

## Assignment 5.4 Submitted by Ali Nasir(2023-KHI-DEG-032) & Saif ur Rehman(2023-KHI-DEG-007):

### Use data from today's Daily Activities

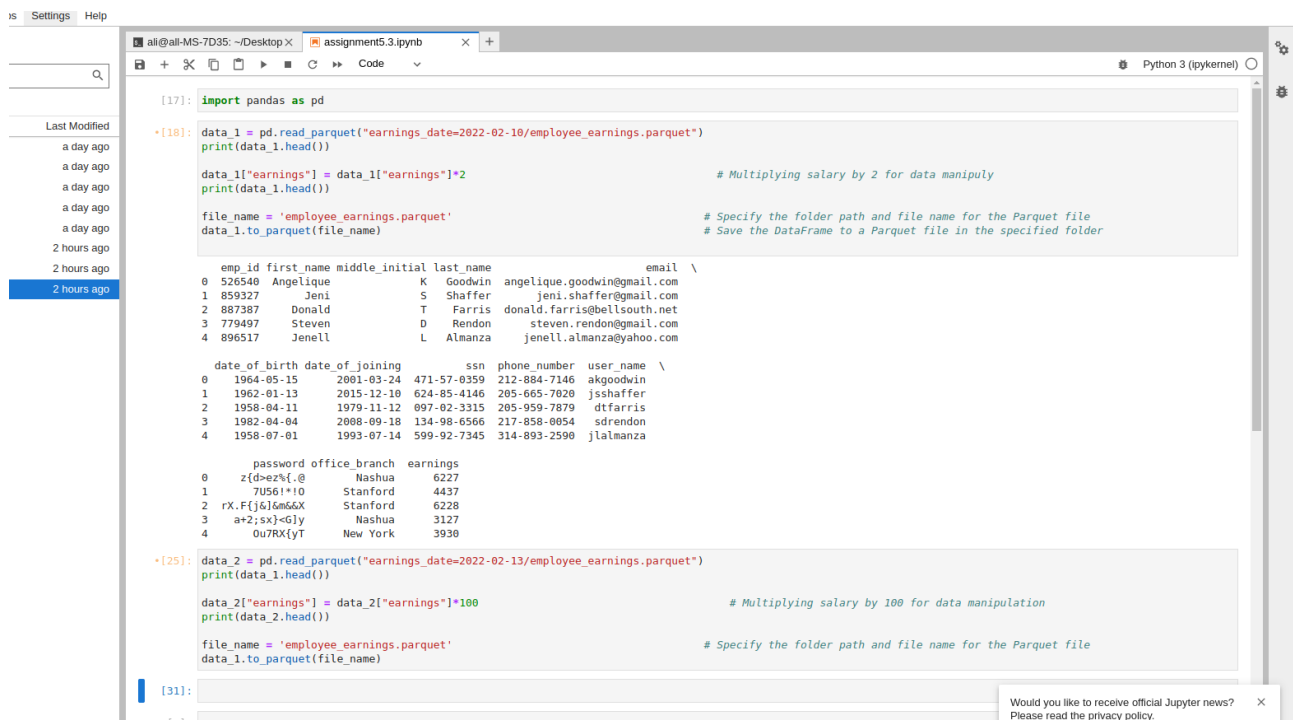
tasks/5\_data\_pipelines/day\_4\_data\_lake/data/output\_data/employee\_earnings

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.

### Solution:

#### Daily task:

- Here is the python script to manipulate the give data salaries.



```
[17]: import pandas as pd

* [18]: data_1 = pd.read_parquet("earnings_date=2022-02-10/employee_earnings.parquet")
print(data_1.head())

data_1["earnings"] = data_1["earnings"]*2
print(data_1.head())

file_name = 'employee_earnings.parquet'
data_1.to_parquet(file_name)

# Multiplying salary by 2 for data manipuly
# Specify the folder path and file name for the Parquet file
# Save the DataFrame to a Parquet file in the specified folder

emp_id first_name middle_initial last_name email \
0 526540 Angelique K Goodwin angelique.goodwin@gmail.com
1 859327 Jeni S Shaffer jeni.shaffer@gmail.com
2 887387 Donald T Farris donald.farris@bellsouth.net
3 779497 Steven D Rendon steven.rendon@gmail.com
4 896517 Jenell L Almanza jenell.almanza@yahoo.com

date_of_birth date_of_joining ssn phone_number user_name \
0 1964-05-15 2001-03-24 471-57-0359 212-884-7146 akgoodwin
1 1962-01-13 2015-12-10 624-85-4146 285-665-7020 jsshaffer
2 1958-04-11 1979-11-12 097-02-3315 205-959-7879 dtfarris
3 1982-04-04 2008-09-18 134-98-6566 217-858-0854 sdrendon
4 1958-07-01 1993-07-14 599-92-7345 314-893-2590 jlalmanza

password office_branch earnings
0 z{d>e2%{.@ Nashua 6227
1 7U56!*!0 Stanford 4437
2 rX.F{j&j&6&6&X Stanford 6228
3 a+2;sx>Gly Nashua 3127
4 0u7RX{yT New York 3930

* [25]: data_2 = pd.read_parquet("earnings_date=2022-02-13/employee_earnings.parquet")
print(data_2.head())

data_2["earnings"] = data_2["earnings"]*100
print(data_2.head())

file_name = 'employee_earnings.parquet'
data_2.to_parquet(file_name)

# Multiplying salary by 100 for data manipulation
# Specify the folder path and file name for the Parquet file

[31]:
```

Would you like to receive official Jupyter news?  
Please read the privacy policy.

- Now we have tried to implement the complex query on the new dataset it is showing me the error because it can not run complex query. Where as these query are successfully run in the athena .

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](#).

Add SQL from templates
Run SQL query

```

1 /* 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 */
2
3
4 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
5 FROM 's3://alinasir-modules-day4/output_data/employee_earnings/earnings_date=2022-02-11/employee_earnings.parquet'
6 WHERE office_branch IN ('New York', 'Scranton')
7 AND
8 (date_diff('year', DATE(date_of_birth), current_date)) > 30;

```

⊗ Unexpected keyword found, KEYWORD:UNKNOWN at line 1, column 20.

### 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.

Download results

Status

⊗ Failed

Close

d queries
Settings
Workgroup: primary

Query 2
Query 3
Query 4
Query 5
Query 6

```

1 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
2 FROM "alinasir_crawler_job"."alinasiroutput_data"
3 WHERE office_branch IN ('New York', 'Scranton')
4 AND
5 (date_diff('year', DATE(date_of_birth), current_date)) > 30;

```

SQL Ln 5, Col 61

Run again
Explain
Cancel
Clear
Create

Reuse query results  
\*Athena engine version 3 only

Query results
Query stats

Completed
Time in queue: 193 ms
Run time: 975 ms
Data scanned: 26.66 KB

Results (46)

Search rows

Copy
Download results

#	emp_id	email	office_branch	age
1	654617	rogelio.woodall@gmail.com	New York	50
2	138911	claudio.heck@aol.com	Scranton	55
3	713294	sammy.dewitt@ibm.com	Scranton	35
4	312726	celine.lumpkin@gmail.com	New York	36
5	551149	michale.colson@comcast.net	Scranton	61
6	402180	allan.bernhardt@gmail.com	New York	61
7	900756	benjamin.doss@gmail.com	Scranton	38

Query 4

```

1
2 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
3 FROM "alinasir_crawler_job"."alinasiroutput_data"
4 GROUP BY office_branch, earnings_date
5 ORDER BY SUM(earnings) desc;
6
7
8

```

SQL Ln 1, Col 1

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 151 ms Run time: 1.105 sec Data scanned: 5.28 KB

Results (28)

Search rows

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_date
1	Nashua	4132	19602	11239.806451612903	348434	2022-02-28
2	Nashua	4132	19602	11239.806451612903	348434	2022-02-9
3	New York	4752	19944	11982.642857142857	335514	2022-02-28
4	New York	4752	19944	11982.642857142857	335514	2022-02-9
5	Scranton	4066	19776	12011.12	300278	2022-02-28
6	Scranton	4066	19776	12011.12	300278	2022-02-9
7	Nashua	2098	9728	6099.8387096774195	189095	2022-02-14

Query 5

```

1 SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
2 FROM (
3 SELECT office_branch as ob, AVG(earnings) AS value FROM "alinasir_crawler_job"."alinasiroutput_data" GROUP BY office_branch, earnings_date
4 ) avg_earnings, "alinasir_crawler_job"."alinasiroutput_data"
5 WHERE office_branch = avg_earnings.ob
6 GROUP BY office_branch;

```

SQL Ln 4, Col 17

Run again Explain Cancel Clear Create

Query results Query stats

Completed Time in queue: 179 ms Run time: 1.083 sec Data scanned: 6.22 KB

Results (4)

Search rows

#	office_branch	earnings_range
1	Scranton	6959.800000000001
2	Stanford	6118.125
3	New York	6367.107142857142
4	Nashua	5619.903225806452

- The final task that we want to perform was to write the query to calculate the change of percentage.

Queries

Settings

Workgroupprimary

Query 2Query 3Query 4Query 5Query 6

1 - WITH earnings\_cte AS (  
2     SELECT  
3         emp\_id,  
4         earnings\_date,  
5         earnings,  
6         LAG(earnings) OVER (PARTITION BY emp\_id ORDER BY earnings\_date) AS previous\_earnings  
7     FROM  
8         "alinasir\_crawler\_job"."alinasiroutput\_data"  
9     WHERE  
10         earnings\_date >= '2022-02-10' -- Specify the given day  
11     )  
12     SELECT  
13         emp\_id,  
14         earnings\_date,  
15         earnings,  
16         previous\_earnings,  
17         (earnings - previous\_earnings) / CAST(previous\_earnings AS double) \* 100 AS percentage\_change  
18  
19     FROM  
20         earnings\_cte  
21     WHERE  
22         previous\_earnings IS NOT NULL  
23     ORDER BY  
24         emp\_id, earnings\_date;  
25  
26     SQL   Ln 15, Col 12

Run again

Explain

Cancel

Clear

Create

Query results

Query stats

Completed

Time in queue: 262 ms   Run time: 1.988 sec   Data scanned: 9.20 KB

Results (600)

Copy

Download results

Search rows

< 1 >

#	emp_id	earnings_date	earnings	previous_earnings	percentage_change
1	138911	2022-02-11	2199	3816	-42.374213836477985
2	138911	2022-02-12	3984	2199	81.17326057298773