Urban Development Initiative - Parks & Open Space Analysis

LACK OF PARKS AND PLAY AREAS

Varun Kumar Atkuri

Table of Contents

1. **Executive Summary**:

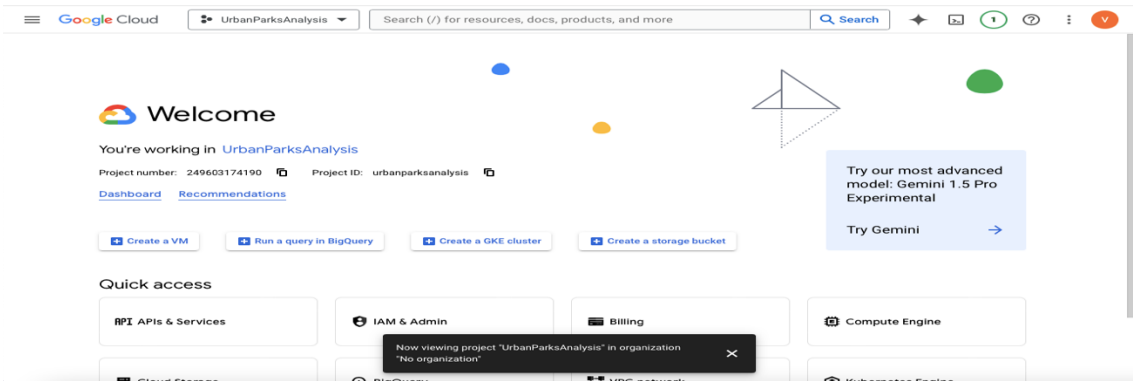
As per the requirements from the stake holders from our organization “Urban Development Initiative”, our data engineer team lead this project to address the question of: 1) park accessibility in urban areas and 2) data analysis. They are components that help achieve a positive impact when it is related to the utilization of urban parks. The beginning of the storages bucket was with the establishment within Google Cloud Platform (GCP) storage for static datasets. With OpenRefine, datasets were cleaned and preprocessed through intensive techniques, guaranteeing the accuracy of data for consequent use. Then, expand the Hadoop ecosystem by including the Cloudera suite of products such as Hive, Spark, and Sentry, which works to enhance data processing capabilities.

The following executions of queries for Big Query, Hive, and Spark have been able to generate observable data insights from the rigorously cleaned data sets. These data therefore help to inform decision-making processes concerning urban park management, and thus stakeholders are enabled with a better understanding of the welcoming issues in order to address them successfully. To carry on the project, deeper steps will be made by applying advanced analytical techniques like machine learning models and spatial analysis so as to draw meaningful conclusions. Algorithms of this kind will help uncover complicated usage patterns in community park spaces yet, the most valuable implications to urban planning policies will follow. By making use of these findings, the project plans to add to the sustainability, user friendliness and reach of urban parks finally, renewing the urban environments’ overall quality of life.

2. **Project Setup in a Cloud Platform:**

We selected GCP among other cloud platforms for our project as Google Cloud Platform (GCP) offers a dependable environment, for our project with features like scalability, managed services, integration capabilities, security measures, compliance adherence and cost efficiency. These aspects contribute to why we choose GCP as the option, for establishing our data infrastructure and performing our analysis.

First, we have created a project with name **UrbanParkAnalysis** in GCP.

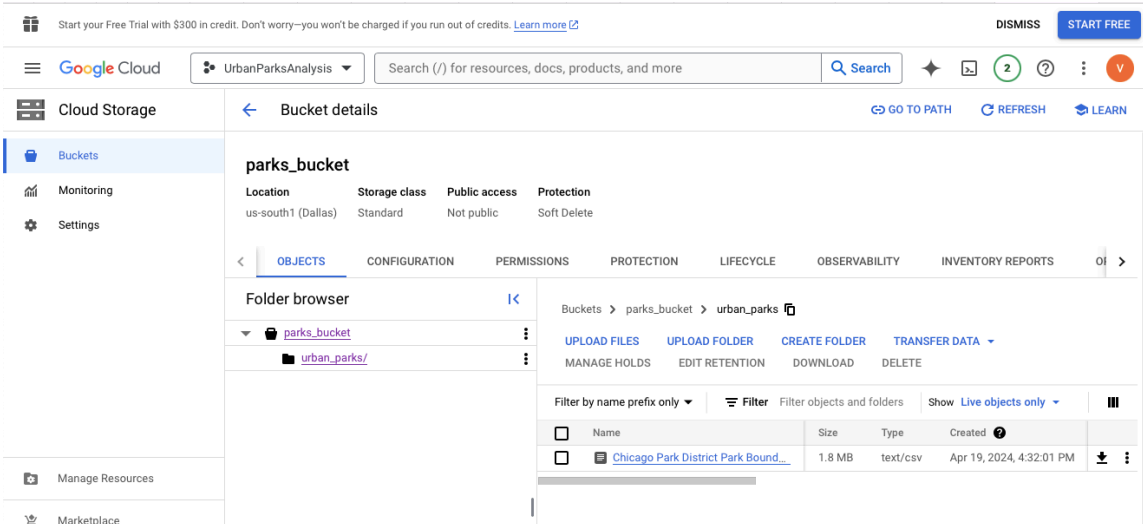


Storage Bucket **parks_bucket** and created a main folder named **urban_parks** are created to store the static datasets in GCP.

We have selected 3 major cities in USA to take few insights from their data regarding the open spaces and parks. The cities are Chicago, Norwalk, New York.

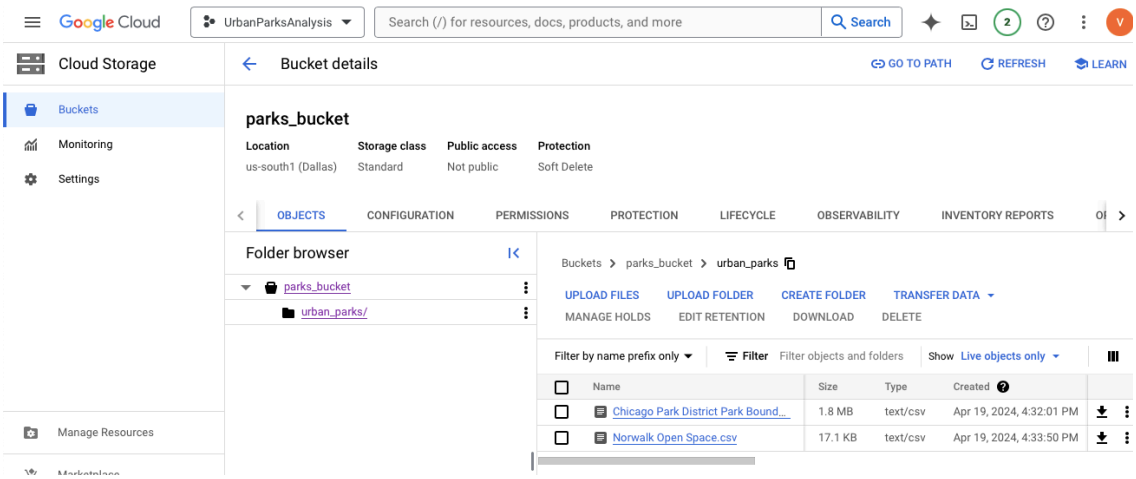
Uploaded the 1st dataset regarding the information about **Chicago** parks and their boundaries into the bucket.

This data collection includes details, about parks located within a city or urban setting. This dataset functions as a list or catalog of the various parks in the urban area. It includes information such as park names, locations, sizes, classifications, and other related attributes. This dataset could be handy for conducting analyses on aspects like park accessibility, usage patterns, distribution and city planning in an environment.

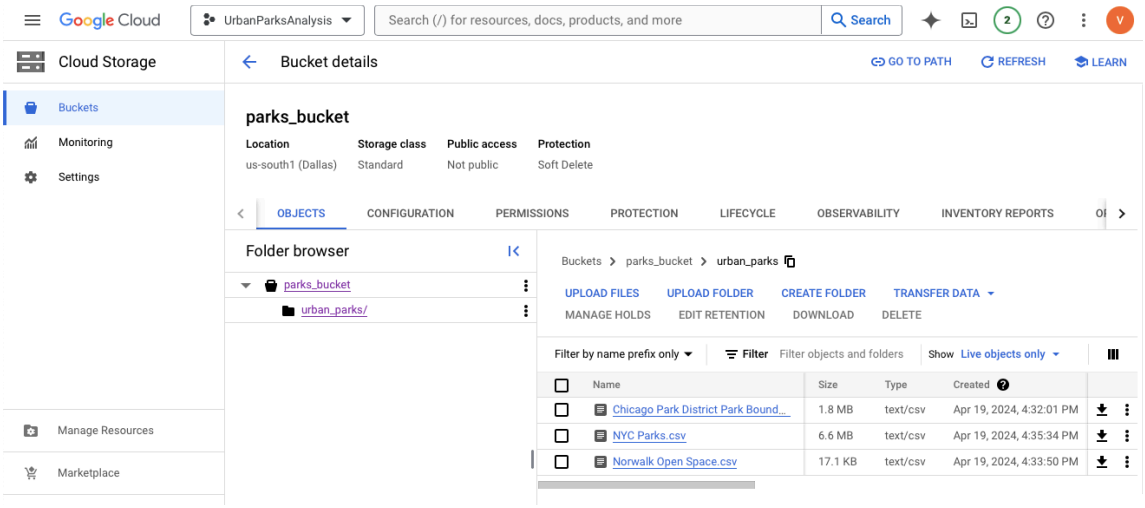


Uploaded the 2nd dataset regarding the information about open spaces in **Norwalk** into the bucket.

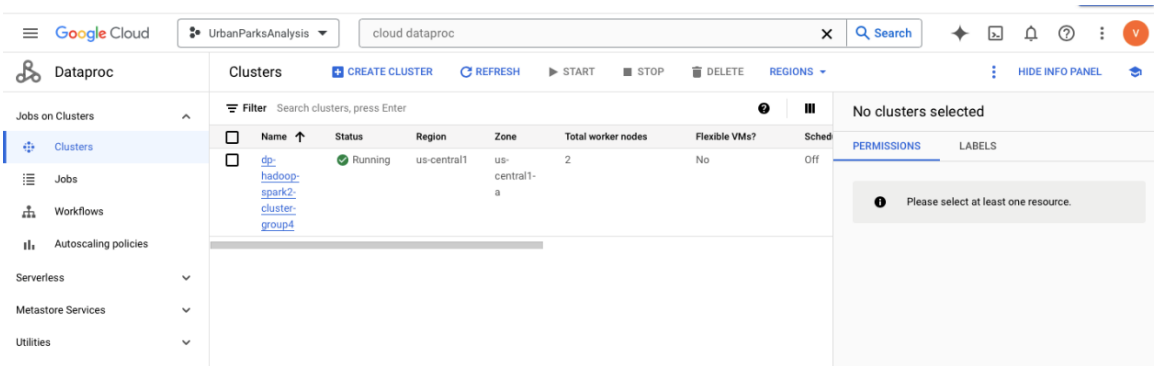
It seems that this dataset offers insights into conservation areas. The dataset covers a variety of spots including conservation areas and golf courses. The ownership status and function codes shed light on restrictions to access and rules regarding use. This dataset could be quite valuable for examining how recreational areas are spread out understanding access, to types of spaces and evaluating how ownership affects access and usage trends.



Uploaded the 3rd dataset regarding the information about parks under NYC Parks into the bucket. The dataset looks like it has some details about parks and recreational places. It has points related to the parks and recreational grounds across the county have been included such as their area size, which jurisdiction they are under, their physical location and which category they fall under.



Setting up Hadoop Ecosystem with Dataproc: In this project, the Dataproc cluster was used for creating spark clusters and managing them, which allows Spark to data processing and analytics done on the cleaned datasets.



3. Data Cleaning and Processing:

For cleaning and processing of data we used OpenRefine tool as it is more cost friendly and reliable.

For all the datasets [Chicago Park District Park Boundaries.csv, Norwalk Open Space.csv, NYC Parks.csv] performed the below steps for cleaning those datasets.

The OpenRefine tool was to clean and preprocess the dataset, removing duplicates and standardizing data formats. Removed few attributes which are irrelevant to the project and made changes to the null values of the zipcode attribute by changing to the relevant zipcode with the help of attributes like ward, jurisdiction, and location attributes. The other string attributes which are null are changed using Facet option by choosing to fill down option. The preprocessed datasets were stored in GCP storage buckets for further analysis.

- **Chicago preprocessed dataset:**

OpenRefine Chicago Park District Park Boundaries Group4 Permalink

617 rows

Show as: **rows** records Show: 5 10 25 50 100 500 **1000** rows

All	PARK_NO	PARK	LOCATION	ZIP	ACRES	WARD	PARK_CLASS	LABEL	Perimeter	Shape_Leng	Shape_Area
1.	2	MCGUANE (JOHN)	2901 S POPLAR AVE	60608	10.3	11	COMMUNITY PARK	McGuane	427.153241288	2,672.32662666	430,341.671424
2.	3	ARMOUR (PHILIP) SQUARE	3309 S SHIELDS AVE	60616	9.05	11	COMMUNITY PARK	Armour Sq	427.153241288	2,484.28300556	391,095.834054
3.	4	FULLER (MELVILLE)	331 W 45TH ST	60609	11.31	3	COMMUNITY PARK	Fuller	427.153241288	2,878.42815048	497,076.662367
4.	5	CORNELL (PAUL) SQUARE	1809 W 50TH ST	60609	8.8	20	COMMUNITY PARK	Cornell Sq	427.153241288	2,462.60759846	385,672.402254
5.	6	RUSSELL (MARTIN) SQUARE	3045 E 83RD ST	60617	10.05	10	COMMUNITY PARK	Russell Sq	427.153241288	2,777.37593622	435,169.113288
6.	7	SHERMAN (JOHN)	1307 W 52ND ST	60609	57.69	20	REGIONAL PARK	Sherman	427.153241288	6,473.49754395	2,513,062.44054
7.	14	DAVIS (DR. NATHAN) SQUARE	4430 S MARSHFIELD AVE	60609	8.91	15	COMMUNITY PARK	Davis (Nathan) Sq	427.153241288	2,465.13667346	386,513.579487
8.	8	OGDEN (WILLIAM)	6500 S RACINE AVE	60636	57.04	17	REGIONAL PARK	Ogden	427.153241288	6,465.96741393	2,518,604.35237
9.	9	HAMILTON (ALEXANDER)	513 W 72ND ST	60621	28.94	6	REGIONAL PARK	Hamilton	427.153241288	4,589.61757141	1,258,959.81997
10.	11	CALUMET	9801 S AVENUE G	60617	181.29	10	CITYWIDE PARK	Calumet	427.153241288	18,437.7291297	7,913,418.27835
11.	21	WASHINGTON (GEORGE)	5531 S DR MARTIN LUTHER KING J	60637	350.49	0	CITYWIDE PARK	Washington (George)	427.153241288	22,537.5345467	15,056,212.332
12.	355	NINEBARK	1447-1453 S HARDING AVE	60623	0.25	24	MINI-PARK	Ninebark	510.790663166	510.790663166	16,072.8552841
13.	10	MARQUETTE (JACQUES)	6734 S KEDZIE AVE	60629	315.63	17	CITYWIDE PARK	Marquette	427.153241288	16,520.0413174	13,731,499.4813
14.	12	BESSEMER (HENRY)	8930 S MUSKEGON AVE	60617	20.27	7	REGIONAL PARK	Bessemer	427.153241288	4,259.42952705	882,731.232161
15.	13	PALMER (POTTER)	200 E 111TH ST	60628	38.44	9	REGIONAL PARK	Palmer	427.153241288	5,156.79114277	1,674,187.25971
16.	146	CHOPIN (FREDERIC)	3420 N LONG AVE	60641	9.28	30	COMMUNITY PARK	Chopin	427.153241288	2,526.0660708	404,291.047908
17.	15	GRAND CROSSING	7655 S INGLESIDE AVE	60619	18.89	8	REGIONAL PARK	Grand Crossing	427.153241288	3,976.46252949	817,500.994948
18.	16	TRUMBULL (LYMAN)	2400 E 105TH ST	60617	18.04	7	REGIONAL PARK	Trumbull	427.153241288	3,728.26082426	785,838.659695
19.	17	MANN (JAMES)	3035 E 130TH ST	60633	18.86	10	REGIONAL PARK	Mann	427.153241288	3,840.19996666	821,525.036631
20.	18	TULEY (MURRAY)	501 E 90TH PL	60619	18.7	9	REGIONAL PARK	Tuley	427.153241288	3,820.06288595	807,732.488843
21.	19	JACKSON (ANDREW)	6401 S STONY ISLAND AVE	60637	551.52	5	MAGNET PARK	Jackson (Andrew)	427.153241288	49,754.537972	23,110,298.3329
22.	22	GAGE (GEORGE)	2415 W 55TH ST	60629	26.5	0	REGIONAL PARK	Gage	427.153241288	8,414.16764662	1,095,017.76667
23.	23	MCKINLEY (WILLIAM)	2210 W PERSHING RD	60609	72.09	12	REGIONAL PARK	McKinley	427.153241288	7,603.11303044	3,125,359.11791
24.	24	GRANT (ULYSSES)	331 E RANDOLPH ST	60605	295.45	0	MAGNET PARK	Grant	427.153241288	55,248.4852471	13,652,838.3136
25.	26	FOSTER (J. FRANK)	1400 W 84TH ST	60620	25.2	21	REGIONAL PARK	Foster	427.153241288	4,258.90148017	1,095,469.76523

- **Norwalk Preprocessed dataset:**

OpenRefine Norwalk Open Space Group4 Permalink

165 rows

Show as: **rows** records Show: 5 10 25 50 100 500 **1000** rows

All	OBJECTID	CATEGORY	NAME	OWNERSHIP	FNCT	DESC_	ACRE_GIS
1.	1	Conservation	Foxboro Conservation Development	Private	P-1	Private, open to public without fee	0.5379
2.	2	General Recreation	Silvermine Golf Course (2 Parcels)	Private	P-3	Private, members or owners only	39.7912
3.	3	Existing Preserved Open Space	Hilltop Conservation Development	Private	P-1	Private, open to public without fee	7.1227
4.	4	General Recreation	Silvermine Golf Course (2 Parcels)	Private	P-3	Private, members or owners only	37.0502
5.	5	General Recreation	Silvermine Golf Course (2 Parcels)	Private	P-3	Private, members or owners only	27.5993
6.	6	Cemetery	Chestnut Hill Road Cemetery	Private	P-C	Private cemetary	0.0704
7.	7	Existing Preserved Open Space	Comstock Hill Open Space	Municipal	M-1	Municipal, open to public without fee	3.5359
8.	8	School	All Saints Catholic School	Private	PS	Private school	28.2371
9.	9	Uncategorized	The Falls Conservation Development	Private	P-3	Private, members or owners only	8.6526
10.	10	Conservation	Riverwalk Easement	Private	P-1	Private, open to public without fee	0.2543
11.	11	Uncategorized	Orchard Lakes Conservation Development	Private	P-3	Private, members or owners only	8.9022
12.	12	Existing Preserved Open Space	Burlington Court Open Space	Municipal	M-1	Municipal, open to public without fee	0.3828
13.	13	Uncategorized	Blue Mountain Conservation Development	Private	P-3	Private, members or owners only	1.5407
14.	14	Cemetery	Broad Street Cemetery	Private	P-C	Private cemetary	12.8178
15.	15	Cemetery	Nursery Cemetery	Private	P-C	Private cemetary	0.1948
16.	16	General Recreation	Quartette Club	Private	P-2	Private, open to public with fee	3.299
17.	17	Cemetery	Union Cemetery	Private	P-C	Private cemetary	16.5599
18.	18	Cemetery	Riverside Cemetery	Private	P-C	Private cemetary	42.0392
19.	19	Existing Preserved Open Space	Betts Brook Park	Municipal	M-1	Municipal, open to public without fee	0.7519
20.	20	Uncategorized	Cannon Brook Conservation Development	Private	P-3	Private, members or owners only	1.6514
21.	21	Existing Preserved Open Space	Fox Run Open Space	Municipal	M-1	Municipal, open to public without fee	3.4114
22.	22	Cemetery	Cemetery (Boston Post Road)	Private	P-C	Private cemetary	1.7385
23.	23	Existing Preserved Open Space	Riverside Park	Municipal	M-1	Municipal, open to public without fee	0.3968
24.	24	Existing Preserved Open Space	Klondike Park	Municipal	M-1	Municipal, open to public without fee	0.0424
25.	25	Cemetery	Weed Avenue Cemetery	Private	P-C	Private cemetary	0.7111
26.	26	Uncategorized	Norwalk Boat Club	Private	P-2	Private, open to public with fee	0.0803

- **New York preprocessed dataset:**

OpenRefine NYC Parks Group4 Permalink

Open... Export Help

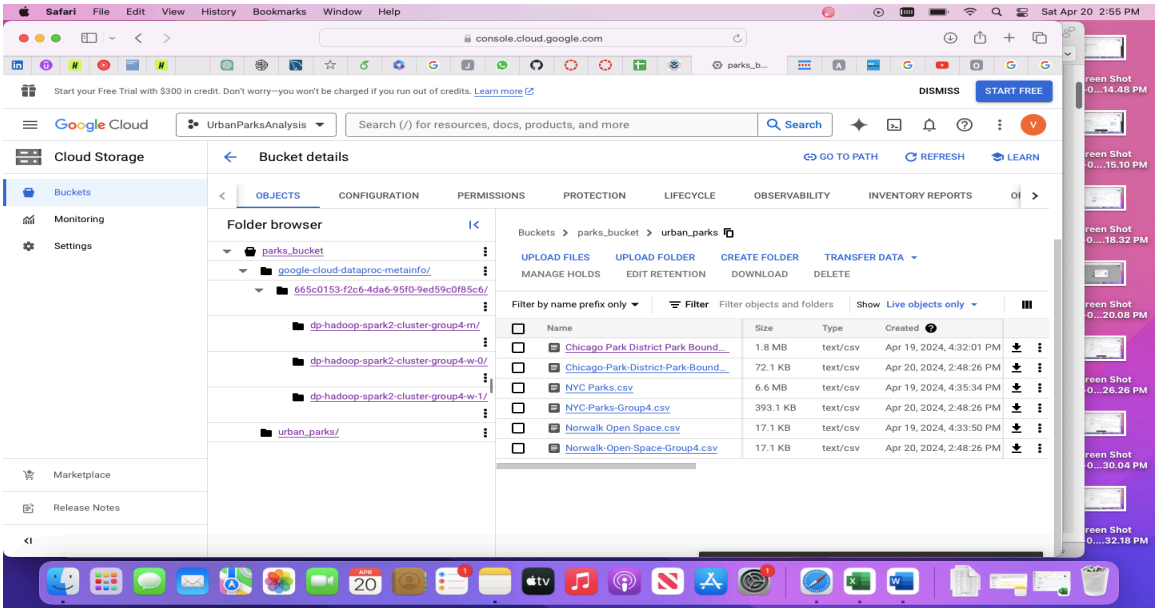
2047 rows

Extensions Wikibase

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first ‹ previous 1 next › last »

	All	ACRES	CLASS	JURISDICTION	LOCATION	MAPPED	NAME311	OBJECTID	OMPPROPID	PARENTID	PERMIT	PERMITDISTRICT	PERMITPARENT	PIP_RATABL
1.		249.389	PARK	DPR/CDOT/SDOT	Astoria Blvd. and 48 St. to Union Tp., Park Drive East.	TRUE	Grand Central Parkway Extension	15,508	Q084A	Q-03	TRUE	Q-03	Q-03	FALSE
2.		9.375	PARK	DPR	W. 19 Rd. bet Jamaica Bay and Cross Bay Blvd.	FALSE	Sunset Cove Park	6,344	Q498	Q-14	FALSE	Q-14	Q-14	FALSE
3.		2.035	PARK	DPR/DOE	251 St. bet. 61 Ave. and 63 Ave.	FALSE	Challenge Playground	6,293	Q346	Q-11	TRUE	Q-11	Q-11	TRUE
4.		3.02	PARK	DPR	Jerome Park Resevoir and Sedgwick Av bet. W	FALSE	Fort Independence Playground	5,732	X020	X-08	TRUE	X-08	X-08	TRUE
5.		63.636	PARK	DPR	Eastern Pkwy. bet. Grand Army Plaza and Ralph Ave.	TRUE	Eastern Parkway	69,228	B029	B-08	TRUE	B-08	B-08	FALSE
6.		326.895	PARK	DPR	Whitestone Exwy. at 13 Ave. to the Linden Blvd and the Belt Pkwy.	TRUE	Cross Island Parkway	69,239	Q135	Q-13	FALSE	Q-13	Q-13	FALSE
7.		10.79	PARK	DPR	169 St., Merrick Blvd., Marne Pl. bet. Linden Blvd., Sayres Ave., and 111 Rd.	TRUE	Archie Spigner Park	6,320	Q051	Q-12	TRUE	Q-12	Q-12	TRUE
8.		3.301	PARK	DPR	Gleason Ave., Watson Ave. bet. Noble Ave., and Rosedale	TRUE	Watson Gleason Playground	6,307	X124	X-09	TRUE	X-09	X-09	TRUE

Uploaded the cleaned datasets into the storage buckets:

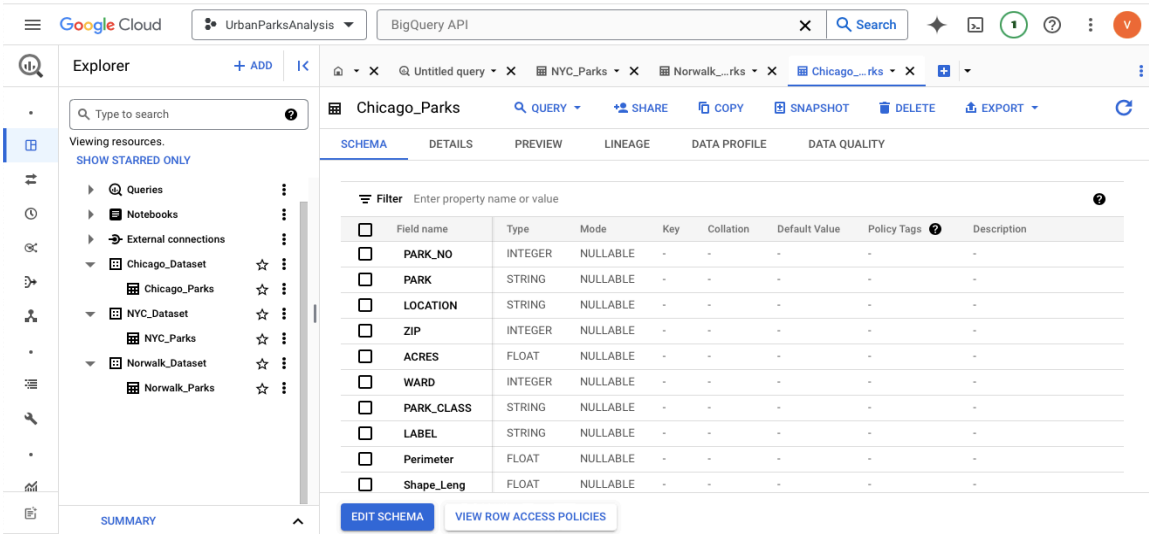


4. Insights on data using Big Query, Hive & Spark:

These primary insights from the datasets will help the data scientists and analyst team for a better understanding about the data.

Big Query: Loaded all the datasets from the storage buckets into the Big Query Studio.

Hive & Spark: Loaded all the datasets into the Hadoop ecosystem from the storage buckets.



Dataset 1: Chicago-Park-District-Park-Boundaries-Group4.csv (Cleaned dataset in the storage bucket) is copied to a table in BigQuery Studio as Chicago_Parks. Ran few queries to verify that we can show meaningful insights form these datasets or not to forward these datasets to the data scientists and analysts' team.

BigQuery: To get to know about the number of parks available in each ward of the Chicago. We ran the below query.

Untitled query RUN SAVE DOWNLOAD

```
1  SELECT WARD, COUNT(*) AS park_count
2  FROM `urbanparksanalysis.Chicago_Dataset.Chicago_Parks`
3  GROUP BY WARD;
4
```

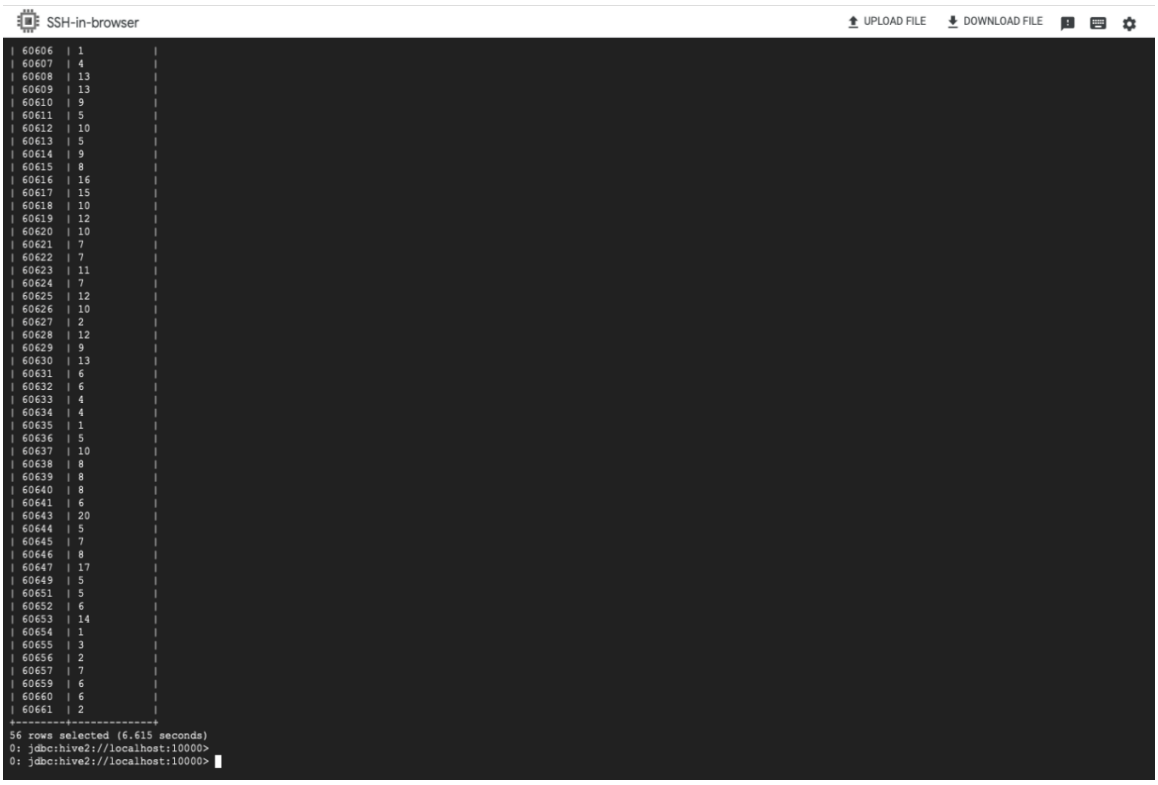
Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXEC

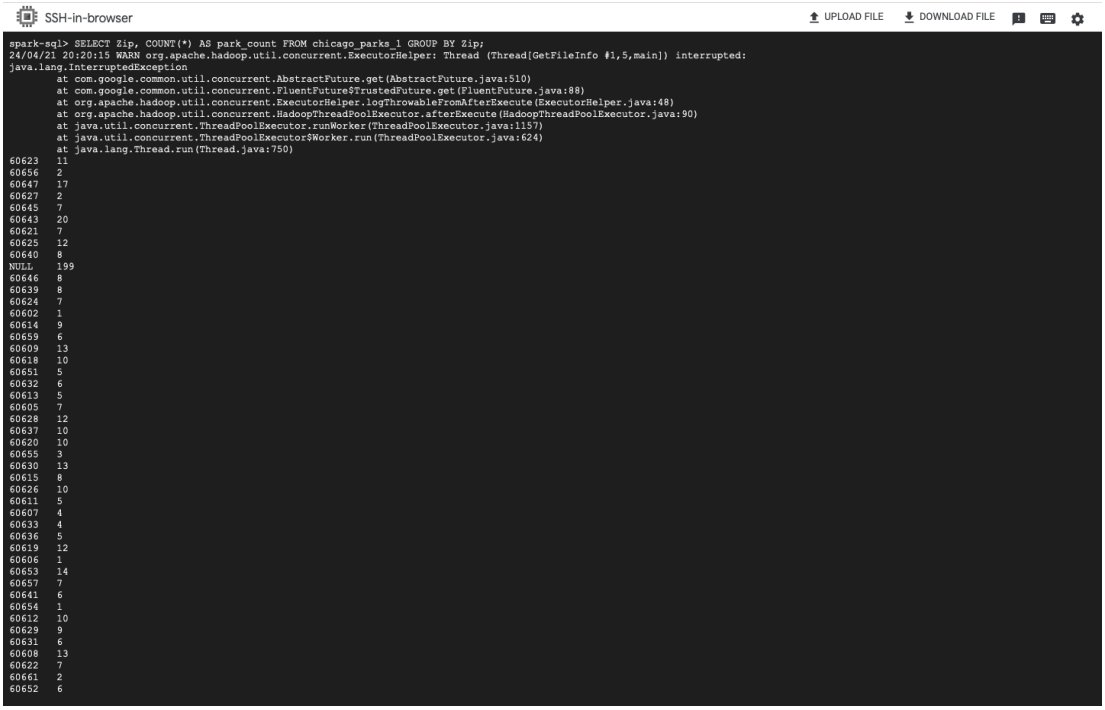
Row	WARD	park_count
1	42	9
2	0	28
3	1	14
4	2	9
5	3	27
6	4	24
7	5	16
8	8	14
9	9	14
10	10	20

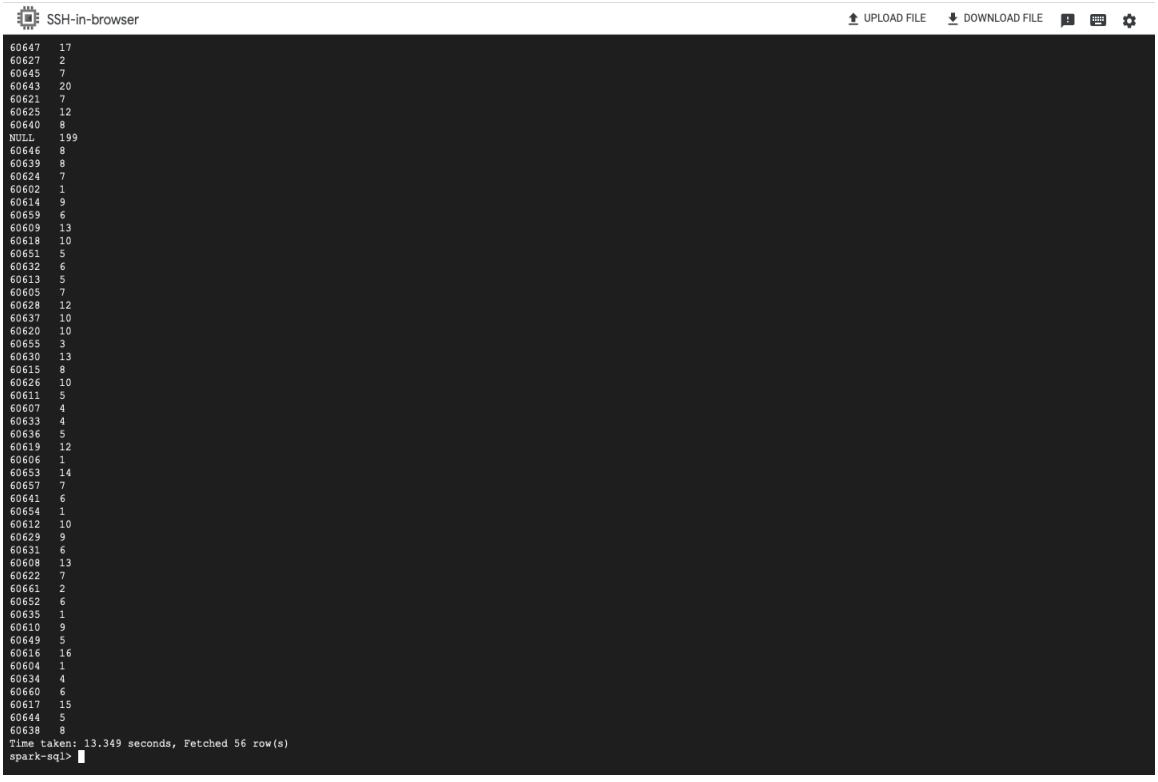
Hive & Spark: What is count of reachable parks in every zip area in Chicago. To know this, we ran the below query in hive.

```
0: jdbc:hive2://localhost:10000> SELECT Zip, COUNT(*) AS park_count F
ROM chicago_parks_1 GROUP BY Zip;
INFO  : Compiling command(queryId=hive_20240421192816_d1d1bf79-be62-4
```



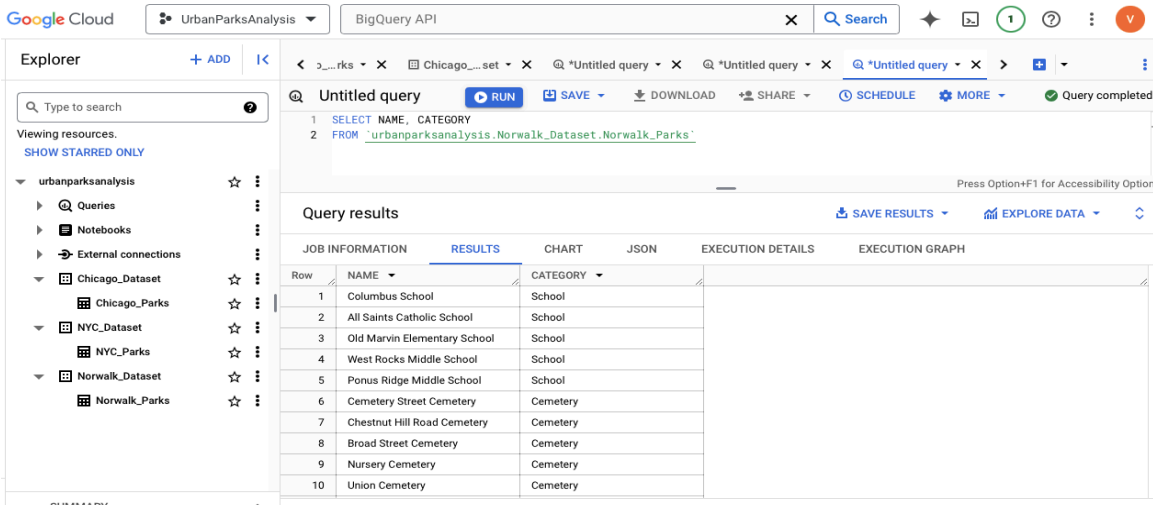
Ran the same query in the spark to validate the query run time in hive and spark. The query took 6.615 seconds in hive while it took double the time of hive which is 13.349 seconds in spark.



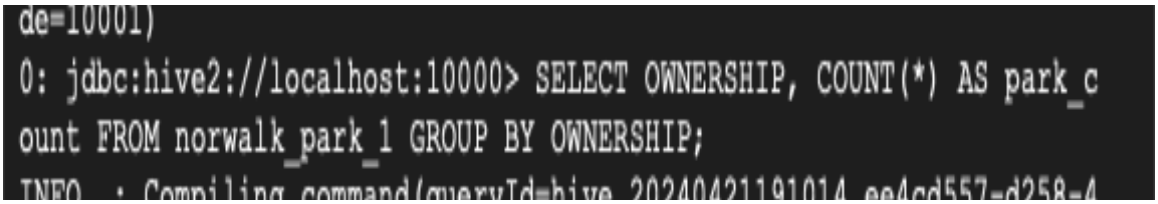


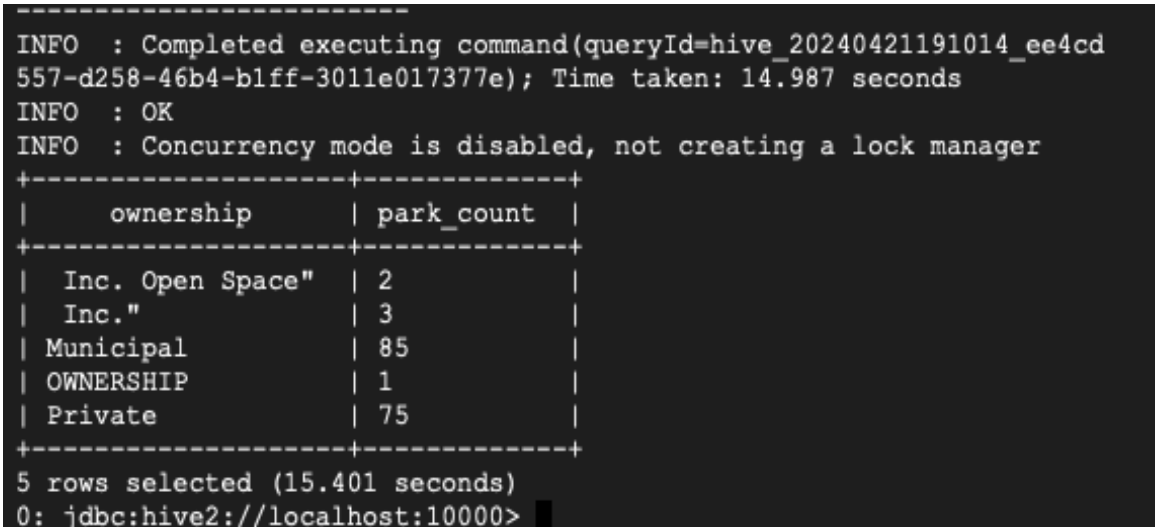
Dataset 2: Norwalk-Open-Space-Group4.csv (Cleaned dataset in the storage bucket) is copied to a table in BigQuery Studio as Norwalk_Parks. Ran few queries to verify that we can display meaningful insights form these datasets or not to forward these datasets to the data scientists and analysts' team.

BigQuery: To get to know about the name and categories of parks available in Norwalk. We ran the below query.



Hive & Spark: To get to know how many parks are ownership under one entity. Ran below query to know it.



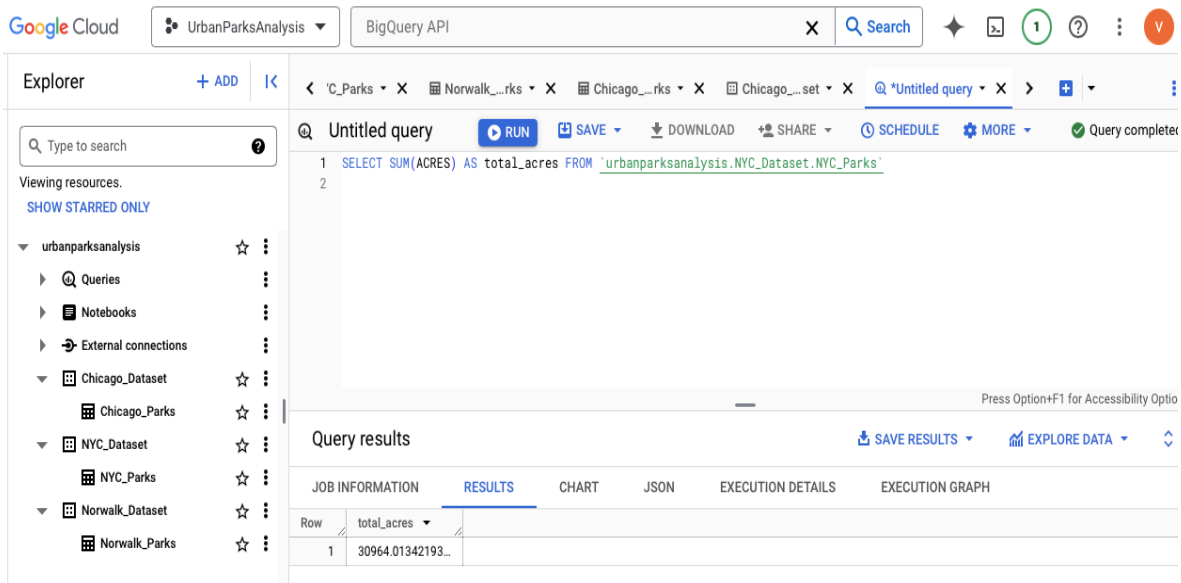


Ran the same query in the spark to validate the query run time in hive and spark. The query took 15.401 seconds in hive while it took only 1.045 seconds in spark.

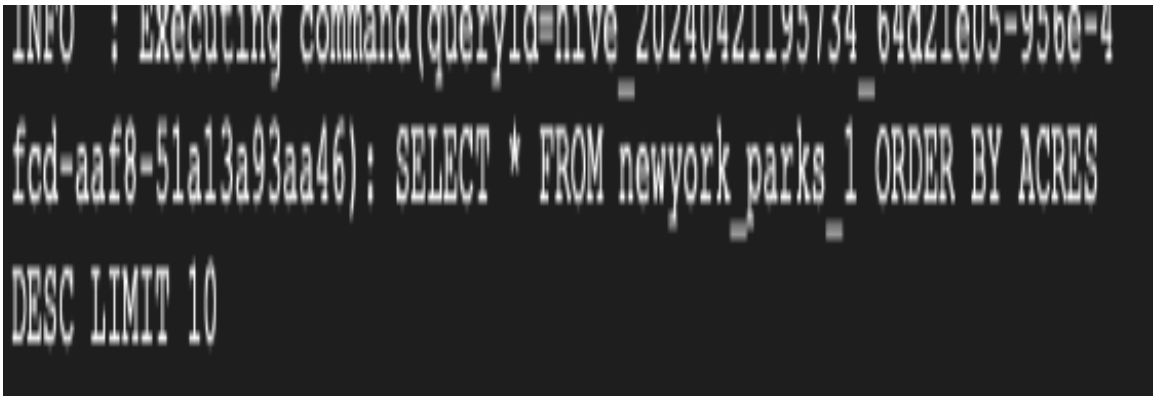
```
spark-sql> SELECT OWNERSHIP, COUNT(*) AS park_count FROM norwalk_park_8 GROUP BY OWNERSHIP;
Error in query: Table or view not found: norwalk_park_8; line 1 pos 46;
'Aggregate ['OWNERSHIP], ['OWNERSHIP, count(1) AS park_count#78L]
+- 'UnresolvedRelation [norwalk_park_8], [], false

spark-sql> SELECT OWNERSHIP, COUNT(*) AS park_count FROM norwalk_park_1 GROUP BY OWNERSHIP;
OWNERSHIP       1
Private 75
 Inc. Open Space"       2
Municipal       85
 Inc."  3
Time taken: 1.045 seconds, Fetched 5 row(s)
spark-sql>
```

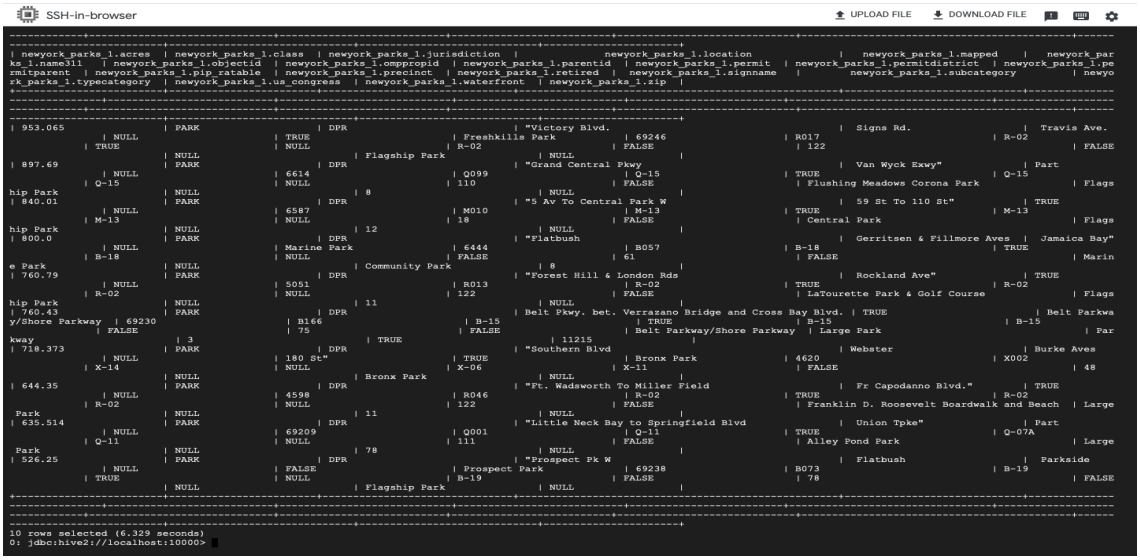
Dataset 3: NYC-Parks-Group4.csv (Cleaned dataset in the storage bucket) is copied to a table in BigQuery Studio as NYC_Parks. Ran few queries to verify that we can show meaningful insights form these datasets or not to forward these datasets to the data scientists and analysts' team.

BigQuery: To get to know total area of parks in the New York. We ran the below query.

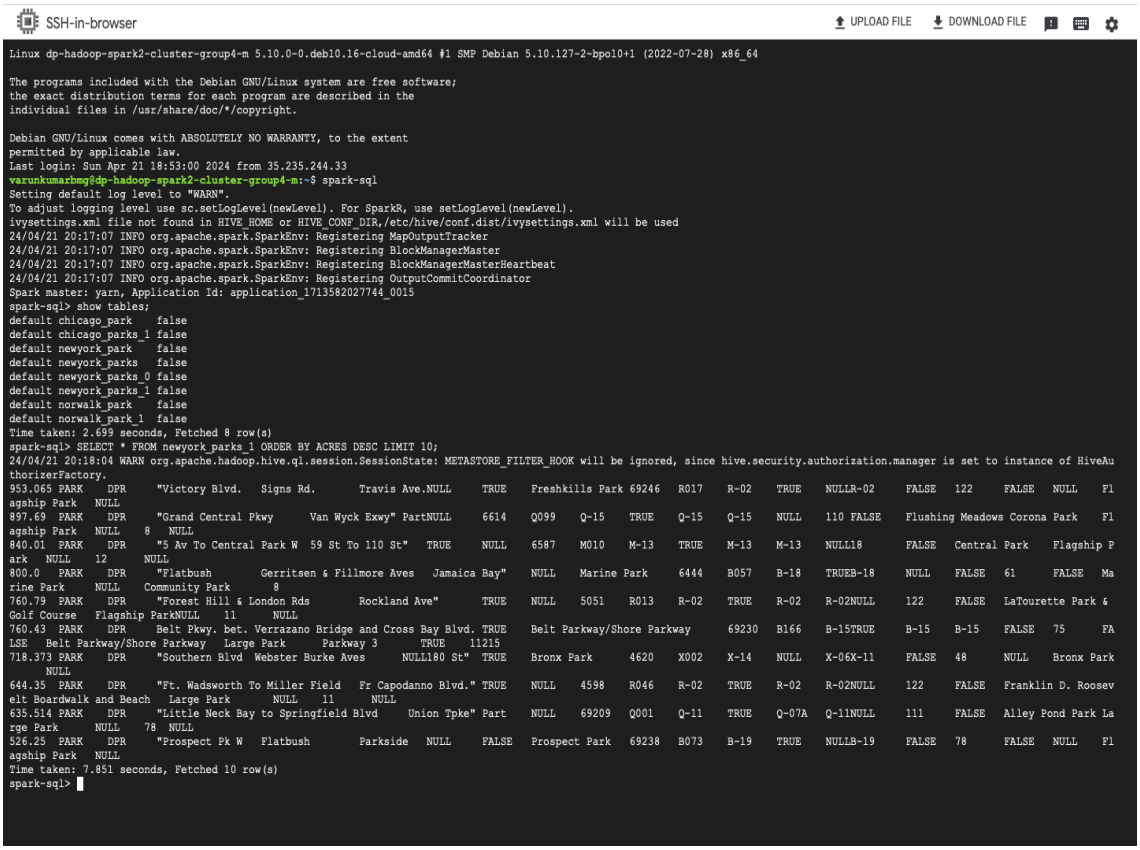


Hive & Spark: To know the 10 largest parks in New York. We ran the below query for that.





Ran the same query in the spark to validate the query run time in hive and spark. The query took 6.329 seconds in hive while it took more time than hive which is 7.851 seconds in spark.


SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Linux dp-hadoop-spark2-cluster-group4-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Apr 21 18:53:00 2024 from 35.235.244.33
varunkumarbmg@dp-hadoop-spark2-cluster-group4-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
24/04/21 20:17:07 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/04/21 20:17:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/04/21 20:17:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/21 20:17:07 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1713582027744_0015
spark-sql> show tables;
default chicago_park false
default chicago_parks_1 false
default newyork_park false
default newyork_parks false
default newyork_parks_0 false
default newyork_parks_1 false
default norwalk_park false
default norwalk_park_1 false
Time taken: 2.699 seconds, Fetched 8 row(s)
spark-sql> SELECT * FROM newyork_parks_1 ORDER BY ACRES DESC LIMIT 10;
24/04/21 20:18:04 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAu
thorizerFactory.
953.065 PARK DPR "Victory Blvd. Signs Rd. Travis Ave.NULL TRUE Freshkills Park 69246 R017 R-02 TRUE NULLR-02 FALSE 122 FALSE NULL Fl
agship Park NULL
897.69 PARK DPR "Grand Central Pkwy Van Wyck Exwy" PartNULL 6614 Q099 Q-15 TRUE Q-15 Q-15 NULL 110 FALSE Flushing Meadows Corona Park Fl
agship Park NULL 8 NULL
840.01 PARK DPR "5 Av To Central Park W 59 St To 110 St" TRUE NULL 6587 M010 M-13 TRUE M-13 M-13 NULL18 FALSE Central Park Flagship P
ark NULL 12 NULL
800.0 PARK DPR "Flatbush Gerritsen & Fillmore Aves Jamaica Bay" NULL Marine Park 6444 B057 B-18 TRUEB-18 NULL FALSE 61 FALSE Ma
rine Park NULL Community Park 8
760.79 PARK DPR "Forest Hill & London Rds Rockland Ave" TRUE NULL 5051 R013 R-02 TRUE R-02 R-02NULL 122 FALSE LaTourette Park &
Golf Course Flagship ParkNULL 11 NULL
760.43 PARK DPR Belt Pkwy. bet. Verrazano Bridge and Cross Bay Blvd. TRUE Belt Parkway/Shore Parkway 69230 B166 B-15TRUE B-15 B-15 FALSE 75 FA
LSE Belt Parkway/Shore Parkway Large Park Parkway 3 TRUE 11215
718.373 PARK DPR "Southern Blvd Webster Burke Aves NULL180 St" TRUE Bronx Park 4620 X002 X-14 NULL X-06X-11 FALSE 48 NULL Bronx Park
NULL
644.35 PARK DPR "Ft. Wadsworth To Miller Field Fr Capodanno Blvd." TRUE NULL 4598 R046 R-02 TRUE R-02 R-02NULL 122 FALSE Franklin D. Roosev
elt Boardwalk and Beach Large Park NULL 11 NULL
635.514 PARK DPR "Little Neck Bay to Springfield Blvd Union Tpke" Part NULL 69209 Q001 Q-11 TRUE Q-07A Q-11NULL 111 FALSE Alley Pond Park La
rge Park NULL 78 NULL
526.25 PARK DPR "Prospect Pk W Flatbush Parkside NULL FALSE Prospect Park 69238 B073 B-19 TRUE NULLB-19 FALSE 78 FALSE NULL Fl
agship Park NULL
Time taken: 7.851 seconds, Fetched 10 row(s)
spark-sql>

5. Conclusion:

The documentation will give a one examination and show the key elements of data cycle for urban park analysis, such as setting of the project, data cleaning and processing, and performing queries in BigQuery, Hive, and Spark. With the help of this strategy, information came to light relative to the park characteristics, ownership types, and relation between acreage. The next phase will be to apply these findings to informed decision-making for the management of urban parks and additionally touch on data science with data scientists team aiming to develop advanced analysis techniques for deeper understandings.

6. References:

[1]. https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Park-Boundaries-curren/ej32-qgdr

[2].https://hub.arcgis.com/datasets/eab879e060034c989421662836de8f74_0/explore?location=41.110978%2C-73.428163%2C12.00&showTable=true

[3].https://data.cityofnewyork.us/Recreation/Parks-Properties/enfh-gkve/about_data

[4]. (Open Al's ChatGPT4, 2024, Generate a logo for urban development initiative non profitable organization.)