Data Warehousing with IBM Cloud Db2 Warehouse

921821104013: M. INDHUJA

PHASE -1 Document Submission

Overview:

This is the phase 1 document for the project "Data Warehousing with IBM Cloud Db2 Warehouse" from IBM on the Naan Mudhalvan Scheme.

Project Title: Data Warehousing with IBM Cloud Db2 Warehouse

Project Description:

Design and set up a robust data warehouse using IBM Cloud Db2 Warehouse. 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!

Project 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:

1. Data Warehouse Structure: Define the schema and structure of the data warehouse to accommodate various data sources

   Sure, a data warehouse schema typically involves a star or snowflake structure. In a star schema, a central fact table is surrounded by dimension tables. In a snowflake schema, dimensions are normalized into multiple related tables. The choice depends on your specific needs and the nature of your data. Make sure to identify key attributes and relationships between data sources to design an efficient structure.

2. Data Integration: Identify data sources and design a strategy to integrate data seamlessly into the data warehouse.

   To identify data sources, assess existing databases, APIs, and applications for relevant information. Prioritize based on business goals. Design a strategy by outlining data extraction methods, transformation processes, and loading mechanisms. Ensure compatibility and consistency for seamless integration into the data warehouse. Regularly review and update the strategy to adapt to evolving data needs.

3. ETL Processes: Plan and implement ETL processes to extract, transform, and load data into the warehouse.

   To design effective ETL processes, start by understanding your data sources, defining transformation logic, and ensuring compatibility with the warehouse schema. Use tools like Apache NiFi or Talend for extraction, transformation, and loading tasks, and regularly monitor and optimize your workflows for efficiency.

4. Data Exploration: Design queries and analysis techniques to empower data architects to explore and analyze data.

Query for Data Profiling:

- **SELECT * FROM dataset WHERE column IS NULL;**
- This helps identify missing values, enabling architects to address data quality issues.

Statistical Analysis Query:

- **SELECT AVG(column), MAX(column), MIN(column) FROM dataset;**
- Provides basic statistics for numerical columns, aiding in understanding data distribution

Correlation Analysis Query:

- **SELECT CORR(column1, column2) FROM dataset;**
- Assists in determining relationships between different columns.

Pattern Recognition Query:

- **SELECT DISTINCT(column) FROM dataset ORDER BY column;**
- Uncovers unique values, patterns, or anomalies within a specific column.

Time-Series Analysis Query:

- **SELECT date_column, SUM(metric) FROM dataset GROUP BY date_column;**
- Useful for exploring trends and patterns over time.

Data Distribution Visualization:

- Utilize histograms or box plots to visually represent the distribution of numerical data.

Cluster Analysis Techniques:

- Implement K-means clustering to identify natural groupings within the dataset

Outlier Detection:

- Leverage statistical methods or algorithms like Isolation Forest to find and investigate outliers.

Data Lineage Query:

- **DESCRIBE EXTENDED dataset;**
- Provides metadata details, helping architects understand the origin and transformations applied to the data.

Query for Schema Evolution:

- SHOW CREATE TABLE table_name;
- Helps track changes in data structure over time.

5. Actionable Insights: Focus on delivering actionable insights by enabling informed decision-making based on data.

   Absolutely, actionable insights empower better decision-making, transforming data into strategic advantages.