

LABORATORIO 2:

Participantes:

- Cabana Cazani, Gabriel
- Larico Cruz, Diego

INGESTA DE DATOS

Nos localizamos en Files view para cargar los archivos csv y almacenarlos en la carpeta Data/raw

The top screenshot shows the Ambari Files view for the 'raj_ops' user. The directory is the root, and it contains 12 files or folders. The table below lists these items:

Name	Size	Last Modified	Owner	Group	Permission
app-logs	--	2018-06-18 10:18	yarn	hadoop	drwxrwxrwx
apps	--	2018-06-18 11:13	hdfs	hdfs	drwxr-xr-x
ats	--	2018-06-18 09:52	yarn	hadoop	drwxr-xr-x
data	--	2025-09-29 16:03	raj_ops	hdfs	drwxr-xr-x
hdp	--	2018-06-18 09:52	hdfs	hdfs	drwxr-xr-x
livy2-recovery	--	2018-06-18 10:11	livy	hdfs	drwx-----
mapred	--	2018-06-18 09:52	mapred	hdfs	drwxr-xr-x
mr-history	--	2018-06-18 09:52	mapred	hadoop	drwxrwxrwx
ranger	--	2018-06-18 10:59	hdfs	hdfs	drwxr-xr-x

The bottom screenshot shows the Ambari Files view for the 'raj_ops' user, specifically the 'data' directory. It contains 1 file or folder:

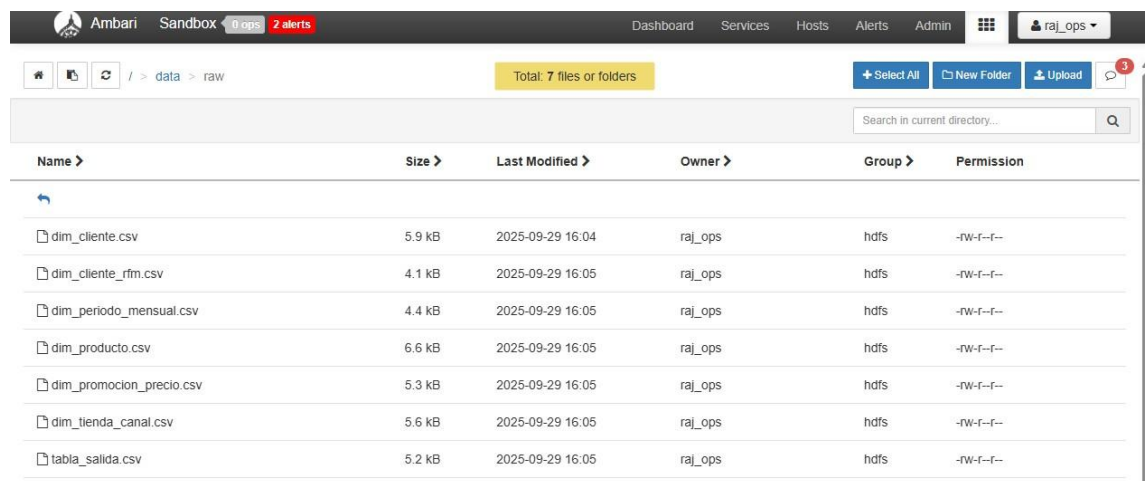
Name	Size	Last Modified	Owner	Group	Permission
raw	--	2025-09-29 16:05	raj_ops	hdfs	drwxr-xr-x

Usamos los siguientes comandos:

- **Hdfs dfs -ls/** para ver las carpetas.
- Para acceder a la carpeta data usamos **hdfs dfs -ls/data**, para acceder a la carpeta raw, luego para subir el archivo csv, usamos **hdfs dfs -put dim_cliente /data/raw**

```
[root@sandbox-hdp ~]# hdfs dfs -ls /
Found 12 items
drwxrwxrwx - yarn      hadoop      0 2018-06-18 15:18 /app-logs
drwxr-xr-x - hdfs      hdfs        0 2018-06-18 16:13 /apps
drwxr-xr-x - yarn      hadoop      0 2018-06-18 14:52 /ats
drwxr-xr-x - raj_ops   hdfs        0 2025-09-29 21:03 /data
drwxr-xr-x - hdfs      hdfs        0 2018-06-18 14:52 /hdp
drwx----- - livy      hdfs        0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x - mapred    hdfs        0 2018-06-18 14:52 /mapred
drwxrwxrwx - mapred    hadoop      0 2018-06-18 14:52 /mr-history
drwxr-xr-x - hdfs      hdfs        0 2018-06-18 15:59 /ranger
drwxrwxrwx - spark     hadoop      0 2025-09-29 22:36 /spark2-history
drwxrwxrwx - hdfs      hdfs        0 2018-06-18 16:06 /tmp
drwxr-xr-x - hdfs      hdfs        0 2018-06-18 16:08 /user
[root@sandbox-hdp ~]# hdfs dfs -ls /data
Found 1 items
drwxr-xr-x - raj_ops   hdfs        0 2025-09-29 21:05 /data/raw
[root@sandbox-hdp ~]# hdfs dfs -put dim_cliente /data/raw
```

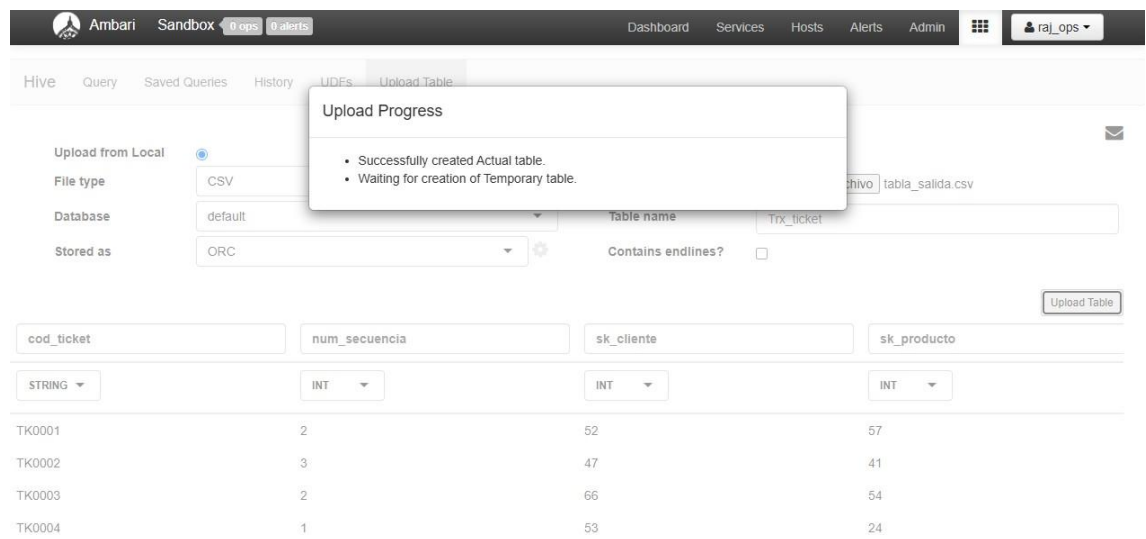
Vemos los archivos cargados en el HDFS



The screenshot shows the Ambari Sandbox interface. At the top, there's a navigation bar with 'Ambari', 'Sandbox', and 'raj_ops' (2 alerts). Below it, a breadcrumb shows the path '/ > data > raw'. A yellow box indicates 'Total: 7 files or folders'. A search bar is present. The main table lists files in HDFS with columns: Name, Size, Last Modified, Owner, Group, and Permission.

Name	Size	Last Modified	Owner	Group	Permission
dim_cliente.csv	5.9 kB	2025-09-29 16:04	raj_ops	hdfs	-rw-r--r--
dim_cliente_rfm.csv	4.1 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--
dim_periodo_mensual.csv	4.4 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--
dim_producto.csv	6.6 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--
dim_promocion_precio.csv	5.3 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--
dim_tienda_canal.csv	5.6 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--
tabla_salida.csv	5.2 kB	2025-09-29 16:05	raj_ops	hdfs	-rw-r--r--

Cargamos la tabla para las consultas sql



The screenshot shows the 'Upload Table' interface in Ambari Sandbox. A modal window titled 'Upload Progress' is open, showing two steps: 'Successfully created Actual table.' and 'Waiting for creation of Temporary table.' Below the modal, the 'Upload from Local' section is active. The 'File type' is set to 'CSV', 'Database' is 'default', and 'Stored as' is 'ORC'. The 'Table name' is 'Trx_ticket'. Below this, there's a table with columns: cod_ticket (STRING), num_secuencia (INT), sk_cliente (INT), and sk_producto (INT). The table contains 4 rows of data.

cod_ticket	num_secuencia	sk_cliente	sk_producto
TK0001	2	52	57
TK0002	3	47	41
TK0003	2	66	54
TK0004	1	53	24

Realizamos consulta básica para ver los 10 primeros registros

SELECT * FROM trx_ticket LIMIT 10;



The screenshot shows the 'Query Editor' interface. It has a tab for 'Worksheet * x' and 'trx_ticket sample x'. The SQL query 'SELECT * FROM trx_ticket LIMIT 10;' is entered in the editor.

Query Process Results (Status: SUCCEEDED)						Save results... ▾
<div> <div>Logs</div> <div>Results</div> </div>						
<div>Filter columns...</div>						<div>previous</div> <div>next</div>
trx_ticket.cod_ticket	trx_ticket.num_secuencia	trx_ticket.sk_cliente	trx_ticket.sk_producto	trx_ticket.sk_tienda	trx_ticl	
TK0001	2	52	57	99	76	
TK0002	3	47	41	47	40	
TK0003	2	66	54	81	68	
TK0004	1	53	24	8	73	
TK0005	4	89	53	2	40	
TK0006	3	13	76	2	4	
TK0007	5	83	69	66	74	
TK0008	4	98	66	14	13	
TK0009	2	79	56	8	2	
TK0010	1	46	36	22	35	

Usamos lo siguiente para poder crear la tabla externa

```
CREATE EXTERNAL TABLE IF NOT EXISTS tabla_salida_external (
  cod_ticket STRING,
  cod_cliente INT,
  cod_sucursal INT,
  cod_fecha INT,
  mto_venta_bruta DOUBLE,
  mto_descuento DOUBLE,
  mto_venta_neta DOUBLE,
  mto_margen DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 'data/raw';
```

Query Editor

Worksheet

```
1 CREATE EXTERNAL TABLE IF NOT EXISTS tabla_salida_external (  
2   cod_ticket STRING,  
3   cod_cliente INT,  
4   cod_sucursal INT,  
5   cod_fecha INT,  
6   mto_venta_bruta DECIMAL(18,2),  
7   mto_descuento DECIMAL(18,2),  
8   mto_venta_neta DECIMAL(18,2),  
9   mto_margen DECIMAL(18,2)  
10 )  
11 ROW FORMAT DELIMITED  
12 FIELDS TERMINATED BY '\t'  
13 STORED AS TEXTFILE  
14 LOCATION '/data/raw';
```

SQL

⚙️

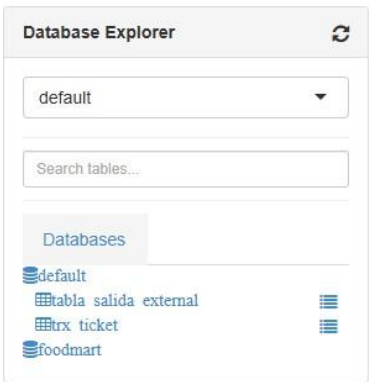
📈

🔗

TEZ

📧

Actualizamos y si se creó la tabla



Vemos el formato de nuestra tabla

Query Editor

Worksheet

```
1 describe formatted tabla_salida_external;
```

col_name	data_type	comment
# col_name	data_type	comment
""	null	null
cod_ticket	string	""
cod_cliente	int	""
cod_sucursal	int	""
cod_fecha	int	""
mto_venta_bruta	decimal(18,2)	""
mto_descuento	decimal(18,2)	""
mto_venta_neta	decimal(18,2)	""
mto_margen	decimal(18,2)	""
""	null	null
# Detailed Table Information	null	null

Luego se puede apreciar que es una tabla externa.

Location:	hdfs://sandbox-hdp.hortonworks.com:8020/data/raw	null
Table Type:	EXTERNAL_TABLE	null
Table Parameters:	null	null
""	EXTERNAL	TRUE