NTD : Public Transit Improvement

The Why Behind Our Every Step

It is more about “Navigating Routes and Connecting Spaces” which evolves from a convenience to a necessity.

- **Accessibility:** (Ensuring convenient access to public transit vehicles)
- **Scarcity:** (Managing the availability of transit’s)
- **Travel apprehension:**(comfort and safety of passengers during transit with well-designed vehicles)
- Addressing the Environmental footprint of public transit vehicles by adopting eco-friendly technologies.

Outline

1. Data Cleaning with Open Refine:

- Use Open Refine Tool to tidy up the data for better quality.

2. Storing Data on GCP Cloud:

- Save the cleaned data in Google Cloud Platform (GCP) for secure storage.

3. Setting up Dataproc Hadoop Cluster:

- Establish a Dataproc Hadoop cluster for batch processing, querying, and streaming data.

4. Creating Sample Queries:

- Develop sample queries to manage and maintain the data effectively.

5. Ensuring Data Quality:

- Prioritize maintaining high data quality for analysts and engineers to analyze, visualize, and interpret the data accurately.

Data Lifecycle

We've gathered and refined data from two government sources using Open Refine. This cleaned data is now securely stored in Google Cloud. To manage and analyze it efficiently, we rely on tools like BigQuery, Hive, and Spark. This streamlined process ensures our data is not only well-organized but also ready for insightful analysis.

Dataset Generation:

Congress established the National Transit Database (NTD) for information and statistics on the transit systems of the United States. The data is collected from beneficiaries of grants from the Federal Transit Administration (FTA).

Dataset Collection

- The 2022 Annual dataset containing data on vehicles operated by each transit agency.
- It gives:
- Age Distribution and the lifetime mileage of each agency's revenue fleets.
- The age distribution of service fleets, by vehicle type.
- The dataset is collected from https://www.transit.dot.gov/ntd/data-product/2022-vehicles
- It contains 2 datasets i.e, Vehicle Age Distribution &Vehile Type count by Agency

Meta Data

- Meta Data of Dataset-Vehicle Age Distribution

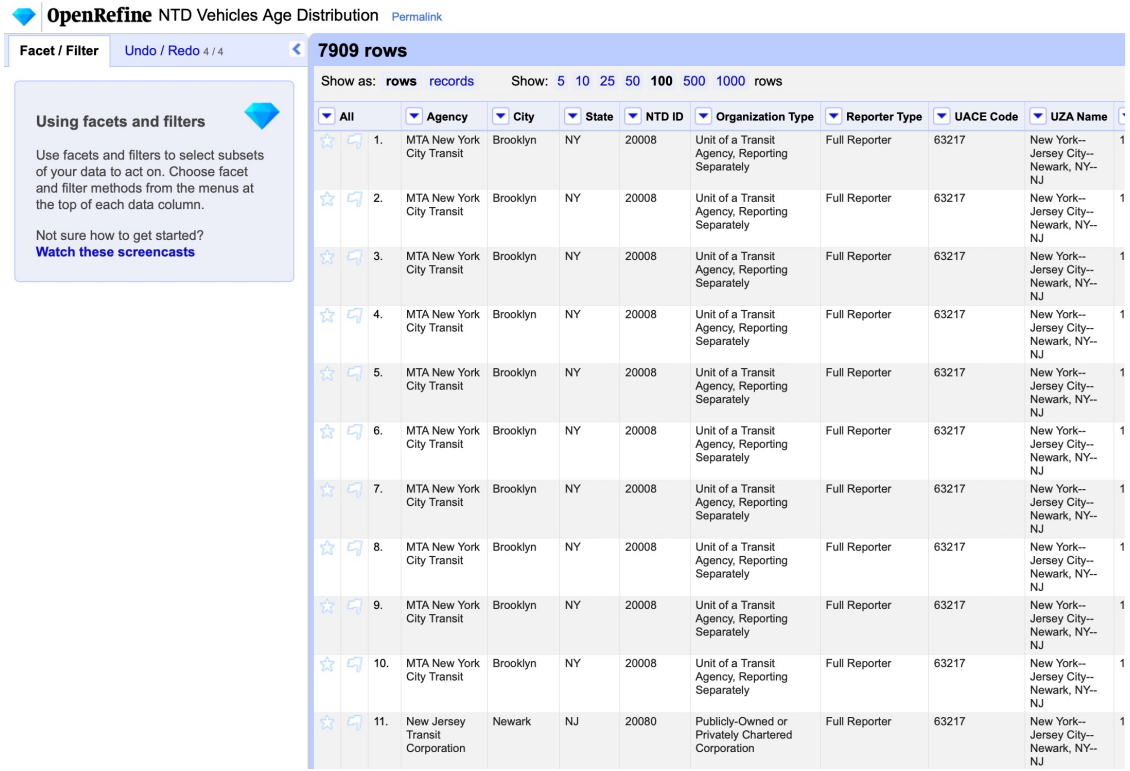
Column Name	Description	Type
Agency	The transit agency's name.	Plain Text
City	The city in which the agency is headquartered.	Plain Text
State	The state in which the agency is headquartered.	Plain Text
NTD ID	A five-digit identifying number for each agency used in the c…	Plain Text
Organization Type	Description of the agency's legal entity.	Plain Text
Reporter Type	The type of NTD report that the agency completed this year.	Plain Text
UACE Code	UACE Code remains consistent across census years.	Plain Text
UZA Name	The name of the agency's Urbanized Area.	Plain Text
Primary UZA Population	The population of the urbanized area primarily served by the…	Number
Agency VOMS	The number of revenue vehicles operated across the whole …	Number
Vehicle Type	The form of passenger conveyance used for revenue operati…	Plain Text

- Meta Data of Dataset-Vehicle Type count by Agency

Column Name	Description	Type
Agency	The transit agency's name.	Plain Text
City	The city in which the agency is headquartered.	Plain Text
State	The state in which the agency is headquartered.	Plain Text
NTD ID	A five-digit identifying number for each agency used in the c…	Plain Text
Organization Type	Description of the agency's legal entity.	Plain Text
Reporter Type	The type of NTD report that the agency completed this year.	Plain Text
UACE Code	UACE Code remains consistent across census years.	Plain Text
UZA Name	The name of the agency's Urbanized Area.	Plain Text
Primary UZA Population	The population of the urbanized area primarily served by the…	Number
Agency VOMS	The number of revenue vehicles operated across the whole …	Number
Bus	Rubber-tired passenger vehicles powered by diesel, gasoline…	Number

Data Processing

- We did Data Processing using Open Refine Tool for both the datasets.
- Null valued rows, are filled down with relevant cells, used Text & Numeric Facet
- City values are clustered based on Key Collision



OpenRefine NTD Vehicles Age Distribution Permalink

Facet / Filter | Undo / Redo 4 / 4

7909 rows

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

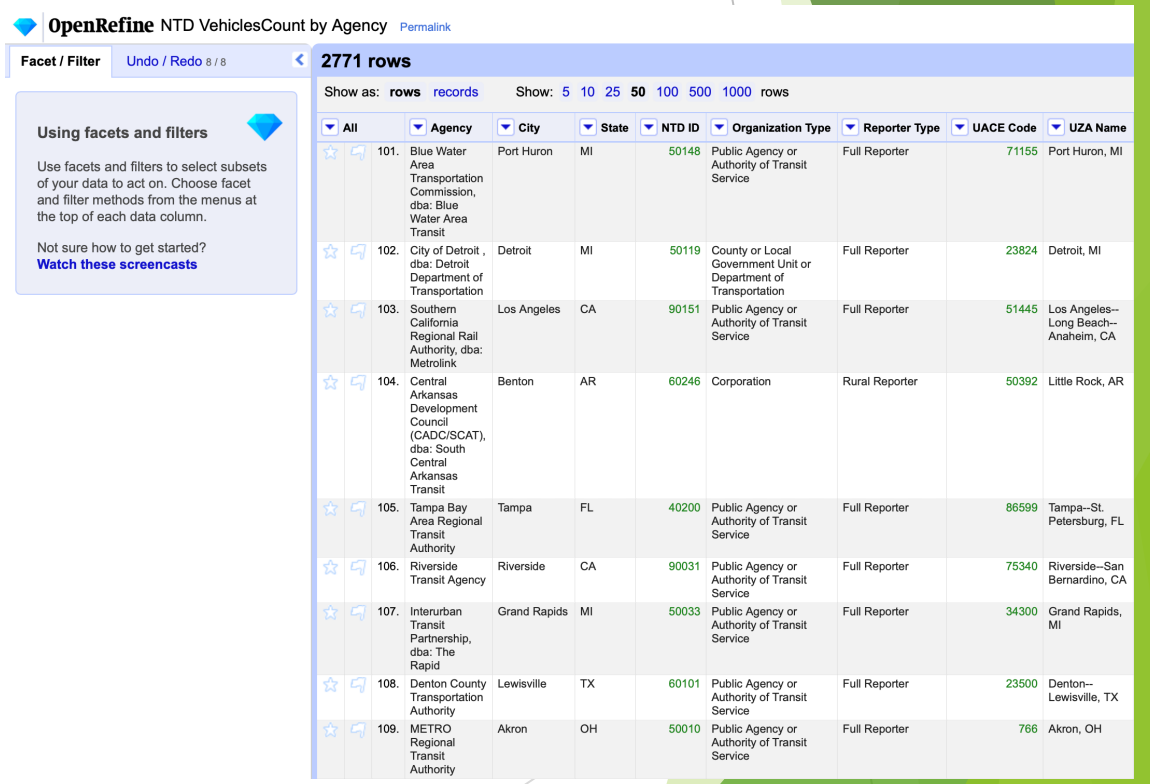
Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	Agency	City	State	NTD ID	Organization Type	Reporter Type	UACE Code	UZA Name
1.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
2.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
3.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
4.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
5.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
6.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
7.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
8.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
9.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
10.	MTA New York City Transit	Brooklyn	NY	20008	Unit of a Transit Agency, Reporting Separately	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ
11.	New Jersey Transit Corporation	Newark	NJ	20080	Publicly-Owned or Privately Chartered Corporation	Full Reporter	63217	New York--Jersey City--Newark, NY--NJ

Refined Dataset : NTD Vehicles Age Distribution



OpenRefine NTD VehiclesCount by Agency Permalink

Facet / Filter | Undo / Redo 8 / 8

2771 rows

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

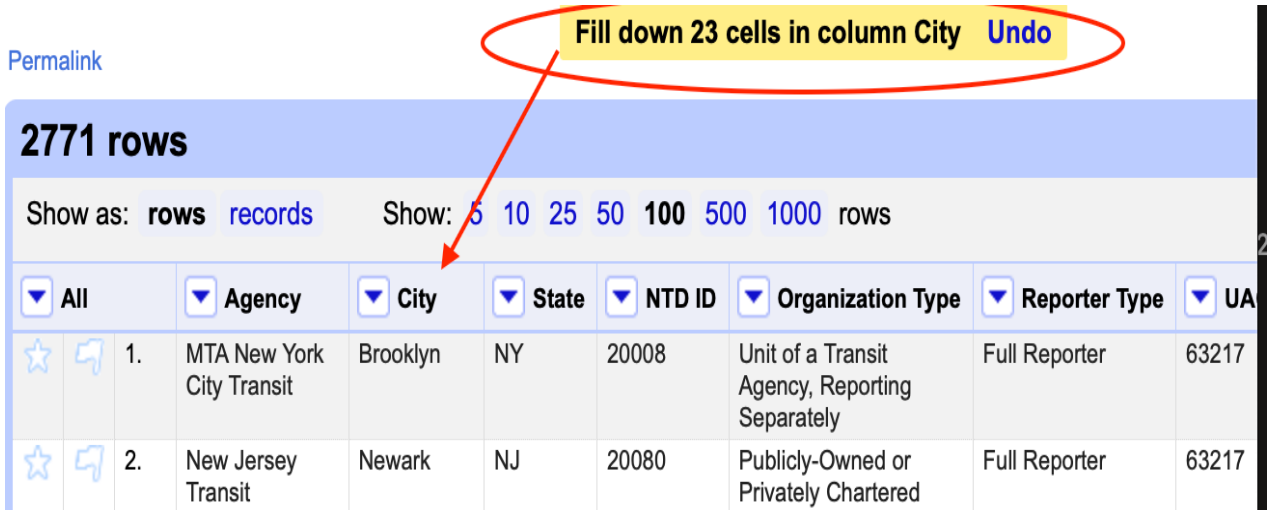
Using facets and filters

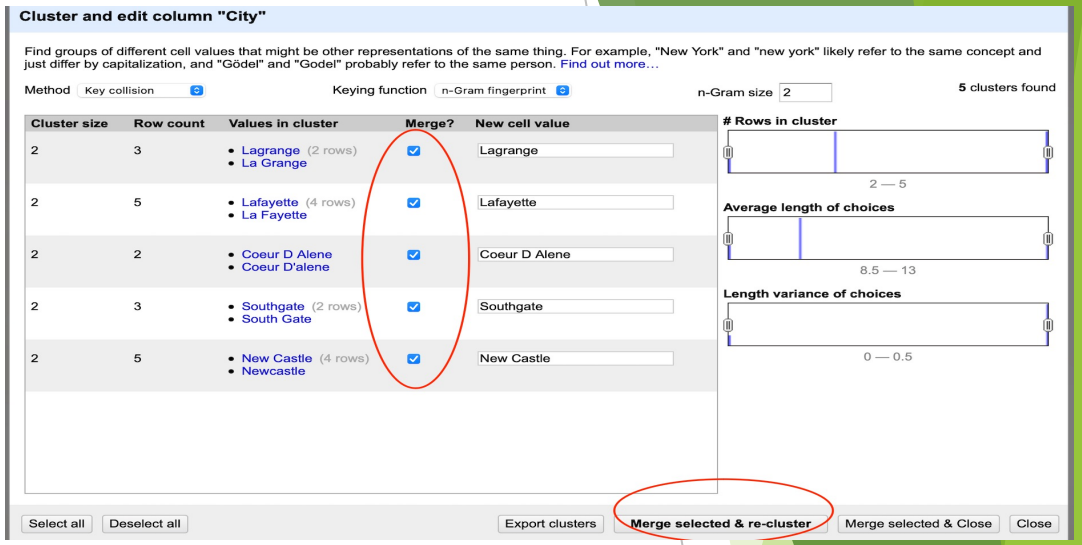
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

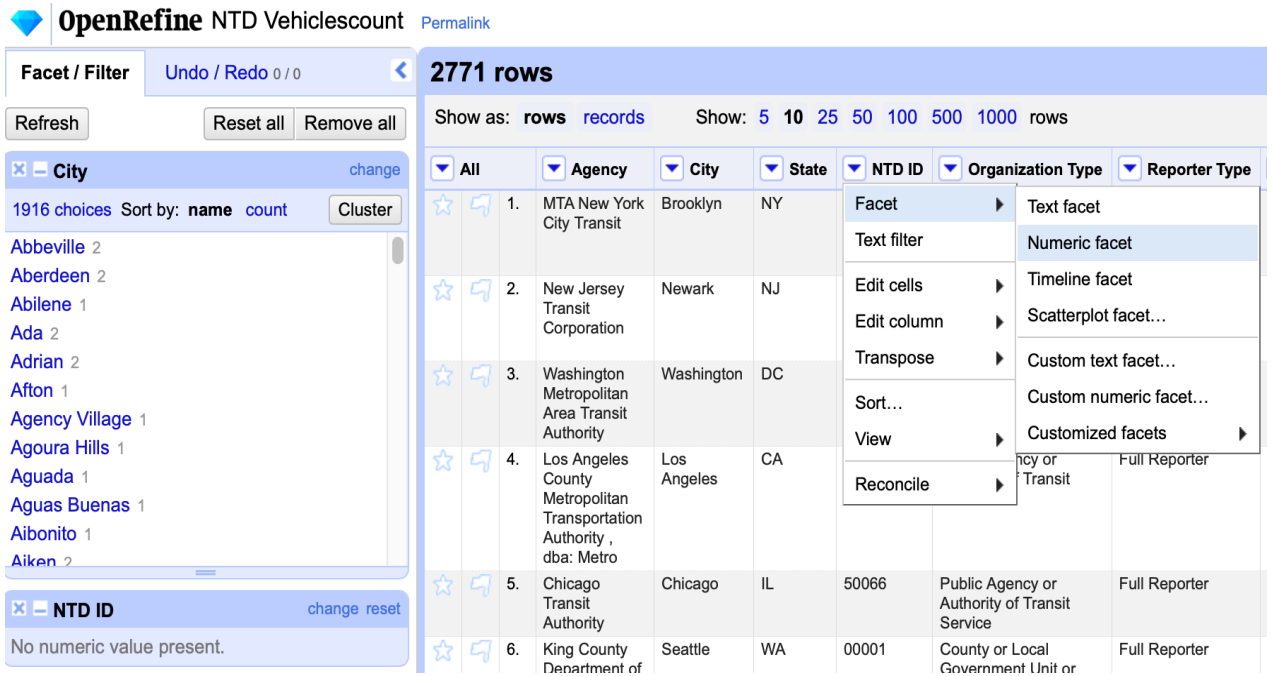
Not sure how to get started?
Watch these screencasts

All	Agency	City	State	NTD ID	Organization Type	Reporter Type	UACE Code	UZA Name
101.	Blue Water Area Transportation Commission, dba: Blue Water Area Transit	Port Huron	MI	50148	Public Agency or Authority of Transit Service	Full Reporter	71155	Port Huron, MI
102.	City of Detroit , dba: Detroit Department of Transportation	Detroit	MI	50119	County or Local Government Unit or Department of Transportation	Full Reporter	23824	Detroit, MI
103.	Southern California Regional Rail Authority, dba: Metrolink	Los Angeles	CA	90151	Public Agency or Authority of Transit Service	Full Reporter	51445	Los Angeles--Long Beach--Anaheim, CA
104.	Central Arkansas Development Council (CADC/SCAT), dba: South Central Arkansas Transit	Benton	AR	60246	Corporation	Rural Reporter	50392	Little Rock, AR
105.	Tampa Bay Area Regional Transit Authority	Tampa	FL	40200	Public Agency or Authority of Transit Service	Full Reporter	86599	Tampa--St. Petersburg, FL
106.	Riverside Transit Agency	Riverside	CA	90031	Public Agency or Authority of Transit Service	Full Reporter	75340	Riverside--San Bernardino, CA
107.	Interurban Transit Partnership, dba: The Rapid	Grand Rapids	MI	50033	Public Agency or Authority of Transit Service	Full Reporter	34300	Grand Rapids, MI
108.	Denton County Transportation Authority	Lewisville	TX	60101	Public Agency or Authority of Transit Service	Full Reporter	23500	Denton--Lewisville, TX
109.	METRO Regional Transit Authority	Akron	OH	50010	Public Agency or Authority of Transit Service	Full Reporter	766	Akron, OH

Refined Dataset: NTD Vehicle Count by Agency


Permalink
Fill down 23 cells in column City
Undo
2771 rows
Show as: rows records
Show: 5 10 25 50 100 500 1000 rows
All
Agency
City
State
NTD ID
Organization Type
Reporter Type
1. MTA New York City Transit | Brooklyn | NY | 20008 | Unit of a Transit Agency, Reporting Separately | Full Reporter | 63217
2. New Jersey Transit | Newark | NJ | 20080 | Publicly-Owned or Privately Chartered | Full Reporter | 63217

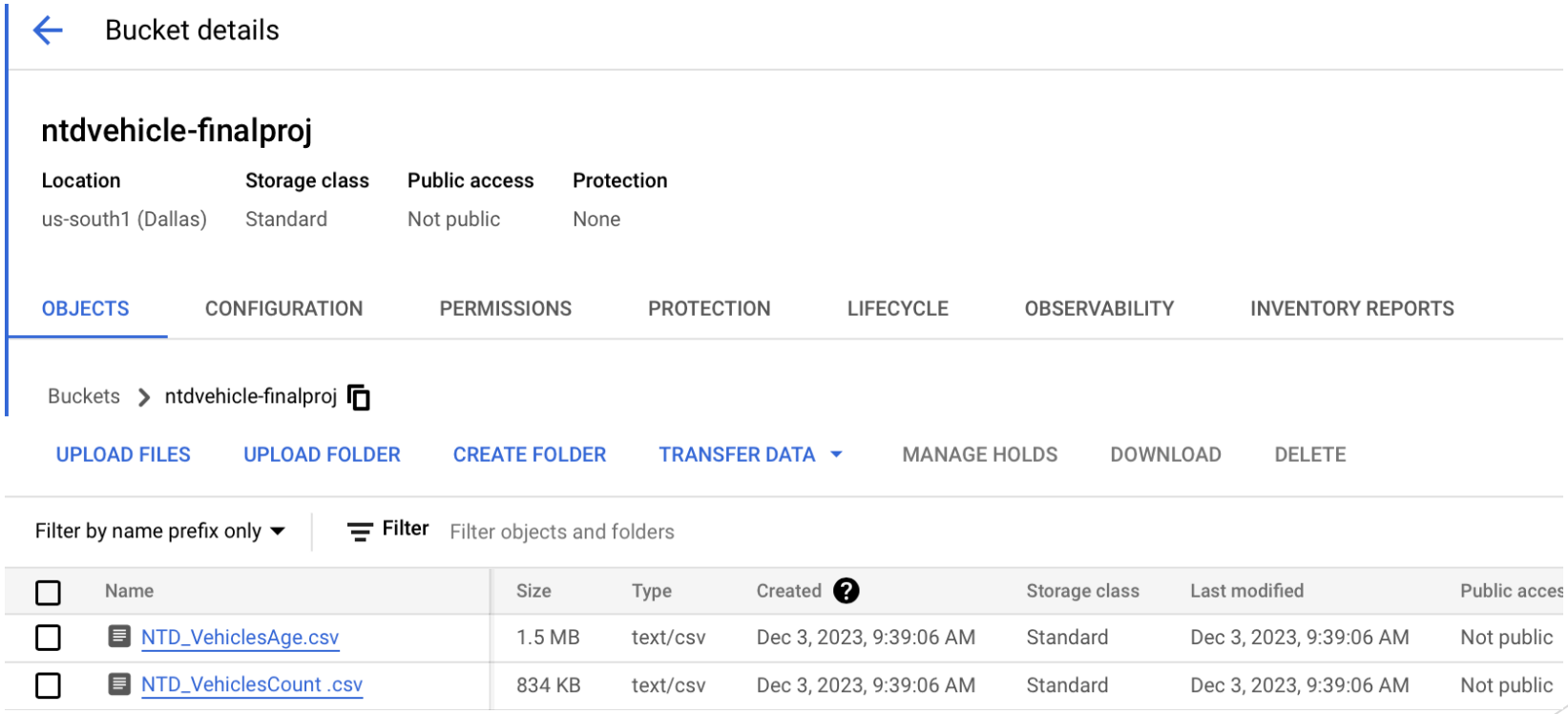

Cluster and edit column "City"
Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more...
Method Key collision
Keying function n-Gram fingerprint
n-Gram size 2
5 clusters found
Cluster size | Row count | Values in cluster | Merge? | New cell value
2 | 3 | Lagrange (2 rows), La Grange | Lagrange
2 | 5 | Lafayette (4 rows), La Fayette | Lafayette
2 | 2 | Coeur D Alene, Coeur D'alene | Coeur D Alene
2 | 3 | Southgate (2 rows), South Gate | Southgate
2 | 5 | New Castle (4 rows), Newcastle | New Castle
Rows in cluster
2 — 5
Average length of choices
8.5 — 13
Length variance of choices
0 — 0.5
Select all
Deselect all
Export clusters
Merge selected & re-cluster
Merge selected & Close
Close


OpenRefine NTD Vehiclescount
Permalink
Facet / Filter
Undo / Redo 0 / 0
Refresh
Reset all
Remove all
City
change
1916 choices Sort by: name count
Cluster
Abbeville 2
Aberdeen 2
Abilene 1
Ada 2
Adrian 2
Afton 1
Agency Village 1
Agoura Hills 1
Aguada 1
Aguas Buenas 1
Aibonito 1
Aiken 2
NTD ID
change reset
No numeric value present.
2771 rows
Show as: rows records
Show: 5 10 25 50 100 500 1000 rows
All
Agency
City
State
NTD ID
Organization Type
Reporter Type
1. MTA New York City Transit | Brooklyn | NY
2. New Jersey Transit Corporation | Newark | NJ
3. Washington Metropolitan Area Transit Authority | Washington | DC
4. Los Angeles County Metropolitan Transportation Authority , dba: Metro | Los Angeles | CA
5. Chicago Transit Authority | Chicago | IL | 50066 | Public Agency or Authority of Transit Service | Full Reporter
6. King County Department of | Seattle | WA | 00001 | County or Local Government Unit or | Full Reporter
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet...
Custom text facet...
Custom numeric facet...
Customized facets


Average Age of Fleet (in Years)
Average Lifetime Miles per Vehicle
4.12 | 118,610
4.12 edit | 118,610
4.12 | 118,610
8.18 | 184,696
6.93 | 231,206
26.07 | 1,391,691
4.59 | 146,004
5.89 | 181,352
5.89 | 181,352
6.96 | 175,524

GCP Cloud Data Storage

- The two datasets Vehicle Age Distribution & Vehicle Type count by Agency are stored in a storage bucket in GCP.



Big Query Studio

Dataset 1 –Vehicles Age

Total Vehicles by state

RUN SAVE QUERY

```
1  SELECT State, SUM(`Total_Vehicles`) AS TotalVehicles
2  FROM `VehiclesAge_2.VehiclesAge`
3  GROUP BY State
4  ORDER BY TotalVehicles DESC
5  LIMIT 5;
```

JOB INFORMATION | RESULTS | CHART | PREVIEW

Row	State	TotalVehicles
1	CA	26187
2	NY	24551
3	IL	11408
4	TX	10030
5	WA	9222

Top Cities with more than 1...les

RUN

```
1  SELECT City, SUM(`Total_Vehicles`) AS TotalVehicles
2  FROM `VehiclesAge_2.VehiclesAge`
3  GROUP BY City
4  HAVING SUM(`Total_Vehicles`) > 1000
5  ORDER BY SUM(`Total_Vehicles`) DESC
6  LIMIT 5;
```

Row	City	TotalVehicles
1	Brooklyn	14358
2	Chicago	6137
3	Los Angeles	5615
4	Newark	5487
5	Washington	4764

Big Query Studio

Dataset 2-VehiclesCount

Count of Agencies by State

```
1  SELECT State, COUNT(*) as AgencyCount
2  FROM `VehiclesCount_1.VehiclesCount`
3  GROUP BY State
4  ORDER BY AgencyCount DESC
5  LIMIT 5;
6
```

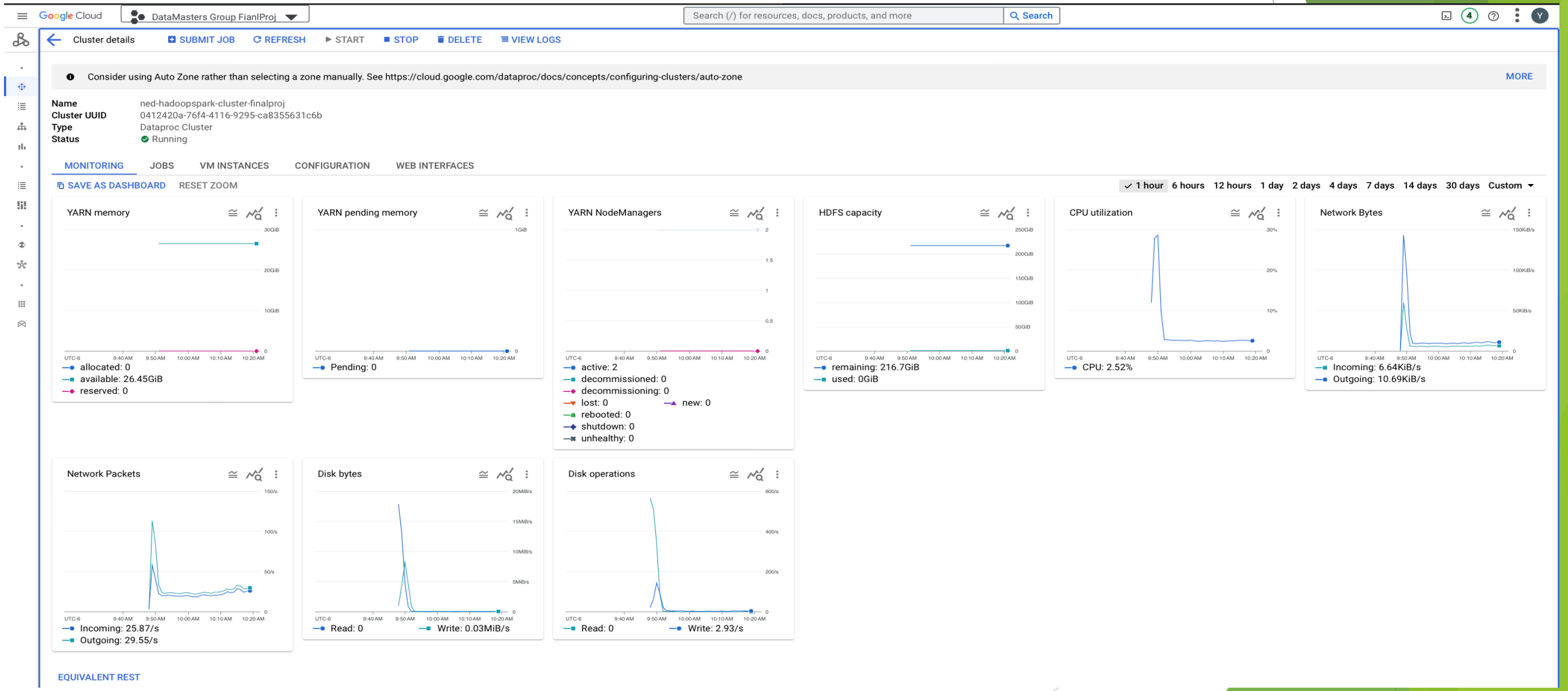
Row	State	AgencyCount
1	CA	218
2	MI	197
3	WI	163
4	OH	153
5	NC	101

List of Agenci... RUN

```
1  SELECT Agency, City
2  FROM `VehiclesCount_1.VehiclesCount`
3  WHERE City = 'Los Angeles';
4
```

Row	Agency	City
1	City of Los Angeles, dba: City of Los Angeles Department of Transportation	Los Angeles
2	Los Angeles County Metropolitan Transportation Authority , dba: Metro	Los Angeles
3	Southern California Regional R...	Los Angeles

Cluster Monitoring



Connecting to Hive

```
yashwanthjilla_unt@ntd-hadoopspark-cluster-finalproj-m:~$ pwd
/home/yashwanthjilla_unt
yashwanthjilla_unt@ntd-hadoopspark-cluster-finalproj-m:~$ beeline -u jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 3.1.3)
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://localhost:10000>
```

Connecting to Spark

```
yashwanthjilla_unt@ned-hadoopspark-cluster-finalproj-m:~$
yashwanthjilla_unt@ned-hadoopspark-cluster-finalproj-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
23/12/03 20:17:24 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/12/03 20:17:24 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/12/03 20:17:25 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/12/03 20:17:25 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1701629121542_0003
```

Hive

Queries & Output for Dataset-1

```
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE IF NOT EXISTS ntd_vehicles
. . . . . .> (ntdvehicles string)
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION '/user/yashwanthjilla_unt/data/NTD_VehiclesCount/';
INFO  : Compiling command(queryId=hive_20231203194153_d45e8c8b-b373-4777-9ce6-e6aa8619ed66):
```

```
INFO  : Completed executing command(q
INFO  : OK
INFO  : Concurrency mode is disabled,
No rows affected (2.787 seconds)
0: jdbc:hive2://localhost:10000>
```

```
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> SELECT * FROM ntd_vehicles LIMIT 1;
```

```
+
+----------------------------------
1 row selected (33.105 seconds)
0: jdbc:hive2://localhost:10000>
```

```
0: jdbc:hive2://localhost:10000> SELECT * FROM ntd_vehicles LIMIT 5;
```

```
+----------------------------------
5 rows selected (11.205 seconds)
0: jdbc:hive2://localhost:10000>
```

Hive Queries & Output for dataset-2

```
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE IF NOT EXISTS ntd_vehiclesAge
. . . . . .> (ntdvehiclesAge STRING)
. . . . . .> ROW FORMAT DELIMITED
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION '/user/yashwanthjilla_unt/data/NTD_VehiclesAge/';
INFO  : Compiling command(queryId=hive_20231203200155_027c678d-fa2f-4731-9877-3d4796397070): C
```

```
INFO  : Completed executing command
INFO  : OK
INFO  : Concurrency mode is disable
No rows affected (0.145 seconds)
0: jdbc:hive2://localhost:10000>
```

```
0: jdbc:hive2://localhost:10000> SELECT COUNT(*) FROM ntd_vehiclesAge;
```

```
+-------+
|  _c0  |
+-------+
| 7910  |
+-------+
1 row selected (17.672 seconds)
0: jdbc:hive2://localhost:10000>
```

```
0: jdbc:hive2://localhost:10000> SELECT AVG(LENGTH(ntdvehiclesAge)) AS avg_length FROM ntd_vehiclesAge;
```

```
+--------------------+
|     avg_length     |
+--------------------+
| 201.6211125158028  |
+--------------------+
1 row selected (6.467 seconds)
0: jdbc:hive2://localhost:10000>
```

Spark Queries for Data Set 1 & 2

```
== SQL ==
CREATE EXTERNAL TABLE IF NOT EXISTS ntd_vehicles
 (ntdvehicles string)ashwanthjilla_unt/data/NTD_VehiclesCount/'
--------------------^^^

spark-sql> CREATE EXTERNAL TABLE IF NOT EXISTS ntd_vehicles
         > (
         >   ntdvehicles STRING
         > )
         > ROW FORMAT DELIMITED
         > STORED AS TEXTFILE
         > LOCATION '/user/yashwanthjilla_unt/data/NTD_VehiclesCount/';
23/12/03 20:22:10 WARN org.apache.hadoop.hive.ql.session.SessionState: M
thorizerFactory.
Time taken: 0.575 seconds
spark-sql>
```

```
spark-sql> SELECT AVG(LENGTH(ntdvehiclesAge)) AS avg_length FROM ntd_vehiclesAge;
201.6211125158028
Time taken: 0.923 seconds, Fetched 1 row(s)
spark-sql>
```

```
spark-sql>
         > SELECT * FROM ntd_vehicles LIMIT 1;
Agency,City,State,NTD ID,Organization Type,Reporter Type,UACE Code,UZA N
with ULB Reported,Articulated Bus >= ULB,Over-the-Road Bus,Over-the-Roac
 Decker Bus >= ULB,School Bus,SchoolBus with ULB Reported,School Bus >=
ile with ULB Reported,Automobile >= ULB,Minivan,Minivan with ULB Reporte
Trolleybus,TrolleyBus with ULB Reported,Trolleybus >= ULB,Heavy Rail Pas
ht Rail Vehicle with ULB Reported,Light Rail Vehicle >= ULB,Commuter Rai
ter Rail Self-Propelled Passenger Car,Commuter Rail Self-Propelled Passe
Reported,Locomotive >= ULB,Automated Guideway Vehicle,Automated Guideway
ey with ULB Reported,Vintage/Historic Trolley >= ULB,Streetcar,Streetcar
norail,Monorail >= ULB,Cable Car,Cable Car with ULB Reported,Cable Car >
orted,Ferryboat >= ULB,Other,Other with ULB Reported,Other >= ULB,Total
 and Other Rubber Tire Vehicles,Trucks and Other Rubber Tire Vehicles >=
Time taken: 5.717 seconds, Fetched 1 row(s)
spark-sql>
```

```
LOCATION '/user/yashwanthjilla_unt/data/NTD_VehiclesAge/'

spark-sql> CREATE EXTERNAL TABLE IF NOT EXISTS ntd_vehiclesAge
         > (
         >   ntdvehiclesAge STRING
         > )
         > ROW FORMAT DELIMITED
         > STORED AS TEXTFILE
         > LOCATION '/user/yashwanthjilla_unt/data/NTD_VehiclesAge/';
Time taken: 0.12 seconds
spark-sql>
```

Hive Vs. Spark Comparison

Dataset	Query Description	Hive Execution Time (sec)	Spark Execution Time (sec)
Dataset-1	CREATE EXTERNAL TABLE	2.787	0.575
Dataset-1	Selecting 1 row	33.105	5.717
Dataset-1	Selecting 5 rows	11.205	0.339
Dataset-2	CREATE EXTERNAL TABLE in VehiclesAge	0.145	0.12
Dataset-2	Select Count all from Vehicle Age	17.672	1.706
Dataset-2	Select Average length	6.467	0.923

Remarks & Key Analysis

- Hive is suitable for traditional data warehousing and SQL queries.
- Spark's in-memory processing makes it a superior choice for tasks involving iterative algorithms and real-time analytics.
- In comparing Hive and Spark based on the provided data and queries, Spark consistently outperforms Hive with faster execution times, particularly for data processing and analytics.
- Developing robust data management through complex queries, deeper multidimensional analysis

Conclusion

Data experts can uncover patterns and relationships to better understand transit usage. This allows for data-driven forecasting and infrastructure planning to efficiently address current problems and proactively meet future needs.

- Predict optimal accessible stop locations via modeling.
- Forecast peak demand periods with predictive analytics.
- Collect and analyze emissions, mileage, fuel efficiency data.
- Model route optimization scenarios to increase sustainability.

Thank You !

- We sincerely thank Dr. Tony Fantasia for providing us the opportunity to expand our knowledge of powerful data tools

CITATIONS:

- https://spark.apache.org/
- https://hive.apache.org/
- https://cloud.google.com/
- https://www.transit.dot.gov/ntd/data-product/2022-vehicles