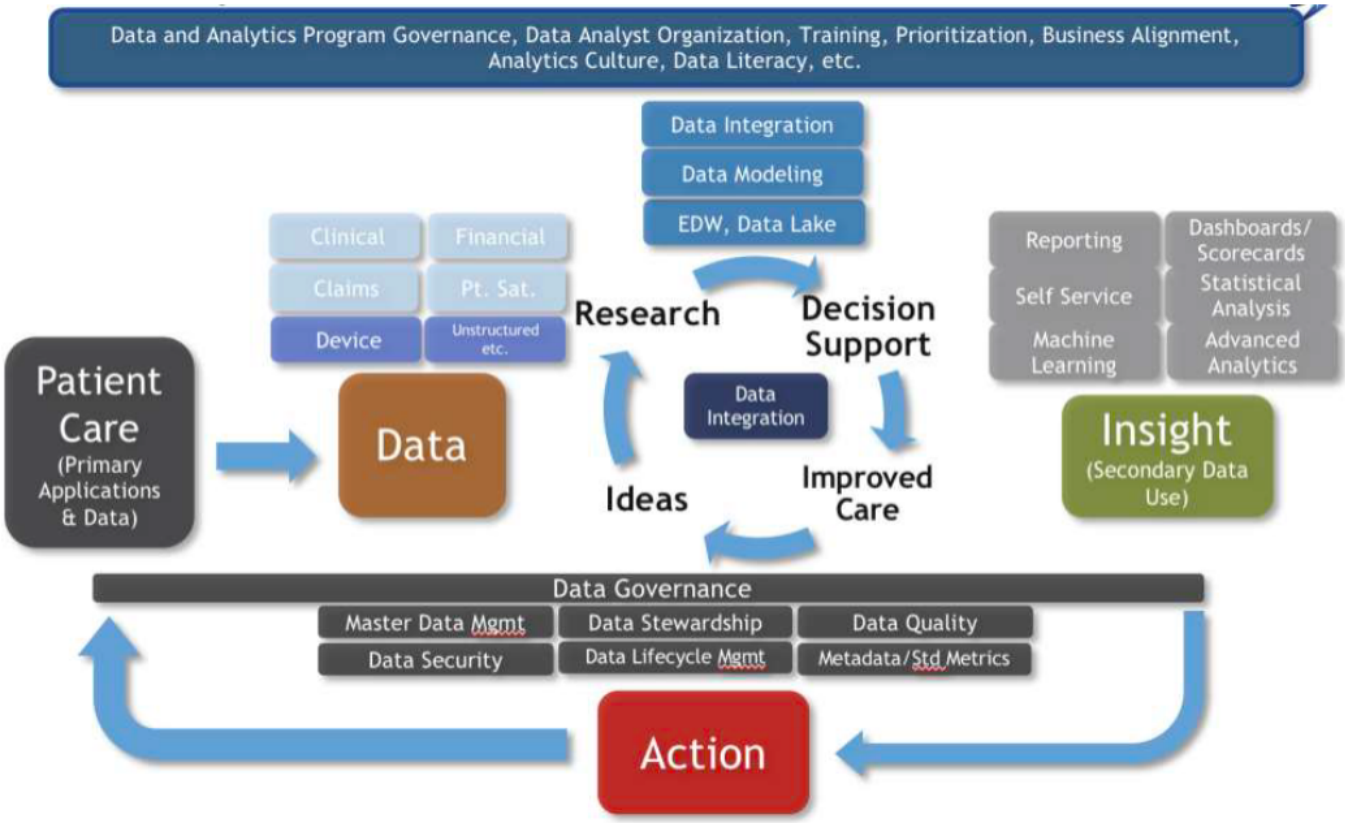



*Figure 2: The healthcare data and analytics process*

## 10 CONCLUSION

A data warehouse is a system that stores data from a company's operational databases as well as external sources. Data warehouse platforms are different from operational databases because they store historical information, making it easier for business leaders to analyze data over a specific period of time. Data warehouse platforms also sort data based on different subject matter, such as customers, products or business activities. Data warehousing improves the speed and efficiency of accessing different data sets and makes it easier for corporate decision-makers to derive insights that will guide the business and marketing strategies that set them

apart from their competitors.This term paper successfully described the basics of a data warehouse, its importance, characteristics, architectural components, advantages, disadvantages, application in healthcare and its implementation in SQL.

## 11 REFERENCES


1. www.javatpoint.com
2. www.sap.com
3. www.tutorialspoint.com
4. www.geeksforgeeks.com
5. www.actian.com
6. www.dataschool.com