

Project Charter

Project members

1. Indra Ariunbold
2. Sravan Kumar Mangalagiri
3. Pavan Kumar Reddy

Project overview

The Data Transformation project centers on leveraging Snowflake and dbt Cloud to enhance and streamline critical business processes within the "Pubs" database. The primary focus is on three key areas: Sales Reporting, Author Royalty, and Employee Factless Fact. The project aims to transform raw data into meaningful insights, providing stakeholders with accurate and accessible information for informed decision-making.

Project Phases

1. Data Profiling and Determining business processes to model

We began by conducting data profiling and defining key business processes. Through a comprehensive analysis of the data sources within the "Pubs" database, we identified data quality issues, assessed data types and relationships, and explored data distributions. Based on this analysis we identified 3 key business processes:

- **Sales reporting:** For each order, we want to identify sales metrics such as total quantity and revenue with their authors, stores it was bought from and time dimensionality. Purpose is to determine what kind of books sell better and which stores are profitable at what season or month.
- **Author royalty:** For each author, we want to determine how much royalty has been paid by publishers. Purpose is to find out which authors are more costly to collaborate with.
- **Employee Analysis:** For each publisher, we want to analyze its employees by their work experience, job title to see which employees at which publishers have resulted in more profitable titles.

This involved setting clear business goals, documenting transformation requirements, and identifying data dependencies. The outputs of this phase include a knowledge of data sources and well-defined business processes that lay a solid foundation for the subsequent data modeling and transformation activities in dbt Cloud.

2. High-level dimensional modeling

In the subsequent phase of our project we transitioned to high-level dimensional modeling. This involved designing a robust dimensional model that aligns with the identified business processes. By applying dimensional modeling principles, we defined key dimensions, such as time and entities specific to each business process. Additionally, we identified fact tables capturing the core business metrics and dimension tables providing context to the data. This high-level modeling approach ensures that the transformed data is structured efficiently for analytical queries, providing us with a user-friendly schema that supports our BI needs. The dimensional model serves as the blueprint for creating dbt models in the subsequent phase, facilitating the seamless translation of business requirements into actionable data transformations.

3. Detail-level dimensional modeling

Following the high-level dimensional modeling, the project progressed into the detail-level dimensional modeling phase. In this stage, we delved deeper into the specifics of each business process to refine and expand upon the initial dimensional model. Key activities included identifying granular attributes, hierarchies, and relationships within each dimension. The detailed dimensional models act as a guide for the development of dbt models, facilitating the transformation of raw data into valuable insights at a granular level. *For this, please refer to the “Dimensional Modeling” Excel spreadsheet that is attached to this project.*

4. Setup of Data Warehouse and dbt Cloud

In the project's implementation phase, we set up the Snowflake data warehouse to house our transformed data and configured dbt Cloud to orchestrate efficient data transformations. We created Snowflake and dbt Cloud accounts, made connections between them.

5. Initial Loads Data Pipeline /ELT/

This phase involved the systematic extraction of data from source systems, loading it into the Snowflake data warehouse, and subsequently applying transformations using dbt Cloud. *For this, please refer to the “Data Loader” PDF file that is attached to this project.*

6. Implementation of business processes

Leveraging detailed dimensional models developed earlier, we created dbt models for each business process. We created sql and yml files for dimension and fact tables as well as OBT. As a result we were able to build our star schemas in Snowflake and have extensive documentation generated by dbt Cloud.

7. Business Intelligence

Following the dbt Cloud implementation, we utilized Business Intelligence (BI) tool, Power BI, to visualize the transformed data. Using our well-defined data warehouse, we were able to design and create Power BI dashboards for each of the business processes. These dashboards serve as user-friendly interfaces, allowing stakeholders to interact with and gain insights from the data easily. By connecting Power BI to the Snowflake data warehouse, we ensure that the dashboards

provide accurate visualizations, enabling informed decision-making and comprehensive analytics. *For this, please refer to the “Sales and Employee BI” and “Royalty BI” Power BI files that are attached to this project.*