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)

Question

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

In [1]:

```
#Importing required libraries

import pandas as pd
import numpy as np
```

In [2]:

```
#loading parquet formate data in df using pandas
df= pd.read_parquet("employee_earnings.parquet")
```

In [3]:

```
df
```

Out[3]:

	emp_id	first_name	middle_initial	last_name	email	date_of_birth
0	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01
...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06

100 rows × 13 columns

In [4]:

```
def generate_data(a):
        # Calculate the minimum and average values of the 'earnings' column
    min_values = df['earnings'].min()
    average_values = df['earnings'].mean()

    for i in range(a):
        # Generate random values for the 'earnings' column within the specified r
        df['earnings'] = np.random.randint(min_values, average_values, size=df.sh
        # Create a filename for each iteration using string formatting
        filename = f"dataframe{i}.parquet"  # Use string formatting to create the
        # Save the DataFrame as a Parquet file with the generated filename
        df.to_parquet(filename)
        return df # Return the DataFrame (optional)
```

In [5]:

```
print(generate_data(2)) # Calling the function and passing the desired number of
```

```
    emp_id first_name middle_initial last_name
email  \
0   526540  Angelique              K   Goodwin   angelique.goodwin@g
mail.com
1   859327       Jeni              S   Shaffer        jeni.shaffer@g
mail.com
2   887387     Donald              T    Farris   donald.farris@bells
outh.net
3   779497     Steven              D    Rendon       steven.rendon@g
mail.com
4   896517     Jenell              L   Almanza      jenell.almanza@y
ahoo.com
..     ...        ...            ...       ...
...
95  549389   Clemente              M     Gould    clemente.gould@hot
mail.com
96  466832      Chang              K     Roden          chang.roden@y
ahoo.com
97  203380     Marvin              R    Nickel          marvin.nickel
@ibm.com
98  915991     Eldora              Y   Tribble  eldora.tribble@earth
link.net
99  289172      Azzie              L    Layman    azzie.layman@hotma
il.co.uk

   date_of_birth date_of_joining          ssn  phone_number  user_na
me  \
0     1964-05-15      2001-03-24  471-57-0359  212-884-7146  akgoodw
in
1     1962-01-13      2015-12-10  624-85-4146  205-665-7020  jsshaff
er
2     1958-04-11      1979-11-12  097-02-3315  205-959-7879   dtfarr
is
3     1982-04-04      2008-09-18  134-98-6566  217-858-0054   sdrend
on
4     1958-07-01      1993-07-14  599-92-7345  314-893-2590  jlalman
za
..           ...             ...          ...           ...
...
95    1961-12-31      1992-10-02  271-17-5467  228-485-0919    cmgou
ld
96    1988-09-07      2010-08-06  074-02-9202  316-256-7851    ckrod
en
97    1986-11-25      2012-10-06  552-99-5545  270-750-7760   mrnick
el
98    1995-05-29      2016-10-17  763-12-2082  236-584-1916  eytribb
le
99    1961-09-06      2004-03-26  637-29-1007  503-456-5899   allaym
an

           password office_branch  earnings
0        z{d>ez%{.@        Nashua      4738
1          7U56!*!O      Stanford      5364
2     rX.F{j&]&m&&X      Stanford      5618
3       a+2;sx}<G]y        Nashua      2520
4          Ou7RX{yT      New York      5206
..              ...           ...       ...
95    m1%+0ojh7VIvJ      Stanford      5524
96  5|Rn]G:#58f$>+S        Nashua      3293
97        8*E[g-_X|      Scranton      5412
98       z>ms?;$8-u        Nashua      5460
99   k<%?%TML.].1ZY      New York      2134

[100 rows x 13 columns]
```

In [6]:

```
df
```

Out[6]:

	emp_id	first_name	middle_initial	last_name	email	date_of_birth
0	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01
...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06

100 rows × 13 columns