IST 722: Data Warehouse

Final Project Report

Dataset: Pubs

A logo with leaves and text

Description automatically generated

**Group Members:**

Hardik Lilani

Rithvik Segu

Kumar Reddy

**Project Overview**

**Introduction**

The "pubs" dataset is designed to manage and analyze the operations of a publishing company. It consists of several related tables:

* **Authors**: Contains personal details about the authors, including their unique identifier, name, and contact information, along with contractual data.
* **Titles**: Lists the books published, including title identifiers, names, the publisher associated, and financial details such as advances, royalties, and sales.
* **Publishers**: Details information about the publishers, like their identifier, name, and location.
* **Sales**: Tracks the sales transactions, detailing which store sold which title, the quantity, and the order information.
* **Stores**: Stores information regarding the retail outlets, including their identifiers and addresses.
* **Discounts**: Provides information on the discount schemes applicable to the stores based on the quantity of the purchase.
* **Royalties Schedule (roysched)**: Outlines the royalty percentage that authors receive based on the sales range of their titles.
* **Title Author**: Links authors to their titles and specifies their royalty agreements.
* **Publisher Information (pub\_info)**: Contains additional details about publishers, potentially including logos and promotional information.
* **Employees and Jobs**: These entities manage information about the company's employees and their job roles, which are linked to the publishers and titles.

This dataset facilitates a comprehensive analysis of the publishing process, from tracking title sales and managing author contracts to handling publisher information and employee details.

**High-Level Dimensional Modeling**

The bus matrix for the "pubs" dataset has been meticulously developed to capture the essential elements of the publishing business process. The matrix identifies three core business processes: Title Sales, Store Sales and Publisher Owned Titles. Each process is associated with a fact table, which is detailed further in the matrix.

1. **Title Sales (fact\_sales)**: This process captures transactional data with each row representing a sale of a particular title. The granularity is well-defined, and the associated facts include quantity, discount amount, and revenue per title.
2. **Store Sales (fact\_store\_sales)**: Representing a periodic snapshot, this fact table holds data at the granularity of one row per book sold in a store on a daily basis. It includes facts such as quantity, discount amount, and revenue per store.
3. **Publisher Owned Titles (fact\_publisher)**: Also a transaction, this table details publisher ownership per author with a granularity of one row per publisher per author. The fact is the participation of the publisher.

**Detail-Level Dimensional Modeling**

The dimension tables identified in the bus matrix play a critical role in contextualizing the facts. These dimensions include:

* **dim\_title**: Provides details on each book or title.
* **dim\_store**: Contains details about the stores where books are sold.
* **dim\_discount**: Describes the discount structures applied to sales.
* **dim\_publisher**: Details publisher information.

**Data Warehouse Implementation**

The "pubs" dataset, when leveraged for data warehousing using Snowflake, dbt (data build tool), and Power BI, creates a robust and scalable analytics ecosystem. Here's how each component fits into the data warehousing solution:

* Initially we had our data in azure data studio and minio S3, we chose snowflake as our data warehousing platform. Snowflake provided a streamlined Extraction and loading process, all the data was stagged in snowflake for further process.
* In the next step we used DBT cloud for transforming the data according to our business process by calculating the sales of each title, revenue generated by each store per day and the number of titles owned by each publisher by using SQL queries. The transformed data was migrated to analytics database.
* Then we integrated the transformed data to Power BI, to perform visualizations on our data. These visualizations provide a deep and meaningful insights to our data. We created a dashboard for each business process

**Changes Made To BI According to Feedback:**

* **Hierarchy added to pubdate column in our third dashboard**
* **In our first dashboard, we changed title\_id to title\_name for better understanding to users**

**Challenges**

During the implementation of our data warehousing solution for the "pubs" dataset, we encountered several challenges:

1. **Limited Data Volume**: The tables within the dataset contained a relatively small number of rows. This restricted the depth of analysis and the robustness of the insights we could derive, as the data did not represent a wide variety of scenarios and outcomes typically used to fine-tune analytics models. We faced a huge challenge while creating the third business process, initially we thought to calculates the revenue received by each author from royalties but we couldn’t get it done so we moved to calculate the number of titles owned by each publisher(factless fact), we did not have a lot of data to play with this, we only had 8 publishers, and only three of them owned titles(16 in total) and those three were from same country.
2. **Data Loading into Snowflake**: We faced difficulties, while loading columns which contained dates in it, tried various methods and finally figured out to initially load it in varchar and then change to date function by using to\_date.
3. **Fact Tables in DBT**: While creating facts in DBT we faced many issues, went back to dimensional modelling(Excel Sheet) made few changes in it, removed the unnecessary dimension tables, added/edited columns in the fact table. We initially loaded the date dimension table into snowflake, but after performing transformations in DBT, we got to know that the date columns in our data was enough, so we did not use this date dimension separately.
4. **Integration with Power BI**: Establishing a connection between DBT and Power BI presented its own set of issues. We missed few relations between the dimension tables and fact tables, we had to connect the key’s from dimension table to fact tables(one to many relationship), which we got to know while performing drill across. We missed a key either in dimension or fact table in couple of instances, but we corrected it by going back to DBT. In our third business process dashboard we did not have a lot of data to play with as mentioned above, so we very simple visualization in it, but we made sure to use drill-across in that dashboard as well. For creating date hierarchy in our pub\_date , we created a calculated column using DAX, were we calculated the year, month and date. Which we later added to our filters in our third dashboard.

These challenges were significant but also provided valuable learning experiences. They underscored the importance of meticulous data preparation and the need for flexible problem-solving approaches when integrating diverse tools within a data warehousing ecosystem.

**Key Learning/Reflection**

* We obtained substantial knowledge about the intricacies of data warehousing, particularly the practicalities of loading and extracting data within the Snowflake environment. This deep dive into operational aspects provided us with hands-on experience that is invaluable for understanding the lifecycle of data within a warehouse.
* Our engagement with the Kimball methodology offered a comprehensive perspective on data warehousing. We learned to appreciate the full spectrum of the process, from understanding the business needs and conceptualizing the data architecture to the actual implementation and data usage for decision-making.
* A pivotal aspect of our learning was mastering the Extract, Load, Transform (ELT) process. We focused on the practical application of dbt to conduct these transformations efficiently. This tool enabled us to manipulate large datasets within the warehouse directly, optimizing performance and scalability.
* The project moved us beyond theoretical knowledge to practical application, particularly in utilizing dbt. We were able to apply the concepts learned in real-world scenarios, enhancing our skills in data transformation and manipulation.
* Through the process of connecting dbt cloud with Power BI, we gained a nuanced understanding of the end-to-end business intelligence process. This included not only the technical aspects of data handling but also the presentation and visualization of data to derive actionable insights.
* The experience has significantly improved our comprehension of data flows and warehousing dynamics. We now better understand how data moves through systems, how it can be transformed to add value, and the importance of data integrity and quality.
* The project taught us the importance of integrating various tools and technologies to build a cohesive data warehousing and business intelligence solution. We learned to navigate and resolve the challenges that arise from such integrations.
* We learned to adapt to the limitations posed by data volume and the necessity of meticulous data preparation. This adaptability has become a key asset in our approach to data warehousing projects.

These learnings contribute to a robust foundation for future projects, ensuring that we are equipped to manage data warehousing endeavors with greater expertise and confidence.

INCASE IF YOU FACE ANY ERROR IN OPENING UP THE GITHUB LINK, THESE ARE OUR DBT CLOUD CREDENTIALS:

Email- datadockers722@gmail.com

Password- DataWarehouse#722  
  
https://cloud.getdbt.com/develop/216373/projects/311465