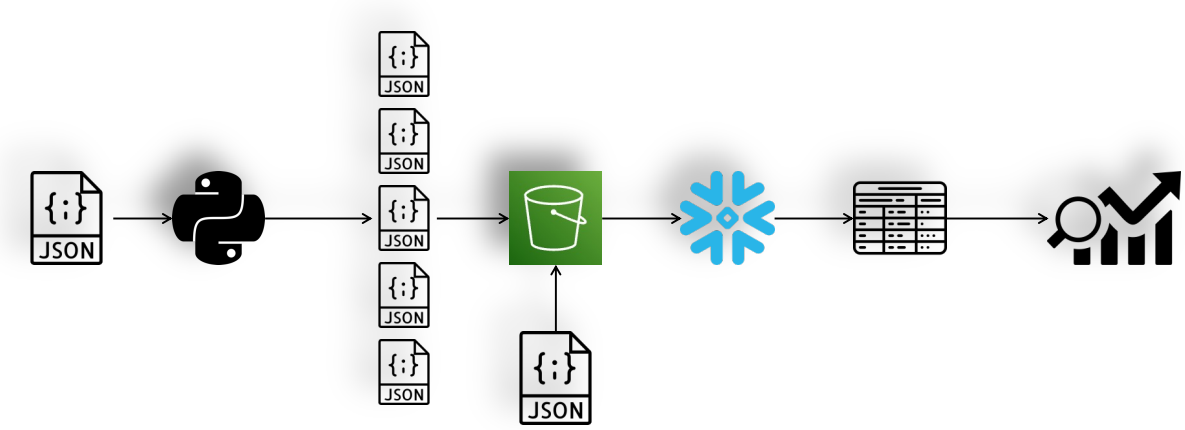
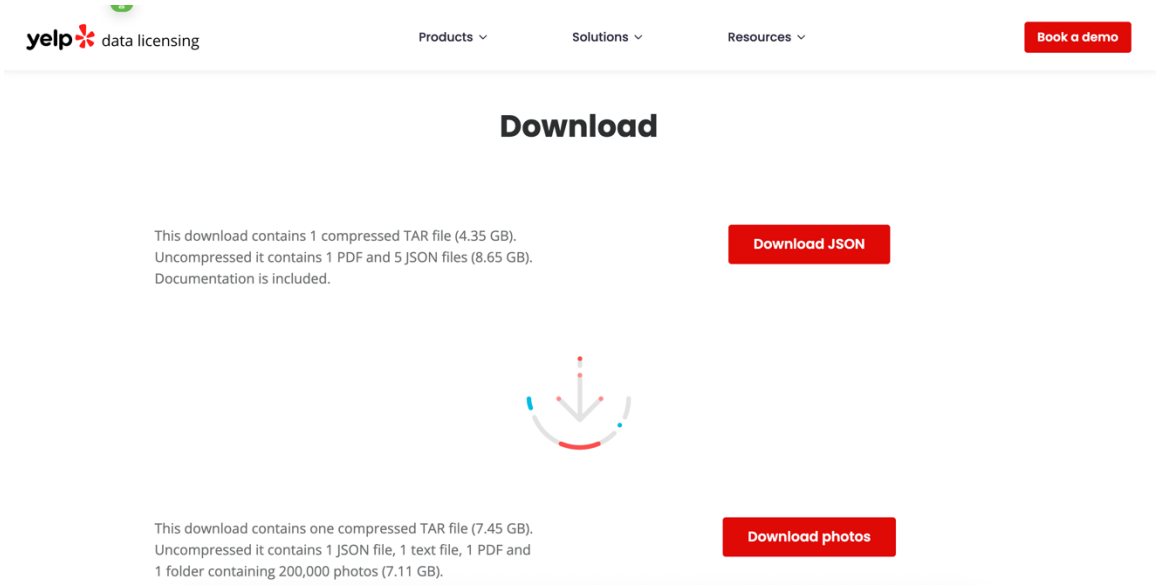
# Data Analytics Pipeline Using Python, AWS and Snowflake



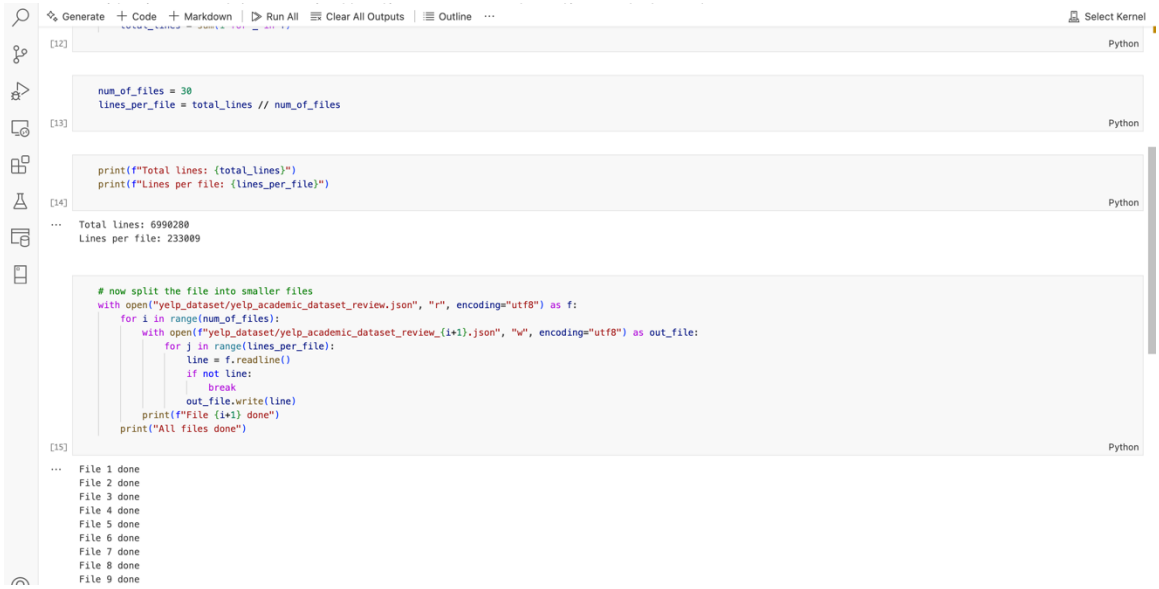

**Steps Involved:**

- **Data Download**: The data was obtained from the Yelp Business Open Dataset.
- **Data Partitioning & AWS S3 Upload**: The dataset was split into 30 parts using Python and uploaded to an AWS S3 bucket.
- **AWS S3 Setup**: The business data was added to the S3 folder, and a new AWS user with read access was created for Snowflake integration.
- **Data Fetch & Transformation in Snowflake**: Data in JSON format was loaded into Snowflake using the VARIANT datatype, and two tables (reviews and businesses) were created from JSON keys.
- **Sentiment Analysis Using UDF**: A User Defined Function (UDF) in Snowflake was created to analyze sentiment of review text using Python's TextBlob, categorizing sentiments as 'Positive', 'Negative', or 'Neutral'.
- **Business Question Answers**: After table creation, 10 business-related questions were answered using SQL queries in Snowflake.

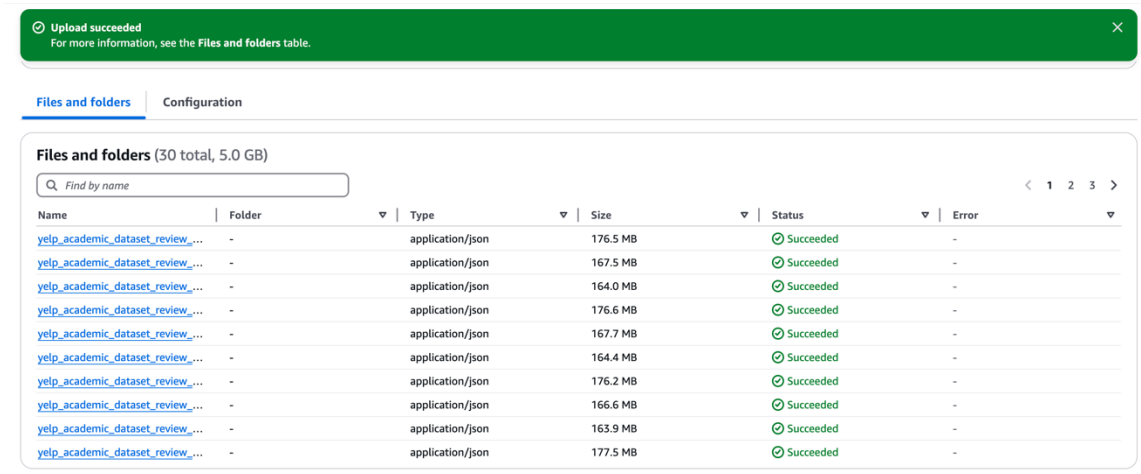
- **YELP Dataset**



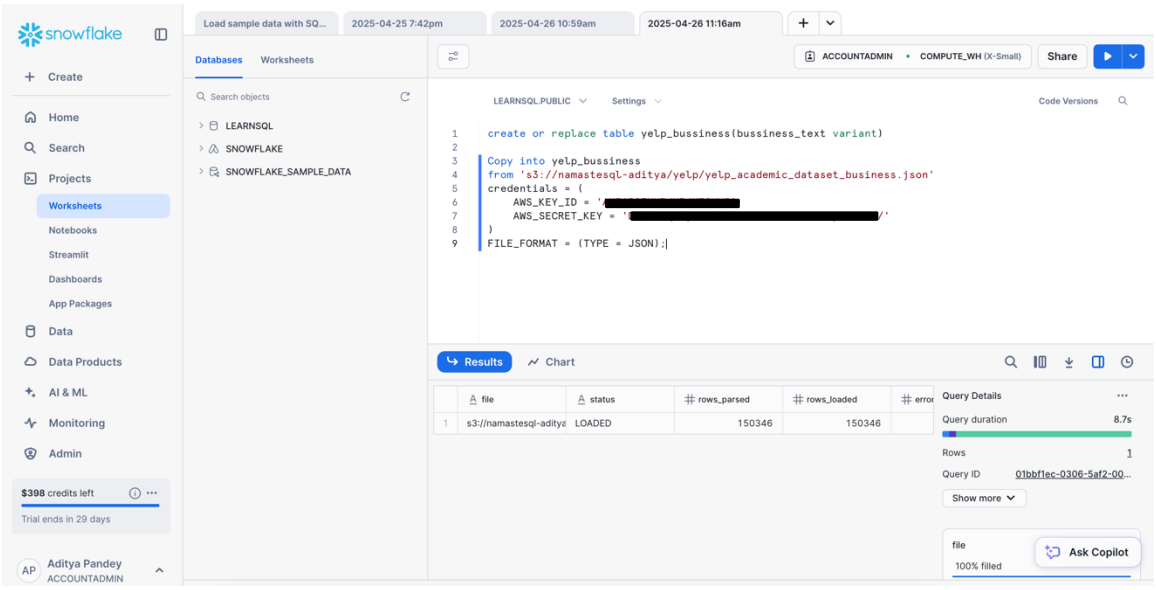

- **Python Data Division**



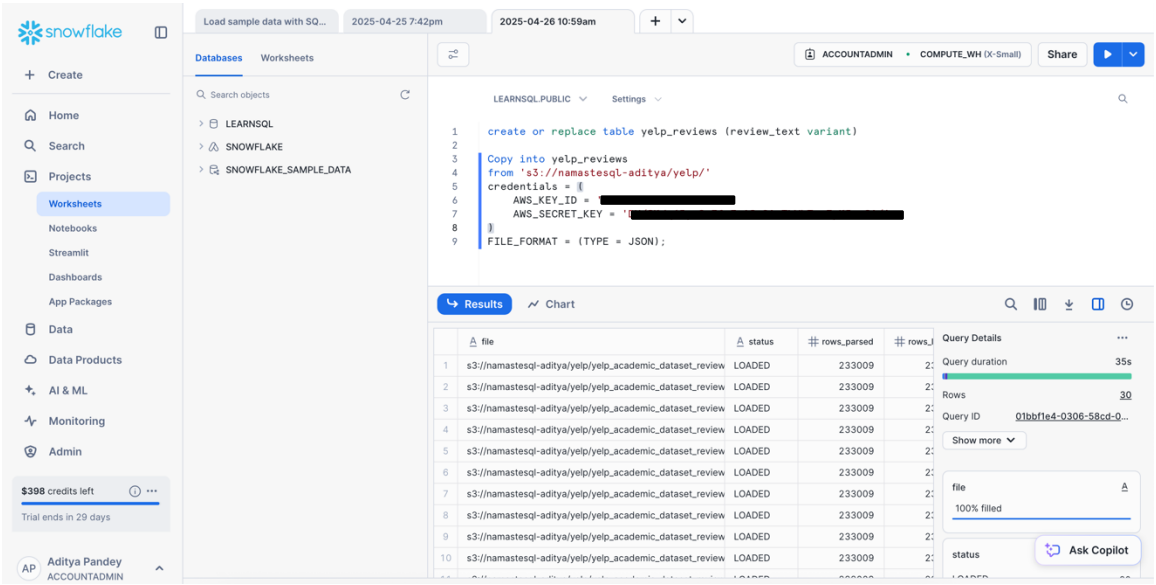
- **AWS Data Upload**



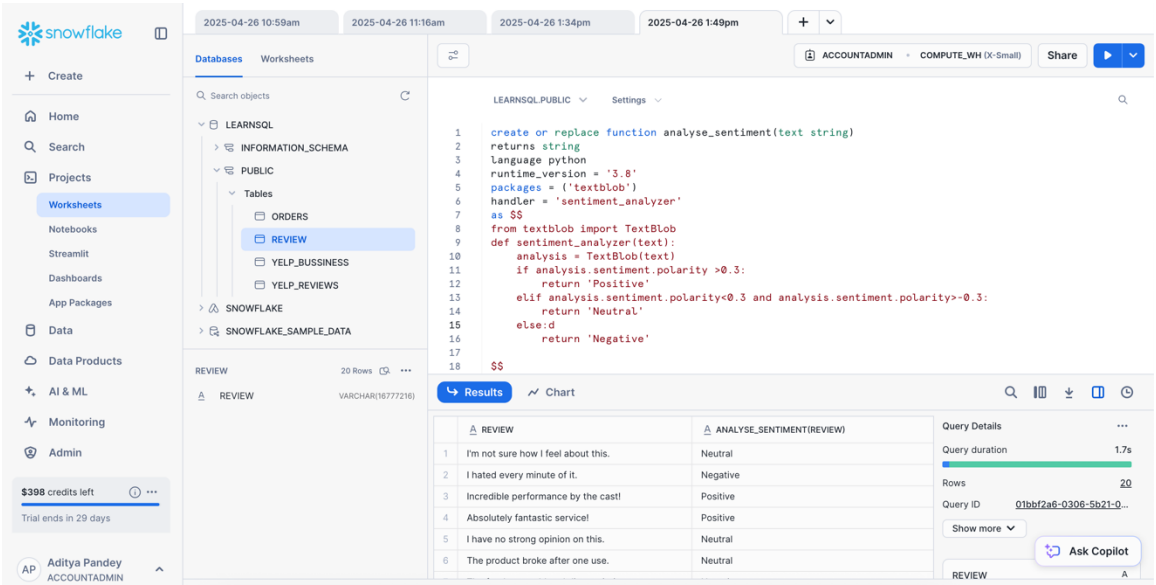

- **Business data fetched from AWS**



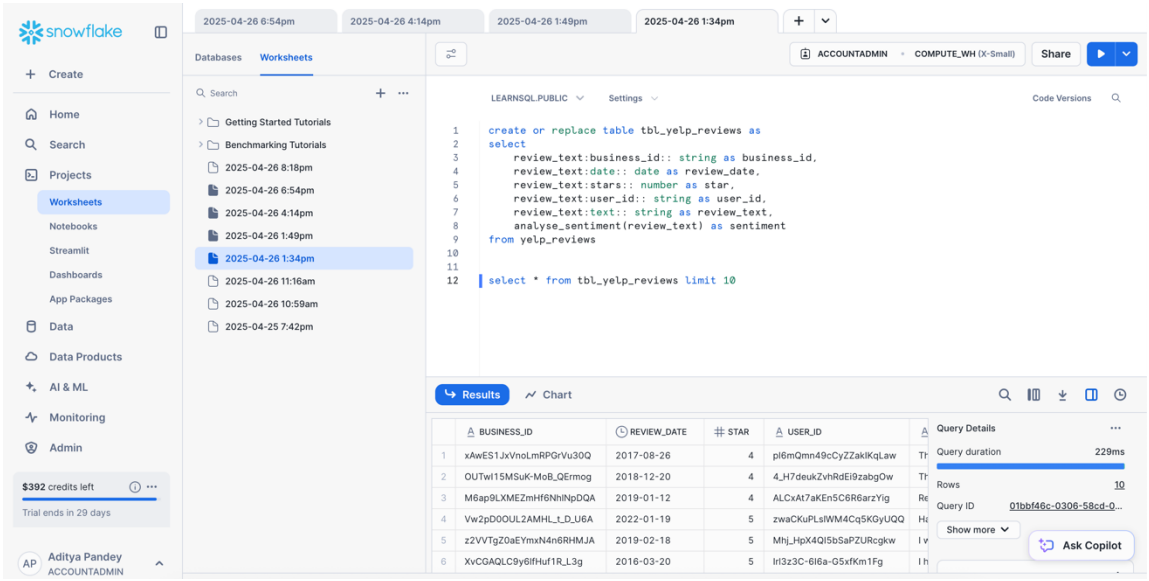
- **Review data fetched from AWS**



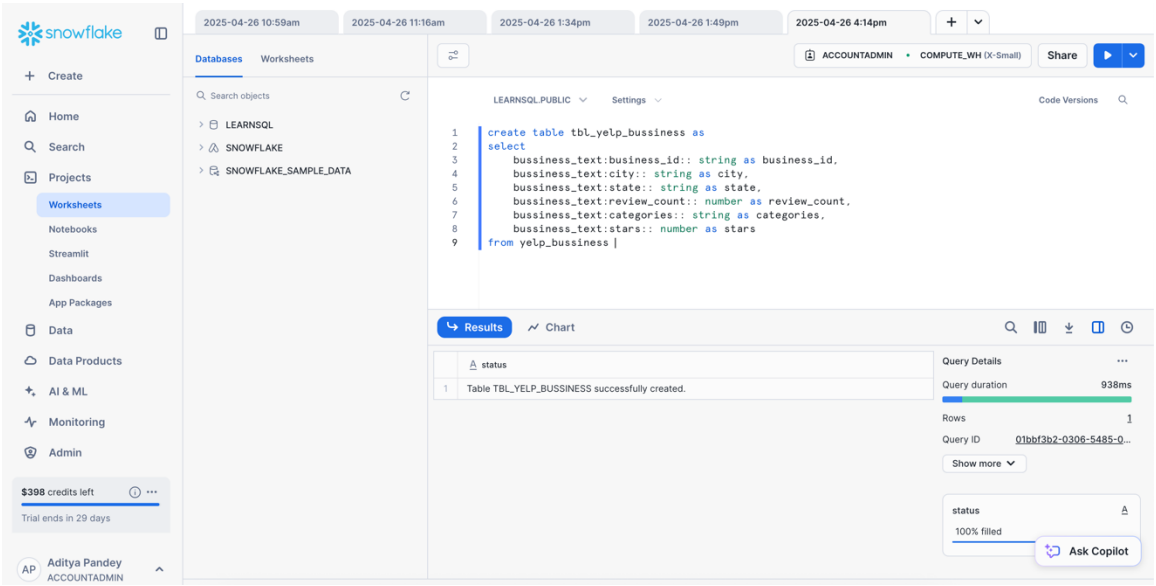

- **UDF**



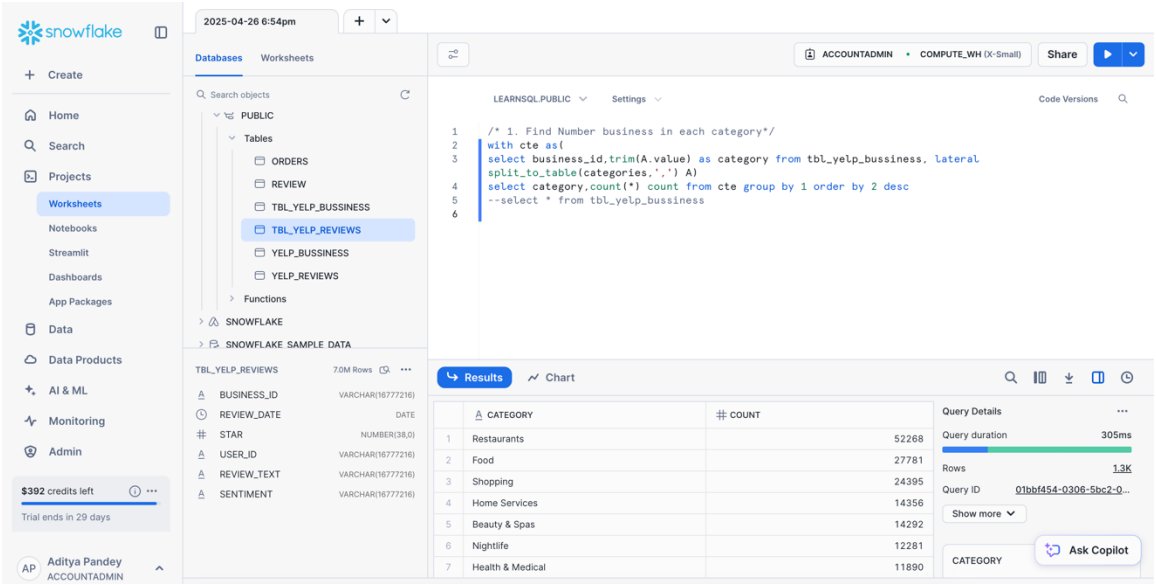
- **JSON to structured table (Review Table)**



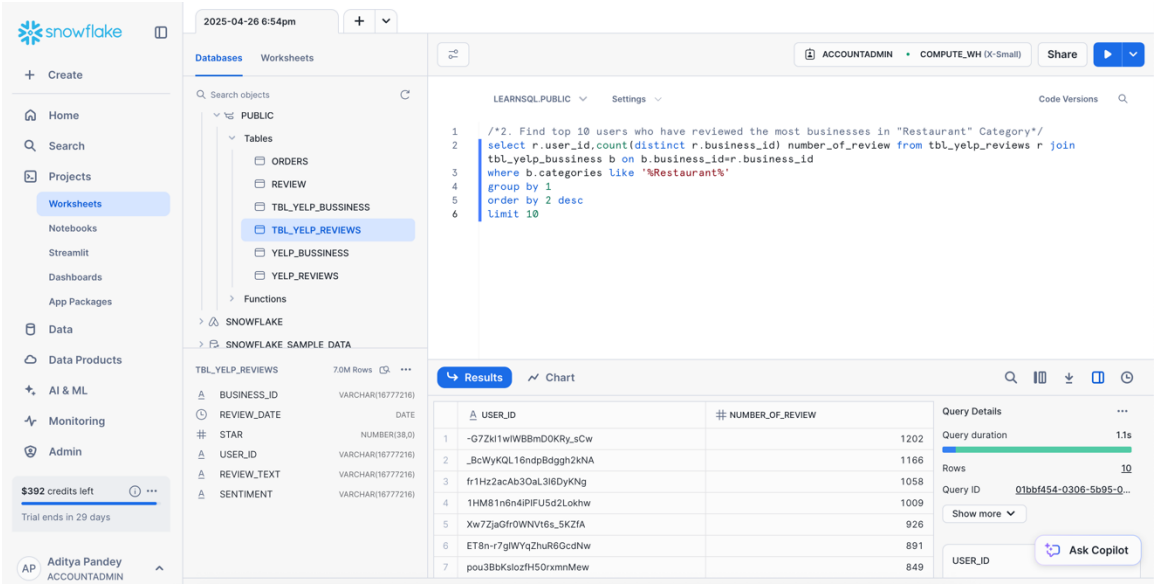

- **JSON to structured table (Business Table)**



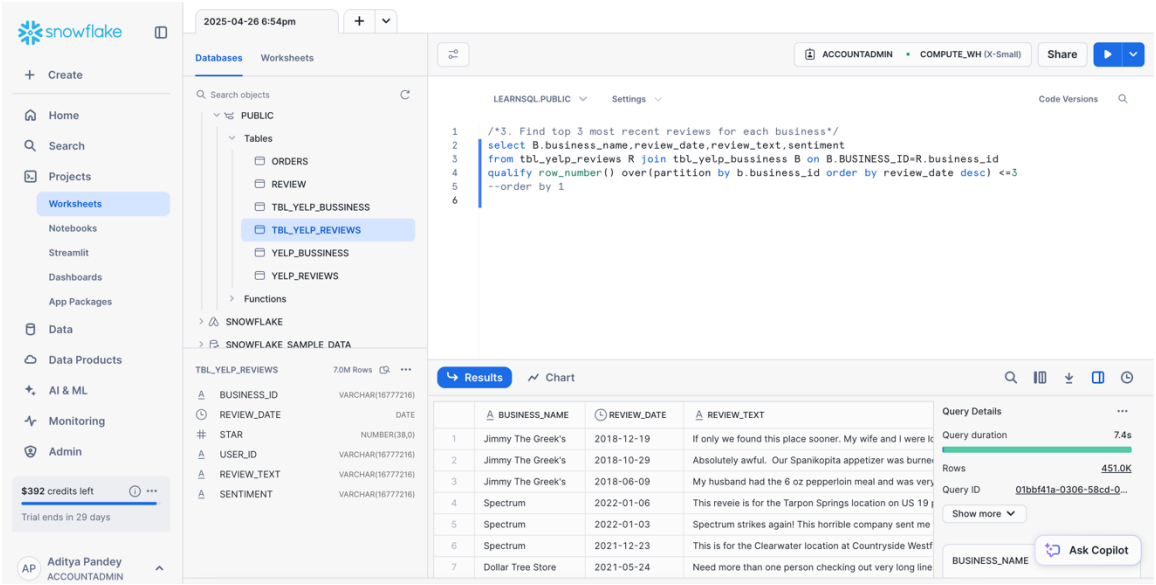
1. Find the Number of Businesses in Each Category



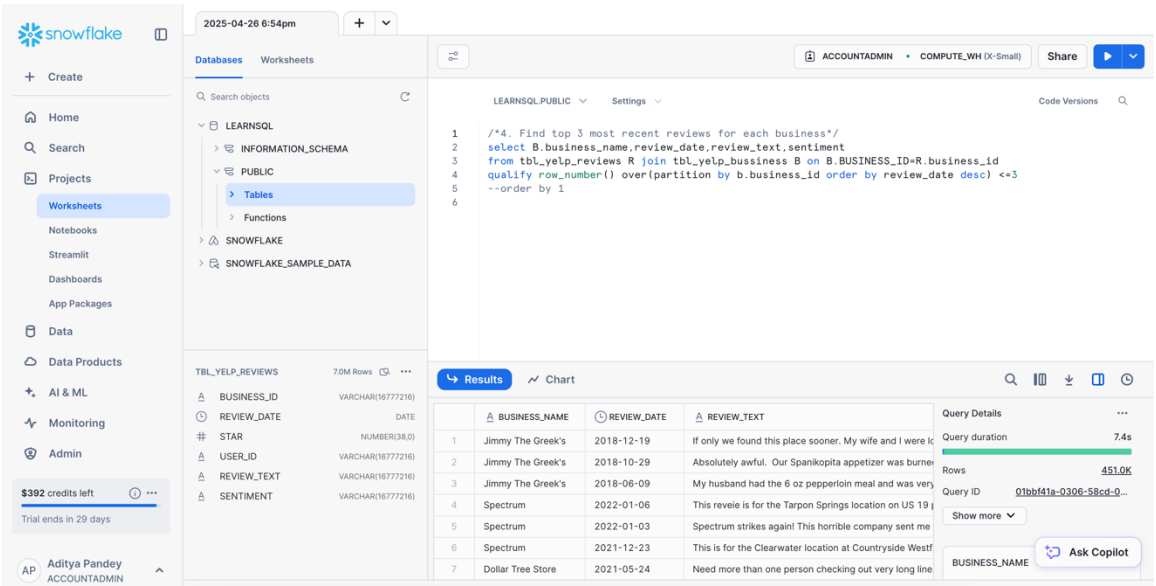

2. Find Top 10 Users Who Have Reviewed the Most Businesses in "Restaurant" Category



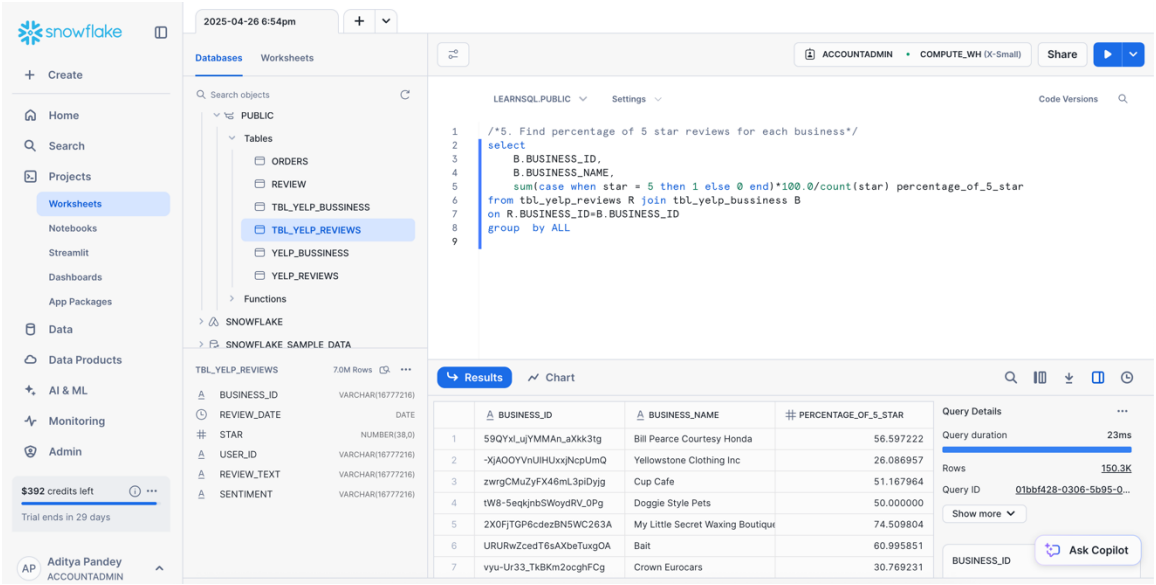
3. Find Most Popular Categories of Business (Based on Number of Reviews)



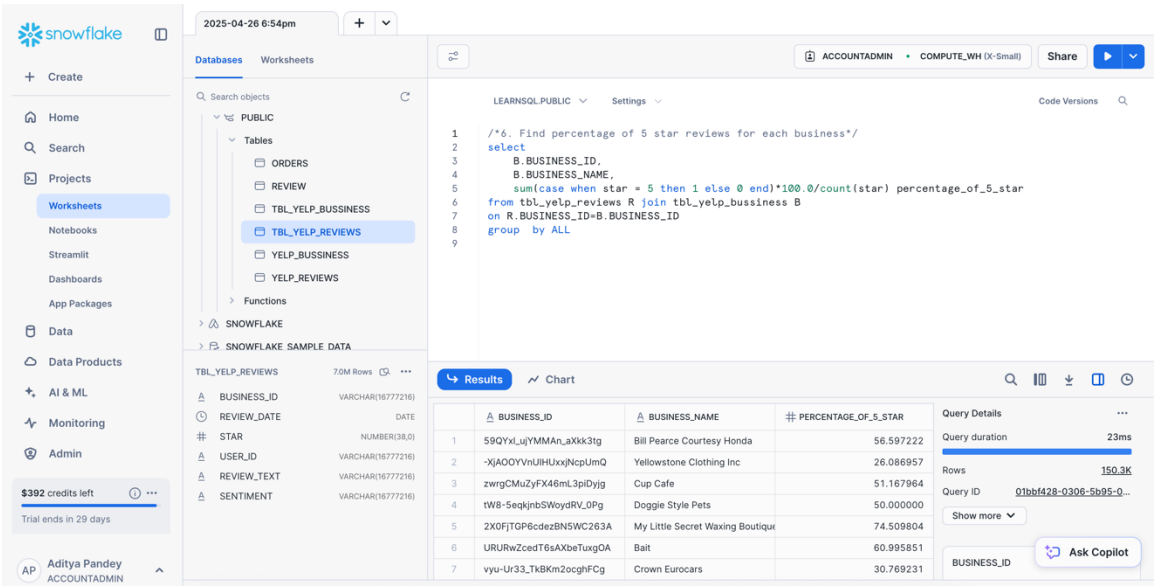

4. Find Top 3 Most Recent Reviews for Each Business



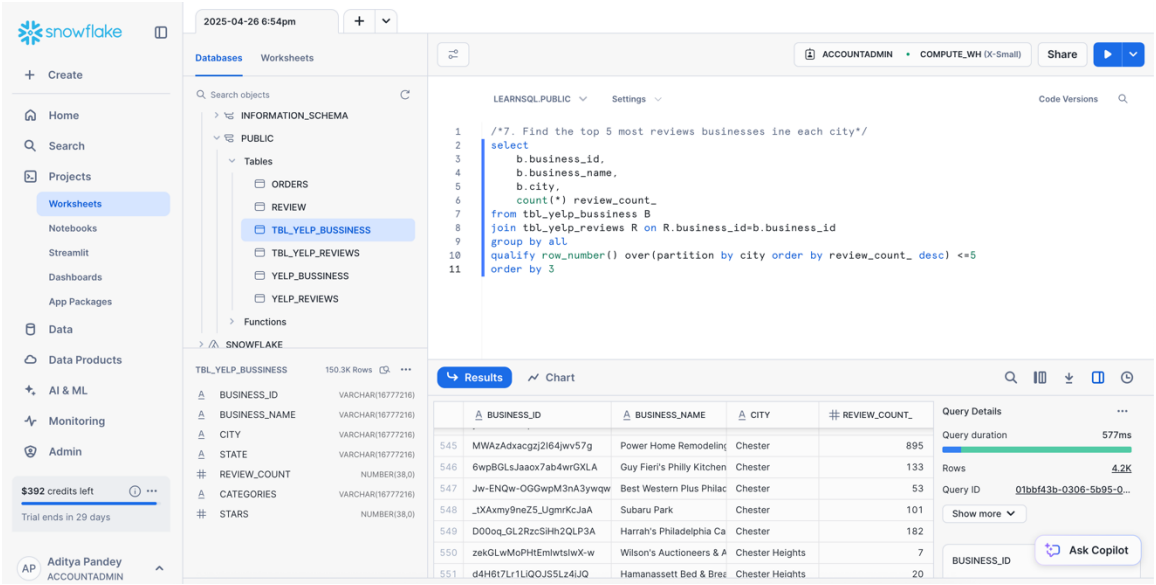
5. Find the Month with the Highest Number of Reviews



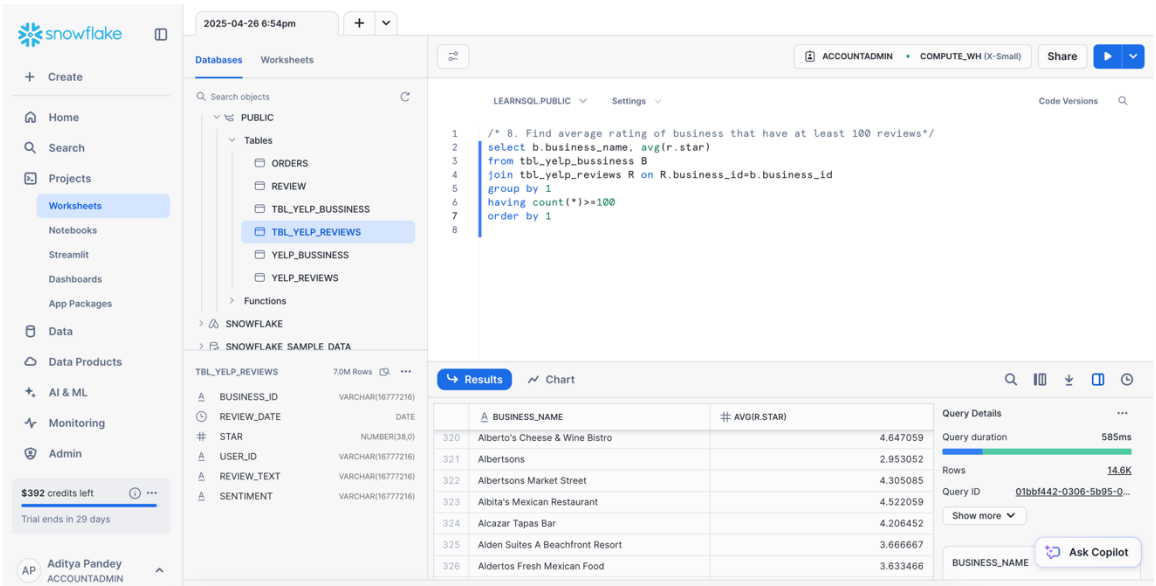

6. Find Percentage of 5-Star Reviews for Each Business



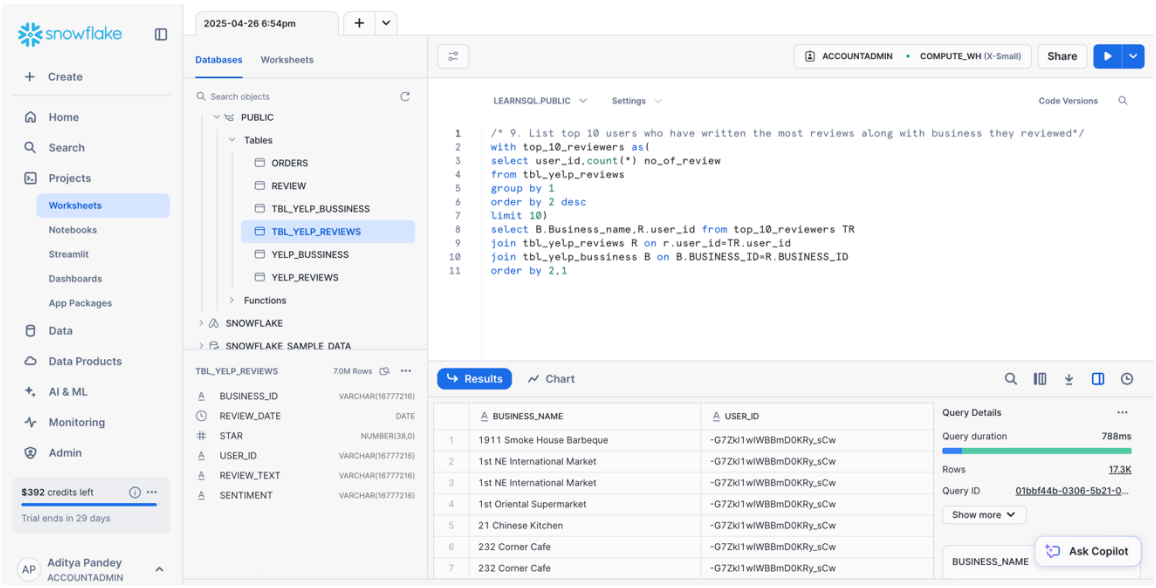
## 7. Find the Top 5 Most Reviewed Businesses in Each City




## 8. Find Average Rating of Businesses that Have At Least 100 Reviews



9. List Top 10 Users Who Have Written the Most Reviews Along with the Businesses They Reviewed




10. Find Top 10 Businesses with the Highest Positive Sentiment Reviews