**Análisis sentimientos**

1. APARTADO A
   1. Siguiendo los pasos que se explican en el tema, crea un clúster Hadoop EMR con 1 master y dos nodos. Selecciona la opción Core Hadoop (versión 7.0.0). No te olvides de seleccionar Sqoop, ya que lo utilizaremos en las prácticas siguientes.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Conexión

Texto

El contenido generado por IA puede ser incorrecto.

1. APARTADO B
   1. Viene en formato Excel. Desde el propio Excel puedes convertirlo a formato ’csv’.

Exportamos

Captura de pantalla de un celular

El contenido generado por IA puede ser incorrecto.

* 1. Crea una carpeta con tu nombre en el directorio user del HDFS de EMR.

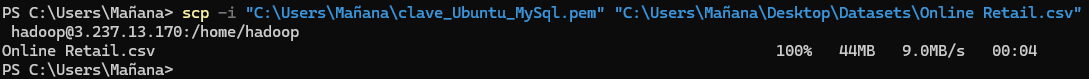
Texto

El contenido generado por IA puede ser incorrecto.

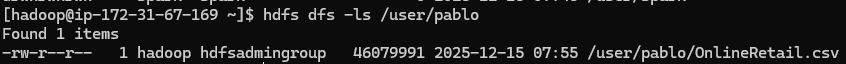
* 1. Crea dentro de él una carpeta llamada ventas y sube a ella el ’csv’ que obtuviste anteriormente.

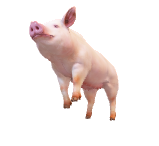


Lo subimos a la machine



Lo subimos a nuestra carpeta personal de hdfs.



1. APARTADO C
   1. Usando Pig
      1. Cargar los datos del dataset en PIG

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

sample = LIMIT data 10;

DUMP sample;

* 1. Consultas:
     1. ¿Cuántos registros tiene la tabla?

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

group\_all = GROUP data ALL;

count\_records = FOREACH group\_all GENERATE COUNT(data) AS total\_registros;

STORE count\_records INTO '/user/pablo/output\_count' USING PigStorage(',');



* + 1. Mostrar registros con cantidades mayores o iguales a cero.

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

data\_qty =

FILTER data BY (int)Quantity >= 0;

STORE data\_qty INTO '/user/pablo/output\_quantity\_ge\_zero' USING PigStorage(';');

Texto

El contenido generado por IA puede ser incorrecto.

* + 1. A partir de la consulta anterior, mostrar registros precio mayor a cero

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

data\_qty =

FILTER data BY (int)Quantity >= 0;

data\_price =

FILTER data\_qty BY (double)REPLACE(UnitPrice, ',', '.') > 0.0;

STORE data\_price INTO '/user/pablo/output\_quantity\_ge\_zero\_price\_gt\_zero' USING PigStorage(';');

Texto

El contenido generado por IA puede ser incorrecto.

* + 1. A partir de la consulta anterior, mostrar solamente los registros con algún valor en el campo CustomerID.

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

data\_qty =

FILTER data BY (int)Quantity >= 0;

data\_price =

FILTER data\_qty BY (double)REPLACE(UnitPrice, ',', '.') > 0.0;

data\_customer =

FILTER data\_price BY CustomerID IS NOT NULL AND CustomerID != '';

STORE data\_customer INTO '/user/pablo/output\_quantity\_price\_customerid' USING PigStorage(';');

Texto

El contenido generado por IA puede ser incorrecto.

* + 1. ¿Cuántos registros tiene la última consulta?

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

data\_qty =

FILTER data BY (int)Quantity >= 0;

data\_price =

FILTER data\_qty BY (double)REPLACE(UnitPrice, ',', '.') > 0.0;

data\_customer =

FILTER data\_price BY CustomerID IS NOT NULL AND CustomerID != '';

group\_all = GROUP data\_customer ALL;

count\_records = FOREACH group\_all GENERATE COUNT(data\_customer) AS total\_registros;

STORE count\_records INTO '/user/pablo/output\_count\_quantity\_price\_customerid' USING PigStorage(',');



* + 1. Almacena la consulta final del punto 4 en un fichero llamado ventas.csv dentro de la carpeta de apartado B.

raw\_data =

LOAD '/user/pablo/OnlineRetail.csv'

USING PigStorage(';')

AS (

InvoiceNo:chararray,

StockCode:chararray,

Description:chararray,

Quantity:chararray,

InvoiceDate:chararray,

UnitPrice:chararray,

CustomerID:chararray,

Country:chararray

);

data =

FILTER raw\_data BY InvoiceNo != 'InvoiceNo';

data\_qty =

FILTER data BY (int)Quantity >= 0;

data\_price =

FILTER data\_qty BY (double)REPLACE(UnitPrice, ',', '.') > 0.0;

data\_customer =

FILTER data\_price BY CustomerID IS NOT NULL AND CustomerID != '';

STORE data\_customer INTO '/user/pablo/temp\_ventas' USING PigStorage(';');

Guardamos el archivo

Texto

El contenido generado por IA puede ser incorrecto.

1. APARTADO D
   1. Crear base de datos

CREATE DATABASE IF NOT EXISTS retail\_db;

* 1. Crear tabla externa sobre los datos RAW (CSV)

CREATE EXTERNAL TABLE IF NOT EXISTS retail\_db.online\_retail (

InvoiceNo STRING,

StockCode STRING,

Description STRING,

Quantity INT,

InvoiceDate STRING,

UnitPrice STRING,

CustomerID STRING,

Country STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ';'

STORED AS TEXTFILE

LOCATION '/user/pablo/ventas'

TBLPROPERTIES ("skip.header.line.count"="1");

* 1. Hive no maneja muy bien el formato de fecha original, conviértelo a d/M/yyyy H:mm
  2. Crea la misma estructura de tabla, pero particionada por año y mes.

CREATE TABLE IF NOT EXISTS retail\_db.online\_retail\_partitioned (

InvoiceNo STRING,

StockCode STRING,

Description STRING,

Quantity INT,

InvoiceDate TIMESTAMP,

UnitPrice DOUBLE,

CustomerID STRING,

Country STRING

)

PARTITIONED BY (year INT, month INT)

STORED AS PARQUET;

* 1. Inserta los registros del punto 1.2 en la tabla particionada.

Usaremos una tabla principal

CREATE EXTERNAL TABLE IF NOT EXISTS retail\_db.ventas\_minimal (

InvoiceNo STRING,

StockCode STRING,

Description STRING,

Quantity INT,

InvoiceDate STRING,

UnitPrice STRING,

CustomerID STRING,

Country STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ';'

STORED AS TEXTFILE

LOCATION '/user/pablo/ventas';

Ejecutamos la carga

INSERT INTO TABLE retail\_db.online\_retail\_partitioned

PARTITION (year, month)

SELECT

InvoiceNo,

StockCode,

Description,

Quantity,

from\_unixtime(unix\_timestamp(InvoiceDate, 'dd/MM/yyyy H:mm')) AS InvoiceDate,

CAST(REPLACE(UnitPrice, ',', '.') AS DOUBLE) AS UnitPrice,

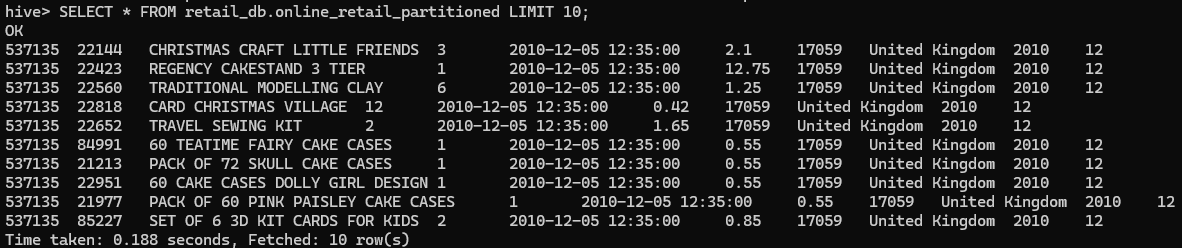
CustomerID,

Country,

YEAR(from\_unixtime(unix\_timestamp(InvoiceDate, 'dd/MM/yyyy H:mm'))) AS year,

MONTH(from\_unixtime(unix\_timestamp(InvoiceDate, 'dd/MM/yyyy H:mm'))) AS month

FROM retail\_db.ventas\_minimal;



1. APARTADO E

Análisis de clientes

* 1. 10 clientes con mayor gasto total

SELECT CustomerID,

SUM(Quantity \* UnitPrice) AS total\_gasto

FROM retail\_db.online\_retail\_partitioned

WHERE CustomerID IS NOT NULL

GROUP BY CustomerID

ORDER BY total\_gasto DESC

LIMIT 10;

Texto

El contenido generado por IA puede ser incorrecto.

* 1. Clientes con más compras (cantidad de facturas)

SELECT CustomerID,

COUNT(DISTINCT InvoiceNo) AS num\_facturas

FROM retail\_db.online\_retail\_partitioned

WHERE CustomerID IS NOT NULL

GROUP BY CustomerID

ORDER BY num\_facturas DESC

LIMIT 10;

Tabla

El contenido generado por IA puede ser incorrecto.

Análisis de productos

* 1. 10 productos más vendidos

SELECT StockCode,

Description,

SUM(Quantity) AS total\_vendido

FROM retail\_db.online\_retail\_partitioned

GROUP BY StockCode, Description

ORDER BY total\_vendido DESC

LIMIT 10;

Texto

El contenido generado por IA puede ser incorrecto.

* 1. 10 productos más rentables (suma de precio unitario por cantidad)

SELECT StockCode,

Description,

SUM(Quantity \* UnitPrice) AS total\_ingresos

FROM retail\_db.online\_retail\_partitioned

GROUP BY StockCode, Description

ORDER BY total\_ingresos DESC

LIMIT 10;

Texto

El contenido generado por IA puede ser incorrecto.

Análisis geográfico

* 1. Países con mayor volumen de ventas

SELECT Country,

SUM(Quantity) AS total\_unidades

FROM retail\_db.online\_retail\_partitioned

GROUP BY Country

ORDER BY total\_unidades DESC;

Texto

El contenido generado por IA puede ser incorrecto.

Análisis temporal

* 1. Ventas totales por mes (suma de precio unitario por cantidad)

SELECT year,

month,

SUM(Quantity \* UnitPrice) AS total\_ventas

FROM retail\_db.online\_retail\_partitioned

GROUP BY year, month

ORDER BY year, month;

Texto

El contenido generado por IA puede ser incorrecto.

* 1. Hora del día con más actividad

SELECT HOUR(InvoiceDate) AS hora,

COUNT(\*) AS num\_transacciones

FROM retail\_db.online\_retail\_partitioned

GROUP BY HOUR(InvoiceDate)

ORDER BY num\_transacciones DESC

LIMIT 1;



1. APARTADO F

SQOOP

* 1. En tu servidor MySQL en la máquina EC2 crea una base de datos y una tabla para almacenar los datos del fichero ventas.csv.

Nos conectamos a la maquina con la base de datos, uno siempre vuelve a donde fue feliz.

Texto

El contenido generado por IA puede ser incorrecto.

Creamos la base de datos

Texto

El contenido generado por IA puede ser incorrecto.

* 1. Exporta con SQOOP los datos de ventas.csv a la tabla que creaste en el punto anterior

Exportamos

sqoop export \

--connect jdbc:mysql://44.212.173.124:3306/retail\_db \

--username admin\_remote \

--password 12345\_Sql \

--table ventas \

--export-dir /user/pablo/ventas \

--input-fields-terminated-by ';' \

--input-lines-terminated-by '\n' \

--columns "InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country" \

--driver com.mysql.cj.jdbc.Driver

Imagen que contiene texto, marcador

El contenido generado por IA puede ser incorrecto.