

**Curso:** “Big Data I: Cloud Computing y almacenamiento masivo de datos”

**Parte ETL:** Diseña un experimento ETL con Impala a tu medida

**Estudiante:** Edgar Pérez Rivera

**Correo:** edjperez@correo.ugr.es

**DNI:** PA1099417

## 1. Estructura y contenido de la base de datos

**Indicaciones:** Localiza una base de datos CSV para utilizar en los ejercicios de ETL, que podrás utilizar tanto en Big Data I como en Big Data II. Debe ser suficientemente diversa en cuanto a representación (tener al menos 10-15 características) y a población (tener más de 5000 registros).

**Resolución:** Para la asignatura y trabajos de la materia, se utilizará la base de datos llamada [Electric Vehicle Population Data](#), esta base de datos es distribuida por el gobierno de los Estados Unidos en el estado de Washington por el departamento de Licencias de Conducir, con el fin de llevar el registro de la utilización y emisión de los coches eléctricos de batería (BEV) y los vehículos eléctricos híbridos enchufables (PHEV) que están actualmente registrados a través del Departamento de Licencias del Estado de Washington (DOL). Distribuido en formato CSV, RDF, JSON y XML, cuenta con 17 características y 174,000 registros de carácter público y con acceso libre, el tipo de licencia de este conjunto de datos es Open Data Commons Open Database License (ODbL) v1.0, con la última actualización el 14 de febrero de 2024.

Los vehículos eléctricos de batería (BEV) y los vehículos eléctricos híbridos enchufables (PHEV) utilizan baterías para alimentar sus motores, cargándose mediante una fuente eléctrica. La elegibilidad para los vehículos de combustible alternativo limpio (CAFV) se determina por los requisitos de combustible y autonomía eléctrica según las leyes estatales de Washington.

Escogí esta base de datos por cumplir con las indicaciones dadas en clase, no es una base de datos repetida como lo es Iris o MNIST, de igual forma es bastante actualizada, cuenta con buena cantidad de características, 17 en específico y la cantidad de registro es aceptable para realizar las asignaciones dadas en la materia.

### Descripción de las columnas:

Característica	Descripción corta	Tipo
VIN (1-10)	Los primeros 10 caracteres del VIN de cada vehículo	String
County	Región geográfica del estado donde reside el propietario	String
City	Ciudad de residencia del propietario registrado	String

Característica	Descripción corta	Tipo
State	Región geográfica asociada con el registro del vehículo	String
Postal Code	Código postal de residencia del propietario registrado	String
Model Year	Año del modelo del vehículo, determinado por el VIN	String
Make	Fabricante del vehículo determinado por el VIN	String
Model	Modelo del vehículo determinado por el VIN	String
Electric Vehicle Type	Indica si es totalmente eléctrico o híbrido enchufable	String
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Clasificación del vehículo como CAFV según requisitos legales	String
Electric Range	Distancia que puede viajar un vehículo con su carga eléctrica	Int
Base MSRP	Precio minorista sugerido por el fabricante (MSRP)	Int
Legislative District	Sección específica del estado de Washington donde reside el propietario	String
DOL Vehicle ID	Número único asignado a cada vehículo por el Departamento de Licencias	String
Vehicle Location	Centro del código postal del vehículo registrado	Float
Electric Utility	Territorios de servicio minorista de energía eléctrica	String
2020 Census Tract	Identificador de zona censal asignado por la Oficina del Censo en 2020	String

Esta base se puede encontrar en la página oficial del estado de Washington, la url es:

[https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2/about\\_data](https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2/about_data)

Alternativamente, se puede encontrar un respaldo en el catálogo de datos del Gobierno de los Estados Unidos, esta alternativa la coloco, ya que la primera dirección en el Cloudoera demoraba en bajar el fichero en la máquina virtual:

<https://catalog.data.gov/dataset/electric-vehicle-population-data>

## 2. Diseña un experimento ETL con Impala a tu medida

**Indicaciones:** Usando como base la presentación "Una herramienta del Ecosistema Hadoop para la consulta estructurada: Impala", que podrás encontrar en esta sección de la asignatura en PRADO, diseña un experimento ETL sobre la base de datos que hayas seleccionado para la experimentación usando la herramienta Impala.

- Carga del fichero y configuración de Impala.

**Resolución:**

La experimentación se centrará en el uso de dos operaciones:

- Incluye una operación de proyección con al menos tres atributos o características pero no todas (hasta 3 puntos).

**Resolución:**

La primera sección de este laboratorio se va a realizar la carga del fichero llamado `Electric_Vehicle_Population_Data` mediante el `hdfs`, luego de la carga solo queda verificar que el fichero esté en la ubicación correcta con el comando `-ls`, la verificación es como recomendación, ya que con las pruebas veía que el fichero es eliminado cuando es usado por `impala` en los casos que repetía el experimento.

```
[cloudera@quickstart ~]$ sudo bash
[root@quickstart cloudera]# su - impala
-bash-4.1$ hdfs dfs -put /var/tmp/materialImpala/Electric_Vehicle_
put
-bash-4.1$
-bash-4.1$
-bash-4.1$
-bash-4.1$ hdfs dfs -ls /user/impala/input
Found 1 items
-rw-r--r--    1 impala supergroup    42030494 2024-03-03 15:49 /user
lation Data.csv
```

La segunda acción es la creación de la base de datos a cargar en impala, donde ya se pueden ejecutar sentencias sql desde el impala-shell, si la base de datos no se ha creado se crea vacía, se procede a verificar si fue creada correctamente.

```

root@quickstart:/home/cloudera
File Edit View Search Terminal Help
[quickstart.cloudera:21000] > CREATE DATABASE IF NOT EXISTS bdcар;
Query: create DATABASE IF NOT EXISTS bdcар
Fetched 0 row(s) in 1.32s
[quickstart.cloudera:21000] > CREATE DATABASE IF NOT EXISTS bdcар
LOCATION '/user/impala/impalastore.db';
Query: create DATABASE IF NOT EXISTS bdcар LOCATION '/user/impala/
impalastore.db'
Fetched 0 row(s) in 0.19s
[quickstart.cloudera:21000] > show databases;
Query: show databases
+-----+-----+
+
| name          | comment
|
+-----+-----+
+
| _impala_builtins | System database for Impala builtin functions
|
| bdcар          |

```



```
[quickstart.cloudera:21000] > LOAD DATA INPATH '/user/impala/input
/Electric_Vehicle_Population_Data.csv' OVERWRITE INTO TABLE Electr
icCar;
Query: load DATA INPATH '/user/impala/input/Electric_Vehicle_Popul
ation_Data.csv' OVERWRITE INTO TABLE ElectricCar
+-----+
| summary |
+-----+
| Loaded 1 file(s). Total files in destination location: 1 |
+-----+
Fetched 1 row(s) in 0.76s
[quickstart.cloudera:21000] > █
```

Consulta de prueba, similar al documento proporcionado en la asignatura.

```
[quickstart.cloudera:21000] > SELECT ModelYear, COUNT(*) AS TotalC
> FROM ElectricCar
> GROUP BY ModelYear;
Query: select ModelYear, COUNT(*) AS TotalCars
FROM ElectricCar
GROUP BY ModelYear
Query submitted at: 2024-03-04 05:41:34 (Coordinator: http://quick
:25000)
Query progress can be monitored at: http://quickstart.cloudera:250
query_id=4944a9ca2c5b8498:e1fd0fca00000000
+-----+-----+
| modelyear | totalcars |
+-----+-----+
| 2014      | 3532      |
| 1999      | 4          |
| 2008      | 21         |
| 2012      | 1638       |
| 2010      | 22         |
| 1997      | 1          |
| 2021      | 19033      |
| 2002      | 2          |
| 2019      | 10913      |
| 2023      | 55284      |
+-----+-----+
```

- Incluye una operación de proyección con al menos tres atributos o características pero no todas (hasta 3 puntos)

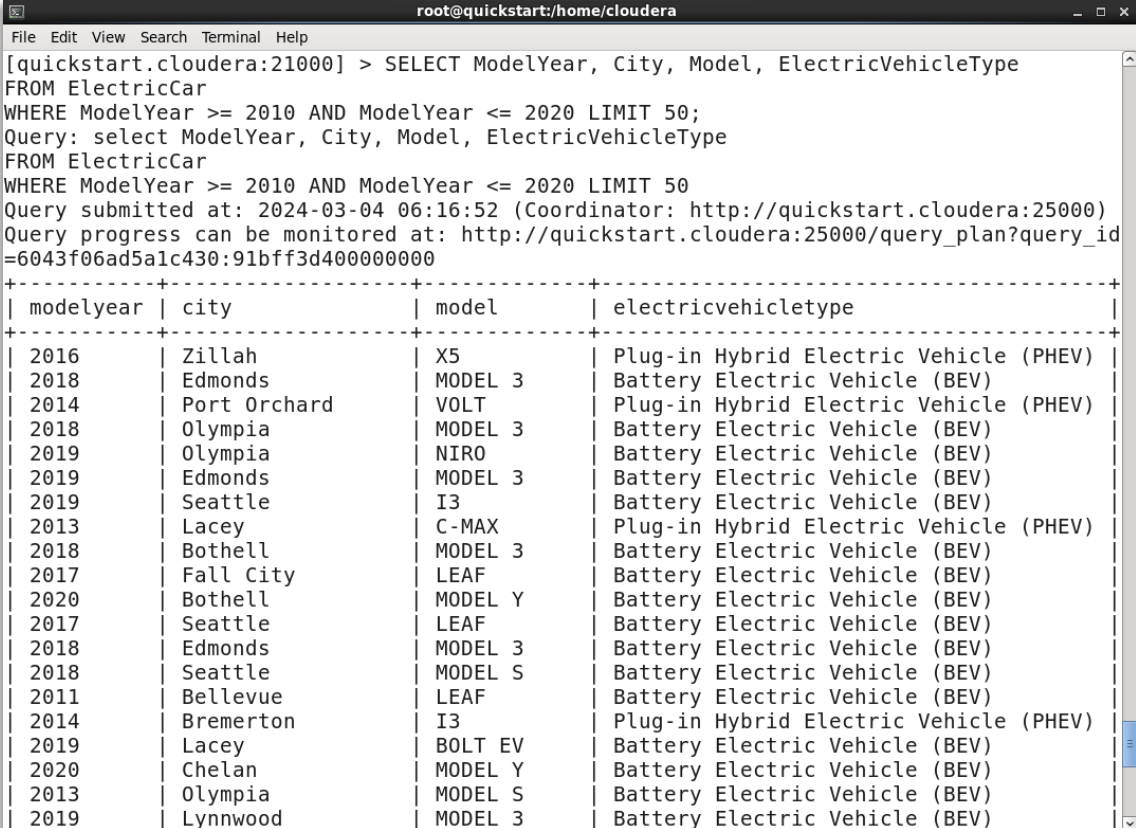
Para este apartado se escogió para la operación de proyección las columnas de City, Model, ElectricVehicleType, a esta consulta se le tuvo que colocar un LIMIT de 50 ya que al testear la operación siempre cargaba los 170,000 registro y se perdía información visual en la consola.

```

root@quickstart:/home/cloudera
File Edit View Search Terminal Help
[quickstart.cloudera:21000] > SELECT City, Model, ElectricVehicleType
FROM ElectricCar
LIMIT 50;
Query: select City, Model, ElectricVehicleType
FROM ElectricCar
LIMIT 50
Query submitted at: 2024-03-04 06:09:13 (Coordinator: http://quickstart.cloudera:25
000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?que
ry_id=847d56877310f2e:5863c606000000000
+-----+-----+-----+
| city          | model      | electricvehicletype |
+-----+-----+-----+
| City          | Model      | Electric Vehicle Type |
| Zillah        | X5         | Plug-in Hybrid Electric Vehicle (PHEV) |
| Edmonds       | MODEL 3    | Battery Electric Vehicle (BEV) |
| Port Orchard  | VOLT       | Plug-in Hybrid Electric Vehicle (PHEV) |
| Bow           | PACIFICA   | Plug-in Hybrid Electric Vehicle (PHEV) |
| Olympia       | MODEL 3    | Battery Electric Vehicle (BEV) |
| Snohomish     | Q5         | Plug-in Hybrid Electric Vehicle (PHEV) |
| Olympia       | NIRO       | Battery Electric Vehicle (BEV) |
| Edmonds       | MODEL 3    | Battery Electric Vehicle (BEV) |
| Seattle       | I3         | Battery Electric Vehicle (BEV) |
| Lacey         | C-MAX      | Plug-in Hybrid Electric Vehicle (PHEV) |
| Bothell       | MODEL 3    | Battery Electric Vehicle (BEV) |
| Fall City     | LEAF       | Battery Electric Vehicle (BEV) |
| Olympia       | TUCSON     | Plug-in Hybrid Electric Vehicle (PHEV) |
| Bothell       | MODEL Y    | Battery Electric Vehicle (BEV) |
| Seattle       | LEAF       | Battery Electric Vehicle (BEV) |

```

- Incluye una operación de selección con una condición simple que solo involucre un operador AND u OR (hasta 2 puntos)  
La segunda operación para realizar es la de AND, se escogió mostrar los modelos de coches entre los años 2010 y 2020.



```

[quickstart.cloudera:21000] > SELECT ModelYear, City, Model, ElectricVehicleType
FROM ElectricCar
WHERE ModelYear >= 2010 AND ModelYear <= 2020 LIMIT 50;
Query: select ModelYear, City, Model, ElectricVehicleType
FROM ElectricCar
WHERE ModelYear >= 2010 AND ModelYear <= 2020 LIMIT 50
Query submitted at: 2024-03-04 06:16:52 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=
=6043f06ad5alc430:91bff3d400000000

```

modelyear	city	model	electricvehicletype
2016	Zillah	X5	Plug-in Hybrid Electric Vehicle (PHEV)
2018	Edmonds	MODEL 3	Battery Electric Vehicle (BEV)
2014	Port Orchard	VOLT	Plug-in Hybrid Electric Vehicle (PHEV)
2018	Olympia	MODEL 3	Battery Electric Vehicle (BEV)
2019	Olympia	NIRO	Battery Electric Vehicle (BEV)
2019	Edmonds	MODEL 3	Battery Electric Vehicle (BEV)
2019	Seattle	I3	Battery Electric Vehicle (BEV)
2013	Lacey	C-MAX	Plug-in Hybrid Electric Vehicle (PHEV)
2018	Bothell	MODEL 3	Battery Electric Vehicle (BEV)
2017	Fall City	LEAF	Battery Electric Vehicle (BEV)
2020	Bothell	MODEL Y	Battery Electric Vehicle (BEV)
2017	Seattle	LEAF	Battery Electric Vehicle (BEV)
2018	Edmonds	MODEL 3	Battery Electric Vehicle (BEV)
2018	Seattle	MODEL S	Battery Electric Vehicle (BEV)
2011	Bellevue	LEAF	Battery Electric Vehicle (BEV)
2014	Bremerton	I3	Plug-in Hybrid Electric Vehicle (PHEV)
2019	Lacey	BOLT EV	Battery Electric Vehicle (BEV)
2020	Chelan	MODEL Y	Battery Electric Vehicle (BEV)
2013	Olympia	MODEL S	Battery Electric Vehicle (BEV)
2019	Lynnwood	MODEL 3	Battery Electric Vehicle (BEV)

- Si incluye una operación de selección más compleja que involucre un operador AND (u OR) combinado con un operador OR (o AND) o con otro operador lógico distinto (hasta 4 puntos).  
Para este apartado se utilizó la operación anterior y se agregó el OR , mostrando los registros de la ciudad de Seattle o Edmonds.

```
root@quickstart:/home/cloudera
File Edit View Search Terminal Help
[quickstart.cloudera:21000] > SELECT City, Model, ElectricVehicleType, COUNT(*) AS TotalCars
FROM ElectricCar
WHERE (City = 'Edmonds' OR City = 'Seattle') AND ModelYear BETWEEN 2010 AND 2020
GROUP BY City, Model, ElectricVehicleType;
Query: select City, Model, ElectricVehicleType, COUNT(*) AS TotalCars
FROM ElectricCar
WHERE (City = 'Edmonds' OR City = 'Seattle') AND ModelYear BETWEEN 2010 AND 2020
GROUP BY City, Model, ElectricVehicleType
Query submitted at: 2024-03-04 06:20:53 (Coordinator: http://quickstart.cloudera:25000)
Query progress can be monitored at: http://quickstart.cloudera:25000/query_plan?query_id=204e1dd02e078b79:ca7808830000000000
+-----+-----+-----+-----+
--+
| city      | model      | electricvehicletype      | totalcars |
+-----+-----+-----+-----+
| Edmonds   | OUTLANDER  | Plug-in Hybrid Electric Vehicle (PHEV) | 2         |
| Edmonds   | XC60       | Plug-in Hybrid Electric Vehicle (PHEV) | 8         |
| Seattle   | X5         | Plug-in Hybrid Electric Vehicle (PHEV) | 96        |
| Edmonds   | LEAF       | Battery Electric Vehicle (BEV)         | 175       |
| Seattle   | I8         | Plug-in Hybrid Electric Vehicle (PHEV) | 8         |
| Edmonds   | RAV4       | Battery Electric Vehicle (BEV)         | 1         |
| Edmonds   | CLARITY    | Plug-in Hybrid Electric Vehicle (PHEV) | 9         |
+-----+-----+-----+-----+
```

Con estos tres apartados se culmina la práctica **Diseña un experimento ETL con Impala a tu medida** implementando el uso de hdfs e impala para la carga, proyección y consulta de información mediante el motor de consulta SQL para realizar consultas interactivas de baja latencia sobre datos almacenados en HDFS.

Referencias

Loading CSV Data into an Impala Table. (n.d.). Retrieved February 15, 2024, from <https://docs.cloudera.com/cdsw/1.10.5/import-data/topics/cdsw-loading-csv-data-into-an-impala-table.html>