



ANALYZING CUSTOMER RATINGS TO IDENTIFY TOP-RATED PRODUCTS (Amazon E-commerce Market 2023)

[DTSC 701/CSCI 626: Intro To Big Data]

RAMIS RAWNAK
GAHYEON BACK
NAFIAH ALAM
RAVI KIRAN REDDY PULIKANDAM

INDEX

1. **Data: Amazon Product Dataset 2023 | 358MB | Kaggle**
2. **Goal: Extract top-rated product from Amazon E-commerce market in 2023**
3. **Technical Approach for big data analytics**
4. **AWS Configuration - S3 Bucket, Glue Crawler and IAM role**
5. **ETL data pipeline - Amazon Glue Visual ETL job**
6. **Queries on Athena**
7. **Conclusion**
8. **Challenges and future solution**

1. DATA USED IN PROJECT



Amazon Product Dataset 2023

Size: 358 GB | Format: .csv | Source: Kaggle

	A	B	C	D	E	F	G	H	I	J	K	L
1	asin	title	imgUrl	productUrl	stars	reviews	price	listPrice	category_id	isBestSeller	boughtInLastMonth	
2	B014TMV5	Sion Softsic	https://m.r	https://ww	4.5	0	139.99	0	104	FALSE	2000	
3	B07GDLQ	Luggage Se	https://m.r	https://ww	4.5	0	169.99	209.99	104	FALSE	1000	
4	B07XSCCY	Platinum El	https://m.r	https://ww	4.6	0	365.49	429.99	104	FALSE	300	
5	B08MVFKG	Freeform H	https://m.r	https://ww	4.6	0	291.59	354.37	104	FALSE	400	
6	B01DJLKZB	Winfield 2	https://m.r	https://ww	4.5	0	174.99	309.99	104	FALSE	400	
7	B07XSCD2F	Maxlite 5 S	https://m.r	https://ww	4.5	0	144.49	0	104	FALSE	500	
8	B07MXF4G	Hard Shell	https://m.r	https://ww	4.5	0	169.99	0	104	FALSE	400	
9	B07H515V	Maxporter	https://m.r	https://ww	4.5	0	299.99	0	104	FALSE	100	
10	B08BXCBCN	Omni 2 Ha	https://m.r	https://ww	4.5	0	112.63	137.04	104	FALSE	500	

1. DATA - (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.

1. DATA - (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.csv 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

GOAL OF THE PROJECT:

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

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.

Why did we choose AWS as our distributed cloud service?

- **Widely Used**: AWS is a popular and widely adopted cloud service with extensive documentation and community support.
- **Comprehensive Services**: AWS provides a variety of services suitable for data storage, processing, and analysis, including S3 for data lake storage, and EMR for running Spark applications.

3. TECHNICAL APPROACH (2)

WHAT SERVICES DID WE CHOOSE?

STORAGE: S3 BUCKET



DATA PIPELINE: GLUE ETL JOBS



QUERIES: ATHENA

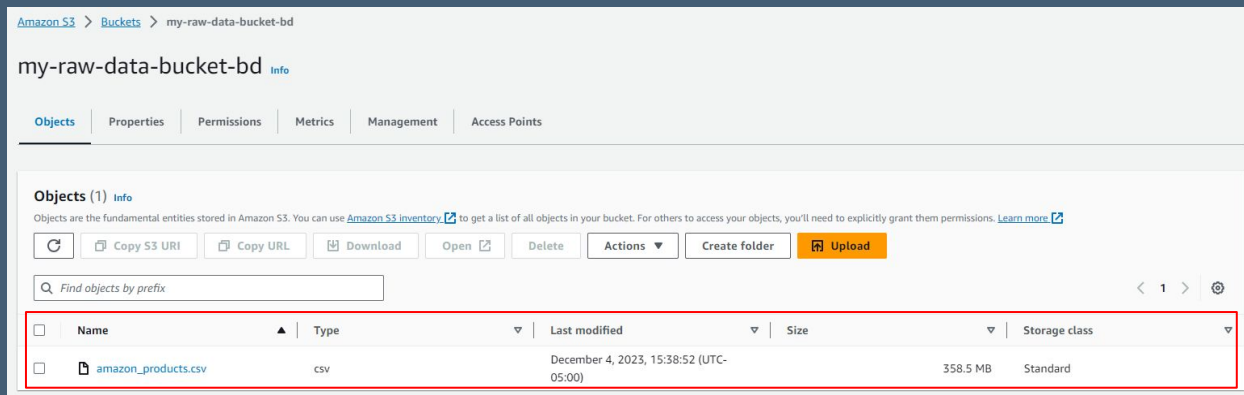


For convenience reason, we choose AWS Glue to design ETL data pipeline. AWS Glue is a fully managed extract, transform, and load (ETL) service that uses Apache Spark under the hood. When you create and run an AWS Glue ETL job, it automatically runs a Spark application in a managed Spark environment.

4. AWS CONFIGURATION (1) S3

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.

S3 Bucket Configuration for Input data



Amazon S3 > Buckets > my-raw-data-bucket-bd

my-raw-data-bucket-bd [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

Objects (1) 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](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	amazon_products.csv	csv	December 4, 2023, 15:38:52 (UTC-05:00)	358.5 MB	Standard

4. AWS Configuration - (2-1) 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.

Glue Crawler Configuration after run

The screenshot displays the AWS Glue console interface for a crawler named 'my-crawler-bd'. The breadcrumb navigation at the top shows 'AWS Glue > Crawlers > my-crawler-bd'. The crawler's status is 'READY', last updated on December 4, 2023, at 20:57:25 UTC. Action buttons include 'Run crawler', 'Edit', and 'Delete'.

Crawler properties

Name my-crawler-bd	IAM role AWSGlueServiceRole-bd	Database my-database-bd	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

► Advanced settings

Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (1)
The list of crawler runs for this crawler.

Filter data: Filter by a date and time range:

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
December 4, 2023 at 20:59:31	December 4, 2023 at 21:00:26	55 s	Completed	-	-

4. AWS CONFIGURATION- (2-2) CRAWLER

After running Crawler, AWS creates table on data catalog of Glue

AWS Glue > Tables > amazon_products_csv

amazon_products_csv

Last updated (UTC)
December 6, 2023 at 23:22:19

Version 0 (Current version) Actions

Table overview Data quality New

Table details Advanced properties

Name amazon_products_csv	Description -	Database my-database-bd	Classification CSV
Location s3://my-raw-data-bucket-bd/amazon_products.csv	Connection -	Deprecated -	Last updated December 6, 2023 at 23:22:19
Input format org.apache.hadoop.mapred.TextInputFormat	Output format org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat	Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	

Schema Partitions Indexes Column statistics - new

Schema (11)

Edit schema as JSON Edit schema

View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key	Comment
1	asin	string	-	-
2	title	string	-	-
3	imgurl	string	-	-
4	producturl	string	-	-
5	stars	double	-	-
6	reviews	bigint	-	-
7	price	double	-	-
8	listprice	double	-	-
9	category_id	bigint	-	-
10	isbestseller	boolean	-	-
11	boughtinlastmonth	bigint	-	-

4. AWS Configuration - (3-1) IAM role

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

Configuration of IAM policy

The screenshot shows the AWS IAM console for the 'my-bd-project' policy. The 'Policy details' section shows it is a 'Customer managed' policy created on December 05, 2023, at 14:50 UTC-05:00. The 'Permissions defined in this policy' section shows a table of permissions for various AWS services, all with 'Full access' and 'All resources'.

Service	Access level	Resource	Request condition
AdSense	Full access	All resources	None
Glue	Full access	All resources	None
QuickSight	Full access	All resources	None
S3	Full access	All resources	None

Configuration of IAM role

The screenshot shows the AWS IAM console for the 'my-bd-glue' role. The 'Summary' section shows it is a role created on December 05, 2023, at 14:55 UTC-05:00. The 'Permissions policies' section shows a table of attached policies, including 'my-bd-project'.

Policy name	Type	Attached entities
my-bd-project	Customer managed	1


4. AWS Config. - (3-2) S3 permission

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.

Configuration of S3 permission

Bucket policy

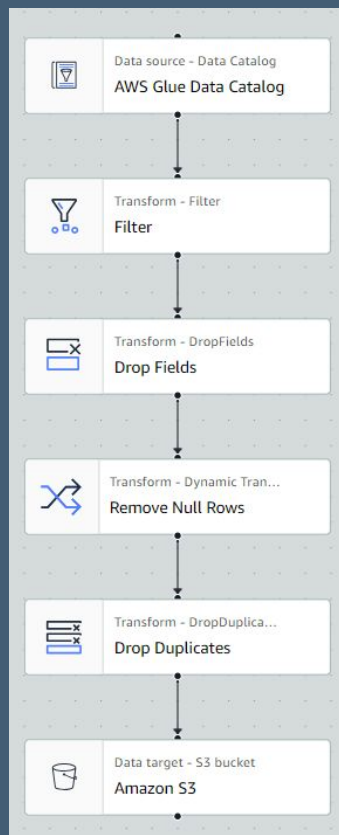
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. [Learn more](#)

 **Public access is blocked because Block Public Access settings are turned on for this bucket**
To determine which settings are turned on, check your Block Public Access settings for this bucket. Learn more about [using Amazon S3 Block Public Access](#)

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": "arn:aws:s3::my-raw-data-bucket-bd/*"
    }
  ]
}
```



5. ETL data pipeline - (1) Visual



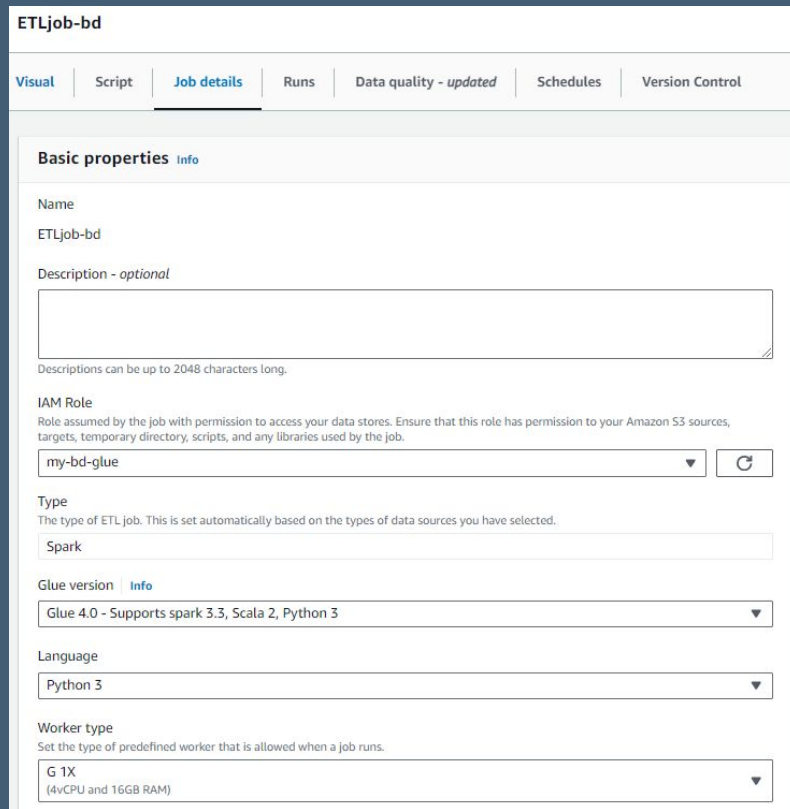
How it works?

- **Data Source:** Data Catalog | Bring the table that is created by Glue crawler and saved in data catalog.
- **Transform in 4 steps:**
Step#1 & 2 will be explained in details following slides.
Step#3 - Drop Duplicates: Eliminate duplicate rows
Step#4 - Drop Null Fields: Remove null value
- **Target:** S3 Bucket | Set up 'data catalog update options' as 'Create a table in the Data Catalog' for query on Athena

5. ETL data pipeline - (2-1) Config.

Before start, set up Glue ETL job detail:

- As mentioned, we utilized AWS Glue to design the 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.



The screenshot displays the AWS Glue console interface for configuring an ETL job. The job is titled 'ETLjob-bd'. The 'Job details' tab is selected, showing the 'Basic properties' section. The configuration includes:

- Name:** ETLjob-bd
- Description:** (Optional, empty text box)
- IAM Role:** my-bd-glue
- Type:** Spark
- Glue version:** Glue 4.0 - Supports spark 3.3, Scala 2, Python 3
- Language:** Python 3
- Worker type:** G 1X (4vCPU and 16GB RAM)

The interface also shows tabs for 'Visual', 'Script', 'Runs', 'Data quality - updated', 'Schedules', and 'Version Control'.

5. ETL data pipeline - (2-2) Config.

Transform Step#1 - Filter:

Extract data only,

Rating: Stars ≥ 4

Qty of Review: Review $\geq 10,000$

Qty of Sales:

Boughtinlastmonth $\geq 10,000$

Transform Step#2 - Drop Fields:

Extract relevant column only for efficient data analysis.

Extracted column: Price, Reviews, starts, isbestseller, title, boughtinlastmonth

The screenshot shows the 'Transform' configuration window for the 'Filter' step. It includes a 'Filter condition' section with 'Global AND' selected. Below this, there are three filter conditions defined: 'stars' is greater than or equal to 4, 'reviews' is greater than or equal to 10000, and 'boughtinlastmonth' is greater than or equal to 10000. Each condition has a 'Remove' button next to it.

Transform

Builds a new output by selecting records from the input data that satisfy a specified predicate function

☒ Global AND
All filter conditions will be applied as a global "AND."

☐ Global OR
All filter conditions will be applied as a global "OR."

Filter condition **Info**
Specify your filter condition by choosing the key, operator, and entering a value.

Key: stars Operation: \geq Value: 4 [Remove]

Key: reviews Operation: \geq Value: 10000 [Remove]

Key: boughtinlastmonth Operation: \geq Value: 10000 [Remove]

[↑Filter]

The screenshot shows the 'Transform' configuration window for the 'Drop Fields' step. It includes a 'Name' field set to 'Drop Fields', a 'Node parents' dropdown, and a 'Filter' button. Below this is a table titled 'DropFields' with columns 'Field' and 'Data type'. The table lists various fields with checkboxes to select which ones to drop. The fields 'category_id', 'imgurl', 'listprice', 'producturl', and 'asin' are selected (checked), while 'price', 'reviews', 'stars', 'isbestseller', 'title', and 'boughtinlastmonth' are not selected.

Transform

Name: Drop Fields

Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

Filter
Filter - Transform

DropFields

	Field	Data type
<input type="checkbox"/>	price	double
<input type="checkbox"/>	reviews	long
<input type="checkbox"/>	stars	double
<input checked="" type="checkbox"/>	category_id	long
<input checked="" type="checkbox"/>	imgurl	string
<input checked="" type="checkbox"/>	listprice	double
<input type="checkbox"/>	isbestseller	boolean
<input checked="" type="checkbox"/>	producturl	string
<input checked="" type="checkbox"/>	asin	string
<input type="checkbox"/>	title	string
<input type="checkbox"/>	boughtinlastmonth	long

[↑Drop Fields]

5. ETL data pipeline - (3) ETL Result

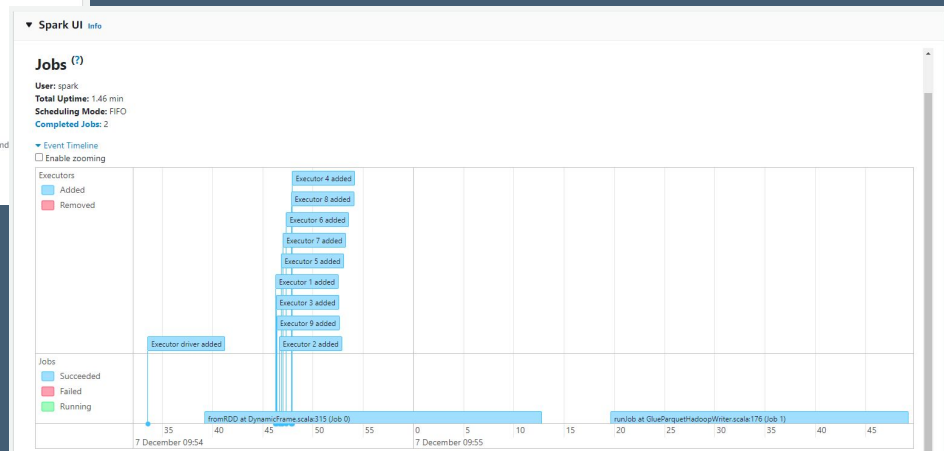
AWS Glue > Monitoring > Job run

Job Run - jr_eb9173dbbaa7521bb9a08ae7904b819bcd5e72d2aba05351c9f1e8f844f6c722

Run details info Rewind job bookmark

jr_eb9173dbbaa7521bb9a08ae7904b819bcd5e72d2aba05351c9f1e8f844f6c722

Job name	Id	Run status	Glue version
my-etl-bd-job	jr_eb9173dbbaa7521bb9a08ae7904b819bcd5e72d2aba05351c9f1e8f844f6c722	Succeeded	4.0
Retry attempt number	Start time (UTC)	End time (UTC)	Start-up time
Initial run	December 06, 2023 4:54:09 AM	December 06, 2023 4:56:01 AM	6 seconds
Execution time	Last modified on (UTC)	Trigger name	Security configuration
1 minute 45 seconds	December 06, 2023 4:56:01 AM	-	-
Timeout	Max capacity	Number of workers	Worker type
2880 minutes	10 DPU's	10	G.1X
Execution class	Log group name	Cloudwatch logs	Performance and debugging recommend
Standard	/aws-glue/jobs	<ul style="list-style-type: none">All logsOutput logsError logs	<ul style="list-style-type: none">View in CloudWatch



[↑Spark UI]

5. ETL data pipeline - (4) ETL Result

Before and After in excel view

—Before (size 358MB)—

	A	B	C	D	E	F	G	H	I	J	K	L
1	asin	title	imgUrl	productUrl	stars	reviews	price	listPrice	category_ic	isBestSeller	boughtInLastMonth	
2	B014TMV5	Sion Softsich	https://m.r	https://ww	4.5	0	139.99	0	104	FALSE	2000	
3	B07GDLCC	Luggage Se	https://m.r	https://ww	4.5	0	169.99	209.99	104	FALSE	1000	
4	B07XSCCZ	Platinum El	https://m.r	https://ww	4.6	0	365.49	429.99	104	FALSE	300	
5	B08MVFKG	Freeform H	https://m.r	https://ww	4.6	0	291.59	354.37	104	FALSE	400	
6	B01DJLKZ	B Winfield 2	https://m.r	https://ww	4.5	0	174.99	309.99	104	FALSE	400	
7	B07XSCD2	F Maxlite 5	https://m.r	https://ww	4.5	0	144.49	0	104	FALSE	500	
8	B07MXF4G	Hard Shell	https://m.r	https://ww	4.5	0	169.99	0	104	FALSE	400	
9	B07H515V	Maxporter	https://m.r	https://ww	4.5	0	299.99	0	104	FALSE	100	
10	B08BXBCN	I Omni 2 Har	https://m.r	https://ww	4.5	0	112.63	137.04	104	FALSE	500	
11	B089K44X	Luggage Se	https://m.r	https://ww	4.4	0	209.99	0	104	FALSE	200	
12	B07QZLHT	I Crew Versa	https://m.r	https://ww	4.7	0	271.99	319.99	104	FALSE	100	
13	B01HIATG	E Chatelet H	https://m.r	https://ww	4.6	0	259.6	413.33	104	FALSE	100	
14	B07QXDJ	E Crew Versa	https://m.r	https://ww	4.5	0	323.99	0	104	FALSE	100	
15	B081JDHN	I Centric 2	https://m.r	https://ww	4.5	0	399.92	699.97	104	FALSE	100	
16	B07DMBG	I Xpedition 3	https://m.r	https://ww	4.3	0	42	0	104	FALSE	300	
17	B08X1BDI	I Stratum 2	https://m.r	https://ww	4.5	0	89.95	139.99	104	FALSE	300	
18	B00GI357	I Women's S	https://m.r	https://ww	4.6	0	164.99	174.99	104	FALSE	200	
19	B084P184	I Ascella X	https://m.r	https://ww	4.5	0	198.41	209.99	104	FALSE	100	
20	B094LVR	C Boren Poly	https://m.r	https://ww	3.9	0	99.99	0	104	FALSE	300	
21	B072Z7K	9 Clear PVC	https://m.r	https://ww	4.2	0	17.99	19.99	104	FALSE	1000	
22	B09WXYG	I Aspire DLX	https://m.r	https://ww	4.6	0	130.29	159.99	104	FALSE	200	
23	B07FDK2	FV Winfield 3	https://m.r	https://ww	4.3	0	366.53	779.97	104	FALSE	50	
24	B07FNSKM	I Oxford Exp	https://m.r	https://ww	4.5	0	119.47	0	104	FALSE	100	
25	B09R35Y	K2 Chatelet	https://m.r	https://ww	4.5	0	276	369.99	104	FALSE	50	
26	B084LP1	JL Anzio Softs	https://m.r	https://ww	4.1	0	98.62	149.99	104	FALSE	300	
27	B0040A6	B1 Amsterdam	https://m.r	https://ww	4.2	0	89.9	121.25	104	FALSE	200	
28	B074ZKX	Z74 Kix Expan	https://m.r	https://ww	4.6	0	78.83	99.99	104	FALSE	300	
29	B08YF6C	N4010 Softsi	https://m.r	https://ww	4.5	0	123.25	144.99	104	FALSE	200	
30	B0812RQ	9 Platinum El	https://m.r	https://ww	4.4	0	399.49	469.99	104	FALSE	50	

—After(size 59KB)—

	A	B	C	D	E	F	G
1	price	reviews	stars	isbestseller	title	boughtinlastmonth	
2	18.99	27263	4.7	TRUE	Shynerk Ba	10000	
3	24.31	28518	4.5	FALSE	Pure Protei	10000	
4	9.71	21854	4.7	TRUE	Edge Shave	30000	
5	18.99	42849	4.6	TRUE	Etekcitcity Inf	10000	
6	8.99	25420	4.4	TRUE	OLANLY Lu:	10000	
7	4.97	98769	4.5	TRUE	NYX PROFE	10000	
8	9.99	12636	4.6	TRUE	e.l.f. Power	40000	
9	8.71	15828	4.7	FALSE	Neutrogena	20000	
10	4.99	20341	4.4	FALSE	essence L	10000	
11	7.2	55040	4.4	FALSE	Lâ€™OrÃ©al	10000	
12	9.5	37968	4.4	FALSE	e.l.f. Porel	10000	
13	16.49	183726	4.6	TRUE	Schick Hydr	10000	
14	15.95	25280	4.6	TRUE	Lilly's Love	10000	
15	44.95	122450	4.3	TRUE	Pharmedoc	10000	
16	15.45	33115	4.6	FALSE	MAJESTIC F	10000	
17	4.44	40005	4.7	TRUE	Bright Start	10000	
18	14.1	12769	4.8	FALSE	Ddrops Org	10000	
19	19.99	16615	4.5	FALSE	Massage Gi	10000	
20	4.99	28988	4.4	FALSE	essence L	10000	
21	6.01	49923	4.6	TRUE	Palmer's Cc	10000	
22	39.99	14705	4.5	TRUE	RAEMAO N	10000	
23	6.47	21095	4.4	FALSE	Dr. Schollâ	10000	
24	14.99	22157	4.5	FALSE	Adams Plus	10000	
25	489.99	24023	4.8	TRUE	Xbox Series	10000	
26	99.95	89852	4.7	TRUE	NOCO Boos	20000	
27	24.99	39441	4.6	FALSE	ANCEL AD3	10000	
28	9.99	12548	4	FALSE	Gel Nail Po	10000	
29	41.95	36945	4.6	FALSE	nbpure Ma	10000	
30	8.93	59015	4.7	TRUE	Garnier Ski	20000	

6. Queries on Athena (1)

Goal: What are the most top-rated products?

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.

SQL to extract the 100 top-rated product list

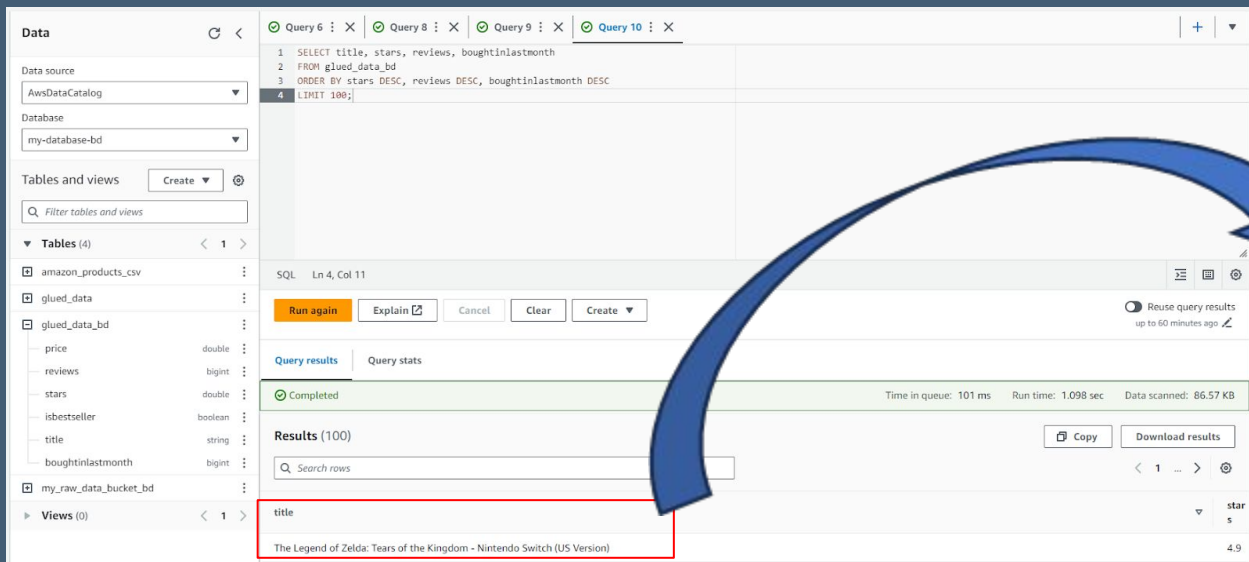
✓ Question#1 :

```
1 SELECT title, stars, reviews, boughtinlastmonth
2 FROM glued_data_bd
3 ORDER BY stars DESC, reviews DESC, boughtinlastmonth DESC
4 LIMIT 100;
```

6. Queries on Athena (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)'



The screenshot shows the AWS Athena console interface. On the left, the 'Data' sidebar lists the data source as 'AwsDataCatalog', the database as 'my-database-bd', and a list of tables including 'amazon_products_csv', 'glued_data', 'glued_data_bd', and 'my_raw_data_bucket_bd'. The 'glued_data_bd' table is expanded, showing columns like 'price', 'reviews', 'stars', 'icbestseller', 'title', and 'boughtinlastmonth'. The main panel displays a SQL query (Query 10) that selects the top 100 products based on stars, reviews, and sales. The query results are shown as a table with 100 rows. The first row is highlighted with a red box, showing the product title 'The Legend of Zelda: Tears of the Kingdom - Nintendo Switch (US Version)'.

```
1 SELECT title, stars, reviews, boughtinlastmonth
2 FROM glued_data_bd
3 ORDER BY stars DESC, reviews DESC, boughtinlastmonth DESC
4 LIMIT 100;
```

SQL Ln 4, Col 11

Run again **Explain** **Cancel** **Clear** **Create**

Query results **Query stats**

Completed Time in queue: 101 ms Run time: 1.098 sec Data scanned: 86.57 KB

Results (100) **Copy** **Download results**

Search rows

title	stars
The Legend of Zelda: Tears of the Kingdom - Nintendo Switch (US Version)	4.9



7. CONCLUSION (1)

List of top 30 top-rated product based on volume of reviews and sales in Sep.2023 in Excel view

	A	B	C	D
1	title	stars	reviews	boughtinlastmonth
2	The Legend of Zelda: Tears of the Kingdom - Nintendo Switch (US Version)	4.9	15084	10000
3	The Legend of Zelda: Tears of the Kingdom - Nintendo Switch (US Version)	4.9	15084	10000
4	Let's Count Soft Book - World of Eric Carle the Very Hungry Caterpillar Baby on the Go Clip Teething Crinkle Soft Sensory Book	4.9	10476	10000
5	Let's Count Soft Book - World of Eric Carle the Very Hungry Caterpillar Baby on the Go Clip Teething Crinkle Soft Sensory Book	4.9	10476	10000
6	PlayStation DualSense Wireless Controller â€” Midnight Black	4.8	90649	10000
7	PlayStation DualSense Wireless Controller â€” Midnight Black	4.8	90649	10000
8	MEGA BLOKS Fisher-Price Toddler Block Toys, Big Building Bag with 80 Pieces and Storage Bag, Blue, Gift Ideas for Kids Age 1+	4.8	77797	10000
9	MEGA BLOKS Fisher-Price Toddler Block Toys, Big Building Bag with 80 Pieces and Storage Bag, Blue, Gift Ideas for Kids Age 1+	4.8	77797	10000
10	Graco TurboBooster 2.0 Backless Booster Car Seat, Denton	4.8	65909	20000
11	Graco TurboBooster 2.0 Backless Booster Car Seat, Denton	4.8	65909	20000
12	Sassy Stacks of Circles Stacking Ring STEM Learning Toy, Age 6+ Months, Multi, 9 Piece Set	4.8	56107	20000
13	Sassy Stacks of Circles Stacking Ring STEM Learning Toy, Age 6+ Months, Multi, 9 Piece Set	4.8	56107	20000
14	Neutrogena Makeup Remover Wipes Singles, Daily Facial Cleanser Towelettes, Gently Removes Oil & Makeup, Alcohol-Free Mi	4.8	54029	60000
15	Neutrogena Makeup Remover Wipes Singles, Daily Facial Cleanser Towelettes, Gently Removes Oil & Makeup, Alcohol-Free Mi	4.8	54029	60000
16	LEGO Classic Medium Creative Brick Box 10696 Building Toy Set - Featuring Storage, Includes Train, Car, and a Tiger Figure, and	4.8	45455	10000
17	LEGO Classic Medium Creative Brick Box 10696 Building Toy Set - Featuring Storage, Includes Train, Car, and a Tiger Figure, and	4.8	45455	10000
18	Infantino Lil' Nibbles Textured Silicone Teether -Sensory Exploration and Teething Relief with Easy to Hold Handles, Orange Cai	4.8	42153	10000
19	Infantino Lil' Nibbles Textured Silicone Teether -Sensory Exploration and Teething Relief with Easy to Hold Handles, Orange Cai	4.8	42153	10000
20	NOW Supplements, Vitamin D-3 5,000 IU, High Potency, Structural Support*, 240 Softgels	4.8	40845	20000
21	NOW Supplements, Vitamin D-3 5,000 IU, High Potency, Structural Support*, 240 Softgels	4.8	40845	20000
22	MunchkinÂ® Float & Play Bubblesâ„C Baby and Toddler Bath Toy, 4 Count	4.8	27099	10000
23	MunchkinÂ® Float & Play Bubblesâ„C Baby and Toddler Bath Toy, 4 Count	4.8	27099	10000
24	Xbox Series X	4.8	24023	10000
25	Xbox Series X	4.8	24023	10000
26	Nature Made Folic Acid 400 mcg (665 mcg DFE), Dietary Supplement for Nervous System Function, 250 Tablets, 250 Day Suppl	4.8	18554	10000
27	Nature Made Folic Acid 400 mcg (665 mcg DFE), Dietary Supplement for Nervous System Function, 250 Tablets, 250 Day Suppl	4.8	18554	10000
28	Fine Ground Celtic Sea Salt â€” (1) 16 Ounce Resealable Bag of Nutritious, Classic Sea Salt, Great for Cooking, Baking, Pickling,	4.8	18110	10000
29	Fine Ground Celtic Sea Salt â€” (1) 16 Ounce Resealable Bag of Nutritious, Classic Sea Salt, Great for Cooking, Baking, Pickling,	4.8	18110	10000
30	Ddrops Organic Baby 400 IU 90 Drops - Daily Vitamin D Liquid for Infants. Supports Teeth & Bone Health. No Preservatives, No	4.8	12769	10000

7. CONCLUSION (2)

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.

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.

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 SOLUTIONS

CHALLENGES FACED:

1. **Challenge:** 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;
```

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

Solution: Must have sufficient fund to afford it.

3. **FUTURE 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



THANKS FOR LISTENING!

ANY QUESTIONS?

Please find all code script, data files and meeting log here: <https://github.com/Duggieeeeey/AWSglue-ETL>

