

# Graded Assignment 5.4

Name: Saad Sameer Khan

Employee#: 2303.KHI.DEG.034

Collaborated with: Mohammad Hamza Asim (2303.KHI.DEG.014)

## Creating data for 2 more days (day 6 & day 7)

We used numpy's random.randint method to generate random earnings data within the range of the maximum and minimum value of the 'earnings' column from one of the day's data. \*ipynb file has been uploaded to git\*

```
Launcher x Assignment_5.4.ipynb x +
100 rows x 13 columns

[27]: 1 df["earnings"].mean()

[27]: 6197.75

[28]: 1 lower_bound = df["earnings"].min()
      2 lower_bound

[28]: 2040

[29]: 1 upper_bound = df["earnings"].max()
      2 upper_bound

[29]: 9954

[30]: 1 np.random.seed(14)
      2 day_6_earnings = np.random.randint(lower_bound, upper_bound, 100)
      3 day_6_earnings

[30]: array([4707, 7504, 3332, 4494, 8511, 5873, 9566, 6882, 8852, 7810, 3272,
          8757, 9696, 7826, 7185, 2511, 9198, 4579, 8728, 6127, 2300, 9346,
          9858, 7672, 6606, 9292, 8317, 9757, 3696, 4189, 8225, 3808, 2271,
          2619, 3783, 7484, 7109, 5285, 5491, 7408, 7150, 8056, 5271, 2817,
          7347, 2067, 7149, 2258, 3875, 4074, 6393, 6004, 6463, 8288, 8887,
          3970, 4133, 4118, 7051, 3641, 5169, 2627, 8389, 8521, 8541, 4869,
          5040, 7495, 3745, 6775, 3341, 4613, 2363, 4149, 6466, 3505, 8646,
          2644, 9446, 5633, 6470, 8572, 4266, 3720, 2525, 8748, 2297, 4227,
          4060, 5993, 6523, 7001, 6148, 9108, 4576, 5513, 7499, 2415, 8744,
          3995])

[31]: 1 np.random.seed(11)
      2 day_7_earnings = np.random.randint(lower_bound, upper_bound, 100)
      3 day_7_earnings

[31]: array([3985, 5815, 7240, 9299, 9545, 6063, 3333, 2372, 7833, 2623, 6986,
          5904, 8805, 7764, 2728, 9624, 5925, 7396, 4725, 7036, 8730, 2180,
          3906, 9245, 7661, 3295, 9676, 5193, 3155, 2448, 4592, 6793, 9787,
          3390, 6819, 3434, 6378, 5816, 9333, 5191, 9816, 5135, 6035, 9022,
          5201, 7523, 7298, 2481, 7187, 6948, 7405, 5254, 5375, 8623, 2375,
          7096, 9888, 2536, 3347, 8432, 3799, 2173, 7853, 9898, 6363, 3763,
          8185, 7291, 8911, 7625, 9788, 5440, 8298, 7832, 8006, 2315, 3985,
          7021, 9298, 2974, 7603, 7433, 7203, 8202, 6645, 8232, 5250, 6439,
          4198, 6296, 3492, 9248, 2304, 8636, 2969, 2425, 3031, 5985, 9588,
          9283])

[32]: 1 day_6_employee_earnings = df.copy()
```

## Uploading new folders to S3 bucket

Next, we uploaded two new folders containing the data created in the previous step, in the output\_data folder in the S3 bucket.

[Amazon S3](#) > [Buckets](#) > [saad-module5-day4](#) > [output\\_data/](#) > [employee\\_earnings/](#) > Upload

### Upload [Info](#)

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

**Files and folders (1 Total, 20.7 KB)**

Remove

Add files

Add folder

All files and folders in this table will be uploaded.

☒

Name

▲

Folder

▼

Type

▼

Size

▼

☒

employee\_earnings.parquet

earnings\_date=2022-02-15/

-

20.7 KB

### Destination

Destination

s3://saad-module5-day4/output\_data/employee\_earnings/

► Destination details

Bucket settings that impact new objects stored in the specified destination.

### Objects (7)

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

< 1 >

⚙

<input type="checkbox"/>	Name	▲	Type	▼	Last modified	▼	Size	▼	Storage class	▼
<input type="checkbox"/>	earnings_date=2022-02-10/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-11/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-12/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-13/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-14/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-15/		Folder		-		-		-	
<input type="checkbox"/>	earnings_date=2022-02-16/		Folder		-		-		-	

# Running crawler

After the new folders were uploaded, we ran the crawler so that it could fetch the new data from the bucket, and update its data catalog, so that we could use the new data in Athena.

### Crawler properties

Name	saad_combined_employee_earnings_crawler	IAM role	<a href="#">saad-glue-role</a>	Database	saad_glue_db	State	READY
Description	-	Security configuration	-	Lake Formation configuration	-	Table prefix	saad_
Maximum table threshold	-						

► Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

### Crawler runs (2)

The list of crawler runs for this crawler.

< 1 > ⚙

	Start time (UTC) ▲	End time (UTC) ▼	Current/last duration ▼	Status ▼	DPU hours ▼	Table changes ▼
<input type="radio"/>	May 18, 2023 at 12:02:29	May 18, 2023 at 12:07:15	04 min 45 s	✔ Completed	0.072	1 table change, 2 partition changes
<input type="radio"/>	May 18, 2023 at 08:06:43	May 18, 2023 at 08:07:47	01 min 04 s	✔ Completed	0.036	1 table change, 5 partition changes

## Re-running queries on updated data

Now we re-ran the queries that were previously run, and observe what changes took place in the results.

### Query 1

*Before:*

Query results

Query stats

Completed

Time in queue: 172 ms

Run time: 958 ms

Data scanned: 19.04 KB

Results (46)

Copy

Download results

Search rows

< 1 >

#	emp_id	email	office_branch	age
1	900756	benjamin.doss@gmail.com	Scranton	38
2	215719	brent.carrillo@aol.com	New York	50
3	530134	mathew.whitfield@gmail.com	New York	36
4	597741	tonya.wilson@aol.com	New York	43
5	391837	cory.hayden@gmail.com	New York	56
6	622405	harrison.hawk@hotmail.co.uk	Scranton	60

*After*

Query results

Query stats

Completed

Time in queue: 158 ms

Run time: 1.422 sec

Data scanned: 26.66 KB

Results (46)

Copy

Download results

Search rows

< 1 >

#	emp_id	email	office_branch	age
1	909018	virgil.trowbridge@aol.com	New York	37
2	878666	elda.champagne@gmail.com	Scranton	39
3	496541	winfred.gonzales@aol.com	Scranton	59
4	976422	jake.espinal@shaw.ca	Scranton	64
5	627298	sterling.erna@hotmail.com	New York	38
6	452163	adalberto.tate@shaw.ca	New York	41

## Query 2

Before:

Query results							Query stats	
✔ Completed							Time in queue: 158 ms	Run time: 858 ms
Data scanned: 3.75 KB								
Results (20)							<a href="#">Copy</a> <a href="#">Download results</a>	
<input type="text" value="Search rows"/>							< 1 > ⚙	
# ▾	office_branch ▾	min_earning s ▾	max_earning s ▾	avg_earnings ▾	total_earning s ▾	earnings_		
1	Nashua	2098	9728	6099.8387096774195	189095	2022-02-		
2	Nashua	2005	9786	6049.451612903225	187533	2022-02-		
3	Nashua	2006	9603	5997.967741935484	185937	2022-02-		
4	New York	2295	9889	6631.285714285715	185676	2022-02-		
5	Nashua	2124	9978	5764.5161290322585	178700	2022-02-		
6	Nashua	2066	9801	5619.903225806452	174217	2022-02-		

After:

Query results							Query stats	
✔ Completed							Time in queue: 213 ms	Run time: 1.016 sec
Data scanned: 5.23 KB								
Results (28)							<a href="#">Copy</a> <a href="#">Download results</a>	
<input type="text" value="Search rows"/>							< 1 > ⚙	
# ▾	office_branch ▾	min_earning s ▾	max_earning s ▾	avg_earnings ▾	total_earning s ▾	earnings_da		
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	New York	2173	9787	6513.178571428572	182369	2022-02-15		
6	New York	2173	9787	6513.178571428572	182369	2022-02-16		

# Query 3

Before:

Query results

Query stats

Completed

Time in queue: 125 ms

Run time: 1.341 sec

Data scanned: 4.42 KB

Results (4)

Copy

Download results

Search rows

< 1 >

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

After:

Query results

Query stats

Completed

Time in queue: 140 ms

Run time: 1.321 sec

Data scanned: 6.18 KB

Results (4)

Copy

Download results

Search rows

< 1 >

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

## Calculating percentage change in earnings

Now running a new query in Athena that calculates the % change in earnings for every employee from day 15 compared to day 12.

Query 2 : X

Query 3 : X

Query 4 : X

Query 1 : X

Query 5 : X

Query 6 : X

+

▼

```
1 WITH day_15 AS (  
2   SELECT emp_id, earnings, earnings_date  
3   FROM "saad_glue_db"."saad_employee_earnings"  
4   WHERE earnings_date = '2022-02-15'  
5 ),  
6 day_12 AS (  
7   SELECT emp_id, earnings, earnings_date  
8   FROM "saad_glue_db"."saad_employee_earnings"  
9   WHERE earnings_date = '2022-02-12'  
10 )  
11 SELECT  
12   day_15.emp_id,  
13   day_15.earnings AS current_earnings,  
14   day_12.earnings AS previous_earnings,  
15   (day_15.earnings - day_12.earnings) / CAST(day_12.earnings AS double) * 100 AS percentage_change  
16 FROM  
17   day_15  
18 JOIN  
19   day_12  
20 ON  
21   day_15.emp_id = day_12.emp_id
```

SQL Ln 21, Col 25

Run again

Explain

Cancel

Clear

Create ▼

Reuse query results

\*Athena engine version 3 only

Here are the results:

Results (100)					Copy	Download results
Search rows					< 1 ... > ⚙	
# ▼	emp_id ▲	current_earnings ▼	previous_earnings ▼	percentage_change ▼		
73	138911	2258	3984	-43.32329317269076		
45	143711	4227	2514	68.13842482100239		
39	147133	8646	2023	327.38507167572914		
96	149972	9108	2686	239.0915860014892		
23	155097	2067	7924	-73.91468955073195		
25	160938	3875	5699	-32.00561502017898		
22	163409	5271	4532	16.306266548984997		
29	170637	3641	8815	-58.695405558706746		