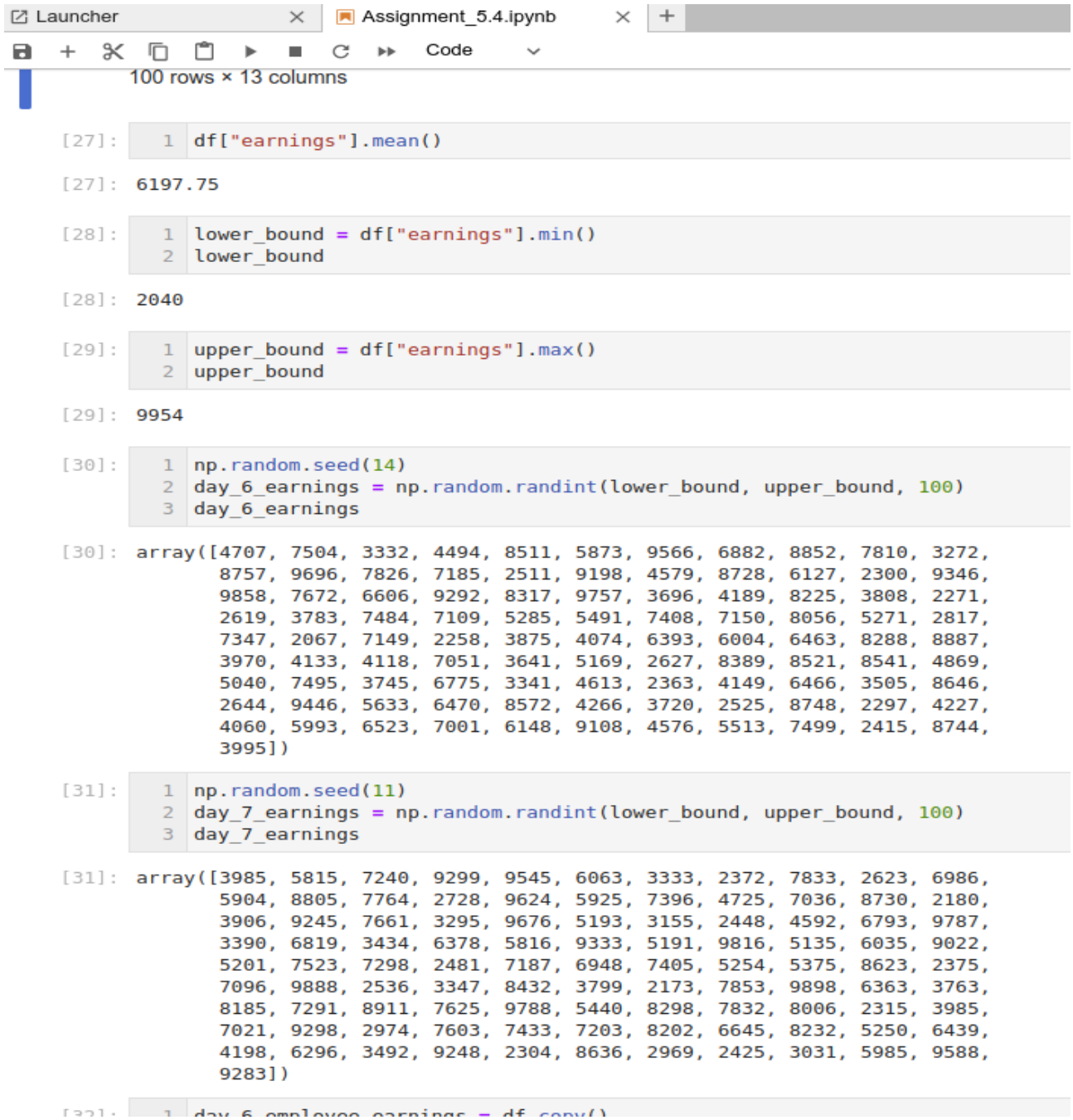
Graded Assignment 5.4

Name: Saad Sameer Khan
Employee#: 2303.KHI.DEG.034
Collaborated with: Mohammad Hamza Asim (2303.KHI.DEG.014)

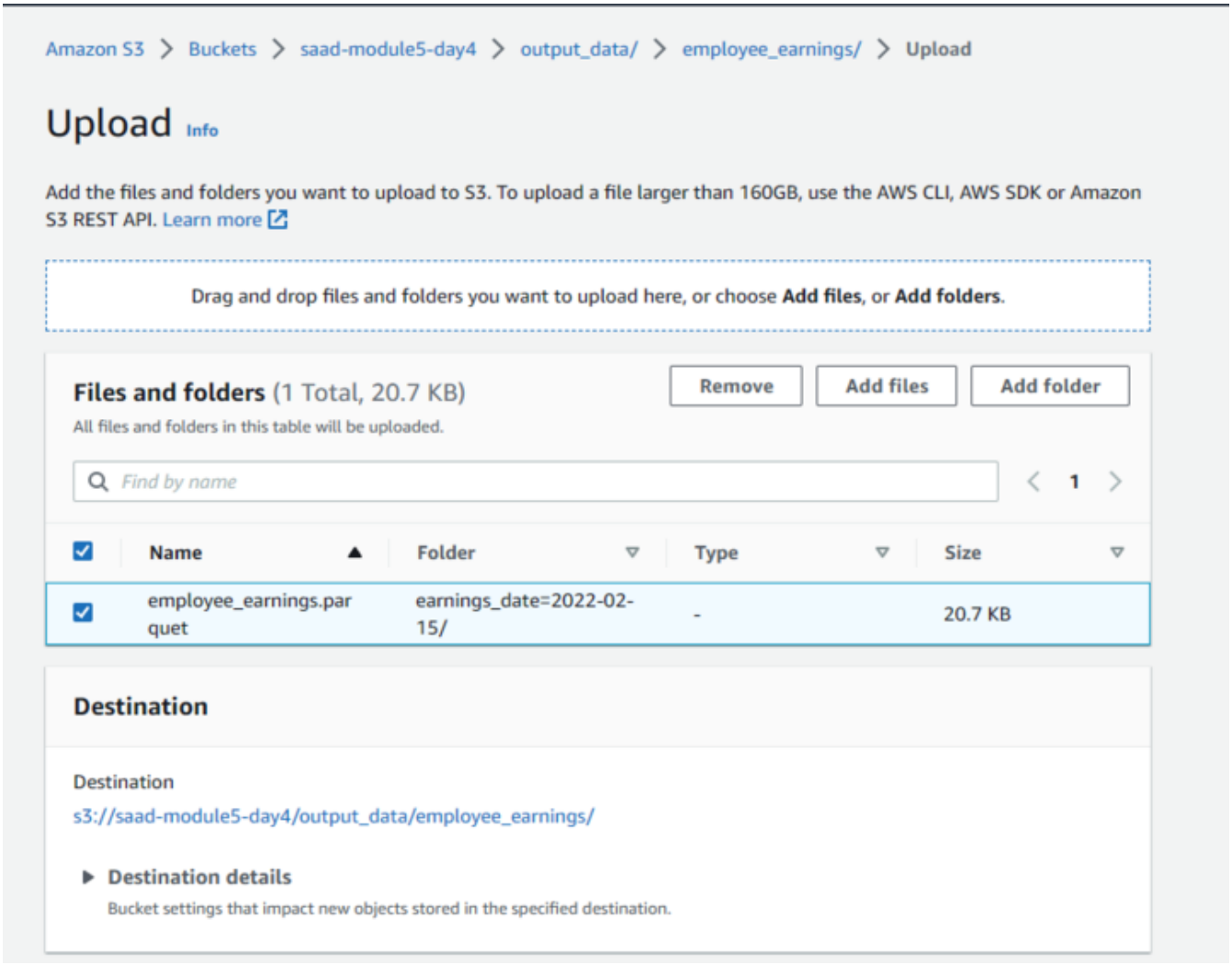
Creating data for 2 more days (day 6 & day 7)

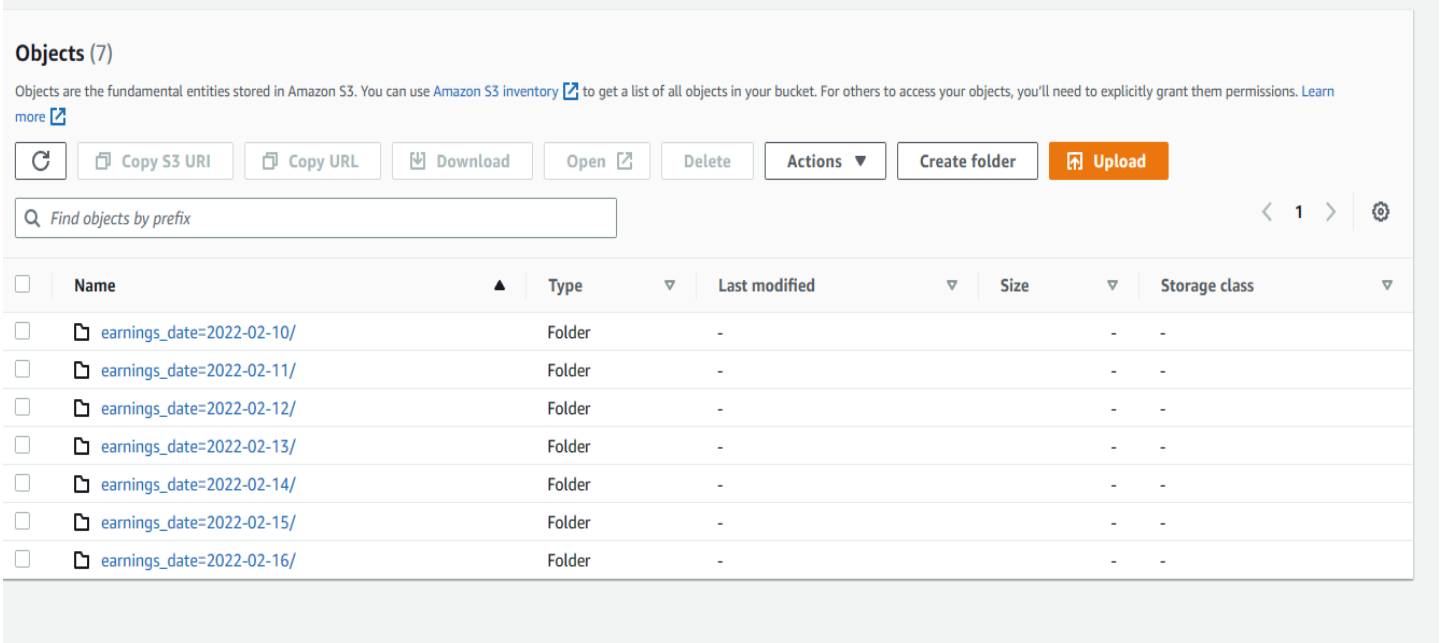
We used numpy's random.randint method to generate random earnings data within the range of the maximum and minimum value of the 'earnings' column from one of the day's data. *ipynb file has been uploaded to git*



Uploading new folders to S3 bucket

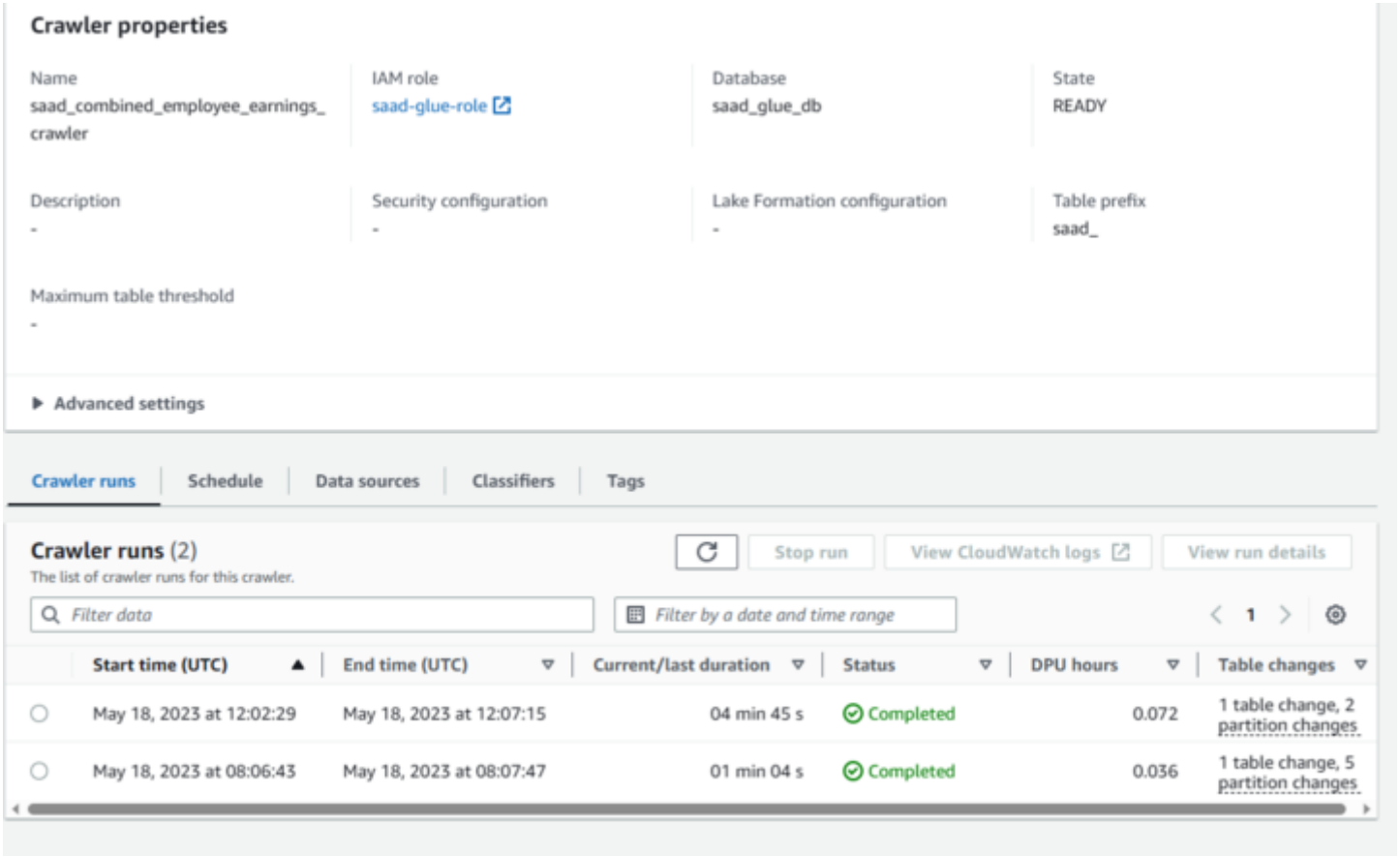
Next, we uploaded two new folders containing the data created in the previous step, in the output_data folder in the S3 bucket.





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.

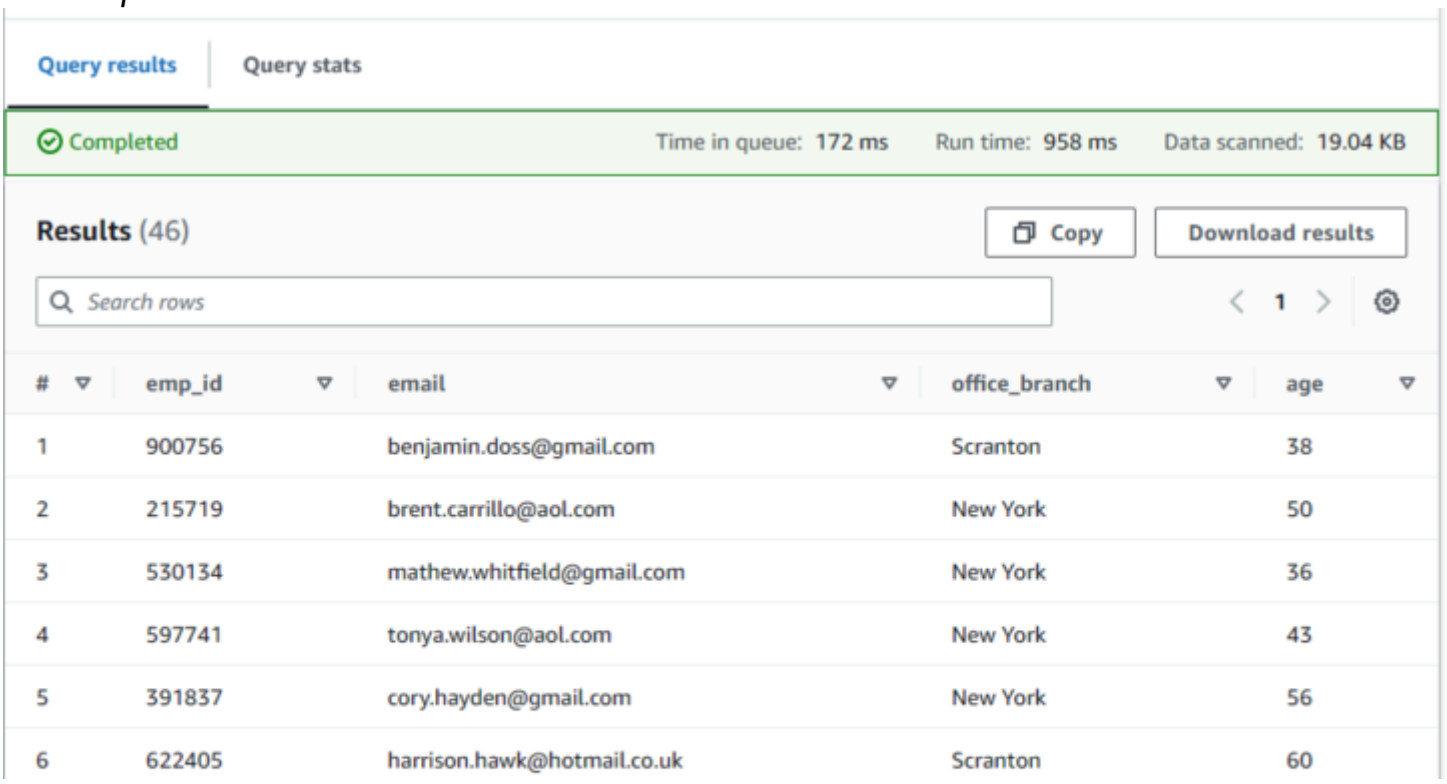


Re-running queries on updated data

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

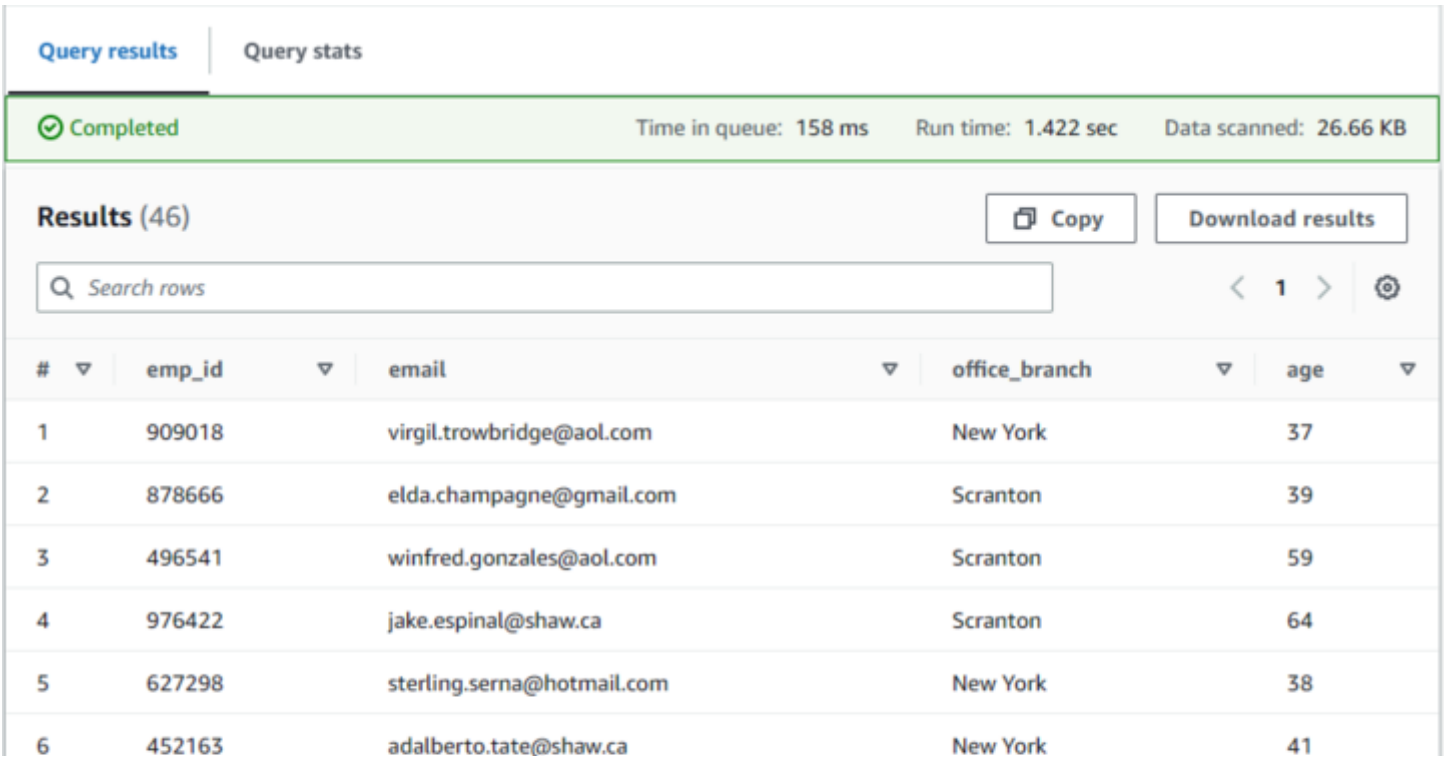
Query 1

Before:



#	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

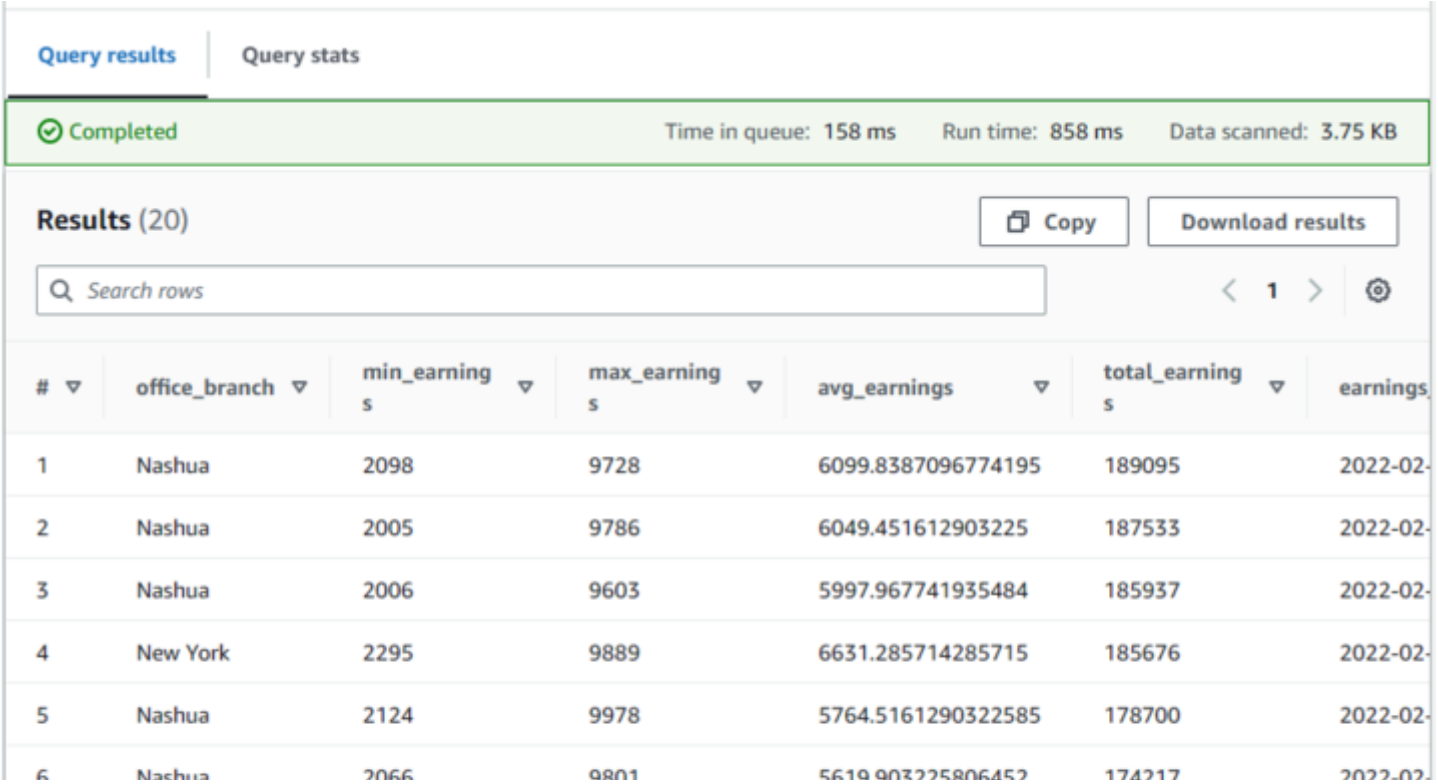
After



#	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.serna@hotmail.com	New York	38
6	452163	adalberto.tate@shaw.ca	New York	41

Query 2

Before:



Query results | Query stats

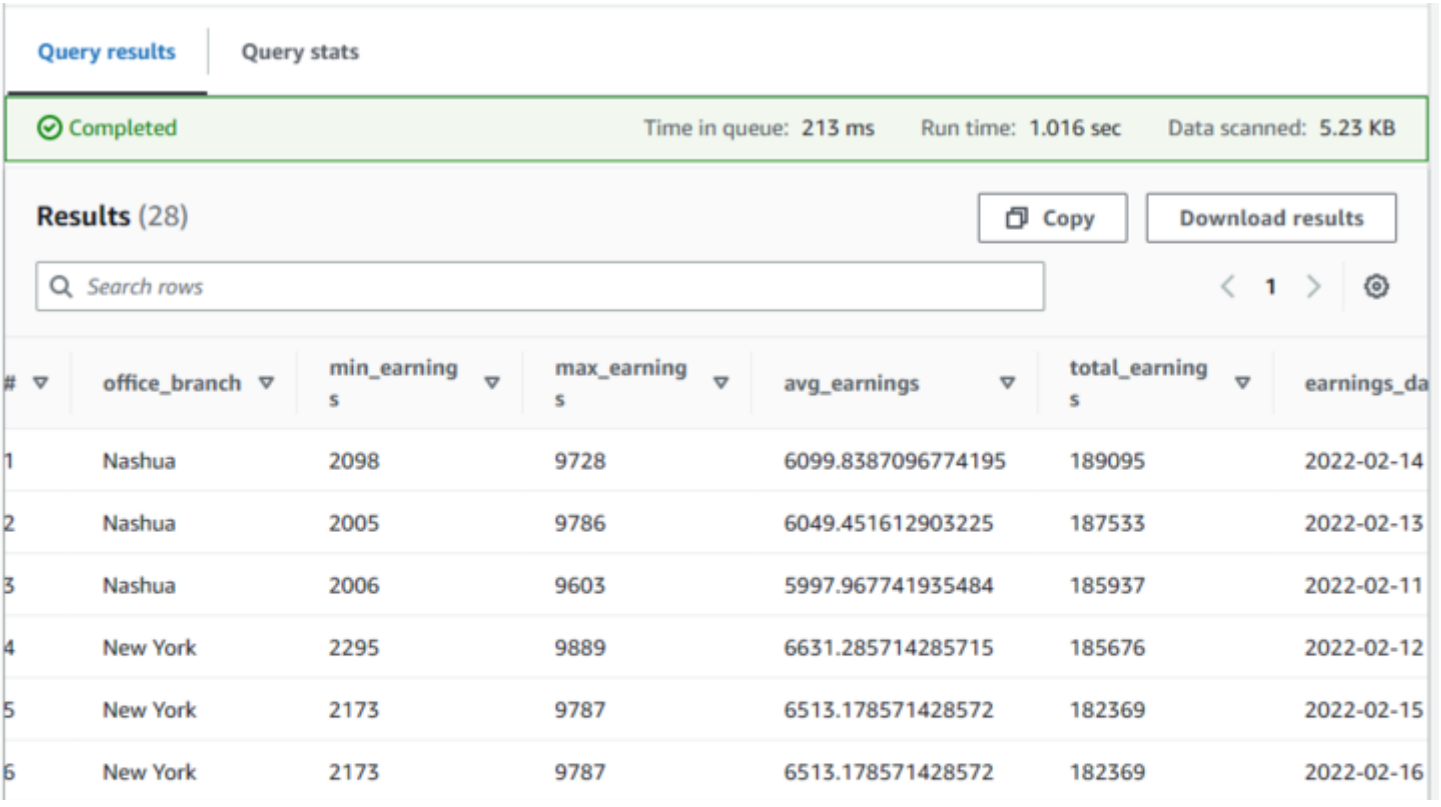
Completed — Time in queue: 158 ms — Run time: 858 ms — Data scanned: 3.75 KB

Results (20) — Copy — Download results

Search rows — 1

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	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

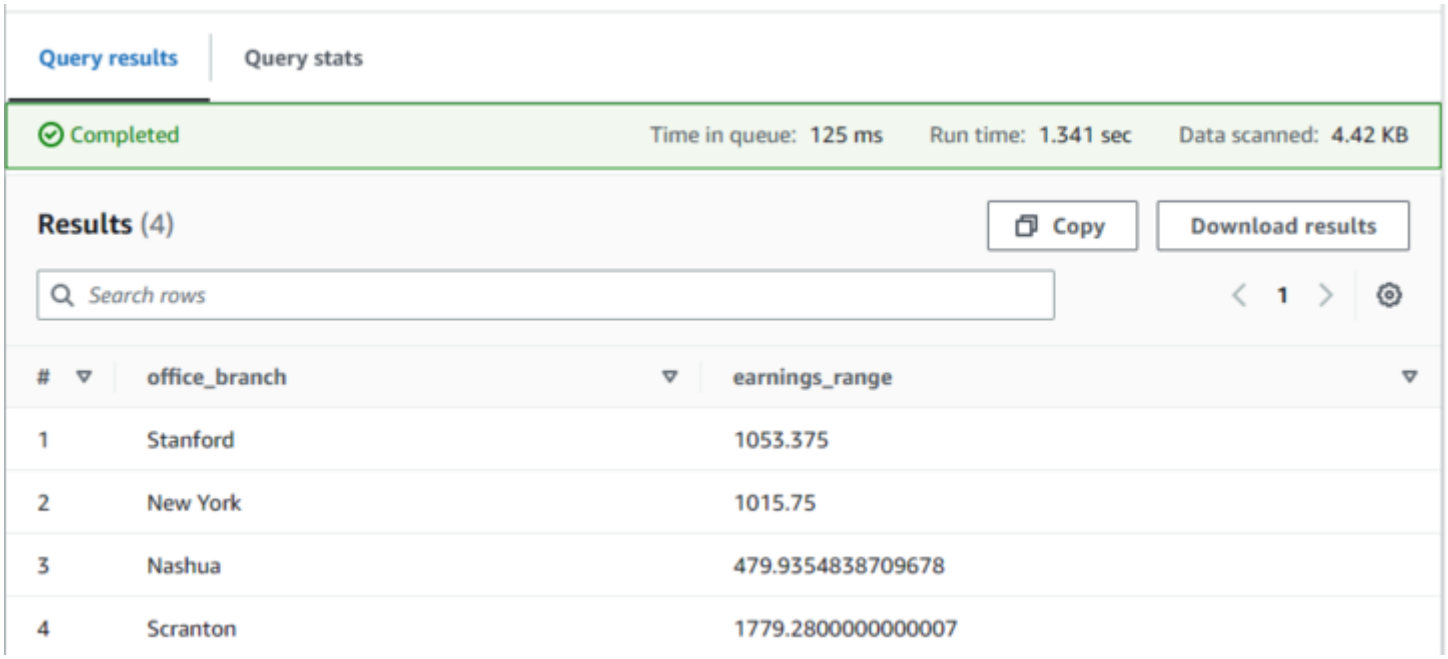
Results (28) — Copy — Download results

Search rows — 1

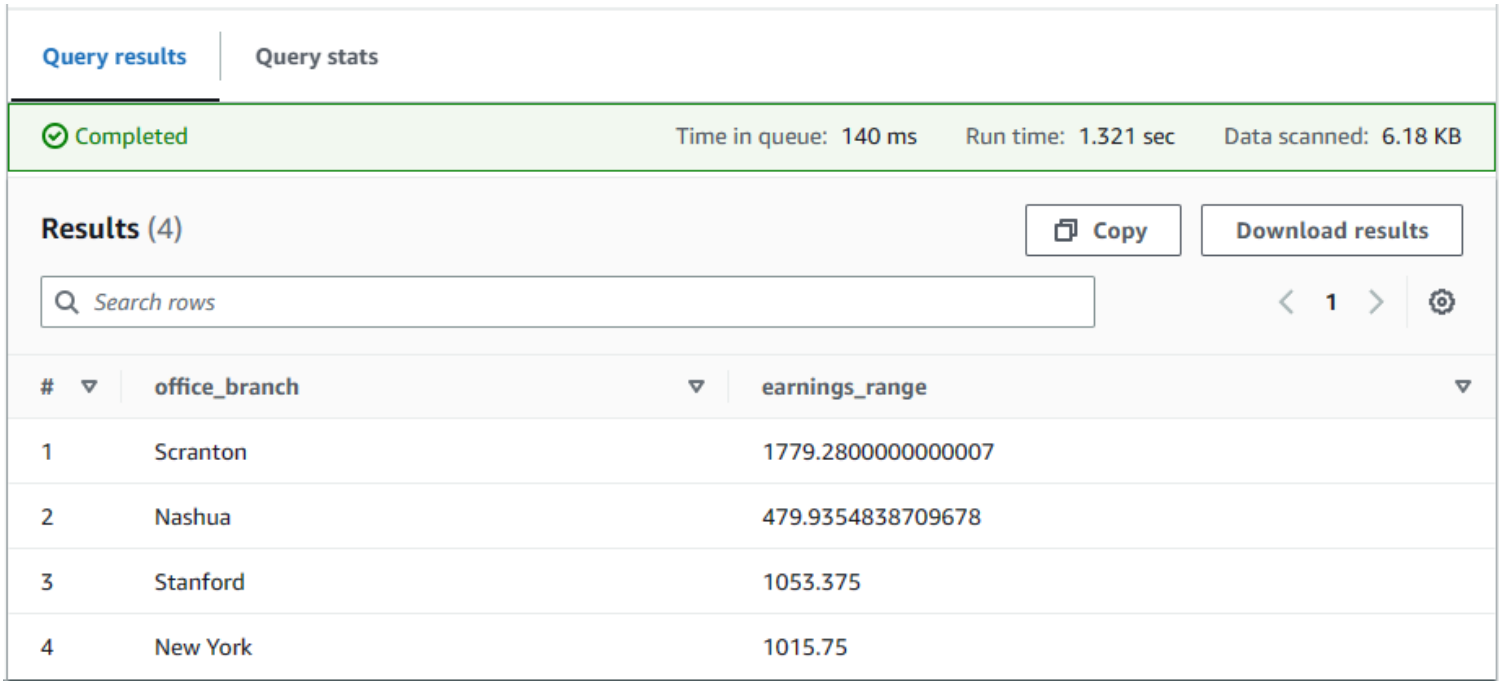
#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_da
1	Nashua	2098	9728	6099.8387096774195	189095	2022-02-14
2	Nashua	2005	9786	6049.451612903225	187533	2022-02-13
3	Nashua	2006	9603	5997.967741935484	185937	2022-02-11
4	New York	2295	9889	6631.285714285715	185676	2022-02-12
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:

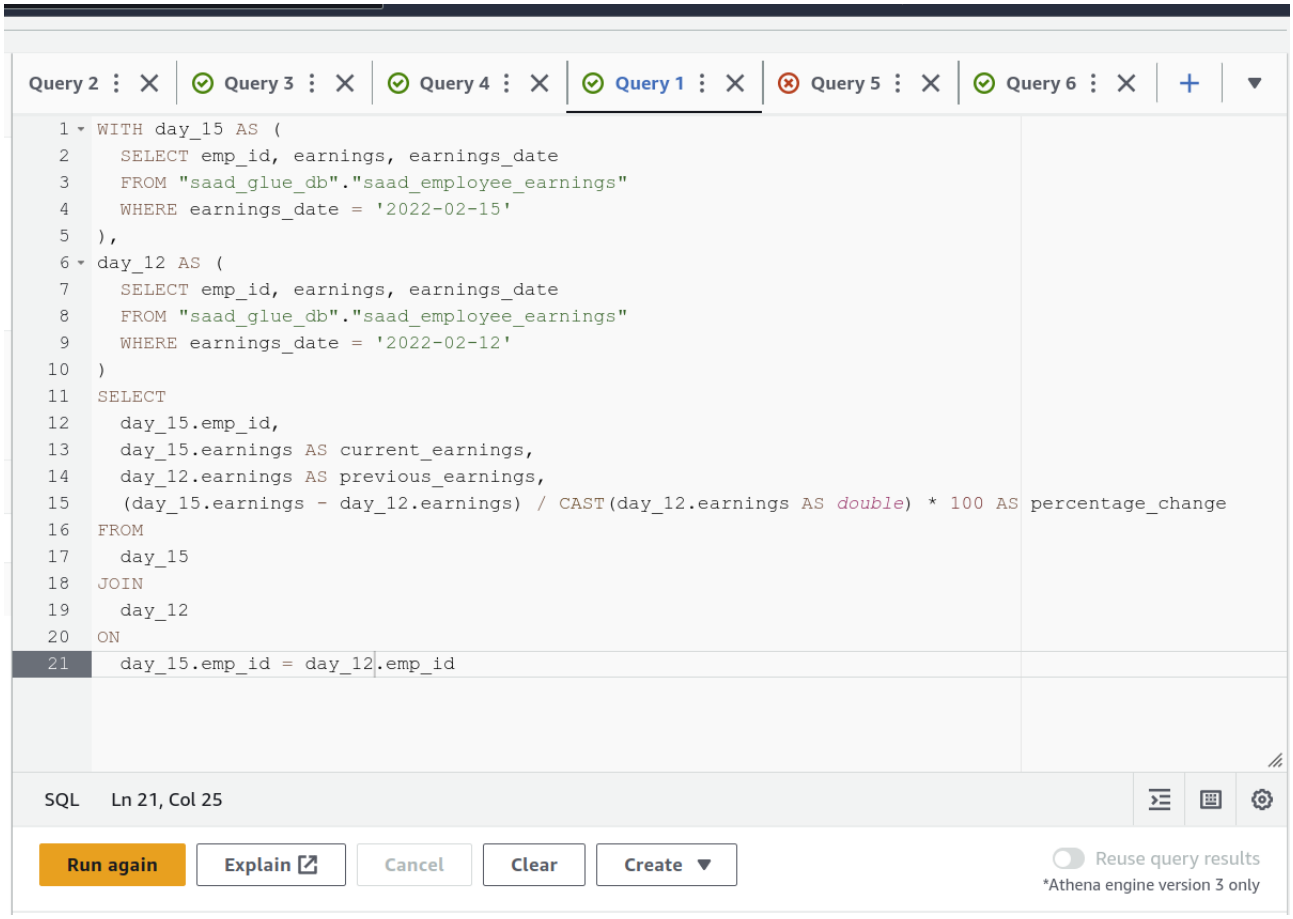


After:

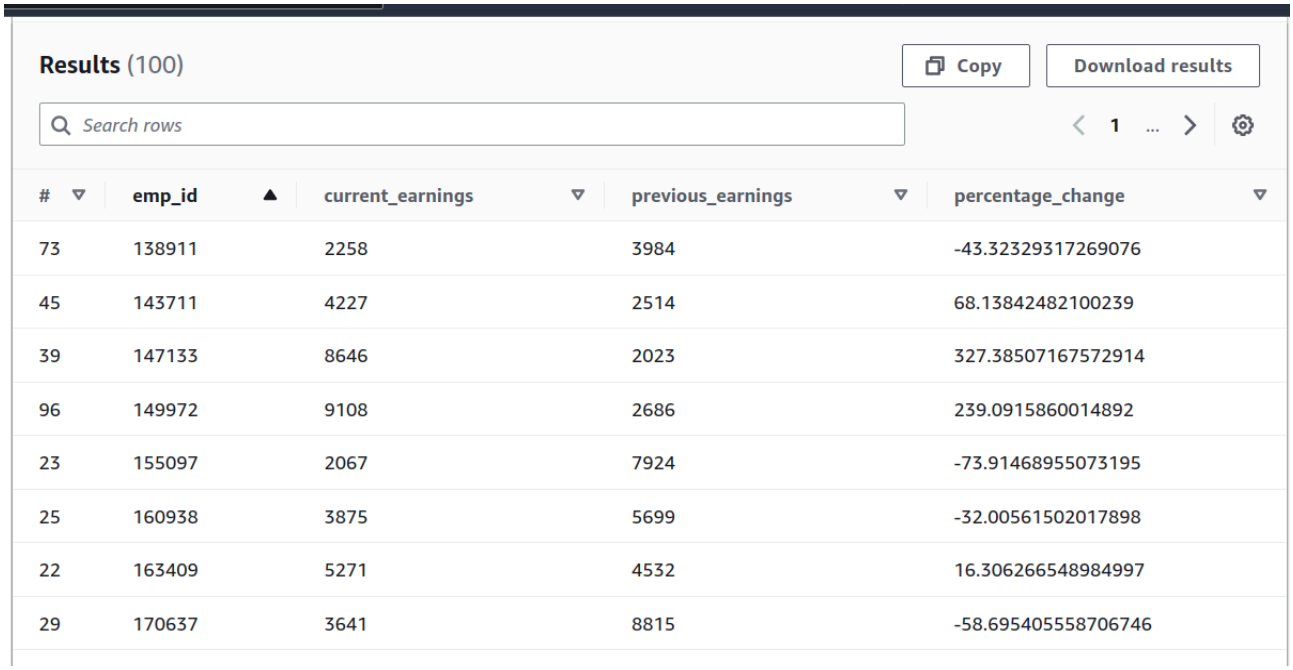


Calculating percentage change in earnings

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



Here are the results:



#	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