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:

```
root@andres-cluster-m:/# hdfs dfs -ls /
Found 5 items
drwxr-xr-x   - root   hadoop          0 2022-02-28 01:55 /andres_leal
drwxrwxrwt   - mapred hadoop          0 2022-02-28 00:20 /hadoop
drwxrwxrwt   - mapred hadoop          0 2022-02-28 02:22 /tmp
drwxrwxrwt   - hdfs   hadoop          0 2022-02-28 00:39 /user
drwx-wx-wx   - hive   hadoop          0 2022-02-28 00:21 /var
root@andres-cluster-m:/# hdfs dfs -mkdir /LEBRI_ANALYTICS
root@andres-cluster-m:/# hdfs dfs -ls /
Found 6 items
drwxr-xr-x   - root   hadoop          0 2022-03-01 02:33 /LEBRI_ANALYTICS
drwxr-xr-x   - root   hadoop          0 2022-02-28 01:55 /andres_leal
drwxrwxrwt   - mapred hadoop          0 2022-02-28 00:20 /hadoop
drwxrwxrwt   - mapred hadoop          0 2022-02-28 02:22 /tmp
drwxrwxrwt   - hdfs   hadoop          0 2022-02-28 00:39 /user
drwx-wx-wx   - hive   hadoop          0 2022-02-28 00:21 /var
root@andres-cluster-m:/#
```

Copiando los archivos a la carpeta en HDFS:

```
root@andres-cluster-m:/# hdfs dfs -copyFromLocal /actividad_3/features.csv /actividad_3/sales.csv /actividad_3/stores.csv /LEBRI_ANALYTICS
root@andres-cluster-m:/# hdfs dfs -ls /LEBRI_ANALYTICS
Found 3 items
-rw-r--r--   2 root hadoop     600478 2022-03-01 02:46 /LEBRI_ANALYTICS/features.csv
-rw-r--r--   2 root hadoop   13264115 2022-03-01 02:46 /LEBRI_ANALYTICS/sales.csv
-rw-r--r--   2 root hadoop        577 2022-03-01 02:46 /LEBRI_ANALYTICS/stores.csv
root@andres-cluster-m:/#
```

Conectándonos a Hive:

```
root@andres-cluster-m:/# beeline -u "jdbc:hive2://andres-cluster-m:10000"
Connecting to jdbc:hive2://andres-cluster-m:10000
Connected to: Apache Hive (version 2.3.7)
Driver: Hive JDBC (version 2.3.7)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.7 by Apache Hive
0: jdbc:hive2://andres-cluster-m:10000>
```

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:

```
0: jdbc:hive2://andres-cluster-m:10000> show tables;
+---------------+
|   tab_name    |
+---------------+
| features      |
| tempfeatures  |
+---------------+
2 rows selected (0.231 seconds)
0: jdbc:hive2://andres-cluster-m:10000> CREATE TABLE IF NOT EXISTS tempsales(
. . . . . . . . . . . . . . . . . . . > store INTEGER,
. . . . . . . . . . . . . . . . . . . > dept STRING,
. . . . . . . . . . . . . . . . . . . > fecha STRING,
. . . . . . . . . . . . . . . . . . . > weekly_sales DOUBLE,
. . . . . . . . . . . . . . . . . . . > isholiday BOOLEAN
. . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . > COMMENT 'informacion historica sobre ventas entre 2010-02-05 y 2012-11-01'
. . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED
. . . . . . . . . . . . . . . . . . . > FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . > TBLPROPERTIES('skip.header.line.count'='1');
No rows affected (0.217 seconds)
0: jdbc:hive2://andres-cluster-m:10000> CREATE TABLE IF NOT EXISTS sales(
. . . . . . . . . . . . . . . . . . . > store INTEGER,
. . . . . . . . . . . . . . . . . . . > dept STRING,
. . . . . . . . . . . . . . . . . . . > fecha TIMESTAMP,
. . . . . . . . . . . . . . . . . . . > weekly_sales DOUBLE,
. . . . . . . . . . . . . . . . . . . > isholiday BOOLEAN
. . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . > COMMENT 'informacion historica sobre ventas entre 2010-02-05 y 2012-11-01';
No rows affected (0.206 seconds)
```

Creando la tabla stores como Managed table también:

```
0: jdbc:hive2://andres-cluster-m:10000> CREATE TABLE IF NOT EXISTS stores(
. . . . . . . . . . . . . . . . . . . > store INTEGER,
. . . . . . . . . . . . . . . . . . . > type STRING,
. . . . . . . . . . . . . . . . . . . > size INTEGER
. . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . > COMMENT 'informacion anonimizada de 45 tiendas'
. . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED
. . . . . . . . . . . . . . . . . . . > FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . > TBLPROPERTIES('skip.header.line.count'='1');
No rows affected (0.211 seconds)
```

```
0: jdbc:hive2://andres-cluster-m:10000> show tables;
+------------------+
|     tab_name     |
+------------------+
| caracteristicas  |
| sales            |
| stores           |
| temp             |
| tempfeatures     |
| tempsales        |
+------------------+
6 rows selected (0.101 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```

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:

```
0: jdbc:hive2://andres-cluster-m:10000> CREATE EXTERNAL TABLE IF NOT EXISTS tempfeatures(
. . . . . . . . . . . . . . . . . . . > store INT,
. . . . . . . . . . . . . . . . . . . > fecha STRING,
. . . . . . . . . . . . . . . . . . . > temperature DOUBLE,
. . . . . . . . . . . . . . . . . . . > fuel_price DOUBLE,
. . . . . . . . . . . . . . . . . . . > markdown1 DOUBLE,
. . . . . . . . . . . . . . . . . . . > markdown2 DOUBLE,
. . . . . . . . . . . . . . . . . . . > markdown3 DOUBLE,
. . . . . . . . . . . . . . . . . . . > markdown4 DOUBLE,
. . . . . . . . . . . . . . . . . . . > markdown5 DOUBLE,
. . . . . . . . . . . . . . . . . . . > cpi DOUBLE,
. . . . . . . . . . . . . . . . . . . > unemployment DOUBLE,
. . . . . . . . . . . . . . . . . . . > isholiday BOOLEAN
. . . . . . . . . . . . . . . . . . . > )
. . . . . . . . . . . . . . . . . . . > COMMENT 'informacion adicional sobre la tienda, el departamento y las caracteristicas de la region donde esta se encuentra para cada fecha'
. . . . . . . . . . . . . . . . . . . > ROW FORMAT DELIMITED
. . . . . . . . . . . . . . . . . . . > FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . > TBLPROPERTIES('skip.header.line.count'='1');
No rows affected (0.289 seconds)
```

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

```
root@andres-cluster-m:/# hdfs dfs -ls /LEBRI_ANALYTICS
Found 3 items
-rw-r--r--   2 root hadoop     600478 2022-03-02 22:56 /LEBRI_ANALYTICS/features.csv
-rw-r--r--   2 root hadoop   13264115 2022-03-02 22:56 /LEBRI_ANALYTICS/sales.csv
-rw-r--r--   2 root hadoop        577 2022-03-02 22:56 /LEBRI_ANALYTICS/stores.csv
root@andres-cluster-m:/#
```

Cargando los datos desde los CSVs hacia las tablas temporales:

```
0: jdbc:hive2://andres-cluster-m:10000> show tables;
+--------------+
|   tab_name   |
+--------------+
| features     |
| sales        |
| stores       |
| tempfeatures |
| tempsales    |
+--------------+
5 rows selected (0.16 seconds)
0: jdbc:hive2://andres-cluster-m:10000> LOAD DATA INPATH '/LEBRI_ANALYTICS/features.csv' INTO TABLE tempfeatures;
No rows affected (0.63 seconds)
0: jdbc:hive2://andres-cluster-m:10000> LOAD DATA INPATH '/LEBRI_ANALYTICS/sales.csv' INTO TABLE tempsales;
No rows affected (0.666 seconds)
0: jdbc:hive2://andres-cluster-m:10000> LOAD DATA INPATH '/LEBRI_ANALYTICS/stores.csv' INTO TABLE stores;
No rows affected (0.434 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```

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:

```
0: jdbc:hive2://andres-cluster-m:10000> ALTER TABLE tempfeatures RENAME TO features;
No rows affected (0.126 seconds)
0: jdbc:hive2://andres-cluster-m:10000> show tables;
+------------+
|  tab_name  |
+------------+
| features   |
| sales      |
| stores     |
| temp       |
| tempsales  |
+------------+
```

Carpeta HDFS luego del cargue:

```
root@andres-cluster-m:/# hdfs dfs -ls /LEBRI_ANALYTICS
root@andres-cluster-m:/#
```

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:

```
0: jdbc:hive2://andres-cluster-m:10000> INSERT INTO TABLE sales
. . . . . . . . . . . . . . . . . . . > SELECT store, dept, from_unixtime(unix_timestamp(fecha, 'dd/mm/yyyy'), 'yyyy-mm-dd'), weekly_sales, isholiday
. . . . . . . . . . . . . . . . . . . > FROM tempsales;
No rows affected (15.519 seconds)
```

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

```
0: jdbc:hive2://andres-cluster-m:10000> SELECT * FROM stores LIMIT 5;
+----------------+---------------+---------------+
| stores.store   | stores.type   | stores.size   |
+----------------+---------------+---------------+
| 1              | A             | 151315        |
| 2              | A             | 202307        |
| 3              | B             | 37392         |
| 4              | A             | 205863        |
| 5              | B             | 34875         |
+----------------+---------------+---------------+
5 rows selected (0.426 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```

Sales

```
0: jdbc:hive2://andres-cluster-m:10000> SELECT * FROM sales LIMIT 5;
+---------------+--------------+--------------------------+----------------------+--------------------+
| sales.store   | sales.dept   |       sales.fecha        | sales.weekly_sales   | sales.isholiday    |
+---------------+--------------+--------------------------+----------------------+--------------------+
| 1             | 1            | 2010-02-05 00:00:00.0    | 24924.5              | false              |
| 1             | 1            | 2010-02-12 00:00:00.0    | 46039.49             | true               |
| 1             | 1            | 2010-02-19 00:00:00.0    | 41595.55             | false              |
| 1             | 1            | 2010-02-26 00:00:00.0    | 19403.54             | false              |
| 1             | 1            | 2010-03-05 00:00:00.0    | 21827.9              | false              |
+---------------+--------------+--------------------------+----------------------+--------------------+
5 rows selected (0.271 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```

Features

```
0: jdbc:hive2://andres-cluster-m:10000> SELECT * FROM features LIMIT 5;
+-----------------+-----------------+-----------------------+-----------------------+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+-----
----------+-----------------------+--------------------+
| features.store  | features.fecha  | features.temperature  | features.fuel_price   | features.markdown1  | features.markdown2  | features.markdown3  | features.markdown4  | features.markdown5  | features.markdown6  | fea
tures.cpi  | features.unemployment  | features.isholiday  |
+-----------------+-----------------+-----------------------+-----------------------+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+-----
----------+-----------------------+--------------------+
| 1               | 05/02/2010      | 42.31                 | 2.572                 | NULL                | NULL                | NULL                | NULL                | NULL                | 211
.0963582   | 8.106                 | false              |
| 1               | 12/02/2010      | 38.51                 | 2.548                 | NULL                | NULL                | NULL                | NULL                | NULL                | 211
.2421698   | 8.106                 | true               |
| 1               | 19/02/2010      | 39.93                 | 2.514                 | NULL                | NULL                | NULL                | NULL                | NULL                | 211
.2891429   | 8.106                 | false              |
| 1               | 26/02/2010      | 46.63                 | 2.561                 | NULL                | NULL                | NULL                | NULL                | NULL                | 211
.3196429   | 8.106                 | false              |
| 1               | 05/03/2010      | 46.5                  | 2.625                 | NULL                | NULL                | NULL                | NULL                | NULL                | 211
.3501429   | 8.106                 | false              |
+-----------------+-----------------+-----------------------+-----------------------+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+-----
----------+-----------------------+--------------------+
5 rows selected (0.2 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```

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.

```
0: jdbc:hive2://andres-cluster-m:10000> DROP VIEW IF EXISTS ventas_volumen_ingresos;
No rows affected (0.064 seconds)
0: jdbc:hive2://andres-cluster-m:10000> CREATE VIEW ventas_volumen_ingresos AS
. . . . . . . . . . . . . . . . . . . . > SELECT s2.store, s2.size, COUNT(*) as num_ventas, SUM(s1.weekly_sales) as ingresos
. . . . . . . . . . . . . . . . . . . . > FROM sales as s1
. . . . . . . . . . . . . . . . . . . . > INNER JOIN stores as s2
. . . . . . . . . . . . . . . . . . . . > ON s1.store = s2.store
. . . . . . . . . . . . . . . . . . . . > GROUP BY s2.store, s2.size
. . . . . . . . . . . . . . . . . . . . > ORDER BY ingresos DESC;
No rows affected (0.404 seconds)
0: jdbc:hive2://andres-cluster-m:10000> SELECT * FROM ventas_volumen_ingresos LIMIT 5;
+-------------------------------+------------------------------+-----------------------------------+---------------------------------+
| ventas_volumen_ingresos.store | ventas_volumen_ingresos.size | ventas_volumen_ingresos.num_ventas | ventas_volumen_ingresos.ingresos |
+-------------------------------+------------------------------+-----------------------------------+---------------------------------+
| 20                            | 203742                       | 10214                             | 3.0139779245999974E8            |
| 4                             | 205863                       | 10272                             | 2.9954395337999994E8            |
| 14                            | 200898                       | 10040                             | 2.889999113399991E8             |
| 13                            | 219622                       | 10474                             | 2.865177037999985E8             |
| 2                             | 202307                       | 10238                             | 2.753824409800004E8             |
+-------------------------------+------------------------------+-----------------------------------+---------------------------------+
5 rows selected (27.218 seconds)
0: jdbc:hive2://andres-cluster-m:10000> show views;
+--------------------------+
|         tab_name         |
+--------------------------+
| sales_avg                |
| sales_std                |
| stores_types_sales       |
| ventas_volumen_ingresos  |
+--------------------------+
4 rows selected (0.195 seconds)
0: jdbc:hive2://andres-cluster-m:10000>
```