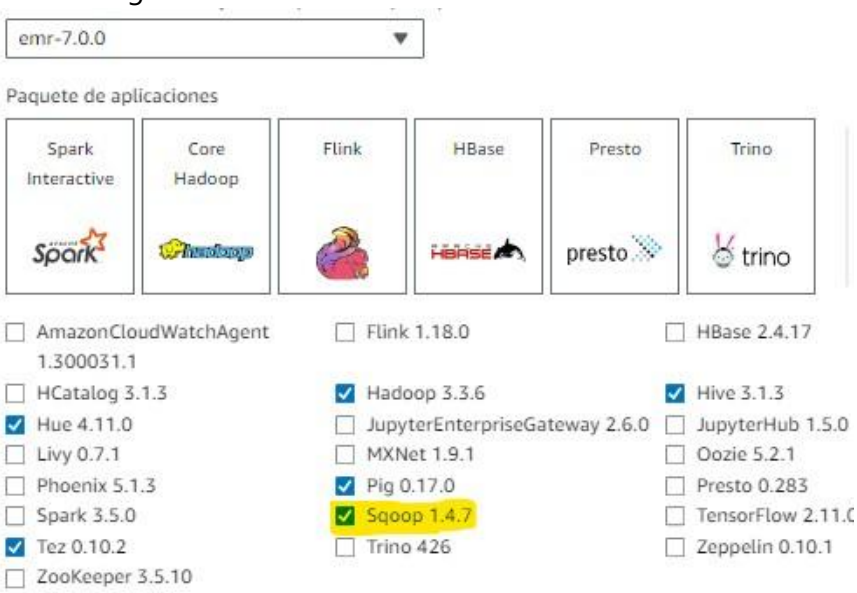
Enviar tareas como Pasos (*Steps*) a clústeres EMR.

CONTENIDO

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.



INTRODUCCIÓN

- Utilizaremos el *dataset* *https://www.kaggle.com/datasets/alimortezaie/online-retail*.

Creamos el Cluster

Crear clúster Información

Nombre y aplicaciones - *obligatorio* Información

Asigne un nombre a su clúster y elija las aplicaciones que desea instalar en él.

Nombre

Clúster Brayan

Versión de Amazon EMR Información

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.

emr-7.0.0

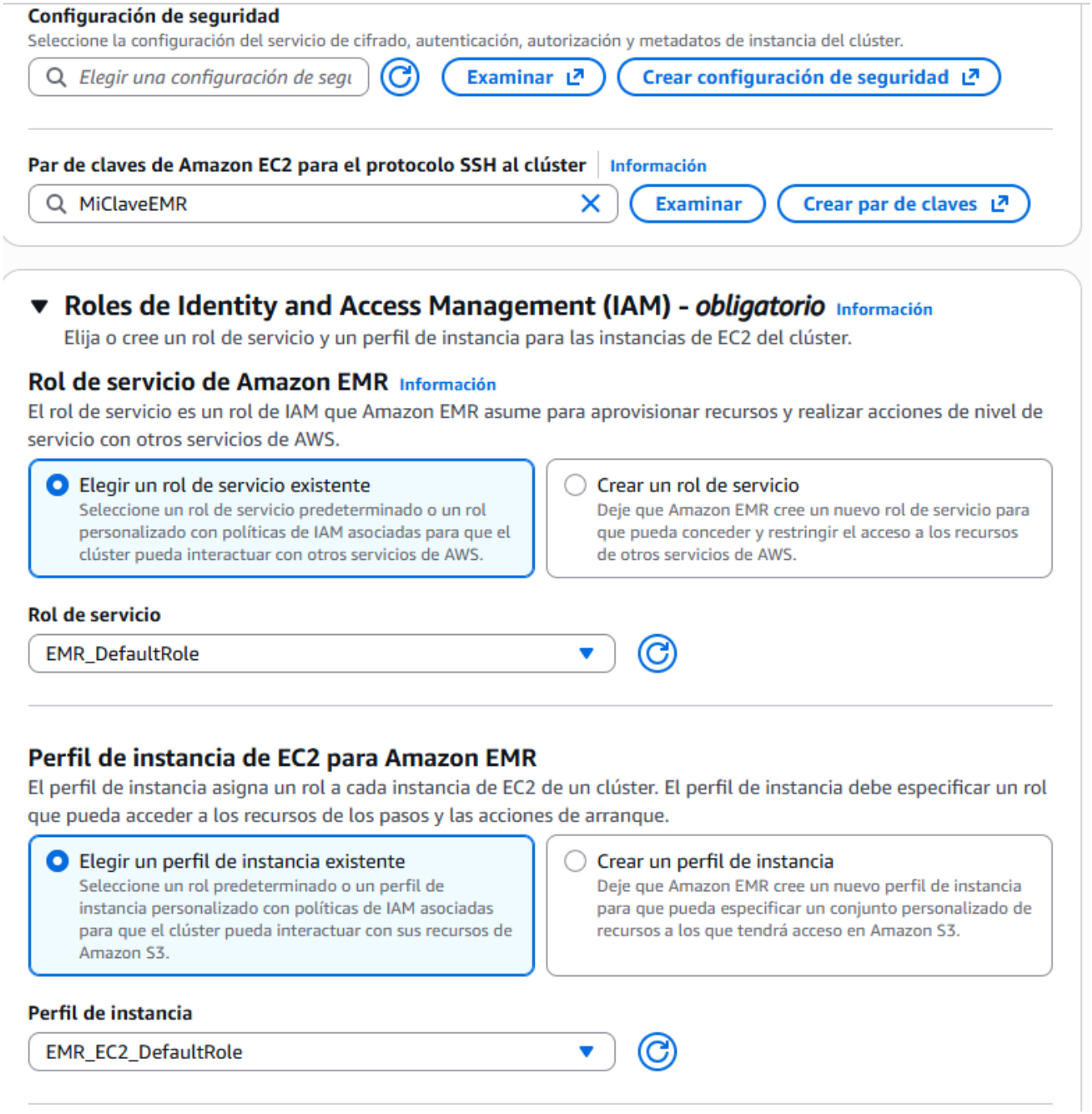
⚠ El soporte para esta versión de EMR finalizará May-01-2026, por lo que ya no podrá recibir soporte técnico. AWS recomienda encarecidamente que ponga en marcha sus cargas de trabajo en la versión más reciente de Amazon EMR para recibir actualizaciones y correcciones críticas para la seguridad. También puede usar el nuevo agente de actualización de Spark para actualizar las aplicaciones existentes en la versión 5.40 o superior a la última versión de EMR. Para obtener más información, consulte Política de soporte estándar de EMR y Actualizaciones de Spark

Paquete de aplicaciones

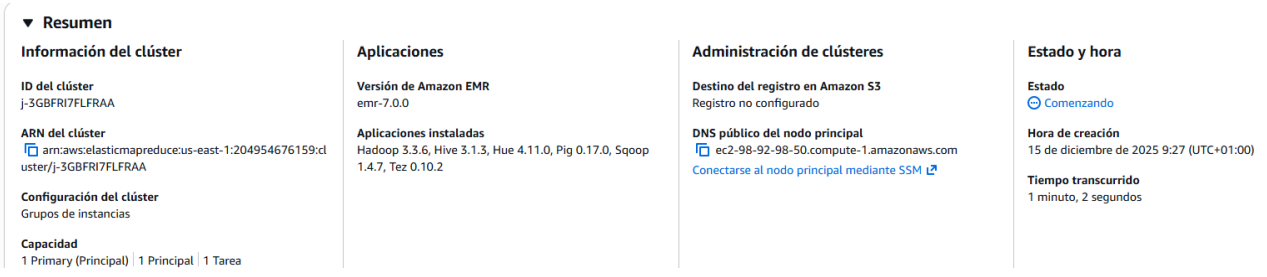
Spark Interactive | Core Hadoop | Flink | HBase | Presto | Trino | Custom

- [ ] AmazonCloudWatchAgent 1.300031.1
- [ ] Flink 1.18.0
- [ ] HBase 2.4.17
- [ ] HCatalog 3.1.3
- [x] Hadoop 3.3.6
- [x] Hive 3.1.3
- [x] Hue 4.11.0
- [ ] JupyterEnterpriseGateway 2.6.0
- [ ] JupyterHub 1.5.0
- [ ] Livy 0.7.1
- [ ] MXNet 1.9.1
- [ ] Oozie 5.2.1
- [ ] Phoenix 5.1.3
- [x] Pig 0.17.0
- [ ] Presto 0.283
- [ ] Spark 3.5.0
- [x] Sqoop 1.4.7
- [ ] TensorFlow 2.11.0
- [x] Tez 0.10.2
- [ ] Trino 426
- [ ] Zeppelin 0.10.1
- [ ] ZooKeeper 3.5.10

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Central	m4.large	1	☐
Nodo1	m4.large	1	☐
Nodo2	m4.large	1	☐



Creado



Nos podemos conectar por SSH al nodo principal

Direcció...	IP elástica	Direccio...	Monitoreo	Nombre del grupo d...	Nombre de...	Hora de lanzamiento	Detalles...
35.170.64.71	–	–	disabled	ElasticMapReduce-slave	MiClaveEMR	2025/12/15 09:28 GMT+1	Linux/UNIX
3.237.40.118	–	–	disabled	ElasticMapReduce-slave	MiClaveEMR	2025/12/15 09:28 GMT+1	Linux/UNIX
98.92.98.50	–	–	disabled	ElasticMapReduce-master	MiClaveEMR	2025/12/15 09:28 GMT+1	Linux/UNIX

```
hadoop@ip-172-31-70-46:~
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '98.92.98.50' (ED25519) to the list of known hosts.
       #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
Last login: Mon Dec 15 08:54:23 2025

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE::::::EEEEEEEEE:::E M::::::::M       M::::::::M R::::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR:::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E::::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::::E  M:::::M  M:::M:::M  M:::::M   R::::::::::::RR
  E::::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E              M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE  M:::::M     MMM     M:::::M   R:::R      R::::R
EE::::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR:::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-70-46 ~]$
```

IU de la aplicación en el nodo principal

Estas requieren que el túnel de SSH esté habilitado.

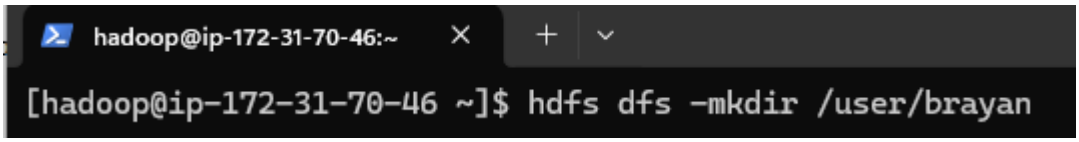
Aplicación	URL de la IU
Administrador de recursos	http://ec2-98-92-98-50.compute-1.amazonaws.com:8088/
Nodo del nombre de HDFS	http://ec2-98-92-98-50.compute-1.amazonaws.com:9870/
Tonalidad	http://ec2-98-92-98-50.compute-1.amazonaws.com:8888/
UI de Tez	http://ec2-98-92-98-50.compute-1.amazonaws.com:8080/tez-ui

HUE
Usuario Brayan
Contraseña Admin123456!

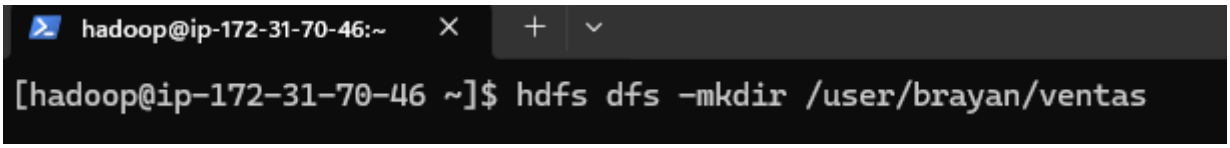
CONTENIDO

APARTADO B

1. **Viene en formato Excel. Desde el propio Excel puedes convertirlo a formato '*csv*'**
2. **Crea una carpeta con tu nombre en el directorio `user` del HDFS de EMR**



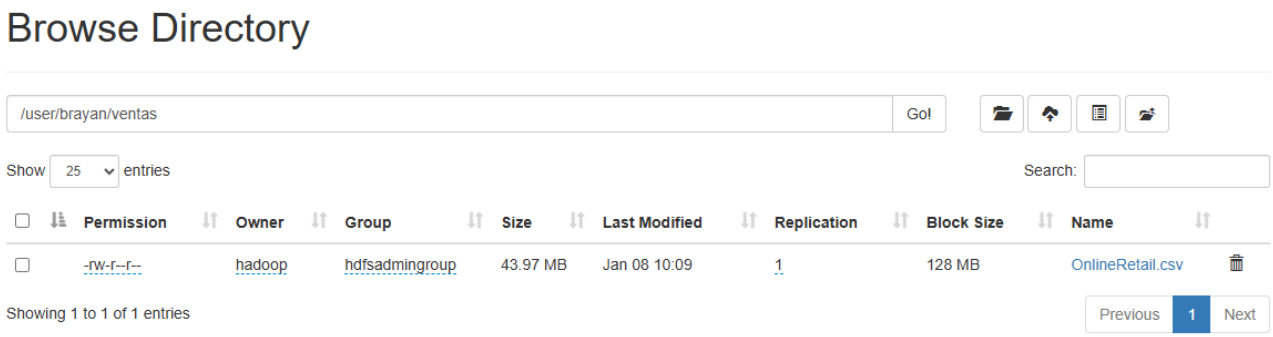
3. **Crea dentro de él una carpeta llamada `ventas` y sube a ella el '*csv*' que obtuviste anteriormente.**




```
PS C:\Users\Mañana\Downloads> scp -i MiClaveEMR.pem OnlineRetail.csv hadoop@ec2-18-206-83-142.compute-1.amazonaws.com:/h
ome/hadoop/
OnlineRetail.csv                                              100%   44MB   8.2MB/s   00:05
[hadoop@ip-172-31-78-134 ~]$ hdfs dfs -put /home/hadoop/OnlineRetail.csv /user/brayan/ventas/
```

Lo movemos de hadoop a directorio ventas

```
[hadoop@ip-172-31-78-134 ~]$ hdfs dfs -put /home/hadoop/OnlineRetail.csv /user/brayan/ventas/
[hadoop@ip-172-31-78-134 ~]$ hdfs dfs -ls /user/brayan/ventas
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup   46102471 2026-01-08 09:09 /user/brayan/ventas/OnlineRetail.csv
[hadoop@ip-172-31-78-134 ~]$
```



CONTENIDO

APARTADO C

USANDO Pig

1. Cargar los datos del *dataset* en PIG

Consultas:

```
grunt> ventas = LOAD '/user/brayan/ventas/OnlineRetail.csv'
>> USING PigStorage('|')
>> AS (
>>     InvoiceNo:chararray,
>>     StockCode:chararray,
>>     Description:chararray,
>>     Quantity:int,
>>     InvoiceDate:chararray,
>>     UnitPrice:double,
>>     CustomerID:chararray,
>>     Country:chararray
>> );
```

2. ¿Cuántos registros tiene la tabla?

```
grunt> total = FOREACH (GROUP ventas ALL) GENERATE COUNT(ventas);
grunt> DUMP total;
```

```
DAG Plan:
Tez vertex scope-118    ->      Tez vertex scope-119,
Tez vertex scope-119

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  Hdfs
BytesRead HdfsBytesWritten Alias        Feature Outputs
scope-118          1          1         541910                   0         541910            64               69
 46102471               0 1-29,total,ventas
scope-119          1          1              0                   1              1            69                0
        0               9 total        GROUP_BY        hdfs://ip-172-31-78-134.ec2.internal:8020/tmp/temp1044652651/tmp
338346828,

Input(s):
Successfully read 541910 records (46102471 bytes) from: "/user/brayan/ventas/OnlineRetail.csv"

Output(s):
Successfully stored 1 records (9 bytes) in: "hdfs://ip-172-31-78-134.ec2.internal:8020/tmp/temp1044652651/tmp338346828"

2026-01-08 09:34:34,329 INFO input.FileInputFormat: Total input files to process : 1
332162 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
2026-01-08 09:34:34,329 INFO util.MapRedUtil: Total input paths to process : 1
(541910)
```

Big Data

3. Mostrar registros con cantidades mayores o iguales a cero.

```
grunt> ventas_q = FILTER ventas BY Quantity >= 0;
grunt> DUMP ventas_q;
```

```
(581586,20685,DOORMAT RED RETROSPOT,10,09/12/2011 12:49,7.08,13113,United Kingdom)
(581587,22631,CIRCUS PARADE LUNCH BOX ,12,09/12/2011 12:50,1.95,12680,France)
(581587,22556,PLASTERS IN TIN CIRCUS PARADE ,12,09/12/2011 12:50,1.65,12680,France)
(581587,22555,PLASTERS IN TIN STRONGMAN,12,09/12/2011 12:50,1.65,12680,France)
(581587,22728,ALARM CLOCK BAKELIKE PINK,4,09/12/2011 12:50,3.75,12680,France)
(581587,22727,ALARM CLOCK BAKELIKE RED ,4,09/12/2011 12:50,3.75,12680,France)
(581587,22726,ALARM CLOCK BAKELIKE GREEN,4,09/12/2011 12:50,3.75,12680,France)
(581587,22730,ALARM CLOCK BAKELIKE IVORY,4,09/12/2011 12:50,3.75,12680,France)
(581587,22367,CHILDRENS APRON SPACEBOY DESIGN,8,09/12/2011 12:50,1.95,12680,France)
(581587,22629,SPACEBOY LUNCH BOX ,12,09/12/2011 12:50,1.95,12680,France)
(581587,23256,CHILDRENS CUTLERY SPACEBOY ,4,09/12/2011 12:50,4.15,12680,France)
(581587,22613,PACK OF 20 SPACEBOY NAPKINS,12,09/12/2011 12:50,0.85,12680,France)
(581587,22899,CHILDREN'S APRON DOLLY GIRL ,6,09/12/2011 12:50,2.1,12680,France)
(581587,23254,CHILDRENS CUTLERY DOLLY GIRL ,4,09/12/2011 12:50,4.15,12680,France)
(581587,23255,CHILDRENS CUTLERY CIRCUS PARADE,4,09/12/2011 12:50,4.15,12680,France)
(581587,22138,BAKING SET 9 PIECE RETROSPOT ,3,09/12/2011 12:50,4.95,12680,France)
grunt>
```

4. A partir de la consulta anterior, mostrar registros precio mayor a cero

```
grunt> ventas_q_p = FILTER ventas_q BY UnitPrice > 0;
```

```
(581585,23328,SET 6 SCHOOL MILK BOTTLES IN CRATE,4,09/12/2011 12:31,3.75,15804,United Kingdom)
(581585,23145,ZINC T-LIGHT HOLDER STAR LARGE,12,09/12/2011 12:31,0.95,15804,United Kingdom)
(581585,22466,FAIRY TALE COTTAGE NIGHT LIGHT,12,09/12/2011 12:31,1.95,15804,United Kingdom)
(581586,22061,LARGE CAKE STAND  HANGING STRAWBERY,8,09/12/2011 12:49,2.95,13113,United Kingdom)
(581586,23275,SET OF 3 HANGING OWLS OLLIE BEAK,24,09/12/2011 12:49,1.25,13113,United Kingdom)
(581586,21217,RED RETROSPOT ROUND CAKE TINS,24,09/12/2011 12:49,8.95,13113,United Kingdom)
(581586,20685,DOORMAT RED RETROSPOT,10,09/12/2011 12:49,7.08,13113,United Kingdom)
(581587,22631,CIRCUS PARADE LUNCH BOX ,12,09/12/2011 12:50,1.95,12680,France)
(581587,22556,PLASTERS IN TIN CIRCUS PARADE ,12,09/12/2011 12:50,1.65,12680,France)
(581587,22555,PLASTERS IN TIN STRONGMAN,12,09/12/2011 12:50,1.65,12680,France)
(581587,22728,ALARM CLOCK BAKELIKE PINK,4,09/12/2011 12:50,3.75,12680,France)
(581587,22727,ALARM CLOCK BAKELIKE RED ,4,09/12/2011 12:50,3.75,12680,France)
(581587,22726,ALARM CLOCK BAKELIKE GREEN,4,09/12/2011 12:50,3.75,12680,France)
(581587,22730,ALARM CLOCK BAKELIKE IVORY,4,09/12/2011 12:50,3.75,12680,France)
(581587,22367,CHILDRENS APRON SPACEBOY DESIGN,8,09/12/2011 12:50,1.95,12680,France)
(581587,22629,SPACEBOY LUNCH BOX ,12,09/12/2011 12:50,1.95,12680,France)
(581587,23256,CHILDRENS CUTLERY SPACEBOY ,4,09/12/2011 12:50,4.15,12680,France)
(581587,22613,PACK OF 20 SPACEBOY NAPKINS,12,09/12/2011 12:50,0.85,12680,France)
(581587,22899,CHILDREN'S APRON DOLLY GIRL ,6,09/12/2011 12:50,2.1,12680,France)
(581587,23254,CHILDRENS CUTLERY DOLLY GIRL ,4,09/12/2011 12:50,4.15,12680,France)
(581587,23255,CHILDRENS CUTLERY CIRCUS PARADE,4,09/12/2011 12:50,4.15,12680,France)
(581587,22138,BAKING SET 9 PIECE RETROSPOT ,3,09/12/2011 12:50,4.95,12680,France)
```

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

```
grunt> ventas_final = FILTER ventas_q_p BY CustomerID IS NOT NULL;
```

```
(581133,90030B,RED KUKUI COCONUT SEED NECKLACE,6,07/12/2011 12:55,1.0,14904,United Kingdom)
(581133,90210C,RED   ACRYLIC FACETED BANGLE,10,07/12/2011 12:55,1.0,14904,United Kingdom)
(581171,POST,POSTAGE,2,07/12/2011 15:02,18.0,12615,France)
(581179,POST,POSTAGE,1,07/12/2011 15:43,240.0,12471,Germany)
(581182,POST,POSTAGE,4,07/12/2011 15:56,28.0,12783,Portugal)
(581183,POST,POSTAGE,4,07/12/2011 16:24,18.0,12569,Germany)
(581184,POST,POSTAGE,2,07/12/2011 16:24,18.0,12569,Germany)
(581221,23444,Next Day Carriage,1,08/12/2011 9:40,15.0,17856,United Kingdom)
(581232,POST,POSTAGE,4,08/12/2011 10:26,40.0,12358,Austria)
(581266,POST,POSTAGE,5,08/12/2011 11:25,18.0,12621,Germany)
(581279,POST,POSTAGE,3,08/12/2011 11:35,18.0,12437,France)
(581336,23444,Next Day Carriage,1,08/12/2011 12:10,15.0,16161,United Kingdom)
(581395,23485,BOTANICAL GARDENS WALL CLOCK ,1,08/12/2011 13:18,25.0,16892,United Kingdom)
(581434,90210B,CLEAR ACRYLIC FACETED BANGLE,10,08/12/2011 16:10,1.0,13599,United Kingdom)
(581434,90210C,RED   ACRYLIC FACETED BANGLE,10,08/12/2011 16:10,1.0,13599,United Kingdom)
(581449,23485,BOTANICAL GARDENS WALL CLOCK ,1,08/12/2011 17:37,25.0,12748,United Kingdom)
(581493,POST,POSTAGE,1,09/12/2011 10:10,15.0,12423,Belgium)
(581494,POST,POSTAGE,2,09/12/2011 10:13,18.0,12518,Germany)
(581570,POST,POSTAGE,1,09/12/2011 11:59,18.0,12662,Germany)
(581574,POST,POSTAGE,2,09/12/2011 12:09,18.0,12526,Germany)
(581578,POST,POSTAGE,3,09/12/2011 12:16,18.0,12713,Germany)
```

6. ¿Cuántos registros tiene la última consulta?

```
DAG Plan:
Tez vertex scope-73

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  Hdfs
BytesRead HdfsBytesWritten Alias        Feature Outputs
scope-73           1         1         541910                  0              0              0               0
 46079956               0 ventas,ventas_q             hdfs://ip-172-31-66-195.ec2.internal:8020/tmp/temp-338222381/tmp
1169883018,

Input(s):
Successfully read 541910 records (46079956 bytes) from: "/user/brayan/ventas/OnlineRetail.csv"

Output(s):
Successfully stored 0 records in: "hdfs://ip-172-31-66-195.ec2.internal:8020/tmp/temp-338222381/tmp1169883018"

2026-01-08 17:39:59,699 INFO input.FileInputFormat: Total input files to process : 1
208600 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
2026-01-08 17:39:59,699 INFO util.MapRedUtil: Total input paths to process : 1
grunt>
```

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

```
grunt> STORE ventas_final INTO '/user/brayan/ventas/ventas.csv'
>> USING PigStorage(';');
```

```
DAG Plan:
Tez vertex scope-153

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  Hdfs
BytesRead HdfsBytesWritten Alias        Feature Outputs
scope-153          1         1         541910                  0           1712              0               0
 46079956           110820 ventas,ventas_q              /user/brayan/ventas/ventas.csv,

Input(s):
Successfully read 541910 records (46079956 bytes) from: "/user/brayan/ventas/OnlineRetail.csv"

Output(s):
Successfully stored 1712 records (110820 bytes) in: "/user/brayan/ventas/ventas.csv"
```

/user/brayan/ventas/ventas.csv Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	hdfsadmingroup	0 B	Jan 08 18:52	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	hdfsadmingroup	108.22 KB	Jan 08 18:52	1	128 MB	part-v000-o000-r-00000

Showing 1 to 2 of 2 entries Previous 1 Next

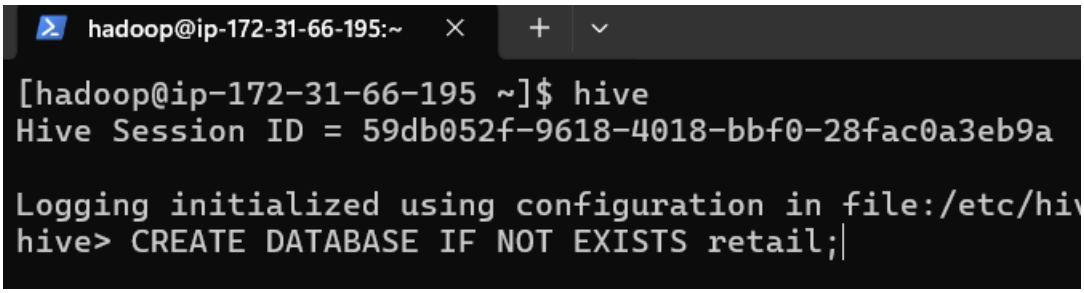
CONTENIDO

APARTADO D

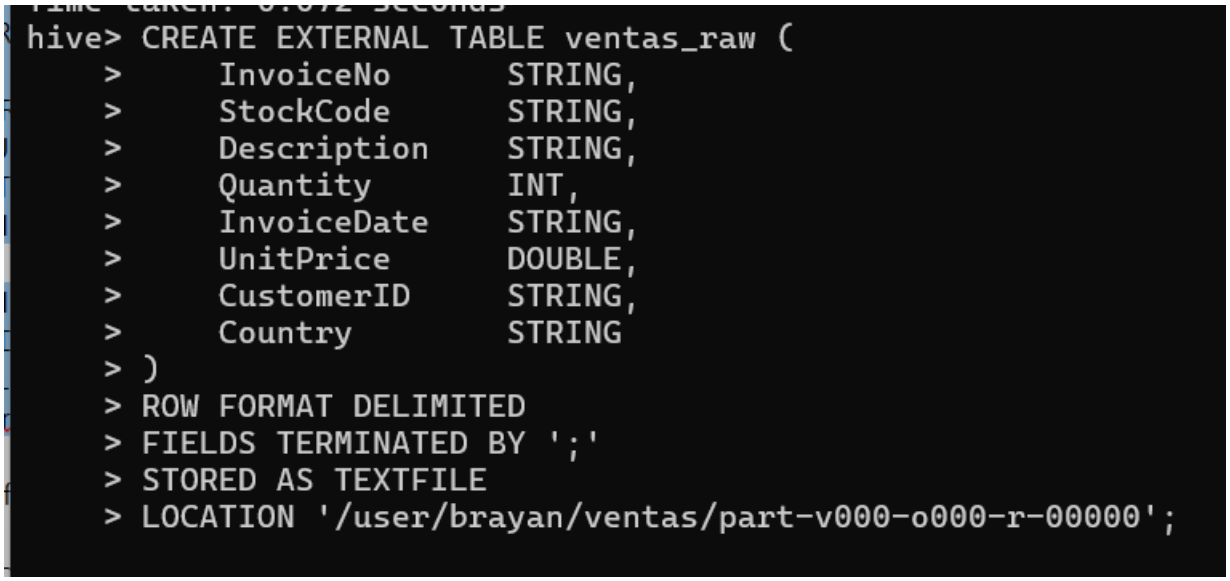
Crear la tabla externa en Hive partiendo del fichero del punto anterior.

Conéctate al nodo maestro (SSH) o usa Hue.

1. Crear base de datos



2. Crear tabla externa sobre los datos RAW (CSV)



3. Hive no maneja muy bien el formato de fecha original, conviértelo a `d/M/yyyy H:mm`

```
hive> INSERT INTO ventas
    > SELECT
    >     InvoiceNo,
    >     StockCode,
    >     Description,
    >     Quantity,
    >     to_timestamp(InvoiceDate, 'dd/MM/yyyy HH:mm'),
    >     UnitPrice,
    >     CustomerID,
    >     Country
    > FROM ventas_raw;
```

4. Crea la misma estructura de tabla pero particionada por año y mes.

```
hive> CREATE TABLE ventas_part (
    >     InvoiceNo      STRING,
    >     StockCode      STRING,
    >     Description    STRING,
    >     Quantity       INT,
    >     InvoiceDate    TIMESTAMP,
    >     UnitPrice      DOUBLE,
    >     CustomerID     STRING,
    >     Country        STRING
    > )
    > PARTITIONED BY (anio INT, mes INT)
    > STORED AS PARQUET;
```

5. Inserta los registros del punto 1.2 en la tabla particionada.

```
hive> INSERT INTO TABLE ventas_part
    > PARTITION (anio, mes)
    > SELECT
    >     InvoiceNo,
    >     StockCode,
    >     Description,
    >     Quantity,
    >     InvoiceDate,
    >     UnitPrice,
    >     CustomerID,
    >     Country,
    >     year(InvoiceDate) AS anio,
    >     month(InvoiceDate) AS mes
    > FROM ventas;
```

CONTENIDO
APARTADO E

Consultas con **HIVE**:

Análisis de clientes

1. 10 clientes con mayor gasto total

```
hive> SELECT CustomerID, SUM(Quantity * UnitPrice) AS gasto
    > FROM ventas_part
    > GROUP BY CustomerID
    > ORDER BY gasto DESC
    > LIMIT 10;
Query ID = hadoop_20260108180328_28579f55-e999-4c5e-96db-8b4eeab4f7ad
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/03  [==========================>>] 100%  ELAPSED TIME: 10.55 s
--------------------------------------------------------------------------------
```

2. Clientes con más compras (cantidad de facturas)

```
hive> SELECT CustomerID, COUNT(DISTINCT InvoiceNo) AS num_facturas
    > FROM ventas_part
    > GROUP BY CustomerID
    > ORDER BY num_facturas DESC
    > LIMIT 10;
Query ID = hadoop_20260108180749_2317a48c-8f75-4361-9571-bf4b95333cfc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 4.06 s
--------------------------------------------------------------------------------
```

Análisis de productos

3. 10 productos más vendidos

```
hive> SELECT Description, SUM(Quantity * UnitPrice) AS ingreso
    > FROM ventas_part
    > GROUP BY Description
    > ORDER BY ingreso DESC
    > LIMIT 10;
Query ID = hadoop_20260108180900_b60334fb-cb9f-424e-849e-0efc5555484f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 1.53 s
--------------------------------------------------------------------------------
```

4. 10 Productos más rentables (suma de precio unitario por cantidad) **Análisis geográfico**

```
hive> SELECT Description, SUM(Quantity * UnitPrice) AS ingreso
    > FROM ventas_part
    > GROUP BY Description
    > ORDER BY ingreso DESC
    > LIMIT 10;
Query ID = hadoop_20260108180900_b60334fb-cb9f-424e-849e-0efc5555484f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 1.53 s
--------------------------------------------------------------------------------
```

5. Países con mayor volumen de ventas

```
hive> SELECT Country, SUM(Quantity * UnitPrice) AS ventas
    > FROM ventas_part
    > GROUP BY Country
    > ORDER BY ventas DESC;
Query ID = hadoop_20260108181040_23924ac0-bce8-48bd-bc98-f19b5d5dedd3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 2.52 s
--------------------------------------------------------------------------------
```

Análisis temporal

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

```
hive> SELECT anio, mes, SUM(Quantity * UnitPrice) AS ventas
    > FROM ventas_part
    > GROUP BY anio, mes
    > ORDER BY anio, mes;
Query ID = hadoop_20260108181123_453a0efa-2b33-4562-8d61-5be9ab151e94
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 2.04 s
--------------------------------------------------------------------------------
```

7. 7. Hora del día con más actividad

```
hive> SELECT hour(InvoiceDate) AS hora, COUNT(*) AS registros
    > FROM ventas_part
    > GROUP BY hour(InvoiceDate)
    > ORDER BY registros DESC;
Query ID = hadoop_20260108181154_79df9c6e-fa93-48cc-9b72-3f119b33576f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1767892880359_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
Reducer 2        container        INITED      2          0        0        2       0       0
Reducer 3        container        INITED      1          0        0        1       0       0
--------------------------------------------------------------------------------
VERTICES: 00/03  [>>--------------------------] 0%    ELAPSED TIME: 2.04 s
--------------------------------------------------------------------------------
```

CONTENIDO

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`.

```
OK
Time taken: 8.184 seconds
hive> CREATE DATABASE retail_mysql;
OK
Time taken: 0.038 seconds
hive>
```

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

```
hive> USE retail_mysql;
OK
Time taken: 0.015 seconds
hive> CREATE TABLE ventas (
    >     InvoiceNo VARCHAR(20),
    >     StockCode VARCHAR(20),
    >     Description VARCHAR(255),
    >     Quantity INT,
    >     InvoiceDate DATETIME,
    >     UnitPrice DOUBLE,
    >     CustomerID VARCHAR(20),
    >     Country VARCHAR(50)
    > );
FAILED: SemanticException [Error 10099]: DATETIME type isn't supported yet. Please use DATE or TIMESTAMP instead
hive>
```