UNIT 5.4 GRADED ASSIGNMENT

Group members

Ifra Saleem (2303.khi.deg.003)
Umaima Siddiqui (2023.KHI.DEG.033)

UNIT 5.4 GRADED ASSIGNMENT

Task:

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).

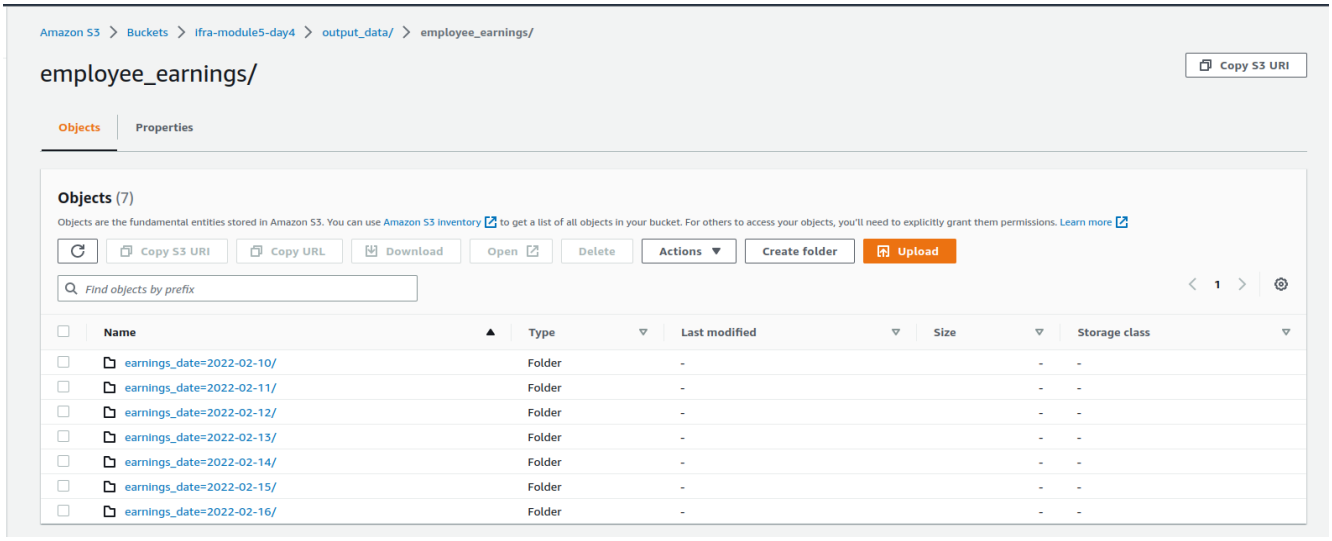
Rerun queries from Task 3 and Task 4 and see how the results change with this new data.

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

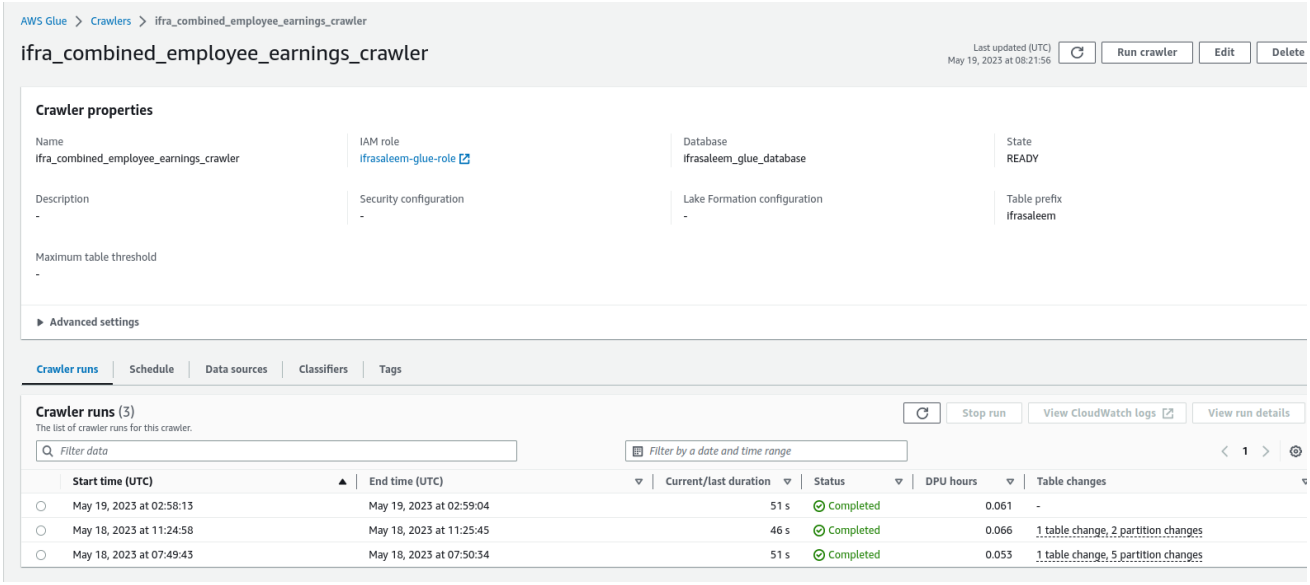
Solution:

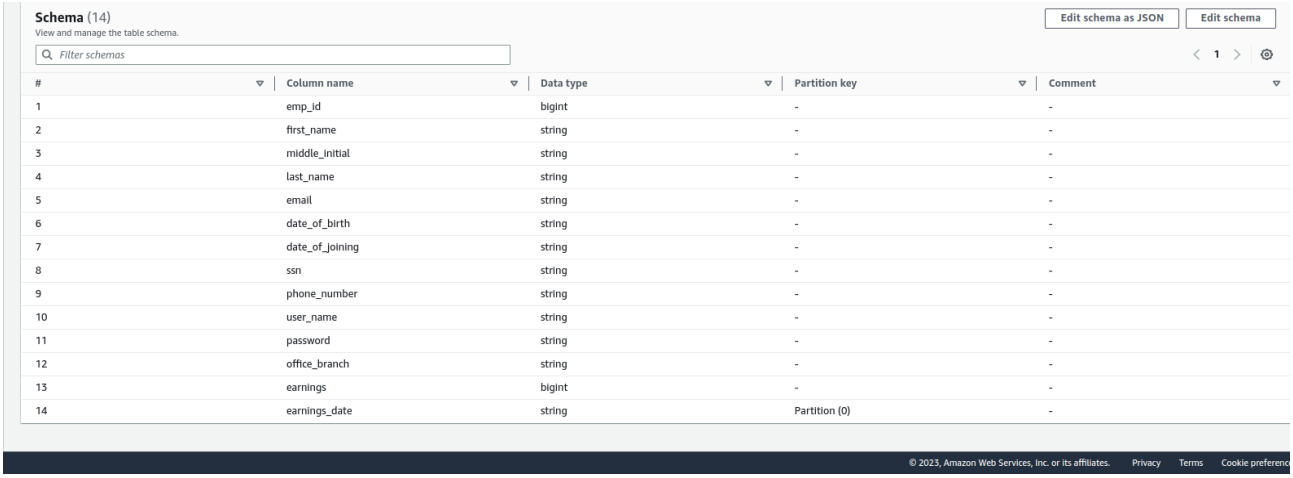
Uploaded two new folders to S3 Bucket.

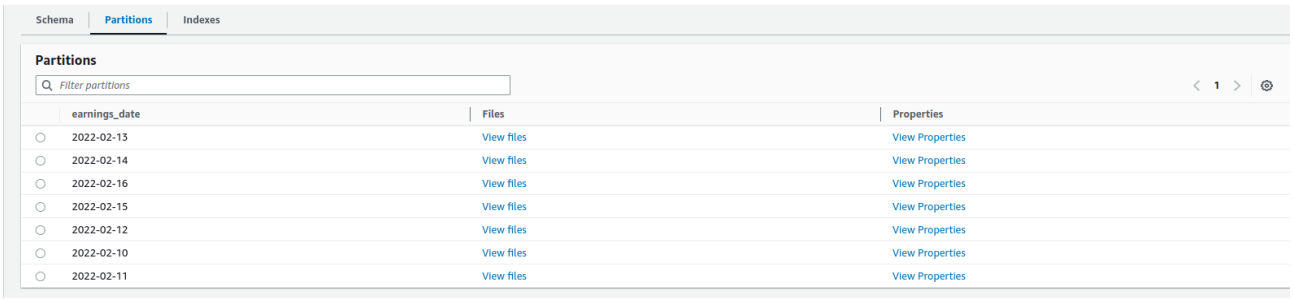
Script file is attached with the assignment which we used to create new data files for two more days.



Run the Crawler again:







Rerun queries from Task 3 and Task 4:

The screenshots are attached for the result of queries of previous data.

Task 3:

Query 1 | Query 1 | Query 2 | Query 3 | Query 14

```
1 SELECT *
2 FROM "ifrasaleem_glue_database"."ifrasaleememployee_earnings";
```

Results (700)

Copy | Download results

Search rows

1 ...

#	emp_id	first_name	middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number	user_name
1	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-03-24	471-57-0359	212-884-7146	akgoodwin
2	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jsshaffer
3	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris
4	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon
5	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza
6	220965	Almeta	Y	Brookins	almeta.brookins@gmail.com	1985-05-08	2017-04-25	109-98-3095	229-238-0915	aybrookins
7	721091	Bobbie	E	Branson	bobbie.branson@hotmail.com	1996-02-10	2017-04-27	047-15-8435	216-849-6986	bebranson
8	633636	Bertram	R	Carlisle	bertram.carlisle@aol.com	1982-12-31	2012-11-02	257-99-3865	215-614-1953	brcarlisle
9	823898	Carlton	E	Leclair	carlton.leclair@cox.net	1986-03-17	2009-01-07	044-15-4027	319-943-1308	celeclair
10	413865	Todd	R	Slater	todd.slater@earthlink.net	1995-06-09	2017-03-07	553-99-2106	218-671-2547	trslater
11	439483	Adalberto	K	Hazel	adalberto.hazel@yahoo.co.in	1959-12-08	1995-04-27	264-99-9423	262-893-5324	akhazel
12	809408	Seymour	B	Nelson	seymour.nelson@hotmail.com	1996-01-21	2017-07-02	119-98-1796	215-821-4109	sbnelson
13	748190	Vicente	D	Dawkins	vicente.dawkins@gmail.com	1994-06-17	2016-07-02	116-98-6696	217-567-3339	vddawkins
14	397283	Rex	Q	Ng	rex.ng@yahoo.com	1982-08-22	2007-01-17	192-86-7823	215-352-5080	rqng
15	553684	Arlena	I	Clough	arlena.clough@yahoo.com	1957-10-15	2015-12-26	032-92-6090	316-737-9183	aiclough
16	495667	Cory	H	Clarke	cory.clarke@shell.com	1980-06-02	2012-04-27	731-28-8642	480-381-5936	chclarke
17	936158	Sofia	Y	Poole	sofia.poole@earthlink.net	1964-12-03	1993-07-18	126-98-9929	215-824-1223	sypoole
18	537591	Samuel	G	Wendt	samuel.wendt@bellsouth.net	1976-11-25	2016-02-24	369-39-1426	229-497-4180	sgwendt
19	909018	Virgil	R	Trowbridge	virgil.trowbridge@aol.com	1986-04-06	2016-11-21	005-13-9026	202-652-7351	vrtrowbridge
20	878666	Elda	Y	Champagne	elda.champagne@gmail.com	1983-05-21	2016-11-08	540-83-9723	218-621-9347	eychampagne

Query 1 | Query 1 | Query 2 | Query 3 | Query 14 | Query 4

```
1 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
2 FROM "ifrasaleem_glue_database"."ifrasaleememployee_earnings"
3 WHERE office_branch IN ('New York', 'Scranton')
4 AND
5 (date_diff('year', DATE(date_of_birth), current_date)) > 30;
6
```

Results (46)

Copy | Download results

Search rows

#	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
7	595558	denisha.fitch@msn.com	Scranton	32
8	314661	charles.quintero@gmail.com	New York	65
9	654617	rogelio.woodall@gmail.com	New York	50
10	138911	claudio.heck@aol.com	Scranton	55
11	713294	sammy.dewitt@ibm.com	Scranton	35
12	312726	celine.lumpkin@gmail.com	New York	36
13	551149	michale.colson@comcast.net	Scranton	61
14	402180	allan.bernhardt@gmail.com	New York	61
15	220965	almeta.brookins@gmail.com	Scranton	38
16	537591	samuel.wendt@bellsouth.net	New York	46
17	767674	irena.dang@gmail.com	New York	51
18	505927	oswaldo.winchester@gmail.com	New York	64
19	432820	myron.marble@gmail.com	New York	42
20	317987	chastity.pineda@shaw.ca	New York	48

Query 1 | Query 1 | Query 2 | Query 3 | Query 14 | Query 4

```
SELECT office_branch, MIN(earnings) as min_earnings, MAX(earnings) as max_earnings, AVG(earnings) as avg_earnings, SUM(earnings) as total_earnings, earnings_date
FROM "ifrasaleem_glue_database"."ifrasaleememployee_earnings"
GROUP BY office_branch, earnings_date
ORDER BY SUM(earnings) desc;
```

Results (28)

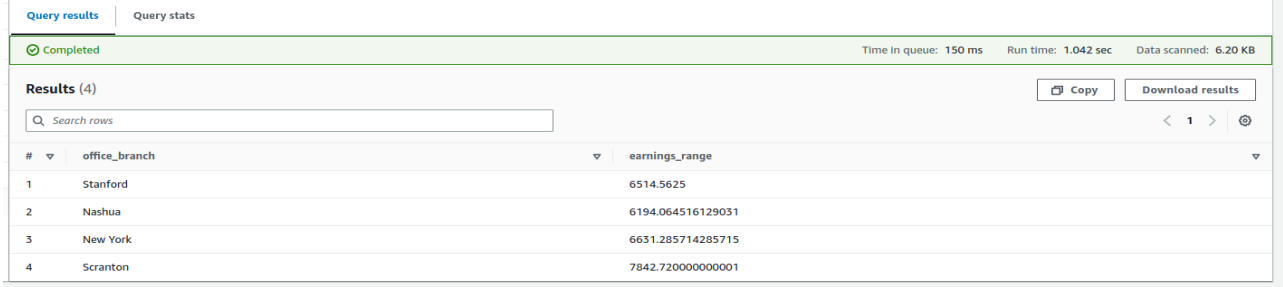
Copy | Download results

Search rows

#	office_branch	min_earnings	max_earnings	avg_earnings	total_earnings	earnings_date
1	Nashua	5234	19135	11813.967741935483	366233	2022-02-16
2	Nashua	4101	18006	11617.870967741936	360154	2022-02-15
3	New York	5989	18901	12246.82142857143	342911	2022-02-16
4	New York	4605	19150	11989.5	335706	2022-02-15
5	Scranton	6444	17504	12894.04	322351	2022-02-16
6	Scranton	4521	18219	11056.88	276422	2022-02-15
7	Stanford	7846	18576	12078.9375	193263	2022-02-15
8	Nashua	2098	9728	6099.838709677419 5	189095	2022-02-14
9	Nashua	2005	9786	6049.451612903225	187533	2022-02-13
10	Stanford	6063	17919	11706.5625	187305	2022-02-16
11	Nashua	2006	9603	5997.967741935484	185937	2022-02-11
12	New York	2295	9889	6631.285714285715	185676	2022-02-12
13	Nashua	2124	9978	5764.5161290322585	178700	2022-02-12
14	Nashua	2066	9801	5619.903225806452	174217	2022-02-10
15	New York	2040	9954	6109.035714285715	171053	2022-02-14
16	Scranton	2788	9916	6830.6	170765	2022-02-13
17	New York	2141	9462	5998.178571428572	167949	2022-02-11
18	New York	2376	9972	5991.321428571428	167757	2022-02-10
19	New York	2195	9734	5615.535714285715	157235	2022-02-13
20	Scranton	2465	9827	6149.72	153743	2022-02-14

Query 1 | Query 1 | Query 2 | Query 3 | Query 14 | Query 4

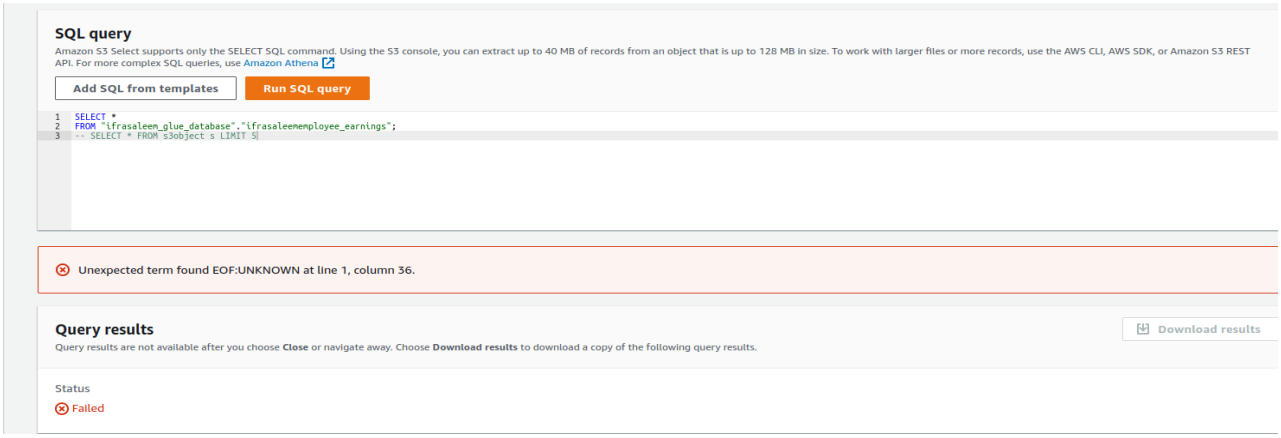
```
SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
FROM (
SELECT office_branch as ob, AVG(earnings) AS value FROM "ifrasaleem_glue_database"."ifrasaleememployee_earnings" GROUP BY office_branch, earnings_date
) avg_earnings, "ifrasaleem_glue_database"."ifrasaleememployee_earnings"
WHERE office_branch = avg_earnings.ob
GROUP BY office_branch;
```

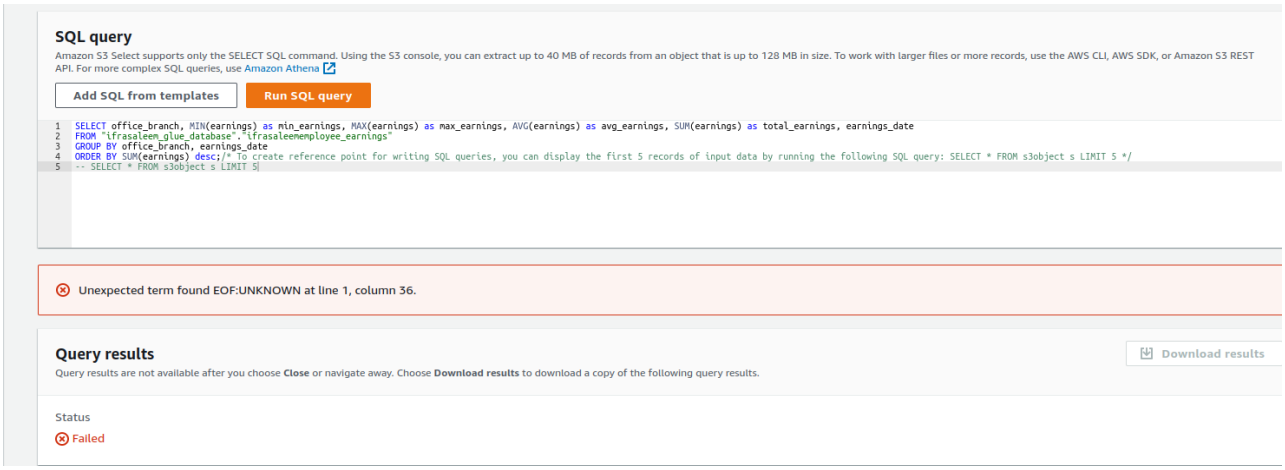


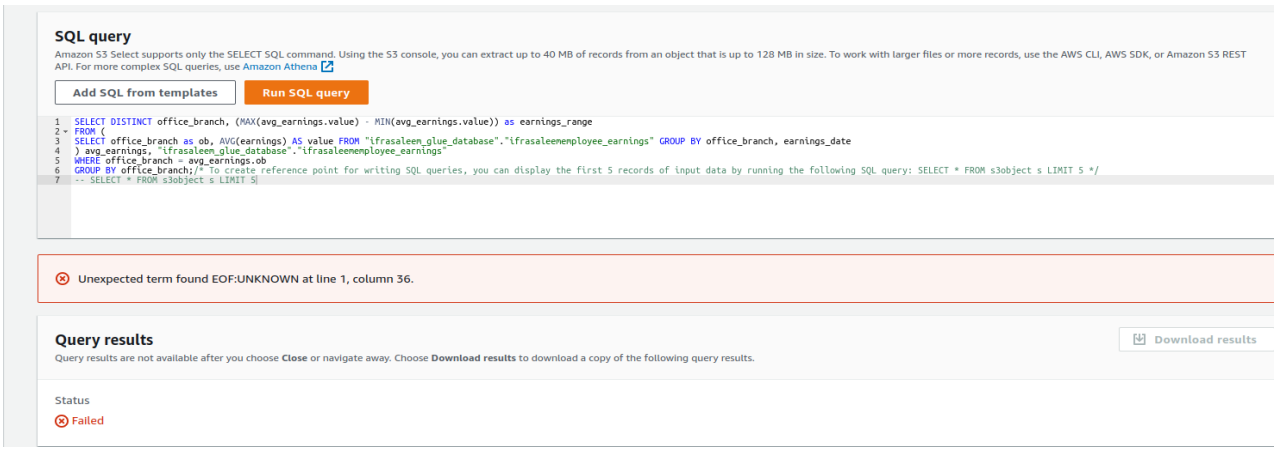
Task 4:

S3 select has some limitations on the type of queries it supports. and complex SQL queries with window functions like LAG are not supported.

S3 Select is primarily designed for simple SQL queries that involve basic filtering and projection operations on CSV, JSON, or Parquet data stored in S3. It does not support advanced SQL features like window functions or subqueries. That's why it is giving errors in running the queries.







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

Formula:

Percentage Change = ((New Value - Old Value) / |Old Value|) * 100

Query 1 | Query 1 | Query 2 | Query 3 | Query 14 | Query 15

```sql
WITH earnings_data AS (
  SELECT emp_id, earnings, earnings_date,
    LAG(earnings) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS previous_earnings,
    LAG(earnings_date) OVER (PARTITION BY emp_id ORDER BY earnings_date) AS previous_earnings_date
  FROM "ifrasaleem_glue_database"."ifrasaleememployee_earnings"
)
SELECT emp_id, earnings_date, earnings, previous_earnings, previous_earnings_date,
  ((earnings - previous_earnings) / ABS(CAST(previous_earnings AS double))) * 100 AS percentage_change
FROM earnings_data
WHERE earnings_date = '2022-02-14';
```

Results (100)

Copy | Download results

Search rows

#	emp_id	earnings_date	earnings	previous_earnings	previous_earnings_date	percentage_change
1	160938	2022-02-14	3469	9033	2022-02-13	-61.596368869699994
2	163409	2022-02-14	5323	7281	2022-02-13	-26.891910451861005
3	170637	2022-02-14	8950	8601	2022-02-13	4.057667713056621
4	233136	2022-02-14	6499	8704	2022-02-13	-25.333180147058826
5	572204	2022-02-14	9168	3962	2022-02-13	131.3982836951035
6	721091	2022-02-14	3557	5042	2022-02-13	-29.45259817532725
7	748190	2022-02-14	6157	3582	2022-02-13	71.88721384701284
8	812053	2022-02-14	7063	5978	2022-02-13	18.149882903981265
9	820109	2022-02-14	8115	7354	2022-02-13	10.348109872178405
10	840113	2022-02-14	8679	2115	2022-02-13	310.35460992907804
11	887387	2022-02-14	8123	4816	2022-02-13	68.66694352159467
12	143711	2022-02-14	8447	9462	2022-02-13	-10.72711900232509
13	147133	2022-02-14	6348	6502	2022-02-13	-2.368501999384805
14	155097	2022-02-14	3945	8825	2022-02-13	-55.297450424929174
15	289172	2022-02-14	5868	4817	2022-02-13	21.818559269254724
16	489275	2022-02-14	9728	4248	2022-02-13	129.00188323917138
17	492527	2022-02-14	6948	6508	2022-02-13	6.76090964961954
18	526254	2022-02-14	6602	7344	2022-02-13	-10.103485838779957
19	633636	2022-02-14	8353	9327	2022-02-13	-10.442800471748686
20	886060	2022-02-14	5307	9157	2022-02-13	-42.04433766517418