

Proyecto Integrador

Data Science

Módulo 4

Realizado Por:

Jovany Zapata Hernández:

Pasos y pantallazos de resultados

Para implementar ejecute

- git clone https://github.com/lopezdar222/herramientas_big_data
- cd herramientas_big_data
- sudo docker-compose -f docker-compose-v1.yml up -d

```
ubuntu@servidor_ubuntu: ~$ sudo docker ps -a
[sudo] password for ubuntu:
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS                                NAMES
ad96857db965   bde2020/hive:2.3.2-postgresql-metastore  "/entrypoint.sh /opt/..."  10 hours ago   Exited (255)  About a minute ago          10000/tcp, 0.0.0.0:9083->9083/tcp, :::9083->9083/tcp, 10002/tcp   hive-server
f9f634d8dace   bde2020/hive:2.3.2-postgresql-metastore  "/entrypoint.sh /opt/..."  10 hours ago   Exited (255)  About a minute ago          10000/tcp, 0.0.0.0:9083->9083/tcp, :::9083->9083/tcp, 10002/tcp   hive-metastore
8c5e735994f3   bde2020/hive-metastore-postgresql:2.3.0  "/docker-entrypoint..."  10 hours ago   Exited (255)  About a minute ago          0.0.0.0:5432->5432/tcp, :::5432->5432/tcp   hive-metastore-postgresql
ea8c04f8e0ac   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8  "/entrypoint.sh /run..."  10 hours ago   Up About a minute (healthy)  0.0.0.0:9864->9864/tcp, :::9864->9864/tcp   datanode
7c79ec142cbc   bde2020/hadoop-resource-manager:2.0.0-hadoop3.2.1-java8  "/entrypoint.sh /run..."  10 hours ago   Up 59 seconds (healthy)      8088/tcp                                     resource-manager
1c1c507d021c   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8  "/entrypoint.sh /run..."  10 hours ago   Up About a minute (healthy)  0.0.0.0:9870->9870/tcp, :::9870->9870/tcp, 0.0.0.0:9010->9010/tcp, :::9010->9010/tcp   namenode
2c580b8c5b53   bde2020/hadoop-node-manager:2.0.0-hadoop3.2.1-java8  "/entrypoint.sh /run..."  10 hours ago   Up About a minute (healthy)  8042/tcp                                     node-manager
7c693ddf4e7b   bde2020/hadoop-history-server:2.0.0-hadoop3.2.1-java8  "/entrypoint.sh /run..."  10 hours ago   Up About a minute (healthy)  8188/tcp                                     history-server
```

Paso 1

1) HDFS

sudo docker exec -it namenode bash

- cd home
- mkdir Datasets
- exit

Nota: este paso no es necesario porque podemos copiar todo el directorio en la ruta destino.

Los 3 puntos anteriores ya no se realizarán puesto que vamos a copiar todo el directorio completo y lo pasaremos a la nueva ruta, así:

- `sudo docker cp ./Datasets namenode:/home/Datasets`

Ubicarse en el contenedor "namenode"

```
sudo docker exec -it namenode bash
```

Crear un directorio en HDFS llamado "/data".

- `hdfs dfs -mkdir -p /data` : verificar como consulta y donde queda el directorio data

Copiar los archivos csv provistos a HDFS:

- `hdfs dfs -put /home/Datasets/* /data`

Nota: Busque dfs.blocksize y dfs.replication en http://<IP_Anfitrion>:9870/conf para encontrar los valores de tamaño de bloque y factor de réplica respectivamente entre otras configuraciones del sistema Hadoop.

The screenshot shows the Hadoop Namenode Overview page in a web browser. The browser address bar shows the URL `192.168.1.11:9870/dfshealth.html#tab-overview`. The page has a green navigation bar with tabs: Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled "Overview 'namenode:9000' (active)". Below the title is a table with the following information:

Started:	Sat Aug 24 20:27:18 -0500 2024
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 10:56:00 -0500 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-1b28a556-e354-4099-a497-fa0198660eb6
Block Pool ID:	BP-1577670739-172.18.0.6-1724549233587

Below the table is a "Summary" section. It contains the following text:

Security is off.
Safemode is off.
44 files and directories, 24 blocks (24 replicated blocks, 0 erasure coded block groups) = 68 total filesystem object(s).
Heap Memory used 32.57 MB of 44.49 MB Heap Memory. Max Heap Memory is 483.38 MB.
Non-Heap Memory used 51.82 MB of 53.06 MB Committed Non-Heap Memory. Max Non-Heap Memory is <unbounded>.

The bottom of the screenshot shows a Windows taskbar with various application icons and a system tray showing the date and time as 8:49 PM on 8/24/2024.

Browsing HDFS

No es seguro 192.168.1.11:9870/explorer.html#/data

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Browse Directory

Show

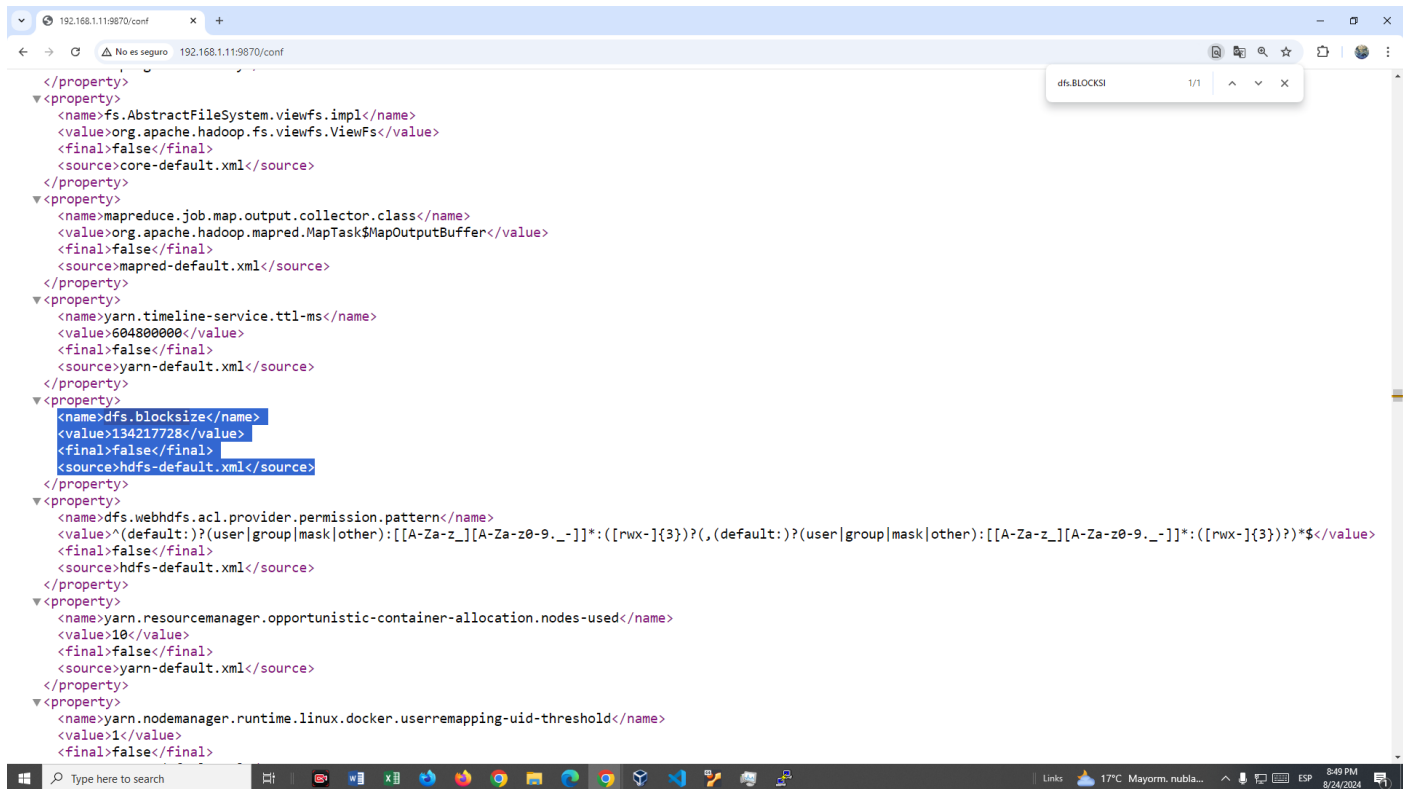
25

entries

Search:

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	4.57 KB	Aug 24 22:12	3	128 MB	Paso02.hql	
<input type="checkbox"/>	-rw-r--r--	root	supergroup	15.93 KB	Aug 24 20:40	3	128 MB	airports.csv	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	calendario	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	canaldeventa	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	cliente	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	compra	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	data_nvo	
<input type="checkbox"/>	drwxr-xr-x	root	supergroup	0 B	Aug 24 20:40	0	0 B	empleado	

dfs.replication se replica por 3 según estándar de Hsdf.



The screenshot shows a web browser window displaying an XML configuration file. The XML contains several property elements. The property `dfs.blocksize` is highlighted with a blue background. Its value is `134217728`, and it is set to `false` for the final configuration. The source of the configuration is `hdfs-default.xml`. Other visible properties include `fs.AbstractFileSystem.viewfs.impl`, `mapreduce.job.map.output.collector.class`, `yarn.timeline-service.ttl-ms`, `dfs.webhdfs.acl.provider.permission.pattern`, `yarn.resourcemanager.opportunistic-container-allocation.nodes-used`, and `yarn.nodemanager.runtime.linux.docker.userremapping-uid-threshold`.

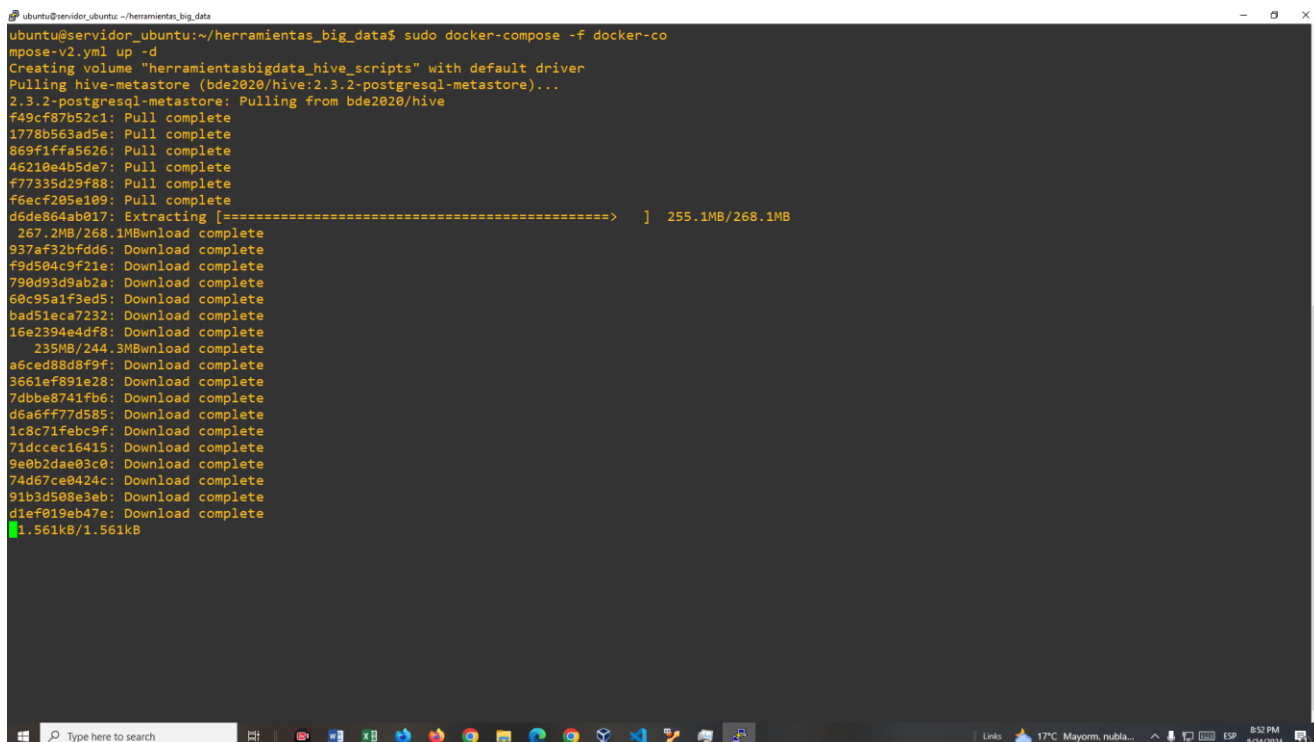
```
</property>
<property>
  <name>fs.AbstractFileSystem.viewfs.impl</name>
  <value>org.apache.hadoop.fs.viewfs.ViewFs</value>
  <final>false</final>
  <source>core-default.xml</source>
</property>
<property>
  <name>mapreduce.job.map.output.collector.class</name>
  <value>org.apache.hadoop.mapred.MapTask$MapOutputBuffer</value>
  <final>false</final>
  <source>mapred-default.xml</source>
</property>
<property>
  <name>yarn.timeline-service.ttl-ms</name>
  <value>60480000</value>
  <final>false</final>
  <source>yarn-default.xml</source>
</property>
<property>
  <name>dfs.blocksize</name>
  <value>134217728</value>
  <final>false</final>
  <source>hdfs-default.xml</source>
</property>
<property>
  <name>dfs.webhdfs.acl.provider.permission.pattern</name>
  <value>^(default:)?(user|group|mask|other):[[A-Za-z_][A-Za-z0-9_-]]*:([rxw-]{3})?((default:)?(user|group|mask|other):[[A-Za-z_][A-Za-z0-9_-]]*:([rxw-]{3})?)*$</value>
  <final>false</final>
  <source>hdfs-default.xml</source>
</property>
<property>
  <name>yarn.resourcemanager.opportunistic-container-allocation.nodes-used</name>
  <value>10</value>
  <final>false</final>
  <source>yarn-default.xml</source>
</property>
<property>
  <name>yarn.nodemanager.runtime.linux.docker.userremapping-uid-threshold</name>
  <value>1</value>
  <final>false</final>
</property>
```

Paso 2

2) Hive

Creamos las tablas y demás desde el el Yaml 2

- `sudo docker-compose -f docker-compose-v2.yml up -d`



The screenshot shows a terminal window with the command `sudo docker-compose -f docker-compose-v2.yml up -d` being executed. The output shows the creation of a volume, pulling the `hive` image, and downloading various components. The process is complete, and the status is `1.561kB/1.561kB`.

```
ubuntu@servidor_ubuntu: ~/herramientas_big_data
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker-compose -f docker-co
mpose-v2.yml up -d
Creating volume "herramientasbigdata_hive_scripts" with default driver
Pulling hive-metastore (bde2020/hive:2.3.2-postgresql-metastore)...
2.3.2-postgresql-metastore: Pulling from bde2020/hive
f49cf87b52c1: Pull complete
1778b563ad5e: Pull complete
869f1ffa5626: Pull complete
46210e4b5de7: Pull complete
f77335d29f88: Pull complete
f6ecf205e109: Pull complete
d6de864ab017: Extracting [=====] 255.1MB/268.1MB
267.2MB/268.1MBwnload complete
937af32bfdd6: Download complete
f9d504c9f21e: Download complete
790d93d9ab2a: Download complete
60c95a1f3ed5: Download complete
bad51eca7232: Download complete
16e2394e4df8: Download complete
235MB/244.3MBwnload complete
a6ced88d8f9f: Download complete
3661ef891e28: Download complete
7dbbe8741fb6: Download complete
d6a6ff77d585: Download complete
1c8c71febc9f: Download complete
71dccec16415: Download complete
9e0b2dae03c0: Download complete
74d67ce0424c: Download complete
91b3d508e3eb: Download complete
d1ef019eb47e: Download complete
1.561kB/1.561kB
```

Resultado final

```
ubuntu@servidor_ubuntu: ~/herramientas_big_data
a6ced88d8f9f: Pull complete
3661ef891e28: Pull complete
7dbbe8741fb6: Pull complete
d6a6ff77d585: Pull complete
1c8c71feb9cf: Pull complete
71dccc16415: Pull complete
9a0b2dae03c0: Pull complete
74d67ce0424c: Pull complete
91b3d508e3eb: Pull complete
d1ef019eb47e: Pull complete
Digest: sha256:620267768985bb57e52a86db9263a354e92d02319d835678852539b21e0895
Status: Downloaded newer image for bde2020/hive:2.3.2-postgresql-metastore
Pulling hive-metastore-postgresql (bde2020/hive-metastore-postgresql:2.3.0)...
2.3.0: Pulling from bde2020/hive-metastore-postgresql
5c90d4a2d1a8: Pull complete
22337bfd13a9: Pull complete
c3961b297acc: Pull complete
5a17453338b4: Pull complete
6364e0d7a283: Pull complete
58c25f5c0dad: Pull complete
f0e675ce88d9: Pull complete
10f26c680a34: Pull complete
873d2c220bff: Pull complete
fd10fb78ded6: Pull complete
ff1356ba118b: Pull complete
8161ea5e47f1: Pull complete
b399213c70b6: Pull complete
08bd4e9a6388: Pull complete
Digest: sha256:9ab91699d1511b874829e6572006cd9d9f1cca413f438b6f21c65b412152bf1
Status: Downloaded newer image for bde2020/hive-metastore-postgresql:2.3.0
resourceManager is up-to-date
Creating hive-metastore-postgresql ...
datanode is up-to-date
historyserver is up-to-date
Creating hive-metastore ...
Creating hive-metastore-postgresql
nodemanager is up-to-date
namenode is up-to-date
Creating hive-server ...
Creating hive-metastore
Creating hive-metastore-postgresql ... done
ubuntu@servidor_ubuntu:~/herramientas_big_data$
```

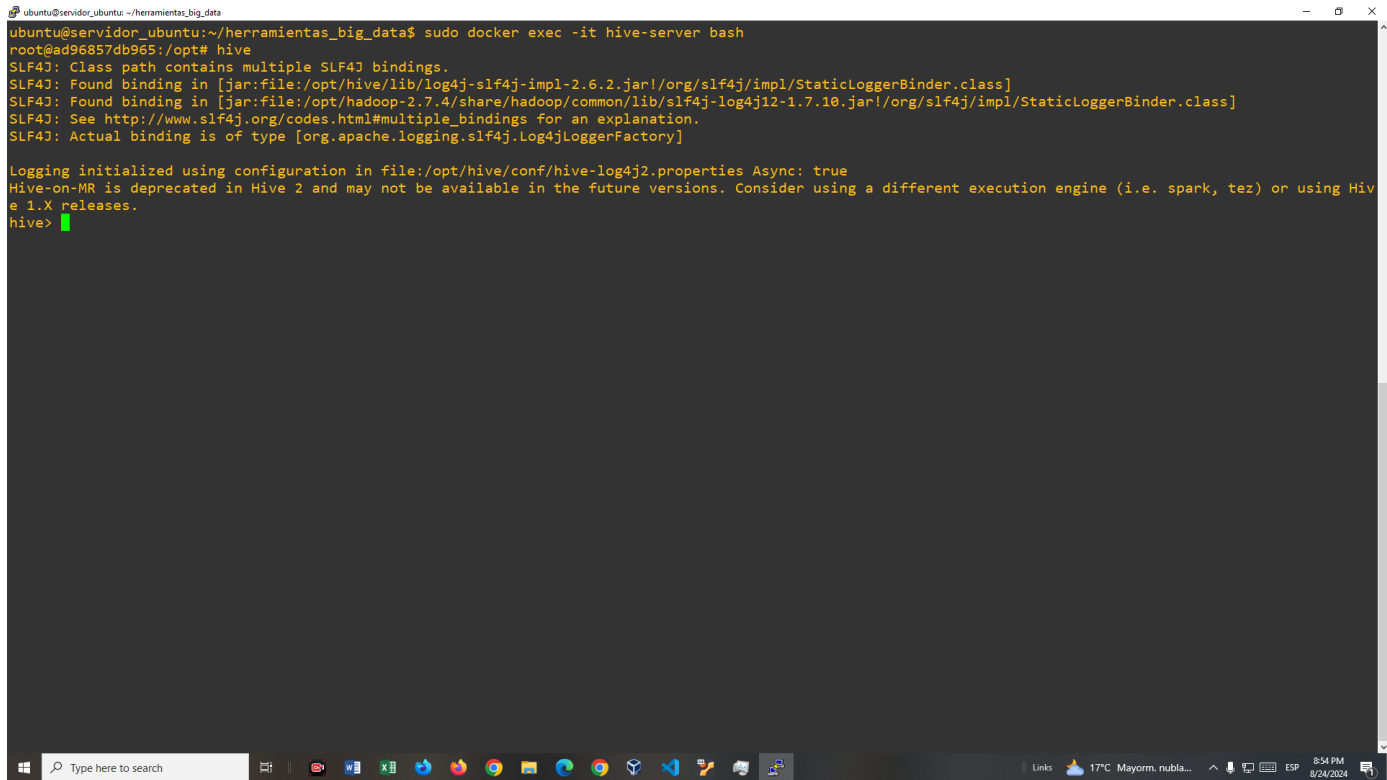
Resultado

```
ubuntu@servidor_ubuntu:~$ sudo docker --version
Docker version 20.10.21, build b0f9bc2f
ubuntu@servidor_ubuntu:~$ sudo docker ps -a
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS
ad96372b0a5   bde2020/hive:2.3.2-postgresql-metastore  "entrypoint.sh /opt/..." 11 hours ago   Exited (255) 39 minutes ago  0.0.0.0:10000->10000/tcp, :::10000->10000/tcp, 10002/tcp
f9f634d8cace   bde2020/hive:2.3.2-postgresql-metastore  "entrypoint.sh /opt/..." 11 hours ago   Exited (255) 39 minutes ago  10000/tcp, 0.0.0.0:9083->9083/tcp, :::9083->9083/tcp, 10002/tcp
8c6e735994f3   bde2020/hive-metastore-postgresql:2.3.0  "/docker-entrypoint...." 11 hours ago   Exited (255) 39 minutes ago  0.0.0.0:5432->5432/tcp, :::5432->5432/tcp
e8b6c4ffebac   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-javab  "/entrypoint.sh /run..." 11 hours ago   Up 39 minutes (healthy)  0.0.0.0:3064->3064/tcp, :::3064->3064/tcp
7c79ee142c8c   bde2020/hadoop-resourceManager:2.0.0-hadoop3.2.1-javab  "/entrypoint.sh /run..." 11 hours ago   Up 38 minutes (healthy)  8088/tcp
1c6c587a923c   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-javab  "/entrypoint.sh /run..." 11 hours ago   Up 39 minutes (healthy)  0.0.0.0:9870->9870/tcp, :::9870->9870/tcp, 0.0.0.0:9010->9000/tcp, :::9010->9000/tcp
1c30b0c3d363   bde2020/hadoop-nodemanager:2.0.0-hadoop3.2.1-javab  "/entrypoint.sh /run..." 11 hours ago   Up 39 minutes (healthy)  8042/tcp
7c693ddfae7b   bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-javab  "/entrypoint.sh /run..." 11 hours ago   Up 39 minutes (healthy)  8188/tcp
ubuntu@servidor_ubuntu:~$ sudo docker service ls
ID            NAME              REPLICAS  IMAGE              PORTS
ubuntu@servidor_ubuntu:~$ sudo docker image ls
REPOSITORY    TAG              IMAGE ID          CREATED           SIZE
bde2020/hive-metastore-postgresql  2.3.0           7ab9e8f93813     4 years ago      279MB
bde2020/hadoop-nodemanager          2.0.0-hadoop3.2.1-javab  4e47dad9d148f     4 years ago      1.37GB
bde2020/hadoop-resourceManager      2.0.0-hadoop3.2.1-javab  36d9a418b5f       4 years ago      1.37GB
bde2020/hadoop-namenode             2.0.0-hadoop3.2.1-javab  839ec11d95f8      4 years ago      1.37GB
bde2020/hadoop-historyserver        2.0.0-hadoop3.2.1-javab  173c52d1f624      4 years ago      1.37GB
bde2020/hadoop-datanode             2.0.0-hadoop3.2.1-javab  df208ee8a7f9      4 years ago      1.37GB
bde2020/hive                        2.3.2-postgresql-metastore  87f5c9f4e2df      6 years ago      1.17GB
ubuntu@servidor_ubuntu:~$ sudo docker volume ls
DRIVER    VOLUME NAME
local     c58548bee154230872c26d19bdc94afab1baef1c0f555ca4b167f5bdc279
local     herramientasbigdata_hadoop_datanode
local     herramientasbigdata_hadoop_datanode
local     herramientasbigdata_hadoop_historyserver
local     herramientasbigdata_hadoop_namenode
local     herramientasbigdata_hive_scripts
ubuntu@servidor_ubuntu:~$ sudo docker network ls
NETWORK ID    NAME                DRIVER    SCOPE
6e8d99ef1276  bridge              bridge    local
3276f941c984  docker_gwbridge     bridge    local
f022939e1a41  herramientasbigdata_default  bridge    local
442040e3e311  host                 host      local
qp3bqbc8ffa  ingress              overlay   swarm
1fcc49d6d46   none                 null      local
ubuntu@servidor_ubuntu:~$
```

Crear tablas en Hive, a partir de los csv ingestados en HDFS.

Para esto, se puede ubicar dentro del contenedor correspondiente al servidor de Hive, y ejecutar desde allí los scripts necesarios

- `sudo docker exec -it hive-server bash`
- `hive`



```
ubuntu@servidor_ubuntu: ~/herramientas_big_data
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker exec -it hive-server bash
root@ad96857db965:/opt# hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

Este proceso de creación las tablas debe poder ejecutarse desde un shell script.

Nota: Para ejecutar un script de Hive, requiere el comando:

`hive -f <script.hql>`

- `hive -f hdfs:///data/paso02.hql` -- función al inicio pero luego no, use la forma siguiente:
- `hive -f Paso02.hql`

Listas las bases de datos creadas

```
ubuntu@servidor_ubuntu: ~
bash: show: command not found
root@ad96857db965:/opt# hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Staticlog
gerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a
different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show databases;
OK
default
integrador
Time taken: 1.051 seconds, Fetched: 2 row(s)
hive> use integrador
> show tables;
FAILED: ParseException line 2:0 missing EOF at 'show' near 'integrador'
hive> show databases;
OK
default
integrador
Time taken: 0.023 seconds, Fetched: 2 row(s)
hive> use integrador;
OK
Time taken: 0.035 seconds
hive> show tables;
OK
calendario
canal_venta
cliente
compra
empleado
gasto
producto
proveedor
sucursal
tipo_gasto
venta
Time taken: 0.051 seconds, Fetched: 11 row(s)
hive> select * from cliente limit 10;
OK
1      HEBER JONI SANTANA      LAS MERAS Y BAT. 24 DE FERRERO 4150 RINCON DEL CAZADOR 42-5
161    58      LOHA VERDE      -58,81858367      -34,38997888
2      Buenos Aires      ANA SAPOLIZA      RUEYREDON Y BURYU RUTAS KM 52.500 S/N B&A LOS POZOS 49-7
578    61      SANTA ROSA      -58,73072751      -34,93085311
3      Buenos Aires      FERNANDO LUIS SARALEGUI CALDERON DE LA BARCA 498      49-3435 15      TORR
ES      -59,1274868      -34,4882159
4      Buenos Aires      MONICA SARASOLA      RUTA 36 KM 45,500 S/N EL PELIGRO      49-2883 29 R
UTA SOL -58,14393954      -34,82052786
5      Buenos Aires      MARIO RAUL SARASUA      492 Y 186 S/N COLONIA URQUIZA 491-6688      34 3
OSE MELCHOR ROMERO      -58,809381      -34,94484721
6      Buenos Aires      PEDRO JESUS SARAVIA      RUTA 2 - KM 44,500 S/N EL PELIGRO      49-2350 18 R
UTA SOL -58,11226426      -35,80786216
7      Buenos Aires      JORGE SARAVIA      VILLARROEL RUTA 3 KM 46500 S/N BARRIO SAN MARIANO 49-5
386    21      VIRREY DEL PINO -58,78894814      -34,86870786
8      Buenos Aires      CARLOS JOSE SARAZOLA      ISLA SANTIAGO S/N ISLA SANTIAGO      623-9935 4
0      ISLA SANTIAGO      -57,88324209      -34,8350311
9      Buenos Aires      OSCAR LUIS SARLO      GARCILAZO DE LA VEGA Y SAN MARTIN S/N SANTA ROSA 4
9-7576 18      SANTA ROSA      -58,75888438      -34,97534955
10     Buenos Aires      JOSE ADOLFO SARZENTO      SEGUIRO SOBERA E/MGALLANES Y P. GALDO S/N SANTA RO
SA      49-7565 58      SANTA ROSA      -58,75283716      -34,95142843
Time taken: 2.453 seconds, Fetched: 10 row(s)
hive>
```

```
ubuntu@servidor_ubuntu: ~
root@ad96857db965:/opt# hive -f /data/paso02.hql
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Staticlog
gerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Could not open input file for reading. (File file:/data/paso02.hql does not exist)
root@ad96857db965:/opt# ^C
root@ad96857db965:/opt# hive -f hdfs:///data/Paso02.hql
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Staticlog
gerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
OK
Time taken: 2.021 seconds
OK
Time taken: 0.054 seconds
OK
Time taken: 0.155 seconds
OK
Time taken: 1.539 seconds
OK
Time taken: 0.03 seconds
OK
Time taken: 0.132 seconds
OK
Time taken: 0.029 seconds
OK
Time taken: 0.096 seconds
OK
Time taken: 0.023 seconds
OK
Time taken: 0.128 seconds
OK
Time taken: 0.021 seconds
OK
Time taken: 0.086 seconds
OK
Time taken: 0.019 seconds
OK
Time taken: 0.112 seconds
OK
Time taken: 0.018 seconds
OK
Time taken: 0.069 seconds
OK
Time taken: 0.031 seconds
```

```
ubuntu@servidor_ubuntu: ~  
OK  
Time taken: 0.112 seconds  
OK  
Time taken: 0.010 seconds  
OK  
Time taken: 0.069 seconds  
OK  
Time taken: 0.031 seconds  
OK  
Time taken: 0.13 seconds  
OK  
Time taken: 0.019 seconds  
OK  
Time taken: 0.091 seconds  
OK  
Time taken: 0.03 seconds  
OK  
Time taken: 0.089 seconds  
OK  
Time taken: 0.031 seconds  
OK  
Time taken: 0.076 seconds  
root@ad96857db965:/opt# show databases;  
bash: show: command not found  
root@ad96857db965:/opt# show database;  
bash: show: command not found  
root@ad96857db965:/opt# hive  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found Binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLog  
gerBinder.class]  
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]  
  
Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true  
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a  
different execution engine (i.e. spark, tez) or using Hive 1.X releases.  
hive> show databases;  
OK  
default  
integrador  
Time taken: 1.051 seconds, Fetched: 2 row(s)  
hive> use integrador  
hive> show tables;  
FAILED: ParseException line 2:0 missing EOF at 'show' near 'integrador'  
hive> show databases;  
OK  
default  
integrador  
Time taken: 0.023 seconds, Fetched: 2 row(s)  
hive> use integrador;  
OK  
Time taken: 0.035 seconds  
hive> show tables;  
OK
```

```
ubuntu@servidor_ubuntu: ~  
hive> show databases;  
OK  
default  
integrador  
Time taken: 1.051 seconds, Fetched: 2 row(s)  
hive> use integrador  
hive> show tables;  
FAILED: ParseException line 2:0 missing EOF at 'show' near 'integrador'  
hive> show databases;  
OK  
default  
integrador  
Time taken: 0.023 seconds, Fetched: 2 row(s)  
hive> use integrador;  
OK  
Time taken: 0.035 seconds  
hive> show tables;  
OK  
calendario  
canal_venta  
cliente  
coopa  
empleado  
gasto  
producto  
proveedor  
sucursal  
tipo_gasto  
venta  
Time taken: 0.051 seconds, Fetched: 11 row(s)  
hive> select * from cliente limit 10;  
OK  
1 HEBER JONI SANTANA LAS HERAS Y BAT. 24 DE FEBRERO 4150 RINCON DEL CAZADOR 42-5  
161 SB LOMA VERDE -58,81850307 -34,30997888  
2 Buenos Aires ANA SAPRIZA PUEYREDON Y DUPLY RUTA3 KM 52.500 S/N BA LOS POZOS 49-7  
578 61 SANTA ROSA -58,73073751 -34,93908311  
3 Buenos Aires FERNANDO LUIS SARALEGUI CALDERON DE LA BARCA 498 49-3435 15 TORR  
ES -59,12794068 -34,43082199  
4 Buenos Aires MANUELA SARASOLA RUTA 36 KM 45,500 S/N EL PELIGRO 49-2883 29 R  
UTA SOL -58,14393954 -34,92852706  
5 Buenos Aires MARIO RAUL SARASUA 492 Y 106 S/N COLONIA URQUIZA 491-4608 34 J  
OSE MELCHOR ROMERO -58,089381 -34,9444471  
6 Buenos Aires PEDRO JESUS SARAVIA RUTA 2 - KM 44,500 S/N EL PELIGRO 49-2350 18 R  
UTA SOL -58,11220426 -35,00786216  
7 Buenos Aires JORGE SARAVIA VILLARROEL RUTA 3 KM 46500 S/N BARRIO SAN MARIANO 49-5  
306 21 VIRREY DEL PINO -58,70894814 -34,86870786  
8 Buenos Aires CARLOS JOSE SARAZOLA ISLA SANTIAGO S/N ISLA SANTIAGO 623-9935 4  
0 ISLA SANTIAGO -57,88154205 -34,8350313  
9 Buenos Aires OSCAR LUIS SARLO GARCILAZO DE LA VEGA Y SAN MARTIN S/N SANTA ROSA 4  
9-7576 18 SANTA ROSA -58,75008438 -34,97534955  
10 Buenos Aires JOSE ADOLFO SARMIENTO SEGUNDO SOMBRERA/MACALLANES Y P. GALDO S/N SANTA RO  
SA 49-7565 58 SANTA ROSA -58,75203716 -34,95142843  
Time taken: 2.453 seconds, Fetched: 10 row(s)  
hive> hive -f Paso03.hql  
hive> select * from cliente limit 10;  
NoViableAltException(240[1])
```


Paso 3

Creamos

```
ubuntu@servidor_ubuntu: ~/herramientas_big_data
ubuntu@servidor_ubuntu:~$ cd herramientas_big_data/1
-bash: cd: herramientas_big_data/1: No such file or directory
ubuntu@servidor_ubuntu:~$ cd herramientas_big_data/
ubuntu@servidor_ubuntu:~/herramientas_big_data$ ls
Datasets      Parquet      Paso01.sh      Paso03.hql      Paso05.py      README.md
Generacion_Ventas.ipynb  Particion_compra.hql  Paso02.hql      Paso04.hql      Paso06_GeneracionVentasNuevasPorDia.py  docker-compose-kafka.y
Mongo         Paso00.sh      Paso02_ConConsultas.hql  Paso04_ConConsulta.hql  Paso06_IncrementalVentas.py  docker-compose-v1.yml
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker cp Particion_compra.hql hive-server:/opt/
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker exec -it hive-server bashsudo docker cp Particion_compra.hql herramientas_big_data/^C
root@ad96857db965:/opt# ls
Particion_compra.hql Paso03.hql Paso04.hql hadoop-2.7.4 hive
root@ad96857db965:/opt# hive -f Particion_compra.hql
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
OK
Time taken: 2.226 seconds
OK
Time taken: 0.043 seconds
OK
Time taken: 0.126 seconds
OK
Time taken: 0.256 seconds
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20240826084222_fdfaa5b1-0ee3-4be2-9291-b5537e6c2022
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2024-08-26 00:42:28,785 Stage-1 map = 0%, reduce = 0%
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
2024-08-26 00:42:30,807 Stage-1 map = 100%, reduce = 0%
Ended Job = job_local1674400364_0001
Stage-4 is selected by condition resolver.
```

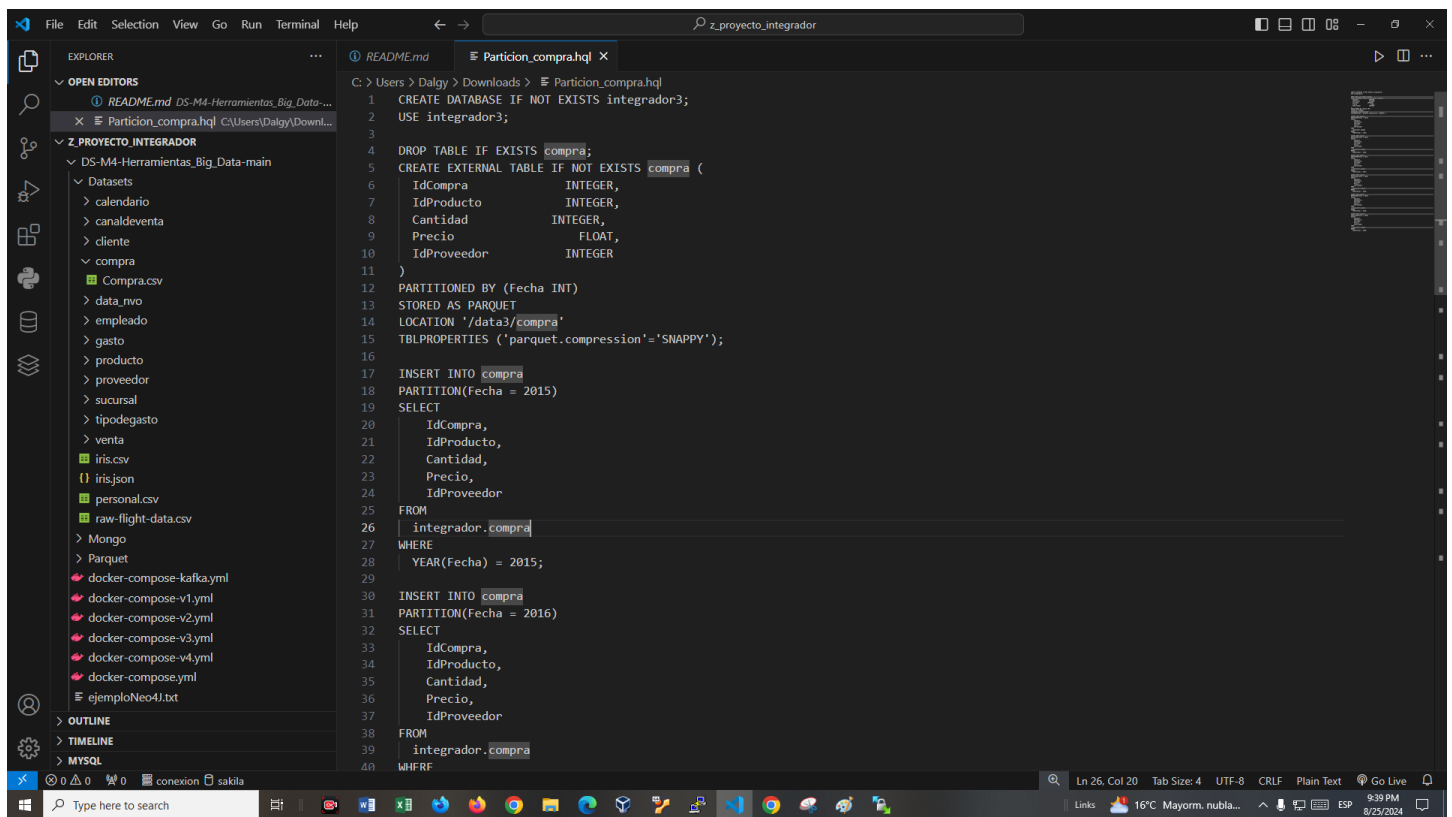
```
ubuntu@servidor_ubuntu: ~/herramientas_big_data
11529 42932 2 109.92 5 2020
11530 42973 4 83.08 6 2020
11531 42974 5 131.36 7 2020
11532 42975 6 143.62 4 2020
11533 42976 5 143.89 1 2020
11534 42977 5 72.95 4 2020
11535 42978 6 88.29 10 2020
11536 42979 2 81.24 10 2020
11537 42980 1 71.67 3 2020
11538 42981 3 35.55 9 2020
11539 42982 4 9.4 6 2020
Time taken: 0.272 seconds, Fetched: 2595 row(s)
hive> use integrador3;
OK
Time taken: 0.016 seconds
hive> select * from compra where fecha = 2020 limit 10;
OK
8945 42832 4 523.27 13 2020
8946 42833 6 621.58 9 2020
8947 42834 6 538.44 9 2020
8948 42835 11 585.74 10 2020
8949 42836 11 553.44 11 2020
8950 42837 14 287.34 8 2020
8951 42839 14 236.09 14 2020
8952 42841 5 221.7 4 2020
8953 42844 18 1367.73 3 2020
8954 42845 24 574.65 5 2020
Time taken: 0.276 seconds, Fetched: 10 row(s)
hive> select * from compra where fecha = 2015 limit 10;
OK
1 42832 13 560.51 12 2015
2 42833 11 497.58 7 2015
3 42834 1 588.5 6 2015
4 42835 9 567.66 14 2015
5 42839 14 231.31 2 2015
6 42840 14 232.07 13 2015
7 42841 8 236.98 4 2015
8 42842 4 255.33 4 2015
9 42845 5 578.61 12 2015
10 42855 1 809.04 6 2015
Time taken: 0.281 seconds, Fetched: 10 row(s)
hive>
```

<https://ed.team/blog/mover-copiar-y-renombrar-directorios-en-linux>

Realizamos nuestro propio script HQL, para crear una nueva base de datos y una tabla para realizarle la partición, con su directorio hdfs, lo pasamos al ecosistema de hadoop en su cluster, y luego a su respectiva ubicación en hive.

Lo corremos:

- **Hive -f Partition_compras**



The screenshot shows a code editor with a file explorer on the left and a terminal at the bottom. The file explorer shows a project named 'Z_PROYECTO_INTEGRADOR' with various datasets and files. The main editor displays an HQL script in a file named 'Partition_compra.hql'. The script creates a database 'integrador3', drops an existing table 'compra', and creates a new external table 'compra' partitioned by 'Fecha' (year). The table has columns: IdCompra (INTEGER), IdProducto (INTEGER), Cantidad (INTEGER), Precio (FLOAT), and IdProveedor (INTEGER). It is stored as PARQUET in the location '/data3/compra' with SNAPPY compression. The script then inserts data from 'integrador.compra' into the new 'compra' table for the years 2015 and 2016.

```
1 CREATE DATABASE IF NOT EXISTS integrador3;
2 USE integrador3;
3
4 DROP TABLE IF EXISTS compra;
5 CREATE EXTERNAL TABLE IF NOT EXISTS compra (
6     IdCompra          INTEGER,
7     IdProducto         INTEGER,
8     Cantidad           INTEGER,
9     Precio             FLOAT,
10    IdProveedor        INTEGER
11 )
12 PARTITIONED BY (Fecha INT)
13 STORED AS PARQUET
14 LOCATION '/data3/compra'
15 TBLPROPERTIES ('parquet.compression'='SNAPPY');
16
17 INSERT INTO compra
18 PARTITION(Fecha = 2015)
19 SELECT
20     IdCompra,
21     IdProducto,
22     Cantidad,
23     Precio,
24     IdProveedor
25 FROM
26     integrador.compra
27 WHERE
28     YEAR(Fecha) = 2015;
29
30 INSERT INTO compra
31 PARTITION(Fecha = 2016)
32 SELECT
33     IdCompra,
34     IdProducto,
35     Cantidad,
36     Precio,
37     IdProveedor
38 FROM
39     integrador.compra
40 WHERE
```

4) SQL

Crear índices para las tablas,

Crear índices en alguna de las tablas cargadas y probar los resultados:

Procedemos a crear uno de los índices de la tabla venta.

```
CREATE INDEX index_fechaentrega ON TABLE venta(Fecha_Entrega)
AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
WITH DEFERRED REBUILD ;
```

Validamos si quedo creado

- **show indexes on venta;**

Listamos la descripción de la tabla completa:

- **describe formatted venta;**

```
integrador
integrador2
integrador3
Time taken: 0.933 seconds, Fetched: 4 row(s)
hive> use integrador2
> use integrador2;
FAILED: ParseException line 2:0 missing EOF at 'use' near 'integrador2'
hive> use integrador2;
OK
Time taken: 0.025 seconds
hive> CREATE INDEX index_fechaentrega ON TABLE venta(Fecha_Entrega)
> AS 'org.apache.hadoop.hive.q1.index.compact.CompactIndexHandler'
> WITH DEFERRED REBUILD
> ;
OK
Time taken: 0.274 seconds
hive> show INDEXES ON venta;
OK
index_venta_sucursal      venta      idsucursal      integrador2__venta_index_venta_sucursal__      compact
index_fechaentrega      venta      fecha_entrega      integrador2__venta_index_fechaentrega__      compact
Time taken: 0.073 seconds, Fetched: 2 row(s)
hive> DESCRIBE FORMATTED venta;
OK
# col_name      data_type      comment
idventa      int
fecha      date
fecha_entrega      date
idcanal      int
idcliente      int
idsucursal      int
idempleado      int
idproducto      int
precio      float
cantidad      int

# Detailed Table Information
Database:      integrador2
Owner:      root
CreateTime:      Mon Aug 26 00:22:12 UTC 2024
LastAccessTime:      UNKNOWN
Retention:      0
```

Muchas gracias y maravillosa practica en donde en grupo resolvimos algunas dudas.

Un abrazo al equipo HENRY.