Designing a Data Warehouse for Amazon Sales Analysis

Overview

- ★ **Bullet Points**:
  - Introduction to the Project
  - Data Exploration and Cleaning
  - Fact and Dimension Table Creation
  - DBT Setup and Configuration
  - Data Preparation for BI
  - Ensuring Data Quality

- ★ **Git_hub_repo: Link**
- ★ **Dashboard Link: Link**

Introduction

This project involves building a data warehouse for an online retail company to optimize the analysis of sales data on Amazon. By utilizing Kimball dimensional modeling, the project aims to create a structured, easy-to-use data environment that supports marketers in analyzing product performance, sales trends, and customer preferences. The data warehouse is designed to facilitate efficient querying and reporting, enhancing decision-making capabilities through clear, actionable insights.

Data Exploration

Initial Observations:

- ★ Dataset Size: **128,975** rows and **24** columns.
- ★ Data Types: Includes `object`, `bool`, `int`, and `float`. Date is in `object` format.
- ★ Missing Values: Found in the following fields: `currency`, `amount`, `courier status`, `ship-city`, `ship-state`, `ship-postal-code`, `ship-country`, `promotion-ids`, and `fulfilled-by`.
- ★ Unwanted Columns: "`Unnamed: 22`" and `index`.
- ★ Duplicate Records: **6**.
- ★ Duplicate Order IDs: **15,431**.

Data Cleaning

- ★ Stripping and typecasting data
- ★ Handling duplicates: Retained the latest Order ID records
- ★ Remove unwanted columns
- ★ Standardized and cleaned date formats (YYYY-MM-DD)
- ★ Cleaned ship-city by removing special characters and numbers
- ★ Created classified-city for major city analytics
- ★ Standardized state names
- ★ Renamed columns to lowercase with underscores

Fact and Dimensions Creation

Fact Table:

- ★ Name: fact_sales
- ★ Key: order_id
- ★ Measures: order_id, date_id, product_id, location_id, promotion_id, fulfilment_id, b2b_id, courier_status_id, fulfilled_by_id, order_status_id, sales_channel_id, ship_service_level_id, qty, amount, currency

Fact and Dimension Creation

Dimension Tables:

- ★ dim_location: location_id, city, state, country, postal_code, classified_city
- ★ dim_product: product_id, sku,category, size, asin, style
- ★ dim_b2b: b2b_id,is_b2b
- ★ dim_courier_status: courier_status_id, courier_status
- ★ dim_date: date_id, date,day, month, year, quarter, season,is_weekend
- ★ dim_fulfilled_by: fulfilled_by_id, fulfilled_by
- ★ dim_fulfilment: fulfilment_id, fulfillment
- ★ dim_order_status: order_status_id, order_status
- ★ dim_sales_channel: sales_channel_id, sales_channel
- ★ dim_shipment_service_level: ship_service_level_id, ship_service_level

Fact and Dimension Creation

Reason:

- ★ Enables future schema changes
- ★ Improves query performance
- ★ Organizes data better
- ★ Increases flexibility and data granularity
- ★ Ensures data quality and consistency

DBT Project Initialization

1. Project Initiation

- Launched a DBT project to streamline data transformation and management.
- Established the project environment and installed necessary tools.

2. Environment Setup

- Installed `dbt-postgres` for PostgreSQL compatibility.
- Configured the environment to ensure smooth integration with the data warehouse.

3. Configuration

- **Profiles Configuration**: Set up `profiles.yml` to define database connection settings.
- **Services Configuration**: Configured `services.yml` to manage DBT

DBT Model Creation and Data Handling

4. Model Creation

- Created SQL models for fact and dimension tables:
  - **Fact Tables**: `fact_sales`
  - **Dimension Tables**: `dim_location`, `dim_product`, etc.
- Designed models to fit the Kimball star schema and support efficient querying.

5. Data Handling

- Configured models to update existing records to prevent data loss.
- Ensured historical data accuracy and integrity.

6. Verification

- Verified data insertions and updates to confirm correctness and completeness.

Data Preparation for use in BI

Direct Queries on Fact and Dimension Tables:

- ★ **Real-Time Data Access**: Always query the latest data for up-to-date insights.
- ★ **Flexibility**: Easily create custom queries without needing pre-defined aggregates.
- ★ **Cost-Efficient**: Avoid extra storage and maintenance costs from additional tables.
- ★ **Simplified Management**: Less complexity in data processing and pipeline.
- ★ **Adaptable**: Quickly adjust to new business questions and evolving needs.
- ★ **Handling Uncertain Requirements**: Made this choice due to incomplete business requirements, allowing for flexibility as needs evolve.

Data Analysis:

- ★ Amazon fulfillment is managed solely by Amazon.
- ★ Merchant fulfillment is handled by "Easy Ship."
- ★ Data is available for March, April, May, and June.
- ★ Most sales occurred during the spring season.
- ★ Merchant fulfillments have the highest number of returns and cancellations.
- ★ Top 3 popular product categories: Kurta, Set, Western Dress.

Data Quality Assurance

- ★ Consistent data cleaning and standardization
- ★ Regular updates and validation checks
- ★ Data integrity maintained through DBT configurations
- ★ Continuous monitoring and adjustments

Thank You