3. Hive caso de estudio (Parte 2)

jueves, 30 de abril de 2020 16:25

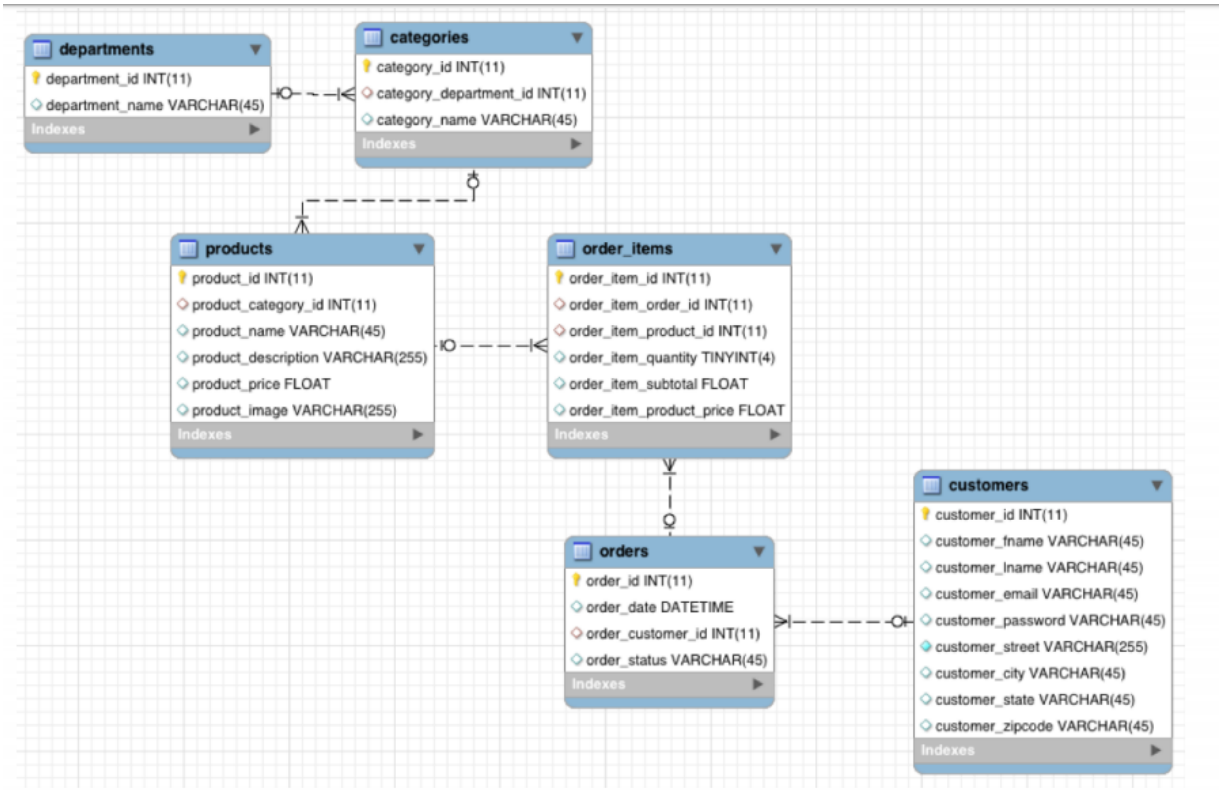
La empresa

Es una tienda de venta de artículos deportivos, que tiene tiendas físicas/presenciales, pero que también tiene sitio de ventas por web.
Ej: Nike, Adidas, Sportline, Foot Locker, etc.

Pregunta de Negocio

- Son los productos más visitados en el sitio web los más vendidos?
- Son los productos más visitados los que hacen parte de los de mayor rentabilidad?

Modelo de datos relacionales



Solución

Actualizamos los datos en hdfs (Ambari)

```
bash-4.2$ cd datasets/
bash-4.2$ hdfs dfs -copyFromLocal * hdfs:///user/jscaicedom/datasets/
copyFromLocal: `hdfs:///user/jscaicedom/datasets/airlines.csv': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/all-news/url.txt': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg/gutenberg-small-en.zip': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg/gutenberg-small-es.zip': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg/gutenberg-txt-es.zip-url.txt': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___LincolnLetters.txt':
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___LincolnsFirstInaugur
exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___LincolnsGettysburgAd
-1863.txt': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___LincolnsInauguralsAd
ctions.txt': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___LincolnsSecondInaugu
 exists
```

```
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___SpeechesandLettersof
65.txt': File exists
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___StateoftheUnionAddre
s
copyFromLocal: `hdfs:///user/jscaicedom/datasets/gutenberg-small/AbrahamLincoln___TheEmancipationProcl
ists
```

- **Cargar los datos operacionales**

Primero creamos la base de datos retail_db en rds
Hacemos conexión con el cliente mysql a la base de datos

```
[ec2-user@ip-172-31-81-128 ~]$ mysql -u admin -p -h database.cy5rftlokiyu.us-eas
t-1.rds.amazonaws.com
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 648
Server version: 5.7.22-log Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> CREATE DATABASE retail_db;
Query OK, 1 row affected (0.01 sec)
```

Hacemos las configuraciones de
retail_db-ddl.sql
retail_db-data.sql

Descargando el repositorio y pasando la configuración

```
[ec2-user@ip-172-31-81-128 ~]$ mysql -u admin -p -h database.cy5rftlokiyu.us-east-1.rds.amazonaws.com retail_db < BigDataLab/big
a/rdbms/retail_db-data.sql
Enter password:
[ec2-user@ip-172-31-81-128 ~]$ mysql -u admin -p -h database.cy5rftlokiyu.us-east-1.rds.amazonaws.com
```

Y revisamos que se hayan creado

```
MySQL [(none)]> use retail_db
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [retail_db]> show tables;
+--------------------+
| Tables_in_retail_db |
+--------------------+
| categories         |
| customers          |
| departments        |
| order_items        |
| orders             |
| products           |
+--------------------+
6 rows in set (0.00 sec)
```

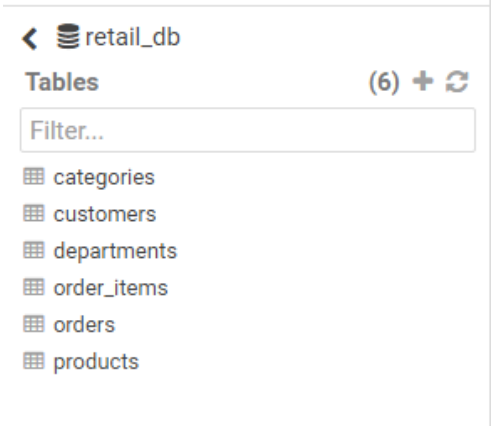
Creamos la base de datos en hue

```
1 CREATE DATABASE retail_db;
```

Transferimos datos vía sqoop al datalake en hive

```
[hadoop@ip-172-31-90-45 ~]$ sqoop import-all-tables --connect jdbc:mysql://database.cy5rft
lokiyu.us-east-1.rds.amazonaws.com:3306/retail_db --username=admin --password=Estella5* --
hive-database retail_db --hive-overwrite --hive-import --warehouse-dir=/tmp/retail_dbtmp -
m 1 --mysql-delimiters
```

Verificamos



- **Procesar**

-- CATEGORIAS MÁS POPULARES DE PRODUCTOS

26.63s Database retail_db

```
SELECT c.category_name, count(order_item_quantity) as count
FROM order_items oi
inner join products p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10
```

	c.category_name	count
1	Cleats	24551
2	Men\'s Footwear	22246
3	Women\'s Apparel	21035
4	Indoor/Outdoor Games	19298
5	Fishing	17325
6	Water Sports	15540
7	Camping & Hiking	13729
8	Cardio Equipment	12487
9	Shop By Sport	10984
10	Electronics	3156

-- TOP 10 DE PRODUCTOS QUE GENERAN GANANCIAS

19.55s Database retail_db Type tex

```
SELECT p.product_id, p.product_name, r.revenue
FROM products p inner join
(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
from order_items oi inner join orders o
on oi.order_item_order_id = o.order_id
where o.order_status <> 'CANCELED'
and o.order_status <> 'SUSPECTED_FRAUD'
group by order_item_product_id) r
on p.product_id = r.order_item_product_id
order by r.revenue desc
limit 10
```

	p.product_id	p.product_name	r.revenue
1	1004	Field & Stream Sportsman 16 Gun Fire Safe	6637668.282318115

2	365	Perfect Fitness Perfect Rip Deck	4233794.3682899475
3	957	Diamondback Women\'s Serene Classic Comfort Bi	3946837.004547119
4	191	Nike Men\'s Free 5.0+ Running Shoe	3507549.2067337036
5	502	Nike Men\'s Dri-FIT Victory Golf Polo	3011600
6	1073	Pelican Sunstream 100 Kayak	2967851.6815185547
7	1014	O\'Brien Men\'s Neoprene Life Vest	2765543.314743042
8	403	Nike Men\'s CJ Elite 2 TD Football Cleat	2763977.4868011475
9	627	Under Armour Girls\' Toddler Spine Surge Runni	1214896.220287323
10	565	adidas Youth Germany Black/Red Away Match Soc	63490

-- SUBIR LOS LOGS AL HDFS: (Creamos la tabla con los datos en S3)

6.91s Database retail_db Type text

```
CREATE EXTERNAL TABLE tmp_access_logs (
        ip STRING,
        fecha STRING,
        method STRING,
        url STRING,
        http_version STRING,
        code1 STRING,
        code2 STRING,
        dash STRING,
        user_agent STRING)
    ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
    WITH SERDEPROPERTIES (
        'input.regex' = '([^ ]*) - - \\[([^\\]]*)\\] "([^\ ]*) ([^\ ]*) ([^\ ]*)" (\\d*) (\\d*) "([
        'output.format.string' = "%1$$s %2$$s %3$$s %4$$s %5$$s %6$$s %7$$s %8$$s %9$$s")
    LOCATION 's3://jscaicedomb/datasets/retail_logs/';
```

```
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20200503042244_26faea22-5b7e-421a-8d15-70c652b67
b8f); Time taken: 3.046 seconds
INFO  : OK
```

-- CREAR DIRECTORIO PARA TABLA EXTERNA CON ETL

Inicio / user / admin / **warehouse**

Name
↑
.
access_logs_etl

0.90s Database retail_db

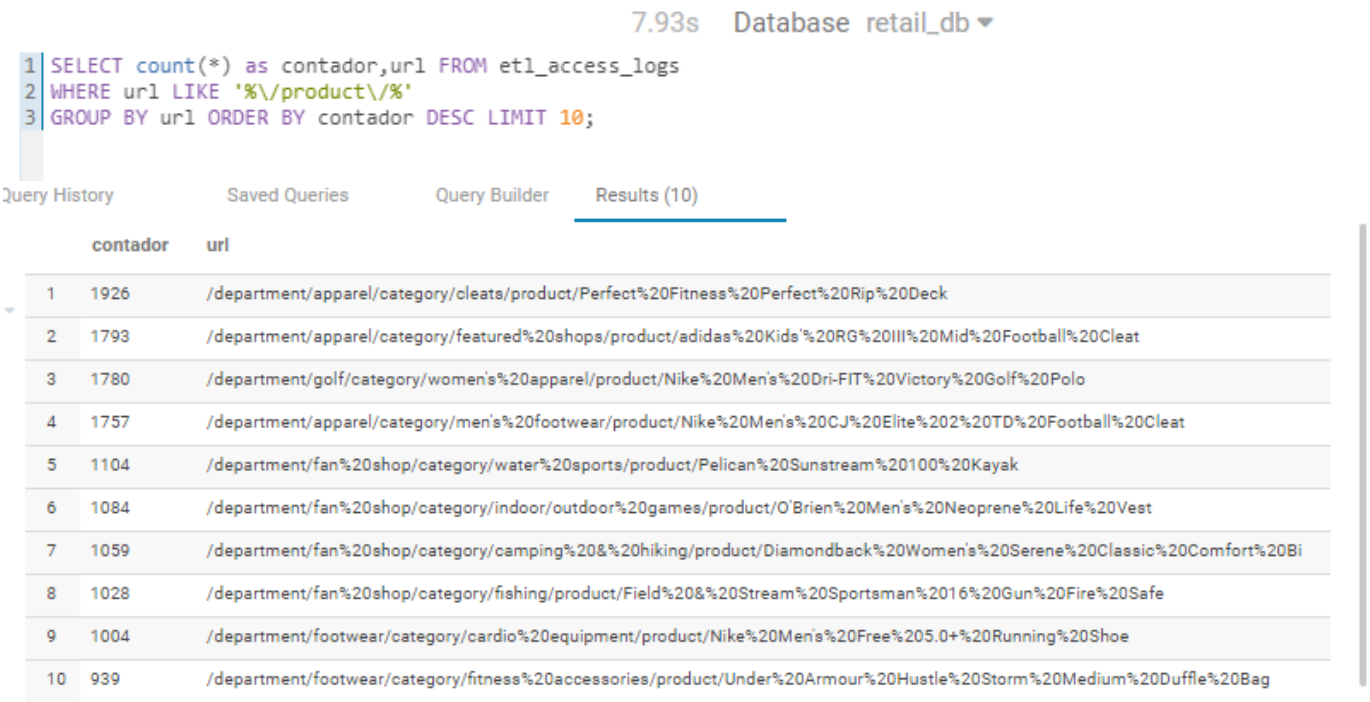
```
CREATE EXTERNAL TABLE etl_access_logs (
        ip STRING,
        fecha STRING,
        method STRING,
        url STRING,
        http_version STRING,
        code1 STRING,
        code2 STRING,
        dash STRING,
        user_agent STRING)
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    LOCATION '/user/admin/warehouse/access_logs_etl/';
```

- **Procesar ETL**

Ejecutamos lo siguiente en el clúster para habilitar permisos

```
$ hdfs dfs -chown -R admin:hdfs /user/admin/
```

--- LOS 10 PRODUCTOS MÁS VISITADOS



	contador	url
1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%20Medium%20Duffle%20Bag

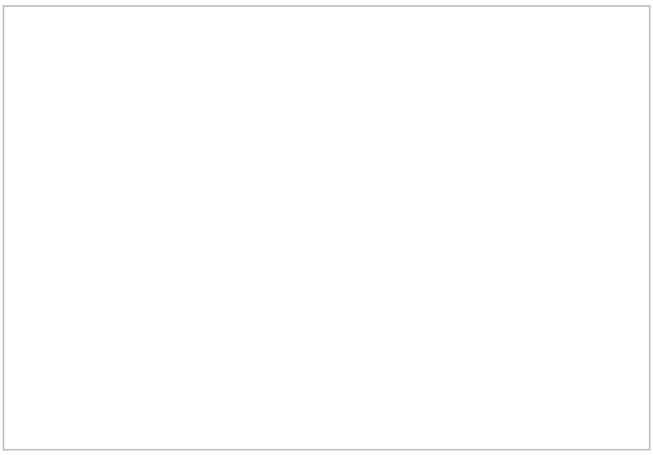
--- LOS 10 PRODUCTOS MÁS VENDIDOS (Código en repositorio github)

p.product_name	count

1	Perfect Fitness Perfect Rip Deck	24515
2	Nike Men\'s CJ Elite 2 TD Football Cleat	22246
3	Nike Men\'s Dri-FIT Victory Golf Polo	21035
4	O\'Brien Men\'s Neoprene Life Vest	19298
5	Field & Stream Sportsman 16 Gun Fire Safe	17325
6	Pelican Sunstream 100 Kayak	15500
7	Diamondback Women\'s Serene Classic Comfort Bi	13729
8	Nike Men\'s Free 5.0+ Running Shoe	12169
9	Under Armour Girls\' Toddler Spine Surge Runni	10617
10	Nike Men\'s Comfort 2 Slide	328

- **Son los productos más visitados en el sitio web los más vendidos?**

  9 productos más vendidos fueron también los más visitados. (Comparé los 2 resultados productos más vendidos y más visitados)



- **Son los productos más visitados los que hacen parte de los de mayor rentabilidad?**

  8 productos más visitados fueron también los de mayor rentabilidad (Comparé los resultados de los más visitados y los que generan más ganancias)