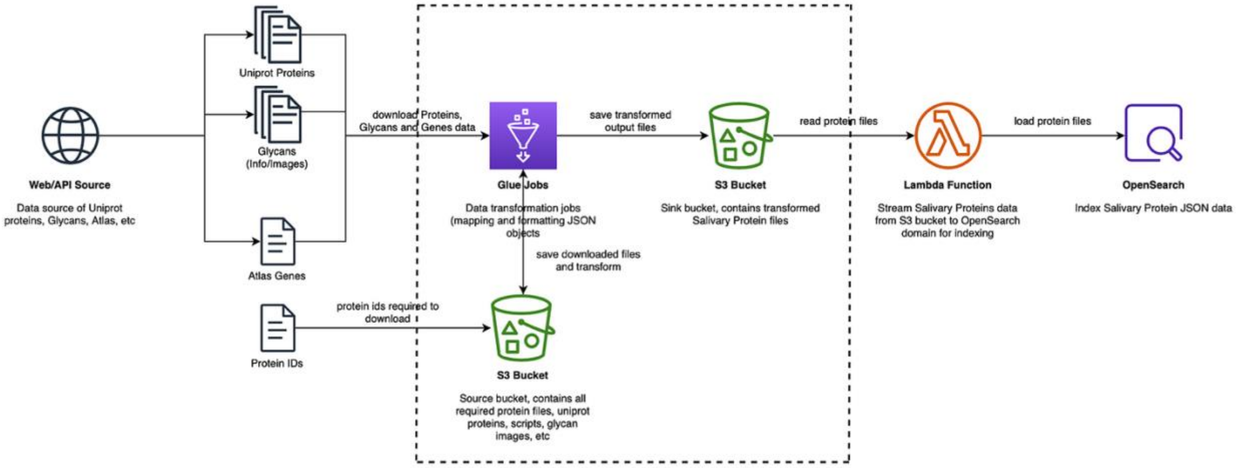
Salivary Proteins Data Pipeline using AWS Glue & OpenSearch

Documentation

AWS Architecture

The cloud architecture of Amazon Web Services utilizes several Glue Python jobs to perform various tasks. These tasks include fetching protein and glycan data along with glycan images from the Uniprot and Glygen websites, retrieving Human Atlas data through an API, and storing all of this information in S3 storage buckets. These files are subsequently employed for additional processing, where necessary salivary protein data is extracted using Glue’s transformation capabilities. Lastly, the obtained salivary protein data is sent to OpenSearch using a Lambda function to facilitate the indexing process.



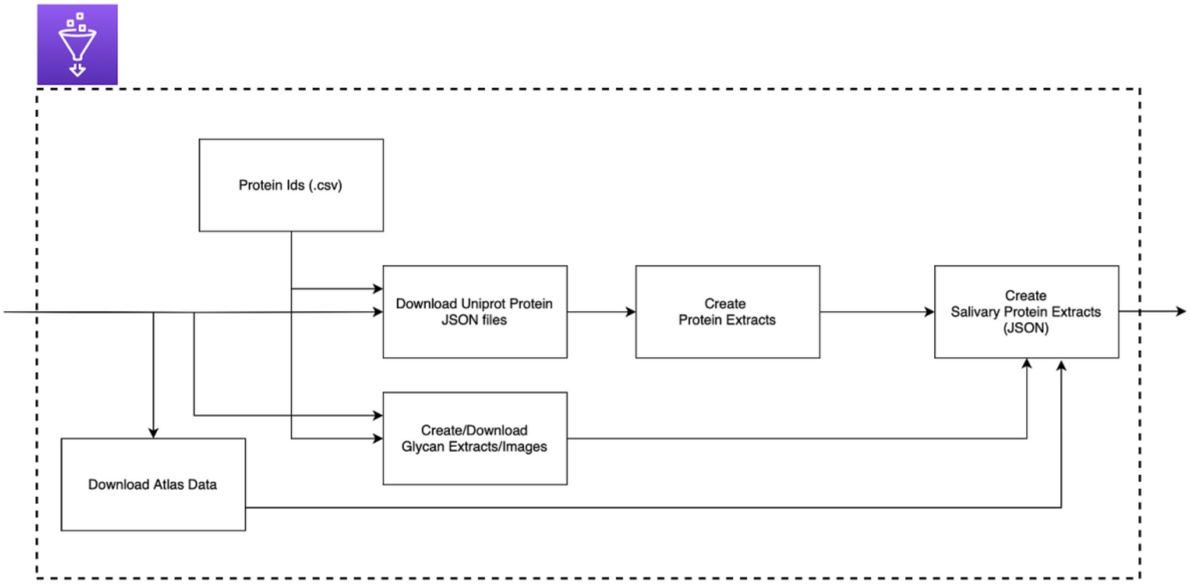
Uniprot Proteins
Glycans
(Info/Images)
Atlas Genes
Protein IDs
Web/API Source
Data source of Uniprot proteins, Glycans, Atlas, etc
download Proteins, Glycans and Genes data
protein ids required to download
Glue Jobs
Data transformation jobs (mapping and formatting JSON objects
save downloaded files and transform
S3 Bucket
Source bucket, contains all required protein files, uniprot proteins, scripts, glycan images, etc
save transformed output files
S3 Bucket
Sink bucket, contains transformed Salivary Protein files
read protein files
Lambda Function
Stream Salivary Proteins data from S3 bucket to OpenSearch domain for indexing
load protein files
OpenSearch
Index Salivary Protein JSON data

Salivary Proteins Data Pipeline—AWS Glue Workflow

AWS Glue Jobs workflow

AWS Glue Jobs simplifies the process of data transformation and ETL by automating many of the tasks involved, allowing users to focus on defining data transformations and extracting insights from their data.

In this workflow, once the pipeline is triggered, three jobs `download-create-protein-extracts, create-glycan extracts, download-glycan-images` run parallelly and once all are completed, the final job `create-salivary-protein-extracts` runs in the end to create salivary proteins JSON files.


Protein Ids (.csv)
Download Uniprot Protein JSON files
Create Protein Extracts
Create Salivary Protein Extracts (JSON)
Create/Download Glycan Extracts/Images
Download Atlas Data

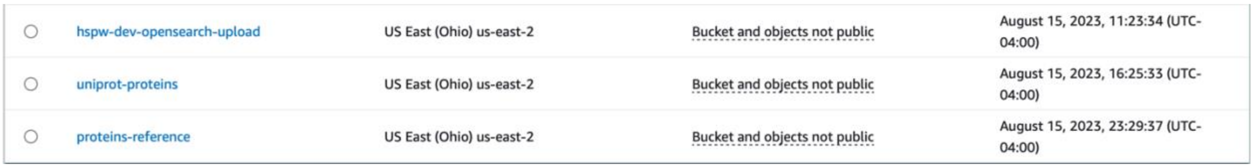
Salivary Proteins Data Pipeline—AWS Glue Jobs Workflow

Prepare AWS environment

Please note, the IAM roles and policies are managed by the AWS Administrator.

1. Create S3 buckets

Amazon Simple Storage Service (Amazon S3) is a scalable object storage service offered by Amazon Web Services (AWS). S3 provides secure, durable, and highly available storage for various types of data, making it suitable for a wide range of use cases.



○	hspw-dev-opensearch-upload	US East (Ohio) us-east-2	Bucket and objects not public	August 15, 2023, 11:23:34 (UTC-04:00)
○	uniprot-proteins	US East (Ohio) us-east-2	Bucket and objects not public	August 15, 2023, 16:25:33 (UTC-04:00)
○	proteins-reference	US East (Ohio) us-east-2	Bucket and objects not public	August 15, 2023, 23:29:37 (UTC-04:00)

S3 buckets—Ohio us-east-2 region

proteins-reference

In this bucket, we will keep all our reference files used in the data pipeline—

`protein_ids.csv`—All *protein ids* we need to process.
`rna_tissue_consensus.tsv`—Atlas *RNA tissue* reference data.
`normal_tissue.tsv`—Atlas *normal tissue* reference data.
`scripts/`—Folder to save all Glue Python job scripts.

uniprot-proteins

Here, all our protein files will be saved—

`uniprot_protein_files/`—Folder to download all uniprot protein files from uniprot website by Glue job `download-create-protein-extracts`.

`protein_extracts/`—Folder to save all protein extracts after downloading and transforming by Glue job `download-create-protein-extracts/create-protein-extracts`.

`glycan_extracts/`—Folder to save all glycan extracts after transforming glycan data by Glue job `create-glycan-extracts`.

`images/`—Folder to save all glycan images, downloading through glygen web API by Glue job `download-glycan-images`.

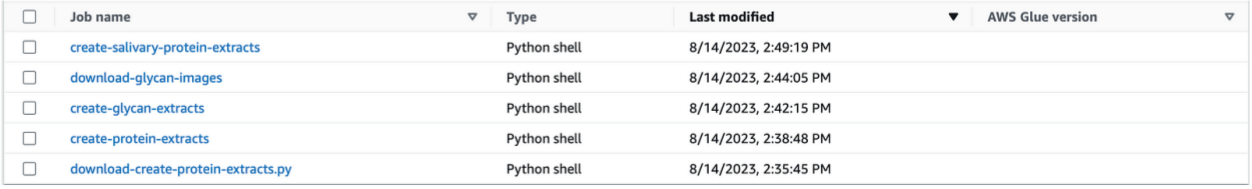
hspw-dev-opensearch-upload

This bucket contains all salivary protein extracts (output files from the data pipeline) which are ready to index in AWS OpenSearch.

`salivary-protein-extracts/`—Folder to save all salivary protein extracts created by job `create-salivary-protein-extracts.`

2. Create Glue Python jobs

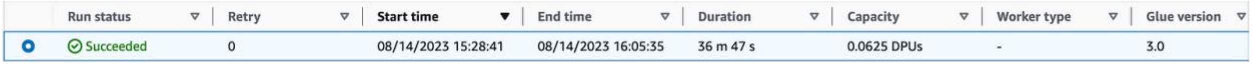
All Python scripts—



□	Job name	Type	Last modified	AWS Glue version
□	create-salivary-protein-extracts	Python shell	8/14/2023, 2:49:19 PM	
□	download-glycan-images	Python shell	8/14/2023, 2:44:05 PM	
□	create-glycan-extracts	Python shell	8/14/2023, 2:42:15 PM	
□	create-protein-extracts	Python shell	8/14/2023, 2:38:48 PM	
□	download-create-protein-extracts.py	Python shell	8/14/2023, 2:35:45 PM	

AWS Glue Python Jobs

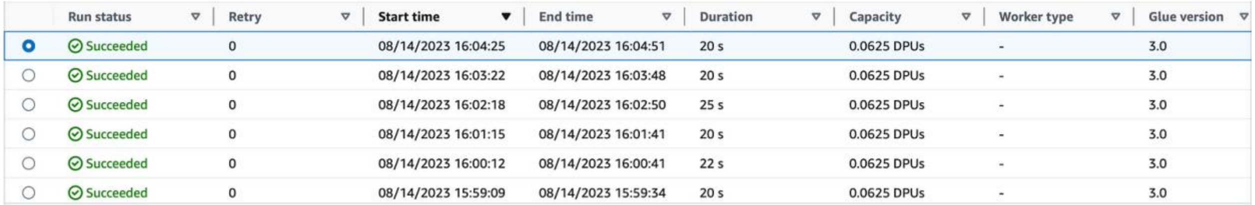
`download-create-protein-extracts` — Job to retrieve *protein ids* from the S3 location `proteins-reference/protein_ids.csv`, simultaneously download corresponding UniProt protein files from the UniProt website using multi-threading and batch processing to expedite the procedure. After each batch, initiate a subsequent task named `create-protein-extracts` passing *batch id* and *protein ids* as job parameters. This task involves generating protein extracts and storing them within the `uniprot-proteins/uniprot_protein_files/` directory. The process incorporates a retry mechanism that reschedules the job at a consistent interval to manage potential errors caused by concurrent job executions.



	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
●	✓ Succeeded	0	08/14/2023 15:28:41	08/14/2023 16:05:35	36 m 47 s	0.0625 DPUs	-	3.0

download-create-protein-extracts—Duration ~ 30 m

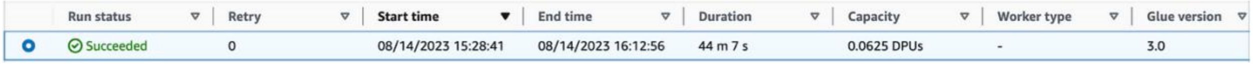
`create-protein-extracts` — This job is triggered by `download-create-protein-extracts` and takes protein ids as job parameters. It creates protein extracts using required protein data and subsequently stores them in the `uniprot-proteins/protein_extracts/` directory.



	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
●	✓ Succeeded	0	08/14/2023 16:04:25	08/14/2023 16:04:51	20 s	0.0625 DPUs	-	3.0
○	✓ Succeeded	0	08/14/2023 16:03:22	08/14/2023 16:03:48	20 s	0.0625 DPUs	-	3.0
○	✓ Succeeded	0	08/14/2023 16:02:18	08/14/2023 16:02:50	25 s	0.0625 DPUs	-	3.0
○	✓ Succeeded	0	08/14/2023 16:01:15	08/14/2023 16:01:41	20 s	0.0625 DPUs	-	3.0
○	✓ Succeeded	0	08/14/2023 16:00:12	08/14/2023 16:00:41	22 s	0.0625 DPUs	-	3.0
○	✓ Succeeded	0	08/14/2023 15:59:09	08/14/2023 15:59:34	20 s	0.0625 DPUs	-	3.0

create-protein-extracts—batch runs—Duration ~ 20 s (30 m)

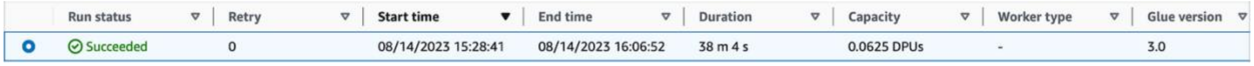
`create-glycan-extracts` — This job involves reading *protein ids*, retrieving glycan data from the glycogen API, and then generating glycan extracts using the necessary glycan information. These extracts are subsequently stored in the `uniprot-proteins/glycan_extracts/` directory.



	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
●	Succeeded	0	08/14/2023 15:28:41	08/14/2023 16:12:56	44 m 7 s	0.0625 DPUs	-	3.0

create-glycan-extracts—Duration ~ 40 m

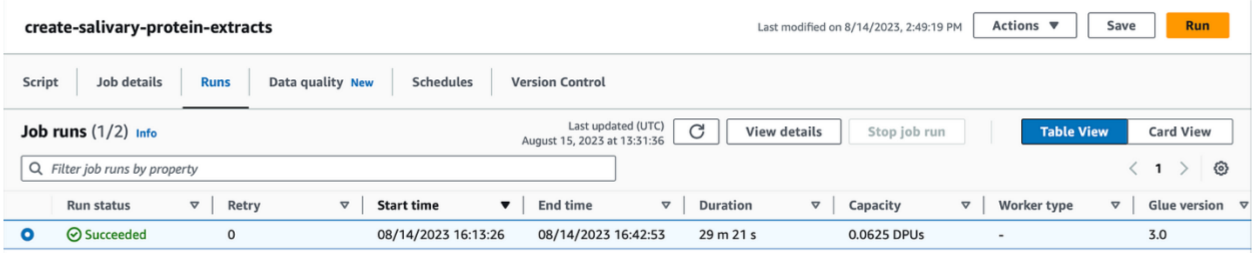
`download-glycan-images`—Job to download all glycan images and save them to `uniprot-proteins/images/` directory reading *protein ids* list.



	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
●	Succeeded	0	08/14/2023 15:28:41	08/14/2023 16:06:52	38 m 4 s	0.0625 DPUs	-	3.0

download-glycan-image—Duration ~ 30 m

`create-salivary-protein-extracts`—The final job is to read the list of *protein ids* and create salivary protein extracts. These extracts encompass essential information obtained from both protein and glycan extracts. Subsequently, these files are saved within the `hspw-dev-opensearch-upload/salivary-protein-extracts/` directory. This action will then activate a Lambda function to facilitate their transfer to the OpenSearch domain index.

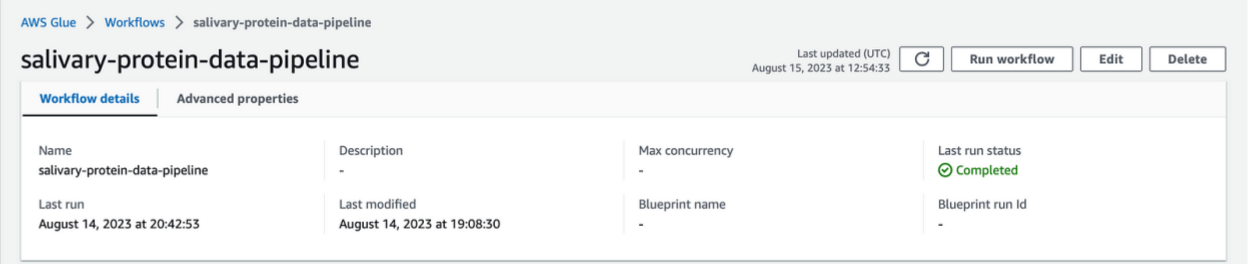

create-salivary-protein-extracts
Last modified on 8/14/2023, 2:49:19 PM
Actions ▼
Save
Run
Script
Job details
Runs
Data quality New
Schedules
Version Control
Job runs (1/2) Info
Last updated (UTC) August 15, 2023 at 13:31:36
View details
Stop job run
Table View
Card View
Filter job runs by property
1

	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue version
●	Succeeded	0	08/14/2023 16:13:26	08/14/2023 16:42:53	29 m 21 s	0.0625 DPUs	-	3.0

create-salivary-protein-extracts—Duration ~ 30 m

3. Create a Glue workflow

Finally, create the Glue workflow to build the data pipeline and execute all jobs with on-demand/scheduled triggers.

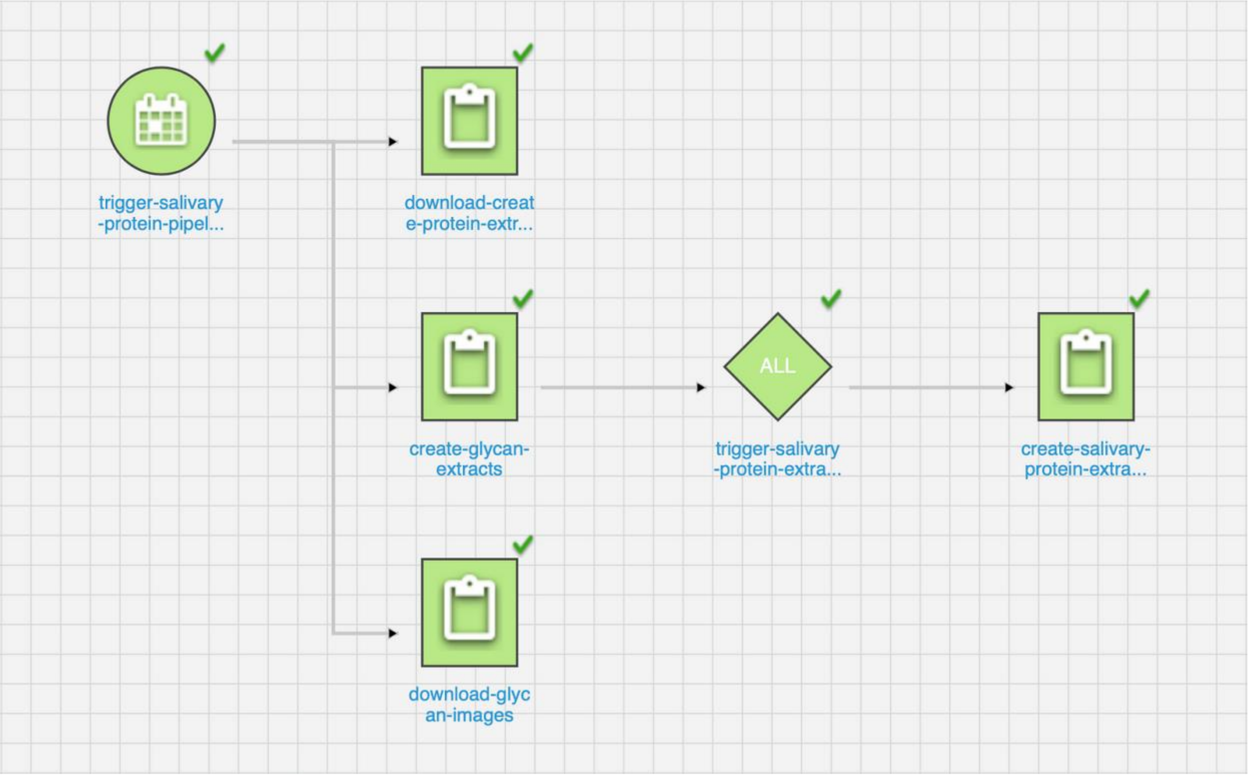


salivary-protein-data-pipeline

Last updated (UTC)
August 15, 2023 at 12:54:33

Name salivary-protein-data-pipeline	Description -	Max concurrency -	Last run status Completed
Last run August 14, 2023 at 20:42:53	Last modified August 14, 2023 at 19:08:30	Blueprint name -	Blueprint run Id -

An AWS Glue workflow is a sequence of interconnected Glue jobs and triggers that automate and orchestrate data processing tasks. This workflow simplifies the process of managing, monitoring, and executing various ETL (Extract, Transform, Load) tasks across your data sources.

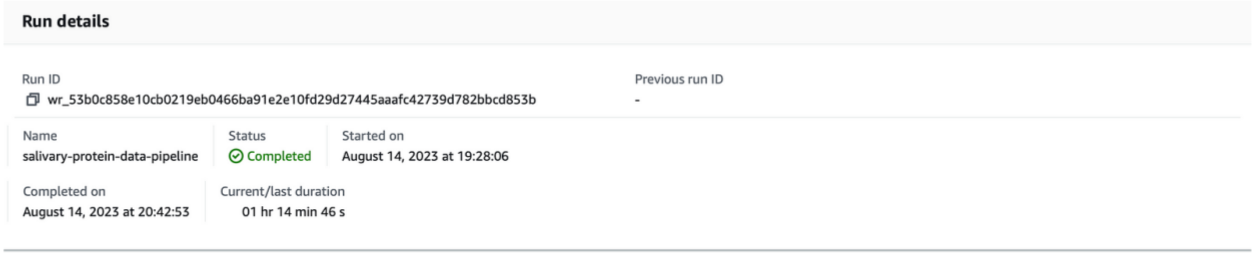


AWS Glue Workflow

`salivary-protein-data-pipeline`—Glue workflow to download and create protein extracts, glycan extracts and images, and salivary protein extracts.

`trigger-salivary-protein-pipeline`—Trigger `salivary-protein-data-pipeline` scheduled monthly on 1st day of every month. It also triggers other parallel jobs—`download-create-protein-extracts, create-glycan extracts,` and `download-glycan-images.`

`trigger-salivary-protein-extracts`—Trigger `create-salivary-protein-extracts` job, after all previously triggered jobs are executed successfully.



Run details

Run ID
wr_53b0c858e10cb0219eb0466ba91e2e10fd29d27445aaafc42739d782bbcd853b

Previous run ID
-

Name
salivary-protein-data-pipeline

Status
Completed

Started on
August 14, 2023 at 19:28:06

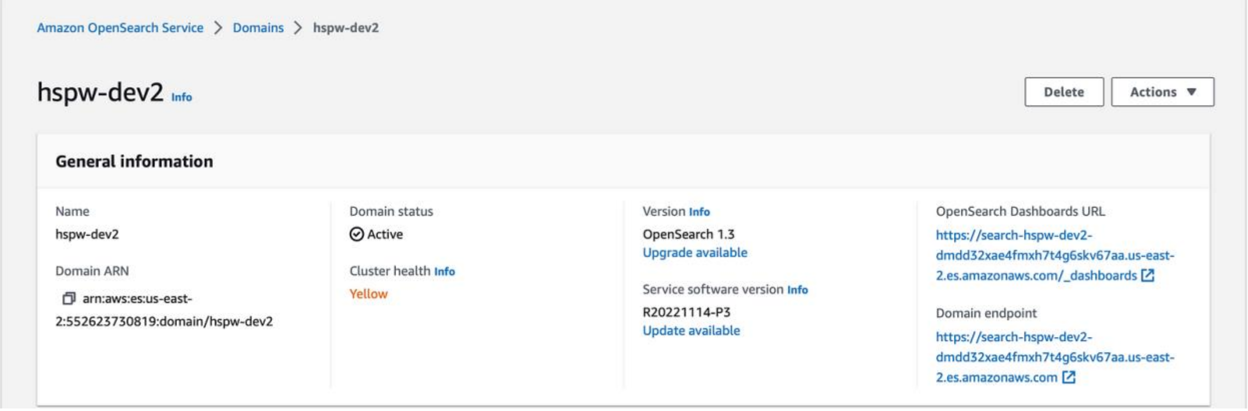
Completed on
August 14, 2023 at 20:42:53

Current/last duration
01 hr 14 min 46 s

Workflow Run Details—Duration ~ 1 hr

4. Create an OpenSearch domain

Amazon OpenSearch (formerly known as Amazon Elasticsearch Service) is a managed service offered by Amazon Web Services (AWS) that provides a scalable and reliable solution for deploying and operating the open-source Elasticsearch and Kibana software for search and analytics.



hspw-dev2 Info

General information

Name
hspw-dev2

Domain ARN
arn:aws:es:us-east-2:552623730819:domain/hspw-dev2

Domain status
Active

Cluster health Info
Yellow

Version Info
OpenSearch 1.3
Upgrade available

Service software version Info
R20221114-P3
Update available

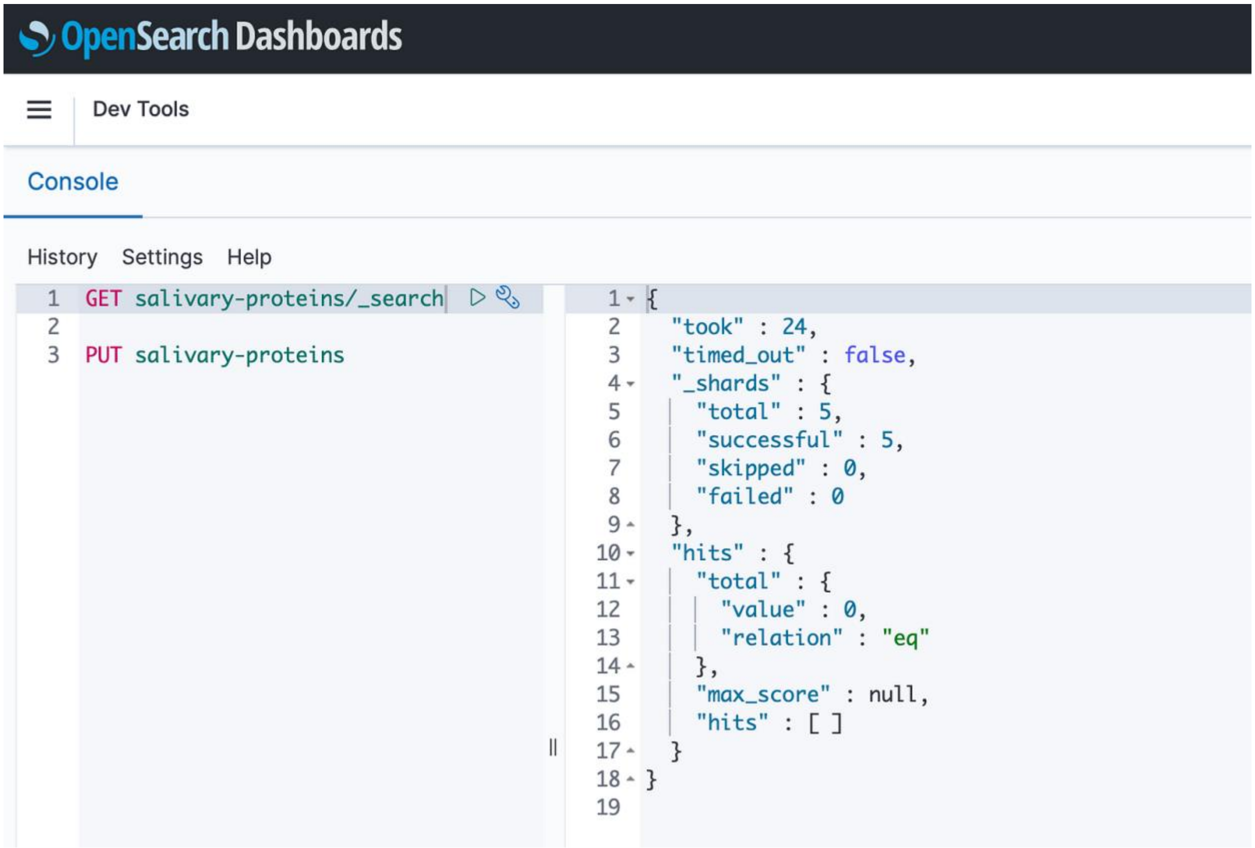
OpenSearch Dashboards URL
https://search-hspw-dev2-dmdd32xae4fmxh7t4g6skv67aa.us-east-2.es.amazonaws.com/_dashboards

Domain endpoint
https://search-hspw-dev2-dmdd32xae4fmxh7t4g6skv67aa.us-east-2.es.amazonaws.com

AWS OpenSearch Domain

OpenSearch domain indexes all salivary protein data and make it available end-to-end for search queries. In our case, accessible by the website.

`hspw-dev2`—OpenSearch service domain to ingest salivary protein data from the S3 bucket and index them.



```
GET salivary-proteins/_search

PUT salivary-proteins
```

```
{
  "took" : 24,
  "timed_out" : false,
  "_shards" : {
    "total" : 5,
    "successful" : 5,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 0,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  }
}
```

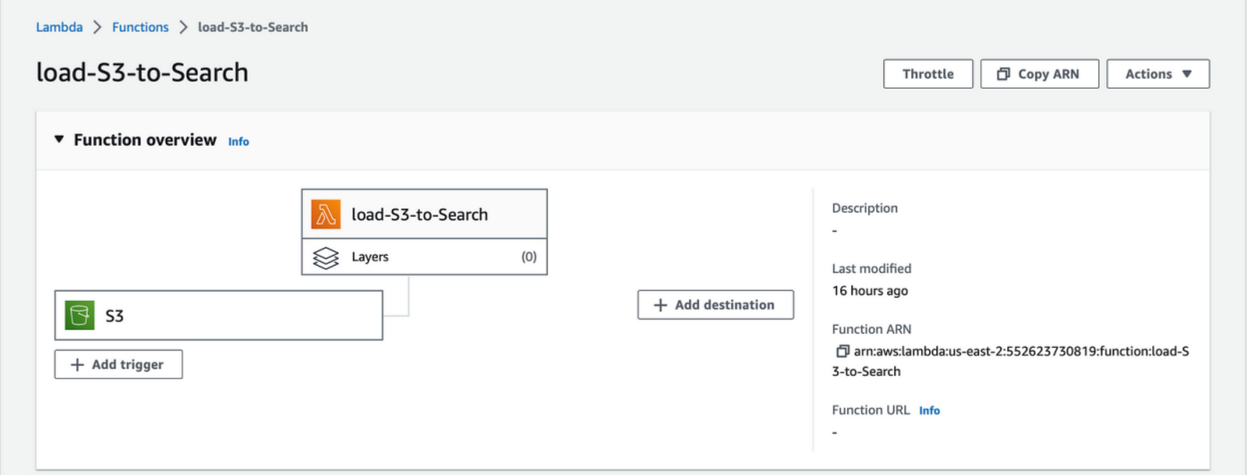
OpenSearch Dashboard

`salivary-proteins`—OpenSearch index which contains all salivary protein data.

5. Deploy a Lambda function

Now, we need a Lambda function to transfer salivary protein data from the S3 bucket to AWS OpenSearch.

AWS Lambda is a serverless computing service provided by Amazon Web Services (AWS) that allows you to run code without provisioning or managing servers. With Lambda, you can execute your code in response to various events and triggers, enabling you to build scalable and cost-effective applications.



load-S3-to-Search

Description
-

Last modified
16 hours ago

Function ARN
arn:aws:lambda:us-east-2:552623730819:function:load-S3-to-Search

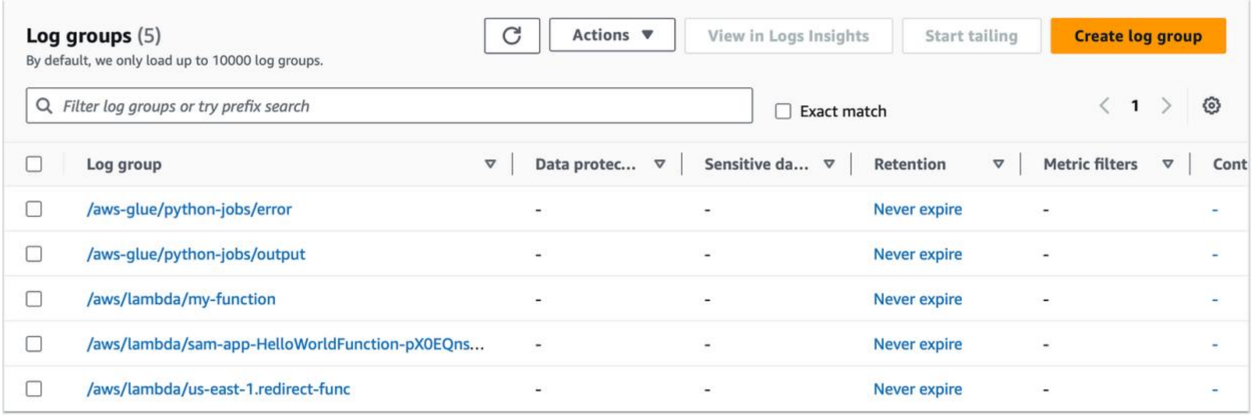
Function URL Info
-

AWS Lambda Function

`load-S3-to-Search`—Lambda function is triggered when any salivary protein extract file is uploaded to the S3 location `hspw-dev-opensearch-upload/salivary-protein-extracts/` which then sends data to OpenSearch index `salivary-proteins`.

CloudWatch Logs

Amazon CloudWatch Logs is a service provided by Amazon Web Services (AWS) that allows you to monitor, store, and analyze log data from various AWS resources and applications. It enables you to collect and centralize log data for better visibility, troubleshooting, and compliance purposes.

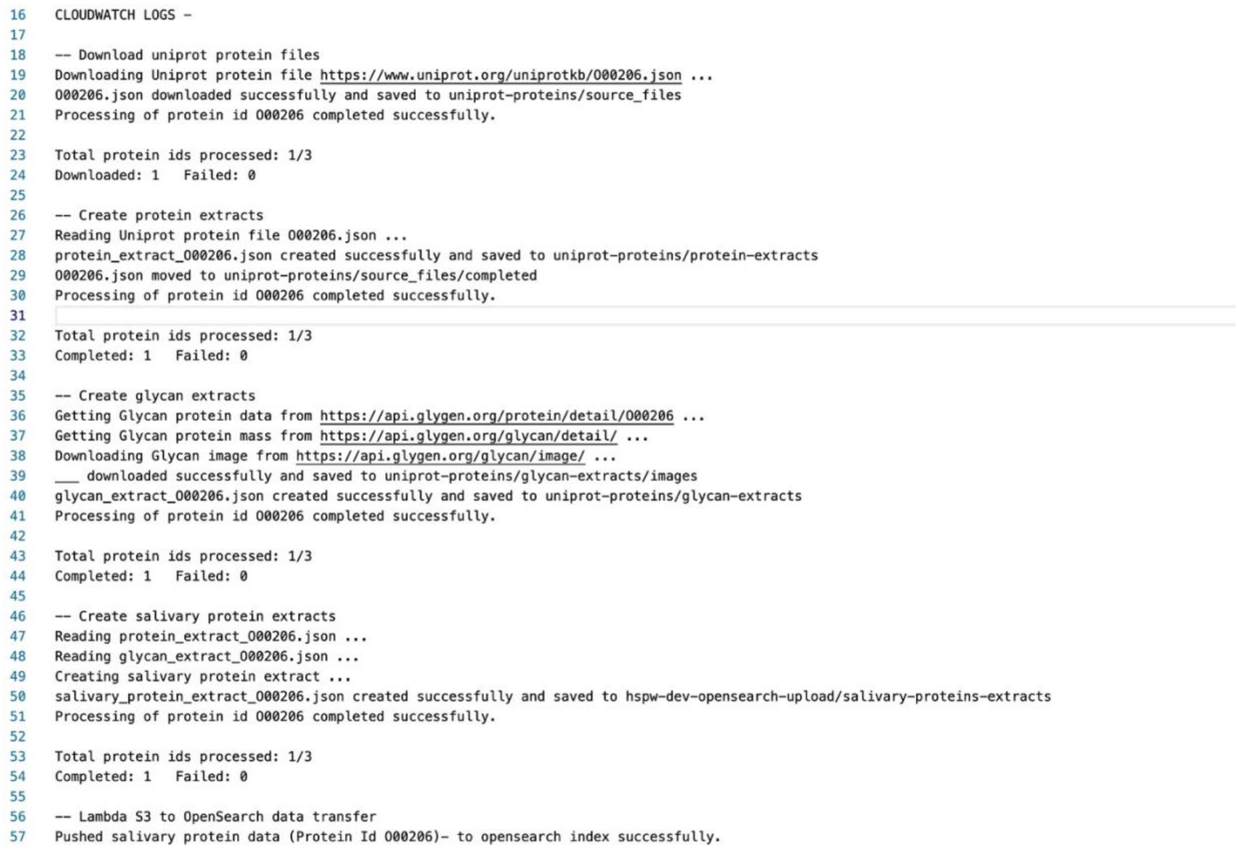


Log groups (5)

By default, we only load up to 10000 log groups.

□	Log group	Data protec...	Sensitive da...	Retention	Metric filters	Cont
□	/aws-glue/python-jobs/error	-	-	Never expire	-	-
□	/aws-glue/python-jobs/output	-	-	Never expire	-	-
□	/aws/lambda/my-function	-	-	Never expire	-	-
□	/aws/lambda/sam-app-HelloWorldFunction-pX0EQns...	-	-	Never expire	-	-
□	/aws/lambda/us-east-1.redirect-func	-	-	Never expire	-	-

CloudWatchLog Groups



```
CLOUDWATCH LOGS -

-- Download uniprot protein files
Downloading Uniprot protein file https://www.uniprot.org/uniprotkb/O00206.json ...
O00206.json downloaded successfully and saved to uniprot-proteins/source_files
Processing of protein id O00206 completed successfully.

Total protein ids processed: 1/3
Downloaded: 1   Failed: 0

-- Create protein extracts
Reading Uniprot protein file O00206.json ...
protein_extract_O00206.json created successfully and saved to uniprot-proteins/protein-extracts
O00206.json moved to uniprot-proteins/source_files/completed
Processing of protein id O00206 completed successfully.

Total protein ids processed: 1/3
Completed: 1   Failed: 0

-- Create glycan extracts
Getting Glycan protein data from https://api.glygen.org/protein/detail/O00206 ...
Getting Glycan protein mass from https://api.glygen.org/glycan/detail/ ...
Downloading Glycan image from https://api.glygen.org/glycan/image/ ...
___ downloaded successfully and saved to uniprot-proteins/glycan-extracts/images
glycan_extract_O00206.json created successfully and saved to uniprot-proteins/glycan-extracts
Processing of protein id O00206 completed successfully.

Total protein ids processed: 1/3
Completed: 1   Failed: 0

-- Create salivary protein extracts
Reading protein_extract_O00206.json ...
Reading glycan_extract_O00206.json ...
Creating salivary protein extract ...
salivary_protein_extract_O00206.json created successfully and saved to hspw-dev-opensearch-upload/salivary-proteins-extracts
Processing of protein id O00206 completed successfully.

Total protein ids processed: 1/3
Completed: 1   Failed: 0

-- Lambda S3 to OpenSearch data transfer
Pushed salivary protein data (Protein Id O00206)- to opensearch index successfully.
```

CloudWatch Logs

Full Code at GitHub

You can get the full code in the JCVenterInstitute GitHub repository.

HSPW-V3/awsjobs at main · JCVenterInstitute/HSPW-V3
*Contribute to JCVenterInstitute/HSPW-V3 development by creating an account on GitHub.*github.com