Data Warehousing with IBM Cloud Db2 Warehouse

Phase 5 : Project Documentation and Submission

Project Objectives :

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 Process :

1. Data Warehouse Structure: Define the schema and structure of the data warehouse to accommodate various data sources.
2. Data Integration: Identify data sources and design a strategy to integrate data seamlessly into the data warehouse.
3. ETL Processes: Plan and implement ETL processes to extract, transform, and load data into the warehouse.
4. Data Exploration: Design queries and analysis techniques to empower data architects to explore and analyze data.
5. Actionable Insights: Focus on delivering actionable insights by enabling informed decision-making based on data.

Development Phases :

Planning and Requirements Gathering :

This phase involves identifying business requirements, understanding data sources, and defining the scope and objectives of the data warehouse.

Data Extraction :

Data is extracted from various source systems, which could include databases, spreadsheets, external data feeds, and more.

Data Transformation :

Implement tools and interfaces for users to query and generate reports from the data warehouse. This may involve business intelligence tools.

Data Access and Security :

Extracted data is transformed and cleaned to ensure consistency and quality. This may involve data cleansing, aggregation, and integration.

Data Loading :

Transformed data is loaded into the data warehouse. There are typically two methods: batch loading (regular updates) and real-time loading (continuous updates).

Data Modeling :

In this phase, the data is organized into a structure that makes it accessible for querying and reporting. Common data models include star schema and snowflake schema.

Metadata Management :

Metadata, which describes the data in the data warehouse, is crucial for understanding the data's meaning and lineage. It should be managed effectively.

Query and Reporting Tools :

Define access controls and security measures to protect the data and ensure that users have the appropriate permissions.

Testing and Quality Assurance :

Rigorously test the data warehouse to ensure that data is accurate, queries perform well, and the system is stable.

User training and Documentation :

Train end-users and provide documentation to help them make the most of the data warehouse.

Monitoring and Optimization :

Continuously monitor the data warehouse's performance and make improvements as needed, such as adding more data sources or refining queries.

Data Warehouse Structure :

- Star Schema
- Snowflake Schema
- Galaxy Schema (Constellation Schema)
- Factless fact table
- Bridge Table
- Hybrid Schema
- Data vault

Data Integration Strategies :

Batch ETL (Extract , Transform , Load) :

- This is one of the most traditional data integration strategies.
- Data is extracted from source systems in batches, transformed to meet the data warehouse's structure and quality standards, and then loaded into the data warehouse.
- ETL tools are often used to automate this process.

Real Time ETL :

- In real-time ETL, data is extracted, transformed, and loaded into the data warehouse on a near real-time basis.
- This strategy is suitable for organizations that require up-to-the-minute data for their reporting and analytics.

Change Data Capture:

- CDC is a technique that identifies and captures changes in source system data since the last extraction.
- It minimizes the amount of data transferred and processed during ETL, making it efficient for large volumes of data.

Data Federation :

- Data federation integrates data from various sources virtually, without physically moving or storing the data in the data warehouse.
- It provides real-time access to data without the need for a dedicated ETL process.

Data Replication :

- Data replication involves copying data from source systems to the data warehouse or data marts.
- This strategy is useful for scenarios where low-latency data access is essential, but it may require additional hardware and storage.

Data Virtualization :

- Data virtualization provides a layer of abstraction that allows data to be accessed from various sources as if it were a single source.
- It simplifies data access and can be especially helpful for quickly integrating new data sources.

Master Data Management :

- MDM focuses on managing the critical master data entities of an organization, such as customer data or product data.
- It ensures that master data is consistent and synchronized across the organization, including the data warehouse.

Data Consolidation :

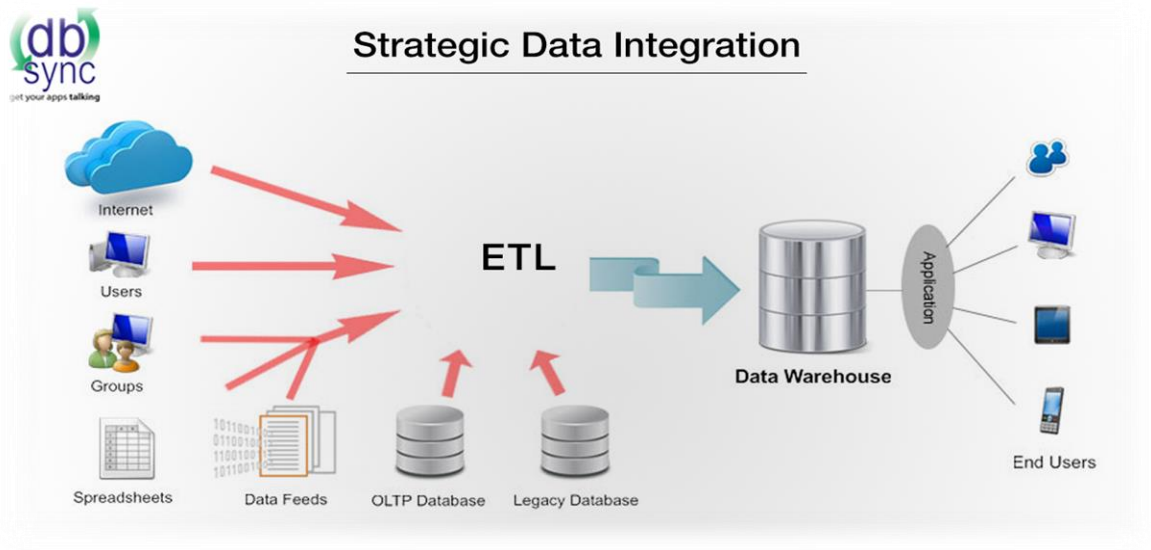
- Data consolidation involves merging multiple data sources into a single source or consolidating data marts into a centralized data warehouse.
- It simplifies data management and ensures a single source of truth.

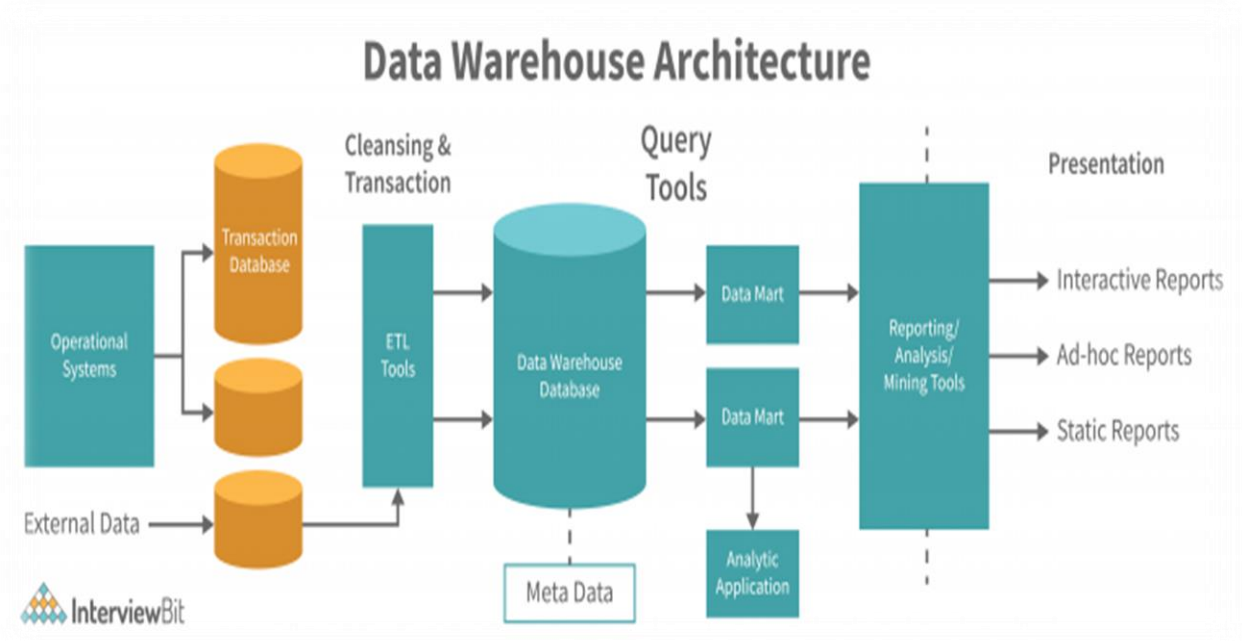
Data Quality and Profiling :

- Data integration strategies often include data quality and profiling processes to ensure that data is accurate, complete, and consistent.
- Data profiling helps identify data quality issues that need to be addressed during the ETL process.

API Integration :

- Application Programming Interfaces (APIs) can be used to integrate data from various sources, especially cloud-based and web-based services.
- This strategy is common in modern data integration processes.


Strategic Data Integration
Internet
Users
Groups
Spreadsheets
Data Feeds
ETL
OLTP Database
Legacy Database
Data Warehouse
Application
End Users


Data Warehouse Architecture
Cleansing & Transaction
Query Tools
Presentation
Operational Systems
Transaction Database
External Data
ETL Tools
Data Warehouse Database
Meta Data
Data Mart
Data Mart
Analytic Application
Reporting/ Analysis/ Mining Tools
Interactive Reports
Ad-hoc Reports
Static Reports
InterviewBit

ETL Processes :

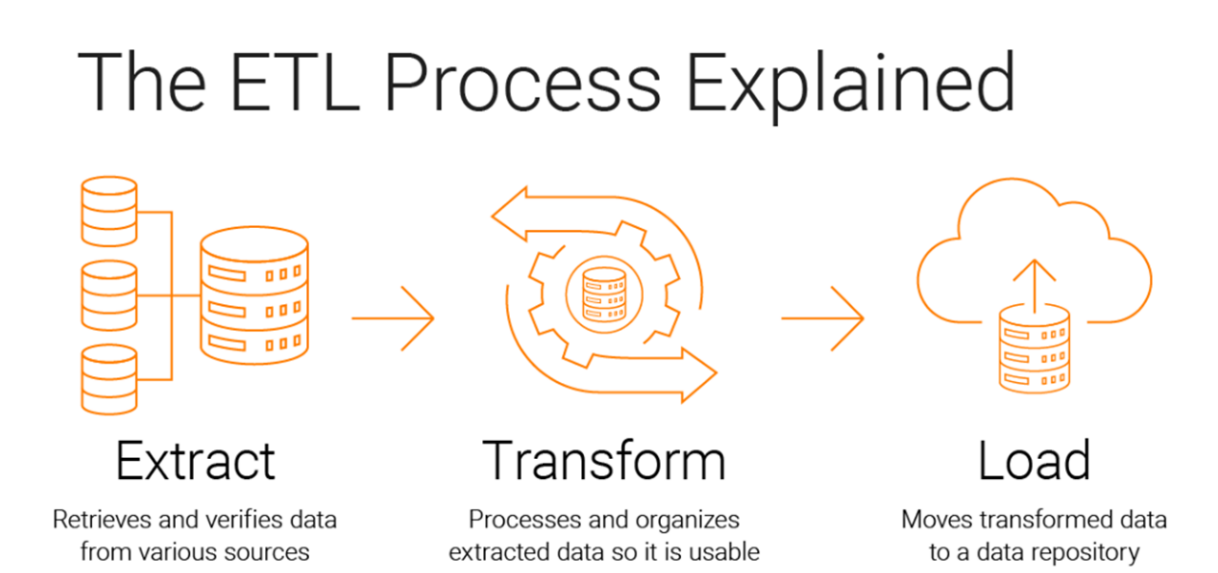
ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse.

Extract:
The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.

Transform:
In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.

Load:
After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.

The ETL process is an iterative process that is repeated as new data is added to the warehouse. The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date. It also helps to ensure that the data is in the format required for data mining and reporting.



Data Exploration Techniques :

SQL Queries:

SQL (Structured Query Language) is commonly used to extract and explore data in a data warehouse. You can write SQL queries to retrieve, filter, and aggregate data to gain insights.

Data Visualization:

Data visualization tools like Tableau, Power BI, and others allow you to create charts, graphs, and dashboards to visualize data patterns and trends. Visualization makes it easier to understand complex data.

Pivot Tables:

In tools like Microsoft Excel or data analysis platforms, pivot tables can help summarize and explore data quickly. You can pivot data by different dimensions to view it from various angles.

Descriptive Statistics:

Calculating basic statistics like mean, median, mode, standard deviation, and percentiles can provide insights into the central tendencies and variability of your data.

Histograms:

Histograms are useful for visualizing the distribution of data values and identifying data skewness or patterns.

Box Plots:

Box plots display the distribution of data and help identify outliers and the overall spread of data.

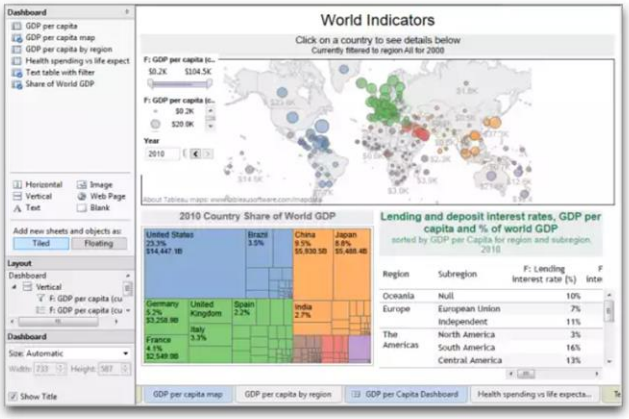
Correlation Analysis:

Analyzing correlations between different attributes in your data can reveal relationships and dependencies between variables.

Clustering and Segmentation:

Techniques like clustering and segmentation can group similar data points together, making it easier to analyze data with common characteristics.

data exploration



data presentation