**Customer Clustering for Enhanced Marketing Strategies in Retail Chain**

1. **Problem Statement**

**Context :**

A successful chain of retail stores operates across various locations, each identified by a unique StoreID. The stores sell products through multiple sales channels, including in-store purchases, online sales, distributors, and wholesale transactions. To optimize marketing strategies, the retail chain collects detailed sales data from each store. This data includes information on customer purchases, sales channels, order details, and financial transactions.

**Business Challenge :**

In the competitive retail industry, understanding customer behavior is crucial for tailoring marketing efforts and maximizing sales. The retail chain faces the challenge of effectively targeting its marketing campaigns due to the diverse nature of its customer base. Without clear customer segments, marketing efforts can be inefficient, resulting in suboptimal business performance. The primary challenge is to segment customers based on their purchasing behavior to provide insights that will help tailor marketing strategies, improve customer engagement, and increase sales.

**Project Goal :**

The goal of this project is to perform a comprehensive RFM (Recency, Frequency, Monetary) analysis to segment customers based on their purchasing behavior. By identifying distinct customer segments, the retail chain aims to develop targeted marketing strategies that cater to the specific needs and preferences of each segment, ultimately enhancing customer satisfaction and driving business growth. This will involve the use of machine learning clustering algorithms to automatically group customers into meaningful segments.

1. **Project Scope :**

**Data Source :**

The dataset used in this project is derived from the retail chain's sales data, which includes detailed information about customer transactions such as order dates, quantities, and monetary values. This data is crucial for understanding customer behavior and preferences.

**Key Metrics :**

**Recency:** Days since the customer's last purchase, calculated using the OrderDate column.

**Frequency:** Total number of purchases made by the customer, derived from the OrderNumber column.

**Monetary:** Total amount of money spent by the customer, calculated using the Order Quantity and Unit Price columns.

**Deliverables :**

**RFM Analysis Results:** Calculation of Recency, Frequency, and Monetary values for each customer.

**Clustering Model and Results:** Segmentation of customers into distinct clusters based on RFM values using machine learning algorithms.

**Tableau Visualizations:** Interactive dashboards and visualizations to present customer segments and insights.

**Comprehensive Documentation:** Detailed documentation of the project, including methodology, analysis, and findings.

**Tools and Technologies :**

**Database:** AWS RDS MySQL and AWS S3 for data storage and processing.

**Programming Language:** Python for data preprocessing and clustering.

**Machine Learning:** Scikit-learn for implementing clustering algorithms.

**Visualization Tool:** Tableau for data visualization.

**Additional Tools:** Pandas, SQLAlchemy.

**Data Understanding :**

Number of Rows: 7,991

Number of Columns: 16

**Columns:**

OrderNumber: Unique order identifier.

Sales Channel: Sales channel (In-Store, Online, etc.).

WarehouseCode: Warehouse identifier.

ProcuredDate: Date of procurement.

OrderDate: Date of order placement.

ShipDate: Date of shipping.

DeliveryDate: Date of delivery.

CurrencyCode: Currency used.

SalesTeamID: Sales team identifier.

CustomerID: Customer identifier.

StoreID: Store identifier.

ProductID: Product identifier.

Order Quantity: Quantity of items ordered.

Discount Applied: Discount applied to the order.

Unit Price: Price per unit.

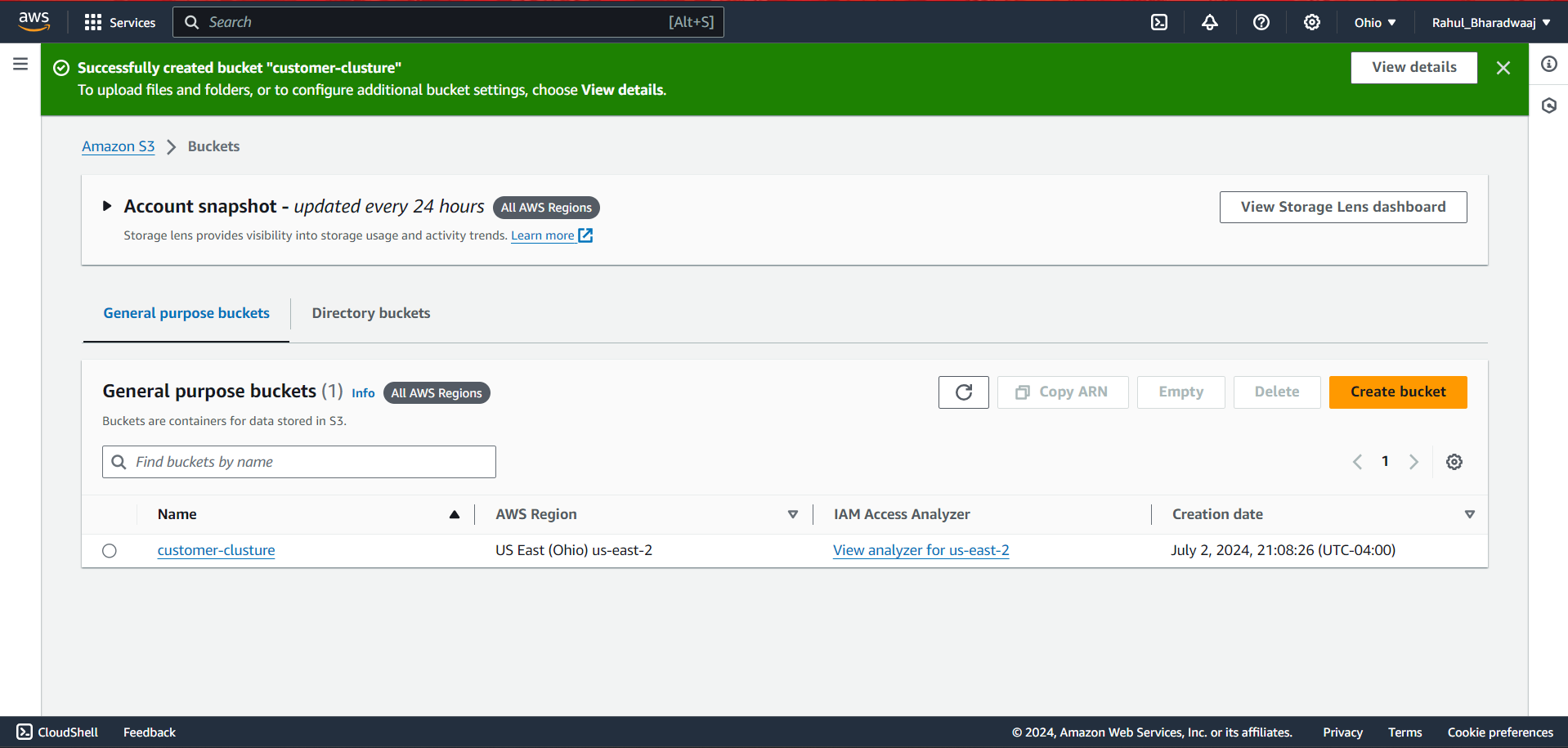
Unit Cost: Cost per unit.

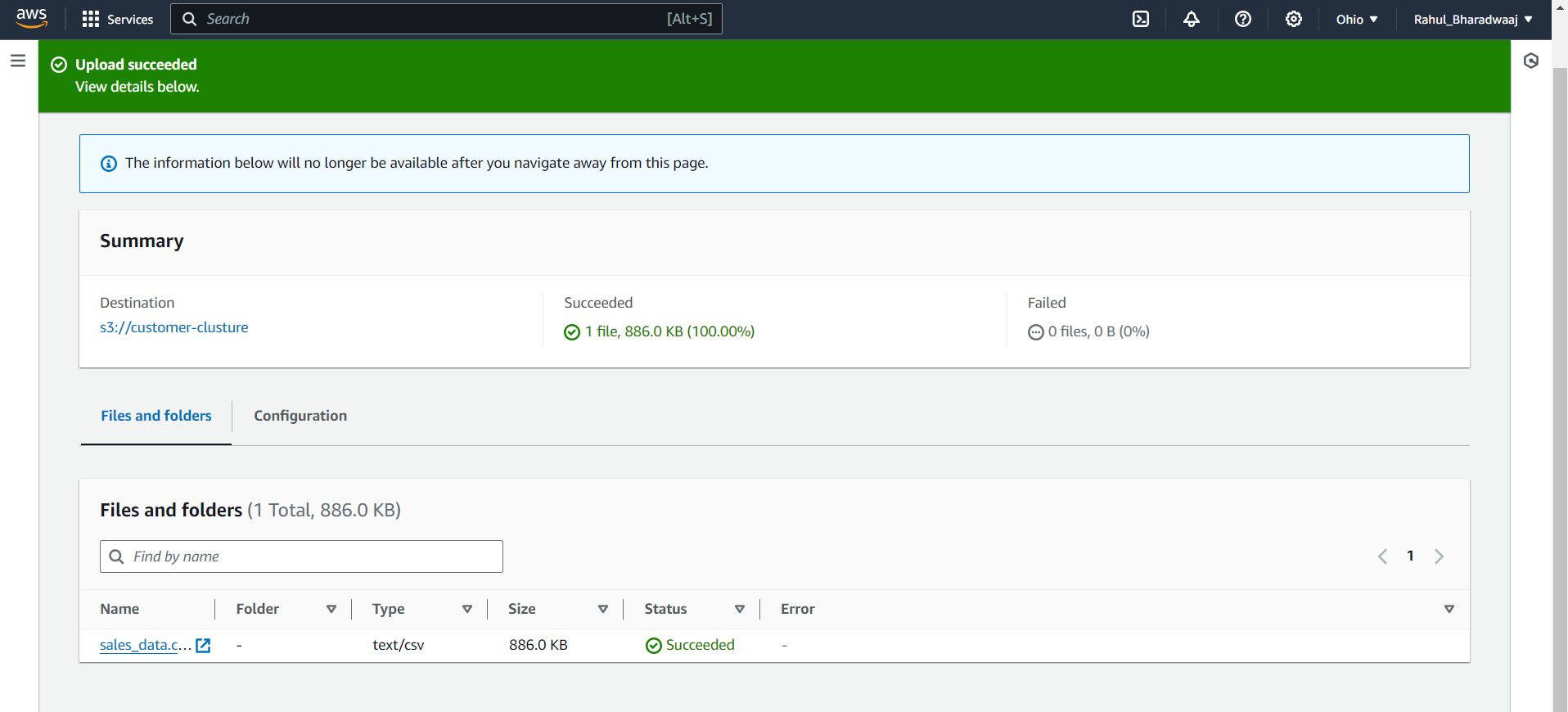
1. **Data Ingestion**

**Step 1: Upload Data to S3**

Created an S3 bucket named customer-cluster.

Uploaded the CSV file sales\_data.csv to the S3 bucket.





**Step 2: AWS RDS MySQL Setup**

Created an RDS Instance

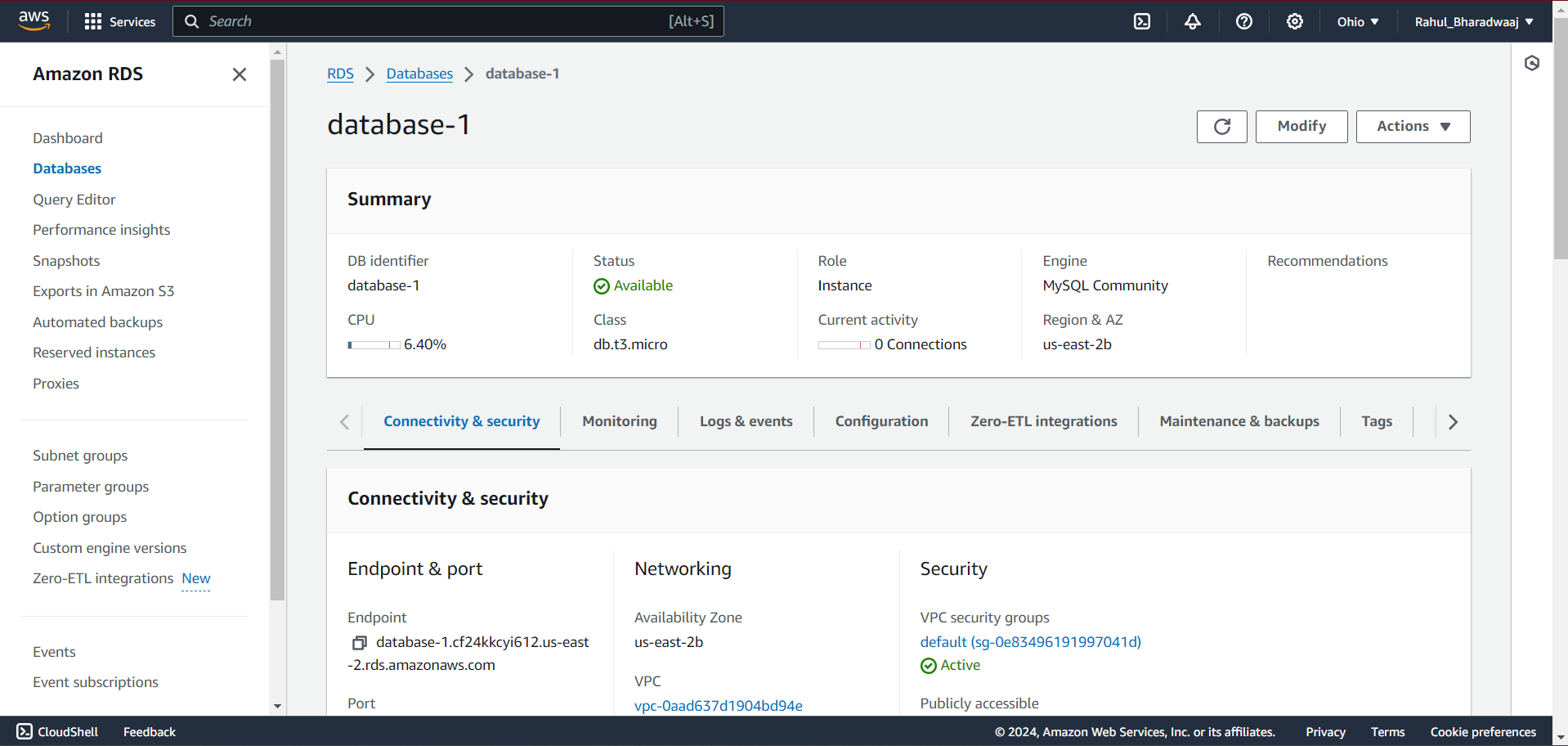
**Engine options:** Select MySQL.

**Templates:** Select "Free tier".

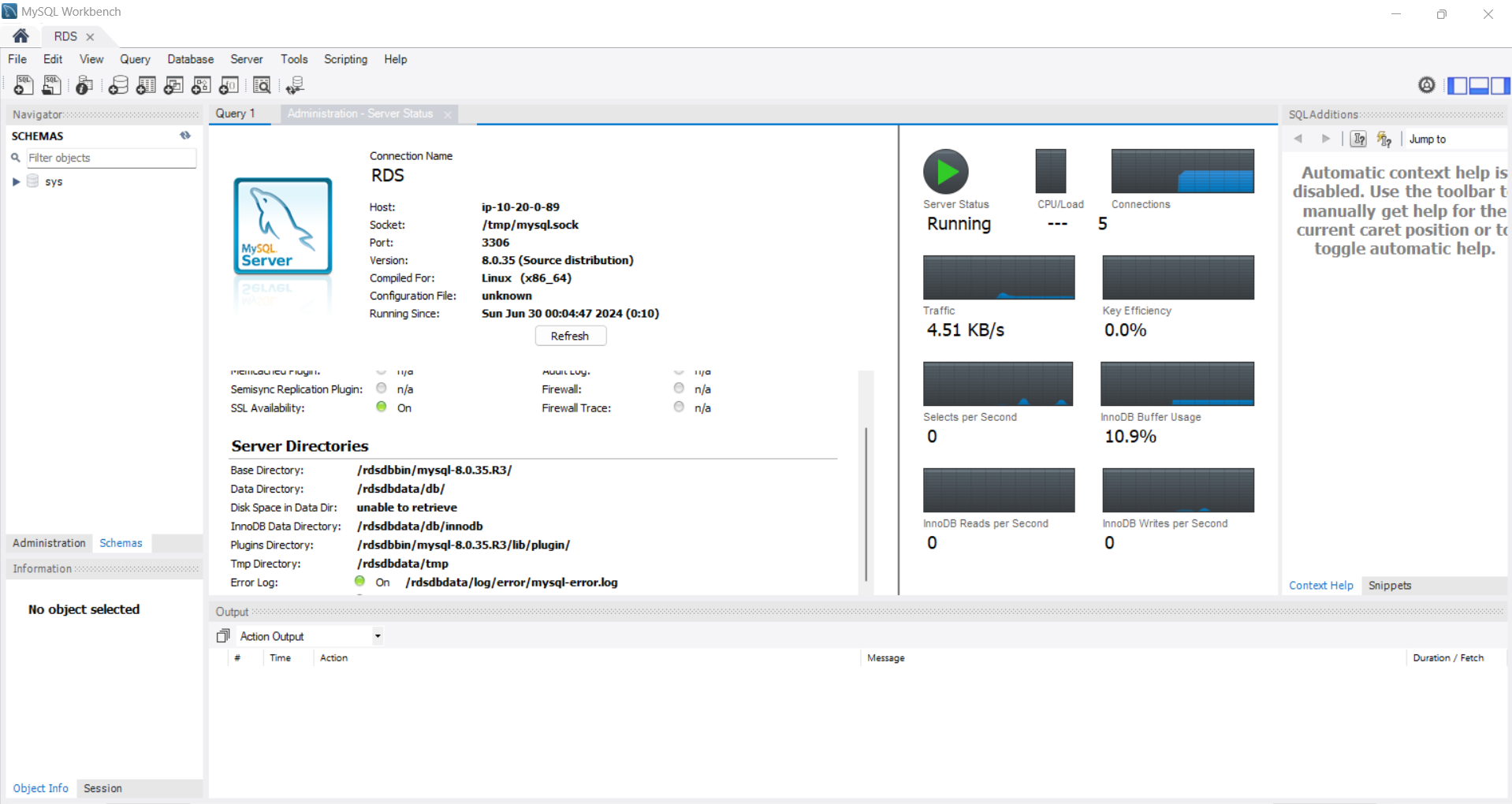
**Instance configuration:** Use default settings for free tier eligibility.

**Storage:** Set storage to 20 GB (free tier limit).

**Connectivity:** Configure VPC, subnet group, and public access as needed.

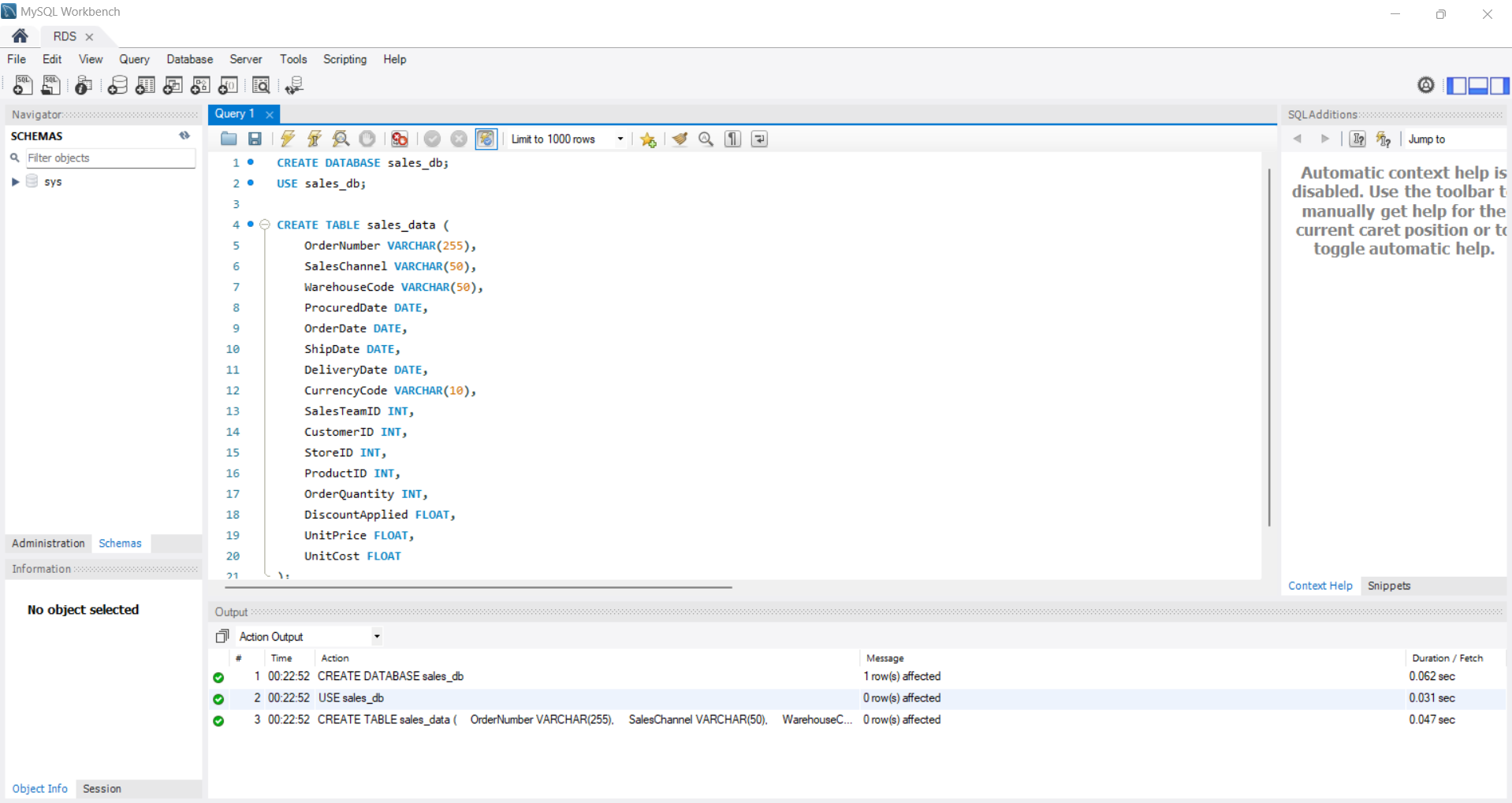


Connected MySQL client (MySQL Workbench ) to AWS RDS MySQL Instance

****

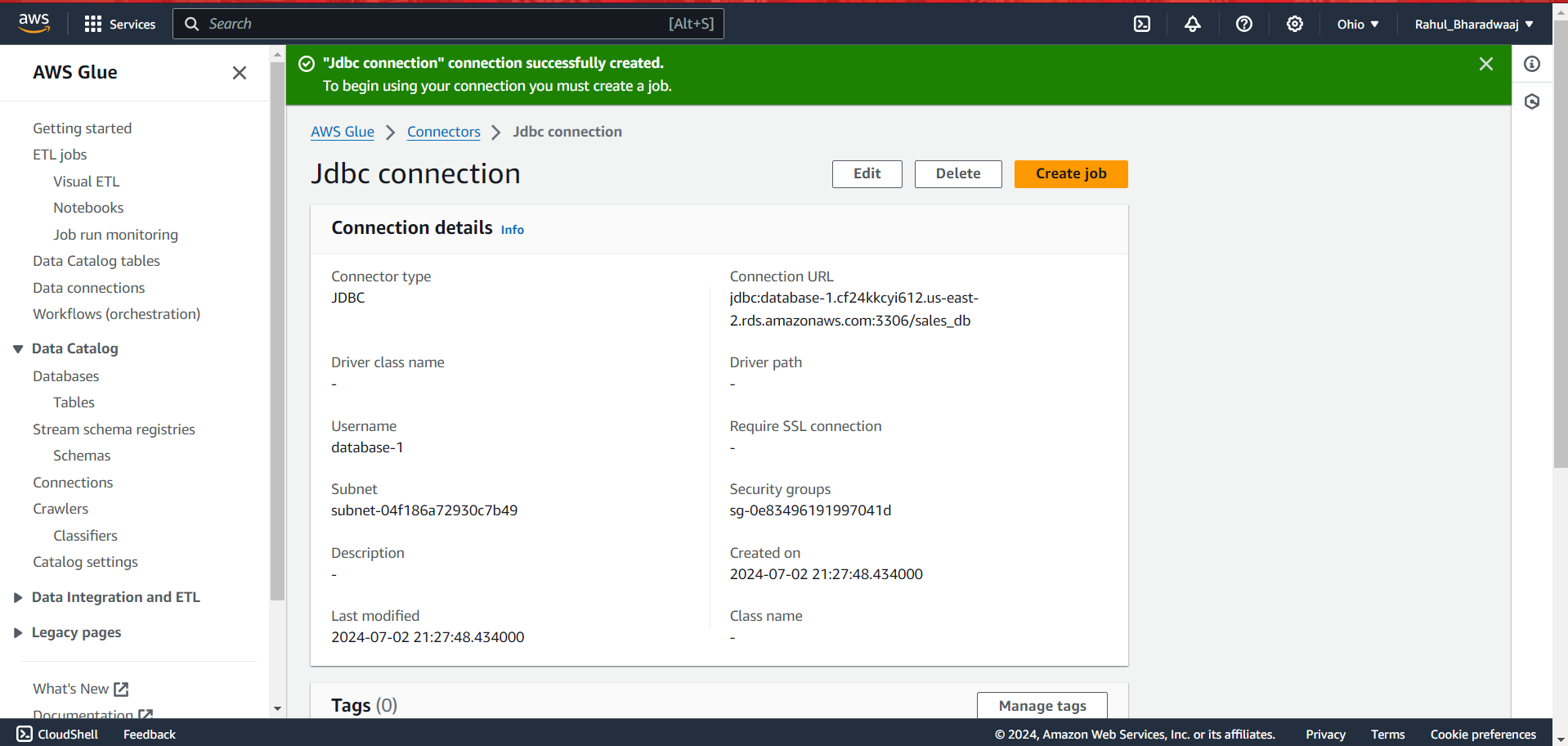
Created a database named sales\_db.

Created a table named sales\_data in the sales\_db database.

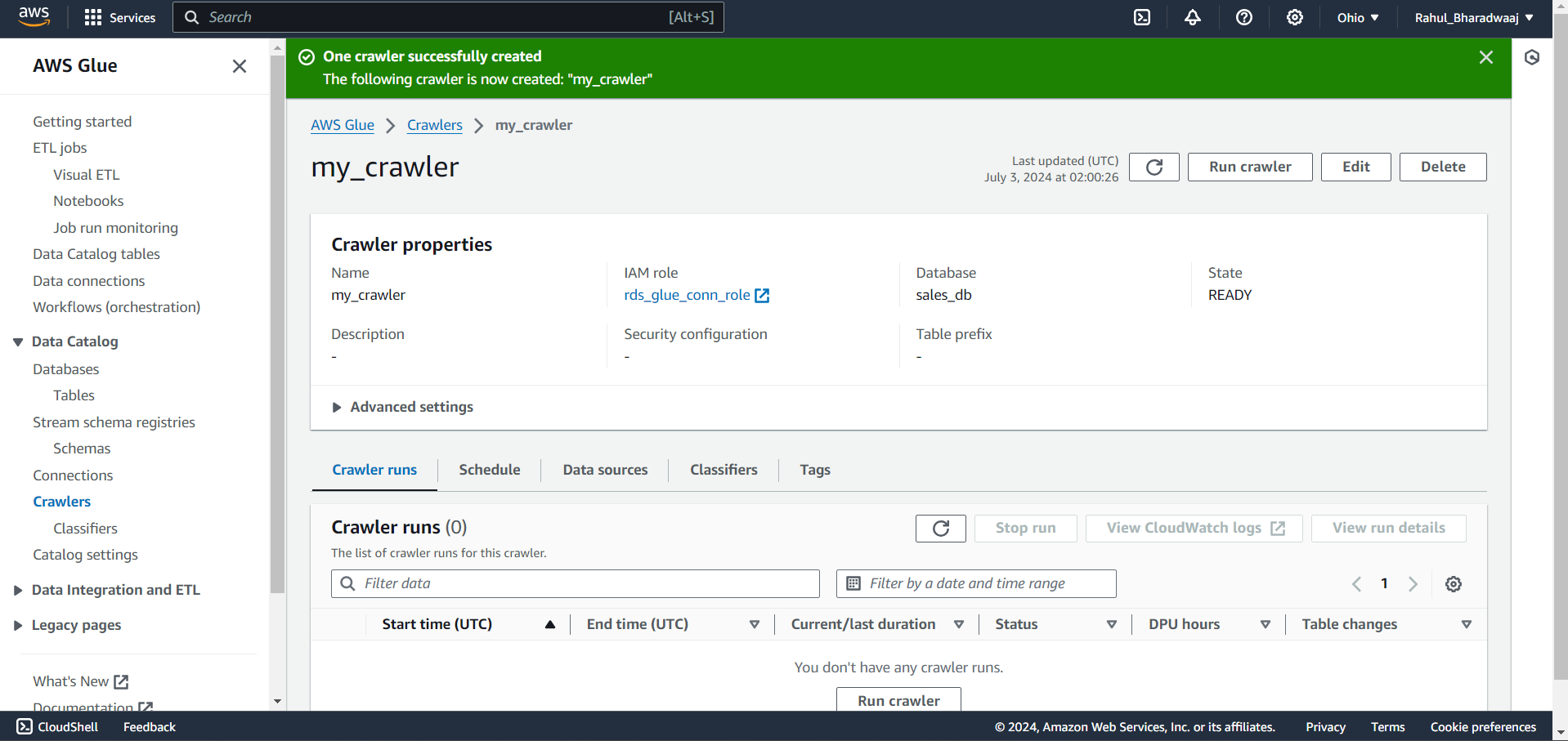


**Step 3: AWS Glue Configuration**

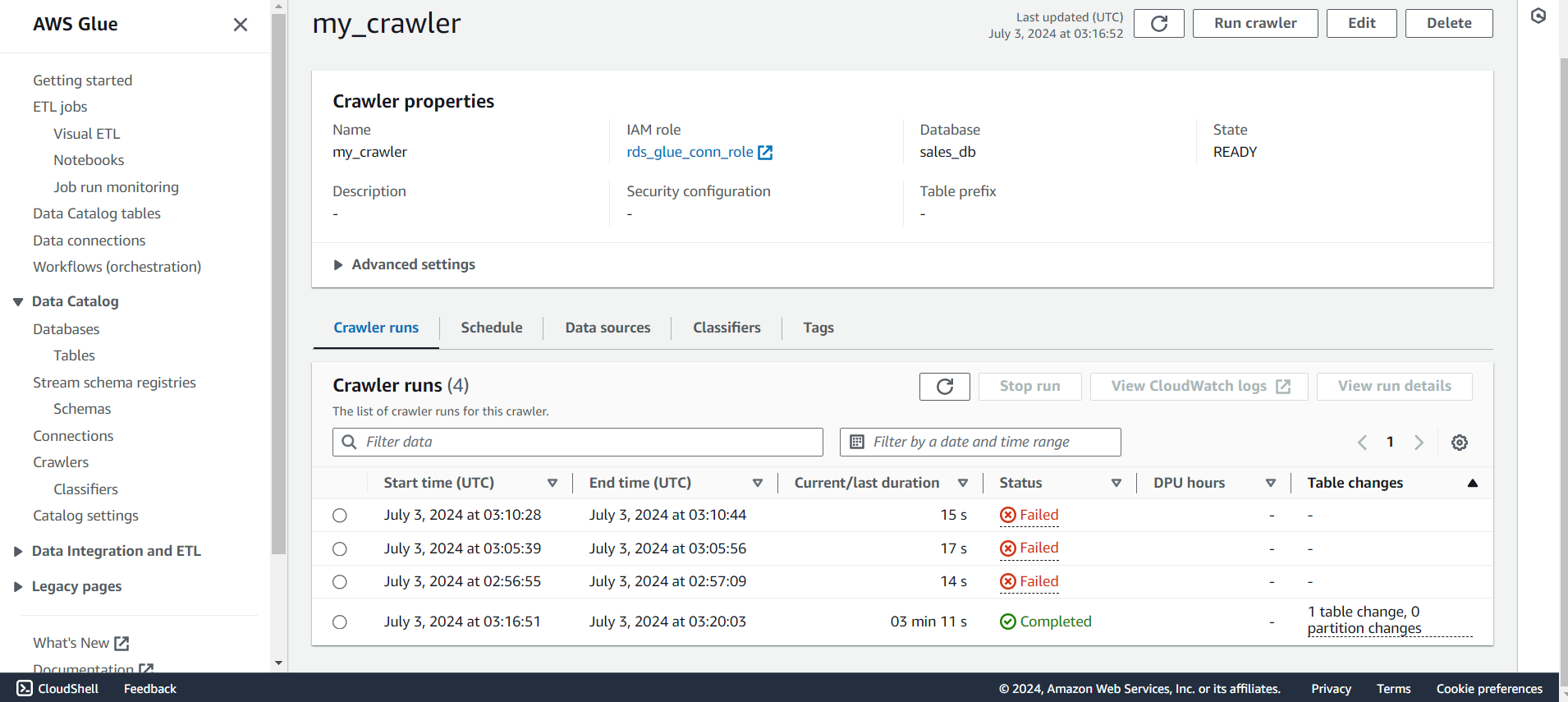
Configured AWS Glue with a JDBC connection to the RDS MySQL database.

****

Created a Glue crawler to catalog the table in the RDS database.



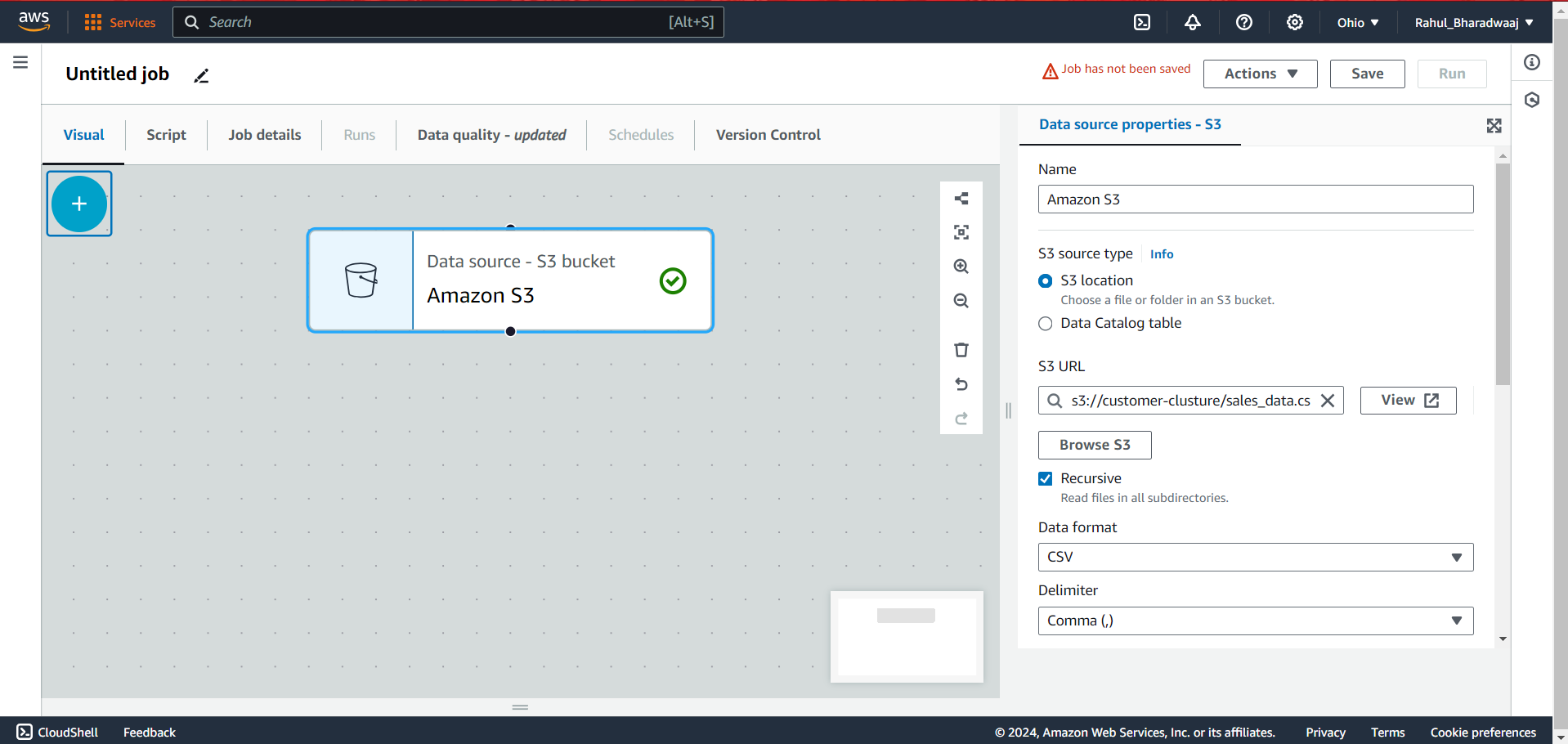
Ran the crawler to identify and catalog the table in the Glue Data Catalog.

****

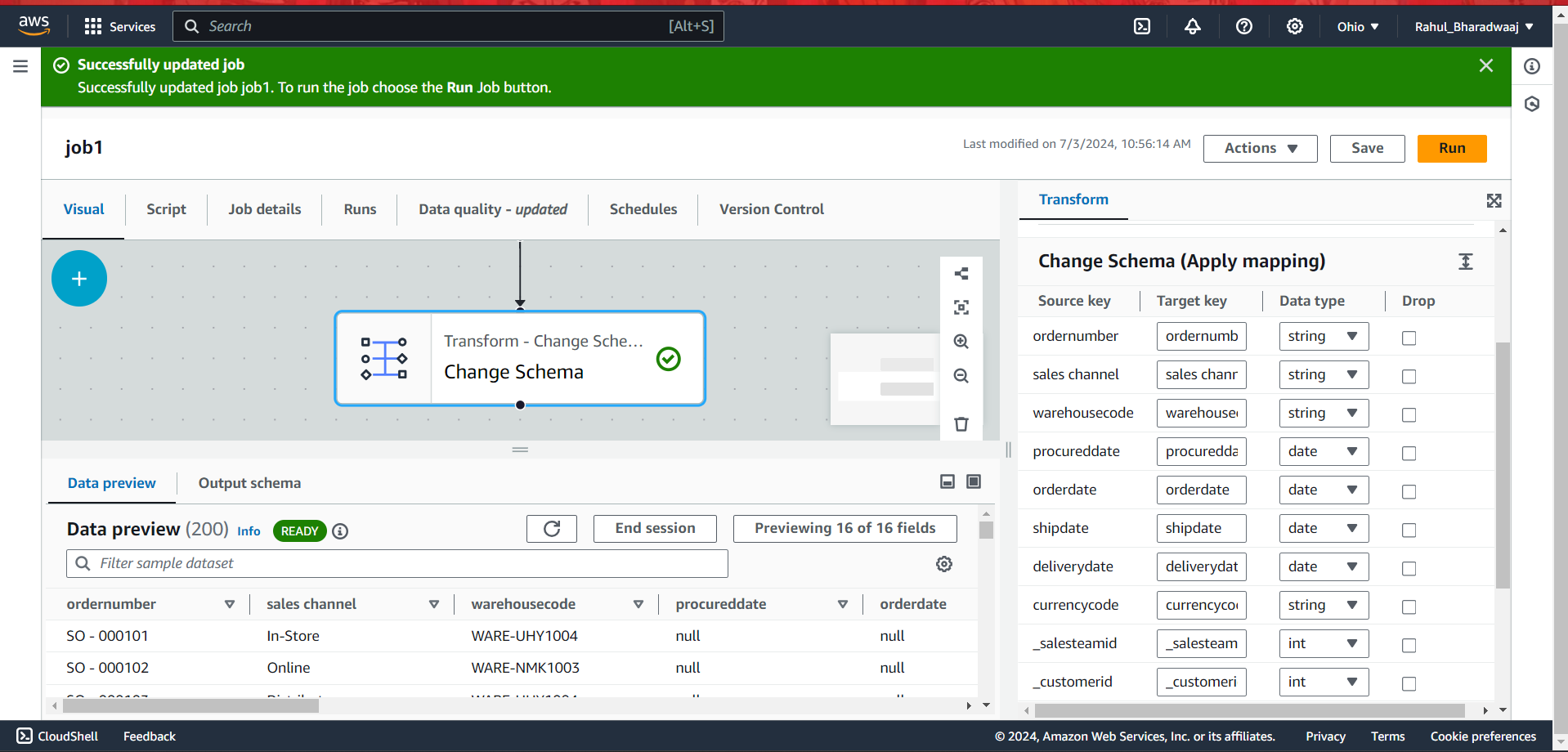
**Step 4: ETL Job Creation**

Created an ETL job in AWS Glue to extract data from the S3 bucket, transform it (if needed), and load it into the RDS MySQL database.

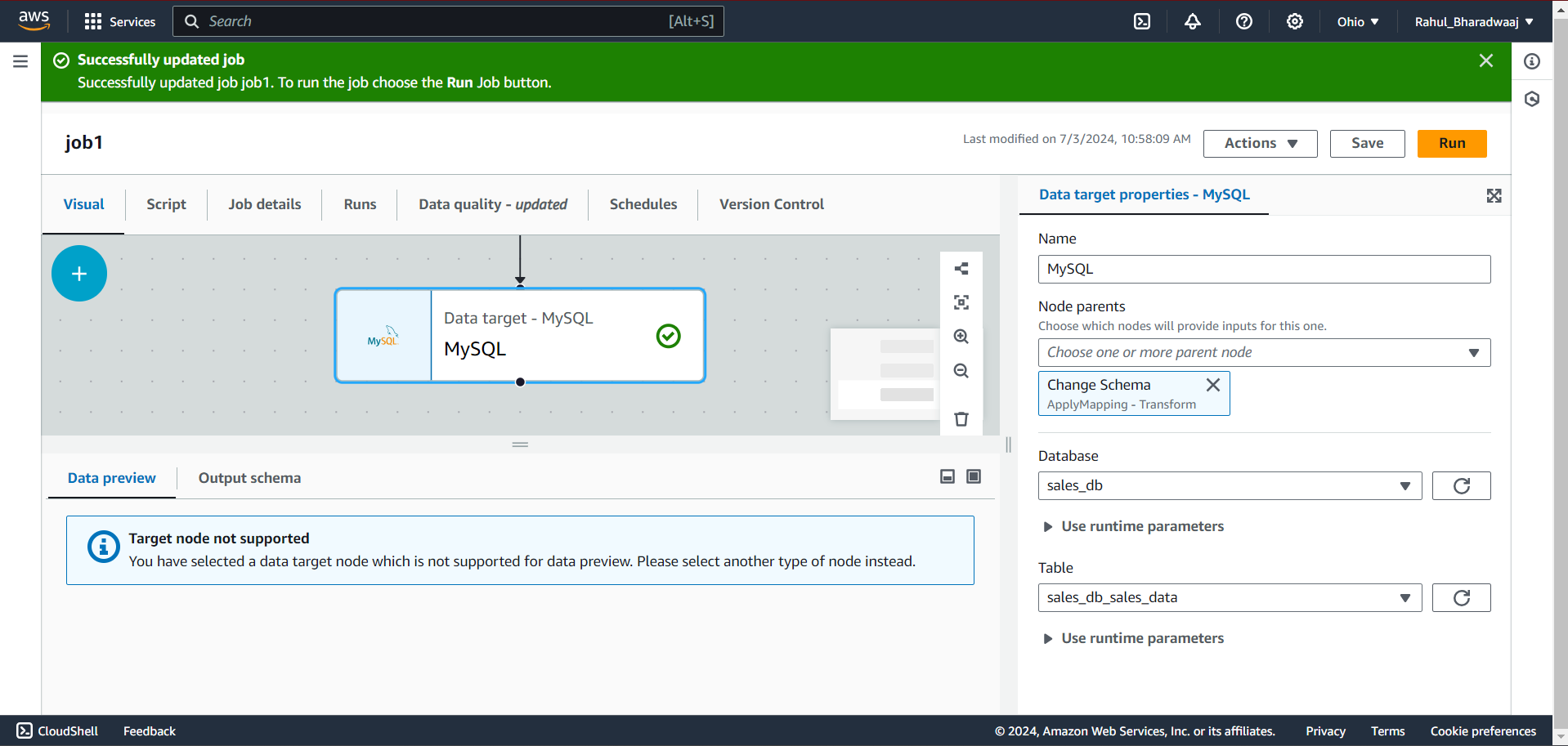
Data Source Configuration



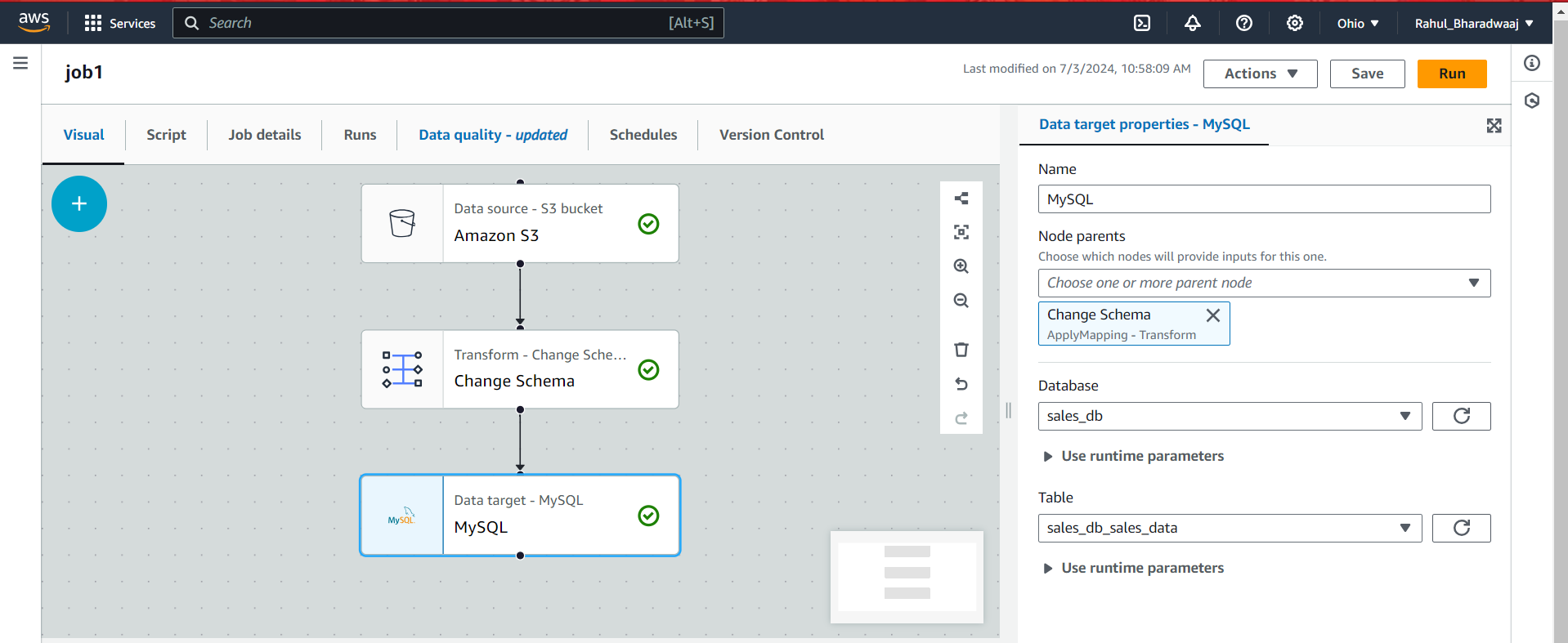
Schema Change Configuration



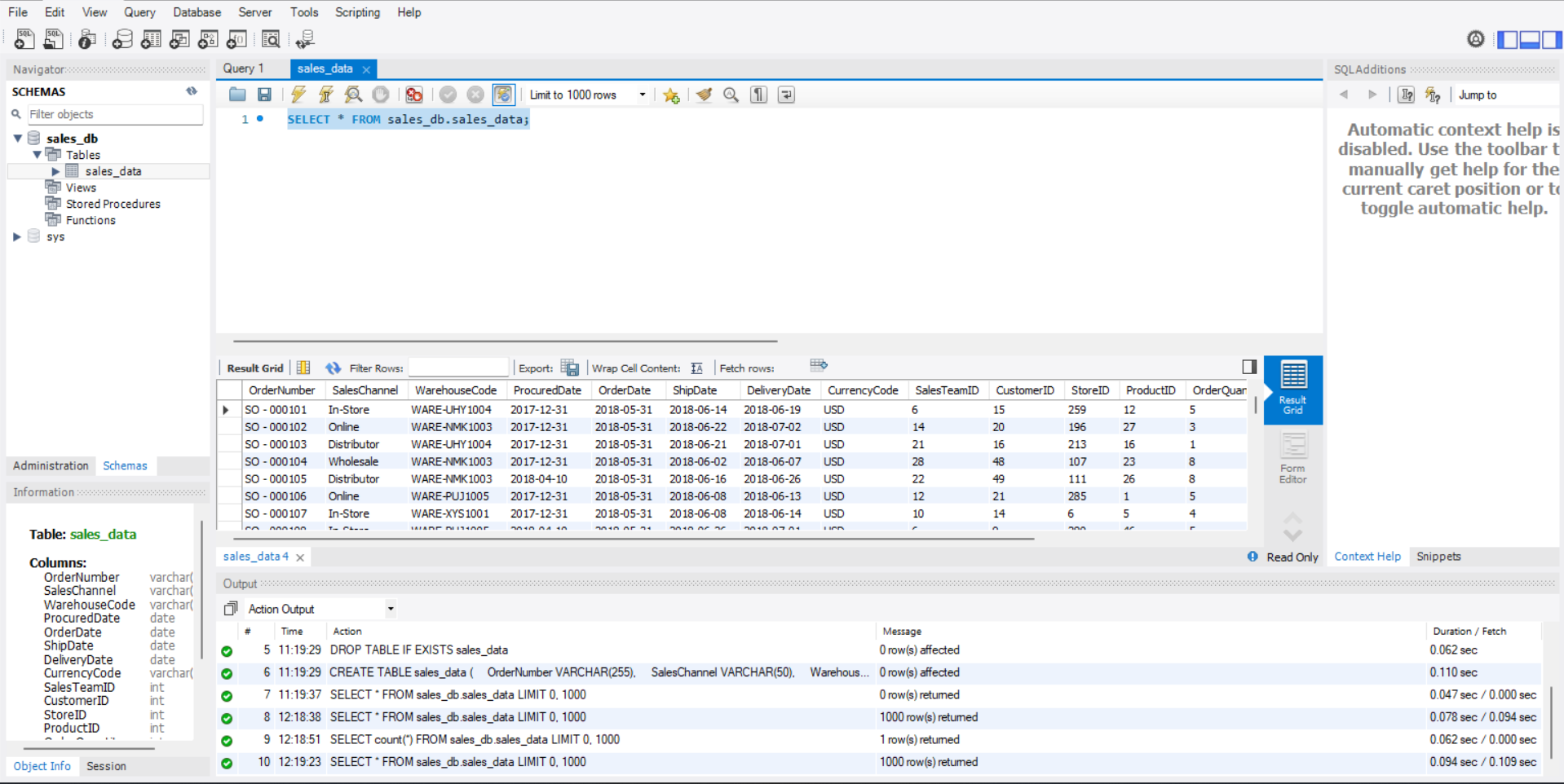
Data Target Configuration



ETL :



Data is successfully ingested into the Database.



1. **Data Analysis ( RFM analysis) :**

**Sales Channel Analysis**

**Question:** Which sales channel has the highest sales volume?

**Insights:**

* The In-Store channel has the highest sales volume with 14,878 orders.
* The Online channel follows with 10,897 orders.
* Distributor and Wholesale channels have 6,287 and 4,100 orders respectively.

**Monthly Sales Analysis**

**Question:** What are the monthly sales trends?

**Insights:**

* Sales show a seasonal pattern with peaks during certain months.
* Highest sales were recorded in December 2018 (1,387 orders) and January 2020 (1,258 orders).
* Lowest sales were observed in May 2018 (39 orders).

**Product Performance**

**Question:** Which products are the top sellers and what are the average discount rates?

**Insights:**

* Product 23 is the top seller with 956 units sold and an average discount of 11.13%.
* Other top-selling products include Product 37 (896 units), Product 8 (879 units), and Product 4 (878 units).
* Discounts range from approximately 10.17% to 12.78% for these products.

**Customer Insights**

**Question:** Who are the top customers based on total spending and order frequency?

**Insights:**

* Customer 12 is the top spender with $2,248,332.40 spent across 210 orders.
* Other significant customers include Customer 29 ($2,112,221.91 across 179 orders) and Customer 17 ($1,962,014.60 across 175 orders).
* High-value customers exhibit frequent purchasing patterns, with top customers placing between 135 to 210 orders.

**Store Sales Analysis**

**Question:** How do sales vary across different store locations?

**Insights:**

* Store 284 has the highest sales volume with 182 orders.
* Other high-performing stores include Store 26 (178 orders) and Store 166 (174 orders).
* Sales performance varies significantly across locations, indicating potential for targeted regional marketing strategies.

**RFM Analysis**

**Question:** How can we segment customers based on Recency, Frequency, and Monetary values?

**Insights:**

* Customers have been segmented into groups based on their purchasing behavior.
* High-value customers (e.g., Customer 12, Customer 29) show high frequency and monetary values.
* Recent purchases are observed from customers like Customer 15 (recency 1286 days, frequency 142, monetary value $1,512,725.99).

1. **Clustering and Analyzing Customer Segments :**

To further enhance the understanding of customer segments, clustering analysis was performed using the RFM values. The goal was to group customers into distinct clusters based on their purchasing behavior.

**Standardizing RFM Values**

To ensure each RFM metric contributes equally to the clustering process, the RFM values were standardized. This step involved transforming the data so that each metric had a mean of 0 and a standard deviation of 1.

**Determining the Optimal Number of Clusters**

The Elbow method was used to determine the optimal number of clusters. By plotting the sum of squared errors (SSE) for a range of cluster numbers, the "elbow point" where the SSE starts to level off was identified, suggesting the optimal number of clusters.

**Performing Clustering**

Using the optimal number of clusters identified from the Elbow method, KMeans clustering was applied to segment the customers. Each customer was assigned a cluster label, indicating their segment.

**Mapping Clusters to Segments**

After clustering, the characteristics of each cluster were analyzed by calculating the mean Recency, Frequency, and Monetary values. Based on these characteristics, clusters were mapped to meaningful segments.

**Analyzing Cluster Characteristics**

The characteristics of each cluster were analyzed to understand the behavior and value of customers in each segment. Key metrics such as average Recency, Frequency, Monetary values, and customer count for each cluster were calculated.

The clustered data, including the cluster labels, was written back to the database for seamless integration and analysis.

**Results**

1. **Cluster 0 - Potential Loyalists:**

Recency: 1287.4

Frequency: 149.88

Monetary: 1,510,802

Customer Count: 25

Description: Customers with moderate recency and frequency, with a significant monetary value. They could become more frequent buyers with the right engagement.

1. **Cluster 1 - Loyal:**

Recency: 1294.0

Frequency: 165.0

Monetary: 1,698,425

Customer Count: 8

Description: High-frequency and high-monetary customers, likely representing the most valuable segment. These are your loyal customers.

**Cluster 2 - At Risk:**

Recency: 1308.5

Frequency: 155.0

Monetary: 1,519,597

Customer Count: 2

Description: Customers with high recency but moderate frequency and monetary values. They are at risk of churning and need re-engagement.

**Cluster 3 - Champions:**

Recency: 1286.733

Frequency: 174.267

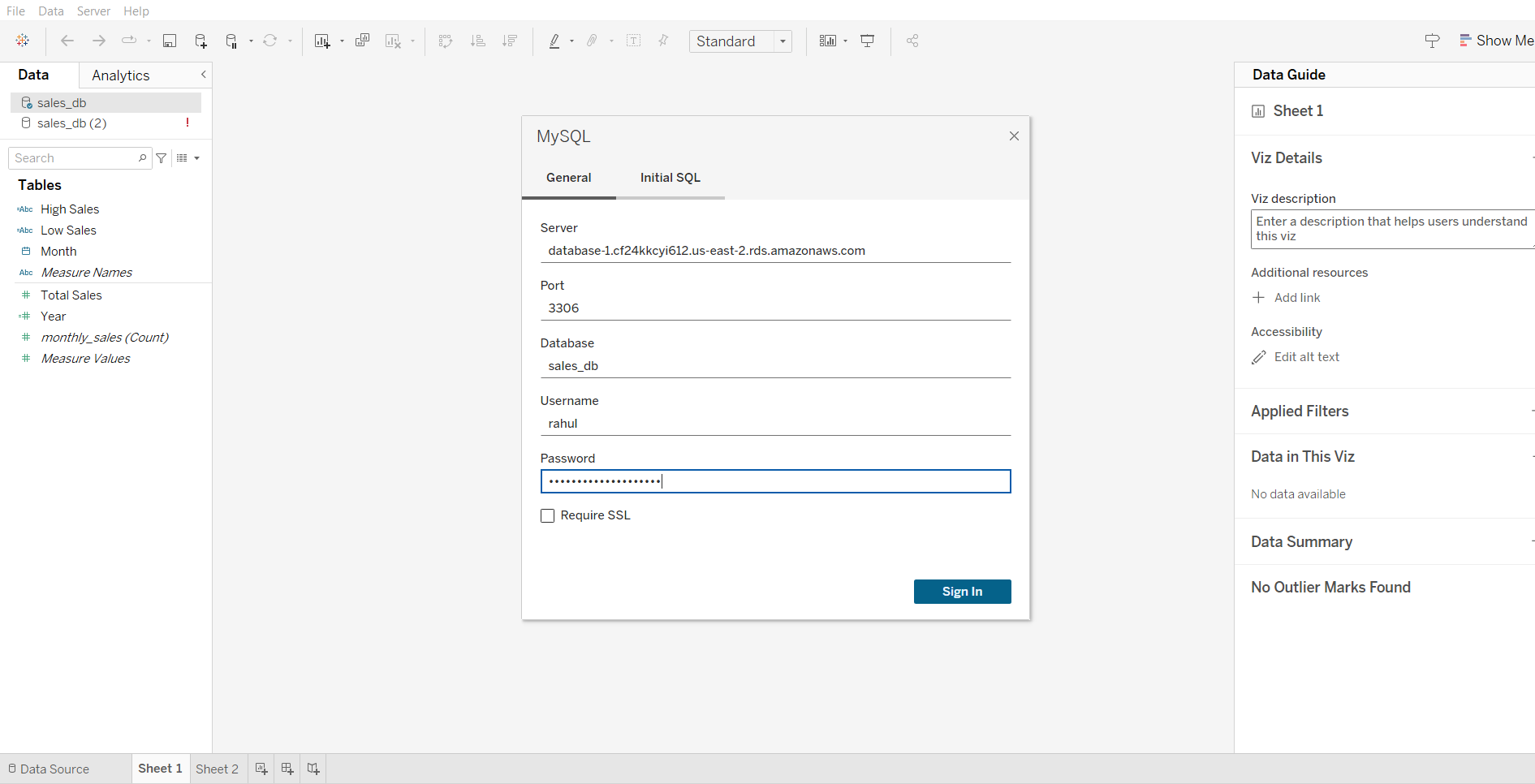
Monetary: 1,886,406

Customer Count: 15

Description: Customers with the highest frequency and monetary values, indicating the most engaged and high-value customers. These are your champions.

**5. Data Visualisation:**

**Initial Connection to AWS RDS:**

****

* Established a live connection between Tableau and AWS RDS for real-time data analysis.
* Documented the setup and configurations with screenshots for verification.

**Dashboard Components and Insights**

**Header and Filters:**

Objective: Provide an overview and interactive filters for users.

Components: Title, date range filter, sales channel filter, customer segment filter.

**KPI Tiles:**

**Objective:** Highlight key performance indicators for quick insights.

**Components:** Total Sales, Average Discount, Total Quantity Sold.

1. **Sales Performance Line Chart**

**Purpose:** To show total sales over time.

**Details:** Displays the total sales for each month

1. **Customer Analysis: RFM**

**Scatter Plot**

**Purpose:** To show the relationship between total spending and order frequency, colored by RFM segments.

**Bar Chart**

**Purpose:** To show the count of customers in each RFM segment.

1. **RFM Analysis Deep Dive**

**Bubble chart**

**Purpose:** To track the evolution of RFM metrics over time and understand the dynamics of customer segments.

1. **Product Performance**

**Treemap**

**Purpose:** To show total quantity sold and average discount per product.

1. **Sales Channel Performance**

**Donut Chart**

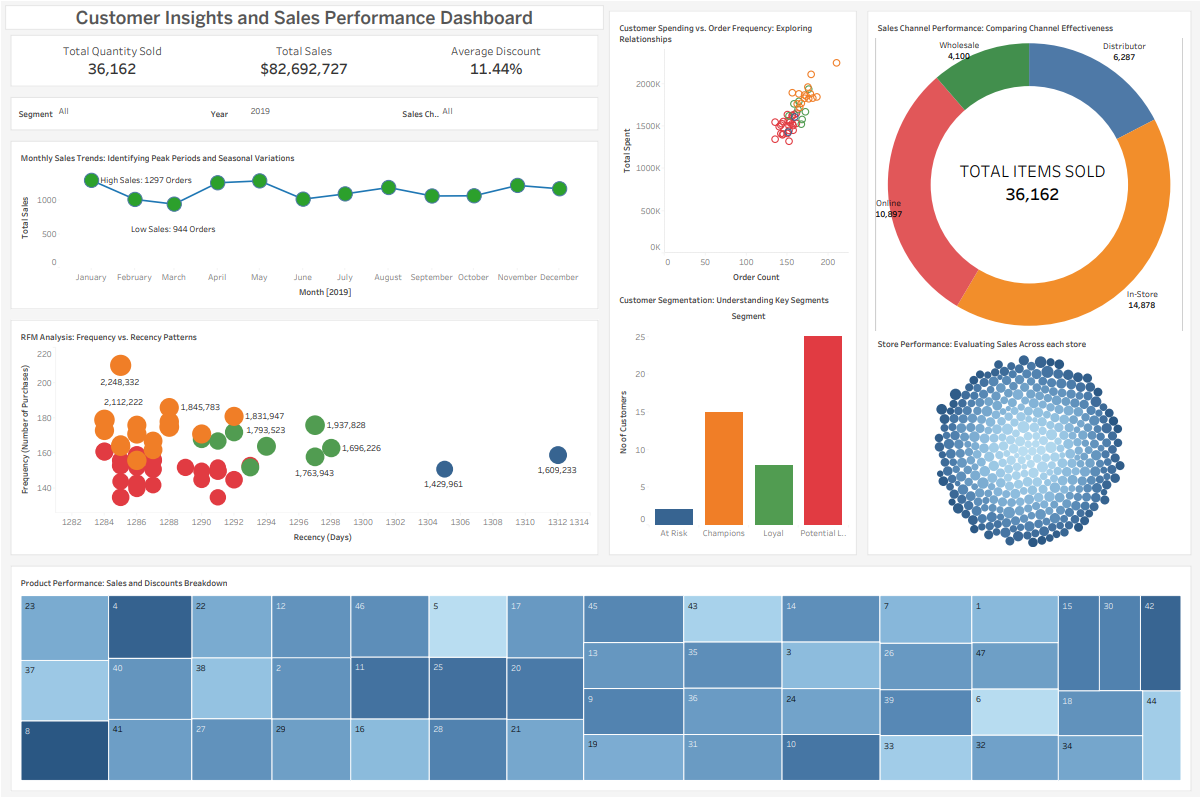
**Purpose:** To compare the performance of different sales channels based on the number of items sold.

1. **Store Performance**

**Stacked Bubble Chart**

**Purpose:** To show the number of items sold by each store.

Dashboard : <https://public.tableau.com/views/Customer_Dashboard_17215683155920/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>



**Conclusion and Insights**

**Overview:**

The analysis aimed to segment customers based on their purchasing behavior using RFM (Recency, Frequency, Monetary) analysis and clustering techniques. The goal was to identify distinct customer segments to optimize marketing strategies, enhance customer engagement, and drive business growth. The data-driven insights from the analysis provide a comprehensive understanding of customer behavior, which is crucial for tailoring marketing efforts effectively.

**Customer Segments and Insights:**

**Cluster 0 - Potential Loyalists:**

**Characteristics:** Customers with moderate recency and frequency, and a significant monetary value.

**Insights:** These customers have the potential to become more frequent buyers with the right engagement strategies.

**Actionable Strategies:**

* **Personalized Communication:** Send personalized emails or messages highlighting new arrivals, special offers, and loyalty programs to encourage more frequent purchases.
* **Exclusive Deals:** Offer exclusive discounts or early access to sales events to make them feel valued and encourage loyalty.
* **Feedback Requests:** Engage with these customers by requesting feedback on their shopping experience and preferences, which can help tailor future communications and offers.

**Cluster 1 - Loyal Customers:**

**Characteristics:** High-frequency and high-monetary customers, representing the most valuable segment.

**Insights:** These are the loyal customers who consistently generate significant revenue.

**Actionable Strategies:**

* **VIP Programs:** Create a VIP or loyalty program with exclusive benefits such as special discounts, priority customer service, and personalized recommendations.
* **Recognition:** Recognize and reward their loyalty through personalized thank-you notes, special gifts, or shout-outs on social media.
* **Engagement Campaigns:** Keep these customers engaged with regular updates on new products, upcoming events, and exclusive previews.

**Cluster 2 - At-Risk Customers:**

**Characteristics:** Customers with high recency but moderate frequency and monetary values.

**Insights:** These customers are at risk of churning and need re-engagement to prevent them from becoming inactive.

**Actionable Strategies:**

* **Re-Engagement Campaigns:** Implement re-engagement campaigns with special offers or incentives to encourage them to return and make purchases.
* **Win-Back Strategies:** Send personalized messages expressing that they are missed and offering a limited-time discount to entice them back.
* **Customer Service Outreach:** Reach out through customer service to address any potential issues or concerns they might have had, showing that their satisfaction is a priority.

**Cluster 3 - Champions:**

**Characteristics:** Customers with the highest frequency and monetary values, indicating the most engaged and high-value customers.

**Insights:** These are the top customers who not only purchase frequently but also spend significantly.

**Actionable Strategies:**

* **Exclusive Events:** Invite these customers to exclusive events, such as product launches or private sales, to make them feel valued and appreciated.
* **Personalized Offers:** Provide highly personalized offers and recommendations based on their purchase history and preferences.
* **Loyalty Reinforcement:** Continuously reinforce their loyalty through special loyalty rewards, such as free gifts with purchases or additional loyalty points.

**Conclusion:**

The customer segmentation analysis provides valuable insights that can help tailor marketing strategies to the specific needs and preferences of each segment. By focusing on personalized communication, exclusive deals, and engagement campaigns, the retail chain can enhance customer satisfaction, foster loyalty, and ultimately drive business growth. The actionable strategies outlined above for each customer segment will help in creating targeted marketing efforts that are more effective and efficient, leading to improved overall performance and customer retention.