Data Warehousing with IBM cloud

Db2 Warehouse

Problem Definition

The project involves designing and setting up a robust data warehouse using IBM cloud Db2 warehouse. The objective is 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.

Overview of IBM Db2 Warehouse on cloud

IBM Db2 Warehouse on Cloud is a fully-managed, elastic cloud data warehouse that delivers independent scaling of storage and compute. It delivers a highly optimized columnar data store, actionable compression data store, actionable compression, and in-memory processing to supercharge your analytics and machine learning workloads.

Data warehouse

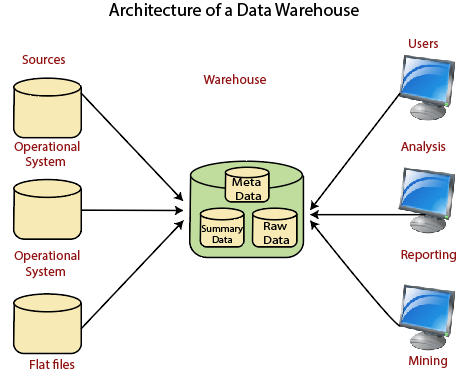
A data warehouse, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning. A data warehouse system enables an organization to run powerful analytics on huge volumes (petabytes and petabytes) of historical data in ways that a standard database cannot.

Data warehousing systems have been a part of business intelligence (BI) solutions for over three decades, a data warehouse might be hosted on a dedicated appliance or in the cloud, and most data warehouses have added analytics capabilities and data visualization and presentation tools

Data warehouse architecture

Data warehouses have a three-tier architecture, which consists of a:

* **Bottom tier:**The bottom tier consists of a data warehouse server, usually a relational database system, which collects, cleanses, and transforms data from multiple data sources through a process known as Extract, Transform, and Load (ETL) or a process known as Extract, Load, and Transform (ELT).
* **Middle tier:**The middle tier consists of an OLAP (i.e. online analytical processing) server which enables fast query speeds. Three types of OLAP models can be used in this tier, which are known as ROLAP, MOLAP and HOLAP. The type of OLAP model used is dependent on the type of database system that exists.
* **Top tier:**The top tier is represented by some kind of front-end user interface or reporting tool, which enables end users to conduct ad-hoc data analysis on their business data.



Data integration

**Data integration is the process of combining data from different sources into a single, unified view. Integration begins with the ingestion process, and includes steps such as cleansing, ETL mapping, and transformation.**

A data integration strategy is a documented plan for how enterprise data sets will be combined and analyzed across the company.

It outlines the different sources of data and how the information will be moved and analyzed. When getting started with a modern data stack, your overall data strategy encompasses data management, data governance ,culture, and architecture. The data integration strategy provides a clear path to breaking down data silos that might exist within your company.

ETL process

1. 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. The process of ETL can be broken down into the following three stages:
2. **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.
3. **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.
4. **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.
5. 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.

### D*ata Exploration*

Data exploration tools make data analysis easier to present and understand through interactive, visual elements, making it easier to share and communicate key insights. Data exploration tools include data visualization software and business intelligence platforms, such as Microsoft Power BI and tableau. Data exploration involves using data visualizations to examine the data at a high level.

**Actionable insights**

Actionable insights are specific, data-driven findings or information that provide clear and practical guidance on how to make informed decisions and take concrete actions. Management may use big data sets with actionable insights to create efficient data-driven activities.

**The steps involves:**

* [Step 1: Strategy](https://www.sisense.com/blog/5-steps-to-data-driven-business-decisions/#strategy)
* [Step 2: Identify key areas](https://www.sisense.com/blog/5-steps-to-data-driven-business-decisions/#identify)
* [Step 3: Data targeting](https://www.sisense.com/blog/5-steps-to-data-driven-business-decisions/#target)
* [Step 4: Collecting and analyzing](https://www.sisense.com/blog/5-steps-to-data-driven-business-decisions/#collect)
* [Step 5: Action Items](https://www.sisense.com/blog/5-steps-to-data-driven-business-decisions/#action)