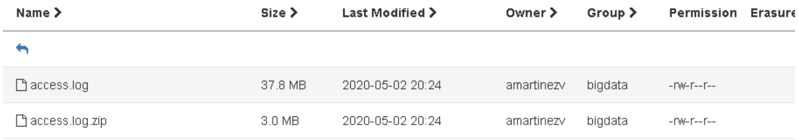
Case retail

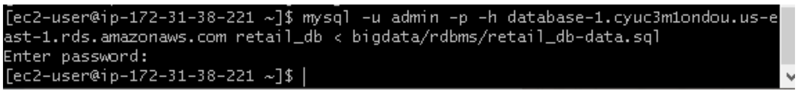
viernes, 1 de mayo de 2020 17:13

- Actualizamos los datos de Shell

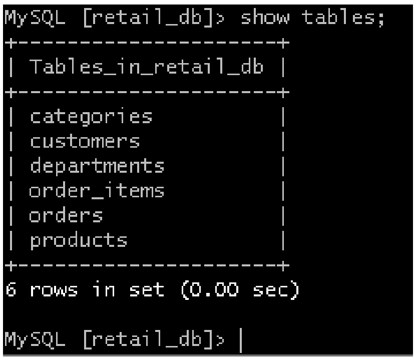


Name	Size	Last Modified	Owner	Group	Permission	Erasure
↰						
access.log	37.8 MB	2020-05-02 20:24	amartinezv	bigdata	-rw-r--r--	
access.log.zip	3.0 MB	2020-05-02 20:24	amartinezv	bigdata	-rw-r--r--	

- Se creo la base de datos retail_db con sus tablas



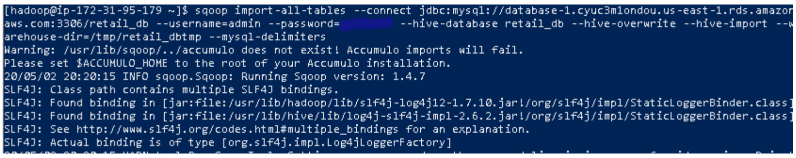
```
[ec2-user@ip-172-31-38-221 ~]$ mysql -u admin -p -h database-1.cyuc3m1ondou.us-e
ast-1.rds.amazonaws.com retail_db < bigdata/rdbms/retail_db-data.sql
Enter password:
[ec2-user@ip-172-31-38-221 ~]$ |
```



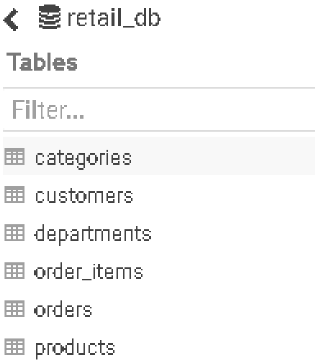
```
MySQL [retail_db]> show tables;
+---------------------+
| Tables_in_retail_db |
+---------------------+
| categories          |
| customers           |
| departments         |
| order_items         |
| orders              |
| products            |
+---------------------+
6 rows in set (0.00 sec)

MySQL [retail_db]> |
```

- Importamos los datos via scoop desde la terminal

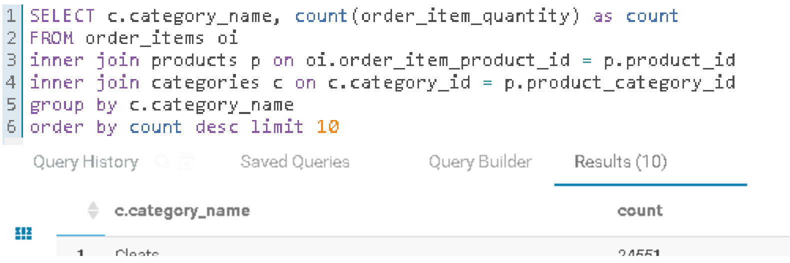


```
[hadoop@ip-172-31-95-179 ~]$ sqoop import-all-tables --connect jdbc:mysql://database-1.cyuc3m1ondou.us-east-1.rds.amazon
aws.com:3306/retail_db --username=admin --password=[redacted] --hive-database retail_db --hive-overwrite --hive-import --w
arehouse-dir=/tmp/retail_dbtmp --mysql-delimiters
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/05/02 20:20:15 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
```



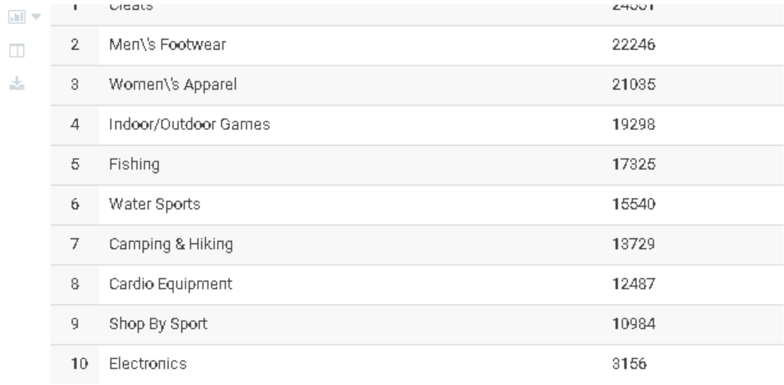
- categories
- customers
- departments
- order_items
- orders
- products

- Categoría más populares de productos:



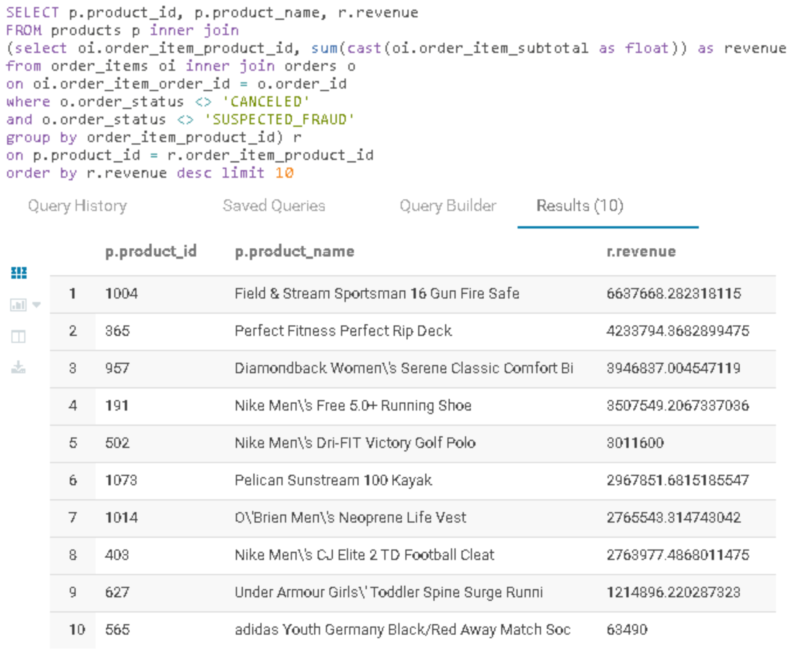
```
SELECT c.category_name, count(order_item_quantity) as count
FROM order_items oi
inner join products p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc limit 10
```

	c.category_name	count
1	Cleats	24551



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

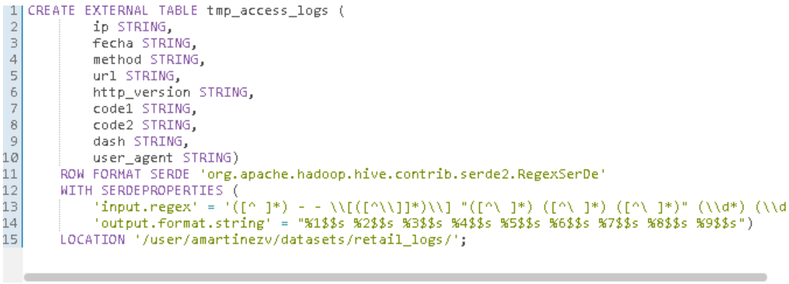
- Top 10 productos que generan ganancias



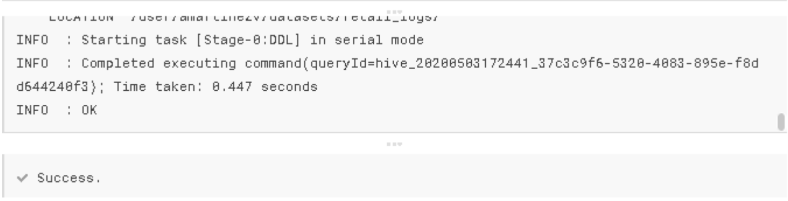
```
SELECT p.product_id, p.product_name, r.revenue
FROM products p inner join
(select oi.order_item_product_id, sum(cast(oi.order_item_subtotal as float)) as revenue
from order_items oi inner join orders o
on oi.order_item_order_id = o.order_id
where o.order_status <> 'CANCELED'
and o.order_status <>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

- Crear tabla y almacenar los logs



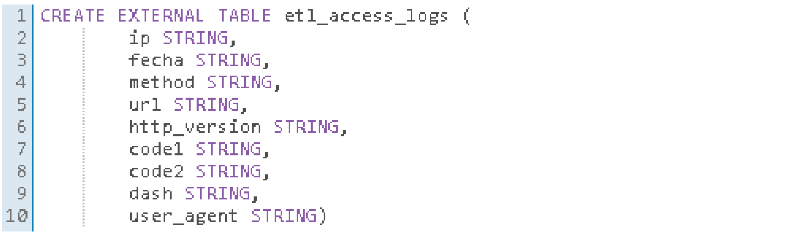
```
CREATE EXTERNAL TABLE tmp_access_logs (
        ip STRING,
        fecha STRING,
        method STRING,
        url STRING,
        http_version STRING,
        code1 STRING,
        code2 STRING,
        dash STRING,
        user_agent STRING)
    ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
    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 '/user/amartinezv/datasets/retail_logs/';
```



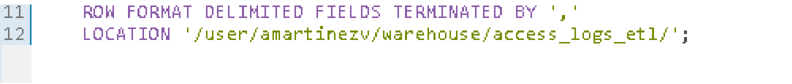
```
    LOCATION '/user/amartinezv/datasets/retail_logs/'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20200503172441_37c3c9f6-5320-4083-895e-f8d
d644240f3); Time taken: 0.447 seconds
INFO  : OK

✓ Success.
```

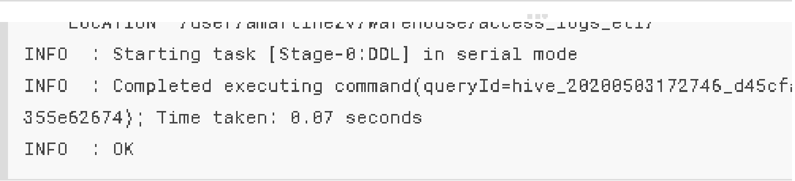
- Crear directorio para tabla externa con etl



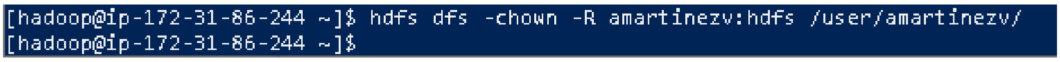
```
CREATE EXTERNAL TABLE etl_access_logs (
        ip STRING,
        fecha STRING,
        method STRING,
        url STRING,
        http_version STRING,
        code1 STRING,
        code2 STRING,
        dash STRING,
        user_agent STRING)
```



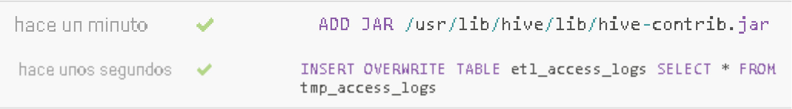
```
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/user/amartinezv/warehouse/access_logs_etl/';
```



```
    LOCATION '/user/amartinezv/warehouse/access_logs_etl/
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20200503172746_d45cf
355e62674); Time taken: 0.07 seconds
INFO  : OK
```

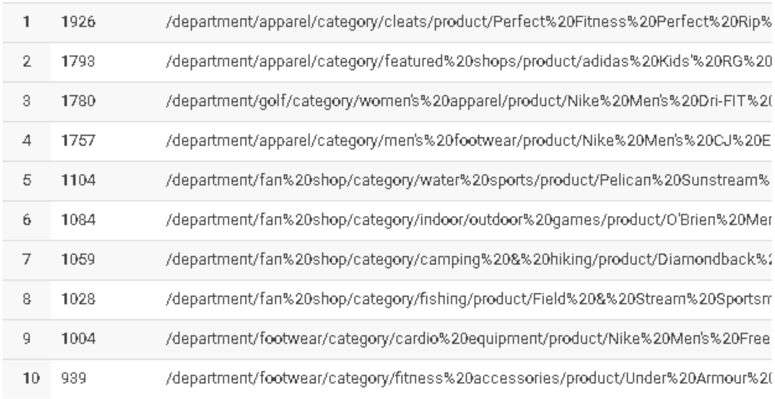


```
[hadoop@ip-172-31-86-244 ~]$ hdfs dfs -chown -R amartinezv:hdfs /user/amartinezv/
[hadoop@ip-172-31-86-244 ~]$
```



hace un minuto	✓	ADD JAR /usr/lib/hive/lib/hive-contrib.jar
hace unos segundos	✓	INSERT OVERWRITE TABLE etl_access_logs SELECT * FROM tmp_access_logs

- Productos más visitados



1	1926	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%
2	1793	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20
3	1780	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%2
4	1757	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20E
5	1104	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%
6	1084	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Me
7	1059	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%
8	1028	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sports
9	1004	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free
10	939	/department/footwear/category/fitness%20accessories/product/Under%20Armour%2

- Preguntas 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?

  Como se puede ver en los resultados obtenidos de los logs los productos más visitados corresponden los que generan más ganancia, por esto son también los que generan más rentabilidad para la empresa.