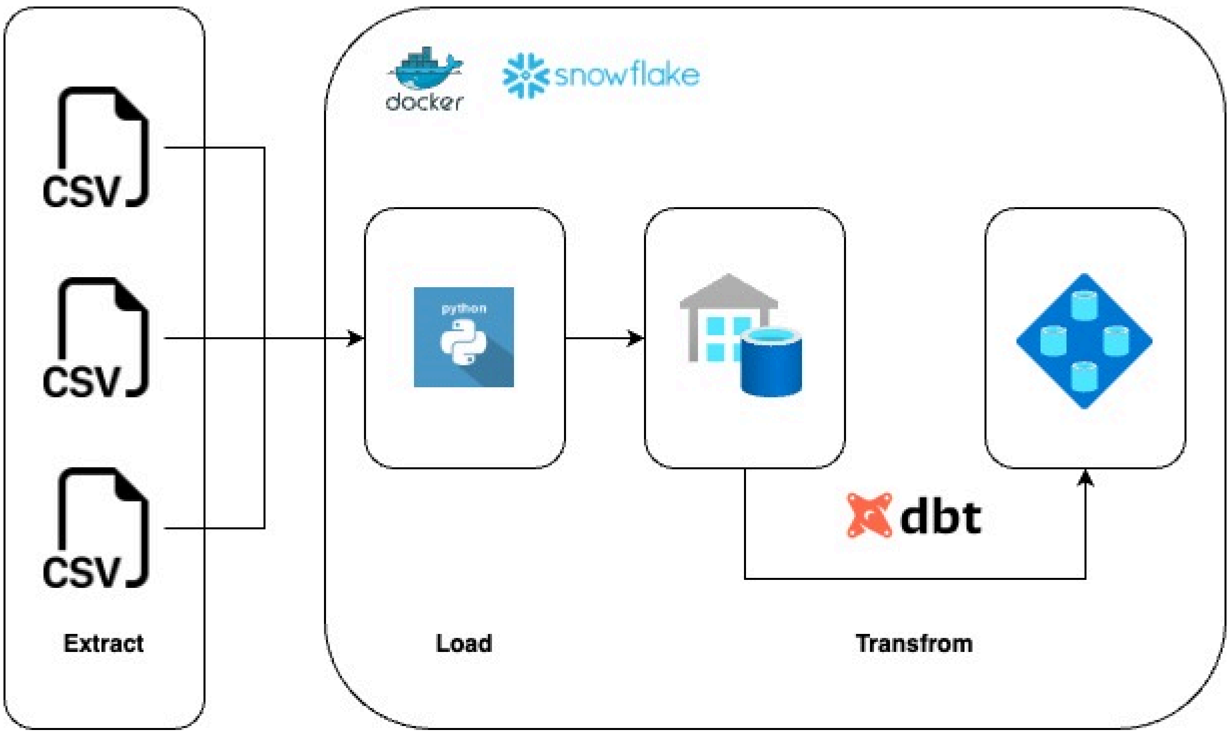
ELT Data Pipeline with Snowflake, dbt, and Docker

Table of Contents

Project Overview

This project is designed to demonstrate an ELT (Extract, Load, Transform) data pipeline using Snowflake as the data warehouse, dbt (data build tool) for transformations, and Docker for containerization. The pipeline extracts data from source systems, loads it into Snowflake, and then transforms it using dbt models.

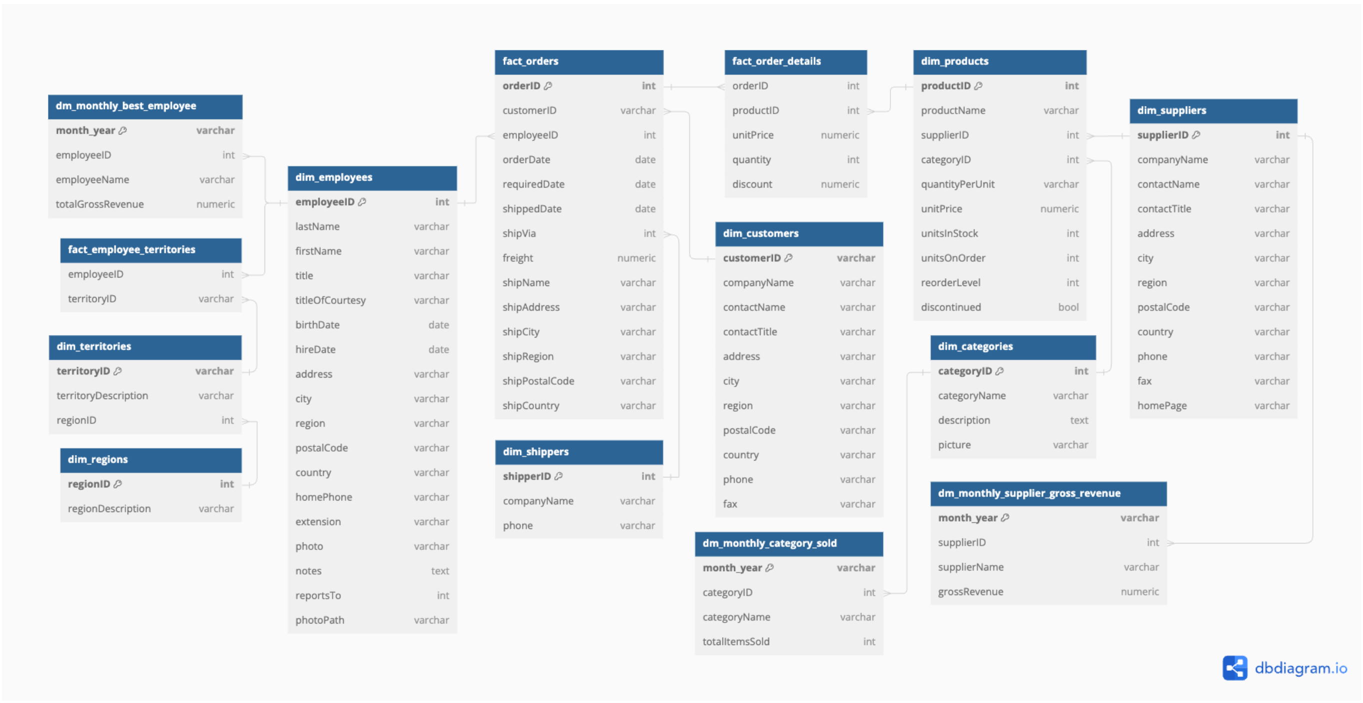
Architecture



1. **Extract**: Data is extracted from source systems.
2. **Load**: Extracted data is loaded into Snowflake.

3. **Transform**: Data is transformed in Snowflake using dbt models.

Data Warehouse Schema



The provided schema represents a relational database designed to store and manage data related to territories, suppliers, shippers, regions, products, customers, categories, employees, orders, and various monthly metrics. source

Prerequisites

- Docker
- Docker Compose
- Snowflake account
- dbt CLI
- Git

Setup and Installation

Clone the Repository

```
git clone https://github.com/lanaahm/elt-snowflake.git
cd elt-snowflake
```

Environment Variables

Create a `.env` file in the project root and add the following environment variables:

```
SNOWFLAKE_ACCOUNT=your_snowflake_account
SNOWFLAKE_USER=your_snowflake_user
SNOWFLAKE_PASSWORD=your_snowflake_password
```

```
SNOWFLAKE_ROLE=your_snowflake_role
SNOWFLAKE_WAREHOUSE=your_snowflake_warehouse
SNOWFLAKE_DATABASE=your_snowflake_database
SNOWFLAKE_SCHEMA=your_snowflake_schema
```

Docker Setup

Build and run the Docker containers:

```
docker-compose build
```

This command will build the Docker images in the docker-compose.yml file.

Usage

Running the Pipeline

To run the entire ELT pipeline:

1. Extract data and load it into Snowflake:

```
docker-compose run --rm snowflake_elt
```

2. Debug dbt connection

```
docker-compose run --rm snowflake_dbt debug
```

3. Run dbt model:

```
docker-compose run --rm snowflake_dbt run
```

4. Generate dbt docs:

```
docker-compose run --rm snowflake_dbt docs generate
```

5. Serve dbt docs:

```
docker-compose run --rm snowflake_dbt docs serve
```

Testing

To run tests on the dbt models:

```
docker-compose run --rm snowflake_dbt test
```

Project Structure

```
elt-snowflake/
│
├── snowflake_dbt/
│   ├── models/
│   ├── snapshots/
│   ├── target/
│   ├── tests/
│   ├── Dockerfile
│   ├── schema.yml
│   └── dbt_project.yml
│
├── snowflake_elt/
│   ├── .env
│   ├── config.py
│   ├── main.py
│   ├── Dockerfile
│   └── requirements.txt
│
├── .editorconfig
├── .gitignore
├── profiles.yml
├── docker-compose.yml
└── README.md
```

- elt-snowflake/: The root directory for the entire project.
- snowflake_dbt/: Directory containing all the dbt-related configurations and files.
  - models/: Directory for dbt models which define the structure of your data transformations.
  - snapshots/: Directory for dbt snapshots, which capture the state of your data at regular intervals.
  - target/: Directory where dbt compiles your SQL files and stores the output.
  - tests/: Directory for dbt tests, ensuring the correctness of your models.
  - Dockerfile: Dockerfile specific to the dbt service, for containerizing the dbt project.
  - schema.yml: YAML file defining the schema for dbt models and tests.
  - dbt_project.yml: Configuration file for the dbt project, specifying project settings and configurations.
- snowflake_elt/: Directory containing Python scripts for data extraction and loading.
  - .env: Environment variable definitions for the Python scripts, such as credentials and configuration settings.
  - config.py: Python configuration file, possibly containing settings and constants used in the ETL process.

    - main.py: Main Python script for executing the ETL pipeline.
    - Dockerfile: Dockerfile specific to the ETL service, for containerizing the ETL scripts.
    - requirements.txt: File listing Python dependencies required for the ETL process.
- .editorconfig: Configuration file for maintaining consistent coding styles across different editors and IDEs.
- .gitignore: File specifying which files and directories should be ignored by Git, preventing them from being committed to the repository.
- profiles.yml: Configuration file for dbt profiles, defining connection details for different environments (e.g., development, production).
- docker-compose.yml: Docker Compose configuration file for defining and running multi-container Docker applications, orchestrating the services defined in the project.
- README.md: Project documentation file, providing an overview of the project, setup instructions, and other relevant information.

Welcome to your new dbt project!

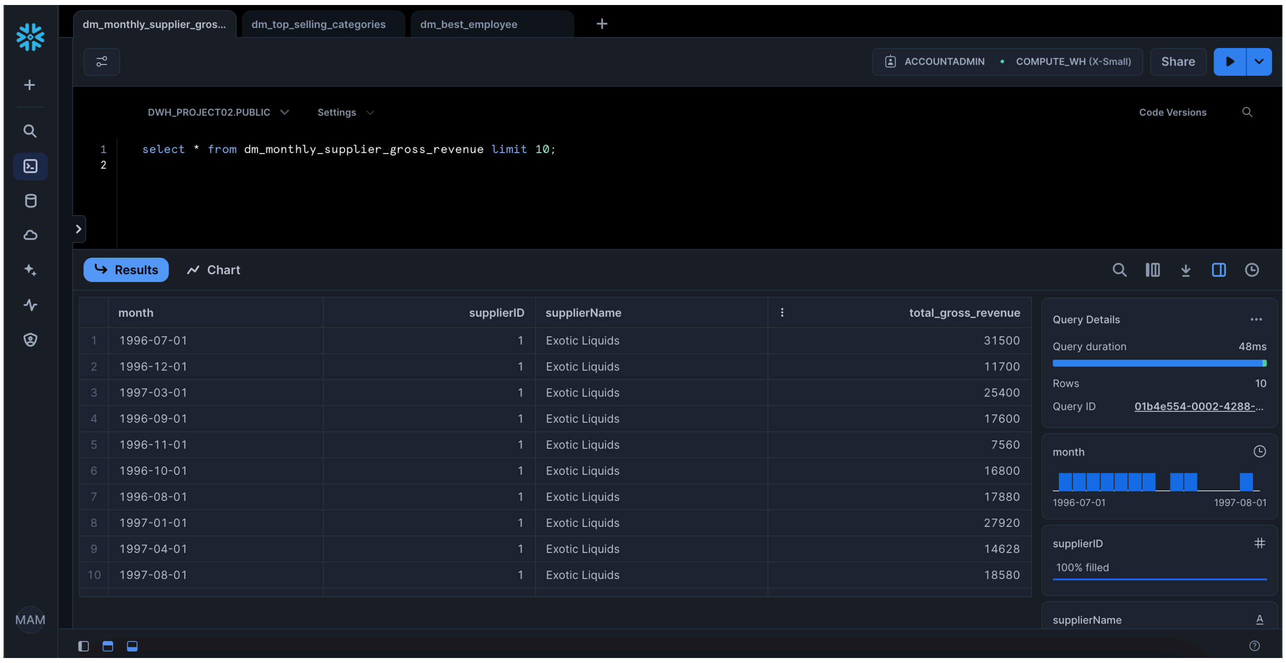
Using the starter project

Try running the following commands:

- dbt run
- dbt test

Results data mart

1. Monthly supplier gross revenue

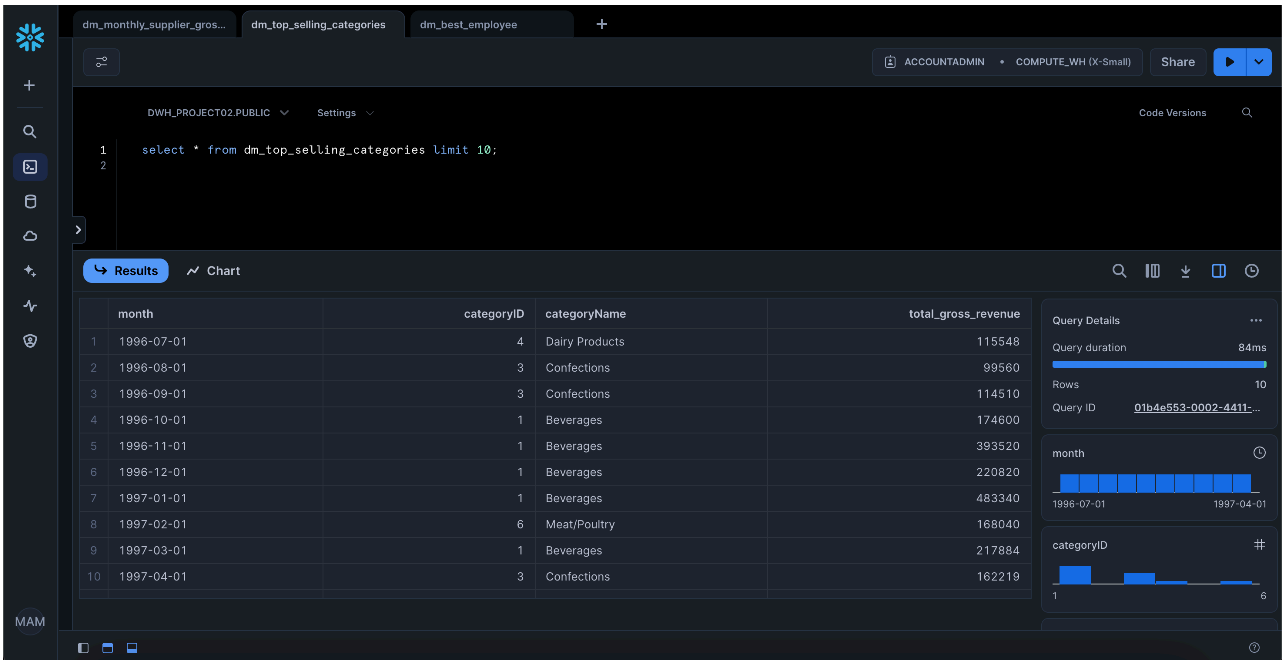


```sql
select * from dm_monthly_supplier_gross_revenue limit 10;
```

	month	supplierID	supplierName	total_gross_revenue
1	1996-07-01	1	Exotic Liquids	31500
2	1996-12-01	1	Exotic Liquids	11700
3	1997-03-01	1	Exotic Liquids	25400
4	1996-09-01	1	Exotic Liquids	17600
5	1996-11-01	1	Exotic Liquids	7560
6	1996-10-01	1	Exotic Liquids	16800
7	1996-08-01	1	Exotic Liquids	17880
8	1997-01-01	1	Exotic Liquids	27920
9	1997-04-01	1	Exotic Liquids	14628
10	1997-08-01	1	Exotic Liquids	18580

Query Details
Query duration 48ms
Rows 10
Query ID 01b4e554-0002-4288-...
month
1996-07-01 1997-08-01
supplierID
100% filled
supplierName

2. Top selling categories

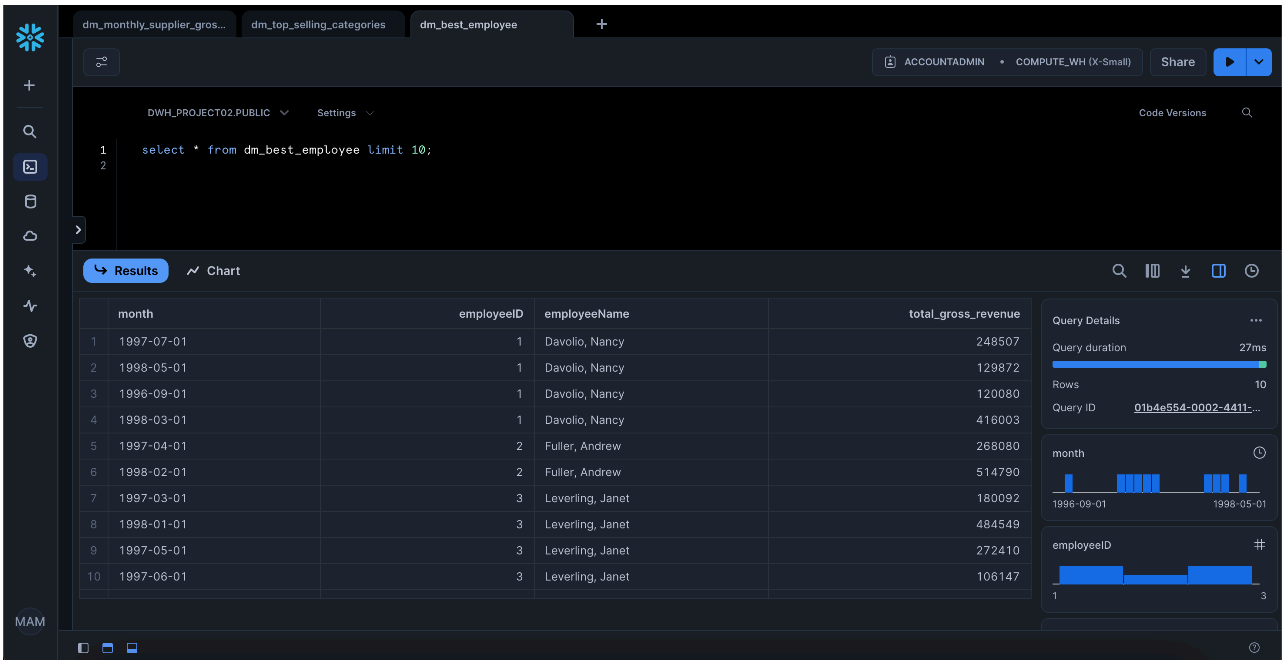


```sql
select * from dm_top_selling_categories limit 10;
```

	month	categoryID	categoryName	total_gross_revenue
1	1996-07-01	4	Dairy Products	115548
2	1996-08-01	3	Confections	99560
3	1996-09-01	3	Confections	114510
4	1996-10-01	1	Beverages	174600
5	1996-11-01	1	Beverages	393520
6	1996-12-01	1	Beverages	220820
7	1997-01-01	1	Beverages	483340
8	1997-02-01	6	Meat/Poultry	168040
9	1997-03-01	1	Beverages	217884
10	1997-04-01	3	Confections	162219

Query Details
Query duration 84ms
Rows 10
Query ID 01b4e553-0002-4411-...
month
1996-07-01 1997-04-01
categoryID
1 6

3. Best employee


dm_monthly_supplier_gros...
dm_top_selling_categories
dm_best_employee
ACCOUNTADMIN • COMPUTE_WH (X-Small)
Share
DWH_PROJECT02.PUBLIC
Settings
Code Versions
Results
Chart
MAM

```
select * from dm_best_employee limit 10;
```

	month	employeeID	employeeName	total_gross_revenue
1	1997-07-01	1	Davolio, Nancy	248507
2	1998-05-01	1	Davolio, Nancy	129872
3	1996-09-01	1	Davolio, Nancy	120080
4	1998-03-01	1	Davolio, Nancy	416003
5	1997-04-01	2	Fuller, Andrew	268080
6	1998-02-01	2	Fuller, Andrew	514790
7	1997-03-01	3	Leverling, Janet	180092
8	1998-01-01	3	Leverling, Janet	484549
9	1997-05-01	3	Leverling, Janet	272410
10	1997-06-01	3	Leverling, Janet	106147

Query Details
Query duration 27ms
Rows 10
Query ID 01b4e554-0002-4411-...

month 1996-09-01 1998-05-01
employeeID 1 3

Resources:

- Learn more about dbt in the docs
- Check out Discourse for commonly asked questions and answers
- Join the chat on Slack for live discussions and support
- Find dbt events near you
- Check out the blog for the latest news on dbt's development and best practices