Project Proposal

Data Warehouse Architecture

Didem B. Aykurt

April 16, 2023

Executive Summary

This proposal presents our organization's comprehensive data warehouse architecture that describes an organization’s data action of collection and storage frame for sorting, cleaning, and organizing the functional data warehouse to enable actionable Business Intelligence (BI) data for effective decision-making. The proposed solution includes a star schema of dimension and fact tables, ETL (Extract, Transform, Load) processing from multiple source production systems, data cleansing activities, a loading plan for data refresh, and strategies to handle challenges related to slowly changing dimensions.

Table of Contents

Goals and Specific Objectives

The goals and specific objectives of the proposed data warehouse architecture require the flow of the components:

- First, collect and combine data from outsources or source production systems into a single store to make the entire data view available for BI analysis.
- Second, to eliminate overlapping data sources and content, certify data quality and accuracy by implementing data cleansing and transformation processes during ETL processing.
- Third, store the collected data within a specific time, provide realization from the past point of view, and sufficiently refresh data for executive and senior-level business analysts to make informed decisions.
- Fourth, guide slowly changing dimensions proficiently and appropriately to rebuild historical data quality and qualify to develop analysis.
- Finally, upgrade business performance by manipulating actionable BI data for strategic decision-making.

Solution

The proposed solution for the data warehouse architecture solution included below components:

Star Schema of Dimension and Fact Tables

The data warehouse will use a star schema to answer business questions by providing insights into the data from the various dimension tables. For example, to answer the first question, the Product Table, Customer Table, and Store Table can be used to get the total sales for each

product. At the same time, the fact table will store the quantitative data such as sales, revenue, and other key performance indicators (KPIs). The fact table will be connected to the dimension tables through foreign key relationships, allowing easy navigation and data analysis.

ETL Processing

The ETL process will extract data from multiple source production systems, transform it to ensure data quality and accuracy and load it into the data warehouse for analysis. The ETL process will involve data extraction from various sources, data cleansing activities such as validation, standardization, and enrichment, and transformation activities such as data aggregation, integration, and denormalization. The ETL process will be automated using tools and technologies to ensure efficiency and scalability.

Data Cleansing Activities

Raw data might contain variable format, missing data, miscoded data, incorrect data, and duplicate data. The ETL process certifies data quality and precision. Therefore, data must be cleaned, transformed, and approved before load data. In this process, handle data validation problems and improve the data.

Loading Plan

The loading plan for the data warehouse will ensure that data is correctly refreshed for effective decision-making. Incremental loading techniques will load only the changed data into the data warehouse, decreasing the processing time and certifying the timely availability of refreshed data. The loading plan will also include data partitioning, indexing, and optimization strategies to improve query performance and minimize data recapture time.

Challenges and Strategies

The following challenges related to slowly changing dimensions may occur in the proposed data warehouse architecture; there are four types of slowly changing dimensions:

- Type 1 Slowly Changing Dimension overwrites the existing data warehouse value with the new value coming from the OLTP system, and this method does not maintain history value.
- Type 2 Slowly Changing Dimension contains a new and old record that uses the record effective date. If the new form enters the old one, sign the non-active effective date.
- Type 3 Slowly Change Dimension contains a new and old description with current value columns holding the new value from the OLTP system. The method comprises limited historical data, such as two previous data pieces.
- Type 4 Slowly Change Dimension has historical data separately linked to the main chronological dimension table. If any new update value into the primary dimension table, the latest data and previous version are stored in the history table. That is great for storing historical data.

Risks and Countermeasures

Determined potential risks and challenges associated with accomplishing the proposed data warehouse architecture solution, such as ETL processing detain, data quality issues, data integration complexities, and business system performance. And if a relative solution and reduction plans to address these risks and ensure smooth implementation and operation of the data warehouse.

Schedule

The set of timelines for the process of putting a decision on the data warehouse architecture solution, including key milestones, deadlines, and dependencies, and verify that the schedule is the practical idea of what can be achieved or expected, taking into point the availability of resources, potential challenges, and other external factors that may impact the timeline.

Deliverable Acceptance and Quality Assurance

Report the standard and process for acceptance action, including data models, ETL processes, data cleansing rules, and loading plans. Additionally, the period importance of quality assurance

measures throughout the performance process, such as data validation, testing, and documentation, to ensure the data warehouse's accuracy, probity, and dependability.

Recommendations/Conclusions

The proposed data warehouse architecture solution will give our organization a stable and improved platform for actionable Business Intelligence (BI) data. Control of activity and combining data from multiple source production systems to realize data cleansing activities and optimize data loading processes. The data warehouse will verify data quality, accuracy, and timeliness for effective decision-making. The slowly changing dimensions will maintain historical data integrity and enable trend analysis. It is recommended that the organization execute this proposed data warehouse architecture solution to improve the full potential of BI data and operate strategic decision-making.

The ETL processing and data cleansing activities will verify data quality, accuracy, and timeliness, while the strategies for handling slowly changing dimensions will maintain historical data integrity. The risks and challenges connected with performance are recognized, solutions are suggested, and a detailed schedule and quality assurance measures are outlined. It is recommended that the organization assume this proposed data warehouse architecture solution to enable effective decision-making and guide strategic business outcomes.

References

- Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The complete guide to dimensional modeling* (3rd edition). Wiley.
- Oracle. (n.d.), 2013. *Understanding Slowly Changing Dimensions*. https://docs.oracle.com/cd/E41507_01/epm91pbr3/eng/epm/phcw/concept_UnderstandingSlowlyChangingDimensions-405719.
- Sebastian Insausti, 2019. *Running Data Warehouse on PostgreSQL*. https://severalnines.com/blog/running-data-warehouse-postgresql/
- Tutorials Point Docs.2022. Data Warehousing – Architecture. https://www.tutorialspoint.com/dwh/dwh_architecture.htm
- Ali U… & Tekhiner F…,2020. *Data warehouse migration tips: preparation and discovery.* https://cloud.google.com/blog/products/data-analytics/how-to-prepare-for-a-data-warehouse-migration-with-google-cloud
- Microsoft Docs. (n.d.), 2023. Plan a data warehouse migration. https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/plan/data-warehouse-migration

Appendix

- The data warehouse schema diagram provides the star schema of dimension and fact tables by sample data.
- ETL process provides a chart showcasing sample data's extraction, transformation, and loading activities.
- The data cleansing rules and validation checks ensure data quality and accuracy with sample data.
- The loading plan draws the data partitioning, indexing, and optimization strategies to improve query performance with sample data.
- The slowly changing dimensions, including Type 1, Type 2, Type 3, and Type 4 scenarios, handle sample strategies.