

## 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).

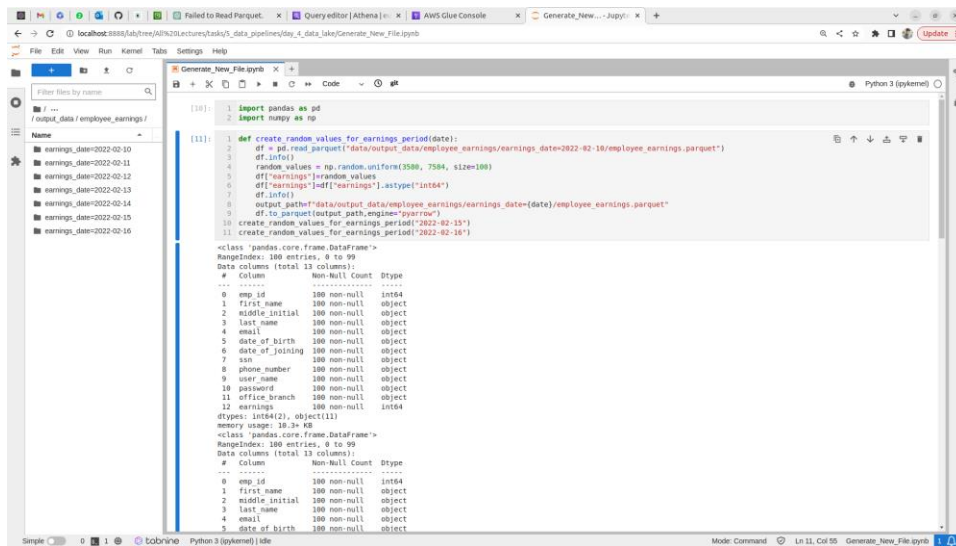
Rerun queries from Task 3 and Task 4 and see how the results change with this new data.

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.



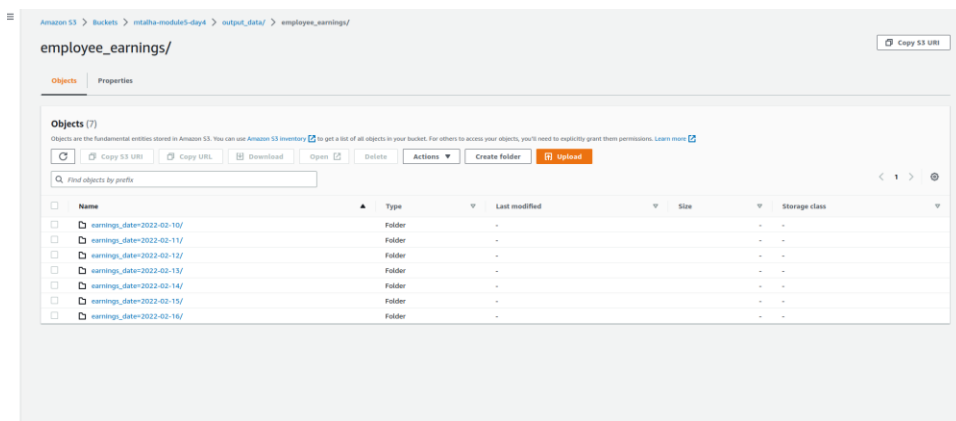
```
[10]: 1 import pandas as pd
2 import numpy as np

[11]: 1 def create_random_values_for_earnings_period(date):
2     df = pd.read_parquet('data/output_data/employee_earnings/earnings_date=2022-02-10/employee_earnings.parquet')
3     df.info()
4     random_values = np.random.uniform(3500, 7500, size=100)
5     df['earnings'] = random_values
6     df['earnings'] = df['earnings'].astype('int64')
7     df.info()
8     output_path = 'data/output_data/employee_earnings/earnings_date=' + date + '/employee_earnings.parquet'
9     df.to_parquet(output_path, engine='pyarrow')
10    create_random_values_for_earnings_period('2022-02-15')
11    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: 30.3+ 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:



## 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	IAM role	Database	State
mt_combined_employee_earnings_crawler	mtalpa-glue-role	mtalpa_glue_database	READY

**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 19, 2023 at 08:01:26	May 19, 2023 at 08:03:28	02 min 02 s	Completed	0.059	1 table change, 5 partition changes
May 19, 2023 at 07:45:46	May 19, 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

Activities Microsoft Edge

ASAP Re... x S3 Man... x emertus... x What is... x S3 Man... x Amazon... x Amazon... x Querye... x emertus... x Categor... x New tab x Settings x

https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-18/query-editor/history/05eaff7b-7af4-4e76-bac6-e4ee7b57d29a

Gmail (26) WhatsApp Poem: Dust if... 4 A.M Study Se... AWS Machine... Google Calendar Fast Style Tran... udacity-deepi... Workshop on... Yes you should... CS231n Winte... Become a tutor

AWS Services Search [All+]

Completed

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

Results (46)

Copy Download results

Search rows

#	emp_id	email	office_branch	age
1	896317	jennell.almanza@yahoo.com	New York	64
2	613636	bertram.carlistead@aol.com	Scranton	40
3	495867	cory.clark@ohail.com	New York	42
4	500905	harris.beavers@ohail.com	Scranton	52
5	482527	hilton.mcgehee@microsoft.com	New York	36
6	512773	clark.harwell@bp.com	Scranton	54
7	403534	clement.hidalgo@gmail.com	New York	63
8	909018	virgil.trowbridge@aol.com	New York	57
9	878666	elida.champagne@gmail.com	Scranton	39
10	496541	witfred.gonzalez@aol.com	Scranton	59
11	976422	jake.espinat@shaw.ca	Scranton	64
12	627298	sterling.berna@hotmail.com	New York	38
13	432165	adalbarto.tate@shaw.ca	New York	41
14	147133	torrence.weller@cox.net	Scranton	63
15	403207	michal.maurer@yahoo.com	Scranton	46
16	397263	rex.ring@yahoo.com	New York	40
17	754455	anastasia.childers@hotmail.com	New York	34

CloudWatch Feedback Language

© 2023 Amazon Web Services, Inc. or its affiliates. Privacy Terms Create performance

Activities Microsoft Edge

ASAP Re... x S3 Man... x emertus... x What is... x S3 Man... x Amazon... x Amazon... x Querye... x emertus... x Categor... x New tab x Settings x

https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-18/query-editor/history/1c6f912-5892-42d3-a948-2e58ff64c85d

Gmail (26) WhatsApp Poem: Dust if... 4 A.M Study Se... AWS Machine... Google Calendar Fast Style Tran... udacity-deepi... Workshop on... Yes you should... CS231n Winte... Become a tutor

AWS Services Search [All+]

Completed

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

Results (20)

Copy Download results

Search rows

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_date
1	Nashua	2088	9728	6099.8387096774195	189095	2022-02-14
2	Nashua	2005	9786	6048.431612603225	187533	2022-02-15
3	Nashua	2006	9605	5997.967741955484	185957	2022-02-11
4	New York	2295	9809	6651.285714285715	185676	2022-02-12
5	Nashua	2124	9978	5764.5161290322565	178700	2022-02-12
6	Nashua	2064	9801	5615.903223806452	174217	2022-02-10
7	New York	2040	9954	6109.635714285715	171055	2022-02-14
8	Scranton	2788	9916	6830.6	170765	2022-02-15
9	New York	2141	9482	5996.178571428572	167949	2022-02-11
10	Nashua	2376	9872	5991.321428571428	167757	2022-02-10
11	New York	2195	9734	5615.535714285715	157235	2022-02-15
12	Scranton	2465	9827	6145.72	153745	2022-02-14
13	Scranton	2025	9846	6065.44	151586	2022-02-12

CloudWatch Feedback Language

© 2023 Amazon Web Services, Inc. or its affiliates. Privacy Terms Create performance

Activities Microsoft Edge

ASAP Re... x S3 Man... x emertus... x What is... x S3 Man... x Amazon... x Amazon... x Querye... x emertus... x Categor... x New tab x Settings x

https://us-east-1.console.aws.amazon.com/athena/home?region=us-east-18/query-editor/history/6dbcb346-07d7-4b83-b09a-22db174d01da

Gmail (26) WhatsApp Poem: Dust if... 4 A.M Study Se... AWS Machine... Google Calendar Fast Style Tran... udacity-deepi... Workshop on... Yes you should... CS231n Winte... Become a tutor

AWS Services Search [All+]

Completed

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

Results (4)

Copy Download results

Search rows

#	office_branch	earnings_range
1	New York	1015.75
2	Stanford	1003.375
3	Nashua	479.9554836709678
4	Scranton	1779.2800000000007

CloudWatch Feedback Language

© 2023 Amazon Web Services, Inc. or its affiliates. Privacy Terms Create performance

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

The first screenshot shows the AWS Glue Console interface with a query editor. The query is: `SELECT * FROM "mtahta_glu_database"."mtahtemployee_earnings";`. The query is completed, and the results are displayed in a table with 700 rows. The columns are: emp\_id, first\_name, middle\_initial, last\_name, email, date\_of\_birth, date\_of\_joining, ssn, and phone\_number. The results show employee information for various individuals.

#	emp_id	first_name	middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number
1	52840	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-05-24	471-57-0559	212-684-7146
2	859327	Jani	S	Shaffer	jani.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

The second screenshot shows the AWS Glue Console interface with a query editor. The query is: `SELECT * FROM "mtahta_glu_database"."mtahtemployee_earnings";`. The query is completed, and the results are displayed in a table with 46 rows. The columns are: emp\_id, email, office\_branch, and age. The results show employee information for various individuals.

#	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.gonzalez@aol.com	Scranton	59
5	976422	jake.espnal@shaw.ca	Scranton	64
6	627296	sterling.erna@hotmail.com	New York	38
7	622405	harrison.hawk@hotmail.co.uk	Scranton	60
8	452163	adalberto.tate@shaw.ca	New York	41
9	147135	tommie.weller@cox.net	Scranton	63
10	595558	denisha.fluhgimson.com	Scranton	32
11	314661	charles.quintens@gmail.com	New York	65
12	403207	michal.mazurek@yahoo.com	Scranton	46
13	896517	jerrell.almanza@yahoo.com	New York	64
14	635636	bertram.carls@bdl.com	Scranton	40
15	823098	carlton.leclair@cox.net	Scranton	37
16	495667	cory.clark@bellsouth.com	New York	42
17	500905	harris.beavers@bellsouth.com	Scranton	52
18	492527	hilton.mcguhee@microsoft.com	New York	36
19	932773	clar.harwell@tsp.com	Scranton	54

The third screenshot shows the AWS Glue Console interface with a query editor. The query is: `SELECT * FROM "mtahta_glu_database"."mtahtemployee_earnings";`. The query is completed, and the results are displayed in a table with 28 rows. The columns are: office\_branch, min\_earnings, max\_earnings, avg\_earnings, total\_earnings, and earnings\_date. The results show employee information for various individuals.

#	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	9605	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	6850.6	170785	2022-02-15
9	New York	2141	9482	5998.178571428572	167949	2022-02-11
10	Nashua	3729	7541	5413.367096774193	167815	2022-02-16
11	New York	2376	9972	5991.321428571428	167757	2022-02-10
12	Nashua	3740	7453	5373.387096774193	166675	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	155743	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	148795	2022-02-16
19	Scranton	3618	7191	5537.84	135446	2022-02-16
20	Scranton	3749	7538	5267.36	131884	2022-02-15

Amazon S3 > Buckets > intalsha-module5-day4 > athena-query-results/ > Unsaved/ > 2023/ > 05/ > 19/

19/ Copy S3 URL

Objects Properties

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](#)

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

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	01e8ac15-a192-4efc-ba6b-04d90e0d47c.csv	csv	May 19, 2023, 09:24:52 (UTC+05:00)	2.3 KB	Standard
<input type="checkbox"/>	01e8ac15-a192-4efc-ba6b-04d90e0d47c.csv.metadata	metadata	May 19, 2023, 09:24:52 (UTC+05:00)	205.0 B	Standard
<input type="checkbox"/>	05b0b0c1-4853-48b7-ba6b-cf22a40264a.csv	csv	May 19, 2023, 10:39:37 (UTC+05:00)	6.2 KB	Standard
<input type="checkbox"/>	05b0b0c1-4853-48b7-ba6b-cf22a40264a.csv.metadata	metadata	May 19, 2023, 10:39:37 (UTC+05:00)	357.0 B	Standard
<input type="checkbox"/>	05a1f5a0-2538-4c99-a980-c3f25c2b3f52.csv	csv	May 19, 2023, 08:00:43 (UTC+05:00)	82.8 KB	Standard
<input type="checkbox"/>	05a1f5a0-2538-4c99-a980-c3f25c2b3f52.csv.metadata	metadata	May 19, 2023, 08:00:43 (UTC+05:00)	757.0 B	Standard
<input type="checkbox"/>	0b9c3ff1c-48c2-437b-8542-5764ac97fdd.csv	csv	May 19, 2023, 06:12:13 (UTC+05:00)	354.0 B	Standard
<input type="checkbox"/>	0b9c3ff1c-48c2-437b-8542-5764ac97fdd.csv.metadata	metadata	May 19, 2023, 06:12:13 (UTC+05:00)	757.0 B	Standard
<input type="checkbox"/>	0b9c3b01-c420-49c7-acc5-3942258d7410.csv	csv	May 19, 2023, 06:04:38 (UTC+05:00)	354.0 B	Standard
<input type="checkbox"/>	0b9c3b01-c420-49c7-acc5-3942258d7410.csv.metadata	metadata	May 19, 2023, 06:04:38 (UTC+05:00)	757.0 B	Standard
<input type="checkbox"/>	0c76546f-8042-4056-9d8e-2cfad7777f.txt	txt	May 19, 2023, 06:15:39 (UTC+05:00)	133.0 B	Standard
<input type="checkbox"/>	0c76546f-8042-4056-9d8e-2cfad7777f.txt.metadata	metadata	May 19, 2023, 06:15:40 (UTC+05:00)	36.0 B	Standard
<input type="checkbox"/>	116c40ec-8243-473b-bd51-cac3d913427b.csv	csv	May 19, 2023, 09:26:17 (UTC+05:00)	149.0 B	Standard
<input type="checkbox"/>	116c40ec-8243-473b-bd51-cac3d913427b.csv.metadata	metadata	May 19, 2023, 09:26:18 (UTC+05:00)	144.0 B	Standard
<input type="checkbox"/>	15afad42-683b-4a29-a371-96e02af1392c.csv	csv	May 19, 2023, 09:55:40 (UTC+05:00)	4.2 KB	Standard
<input type="checkbox"/>	15afad42-683b-4a29-a371-96e02af1392c.csv.metadata	metadata	May 19, 2023, 09:55:40 (UTC+05:00)	297.0 B	Standard
<input type="checkbox"/>	27c509ab-050d-4807-4c30-96a20674d4a.csv	csv	May 19, 2023, 06:03:17 (UTC+05:00)	35.2 KB	Standard

IntelliJ Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Create preferences

## 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.

Amazon Athena > Query editor

EditorRecent queriesSaved queriesSettings

Workgroupprimary

Data source

AthenaDataCatalog

Database

mtaisha\_glue\_database

Tables and views

Filter tables and views

Tables (5)

employee\_earnings

mtaisha\_earnings

mtaisha\_employee\_earnings

mtaisha\_location\_location

mtaishaemployee\_earnings

Views (0)

1 - WITH current\_data AS (  
2 SELECT emp\_id, earnings, earnings\_date, first\_name, last\_name  
3 FROM "mtaisha\_glue\_database"."mtaishaemployee\_earnings"  
4 WHERE earnings\_date = '2022-02-14' -- Current date  
5 ),  
6 previous\_data AS (  
7 SELECT emp\_id, earnings, earnings\_date  
8 FROM "mtaisha\_glue\_database"."mtaishaemployee\_earnings"  
9 WHERE earnings\_date = '2022-02-13' -- Yesterday's date  
10 )  
11 SELECT  
12 current\_data.emp\_id, current\_data.first\_name, current\_data.last\_name,  
13 current\_data.earnings AS current\_earnings,  
14 previous\_data.earnings AS previous\_earnings,  
15 (current\_data.earnings - previous\_data.earnings) / CAST(previous\_data.earnings AS double) \* 100 AS percentage\_change  
16 FROM  
17 current\_data  
18 JOIN  
19 previous\_data  
20 ON  
21 current\_data.emp\_id = previous\_data.emp\_id

SQLLn 14, Col 47

Run

Explain

Cancel

Clear

Create

Reuse query results

up to 60 minutes ago

Query results

Query state

Query results

Query state

Completed

Time in queue: 125 msRun time: 896 msData scanned: 4.57 KB

Results (100)

Copy

Download results

Search rows

#	emp_id	first_name	last_name	current_earnings	previous_earnings	percentage_change
1	526540	Angelique	Goodwin	2716	2843	-4.467112205416813
2	859327	Jeni	Shaffer	8357	7280	14.793956043956044
3	887387	Donald	Farris	8123	4816	68.66694352159467
4	779497	Steven	Rendon	8297	2466	236.455798064558
5	896517	Jamell	Almanza	2057	8862	-76.78853531934101
6	230965	Almeta	Brookins	9378	6721	39.532807617914
7	721091	Bobbie	Branson	3557	5042	-29.45259817532725
8	633636	Bertram	Carlisle	8353	9527	-10.442800471748686
9	823898	Carlton	Leclair	4977	5753	-13.48861463354217
10	415885	Todd	Slater	7272	3816	90.56603773584906
11	439483	Adalberto	Hazel	4101	5424	19.72196261682243
12	809408	Seymour	Nelson	4843	8081	-40.06929835416409
13	748190	Vicente	Dawkins	6157	5582	71.88721384701284
14	397283	Rex	Ng	8794	2662	230.35311795642372
15	553684	Arlena	Clough	2279	2005	13.665835411471332
16	495667	Cory	Clarke	9462	5357	76.62871009835597

© 2021, Amazon Web Services, Inc. or its affiliates. PrivacyTermsCreate preferences