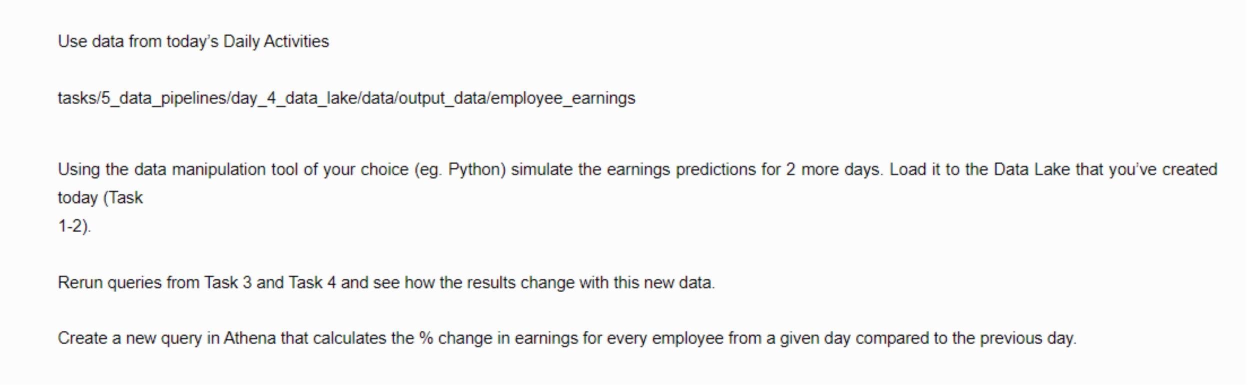
***Name: Osama Abdul Razzak(2303.KHI.DEG.029)***
***Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)***
***Peer Name: M Humza Moeen(2303.KHI.DEG.019)***

# Assignment 5.4



Use data from today's Daily Activities

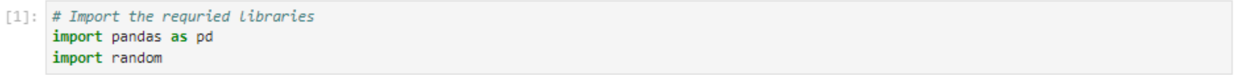
tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings

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).

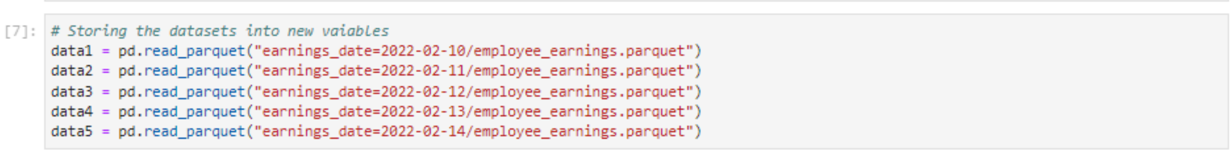
Rerun queries from Task 3 and Task 4 and see how the results change with this new data.

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

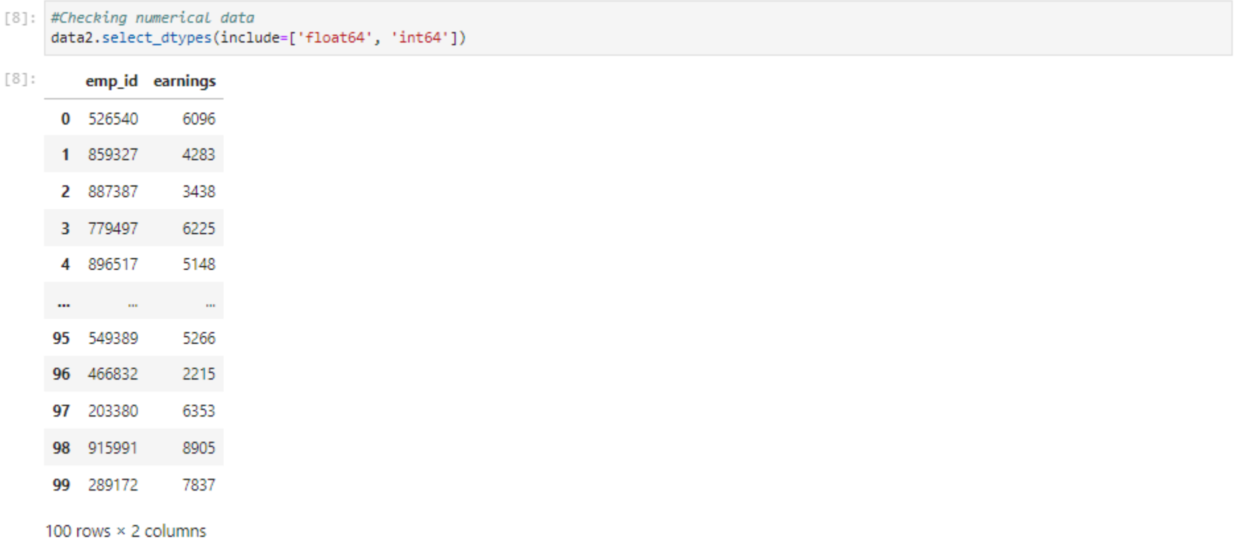
I choose pandas as data manupulation tool, for simulating the prediction for two more days



```
[1]: # Import the requried libraries
     import pandas as pd
     import random
```



```
[7]: # Storing the datasets into new vaiables
     data1 = pd.read_parquet("earnings_date=2022-02-10/employee_earnings.parquet")
     data2 = pd.read_parquet("earnings_date=2022-02-11/employee_earnings.parquet")
     data3 = pd.read_parquet("earnings_date=2022-02-12/employee_earnings.parquet")
     data4 = pd.read_parquet("earnings_date=2022-02-13/employee_earnings.parquet")
     data5 = pd.read_parquet("earnings_date=2022-02-14/employee_earnings.parquet")
```



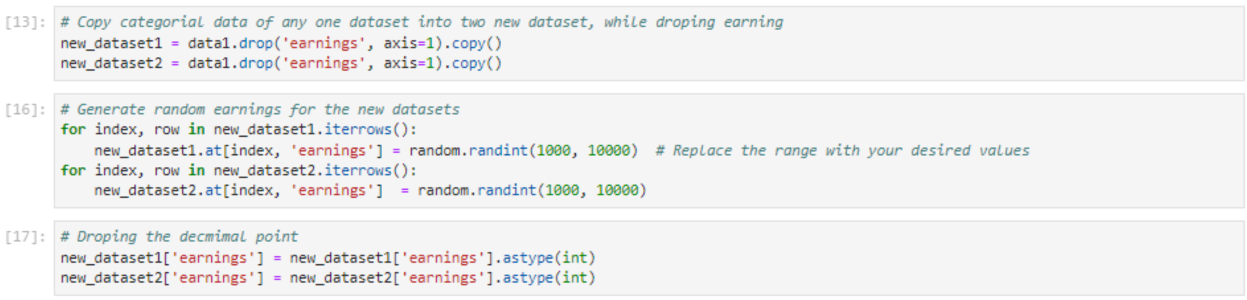
```
[8]: #Checking numerical data
     data2.select_dtypes(include=['float64', 'int64'])
```

[8]:

| | emp_id | earnings |
|---|---|---|
| **0** | 526540 | 6096 |
| **1** | 859327 | 4283 |
| **2** | 887387 | 3438 |
| **3** | 779497 | 6225 |
| **4** | 896517 | 5148 |
| **...** | ... | ... |
| **95** | 549389 | 5266 |
| **96** | 466832 | 2215 |
| **97** | 203380 | 6353 |
| **98** | 915991 | 8905 |
| **99** | 289172 | 7837 |

100 rows × 2 columns

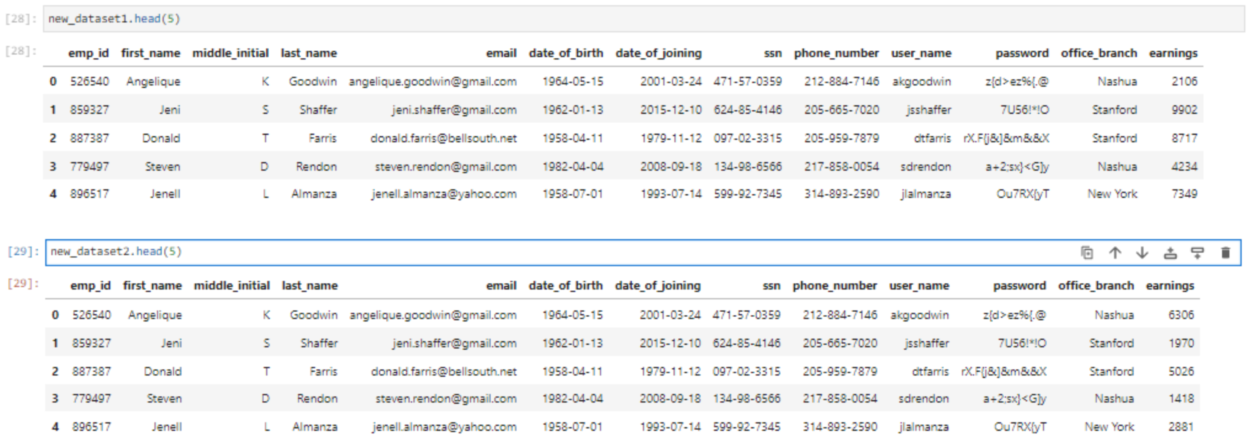
***Name: Osama Abdul Razzak(2303.KHI.DEG.029)***
***Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)***
***Peer Name: M Humza Moeen(2303.KHI.DEG.019)***



```
[13]: # Copy categorial data of any one dataset into two new dataset, while droping earning
      new_dataset1 = data1.drop('earnings', axis=1).copy()
      new_dataset2 = data1.drop('earnings', axis=1).copy()

[16]: # Generate random earnings for the new datasets
      for index, row in new_dataset1.iterrows():
          new_dataset1.at[index, 'earnings'] = random.randint(1000, 10000)  # Replace the range with your desired values
      for index, row in new_dataset2.iterrows():
          new_dataset2.at[index, 'earnings']  = random.randint(1000, 10000)

[17]: # Droping the decmimal point
      new_dataset1['earnings'] = new_dataset1['earnings'].astype(int)
      new_dataset2['earnings'] = new_dataset2['earnings'].astype(int)
```



```
[28]: new_dataset1.head(5)
```

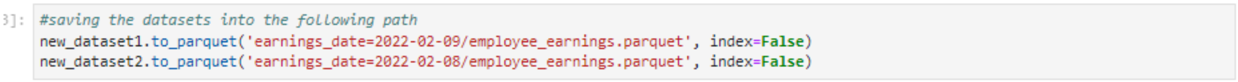
[28]:

| | emp_id | first_name | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password | office_branch | earnings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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%{.@ | Nashua | 2106 |
| **1** | 859327 | Jeni | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!*!O | Stanford | 9902 |
| **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 | Stanford | 8717 |
| **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}<G}y | Nashua | 4234 |
| **4** | 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 | 7349 |

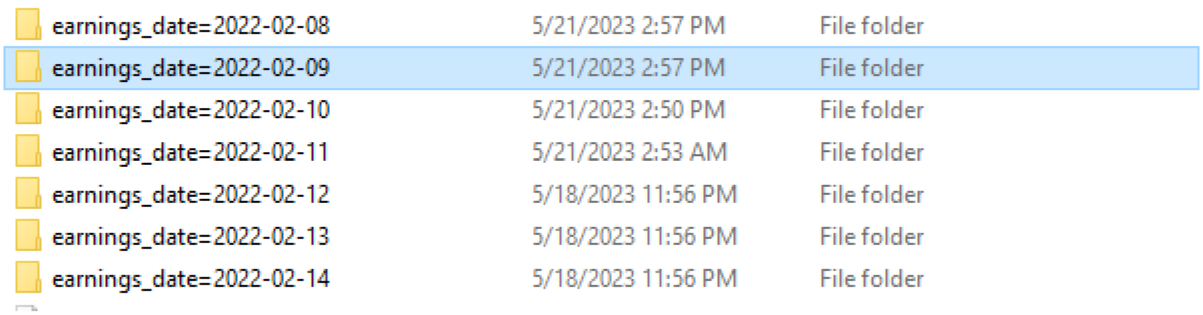
```
[29]: new_dataset2.head(5)
```

[29]:

| | emp_id | first_name | middle_initial | last_name | email | date_of_birth | date_of_joining | ssn | phone_number | user_name | password | office_branch | earnings |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **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%{.@ | Nashua | 6306 |
| **1** | 859327 | Jeni | S | Shaffer | jeni.shaffer@gmail.com | 1962-01-13 | 2015-12-10 | 624-85-4146 | 205-665-7020 | jsshaffer | 7U56!*!O | Stanford | 1970 |
| **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 | Stanford | 5026 |
| **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}<G}y | Nashua | 1418 |
| **4** | 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 | 2881 |



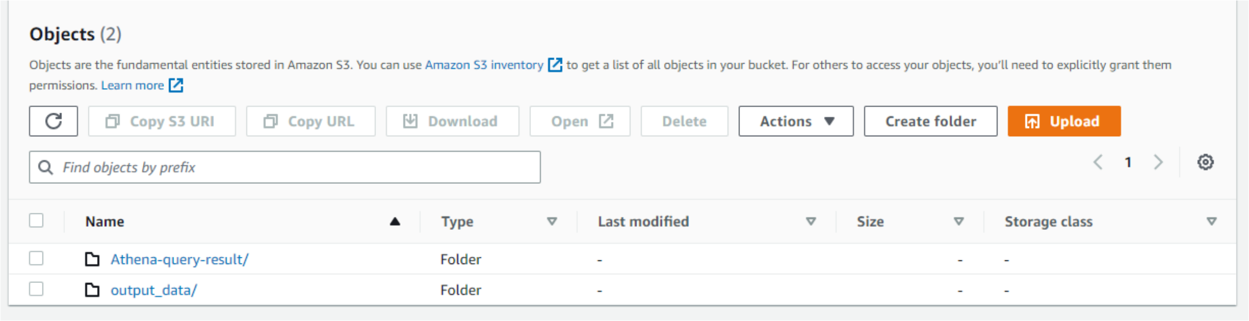
```
3]: #saving the datasets into the following path
    new_dataset1.to_parquet('earnings_date=2022-02-09/employee_earnings.parquet', index=False)
    new_dataset2.to_parquet('earnings_date=2022-02-08/employee_earnings.parquet', index=False)
```



| | | |
|---|---|---|
| earnings_date=2022-02-08 | 5/21/2023 2:57 PM | File folder |
| earnings_date=2022-02-09 | 5/21/2023 2:57 PM | File folder |
| earnings_date=2022-02-10 | 5/21/2023 2:50 PM | File folder |
| earnings_date=2022-02-11 | 5/21/2023 2:53 AM | File folder |
| earnings_date=2022-02-12 | 5/18/2023 11:56 PM | File folder |
| earnings_date=2022-02-13 | 5/18/2023 11:56 PM | File folder |
| earnings_date=2022-02-14 | 5/18/2023 11:56 PM | File folder |

*Name: Osama Abdul Razzak(2303.KHI.DEG.029)*
*Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)*
*Peer Name: M Humza Moeen(2303.KHI.DEG.019)*

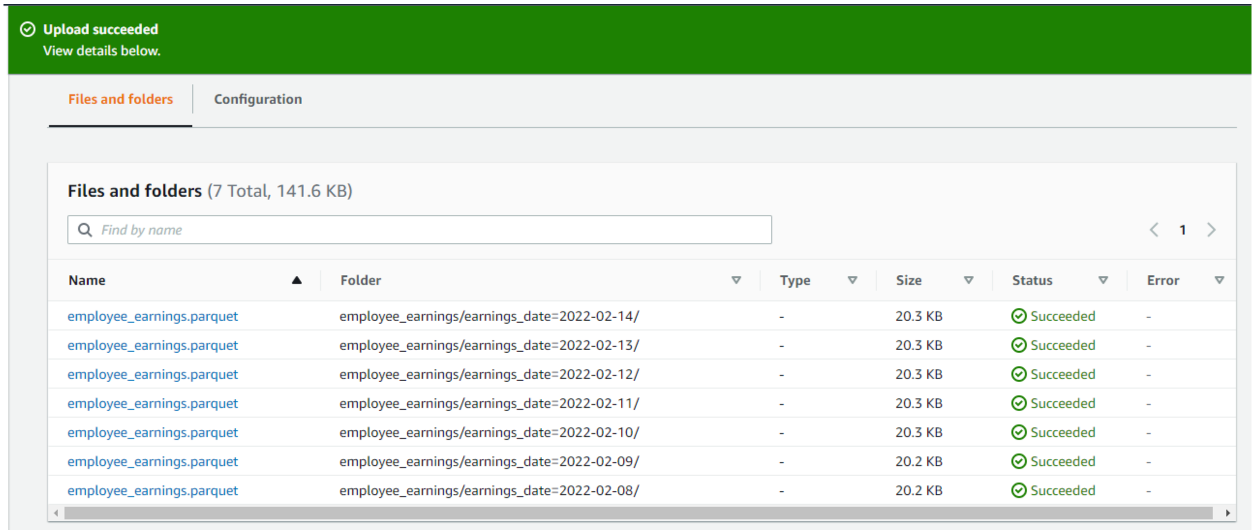
Now putting the dataset into s3 bucket and for saving query we create another folder 'Athena query result'



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

| Name | Type | Last modified | Size | Storage class |
|---|---|---|---|---|
| Athena-query-result/ | Folder | - | - | - |
| output_data/ | Folder | - | - | - |

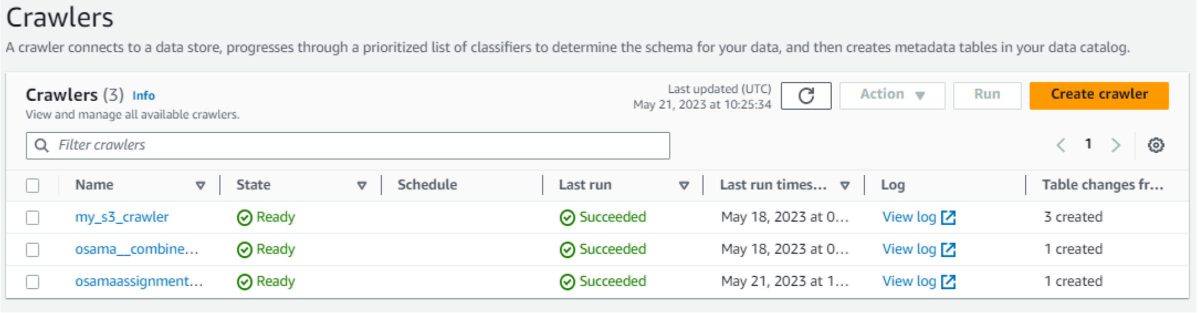


Upload succeeded
View details below.

Files and folders (7 Total, 141.6 KB)

| Name | Folder | Type | Size | Status | Error |
|---|---|---|---|---|---|
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-14/ | - | 20.3 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-13/ | - | 20.3 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-12/ | - | 20.3 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-11/ | - | 20.3 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-10/ | - | 20.3 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-09/ | - | 20.2 KB | Succeeded | - |
| employee_earnings.parquet | employee_earnings/earnings_date=2022-02-08/ | - | 20.2 KB | Succeeded | - |

Then we create crawler name "osamaassignment_combined_earning_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 (3)

View and manage all available crawlers.

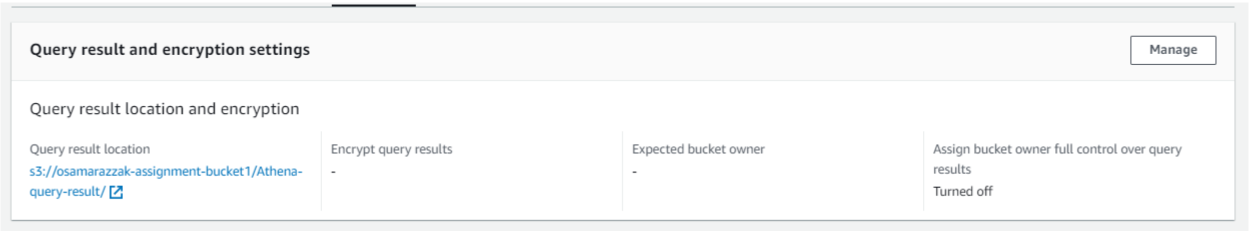
Last updated (UTC)
May 21, 2023 at 10:25:34

| Name | State | Schedule | Last run | Last run times... | Log | Table changes fr... |
|---|---|---|---|---|---|---|
| my_s3_crawler | Ready | | Succeeded | May 18, 2023 at 0... | View log | 3 created |
| osama__combine... | Ready | | Succeeded | May 18, 2023 at 0... | View log | 1 created |
| osamaassignment... | Ready | | Succeeded | May 21, 2023 at 1... | View log | 1 created |

***Name: Osama Abdul Razzak(2303.KHI.DEG.029)***
***Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)***
***Peer Name: M Humza Moeen(2303.KHI.DEG.019)***

Now, we move to Athena

First, we provide the for saving our query result

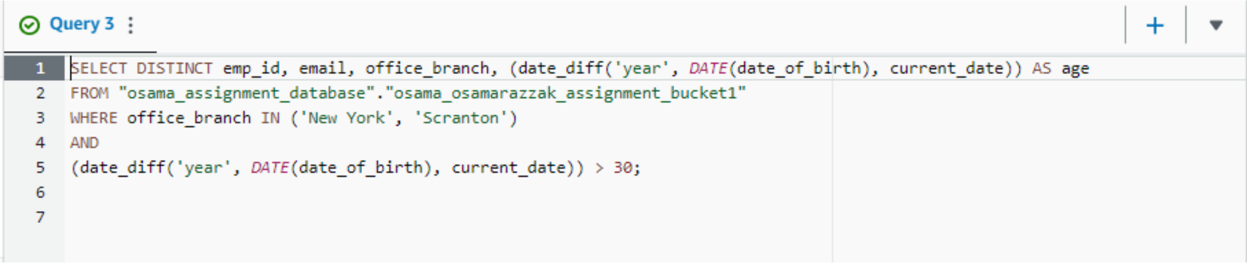


**Query result and encryption settings**

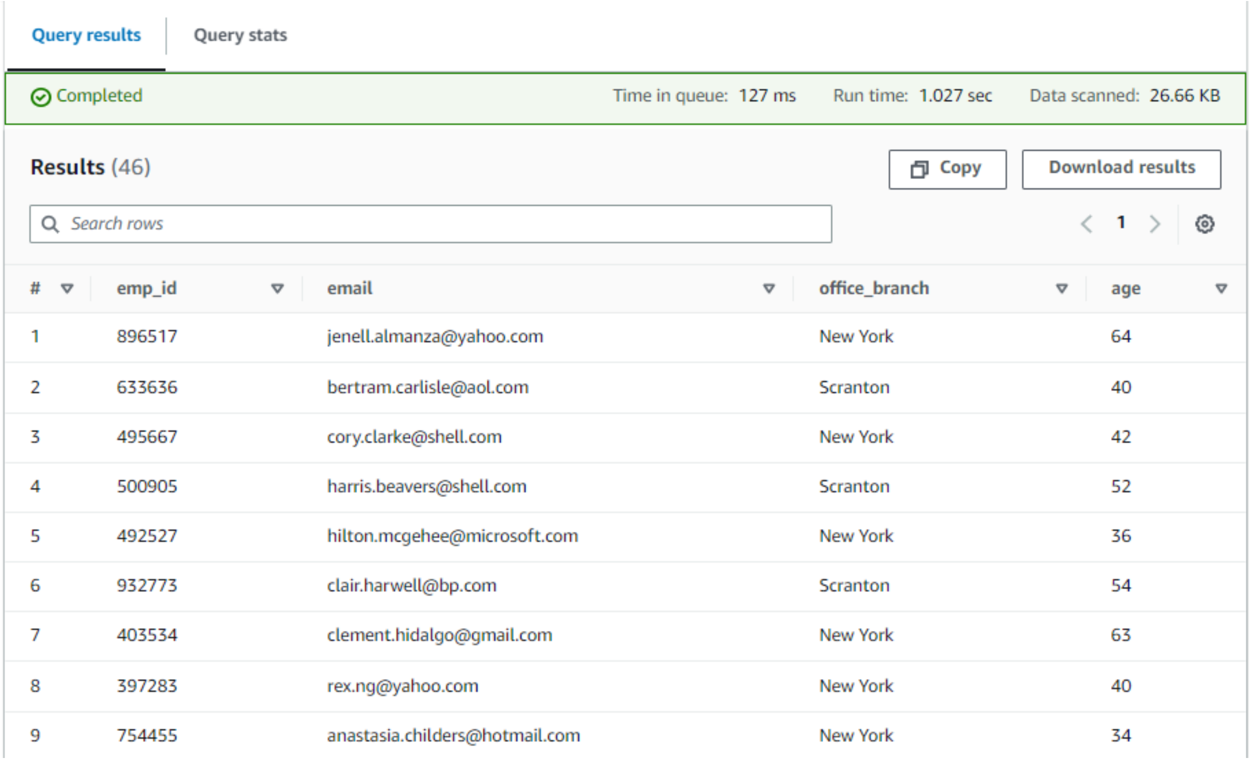
Query result location and encryption

| Query result location | Encrypt query results | Expected bucket owner | Assign bucket owner full control over query results |
|---|---|---|---|
| s3://osamarazzak-assignment-bucket1/Athena-query-result/ | - | - | Turned off |

## Query # 1



```
SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
FROM "osama_assignment_database"."osama_osamarazzak_assignment_bucket1"
WHERE office_branch IN ('New York', 'Scranton')
AND
(date_diff('year', DATE(date_of_birth), current_date)) > 30;
```



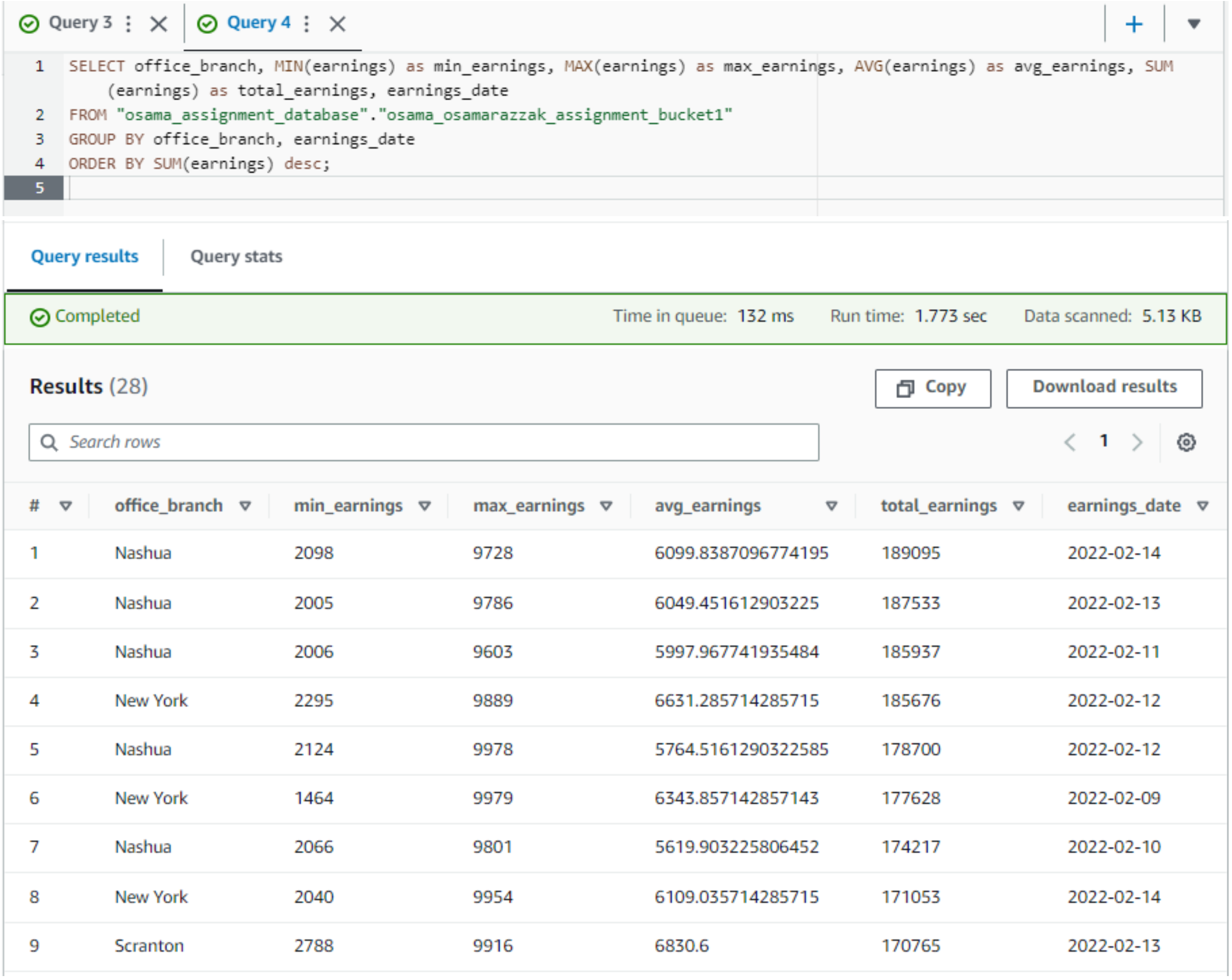
Completed    Time in queue: 127 ms    Run time: 1.027 sec    Data scanned: 26.66 KB

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 | 397283 | rex.ng@yahoo.com | New York | 40 |
| 9 | 754455 | anastasia.childers@hotmail.com | New York | 34 |

***Name: Osama Abdul Razzak(2303.KHI.DEG.029)***<br>
***Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)***<br>
***Peer Name: M Humza Moeen(2303.KHI.DEG.019)***

## Query # 2



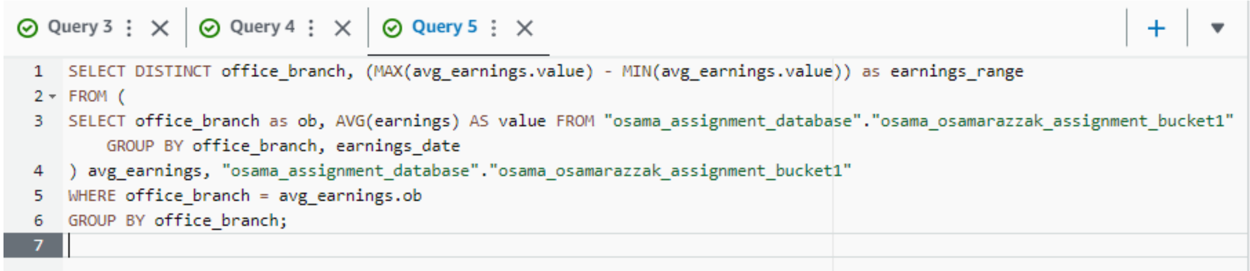
```sql
SELECT office_branch, MIN(earnings) as min_earnings, MAX(earnings) as max_earnings, AVG(earnings) as avg_earnings, SUM
    (earnings) as total_earnings, earnings_date
FROM "osama_assignment_database"."osama_osamarazzak_assignment_bucket1"
GROUP BY office_branch, earnings_date
ORDER BY SUM(earnings) desc;
```

Completed  Time in queue: 132 ms  Run time: 1.773 sec  Data scanned: 5.13 KB

Results (28)

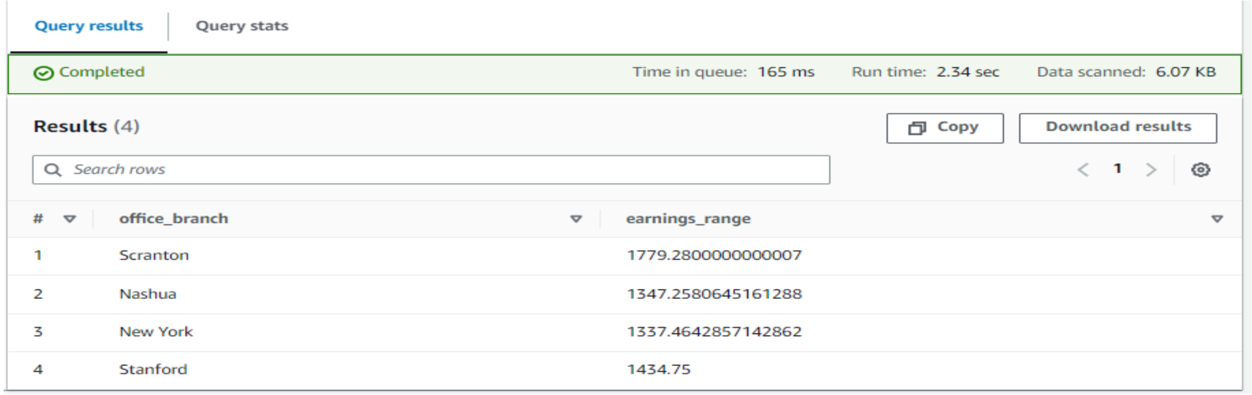
| # | office_branch | min_earnings | max_earnings | avg_earnings | total_earnings | earnings_date |
|---|---|---|---|---|---|---|
| 1 | Nashua | 2098 | 9728 | 6099.8387096774195 | 189095 | 2022-02-14 |
| 2 | Nashua | 2005 | 9786 | 6049.451612903225 | 187533 | 2022-02-13 |
| 3 | Nashua | 2006 | 9603 | 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 | New York | 1464 | 9979 | 6343.857142857143 | 177628 | 2022-02-09 |
| 7 | Nashua | 2066 | 9801 | 5619.903225806452 | 174217 | 2022-02-10 |
| 8 | New York | 2040 | 9954 | 6109.035714285715 | 171053 | 2022-02-14 |
| 9 | Scranton | 2788 | 9916 | 6830.6 | 170765 | 2022-02-13 |

## Query # 3



```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 "osama_assignment_database"."osama_osamarazzak_assignment_bucket1"
    GROUP BY office_branch, earnings_date
) avg_earnings, "osama_assignment_database"."osama_osamarazzak_assignment_bucket1"
WHERE office_branch = avg_earnings.ob
GROUP BY office_branch;
```

***Name: Osama Abdul Razzak(2303.KHI.DEG.029)***
***Peer Name: Rahima Siddiqui(2303.KHI.DEG.030)***
***Peer Name: M Humza Moeen(2303.KHI.DEG.019)***

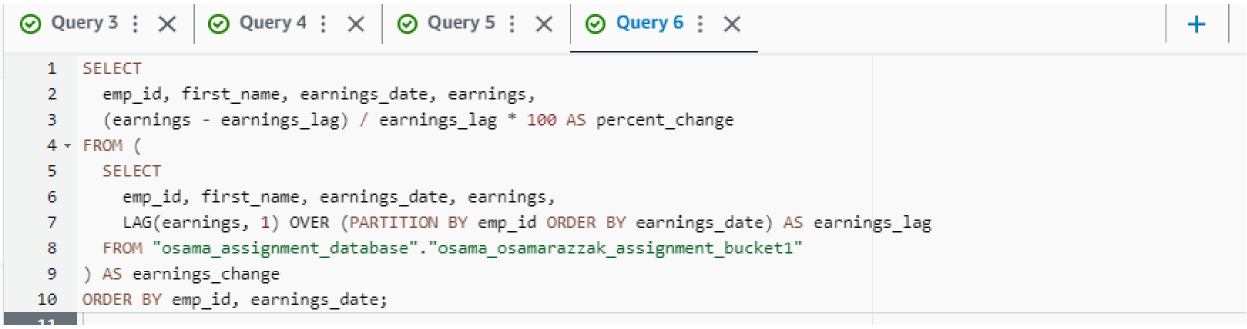


Completed Time in queue: 165 ms Run time: 2.34 sec Data scanned: 6.07 KB

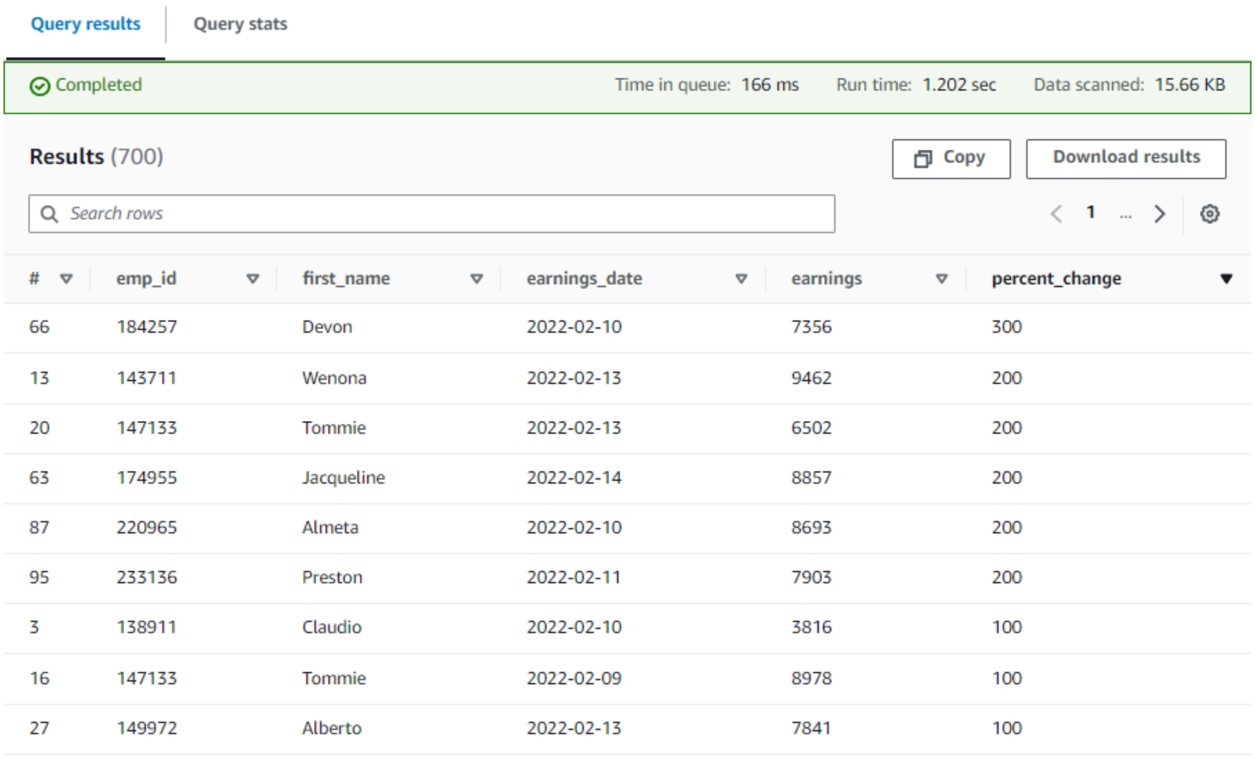
**Results** (4)

| # | office_branch | earnings_range |
|---|---|---|
| 1 | Scranton | 1779.2800000000007 |
| 2 | Nashua | 1347.2580645161288 |
| 3 | New York | 1337.4642857142862 |
| 4 | Stanford | 1434.75 |

## Query # 4



```
SELECT
  emp_id, first_name, earnings_date, earnings,
  (earnings - earnings_lag) / earnings_lag * 100 AS percent_change
FROM (
  SELECT
    emp_id, first_name, earnings_date, earnings,
    LAG(earnings, 1) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS earnings_lag
  FROM "osama_assignment_database"."osama_osamarazzak_assignment_bucket1"
) AS earnings_change
ORDER BY emp_id, earnings_date;
```



Completed Time in queue: 166 ms Run time: 1.202 sec Data scanned: 15.66 KB

**Results** (700)

| # | emp_id | first_name | earnings_date | earnings | percent_change |
|---|---|---|---|---|---|
| 66 | 184257 | Devon | 2022-02-10 | 7356 | 300 |
| 13 | 143711 | Wenona | 2022-02-13 | 9462 | 200 |
| 20 | 147133 | Tommie | 2022-02-13 | 6502 | 200 |
| 63 | 174955 | Jacqueline | 2022-02-14 | 8857 | 200 |
| 87 | 220965 | Almeta | 2022-02-10 | 8693 | 200 |
| 95 | 233136 | Preston | 2022-02-11 | 7903 | 200 |
| 3 | 138911 | Claudio | 2022-02-10 | 3816 | 100 |
| 16 | 147133 | Tommie | 2022-02-09 | 8978 | 100 |
| 27 | 149972 | Alberto | 2022-02-13 | 7841 | 100 |