# Data Warehousing with IBM Cloud Db2 Warehouse

## Cloud Computing

## Nalaya Thiran Project

**by**


| | |
|---|---|
| **Aashif Ali N** | **211521104001** |
| K.S.Nithish Kumar | 211521104099 |
| K.Prabakar | 211521104108 |
| N.Maheshwaran | 211521104084 |
| P.Praveen Kumar | 211521104114 |

from

Panimalar Institute of Technology

B.E

Computer Science Engineering

# ABSTRACT

Data warehousing is the act of gathering, organising, and managing data from various sources in order to deliver useful business insights and projections to users. It is a relational database management system (RDBMS) design that is used to satisfy the needs of transaction processing systems. It is a database that stores information geared towards meeting decision-making needs. Data warehousing entails data cleaning, integration, and consolidation. The data in the data warehouse is not the same as the data in the operating environment. A data warehouse is built by combining data from disparate sources to support analytical reporting, structured and/or ad hoc queries, and decision making. This project will involve designing, creating, and putting into use a cutting-edge data warehousing system that is specifically customised to the needs of our organisation. It includes a number of crucial elements, including as data storage, governance, security, and user access as well as data extraction, transformation, and loading (ETL).The successful execution of the Data Warehousing Project will equip our organization with a powerful platform for data-driven decision-making, enabling us to unlock valuable insights, optimize operations, and gain a competitive edge in today's dynamic business environment.This project signifies our commitment to leveraging data as a strategic asset to drive innovation and growth.

# INTRODUCTION :

In today's world, organisations are always looking for new methods to leverage the power of data to make wise decisions and gain a competitive edge in today's data-driven world. By offering a centralised location for the storage, processing, and analysis of enormous amounts of data, data warehousing plays a significant part in this process. To ensure that an organization's unique demands and objectives are met, however, the creation of an efficient data warehouse necessitates thorough problem description and design thinking. The process for problem definition and design thinking for data warehousing is described in this document.

# Problem Definition :

Data warehousing is a complex and challenging process, and there are many potential problems that can arise. Some of the most common problems include:

- Identifying Business Goals: State the organization's business goals and objectives clearly. Determine how data warehousing can help you achieve these objectives, such as better decision-making, better customer insights, or better operations.
- Problems with data warehouses: Although Gartner predicts that up to 80% of an organization's data is unstructured, traditional data warehouses can only store clean, highly structured data. They do not support communications frameworks like HL7, JSON, and XML, as well as unstructured data types like text, photos, and IoT data.
- Performance: Data warehouses must be able to efficiently handle high amounts of queries and transactions despite their potential size and complexity.
- Cost: Data warehousing can be expensive, both in terms of hardware and software costs, as well as the costs of implementing and maintaining a data warehouse.

# Design Thinking :

When it comes to data warehousing, design thinking can be used to create a data warehouse that meets the needs of the users. Here are some steps that can be followed:

- Empathise:
  - To understand the needs and pain points of end users, such as analysts, managers, and executives, conduct interviews and workshops with them.
  - To represent various user groups and their demands within the data warehouse, create user personas.
- Define the problem:
  - The next step is to outline the problem that needs to be solved when the users' needs have been understood.
  - Clarifying the data warehouse's objectives, outlining the specific issues that it will be used to solve, and establishing the project's parameters are possible steps in this process.
- Ideate:
  - Identify potential architectures, features, and functionalities for data warehouses that would address the user stories that have been defined.
  - To develop creative ideas, promote collaboration across functional boundaries.
- Prototype:
  - After generating a number of concepts, the following step is to prototype the most promising solutions.
  - This entails developing a functioning model of the solution in order for it to be tested and assessed.
- Test:
  - Test the prototype with users and stakeholders to get feedback.
  - This feedback can be used to refine the solution.

- Data sources:
  - Determine and document all data sources that will be fed into the data warehouse. Databases, external APIs, flat files, CRM systems, and other sources can all be used.
  - Understand these sources' data formats, schemas, and data quality.
- Data integration:
  - Data warehouses frequently integrate data from several sources, which can be challenging and complex.
  - It's possible for data from several sources to have various formats, structures, and degrees of quality.
- ETL Processes:
  - ETL stands for Extract, Transform, and Load. It is a data integration process that combines data from multiple sources into a single, consistent data store that is loaded into a data warehouse or other target system.
  - Data Extraction (E):
    - Create procedures for data extraction from source systems. Setting up ETL (Extract, Transform, Load) pipelines or data integration technologies is frequently required for this.
    - Ensure fast, trustworthy, and minimally disruptive data extraction from source systems.
  - Data Transformation (T):
    - As data goes from the source to the data warehouse, it must be transformed and cleaned.
    - This could include data cleansing, enrichment, aggregation, and format changes.
    - During this stage, apply business rules, calculations, and data quality checks.
  - Data Loading (L):
    - Load the transformed data into the data warehouse. Loading can be done in different ways, such as batch processing or real-time streaming, depending on the organization's needs.
    - Ensure that the loading process is optimized for performance and scalability.