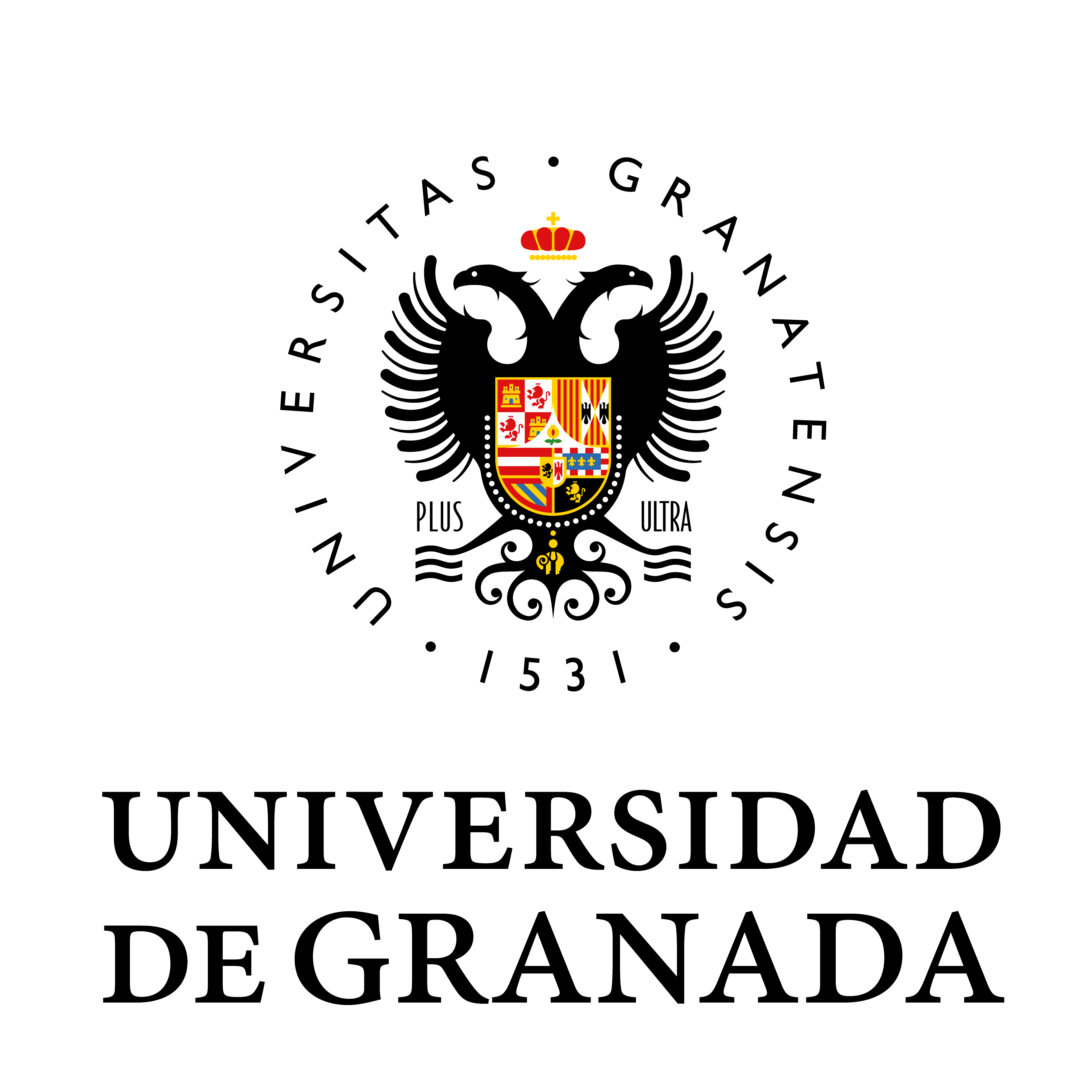
**Universidad de Granada**

**Máster Universitario en Ciencia de Datos e Ingeniería de Computadores**

**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*](https://data.wa.gov/Transportation/Electric-Vehicle-Population-Data/f6w7-q2d2/about_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 |
| 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>

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>

1. **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.

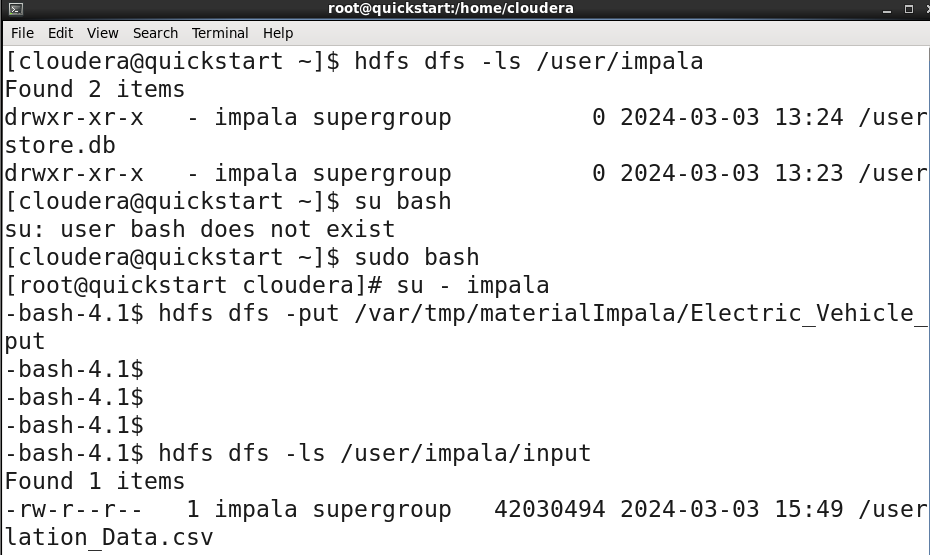
**Resolución:**

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\_Populartion\_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.



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.

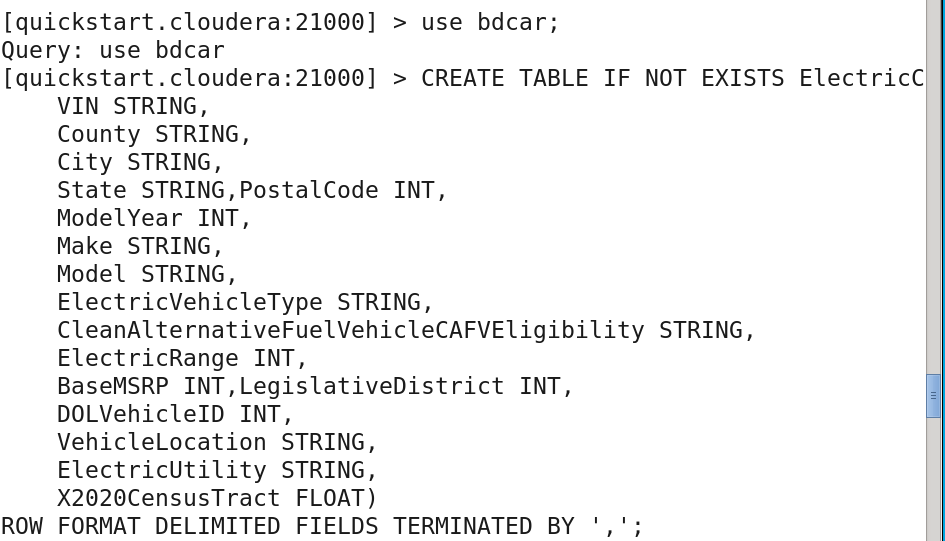
Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente con confianza media

Una vez creada la base de datos, se procede a crear la o las tablas a utilizar en ella, de este modo aquí se nombre las columnas y se dan las indicaciones de formateo de datos.



Carga de los datos del fichero a la base de datos.

Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente con confianza media

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

Tabla

Descripción generada automáticamente con confianza media

* 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.

Tabla

Descripción generada automáticamente

* 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.

Tabla

Descripción generada automáticamente

* 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.

Tabla

Descripción generada automáticamente

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>