ETL Setup Guide

This guide will help you set up and run the ETL pipeline for the BrazilRetail-BI project.

Overview

The project now supports two ETL environments:

1. **Local ETL (**`etl_local/`**)**: Loads data into a local PostgreSQL database using SQLAlchemy.
2. **Production ETL (**`etl_prod/`**)**: Loads data directly into Supabase using the Supabase Python client.

Prerequisites

- Python 3.8+
- PostgreSQL database (for Local ETL)
- Supabase project (for Production ETL)
- Dataset downloaded (see dataset_setup.md)

Step 1: Install Dependencies

Install the required Python packages:

```
pip install -r requirements.txt
pip install supabase  # Required for Production ETL
```

Step 2: Configure Environment

Update the `.env` file with your credentials:

```
# For Local ETL

DATABASE_URL=postgresql://your_username:your_password@localhost:5432/brazilretail_bi
# For Production ETL
SUPABASE_URL=https://your-project.supabase.co
SUPERKEY=your-service-role-key
```

Step 3: Run Local ETL

1. Create the local database schema:

   ```
   python db_schema/create_schema.py
   ```

2. Setup constraints and indexes (Guardrails):

   ```
   python db_schema/setup_constraints.py
   ```

3. Run the local ETL pipeline:

   ```
   python -m etl_local.main
   ```

Step 4: Run Production ETL

The production ETL loads data directly into Supabase and supports incremental updates.

Important: Ensure your Supabase tables are created. You can run `db_schema/create_schema.py` with your Supabase `DATABASE_URL` set in `.env` if they don't exist.

1   Setup constraints and indexes (Guardrails): Ensure `DATABASE_URL` in `.env` points to your Supabase instance, then run:

```
python db_schema/setup_constraints.py
```

2   Run full load (initial setup):

```
python -m etl_prod.main --full-reload
```

3   Run incremental update (updates existing records and inserts new ones):

```
python -m etl_prod.main --incremental
```

4   Update specific tables only:

```
python -m etl_prod.main --incremental --tables orders
order_items
```

Troubleshooting

- **Supabase errors**: Verify `SUPABASE_URL` and `SUPERKEY` in `.env`. Ensure the service key has bypass RLS permissions if needed (usually it does).
- **Database connection errors**: Verify your `DATABASE_URL` in `.env` for local ETL.
- **Missing data files**: Ensure all CSV files are in the `data/` directory.

Next Steps

Once ETL is complete, you can:

- Set up Metabase for dashboard creation
- Run data analysis and create reports