DATA WAREHOUSING WITH IBM CLOUD DB2 WAREHOUSE

PHASE 5

PROJECT OBJECTIVES :

start building the data warehouse using IBM cloud Db warehouse define the schema and structure of the data warehouse tables Identify data sources (e.g., CSV files, databases)and design a strategy to integrate them into the data warehouse. Implement ETL processes to extract, transform, and load data into the data warehouse.

INTRODUCTION :

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.

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 :

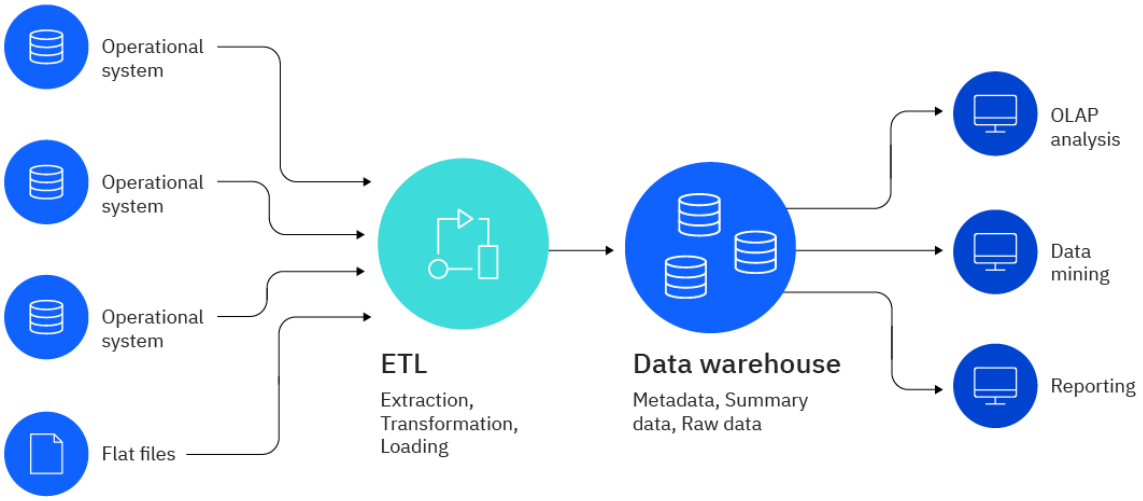
A typical data warehouse has four main components: a central database, ETL (extract, transform, load) tools, metadata, and access tools. All of these components are engineered for speed so that you can get results quickly and analyze data on the fly.

o Central database

o Data integration

o Metadata

o Data warehouse access 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.

COMPONENTS OF DATAWAREHOUSE ARCHITECTURE:

1.ETL:

When database analysts want to move data from a data source into their data warehouse, this is the process they use. In short, ETL converts data into a usable format so that once it's in the data warehouse, it can be analyzed/queried/etc.

2.Metadata:

Metadata is data about data. Basically, it describes all of the data that's stored in a system to make it searchable. Some examples of metadata include authors, dates or locations of an article, create date of a file, the size of a file, etc. Think of it like the titles of a column in a spreadsheet. Metadata allows you to organize your data to make it usable, so you can analyze it to create dashboards and reports.

3.SQL query processing:

SQL is the de facto standard language for querying your data. This is the language that analysts use to pull out insights from their data stored in the data warehouse. Typically data warehouses have proprietary SQL query processing technologies tightly coupled with the compute. This allows for very high performance when it comes to your analytics. One thing to note, however, is that the cost of a data warehouse can start getting expensive the more data and SQL compute resources you have.

EXAMPLE:

```
SELECT c.customer_name, SUM(o.total_amount) AS total_spent

FROM customers c JOIN orders o ON c.customer_id = o.customer_id

GROUP BY c.customer_name

ORDER BY total_spent DESC;
```

4.Data layer:

The data layer is the access layer that allows users to actually get to the data. This is typically where you'd find a data mart. This layer partitions segments of your data out depending on who you want to give access to, so you can get very granular across your organization. For instance, you may not want to give your sales team access to your HR team's data, and vice versa.

IMPLEMENTATION OF ETL PROCESS:

1. Extract Data:

To extract data, you can use various libraries in Python. In this example, we'll use the pandas library to read data from a CSV file.

2. Transform Data:

Perform data transformations as needed. For this example, let's convert a column to uppercase:

3. Load Data into Db2 Warehouse:

To load data into Db2 Warehouse, you can use Python libraries like ibm_db or SQLAlchemy. Make sure to replace the placeholders with your actual database connection details.

DATA EXPLORATION USING SQL QUERIES:

Now, let's enable data architects to explore and analyze data within Db2 Warehouse using SQL queries and analysis techniques.

1. Connect to Db2 Warehouse:

Data architects can use Python and libraries like ibm_db to connect to Db2 Warehouse.

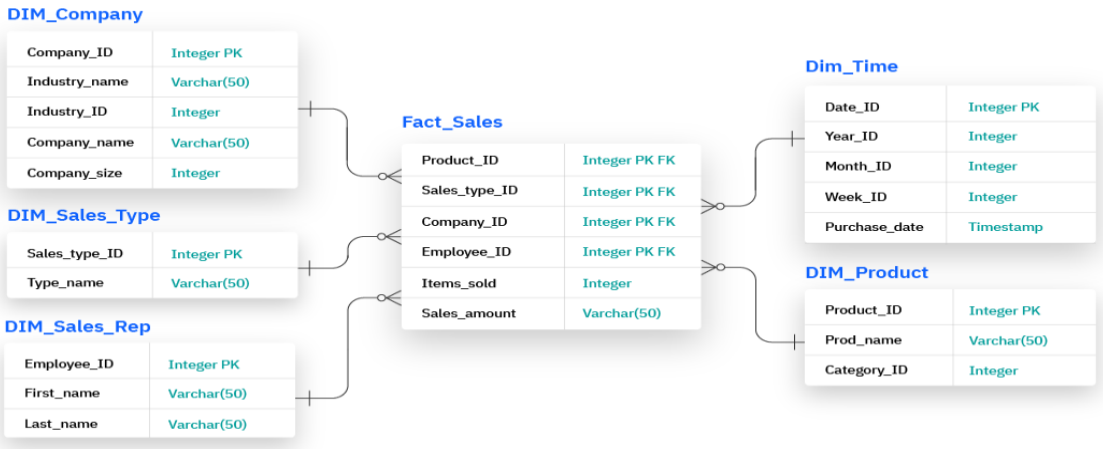
2. Run SQL Queries:

Data architects can execute SQL queries to analyze the data. For example, to retrieve data from a table and calculate the average of a column:

3. Generate Analysis Outputs:

Data architects can use Python libraries like pandas, matplotlib, or seaborn to generate visualizations and insights from the data. For example, creating a bar chart of the results

SCHEMAS IN DATA WAREHOUSE:



EXAMPLE DATASET:

1	ID	LOTAREA	BLDGTYPE	HOUSESTYLE	OVERALLCOND	YEARBUILT	ROOFSTYLE	EXTERCOND	FOUNDATION	BSMTCOND	HEATING	HEATINGQC	CENTRALAIR	ELECTRICAL	FUL
2	1	8450	1Fam	2Story	5	2003	Gable	TA	PConc	TA	GasA	Ex	Y	SBrkr	2
3	2	9600	1Fam	1Story	8	1976	Gable	TA	CBlock	TA	GasA	Ex	Y	SBrkr	2
4	3	11250	1Fam	2Story	5	2001	Gable	TA	PConc	TA	GasA	Ex	Y	SBrkr	2
5	4	9550	1Fam	2Story	5	1915	Gable	TA	BrkTil	Gd	GasA	Gd	Y	SBrkr	1
6	5	14260	1Fam	2Story	5	2000	Gable	TA	PConc	TA	GasA	Ex	Y	SBrkr	2
7	6	14115	1Fam	1.5Fin	5	1993	Gable	TA	Wood	TA	GasA	Ex	Y	SBrkr	1
8	7	10084	1Fam	1Story	5	2004	Gable	TA	PConc	TA	GasA	Ex	Y	SBrkr	2
9	8	10382	1Fam	2Story	6	1973	Gable	TA	CBlock	TA	GasA	Ex	Y	SBrkr	2
10	9	6120	1Fam	1.5Fin	5	1931	Gable	TA	BrkTil	TA	GasA	Gd	Y	FuseF	2
11	10	7420	2fmCon	1.5Unf	6	1939	Gable	TA	BrkTil	TA	GasA	Ex	Y	SBrkr	1
12	11	11200	1Fam	1Story	5	1965	Hip	TA	CBlock	TA	GasA	Ex	Y	SBrkr	1
13	12	11924	1Fam	2Story	5	2005	Hip	TA	PConc	TA	GasA	Ex	Y	SBrkr	3
14	13	12968	1Fam	1Story	6	1962	Hip	TA	CBlock	TA	GasA	TA	Y	SBrkr	1
15	14	10652	1Fam	1Story	5	2006	Gable	TA	PConc	TA	GasA	Ex	Y	SBrkr	2
16	15	10920	1Fam	1Story	5	1960	Hip	TA	CBlock	TA	GasA	TA	Y	SBrkr	1
17	16	6120	1Fam	1.5Unf	8	1929	Gable	TA	BrkTil	TA	GasA	Ex	Y	FuseA	1
18	17	11241	1Fam	1Story	7	1970	Gable	TA	CBlock	TA	GasA	Ex	Y	SBrkr	1

Conclusion:

In this solution we successfully designed and set up a data warehouse using IBM cloud Db2 warehouse .we designed a data warehouse structure, integrated data from various sources , performed ETL processes , and enabled data exploration.the solution empowers data architects to explore, analyze and deliver actionable data for informed decision-making , contributing to unlocking valuable business insights and driving informed decisions.

PREPARED BY:

1.SHAMSHUNNAFIYA N

2.RAJESHWARI R

3.PAVITHRA A

4.BHUVANESHWARI P