

Batch Time Analysis of Transactional Data

Steps to perform:

1. Create an S3 bucket with a unique name and upload the CSV file to the S3 bucket (ensure that the file is in UTF-8 format only)

The image shows two screenshots of the AWS Management Console. The top screenshot is the 'Create bucket' page. The bottom screenshot is the 'Buckets' list page.

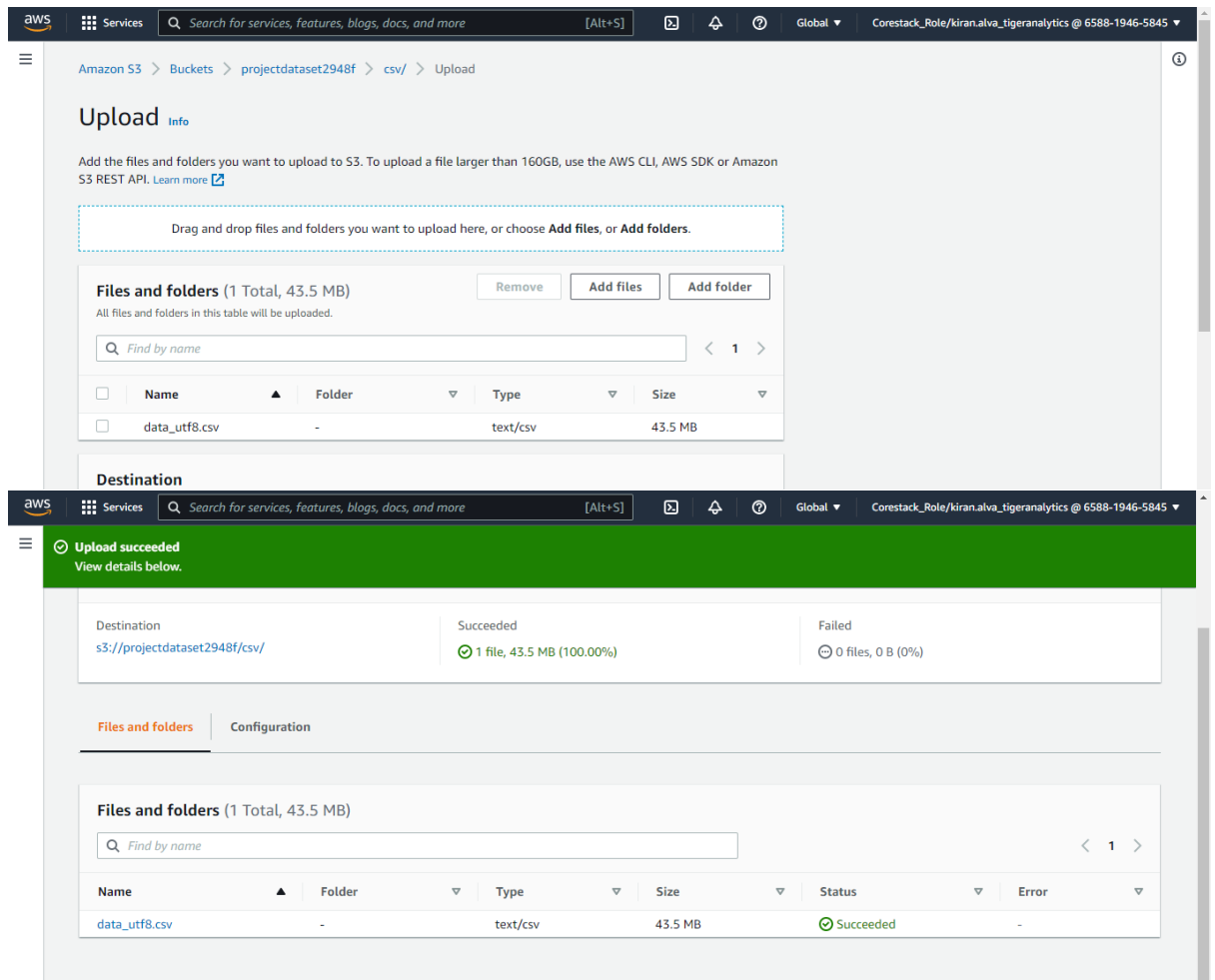
Create bucket page:

- Page title: Create bucket
- Sub-header: Buckets are containers for data stored in S3. [Learn more](#)
- Section: General configuration
- Form fields:
 - Bucket name: projectdataset2948f
 - AWS Region: US East (N. Virginia) us-east-1
- Text: Copy settings from existing bucket - optional. Only the bucket settings in the following configuration are copied.
- Button: Choose bucket

Buckets page:

- Page title: Buckets
- Section: Account snapshot
- Text: Storage lens provides visibility into storage usage and activity trends. [Learn more](#)
- Button: View Storage Lens dashboard
- Section: Buckets (4)
- Text: Buckets are containers for data stored in S3. [Learn more](#)
- Search bar: Find buckets by name
- Buttons: Refresh, Copy ARN, Empty, Delete, Create bucket
- Table:

	Name	AWS Region	Access	Creation date
<input type="radio"/>	aws-glue-assets-658819465845-us-east-1			August 5, 2022, 14:28:14 (UTC+05:30)
<input type="radio"/>	gluebucket1213556			August 5, 2022, 14:22:50 (UTC+05:30)
<input type="radio"/>	kiranalva777			August 5, 2022, 15:27:40 (UTC+05:30)



The screenshot shows the AWS S3 console interface for uploading a file. The breadcrumb navigation indicates the path: Amazon S3 > Buckets > projectdataset2948f > csv/ > Upload.

Upload

Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (1 Total, 43.5 MB)

All files and folders in this table will be uploaded.

Find by name

<input type="checkbox"/>	Name	Folder	Type	Size
<input type="checkbox"/>	data_utf8.csv	-	text/csv	43.5 MB

Destination

Upload succeeded
View details below.

Destination	Succeeded	Failed
s3://projectdataset2948f/csv/	1 file, 43.5 MB (100.00%)	0 files, 0 B (0%)

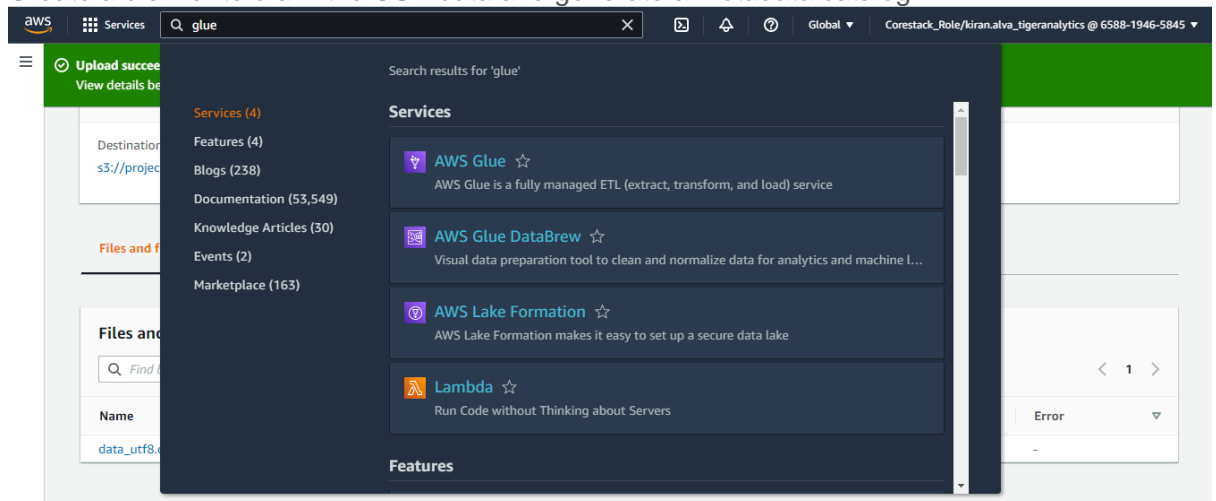
Files and folders | **Configuration**

Files and folders (1 Total, 43.5 MB)

Find by name

Name	Folder	Type	Size	Status	Error
data_utf8.csv	-	text/csv	43.5 MB	Succeeded	-

2. Create a crawler to crawl the CSV data and generate a metadata catalog



The screenshot shows the AWS console search results for 'glue'. The search bar at the top contains 'glue'.

Search results for 'glue'

Services (4)

- AWS Glue** ☆
AWS Glue is a fully managed ETL (extract, transform, and load) service
- AWS Glue DataBrew** ☆
Visual data preparation tool to clean and normalize data for analytics and machine l...
- AWS Lake Formation** ☆
AWS Lake Formation makes it easy to set up a secure data lake
- Lambda** ☆
Run Code without Thinking about Servers

Features

Features (4)

Blogs (238)

Documentation (53,549)

Knowledge Articles (30)

Events (2)

Marketplace (163)

The image displays three sequential screenshots of the AWS Glue console interface, illustrating the steps to add a new crawler.

Top Screenshot: Crawler Overview
The main page shows the 'Crawlers' section. A description states: 'A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.' Below this, there are buttons for 'Add crawler', 'Run crawler', and 'Action'. A search bar is present with the text 'Filter by tags and attributes'. A table header is visible with columns: Name, Schedule, Status, Logs, Last runtime, Median runtime, Tables updated, and Tables added. A message indicates 'You don't have any crawlers yet' with an 'Add crawler' button.

Middle Screenshot: Add crawler wizard
The 'Add crawler' wizard is shown. The first step is 'Crawler info', which is selected. The 'Crawler name' field is filled with 'project'. Below it, there is a section for 'Tags, description, security configuration, and classifiers (optional)'. A 'Next' button is visible at the bottom right.

Bottom Screenshot: Choose S3 path dialog
A dialog box titled 'Choose S3 path' is open. It shows a tree view of S3 buckets and folders. The path 's3://projectdataset2948f/csv' is selected. A 'Select' button is at the bottom. The background shows the 'Add crawler' wizard with the 'Data store' step selected.

Add database

Database name

► Description and location (optional)

Resource link name

Shared database suggestions

Shared database

Shared database owner account ID

Create

Add crawler

- ✓ Crawler info
 - project
- ✓ Crawler source type
 - Data stores
- ✓ Data store
 - S3: s3://projectdata...
- ✓ IAM Role
 - arn:aws:iam::658819465845:role/service-role/AWSGlueServiceRole-kiki
- ✓ Schedule
 - Run on demand
- ✓ Output
 - projectdb
- Review all steps

Crawler info

Name	project
Tags	-

Data stores




Data store	S3
Include path	s3://projectdataset2948f/csv
Connection	
Exclude patterns	

IAM role

IAM role	arn:aws:iam::658819465845:role/service-role/AWSGlueServiceRole-kiki
----------	---

services, features, blogs, docs, and more

[Alt+S]



N. Virginia

Corestack_Role/kiran.alva_tigeranalytics @ 6588-1946-5845

Add crawler

✓ Crawler info

project

✓ Crawler source type

Data stores

✓ Data store

S3: s3://projectdata...

✓ IAM Role

arn:aws:iam::658819465845:role/service-role/AWSGlueServiceRole-kiki

✓ Schedule

Run on demand

✓ Output

projectdb

○ Review all steps

IAM role

IAM role

arn:aws:iam::658819465845:role/service-role/AWSGlueServiceRole-kiki

Schedule

Schedule

Run on demand

Output

Database

projectdb

Prefix added to tables (optional)

Create a single schema for each S3 path

false

Table level (optional)

► Configuration options

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler **project** was created to run on demand. [Run it now?](#)

[User preferences](#)

Add crawler

Run crawler

Action

Filter by tags and attributes

Showing: 1 - 1

<input checked="" type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input checked="" type="checkbox"/>	project		Ready		0 secs	0 secs	0	0

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

[User preferences](#)

Add crawler

Run crawler

Action

Filter by tags and attributes

Showing: 1 - 1

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	project		1min elapsed		1 min	1 min	0	1

services, features, blogs, docs, and more [Alt+S] N. Virginia Corestack_Role/kiran.alva_tigeranalytics @ 6588-1946-5845

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler "project" completed and made the following changes: 1 tables created, 0 tables updated. See the tables created in database [projectdb](#).

[User preferences](#)


[Add crawler](#) [Run crawler](#) [Action](#) Showing: 1 - 1

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	project		Ready	Logs	1 min	1 min	0	1

3. Create a Glue job to transform the data into the Parquet format as CSV is not optimal for data warehouse queries

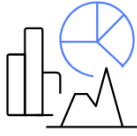
aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia Corestack_Role/kiran.alva_tigeranalytics @ 6588-1946-5845

Getting started




Create and manage jobs
Visually author, run, view, and edit your AWS Glue jobs.

[View jobs](#)



Monitor job runs
Diagnose, debug, and check the status of your AWS Glue jobs.

[View monitoring](#)



Connect external data stores
Create and manage connections to external data stores for AWS Glue Studio.

[View connectors](#)

aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia Corestack_Role/kiran.alva_tigeranalytics @ 6588-1946-5845

Introducing the new job authoring experience using Jupyter notebooks
Write your own code in a Jupyter Notebook for interactive development.

AWS Glue Studio > Jobs

Jobs [Info](#)

Create job [Info](#)

[Create](#)

☐ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

☒ Visual with a blank canvas
Author using an interactive visual interface.

☐ Spark script editor
Write or upload your own Spark code.

☐ Python Shell script editor
Write or upload your own Python shell script.

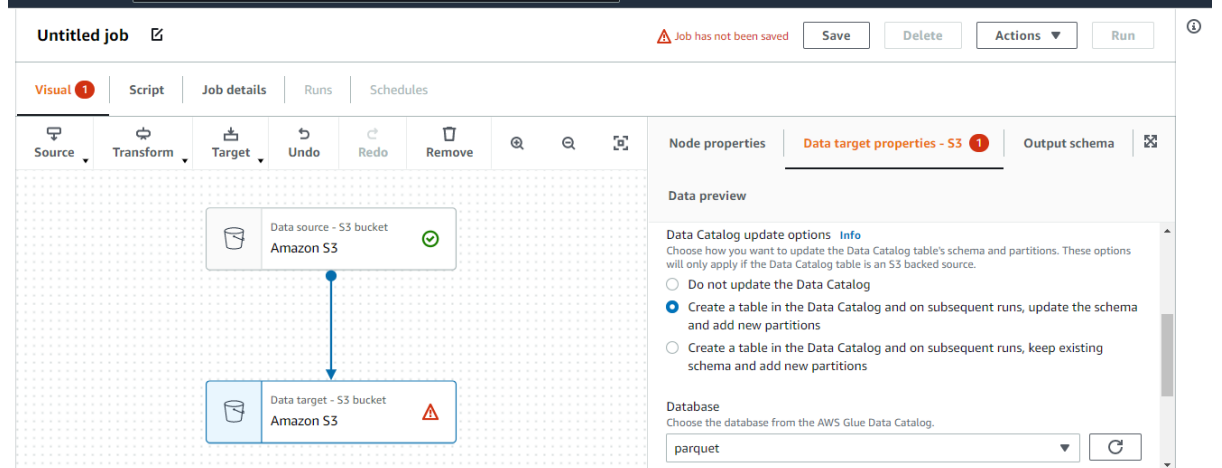
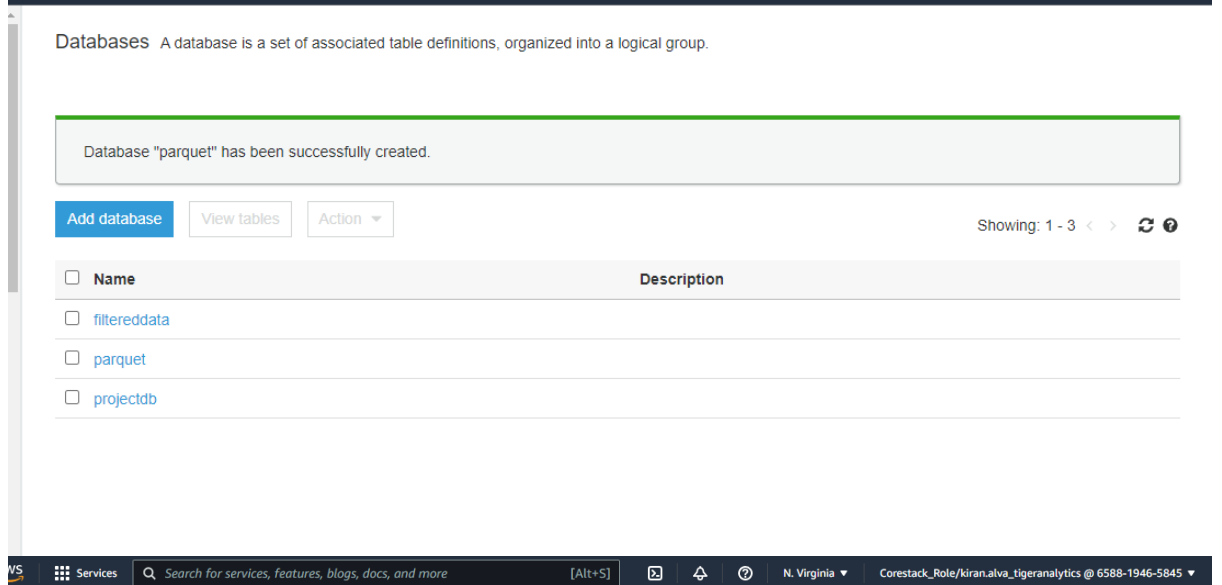
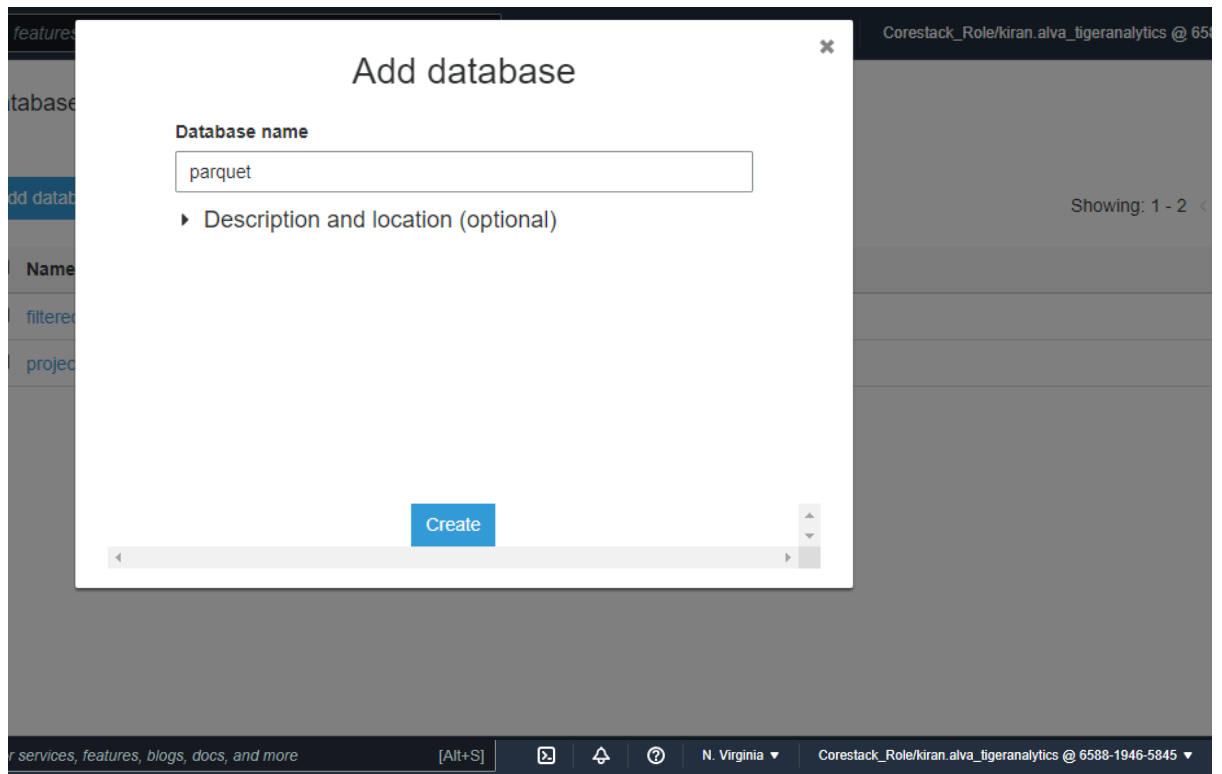
☐ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

Your jobs (2) [Info](#)

[Refresh](#) [Actions](#) [Run job](#)

The screenshots illustrate the configuration of a crawler job in the AWS Glue console. The first screenshot shows the 'Data source properties - S3' tab, where the S3 URL is set to 's3://projectdataset2948f/csv/da'. The second screenshot shows the 'Data format' set to 'CSV' and the 'Delimiter' set to 'Comma (,)'. The third screenshot shows the 'Data target properties - S3' tab, where the 'Format' is set to 'Parquet' and the 'S3 Target Location' is set to 's3://projectdataset2948f/parquet/'.

4. Add another crawler to crawl the Parquet data files to generate the metadata catalog of the Parquet file in order to query it with Athena



The image displays three sequential screenshots of the AWS Glue console interface, showing the configuration and execution of a job named "Project -AWS".

Top Screenshot: Job Configuration (Untitled job)

- Visual Tab:** Shows a workflow with a "Data source - S3 bucket Amazon S3" node connected to a "Data target - S3 bucket Amazon S3" node.
- Node properties:** The "Data target properties - S3" tab is active.
- Data preview:** Shows the configuration for the data target:
 - Choose the database from the AWS Glue Data Catalog: **parquet**
 - Table name: **datautf**
 - Partition keys - optional: **Add a partition key**

Middle Screenshot: Job Configuration (Project -AWS)

- Visual Tab:** Shows the same workflow as the top screenshot.
- Node properties:** The "Data target properties - S3" tab is active.
- Data preview:** Shows the configuration for the data target:
 - Format: **Parquet**
 - Compression Type: **Snappy**
 - S3 Target Location: **Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).**

Bottom Screenshot: Job Execution (Project -AWS)

- Visual Tab:** Shows the same workflow as the top screenshot.
- Runs Tab:** The "Recent job runs (1)" section is active, showing a single run with the following details:
 - Job name: **Project -AWS**
 - Id: **jr_c75131bada2b8ad9cd66b060a1a78287f8dd3835ef6a32a058473ca4a5216b1a**
 - Run status: **Running**
 - Glue version: **3.0**

Project -AWS

Last modified on 8/5/2022, 4:27:21 PM

Save Delete Actions Run

Visual Script Job details **Runs** Schedules

Recent job runs (1) Info

Find runs

August 05, 2022 4:27:45 PM

Rewind job bookmark

Job name	Id	Run status	Glue version
Project -AWS	jr_c75131bada2b8ad9cd66b060a1a78287f8dd3835ef6a32a058473ca4a5216b1a	Succeeded	3.0

Retry attempt number	Start time	End time	Start-up time
Initial run	August 05, 2022 4:27:45 PM	August 05, 2022 4:28:44 PM	7 seconds

Execution time	Last modified on	Trigger name	Security configuration

5. Query the data to identify the best-selling item and countries where customers have bought the most-sold item using Athena

Search results for 'athena'

Services (20) See all 20 results

Features (138) See all 138 results

Features

- Athena ☆ Query Data in S3 using SQL
- FreeRTOS ☆ FreeRTOS is an IoT Operating System for Microcontrollers
- Storage Gateway ☆ Hybrid Storage Integration
- WorkSpaces ☆ Desktops in the Cloud

Showing: 1 - 3

Editor Recent queries Saved queries Settings

Workgroup primary

Data

Data source: AwsDataCatalog

Database: parquet

Tables and views

Filter tables and views

Tables (1)

datautf

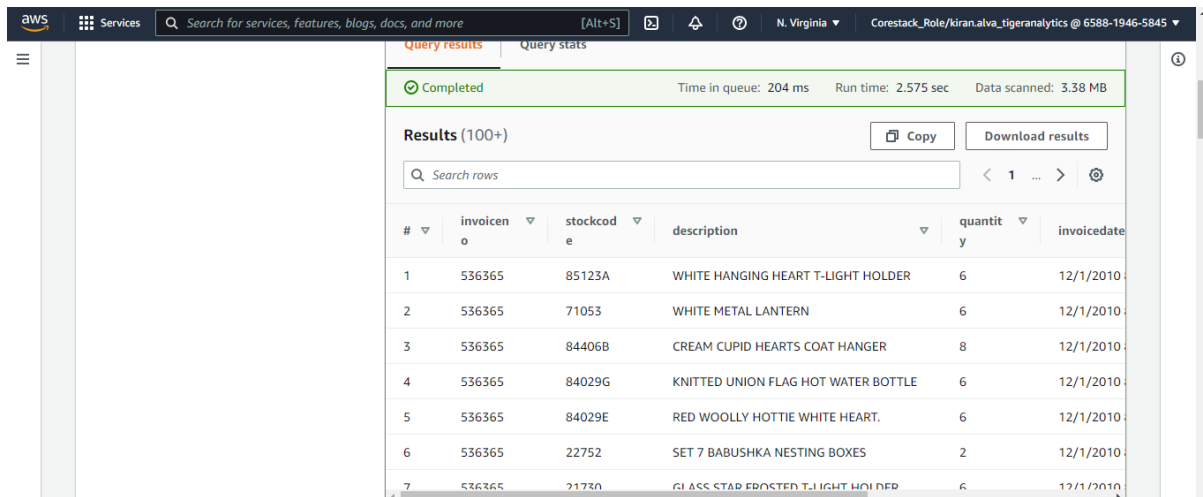
Views (0)

Query 1

```
1 select * from datautf;
```

SQL Ln 1, Col 22

Run again Explain Cancel Save Clear Create



Query results | Query stats

Completed Time in queue: 204 ms Run time: 2.575 sec Data scanned: 3.38 MB

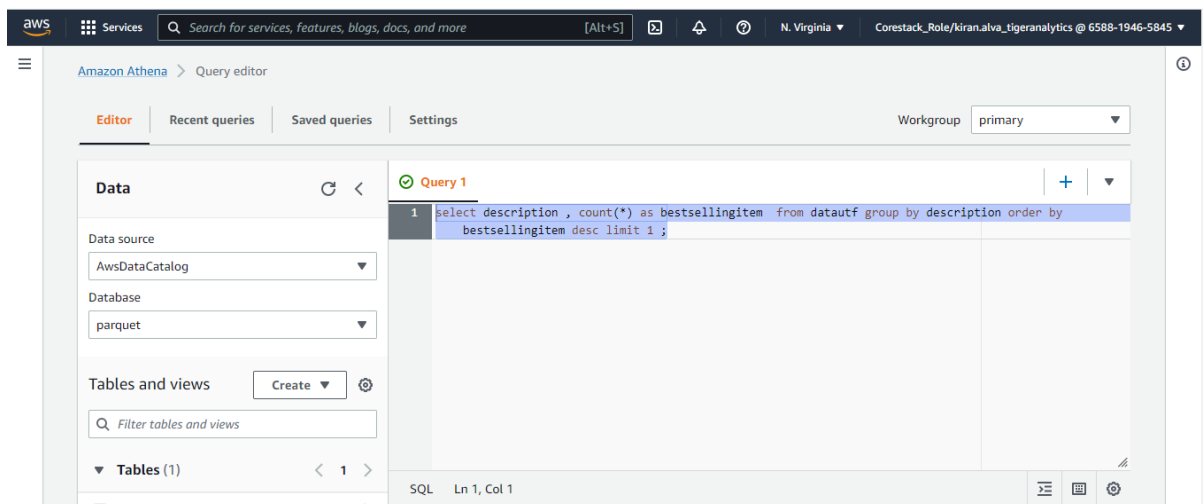
Results (100+) Copy Download results

Search rows

#	invoice no	stockcode	description	quantity	invoicedate
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010
2	536365	71053	WHITE METAL LANTERN	6	12/1/2010
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010
7	536365	21720	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010

Best Selling Item

select description , count(*) as bestsellingitem from datautf group by description order by bestsellingitem desc limit 1 ;



Amazon Athena > Query editor

Editor | Recent queries | Saved queries | Settings

Workgroup: primary

Data

Data source: AwsDataCatalog

Database: parquet

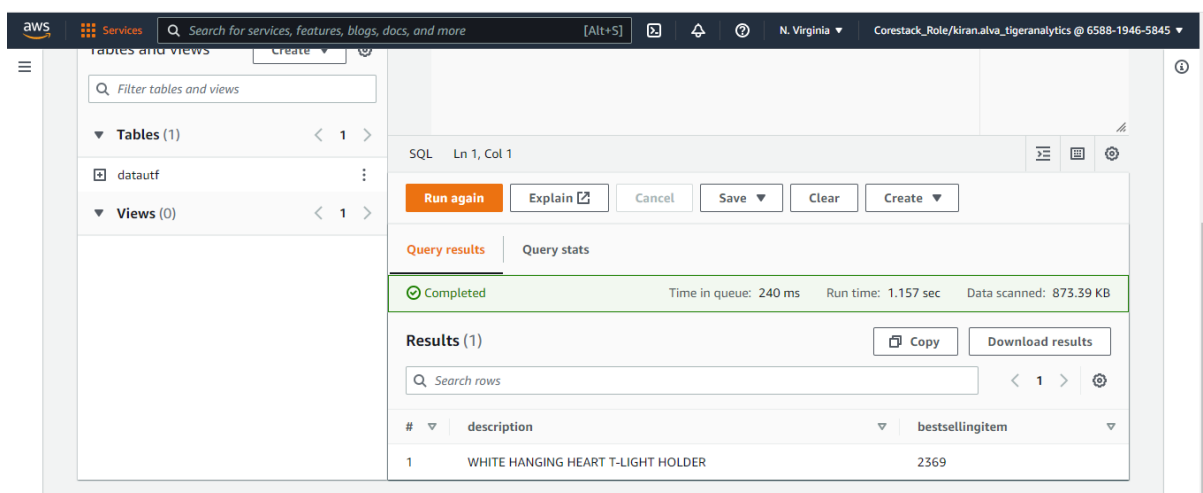
Tables and views: Create Filter tables and views

Tables (1): datautf

Query 1

```
1 select description , count(*) as bestsellingitem from datautf group by description order by bestsellingitem desc limit 1 ;
```

SQL Ln 1, Col 1



Query results | Query stats

Completed Time in queue: 240 ms Run time: 1.157 sec Data scanned: 873.39 KB

Results (1) Copy Download results

Search rows

#	description	bestsellingitem
1	WHITE HANGING HEART T-LIGHT HOLDER	2369

select description, country from datautf where description = (select description from datautf group by description order by count(*) desc limit 1) ;

The screenshot shows the Amazon Athena Query Editor interface. The top navigation bar includes the AWS logo, a search bar, and the user's profile. The main interface has a sidebar on the left with 'Data' and 'Tables (1)' sections. The 'Data' section shows the 'Data source' as 'AwsDataCatalog' and the 'Database' as 'parquet'. The 'Query 1' editor shows the SQL query: `select description, country from datautf where description = (select description from datautf group by description order by count(*) desc limit 1) ;`. The status bar at the bottom indicates 'SQL Ln 2, Col 151'.

The screenshot shows the Amazon Athena Results page. The top navigation bar is the same as the previous screenshot. The main interface shows the 'Results (100+)' section. The results are displayed in a table with columns '#', 'description', and 'country'. The first 10 rows of results are shown, all with the same description: 'WHITE HANGING HEART T-LIGHT HOLDER' and country: 'United Kingdom'.

#	description	country
1	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
2	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
3	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
4	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
5	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
6	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
7	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
8	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
9	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
10	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom

The screenshot shows the Amazon Athena Results page, displaying rows 91 to 100 of the query results. The results are displayed in a table with columns '#', 'description', and 'country'. All rows have the same description: 'WHITE HANGING HEART T-LIGHT HOLDER' and country: 'United Kingdom'.

#	description	country
91	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
92	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
93	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
94	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
95	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
96	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
97	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
98	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
99	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
100	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom