Members:

1. Syed Mohammad Anjil Hussain Rizvi (2303.KHI.DEG.031)
2. Huzefa Anver (2303.KHI.DEG.002)

Simulate earning predictions for two more days:

Manipulations to the data are made in Untitled.ipynb file. It can be found inside the same folder as this pdf file.

Data has been manipulated using pandas. The manipulation we are talking about Is simply multiplying the earnings column with 4.

Parquet file loaded:

taking a parquet file and manipulating its earings columnn to produce a new parquet file

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

[2]: df_1 = pd.read_parquet('/home/syedmohammadanjilhussainrizvi/Desktop/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnir

[3]: df_1
```

[3]:

	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	6227
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jsshaffer	7U56!*IO	Stanford	4437
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	6228
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	3127
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	3930
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31	1992-10-02	271-17-5467	228-485-0919	cmgould	m1%+0ojh7VlvJ	Stanford	4052
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	074-02-9202	316-256-7851	ckroden	5[Rn]G:#58f$>+S	Nashua	2886
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25	2012-10-06	552-99-5545	270-750-7760	mrnickel	8*E[g-_X|	Scranton	8586
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29	2016-10-17	763-12-2082	236-584-1916	eytribble	z>ms?;$8-u	Nashua	6493
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06	2004-03-26	637-29-1007	503-456-5899	allayman	k<%?%TML.].1ZY	New York	5261

100 rows × 13 columns

Accessing the earnings column:

```
[4]: earnings= df_1.loc[:,'earnings']
     earnings

[4]: 0     6227
     1     4437
     2     6228
     3     3127
     4     3930
           ...
     95    4052
     96    2886
     97    8586
     98    6493
     99    5261
     Name: earnings, Length: 100, dtype: int64
```

Manipulating the earnings column

```
[12]: new_earnings = earnings * 4

[13]: df_1.loc[:,'earnings'] = new_earnings
      df_1
```

[13]:

	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	99632
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jsshaffer	7U56I*IO	Stanford	70992
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F[jã]&m&&X	Stanford	99648
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	50032
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	62880
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31	1992-10-02	271-17-5467	228-485-0919	cmgould	m1%+0ojh7VlvJ	Stanford	64832
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	074-02-9202	316-256-7851	ckroden	5[Rn]G:#58f$>+S	Nashua	46176
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25	2012-10-06	552-99-5545	270-750-7760	mrnickel	8*E[g-_X]	Scranton	137376
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29	2016-10-17	763-12-2082	236-584-1916	eytribble	z>ms?;$8-u	Nashua	103888
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06	2004-03-26	637-29-1007	503-456-5899	allayman	k<%?%TML.].1ZY	New York	84176

100 rows × 13 columns

Creating a new parquet file:

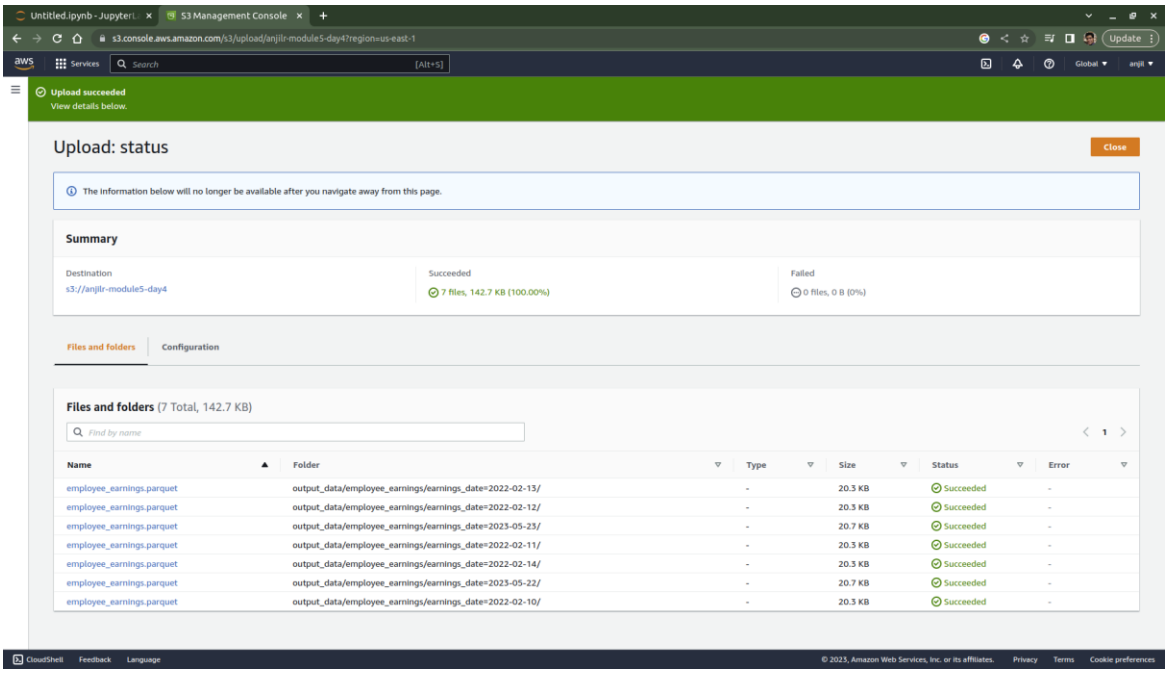
```
[16]: path = "/home/syedmohammadanjilhussainrizvi/Desktop/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings/earnings_dat

[17]: df_1.to_parquet(path)
```

All of these steps were repeated on another existing parquet file to create one more parquet file for another day.

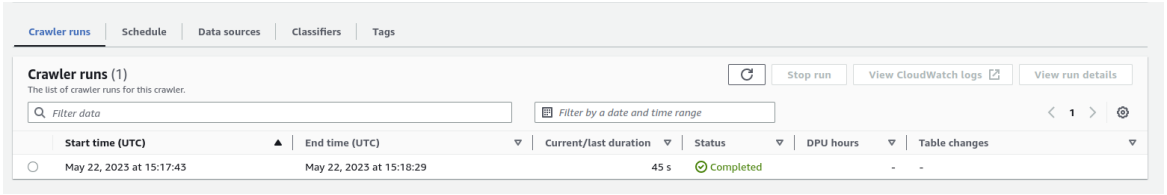
Data lake preparation:

1. Contents loaded into an s3 bucket:

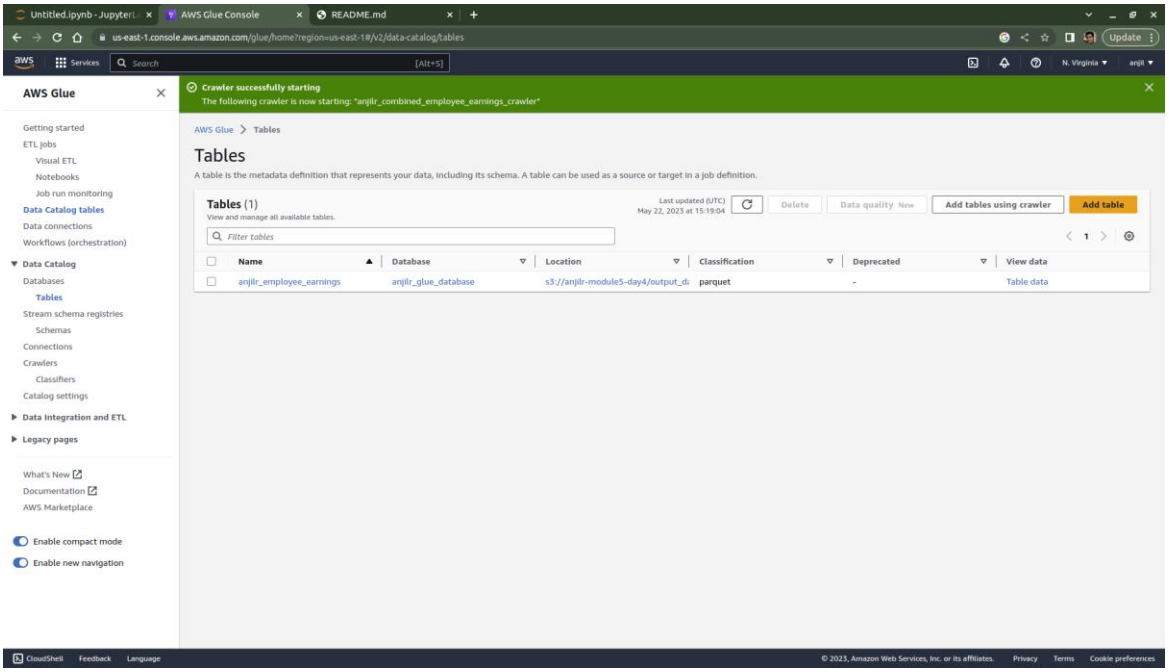


(folders with year 2023 in the title are the new folders containing new parquet files)

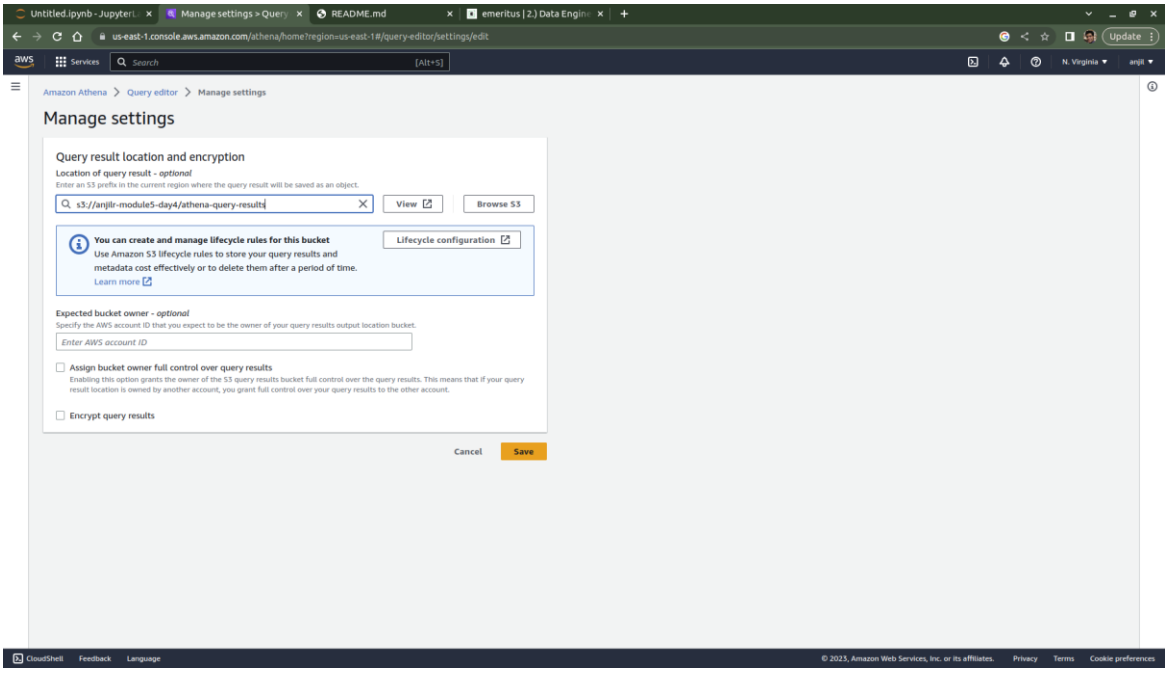
2. Crawler created and spun up:



3. We can see that it has produced our table that we are going to query from:

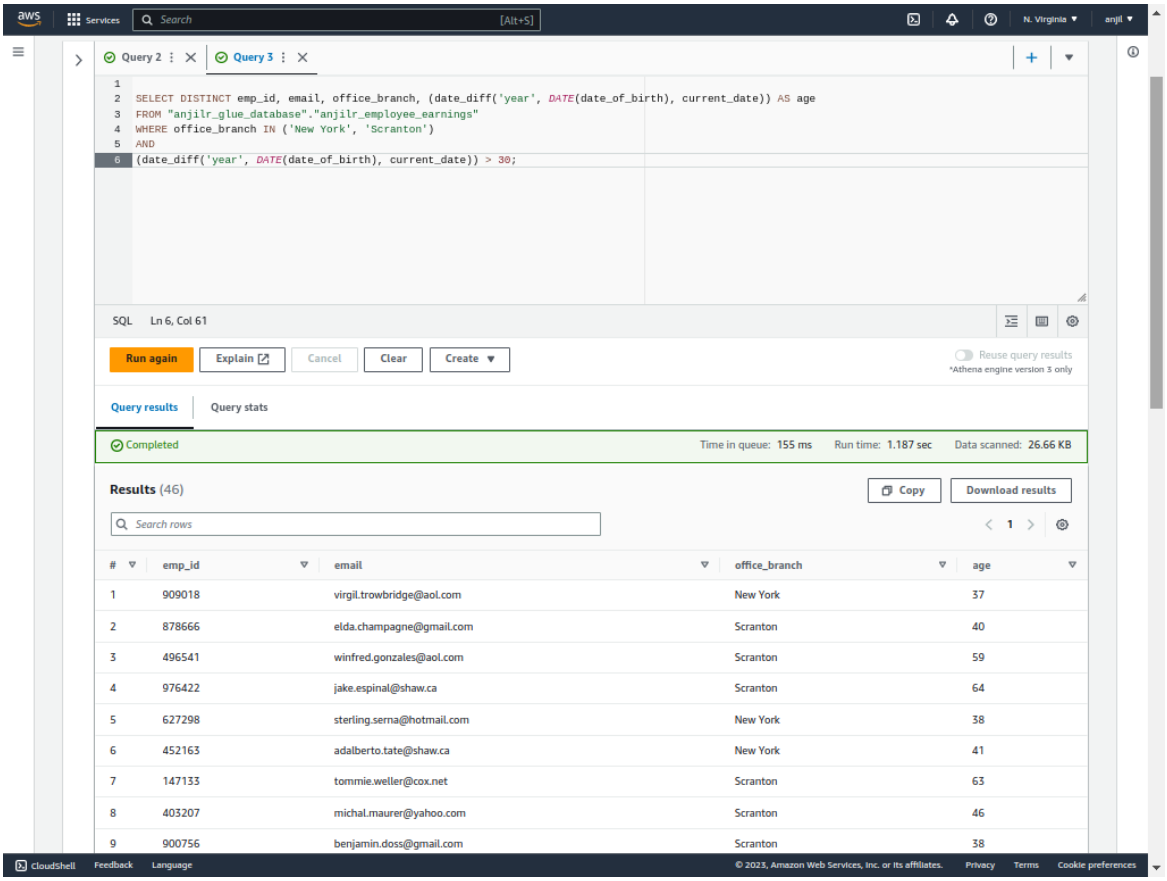


4. Query result location set inside s3 bucket:



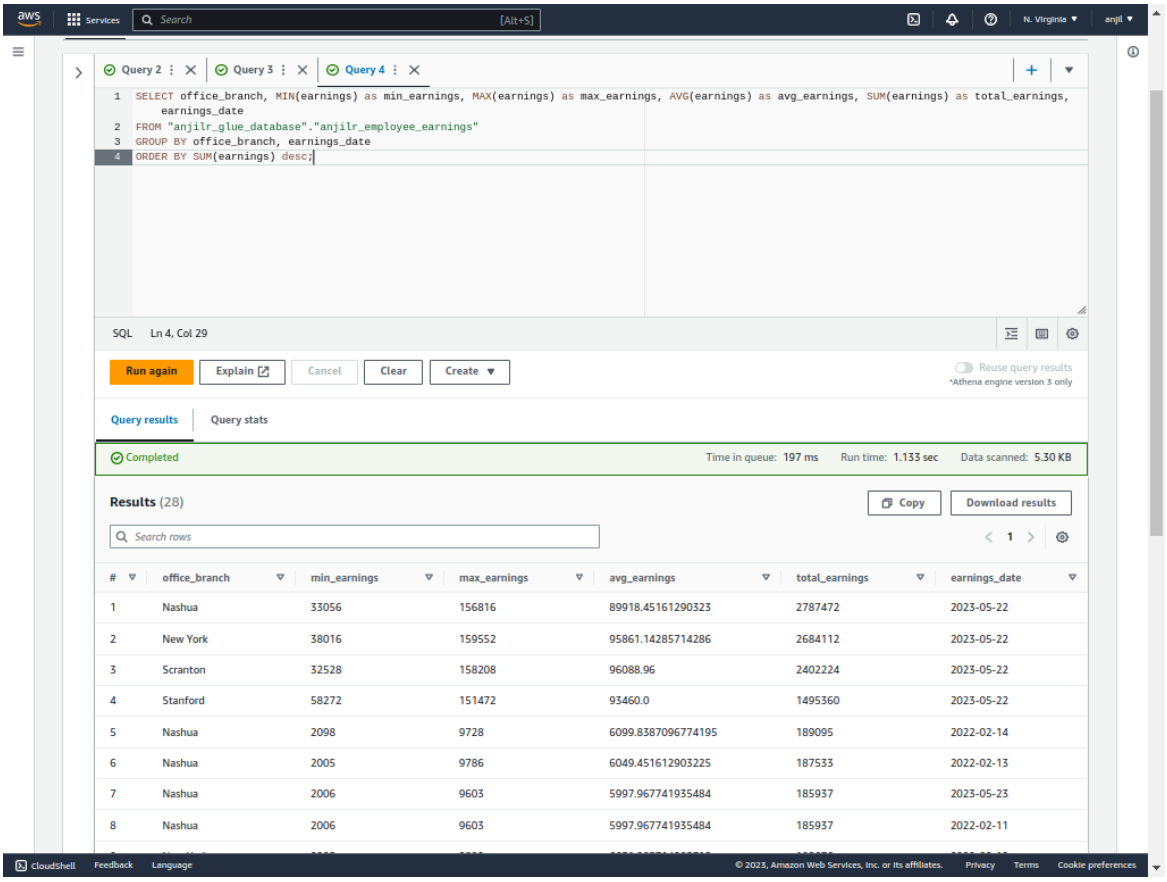
Querying using Athena:

All employees from offices 'New York' and 'Scranton' with age > 30:



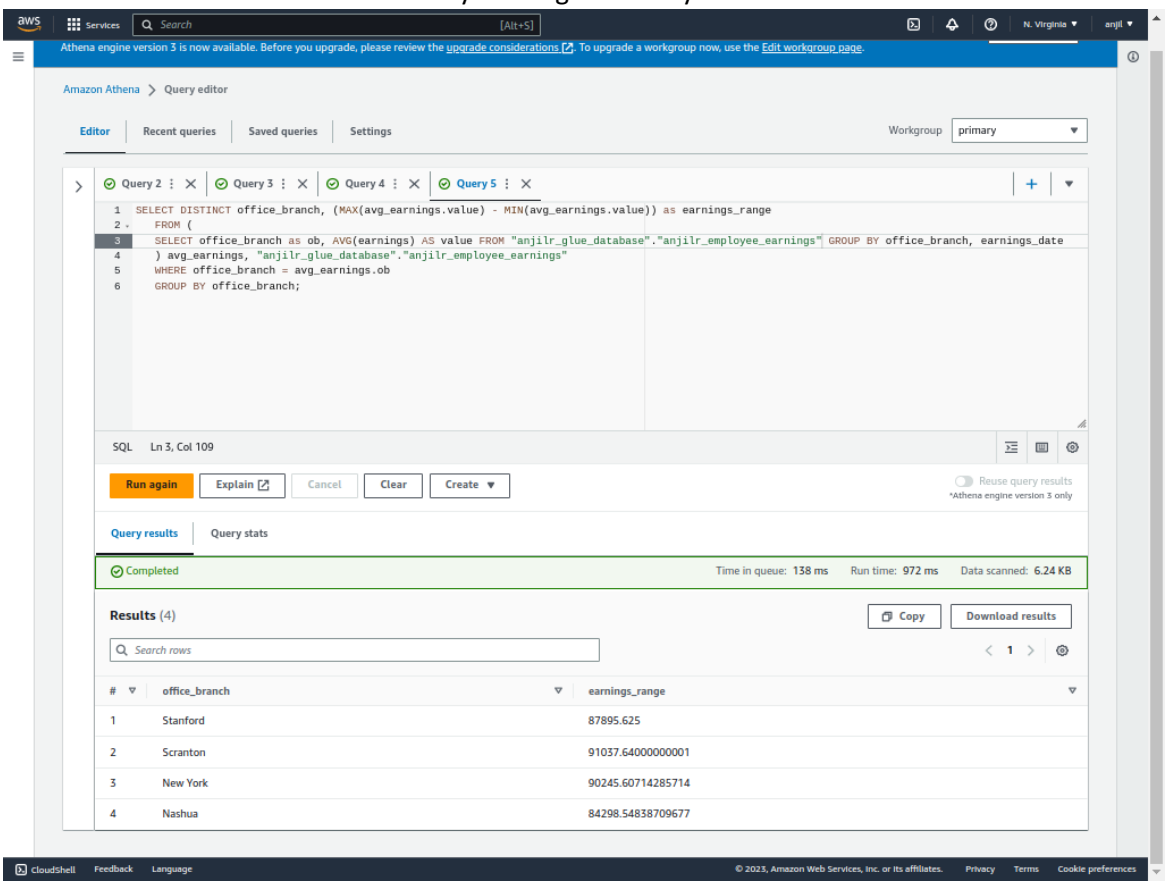
(no changes anticipated between this run and queries run before manipulation of data)

Min, max, average and total earnings for each office and each day - sorted by total earnings, highest to lowest:



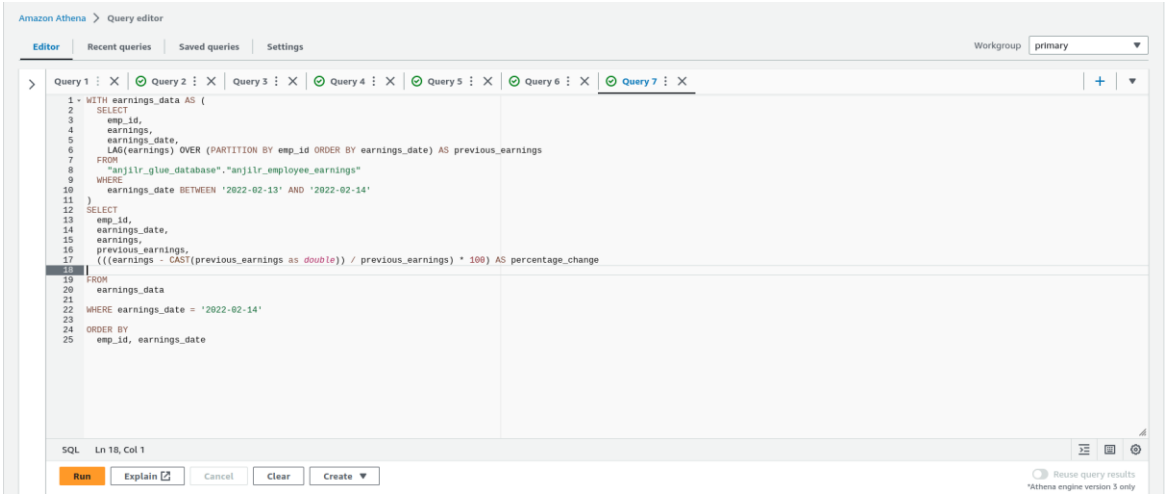
(in this one however we can expect different query results between this run and the same query run before data manipulation; this query is related to earnings which is why we can expect a difference. In fact, we can already see the data points from new parquet files being reflected in the first few rows of the query)

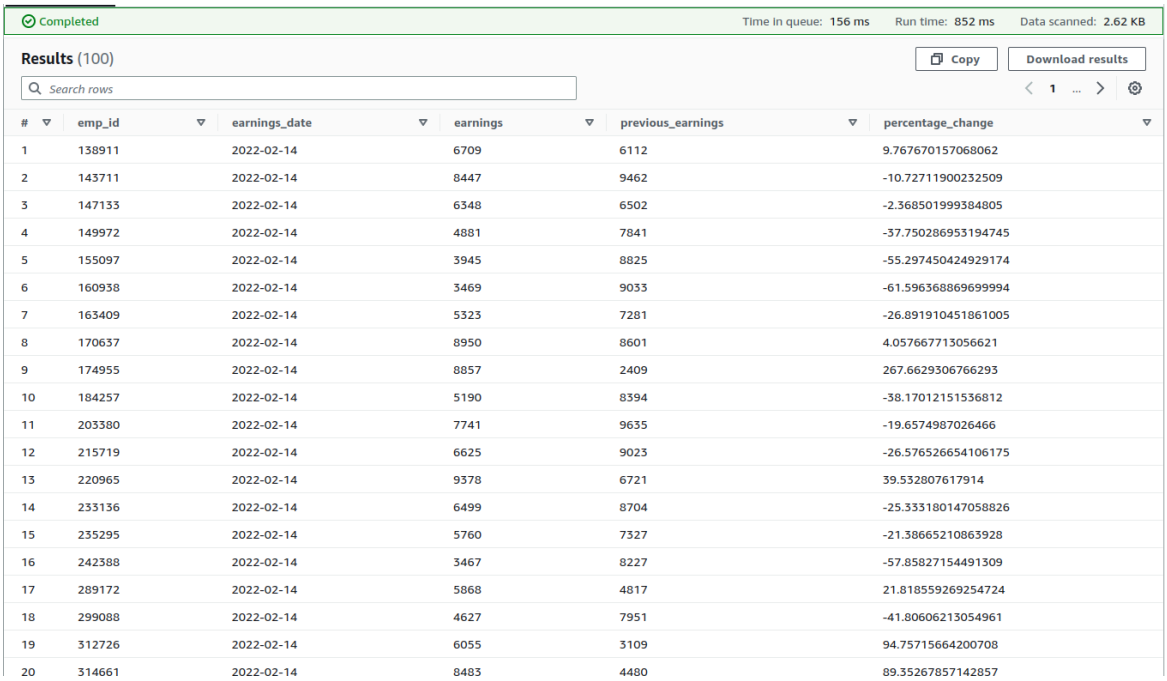
Difference between worst and best day earnings for every office branch:



Assignment Query:

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





Completed — Time in queue: 156 ms — Run time: 852 ms — Data scanned: 2.62 KB

Results (100)

Copy | Download results

Search rows

1 ...

#	emp_id	earnings_date	earnings	previous_earnings	percentage_change
1	138911	2022-02-14	6709	6112	9.767670157068062
2	143711	2022-02-14	8447	9462	-10.72711900232509
3	147133	2022-02-14	6348	6502	-2.368501999384805
4	149972	2022-02-14	4881	7841	-37.750286953194745
5	155097	2022-02-14	3945	8825	-55.297450424929174
6	160938	2022-02-14	3469	9033	-61.596368869699994
7	163409	2022-02-14	5323	7281	-26.891910451861005
8	170637	2022-02-14	8950	8601	4.057667713056621
9	174955	2022-02-14	8857	2409	267.6629306766293
10	184257	2022-02-14	5190	8394	-38.17012151536812
11	203380	2022-02-14	7741	9635	-19.6574987026466
12	215719	2022-02-14	6625	9023	-26.576526654106175
13	220965	2022-02-14	9378	6721	39.532807617914
14	233136	2022-02-14	6499	8704	-25.333180147058826
15	235295	2022-02-14	5760	7327	-21.38665210863928
16	242388	2022-02-14	3467	8227	-57.85827154491309
17	289172	2022-02-14	5868	4817	21.818559269254724
18	299088	2022-02-14	4627	7951	-41.80606213054961
19	312726	2022-02-14	6055	3109	94.75715664200708
20	314661	2022-02-14	8483	4480	89.35267857142857