ITCS-6190
Cloud Computing for Data Analytics
LLM - Detect AI Generated Text

Team Members

Apoorve Bhargava
Soumik Paul
Juhi Jadhav
Raj Kumar
Sai Kumar Reddy, Bommareddy

Dataset: **LLM - Detect AI Generated Text**

AWS Services Used
AWS S3
AWS Athena
AWS Glue
AWS Quicksight
AWS Sagemaker

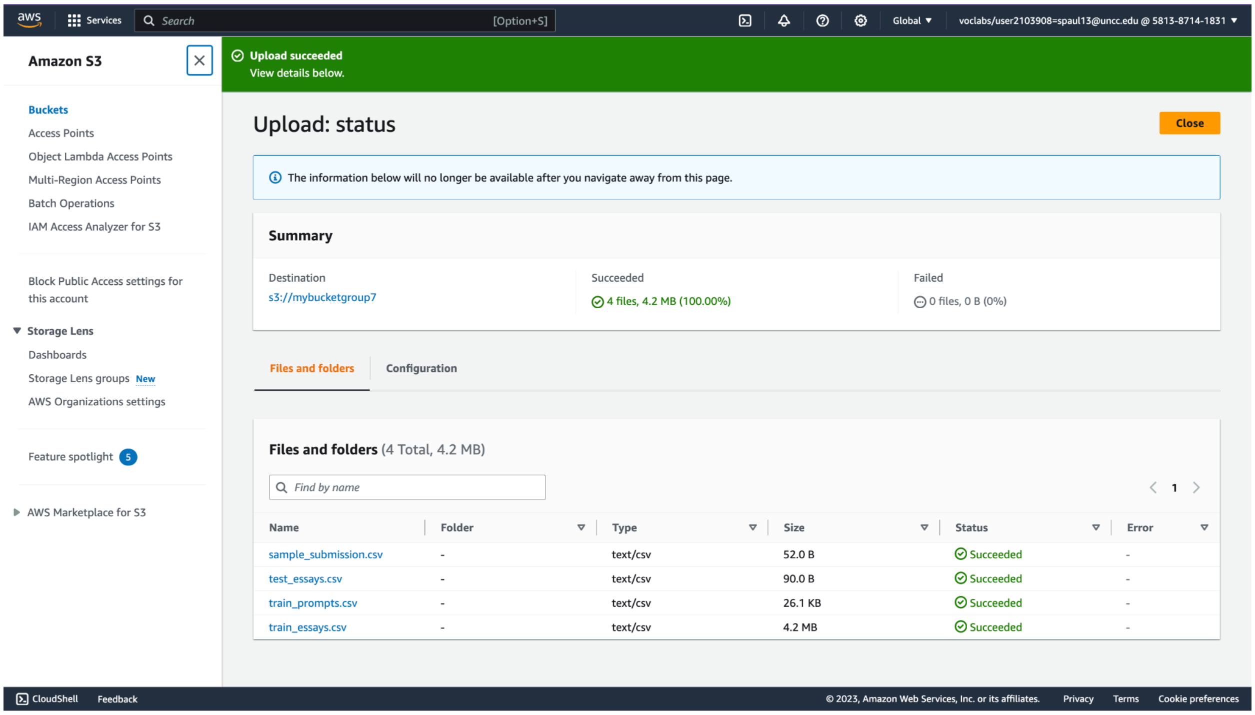
About the data

The LLM (Large Language Model) Detect AI Generated Text dataset is likely designed for training models to identify and differentiate between human-generated and AI-generated text. It encompasses a diverse collection of text samples, featuring instances of both human and machine-generated content. This dataset aims to enhance the capabilities of language models in discerning the unique patterns and characteristics associated with AI-generated text, aiding in the development of effective detection algorithms. Researchers and developers use such datasets to advance the field of natural language processing by creating models that can reliably identify content produced by sophisticated language models like GPT (Generative Pre-trained Transformer).

There are 4 CSV files in the dataset with respective columns

1. **Train_Essays.csv: Id, prompt_id, text, generated**
2. **Train_Prompts.csv: prompt_id, prompt_name, instructions, source_text**
3. **Test_essays.csv : id, prompt_id, text**
4. **Sample_submission.csv : id, generated**

Creation of Bucket and storing the dataset into the bucket



Upload succeeded
View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

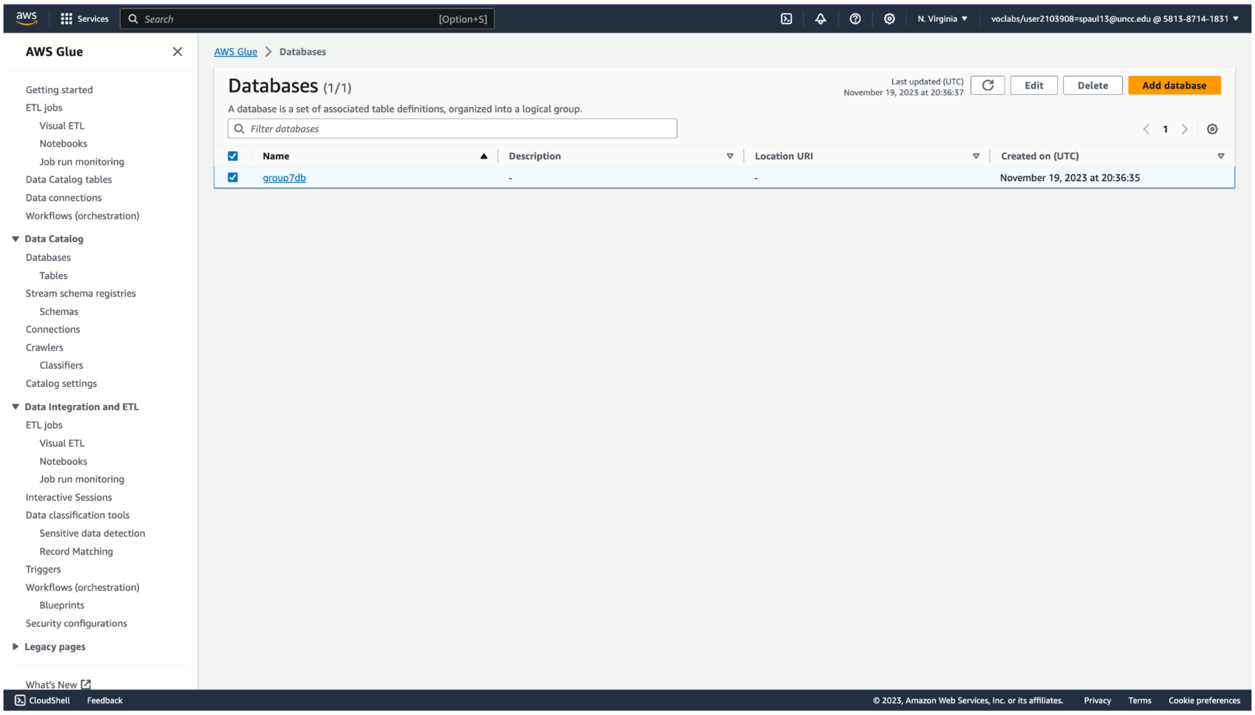
Destination	Succeeded	Failed
s3://mybucketgroup7	4 files, 4.2 MB (100.00%)	0 files, 0 B (0%)

Files and folders (4 Total, 4.2 MB)

Name	Folder	Type	Size	Status	Error
sample_submission.csv	-	text/csv	52.0 B	Succeeded	-
test_essays.csv	-	text/csv	90.0 B	Succeeded	-
train_prompts.csv	-	text/csv	26.1 KB	Succeeded	-
train_essays.csv	-	text/csv	4.2 MB	Succeeded	-

Data Preparation using AWS Glue

Step 1: AWS Glue is used to fetch the data from S3 to AWS Glue

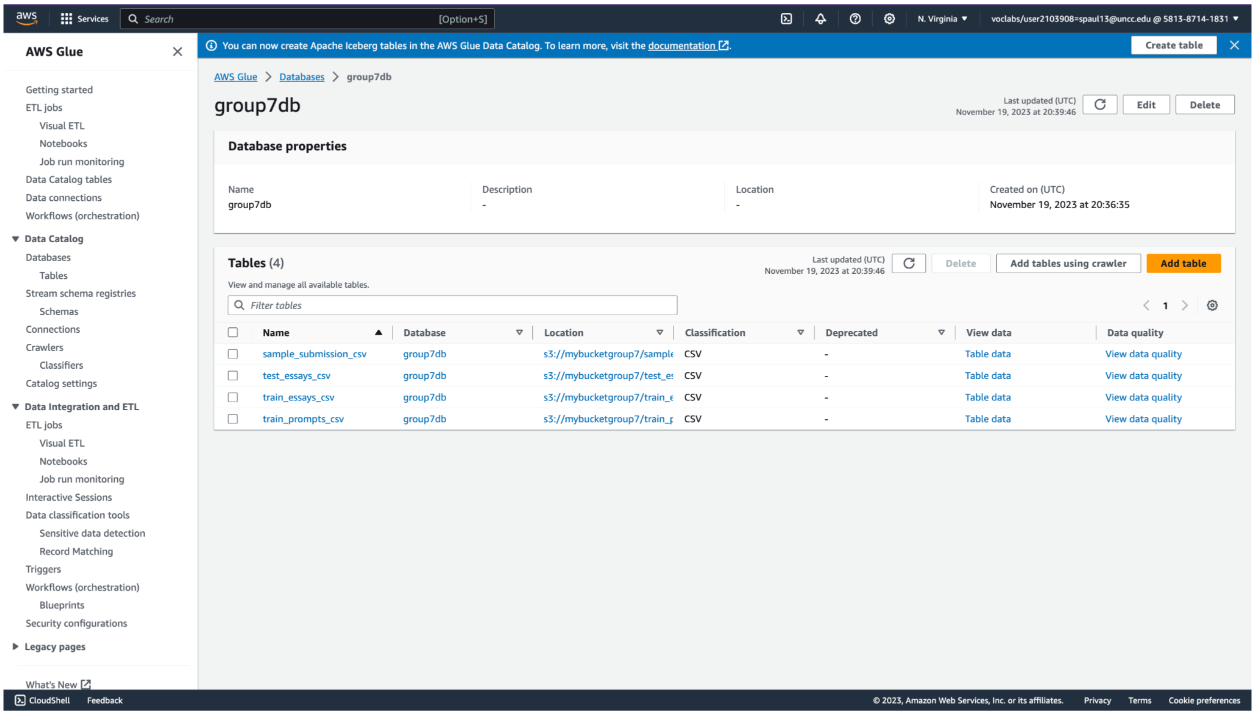


Databases (1/1)

Last updated (UTC)
November 19, 2023 at 20:36:37

A database is a set of associated table definitions, organized into a logical group.

☑	Name	Description	Location URI	Created on (UTC)
☑	group7db	-	-	November 19, 2023 at 20:36:35



group7db

Last updated (UTC)
November 19, 2023 at 20:39:46

Database properties

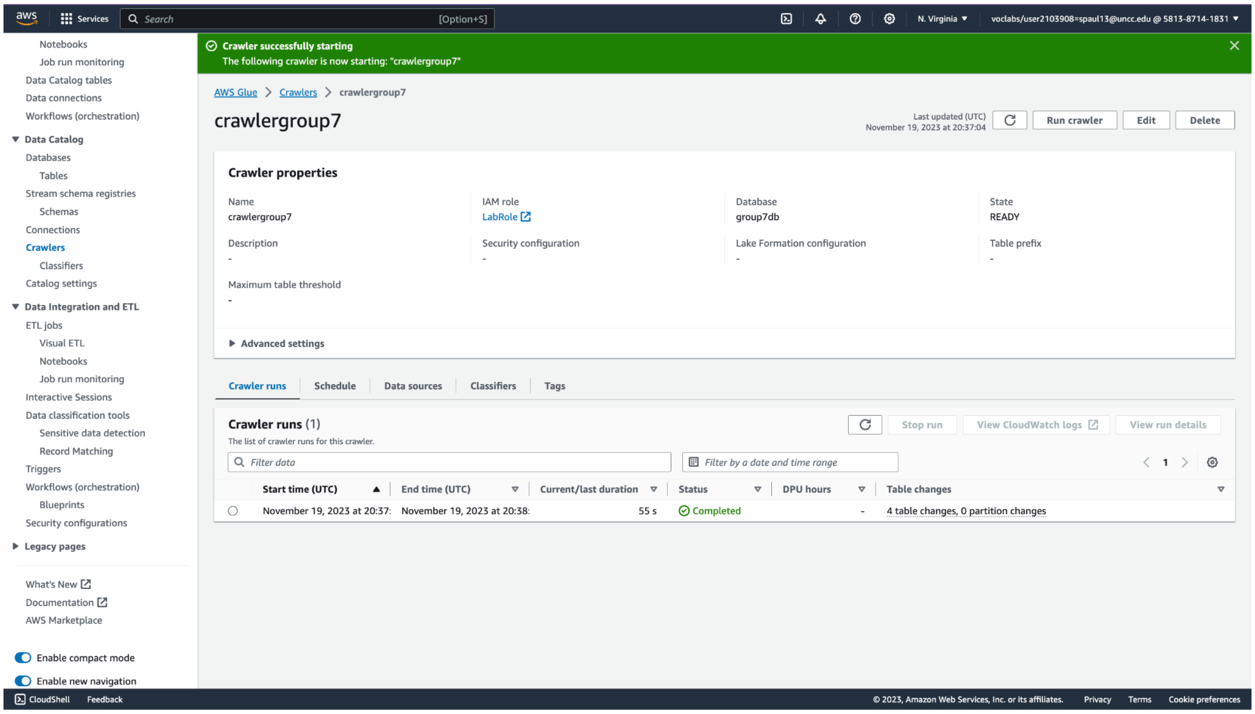
Name	Description	Location	Created on (UTC)
group7db	-	-	November 19, 2023 at 20:36:35

Tables (4)

Last updated (UTC)
November 19, 2023 at 20:39:46

View and manage all available tables.

□	Name	Database	Location	Classification	Deprecated	View data	Data quality
□	sample_submission_csv	group7db	s3://mybucketgroup7/sample	CSV	-	Table data	View data quality
□	test_essays_csv	group7db	s3://mybucketgroup7/test_es	CSV	-	Table data	View data quality
□	train_essays_csv	group7db	s3://mybucketgroup7/train_e	CSV	-	Table data	View data quality
□	train_prompts_csv	group7db	s3://mybucketgroup7/train_p	CSV	-	Table data	View data quality



Crawler successfully starting
The following crawler is now starting: "crawlergroup7"

AWS Glue > Crawlers > crawlergroup7

crawlergroup7

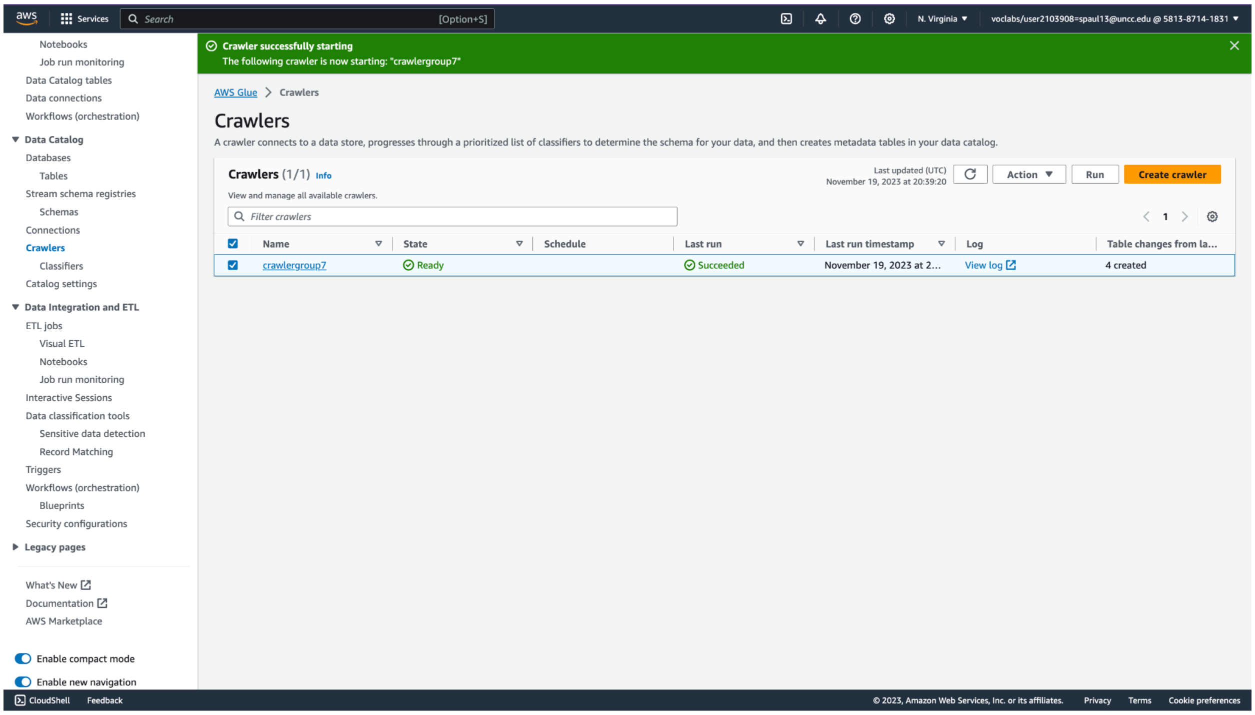
Last updated (UTC)
November 19, 2023 at 20:37:04

Crawler properties

Name	IAM role	Database	State
crawlergroup7	LabRole	group7db	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	-
Maximum table threshold			
-			

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

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	November 19, 2023 at 20:37:	November 19, 2023 at 20:38:	55 s	✓ Completed	-	4 table changes, 0 partition changes



Crawler successfully starting
The following crawler is now starting: "crawlergroup7"

AWS Glue > Crawlers

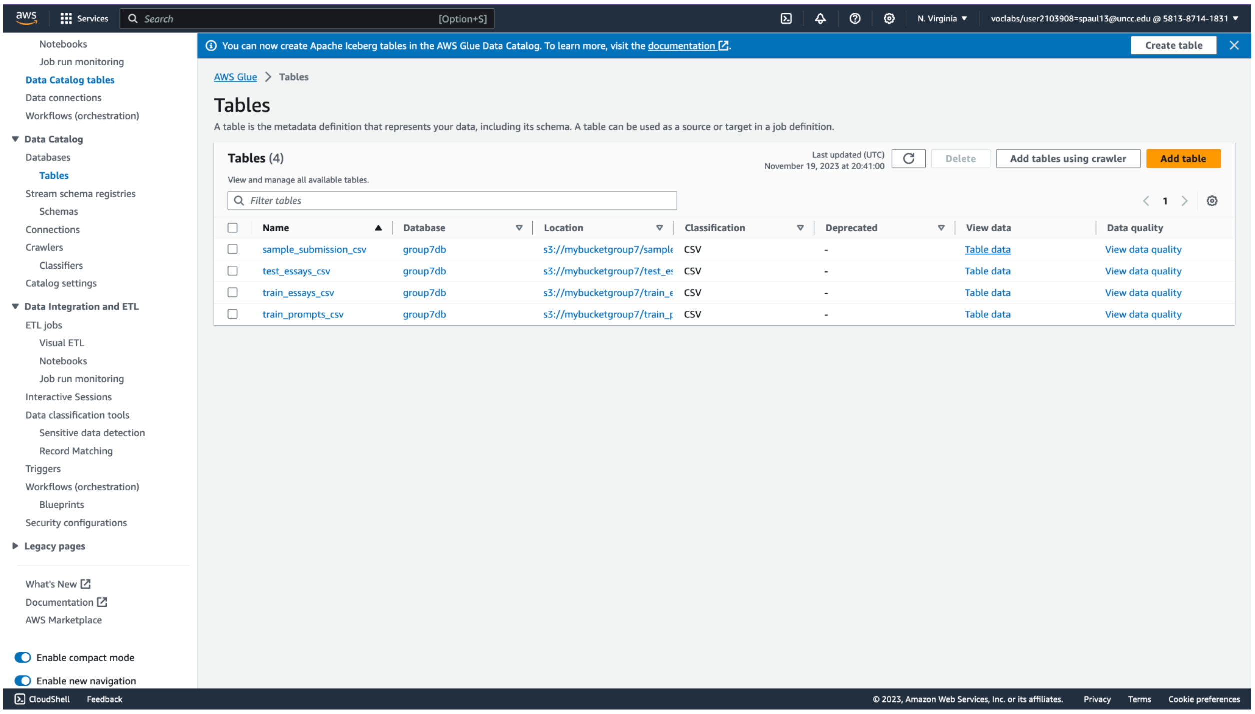
Crawlers

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.

Crawlers (1/1)
View and manage all available crawlers.

Last updated (UTC)
November 19, 2023 at 20:39:20

☑	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from la...
☑	crawlergroup7	✓ Ready		✓ Succeeded	November 19, 2023 at 2...	View log	4 created



You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the documentation.

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

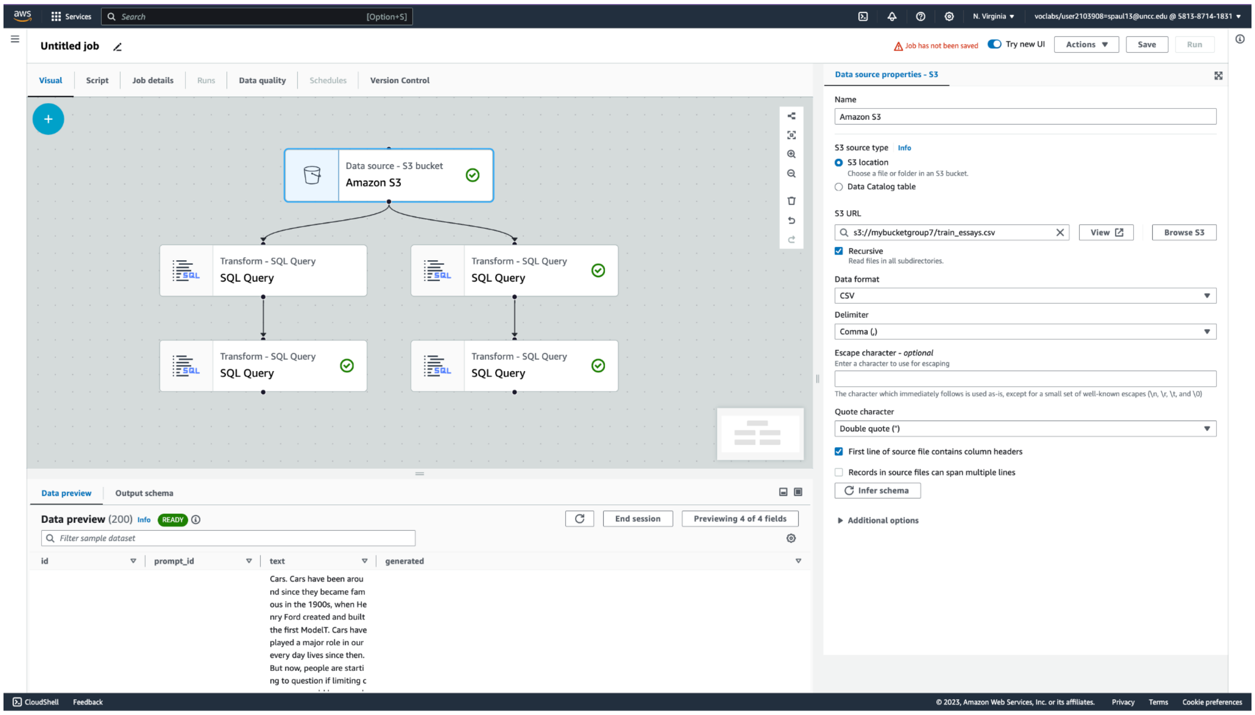
Tables (4)

View and manage all available tables.

Last updated (UTC)
November 19, 2023 at 20:41:00

Name	Database	Location	Classification	Deprecated	View data	Data quality
sample_submission_csv	group7db	s3://mybucketgroup7/sampl	CSV	-	Table data	View data quality
test_essays_csv	group7db	s3://mybucketgroup7/test_es	CSV	-	Table data	View data quality
train_essays_csv	group7db	s3://mybucketgroup7/train_e	CSV	-	Table data	View data quality
train_prompts_csv	group7db	s3://mybucketgroup7/train_p	CSV	-	Table data	View data quality

Data Transformation using Visual ETL



Data source properties - S3

Name: Amazon S3

S3 source type
● S3 location
Choose a file or folder in an S3 bucket.
○ Data Catalog table

S3 URL: s3://mybucketgroup7/train_essays.csv

☑ Recursive
Read files in all subdirectories.

Data format: CSV

Delimiter: Comma (,)

Escape character - *optional*
Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

Quote character: Double quote (")

☑ First line of source file contains column headers

□ Records in source files can span multiple lines

Infer schema

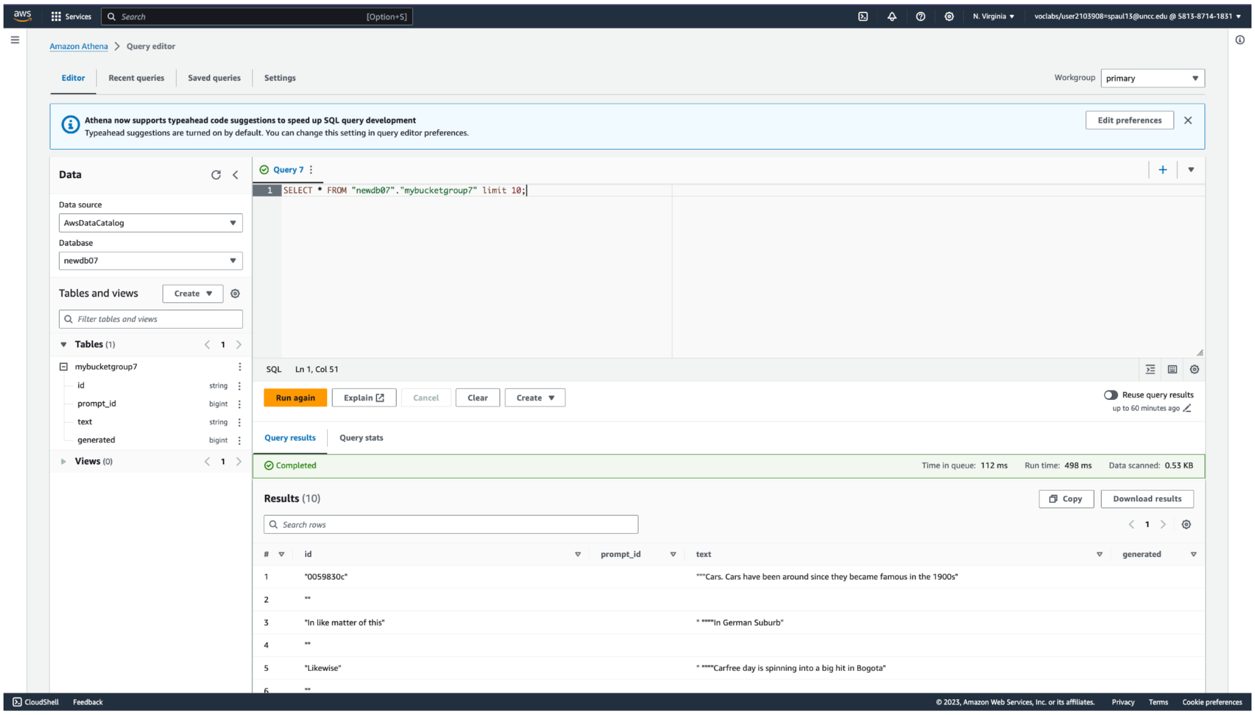
Additional options

Data preview (200) READY

id	prompt_id	text	generated
		Cars. Cars have been around since they became famous in the 1900s, when Henry Ford created and built the first ModelT. Cars have played a major role in our every day lives since then. But now, people are starting to question if limiting c	

Step 2: Preview the data on AWS Athena

Run a few SQL queries on AWS Athena to understand the datasets.



Workgroup: primary

Athena now supports typeahead code suggestions to speed up SQL query development
Typeahead suggestions are turned on by default. You can change this setting in query editor preferences.

Data source: AwsDataCatalog

Database: newdb07

mybucketgroup7	
id	string
prompt_id	bigint
text	string
generated	bigint

```
SELECT * FROM "newdb07"."mybucketgroup7" limit 10;
```

Ln 1, Col 51

Reuse query results
up to 60 minutes ago

Completed   Time in queue: 112 ms   Run time: 498 ms   Data scanned: 0.53 KB

Results (10)

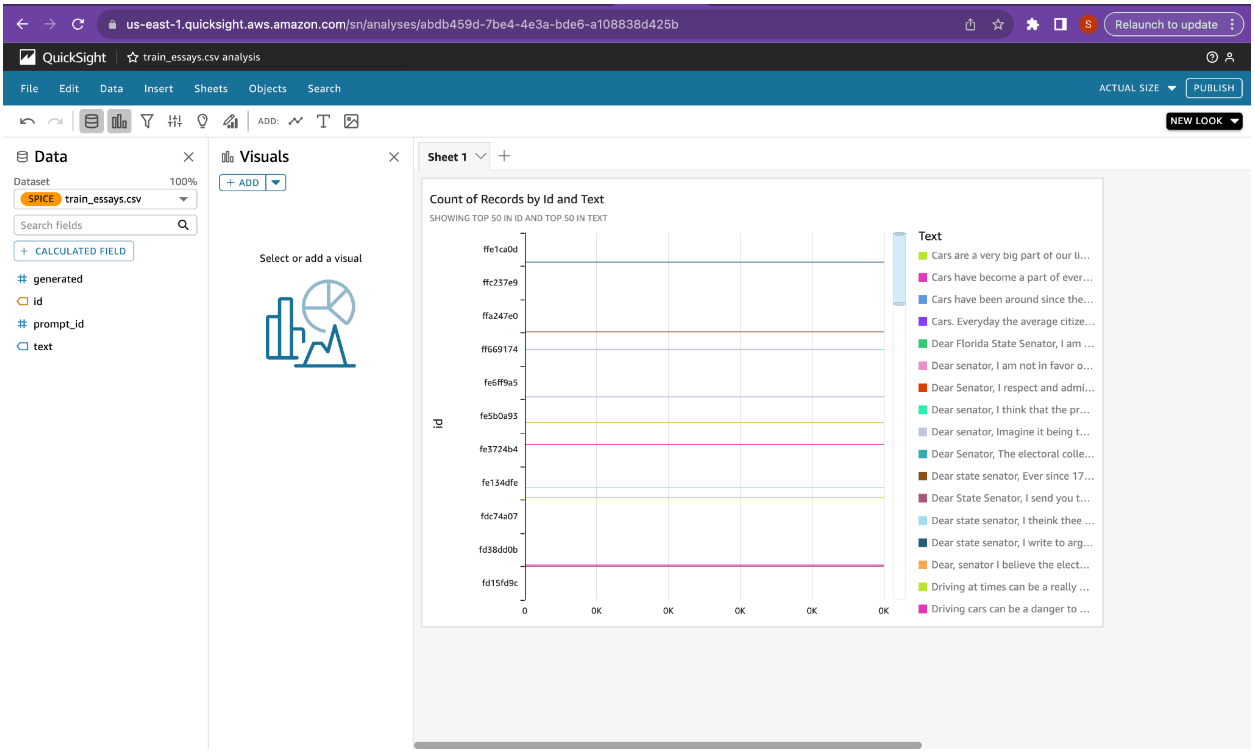
#	id	prompt_id	text	generated
1	"0059830c"		"""Cars. Cars have been around since they became famous in the 1900s"	
2	""			
3	"In like matter of this"		" """"In German Suburb"	
4	""			
5	"Likewise"		" """"Carfree day is spinning into a big hit in Bogota"	
6	""			

© 2023, Amazon Web Services, Inc. or its affiliates.

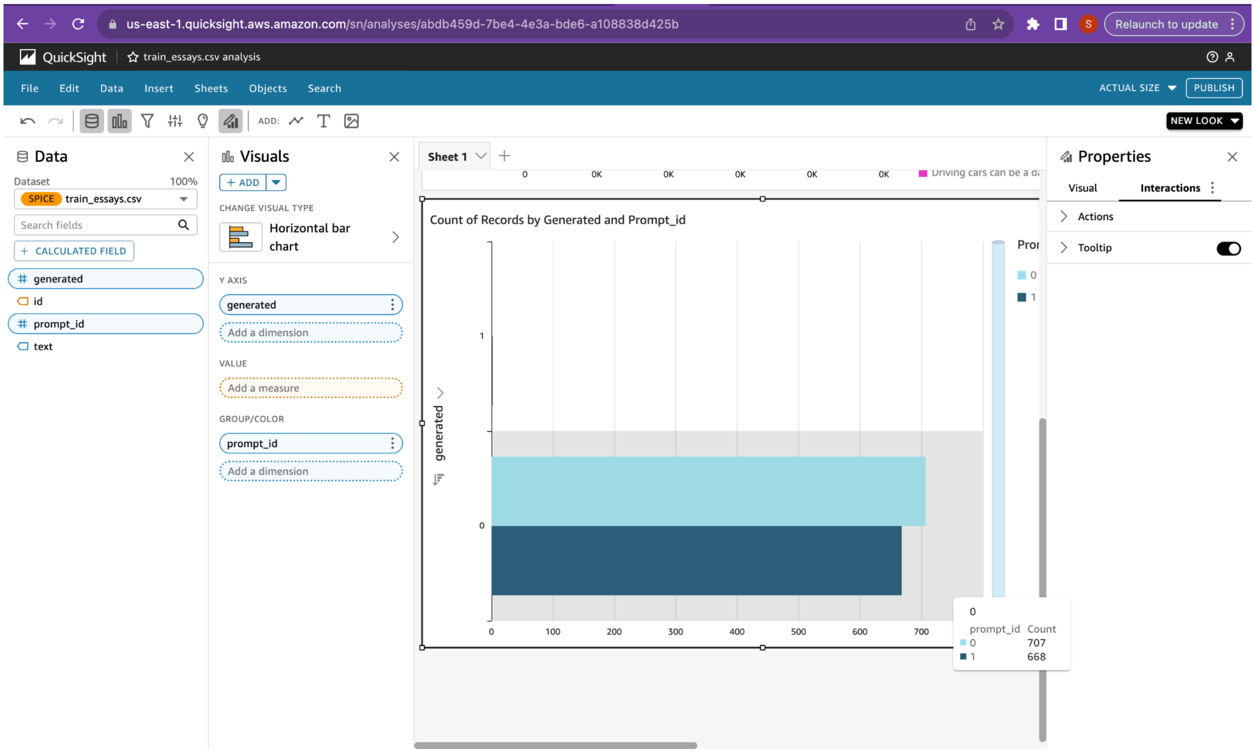
Data Visualization on Quick Sight

The dataset has been visualized on AWS Quick Sight to get a better understanding of the data

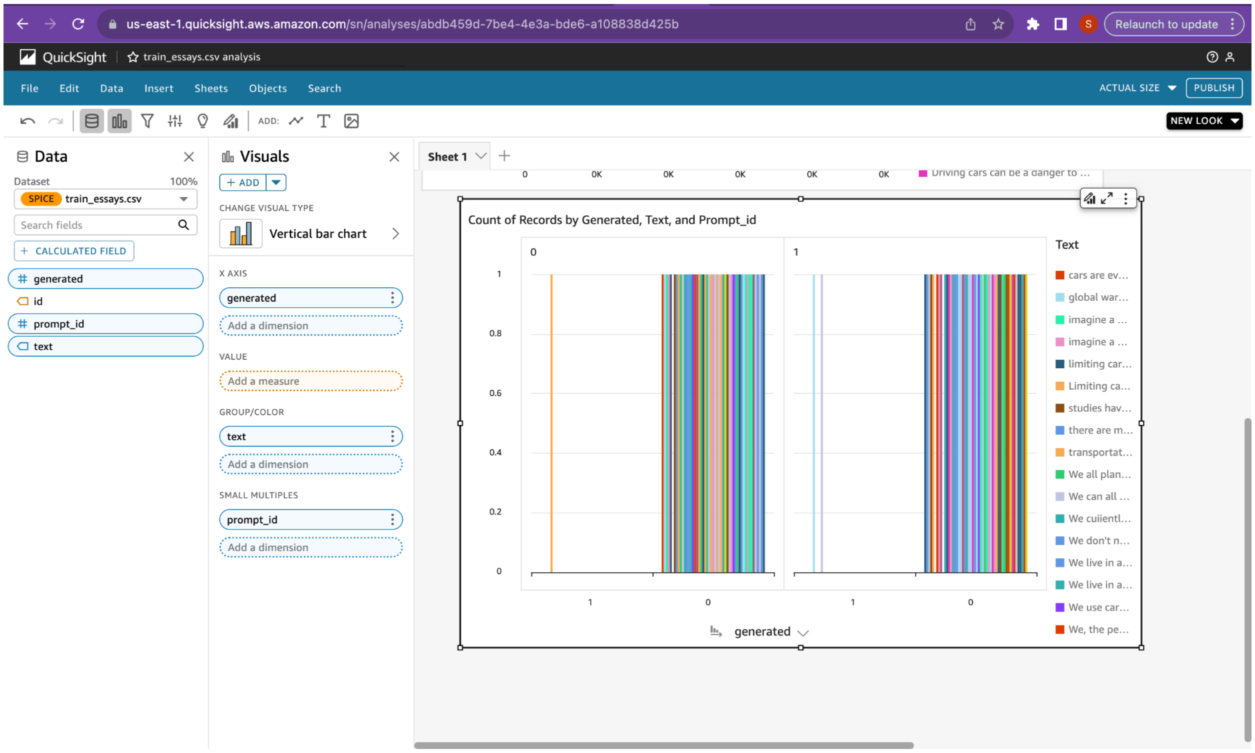
Count of Records by IDs and Text


us-east-1.quicksight.aws.amazon.com/sn/analyses/abdb459d-7be4-4e3a-bde6-a108838d425b
Relaunch to update
QuickSight
train_essays.csv analysis
File
Edit
Data
Insert
Sheets
Objects
Search
ACTUAL SIZE
PUBLISH
ADD:
NEW LOOK
Data
Dataset
100%
SPICE
train_essays.csv
Search fields
+ CALCULATED FIELD
generated
id
prompt_id
text
Visuals
+ ADD
Select or add a visual
Sheet 1
Count of Records by Id and Text
SHOWING TOP 50 IN ID AND TOP 50 IN TEXT
id
ffe1ca0d
ffc237e9
ffa247e0
ff669174
fe6ff9a5
fe5b0a93
fe3724b4
fe134dfe
fdc74a07
fd38dd0b
fd15fd9c
0
0K
0K
0K
0K
0K
Text
Cars are a very big part of our li…
Cars have become a part of ever…
Cars have been around since the…
Cars. Everyday the average citize…
Dear Florida State Senator, I am …
Dear senator, I am not in favor o…
Dear Senator, I respect and admi…
Dear senator, I think that the pr…
Dear senator, Imagine it being t…
Dear Senator, The electoral colle…
Dear state senator, Ever since 17…
Dear State Senator, I send you t…
Dear state senator, I theink thee …
Dear state senator, I write to arg…
Dear, senator I believe the elect…
Driving at times can be a really …
Driving cars can be a danger to …

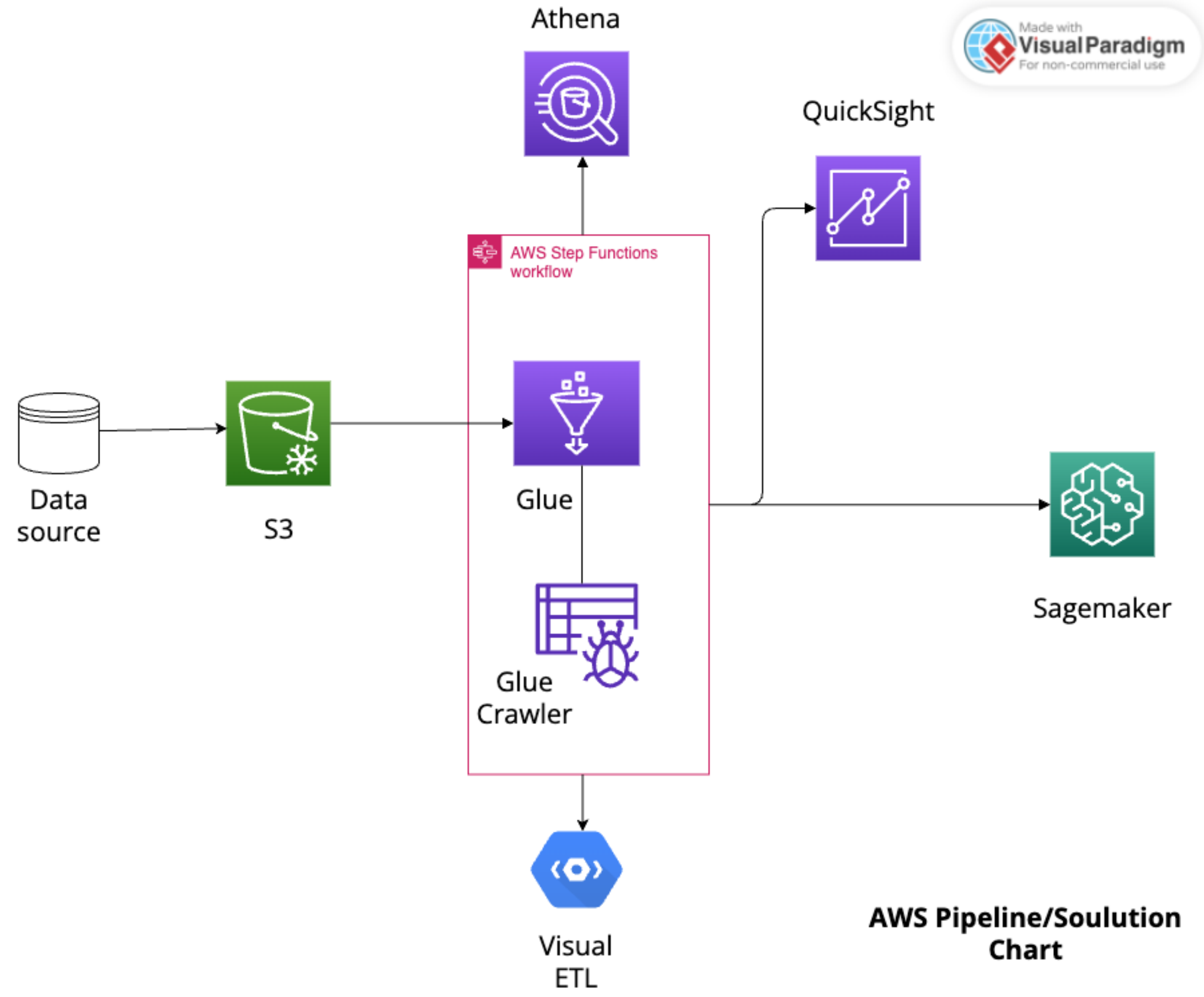
Count of Records by Generated and Prompt_Ids



Count of Records by Generated Text, and Prompt_Ids



AWS Pipeline Solution Chart


Athena
Made with Visual Paradigm For non-commercial use
QuickSight
AWS Step Functions workflow
Data source
S3
Glue
Sagemaker
Glue Crawler
Visual ETL
AWS Pipeline/Soulution Chart