Name: - Muhammad Huzaifa Waseem (2303-KHI-DEG-021)

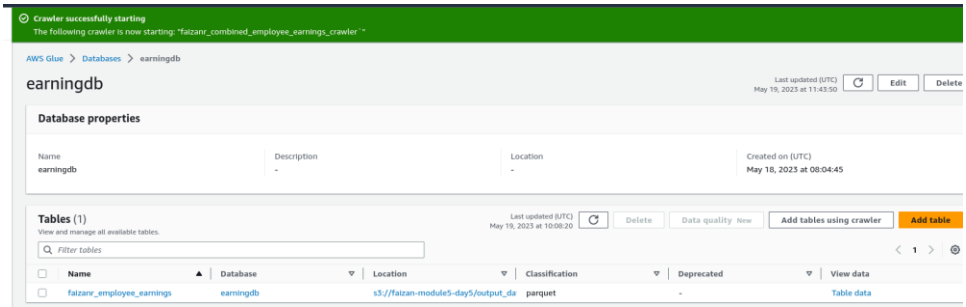
Pair Partner 1: - Muhammad Faizan Rafique (2303.005.KHI.DEG)

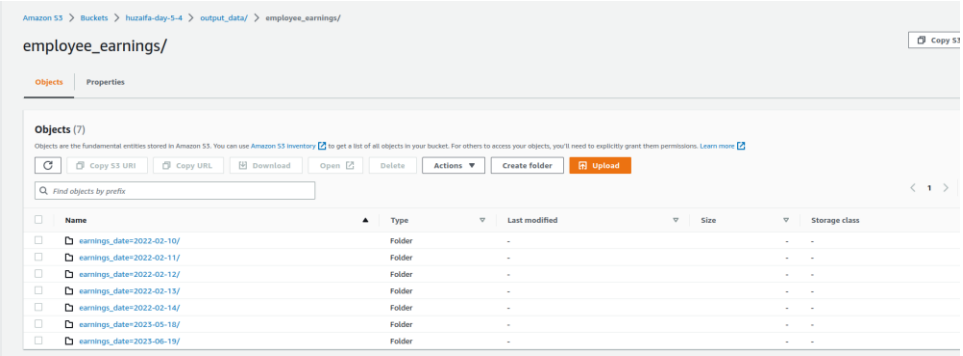
Pair Partner 2: - Syed Muhammad Hammad Irshad(2303.KHI.DEG.032)

UNIT 5.4:

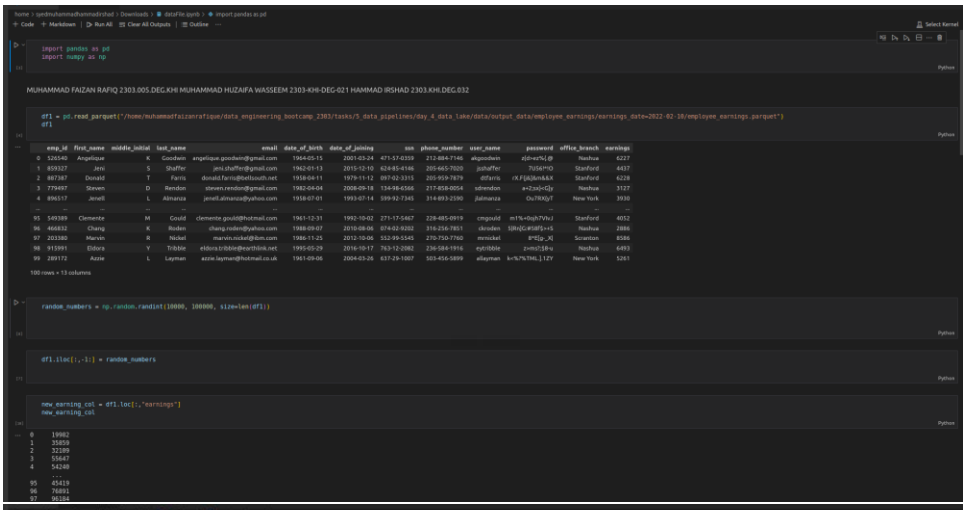
Assignment

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).



```
import pandas as pd
import numpy as np
```

MUHAMMAD FAIZAN RAFIQ 2303.005.DEG.KHI MUHAMMAD HUZAIFA WASSEEM 2303-KHI-DEG-021 HAMMAD IRSHAD 2303.KHI.DEG.032

```
df1 = pd.read_parquet("/home/muhammadfaizanrafique/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings/earnings_date=2022-02-10/employee_earnings.parquet")
df1
```

	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	jishaffer	7)(5E!*!O	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[A]&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;sa)=Gjv	Nashua	3127
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX0jT	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%+0sjh7VkJ	Stanford	4052
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	874-02-9202	316-256-7851	ckroden	5[Rh]C:#58F5=+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[p_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(!18=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=%7%TML;1ZY	New York	5261

100 rows × 13 columns

```
random_numbers = np.random.randint(10000, 100000, size=len(df1))
```

```
df1.iloc[:,-1] = random_numbers
```

```
new_earning_col = df1.loc[:,"earnings"]
new_earning_col
```

```
0    19982
1    35859
2    32189
3    55647
4    54240
     ...
95   45419
96   76891
97   96184
Name: earnings, Length: 100, dtype: int64
```

```
df1.head()
```

	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	19219
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jishaffer	7)(5E!*!O	Stanford	21040
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F[A]&m&&X	Stanford	60196
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon	a+2;sa)=Gjv	Nashua	29960
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX0jT	New York	37831

```
path = '/home/muhammadfaizanrafique/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings/earnings_date=2023-05-18/employee_earnings.parquet'
df1.to_parquet(path)
```

```
df2 = pd.read_parquet('/home/muhammadfaizanrafique/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings/earnings_date=2022-02-12/employee_earnings.parquet')
df2
```

	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	6176
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jishaffer	7)(5E!*!O	Stanford	8145
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F[A]&m&&X	Stanford	5278
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon	a+2;sa)=Gjv	Nashua	9547
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX0jT	New York	4138
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31	1992-10-02	271-17-5467	228-485-0919	cmgould	m1%+0sjh7VkJ	Stanford	3402
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	874-02-9202	316-256-7851	ckroden	5[Rh]C:#58F5=+S	Nashua	5373
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25	2012-10-06	552-99-5545	270-750-7760	mrnickel	8"E[p_X]	Scranton	7284
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29	2016-10-17	763-12-2082	236-584-1916	eytribble	z=ms(!18=u	Nashua	2810
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06	2004-03-26	637-29-1007	503-456-5899	allayman	k=%7%TML;1ZY	New York	2295

100 rows × 13 columns

```
random_numbers = np.random.randint(10000, 100000, size=len(df2))
df1.iloc[:,-1] = random_numbers
new_earning_col = df1.loc[:,'earnings']
new_earning_col
```

```
0    98088
1    14329
2    78158
3    96355
4    54339
     ...
95   28349
96   73849
97   18889
98   94604
99   96935
Name: earnings, Length: 100, dtype: int64
```

```
df2.head()
df2
```

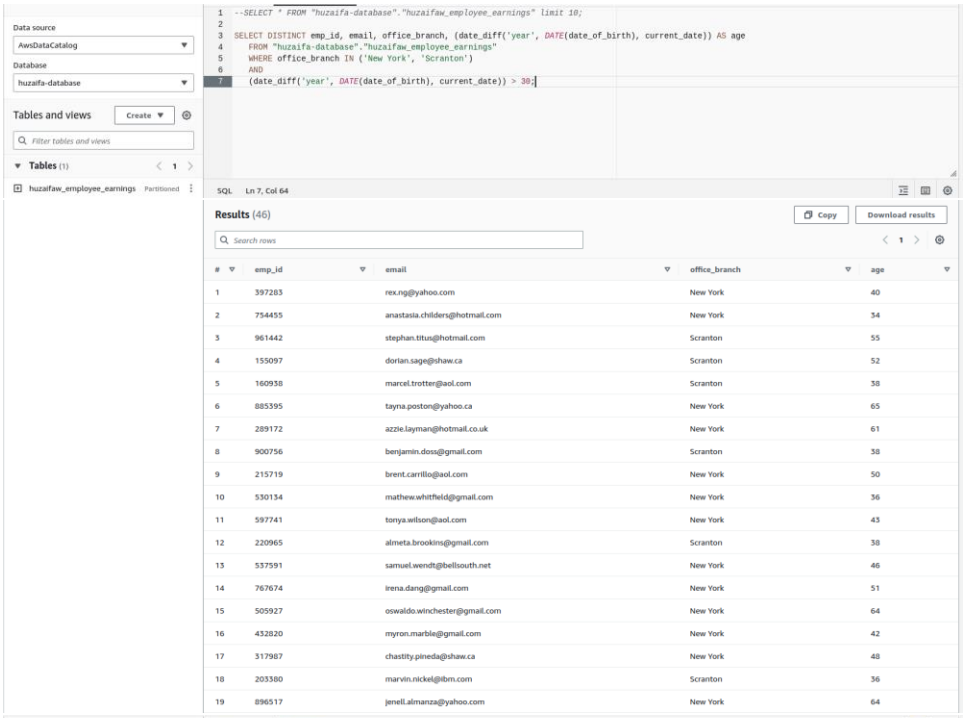
	emp_id	first_name	middle_initial	last_name	email	date_of_birth	date_of_joining	ssn	phone_number	user_name	password	office_branch	earnings
0	526540	Angelique	K	Goodwin	angelique.goodwin@gmail.com	1964-05-15	2001-03-24	471-57-0359	212-884-7146	akgoodwin	z@d=ez%(@	Nashua	6176
1	859327	Jeni	S	Shaffer	jeni.shaffer@gmail.com	1962-01-13	2015-12-10	624-85-4146	205-665-7020	jishaffer	7)(5E!*!O	Stanford	8145
2	887387	Donald	T	Farris	donald.farris@bellsouth.net	1958-04-11	1979-11-12	097-02-3315	205-959-7879	dtfarris	rX.F[A]&m&&X	Stanford	5278
3	779497	Steven	D	Rendon	steven.rendon@gmail.com	1982-04-04	2008-09-18	134-98-6566	217-858-0054	sdrendon	a+2;sa)=Gjv	Nashua	9547
4	896517	Jenell	L	Almanza	jenell.almanza@yahoo.com	1958-07-01	1993-07-14	599-92-7345	314-893-2590	jlalmanza	Ou7RX0jT	New York	4138
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	549389	Clemente	M	Gould	clemente.gould@hotmail.com	1961-12-31	1992-10-02	271-17-5467	228-485-0919	cmgould	m1%+0sjh7VkJ	Stanford	3402
96	466832	Chang	K	Roden	chang.roden@yahoo.com	1988-09-07	2010-08-06	874-02-9202	316-256-7851	ckroden	5[Rh]C:#58F5=+S	Nashua	5373
97	203380	Marvin	R	Nickel	marvin.nickel@ibm.com	1986-11-25	2012-10-06	552-99-5545	270-750-7760	mrnickel	8"E[p_X]	Scranton	7284
98	915991	Eldora	Y	Tribble	eldora.tribble@earthlink.net	1995-05-29	2016-10-17	763-12-2082	236-584-1916	eytribble	z=ms(!18=u	Nashua	2810
99	289172	Azzie	L	Layman	azzie.layman@hotmail.co.uk	1961-09-06	2004-03-26	637-29-1007	503-456-5899	allayman	k=%7%TML;1ZY	New York	2295

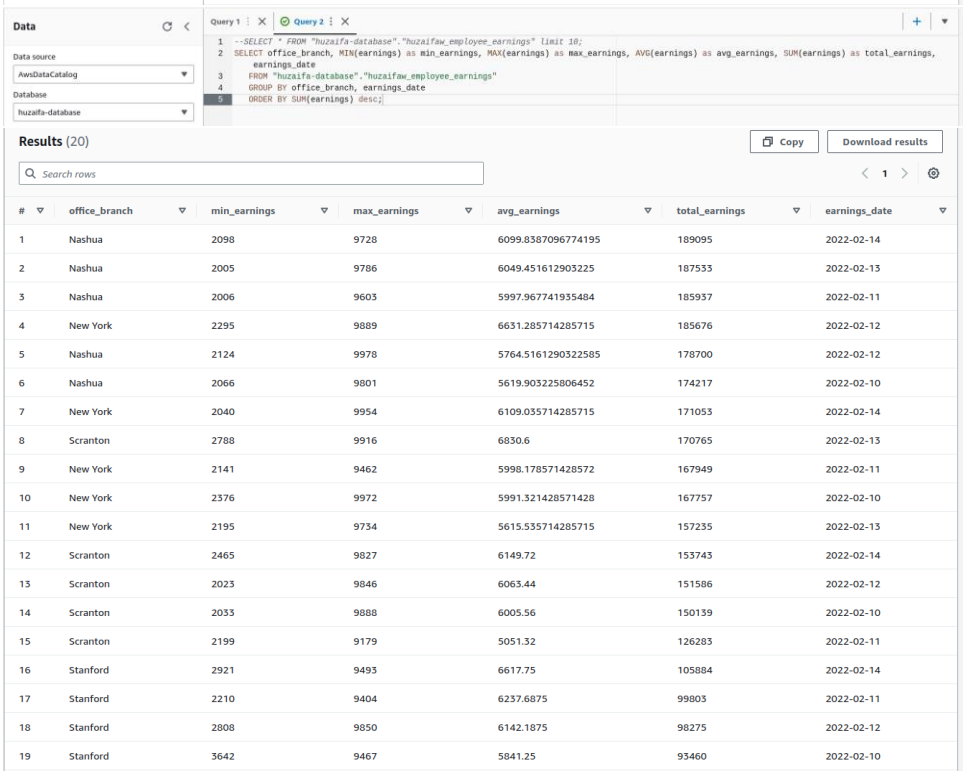
100 rows × 13 columns

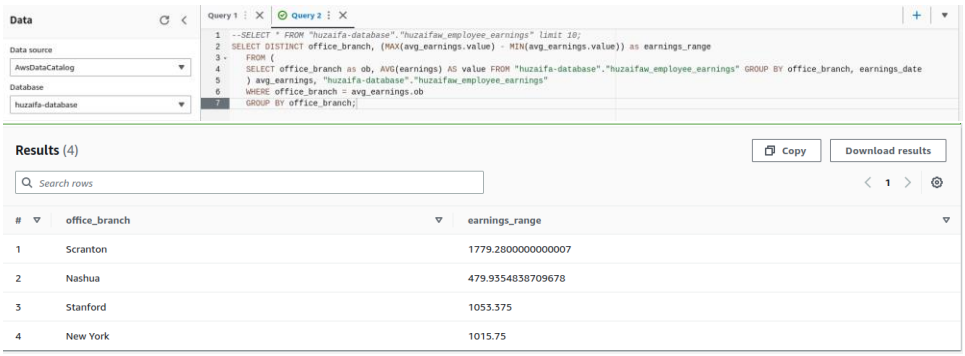
```
new_path = '/home/muhammadfaizanrafique/data_engineering_bootcamp_2303/tasks/5_data_pipelines/day_4_data_lake/data/output_data/employee_earnings/earnings_date=2023-06-10/employee_earnings.parquet'
df2.to_parquet(new_path)
```

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

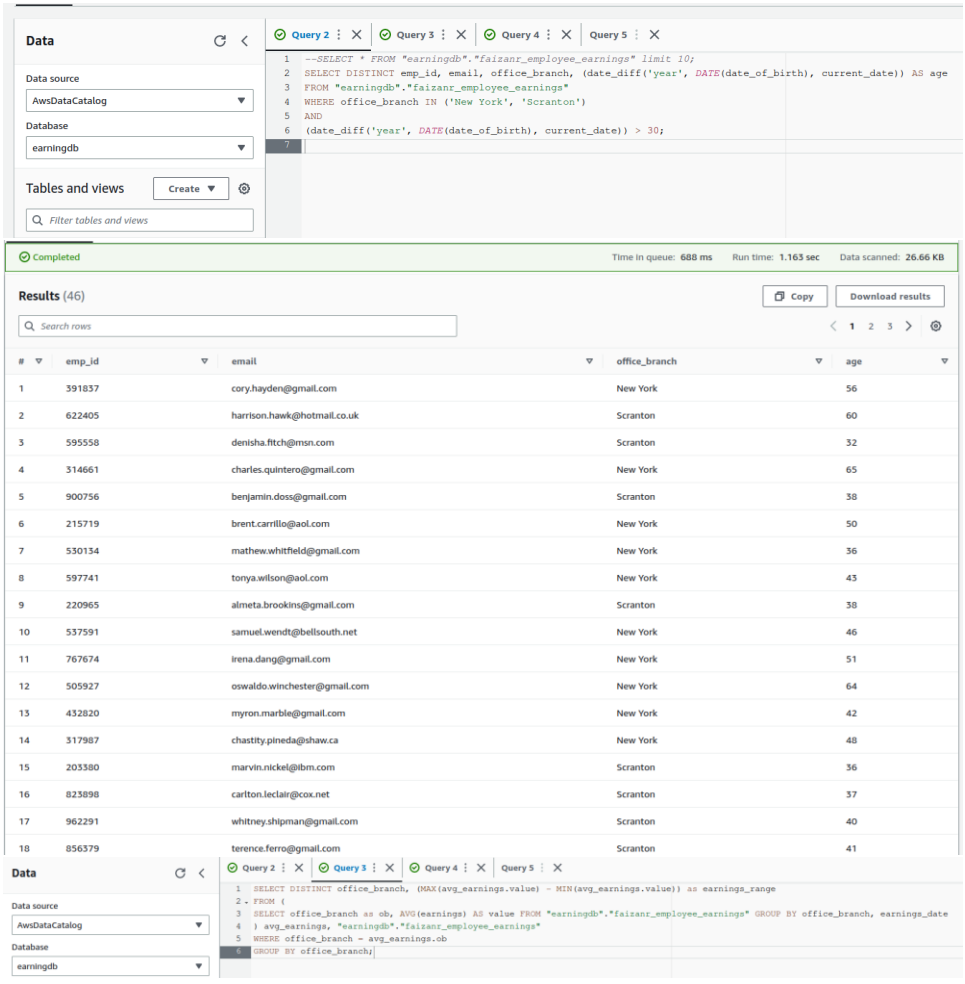
PREVIOUS DATA:

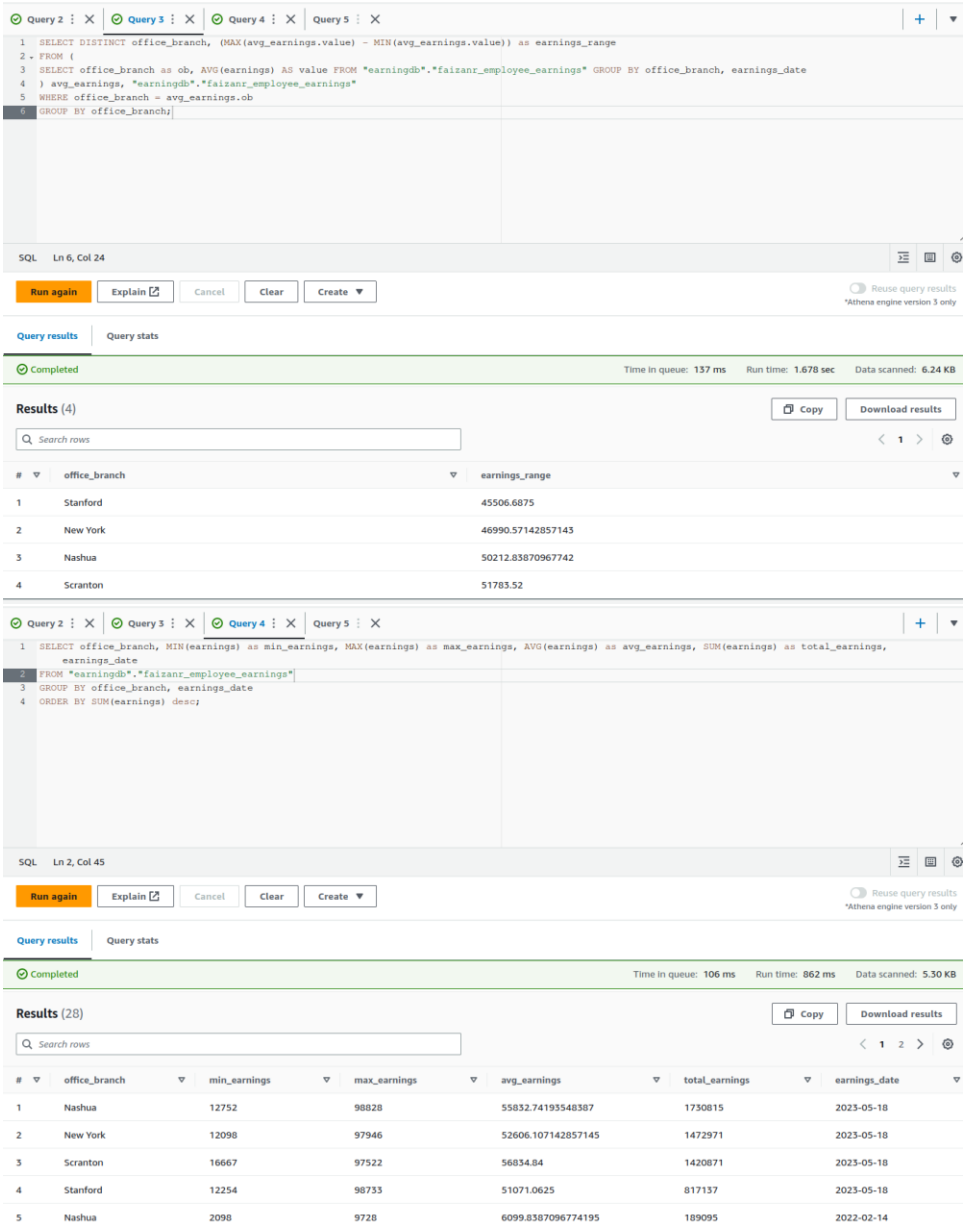

Data source
AwsDataCatalog
Database
huzaifa-database
Tables and views
Create
Filter tables and views
Tables (1)
huzaifaw_employee_earnings
Partitioned
1 --SELECT * FROM "huzaifa-database"."huzaifaw_employee_earnings" limit 10;
2
3 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
4 FROM "huzaifa-database"."huzaifaw_employee_earnings"
5 WHERE office_branch IN ('New York', 'Scranton')
6 AND
7 (date_diff('year', DATE(date_of_birth), current_date)) > 30;
SQL
Ln 7, Col 64
Results (46)
Copy
Download results
Search rows
emp_id email office_branch age
1 397283 rex.ng@yahoo.com New York 40
2 754455 anastasia.childers@hotmail.com New York 34
3 961442 stephan.titus@hotmail.com Scranton 55
4 155097 dorian.sage@shaw.ca Scranton 52
5 160938 marcel.trotter@aol.com Scranton 38
6 885395 tayna.poston@yahoo.ca New York 65
7 289172 azzie.layman@hotmail.co.uk New York 61
8 900756 benjamin.doss@gmail.com Scranton 38
9 215719 brent.carrillo@aol.com New York 50
10 530134 mathew.whitfield@gmail.com New York 36
11 597741 tonya.wilson@aol.com New York 43
12 220965 almeta.brookins@gmail.com Scranton 38
13 537591 samuel.wendt@bellsouth.net New York 46
14 767674 irena.dang@gmail.com New York 51
15 505927 oswaldo.winchester@gmail.com New York 64
16 452820 myron.marble@gmail.com New York 42
17 317987 chastity.pineda@shaw.ca New York 48
18 203380 marvin.nickel@ibm.com Scranton 36
19 896517 jenell.almanza@yahoo.com New York 64


Data
Data source
AwsDataCatalog
Database
huzaifa-database
Query 1
Query 2
1 --SELECT * FROM "huzaifa-database"."huzaifaw_employee_earnings" limit 10;
2 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
3 FROM "huzaifa-database"."huzaifaw_employee_earnings"
4 GROUP BY office_branch, earnings_date
5 ORDER BY SUM(earnings) desc;
Results (20)
Copy
Download results
Search rows
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 9603 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 6830.6 170765 2022-02-13
9 New York 2141 9462 5998.178571428572 167949 2022-02-11
10 New York 2376 9972 5991.321428571428 167757 2022-02-10
11 New York 2195 9734 5615.535714285715 157235 2022-02-13
12 Scranton 2465 9827 6149.72 153743 2022-02-14
13 Scranton 2023 9846 6063.44 151586 2022-02-12
14 Scranton 2033 9888 6005.56 150139 2022-02-10
15 Scranton 2199 9179 5051.32 126283 2022-02-11
16 Stanford 2921 9493 6617.75 105884 2022-02-14
17 Stanford 2210 9404 6237.6875 99803 2022-02-11
18 Stanford 2808 9850 6142.1875 98275 2022-02-12
19 Stanford 3642 9467 5841.25 93460 2022-02-10

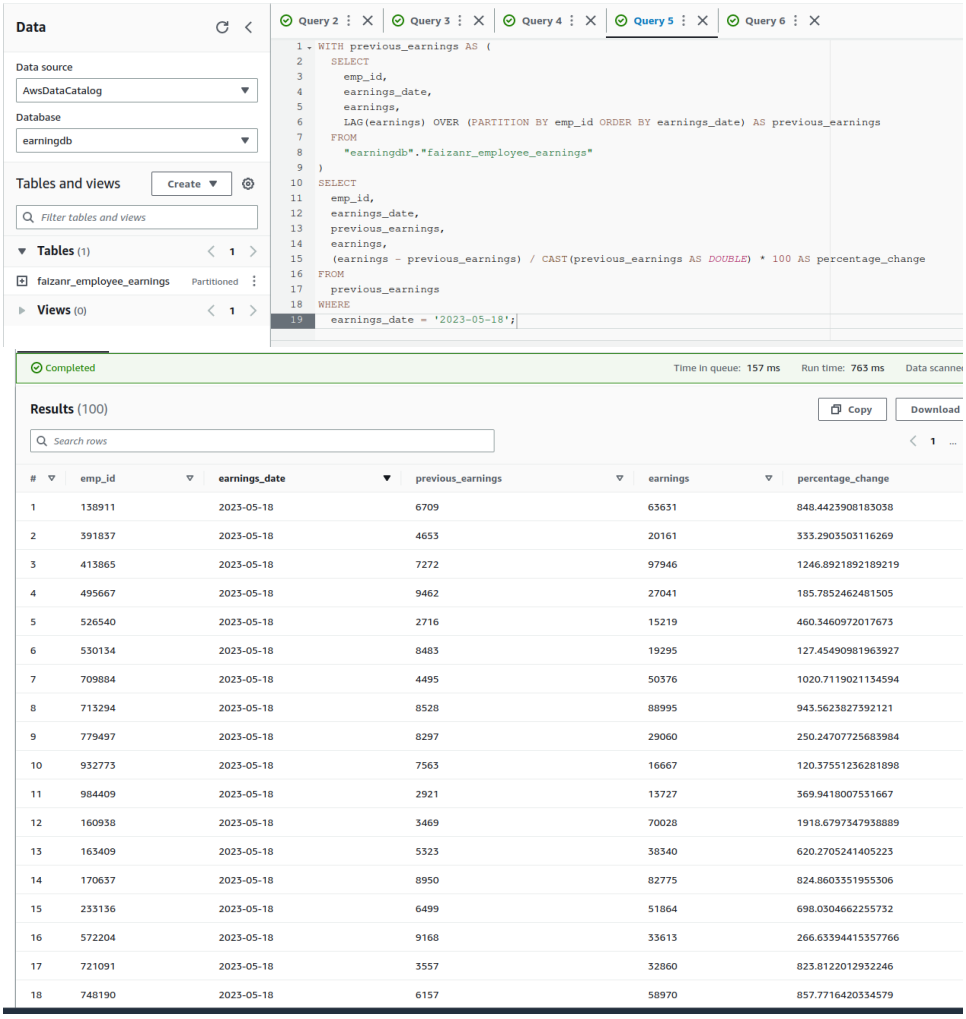

Data
Data source
AwsDataCatalog
Database
huzaifa-database
Query 1
Query 2
1 --SELECT * FROM "huzaifa-database"."huzaifaw_employee_earnings" limit 10;
2 SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
3 FROM (
4 SELECT office_branch as ob, AVG(earnings) AS value FROM "huzaifa-database"."huzaifaw_employee_earnings" GROUP BY office_branch, earnings_date
5) avg_earnings, "huzaifa-database"."huzaifaw_employee_earnings"
6 WHERE office_branch = avg_earnings.ob
7 GROUP BY office_branch;
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

After Data:


Data
Data source
AwsDataCatalog
Database
earningdb
Tables and views
Create
Filter tables and views
Query 2
Query 3
Query 4
Query 5
1 --SELECT * FROM "earningdb"."faizanr_employee_earnings" limit 10;
2 SELECT DISTINCT emp_id, email, office_branch, (date_diff('year', DATE(date_of_birth), current_date)) AS age
3 FROM "earningdb"."faizanr_employee_earnings"
4 WHERE office_branch IN ('New York', 'Scranton')
5 AND
6 (date_diff('year', DATE(date_of_birth), current_date)) > 30;
7
Completed
Time in queue: 688 ms
Run time: 1.163 sec
Data scanned: 26.66 KB
Results (46)
Copy
Download results
Search rows
1 2 3
emp_id email office_branch age
1 391837 cory.hayden@gmail.com New York 56
2 622405 harrison.hawk@hotmail.co.uk Scranton 60
3 595558 denisha.fitch@msn.com Scranton 32
4 314661 charles.quintero@gmail.com New York 65
5 900756 benjamin.doss@gmail.com Scranton 38
6 215719 brent.carrillo@aol.com New York 50
7 530134 mathew.whitfield@gmail.com New York 36
8 597741 tonya.wilson@aol.com New York 43
9 220965 almeta.brookins@gmail.com Scranton 38
10 537591 samuel.wendt@bellsouth.net New York 46
11 767674 irena.dang@gmail.com New York 51
12 505927 oswaldo.winchester@gmail.com New York 64
13 432820 myron.marble@gmail.com New York 42
14 317987 chastity.pineda@shaw.ca New York 48
15 203380 marvin.nickel@ibm.com Scranton 36
16 823898 carlton.leclair@cox.net Scranton 37
17 962291 whitney.shipman@gmail.com Scranton 40
18 856379 terence.ferro@gmail.com Scranton 41
Data
Data source
AwsDataCatalog
Database
earningdb
Query 2
Query 3
Query 4
Query 5
1 SELECT DISTINCT office_branch, (MAX(avg_earnings.value) - MIN(avg_earnings.value)) as earnings_range
2 FROM (
3 SELECT office_branch as ob, AVG(earnings) AS value FROM "earningdb"."faizanr_employee_earnings" GROUP BY office_branch, earnings_date
4) avg_earnings, "earningdb"."faizanr_employee_earnings"
5 WHERE office_branch = avg_earnings.ob
6 GROUP BY office_branch;



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



#	emp_id	earnings_date	previous_earnings	earnings	percentage_change
1	138911	2023-05-18	6709	63631	848.4423908183038
2	391837	2023-05-18	4653	20161	333.2903503116269
3	413865	2023-05-18	7272	97946	1246.8921892189219
4	495667	2023-05-18	9462	27041	185.7852462481505
5	526540	2023-05-18	2716	15219	460.3460972017673
6	530134	2023-05-18	8483	19295	127.45490981963927
7	709884	2023-05-18	4495	50376	1020.7119021134594
8	713294	2023-05-18	8528	88995	943.5623827392121
9	779497	2023-05-18	8297	29060	250.24707725683984
10	932773	2023-05-18	7563	16667	120.37551236281898
11	984409	2023-05-18	2921	13727	369.9418007531667
12	160938	2023-05-18	3469	70028	1918.6797347938889
13	163409	2023-05-18	5323	38340	620.2705241405223
14	170637	2023-05-18	8950	82775	824.8603351955306
15	233136	2023-05-18	6499	51864	698.0304662255732
16	572204	2023-05-18	9168	33613	266.63394415357766
17	721091	2023-05-18	3557	32860	823.8122012932246
18	748190	2023-05-18	6157	58970	857.7716420334579