DATA WAREHOUSING WITH IBM CLOUD DB2 WAREHOUSE

INTRODUCTION:

Data Warehouse is a relational database management system (RDBMS) construct to meet the requirement of transaction processing systems. It can be loosely described as any centralized data repository which can be queried for business benefits. It is a group of decision support technologies, targets to enabling the knowledge worker (executive, manager, and analyst) to make superior and higher decisions. So, Data Warehousing support architectures and tool for business executives to systematically organize, understand and use their information to make strategic decisions.

Data and analytics have become indispensable to businesses to stay competitive. Business users rely on reports, dashboards, and analytics tools to extract insights from their data, monitor business performance, and support decision making. Data warehouses power these reports, dashboards, and analytics tools by storing data efficiently to minimize the input and output (I/O) of data and deliver query results quickly to hundreds and thousands of users concurrently.

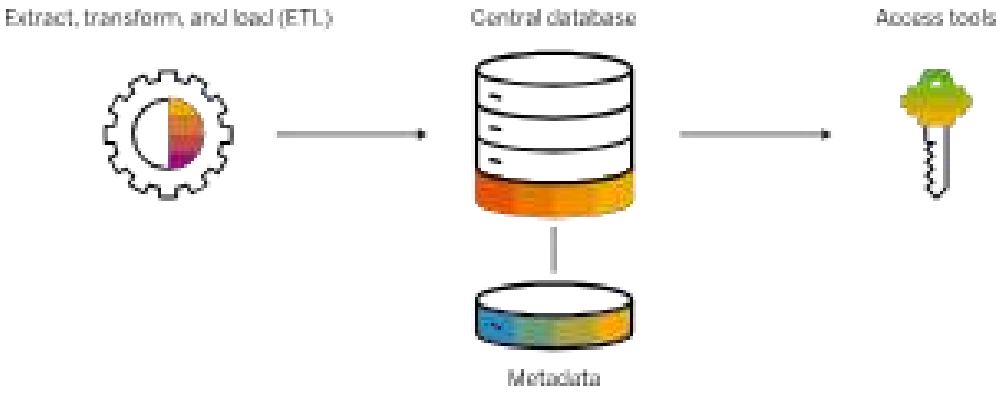
Data Warehouse environment contains an extraction, transportation, and loading (ETL) solution, an online analytical processing (OLAP) engine, customer analysis tools, and other applications that handle the process of gathering information and delivering it to business users. A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources. A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.

OBJECTIVE:

The main objective of this project is to bring together data from various sources to unlock valuable business insights. Perform advanced data integration and transformation effortlessly. Empower data architects to explore, analyze, and deliver actionable data for informed decision-making. The goal of this data warehouse project is to create a trove of historical data that can be retrieved and analyzed to provide useful insight into the organization's operations. A data warehouse is a vital component of business intelligence.

Data Warehouse Structure:

A typical data warehouse has four main components: a central database, ETL (extract, transform, load) tools, metadata, and access tools. All of these components are engineered for speed so that you can get results quickly and analyze data on the fly.



- **Central database:** A database serves as the foundation of your data warehouse. Traditionally, these have been standard relational databases running on premise or in the cloud. But because of Big Data, the need for true, real-time performance, and a drastic reduction in the cost of RAM, in-memory databases are rapidly gaining in popularity.

- **Data integration:** Data is pulled from source systems and modified to align the information for rapid analytical consumption using a variety of data integration approaches such as ETL (extract, transform, load) and ELT as well as real-time data replication, bulk-load processing, data transformation, and data quality and enrichment services.
- **Metadata:** Metadata is data about your data. It specifies the source, usage, values, and other features of the data sets in your data warehouse. There is business metadata, which adds context to your data, and technical metadata, which describes how to access data – including where it resides and how it is structured.
- **Data warehouse access tools:** Access tools allow users to interact with the data in your data warehouse. Examples of access tools include: query and reporting tools, application development tools, data mining tools, and OLAP tools.

DATA INTEGRATION:

Data integration is the process of combining and harmonizing diverse datasets from various sources, formats, and structures into a unified and coherent format, making it accessible and usable for analysis, reporting, and decision-making. It involves extracting, transforming, and loading (ETL) data from disparate sources, ensuring data quality and consistency, and creating a consolidated, integrated data repository that provides a holistic view of information, enabling organizations to gain valuable insights and improve their overall data-driven operations and strategies.

- **Data Source Identification** : Identify all relevant data sources, including databases, APIs, files, and streams and Understand the types of data (structured, semi-structured, unstructured) each source contains.

- **Data Quality Assessment:** Assess data source quality, considering factors like accuracy, completeness, consistency, and timeliness and Prioritize data sources with high-quality and reliable data.
- **ETL Process Design**: Develop ETL processes for data extraction, transformation, and loading. Implement incremental extraction and data validation to ensure data quality.
- **Data Modeling and Storage**: Design a data model for the data warehouse, defining tables, relationships, and schema. Optimize data storage with appropriate indexing and partitioning strategies.

ETL Processes:

ETL, which stands for Extract, Transform, Load, is a crucial data integration process used in modern organizations to move and manipulate data from various sources into a target database or data warehouse.

- **Extraction**: In this initial step, data is extracted from diverse sources such as databases, flat files, APIs, or external systems. The extraction process may involve both structured and unstructured data. Data extraction methods can be incremental or full, depending on the need. Extracted data is often in its raw, unprocessed form, and it may require cleansing and validation to ensure its accuracy and integrity.
- **Transformation**: After extraction, the data undergoes a series of transformations to prepare it for analysis or reporting. Transformation tasks can include data cleaning, validation, enrichment, and aggregation. This step may also involve data normalization, where data is standardized to a common format, and data enrichment, where additional information is added to enhance its value. Transformations are typically defined by business rules and logic to ensure that the data meets the desired quality and format standards.

- **Loading**: Once the data is extracted and transformed, it is loaded into the target database or data warehouse. Loading can be done in different ways, including batch processing or real-time streaming, depending on the requirements of the organization. Loading data into a centralized repository allows for efficient querying and reporting, making it accessible for various analytics and business intelligence applications.

Data Exploration:

Data exploration is a crucial step in the data analysis process. It involves examining and summarizing the main characteristics of a dataset to better understand its structure, identify patterns, detect outliers, and gather insights that can guide further analysis.

- **Statistical Analysis and Visualization Queries:** Data architects should design SQL queries or use data analysis tools to perform descriptive statistics on the dataset. These queries can help identify key summary statistics such as mean, median, mode, standard deviation, and percentiles. Additionally, data architects can generate various visualization queries to create charts, graphs, and plots that provide a visual representation of the data's distribution, trends, and outliers.
- **Data Profiling Queries:** Data profiling queries are essential for understanding the quality and structure of the data. Data architects can design SQL queries to count missing values, identify duplicates, and assess data consistency. Queries for calculating data skewness, kurtosis, and cardinality of columns can help uncover data anomalies and guide data cleaning efforts. Data profiling queries often serve as a foundation for data cleansing and transformation tasks.

- **Exploratory Data Analysis (EDA):** Data architects can employ EDA techniques to gain deeper insights into the data. This involves running queries to calculate correlation coefficients between different variables, generating scatter plots and heatmaps to visualize relationships, and performing hypothesis testing to assess statistical significance.

empowering data architects to explore and analyze data involves designing queries for statistical analysis, data profiling, and exploratory data analysis. These techniques help uncover insights, assess data quality, and lay the groundwork for effective data modeling and decision-making.

Actionable Insights:

Actionable insights refer to meaningful and practical information extracted from data or analysis that can be used to make informed decisions or take specific actions. These insights are valuable because they provide guidance on how to improve processes, solve problems, or achieve objectives.

- **Data Collection and Analysis:** Start by collecting relevant data from various sources and conducting thorough analysis. Ensure the data is accurate, up-to-date, and aligned with your objectives. Use data analysis tools and techniques to identify patterns, trends, and anomalies.
- **Clear and Specific Recommendations:** Transform your analysis into clear and specific recommendations. Avoid vague or generic insights. Clearly state what actions should be taken, why they are necessary, and the expected impact on the desired outcomes.
- **Communication and Visualization:** Present your insights in a concise and easily understandable manner. Use data visualization techniques like charts, graphs, or dashboards to make complex information more accessible. Tailor

your communication to the audience, whether it's executives, managers, or frontline staff.

- **Feedback and Iteration:** Establish a feedback loop to track the implementation of recommended actions and measure their effectiveness. Continuously assess whether the insights are leading to the desired outcomes, and be prepared to iterate and adjust strategies as needed.

Applications:

A robust data warehouse can have various applications across different industries and functions.

- **Business Intelligence and Reporting**: A robust data warehouse serves as a centralized repository for all organizational data, enabling easy access to historical and current data. This data can be used for generating business intelligence reports and dashboards, helping organizations make informed decisions based on data-driven insights.
- **Data Integration**: Data warehouses can integrate data from various sources, including databases, spreadsheets, and external data feeds. This integration simplifies data management and provides a unified view of the organization's data, making it easier to analyze and extract valuable information.
- **Advanced Analytics and Data Mining**: Data warehouses provide a structured and optimized environment for advanced analytics, data mining, and machine learning. Businesses can use this capability to uncover patterns, trends, and correlations in their data, leading to improved forecasting, customer segmentation, and anomaly detection.
- **Historical Data Storage**: Robust data warehouses store historical data, allowing organizations to track performance over time and conduct historical

analysis. This is valuable for compliance, auditing, and understanding long-term trends and patterns in data.

- **Operational Decision Support**: Data warehouses support operational decision-making by providing quick access to up-to-date data. This is especially important in industries where real-time decision-making is critical, such as e-commerce, supply chain management, and financial services.
- **Customer Relationship Management (CRM):** Data warehouses can consolidate customer data from various sources to provide a comprehensive view of customer behavior and preferences. CRM teams can use data warehouses to segment customers based on demographics, purchase history, and behavior for targeted marketing efforts.

CONCLUSION:

In conclusion, the robust data warehouse paper has successfully addressed our organization's data management needs. It has enhanced data accessibility, reliability, and performance, enabling better-informed decision-making. The project's scalability ensures future growth and adaptability to evolving data requirements. With improved data quality and integration, our organization is better positioned to achieve its strategic goals. This data warehouse project represents a vital asset in our quest for data-driven excellence.

Presented by,

A. Pavithra

K. Rajeshwari

P. Bhuvaneshwari

N. Shamshunnafiya