ACTIVIDAD 3

ANÁLISIS EXPLORATORIO CON APACHE HIVE SOBRE HDFS

PRESENTADO A:

MsC PABLO VILLACORTA IGLESIAS

AUTORES:

ANDRÉS FELIPE LEAL MORA

JUAN DAVID ESCOBAR ESCOBAR

JUAN MANUEL BAUTISTA CORREA

WILLIAM RAMIRO RIOS HENAO

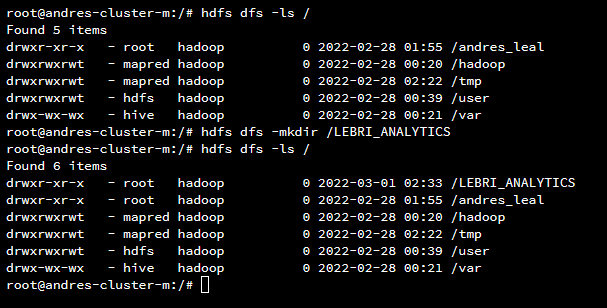
UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MÁSTER EN VISUALIZACION Y PROCESAMIENTO DE DATOS MASIVOS

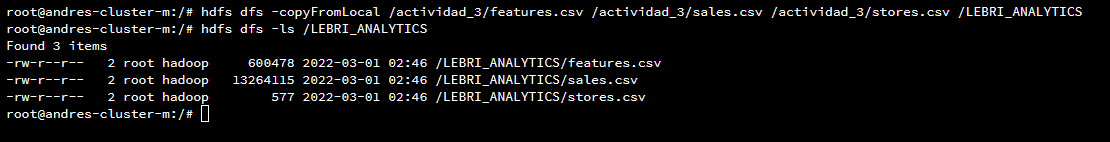
INGENIERÍA PARA EL PROCESADO MASIVO DE DATOS

MARZO, 2022

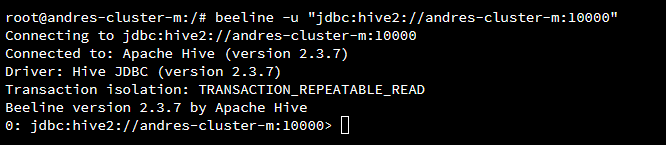
**Creando el folder del equipo:**



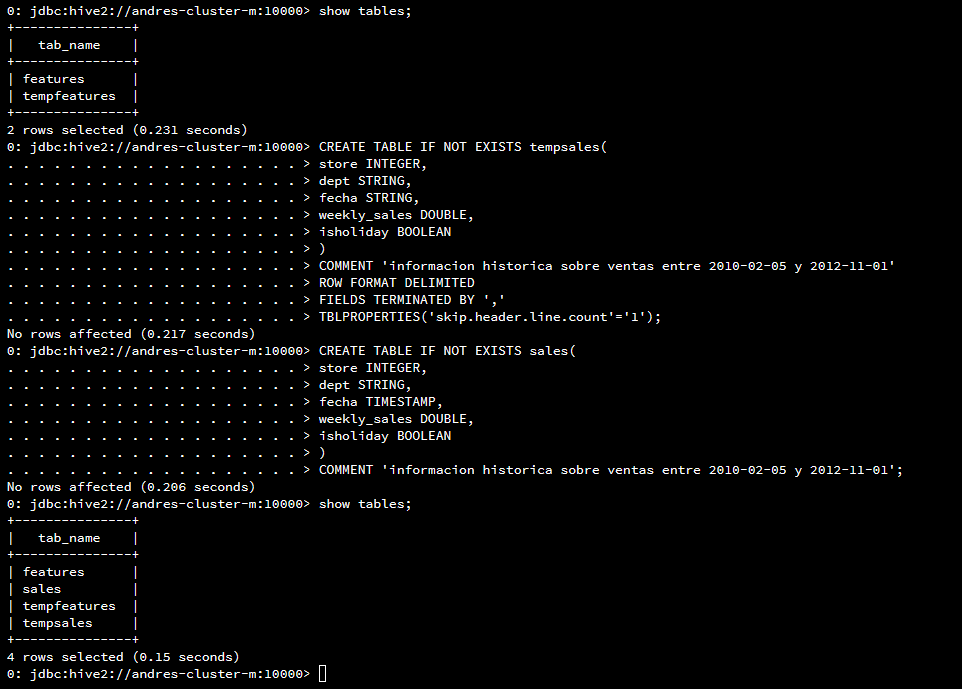
**Copiando los archivos a la carpeta en HDFS:**



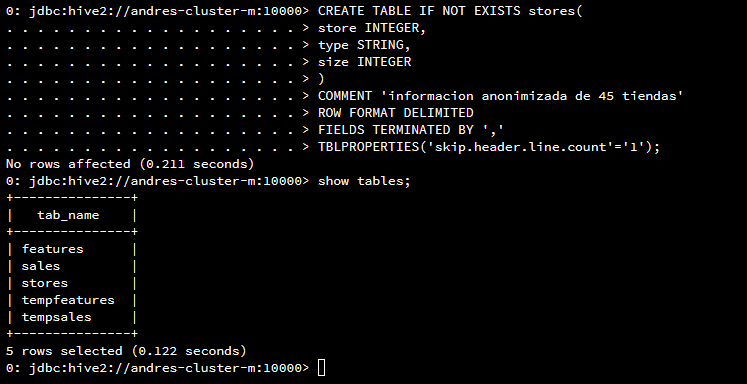
**Conectándonos a Hive:**

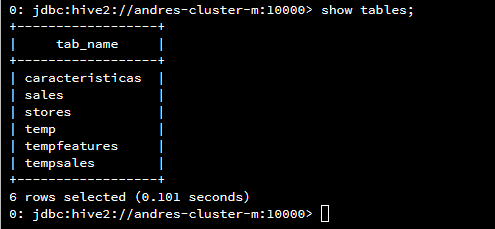


**Creando la tabla sales, como Managed table**. Debido al formato de fecha que tiene el archivo, creamos una tabla temporal que usaremos para castear la columna fecha al formato que queremos en la tabla features final, creamos ambas tablas, la temporal y la final:

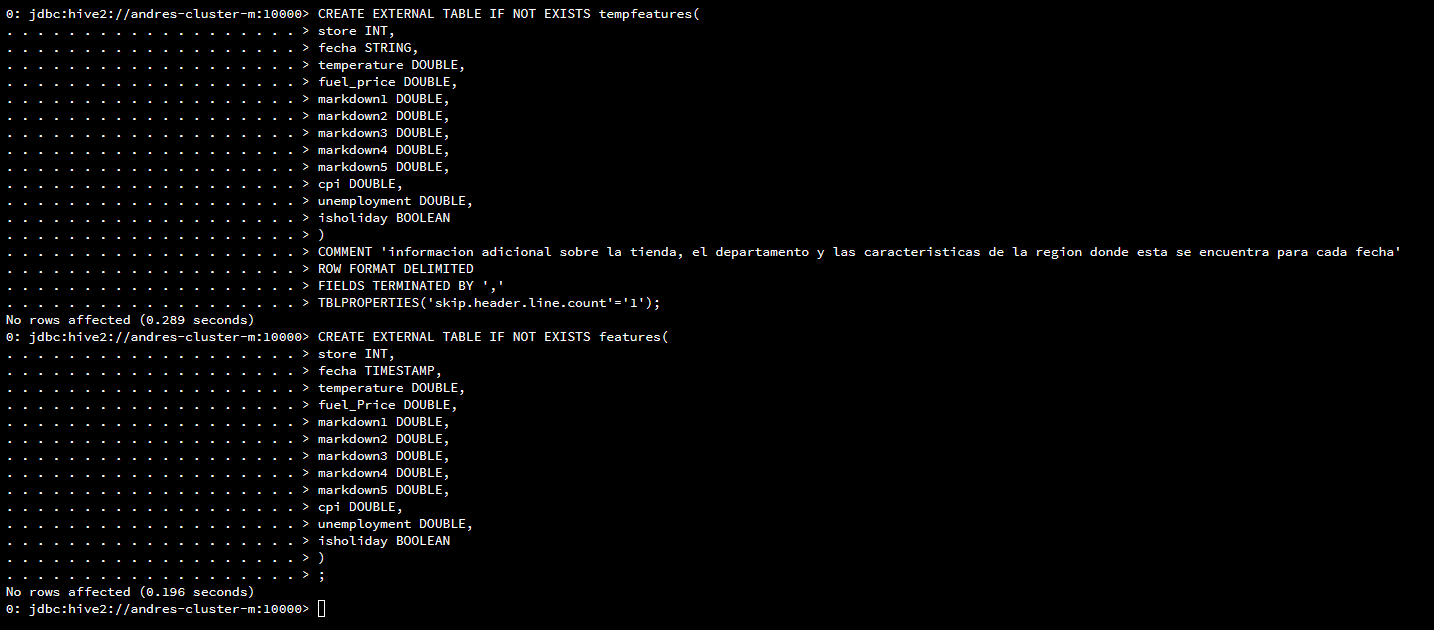


**Creando la tabla stores como Managed table también:**

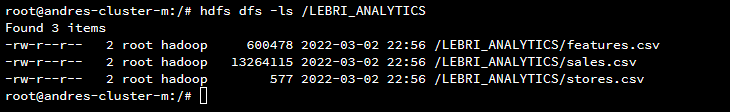




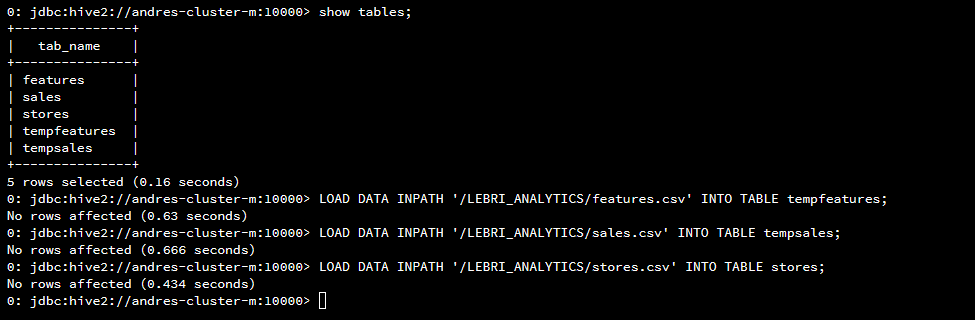
Creando la tabla features como tabla externa. Debido al formato de fecha que tiene el archivo, creamos una tabla temporal que usaremos para castear la columna fecha al formato que queremos en la tabla features final, creamos ambas tablas, la temporal y la final:



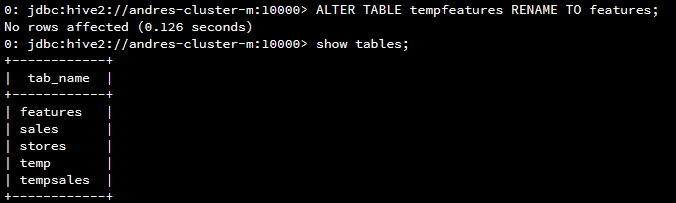
Previo al cargue de datos desde los CSVs hacia las tablas creadas:



**Cargando los datos desde los CSVs hacia las tablas temporales:**



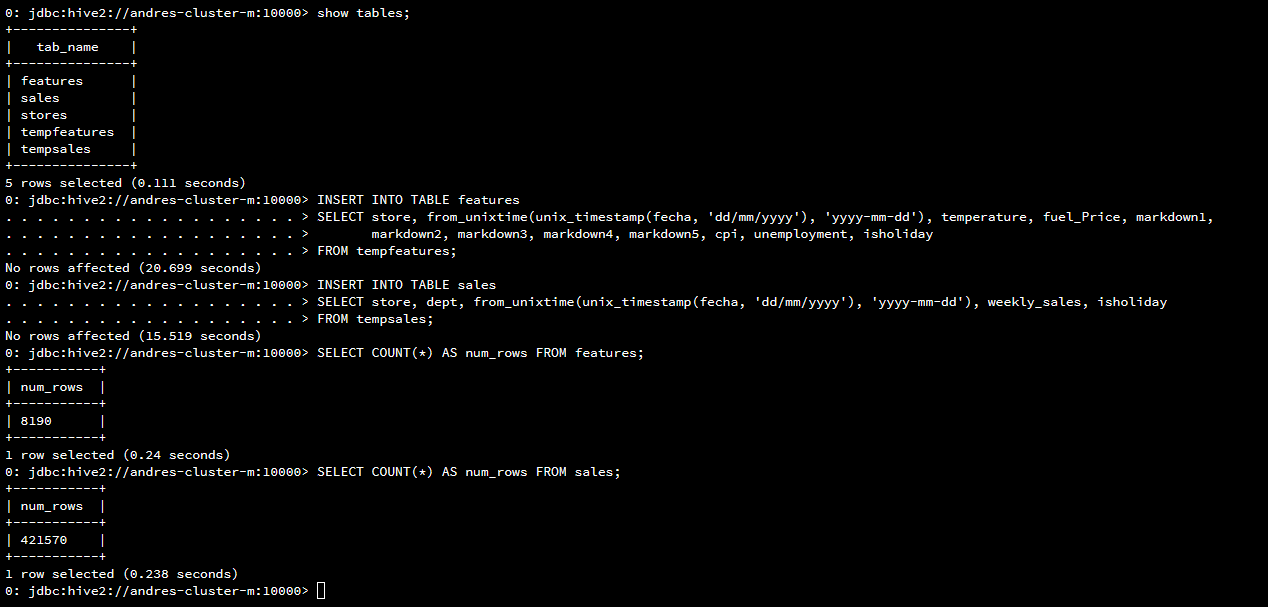
Dado que no se logra insertar los datos desde la tabla tempfeatures hacia una tabla features que sí tuviera la fecha como TIMESTAMP, se decide utilizar los datos de fecha como STRING, y simplemente cambiar el nombre de la tabla tempfeatures a features:



Carpeta HDFS luego del cargue:



Insertando los datos desde las tablas temporales hacia las finales, con el casteo del formato de fecha. Y contando las filas en ambas tablas, features y sales:

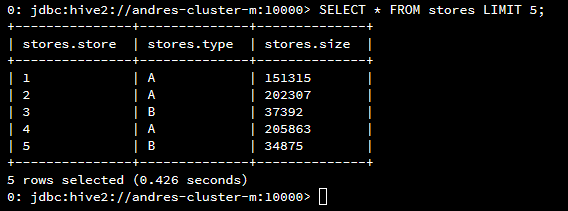


**Contando el número de filas en las tablas:**

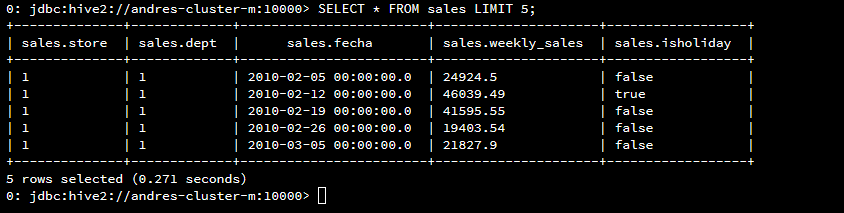
* Features: 8190 filas
* Sales: 421570 filas
* Stores: 45 filas

**Dando un vistazo a las primeras 5 filas de las tablas creadas:**

Stores



Sales



Features



**Revisando los valores máximo y mínimo de las variables numéricas de nuestras tablas, se encontraron los siguientes rangos:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tabla** | **Variable** | **Min** | **Max** |
| stores | size | 34875 | 219622 |
| sales | weekly\_sales | -4988.94 | 693099.36 |
| features | temperature | -7.29 | 101.95 |
| features | fuel\_Price | 2.472 | 4.468 |
| features | cpi | 126.064 | 228.976 |
| features | unemployment | 3.684 | 14.313 |

**Estudiando las diferentes categorías de las principales variables categóricas (Apartado 11. en el .txt):**

* Contamos en nuestros datos con información de 45 tiendas en total, identificadas con enteros del 1 al 45.
* Existen 3 tipos de tienda, A, B y C; perteneciendo 22, 17 y 6 tiendas, respectivamente, a cada categoría.
* Existen 81 tipos diferentes de departamentos, en las tiendas, identificados con números enteros entre 1 y 99. Los 3 tipos de departamento con mayor número de registros de ventas son los depts 1, 10 y 13, con 6435 cada uno; de igual forma, los 3 con menos registros son los depts 65, 39 y 43 con 143, 16 y 12, respectivamente.
* Las 3 tiendas con mayor número de registros de ventas son la 13, 10 y 4, con 10474, 10315 y 10272, respectivamente. Así mismo, las 3 con menos registros de venta son las tiendas 43, 33 y 36, con 6751, 6487 y 6222.

Entendiendo que en los datos sales se encuentran todos los registros del periodo de tiempo mencionado en el anexo (2010-02-05 y 2012-11-01), las conclusiones arriba mencionadas nos hablan entonces no solamente de cantidad de registros, nos hablan de los departamentos y tiendas con mayor y menor volumen de ventas, en dicho periodo de tiempo.

**Revisando valores anómalos (Apartado 12. en el .txt) encontramos la cantidad de valores nulos y negativos en variables continuas relevantes:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tabla** | **Variable** | **Nulos** | **Negativos** | **Filas de la tabla** |
| sales | weekly\_sales | 0 | 1285 | 421570 |
| features | temperature | 0 | 4 | 8190 |
| features | fuel\_Price | 0 | 0 | 8190 |
| features | cpi | 585 | 0 | 8190 |
| features | unemployment | 585 | 0 | 8190 |

**Creando vistas con agregaciones que generen valor (Apartado 14. en el .txt):**

Proponemos un par de vistas calculando la media y desviación estándar de las ventas (aquellas mayores a 0), que nos permiten agregar un análisis más a la búsqueda de valores anómalos:

Adicional a registros con nulos y negativos, encontramos que existen registros con valores anormalmente altos, respecto a todo el conjunto, en las ventas semanales, 9131 ventas superaron por 3 desviaciones estándar la media del conjunto. También se encontró en esta columna particular, con Media = 16030.32 y Desv. Estándar = 22728.47, que presente una variabilidad altísima, la desviación del conjunto es cerca de 1.3 veces el valor de la media misma, de modo que la existencia de extremos en esta columna es de esperar.

**Creando una vista que genere valor (Apartado 13. en el .txt):**

Pensando en la importancia que podrían tener análisis de ventas por tipo de tienda, y potenciales análisis relacionados con tamaño de las distintas tiendas (que seguro podrá relacionarse con costos asociados, o relacionarse con el flujo de clientes que la frecuentan), vemos relevante cruzar la información disponible en las tablas stores y sales.

**Creando una vista de agrupación y agregación que agregue valor (Apartado 15. en el .txt):**

Encontramos relevante poder contrastar volumen de ventas y cantidad de dinero generado por las tiendas, buscando entender si efectivamente las tiendas con más ventas son aquellas que más revenue están aportando a la compañía.

De esta vista, se puede ver que es la tienda 20 la que más revenue ha generado, sin ser la tienda de mayor tamaño, y sin ser la tienda con mayor volumen de ventas. Así mismo, vemos que muy cerca se encuentra la tienda 4 (mayor en tamaño), en el segundo lugar en cantidad de revenue generado, teniendo incluso una mayor cantidad de ventas.

Así mismo, esta vista nos permitiría realizar análisis sobre las tiendas de menor revenue, y si se llegan a obtener datos claros de costos, realizar un análisis integrado de volumen de ventas – ingresos – costos, entre todas las tiendas y tipos de tienda que tiene la compañía.

