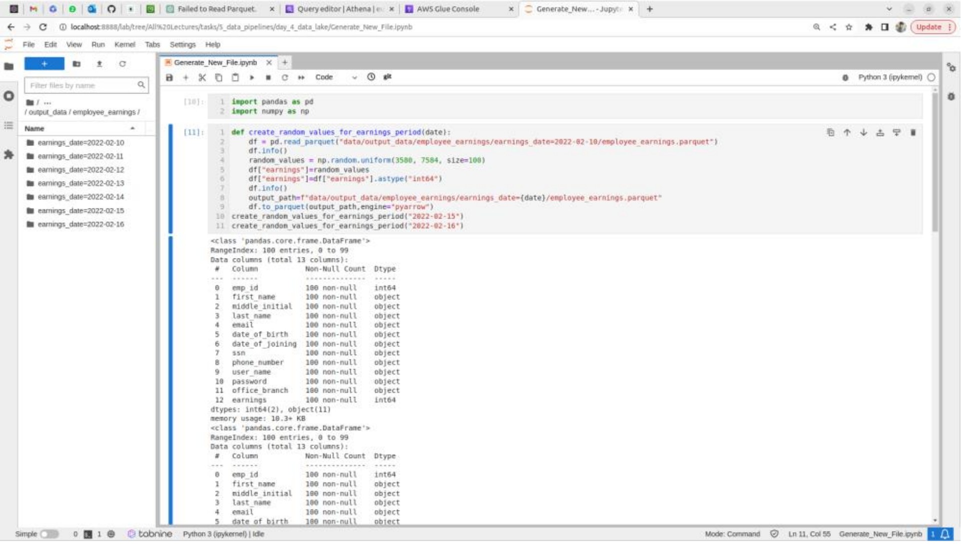
Problem Statement:

Using the data manipulation tool of your choice (eg. Python) simulate the earnings predictions for 2 more days. Load it to the Data Lake that you've created today (Task 1-2).
Rerun queries from Task 3 and Task 4 and see how the results change with this new data.
Create a new query in Athena that calculates the % change in earnings for every employee from a given day compared to the previous day.

Solution:

Creating the new data using this script

We firstly analyzed the given data and based on that data we created the new data using a suitable range of random data.

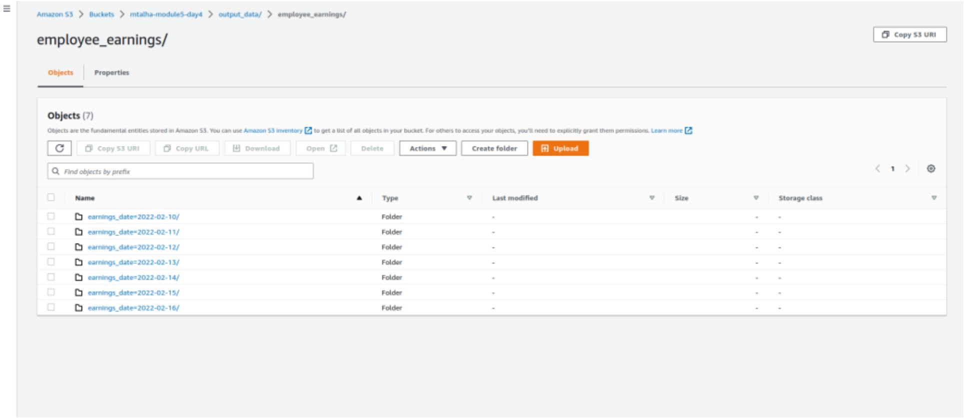


```
[10]: import pandas as pd
      import numpy as np
```

```
[11]: def create_random_values_for_earnings_period(date):
          df = pd.read_parquet("data/output_data/employee_earnings/earnings_date=2022-02-10/employee_earnings.parquet")
          df.info()
          random_values = np.random.uniform(3580, 7584, size=100)
          df["earnings"]=random_values
          df["earnings"]=df["earnings"].astype("int64")
          df.info()
          output_path=f"data/output_data/employee_earnings/earnings_date={date}/employee_earnings.parquet"
          df.to_parquet(output_path,engine="pyarrow")
      create_random_values_for_earnings_period("2022-02-15")
      create_random_values_for_earnings_period("2022-02-16")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   emp_id           100 non-null    int64
 1   first_name       100 non-null    object
 2   middle_initial   100 non-null    object
 3   last_name        100 non-null    object
 4   email            100 non-null    object
 5   date_of_birth    100 non-null    object
 6   date_of_joining  100 non-null    object
 7   ssn              100 non-null    object
 8   phone_number     100 non-null    object
 9   user_name        100 non-null    object
 10  password         100 non-null    object
 11  office_branch    100 non-null    object
 12  earnings         100 non-null    int64
dtypes: int64(2), object(11)
memory usage: 10.3+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   emp_id           100 non-null    int64
 1   first_name       100 non-null    object
 2   middle_initial   100 non-null    object
 3   last_name        100 non-null    object
 4   email            100 non-null    object
 5   date_of_birth    100 non-null    object
```

Task 1: Loading the Data into S3:

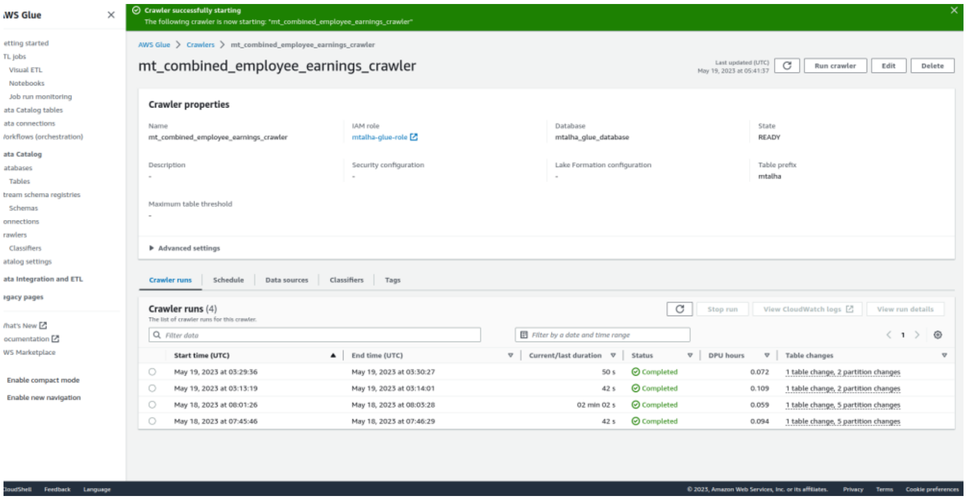


Objects (7)

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	-	-	-

Task2: Running the Glue Crawler on the new data:



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

mt_combined_employee_earnings_crawler

Last updated (UTC)
May 19, 2023 at 05:41:37

Crawler properties

Name: mt_combined_employee_earnings_crawler
IAM role: mtalha-glue-role
Database: mtalha_glue_database
State: READY
Description: -
Security configuration: -
Lake Formation configuration: -
Table prefix: mtalha
Maximum table threshold: -

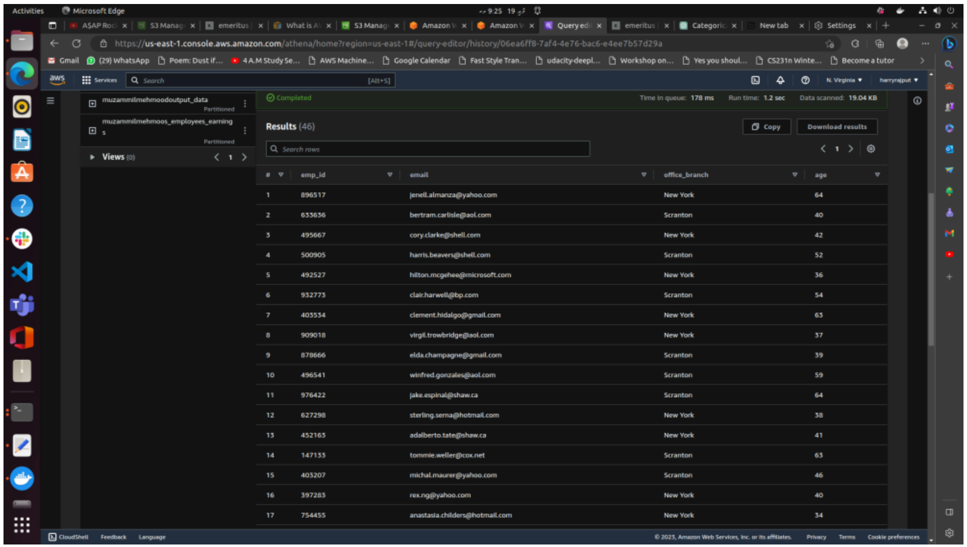
▶ Advanced settings

Crawler runs (4)
The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	May 19, 2023 at 03:29:36	May 19, 2023 at 03:30:27	50 s	Completed	0.072	1 table change, 2 partition changes
○	May 19, 2023 at 03:13:19	May 19, 2023 at 03:14:01	42 s	Completed	0.109	1 table change, 2 partition changes
○	May 18, 2023 at 08:01:26	May 18, 2023 at 08:03:28	02 min 02 s	Completed	0.059	1 table change, 5 partition changes
○	May 18, 2023 at 07:45:46	May 18, 2023 at 07:46:29	42 s	Completed	0.094	1 table change, 5 partition changes

Task 3: Rerunning the previous query to see the change:

These are the previous query result



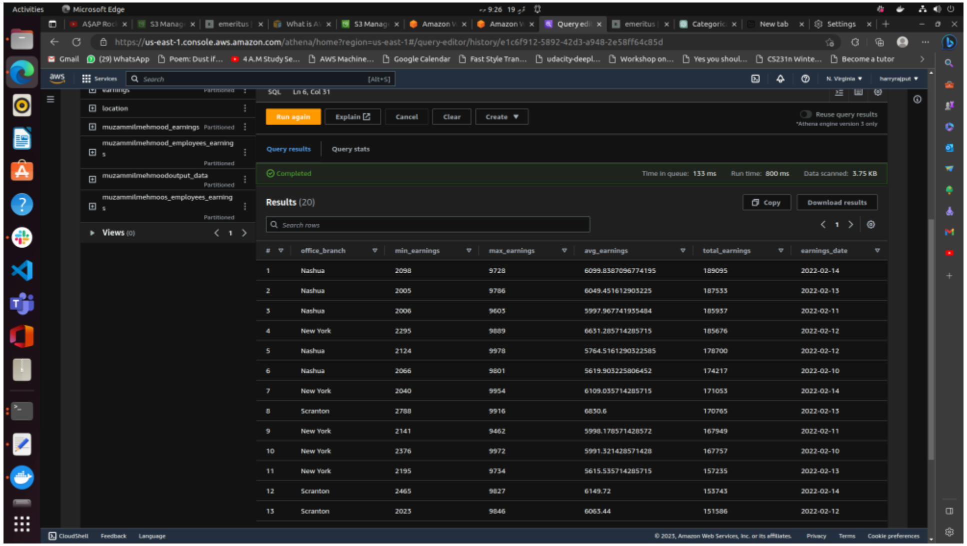
https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/06ea6ff8-7af4-4e76-bac6-e4ee7b57d29a

Completed Time in queue: 178 ms Run time: 1.2 sec Data scanned: 19.04 KB

Results (46)

#	emp_id	email	office_branch	age
1	896517	jenell.almanza@yahoo.com	New York	64
2	633636	bertram.carlisle@aol.com	Scranton	40
3	495667	cory.clarke@shell.com	New York	42
4	500905	harris.beavers@shell.com	Scranton	52
5	492527	hilton.mcgehee@microsoft.com	New York	36
6	932773	clair.harwell@bp.com	Scranton	54
7	403534	clement.hidalgo@gmail.com	New York	63
8	909018	virgil.trowbridge@aol.com	New York	37
9	878666	elda.champagne@gmail.com	Scranton	39
10	496541	winfred.gonzales@aol.com	Scranton	59
11	976422	jake.espinal@shaw.ca	Scranton	64
12	627298	sterling.serna@hotmail.com	New York	38
13	452163	adalberto.tate@shaw.ca	New York	41
14	147133	tommie.weller@cox.net	Scranton	63
15	403207	michal.maurer@yahoo.com	Scranton	46
16	397283	rex.ng@yahoo.com	New York	40
17	754455	anastasia.childers@hotmail.com	New York	34

muzammilmehmoodoutput_data
muzammilmehmoos_employees_earnings
Views (0)
Copy
Download results
© 2023, Amazon Web Services, Inc. or its affiliates.



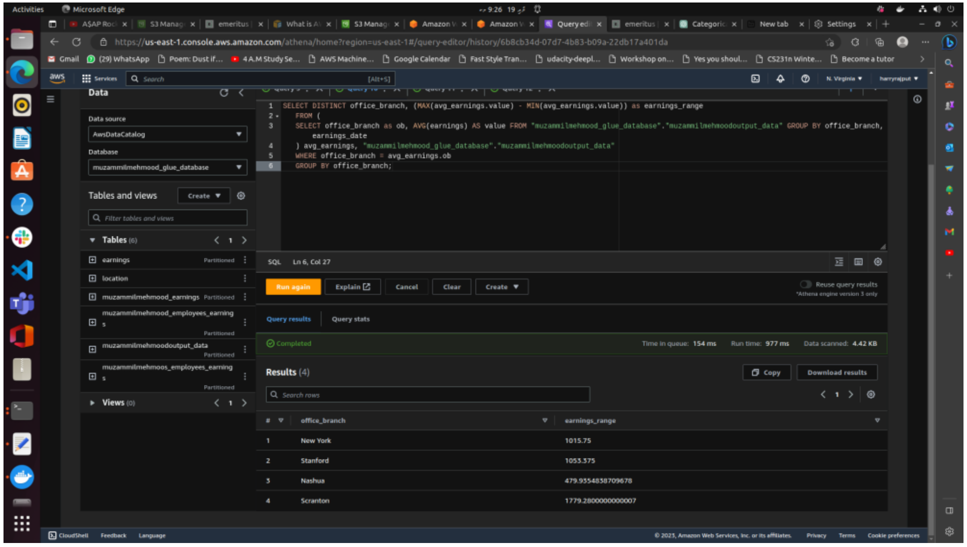
https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/e1c6f912-5892-42d3-a948-2e58ff64c85d

SQL Ln 6, Col 31

Completed Time in queue: 133 ms Run time: 800 ms Data scanned: 3.75 KB

Results (20)

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_date
1	Nashua	2098	9728	6099.8387096774195	189095	2022-02-14
2	Nashua	2005	9786	6049.451612903225	187533	2022-02-13
3	Nashua	2006	9603	5997.967741935484	185937	2022-02-11
4	New York	2295	9889	6631.285714285715	185676	2022-02-12
5	Nashua	2124	9978	5764.5161290322585	178700	2022-02-12
6	Nashua	2066	9801	5619.903225806452	174217	2022-02-10
7	New York	2040	9954	6109.035714285715	171053	2022-02-14
8	Scranton	2788	9916	6830.6	170765	2022-02-13
9	New York	2141	9462	5998.178571428572	167949	2022-02-11
10	New York	2376	9972	5991.321428571428	167757	2022-02-10
11	New York	2195	9734	5615.535714285715	157235	2022-02-13
12	Scranton	2465	9827	6149.72	153743	2022-02-14
13	Scranton	2023	9846	6063.44	151586	2022-02-12



https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/6b8cb34d-07d7-4b83-b09a-22db17a401da

Data source: AwsDataCatalog

Database: muzammilmehmood_glue_database

```sql
SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
  FROM (
  SELECT office_branch as ob, AVG(earnings) AS value FROM "muzammilmehmood_glue_database"."muzammilmehmoodoutput_data" GROUP BY office_branch,
     earnings_date
  ) avg_earnings, "muzammilmehmood_glue_database"."muzammilmehmoodoutput_data"
  WHERE office_branch = avg_earnings.ob
  GROUP BY office_branch;
```

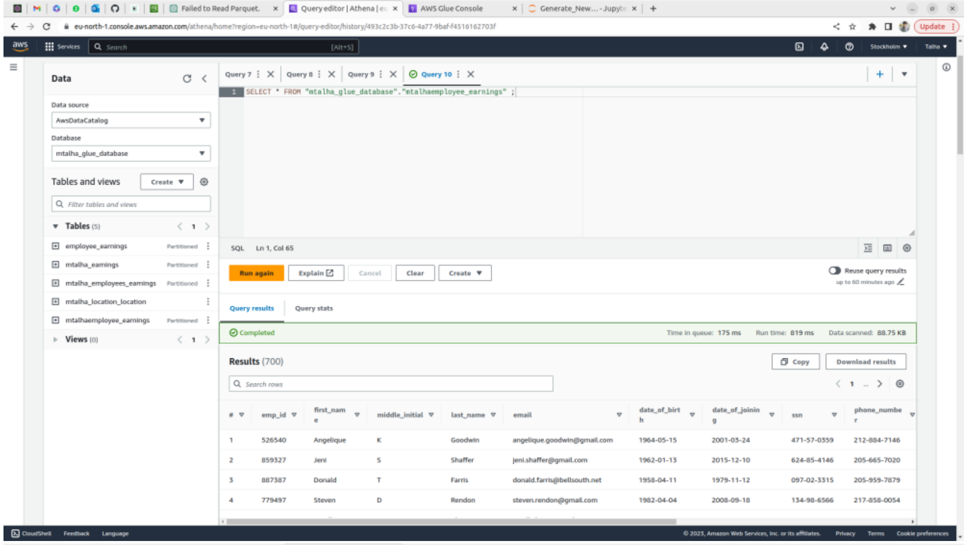
SQL Ln 6, Col 27

Completed Time in queue: 154 ms Run time: 977 ms Data scanned: 4.42 KB

Results (4)

#	office_branch	earnings_range
1	New York	1015.75
2	Stanford	1053.375
3	Nashua	479.9354838709678
4	Scranton	1779.2800000000007

Now let me show you the result of the previous query on new data



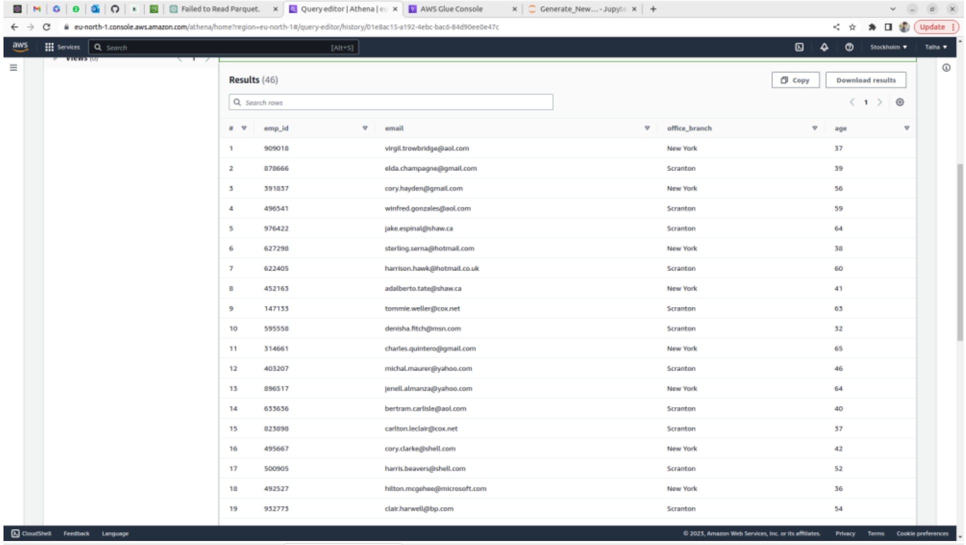
```
SELECT * FROM "mtalha_glue_database"."mtalhaemployee_earnings" ;
```

SQL Ln 1, Col 65

Completed — Time in queue: 175 ms — Run time: 819 ms — Data scanned: 88.75 KB

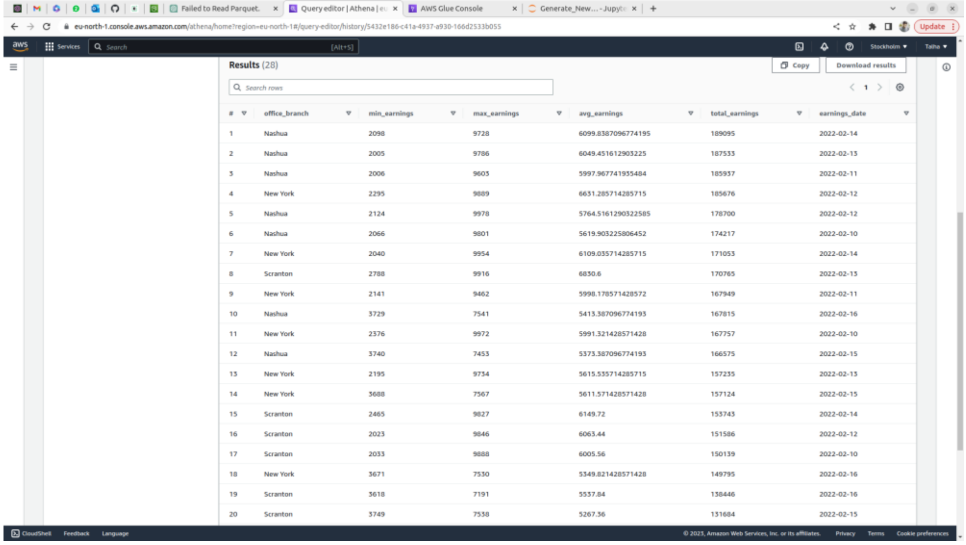
Results (700)

#	emp_id	first_name	middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number
1	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-03-24	471-57-0359	212-884-7146
2	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020
3	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879
4	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054



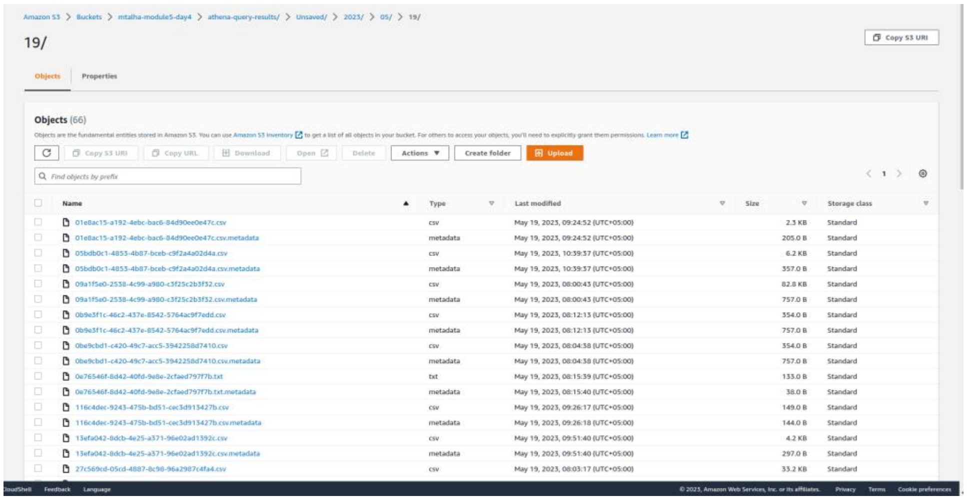
Results (46)

#	emp_id	email	office_branch	age
1	909018	virgil.trowbridge@aol.com	New York	37
2	878666	elda.champagne@gmail.com	Scranton	39
3	391837	cory.hayden@gmail.com	New York	56
4	496541	winfred.gonzales@aol.com	Scranton	59
5	976422	jake.espinal@shaw.ca	Scranton	64
6	627298	sterling.serna@hotmail.com	New York	38
7	622405	harrison.hawk@hotmail.co.uk	Scranton	60
8	452163	adalberto.tate@shaw.ca	New York	41
9	147133	tommie.weller@cox.net	Scranton	63
10	595558	denisha.fitch@msn.com	Scranton	32
11	314661	charles.quintero@gmail.com	New York	65
12	403207	michal.maurer@yahoo.com	Scranton	46
13	896517	jenell.almanza@yahoo.com	New York	64
14	633636	bertram.carlisle@aol.com	Scranton	40
15	823898	carlton.leclair@cox.net	Scranton	37
16	495667	cory.clarke@shell.com	New York	42
17	500905	harris.beavers@shell.com	Scranton	52
18	492527	hilton.mcgehee@microsoft.com	New York	36
19	932773	clair.harwell@bp.com	Scranton	54



Results (28)

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_date
1	Nashua	2098	9728	6099.8387096774195	189095	2022-02-14
2	Nashua	2005	9786	6049.451612903225	187533	2022-02-13
3	Nashua	2006	9603	5997.967741935484	185937	2022-02-11
4	New York	2295	9889	6631.285714285715	185676	2022-02-12
5	Nashua	2124	9978	5764.5161290322585	178700	2022-02-12
6	Nashua	2066	9801	5619.903225806452	174217	2022-02-10
7	New York	2040	9954	6109.035714285715	171053	2022-02-14
8	Scranton	2788	9916	6830.6	170765	2022-02-13
9	New York	2141	9462	5998.178571428572	167949	2022-02-11
10	Nashua	3729	7541	5413.387096774193	167815	2022-02-16
11	New York	2376	9972	5991.321428571428	167757	2022-02-10
12	Nashua	3740	7453	5373.387096774193	166575	2022-02-15
13	New York	2195	9734	5615.535714285715	157235	2022-02-13
14	New York	3688	7567	5611.571428571428	157124	2022-02-15
15	Scranton	2465	9827	6149.72	153743	2022-02-14
16	Scranton	2023	9846	6063.44	151586	2022-02-12
17	Scranton	2033	9888	6005.56	150139	2022-02-10
18	New York	3671	7530	5349.821428571428	149795	2022-02-16
19	Scranton	3618	7191	5537.84	138446	2022-02-16
20	Scranton	3749	7538	5267.36	131684	2022-02-15



19/

Objects (66)

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

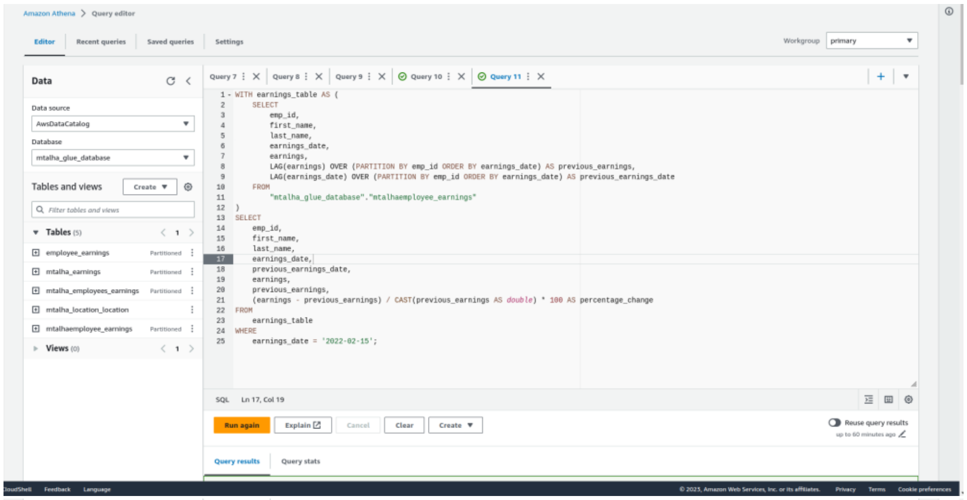
□	Name	Type	Last modified	Size	Storage class
□	01e8ac15-a192-4ebc-bac6-84d90ee0e47c.csv	csv	May 19, 2023, 09:24:52 (UTC+05:00)	2.3 KB	Standard
□	01e8ac15-a192-4ebc-bac6-84d90ee0e47c.csv.metadata	metadata	May 19, 2023, 09:24:52 (UTC+05:00)	205.0 B	Standard
□	05bdb0c1-4853-4b87-bceb-c9f2a4a02d4a.csv	csv	May 19, 2023, 10:39:37 (UTC+05:00)	6.2 KB	Standard
□	05bdb0c1-4853-4b87-bceb-c9f2a4a02d4a.csv.metadata	metadata	May 19, 2023, 10:39:37 (UTC+05:00)	357.0 B	Standard
□	09a1f5e0-2538-4c99-a980-c3f25c2b3f32.csv	csv	May 19, 2023, 08:00:43 (UTC+05:00)	82.8 KB	Standard
□	09a1f5e0-2538-4c99-a980-c3f25c2b3f32.csv.metadata	metadata	May 19, 2023, 08:00:43 (UTC+05:00)	757.0 B	Standard
□	0b9e3f1c-46c2-437e-8542-5764ac9f7edd.csv	csv	May 19, 2023, 08:12:13 (UTC+05:00)	354.0 B	Standard
□	0b9e3f1c-46c2-437e-8542-5764ac9f7edd.csv.metadata	metadata	May 19, 2023, 08:12:13 (UTC+05:00)	757.0 B	Standard
□	0be9cbd1-c420-49c7-acc5-3942258d7410.csv	csv	May 19, 2023, 08:04:38 (UTC+05:00)	354.0 B	Standard
□	0be9cbd1-c420-49c7-acc5-3942258d7410.csv.metadata	metadata	May 19, 2023, 08:04:38 (UTC+05:00)	757.0 B	Standard
□	0e76546f-8d42-40fd-9e8e-2cfaed797f7b.txt	txt	May 19, 2023, 08:15:39 (UTC+05:00)	133.0 B	Standard
□	0e76546f-8d42-40fd-9e8e-2cfaed797f7b.txt.metadata	metadata	May 19, 2023, 08:15:40 (UTC+05:00)	38.0 B	Standard
□	116c4dec-9243-475b-bd51-cec3d913427b.csv	csv	May 19, 2023, 09:26:17 (UTC+05:00)	149.0 B	Standard
□	116c4dec-9243-475b-bd51-cec3d913427b.csv.metadata	metadata	May 19, 2023, 09:26:18 (UTC+05:00)	144.0 B	Standard
□	13efa042-8dcb-4e25-a371-96e02ad1392c.csv	csv	May 19, 2023, 09:51:40 (UTC+05:00)	4.2 KB	Standard
□	13efa042-8dcb-4e25-a371-96e02ad1392c.csv.metadata	metadata	May 19, 2023, 09:51:40 (UTC+05:00)	297.0 B	Standard
□	27c569cd-05cd-4887-8c98-96a2987c4fa4.csv	csv	May 19, 2023, 08:03:17 (UTC+05:00)	33.2 KB	Standard

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

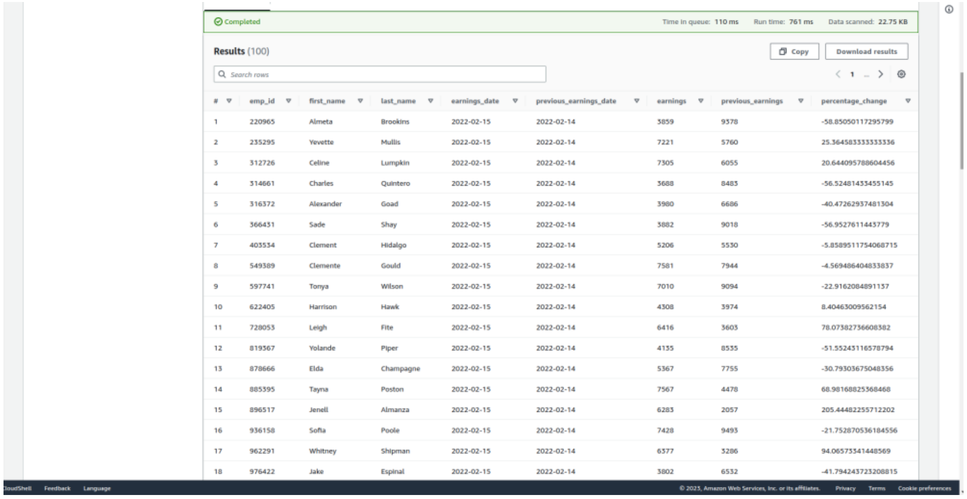
Task4: Querying the data using S3 Select

Due to the limitation of s3 select and the complexity of your queries, we were unable to run the queries directly on s3 select as it operates on the single object at a time, while Athena allows you to run queries across multiple objects and supports more complex queries and it would need multiple subsets of data, which cause errors in s3 select.

Create a new query in Athena that calculates the % change in earnings for every employee from a given day compared to the previous day.



```sql
WITH earnings_table AS (
    SELECT
        emp_id,
        first_name,
        last_name,
        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
        "mtalha_glue_database"."mtalhaemployee_earnings"
)
SELECT
    emp_id,
    first_name,
    last_name,
    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';
```



Results (100)

#	emp_id	first_name	last_name	earnings_date	previous_earnings_date	earnings	previous_earnings	percentage_change
1	220965	Almeta	Brookins	2022-02-15	2022-02-14	3859	9378	-58.85050117295799
2	235295	Yevette	Mullis	2022-02-15	2022-02-14	7221	5760	25.364583333333336
3	312726	Celine	Lumpkin	2022-02-15	2022-02-14	7305	6055	20.644095788604456
4	314661	Charles	Quintero	2022-02-15	2022-02-14	3688	8483	-56.52481433455145
5	316372	Alexander	Goad	2022-02-15	2022-02-14	3980	6686	-40.47262937481304
6	366431	Sade	Shay	2022-02-15	2022-02-14	3882	9018	-56.9527611443779
7	403534	Clement	Hidalgo	2022-02-15	2022-02-14	5206	5530	-5.8589511754068715
8	549389	Clemente	Gould	2022-02-15	2022-02-14	7581	7944	-4.569486404833837
9	597741	Tonya	Wilson	2022-02-15	2022-02-14	7010	9094	-22.9162084891137
10	622405	Harrison	Hawk	2022-02-15	2022-02-14	4308	3974	8.40463009562154
11	728053	Leigh	Fite	2022-02-15	2022-02-14	6416	3603	78.07382736608382
12	819367	Yolande	Piper	2022-02-15	2022-02-14	4135	8535	-51.55243116578794
13	878666	Elda	Champagne	2022-02-15	2022-02-14	5367	7755	-30.79303675048356
14	885395	Tayna	Poston	2022-02-15	2022-02-14	7567	4478	68.98168825368468
15	896517	Jenell	Almanza	2022-02-15	2022-02-14	6283	2057	205.44482255712202
16	936158	Sofia	Poole	2022-02-15	2022-02-14	7428	9493	-21.752870536184556
17	962291	Whitney	Shipman	2022-02-15	2022-02-14	6377	3286	94.06573341448569
18	976422	Jake	Espinal	2022-02-15	2022-02-14	3802	6532	-41.794243723208815