

NAME: Mohammad Hussam (2303.KHI.DEG.020)
MAVIA ALAM KHAN (2303.KHI.DEG.017)
PAIRING WITH :
ASSIGNMENT NO : 5.4

Here we run the task 3 queries

Amazon Athena > Query editor

Editor | Recent queries | Saved queries | Settings

Workgroup: primary

Data

Data source: AwsDataCatalog

Database: mavia-glue-database

Tables and views

Filter tables and views

Tables (8)

alam_employee_earnings Partitioned

mavia_earnings_1_csv

Query 1 : X Query 2 : X Query 3 : X

```
1 -- SELECT * FROM "mavia-glue-database"."alam_employee_earnings" limit 10;
2
3 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
4 FROM "mavia-glue-database"."alam_employee_earnings"
5 WHERE office_branch IN ('New York', 'Scranton')
6 AND
7 (date_diff('year', DATE(date_of_birth), current_date)) > 30;
```

SQL Ln 7, Col 61

Amazon Athena > Query editor

Editor | Recent queries | Saved queries | Settings

Workgroup: primary

Data

Data source: AwsDataCatalog

Database: mavia-glue-database

Tables and views

Filter tables and views

Tables (8)

alam_employee_earnings Partitioned

mavia_earnings_1_csv

Query 1 : X Query 2 : X Query 3 : X

```
1 -- SELECT * FROM "mavia-glue-database"."alam_employee_earnings" limit 10;
2
3 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
4 FROM "mavia-glue-database"."alam_employee_earnings"
5 WHERE office_branch IN ('New York', 'Scranton')
6 AND
7 (date_diff('year', DATE(date_of_birth), current_date)) > 30;
```

SQL Ln 7, Col 61

mavia_employee_earnings Partitioned

mavia_locations

mavia_locations_csv

mavia_output_data Partitioned

mavialkhanearnings Partitioned

Views (0)

Query results | Query stats

Completed

Time in queue: 109 ms Run time: 924 ms Data scanned: 26.13 KB

Results (46)

Search rows

| # | emp_id | email | office_branch | age |
|---|--------|----------------------------|---------------|-----|
| 1 | 900756 | benjamin.doss@gmail.com | Scranton | 38 |
| 2 | 654617 | rogerio.woodall@gmail.com | New York | 50 |
| 3 | 138911 | claudio.heck@aol.com | Scranton | 55 |
| 4 | 713294 | sammy.dewitt@ibm.com | Scranton | 35 |
| 5 | 215719 | brent.carrillo@aol.com | New York | 50 |
| 6 | 312726 | celine.lumpkin@gmail.com | New York | 36 |
| 7 | 530134 | mathew.whitfield@gmail.com | New York | 36 |

Tables and views

Create

Filter tables and views

▼ Tables (8)

alam_employee_earnings

mavia_earnings_1_csv

mavia_earnings_2_csv

mavia_employee_earnings

mavia_locations

mavia_locations_csv

mavia_output_data

maviakhanearnings

Views (0)

```

13
14 SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
15 - FROM (
16 SELECT office_branch as ob, AVG(earnings) AS value FROM "mavia-glue-database"."mavia_employee_earnings" GROUP BY office_branch, earnings_date
17 ) avg_earnings, "mavia-glue-database"."alam_employee_earnings"
18 WHERE office_branch = avg_earnings.ob
19 GROUP BY office_branch;

```

SQL Ln 18, Col 38

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 151 ms Run time: 1.348 sec Data scanned: 4.69 KB

Results (4)

Search rows

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

Tables and views

Create

Filter tables and views

▼ Tables (8)

alam_employee_earnings

mavia_earnings_1_csv

mavia_earnings_2_csv

mavia_employee_earnings

mavia_locations

mavia_locations_csv

mavia_output_data

maviakhanearnings

Views (0)

```

13
14 SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
15 - FROM (
16 SELECT office_branch as ob, AVG(earnings) AS value FROM "mavia-glue-database"."mavia_employee_earnings" GROUP BY office_branch, earnings_date
17 ) avg_earnings, "mavia-glue-database"."alam_employee_earnings"
18 WHERE office_branch = avg_earnings.ob
19 GROUP BY office_branch;

```

SQL Ln 18, Col 38

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 151 ms Run time: 1.348 sec Data scanned: 4.69 KB

Results (4)

Search rows

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

Tables and views

Create

Filter tables and views

▼ Tables (8)

alam_employee_earnings

mavia_earnings_1_csv

mavia_earnings_2_csv

mavia_employee_earnings

mavia_locations

mavia_locations_csv

mavia_output_data

maviakhanearnings

Views (0)

```

8 SELECT office_branch, MIN(earnings) as min_earnings, MAX(earnings) as max_earnings, AVG(earnings) as avg_earnings, SUM(earnings) as total_earnings
9 , earnings_date
10 FROM "mavia-glue-database"."alam_employee_earnings"
11 GROUP BY office_branch, earnings_date
12 ORDER BY SUM(earnings) desc;

```

SQL Ln 12, Col 29

Run again Explain Cancel Clear Create

Reuse query results up to 60 minutes ago

Query results Query stats

Completed Time in queue: 548 ms Run time: 791 ms Data scanned: 5.24 KB

Results (28)

Search rows

| # | 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 | 2017 | 9614 | 6008.967741935484 | 186278 | 2022-02-16 |
| 4 | Nashua | 2006 | 9603 | 5997.967741935484 | 185937 | 2022-02-11 |
| 5 | New York | 2295 | 9889 | 6631.285714285715 | 185676 | 2022-02-12 |
| 6 | Nashua | 2124 | 9978 | 5764.5161290322585 | 178700 | 2022-02-12 |
| 7 | Nashua | 2076 | 9811 | 5629.90322306452 | 174327 | 2022-02-15 |

We calculates the % change in earnings for every employee from a given day compared to the previous day.

Query 1 : X Query 4 : X Query 5 : X Query 6 : X

```

1 - WITH earnings_change AS (
2   SELECT
3     emp_id,
4     earnings,
5     earnings_date,
6     LAG(earnings) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS previous_earnings
7   FROM
8     "mavia-glue-database"."alam_employee_earnings"
9 )
10 SELECT
11   emp_id,
12   earnings_date,
13   earnings,
14   previous_earnings,
15   CASE

```

alam_employee_earnings Partitioned

Data source: AwsDataCatalog Database: mavia-glue-database

Tables and views: Filter tables and views

Tables (8):

- alam_employee_earnings Partitioned
- mavia_earnings_1_csv
- mavia_earnings_2_csv
- mavia_employee_earnings Partitioned
- mavia_locations
- mavia_locations_csv
- mavia_output_data Partitioned
- maviakhanearnings Partitioned

Views (0)

```

11   emp_id,
12   earnings_date,
13   earnings,
14   previous_earnings,
15   CASE
16     WHEN previous_earnings IS NOT NULL THEN (earnings - previous_earnings) / cast(previous_earnings as double) * 100
17     ELSE NULL
18   END AS percentage_change
19 FROM
20   earnings_change
21 WHERE
22   earnings_date = '2022-02-16' -- Replace with your desired date
23 ORDER BY
24   emp_id, earnings_date;
25

```

SQL Ln 24, Col 25

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 142 ms Run time: 747 ms Data scanned: 9.16 KB

Results (100)

Search rows

| # | emp_id | earnings_date | earnings | previous_earnings | percentage_change |
|---|--------|---------------|----------|-------------------|---------------------|
| 1 | 138911 | 2022-02-16 | 2210 | 3826 | -42.237323575535804 |
| 2 | 143711 | 2022-02-16 | 4431 | 2831 | 56.51713175556341 |
| 3 | 147133 | 2022-02-16 | 3422 | 5088 | -32.7437106918239 |
| 4 | 149972 | 2022-02-16 | 7918 | 5353 | 47.91705585652905 |
| 5 | 155097 | 2022-02-16 | 2703 | 5957 | -54.62481114655028 |

Here we create s3 bucket and create 2 folder one is employee earning and second is athena-query-result. In employee earning we stored a dataset

Amazon S3 Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

AWS Organizations settings

Feature spotlight

Amazon S3 > Buckets > mavia-module5-day4

mavia-module5-day4

Objects Properties Permissions Metrics Management Access Points

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](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

| Name | Type | Last modified | Size | Storage class |
|-----------------------|--------|---------------|------|---------------|
| athena-query-results/ | Folder | - | - | - |
| employee_earnings/ | Folder | - | - | - |

We create a crawler

Crawlers successfully starting
The following crawlers are now starting: "maviakhan_combined_employee_earnings_crawler", "mavia_combined_employee_earnings_crawler"

AWS Glue > Crawlers > maviakhan_combined_employee_earnings_crawler

maviakhan_combined_employee_earnings_crawler Last updated (UTC) May 22, 2023 at 09:52:44 Run crawler Edit Delete

Crawler properties

| | | | | | | | |
|-------------------------|--|------------------------|---------------------|------------------------------|---------------------|--------------|-------|
| Name | maviakhan_combined_employee_earnings_crawler | IAM role | maviakhan-glue-role | Database | mavia-glue-database | State | READY |
| Description | - | Security configuration | - | Lake Formation configuration | - | Table prefix | alam_ |
| Maximum table threshold | - | | | | | | |

▶ Advanced settings

Crawler runs (1) Stop run View CloudWatch logs View run details

The list of crawler runs for this crawler.

Filter data Filter by a date and time range

| | Start time (UTC) | End time (UTC) | Current/last duration | Status | DPU hours | Table changes |
|---|--------------------------|--------------------------|-----------------------|-----------|-----------|-------------------------------------|
| ○ | May 22, 2023 at 04:42:35 | May 22, 2023 at 04:43:24 | 49 s | Completed | 0.068 | 1 table change, 7 partition changes |

Using pandas we read a dataset

View Run Kernel Tabs Settings Help

Untitled.ipynb Python 3 (pykernel)

```
[54]: import pandas as pd
import os

[55]: df=pd.read_parquet("output_data/employee_earnings/earnings_date=2022-02-10/employee_earnings.parquet")

[56]: df.head()
```

| | emp_id | first_name | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password |
|---|--------|------------|----------------|-----------|-----------------------------|---------------|-----------------|-------------|--------------|-----------|---------------|
| 0 | 526540 | Angelique | K | Goodwin | angelique.goodwin@gmail.com | 1964-05-15 | 2001-03-24 | 471-57-0359 | 212-884-7146 | akgoodwin | z[d>ez%{. @ |
| 1 | 859327 | Jeni | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!""O |
| 2 | 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 |
| 3 | 779497 | Steven | D | Rendon | steven.rendon@gmail.com | 1982-04-04 | 2008-09-18 | 134-98-6566 | 217-858-0054 | sdrendon | a+2;sx)<Gjy |
| 4 | 896517 | Jenell | L | Almanza | jenell.almanza@yahoo.com | 1958-07-01 | 1993-07-14 | 599-92-7345 | 314-893-2590 | jialmanza | Ou7RX(yT |

```
[57]: df["earnings"]=df["earnings"]+10
```

Here we change the value of earning column

es by name

ut_data / employee_earnings /

Last Modified

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

6 hours ago

ngs_da...

6 hours ago

Untitled.ipynb

Python 3 (ipykernel)

[57]:

df["earnings"]=df["earnings"]+10

[58]:

df

[58]:

| | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password | office_branch | earnings |
|-----|----------------|-----------|-----------------------------|---------------|-----------------|-------------|--------------|-----------|------------------|---------------|----------|
| | K | Goodwin | angelique.goodwin@gmail.com | 1964-05-15 | 2001-03-24 | 471-57-0359 | 212-884-7146 | akgoodwin | z(d-ez%{.@ | Nashua | 6237 |
| | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!*IO | Stanford | 4447 |
| | T | Farris | donald.farris@bellsouth.net | 1958-04-11 | 1979-11-12 | 097-02-3315 | 205-959-7879 | dtfarris | rX.F{j&j&m&&X | Stanford | 6238 |
| | D | Rendon | steven.rendon@gmail.com | 1982-04-04 | 2008-09-18 | 134-98-6566 | 217-858-0054 | sdrendon | a+2:~xj<Gjy | Nashua | 3137 |
| | L | Almanza | jenell.almanza@yahoo.com | 1958-07-01 | 1993-07-14 | 599-92-7345 | 314-893-2590 | jialmanza | Ou7RX{yT | New York | 3940 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | M | Gould | clemente.gould@hotmail.com | 1961-12-31 | 1992-10-02 | 271-17-5467 | 228-485-0919 | cmgould | m1%+0o~h7VlvJ | Stanford | 4062 |
| | K | Roden | chang.roden@yahoo.com | 1988-09-07 | 2010-08-06 | 074-02-9202 | 316-256-7851 | ckroden | 5jRn{G:~58f\$~+S | Nashua | 2896 |

We create a directory and stored a new parquet file

Last Modified

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

6 hours ago

ngs_da...

6 hours ago

[59]:

directory = 'output_data/employee_earnings/earnings_date=2022-02-10'
os.makedirs(directory, exist_ok=True)

[60]:

df.to_parquet('output_data/employee_earnings/earnings_date=2022-02-15/new_dataset.parquet', index=False)

[49]:

df_1=pd.read_parquet("output_data/employee_earnings/earnings_date=2022-02-11/employee_earnings.parquet")

[50]:

df_1.head()

[50]:

| | emp_id | first_name | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password | of |
|---|--------|------------|----------------|-----------|-----------------------------|---------------|-----------------|-------------|--------------|-----------|---------------|----|
| 0 | 526540 | Angelique | K | Goodwin | angelique.goodwin@gmail.com | 1964-05-15 | 2001-03-24 | 471-57-0359 | 212-884-7146 | akgoodwin | z(d-ez%{.@ | |
| 1 | 859327 | Jeni | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!*IO | |
| 2 | 887387 | Donald | T | Farris | donald.farris@bellsouth.net | 1958-04-11 | 1979-11-12 | 097-02-3315 | 205-959-7879 | dtfarris | rX.F{j&j&m&&X | |
| 3 | 779497 | Steven | D | Rendon | steven.rendon@gmail.com | 1982-04-04 | 2008-09-18 | 134-98-6566 | 217-858-0054 | sdrendon | a+2:~xj<Gjy | |
| 4 | 896517 | Jenell | L | Almanza | jenell.almanza@yahoo.com | 1958-07-01 | 1993-07-14 | 599-92-7345 | 314-893-2590 | jialmanza | Ou7RX{yT | |

[51]:

df_1["earnings"]=df_1["earnings"]+11

[52]:

df_1

Last Modified

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

4 days ago

ngs_da...

6 hours ago

ngs_da...

6 hours ago

[52]:

df_1

[52]:

| | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password | office_branch | earnings |
|-----|----------------|-----------|------------------------------|---------------|-----------------|-------------|--------------|-----------|------------------|---------------|----------|
| | K | Goodwin | angelique.goodwin@gmail.com | 1964-05-15 | 2001-03-24 | 471-57-0359 | 212-884-7146 | akgoodwin | z(d-ez%{.@ | Nashua | 6107 |
| | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!*IO | Stanford | 4294 |
| | T | Farris | donald.farris@bellsouth.net | 1958-04-11 | 1979-11-12 | 097-02-3315 | 205-959-7879 | dtfarris | rX.F{j&j&m&&X | Stanford | 3449 |
| | D | Rendon | steven.rendon@gmail.com | 1982-04-04 | 2008-09-18 | 134-98-6566 | 217-858-0054 | sdrendon | a+2:~xj<Gjy | Nashua | 6236 |
| | L | Almanza | jenell.almanza@yahoo.com | 1958-07-01 | 1993-07-14 | 599-92-7345 | 314-893-2590 | jialmanza | Ou7RX{yT | New York | 5159 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | M | Gould | clemente.gould@hotmail.com | 1961-12-31 | 1992-10-02 | 271-17-5467 | 228-485-0919 | cmgould | m1%+0o~h7VlvJ | Stanford | 5277 |
| | K | Roden | chang.roden@yahoo.com | 1988-09-07 | 2010-08-06 | 074-02-9202 | 316-256-7851 | ckroden | 5jRn{G:~58f\$~+S | Nashua | 2226 |
| | R | Nickel | marvin.nickel@ibm.com | 1986-11-25 | 2012-10-06 | 552-99-5545 | 270-750-7760 | mrrnickel | 8*E[g_-X] | Scranton | 6364 |
| | Y | Tribble | eldora.tribble@earthlink.net | 1995-05-29 | 2016-10-17 | 763-12-2092 | 236-584-1916 | eytribble | z>ms7;\$8-u | Nashua | 8916 |

[1] .

⊗ Failed