Práctica Hive

1. A partir de los datos (CSV) de Padrón de Madrid (<https://datos.madrid.es/egob/catalogo/200076-1-padron.csv>) llevar a cabo lo siguiente:
   1. Crear Base de datos **datos\_padron**
   2. Crear tabla padron\_txt con todos los campos del fichero CSV y cargar los datos mediante el comando LOAD DATA LOCAL INPATH. La tabla tendrá formato texto y tendrá como delimitador de campo el carácter ';' y los campos estarán encerrados en comillas dobles '”’ y se deberá omitir la cabecera del fichero de datos al crear la tabla
   3. (OPCIONAL) Crear tabla padron\_txt2 que haga **trim** sobre los datos de algunas columnas con la finalidad de eliminar los espacios en blanco innecesarios.
   4. Investigar y entender la diferencia de incluir la palabra LOCAL en el comando LOAD DATA
2. ¿Qué es CTAS?
   1. Crear Tabla (en Hive) padron\_parquet (cuyos datos serán almacenados en el formato columnar **parquet**) a partir de la tabla padron\_txt mediante un CTAS.
   2. (OPCIONAL) Crear Tabla (en Hive) padron\_parquet2 (cuyos datos serán almacenados en el formato columnar **parquet**) a partir de la tabla padron\_txt2 mediante un CTAS.
   3. Investigar en que consiste el formato columnar parquet y las ventajas de trabajar con este tipo de formatos.
   4. Comparar el tamaño de los ficheros de los datos de las tablas padron\_txt (CSV) y padron\_parquet (alojados en hdfs cuya ruta se puede obtener de la propiedad location de cada tabla por ejemplo haciendo SHOW CREATE TABLE)
   5. (OPCIONAL) Comparar el tamaño de los ficheros de los datos de las tablas padron\_txt (CSV), padron\_txt2, padron\_parquet y padron\_parquet2 (alojados en hdfs cuya ruta se puede obtener de la propiedad location de cada tabla por ejemplo haciendo SHOW CREATE TABLE)
3. Impala
   1. ¿Qué es impala?
   2. ¿En qué se diferencia a Hive?
   3. Comando INVALIDATE METADATA, ¿en qué consiste?
   4. Hacer invalidate metadata en Impala de Base de datos **datos\_padron**
   5. Calcular el total de EspanolesHombres, EspanolesMujeres, ExtranjerosHombresy ExtranjerosMujeres agrupado por DESC\_DISTRITO y DESC\_BARRIO.
      1. Llevar a cabo la consulta en Hive en las tablas padron\_txt y padron\_parquet. ¿Alguna conclusión?
      2. Llevar a cabo la consulta en Impala en las tablas padron\_txt y padron\_parquet. ¿Alguna conclusión?
      3. ¿Se percibe alguna diferencia de rendimiento entre Hive e Impala?
4. Particionamiento
   1. Crear tabla (Hive) **padron\_particionado** particionada por los campos DESC\_DISTRITO y DESC\_BARRIO cuyos datos estén en formato parquet.
   2. Insertar datos (en cada partición) dinámicamente (con Hive) en la tabla recién creada a partir de un select de la tabla **padron\_parquet**
   3. Hacer invalidate metadata en Impala de Base de datos **padron\_particionado**
   4. Calcular el total de EspanolesHombres, EspanolesMujeres, ExtranjerosHombresy ExtranjerosMujeres agrupado por DESC\_DISTRITO y DESC\_BARRIO para los distritos CENTRO, LATINA, CHAMARTIN, TETUAN, VICALVARO y BARAJAS.
      1. Llevar a cabo la consulta en Hive en las tablas padron\_parquet y padron\_particionado. ¿Alguna conclusión?
      2. Llevar a cabo la consulta en Impala en las tablas padron\_parquet y padron\_particionado. ¿Alguna conclusión?
   5. Hacer consultas de agregación (Max, Min, Avg, Count) tal cual el ejemplo anterior con las 3 tablas (padron\_txt, padron\_parquet y padron\_particionado) y comparar rendimientos tanto en Hive como en impala y sacar conclusiones.

Información:

* <https://cwiki.apache.org/confluence/collector/pages.action?key=Hive>
* <https://cwiki.apache.org/confluence/display/Hive/GettingStarted#GettingStarted-CreatingHiveTables>
* <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL>
* <https://cwiki.apache.org/confluence/display/Hive/DynamicPartitions>