

Tutorial Big Data CLEI 2019



Ciudad de Panamá, Panamá

Por: Edwin Montoya
emontoya@eafit.edu.co

Hive y Sqoop

Caso de estudio

- Amazon EMR –
- Amazon RDS / MySQL –

Edwin Montoya – emontoya@eafit.edu.co

2019

Datos y scripts en:

<https://github.com/edwinmontoya/tutorialbigdataclei2019.git>

Datos específicos:

Github:/datasets/retail_logs/

Amazon S3: s3://emontoyapublic/datasets/retail_logs/

Amazon RDS/MySQL (más adelante se explica)

Scripts:

Github:/02-hive/case_retail_db* (para Amazon EMR)
/rdbms/retail_db* (para Amazon RDS)

Pasos

La empresa

Define las PREGUNTAS DE NEGOCIO

Identificar las fuentes de datos

ETL

Procesamiento - Analítica

Resultado -> Aplicación

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?

Identificar los datos

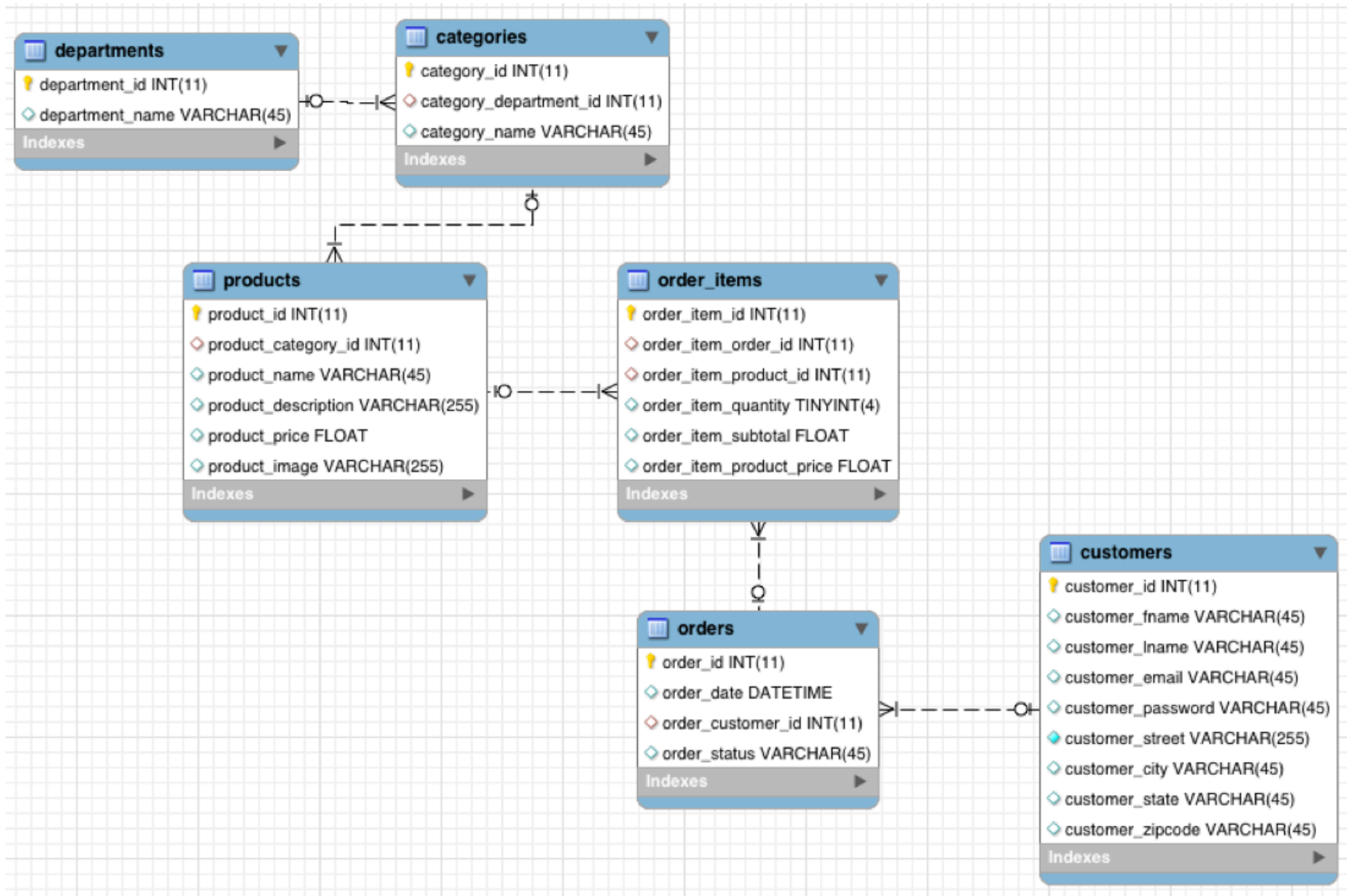
¿Dónde están los Datos?

La primera línea de trabajo, son los datos operativos de ventas reales de la compañía de los productos.

Estos datos se encuentran almacenados en RDBMBS, en este caso MySQL. Luego, hay que entender el modelo de datos.

Logs de navegación del sitio principal de comercio electrónico.

Modelo de datos relacionales



Carga de datos al DHW

Base de datos Operacionales

Debemos ingestar los datos operacionales de MySQL en HDFS y Hive.

Para ello, utilizamos **HUE o sqoop desde Shell**

Info de los datos operacionales:

- **Base de datos:** Amazon RDS/MySQL
- **IP_server_DB:** database-1.cj1yhistqein.us-east-2.rds.amazonaws.com
- **Database:** retail_db
- **Username:** retail_dba
- **Password:** retail_dba

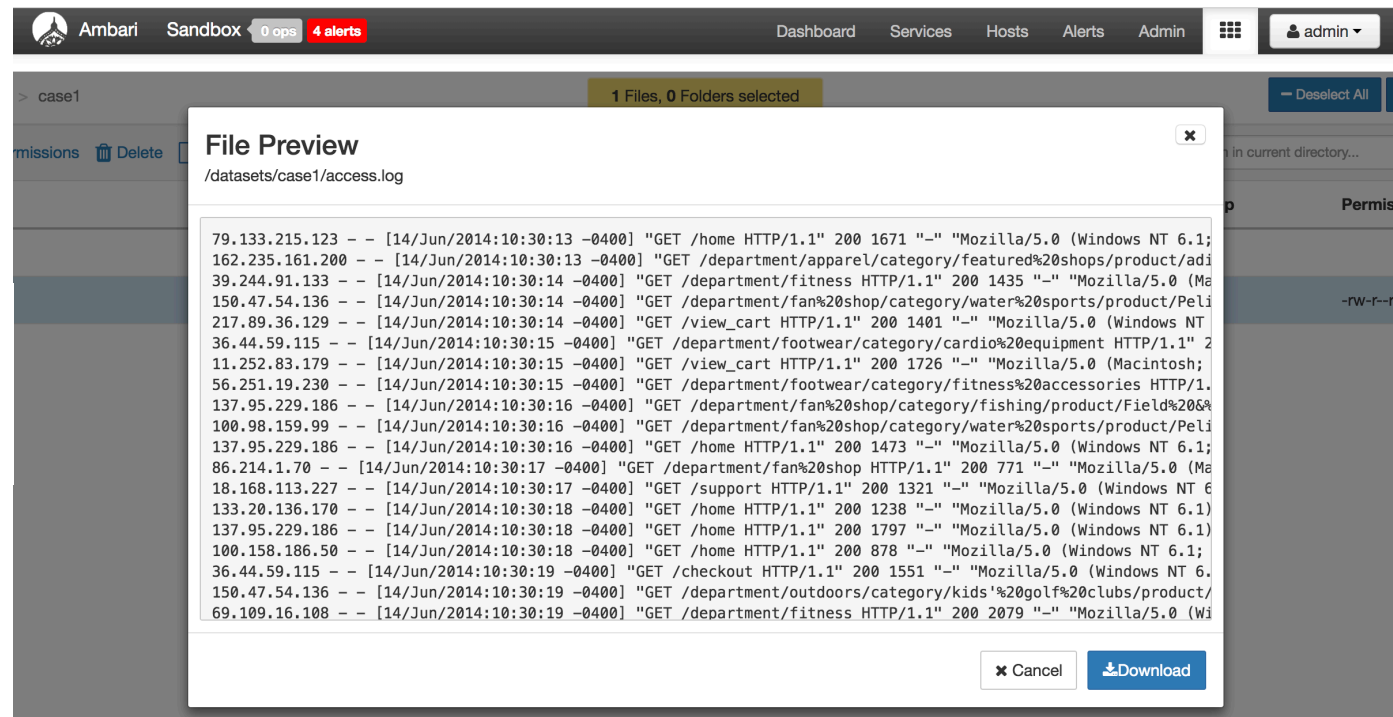
TAREA: Carga de datos al DHW

Logs de navegación

Datasets

File Preview

/datasets/case1/access.log



The screenshot shows the Ambari web interface. At the top, there's a navigation bar with 'Ambari', 'Sandbox', '0 ops', and '4 alerts'. Below this, a 'File Preview' modal window is open, displaying the contents of the file '/datasets/case1/access.log'. The log entries are formatted as follows:

```
79.133.215.123 - - [14/Jun/2014:10:30:13 -0400] "GET /home HTTP/1.1" 200 1671 "-" "Mozilla/5.0 (Windows NT 6.1; 162.235.161.200 - - [14/Jun/2014:10:30:13 -0400] "GET /department/apparel/category/featured%20shops/product/adi 39.244.91.133 - - [14/Jun/2014:10:30:14 -0400] "GET /department/fitness HTTP/1.1" 200 1435 "-" "Mozilla/5.0 (Ma 150.47.54.136 - - [14/Jun/2014:10:30:14 -0400] "GET /department/fan%20shop/category/water%20sports/product/Peli 217.89.36.129 - - [14/Jun/2014:10:30:14 -0400] "GET /view_cart HTTP/1.1" 200 1401 "-" "Mozilla/5.0 (Windows NT 36.44.59.115 - - [14/Jun/2014:10:30:15 -0400] "GET /department/footwear/category/cardio%20equipment HTTP/1.1" 2 11.252.83.179 - - [14/Jun/2014:10:30:15 -0400] "GET /view_cart HTTP/1.1" 200 1726 "-" "Mozilla/5.0 (Macintosh; 56.251.19.230 - - [14/Jun/2014:10:30:15 -0400] "GET /department/footwear/category/fitness%20accessories HTTP/1. 137.95.229.186 - - [14/Jun/2014:10:30:16 -0400] "GET /department/fan%20shop/category/fishing/product/Field%20& 100.98.159.99 - - [14/Jun/2014:10:30:16 -0400] "GET /department/fan%20shop/category/water%20sports/product/Peli 137.95.229.186 - - [14/Jun/2014:10:30:16 -0400] "GET /home HTTP/1.1" 200 1473 "-" "Mozilla/5.0 (Windows NT 6.1; 86.214.1.70 - - [14/Jun/2014:10:30:17 -0400] "GET /department/fan%20shop HTTP/1.1" 200 771 "-" "Mozilla/5.0 (Ma 18.168.113.227 - - [14/Jun/2014:10:30:17 -0400] "GET /support HTTP/1.1" 200 1321 "-" "Mozilla/5.0 (Windows NT 6 133.20.136.170 - - [14/Jun/2014:10:30:18 -0400] "GET /home HTTP/1.1" 200 1238 "-" "Mozilla/5.0 (Windows NT 6.1) 137.95.229.186 - - [14/Jun/2014:10:30:18 -0400] "GET /home HTTP/1.1" 200 1797 "-" "Mozilla/5.0 (Windows NT 6.1) 100.158.186.50 - - [14/Jun/2014:10:30:18 -0400] "GET /home HTTP/1.1" 200 878 "-" "Mozilla/5.0 (Windows NT 6.1; 36.44.59.115 - - [14/Jun/2014:10:30:19 -0400] "GET /checkout HTTP/1.1" 200 1551 "-" "Mozilla/5.0 (Windows NT 6. 150.47.54.136 - - [14/Jun/2014:10:30:19 -0400] "GET /department/outdoors/category/kids'%20golf%20clubs/product/ 69.109.16.108 - - [14/Jun/2014:10:30:19 -0400] "GET /department/fitness HTTP/1.1" 200 2079 "-" "Mozilla/5.0 (Wi
```

At the bottom of the preview window, there are two buttons: 'Cancel' and 'Download'.

Inspira Crea Transforma

UNIVERSIDAD
EAFIT[®]

Cargar los datos operacionales

Sqoop import

1. CREA TU PROPIA BASE DE DATOS, Ej: 'emontoyadb'

```
CREATE DATABASE emontoyadb;
```

2. IMPORTA LOS DATOS AL DATALAKE (hdfs y/o Hive)

```
sqoop import-all-tables \
```

```
  --connect jdbc:mysql:// database-1.cj1yhistqein.us-east-2.rds.amazonaws.com  
:3306/retail_db \
```

```
  --username=retail_dba --password= retail_dba \
```

```
  --hive-database emontoyadb \
```

```
  --hive-overwrite \
```

```
  --warehouse-dir=/user/emontoya/retail_db/ \
```

```
  --hive-import \
```

```
  --mysql-delimiters -m 1
```

Query

Search saved documents...

Jobs

admin

Sqoop 1

Add a name...

Add a description...

+ ↺

```
1 import-all-tables --connect jdbc:mysql://database-1.cj1yhistqein.us-east-2.rds.amazonaws.com:3306/retail_db \  
2   --username=retail_dba --password=retail_dba --hive-database retail_db \  
3   --hive-overwrite --hive-import --warehouse-dir=/tmp/retail_dbtmp -m 1 --mysql-delimiters
```

4m, 34s ⚙ ?

Procesar

Una vez cargados los datos, se procesan con:

Hive:

-- **CATEGORIAS MÁS POPULARES DE PRODUCTOS**

The screenshot shows the Apache Hue web interface. On the left, the 'Databases' sidebar lists 'default' and 'retail_db'. The main panel displays a Hive query in the 'Query' editor, which has been executed. The query is as follows:

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

Below the query editor, the execution log shows the query completed successfully in 17.218 seconds. At the bottom, the 'Results (10)' tab is active, displaying a table with the following data:

	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

script

```
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;
```

resultado







✓ Execute	Save As	Insert UDF ▾	Visual Explain
RESULTS	LOG	VISUAL EXPLAIN	TEZ UI
Filter columns	×		
c.category_name	count		
Cleats	24551		
Men\'s Footwear	22246		
Women\'s Apparel	21035		
Indoor/Outdoor Games	19298		
Fishing	17325		
Water Sports	15540		
Camping & Hiking	13729		
Cardio Equipment	12487		
Shop By Sport	10984		
Electronics	3156		

Inspira Crea Transforma

top 10 de productos que generan ganancias

Query

Search saved documents...

  Hive  Add a name... Add a description...   

+ ↺

11.31s Database retail_db ▾ Type text ▾ ⚙ ?

▶ ▾

1 SELECT p.product_id, p.product_name, r.revenue

2 FROM products p inner join

3 (select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue

4 from order_items oi inner join orders o

5 on oi.order_item_order_id = o.order_id

6 where o.order_status <> 'CANCELED'

7 and o.order_status <> 'SUSPECTED_FRAUD'

8 group by order_item_product_id) r

9 on p.product_id = r.order_item_product_id

10 order by r.revenue desc

11 limit 10

📖 ▾

Inspira Crea Transforma

UNIVERSIDAD
EAFIT¹⁶
®

script

```
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;
```

resultados

Query History

Saved Queries

Query Builder

Results (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

Correlación

Como usted es una persona de datos muy inteligente, se da cuenta de que otra pregunta comercial interesante sería: ¿son los productos más vistos también los más vendidos?

Ingestar, almacenar, y procesar **logs de eventos** de servidores web.

Una de las tecnologías que se usan para esto es FLUME. Flume es un framework de ingesta en tiempo real hacia HDFS.

Permite enrutar, filtrar, agregar, etc en una plataforma de procesamiento escalable.

Mientras tanto, se cargaran los logs en HDFS y procesaran con Hive

cargar los datos en HDFS en Hive como una tabla externa

Datos previamente cargados en:

/datasets/retail_logs/access.log


-- CREAR DIRECTORIO PARA TABLA EXTERNA CON ETL

\$ hdfs dfs -mkdir /user/<username>/warehouse/access_logs_etl




Crear tablas ETL - intermedia

Query

Q Search saved documents...

 Hive

Add a name... Add a description...



0.78s Database retail_db Type text

+ ↺

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

```
LOCATION /user/admin/datasets/retail_logs/  
INFO : Starting task [Stage-0:DDL] in serial mode  
INFO : Completed executing command(queryId=hive_20191004051642_fdc4add9-c863-48f7-8730-c4653e2e096a); Time taken:  
0.061 seconds  
INFO : OK
```

script

use <username>;

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' = '([^\]*) - - \\[([^\]*)\] ([^\]*) ([^\]*) ([^\]*) ([^\]*) ([^\]*) ([^\]*) ([^\]*) ([^\]*)',

'output.format.string' = "%1\$s %2\$s %3\$s %4\$s %5\$s %6\$s %7\$s %8\$s %9\$s")

LOCATION '/user/<username>/datasets/retail_logs/';

Crear tablas ETL - final

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/<username>/warehouse/access_logs_etl';


Query

Q Search saved documents...

 Hive

 Add a name... Add a description...





1 CREATE EXTERNAL TABLE etl_access_logs (
2 ip STRING,
3 fecha STRING,
4 method STRING,
5 url STRING,
6 http_version STRING,
7 code1 STRING,
8 code2 STRING,
9 dash STRING,
10 user_agent STRING)
11 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
12 LOCATION '/user/admin/warehouse/access_logs_etl/';

8.22s Da

Procesar ETL

ADD JAR /usr/lib/hive/lib/hive-contrib.jar;

INSERT OVERWRITE TABLE etl_access_logs SELECT * FROM tmp_access_logs;

Ahora sí, la pregunta de negocio

--- MUESTRE LOS PRODUCTOS MÁS VISITADOS

SELECT count(*) as contador,url

FROM etl_access_logs

WHERE url LIKE '%\product\%'

GROUP BY url

ORDER BY contador

DESC LIMIT 10;

Resultados

Query

Search saved documents...

2) + ↺

5.99s Database retail_db ▾ Type text ▾ ⚙ ?

1 SELECT count(*) as contador,url FROM etl_access_logs

2 WHERE url LIKE '%\product\%'

3 GROUP BY url ORDER BY contador DESC LIMIT 10;

▶ ▾

📖 ▾

INFO : Map 1: 1/1 Reducer 2: 0/1 Reducer 3: 0/1 ***

INFO : Map 1: 1/1 Reducer 2: 1/1 Reducer 3: 1/1

INFO : Completed executing command(queryId=hive_20191004052155_80b90096-330b-46ba-bd16-4cec37653ee3); Time taken:

5.126 seconds

INFO : OK

Query History

Saved Queries

Query Builder

Results (10)

🔍

📊 ▾

📄

⬇

	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%20Footb
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20C
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20
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%2
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Sere
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Runnin
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%20Hustle%20Storm%

UNIVERSIDAD
EAFIT²¹

Conclusiones de analítica de este caso