Contents

[AWS Comprehend Tutorial 3](#_Toc60052610)

[Creating an administrator IAM user and group (console) 3](#_Toc60052611)

[Tutorial: Analyzing Insights from Customer Reviews with Amazon Comprehend 7](#_Toc60052612)

[TL;DR 7](#_Toc60052613)

[Running Analysis Jobs on Documents in Amazon S3 8](#_Toc60052614)

[Tutorial: Custom entities analysis 12](#_Toc60052615)

[TL;DR 12](#_Toc60052616)

[Train the recogniser 12](#_Toc60052617)

[STEPS 12](#_Toc60052618)

[Analyse the results 15](#_Toc60052619)

[Prepare the output 15](#_Toc60052620)

[View the output 15](#_Toc60052621)

[Observations 16](#_Toc60052622)

[Preparing the Amazon Comprehend Output for Data Visualization 17](#_Toc60052623)

[TL;DR 17](#_Toc60052624)

[Load the Data into an AWS Glue Data Catalog 19](#_Toc60052625)

[TL;DR 19](#_Toc60052626)

[Prepare the Data for Analysis 24](#_Toc60052627)

[TL;DR 24](#_Toc60052628)

[Visualizing Amazon Comprehend Output in Amazon QuickSight 27](#_Toc60052629)

[TL;DR 27](#_Toc60052630)

[Give Amazon QuickSight Access 27](#_Toc60052631)

[Import the Datasets 28](#_Toc60052632)

[Create a Sentiment Visualization 31](#_Toc60052633)

[Publish a Dashboard 33](#_Toc60052634)

**Document Title:** AWS Comprehend Tutorial.docx By: Arapaut V. Sivaprasad

**Created Date**: 22-Dec-20 **Last Modified**: 28 Dec. 20

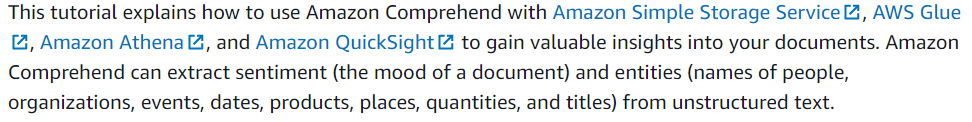
**Purpose**: To learn and document the process of creating an AWS IAM user account, do a Comprehend Analysis, prepare the results data using AWS Glue and AWS Athena and visualise the results using AWS QuickSight.

**DISCLAIMER**

This document is for the personal use of the author and is not structured as a tutorial document for others. Details are probably incomplete, cryptic and understandable only to the author. For more details, use the AWS URLs where the processes are described. The info at these URLs may have changed after this document was created. Usernames and passwords in the doc are abbreviated for security, and the full passwords cannot be guessed.

# AWS Comprehend Tutorial

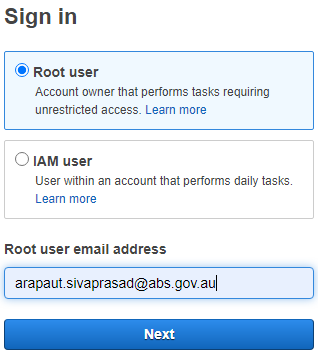
<https://docs.aws.amazon.com/comprehend/latest/dg/tutorial-reviews.html>



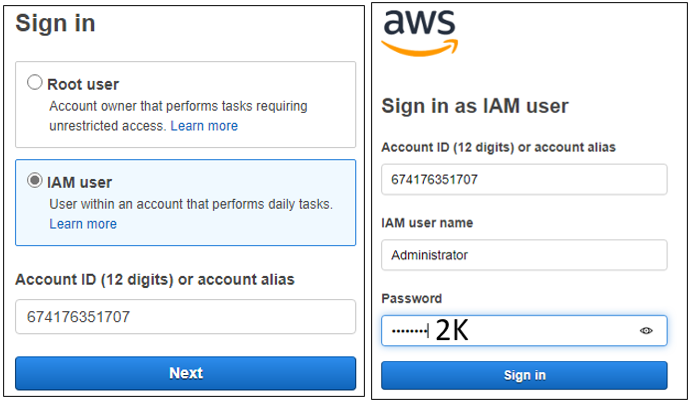
## Creating an administrator IAM user and group (console)

<https://docs.aws.amazon.com/IAM/latest/UserGuide/getting-started_create-admin-group.html>

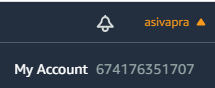
* Sign in to the [IAM console](https://console.aws.amazon.com/iam/)
  + Sign in using the AWS account details…
  + <https://aws.amazon.com/console/>
  + 



IAM User ID: **674176351707** - This is the user ID of the root user. An Identity and Account Management (IAM) user, ‘Administrator’, is created as described below. Either can be used to login, but IAM is safer than the root user for most actions. You can create multiple IAM users with varying privileges.

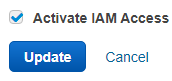


* On the navigation bar, choose your account name, and then choose My Account.



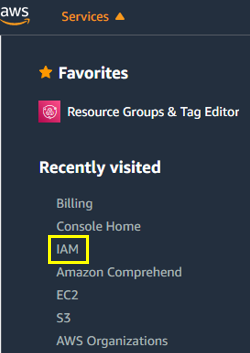
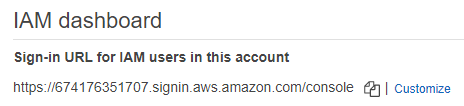
* Next to IAM User and Role Access to Billing Information, choose Edit.

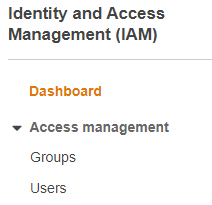
* Select the check box to **Activate IAM Access** and choose **Update**.

IAM user/role access to billing information is activated.

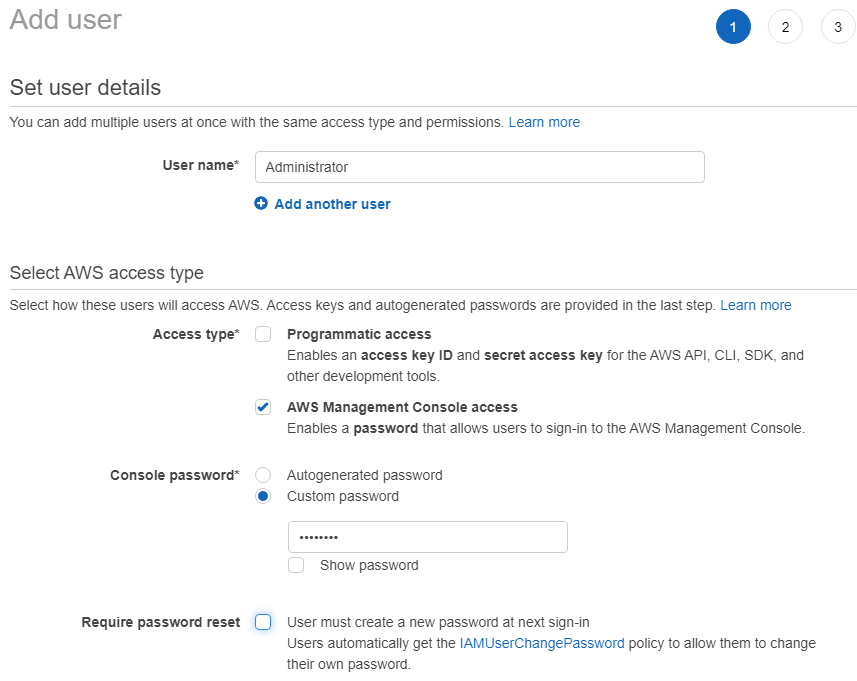
* On the navigation bar, choose **Services** and then **IAM** to return to the IAM dashboard.

* In the navigation pane, choose Users and then choose Add user.

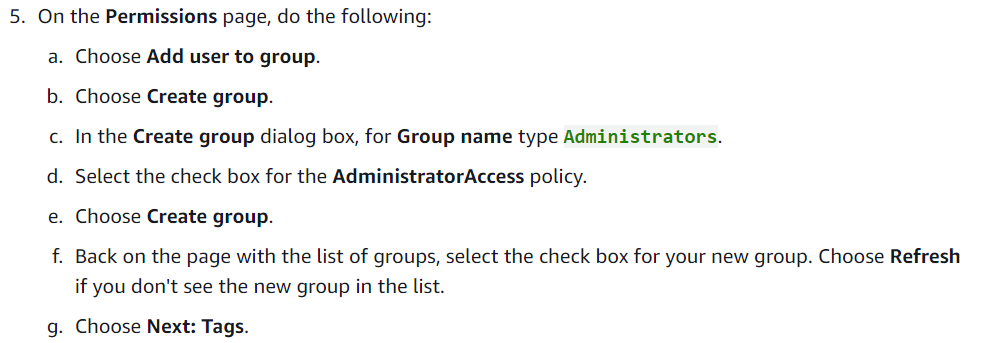
* On the **Details** page, do the following:
  + For **Username**, type **Administrator**.
  + Select the check box for **AWS Management Console access**, select **Custom password**, and then type your new password in the text box.
  + By default, AWS forces the new user to create a new password when first signing in. You can optionally clear the check box next to **User must create a new password at next sign-in** to allow the new user to reset their password after they sign in.

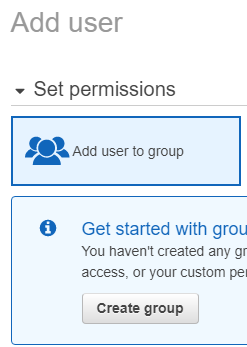
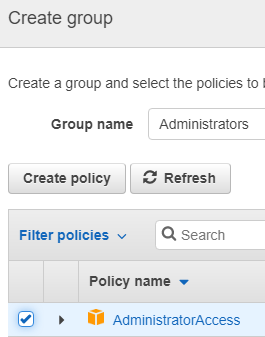


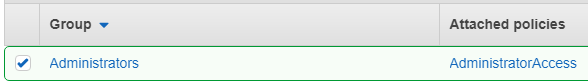
2K

* + Choose **Next: Permissions**.





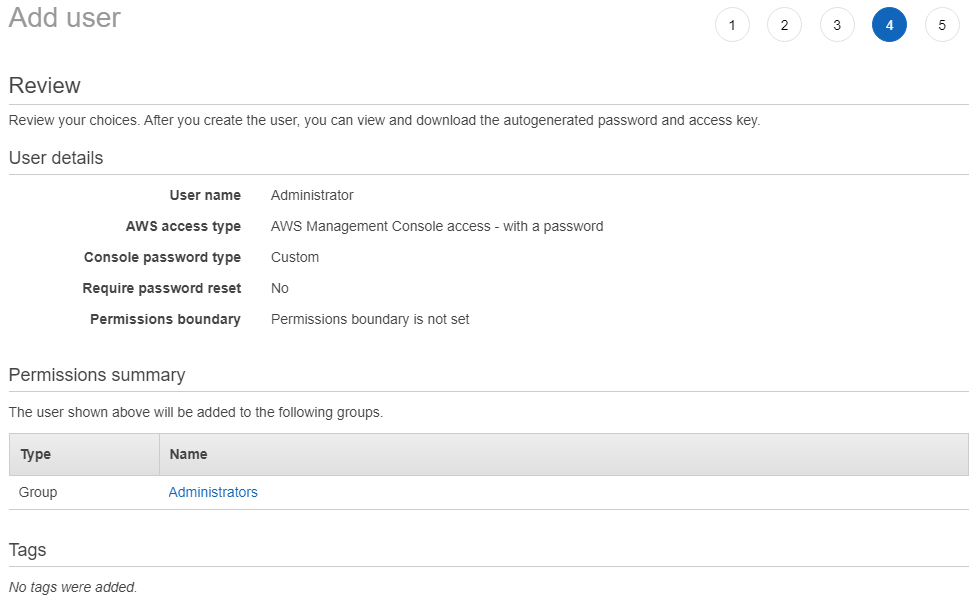
  

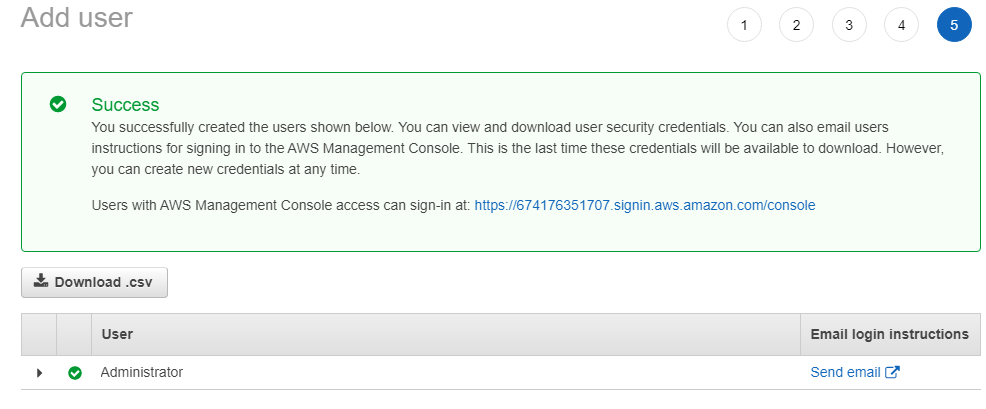
(Skip this)



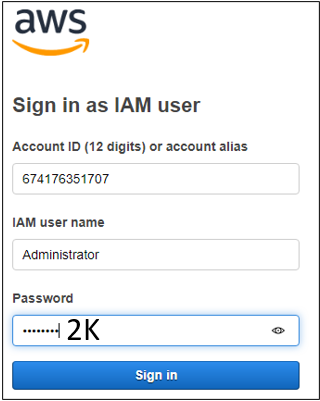
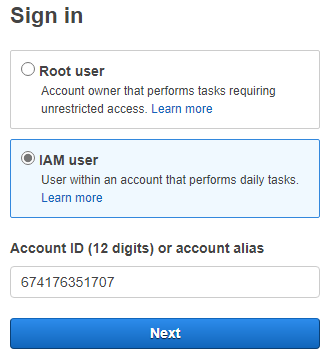








IAM User ID: 674176351707



\_\_\_END OF SECTION\_\_\_

## Tutorial: Analyzing Insights from Customer Reviews with Amazon Comprehend

<https://docs.aws.amazon.com/comprehend/latest/dg/tutorial-reviews.html>

### TL;DR

* Create the input CSV file as one document per line. No commas.
* Upload the CSV to *s3://asivapra-abs/input/*
* Launch ‘[Amazon Comprehend](https://ap-southeast-2.console.aws.amazon.com/comprehend/home?region=ap-southeast-2#welcome)’
  + [Analysis jobs](https://ap-southeast-2.console.aws.amazon.com/comprehend/v2/home?region=ap-southeast-2#analysis) – Opens [*Amazon Comprehend*](https://ap-southeast-2.console.aws.amazon.com/comprehend/v2/home?region=ap-southeast-2#welcome)*/*[*Analysis jobs*](https://ap-southeast-2.console.aws.amazon.com/comprehend/v2/home?region=ap-southeast-2#analysis)
    - Create Job
      * Job Settings
        + Name – a recognizable name that will appear in [*Amazon Comprehend*](https://ap-southeast-2.console.aws.amazon.com/comprehend/v2/home?region=ap-southeast-2#welcome)*/*[*Analysis jobs*](https://ap-southeast-2.console.aws.amazon.com/comprehend/v2/home?region=ap-southeast-2#analysis)
        + Analysis type

Sentiment

* + - * + Input data

My documents

S3 location

s3://asivapra-abs/input/inputfile.csv

Input format

One document per line

* + - * + Output data

S3 Location

s3://asivapra-abs/output/

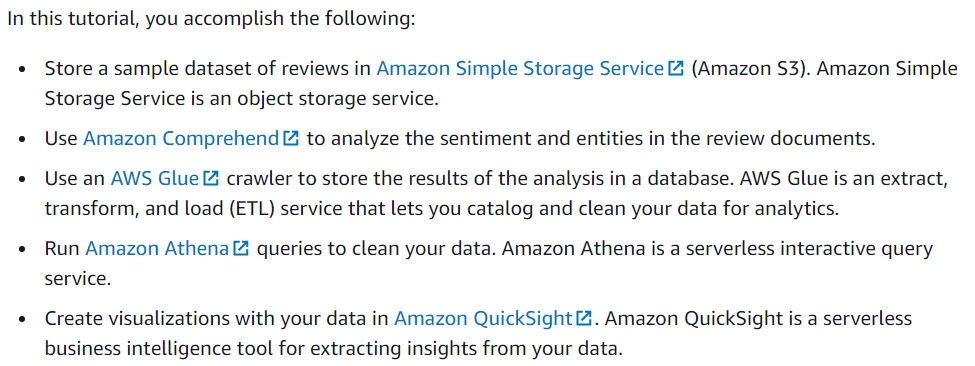
* + - * + Access permissions

Use an existing IAM role

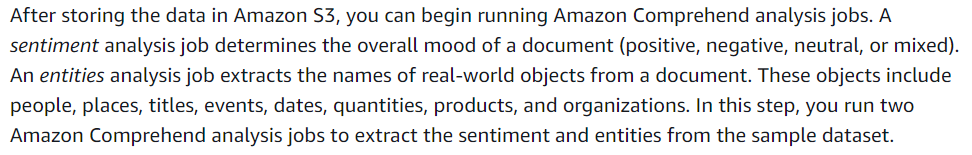


* + - * 
    - Wait 6-7 min for the job to finish

\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



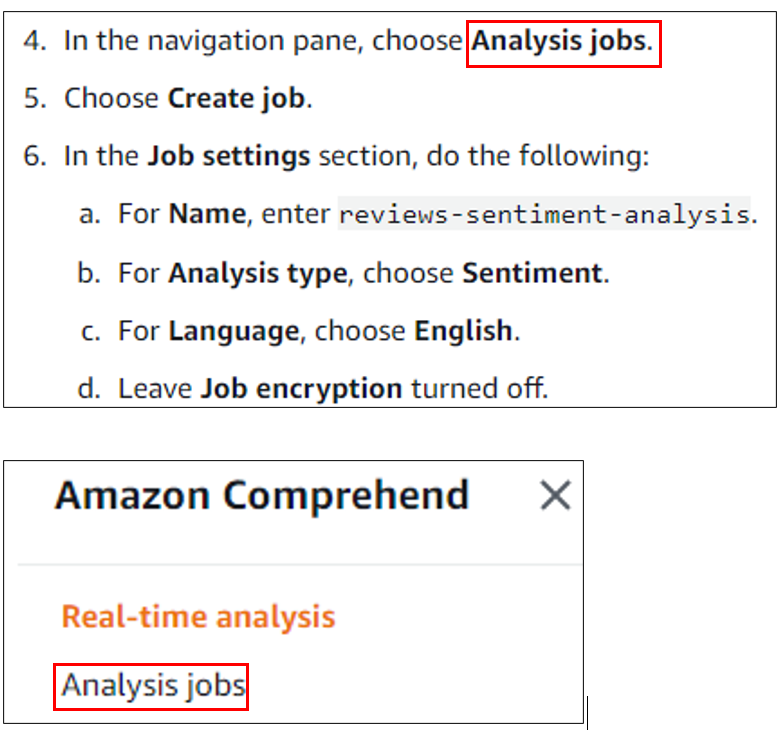
### Running Analysis Jobs on Documents in Amazon S3

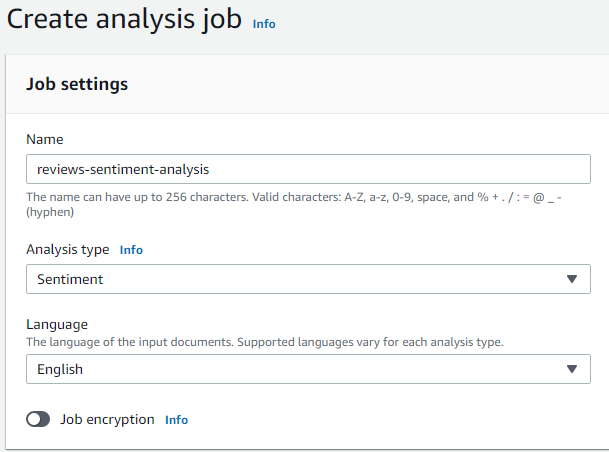


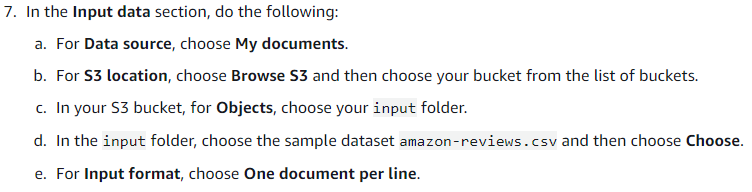
STEPS

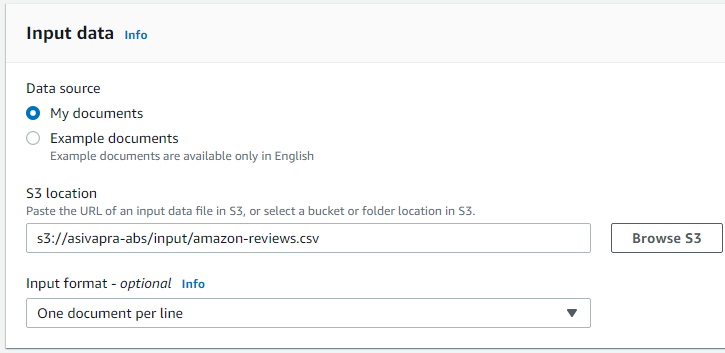
1. Open the Amazon Comprehend console at <https://console.aws.amazon.com/comprehend/>
2. Choose Launch Amazon Comprehend.

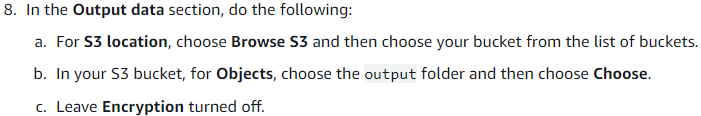


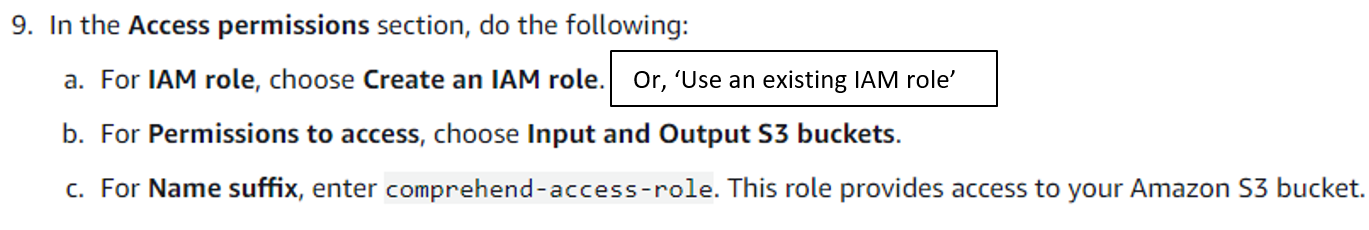


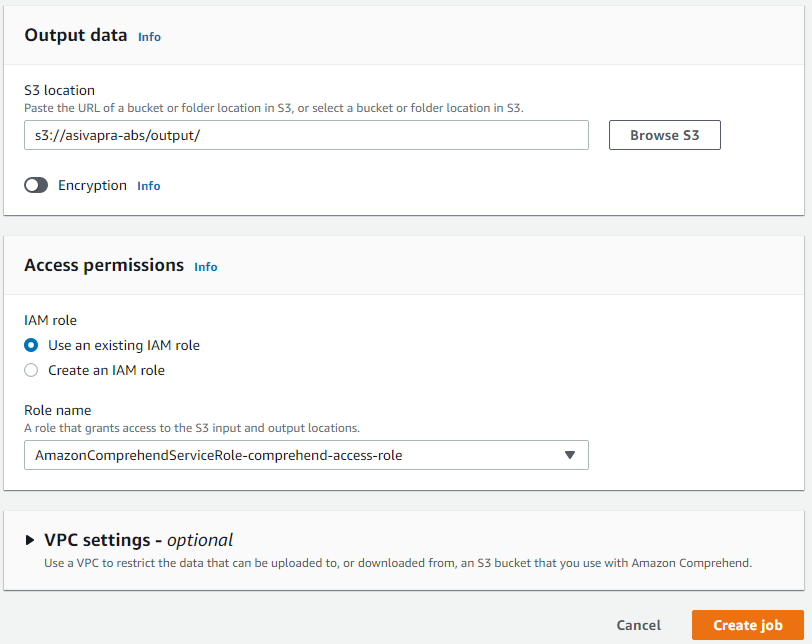


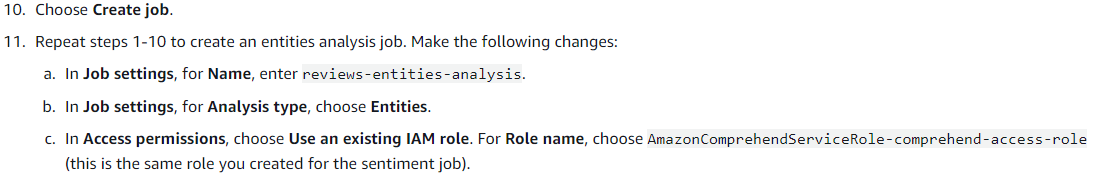


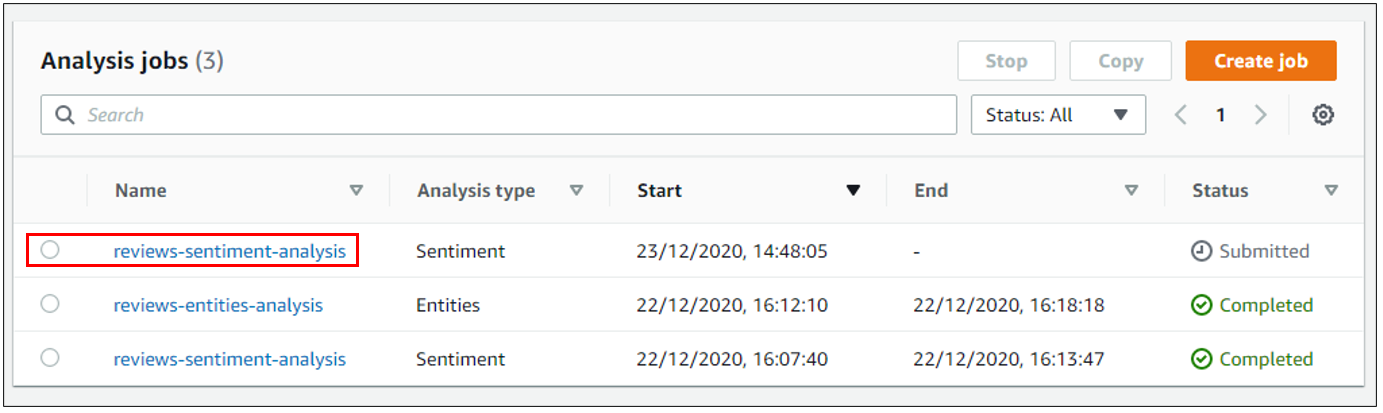












Status will change to ‘Completed’ when the job finishes. It takes several minutes.

\_\_\_END OF SECTION\_\_\_

## Tutorial: Custom entities analysis

### TL;DR

* Create a CSV file of text and type.
  + Multiple words in text column are permitted, but single word (or words connected with underscore) for type.
  + Type must be all upper case.
* Train the entity recogniser
* Use in [Analysis](#_Running_Analysis_Jobs) as before.
* [Download, extract and upload](#_Preparing_the_Amazon) the results
* Load the data into [AWS Glue](#_Load_the_Data) 
  + This will create a database table as e.g. avs\_674176351707\_ner\_f35b6c257ee4a2405f3cda251036a4c7
* View this table in [Athena](https://ap-southeast-2.console.aws.amazon.com/athena/home?region=ap-southeast-2#query) by the Database query:
  + SELECT \* FROM "comprehend-results"."avs\_674176351707\_ner\_f35b6c257ee4a2405f3cda251036a4c7";
    - (This will be listed under ‘Tables’)
  + Download as CSV

\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

To look for words and phrases that are specific for a company, it is necessary to train a model for custom entities. For example, to look for medicine brands in a list of product names, one must train a model using brand names. To look for profanity in user feedback, it is necessary to train a model using common words and phrases that are undesired.

### Train the recogniser

A recogniser is a list of text/type combination that will be used to analyse a document. This is a CSV document as below:

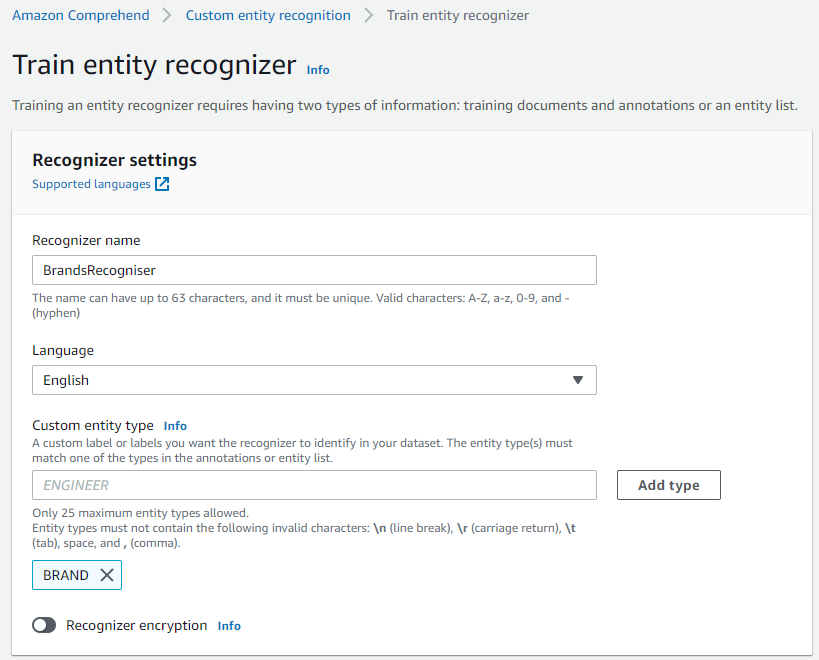
|  |  |
| --- | --- |
| Text | Type |
| Aerogard | BRAND |
| Paracetamol | MEDICINE |
| Aspirin | MEDICINE |
| Aspirin | BRAND |

Input data should not contain **duplicates**. Presence of duplicate samples might result into test set contamination and therefore negatively affect training process, model metrics, and behaviour.

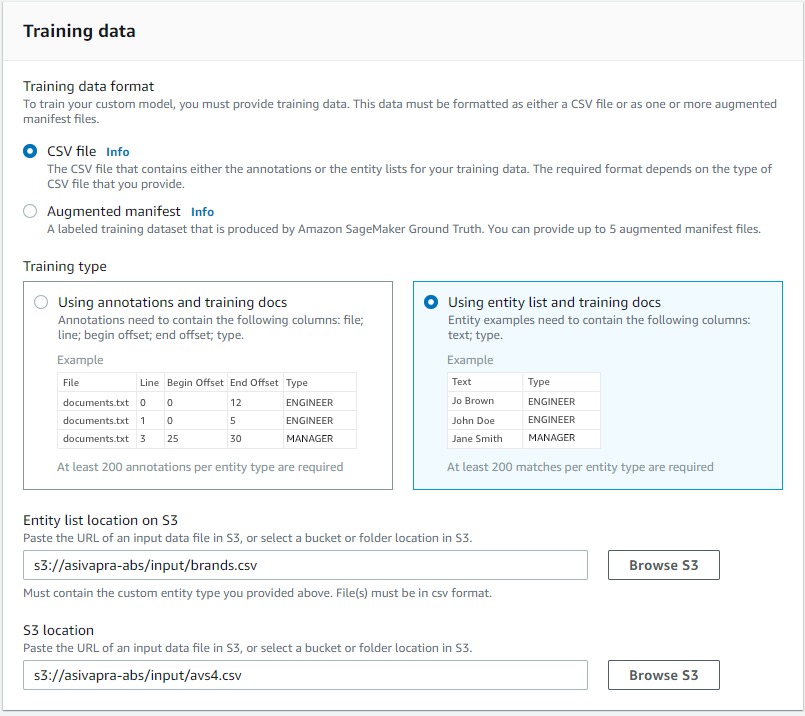
A maximum of 25 entity types can be used. Must have at least 200 text lines in the training doc.

### STEPS

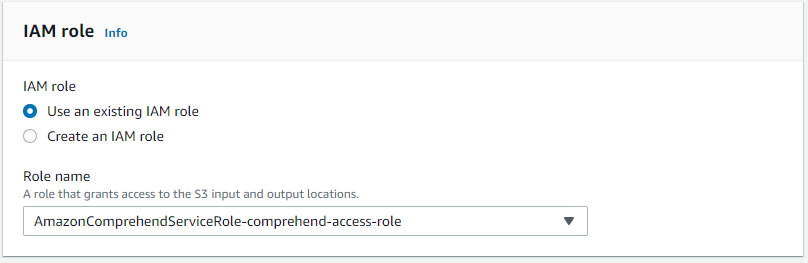
* [Train Entity Recogniser](https://ap-southeast-2.console.aws.amazon.com/comprehend/home?region=ap-southeast-2#create-custom-entity-recognizer)
* Give a unique name for ‘Recogniser name’
* Enter the type names (e.g. BRAND) IN ‘Custom entity type’. These must be all upper case. Underscores are allowed.



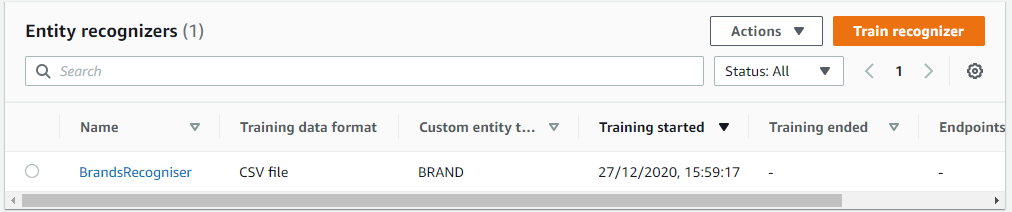
* Choose ‘CSV file’ as the training data format.
* Choose ‘Using entity list and training docs.
  + The entity list is the CSV file given earlier.
  + Training data is a list of text, one per line. e.g. product list



* Choose an IAM role and click ‘Train’







* Takes around 20 min to complete.

### Analyse the results

After the Entity Recogniser is ready, it can be used for analysing the result just as you would for ‘sentiment’ and ‘entities’ (See [details](#_Tutorial:_Analyzing_Insights) here)

### Prepare the output

To prepare the output, use [AWS Glue](#_Load_the_Data) as described below.

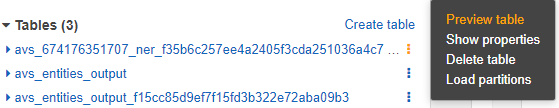
### View the output

There is no visual display using AWS QuickSight. Instead, run some queries in Athena and see which entities your custom annotator picks up.

**Query:**

* SELECT \* FROM "comprehend-results"."avs\_674176351707\_ner\_f35b6c257ee4a2405f3cda251036a4c7";

The table name is long and different each time. The above query can be pre-loaded by clicking the ‘Preview table’ link on the tables listed on the left. Then, remove the ‘limit 10’ from the end of the query to display the full table.



Download the output in a CSV file by clicking the first icon: 

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| col0 | col1 | col2 | col3 | col4 | col5 | col6 |
| {"Entities": [{"BeginOffset": 0 | "EndOffset": 6 | "Score": 0.9999970197767496 | "Text": "Palmer" | "Type": "BRAND"}] | "File": "avs4.csv" | "Line": 157} |
| {"Entities": [{"BeginOffset": 0 | "EndOffset": 4 | "Score": 0.9999983310727032 | "Text": "Olay" | "Type": "BRAND"}] | "File": "avs4.csv" | "Line": 181} |
| {"Entities": [{"BeginOffset": 0 | "EndOffset": 6 | "Score": 0.9999995231630692 | "Text": "Medela" | "Type": "BRAND"}] | "File": "avs4.csv" | "Line": 177} |

The above table may now be analysed with a custom Python program to find the word(s) in the original data file (avs4.csv) by following the BeginOffset, EndOffset, text and line number (col 0, 1, 3, 6) values.

### Observations

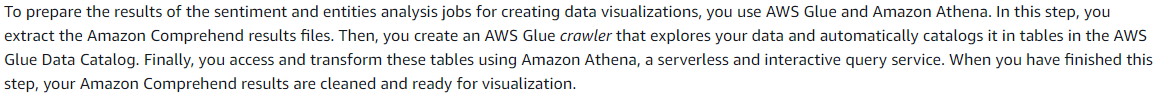
* Picks up more than one brand in a single line if present.
* Seems to omit the second word in the brand if it is a one letter. E.g. “Baby U” picks only “Baby”
* 2 and 3 letter words, if on their own, are picked up. E.g. 3M

## Preparing the Amazon Comprehend Output for Data Visualization

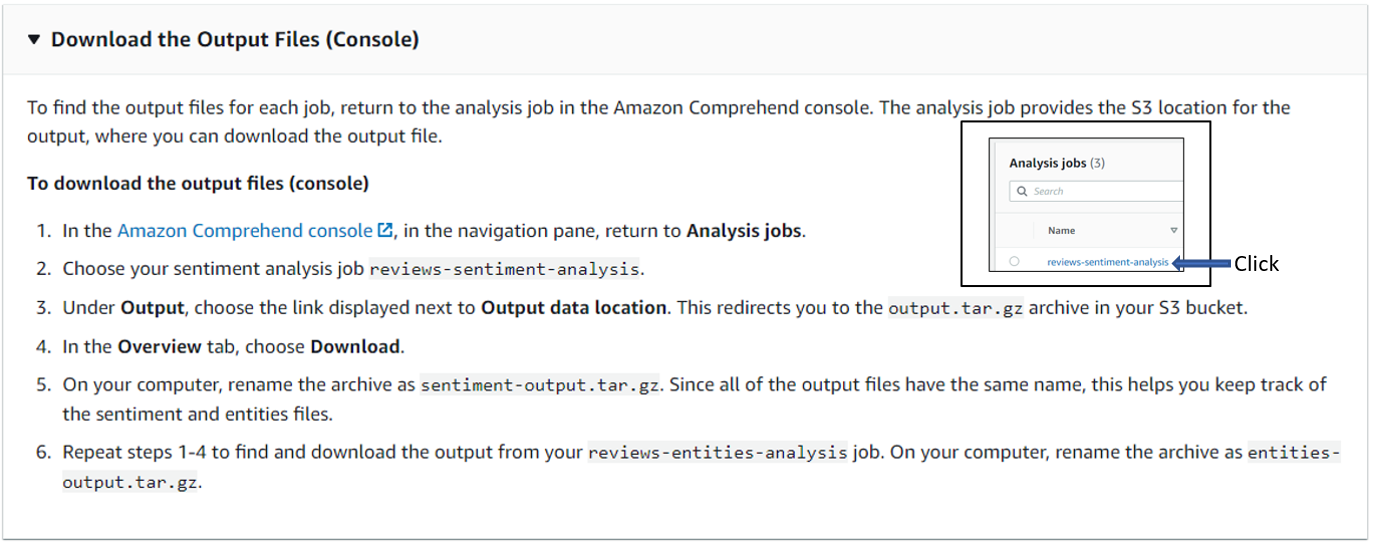
### TL;DR

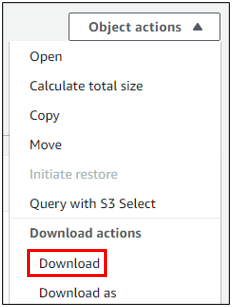
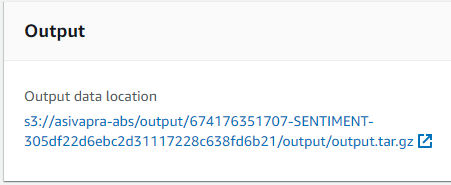
* Click the job link on ‘Amazon Comprehend/Analysis jobs’
* Click the link under ‘Output data location’
* Click ‘Object actions/ download’
* Rename the downloaded ‘output.tar.gz’ to something identifiable (e.g. avs\_output.tar.gz)
* Unzip the ‘output’ into the same directory and rename it as ‘sentiment-output’
  + The output file name is unchangeable. This requires more investigation.
* Upload to ‘sentiment-results’ (s3://asivapra-abs/sentiment-results/)
  + The folder name is unchangeable. This requires more investigation.

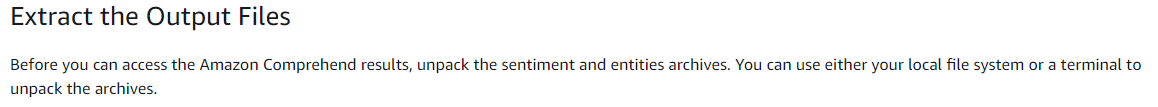
\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



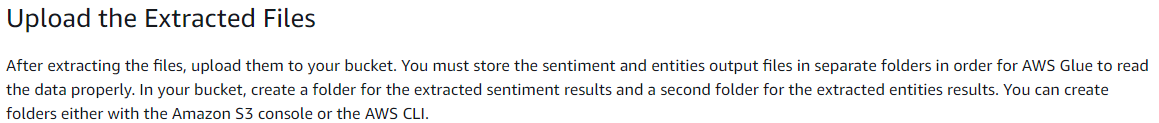
Open the Amazon Comprehend console at <https://console.aws.amazon.com/comprehend/>



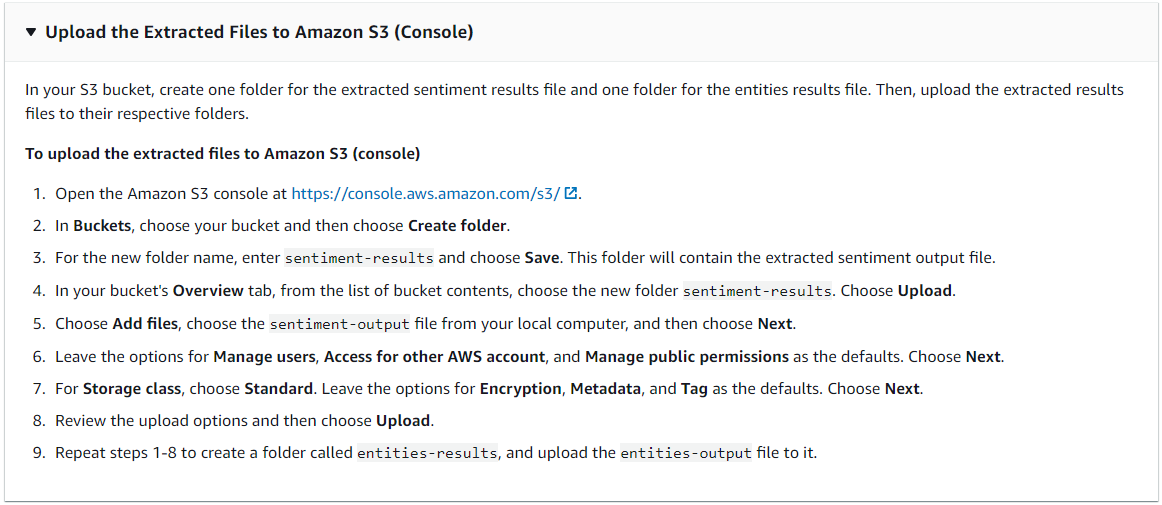




NOTE: The file will always be downloaded as ‘output.tar.gz’. Must rename it to something meaningful to avoid overwriting. Unzipping the \*.gz file is a hassle on the work PC until WinZip is installed by the service desk. The NGD is a virtual Windows within Windows and cannot access local disks including a USB. Hence, transferring files to and from other computers can only be done through S3 or email.



Note: The extracted files must be uploaded to S3 in specified folders. Given below are the folder names, which can be anything. Must remember the names to be given to the AWS programs that work on them.



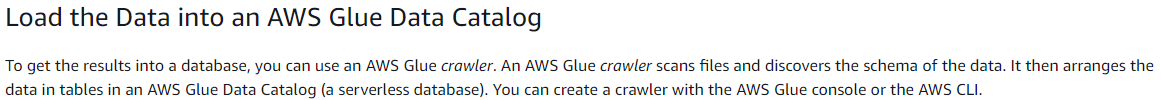
\_\_\_END of SECTION\_\_\_\_\_\_\_\_\_\_\_

## Load the Data into an AWS Glue Data Catalog

### TL;DR

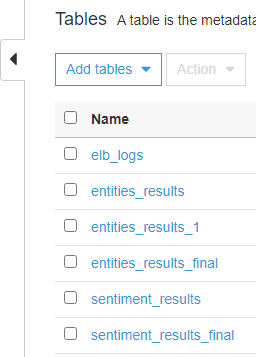
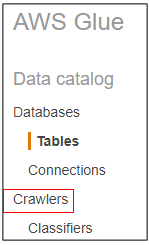
* Open AWS Glue (<https://ap-southeast-2.console.aws.amazon.com/glue/>)
* Crawlers | Add Crawler
* Or, click an existing crawler and Edit it.
  + Crawler name: an identifiable name
  + Crawler source type
    - Data store
  + Repeat crawls of S3 data stores
    - Crawl all folders
  + Choose a data store
    - S3
  + Connection
    - Skip it
  + Crawl data in
    - Specified path in my account
  + Include path
    - s3://asivapra-abs/sentiment-results/
    - The path name is unchangeable. This requires more investigation.
  + Add another data store
    - No
  + Choose an IAM role
    - Choose an existing IAM role
    - AWSGlueServiceRole-glue-access-role
  + Create a schedule for this crawler
    - Frequency
      * Run on demand
  + Configure the crawler's output
    - Database
      * Comprehend-results
  + Finish
  + Crawlers
    - Select the newly created crawler
    - Click ‘Run Crawler’
      * ~1 min to finish

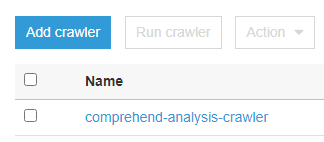
\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



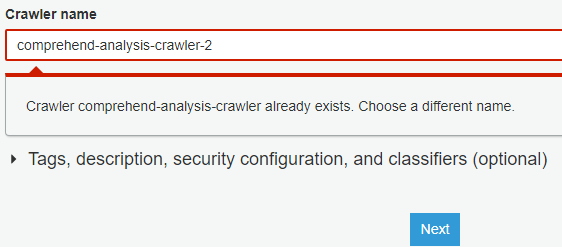
Open the AWS Glue console at <https://console.aws.amazon.com/glue/>.

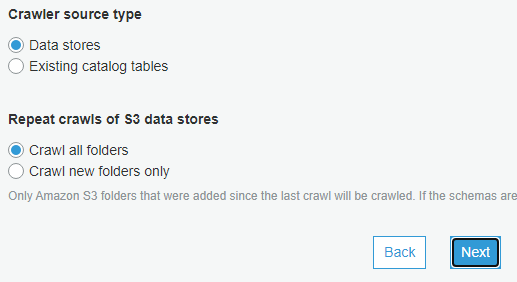




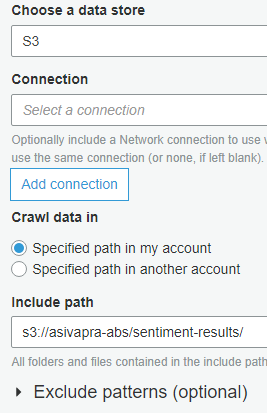
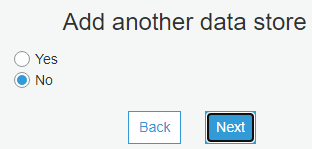




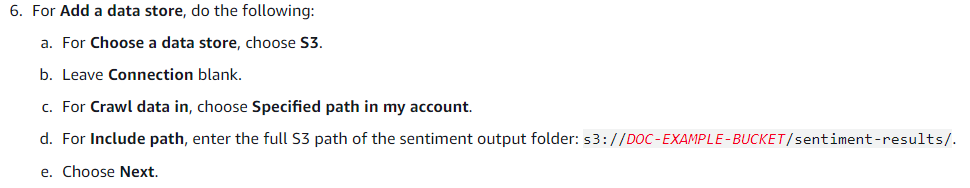






The ‘Include path’ must be named ‘sentiment-results’ for sentiment analysis. Also, the output in there must be called ‘sentiment-output’. This requires verification.



NOTES: For (d) above: The URI for the folder can be found as below.

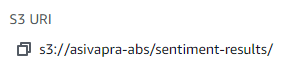
1. Access the S3 management console

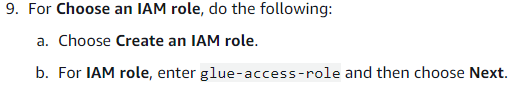
[S3 Management Console (amazon.com)](https://s3.console.aws.amazon.com/s3/buckets/asivapra-abs?region=ap-southeast-2&tab=objects)

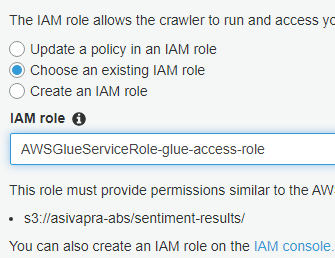
1. Click on the folder:



1. Copy the S3 URI

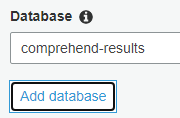
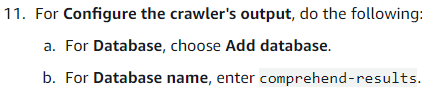




Note: Using an existing IAM role

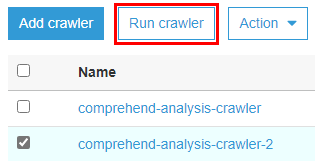


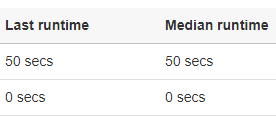
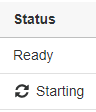




NOTE: Click ‘Add database’ if the database does not exist.



NOTE: It will take several minutes to finish.



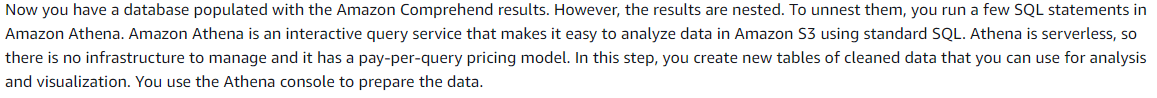
\_\_\_END OF SECTION\_\_\_

## Prepare the Data for Analysis

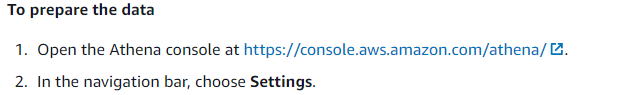
### TL;DR

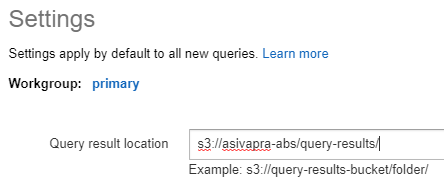
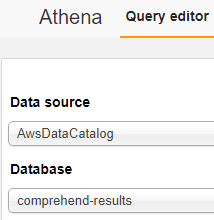
* Open the Athena console at [https://console.aws.amazon.com/athena/](https://console.aws.amazon.com/athena/home)
* Insert the SQL query:
  + - CREATE TABLE sentiment\_results\_final AS
    - SELECT file, line, sentiment,
    - sentimentscore.mixed AS mixed,
    - sentimentscore.negative AS negative,
    - sentimentscore.neutral AS neutral,
    - sentimentscore.positive AS positive
    - FROM avs\_sentiment\_results
* There should be two tables in the database now.
  + **Database**
    - Comprehend-results
      * avs\_sentiment\_results
      * sentiment\_results\_final

\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



* Open the Athena console at [https://console.aws.amazon.com/athena/](https://console.aws.amazon.com/athena/home).











NOTE: The table must not exit already.

CREATE TABLE sentiment\_results\_final AS

SELECT file, line, sentiment,

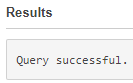
sentimentscore.mixed AS mixed,

sentimentscore.negative AS negative,

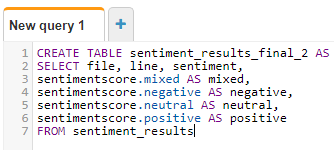
sentimentscore.neutral AS neutral,

sentimentscore.positive AS positive

FROM sentiment\_results









CREATE TABLE entities\_results\_1 AS

SELECT file, line, nested FROM entities\_results

CROSS JOIN UNNEST(entities) as t(nested)



CREATE TABLE entities\_results\_final\_2 AS

SELECT file, line,

nested.beginoffset AS beginoffset,

nested.endoffset AS endoffset,

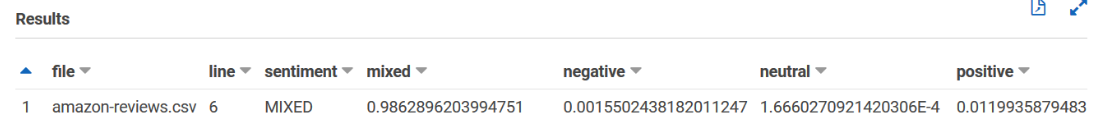
nested.score AS score,

nested.text AS entity,

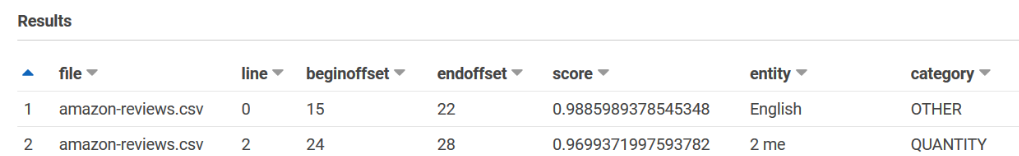
nested.type AS category

FROM entities\_results\_1

Your **sentiment\_results\_final** table should look like the following:



Your **entities\_results\_final** table should look like the following:



\_\_\_END OF SECTION\_\_\_

## Visualizing Amazon Comprehend Output in Amazon QuickSight

### TL;DR

* + Start [QuickSight](•%09https:/ap-southeast-2.quicksight.aws.amazon.com/sn/start/analyses)
  + Datasets
    - New Dataset
    - Athena
      * sentiment\_results\_final
    - Create datasource
      * Database
        + sentiment\_results\_final
        + Import to SPICE for quicker analysis

Visialise

Sentiment

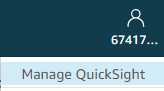
Pie Chart

\_\_\_\_End of TL;DR\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

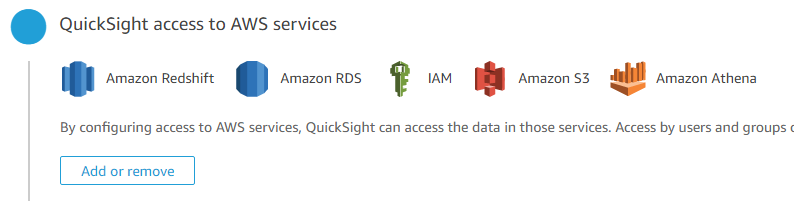
After storing the Amazon Comprehend results in tables, you can connect to and visualize the data with Amazon QuickSight.

### Give Amazon QuickSight Access

1. Open the [Amazon QuickSight console](https://quicksight.aws.amazon.com/sn/start).
2. Choose your profile name in the navigation bar and choose **Manage QuickSight**.

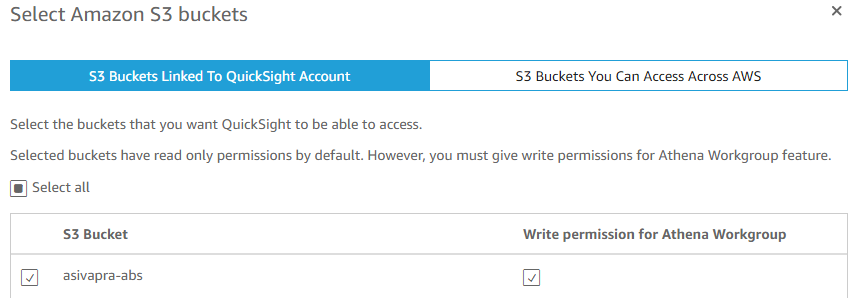


1. Choose Security and permissions.



1. Choose Amazon S3.



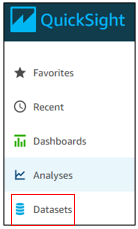




### Import the Datasets

Before creating visualizations, you must add the sentiment and entities datasets to Amazon QuickSight.

1. Open the [Amazon QuickSight console](https://quicksight.aws.amazon.com/sn/start).

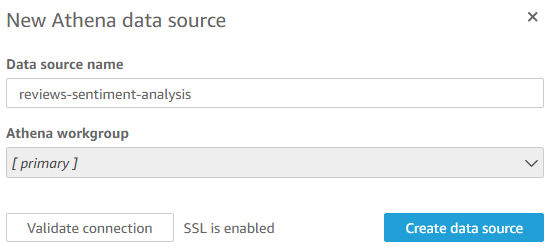
In the navigation bar, in **Datasets**, choose **New dataset**.



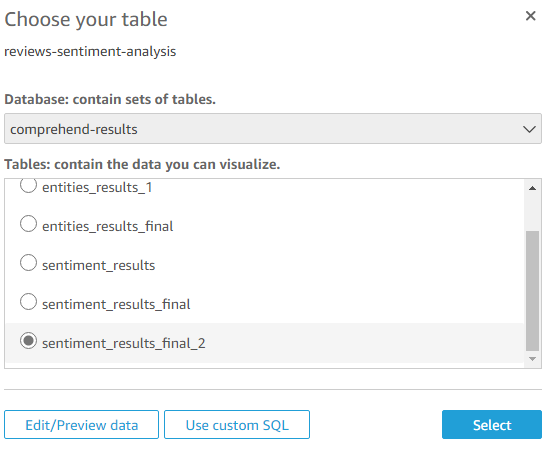
For **Create a Data Set**, choose **Athena**.



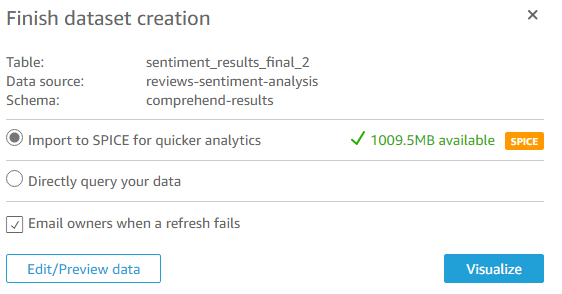
1. For Data source name, enter reviews-sentiment-analysis and choose Create data source.

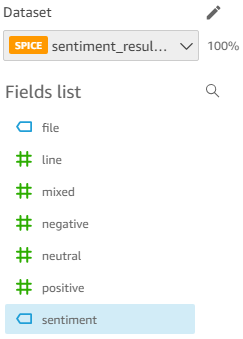


1. For Database, choose the database **comprehend-results**.
2. For **Tables**, choose the sentiment table sentiment\_results\_final\_2 and then choose **Select**.

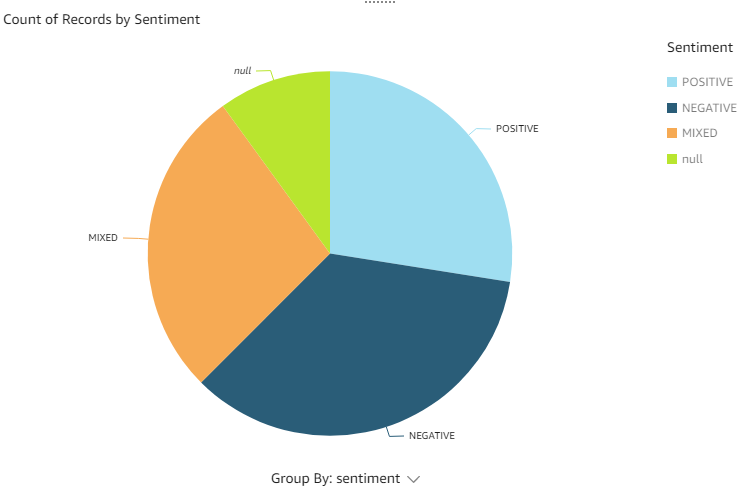


1. Choose Import to SPICE for quicker analytics and choose Visualize.



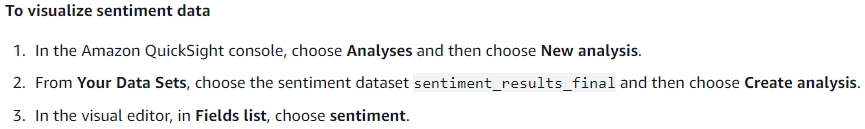


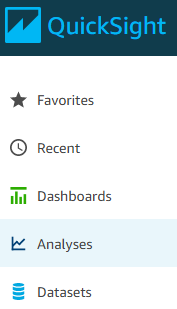


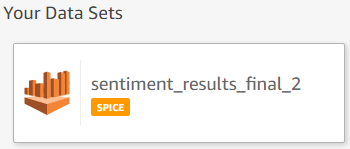


### Create a Sentiment Visualization

Once the above steps are completed, it is possible to do new analyses directly as below.

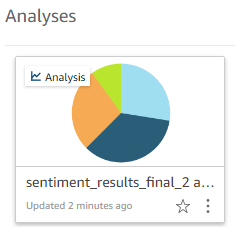




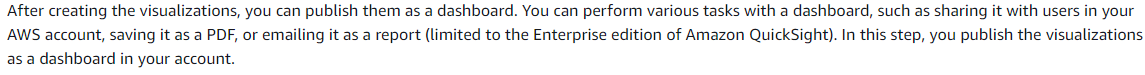




Or click on the displayed charts for a re-display.

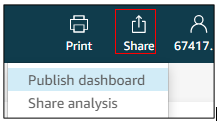


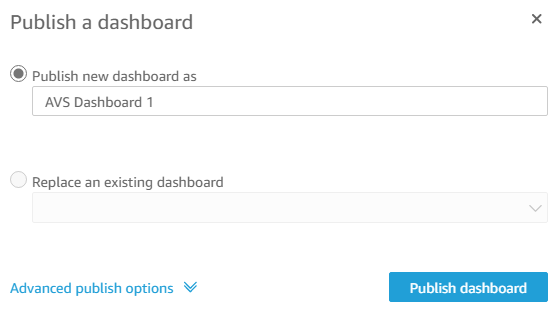
### Publish a Dashboard



**To publish your dashboard**

1. In the navigation bar, choose **Share**.





NOTE: The publishing does not seem to work as described. Maybe missing something.