Project 7: Data Warehousing with IBM Cloud Db2 Warehouse

Phase 1: Problem Definition and Design Thinking Document

Problem Definition:

Our project aims to tackle the challenge of modernizing our data management by designing and implementing a robust data warehouse using IBM Cloud Db2 Warehouse. This project is driven by the need to harness the potential of the organization's data, coming from diverse sources, and to empower data architects with the tools needed for insightful data analysis and informed decision-making.

The project encompasses the following key objectives:

i. Data Warehouse Structure: We will define a flexible and scalable schema and structure for our data warehouse that can efficiently handle data from various sources.
ii. Data Integration: The project will involve identifying key data sources within our organization and devising a strategy for integrating this data seamlessly into the data warehouse.
iii. ETL (Extract, Transform, Load) Processes: To ensure data accuracy and relevance, we will develop robust ETL processes for extracting data from source systems, transforming it into a usable format, and loading it efficiently into the data warehouse.
iv. Data Exploration: We aim to create an intuitive and interactive environment that allows data architects to explore and analyze the data effectively.
v. Actionable Insights: Ultimately, our goal is to deliver actionable insights from the data analysis, enabling our organization to make data-driven decisions.

Design Thinking:

➢ **Data Warehouse Structure**

To define the data warehouse structure, we will take the following practical steps:

- Data Profiling: We will conduct thorough data profiling to gain insights into the data's characteristics, including data types, relationships, and quality.
- Entity-Relationship Diagram (ERD): We will create an ERD to visualize the data model, making it easier to design schemas and tables.
- Performance Optimization: Our team will consider performance optimization techniques, such as data partitioning and indexing, to ensure efficient querying.

➢ **Data Integration**

For effective data integration, we will follow a realistic approach:

- Source Identification: We will identify all potential data sources within the organization, including databases, third-party APIs, and legacy systems.
- Data Extraction Strategy: We will determine data extraction methods and frequency, aligning them with our organization's data needs.
- Transformation Rules: We will establish clear data transformation rules and procedures to maintain data consistency and quality.

- Quality Checks: During integration, data validation and quality checks will be implemented to detect and mitigate data anomalies.

- **ETL Processes**

Pragmatic ETL processes will be designed and implemented as follows:

- ETL Tool Selection: We will select ETL tools or custom scripting languages based on our organization's resources and expertise.
- Data Profiling: Our team will perform data profiling to gain a deep understanding of data quality and structure.
- Data Lineage Tracking: We will ensure data lineage tracking for auditability and to facilitate troubleshooting.
- Monitoring and Error Handling: ETL processes will be monitored for performance, and error handling mechanisms will be in place to address issues promptly.

- **Data Exploration**

To promote practical data exploration, we will adopt these strategies:

- User-Centric Design: Our user interface using flask will prioritize user-friendliness, making it easier for data architects to interact with the data warehouse.
- Query Building: We will provide user-friendly query-building capabilities to empower data architects to create custom queries.
- Data Visualization: Implementing data visualization components, such as interactive charts and graphs, will enhance the data exploration experience.
- User Guide: We will develop a comprehensive user guide to assist data architects in making the most of data exploration features.

- **Actionable Insights**

Our approach to delivering actionable insights will be grounded in reality:

- Analysis Templates: We will create predefined analysis templates for common use cases to expedite insights generation.
- Alerting Mechanisms: Critical data thresholds will trigger alerting mechanisms to keep stakeholders informed and proactive.
- Continuous Improvement: We commit to continuous assessment and refinement of our analysis methods based on feedback and evolving business requirements.