DTSC 701/CSCI 626 : Intro to Big Data

**Analyzing Customer Ratings to Identify Top-rated Products**

* Amazon E-commerce Market in 2023

Team member: Ramis Rawnak, Gahyeon Back, Nafiah Alam, Ravi Kiran Reddy Pulikadam

horizontal line

# 

# Index

1. Data

2. Goal of the Project

3. Technical Approach for Big Data Analytics

4. AWS Conﬁguration - S3 Bucket, Glue Crawler and IAM role

5. ETL Data Pipeline

6. Queries on Athena

7. Conclusion

8. Challenges and Future Analysis Idea

## 

## 1. Data

DATA USED IN PROJECT: Amazon Product Dataset 2023 from Kaggle  
Data description: The size of dataset is 358 GB. The format of data is .csv.

### (1) Key Features

* PRODUCT RATINGS: It highlights the importance of product ratings as a primary indicator of customer satisfaction and product quality.
* QUANTITY OF REVIEWS: :It indicates that the number of reviews associated with a product provides valuable insights into the level of customer engagement and the overall popularity of the product.
* PRODUCT OF CATEGORIES :It suggests that analyzing ratings by product category can reveal trends and patterns in customer preferences across different product types
* PRODUCT PRICES:It suggests that understanding the relationship between customer ratings and product prices can provide insights into pricing strategies and customer value perception.
* PRODUCT DESCRIPTIONS: It suggests that analyzing product descriptions alongside ratings can provide a more comprehensive understanding of customer feedback and identify areas for product improvement.

### (2) Columns

* asin: type-string | Product ID from Amazon.
* title: type-string | Title of the product.
* imUrl: type-string | Url of the product image.
* product URL: type-string | Url of the product.
* stars: type-ﬂoat | Product rating, If 0, no ratings were found.
* reviews: type-int | Number of reviews, If 0, no reviews were found.
* price: type-ﬂoat | currency-USD | Buy now price of the product. If 0, price was unavailable.
* ListPrice: type-ﬂoat | currency- USD | Original price of product before discount. If 0, unavailable.
* category\_id: type-int | Use the amazon\_categories.cvs at dataset to ﬁnd the actual category
* name.
* isBestSeller: type-boolean | Whether the product had the Amazon BestSeller status or not.
* BoughtinLastMonth: type-int | Sales quantity of the product occurred in Sep 2023 ONLY.

## 

## 2. Goal

### (1) Goal

Analyze customer ratings to identify and extract the top-rated products list.

### (2) Significance/Relevance

Analyzing customer ratings is a crucial aspect of understanding market dynamics. It provides businesses with actionable data to make informed decisions, tailor their offerings to meet customer expectations, and ultimately drive success in a competitive market.

## 3. Technical Approach

### (1) Why did we choose Data Lake?

* Flexibility: Data lakes can store raw, unstructured, and structured data, providing flexibility in handling various data formats.
* Scalability: Data lakes, especially cloud-based solutions, can scale horizontally to accommodate growing datasets.
* Cost-Efficiency: Cloud-based data lakes often offer a pay-as-you-go model, making it cost-effective for storing large amounts of data.

### (2) What Services did we choose?

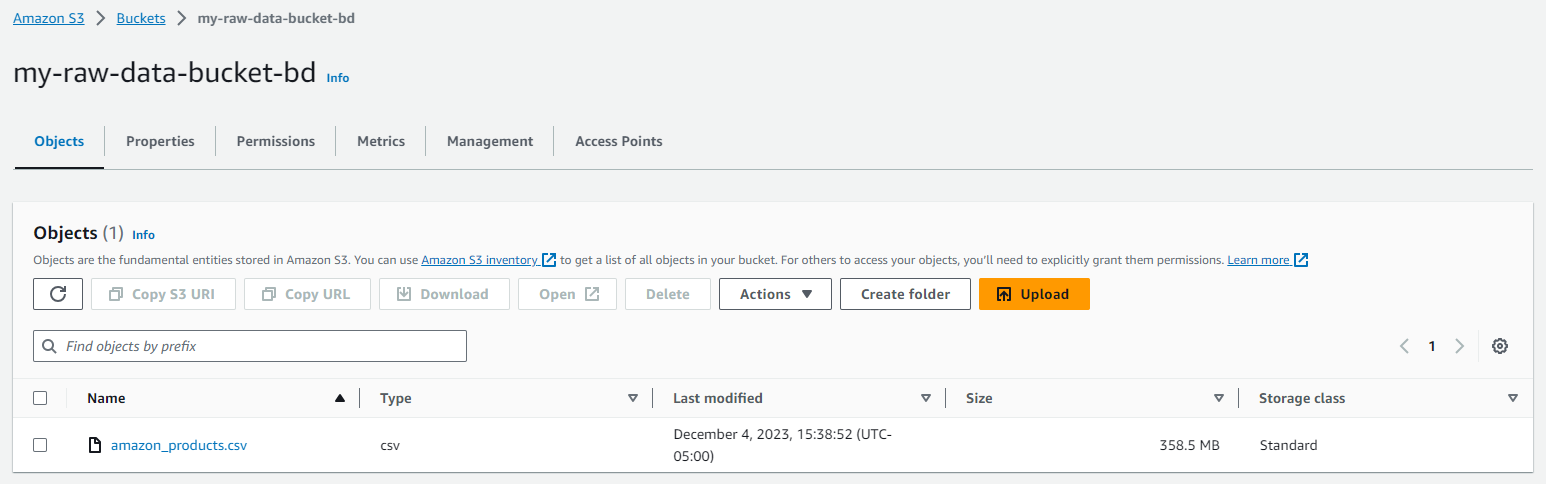
* STORAGE: S3 BUCKET
* DATA PIPELINE: GLUE ETL JOBS
* QUERIES: ATHENA

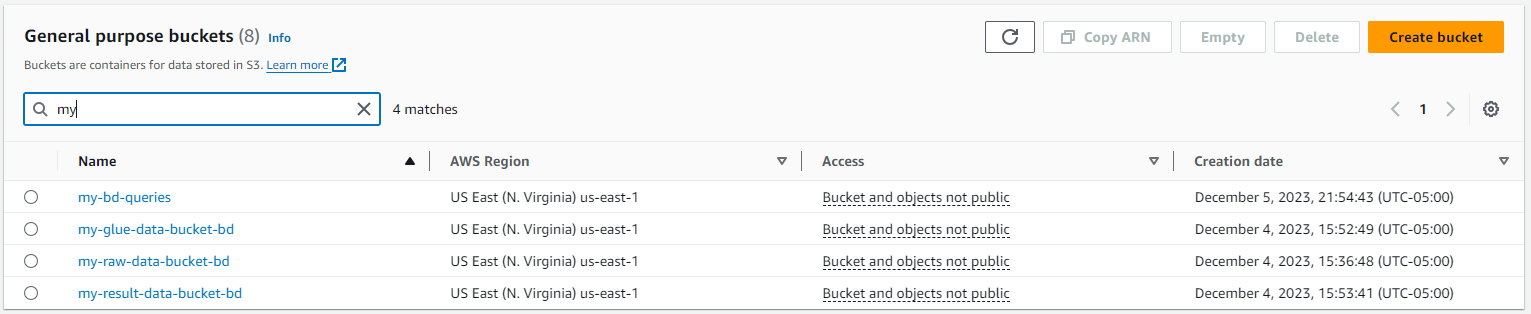
For convenience, we choose AWS Glue to design ETL data pipeline. AWS Glue is a fully managed extract, transform, and load (ETL) service that uses Apache Spark under the hood. When you create and run an AWS Glue ETL job, it automatically runs a Spark application in a managed Spark environment.

## 4. AWS Configuration

### (1) S3 Bucket

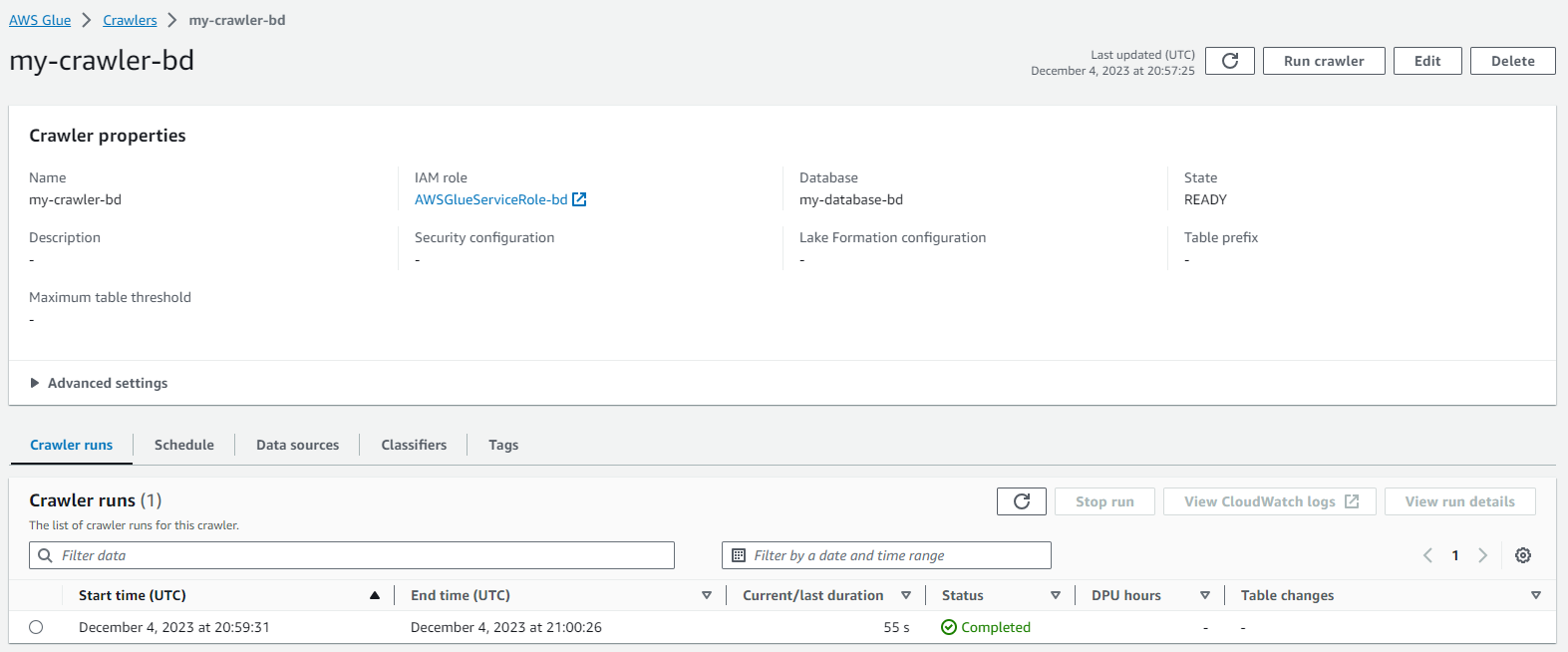
Created S3 buckets using the AWS Management Console. S3 is used as a storage solution for our data. Uploading the the .csv file to S3 make the data accessible for processing by various AWS services



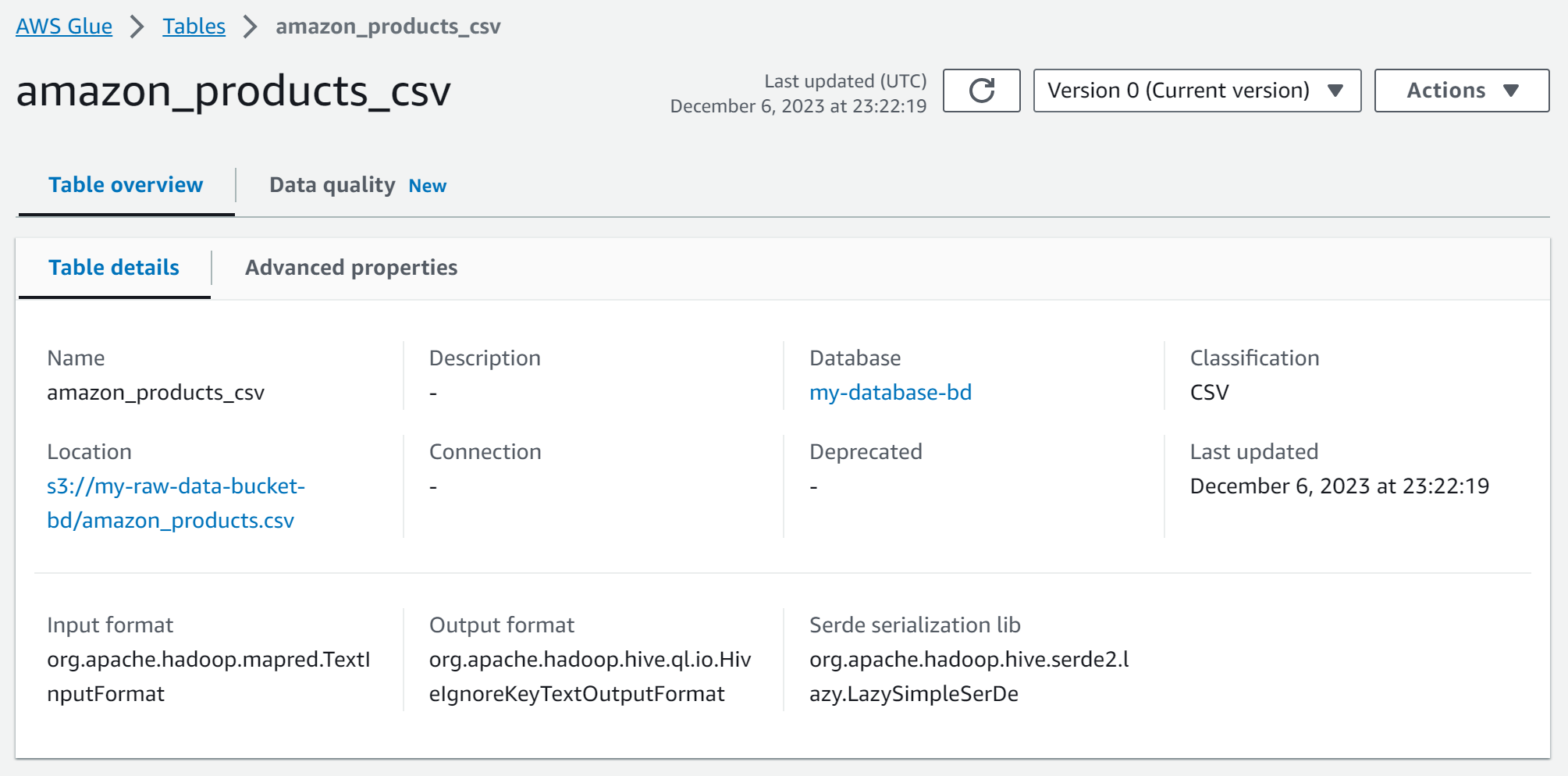


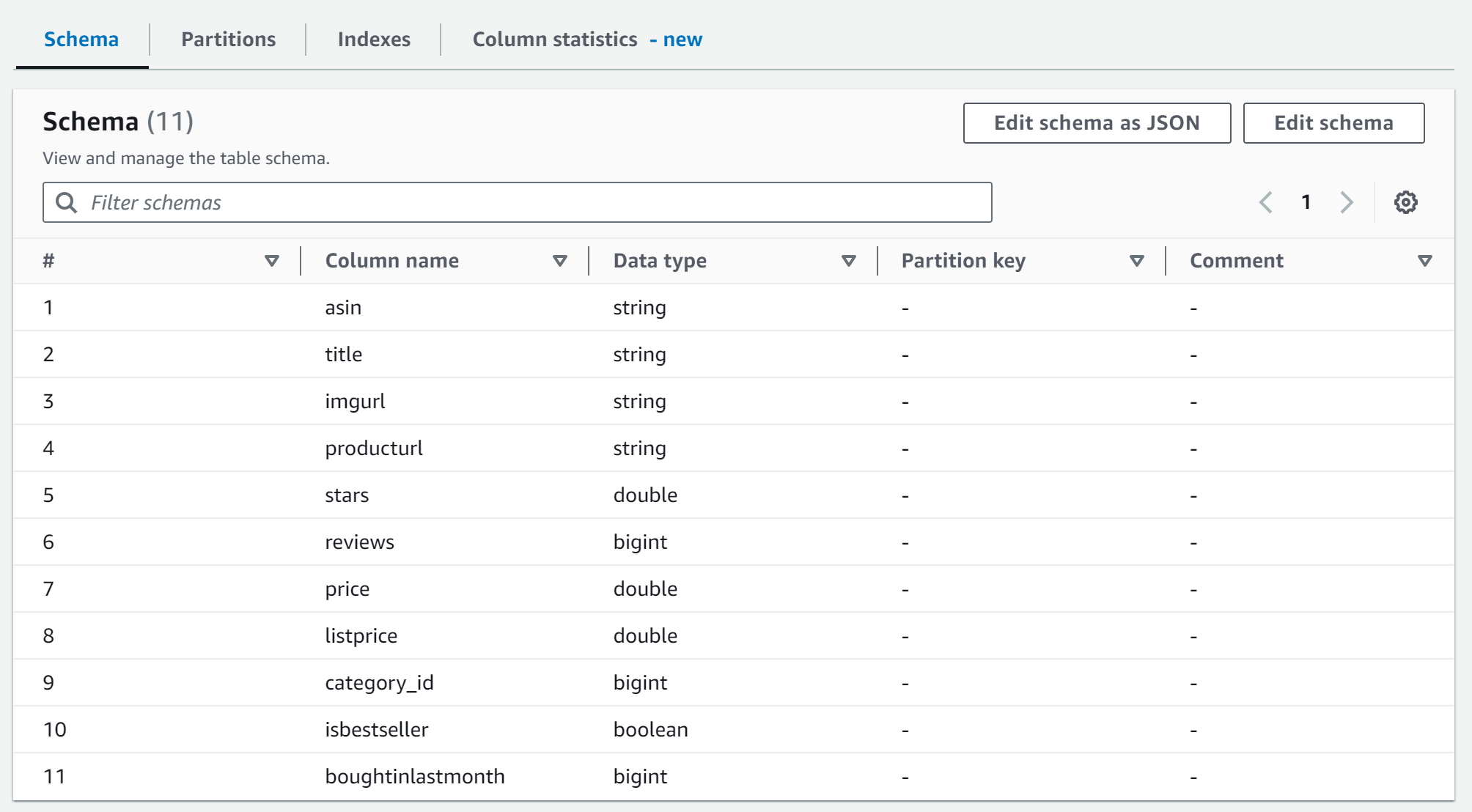
### (2) Glue Crawler

Create a Glue Crawler using AWS Glue Console. Specify the S3 bucket and path where the data is stored. Run the Glue Crawler to discover the schema and metadata of the data. The Glue Crawler is used to automatically infer the schema and metadata of the data. This is important for the subsequent steps where you’ll be processing and transforming the data



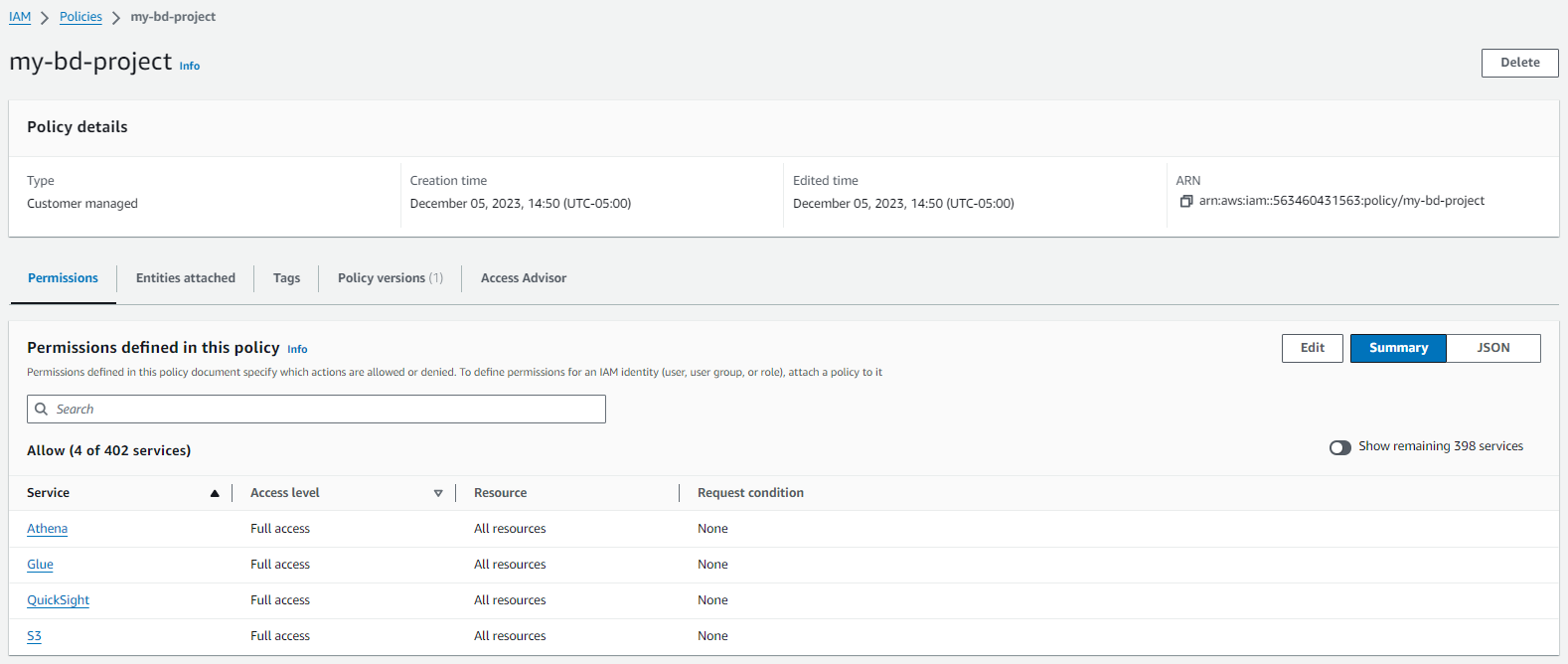
Running an AWS Glue Crawler automatically creates or updates tables in the Glue Data Catalog, streamlining the process of managing metadata and making your data easily accessible for analysis and transformation.

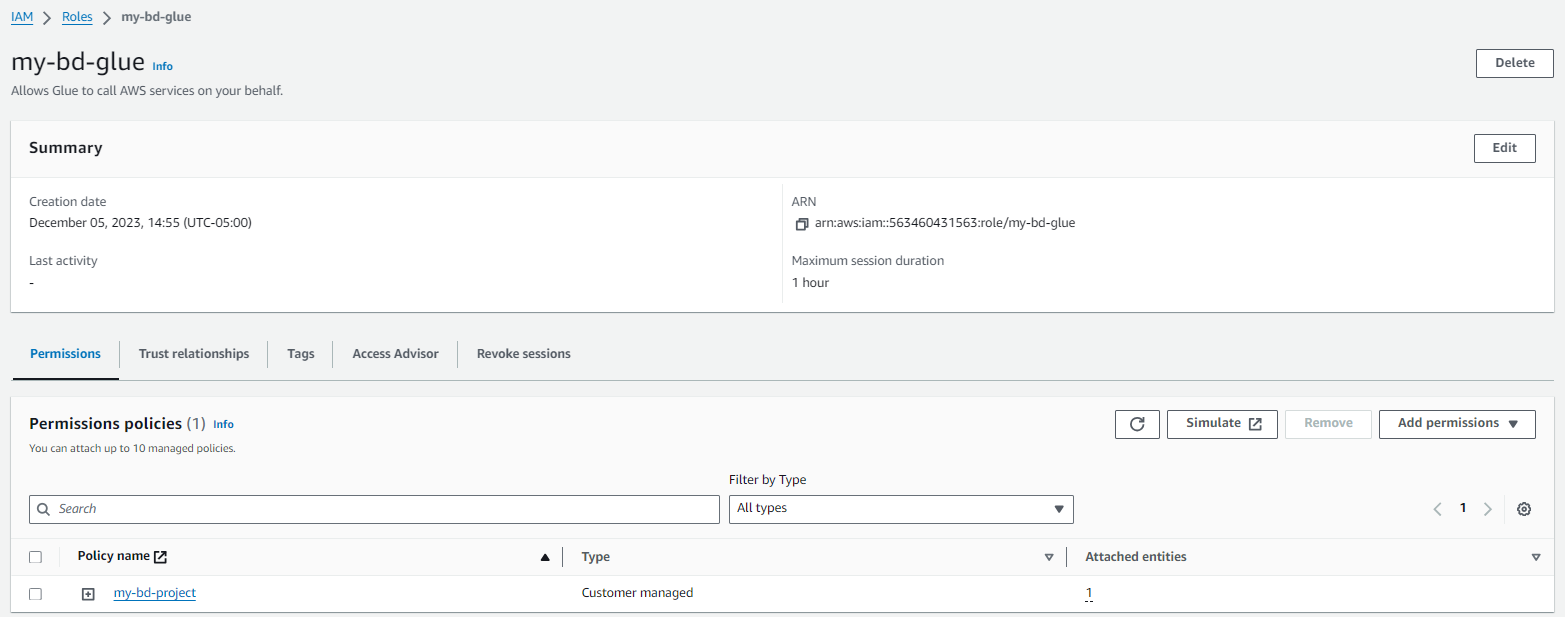




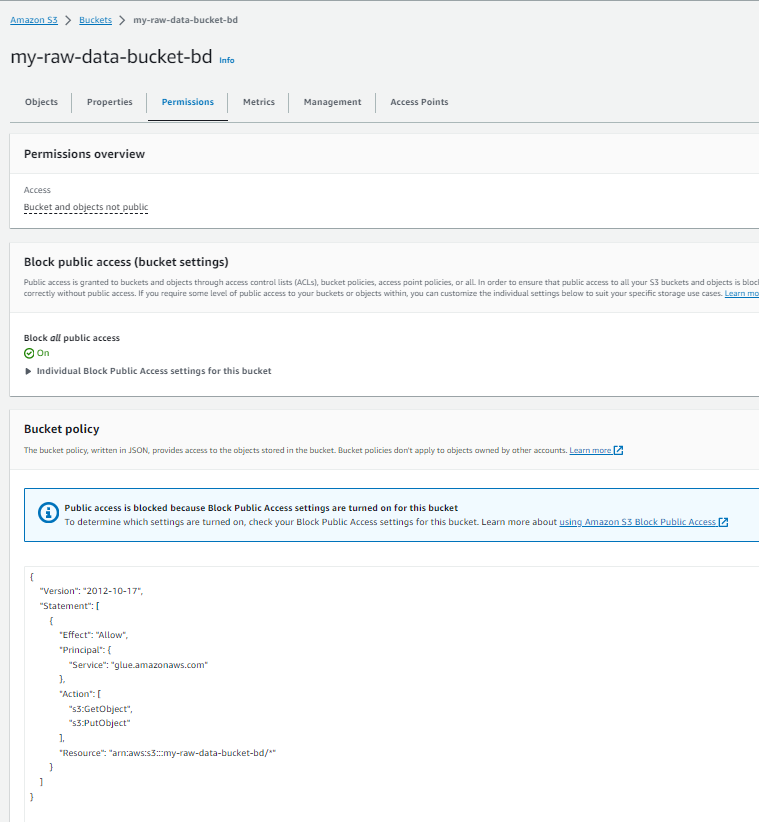
### (3) IAM role and S3 permission

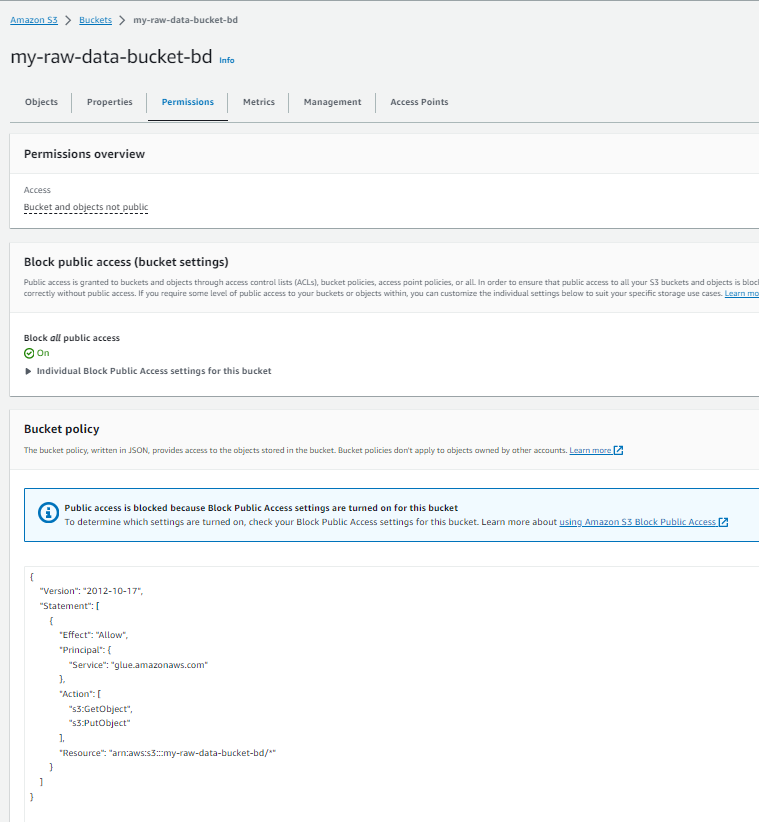
Create an IAM role with the necessary permissions for your Glue ETL job to access S3 and other AWS services. The IAM role ensures that your Glue ETL job has the necessary permissions to interact with other AWS services.





Ensure that the S3 bucket has the correct permissions, allowing the IAM role to read and write data. Proper S3 bucket permissions are crucial for reading and writing data during the ETL process.





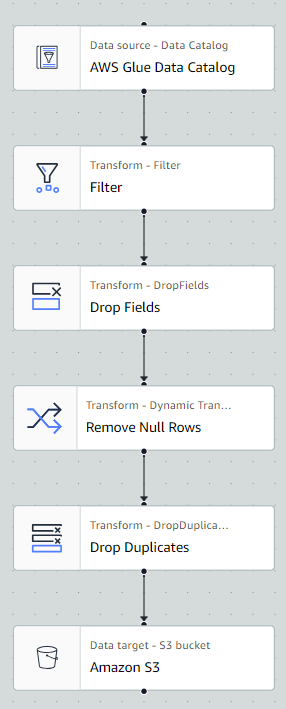
## 5. ETL data pipeline

Design Visual Glue ETL job:

* Use the AWS Glue Console to create a new ETL job.
* Define the source (Data Catalog table from glue crawler), target (S3), and transformations needed using a visual interface or Glue Python/Spark scripts.
* Configure the job to output the data in .csv format.

The Glue ETL job is designed to transform and move data from source to target. The visual interface simplifies the ETL process, and you can specify the output format as .csv for your use case.

### (1) Visual

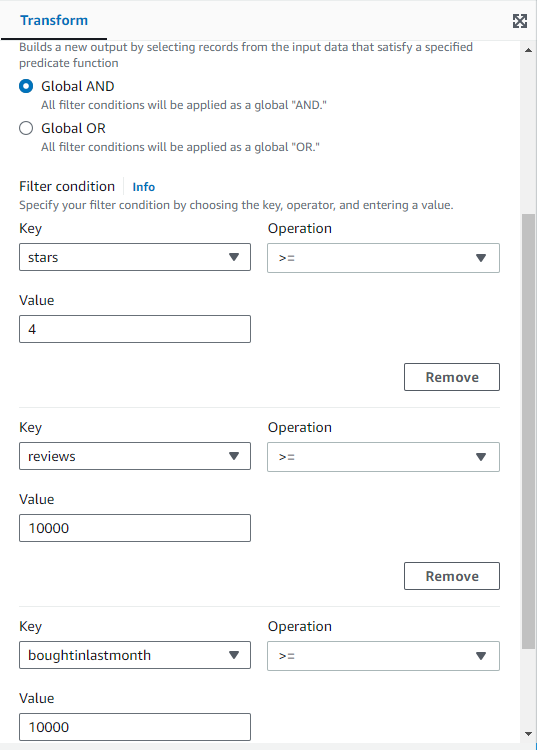


### (2) Configuration

A. Data Source: Data Catalog - Bring the table that is created by Glue crawler and saved in the data catalog.

B. Transform in 4 steps:

Step#1 - Filter: Extract data only, Rating: Stars >= 4 & Qty of Review: Review >=10,000 & Qty of Sales: Boughtinlastmonth >=10,000

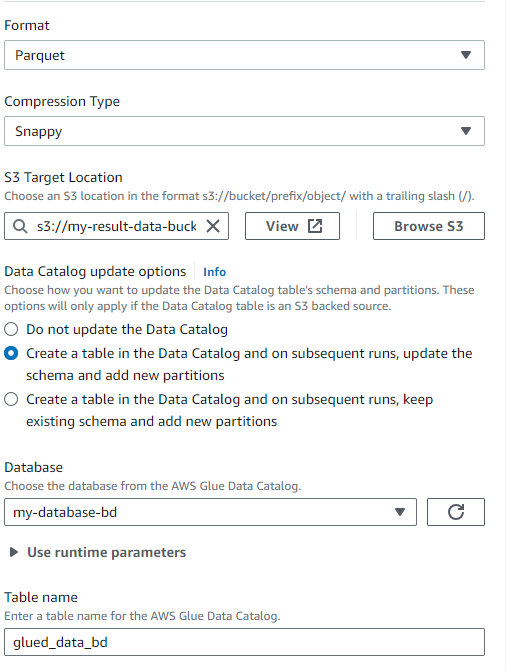


Step#2 - Drop Fields: Extract relevant column only for efficient data analysis. Extracted column: Price, Reviews, starts, isbestseller, title, boughtinlastmonth.

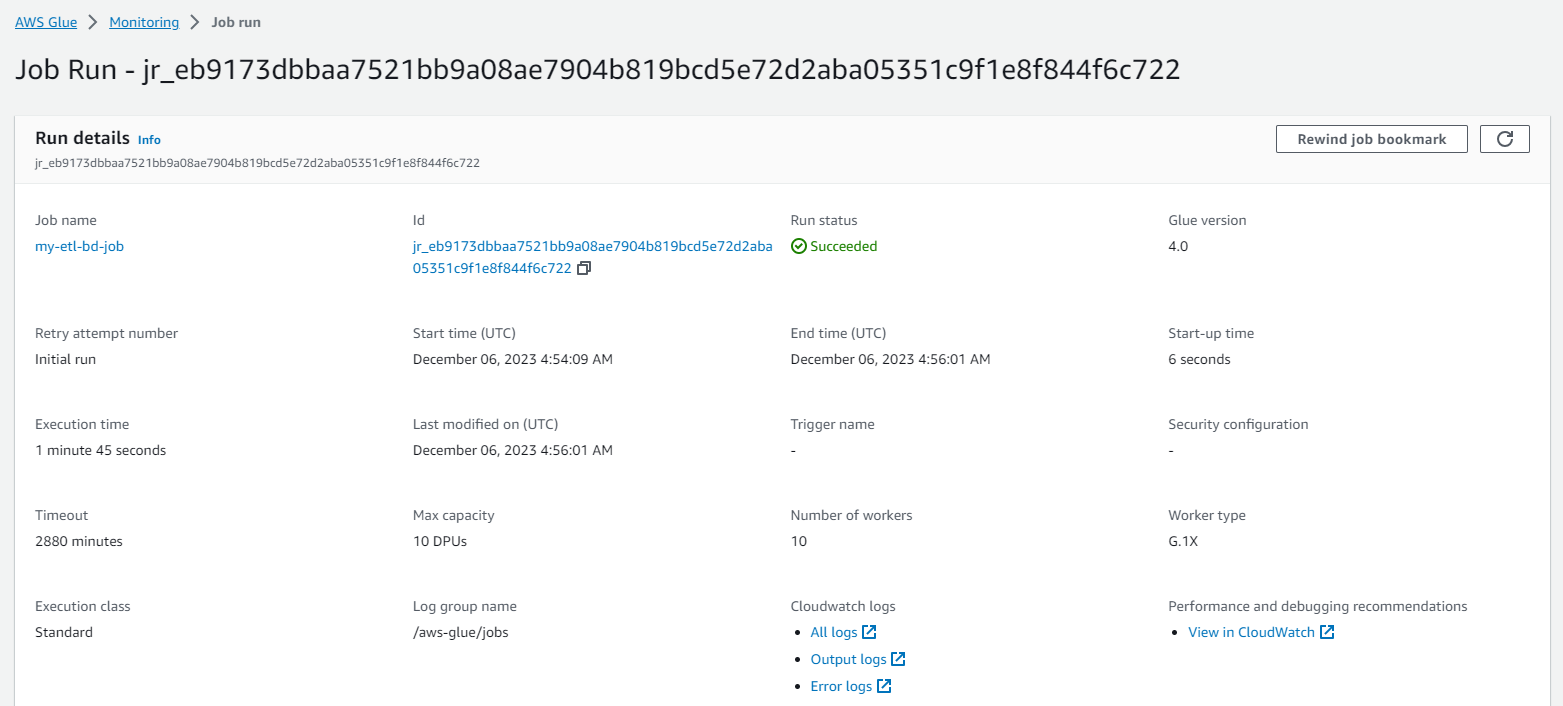
Step#3 - Drop Duplicates: Eliminate duplicate rows

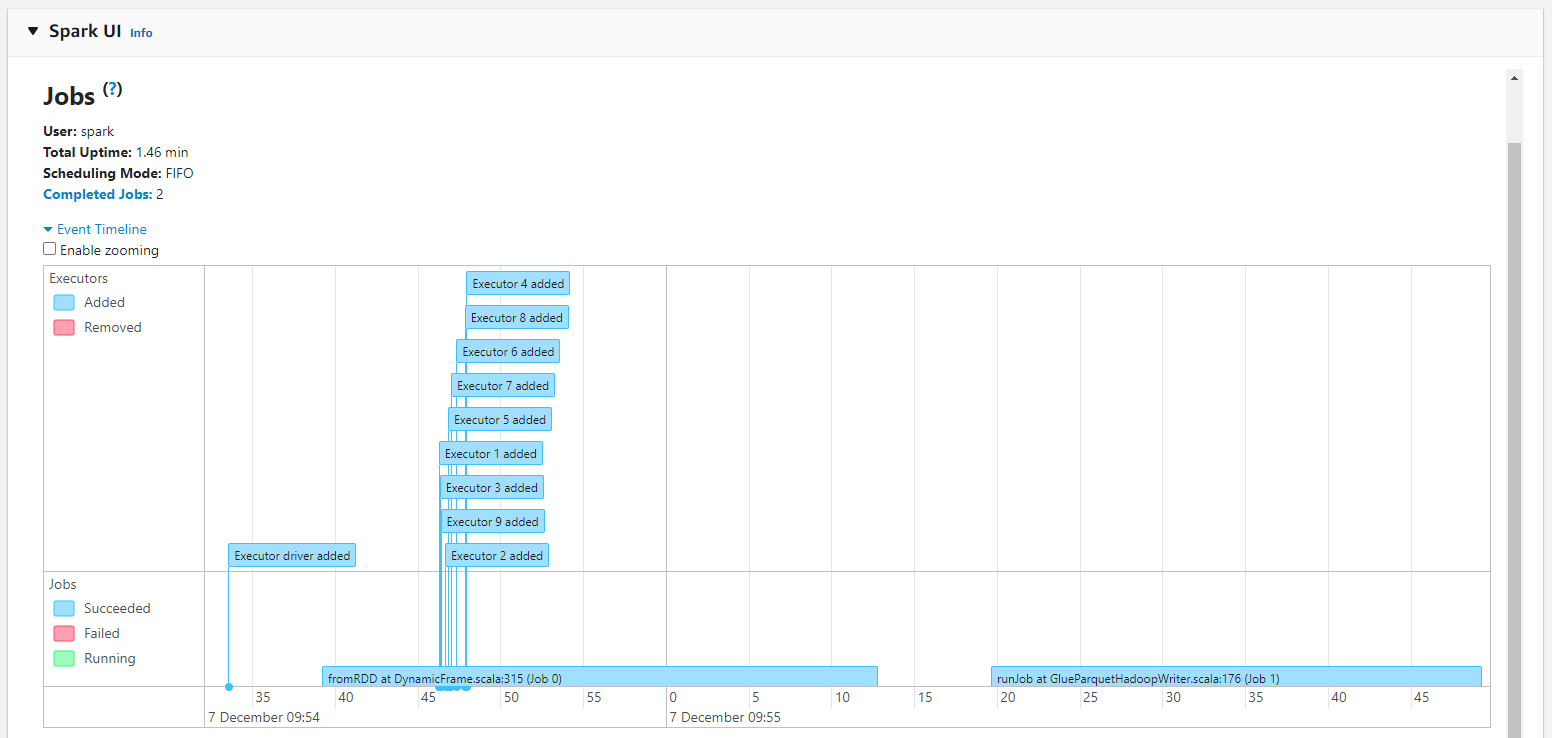
Step#4 - Drop Null Fields:Remove null value

C. Target: S3 Bucket - Set up ‘data catalog update options’ as ‘Create a table in the Data Catalog’ for query on Athena



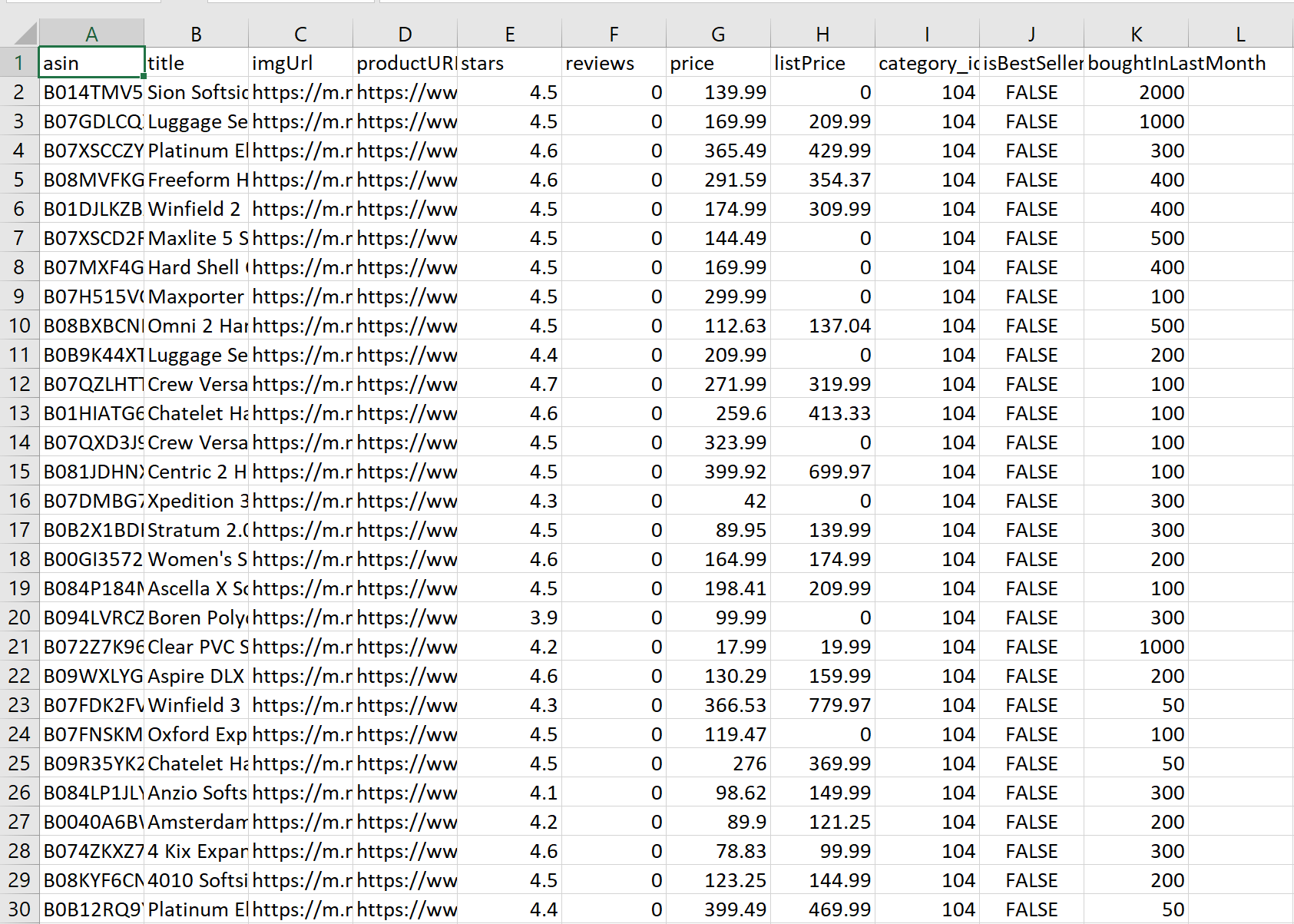
### (3) Result



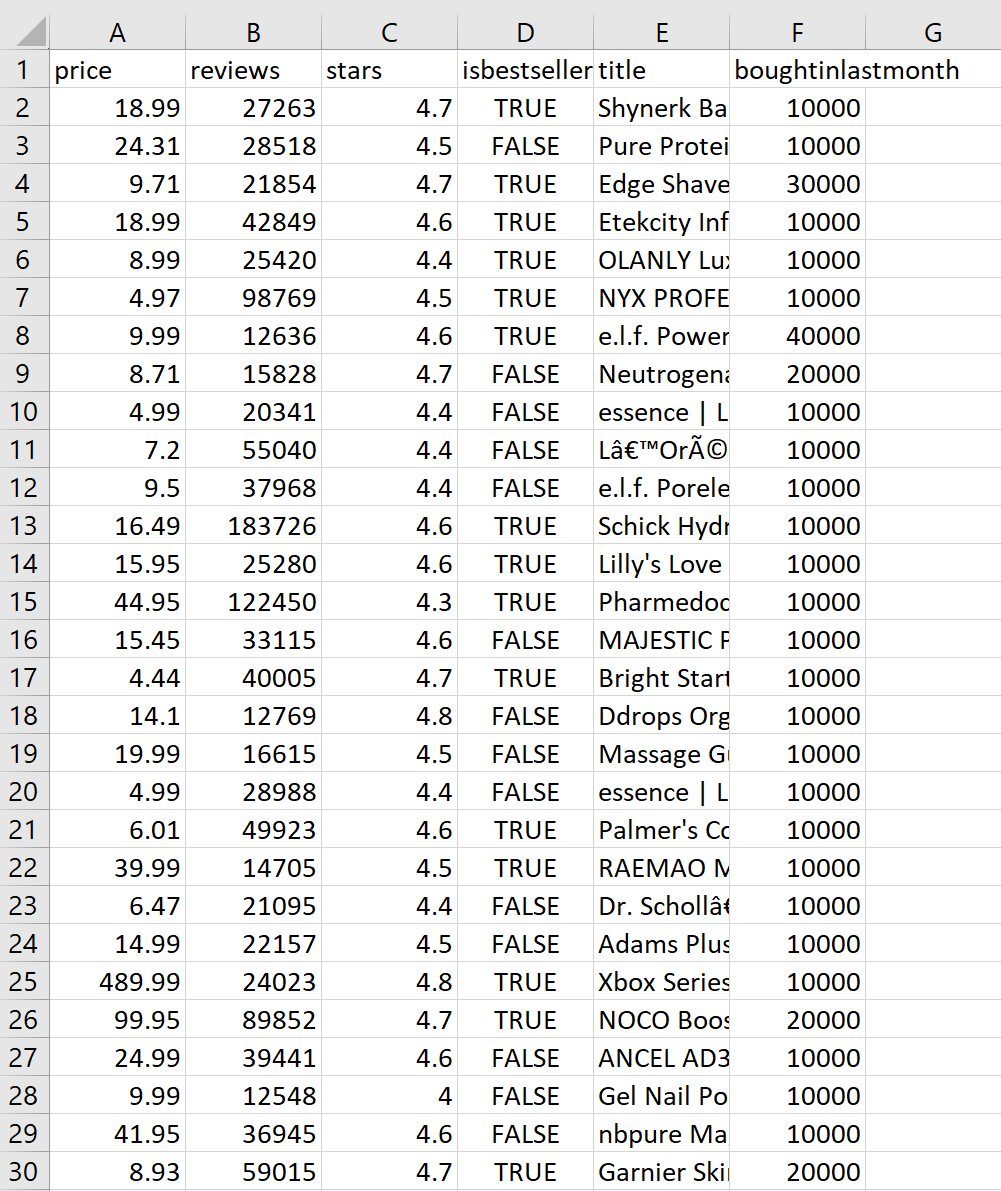


### (3) Result in Excel view

Raw data, the size of the .csv file is 358GB.



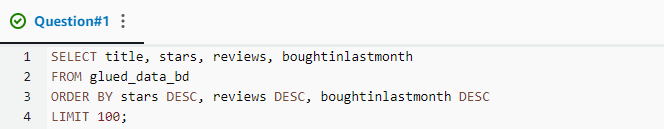
After data preprocessing, the size of the .csv file decrease to 59KB.



## 6. Queries on Athena

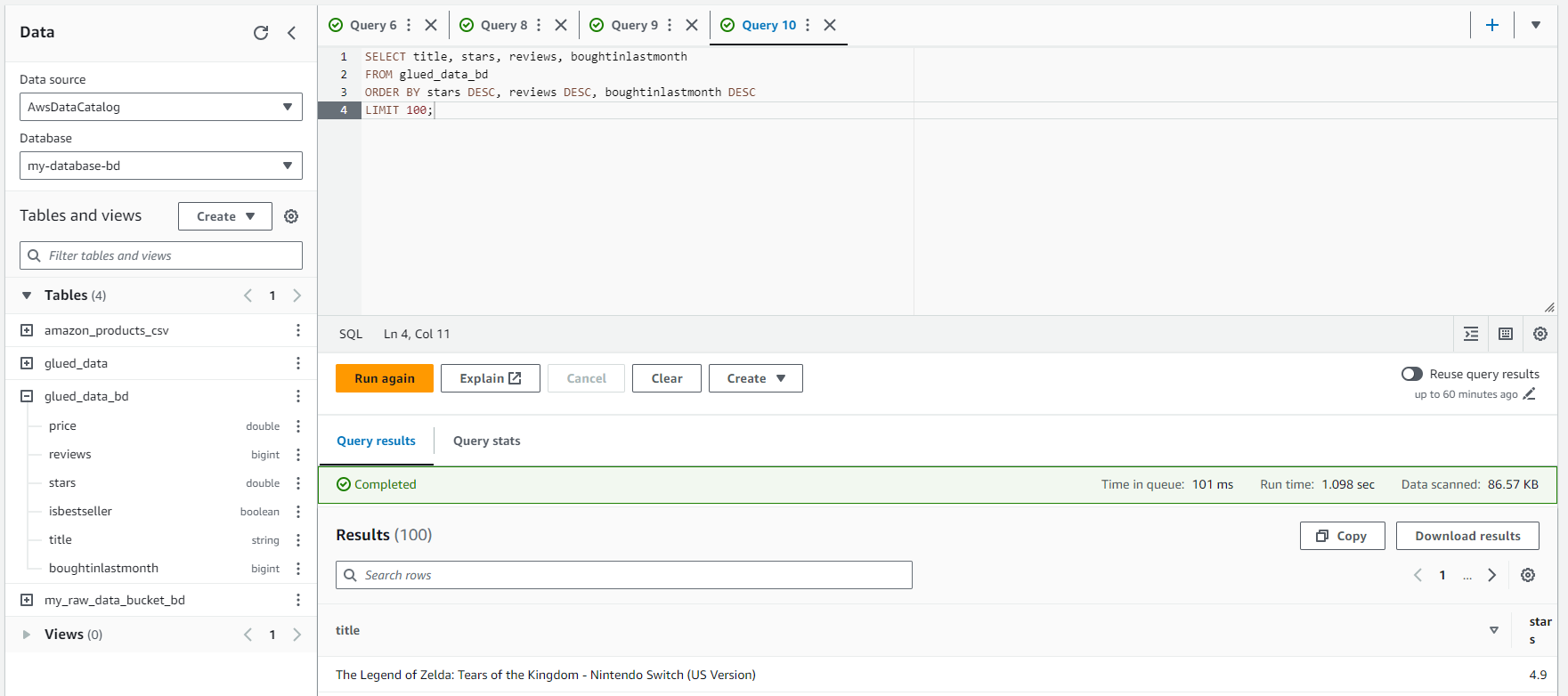
### (1) Queries

To retrieve the top 100 most highly rated products based on quantity of actual reviews and sales volume, write query selects all columns from the preprocessed data and orders the results based on the rating points ﬁrst in descending order, then by qty of reviews and sales both in descending order as well.



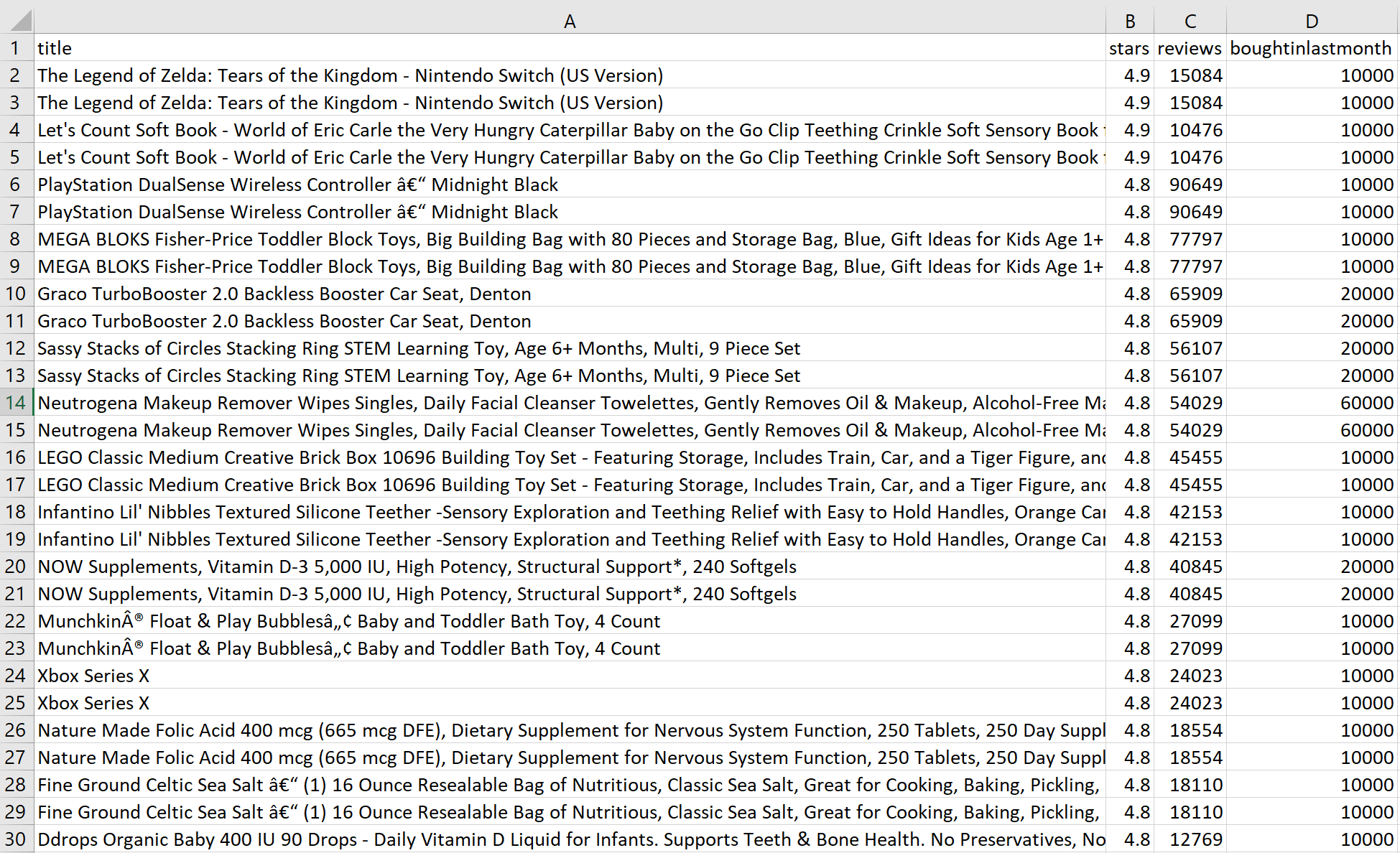
### (2) Result

The most highly rated product based on reviews and sales in 2023 was ‘The Legend of Zelda: Tears of the Kingdom - Nintendo Switch (US Version)’



### (2) Result in excel view

The top 30 top-rated product lists on Amazon E-commerce market in 2023.



## 7. Conclusion

**(1) Top-rated Products**

* Our analysis has revealed a list of top-rated products based on customer ratings.
* These ﬁndings are crucial for businesses aiming to understand customer preferences and make data-driven decisions to enhance their product offerings.

**(2) Key Insights**

* In conclusion, the analysis has provided key insights into customer preferences and top-rated products.
* Businesses can use this information to tailor their strategies, improve product offerings, and ultimately enhance customer satisfaction.

**(3) Business Significance**

* The ﬁndings have signiﬁcant implications for businesses, allowing them to align their products more closely with customer expectations.
* This alignment is crucial for staying competitive in the market and fostering long-term customer loyalty.

## 8. Challenges and Future Analysis Ideas

**(1) Challenges**

* Problem#1: After Glue ETL job, we need to load table Metadata since AWS automatically saves the table as a set of schema and metadata separately.

⇒ Solution: The command below discovers the partitions in data and updates Athena metadata accordingly.

MSCK REPAIR TABLE your\_table\_name;

* Problem#2: Wanted to visualize more through Quicksight (dashboard visualization tool). But, it is $24/month. So, we used only Athena.

⇒ Solution: Must have sufficient funds to afford it.

**(2) Feature Advanced Analysis Idea**

* Classify the categories of top rated product and visualize with pie chart to get some Amazon e-commerce market insight in 2023

## [Resources]

<https://github.com/Duggieeeey/AWSglue-ETL>