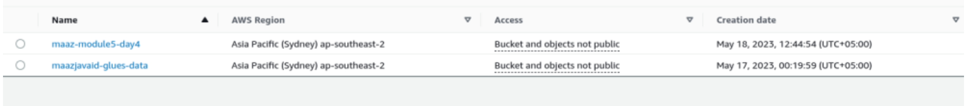
Assignment 5.4 Data Lake

Peer Members:

- Syed Muhammad Raqim Ali Shah (2303.KHI.DEG.008)
- Maaz Javaid Siddique (2303.KHI.DEG.004)
- Qadeer Hussain (2303.KHI.DEG.006)

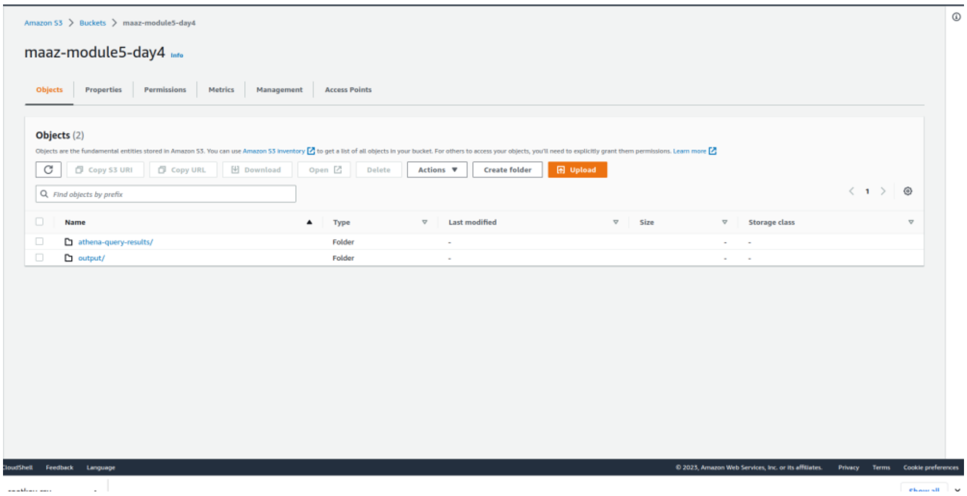
Answer:

S3 Bucket:



	Name ▲	AWS Region	Access	Creation date
○	maaz-module5-day4	Asia Pacific (Sydney) ap-southeast-2	Bucket and objects not public	May 18, 2023, 12:44:54 (UTC+05:00)
○	maazjavaid-glues-data	Asia Pacific (Sydney) ap-southeast-2	Bucket and objects not public	May 17, 2023, 00:19:59 (UTC+05:00)

Folders create in a bucket:



Amazon S3 > Buckets > maaz-module5-day4

maaz-module5-day4 Info

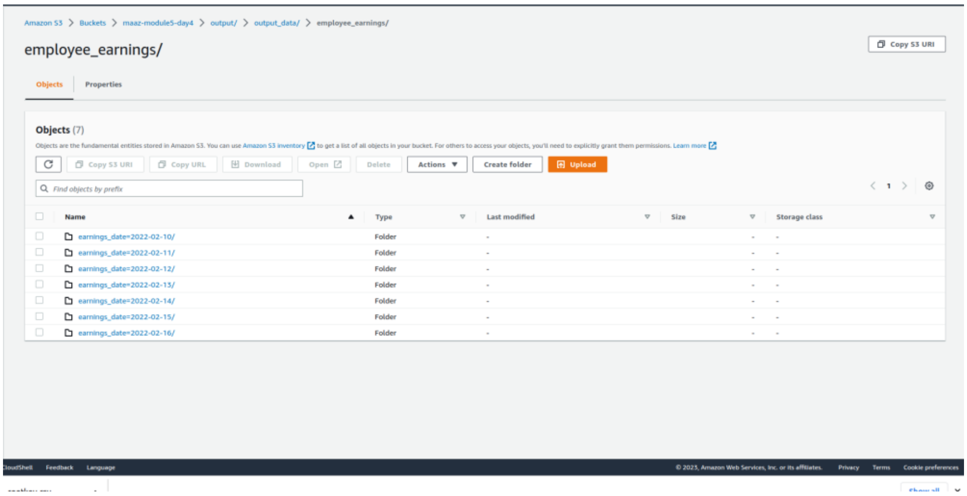
Objects (2)

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

Find objects by prefix

□	Name ▲	Type	Last modified	Size	Storage class
□	athena-query-results/	Folder	-	-	-
□	output/	Folder	-	-	-

Data we add in the output folder:



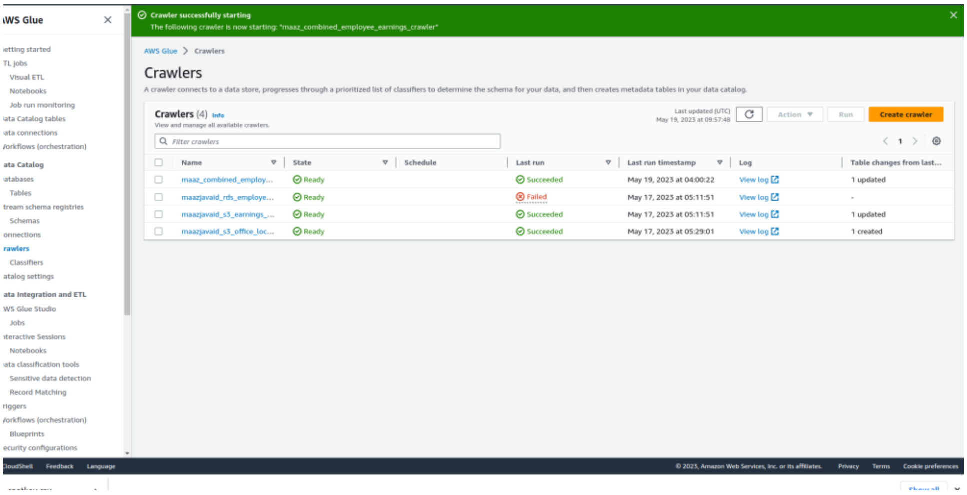
employee_earnings/

Objects (7)

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

□	Name	Type	Last modified	Size	Storage class
□	earnings_date=2022-02-10/	Folder	-	-	-
□	earnings_date=2022-02-11/	Folder	-	-	-
□	earnings_date=2022-02-12/	Folder	-	-	-
□	earnings_date=2022-02-13/	Folder	-	-	-
□	earnings_date=2022-02-14/	Folder	-	-	-
□	earnings_date=2022-02-15/	Folder	-	-	-
□	earnings_date=2022-02-16/	Folder	-	-	-

Here i create a crawler:



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

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.

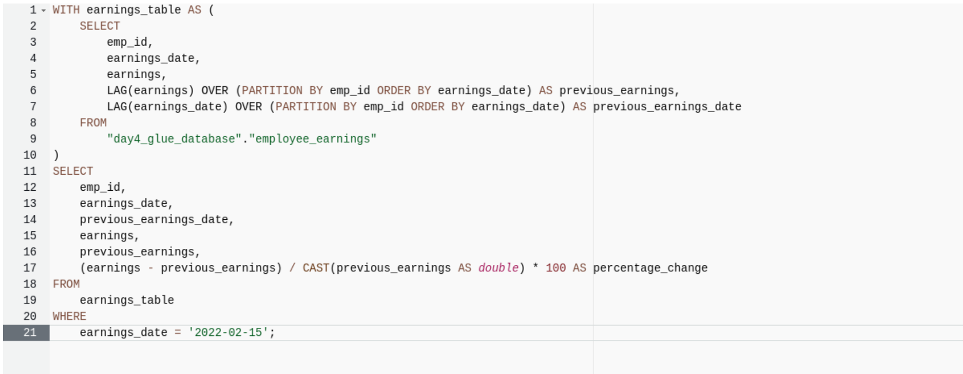
Crawlers (4)

View and manage all available crawlers.

Last updated (UTC)
May 19, 2023 at 09:57:48

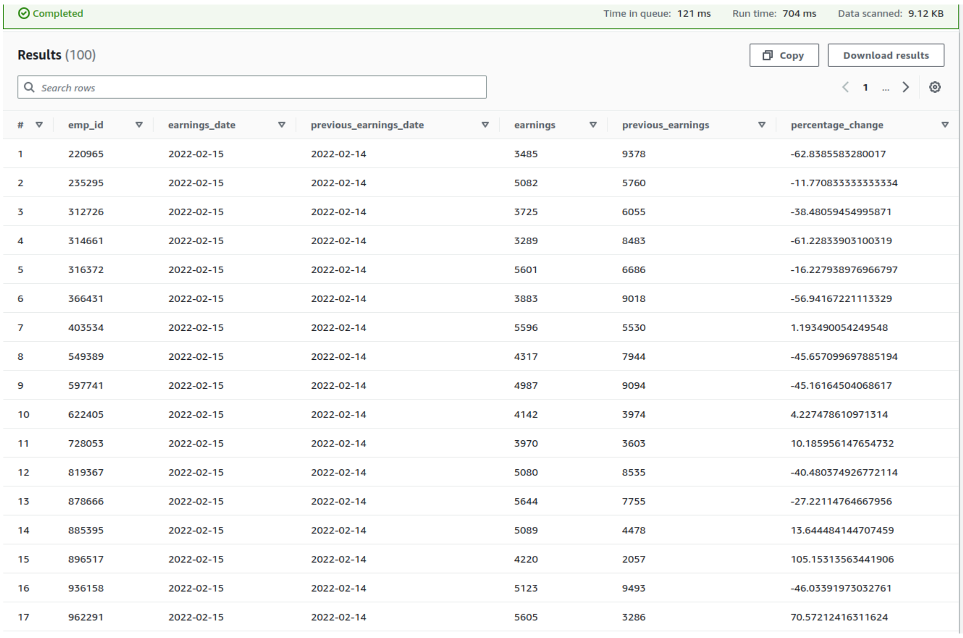
□	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last...
□	maaz_combined_employ...	Ready		Succeeded	May 19, 2023 at 04:00:22	View log	1 updated
□	maazjavaid_rds_employe...	Ready		Failed	May 17, 2023 at 05:11:51	View log	-
□	maazjavaid_s3_earnings_...	Ready		Succeeded	May 17, 2023 at 05:11:51	View log	1 updated
□	maazjavaid_s3_office_loc...	Ready		Succeeded	May 17, 2023 at 05:29:01	View log	1 created

Athena Query editor:



```
WITH earnings_table AS (
    SELECT
        emp_id,
        earnings_date,
        earnings,
        LAG(earnings) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS previous_earnings,
        LAG(earnings_date) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS previous_earnings_date
    FROM
        "day4_glue_database"."employee_earnings"
)
SELECT
    emp_id,
    earnings_date,
    previous_earnings_date,
    earnings,
    previous_earnings,
    (earnings - previous_earnings) / CAST(previous_earnings AS double) * 100 AS percentage_change
FROM
    earnings_table
WHERE
    earnings_date = '2022-02-15';
```

Output Table:

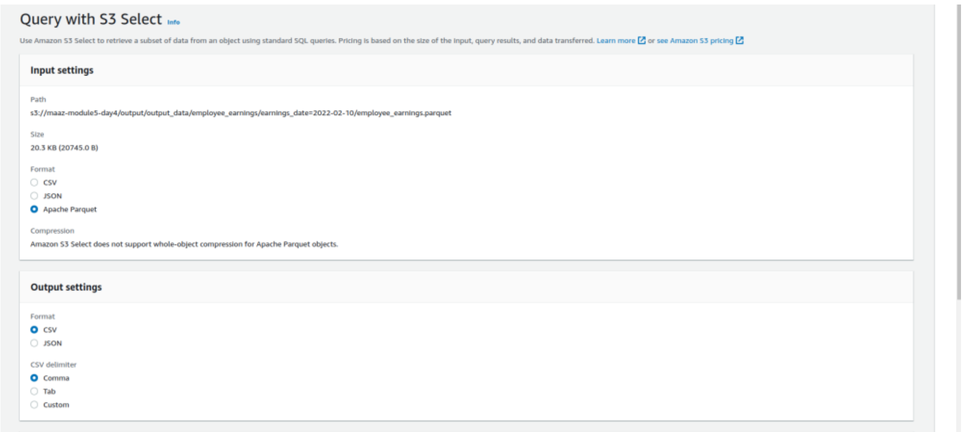


Completed — Time in queue: 121 ms — Run time: 704 ms — Data scanned: 9.12 KB

Results (100)

#	emp_id	earnings_date	previous_earnings_date	earnings	previous_earnings	percentage_change
1	220965	2022-02-15	2022-02-14	3485	9378	-62.8385583280017
2	235295	2022-02-15	2022-02-14	5082	5760	-11.770833333333334
3	312726	2022-02-15	2022-02-14	3725	6055	-38.48059454995871
4	314661	2022-02-15	2022-02-14	3289	8483	-61.22833903100319
5	316372	2022-02-15	2022-02-14	5601	6686	-16.227938976966797
6	366431	2022-02-15	2022-02-14	3883	9018	-56.94167221113329
7	403534	2022-02-15	2022-02-14	5596	5530	1.193490054249548
8	549389	2022-02-15	2022-02-14	4317	7944	-45.657099697885194
9	597741	2022-02-15	2022-02-14	4987	9094	-45.16164504068617
10	622405	2022-02-15	2022-02-14	4142	3974	4.227478610971314
11	728053	2022-02-15	2022-02-14	3970	3603	10.185956147654732
12	819367	2022-02-15	2022-02-14	5080	8535	-40.480374926772114
13	878666	2022-02-15	2022-02-14	5644	7755	-27.22114764667956
14	885395	2022-02-15	2022-02-14	5089	4478	13.644484144707459
15	896517	2022-02-15	2022-02-14	4220	2057	105.15313563441906
16	936158	2022-02-15	2022-02-14	5123	9493	-46.03391973032761
17	962291	2022-02-15	2022-02-14	5605	3286	70.57212416311624

Rerun queries from Task 3 and Task 4 and see how the results change with this new data.



Query with S3 Select Info

Use Amazon S3 Select to retrieve a subset of data from an object using standard SQL queries. Pricing is based on the size of the input, query results, and data transferred. Learn more or see Amazon S3 pricing

Input settings

Path
s3://maaz-module5-day4/output/output_data/employee_earnings/earnings_date=2022-02-10/employee_earnings.parquet

Size
20.3 KB (20745.0 B)

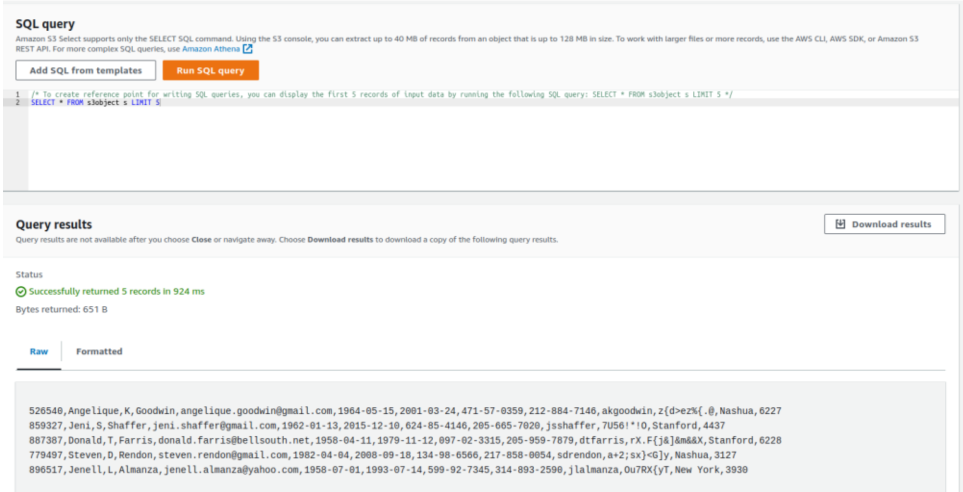
Format
○ CSV
○ JSON
● Apache Parquet

Compression
Amazon S3 Select does not support whole-object compression for Apache Parquet objects.

Output settings

Format
● CSV
○ JSON

CSV delimiter
● Comma
○ Tab
○ Custom



SQL query

Amazon S3 Select supports only the SELECT SQL command. Using the S3 console, you can extract up to 40 MB of records from an object that is up to 128 MB in size. To work with larger files or more records, use the AWS CLI, AWS SDK, or Amazon S3 REST API. For more complex SQL queries, use Amazon Athena

```
/* To create reference point for writing SQL queries, you can display the first 5 records of input data by running the following SQL query: SELECT * FROM s3object s LIMIT 5 */
SELECT * FROM s3object s LIMIT 5
```

Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy of the following query results.

Status
Successfully returned 5 records in 924 ms
Bytes returned: 651 B

```
526540,Angelique,K,Goodwin,angelique.goodwin@gmail.com,1964-05-15,2001-03-24,471-57-0359,212-884-7146,akgoodwin,z{d>ez%{.@,Nashua,6227
859327,Jeni,S,Shaffer,jeni.shaffer@gmail.com,1962-01-13,2015-12-10,624-85-4146,205-665-7020,jsshaffer,7U56!*!O,Stanford,4437
887387,Donald,T,Farris,donald.farris@bellsouth.net,1958-04-11,1979-11-12,097-02-3315,205-959-7879,dtfarris,rX.F{j&]&m&&X,Stanford,6228
779497,Steven,D,Rendon,steven.rendon@gmail.com,1982-04-04,2008-09-18,134-98-6566,217-858-0054,sdrendon,a+2;sx}<G]y,Nashua,3127
896517,Jenell,L,Almanza,jenell.almanza@yahoo.com,1958-07-01,1993-07-14,599-92-7345,314-893-2590,jlalmanza,Ou7RX{yT,New York,3930
```

Query is not run in the S3 bucket because s3 is a file storage service it's not a query editor. It accepts a very basic query. If I apply multi functions query it fails the query.