IBM NAAN MUDHALVAN

CLOUD APPLICATION DEVELOPMENT

PROJECT

Data Warehousing with IBM Cloud Db2 Warehouse

Phase 1:

Problem Definition:

The project involves designing and setting up a robust data warehouse using IBM Cloud Db2 Warehouse. The objective is to bring together data from various sources, perform advanced data integration and transformation, and provide data architects with the tools to explore, analyze, and deliver actionable data for informed decision-making. This project encompasses defining the data warehouse structure, integrating data sources, performing ETL (Extract, Transform, Load) processes, and enabling data analysis.

Design Thinking:

Data Warehouse Architecture:

The single tier Data Warehouse architecture is composed of a single hardware layer. This hardware layer is composed of a single hardware layer. There are three approaches to creating a data warehouse layer: Single tier, two-tier, and three-tier.

Single-tier architecture:

A single-layer structure aimed at keeping data space minimal. This structure is rarely used in real life.

Two-tier architecture:

Data warehouse is the aggregation of data in a format that is easy to transform and load into a database. Data warehouses can be implemented in a number of different ways, and it is important to pick the right one for your business needs. The most important thing to consider is scalability. If you want to store large amounts of data in a small amount of space, then you should consider using a data warehouse.

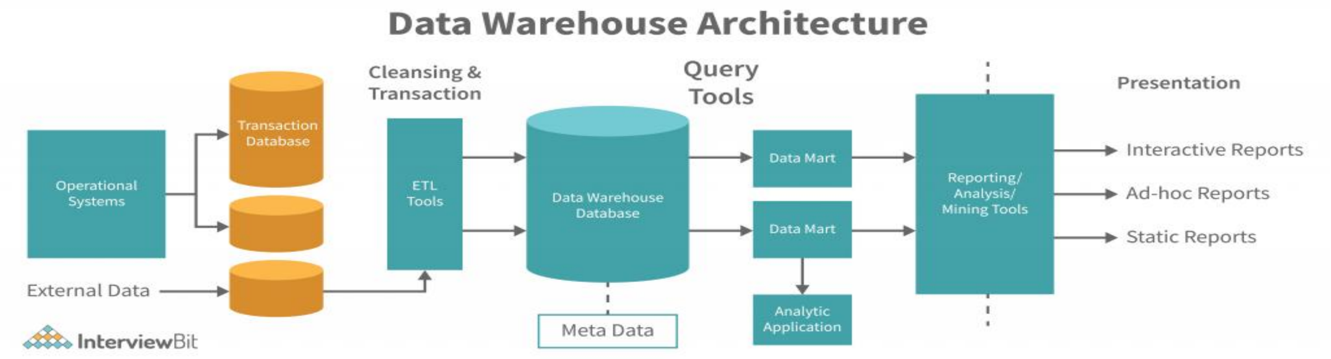
Three-Tier Data Warehouse Architecture:

The Top, Middle, and Bottom Tiers of this Architecture of Data Warehouse are collectively referred to as the Top Tier.

1. The bottom tier of the Datawarehouse is a relational database system. This database system typically contains a relational database system. Back-end tools clean, transform, and load data into this layer.

2. A middle tier OLAP server is either ROLAP or MOLAP-based. It abstracts OLAP from the end user by serving as a middle tier OLAP server. Data warehouses that facilitate end-user interaction with the database and middle tier OLAP servers that abstract OLAP from the end user are known as middle tier OLAP servers.

3. The front-end client layer of the top-tier is important because it is the first point of interaction with the data. It is where data is presented to the end user, and decisions are made with the data. The front-end client layer of top-tier must work with real-time data and must be able to process data quickly. It is also important to work with data that is in a format that top-tier can understand and use. Typically, top-tier data is in a relational database format, but it could be a file or a stream. Top-tier data must be well-structured, must be validated, and must be structured in a way that allows for easier data profiling and analytics



Components of Data Warehouse

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.

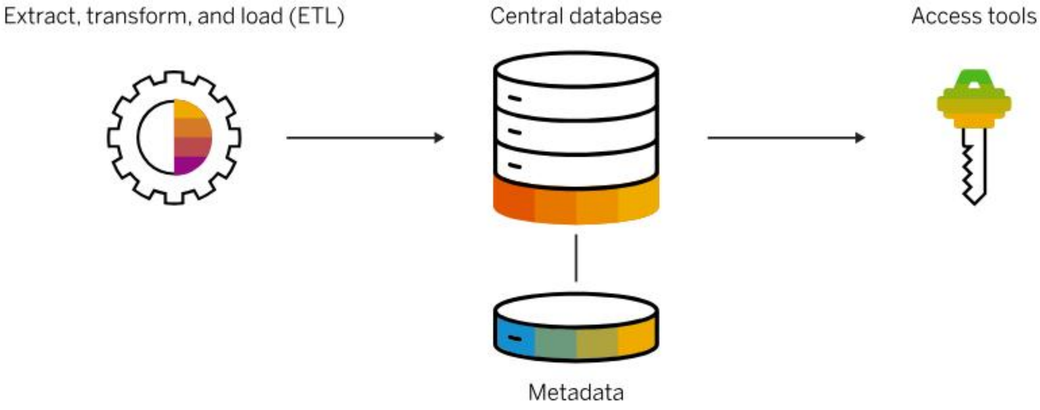


Diagram showing the components of a data warehouse.

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:

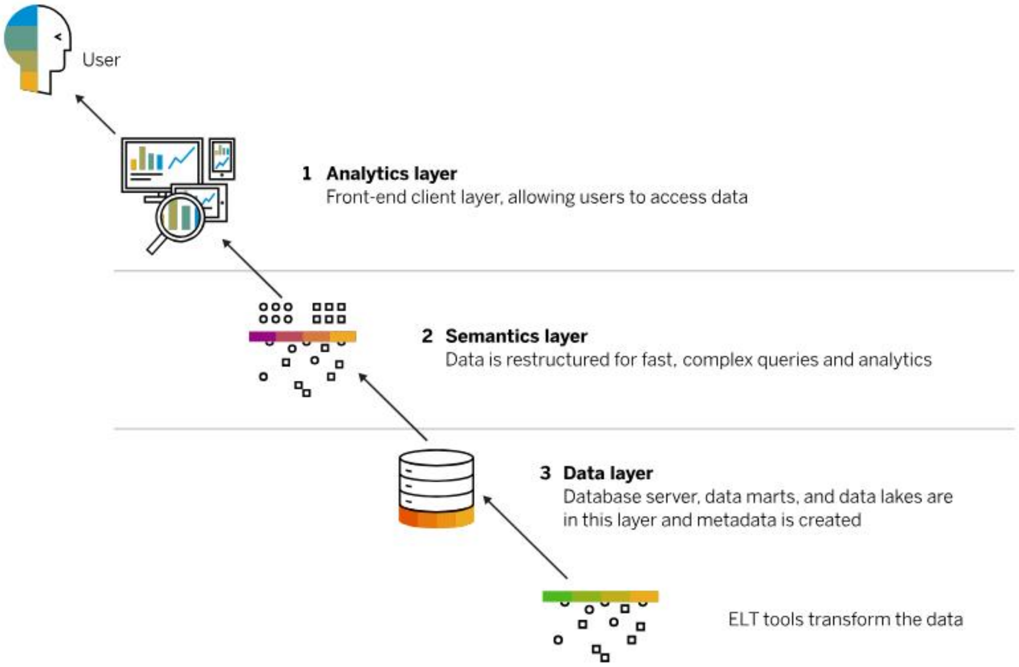
Data is pulled from source systems and modified to align the information for rapid analytical consumption using a variety of data integration approaches such as ETL (extract, transform, load) and ELT as well as real-time data replication, bulk-load processing, data transformation, and data quality and 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.


User
1 Analytics layer
Front-end client layer, allowing users to access data
2 Semantics layer
Data is restructured for fast, complex queries and analytics
3 Data layer
Database server, data marts, and data lakes are in this layer and metadata is created
ELT tools transform the data

Data Warehouse Integration

Data warehouse integration combines data from several sources into a single, unified warehouse. The data warehouse can be accessed by any department within an organization, and the data can be easily structured into spreadsheets or tables for research and analysis purposes.

Think of your data integration strategy as the maestro of an orchestra. Each instrument (or data source in our case) plays its part, but the magic truly happens when they all come together in harmony, creating a symphony of actionable business insights.

- **Enhances data accuracy and consistency:** A robust data integration strategy ensures all departments use the same data, promoting cohesive decision-making across the organization.
- **Increases operational efficiency:** Automating data collection and aggregation allows your team to focus on value-adding tasks like data analysis and strategy development.
- **Enables faster, informed decisions:** With a comprehensive view of all your data, you can quickly identify trends, forecast outcomes, and gain a competitive edge.
- **Improves customer experience:** Integrating data from various touchpoints provides a holistic view of your customer's journey, helping to tailor offerings, enhance satisfaction, and foster loyalty.
- **Lays foundation for advanced technologies:** A strategic approach to data integration prepares your business for utilizing AI and machine learning, which can generate accurate predictions, provide deeper insights, and drive growth.

The Data Integration Process: Step-by-Step

1. Data Extraction: Gathering Insights from Various Sources

2. Data Transformation: Shaping Raw Data into Valuable Insights

3. Data Delivery: Establishing Trust Between Systems

4. Data Loading: The Bridge to Seamless Data Flow

5. Data Validation: Ensuring Accuracy and Reliabiliy

6. Process optimization: Iterative improvements for sustained success

Implementing Data Integration Strategy

Analyze Data Sources:

Begin by listing all data sources you plan to use. This includes internal databases, external APIs, SaaS applications, and other third-party sources. Ensure these sources are compatible with each other and can scale with your growth.

Set Clear Goals and Metrics:

Identify the specific purpose of the data integration. Are you looking to improve customer experience, automate operations, or gain a competitive edge? Setting clear goals will help you stay focused on what's important and choose the right approach for your needs.

Design Data Model:

This involves creating a detailed data model to define the structure and flow of data between different sources. You need to consider factors like data organization, data latency, and accessibility.

Choose the Right Approach and Platform:

Different data integration approaches suit different business needs. It's important to choose an approach that aligns with your business goals. The platform should be scalable, secure, and easy to use.

Simplicity is Key:

Minimizing complexities in data integration techniques can help speed up the process and reduce errors.

Ensure Data Quality and Security:

Data quality management and enhanced security measures are crucial in big data integration.

Conduct Thorough Testing:

Testing the data integration process helps identify and rectify issues before they impact business operations.

Embrace Flexibility and Adaptability:

Dynamic data landscapes require flexibility and adaptability. Be prepared to adjust your strategy as business needs, technologies, and data sources evolve.

Monitor and Optimize:

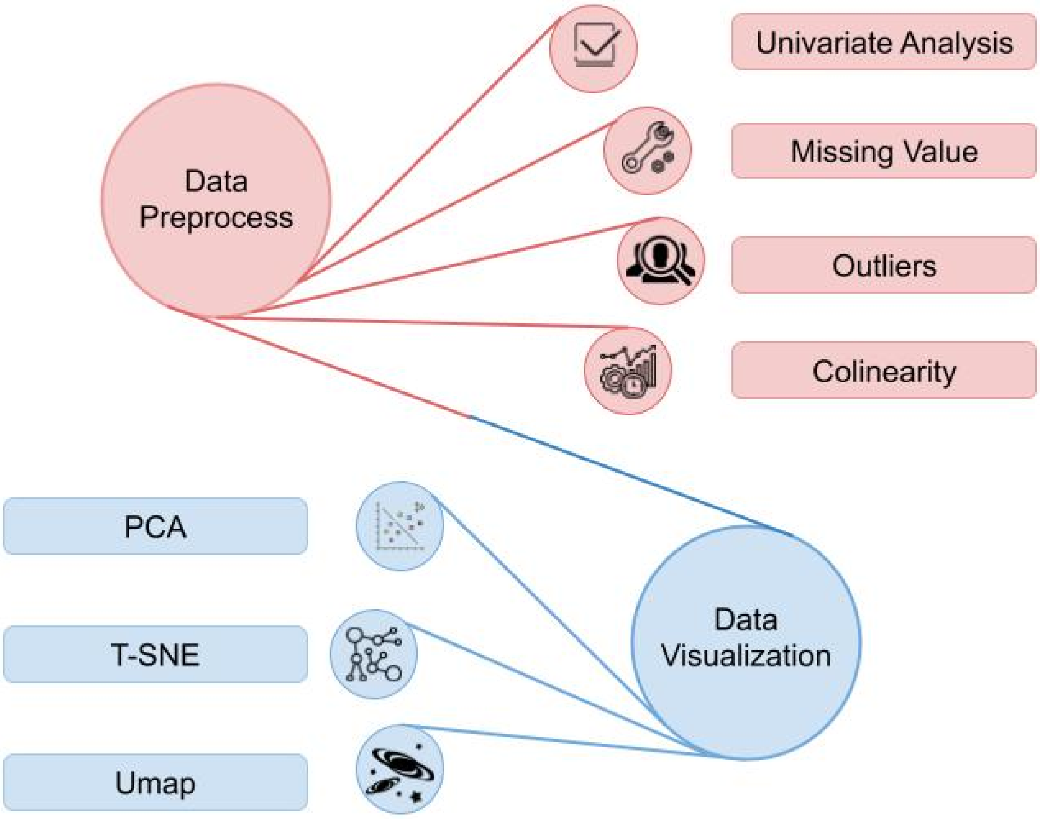
Constant monitoring will help identify potential issues, allowing you to optimize your data integration strategy for better performance.

Remember, a successful data integration strategy requires careful planning, execution, and ongoing management. It's not just about the technology but also how it's used to meet business objectives.

ETL Processes:
Extract, transform, and load (ETL) works by moving data from the source system to the destination system at periodic intervals. The ETL process works in three steps:

1. Extract the relevant data from the source database
2. Transform the data so that it is better suited for analytics
3. Load the data into the target database

Data Exploration
Data exploration, also known as exploratory data analysis (EDA), is a process where users look at and understand their data with statistical and visualization methods. This step helps identifying patterns and problems in the dataset, as well as deciding which model or algorithm to use in subsequent steps.



Actionable insights:

Actionable insights are conclusions drawn from data that can be turned directly into an action or a response. The data informing the insights can be structured or unstructured, quantitative or qualitative.