As part of the Data Engineering Capstone Project from Coursera, I assumed the role of the Associate Data Warehouse Engineer who has recently joined an e-commerce organization named SoftCart.

Sofcart uses a hybrid architecture, with some of its databases on premises and some on cloud.

TOOLS AND TECHNOLOGIES:

* OLTP database – MySQL
* NoSQL database – MongoDB
* Production Data warehouse – DB2 on Cloud
* Staging Data warehouse – PostgreSQL
* Big data platform – Hadoop
* Big data analytics platform – Spark
* Business Intelligence Dashboard – IBM Cognos Analytics
* Data Pipelines – Apache Airflow

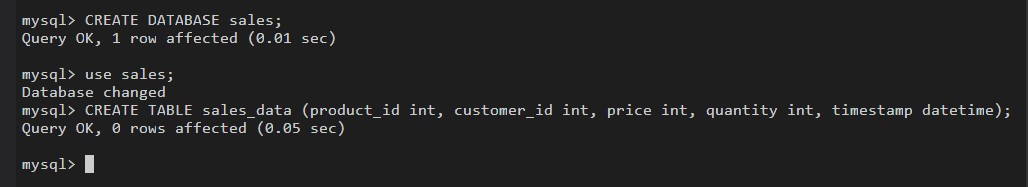
PROCESS:

* SoftCart’s online presence is primarily through its website, which customers access using a variety of devices like laptops, mobiles and tablets.
* All the catalog data of the products is stored in the MongoDB NoSQL server.
* All the transactional data like inventory and sales are stored in the MySQL database server.
* SoftCart’s webserver is driven entirely by these two databases.
* Data is periodically extracted from these two databases and put into the staging data warehouse warehouse running on PostgreSQL.
* The production data warehouse is on the cloud instance of IBM DB2 server.
* BI teams connect to the IBM DB2 for operational dashboard creation. IBM Cognos Analytics is used to create dashboards.
* SoftCart uses Hadoop cluster as its big data platform where all the data is collected for analytics purposes.
* Spark is used to analyze the data on the Hadoop cluster.
* To move data between OLTP, NoSQL and the data warehouse, ETL pipelines are used and these run on Apache Airflow.

Your company needs you to design a data platform that uses MySQL as an OLTP database. You will be using MySQL to store the OLTP data.

DESIGN THE OLTP DATABASE

Create a database named sales:

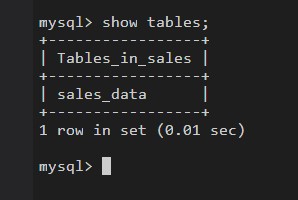


LOAD THE DATA

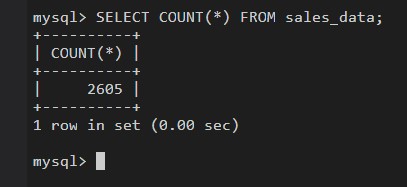
Import the data from a csv file using phpMyAdmin



List the tables in the database

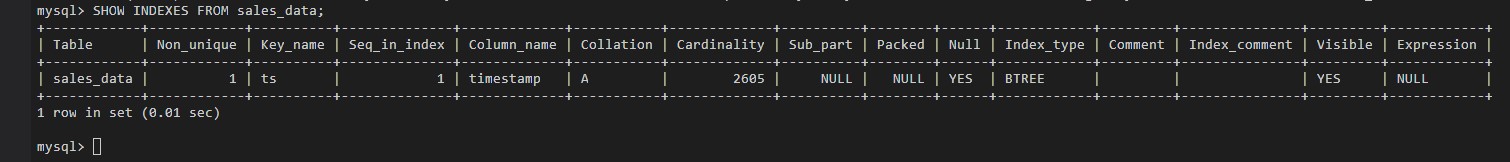


Write a query to find out the count of records in the table sales\_data

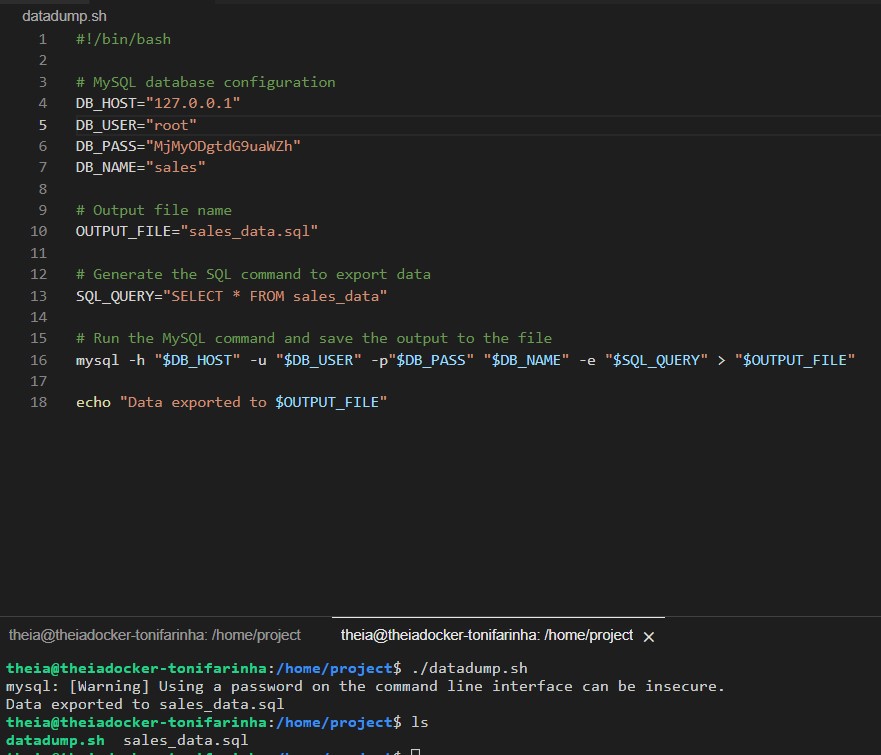


SET UP ADMIN TASKS

Create an index named ts on the timestamp field. List indexes on the table sales\_data.



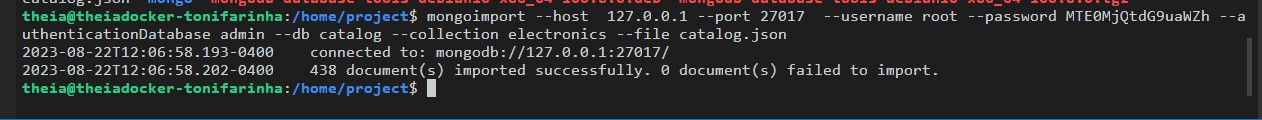
Write a bash script that exports all the rows in the sales\_daya table to a file named sales\_data.sql



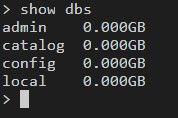
Your company needs you to design a data platform that uses MongoDB as a NoSQL database. You will be using MongoDB to store the e-commerce catalog data.

WORKING WITH MONGODB

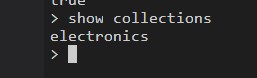
Import ‘catalagog.json’into MongoDB server into a database named ‘catalog’ and a collection named ‘electronics’



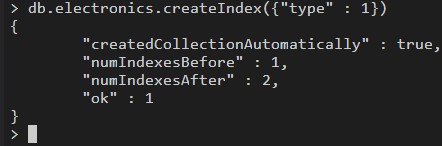
List all of the databases



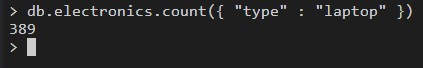
List out all of the collections in the database catalog



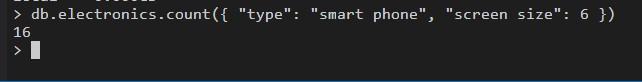
Create an index on the field ‘type’

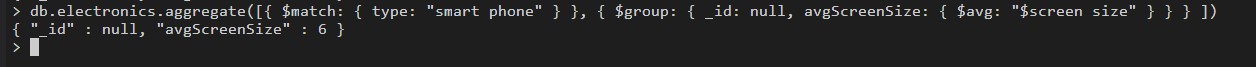


Write a query to find the count of laptops

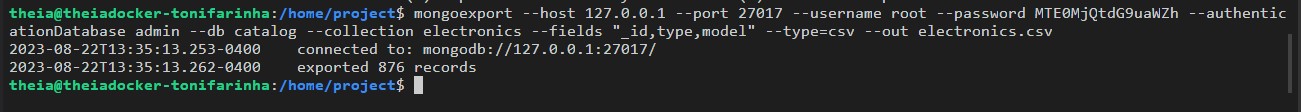


Write a query to find the number of smart phones with screen size of 6 inches



Write a query to find out the average screen size of smart phones

Export the fields ‘-id’, ‘type’, ‘model’, from the ‘electronics’collection into a file named electronics.csv



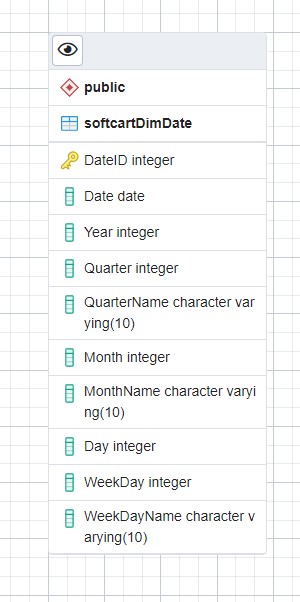
The company retails download only items like E-Books, Movies, Songs etc. The company has international presence and customers from all over the world. The company would like to create a data warehouse so that it can create reports like

* total sales per year per country
* total sales per month per category
* total sales per quarter per country
* total sales per category per country

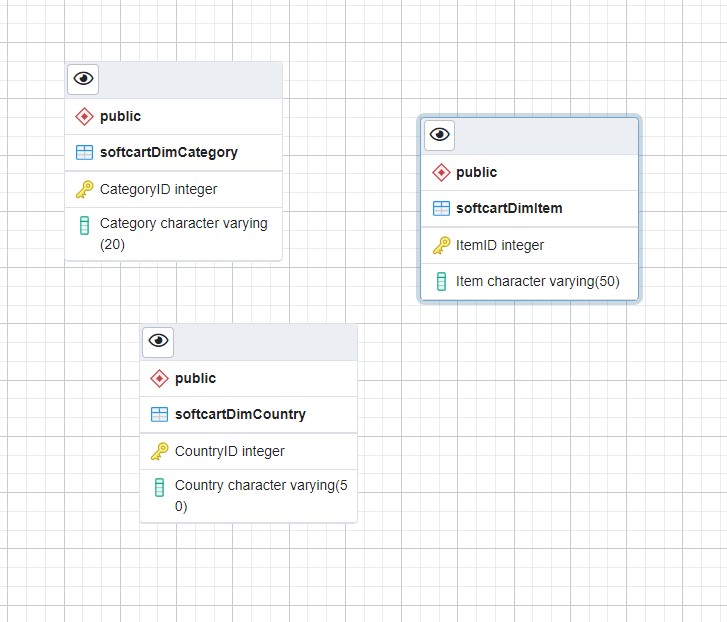
You will use your data warehousing skills to design and implement a data warehouse for the company.

DESIGN A DATA WAREHOUSE

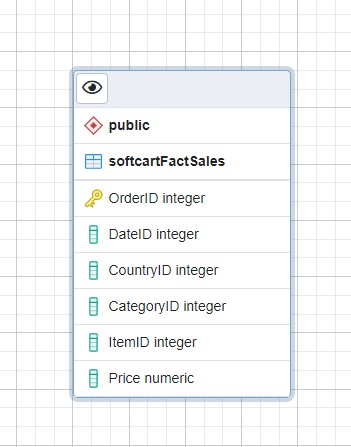
Design the dimension table softcartDimDate



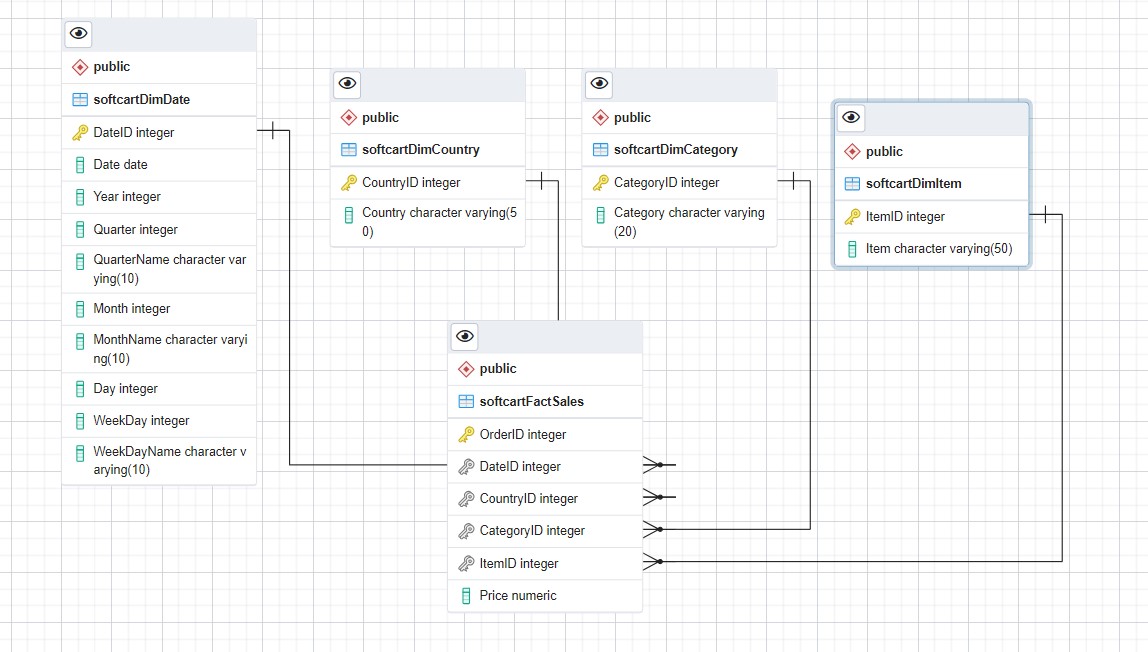
Design the dimension tables softcartDimCategory, softcartDimItem and softcartDimCountry



Design the fact table softcartFactSales

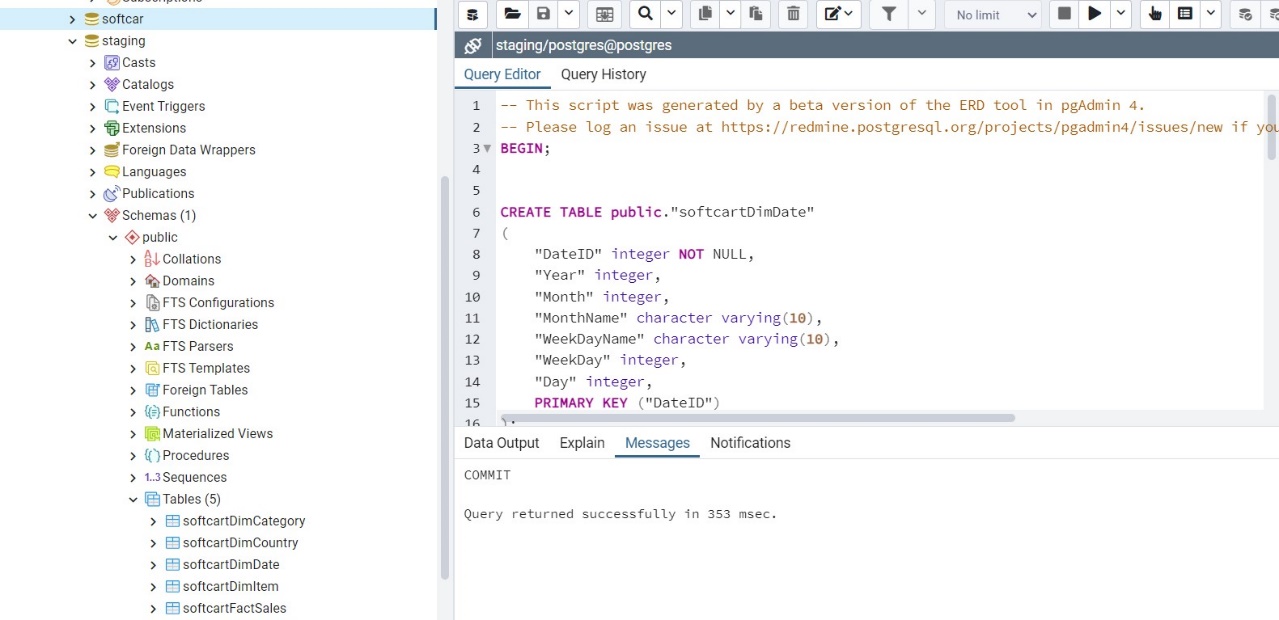


Design the relationships



CREATE THE SCHEMA

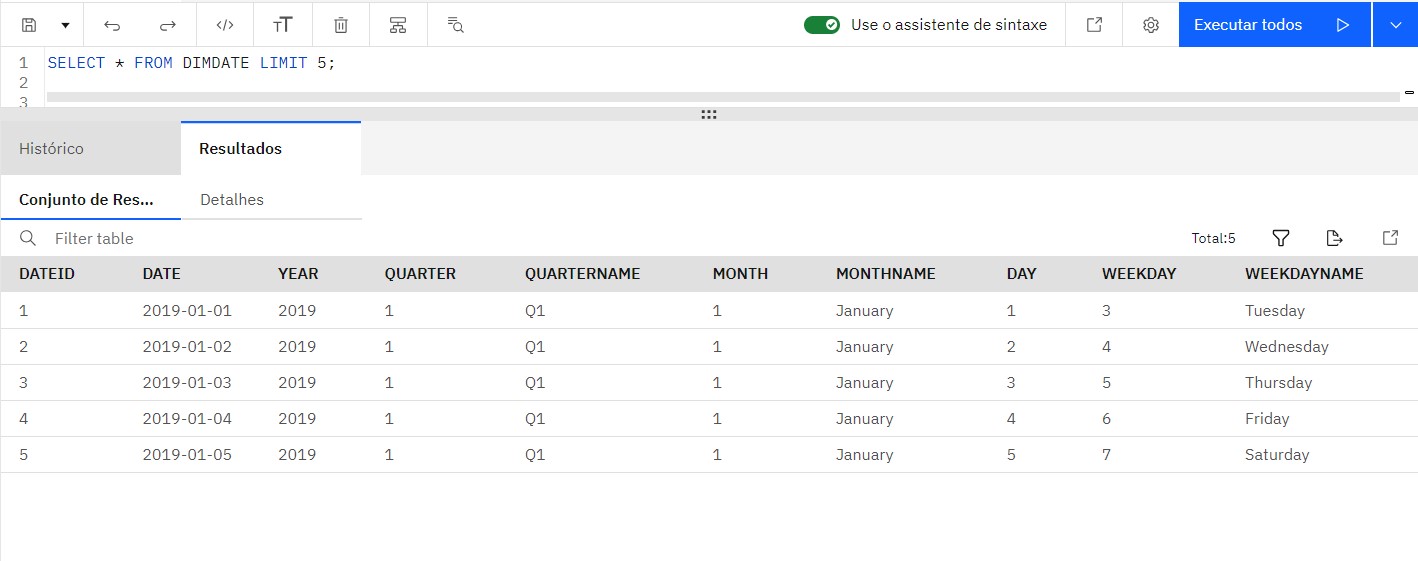
Download the schema sql from ERD tool and create the schema in a database named ‘staging’



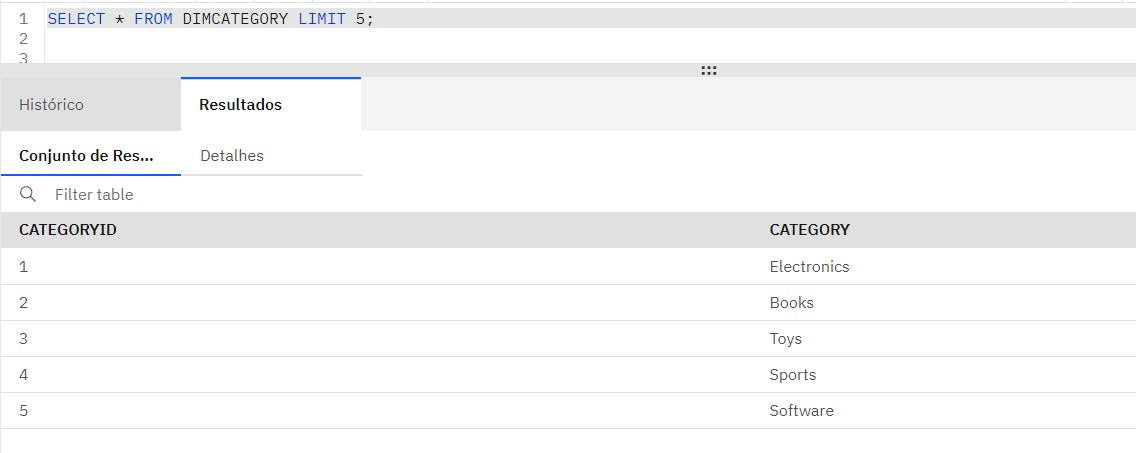
The company retails download only items like E-Books, Movies, Songs etc. The company has international presence and customers from all over the world. You have designed the schema for the data warehouse in the previous assignment. Data engineering is a team game. Your senior data engineer reviewed your design. Your schema design was improvised to suit the production needs of the company. In this assignment you will generate reports out of the data in the data warehouse using DB2

LOAD DATA INTO THE DATA WAREHOUSE

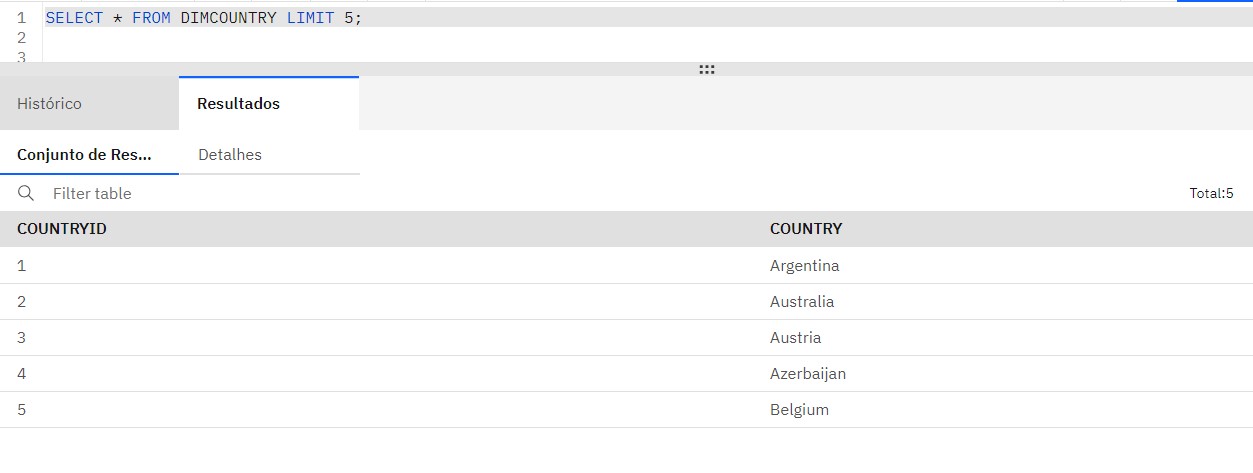
Load data into the dimension table DimDate



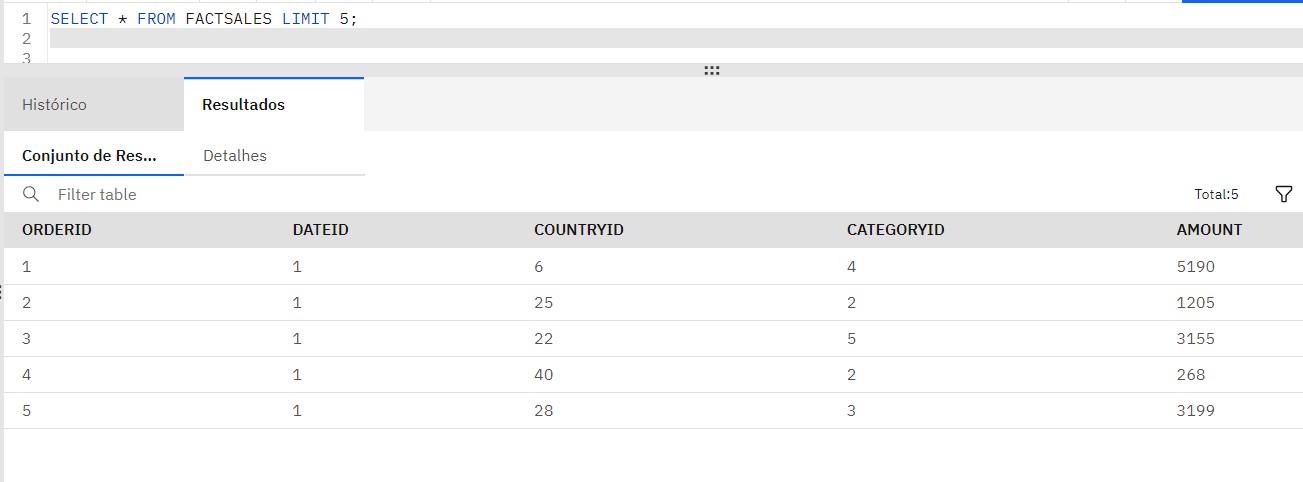
Load data into the dimension table DimCategory



Load data into the dimension table DimCountry

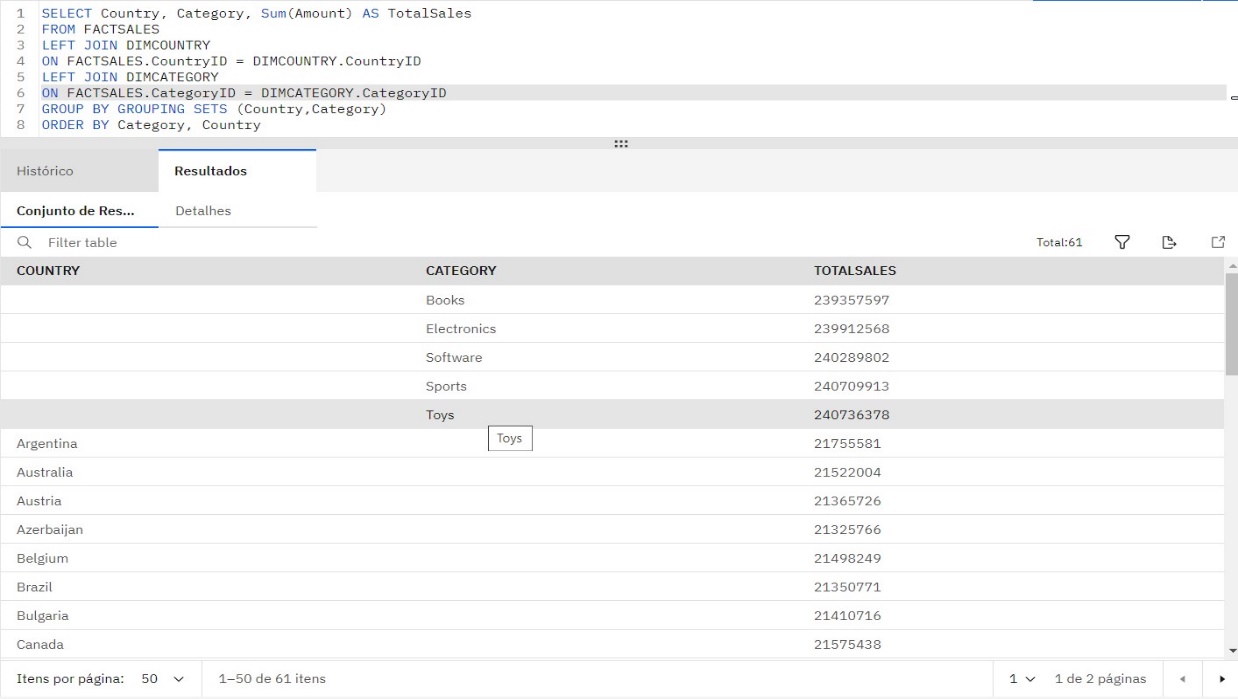


Load data into the fact table FactSales

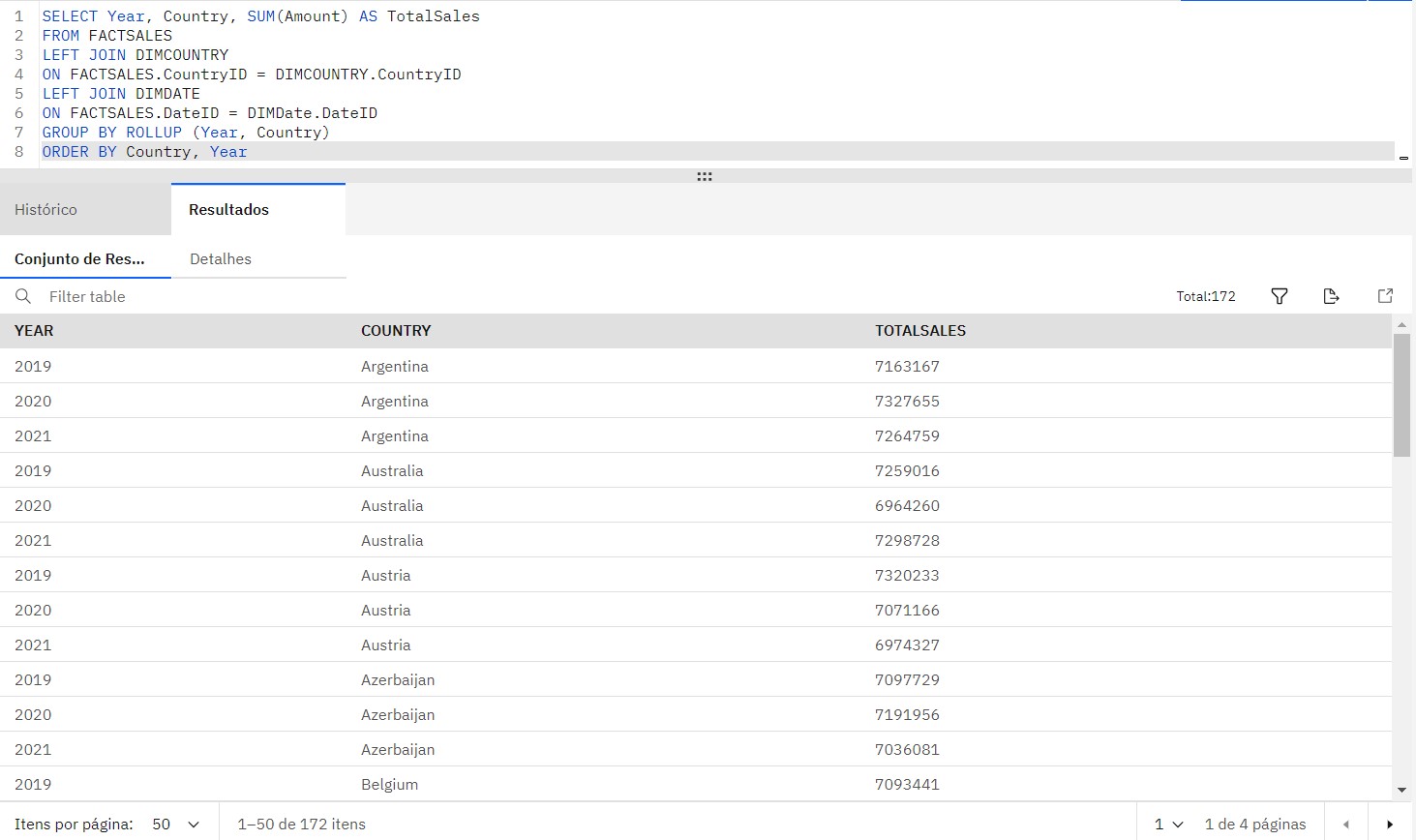


QUERIES FOR DATA ANALYTICS

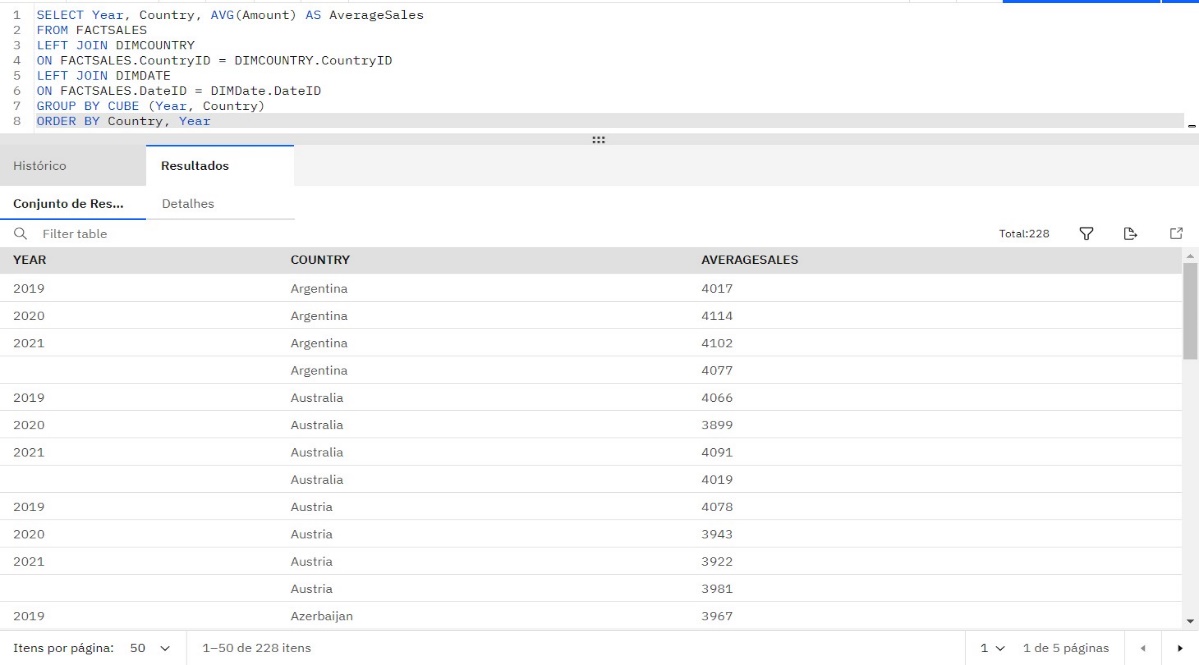
Create a grouping sets query using the columns country, category, totalsales



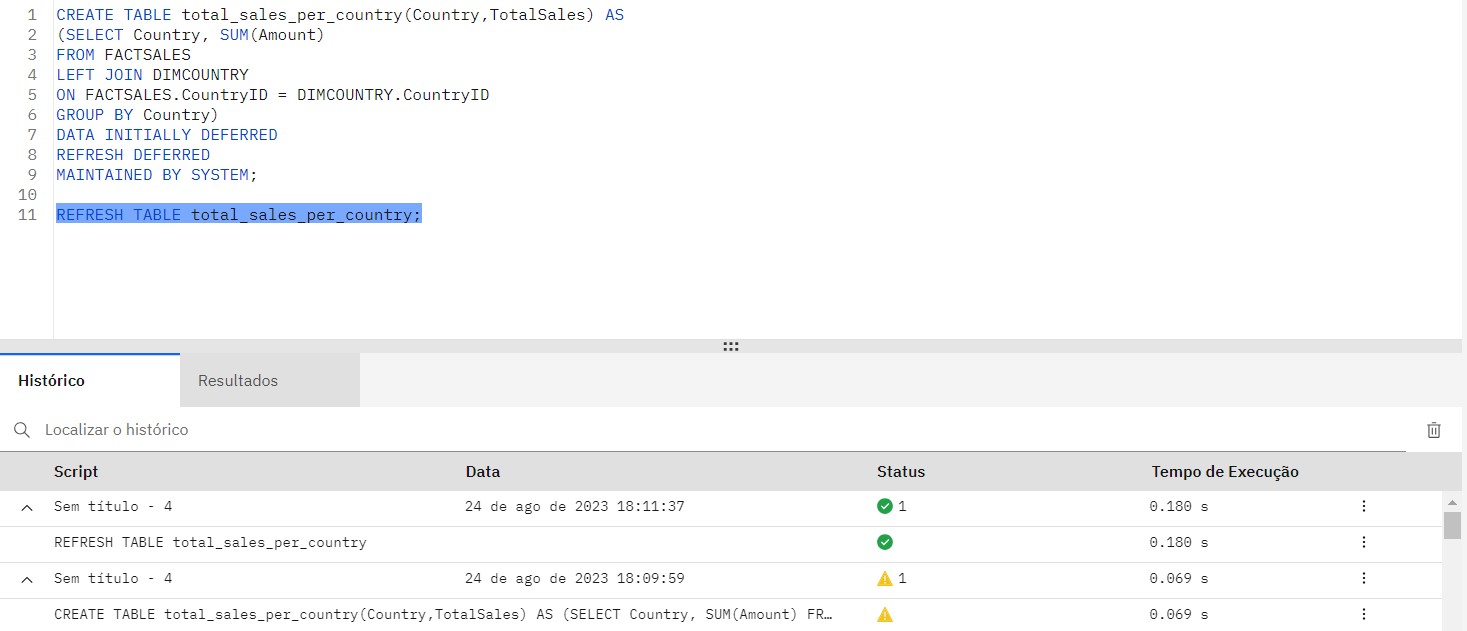
Create a rollup query using the columns year, country, and totalsales



Create a cube query using the columns year, country, and average sales



Create an MQT named total\_sales\_per\_country that has the columns country and total\_sales

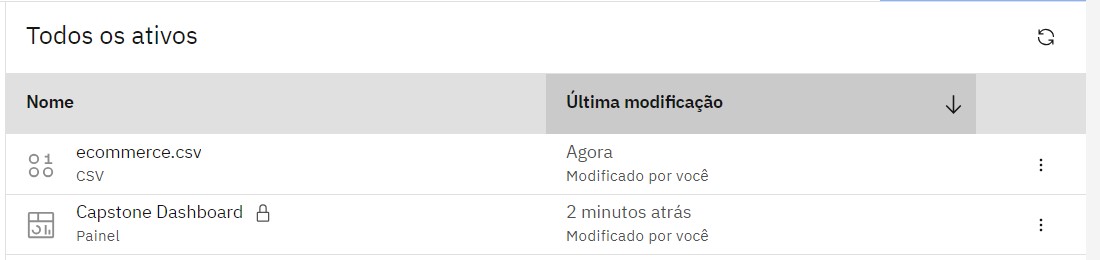


Your company has finished setting up a data warehouse. Now you are assigned the responsibility to design a reporting dashboard that reflects the key metrics of the business.

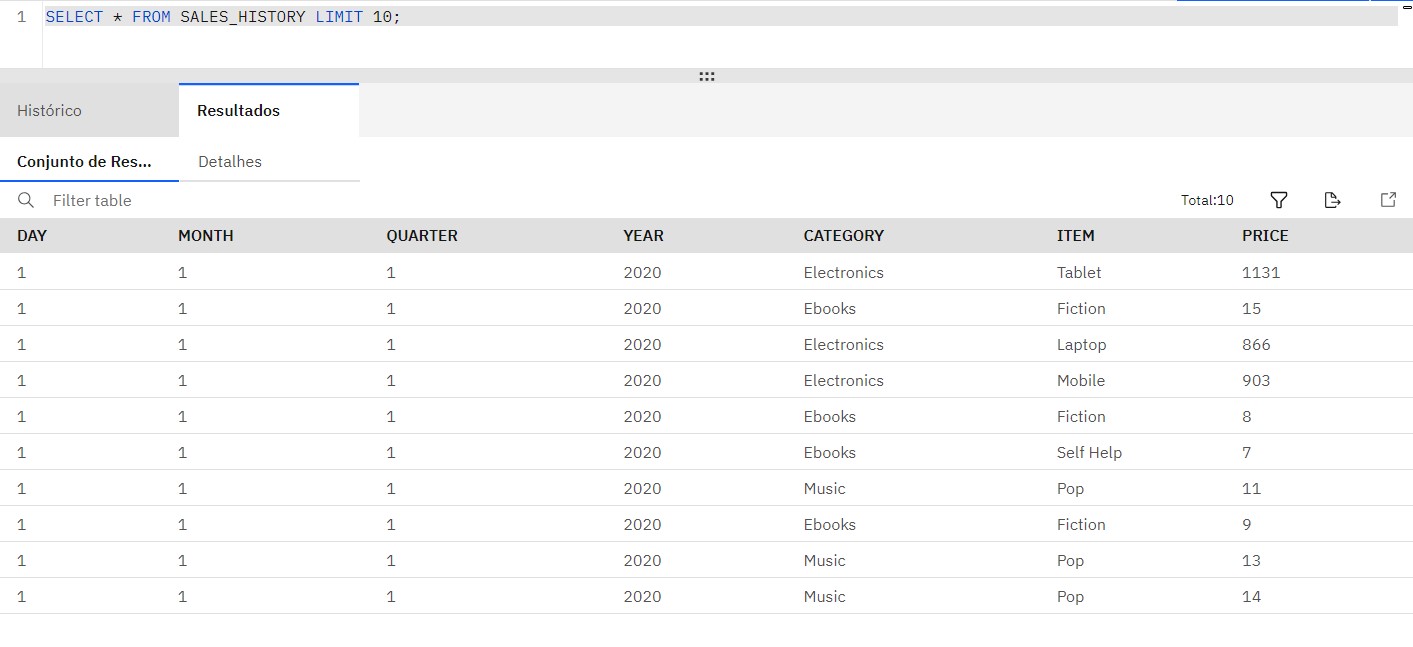
LOAD DATA INTO THE DATA WAREHOUSE

Import data in the downloaded file ecommerce.csv into a table named sales\_history

Upload the downloaded csv file ecommerce.csv into Cognos Analytics and take a screenshot of successful loading of the CSV file into Cognos.



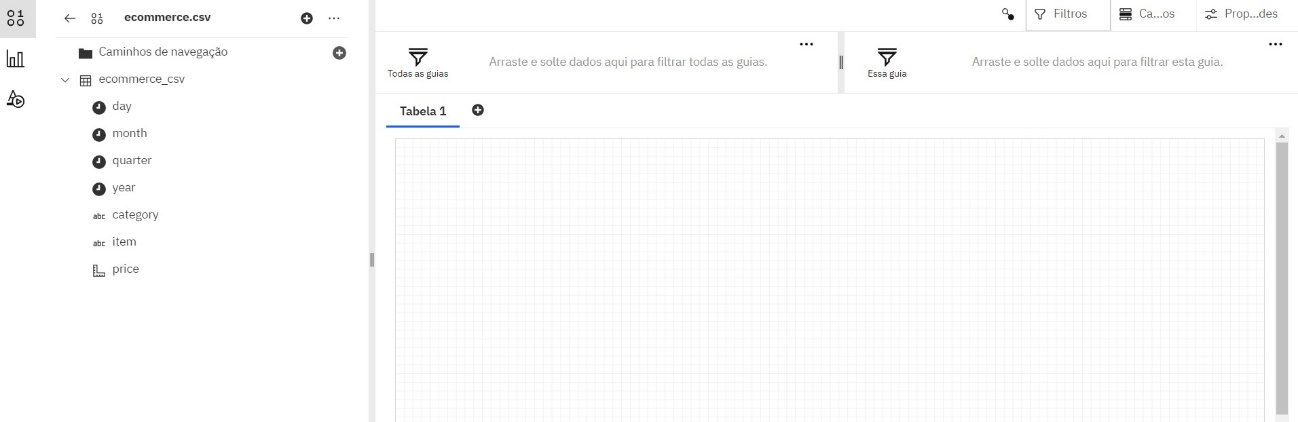
List the first 10 rows in the table sales\_history



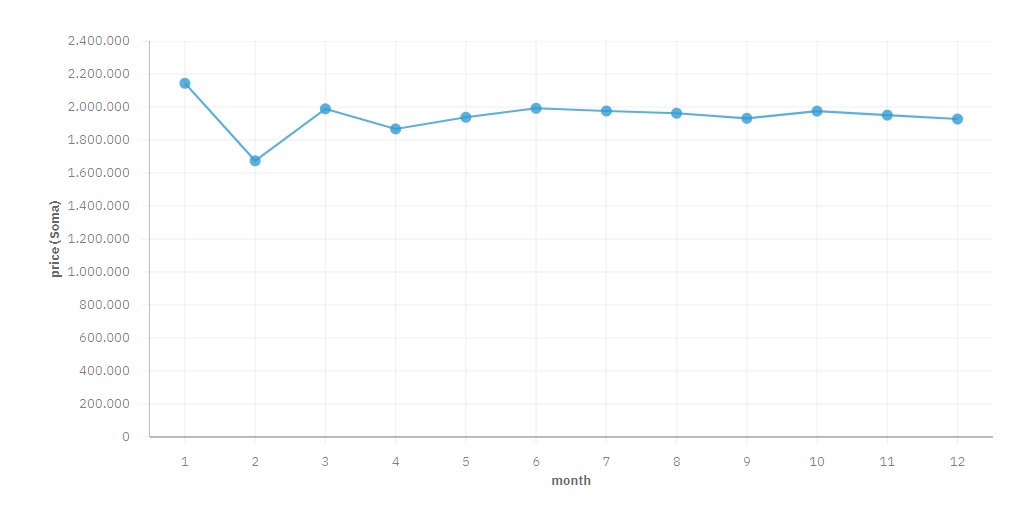
ACCESSING THE DATA SOURCE IN COGNOS

Create a data source in Cognos that points to the table sales\_history in your IBM DB2 database.

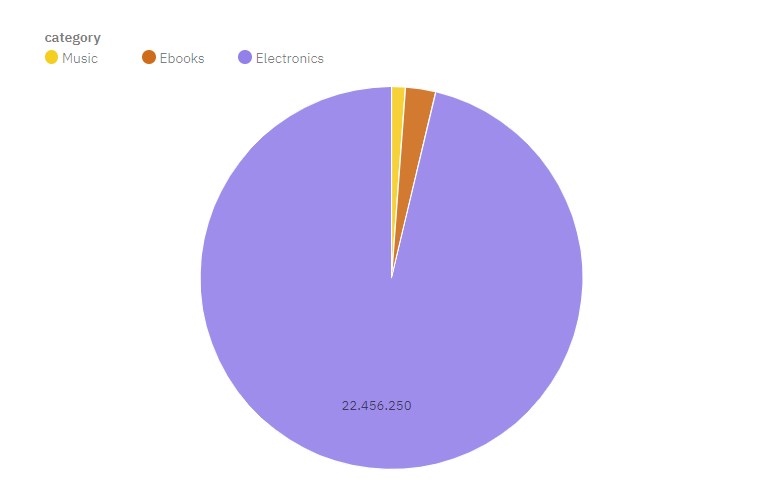
Take a screenshot of the datasource in the Cognos Analytics workspace.



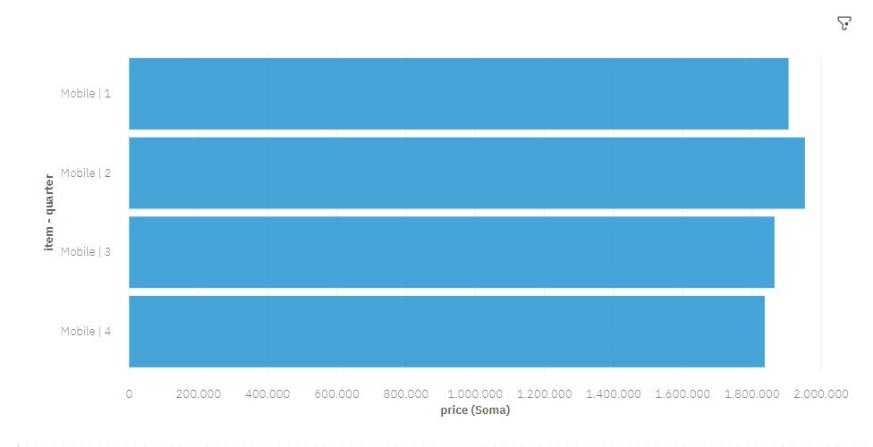
Create a line chart of month wise total sales for the year 2020.



Create a pie chart of category wise total sales.



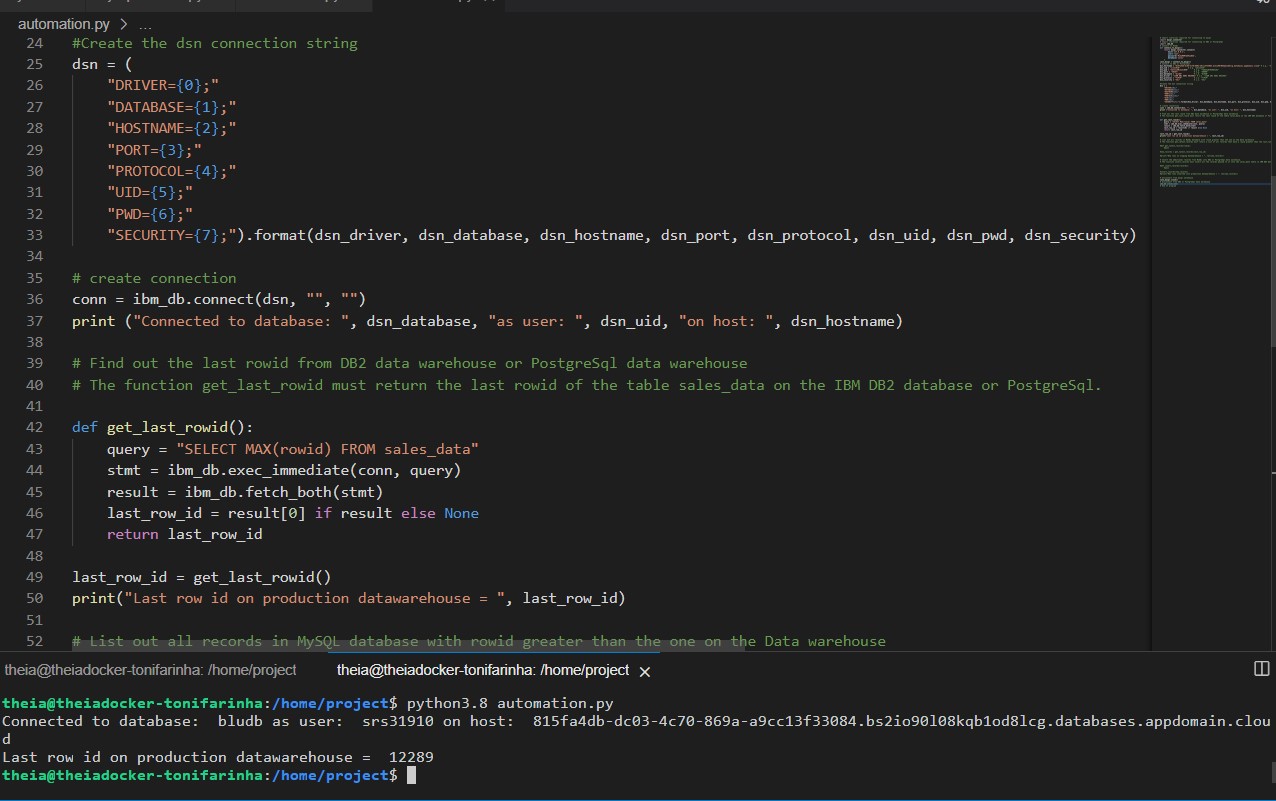
Create a bar chart of Quarterly sales of mobile phones



You need to keep data synchronized between different databases/data warehouses as a part of your daily routine. One task that is routinely performed is the sync up of staging data warehouse and production data warehouse. Automating this sync up will save you a lot of time and standardize your process. You will be given a set of python scripts to start with. You will use/modify them to perform the incremental data load from MySQL server which acts as a staging warehouse to the IBM DB2 or PostgreSQL which is a production data warehouse. This script will be scheduled by the data engineers to sync up the data between the staging and production data warehouse.

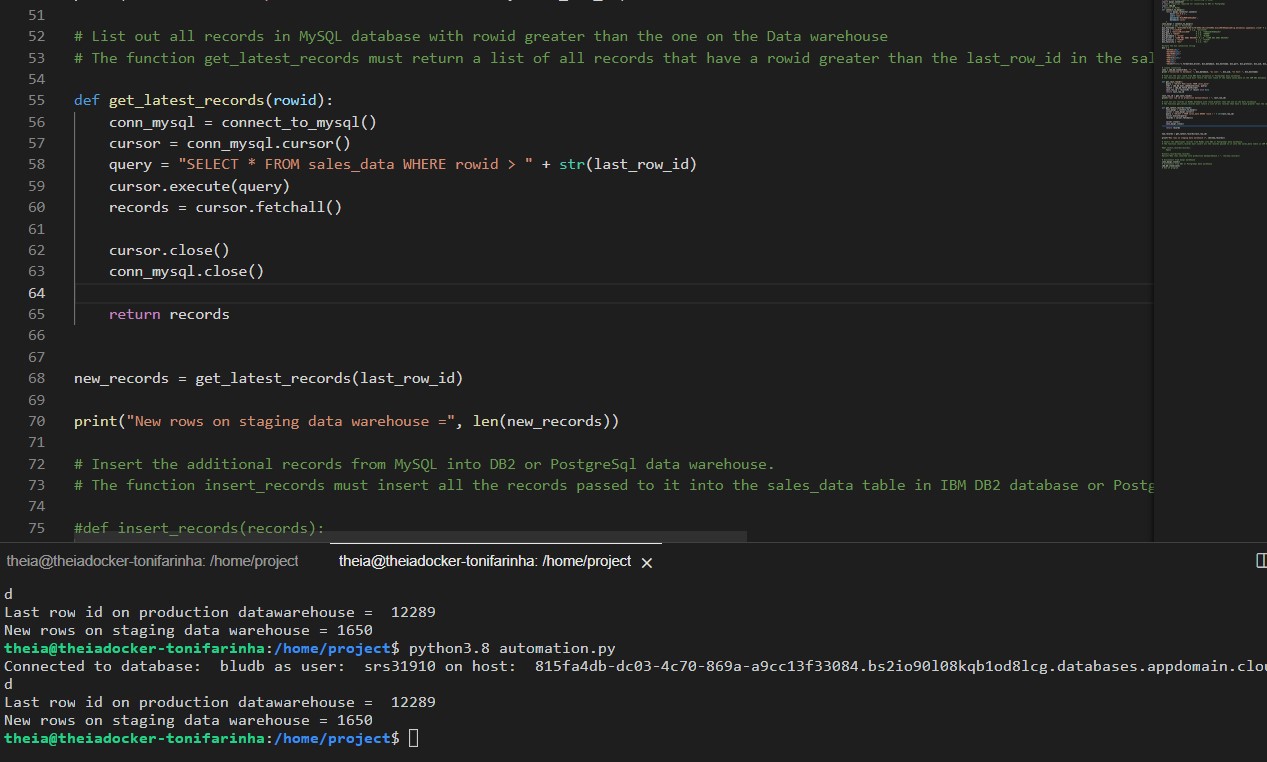
AUTOMATE LOADING OF INCREMENTAL DATA INTO THE DATA WAREHOUSE

In the program automation.py implement the function get\_last\_rowid()

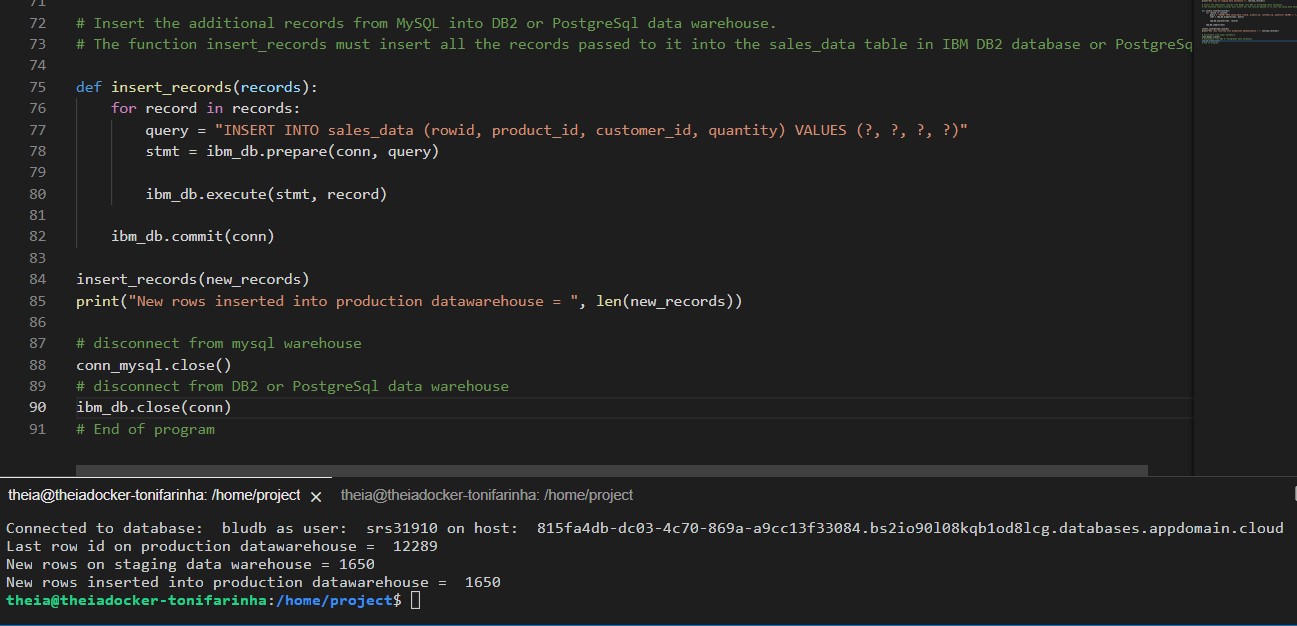


In the program automation.py implement the function get\_latest\_records()

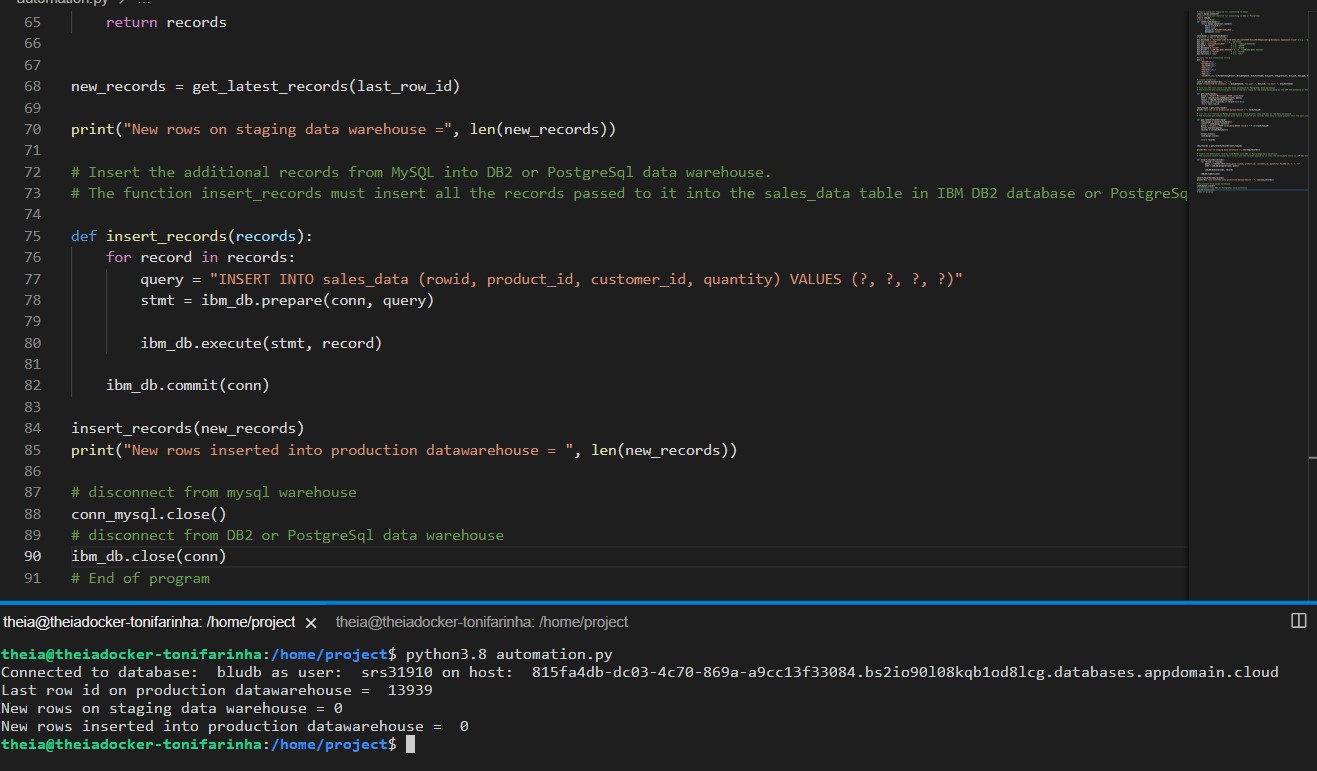
This function must connect to the MySQL database and return all records later than the given last\_rowid.



In the program automation.py implement the function insert\_records()



Run the program automation.py and test if the synchronization is happening as expected.

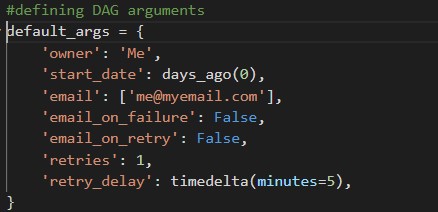


Write a pipeline that analyzes the web server log file, extracts the required lines(ending with html) and fields(time stamp, size ) and transforms (bytes to mb) and load (append to an existing file.)

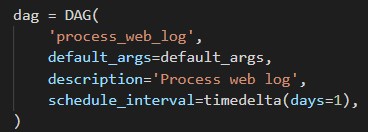
CREATE A DAG

Create a DAG with these arguments:

Owner, start\_date, email



Create a DAG named process\_web\_log that runs daily.



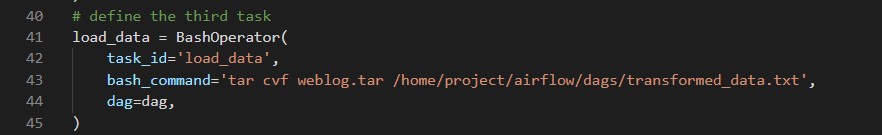
Create a task named extract\_data. This task should extract the ipaddress field from the web server log file and save it into a file named extracted\_data.txt



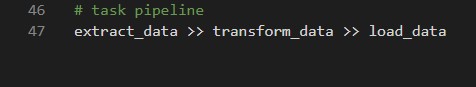
Create a task named transform\_data. This task should filter out all the occurrences of ipaddress “198.46.149.143” from extracted\_data.txt and save the output to a file named transformed\_data.txt.



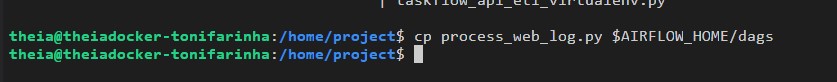
Create a task named load\_data. This task should archive the file transformed\_data.txt into a tar file named weblog.tar.



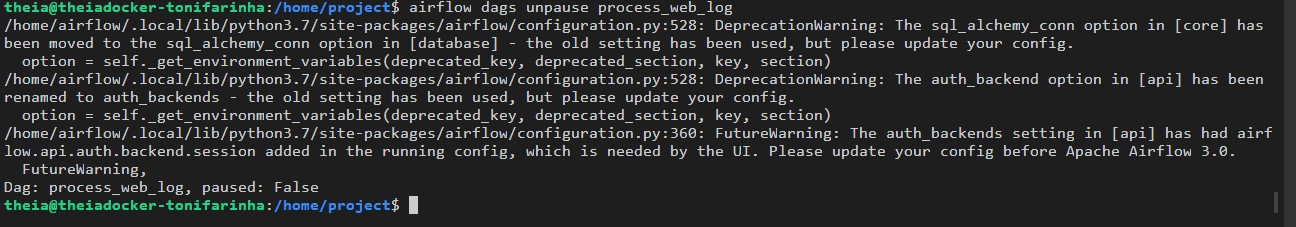
Define the task pipeline as per the details given below: first task extract data, second task transform data, third task load\_data



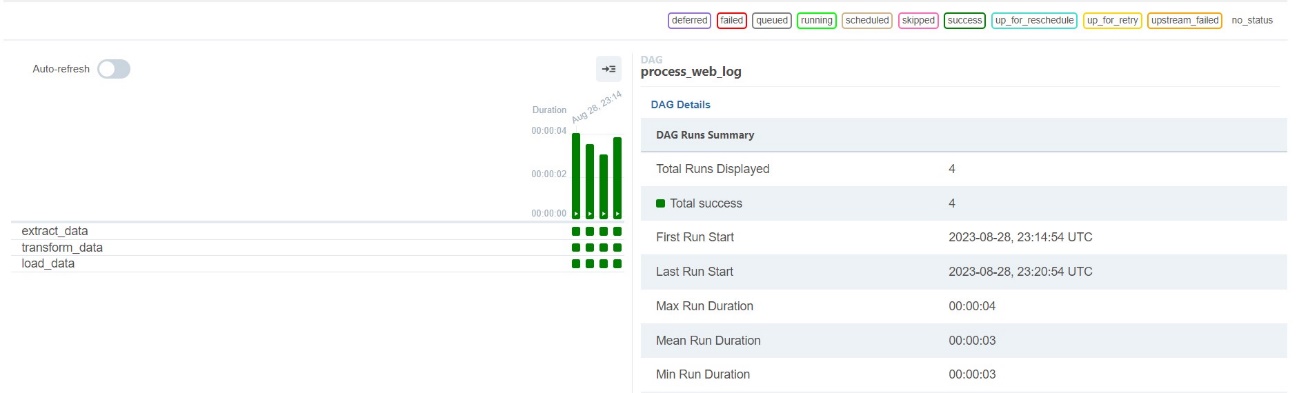
GETTING THE DAG OPERATIONAL

Submit the DAG

Unpause the DAG

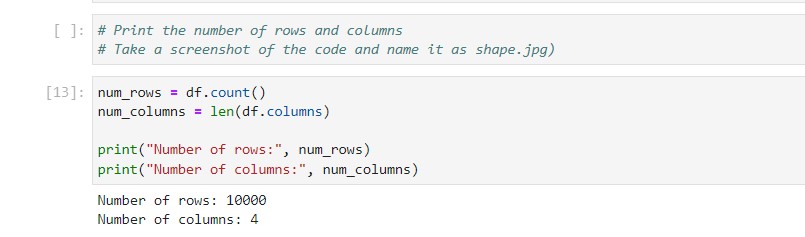


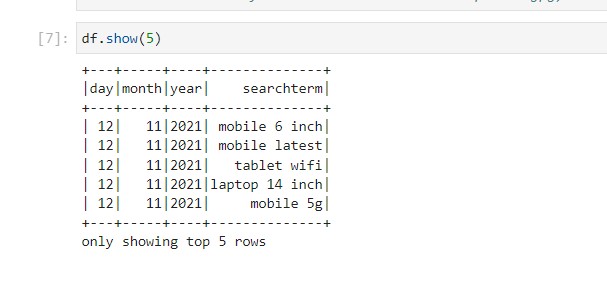
Monitor the DAG



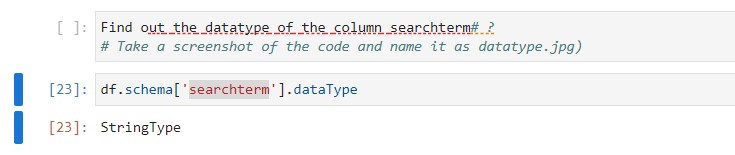
ANALYSE SEARCH TERMS ON THE E-COMMERCE WEB SERVER

Start a Spark session, download the search term dataset, load the csv into a spark dataframe, print the number of rows and columns

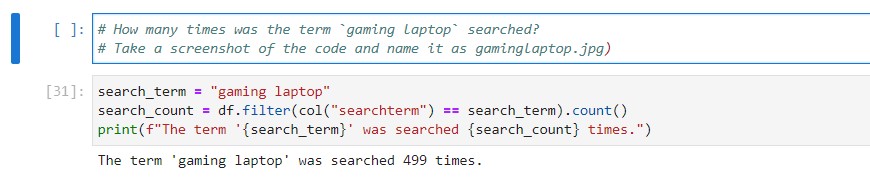


Print the top 5 rows

Find out the datatype of the column searchterm



How many times was the term `gaming laptop` searched?



Print the top 5 most frequently used search terms



Load the sales forecast model.



# Using the sales forecast model, predict the sales for the year of 2023

