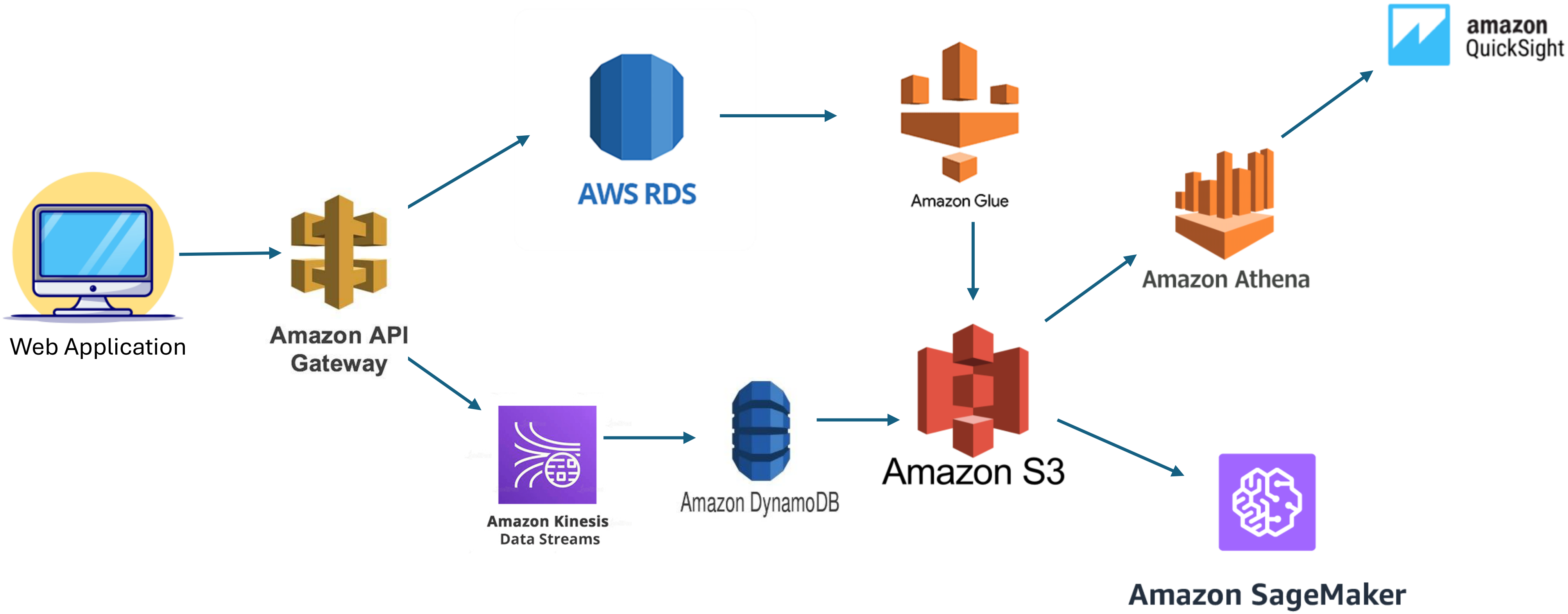
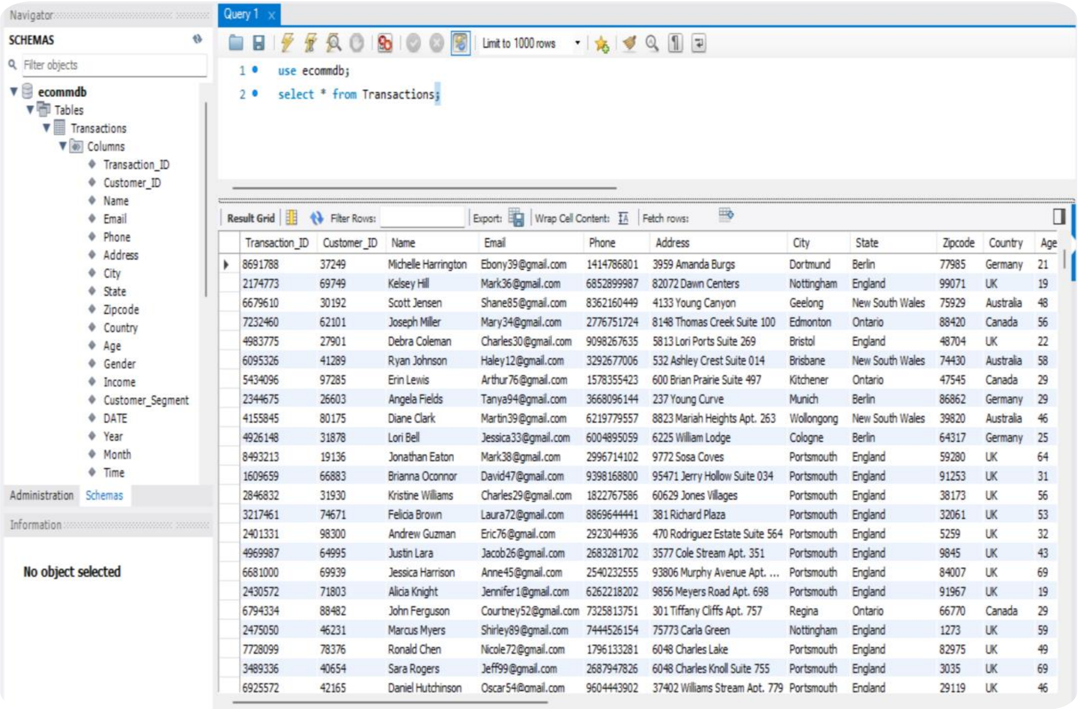
Data Engineering Ecommerce End to End Project

The E-commerce Data Pipeline and Analytics project is an end-to-end solution leveraging AWS (Kinesis, Glue, S3, Athena, QuickSight, SageMaker) to ingest, transform, and analyze transactional and clickstream data. It delivers real-time insights, dashboards, and personalized recommendations, enhancing customer engagement and business decisions.

Process Flow Chart



AWS RDS



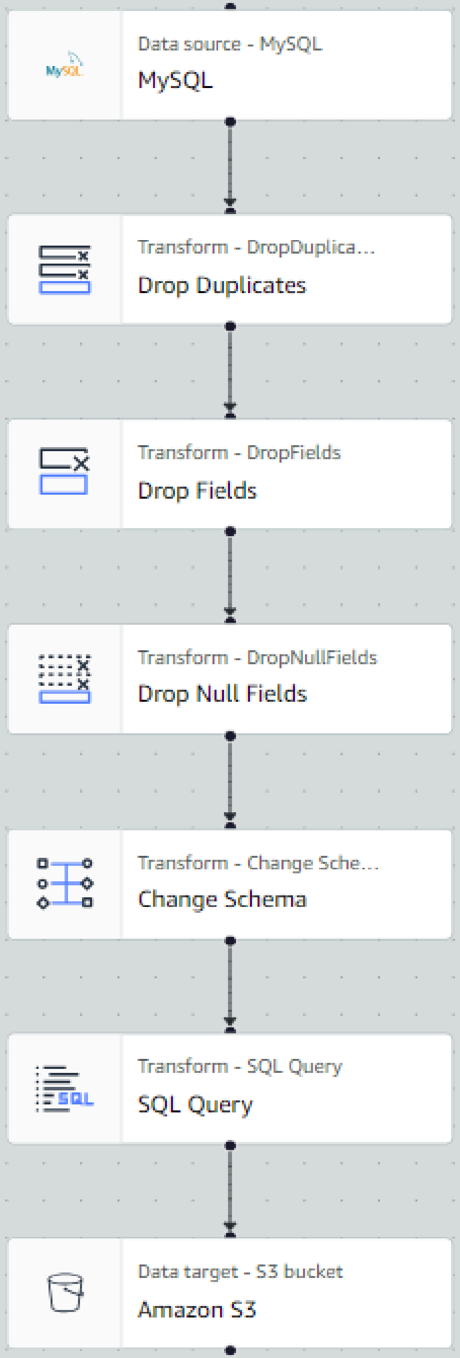
```
use ecommdb;
select * from Transactions;
```

Transaction_ID	Customer_ID	Name	Email	Phone	Address	City	State	Zipcode	Country	Age
8691788	37249	Michelle Harrington	Ebony39@gmail.com	1414786801	3959 Amanda Burgs	Dortmund	Berlin	77985	Germany	21
2174773	69749	Kelsey Hill	Mark36@gmail.com	6852899987	82072 Dawn Centers	Nottingham	England	99071	UK	19
6679610	30192	Scott Jensen	Shane85@gmail.com	8362160449	4133 Young Canyon	Geelong	New South Wales	75929	Australia	48
7232460	62101	Joseph Miller	Mary34@gmail.com	2776751724	8148 Thomas Creek Suite 100	Edmonton	Ontario	88420	Canada	56
4983775	27901	Debra Coleman	Charles30@gmail.com	9098267635	5813 Lori Ports Suite 269	Bristol	England	48704	UK	22
6095326	41289	Ryan Johnson	Haley12@gmail.com	3292677006	532 Ashley Crest Suite 014	Brisbane	New South Wales	74430	Australia	58
5434096	97285	Erin Lewis	Arthur76@gmail.com	1578355423	600 Brian Prairie Suite 497	Kitchener	Ontario	47545	Canada	29
2344675	26603	Angela Fields	Tanya94@gmail.com	3668096144	237 Young Curve	Munich	Berlin	86862	Germany	29
4155845	80175	Diane Clark	Martin39@gmail.com	6219779557	8823 Mariah Heights Apt. 263	Wollongong	New South Wales	39820	Australia	46
4926148	31878	Lori Bell	Jessica33@gmail.com	6004895059	6225 William Lodge	Cologne	Berlin	64317	Germany	25
8493213	19136	Jonathan Eaton	Mark38@gmail.com	2996714102	9772 Sosa Coves	Portsmouth	England	59280	UK	64
1609659	66883	Brianna Oconnor	David47@gmail.com	9398168800	95471 Jerry Hollow Suite 034	Portsmouth	England	91253	UK	31
2846832	31930	Kristine Williams	Charles29@gmail.com	1822767586	60629 Jones Villages	Portsmouth	England	38173	UK	56
3217461	74671	Felicia Brown	Laura72@gmail.com	8869644441	381 Richard Plaza	Portsmouth	England	32061	UK	53
2401331	98300	Andrew Guzman	Eric76@gmail.com	2923044936	470 Rodriguez Estate Suite 564	Portsmouth	England	5259	UK	32
4969987	64995	Justin Lara	Jacob26@gmail.com	2683281702	3577 Cole Stream Apt. 351	Portsmouth	England	9845	UK	43
6681000	69939	Jessica Harrison	Anne45@gmail.com	2540232555	93806 Murphy Avenue Apt. ...	Portsmouth	England	84007	UK	69
2430572	71803	Alicia Knight	Jennifer1@gmail.com	6262218202	9856 Meyers Road Apt. 698	Portsmouth	England	91967	UK	19
6794334	88482	John Ferguson	Courtney52@gmail.com	7325813751	301 Tiffany Cliffs Apt. 757	Regina	Ontario	66770	Canada	29
2475050	46231	Marcus Myers	Shirley89@gmail.com	7444526154	75773 Carla Green	Nottingham	England	1273	UK	59
7728099	78376	Ronald Chen	Nicole72@gmail.com	1796133281	6048 Charles Lake	Portsmouth	England	82975	UK	49
3489336	40654	Sara Rogers	Jeff99@gmail.com	2687947826	6048 Charles Knoll Suite 755	Portsmouth	England	3035	UK	69
6925572	42165	Daniel Hutchinson	Oscar54@gmail.com	9604443902	37402 Williams Stream Apt. 779	Portsmouth	England	29119	UK	46

- The web application sends transactional data to secure RESTful endpoints created in AWS API Gateway, ensuring reliable data reception.
- These endpoints are integrated with AWS Lambda functions, facilitating the data processing.
- The processed data is seamlessly and securely loaded into Amazon RDS.

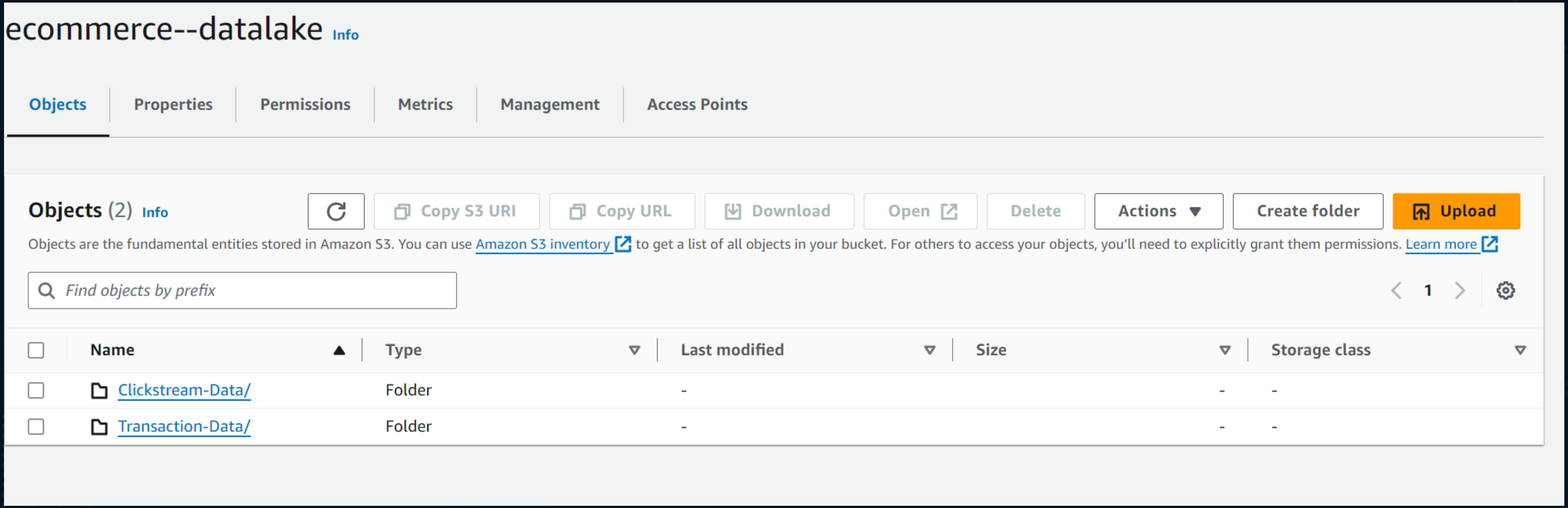
AWS GLUE

- **Extract**: Data is extracted from the MySQL database containing e-commerce transactions.
- **Transform**:
  - Drop duplicates to maintain data quality.
  - Drop unnecessary fields to reduce data size.
  - Remove null fields to ensure completeness.
  - Change schema to standardize data structure.
  - Apply custom SQL queries for complex transformations.
- **Load**: The cleaned and transformed data is loaded into Amazon S3, ready for further use.



S3 Data Lake

- The primary objective of the S3 data lake is to consolidate various data sources into a single, scalable storage solution. This allows for:
- **Enhanced Data Integration**: Combining transactional and clickstream data for a unified view.
- **Improved Data Accessibility**: Making data easily accessible for analytics and machine learning.
- **Scalable and Cost-Effective Storage**: Leveraging the scalability and cost-efficiency of Amazon S3.



ecommerce--datalake Info

Objects (2) Info

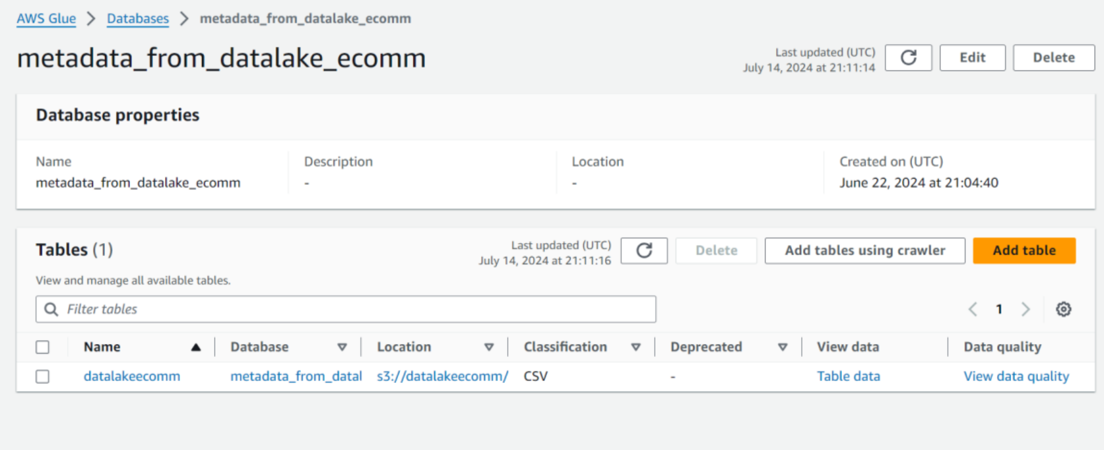
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	Clickstream-Data/	Folder	-	-	-
□	Transaction-Data/	Folder	-	-	-

AWS Athena

AWS Glue Data Catalog was used to create and manage metadata for all transactional data stored in S3.

Glue Crawlers were configured to automatically scan the S3 bucket, infer the schema of new or modified data, and update the catalog



AWS Glue > Databases > metadata_from_datalake_ecomm

metadata_from_datalake_ecomm

Last updated (UTC)
July 14, 2024 at 21:11:14

Database properties

Name	Description	Location	Created on (UTC)
metadata_from_datalake_ecomm	-	-	June 22, 2024 at 21:04:40

Tables (1)

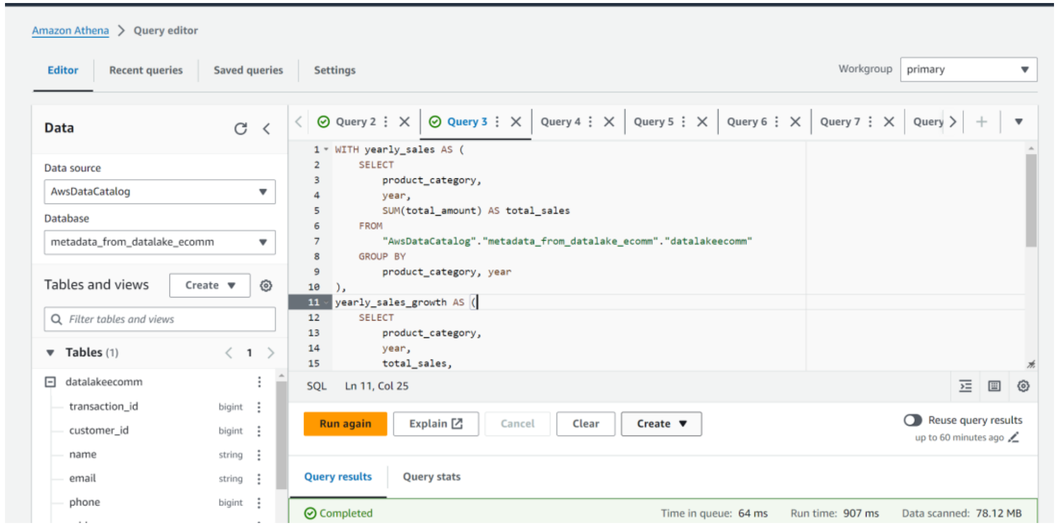
Last updated (UTC)
July 14, 2024 at 21:11:16

View and manage all available tables.

	Name	Database	Location	Classification	Deprecated	View data	Data quality
☐	datalakeecomm	metadata_from_datal	s3://datalakeecomm/	CSV	-	Table data	View data quality

Amazon Athena was used to perform interactive SQL queries on the data stored in S3, leveraging the metadata defined in the Glue Data Catalog.

Athena's integration with Glue allowed it to utilize the schema information for efficient query execution. This enabled quick analysis of large datasets without the need to manage any infrastructure.



Amazon Athena > Query editor

Workgroup: primary

Data source: AwsDataCatalog

Database: metadata_from_datalake_ecomm

Tables (1)

datalakeecomm	
transaction_id	bigint
customer_id	bigint
name	string
email	string
phone	bigint

```
WITH yearly_sales AS (
    SELECT
        product_category,
        year,
        SUM(total_amount) AS total_sales
    FROM
        "AwsDataCatalog"."metadata_from_datalake_ecomm"."datalakeecomm"
    GROUP BY
        product_category, year
),
yearly_sales_growth AS (
    SELECT
        product_category,
        year,
        total_sales,
```

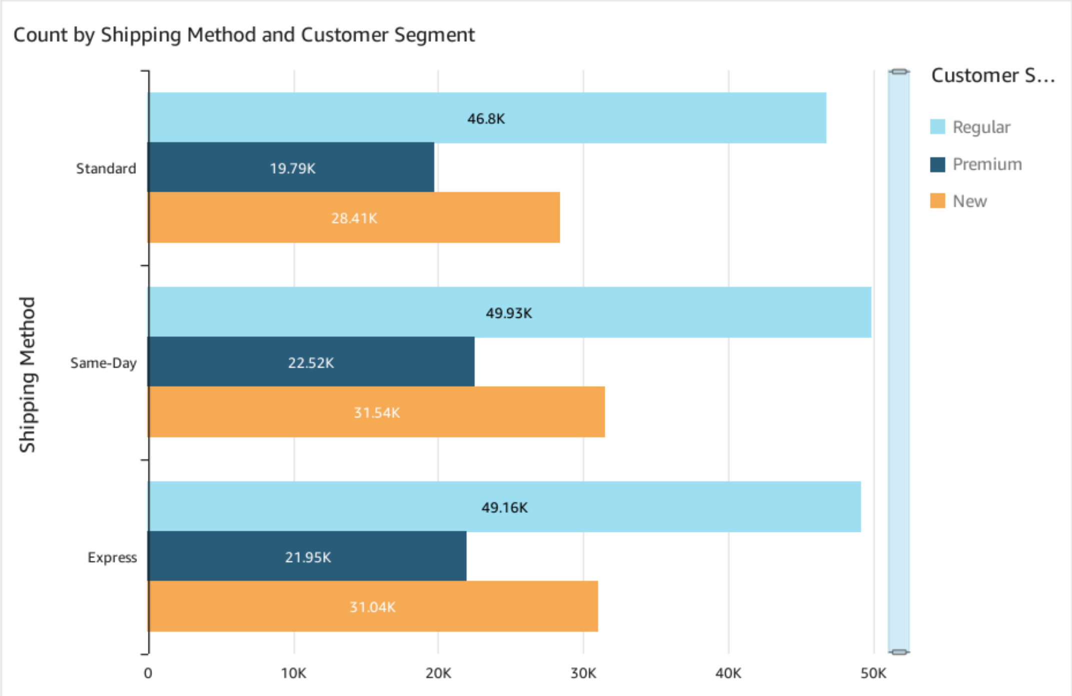
Ln 11, Col 25

Reuse query results
up to 60 minutes ago

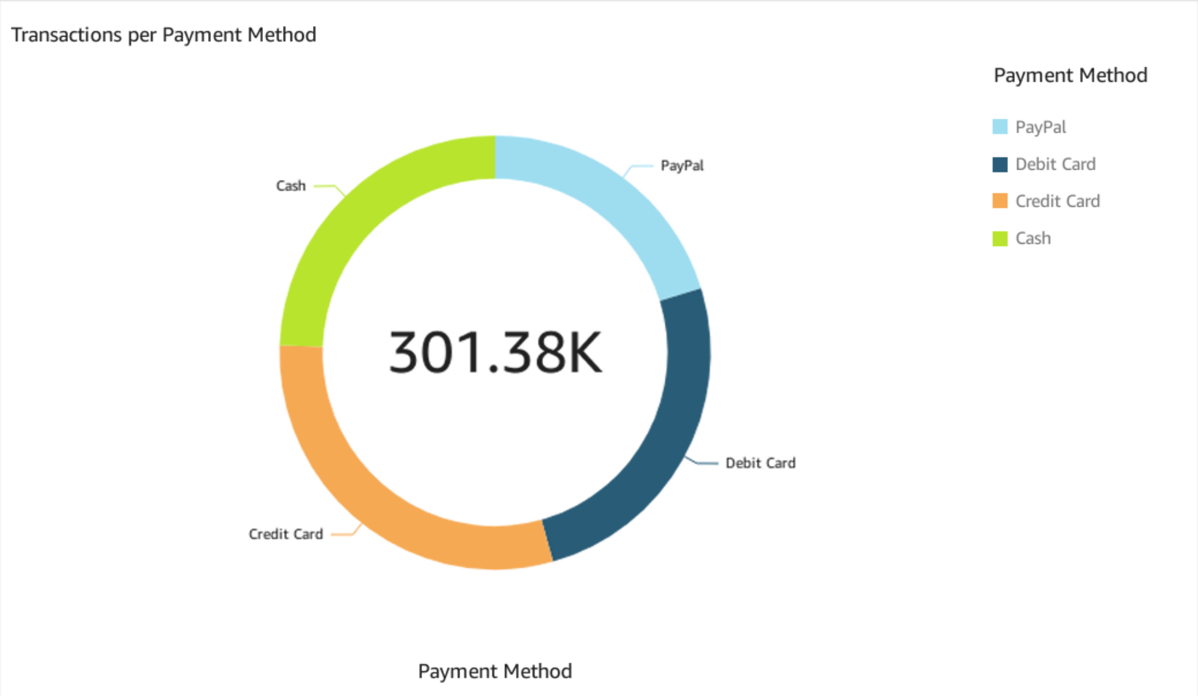
Completed   Time in queue: 64 ms   Run time: 907 ms   Data scanned: 78.12 MB

AWS Quicksight

- Count by Shipping Method and Customer Segment:
  - Regular customers have the most transactions, especially with Standard shipping.
  - Premium customers follow, with New customers having the least transactions.
- Transactions per Payment Method:
  - Credit Card is the most used payment method, followed by Debit Card and PayPal.
  - Cash transactions are the least common.



Count by Shipping Method and Customer Segment



Transactions per Payment Method

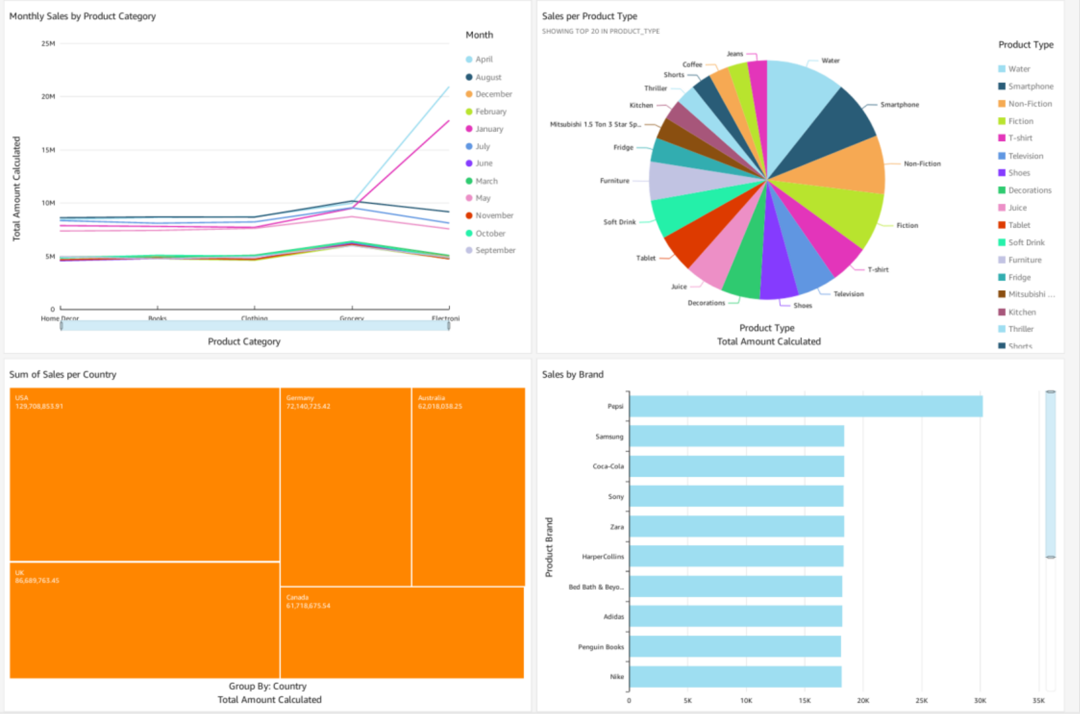
AWS Quicksight

- Monthly Sales by Product Category:
  - Sales peak in December, especially for Electronics, Grocery, and Clothing.
  - Home Decor and Books have lower sales.

- Sales per Product Type:
  - Top-selling items: Smartphones, Televisions, Furniture.
  - Notable sales in consumables like Water and Juice.
  - Lower sales in Shorts and Thriller books.

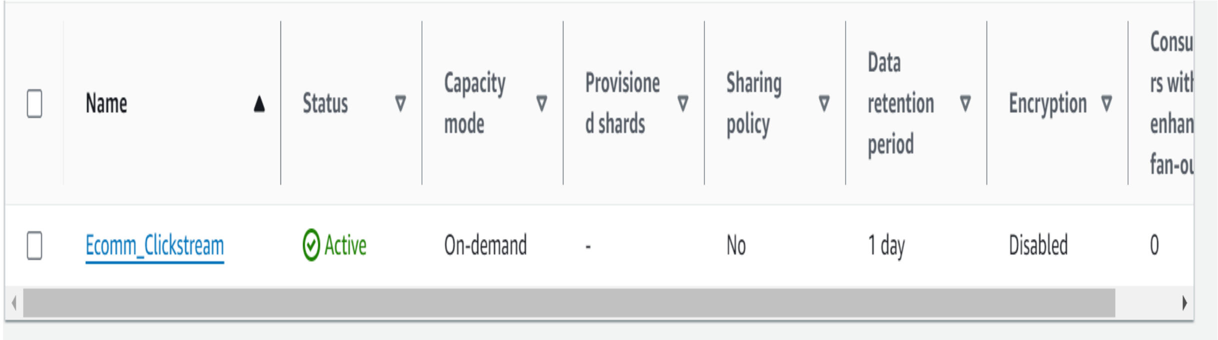
- Sum of Sales per Country:
  - USA leads in total sales, followed by the UK and Germany.
  - Australia and Canada have the least sales.

- Sales by Brand:
  - Leading brands: Nike, Adidas, Sony, Samsung.
  - Strong performance in beverages (Coca-Cola, Pepsi) and books (Penguin Books, HarperCollins).
  - Consistent but moderate sales for retail brands (Bed Bath & Beyond, Zara).



Kinesis

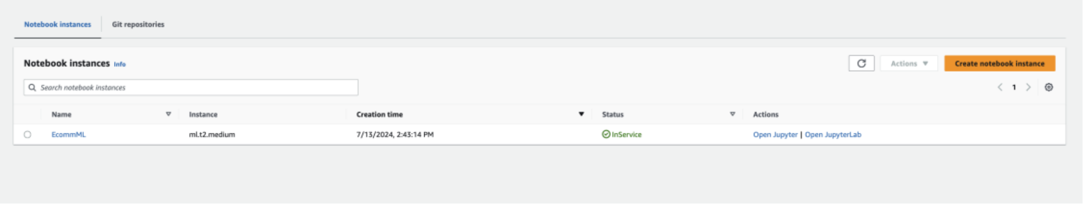
- Real-time Ingestion and Scalability: Collects and ingests real-time clickstream data from the web application, automatically scaling to handle high data volumes and ensuring data durability through replication across multiple Availability Zones.
- Processing and Storage: Processes clickstream data from Kinesis, generates user sessions based on a 30-minute inactivity rule, and writes the processed data to Amazon DynamoDB, scaling automatically with the data load and ensuring cost efficiency by charging only for actual processing time.



□	Name ▲	Status ▽	Capacity mode ▽	Provisioned shards ▽	Sharing policy ▽	Data retention period ▽	Encryption ▽	Consu rs wit enhan fan-ou
□	Ecomm_Clickstream	✓ Active	On-demand	-	No	1 day	Disabled	0

Sagemaker ML

- Developed Recommendation System: Leveraged Amazon SageMaker to create a recommendation system using clickstream data stored in Amazon DynamoDB.
- Data Preparation: Extracted and transformed user activity data, including session details and transaction IDs, into a format suitable for machine learning.
- Exploratory Data Analysis (EDA): Used Jupyter notebooks on SageMaker to perform EDA and visualizations, gaining insights into user behavior.
- Model Training: Trained a machine learning model on SageMaker to analyze user behavior and generate personalized recommendations.
- Real-time Recommendations: Deployed the model to provide real-time recommendations, continuously updating with new data.
- Enhanced User Engagement: Delivered tailored content to users, enhancing engagement and satisfaction with the web application.



Notebook instances Info

	Name	Instance	Creation time	Status	Actions
○	EcommML	ml.t2.medium	7/13/2024, 2:43:14 PM	InService	Open Jupyter | Open JupyterLab