Miguel_Diaz_ClassWork_Sesson_8

Cloud Computing

Athena SQL

1- Creating tables:

```
CREATE EXTERNAL TABLE IF NOT EXISTS `aiml_class`.`LRS_miguel_data` (
    `route_id` varchar(50), `segment_start_latitude` float,
    `segment_start_longitude` float, `segment_end_latitude` float,
    `segment_end_longitude` float, `segment_start_measure` float,
    `segment_end_measure` float, `geohash_lrs` varchar(10)
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('field.delim' = ',')
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io
    .HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://reactorlab/Miguel/LRS_data/'
TBLPROPERTIES ('classification' = 'csv')
```

```
CREATE EXTERNAL TABLE IF NOT EXISTS `aiml_class`.`Miguel_gps_data` (
    `datapointid` varchar(50), `journeyid` varchar(50), `latitude`
    float, `longitude` float, `month` int, `day` int, `hour` int,
    `geohash` varchar(10), `speed` float
) COMMENT "Miguel data class SQL AWS"
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('field.delim' = ',')
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io
    .HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://reactorlab/Miguel/GPS_data/'
TBLPROPERTIES ('classification' = 'csv')
```

GPS

SQL Ln 1, Col 55

Run again | Explain | Cancel | Clear | Create

Reuse query results
up to 60 minutes ago

Query results | Query stats

Completed — Time in queue: 116 ms — Run time: 629 ms — Data scanned: 955.76 KB

Results (10)

Copy | Download results

Search rows

#	datapointid	journeyid	latitude	longitude	month	day	hour	geohash	speed
1	757baad3-eec2-4cc1-9c9d-b1aef1d09acc	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503708	-94.32624	10	1	5	9zkuje	21.88
2	bafd0a8f-8fe5-4fbf-8490-d70f3866b526	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503708	-94.32593	10	1	5	9zkuje	32.25
3	4eb9fc5f-6541-4673-94af-d818e492c8e9	53ca9cce2fb74c2048a1f436c19e01b786086727	41.5037	-94.32553	10	1	5	9zkuje	40.32
4	82e68eb8-f6a2-4417-a317-6efc043fab88	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503696	-94.32508	10	1	5	9zkuje	44.92
5	d682dfbd-0304-4fe4-8859-3acbe4808a2a	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503696	-94.32461	10	1	5	9zkuje	47.23
6	7a546536-42fe-418e-aa5b-823373254a87	53ca9cce2fb74c2048a1f436c19e01b786086727	41.50369	-94.324135	10	1	5	9zkuje	47.23
7	d577364d-c36e-4151-8d6d-5fce79a3c8e3	53ca9cce2fb74c2048a1f436c19e01b786086727	41.50368	-94.32365	10	1	5	9zkuje	48.38
8	ae9ed02e-cfa0-456c-addf-f579c998c056	53ca9cce2fb74c2048a1f436c19e01b786086727	41.50368	-94.32314	10	1	5	9zkuje	49.53
9	29dc09a2-c357-42d6-8a3a-109858392c54	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503685	-94.32264	10	1	5	9zkuje	49.53
10	36240474-41bb-4e0b-a0bf-3586cee433c9	53ca9cce2fb74c2048a1f436c19e01b786086727	41.50369	-94.32214	10	1	5	9zkuje	48.38

LRS

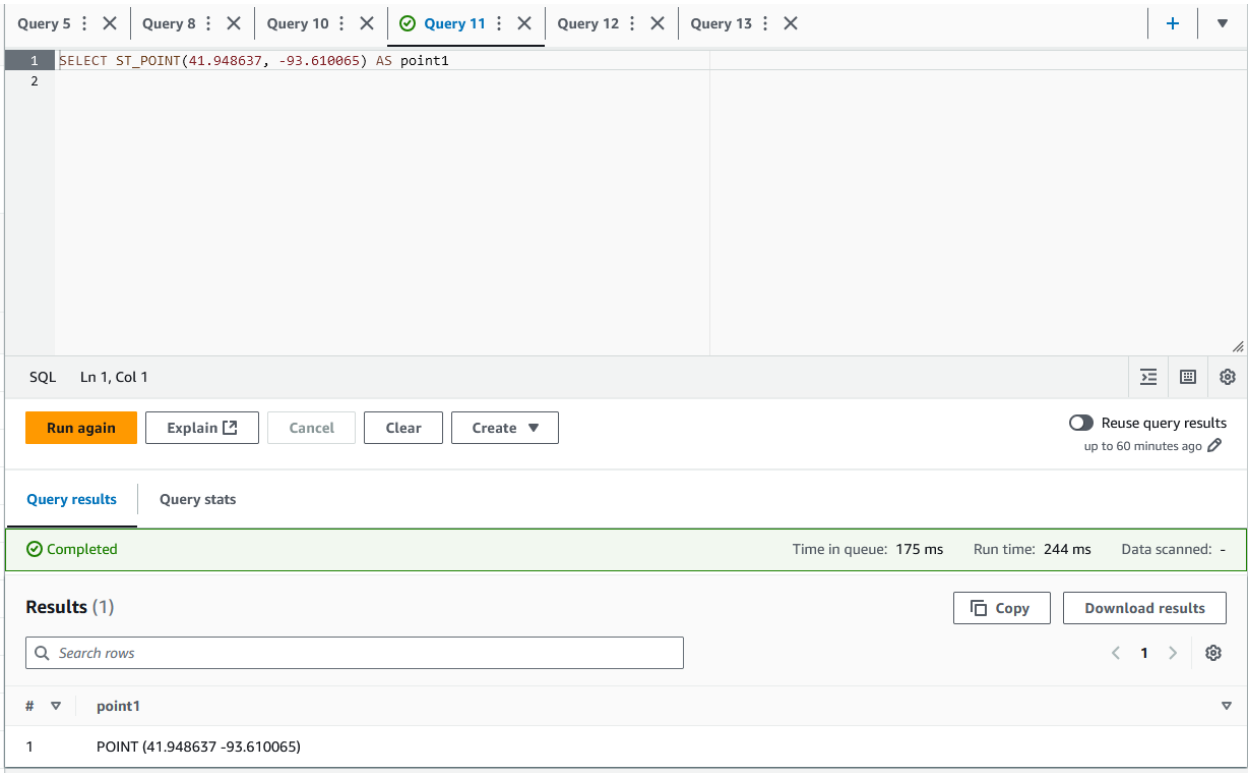
Results (10)

Copy | Download results

Search rows

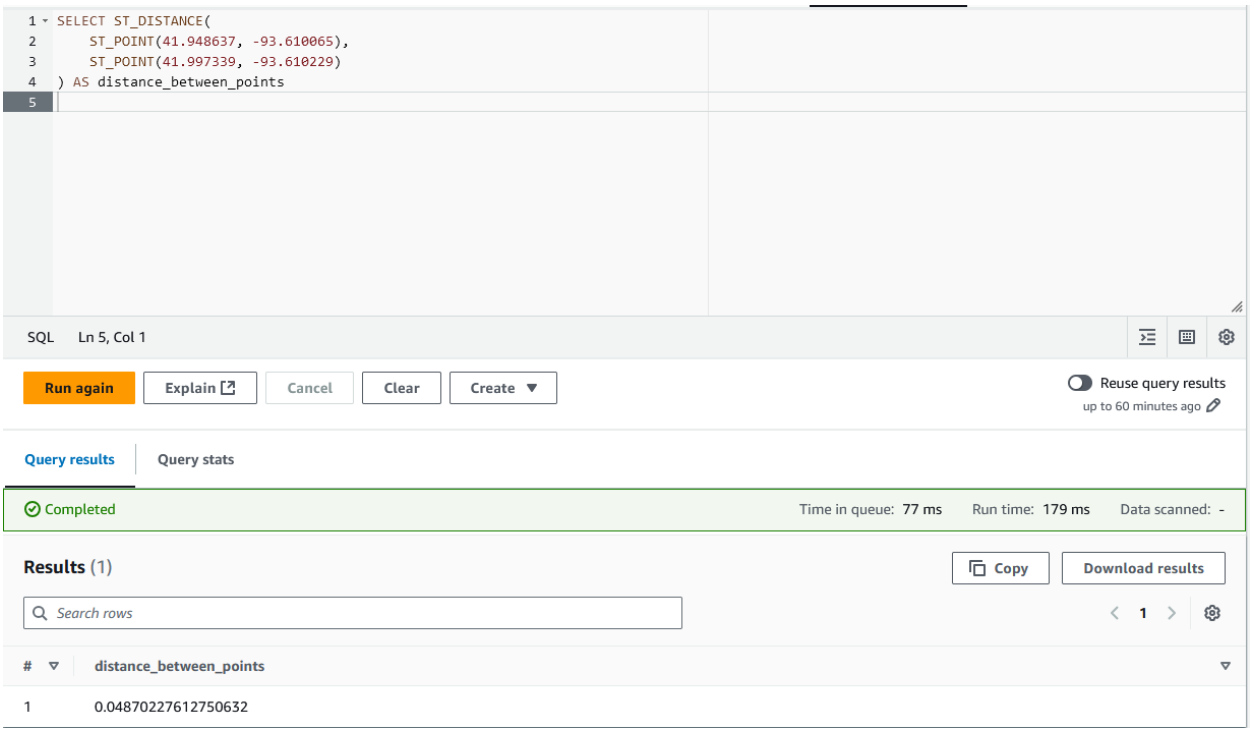
#	route_id	segment_start_latitude	segment_start_longitude	segment_end_latitude	segment_end_longitude	segment_start_measure	segment_end_measure	geohash_lrs
1	C000146110E	41.20839	-94.63829	41.208378	-94.63685	0.075	0.151	9zkdqm
2	C000146500E	41.446068	-94.31337	41.446068	-94.312485	14.626	14.672	9zkgvc
3	C000146500E	41.446053	-94.30814	41.44605	-94.3067	14.897	14.972	9zkgvc
4	C000146500E	41.446056	-94.30959	41.446053	-94.30814	14.822	14.897	9zkgvc
5	C000146500E	41.44606	-94.311035	41.446056	-94.30959	14.747	14.822	9zkgvc
6	C000146500E	41.44604	-94.30083	41.446033	-94.299385	15.277	15.352	9zkgy1
7	C000146500E	41.446033	-94.299065	41.44603	-94.297615	15.369	15.444	9zkgy1
8	C000146500E	41.44602	-94.29477	41.446014	-94.29333	15.592	15.667	9zkgy3
9	C000146500E	41.44593	-94.282814	41.445942	-94.281364	16.212	16.287	9zkgy9
10	C000146500E	41.44605	-94.26183	41.446037	-94.26084	17.302	17.353	9zkgz1

1.Make point feature from Point 1: 41.948637, -93.610065using ST_POINT

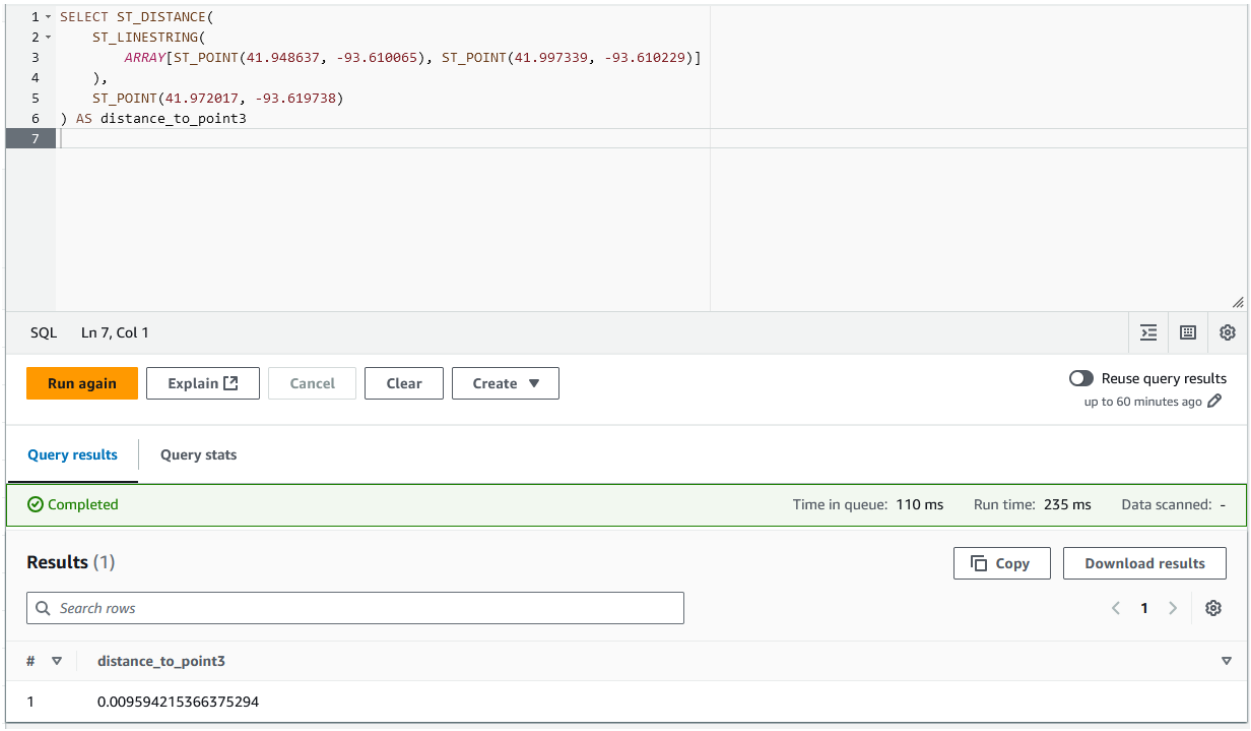


2.Find distance (ST_DISTANEC) between points below: (Ans: 0.048702)

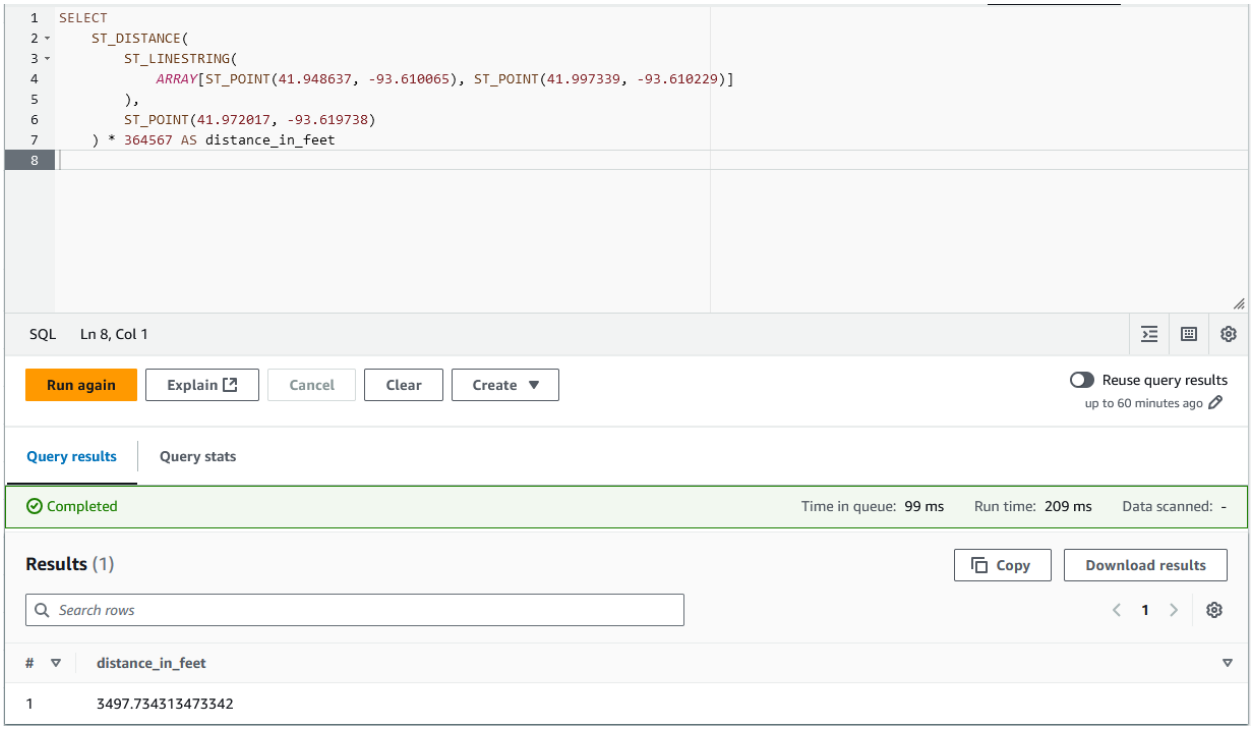
•Point 1: 41.948637, -93.610065, Point 2 : 41.997339, -93.610229



3.Create a line string (ST_LINESTRING) from above points and calculate distance from Point 3 : 41.972017, -93.619738 (Ans: 0.0095942)



4.Get it in feet by multiplying 364567 (Ans: 3497.7 ft)



5.Make point geometry from latitude and longitude in gps_data table

```
1  SELECT ST_POINT(latitude, longitude) AS point_geom
2  FROM miguel_gps_data
3
```

SQL Ln 2, Col 21

Run again | Explain | Cancel | Clear | Create

Reuse query results
up to 60 minutes ago

Query results | Query stats

Completed — Time in queue: 139 ms — Run time: 1.917 sec — Data scanned: 27.64 MB

Results (244,614)

Copy | Download results

Search rows

1 ...

#	point_geom
1	POINT (41.50370788574219 -94.32624053955078)
2	POINT (41.50370788574219 -94.325927734375)
3	POINT (41.503700256347656 -94.32553100585938)
4	POINT (41.50369644165039 -94.32508087158203)
5	POINT (41.50369644165039 -94.3246078491211)
6	POINT (41.50368881225586 -94.32413482666016)
7	POINT (41.50368118286133 -94.32364654541016)
8	POINT (41.50368118286133 -94.3231430053711)
9	POINT (41.503684997558594 -94.32263946533203)
74	POINT (41.49679183959961 -94.31802368164062)
75	POINT (41.49679183959961 -94.31810760498047)
76	POINT (41.49679183959961 -94.31818389892578)
77	POINT (41.49676513671875 -94.31824493408203)
78	POINT (41.49672317504883 -94.31826782226562)
79	POINT (41.49668502807617 -94.3182601928711)
80	POINT (41.496639251708984 -94.3182601928711)
81	POINT (41.49658203125 -94.3182601928711)
82	POINT (41.496524810791016 -94.31825256347656)
83	POINT (41.49646759033203 -94.31825256347656)
84	POINT (41.49641036987305 -94.31824493408203)
85	POINT (41.496360778808594 -94.31822204589844)

85	POINT (41.496360778808594 -94.31822204589844)
86	POINT (41.496337890625 -94.31818389892578)
87	POINT (41.496334075927734 -94.31816101074219)
88	POINT (41.496334075927734 -94.31814575195312)
89	POINT (41.496341705322266 -94.31812286376953)
90	POINT (41.49634552001953 -94.318115234375)
91	POINT (41.49634552001953 -94.318115234375)
92	POINT (41.49634552001953 -94.318115234375)
93	POINT (41.49314498901367 -94.24730682373047)
94	POINT (41.493186950683594 -94.24610137939453)
95	POINT (41.49323272705078 -94.24488830566406)
96	POINT (41.4932746887207 -94.24365997314453)
97	POINT (41.49332046508789 -94.24241638183594)
98	POINT (41.300865173339844 -94.35681915283203)
99	POINT (41.30038833618164 -94.35682678222656)
100	POINT (41.29983139038086 -94.35682678222656)

6.Create a buffer (ST_BUFFER) around 41.502954, -94.295980 for 0.05 units and filter all datapoints (ST_CONTAINS) from gps_data within the buffer (Ans: 54,581 rows)

```
WITH buffered_area AS (
    SELECT ST_BUFFER(ST_POINT(41.502954, -94.295980), 0.05) AS buffer
)
SELECT *
FROM miguel_gps_data
WHERE ST_CONTAINS(
    (SELECT buffer FROM buffered_area),
    ST_POINT(latitude, longitude)
)
```

SQL Ln 4, Col 9

Run again | Explain | Cancel | Clear | Create

Reuse query results up to 60 minutes ago

Query results | Query stats

Completed — Time in queue: 109 ms — Run time: 1.928 sec — Data scanned: 27.64 MB

Results (54,581)

Copy | Download results

Search rows

1 ...

	journeyid	latitude	longitude	month	day	hour	geohash	speed
-ae15-b7b50c8a9cb1	492d473589eb3d15e52dd67ea59d102d1f3fa08d	41.493145	-94.24731	10	1	1	9zkup3	123.26
9c9d-b1aef1d09acc	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503708	-94.32624	10	1	5	9zkuje	21.88
490-d70f3866b526	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503708	-94.32593	10	1	5	9zkuje	32.25
94af-d818e492c8e9	53ca9cce2fb74c2048a1f436c19e01b786086727	41.5037	-94.32553	10	1	5	9zkuje	40.32
-a317-6efc043fab88	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503696	-94.32508	10	1	5	9zkuje	44.92
-8859-3acbe4808a2a	53ca9cce2fb74c2048a1f436c19e01b786086727	41.503696	-94.32461	10	1	5	9zkuje	47.23

Part 7:

Objective: Assign all points to nearest route_id

Logic:

•Find distance from each point in gps_data to all line strings in the lrs_data (start coordinates and end coordinates)

```
select *, st_distance(
                st_point(a.longitude, a.latitude),
                st_linestring(array[
                        st_point(b.segment_start_longitude, b.segment_start_latitude),
                        st_point(b.segment_end_longitude,b.segment_end_latitude)])) distance from "miguel_gps_data"  a
cross join (select * from "aiml_class"."lrs_miguel_data" where not segment_start_latitude = segment_end_latitude and not segment_start_longitude
    = segment_end_longitude)  b
where a.day =1 and a.hour = 1
```

SQL Ln 1, Col 1

Run again | Explain | Cancel | Clear | Create

Reuse query results
up to 60 minutes ago

Query results | Query stats

Completed | Time in queue: 106 ms | Run time: 8 min 31.474 sec | Data scanned: 29.34 MB

Results (55,732,280)

Copy | Download results

Search rows

1 ...

de	segment_end_latitude	segment_end_longitude	segment_start_measure	segment_end_measure	geohash_lrs	distance
	41.208378	-94.63685	0.075	0.151	9zkdqm	0.4825298007254618
	41.44605	-94.3067	14.897	14.972	9zkgvc	0.07580107906929805
	41.446053	-94.30814	14.822	14.897	9zkgvc	0.07693382190289554
	41.446056	-94.30959	14.747	14.822	9zkgvc	0.07808284814475884
	41.446033	-94.299385	15.277	15.352	9zkgy1	0.07022562430007366
	41.44603	-94.297615	15.369	15.444	9zkgy1	0.06892584199992573
	41.446014	-94.29333	15.592	15.667	9zkgy3	0.06587244609901427
	41.445942	-94.281364	16.212	16.287	9zkgy9	0.058206962952549124
	41.446037	-94.26084	17.302	17.353	9zkgz1	0.04901345739020325
	41.44611	-94.244545	18.154	18.199	9zkgz3	0.04708118899213232
	41.273144	-94.6911	0.463	0.484	9zkdv3	0.49533208444166066
	41.27306	-94.579926	3.711	3.786	9zkdz9	0.3988391644739184
	41.272835	-94.437935	5.206	5.281	9zkffc	0.2913342466444313
	41.272766	-94.415985	6.341	6.424	9zkfg3	0.2775233544373604
	41.27273	-94.38189	8.101	8.198	9zkfu1	0.2582526608336776
	41.272743	-94.361374	9.191	9.266	9zkfu9	0.24816977200672993

•Filter the nearest one for each point through a nested query

Column rn, created to filter, based on requested criteria.

```
WITH distance_calculations AS (SELECT a.*,b.route_id,
        ST_DISTANCE(ST_POINT(a.longitude, a.latitude), ST_LINESTRING(ARRAY[
                    ST_POINT(b.segment_start_longitude, b.segment_start_latitude),
                    ST_POINT(b.segment_end_longitude, b.segment_end_latitude)])) AS distance
    FROM
        "miguel_gps_data" a
    CROSS JOIN
        (SELECT * FROM "aiml_class"."lrs_miguel_data" WHERE NOT segment_start_latitude = segment_end_latitude
           AND NOT segment_start_longitude = segment_end_longitude) b
    WHERE a.day = 1 AND a.hour = 1)

SELECT *
FROM (SELECT *,ROW_NUMBER() OVER (PARTITION BY a.latitude, a.longitude ORDER BY distance ASC) AS rn
    FROM distance_calculations AS a) AS nearest_distances WHERE rn = 1
```

SQL Ln 15, Col 1

Run | Explain | Cancel | Clear | Create

Reuse query results up to 60 minutes ago

Query results | Query stats

Completed — Time in queue: 97 ms — Run time: 2 min 0.615 sec — Data scanned: 29.34 MB

Results (2,401)

Copy | Download results

Search rows

1 ...

	latitude	longitude	month	day	hour	geohash	speed	route_id	distance	rn
102d1f3fa08d	41.493145	-94.24731	10	1	1	9zkup3	123.26	S001910080E	1.6774068446643738E-5	1
102d1f3fa08d	41.49332	-94.24242	10	1	1	9zkup3	123.26	S001910080E	1.1877735251541415E-5	1
d68fc060a1	41.29983	-94.35683	10	1	1	9zkfuy	73.72	C000141430N	7.327794155997675E-4	1
d68fc060a1	41.297325	-94.35683	10	1	1	9zkfuv	96.76	C000141430N	6.241571663807685E-6	1
d68fc060a1	41.29659	-94.356834	10	1	1	9zkfuv	97.91	C000141430N	2.812540281221893E-5	1
d68fc060a1	41.295116	-94.35686	10	1	1	9zkfuv	97.91	C000141430N	6.424120590218035E-6	1
d68fc060a1	41.293583	-94.35684	10	1	1	9zkfuv	103.67	C000141430N	1.1589858474339801E-5	1
d68fc060a1	41.291943	-94.35685	10	1	1	9zkfuu	111.74	C000141430N	6.568541883716239E-6	1
d68fc060a1	41.283905	-94.35684	10	1	1	9zkfug	94.46	C000141430N	1.0789593218788873E-5	1