Analysis and Summary of Product Reviews Project

Report

Designing a Fundamental Data Warehouse Schema for E-commerce

Instructor: Nguyễn Ngọc Thiện
Implementer: Nguyễn Phúc Nhân

Sep 2024, Ho Chi Minh City

Table of contents

1 Introduction

1.1 Objectives

The objectives of project are outlined below:

- Design and implement data warehouse to centralize and streamline data from an e-commerce platform
- Analyze product reviews to understand customer sentiment and product performance, driving business strategies and improving customer satisfaction
- Use data to train language models, analyze customer reviews for deeper insights into user feedback.

1.2 Project Scope

- The scope of the project encompasses a case study focused on analyzing and summarizing user feedback on products from an e-commerce platform.

- This will help users save time by providing concise summaries of multiple reviews, allowing them to quickly assess product quality and make informed purchasing decisions.

2 Data Warehouse Architechture and Implementation

2.1 Dimension Modeling

Dimension modeling is a database design methodology used in data warehouses to organize and structure data for easy querying and analysis. It is widely adopted in decision support systems, especially for reporting and data analysis.

Dimension modeling primarily consists of two types of tables:

2.1.1 Fact Table:

- Contains quantitative data related to business events, such as sales, revenue, and product quantities sold.

- Each record in the fact table typically contains foreign keys that link to dimension tables.

- The data in fact tables is usually numerical, like sales amounts, revenue, or customer ratings.

2.1.2 Dimension Table:

- Contains descriptive attributes that provide context for the data in the fact table. These attributes are relatively static and provide qualitative information.

- Examples include product information (name, category, brand), customer details (name, region, registration date), or time (day, month, quarter, year).

2.1.3 Star Schema vs. Snowflake Schema in Dimension Modeling

Dimension modeling is usually organized in either a **star schema** or a **snowflake schema**:

- **Star Schema**: This is the simplest model, where the fact table is at the center, and the dimension tables surround it. The fact table is linked directly to each dimension table through foreign keys.

- **Snowflake Schema**: This is a more complex version of the star schema, where the dimension tables are further normalized into smaller tables, creating a structure that resembles a snowflake.

2.1.4 Benefits of Dimension Modeling:

- **Easy Querying**: Dimension modeling is optimized for querying and reporting, making it easier to run complex queries.
- **High Performance**: The structure of dimension modeling improves performance in analytical systems.
- **Intuitive and Understandable**: Designed for end users, dimension modeling is user-friendly, making it easier to navigate without deep technical knowledge of the database.

Dimension modeling is a powerful tool for organizing and analyzing large datasets in data warehouses, making it particularly useful for businesses aiming to derive insights from complex data.

2.2 Slowly Changing - Dim Table

Name of Table	Attributes	Purpose

Name of Table	Attributes	Purpose
dim_Product	• product_id • name • description • category_id • brand_id • price • discount • rating_average • review_count	This table stores comprehensive details about each product, including its category and brand affiliations. By analyzing this data, businesses can gain insights into product performance, customer preferences, and the impact of pricing strategies on sales and reviews.
dim_Customer	• customer_id • name • full_name • region • created_time • joined_time • total_reviews • total_thanks	Contains customer information, allowing segmentation and analysis of customer behavior and demographics.
dim_Seller	• seller_id • name • is_best_store	Holds seller details, facilitating the evaluation of seller performance and reputation.

Name of Table	Attributes	Purpose
dim_Category	• category_id • name	Categorizes products, aiding in the analysis of product categories.
dim_Brand	• brand_id • name	Stores brand information, enabling brand performance analysis.
dim_Date	• date_id • date • day • month • year • quarter • is_weekend • is_holiday	Provides date-related attributes for time-based analysis.

2.3 Event Record - Fact Table

Name of Table	Attributes	Purpose

Name of Table	Attributes	Purpose
fact_Comment	• comment_id • review_id • customer_id • content • score • is_reported	Stores comments on reviews, allowing sentiment analysis and quality control of reviews.
fact_Sentiment	• sentiment_id • review_id • sentiment_type	Captures sentiment analysis results, aiding in understanding customer sentiment.
fact_Sales	• sale_id • product_id • customer_id • seller_id • date_id • quantity • price • discount • total_amount	Records sales transactions, enabling sales performance analysis.

Name of Table	Attributes	Purpose
fact_Reviews	• review_id • product_id • customer_id • seller_id • date_id • rating • content • helpful_votes • helpful_count • summary_id	Stores product reviews, facilitating detailed review analysis.
fact_ReviewSummary	• summary_id • product_id • summary_text • positive_points • negative_points • average_rating • total_reviews • last_updated	Summarizes reviews, providing a consolidated view of product feedback.

2.4 Model Diagram

The diagram below illustrates the star schema used in the product review analysis data mart for an e-commerce platform.

The fact tables capture key transactions and review data, while the dimension tables provide descriptive information necessary for analysis, such as product details, customer data, and dates.

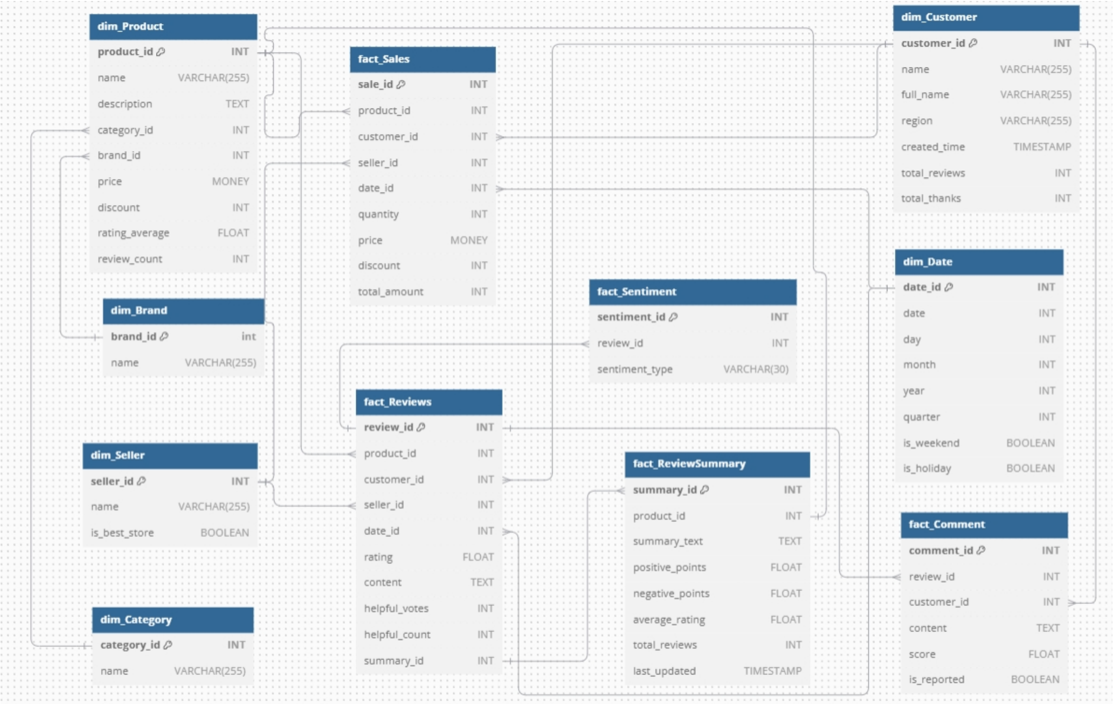
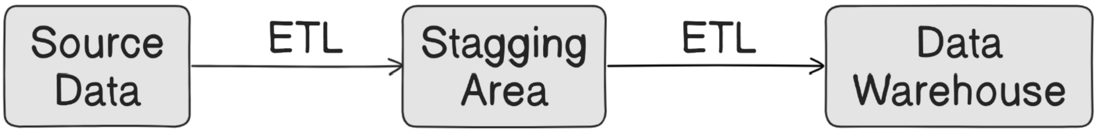


Figure 1: Star Schema Model for E-commerce Product Review Analysis

3 ETL Method And Strategy

3.1 ETL Enviroment

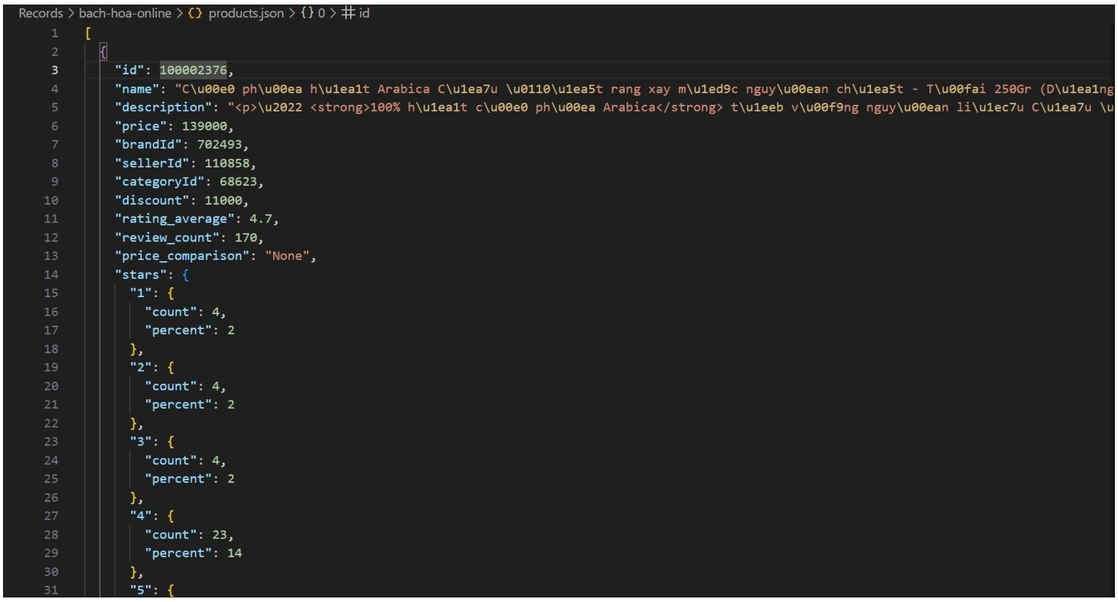
Before extraction, two databases were created: staging and data warehouse. These were partitioned to separate raw data from clean, prepared data, following a two-phase ETL approach. Source data is first extracted, transformed, and loaded into staging tables, then further examined and transformed again before loading into the data warehouse.



This iterative process helped improve data quality. Initial ETL runs revealed issues, which were resolved by adjusting the ETL packages. Once tested and validated, ETL workflows were created for each data source, with a separate ETL package to load the Facts table in the data warehouse.

3.2 ETL method in this project

The source data for this project was crawled from Tiki's API, a popular e-commerce platform with a public API (Tiki's API Documentation). Using Python, requests were sent to the API, and the responses were saved in JSON format. The next step involved cleaning and selecting the necessary data, focusing on product reviews and other relevant information, which was stored in the staging area as raw data.



Records > bach-hoa-online > {} products.json > {} 0 > # id

```
[
  {
    "id": 100002376,
    "name": "C\u00e0 ph\u00ea h\u1ea1t Arabica C\u1ea7u \u0110\u1ea5t rang xay m\u1ed9c nguy\u00ean ch\u1ea5t - T\u00fai 250Gr (D\u1ea1ng
    "description": "<p>\u2022 <strong>100% h\u1ea1t c\u00e0 ph\u00ea Arabica</strong> t\u1eeb v\u00f9ng nguy\u00ean li\u1ec7u C\u1ea7u \u
    "price": 139000,
    "brandId": 702493,
    "sellerId": 110858,
    "categoryId": 68623,
    "discount": 11000,
    "rating_average": 4.7,
    "review_count": 170,
    "price_comparison": "None",
    "stars": {
      "1": {
        "count": 4,
        "percent": 2
      },
      "2": {
        "count": 4,
        "percent": 2
      },
      "3": {
        "count": 4,
        "percent": 2
      },
      "4": {
        "count": 23,
        "percent": 14
      },
      "5": {
```

Figure 2: Example Raw Data for Product Infomations

Once the data warehouse was designed, data was loaded into it, with debugging performed as needed. After successfully initializing and populating the warehouse with the initial data, the data was then transferred to the data mart for analysis. The analysis results were stored and regularly updated in the data warehouse for future use.

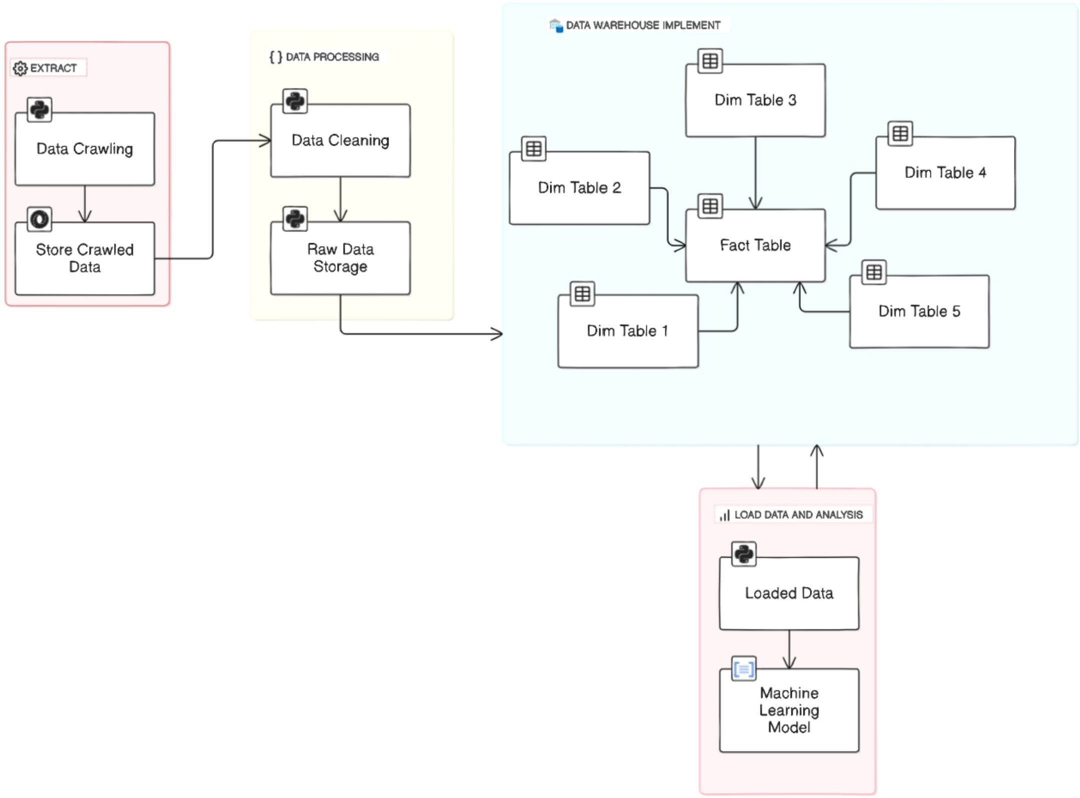


Figure 3: ETL Workflow

References

[1] Inmon, W. H. (2005). Building the data warehouse (4 Edition). John Wiley & Sons

[2] Rainardi, V. (2008). Building a Data Warehouse: With Examples in SQL Server. Apress

[3] Kumar, P., & Kavita, D. (2021). Data Warehouse Concept and Its Usage. Sri Sivasubramaniya Nadar College of Engineering, Jagannath University.

[4] Rahman, N., & Rutz, D. (2019). Building data warehouses using automation: Concepts, methodologies, tools, and applications. In book Rapid Automation (pp. 735-759).