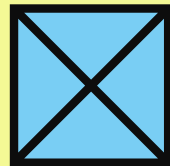


PROYECTO DE
ESPECIALIZACION

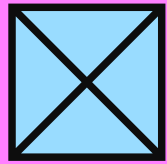
BIG DATA

TAREA NO
GRUPAL



INTEGRANTES

- JENRY LUIS
- WILBERD ESPINO
- DIEGO BERNALES



PASOS A SEGUIR

CASO DE
USO

CREACION
DEL
DATALAKE

BIG
QUERY

REAL
TIME

1

3

5

7

2

4

6

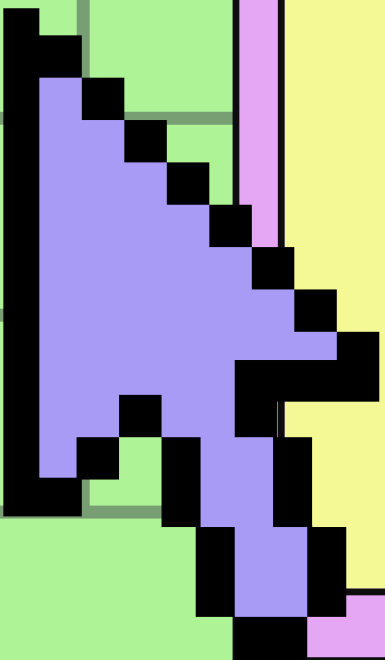
8

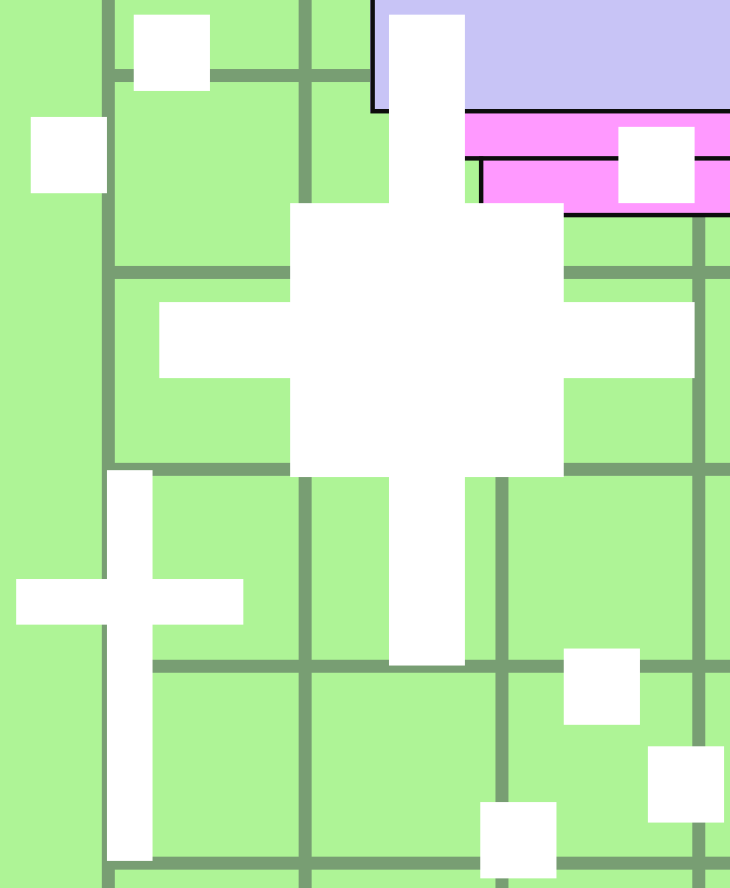
ARQUITECTURA

POBLACION
DEL
DATALAKE

DATA
STUDIO

CONCLUSIONES

- 
- Minimarket dedicado a la venta retail de productos de necesidad básica, alimentación, limpieza, entre otros.
 - Cuenta con 4 sucursales de los cuáles 1 está ubicado en una zona comercial de alta demanda.
 - Cuenta con 20 vendedores en total.



CASO DE USO

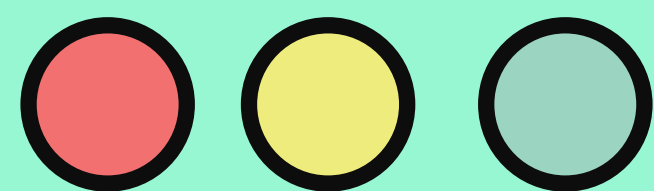


OBJETIVO:

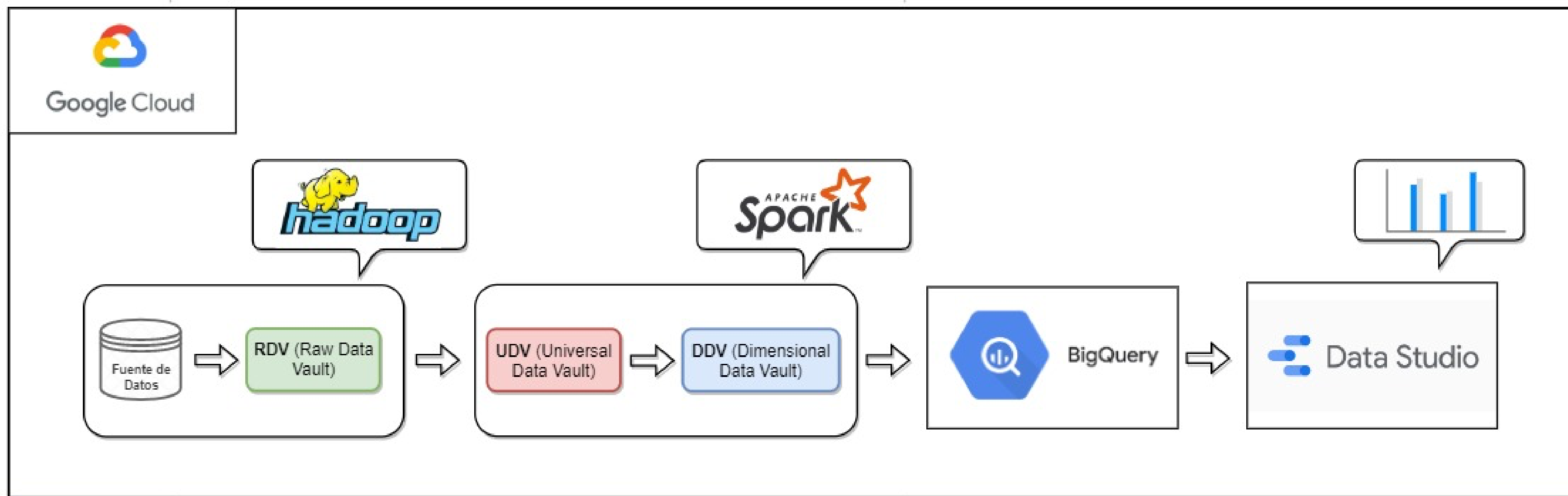
- Desarrollar una arquitectura Big Data en la nube para el procesamiento de la información de las ventas de la cadena de minimarkets.
- Generar un flujo de datos que permita realizar un análisis de la información y generar reportes visuales para optimizar la toma de decisiones.
- Generar un proceso en tiempo real para ver los registros de ventas y obtener los productos mas vendidos en el momento.

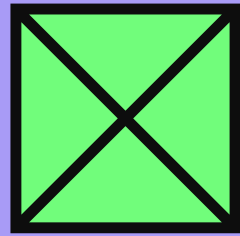


CASO DE
USO



ARQUITECTURA





CREACION DEL DATALAKE

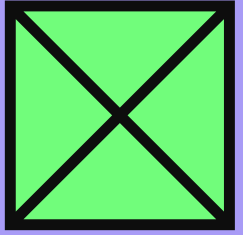


Ejecutamos los siguientes comandos en hdfs:

```
#hdfs dfs -mkdir /user/diego/datalake/rdv  
#hdfs dfs -mkdir /user/diego/datalake/udv  
#hdfs dfs -mkdir /user/diego/datalake/ddv
```

 carga_rdv.sql

```
1 CREATE DATABASE IF NOT EXISTS RDV LOCATION '/user/diego/datalake/rdv';  
2 CREATE DATABASE IF NOT EXISTS UDV LOCATION '/user/diego/datalake/udv';  
3 CREATE DATABASE IF NOT EXISTS DDV LOCATION '/user/diego/datalake/ddv';  
4
```



POBLACION DEL DATALAKE



```
5 CREATE TABLE IF NOT EXISTS RDV.PROY_VENTAS
6 (
7   id_empleado STRING,
8   id_producto STRING,
9   id_cliente STRING,
10  id_ubigeo STRING,
11  id_tiempo STRING,
12  id_venta STRING,
13  cantidad STRING,
14  Precio_producto STRING,
15  totalB STRING,
16  utilidad STRING,
17  igv_rec STRING
18 )
19 ROW FORMAT DELIMITED
20 FIELDS TERMINATED BY ','
21 LINES TERMINATED BY '\n'
22 STORED AS TEXTFILE
23 LOCATION '/user/diego/datalake/rdv/proy_ventas'
24 TBLPROPERTIES (
25   'skip.header.line.count'='1'
26 );
27 --Cargamos el archivo en la ruta de la tabla
28 LOAD DATA LOCAL INPATH '/home/diego/dataset/proyecto/ventas.csv' INTO TABLE RDV.PROY_VENTAS;
```



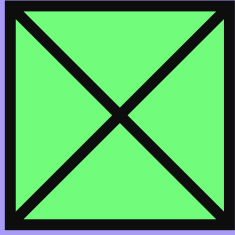
Capa RDV

Los archivos utilizados:

- carga_rdv.sql
- proy_carga_rdv.sh

proy_carga_rdv.sh

```
^ v ## [HIVE] ^
1 #hdfs dfs -mkdir /user/diego/datalake/rdv
2 #hdfs dfs -mkdir /user/diego/datalake/udv
3 #hdfs dfs -mkdir /user/diego/datalake/ddv
4
5 ## [HIVE]
6 beeline -u jdbc:hive2:// -f /home/diego/scripts/proyecto/carga_rdv.sql
7 wait
```

POBLACION DEL DATALAKE

×

—

proy_proceso_udv.py

```
19
20 #####VENTAS #####
21 dfventas = spark.sql("""select * from rdv.proy_ventas""")
22
23 dfventas=dfventas.withColumn("id_emp", col("id_empleado").cast("Integer"))\
24     .withColumn("id_prod", col("id_producto").cast("Integer"))\
25     .withColumn("id_cli", col("id_cliente").cast("Integer"))\
26     .withColumn("id_ubi", col("id_ubigeo").cast("Integer"))\
27     .withColumn("id_tiem", col("id_tiempo").cast("Integer"))\
28     .withColumn("id_vent", col("id_venta").cast("Integer"))\
29     .withColumn("cant_prod", col("cantidad").cast("Integer"))\
30     .withColumn("precio_prod", col("precio_producto").cast("Double"))\
31     .withColumn("total_venta", col("totalb").cast("Double"))\
32     .withColumn("utilidades", col("utilidad").cast("Double"))\
33     .withColumn("igv", col("igv_rec").cast("Double"))
34
35
36 dfventas=dfventas.drop("id_empleado","id_producto","id_cliente","id_ubigeo","id_tiempo","id_venta","cantidad","precio_producto","totalb","utilidad","igv_re
37
38 dfventas=dfventas.withColumn("fechaactualizacion", lit('2022-12-29'))
39
40 dfventas=dfventas.filter(col("id_emp")>0)
41
42 dfventas.show(5)
43
44 escribirenhivesinparticiones(dfventas,tableName = 'hm_proy_ventas',dbName = 'udv', location = '/user/diego/datalake/udv/')
45
```

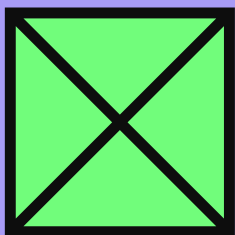
<

>

Capa UDV

Archivo utilizado:

- proy_proceso_udv.py



POBLACION DEL DATALAKE



```
proy_proceso_ddv.py
Edited

23 df_ventas=spark.sql("""select * from udv.hm_proy_ventas""")
24
25 df_empleados = spark.sql("""select * from udv.hm_proy_empleados""")
26
27 df_ventas_final=df_ventas.join(df_empleados, on = "id_emp", how = "left")
28
29
30 df_productos = spark.sql("""select * from udv.hm_proy_productos""")
31
32 df_ventas_final = df_ventas_final.join(df_productos, on = "id_prod", how = "left")
33
34
35 df_clientes = spark.sql("""select * from udv.hm_proy_clientes""")
36
37 df_ventas_final = df_ventas_final.join(df_clientes, on = "id_cli", how = "left")
38
39
40 df_ubigeo = spark.sql("""select * from udv.hm_proy_ubigeo""")
41
42 df_ventas_final = df_ventas_final.join(df_ubigeo, on = "id_ubi", how = "left")
43
44
45 df_tiempo = spark.sql("""select * from udv.hm_proy_tiempo""")
46
47 df_ventas_final = df_ventas_final.join(df_tiempo, on = "id_tiem", how = "left")
48
49
50 ventas = df_ventas_final.select("id_emp","nom_emp","cod_prod","desc_prod","cant_prod","nombre_cli","gen_cli","sucursal","codmes","annio","mes","dia","trimes","total_venta","igv","utilidades")
51
52 ventas=ventas.withColumn("fechaactualizacion", lit('2022-12-29')) #simulando fecha actual de ejecucion
53
54 ventas.show(5)
55
56 escribirenhiveconparticiones(ventas,tableName = 'hmproy_ventas_prod', partitionField = 'codmes', dropOld = True ,dbName = 'ddv', location = '/user/diego/datalake/ddv/')
```



Capa DDV

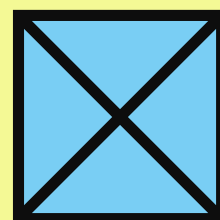
Archivo utilizado:

- proy_proceso_ddv.py

BIG QUERY

carga_bigQuery.py

```
1 from __future__ import print_function
2 import tempfile
3 from pyspark.sql import SparkSession
4
5 spark = SparkSession.builder.appName('Proyecto BigData').getOrCreate()
6
7 df1=spark.sql("""select * from ddv.hmproy_ventas_prod""")
8
9 df1.write \
10 .format("bigquery") \
11 .option("writeMethod", "direct") \
12 .save("dwh.proy_ventas")
13
14
```



```
## spark-submit --jars spark-3.1-bigquery-0.27.0-preview.jar ./carga_bigQuery.py
```

Explorer

+ ADD DATA

IK

Type to search

Viewing pinned projects.

studious-bit-364400

External connections

dwh

categoria_yanbal

proy_ventas

tbi-vuelos

tickets

ventas_yanbal

proy_ventas

*Unsaved query 2

RUN

SAVE

SHARE

SCHEDULE

MORE

```
1 SELECT * FROM 'studious-bit-364400.dwh.proy_ventas' LIMIT 20
```

Query results

SAVE RESULTS

JOB INFORMATION

RESULTS

JSON

EXECUTION DETAILS

Row	id_emp	nom_emp	cod_prod	desc_prod	cant_prod	nombre_cli	gen_cli
1	2	Quiles Puig Maria-Jose	HCP010401	Cable Audio Auxiliar Usams Ne...	2	JOSSELYN	F
2	15	Ventura Castillo Emma Violeta	L010104	Yogurt Vive Day Bebible Fresa ...	7	MARIA ANTONIA	F
3	14	Olmo Alba Jairo Diego	PS010706	Pop Corn Movie Pop natural x ...	5	JUSTINA	F
4	19	Silva Oviedo Miriam Lidia	B020210	Bebida Energizante Red Bull 35...	5	ALEX ROBERTO	M
5	2	Quiles Puig Maria-Jose	A010401	Gelatina Universal Fresa Diet x ...	6	ALEX ROBERTO	M
6	16	Noguera Cruz Pamela	PB010102	Turron PYC x 500 Gr	2	YSABEL	F
7	10	Sanchez Bartolome Aurelia Ariel	B020210	Bebida Energizante Red Bull 35...	8	ALEX ROBERTO	M
8	6	Alcazar Garces Saray Joaquina	PS010102	Camote Villa Natura x 150 g	1	CESAR OSWALDO ALEXIS	M
9	1	Hernando Mesequer Antonia R...	HCP010401	Cable Audio Auxiliar Usams Ne...	6	JOSSELYN	F
10	18	San-Martin Sanchez Alina Naira	PB010102	Turron PYC x 500 Gr	5	YSABEL	F
11	20	Mesequer Cabrera Antonia	PS010102	Camote Villa Natura x 150 g	5	CESAR OSWALDO ALEXIS	M
12	18	San-Martin Sanchez Alina Naira	PS010706	Pop Corn Movie Pop natural x ...	10	JUSTINA	F
13	7	Reyes Palma Julião Alvaro	A010401	Gelatina Universal Fresa Diet x ...	3	ALEX ROBERTO	M
14	12	Castilla Serafin Luis Noe	L010104	Yogurt Vive Day Bebible Fresa ...	5	MARIA ANTONIA	F
15	13	Da Silva Torres Xavier Tomás	A010201	Conserva de Duraznos en Mita...	7	ESTHER OFELIA	F
16	6	Alcazar Garces Saray Joaquina	PB010406	Panetón D'Onofrio Chocotón 5...	1	FRANCELLIS VICTOR	M
17	18	San-Martin Sanchez Alina Naira	PS010302	Mani Cervezero Granuts x 180 g	7	REYNA	F
18	9	Codina Calderon Salome Alba	C010501	Combo 01 Pollo Rostizado Ta...	4	DIANA VALERIA	F

Load more

DATA STUDIO

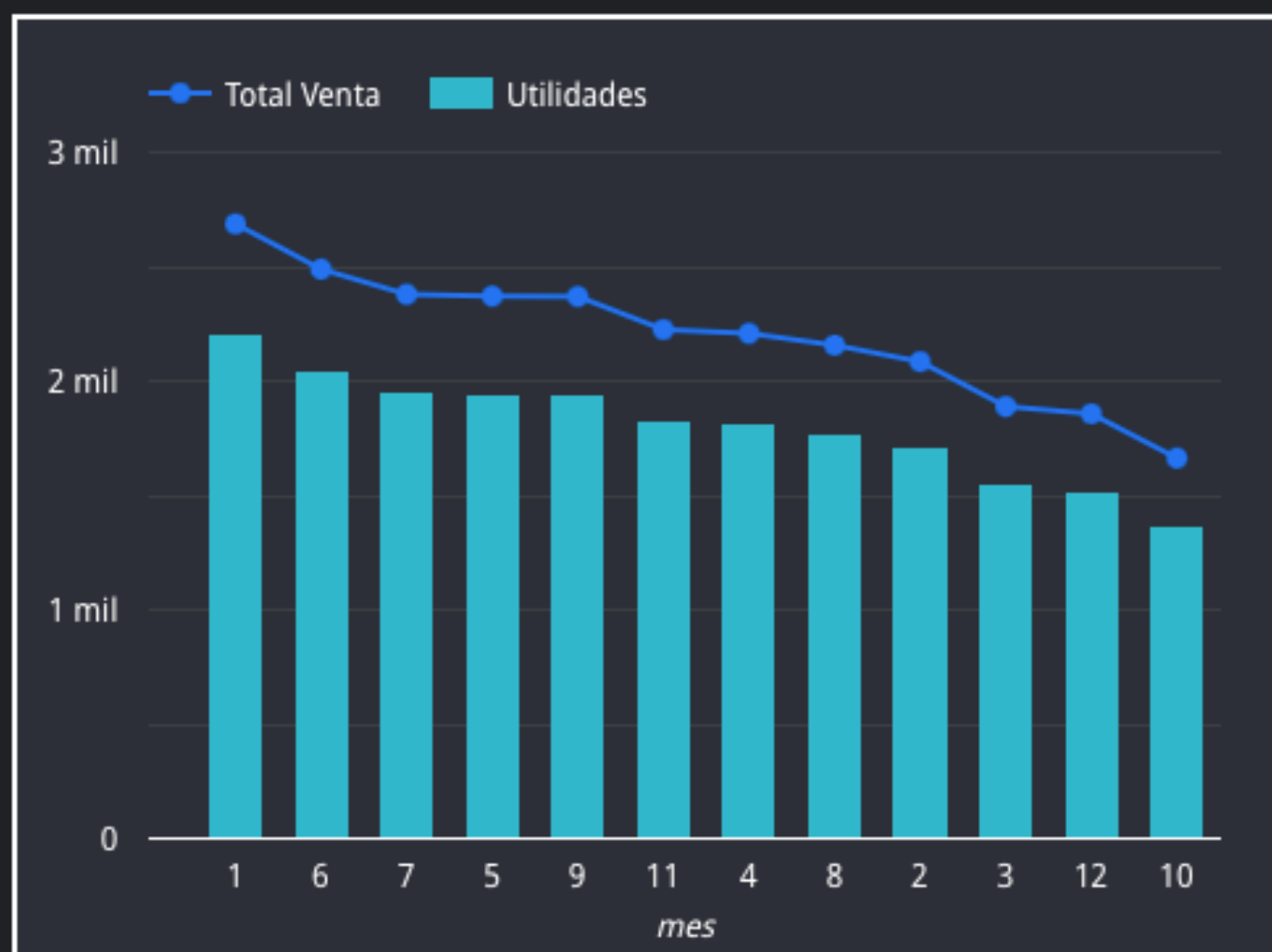


Total Ventas
26,4 mil

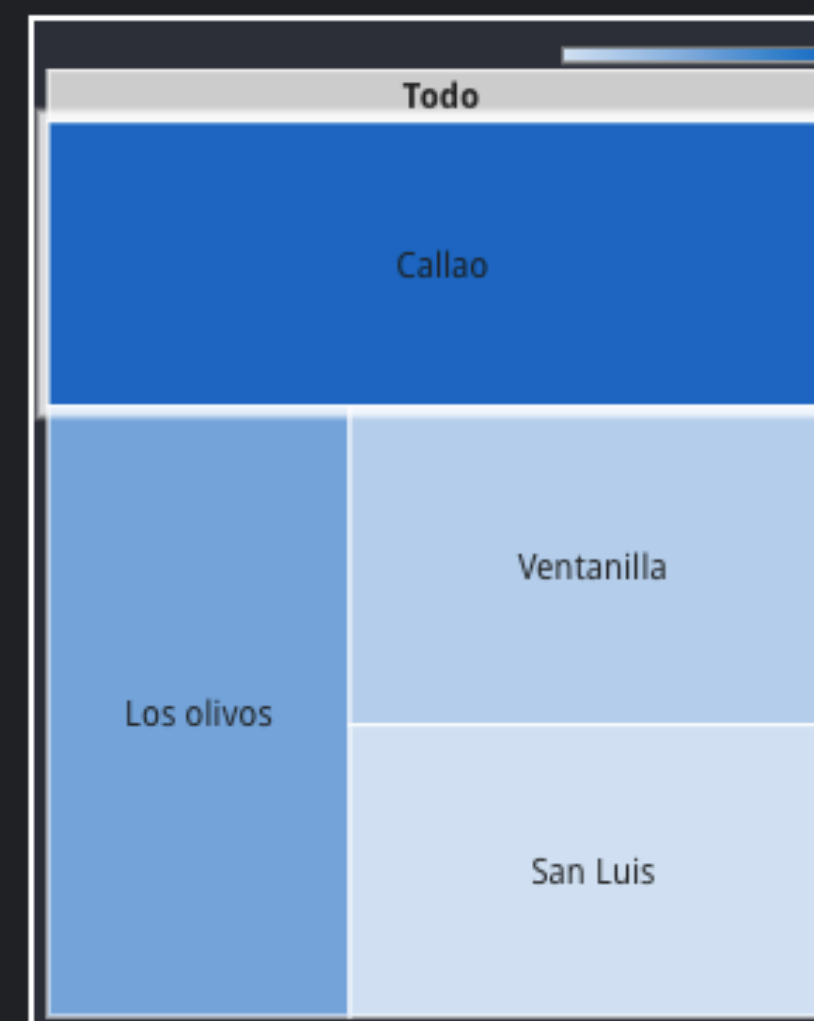
Total Neto
21,6 mil

Mes 1 12

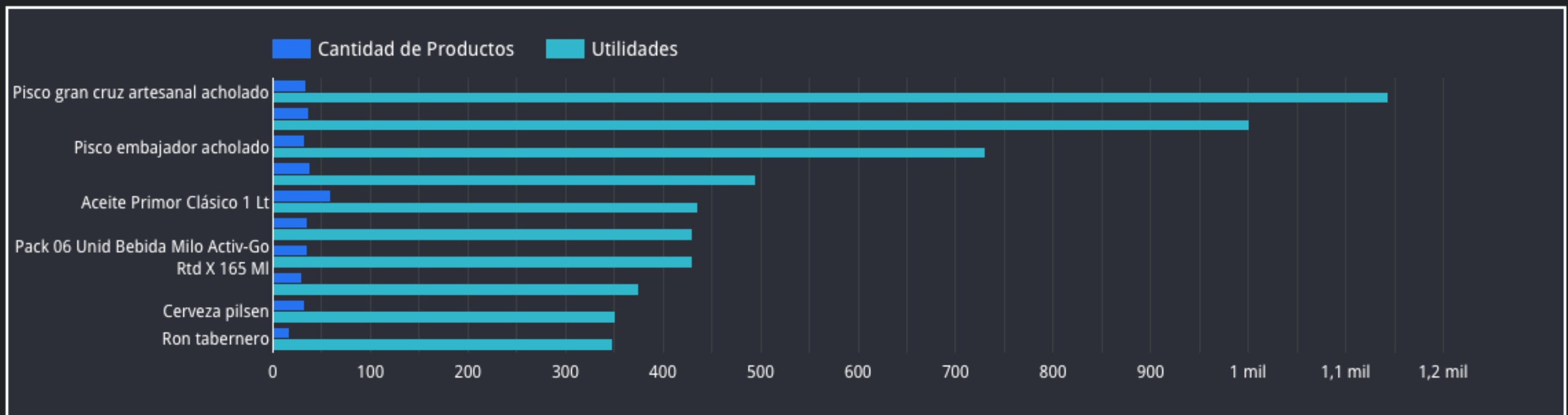
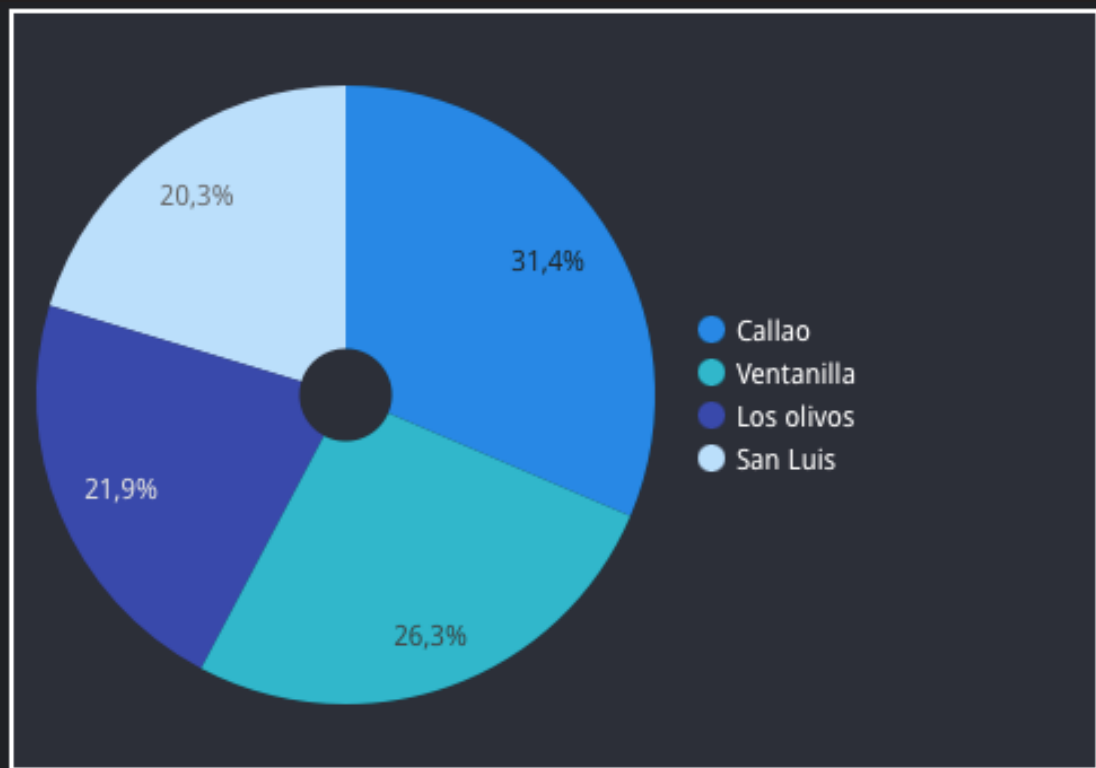
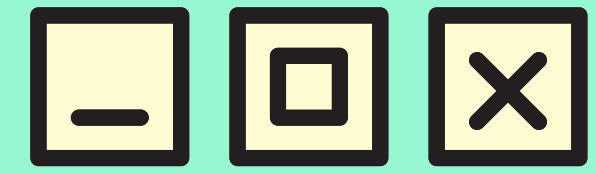
Año

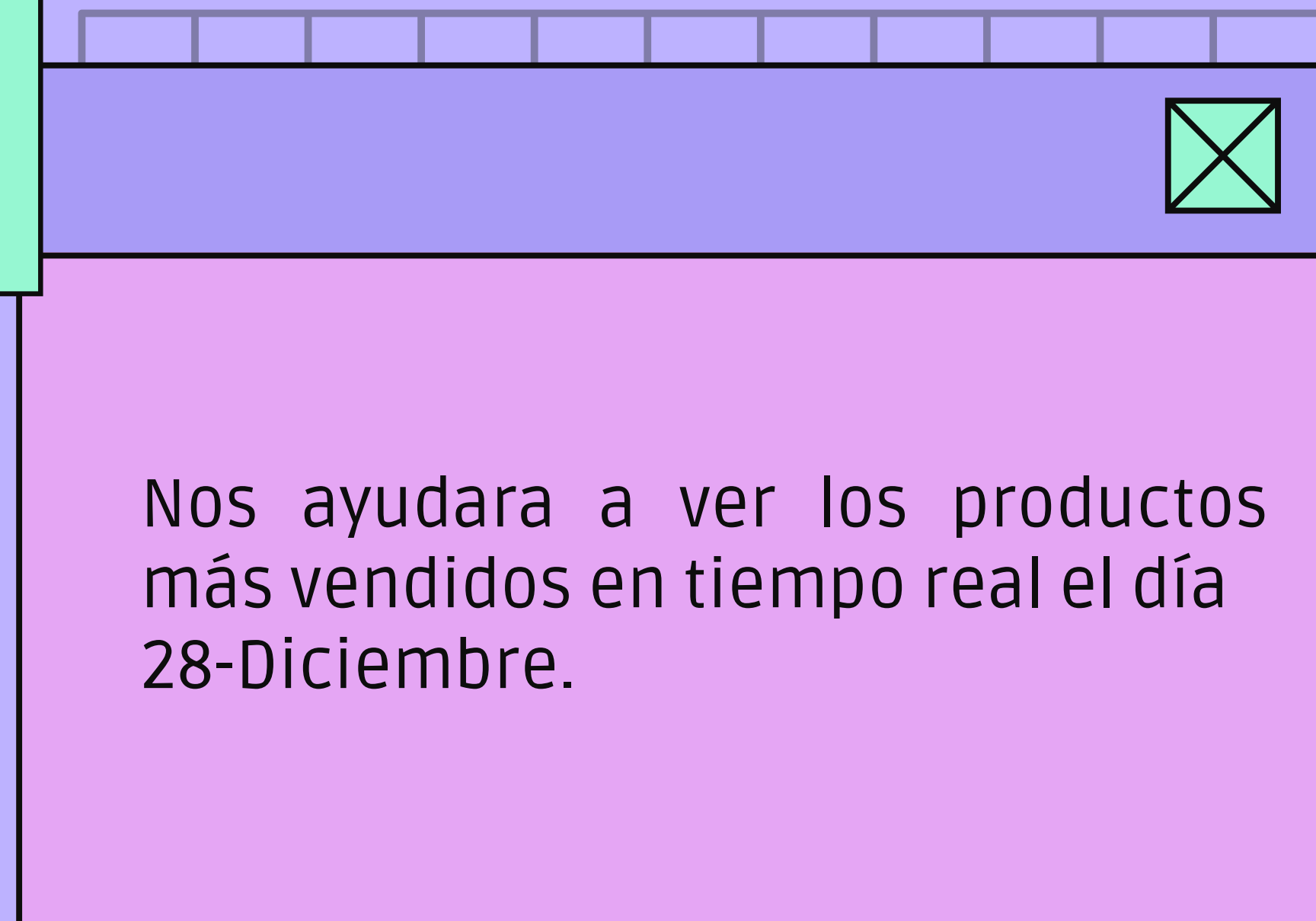
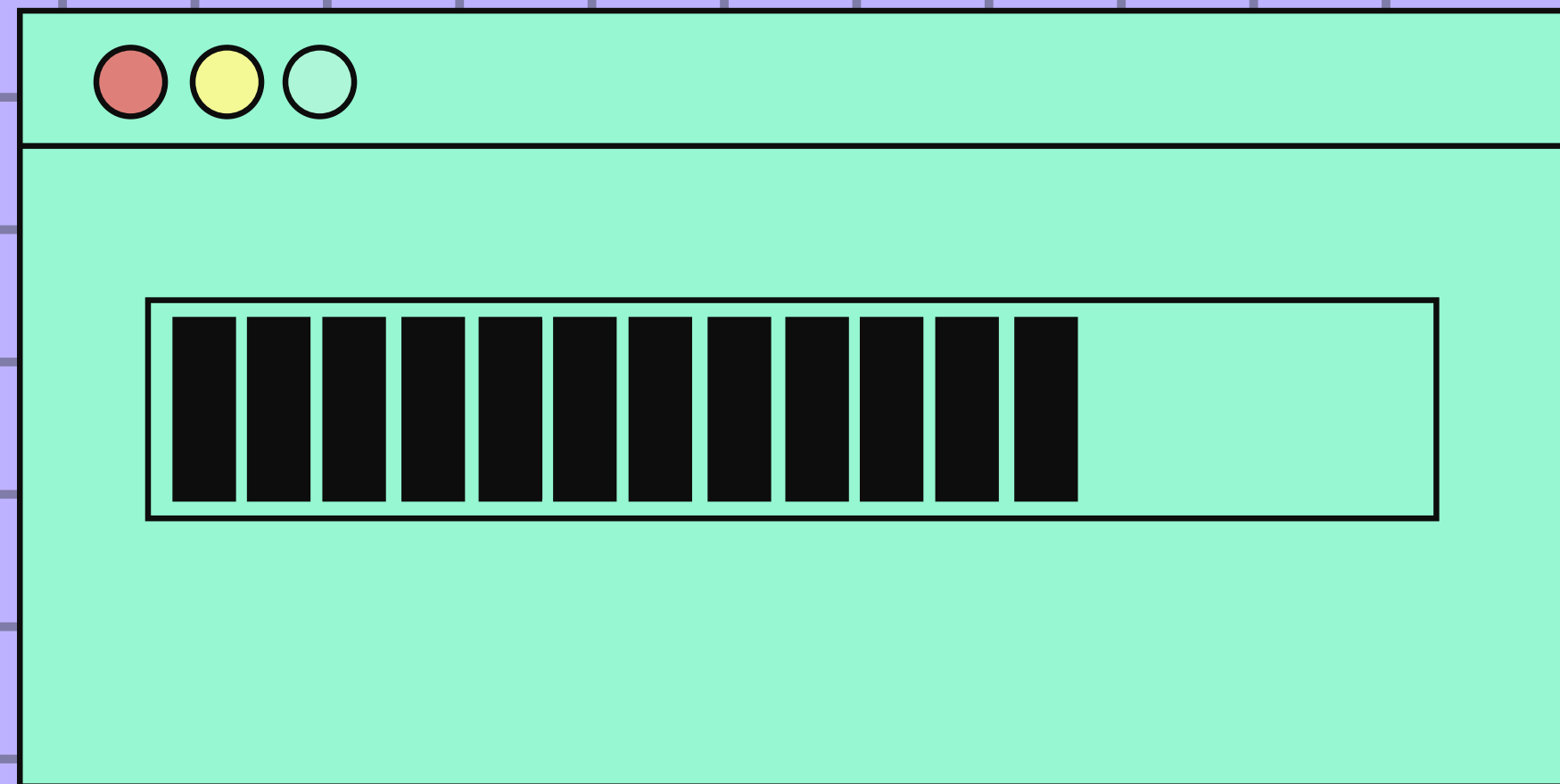


	Nombre Empl...	Total Venta ▾
1.	Ureña Viera Samar...	1.924,6
2.	San-Martin Sanche...	1.795,4
3.	Codina Calderon Sa...	1.690,6
4.	Castilla Serafin Luis...	1.649,4
5.	Ventura Castillo Em...	1.601,4
6.	Meseguer Cabrera ...	1.515
7.	Da Silva Torres Xavi...	1.481,4
8.	Noguera Cruz Pamela	1.393,6
9.	Alcazar Garces Sara...	1.345,3
10.	Campoy Noguera N...	1.340



DATA STUDIO

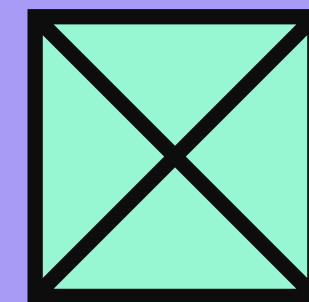
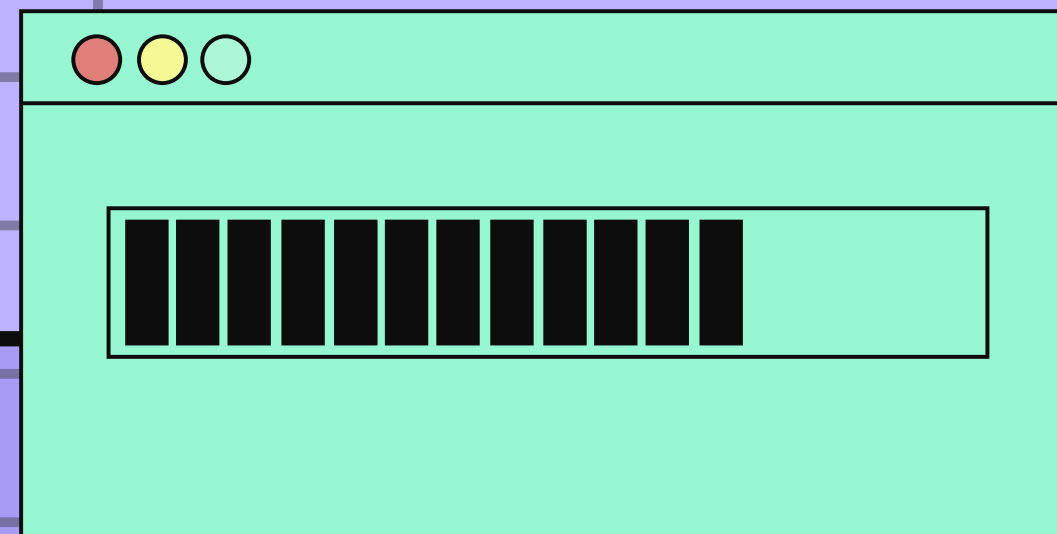




Nos ayudara a ver los productos
más vendidos en tiempo real el día
28-Diciembre.

REAL TIME

REAL TIME



Información de ventas actuales

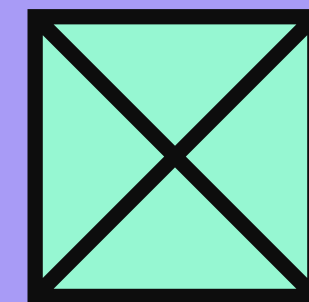
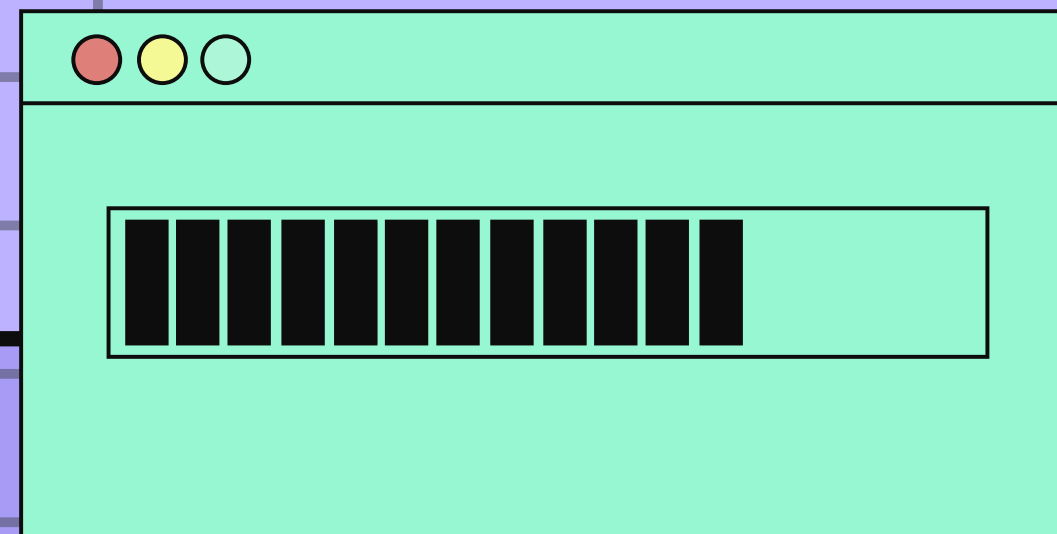
```
0: jdbc:hive2://> SELECT descripcion_producto, SUM(cantidad) from PROY_TIEMPOREAL.RESULTS1 group by descripcion_producto;
Query ID = diego_20221015011955_6899e688-d6e6-4b8e-b7a3-76df92e76872
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665794587624_0001)
```

OK

descripcion_producto	_c1
Crisinos Soy Diet Integrales Chia x 80 gr	5
Lomo Saltado Tambo Tugou	6
Maní Con Pasas Villa Natura 80 g	10
Pack Laive Queso Fundido + Jamonada/Pollo x 170 Gr	7
Papel Higiénico Noble Doble Hoja 2 und	8
Pila Duracell AA 2 und	6
Wafer Mega Gn Sabor Vainilla x 61 Gr	8

7 rows selected (6.579 seconds)

REAL TIME



Agregando el archivo ventas1.csv

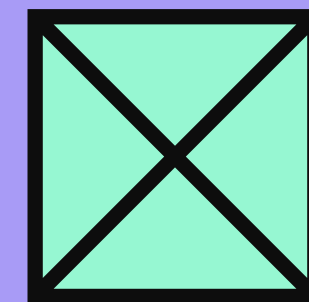
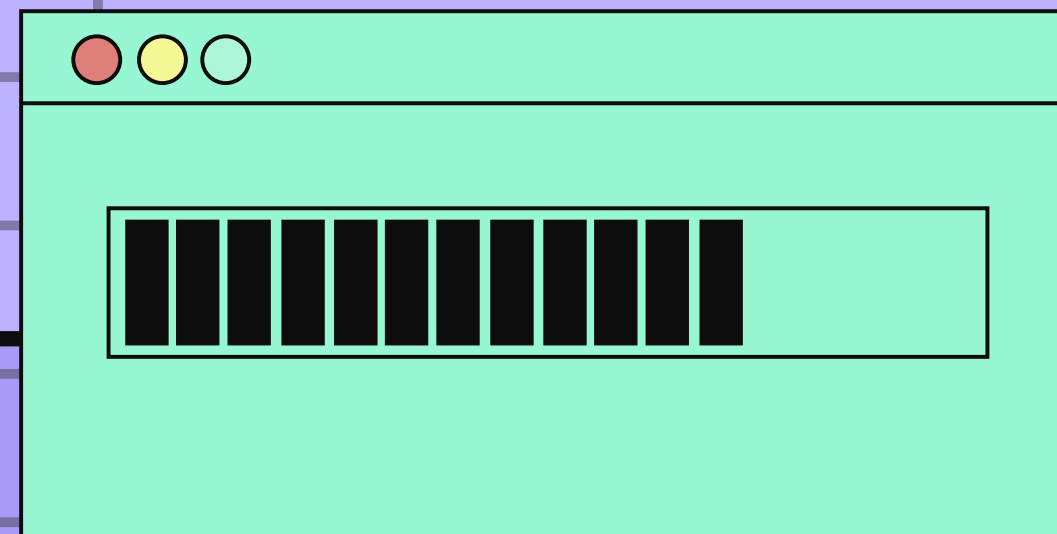
```
0: jdbc:hive2://> SELECT descripcion_producto, SUM(cantidad) from PROY_TIEMPOREAL.RESULTS1 group by descripcion_producto;
Query ID = diego_20221015012442_09495633-44b4-48ff-a9e8-f34b82e7620d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665794587624_0001)
```

OK

descripcion_producto	_c1
Chocoton La Vita Panettiere Mini X 70 Gr	2
Crisinos Soy Diet Integrales Chia x 80 gr	5
Lomo Saltado Tambo Tugou	6
Maní Con Pasas Villa Natura 80 g	10
Pack Laive Queso Fundido + Jamonada/Pollo x 170 Gr	7
Panetón D'Onofrio Chocotón 500 g	6
Panetón Todinnito x 100 gr	3
Panetón Todinno pack caja 900 gr	2
Papel Higiénico Noble Doble Hoja 2 und	8
Pila Duracell AA 2 und	6
Wafer Mega Gn Sabor Vainilla x 61 Gr	8

11 rows selected (11.756 seconds)

REAL TIME



Agregando el archivo ventas2.csv

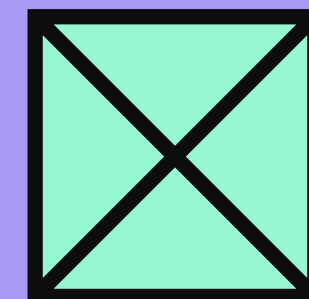
```
0: jdbc:hive2://> SELECT descripcion_producto, SUM(cantidad) from PROY_TIEMPOREAL.RESULTS1 group by descripcion_producto;
Query ID = diego_20221015012603_3e92b6af-d51a-46b3-936e-ec27754e8a38
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665794587624_0001)
```

OK

descripcion_producto	_c1
Chocoton La Vita Panettiere Mini X 70 Gr	8
Crisinos Soy Diet Integrales Chia x 80 gr	5
Lomo Saltado Tambo Tugou	6
Maní Con Pasas Villa Natura 80 g	10
Pack Laive Queso Fundido + Jamonada/Pollo x 170 Gr	7
Panetón D'Onofrio Chocotón 500 g	12
Panetón Todinnito x 100 gr	3
Panetón Todinno pack caja 900 gr	2
Papel Higiénico Noble Doble Hoja 2 und	8
Pila Duracell AA 2 und	6
Wafer Mega Gn Sabor Vainilla x 61 Gr	8

11 rows selected (9.012 seconds)

REAL TIME



Agregando el archivo ventas3.csv

```
0: jdbc:hive2://> SELECT descripcion_producto, SUM(cantidad) from PROY_TIEMPOREAL.RESULTS1 group by descripcion_producto;
Query ID = diego_20221015012708_f838d724-52a4-464f-b731-f7658b53e814
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1665794587624_0001)
```

OK

descripcion_producto	_c1
Chocoton La Vita Panettiere Mini X 70 Gr	23
Crisinos Soy Diet Integrales Chia x 80 gr	5
Lomo Saltado Tambo Tugou	6
Maní Con Pasas Villa Natura 80 g	10
Pack Laive Queso Fundido + Jamonada/Pollo x 170 Gr	7
Panetón D'Onofrio Chocotón 500 g	12
Panetón Todinnito x 100 gr	3
Panetón Todinno pack caja 900 gr	2
Papel Higiénico Noble Doble Hoja 2 und	8
Pila Duracell AA 2 und	6
Wafer Mega Gn Sabor Vainilla x 61 Gr	8

11 rows selected (9.079 seconds)

CONCLUSIONES

- Se gestionó de manera efectiva una solución de Big Data en la cual el principal enfoque estuvo en los datos brindados por la empresa empleando Apache Spark y Apache Hadoop. Y como resultado obtuvimos un reporte con los datos procesados previamente que ayudará en la toma de decisiones de la empresa.
- Se implementó una solución de tiempo real que actualiza las ventas con cada carga de archivos nuevos.