UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA

DIRECCIÓN DE POSGRADO

DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES

SEGUNDA VERSIÓN

LABORATORIO HIVE

NOMBRE : HEBERT JUAN DE DIOS DELGADILLO FERNANDEZ
DOCENTE : DANNY LUIS HUANCA SEVILLA

Cochabamba – Bolivia
2023

Laboratorio HIVE

Asegurarse de tener los servicios de Hadoop y hive arriba

1. Ejecutar IniciarHadoop – levanta los datanodes y los namenodes.

```
(base) curso@cursobigdata:~$ iniciarHadoop
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-curso-namenode-cursobigdata.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-curso-datanode-cursobigdata.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-curso-secondarynamenode-cursobig
data.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-curso-resourcemanager-cursobigdata.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-curso-nodemanager-cursobigdata.out
```

2. Verificar con jps

```
(base) curso@cursobigdata:~$ jps
1937 DataNode
2181 SecondaryNameNode
2344 ResourceManager
2520 NodeManager
1736 NameNode
2826 Jps
```

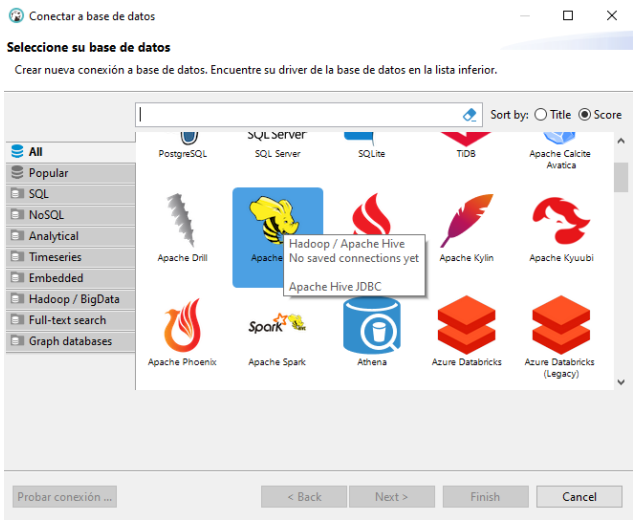
3. Ejecutar iniciarHive - Levanta el metastore

```
(base) curso@cursobigdata:~$ iniciarHive
2023-09-01 12:13:41: Starting Hive Metastore Server
```

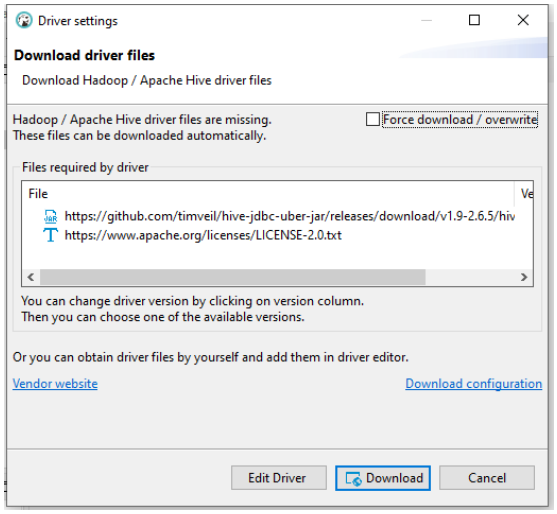
4. Levantar el servicio para conectarse por el puerto 10000 en otra sesión ssh

```
(base) curso@cursobigdata:~$ hiveserver2
2023-09-01 12:13:53: Starting HiveServer2
Hive Session ID = 5a9f8231-d91d-422b-9bc2-b44aa83581e2
Hive Session ID = 7fb11ecb-8c56-4f28-8cb6-a02368043c3a
Hive Session ID = 2ff77db0-7e46-4eef-a1a1-dde92814a85a
Hive Session ID = f1335ec0-be12-4d7f-87a5-a5456daa66d1
```

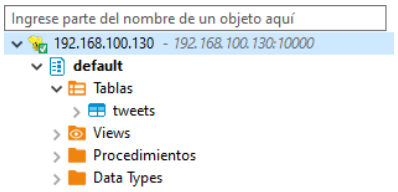
5. Conectarse con el cliente (aginity o dbeaver)



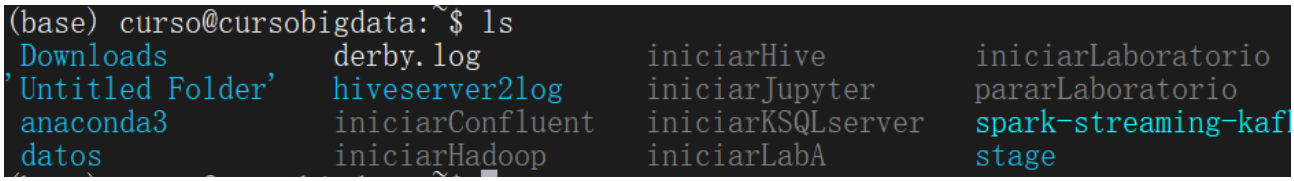
6. Crear la conexión bajando el driver que pida



Ya tenemos la conexión

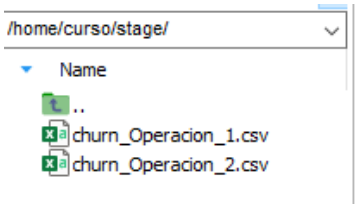


7. Verificamos que se tenga la carpeta stage en el filesystem local, sino se encuentra, entonces crearmos con mkdir stage



En este caso contábamos con la carpeta stage.

8. Crear una transformación que deposite datos en el filesystem local del servidor en la carpeta **stage,** los archivos de la carpeta Telco, son dos archivos csv.



9. Subir los arhivos del filesystem local (carpeta **stage**) al directorio en el HDFS que se mapeará como tabla externa.

Creamos la carpeta, subimos los datos y verificamos el primer archivo:

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -mkdir -p /TABLAEXTERNA/tablatelco1
(base) curso@cursobigdata:~/stage$ hdfs dfs -put churn_Operacion_1.csv /TABLAEXTERNA/tablatelco1
(base) curso@cursobigdata:~/stage$ hdfs dfs -ls /TABLAEXTERNA/tablatelco1
Found 1 items
-rw-r--r--   1 curso supergroup     977501 2023-09-01 12:28 /TABLAEXTERNA/tablatelco1/churn_Operacion_1.csv
```

Creamos la carpeta, subimos los datos y verificamos el segundo archivo:

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -mkdir -p /TABLAEXTERNA/tablatelco2
(base) curso@cursobigdata:~/stage$ hdfs dfs -put churn_Operacion_2.csv /TABLAEXTERNA/tablatelco2
(base) curso@cursobigdata:~/stage$ hdfs dfs -ls /TABLAEXTERNA/tablatelco2
Found 1 items
-rw-r--r--   1 curso supergroup     380191 2023-09-01 12:29 /TABLAEXTERNA/tablatelco2/churn_Operacion_2.csv
```

10. Creamos una tabla externa desde hive utilizando algún cliente que pueda conectarse con Hive. Esta tabla externa debe apuntar al directorio donde se encuentran los datos en elHDFS.

- Se usará DBeaver

```
CREATE TABLE TELCO1_EXT(CUSTOMERID String,

GENDER String,  String, DEPENDENTS String,

TENURE float, SENIORCITIZEN float, PHONESERVICE String, MULTIPLELINES String,
INTERNETSERVICE String, ONLINESECURITY String, ONLINEBACKUP String, DEVICEPROTECTION
String,TECHSUPPORT String, STREAMINGTV String, STREAMINGMOVIES String,

CONTRACT String, PAPERLESSBILLING String,PAYMENTMETHOD String, MONTHLYCHARGES float,
TOTALCHARGES float,

CHURN String

)

comment 'datos tabla telco1'

row format delimited fields terminated by ','stored as textfile

location '/TABLA_EXTERNA/tablatelco1'

;
```

Se hizo una pequeña corrección dado que la creación de la tabla estaba desordenada:

```
CREATE TABLE TELCO1_EXT
 (CUSTOMERID String,
 GENDER String,
 SENIORCITIZEN float,
 PARTNER String,
 DEPENDENTS String,
 TENURE float,
 PHONESERVICE String,
 MULTIPLELINES String,
 INTERNETSERVICE String,
 ONLINESECURITY String, S
 ONLINEBACKUP String,
 DEVICEPROTECTION String,
 TECHSUPPORT String,
 STREAMINGTV String,
 STREAMINGMOVIES String,
 CONTRACT String,
 PAPERLESSBILLING String,
 PAYMENTMETHOD String,
 MONTHLYCHARGES float,
 TOTALCHARGES float,
 CHURN String
 )
```

11. Insertar los datos de la tabla externa a una tabla hive.

```
create table telco1
as select *
from telco1_ext
```

Estadísticas 1

Name	Value
Updated Rows	-1
Query	create table telco1
	as select *
	from telco1_ext
Start time	Fri Sep 01 09:30:00 BOT 2023
Finish time	Fri Sep 01 09:30:01 BOT 2023

12. Verificamos los tamaños de los archivos en ambos directorios en el de la tabla externa y elcreado en HIVE.

De la tabla externa almacenada en hdfs:

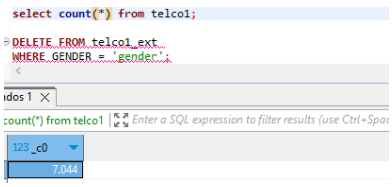
```
(base) curso@cursobigdata:~/stage$ hdfs dfs -du -h /TABLAEXTERNA/tablatelco1
954.6 K  /TABLAEXTERNA/tablatelco1/churn_Operacion_1.csv
```

De la tabla creada por hive:

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -du -h /user/hive/warehouse/telco1
976.6 K  /user/hive/warehouse/telco1/000000_0
```

¿Cuántos registros tiene la tabla telco1?

Tiene 7043 registros en total, la tabla devuelve 7044, pero cargo los headers como si fuera datos y no deja eliminarlos:



13. Generar un procedimiento similar para subir datos a la tabla telco2. ¿Cuántos registrostiene la tabla telco2?

```
CREATE TABLE TELCO2_EXT(
state String,
account float,
area_code float,
phone_number String,
international_plan String,
voice_mail_plan String,
number_vmail_messages float,
total_day_minutes float,
total_day_calls float,
total_day_charge float,
total_eve_minutes float,
total_eve_calls float,
total_eve_charge float,
total_night_minutes float,
total_night_calls float,
total_night_charge float,
total_intl_minutes float,
total_intl_calls float,
total_intl_charge float,
number_customer_service_calls float,
churn String
```

Value
-1
CREATE TABLE TELCO2_EXT(
state String,
account float,
area_code float,
phone_number String,
international_plan String,
voice_mail_plan String,
number_vmail_messages float,
total_day_minutes float,
total_day_calls float,
total_day_charge float,
total_eve_minutes float,
total_eve_calls float,
total_eve_charge float,
total_night_minutes float,
total_night_calls float,

```
create table telco2
as select *
from telco2_ext

select * from telco2
```

from telco2 | Enter a SQL expression to filter results (use Ctrl+Space)

state	account	area_code	phone_number	international_plan	voice_m
NJ	137	415	358-1921	no	no
OH	84	408	375-9999	yes	no
OK	75	415	330-6626	yes	no
AL	118	510	391-8027	yes	no
MA	121	510	355-9993	no	yes
MO	147	415	329-9001	yes	no
LA	117	408	335-4719	no	no
WV	141	415	330-8173	yes	yes
IN	65	415	329-6603	no	no
RI	74	415	344-9403	no	no
IA	168	408	363-1107	no	no

La tabla tiene 3333 registros, aparece 3334 ya que toma las cabeceras como datos

```
select count(*) from telco2;
```

Resultados 1

select count(*) from telco2 | Enter a SQL expression to

Grilla

	123 _c0
1	3.334

14. Duplique la tabla telco1 creando una tabla telco3 en base a la tabla telco1

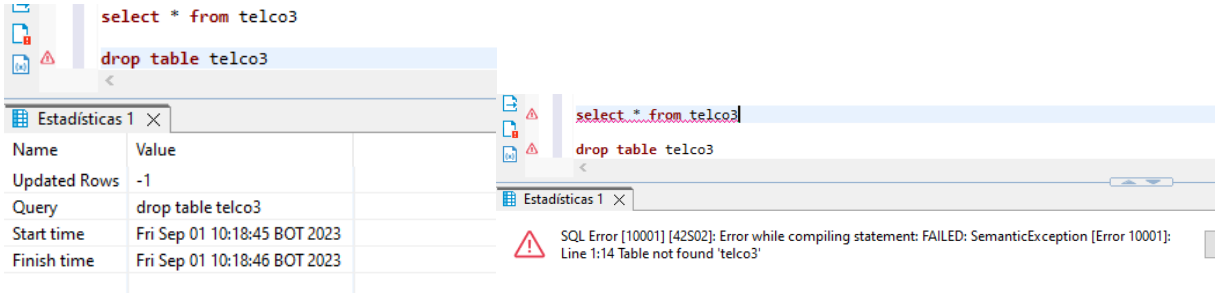
```
create table telco3
as select *
from telco1_ext

select * from telco3
```

esultados 1

elect * from telco3 | Enter a SQL expression to filter results (use Ctrl+Space)

	customerid	gender	seniorcitizen	partner	dependents
	7590-VHVEG	Female	0	Yes	No
	5575-GNVDE	Male	0	No	No
	3668-QPYBK	Male	0	No	No
	7795-CFOCW	Male	0	No	No
	9237-HQITU	Female	0	No	No
	9305-CDSKC	Female	0	No	No
	1452-KIOVK	Male	0	No	Yes
	6713-OKOMC	Female	0	No	No
0	7892-POOKP	Female	0	Yes	No
1	6388-TABGU	Male	0	No	Yes
2	9763-GRSKD	Male	0	Yes	Yes

15. Borre la tabla telco3



La tabla fue borrada exitosamente, sale error porque ya no encuentra la tabla.

16. Ingrese por ssh a la máquina virtual hadoop. ¿Como se encuentran almacenadas las tablas?Verifique la dirección de user/hive/warehouse

```
(base) curso@cursobigdata: ~/stage$ hdfs dfs -du -h /user/hive/warehouse/
0          /user/hive/warehouse/src
976.6 K    /user/hive/warehouse/telco1
370.9 K    /user/hive/warehouse/telco2
37         /user/hive/warehouse/tweets
```

Estan almacenadas como archivos del cluster de hadoop, dentro de cada una de ellas están almacenadas por partes como la tabla 2 mostrada a continuación:

```
(base) curso@cursobigdata: ~/stage$ hdfs dfs -du -h /user/hive/warehouse/telco2
370.9 K  /user/hive/warehouse/telco2/000000_0
```

Es como un archivo dentro de Hive que sabe como interpretar la table mediante ese archivo.

17. Desde la línea de comandos inicie el cliente hive. Verifique las bases de datos existentes y cree una base que se denomine prueba.

Iniciamos el cliente Hive:

```
(base) curso@cursobigdata: /stage$ hive
Hive Session ID = e5eb6f09-a17d-4676-9aae-36189cf7fc6f

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-3.1.2.jar!/hive-log4j2.pro
perties 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 Session ID = a439c4f5-c037-436f-b746-c44ed2b1739f
```

Solo esta la base de datos “default”

```
hive> show databases;
OK
default
Time taken: 0.693 seconds, Fetched: 1 row(s)
```

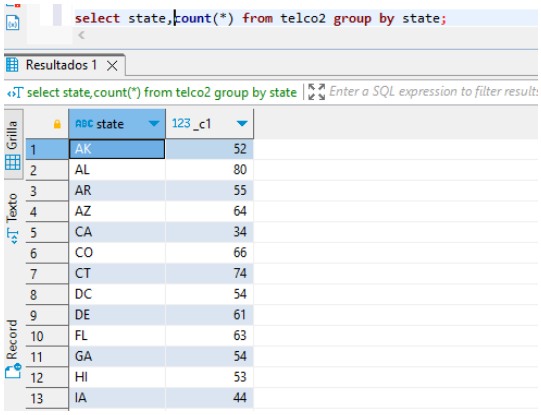
Y las tablas de esta base de datos son:

```
hive> use default
    > ;
OK
Time taken: 0.026 seconds
hive> show tables
    > ;
OK
telco1
telco1_ext
telco2
telco2_ext
tweets
Time taken: 0.039 seconds, Fetched: 5 row(s)
```

Creamos la base de datos prueba y verificamos que se creó correctamente:

```
hive> create database prueba;
OK
Time taken: 0.104 seconds
hive> show databases;
OK
default
prueba
Time taken: 0.028 seconds, Fetched: 2 row(s)
```

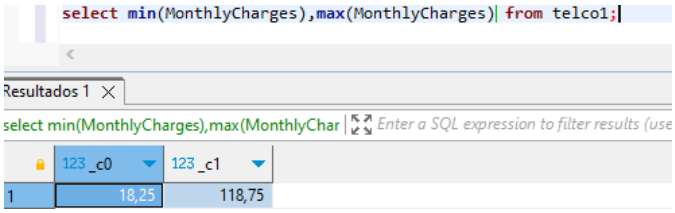
18. Ejecute un query que permita visualizar la distribución de estados en la base telco1(conteo).



Esta tabla nos muestra un conteo de la distribución de estados de la tabla telco 2.

19. Obtenga el mínimo y máximo de una variable cuantitativa de la tabla telco1.

Obtuvimos el mínimo y el máximo de la variable MonthlyCharges:

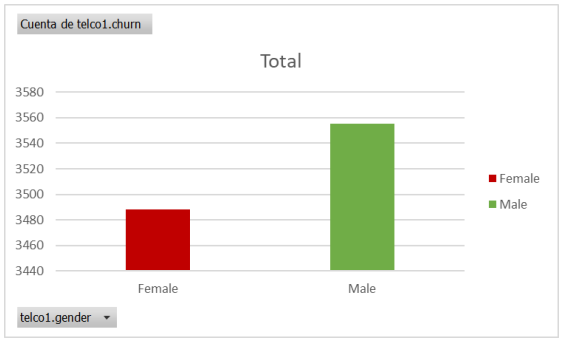


20. Conecte a Excel y mediante una tabla dinámica en base a telco1 indique la distribución delgénero respecto de la variable de interés churn.

La tabla que muestra la distribución de genero:

Etiquetas de fila	Cuenta de telco1.churn
Female	3488
Male	3555
Total general	**7043**

Acompañada de una grafica de diagrama de barras de la misma distribución:



21. Inserte la tabla telco2 mediante una tabla dinámica en Excel y proporcione algunas vistas del comportamiento de la variable churn respecto de otras variables cuantitativas

total_day_minutes

total_day_calls

total_day_charge

total_eve_minutes

total_eve_calls

total_eve_charge

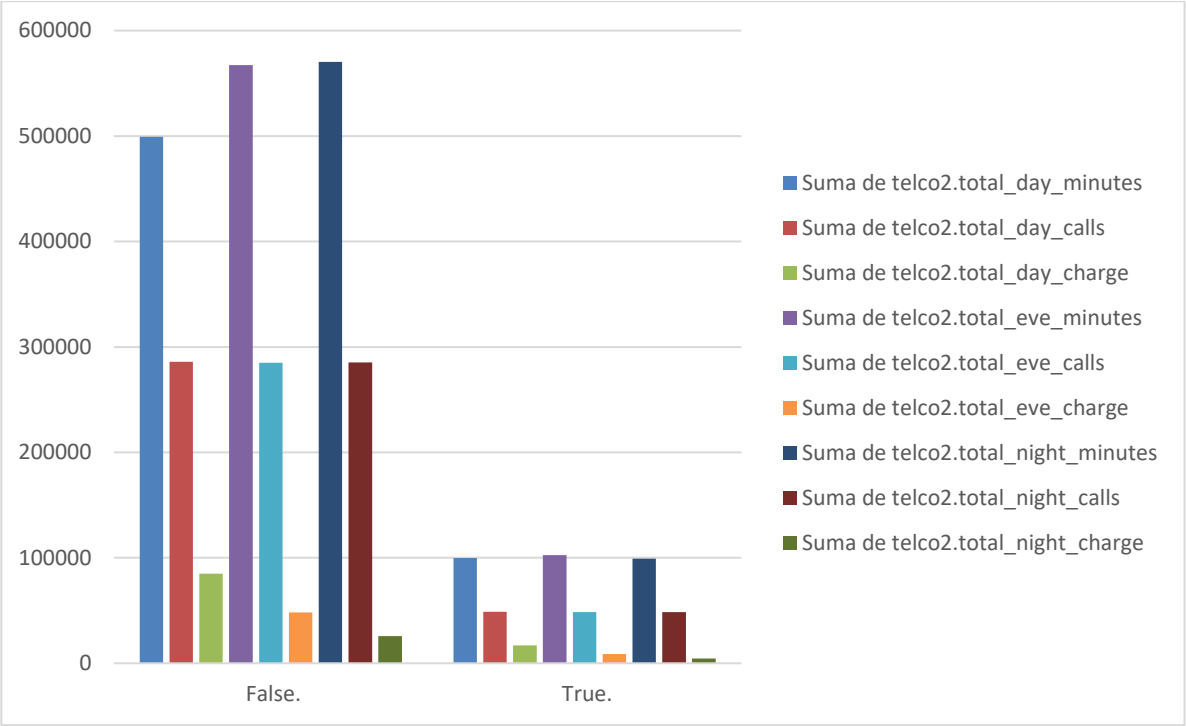
total_night_minutes

total_night_calls

total_night_charge

La tabla que usamos agrupa el total de cada uno de los campos requeridos con la variable churn:

Etiquetas de fila	Suma de telco2.total_day_minutes	Suma de telco2.total_day_calls	Suma de telco2.total_day_charge	Suma de telco2.total_eve_minutes	Suma de telco2.total_eve_calls
False.	499250.8999	285807	84874.20005	567273.3997	285110
True.	99939.49972	48945	16989.97001	102594.1	48571
Total general	**599190.3996**	**334752**	**101864.1701**	**669867.4996**	**333681**



22. ¿Existe algún patrón que indique que alguna de estas variables explica mejor la variable objetivo churn?

Si, que aquellos que se dieron de baja del servicio realizaban menos llamadas que aquellos que se quedaron activos.

23. Realice el mismo análisis respecto de variables cualitativas como:

state

area_code

international_plan

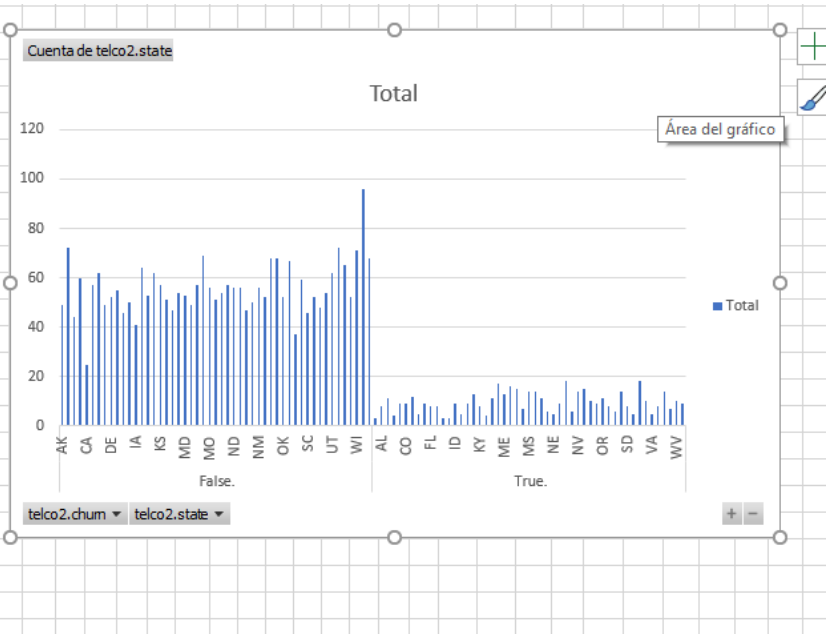
voice_mail_plan

La siguiente tabla nos agrupa la cantidad de personas que se dieron de baja y no por estado, código de distrito y si tenia plan internacional y de voz en el orden respectivo.

Cuenta de telco2.churn	Etiquetas de columna		
Etiquetas de fila	False.	True.	Total general
AK	49	3	52
AL	72	8	80
AR	44	11	55
AZ	60	4	64
CA	25	9	34
CO	57	9	66
CT	62	12	74
408	20	2	22
415	30	9	39
no	29	7	36
no	19	6	25
yes	10	1	11
yes	1	2	3
no	1	2	3
510	12	1	13
no	8	1	9
no	5	1	6
yes	3		3
yes	4		4
no	3		3
yes	1		1
DC	49	5	54
408	12	2	14
415	24	3	27
no	22	3	25

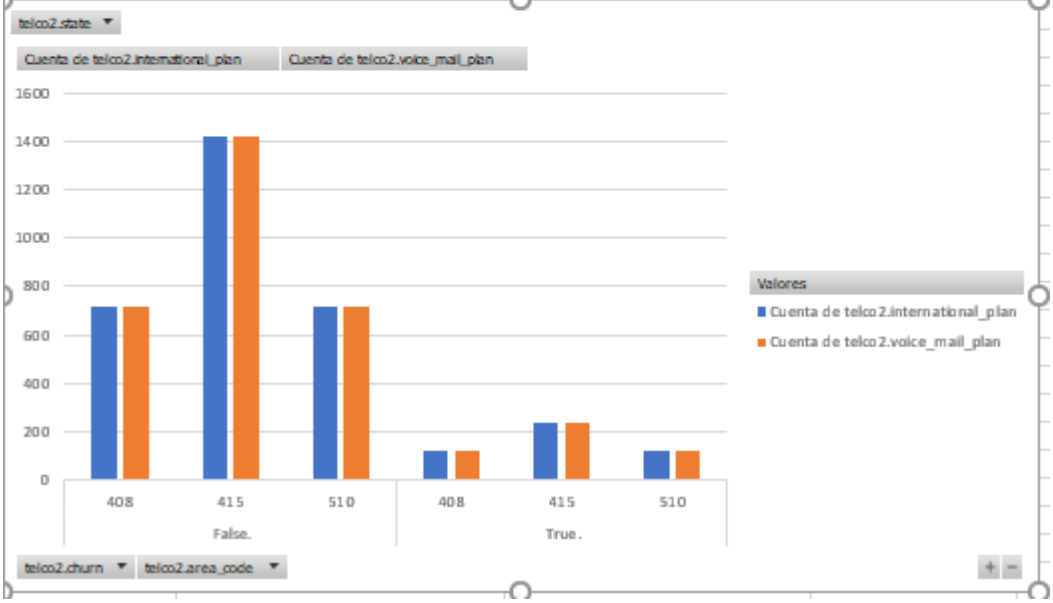
La siguiente tabla nos muestra la cantidad de personas que se dieron de baja agrupado por estados.

Etiquetas de fila	Cuenta de telco2.state
False.	2850
AK	49
AL	72
AR	44
AZ	60
CA	25
CO	57
CT	62
DC	49
DE	52
FL	55
GA	46
HI	50
IA	41
ID	64
IL	53
IN	62
KS	57
KY	51
LA	47
MA	54
MD	53
ME	49



La siguiente tabla y gráficos muestran la cantidad de personas con plan de voz e internacional filtrada por estados, actualmente se muestran de todos los estados.

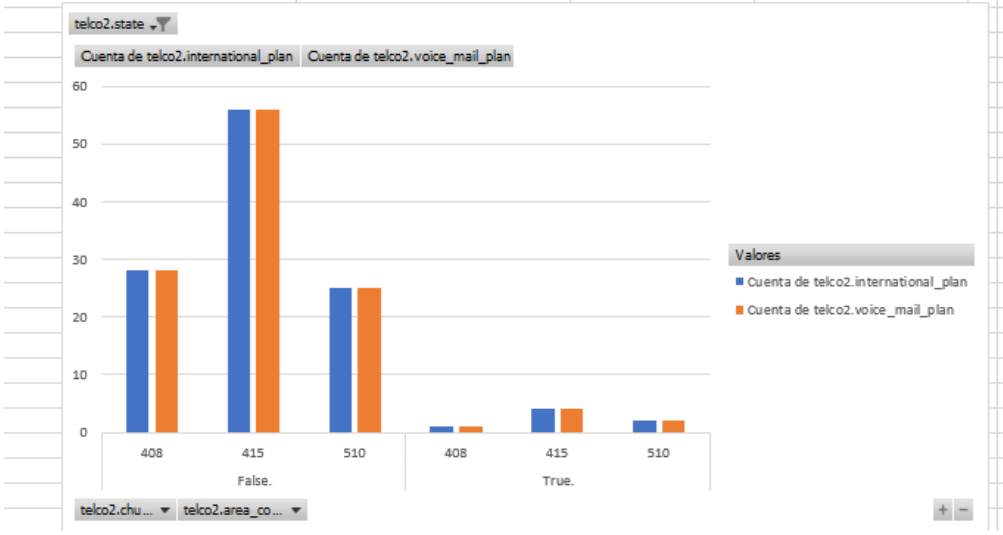
telco2.state	(Todas)	
Etiquetas de	Cuenta de telco2.internation	Cuenta de telco2.voice_mail_plan
False.	**2850**	**2850**
408	716	716
415	1419	1419
510	715	715
True.	**483**	**483**
408	122	122
415	236	236
510	125	125
Total general	**3333**	**3333**



La siguiente tabla es la misma que la anterior pero filtrada solo para 2 estados (AK y AZ):

(Varios elementos)

Cuenta de telco2.international_plan	Cuenta de telco2.voice_mail_plan
109	**109**
28	28
56	56
25	25
7	**7**
1	1
4	4
2	2
116	**116**



Conclusiones

El laboratorio nos enseño a como poder almacenar grandes cantidades de datos en una base que acepta comando SQL lo cual es bastante interesante, ya que podríamos aplicarlo a cualquier empresa que empezó a crecer y que una base de datos relacional normal no es suficiente, ademas se conecta con el laboratorio 1 ya que hive trabaja sobre el entorno de Hadoop y es muy interesante ver como sobre un sistema de archivos gigante que esta formado por varias computadoras puede usarse para almacenar grandes cantidades de datos.