Practica integradora proyecto datascience modulo 4

Realizado por Franco Batista

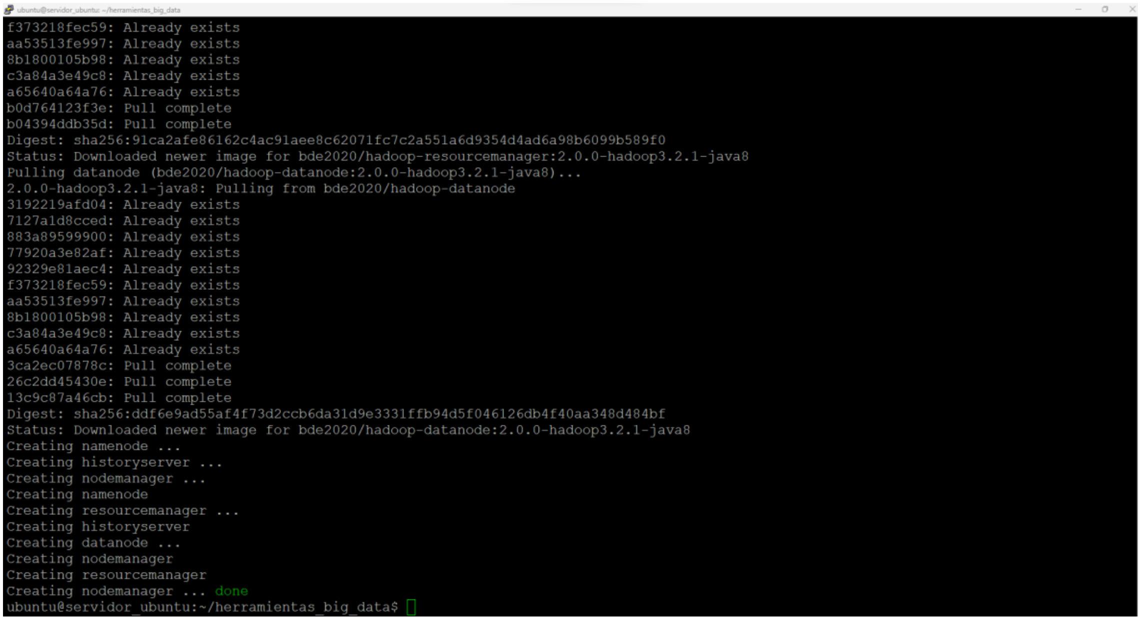
Paso 1

Cargamos el directorio con los archivos desde github:

https://github.com/lopezdar222/herramientas_big_data

Cargamos los archivos de configuracion Docker compouse.yml

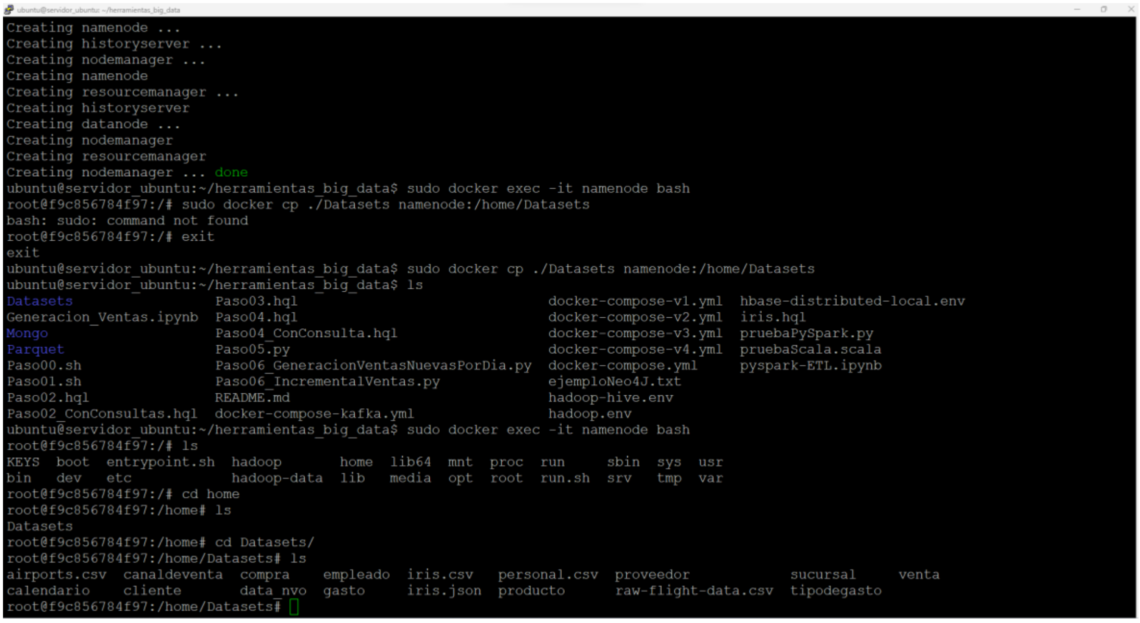
sudo docker-compose -f docker-compose-v1.yml up –d



```
f373218fec59: Already exists
aa53513fe997: Already exists
8b1800105b98: Already exists
c3a84a3e49c8: Already exists
a65640a64a76: Already exists
b0d764123f3e: Pull complete
b04394ddb35d: Pull complete
Digest: sha256:91ca2afe86162c4ac91aee8c62071fc7c2a551a6d9354d4ad6a98b6099b589f0
Status: Downloaded newer image for bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8
Pulling datanode (bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8)...
2.0.0-hadoop3.2.1-java8: Pulling from bde2020/hadoop-datanode
3192219afd04: Already exists
7127a1d8cced: Already exists
883a89599900: Already exists
77920a3e82af: Already exists
92329e81aec4: Already exists
f373218fec59: Already exists
aa53513fe997: Already exists
8b1800105b98: Already exists
c3a84a3e49c8: Already exists
a65640a64a76: Already exists
3ca2ec07878c: Pull complete
26c2dd45430e: Pull complete
13c9c87a46cb: Pull complete
Digest: sha256:ddf6e9ad55af4f73d2ccb6da31d9e3331ffb94d5f046126db4f40aa348d484bf
Status: Downloaded newer image for bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
Creating namenode ...
Creating historyserver ...
Creating nodemanager ...
Creating namenode
Creating resourcemanager ...
Creating historyserver
Creating datanode ...
Creating nodemanager
Creating resourcemanager
Creating nodemanager ... done
ubuntu@servidor_ubuntu:~/herramientas_big_data$ 
```

sudo docker exec -it namenode bash

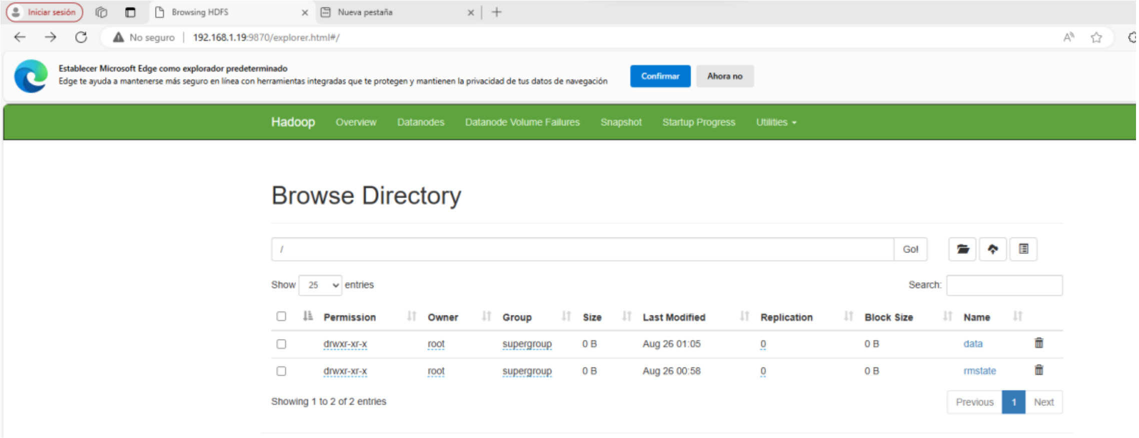
sudo docker cp ./Datasets namenode:/home/Datasets



```
Creating namenode ...
Creating historyserver ...
Creating nodemanager ...
Creating namenode
Creating resourcemanager ...
Creating historyserver
Creating datanode ...
Creating nodemanager
Creating resourcemanager
Creating nodemanager ... done
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker exec -it namenode bash
root@f9c856784f97:/# sudo docker cp ./Datasets namenode:/home/Datasets
bash: sudo: command not found
root@f9c856784f97:/# exit
exit
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker cp ./Datasets namenode:/home/Datasets
ubuntu@servidor_ubuntu:~/herramientas_big_data$ ls
Datasets                 Paso03.hql                              docker-compose-v1.yml  hbase-distributed-local.env
Generacion_Ventas.ipynb  Paso04.hql                              docker-compose-v2.yml  iris.hql
Mongo                    Paso04_ConConsulta.hql                  docker-compose-v3.yml  pruebaPySpark.py
Parquet                  Paso05.py                               docker-compose-v4.yml  pruebaScala.scala
Paso00.sh                Paso06_GeneracionVentasNuevasPorDia.py  docker-compose.yml     pyspark-ETL.ipynb
Paso01.sh                Paso06_IncrementalVentas.py             ejemploNeo4J.txt
Paso02.hql               README.md                               hadoop-hive.env
Paso02_ConConsultas.hql  docker-compose-kafka.yml                hadoop.env
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker exec -it namenode bash
root@f9c856784f97:/# ls
KEYS  boot  entrypoint.sh  hadoop       home  lib64  mnt  proc  run     sbin  sys  usr
bin   dev   etc            hadoop-data  lib   media  opt  root  run.sh  srv   tmp  var
root@f9c856784f97:/# cd home
root@f9c856784f97:/home# ls
Datasets
root@f9c856784f97:/home# cd Datasets/
root@f9c856784f97:/home/Datasets# ls
airports.csv  canaldeventa  compra    empleado  iris.csv   personal.csv  proveedor           sucursal     venta
calendario    cliente       data_nvo  gasto     iris.json  producto      raw-flight-data.csv  tipodegasto
root@f9c856784f97:/home/Datasets# 
```

hdfs dfs -mkdir -p /data

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



Iniciar sesión
Browsing HDFS
Nueva pestaña
No seguro
192.168.1.19:9870/explorer.html#/
Establecer Microsoft Edge como explorador predeterminado
Edge te ayuda a mantenerse más seguro en línea con herramientas integradas que te protegen y mantienen la privacidad de tus datos de navegación
Confirmar
Ahora no
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

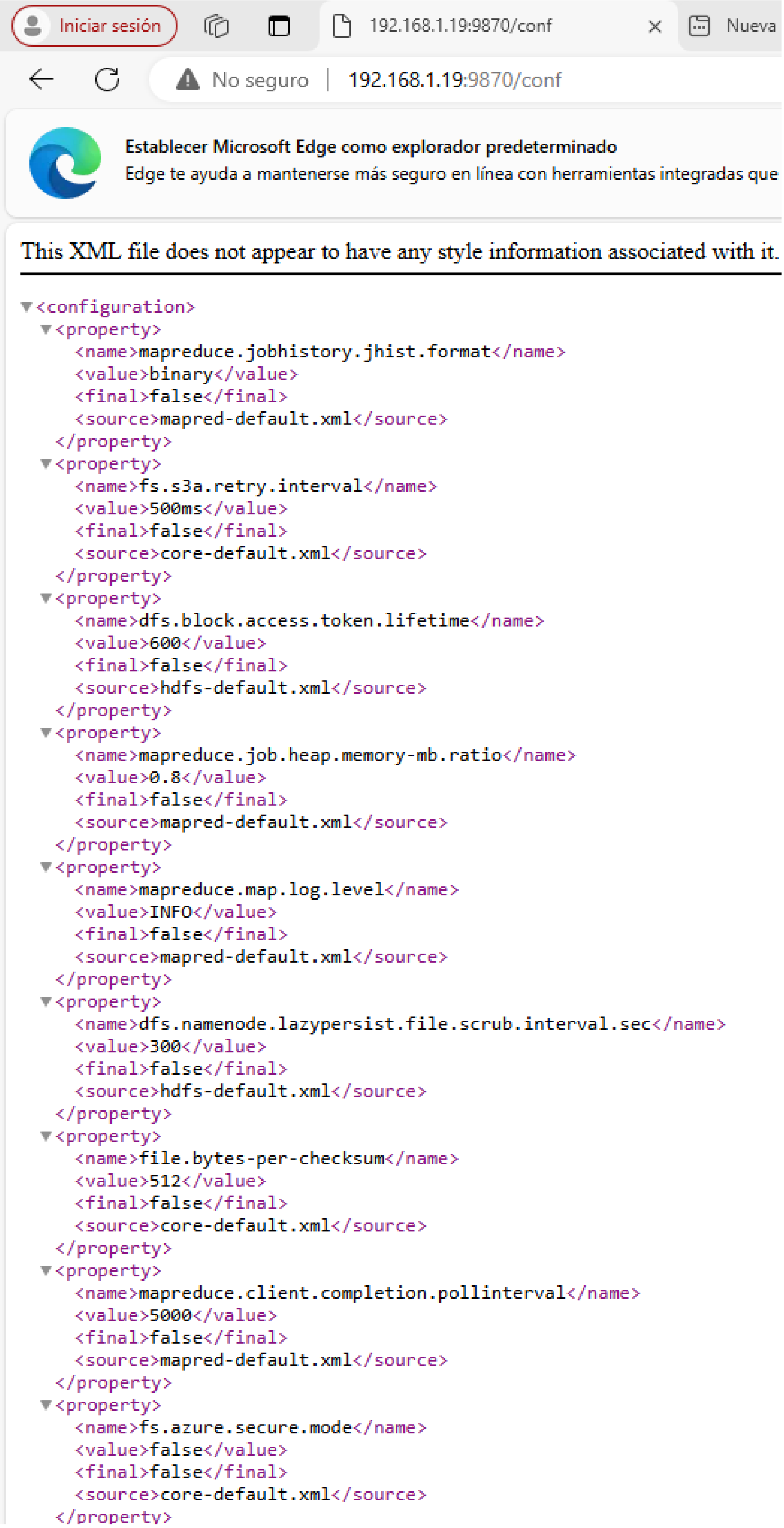
Browse Directory

/ Go!

Show 25 entries   Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:05	0	0 B	data
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 00:58	0	0 B	rmstate

Showing 1 to 2 of 2 entries



Establecer Microsoft Edge como explorador predeterminado

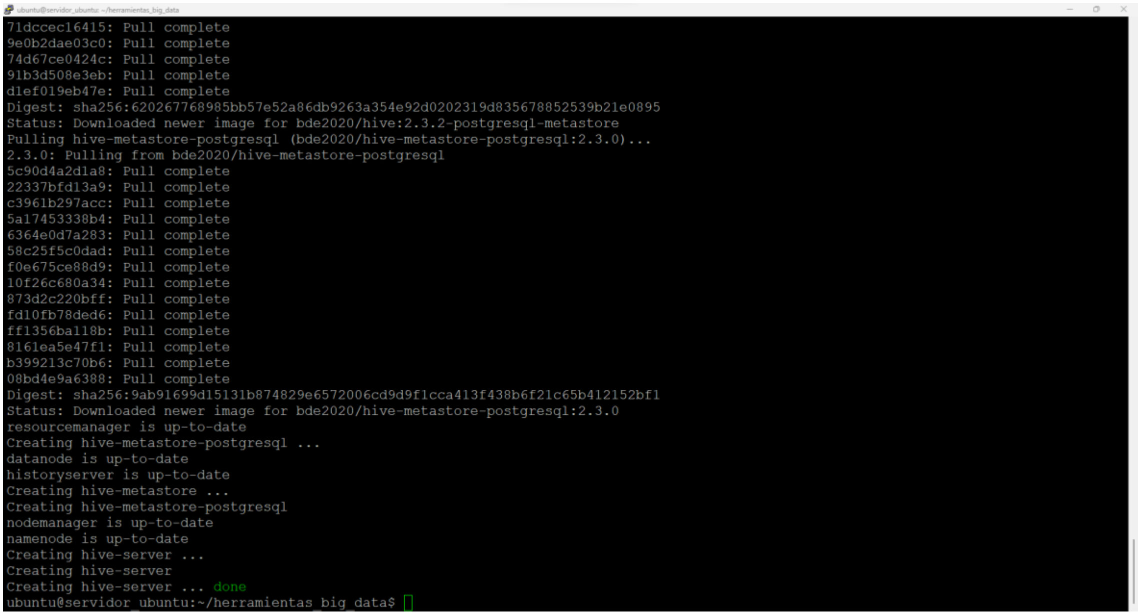
Edge te ayuda a mantenerse más seguro en línea con herramientas integradas que

This XML file does not appear to have any style information associated with it.

```
▼<configuration>
  ▼<property>
      <name>mapreduce.jobhistory.jhist.format</name>
      <value>binary</value>
      <final>false</final>
      <source>mapred-default.xml</source>
    </property>
  ▼<property>
      <name>fs.s3a.retry.interval</name>
      <value>500ms</value>
      <final>false</final>
      <source>core-default.xml</source>
    </property>
  ▼<property>
      <name>dfs.block.access.token.lifetime</name>
      <value>600</value>
      <final>false</final>
      <source>hdfs-default.xml</source>
    </property>
  ▼<property>
      <name>mapreduce.job.heap.memory-mb.ratio</name>
      <value>0.8</value>
      <final>false</final>
      <source>mapred-default.xml</source>
    </property>
  ▼<property>
      <name>mapreduce.map.log.level</name>
      <value>INFO</value>
      <final>false</final>
      <source>mapred-default.xml</source>
    </property>
  ▼<property>
      <name>dfs.namenode.lazypersist.file.scrub.interval.sec</name>
      <value>300</value>
      <final>false</final>
      <source>hdfs-default.xml</source>
    </property>
  ▼<property>
      <name>file.bytes-per-checksum</name>
      <value>512</value>
      <final>false</final>
      <source>core-default.xml</source>
    </property>
  ▼<property>
      <name>mapreduce.client.completion.pollinterval</name>
      <value>5000</value>
      <final>false</final>
      <source>mapred-default.xml</source>
    </property>
  ▼<property>
      <name>fs.azure.secure.mode</name>
      <value>false</value>
      <final>false</final>
      <source>core-default.xml</source>
    </property>
```

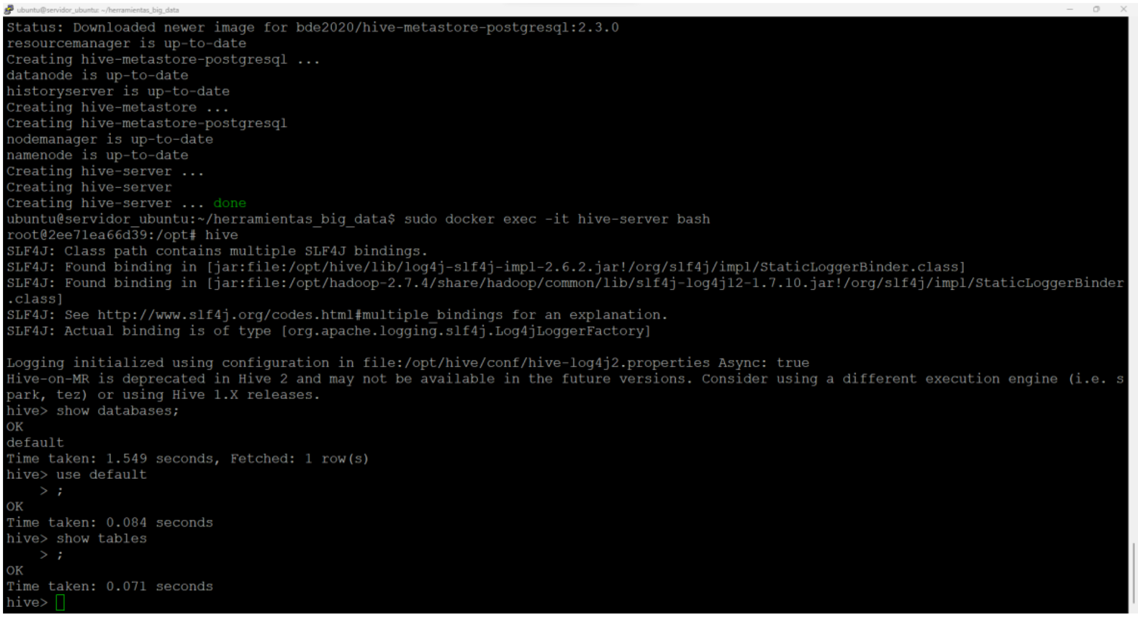
Paso 2

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



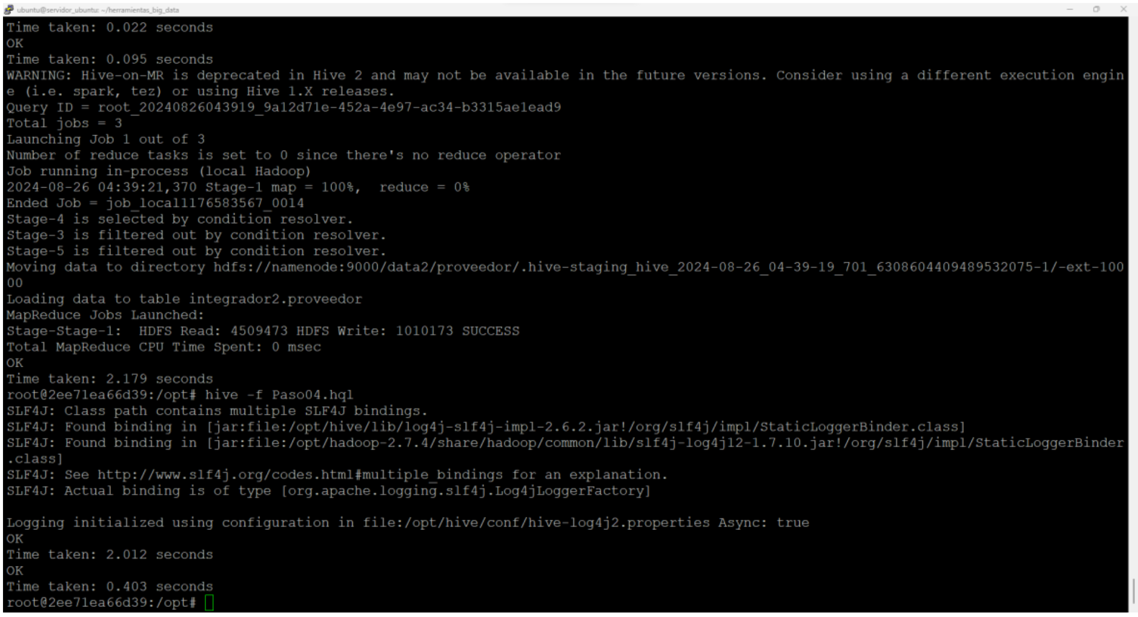
```
71dccec16415: Pull complete
9e0b2dae03c0: Pull complete
74d67ce0424c: Pull complete
91b3d508e3eb: Pull complete
d1ef019eb47e: Pull complete
Digest: sha256:620267768985bb57e52a86db9263a354e92d0202319d835678852539b21e0895
Status: Downloaded newer image for bde2020/hive:2.3.2-postgresql-metastore
Pulling 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: sha256:9ab91699d15131b874829e6572006cd9d9f1cca413f438b6f21c65b412152bf1
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-server
Creating hive-server ... done
ubuntu@servidor_ubuntu:~/herramientas_big_data$
```

se Crean tablas en Hive, tomando como referencia los csv ingestados en HDFS.



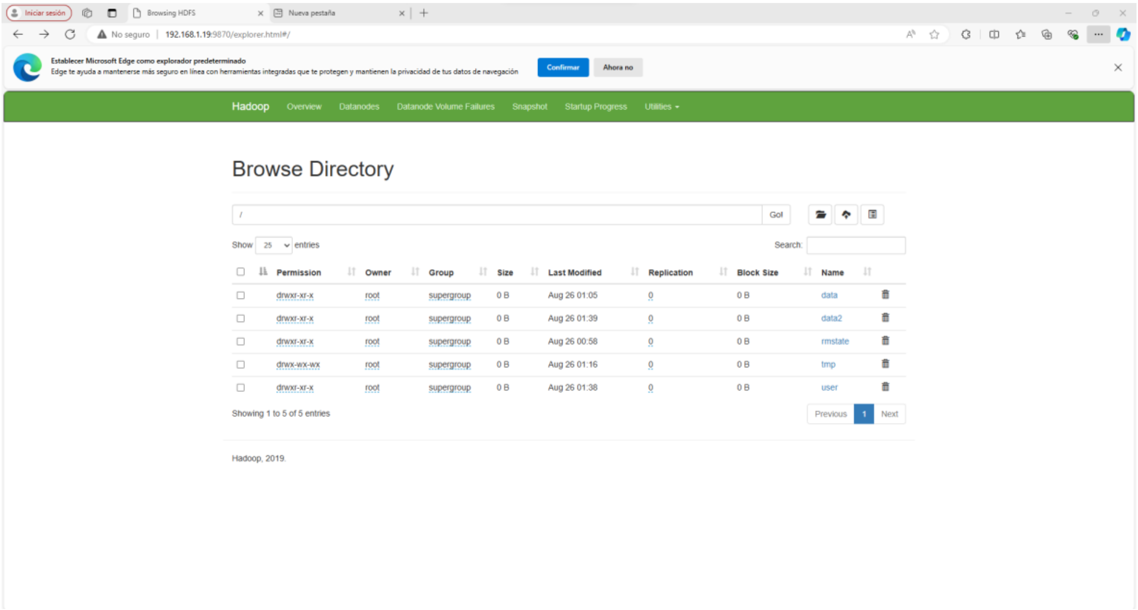
```
Status: Downloaded newer image for bde2020/hive-metastore-postgresql:2.3.0
resourcemanager is up-to-date
Creating hive-metastore-postgresql ...
datanode is up-to-date
historyserver is up-to-date
Creating hive-metastore ...
Creating hive-metastore-postgresql
nodemanager is up-to-date
namenode is up-to-date
Creating hive-server ...
Creating hive-server
Creating hive-server ... done
ubuntu@servidor_ubuntu:~/herramientas_big_data$ sudo docker exec -it hive-server bash
root@2ee71ea66d39:/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. s
park, tez) or using Hive 1.X releases.
hive> show databases;
OK
default
Time taken: 1.549 seconds, Fetched: 1 row(s)
hive> use default
    > ;
OK
Time taken: 0.084 seconds
hive> show tables
    > ;
OK
Time taken: 0.071 seconds
hive>
```



```
Time taken: 0.022 seconds
OK
Time taken: 0.095 seconds
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engin
e (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20240826043919_9a12d71e-452a-4e97-ac34-b3315ae1ead9
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 04:39:21,370 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local1176583567_0014
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:9000/data2/proveedor/.hive-staging_hive_2024-08-26_04-39-19_701_6308604409489532075-1/-ext-100
00
Loading data to table integrador2.proveedor
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 4509473 HDFS Write: 1010173 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.179 seconds
root@2ee71ea66d39:/opt# hive -f Paso04.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: 2.012 seconds
OK
Time taken: 0.403 seconds
root@2ee71ea66d39:/opt# 
```



Establecer Microsoft Edge como explorador predeterminado
Edge te ayuda a mantenerse más seguro en línea con herramientas integradas que te protegen y mantienen la privacidad de tus datos de navegación

Browse Directory

/

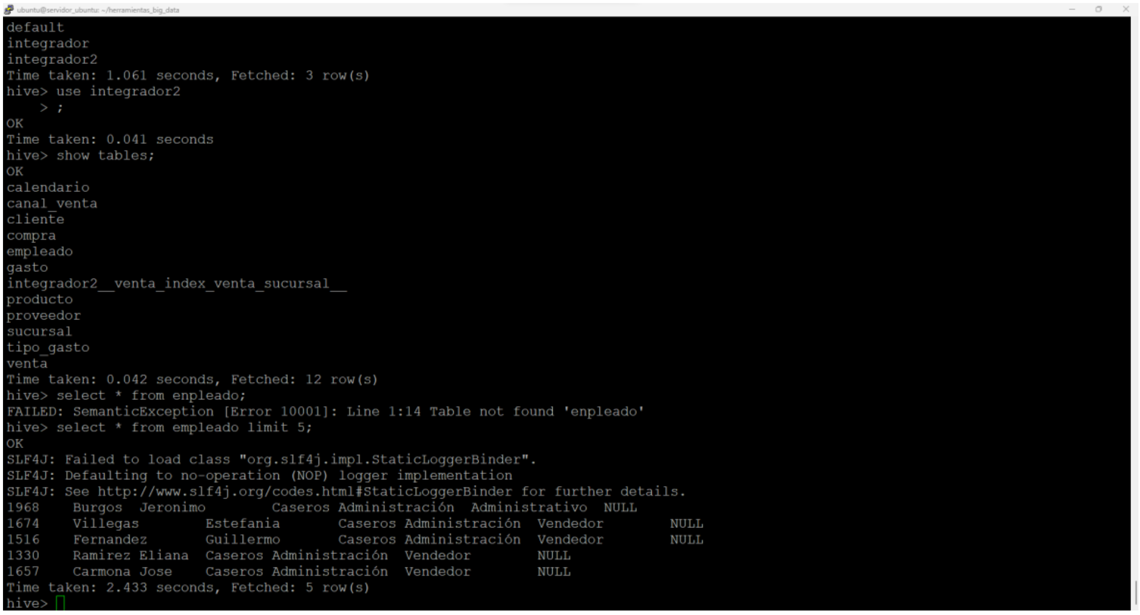
Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:05	0	0 B	data
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:39	0	0 B	data2
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 00:58	0	0 B	rmstate
□	drwx-wx-wx	root	supergroup	0 B	Aug 26 01:16	0	0 B	tmp
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:38	0	0 B	user

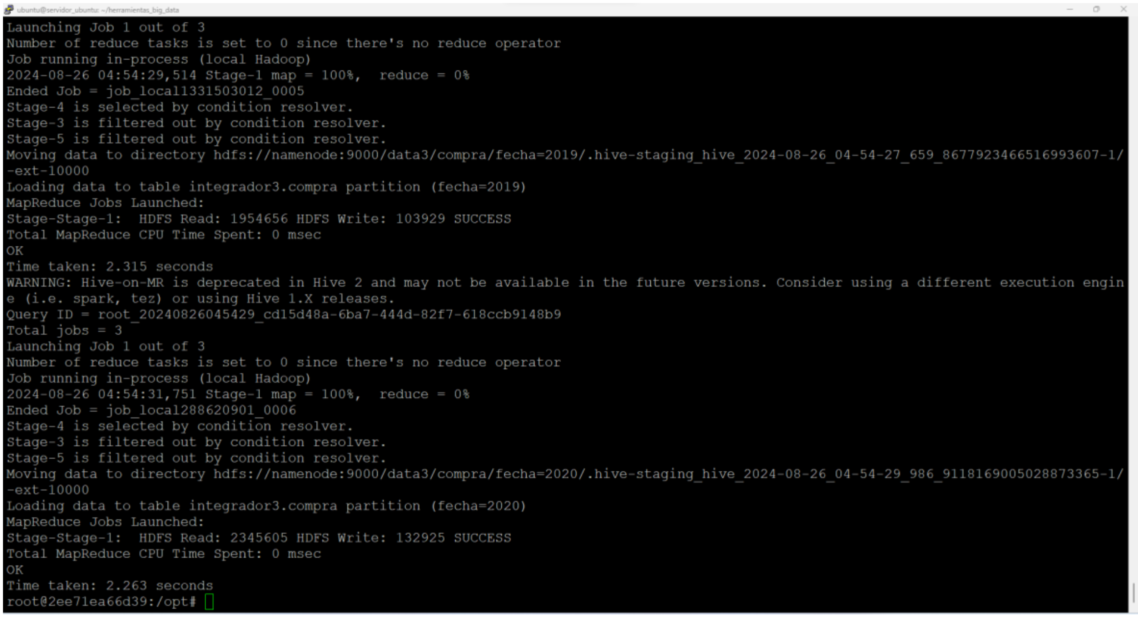
Showing 1 to 5 of 5 entries

Hadoop, 2019.

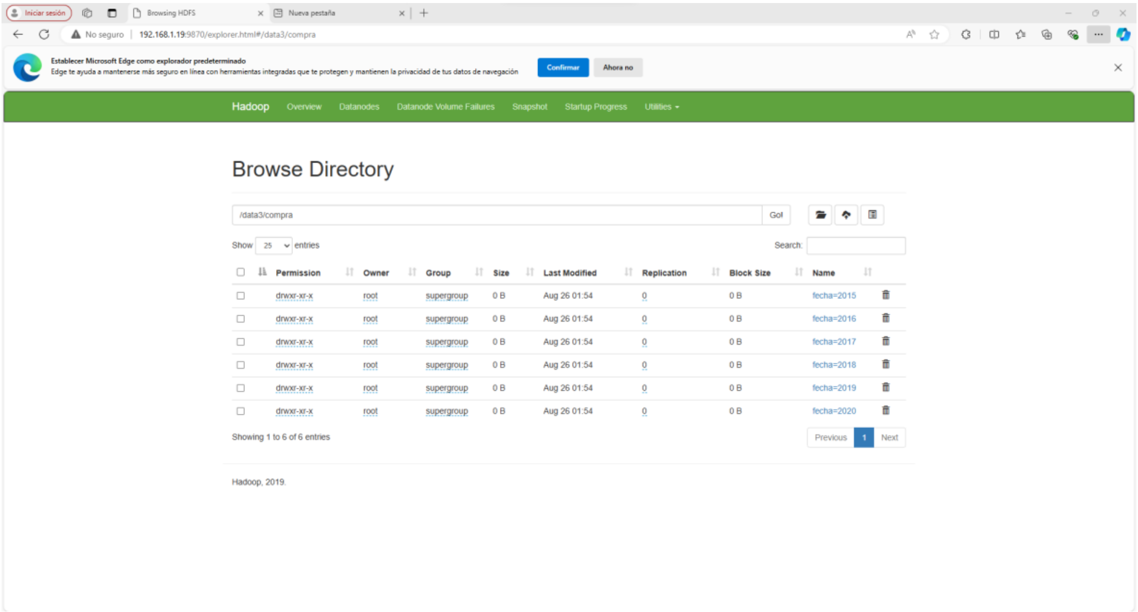


```
default
integrador
integrador2
Time taken: 1.061 seconds, Fetched: 3 row(s)
hive> use integrador2
    > ;
OK
Time taken: 0.041 seconds
hive> show tables;
OK
calendario
canal_venta
cliente
compra
empleado
gasto
integrador2__venta_index_venta_sucursal__
producto
proveedor
sucursal
tipo_gasto
venta
Time taken: 0.042 seconds, Fetched: 12 row(s)
hive> select * from enpleado;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'enpleado'
hive> select * from empleado limit 5;
OK
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.
1968    Burgos  Jeronimo        Caseros Administración  Administrativo  NULL
1674    Villegas        Estefania       Caseros Administración  Vendedor        NULL
1516    Fernandez       Guillermo       Caseros Administración  Vendedor        NULL
1330    Ramirez Eliana  Caseros Administración  Vendedor        NULL
1657    Carmona Jose    Caseros Administración  Vendedor        NULL
Time taken: 2.433 seconds, Fetched: 5 row(s)
hive> 
```

Paso 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 04:54:29,514 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local1331503012_0005
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:9000/data3/compra/fecha=2019/.hive-staging_hive_2024-08-26_04-54-27_659_8677923466516993607-1/
-ext-10000
Loading data to table integrador3.compra partition (fecha=2019)
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 1954656 HDFS Write: 103929 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.315 seconds
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engin
e (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20240826045429_cd15d48a-6ba7-444d-82f7-618ccb9148b9
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 04:54:31,751 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local288620901_0006
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://namenode:9000/data3/compra/fecha=2020/.hive-staging_hive_2024-08-26_04-54-29_986_9118169005028873365-1/
-ext-10000
Loading data to table integrador3.compra partition (fecha=2020)
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 2345605 HDFS Write: 132925 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.263 seconds
root@2ee71ea66d39:/opt# 
```



Iniciar sesión
Browsing HDFS
Nueva pestaña
No seguro
192.168.1.19:9870/explorer.html#/data3/compra
Establecer Microsoft Edge como explorador predeterminado
Edge te ayuda a mantenerse más seguro en línea con herramientas integradas que te protegen y mantienen la privacidad de tus datos de navegación
Confirmar
Ahora no
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

Browse Directory

/data3/compra Go!

Show 25 entries Search:

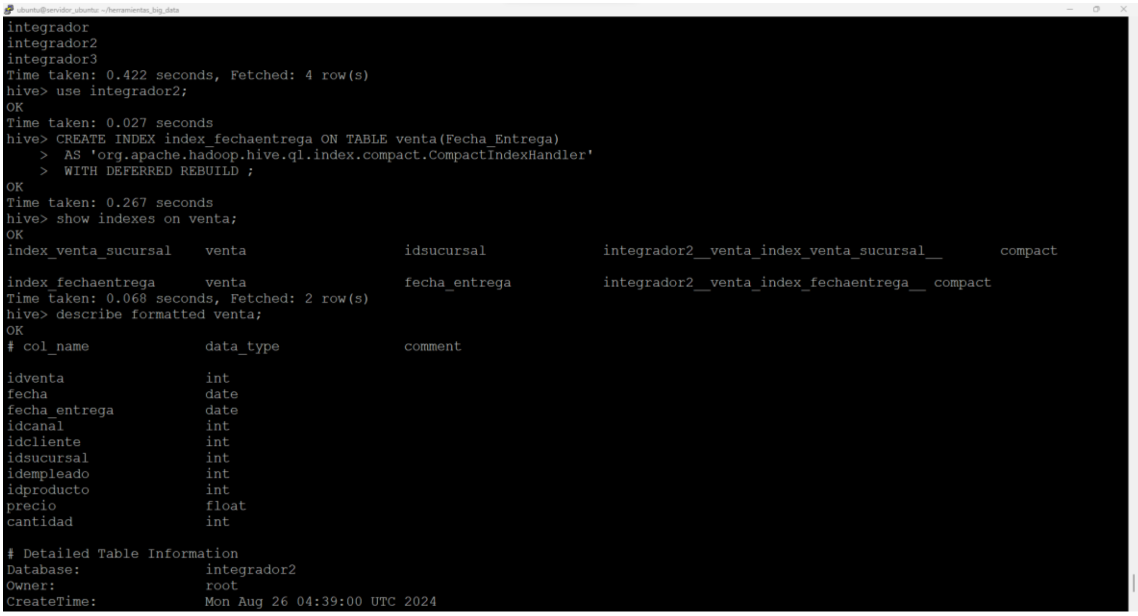
□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2015
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2016
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2017
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2018
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2019
□	drwxr-xr-x	root	supergroup	0 B	Aug 26 01:54	0	0 B	fecha=2020

Showing 1 to 6 of 6 entries

Previous 1 Next

Hadoop, 2019.

Paso 4



```
integrador
integrador2
integrador3
Time taken: 0.422 seconds, Fetched: 4 row(s)
hive> use integrador2;
OK
Time taken: 0.027 seconds
hive> CREATE INDEX index_fechaentrega ON TABLE venta(Fecha_Entrega)
    > AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
    > WITH DEFERRED REBUILD ;
OK
Time taken: 0.267 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: 0.068 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 04:39:00 UTC 2024
```