# IST 722: Data Warehouse

# Final Project Report

## Database: Chinook

**Team 8 Group Members**

Vishnu Charugundla

Rishikesh Ramesh

Pranav Mahesh Mekal

**Introduction:**

In today's data-driven world, organizations rely on robust data warehousing systems to extract actionable insights and make informed decisions. This project focuses on leveraging the Chinook database to design and implement a data warehouse solution. The Chinook database, a sample database that emulates a digital media store, provides a rich dataset with information on customers, invoices, artists, albums, and tracks. This comprehensive dataset serves as an excellent foundation for understanding and applying core data warehouse concepts.

The primary objective of this project is to transform the transactional data within the Chinook database into a dimensional data warehouse model, enabling efficient querying and analytics. By employing dimensional modeling techniques, Extract-Transform-Load (ETL) processes, and business intelligence tools, this project seeks to deliver a scalable and actionable business intelligence solution. Through the design and implementation of star schemas, the project captures key business processes such as sales analysis and customer segmentation, which are essential for making data-driven decisions in a real-world digital media store.

Building a business intelligence (BI) solution on top of this data warehouse aims to provide actionable insights through advanced data analysis. BI tools empower stakeholders to explore sales patterns, customer behavior, and inventory management by executing complex queries. The solution enables the creation of real-time reports and dashboards that support monitoring of company operations. Additionally, it facilitates multidimensional data analysis across various aspects, such as time periods, geographic regions, and customer profiles. Transforming the Chinook database into a data warehouse, coupled with a BI solution, enhances strategic decision-making. This approach helps the company streamline operations, improve customer engagement, and drive increased sales.

**Project Timeline**

1. Weeks 1 and 2: Requirements and Planning
   1. Data Profiling: Analyze the source data for quality, consistency, and completeness.
   2. Business Process Identification: Define key business processes to be captured in the data warehouse.
   3. Bus Matrix Creation: Identify fact and dimension tables, and map business processes to data needs.
2. Weeks 3 and 4: Initial Design
   1. Dimensional Modeling: Design star and snowflake schemas based on business processes.
   2. Source-to-Target Mapping: Document mappings between source data and target schema.
   3. Technical Architecture Design: Plan the infrastructure setup, including database platforms and tools.
3. Weeks 5, 6, and 7: Implementation
   1. Database Setup: Configure the database environment (e.g., Snowflake, BigQuery).
   2. Schema Creation: Create fact and dimension tables as per the design.
   3. ETL Development: Develop ETL/ELT pipelines using dbt, Apache Airflow, or other tools.
4. Weeks 8, 9, and 10: Analytics and Testing
   1. Dashboard Development: Build visualizations and dashboards using tools like Power BI or Tableau.
   2. Testing and Validation: Perform unit, integration, and end-to-end testing of the data pipelines and analytics.
   3. Documentation: Create comprehensive technical and functional documentation.
5. Weeks 11 and 12: Finalization and Delivery
   1. Final Documentation: Compile all project deliverables, including technical specs and user guides.
   2. Project Delivery: Deliver the final project to stakeholders, including a walk-through session.

**Technical Stack**

The project will leverage the following tools and technologies:

* **Data Warehouse**: Snowflake
* **ETL Tool**: dbT
* **BI Tool**: Tableau

**Functional Requirements**

|  |  |
| --- | --- |
| Category | Requirement |
| Data Collection | Collect data from multiple sources, including customer purchases, album sales, and invoice payments. |
| Data Processing | Process and clean data to ensure consistency and integrity before storage.  Transform source data into dimensional models (fact and dimension tables). |
| ETL Pipeline | Automate ETL pipelines to extract, transform, and load data using dbT. |
| Data Storage | Store processed data in a cloud-based data warehouse (Snowflake).  Maintain dimensional models for efficient querying and analytics. |
| Reporting | Develop dashboards in Tableau for monitoring KPIs and trends.  Generate detailed reports for business stakeholders. |

**Non-Functional Requirements**

|  |  |
| --- | --- |
| Category | Requirement |
| Scalability | Support increased data volume as the business expands its processes and customer base. |
| Performance | Ensure dashboards render complex queries within 5 seconds to maintain user engagement. |
| Reliability | Build robust pipeline mechanisms to automatically retry and recover from errors during ETL processes. |
| Data Security | Enable granular access control at schema, table, and column levels to safeguard sensitive data. |
| Data Accuracy | Ensure data transformations maintain accuracy and completeness across all processing stages. |
| Usability | Design dashboards with actionable insights, using intuitive layouts and minimal training requirements.  Provide step-by-step navigation or tooltips for users accessing advanced analytics features. |
| Maintainability | Use modular ETL scripts and configurations to facilitate quick updates and debugging. |

**Key Business Processes Based**

**1. Sales of Songs**

- Objective: Track individual song sales performance

- Details:

- Monitor monthly quantity and sales amounts

- Track total revenue and discounts

- Analyze single song performance metrics

**2. Album Sales**

- Objective: Evaluate album-level performance

- Details:

- Track total amount per album sale

- Monitor revenue per album

- Analyze quantity and discount impacts

**3. Customer Purchase Analysis**

- Objective: Understand customer buying patterns

- Details:

- Track total amount and number of items per purchase

- Monitor total tracks and albums bought

- Analyze invoice status and customer behavior

**4. Invoice Management**

- Objective: Monitor payment processing

- Details:

- Track payment amounts and dates

- Monitor payment methods

- Manage outstanding balances

**5. Artist Performance Tracking**

- Objective: Evaluate artist success metrics

- Details:

- Monitor total tracks and albums sold

- Track revenue per artist

- Analyze average discounts

**6. Invoice Lifecycle Management**

- Objective: Track invoice process flow

- Details:

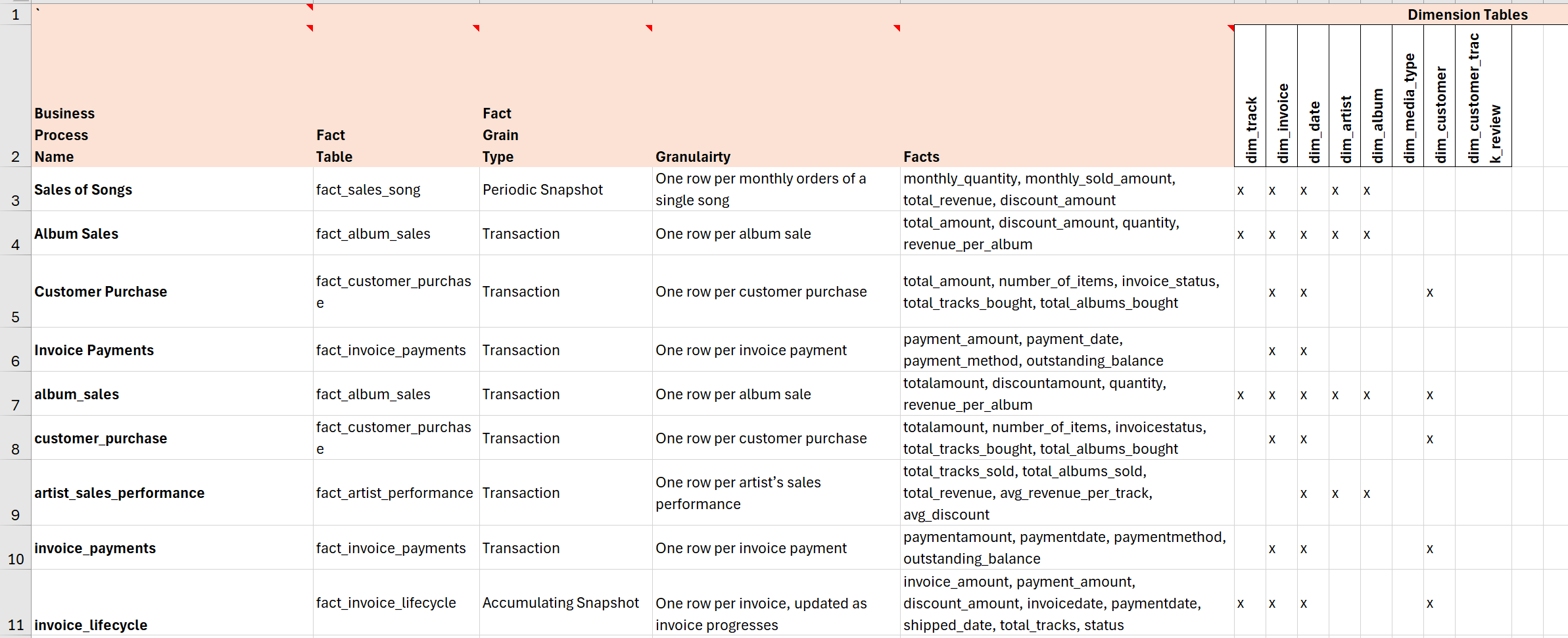
- Monitor invoice amount and payment status

- Track discount applications

- Manage shipment dates and status updates

**High-level dimensional modeling**

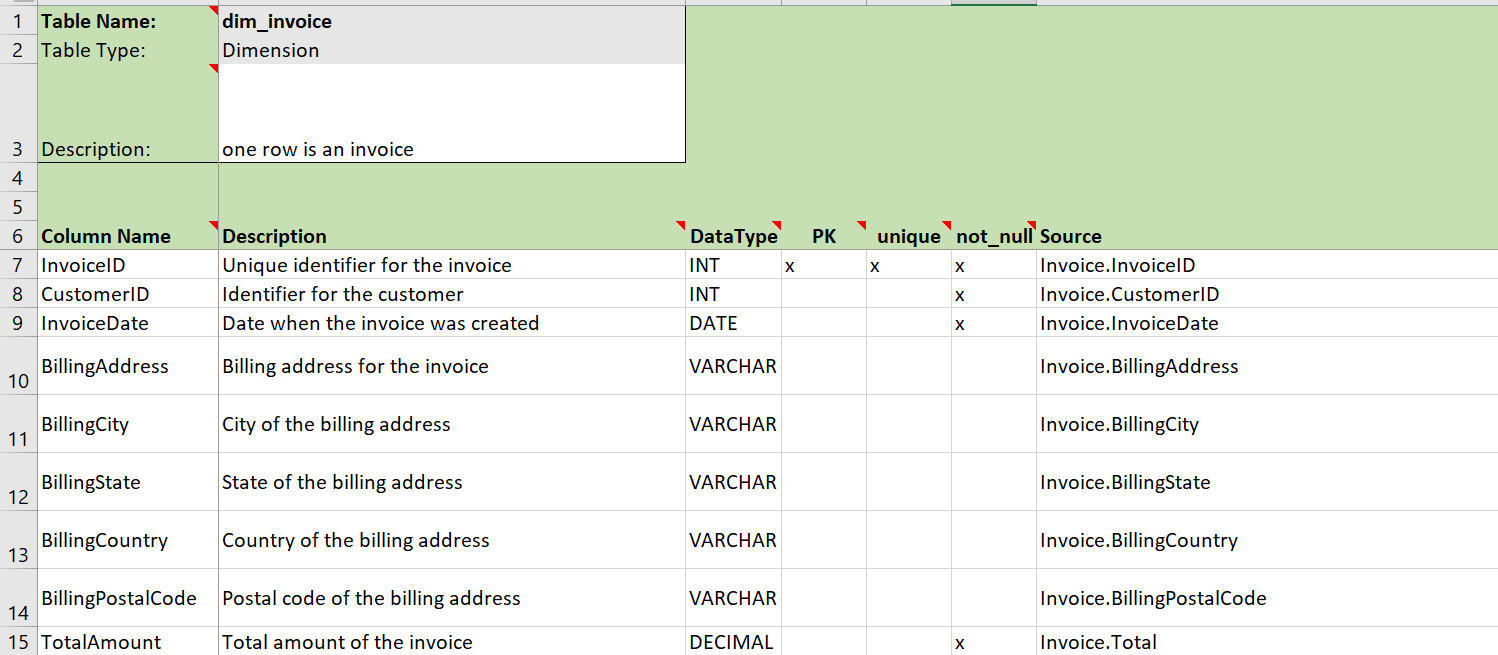
**Bus matrix**

****

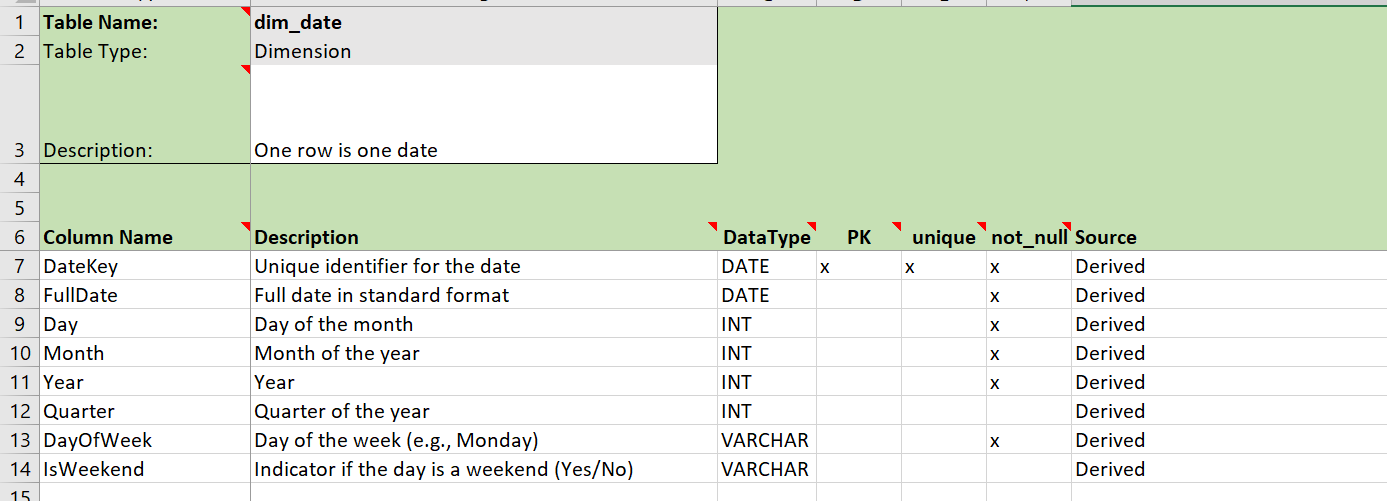
The bus matrix illustrates the key fact tables and their relationships with dimension tables in our data warehouse. It shows six distinct business processes: sales of songs (periodic snapshots), album sales (transactions), customer purchases, invoice payments, artist performance, and invoice lifecycle tracking (accumulating snapshot). Each process connects to relevant dimensions like tracks, artists, and customers, enabling comprehensive analysis of sales, customer behavior, and financial metrics. The matrix's granularity varies from individual song sales to aggregated invoice lifecycles, providing both detailed and summary-level analytical capabilities.

**Detail-level dimensional modelling**

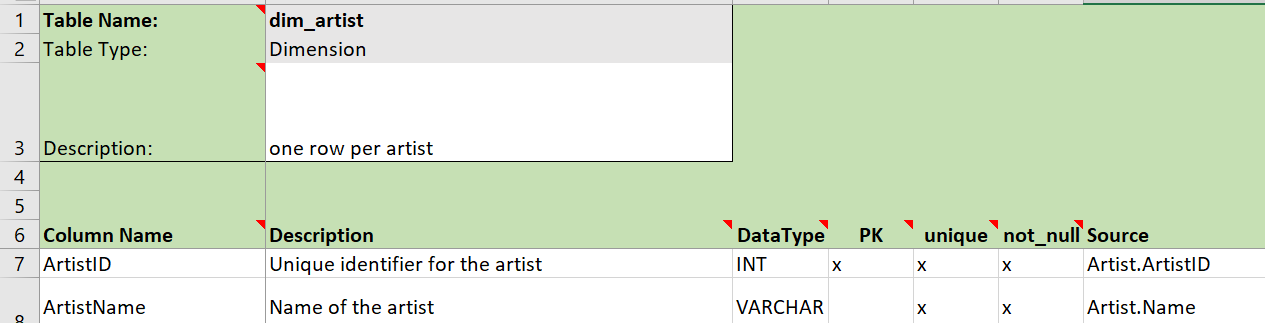
**dim\_invoice**

****

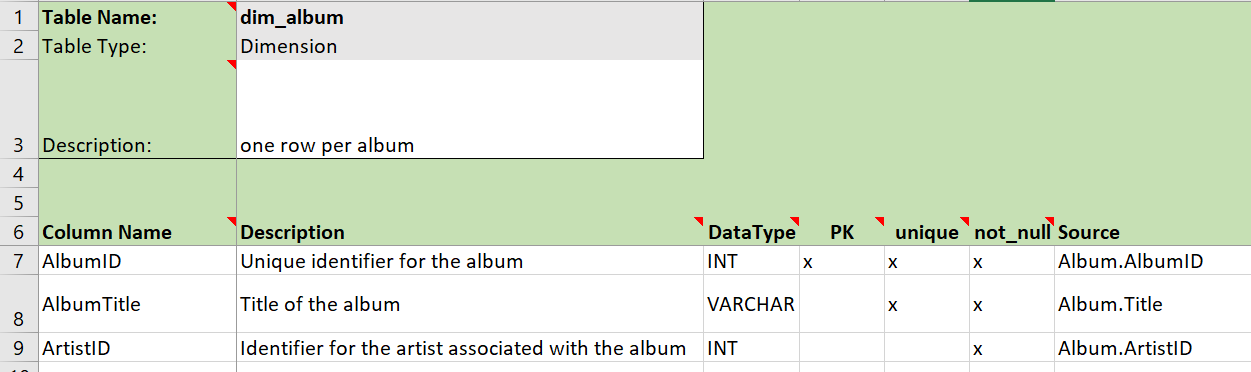
**dim\_date**

****

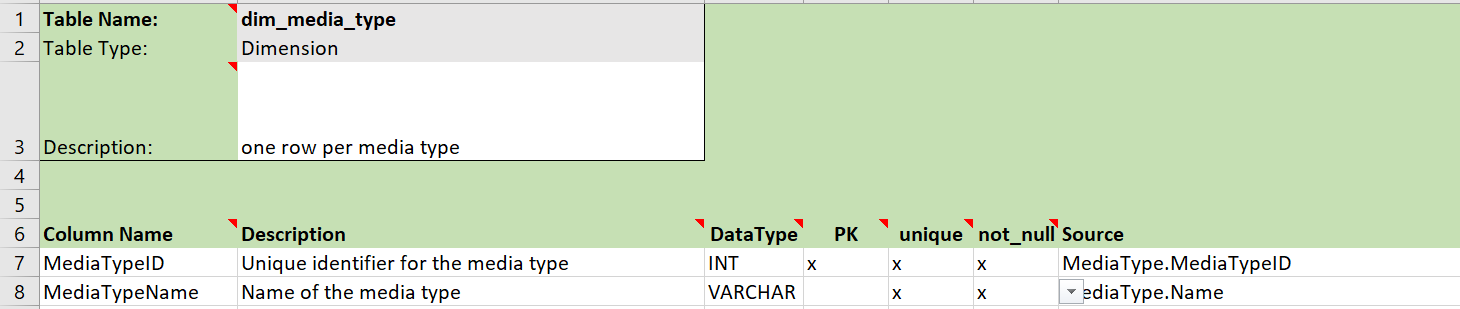
**dim\_artist**

****

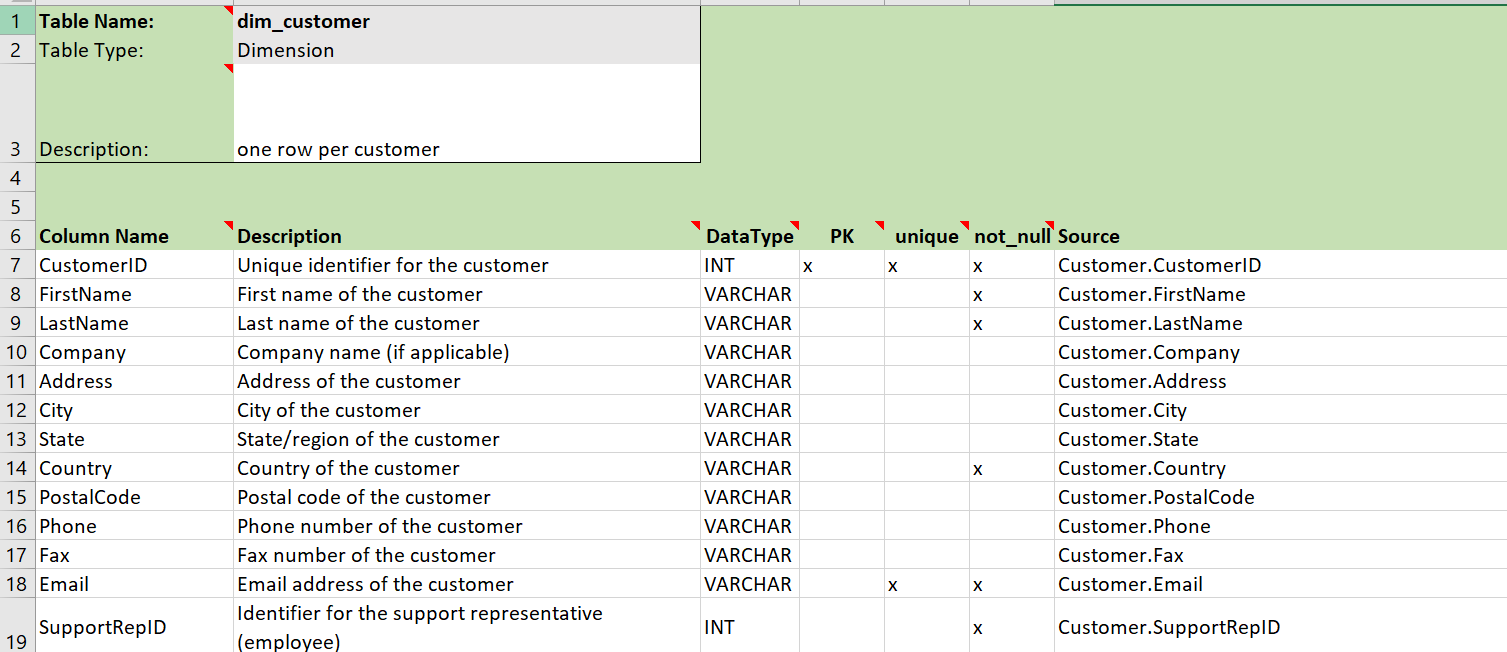
**dim\_album**

****

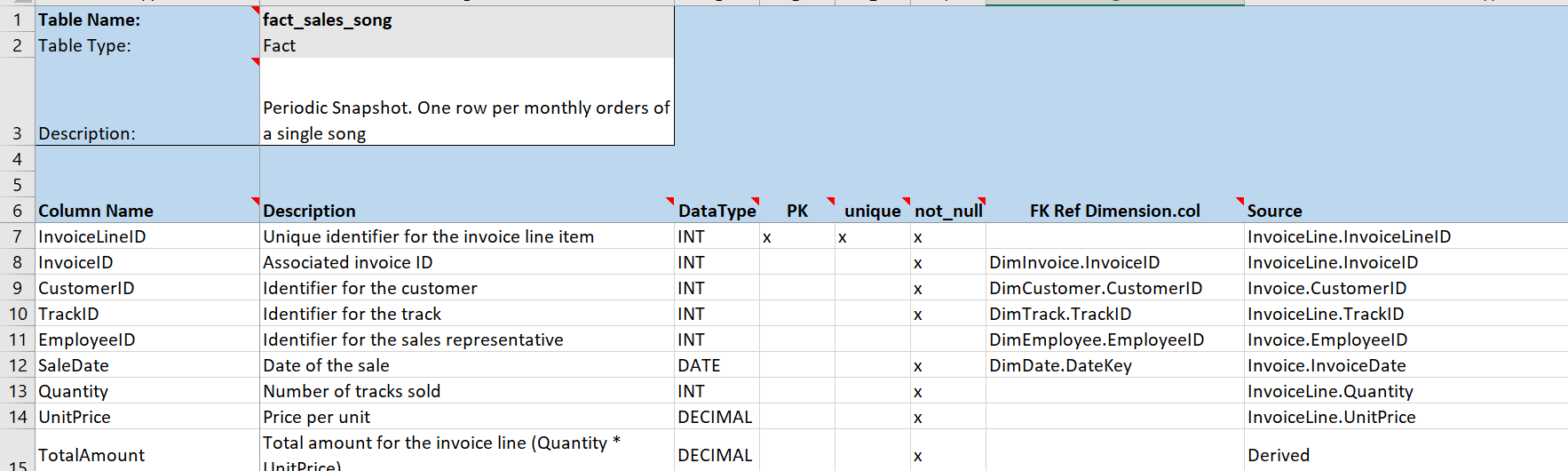
**dim\_media\_type**

****

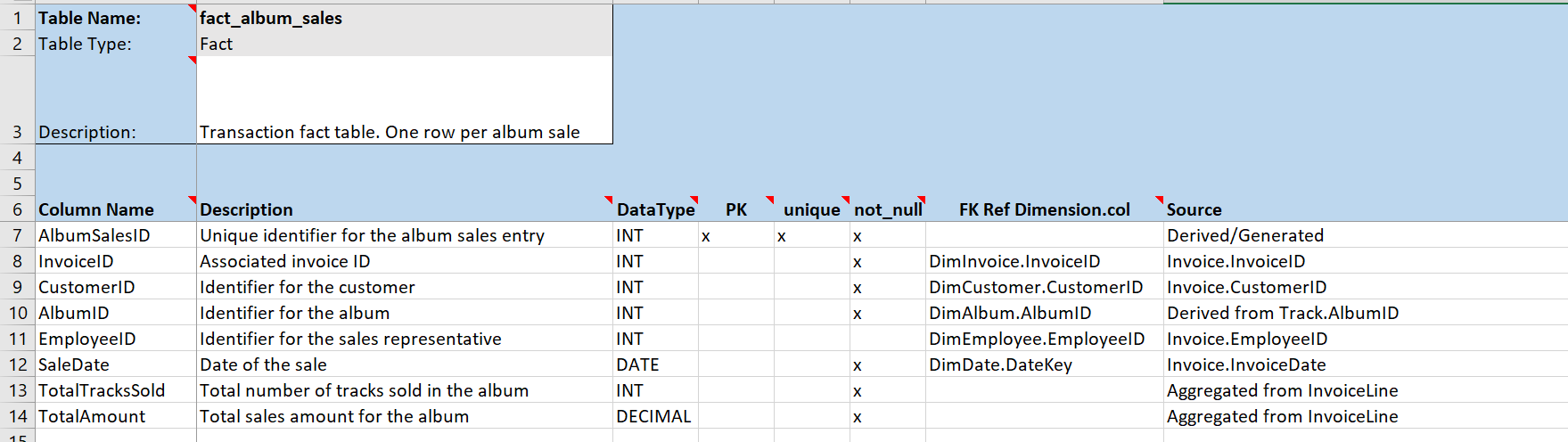
**dim\_customer**

****

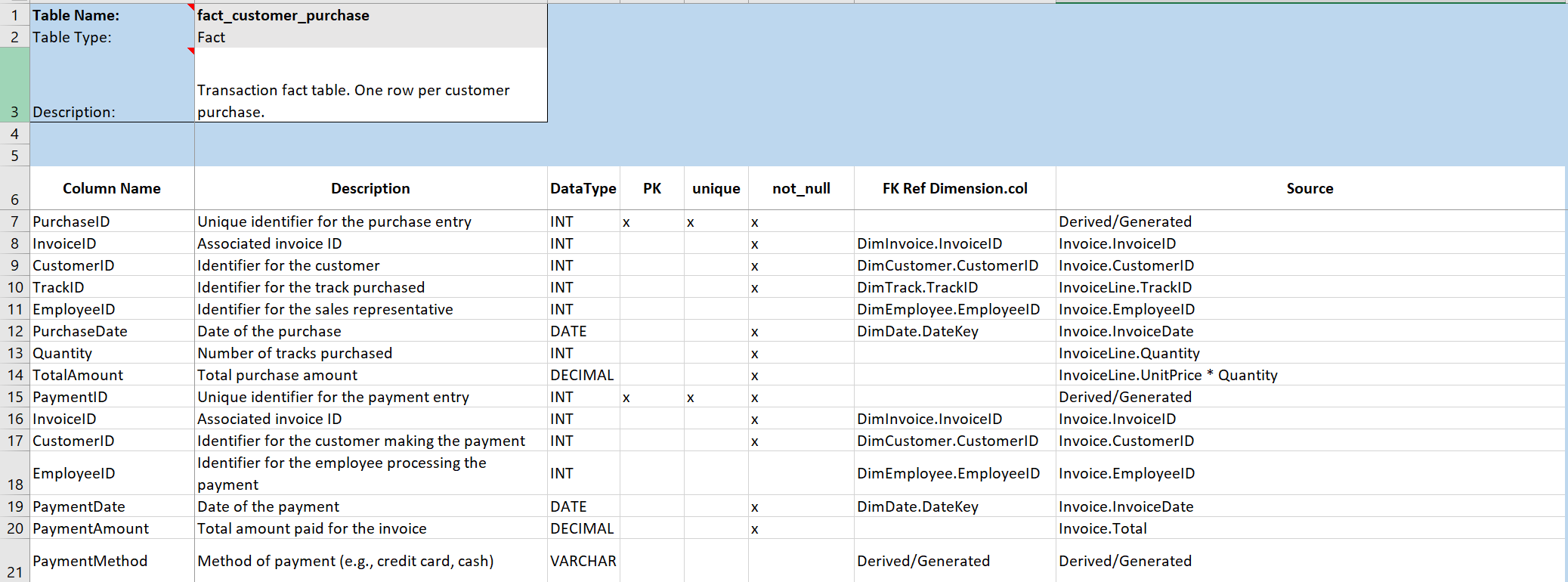
**fact\_sales\_song**

****

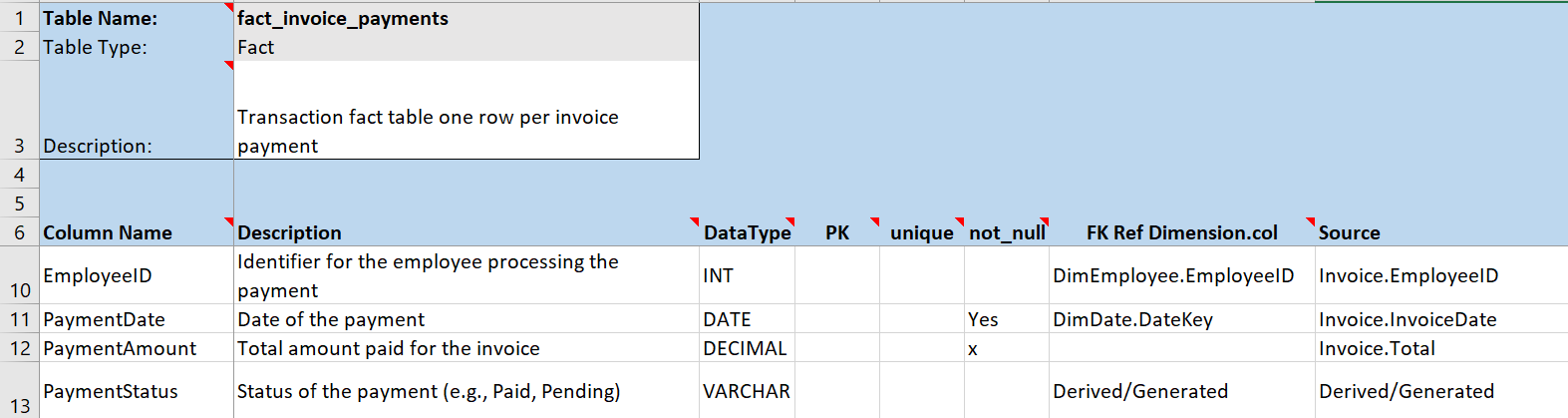
**fact\_album\_sales**

****

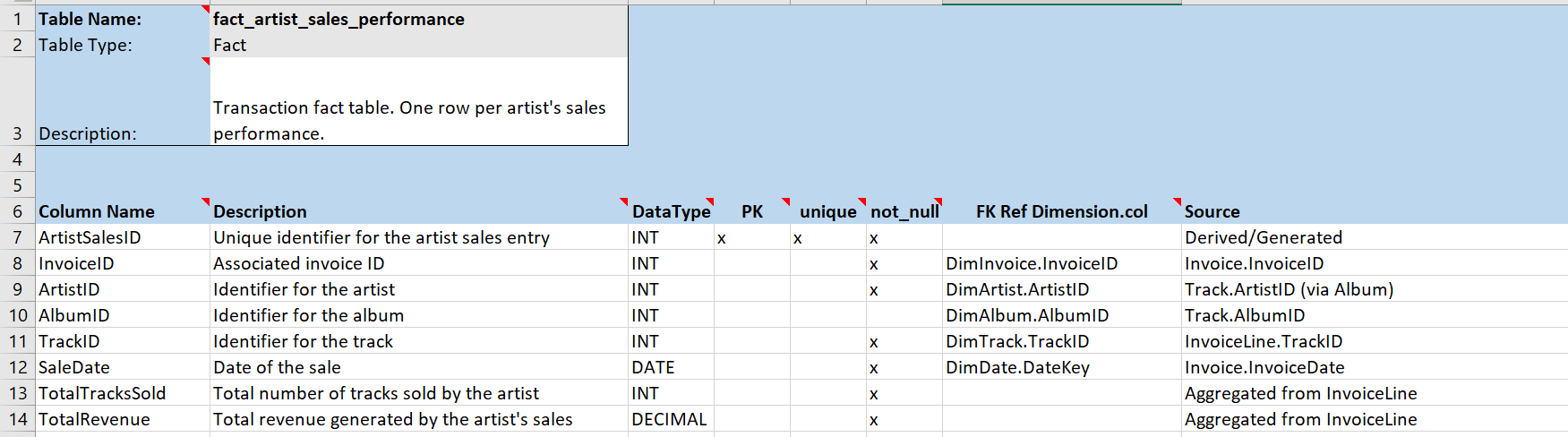
**fact\_customer\_purchase**

****

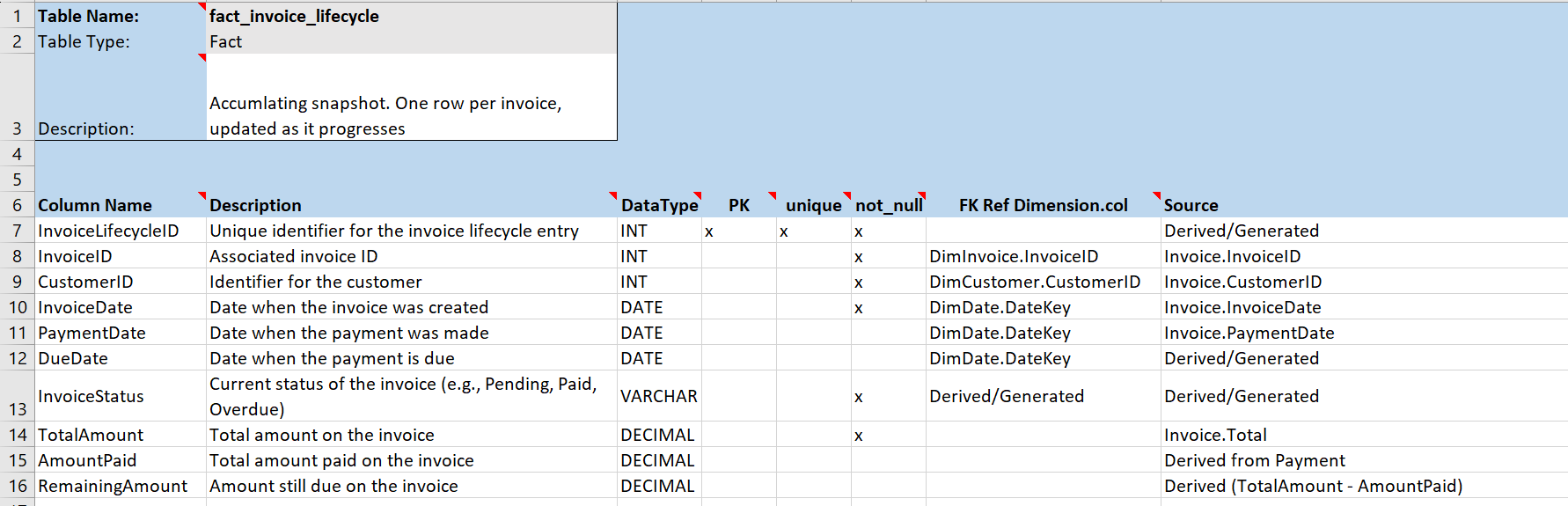
**fact\_invoice\_payments**

****

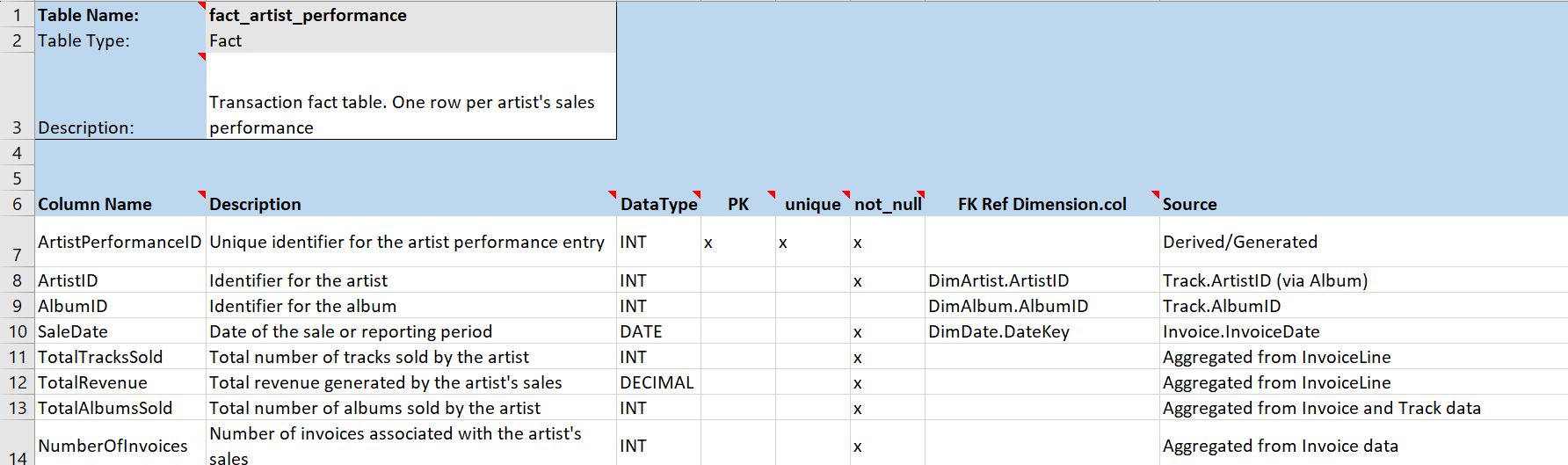
**fact\_artist\_sales\_performance**

****

**fact\_invoice\_lifecycle**

****

**fact\_artist\_performance**

****

**Data Warehouse Implementation**

We have our data in Snowflake. We have used dbt for ETL implementation. Stored the dimension tables and fact tables in dbt.

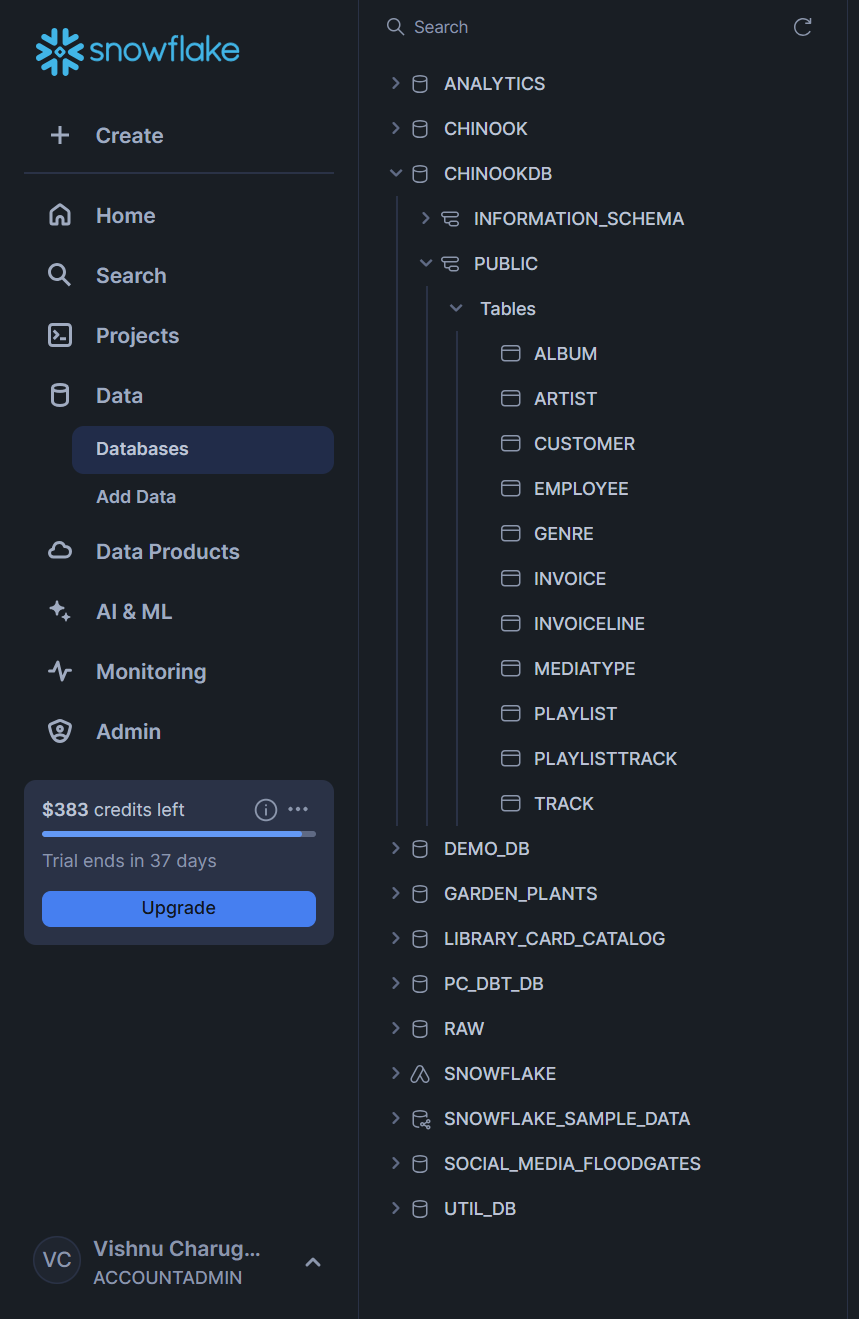
**1. Overview**

* Platform: Snowflake Data Warehouse
* ETL Tool: dbt (data build tool)
* Architecture: Kimball dimensional modeling
* Implementation Pattern: Star Schema

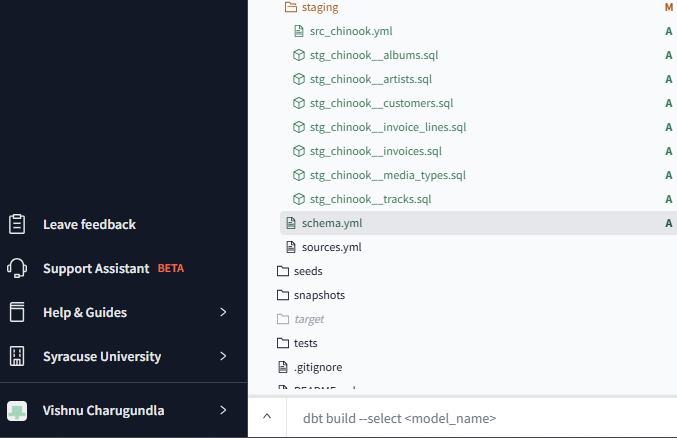
**2. Data Architecture**

* **Source Layer (Snowflake)**
  + Database: CHINOOKDB
  + Schema: PUBLIC
  + Source Tables: ALBUM, ARTIST, CUSTOMER, EMPLOYEE, GENRE, INVOICE, INVOICELINE, MEDIATYPE, PLAYLIST, PLAYLISTTRACK, TRACK

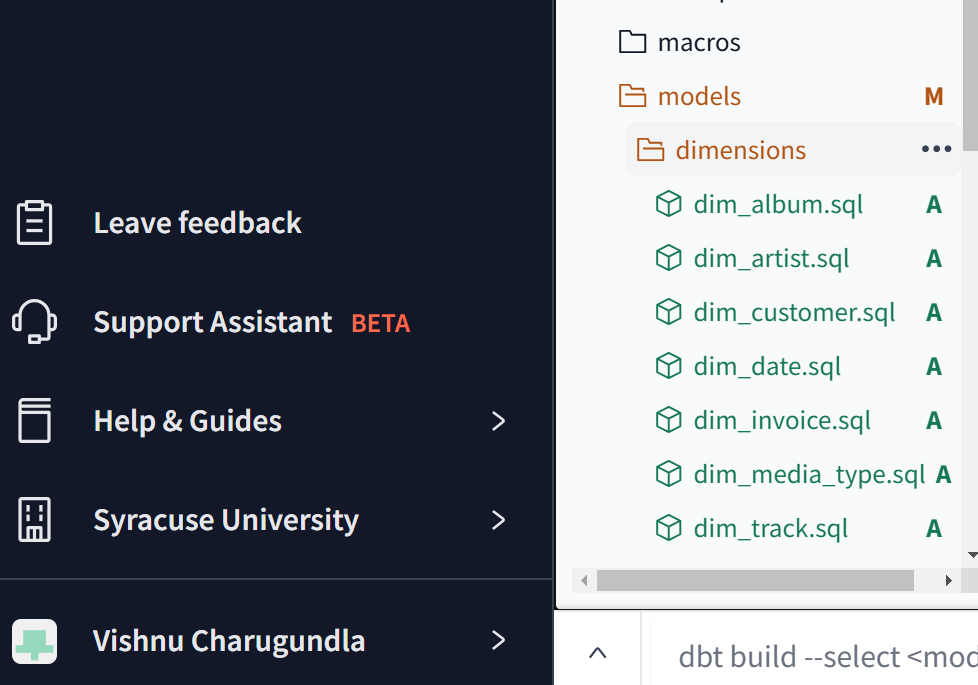
**Data residing in Snowflake**



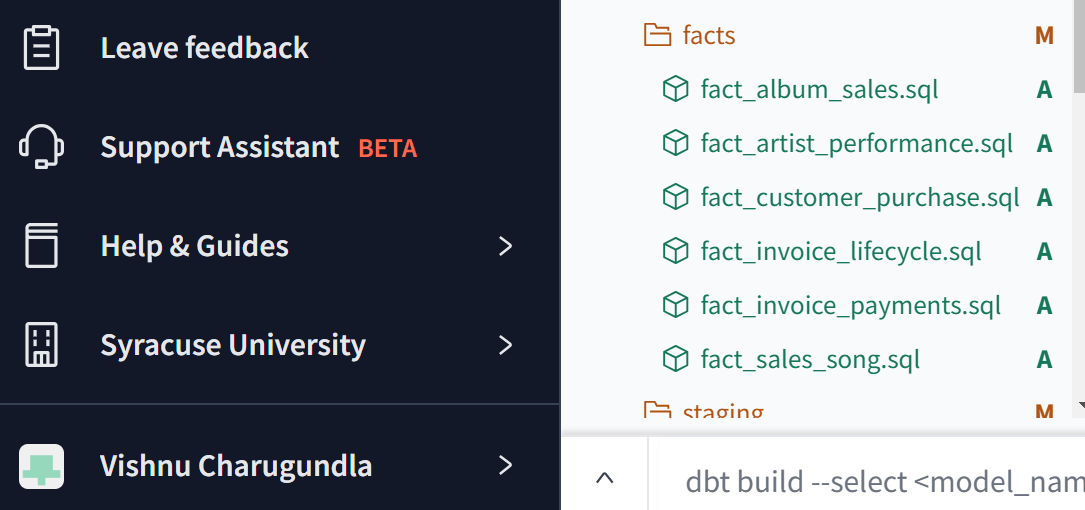
* **Staging Layer (dbt)**
  + Clean and standardized source data
  + Consistent naming conventions
  + Initial data quality checks
  + Key staging models:
    - stg\_chinook\_\_albums
    - stg\_chinook\_\_artists
    - stg\_chinook\_\_customers
    - stg\_chinook\_\_invoice\_lines
    - stg\_chinook\_\_invoices
    - stg\_chinook\_\_tracks
    - stg\_chinook\_\_media\_types



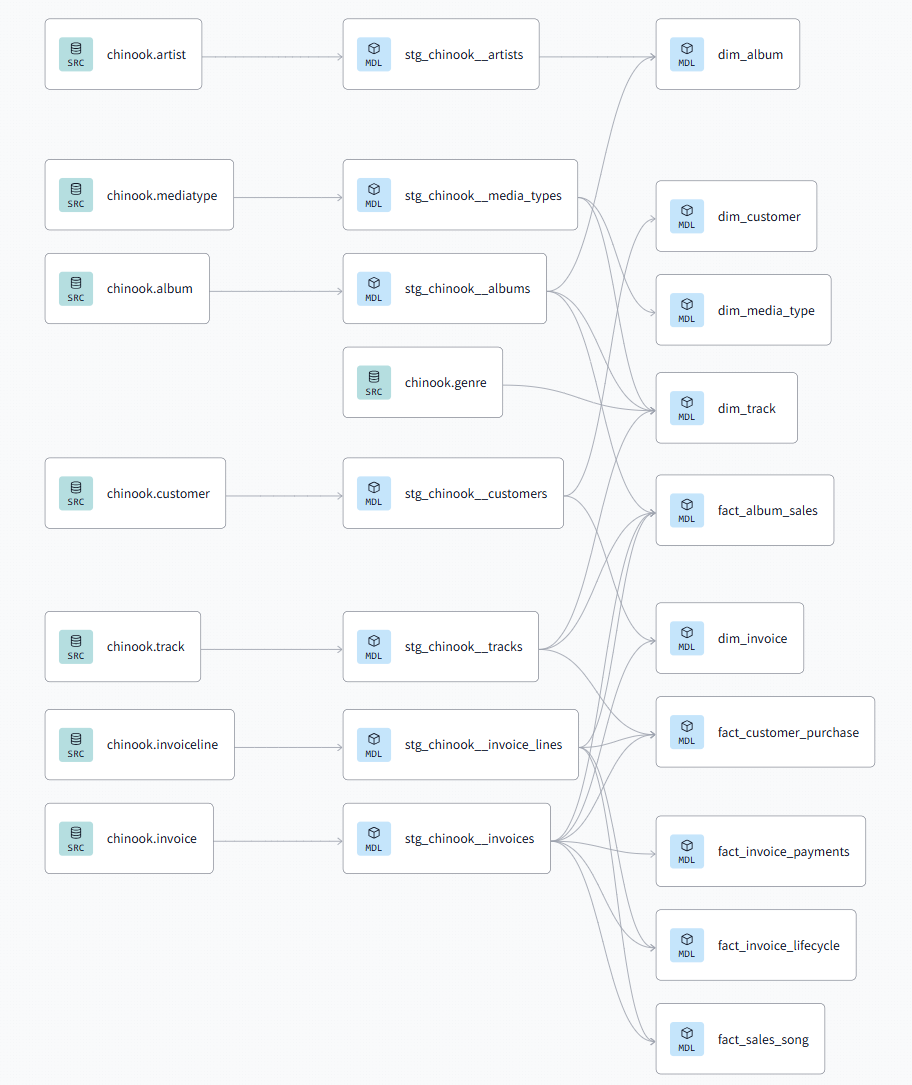
* **Dimensional Layer (dbt)**
  + Conformed dimensions following Kimball methodology
  + Implemented dimensions:
    - dim\_date (Time dimension)
    - dim\_customer (Customer information)
    - dim\_album (Album details)
    - dim\_artist (Artist information)
    - dim\_track (Track information)
    - dim\_invoice (Invoice details)
    - dim\_media\_type (Media type classification)



* **Fact Tables (dbt)**
  + Implemented fact tables:
    - fact\_sales\_song (Song-level sales transactions)
    - fact\_album\_sales (Album-level sales aggregations)
    - fact\_customer\_purchase (Customer purchase patterns)
    - fact\_invoice\_payments (Payment tracking)
    - fact\_artist\_performance (Artist performance metrics)
    - fact\_invoice\_lifecycle (Invoice status tracking)



**Lineage: Source to target maps**

****

**Data Flow**

1. Source data resides in Snowflake CHINOOKDB
2. dbt staging models clean and standardize the data
3. Dimension tables are built from staging models
4. Fact tables integrate with dimensions using surrogate keys
5. Final tables are materialized in Snowflake for analytics

**Key Features**

* Surrogate key implementation for all dimensions
* Standardized naming conventions
* Automated testing for data quality
* Documentation and lineage tracking
* Incremental processing capability
* Version control and change management

**Implementation Best Practices**

* All tables are properly documented
* Data quality tests implemented
* Consistent naming conventions followed
* Surrogate keys for all dimensions
* Proper referential integrity maintained

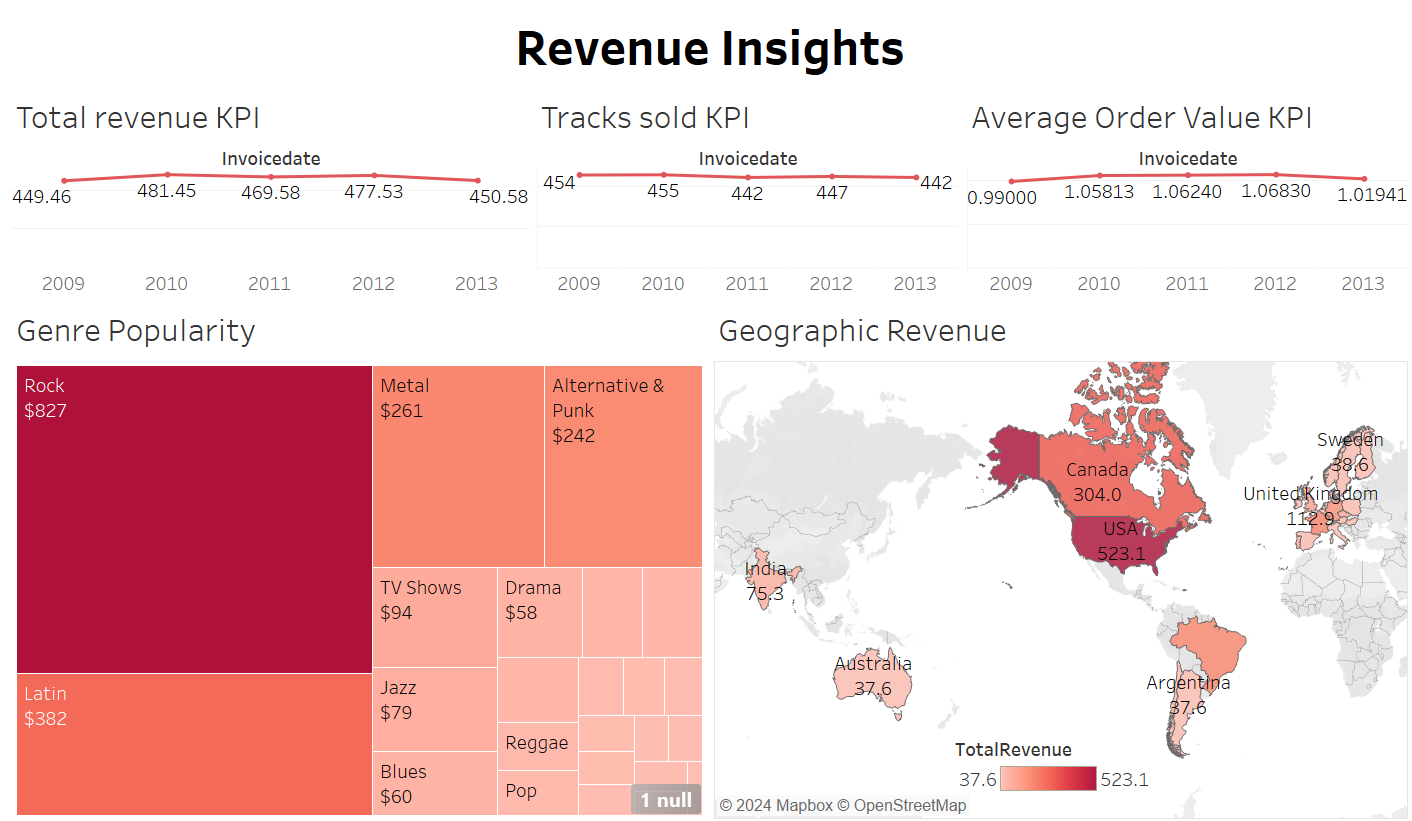
**Data Quality Measures**

* Unique key constraints
* Not-null constraints for critical fields
* Referential integrity checks
* Data freshness validation
* Value range check

**BI implementation:**

Implemented Tableau Dashboards for the BI section of our Datawarehouse project.

**Revenue Insights Dashboard Analysis**

****

**Key Metrics:**

**- Total Revenue KPI (2009-2013):** Shows revenue trend peaking at $481.45 in 2010

**- Tracks Sold KPI:** Consistent performance averaging ~450 tracks annually

**- Average Order Value:** Slight upward trend from $0.99 to $1.07

**Key Visualizations:**

**1. Genre Revenue Tree Map**

- Rock dominates ($827)

- Latin second highest ($382)

- Other genres show varying performance

**2. Geographic Revenue Distribution**

- USA leads ($523.1)

- Canada second ($304.0)

- Strong presence in European markets

**Business Value:**

- Tracks revenue performance across time

- Identifies top-performing genres

- Shows geographic market strengths

- Enables data-driven decisions for inventory and market expansion

**KPIs Monitored:**

- Revenue trends

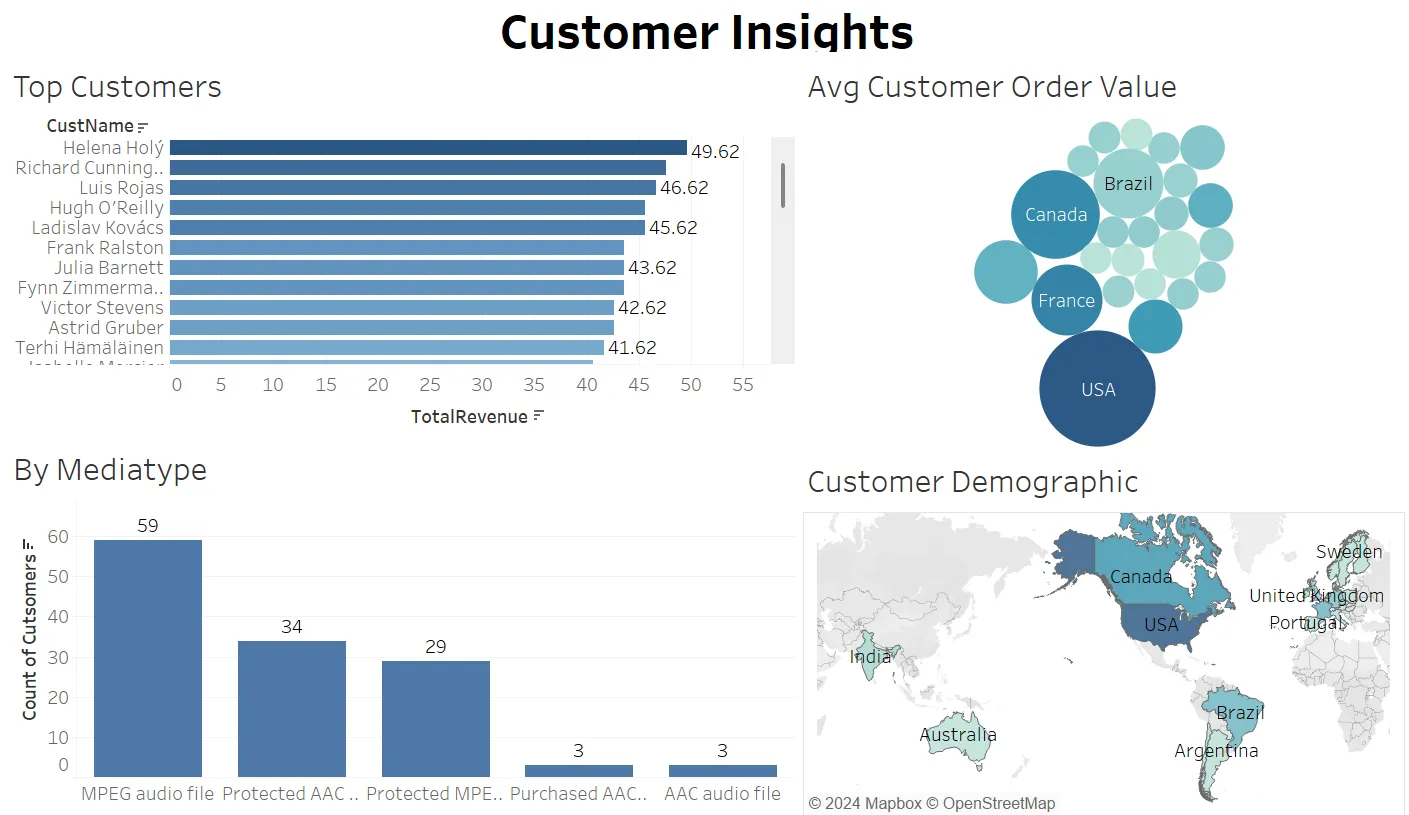
- Sales volume

- Average order value

- Geographic distribution

- Genre performance

**Customer Insights Dashboard Analysis:**



**Key Visualizations:**

**1. Top Customers:**

- Helena Holy leads with $49.62 revenue

- Top 10 customers range from $41-50 in spending

**2. Media Type Distribution:**

- MPEG audio dominates (59 customers)

- Protected AAC second (34 customers)

- Shows format preferences

**3. Customer Demographics:**

- Bubble chart showing order value by country

- Geographic distribution map

- USA and Canada lead customer base

**Business Value:**

- Identifies high-value customers

- Shows preferred media formats

- Reveals geographic customer concentrations

- Enables targeted marketing strategies

**KPIs Tracked:**

- Customer lifetime value

- Media format adoption

- Geographic distribution

- Average order value by region

**Project Roster:**

**1. Bus Matrix Development**

Team Member: Vishnu Charugundla, Rishikesh Ramesh

Details: Created comprehensive business process mapping for Chinook Music Store, identifying key processes like sales tracking, customer management, and inventory control.

**2. Detailed Dimensional Modeling**

Team Member: Vishnu Charugundla, Pranav Mahesh Mekal

Details: Designed and implemented Detailed Dimensional Modeling for Chinook database, including fact tables for sales and dimensions for customers, products, and time.

**3. DBT Cloud Setup**

Team Member: Vishnu Charugundla, Pranav Mahesh Mekal

Established DBT Cloud environment, configured version control, and set up development workflows for data transformation.

**4. Snowflake Data Loading**

Team Members: Vishnu Charugundla, Rishikesh Ramesh

Details: Successfully loaded Chinook database into Snowflake, configured warehouses, and established data quality checks.

**5. DBT and Snowflake Integration**

Team Members: Vishnu Charugundla, Pranav Mahesh Mekal

Details: Integrated DBT models with Snowflake, creating staging, dimension, and fact models for efficient data transformation.

**6. Tableau Visualization**

Team Members: Pranav Mahesh Mekal, Rishikesh Ramesh

Details: Created interactive dashboards for sales analytics, customer insights, and product performance metrics.

**7. Report Writing**

Team Members: Vishnu Charugundla, Rishikesh Ramesh

Details: Documented project methodology, technical implementation, and business process analysis in comprehensive report.

**8. Presentation**

Team Members: Vishnu Charugundla

Details: Developed presentation showcasing information about data and other important insights.