Προχωρημένα Θέματα Βάσεων Δεδομένων
Εξαμηνιαία Εργασία

Ομάδα 37
Βασίλειος Βρεττός - el18126,
Νικόλαος Παγώνας - el18175

1   Εγκατάσταση Συστημάτων

Ολόκληρη η διαδικασία εγκατάστασης των Apache Spark και Hadoop DFS παρουσιάζεται στο Github Repository της ομάδας, με script εγκατάστασης και μικρές επεξηγήσεις (Github Link).

Η δημιουργία των RDD και Dataframe χρησιμοποιώντας τα δεδομένα που μας δίνονται (parquet files και zone lookup csv) είναι μια πολύ απλή διαδικασία αφού δημιουργήσουμε το SparkSession στην Scala.

2   Ερωτήματα/Queries

Σε αυτό το σημείο, παρουσιάζουμε τα ερωτήματα που κληθήκαμε να απαντήσουμε, το SQL query που τα απαντάει, το αποτέλεσμα του query και ο χρόνος εκτέλεσης τους.

3 Αποτελέσματα

3.1 Με έναν worker

3.1.1 Query 1

Pickup_datetime	Dropoff_datetime	Pickup_zone	Dropoff_zone	Trip_distance	Total_amount	Tip_amount
2022-03-17T12:27:47.000+02:00	2022-03-17T12:27:58.000+02:00	Battery Park	Battery Park	0.0	45.8	40.0

3.1.2 Query 2

Month	Pickup_zone	Dropoff_zone	Tolls_amount	Pickup_datetime	Dropoff_datetime	Trip_distance	Total_amount	Tip_amount
1	JFK Airport	Schuylerville/Edgewater Park	82.45	2022-01-18T23:51:55.000+02:00	2022-01-19T00:16:19.000+02:00	16.79	129.25	0.0
2	TriBeCa/Civic Center	TriBeCa/Civic Center	77.0	2022-02-23T22:09:57.000+02:00	2022-02-23T22:10:07.000+02:00	0.0	132.3	0.0
3	Midtown East	Sutton Place/Turtle Bay North	139.0	2022-03-26T05:02:51.000+02:00	2022-03-26T05:03:45.000+02:00	0.1	145.3	0.0
4	West Village	West Village	911.87	2022-04-29T04:31:21.000+03:00	2022-04-29T04:32:30.000+03:00	0.0	918.67	0.0
5	Upper West Side South	West Chelsea/Hudson Yards	813.75	2022-05-21T16:47:48.000+03:00	2022-05-21T17:05:47.000+03:00	2.4	845.55	0.0
6	Financial District North	Financial District North	77.0	2022-06-23T20:43:16.000+03:00	2022-06-23T20:48:14.000+03:00	0.6	91.31	0.01
6	Lincoln Square East	JFK Airport	800.09	2022-06-12T16:51:46.000+03:00	2022-06-12T17:56:48.000+03:00	22.0	870.89	0.0

3.1.3 Query 3 (SQL)

Month	Starting_day	Average_distance	Average_amount
1	1	5.394411652119241	19.565880333233707
1	16	5.053523299628317	18.849869490208068
2	1	6.224118794096238	19.21423910463858
2	16	5.809556810665932	19.878940350438274
3	1	6.442990326813408	20.296908557261577
3	16	5.504069666391406	20.755608446018496
4	1	5.618632523223211	21.120098919973096
4	16	5.546105056180586	21.037355457801407
5	1	6.198675369065214	21.518062994494297
5	16	7.741069199080349	22.31758523166161
6	1	6.2602549221727815	22.04635959581888
6	16	6.111722013164792	21.855333058482078

3.1.4 Query 3 (RDD)

Month	Starting_day	Average_distance	Average_amount
1	1	5.394405	19.573868
1	16	5.053545	18.85672
2	1	6.2240744	19.214857
2	16	5.809599	19.879427
3	1	6.443038	20.293175
3	16	5.5040126	20.75156
4	1	5.618543	21.117859
4	16	5.546019	21.03438
5	1	6.1986613	21.515226
5	16	7.7409587	22.313637
6	1	6.2602882	22.04124
6	16	6.1116147	21.851416

3.1.5 Query 4

Max_passenger_count	Hour	Weekday	Rank
9.0	1	0	1
9.0	20	0	2
9.0	17	0	3
9.0	11	1	1
9.0	19	1	2
8.0	20	1	3
8.0	20	2	1
8.0	12	2	2
8.0	16	2	3
9.0	21	3	1
8.0	17	3	2
8.0	1	3	3
9.0	18	4	1
8.0	16	4	2
8.0	19	4	3
8.0	21	5	1
8.0	18	5	2
8.0	22	5	3
9.0	16	6	1
8.0	23	6	2
8.0	22	6	3

3.1.6 Query 5

Day	Month	Percentage	Rank
29	1	21.836209686979053	1
15	1	19.702726107621427	2
21	1	19.58041382314713	3
30	1	19.572931292565993	4
22	1	19.568150619535107	5
4	2	19.824773873877216	1
5	2	19.724573811111668	2
10	2	19.64430000372954	3
9	2	19.601864761149884	4
6	2	19.6000134462146	5
9	3	19.818872357026137	1
30	3	19.641567232750592	2
12	3	19.581624438920127	3
11	3	19.570233566718244	4
10	3	19.56733343581587	5
7	4	19.456269554817997	1
1	4	19.439107330878002	2
6	4	19.3991639831137	3
27	4	19.306388414643607	4
5	4	19.273309702970383	5
4	5	19.45314477000875	1
12	5	19.43960396912661	2
11	5	19.28322707402209	3
10	5	19.230092123348506	4
6	5	19.22714433407941	5
16	6	19.308355808593834	1
23	6	19.247823870672885	2
8	6	19.17339590901764	3
9	6	19.168779190113714	4
17	6	19.137251898695766	5

3.2 Με δύο workers

3.2.1 Query 1

Pickup_datetime	Dropoff_datetime	Pickup_zone	Dropoff_zone	Trip_distance	Total_amount	Tip_amount
2022-03-17T12:27:47.000+02:00	2022-03-17T12:27:58.000+02:00	Battery Park	Battery Park	0.0	45.8	40.0

3.2.2 Query 2

Month	Pickup_zone	Dropoff_zone	Tolls_amount	Pickup_datetime	Dropoff_datetime	Trip_distance	Total_amount	Tip_amount
1	JFK Airport	Schuylerville/Edgewater Park	82.45	2022-01-18T23:51:55.000+02:00	2022-01-19T00:16:19.000+02:00	16.79	129.25	0.0
2	TriBeCa/Civic Center	TriBeCa/Civic Center	77.0	2022-02-23T22:09:57.000+02:00	2022-02-23T22:10:07.000+02:00	0.0	132.3	0.0
3	Midtown East	Sutton Place/Turtle Bay North	139.0	2022-03-26T05:02:51.000+02:00	2022-03-26T05:03:45.000+02:00	0.1	145.3	0.0
4	West Village	West Village	911.87	2022-04-29T04:31:21.000+03:00	2022-04-29T04:32:30.000+03:00	0.0	918.67	0.0
5	Upper West Side South	West Chelsea/Hudson Yards	813.75	2022-05-21T16:47:48.000+03:00	2022-05-21T17:05:47.000+03:00	2.4	845.55	0.0
6	Financial District North	Financial District North	77.0	2022-06-23T20:43:16.000+03:00	2022-06-23T20:48:14.000+03:00	0.6	91.31	0.01
6	Lincoln Square East	JFK Airport	800.09	2022-06-12T16:51:46.000+03:00	2022-06-12T17:56:48.000+03:00	22.0	870.89	0.0

3.2.3 Query 3 (SQL)

Month	Starting_day	Average_distance	Average_amount
1	1	5.394411652119241	19.565880333233707
1	16	5.053523299628317	18.849869490208068
2	1	6.224118794096238	19.21423910463858
2	16	5.809556810665932	19.878940350438274
3	1	6.442990326813408	20.29690855726158
3	16	5.504069666391406	20.755608446018496
4	1	5.618632523223211	21.1200989199731
4	16	5.546105056180586	21.037355457801407
5	1	6.198675369065214	21.518062994494297
5	16	7.741069199080349	22.31758523166161
6	1	6.2602549221727815	22.04635959581888
6	16	6.111722013164792	21.855333058482078

3.2.4 Query 3 (RDD)

Month	Starting_day	Average_distance	Average_amount
1	1	5.394405	19.573868
1	16	5.053545	18.85672
2	1	6.2240744	19.214857
2	16	5.809599	19.879427
3	1	6.443038	20.293175
3	16	5.5040126	20.75156
4	1	5.618543	21.117859
4	16	5.546019	21.03438
5	1	6.1986613	21.515226
5	16	7.7409587	22.313637
6	1	6.2602882	22.04124
6	16	6.1116147	21.851416

3.2.5 Query 4

Max_passenger_count	Hour	Weekday	Rank
9.0	1	0	1
9.0	20	0	2
9.0	17	0	3
9.0	11	1	1
9.0	19	1	2
8.0	20	1	3
8.0	20	2	1
8.0	12	2	2
8.0	16	2	3
9.0	21	3	1
8.0	17	3	2
8.0	1	3	3
9.0	18	4	1
8.0	16	4	2
8.0	19	4	3
8.0	21	5	1
8.0	18	5	2
8.0	22	5	3
9.0	16	6	1
8.0	23	6	2
8.0	22	6	3

3.2.6 Query 5

Day	Month	Percentage	Rank
29	1	21.836209686979053	1
15	1	19.702726107621427	2
21	1	19.58041382314713	3
30	1	19.572931292565993	4
22	1	19.568150619535107	5
4	2	19.824773873877216	1
5	2	19.724573811111668	2
10	2	19.64430000372954	3
9	2	19.601864761149884	4
6	2	19.6000134462146	5
9	3	19.818872357026137	1
30	3	19.641567232750592	2
12	3	19.581624438920127	3
11	3	19.570233566718244	4
10	3	19.56733343581587	5
7	4	19.456269554817997	1
1	4	19.439107330878002	2
6	4	19.3991639831137	3
27	4	19.306388414643607	4
5	4	19.273309702970383	5
4	5	19.45314477000875	1
12	5	19.43960396912661	2
11	5	19.28322707402209	3
10	5	19.230092123348506	4
6	5	19.22714433407941	5
16	6	19.308355808593834	1
23	6	19.247823870672885	2
8	6	19.17339590901764	3
9	6	19.168779190113714	4
17	6	19.137251898695766	5

4 Χρόνοι

4.1 Με έναν worker

Query	Time (sec)
Q1	14.565511047
Q2	35.266391669
Q3 (SQL)	8.252414362
Q3 (RDD)	27.617137578
Q4	9.212746264
Q5	10.044835744

4.2 Με δύο workers

Query	Time (sec)
Q1	15.747357043
Q2	31.553428541
Q3 (SQL)	7.263829931
Q3 (RDD)	28.64680576
Q4	9.302175079
Q5	8.537400986

4.1 Με έναν worker