I.U. PASCUAL BRAVO

ET-0155 – Fundamentos de Big Data – Grupo 0100

Periodo 2024-2

Profesor: Jaime E Soto U

INFORME

Caso de Estudio: Empresa “Gaseosas Poderosas”

Solución de un problema de Big data a través de un proceso ETL/ELT

Grupo 2

Harlan Santiago Enciso Riaño

Miguel Angeles Rojas Pabon

Maria Camila Rodriguez Ortiz

Institución Universitaria Pascual Bravo

Facultad de Ingeniería

Medellín

2025

I.U. PASCUAL BRAVO
ET-0155 – Fundamentos de Big Data – Grupo 0100
Periodo 2024-2
Profesor: Jaime E Soto U

ANÁLISIS DEL CASO DE ESTUDIO

A continuación, se presenta el caso de la empresa “Gaseosas Poderosas”, una franquicia especializada en la venta de refrescos económicos. En la actualidad, su sistema de ventas es un archivo de hoja de cálculo, por lo que existen problemas de duplicados, inconsistencias y desorden. Además, se envían los reportes diarios de cada punto de venta como archivos separados; esto impide realizar un análisis de toda la información, sin mencionar que se encuentra desactualizada. Por lo tanto, el presente proyecto tendrá como objetivo la implementación del proceso ETL: extracción, transformación y carga de datos, para llevar registros de ventas a una base de datos relacional PostgreSQL, mediante la cual se garantiza la integridad, coherencia y disponibilidad de la información para tomar decisiones. Para la etapa de extracción, se han identificado datos de regiones, departamentos, municipios y transacciones de ventas; se ha detectado el planteamiento de los problemas relativos al formato de la fecha, valores faltantes, negativos, limpieza de datos y asignación de datos. En cuanto a la carga, los datos se almacenan en tablas estandarizadas como departamentos, municipios, productos, transacciones y regiones, se definen las relaciones entre ellas y una vista conjunta para consultas. Después de implementar esta solución, el informe se podrá desarrollar el panel de control, análisis de ventas e informes y gráficos de Pareto, lo que mejorará la eficiencia en marketing, compra y ventas.

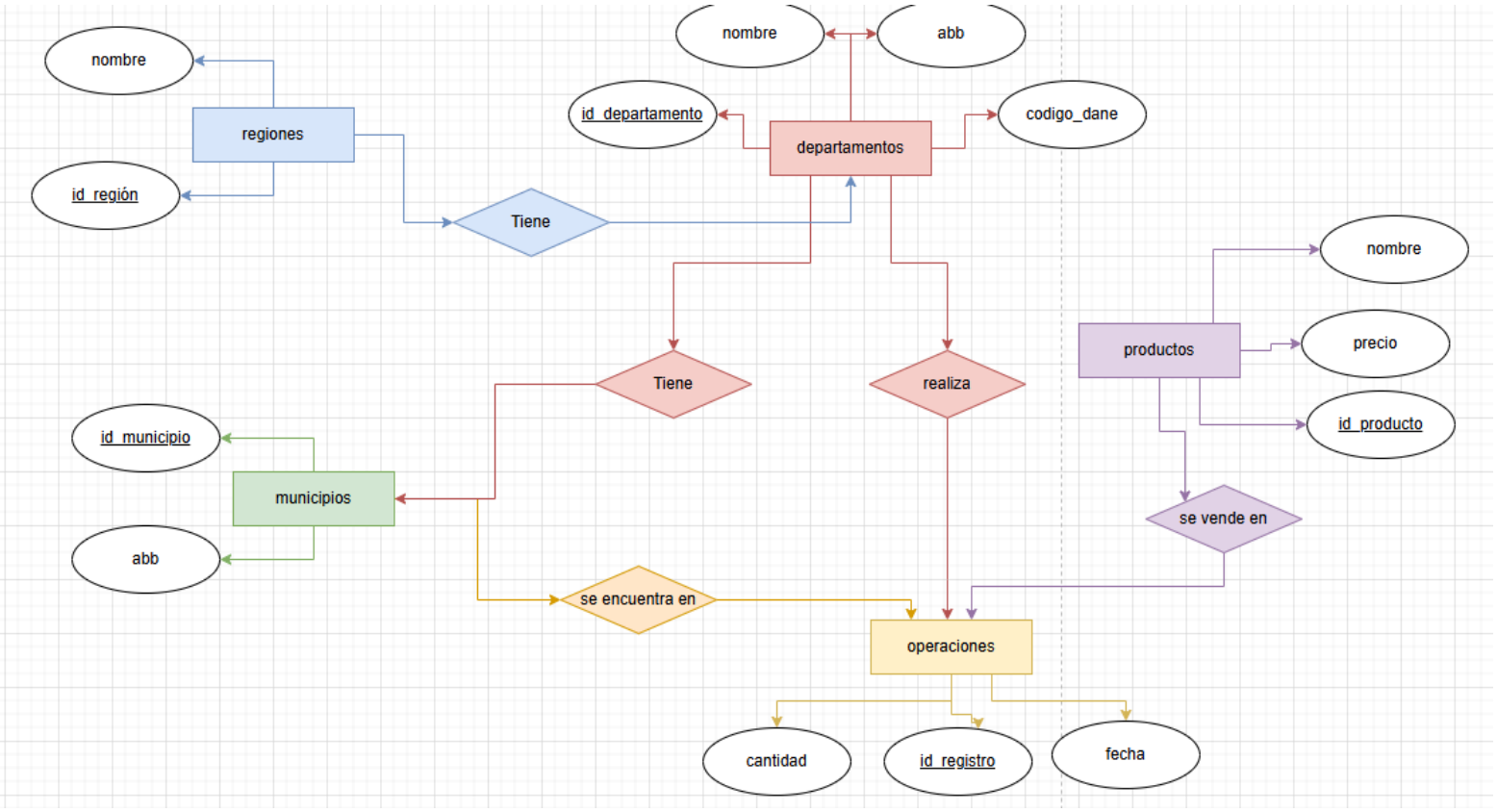
I.U. PASCUAL BRAVO

ET-0155 – Fundamentos de Big Data – Grupo 0100

Periodo 2024-2

Profesor: Jaime E Soto U

DIAGRAMA CHEN



DICCIONARIO DE DATOS

Tabla: temporal

Campo	Tipo de Dato	Tamaño	PK	FK	Descripción
codigo_dep	CHARACTER VARYING	10	✔	—-	Codigo del departamento
codigo_mun	CHARACTER VARYING	10	—-	—-	Codigo del municipio
codigo_region	INTEGER	—--	—--	—--	Codigo de la region
departamento	text	—-	—--	—--	Nombre del departamento
municipio	text	—---	—---	—--	Nombre del

					municipio
region	text	———	——	——	Nombre de la region

Tabla: departamentos

Campo	Tipo de Dato	Tamaño	PK	FK	Descripción
id_departamento	INTEGER	———	✔	——	Identificador del departamento
nombre	CHARACTER VARYING	70	——	———	Nombre del departamento
abb	CHARACTER VARYING	3	——	———	Abreviatura del departamento
codigo_dane	CHARACTER VARYING	10	———	———	Codigo DANE
codigo_region	INTEGER	———	———	———	Relación con la tabla region
población	INTEGER	——	———	———	Población del departamento

Tabla: municipios

Campo	Tipo de Dato	Tamaño	PK	FK	Descripción
id municipio	INTEGER	——	✔	——	Identificador del municipio
id_departamento	INTEGER	——	——	✔	Relación con la tabla departamento

nombre	CHARACTER VARYING	70	—--	—--	Nombre del municipio
abb	CHARACTER VARYING	5	—--	—--	Abreviatura del municipio
codigo_dane	CHARACTER VARYING	10	—--	—--	Codigo DANE
población	INTEGER	—--	—--	—--	Población del municipio

Tablas: productos

Campo	**Tipo de dato**	**Tamaño**	**PK**	**FK**	**Descripción**
id_producto	INTEGER	—--	✔	—--	Identificador del producto
nombre	CHARACTER VARYING	20	—--	—--	Nombre del producto
precio	INTEGER	—--	—--	—--	Precio unitario del producto

Tablas: operaciones

Campo	**Tipo de dato**	**Tamaño**	**PK**	**FK**	**Descripción**
id_registro	INTEGER	—--	✔	—--	Identificador de la operación
id_departamento	INTEGER	—--	—--	✔	Relación de la tabla departamentos
id_municipio	INTEGER	—--	—--	✔	Relación de la tabla municipios
id_producto	INTEGER	—--	—--	✔	Relación de la tabla productos
fecha	CHARACTER	10	—--	—--	Fecha de la

	VARYING				operación (AAAA-MM-DD)
cantidad	INTEGER	—--	—--	—--	Cantidad vendida
estado	CHARACTER VARYING	1	—--	—--	Estado de la operación (F o V)

Tabla: regiones

Campo	Tipo de dato	Tamaño	PK	FK	Descripción
id_region	INTEGER	—--	✔	—--	Identificador de la region
nombre	CHARACTER VARYING	20	—--	—--	Nombre de la region

4. Algoritmo (Python) actualizado con solución al problema de transformación del formato de fecha.

¿Qué hace el algoritmo ETL?

El algoritmo implementa un proceso de Extracción, Transformación y Carga (ETL) para procesar datos geográficos de Colombia. Su objetivo principal es normalizar y estructurar información sobre departamentos y municipios colombianos, convirtiéndola desde un formato plano CSV hacia una base de datos relacional estructurada.

¿De cuáles fuentes se obtienen los datos?

Los datos provienen de una fuente principal:
Archivo CSV: colombia-dane-departamentos.csv
Origen de los datos: DANE (Departamento Administrativo Nacional de Estadística de Colombia)

URL de referencia:
https://www.datos.gov.co/Mapas-Nacionales/Departamentos-y-municipios-de-Colombia/xdk5-pm3f/data

¿Qué procesos realiza?
El algoritmo ejecuta los siguientes procesos en secuencia:

1. Fase de Extracción:
Lee el archivo CSV fila por fila
Omite la cabecera del archivo
Extrae datos de región, departamento y municipio

2. Fase de Transformación:

- Mapeo de regiones: Convierte nombres de regiones a códigos numéricos (1-6)
- Generación de IDs únicos:
- Departamentos: Combina código país (57) + código departamento con formato de 2 dígitos
- Municipios: Combina ID departamento + contador secuencial con formato de 3 dígitos
- Validación de datos: Verifica longitud de códigos
- Normalización: Trunca nombres largos a 50 caracteres.

3. Fase de Carga:

- Carga temporal: Inserta todos los datos en una tabla temporal
- Carga departamentos: Agrupa y carga departamentos únicos con sus nuevos códigos
- Carga municipios: Carga municipios ordenados con IDs secuenciales por departamento

4. Procesos Auxiliares:

- Limpieza de tablas (TRUNCATE)
- Manejo de transacciones
- Control de errores y excepciones
- Logging detallado del proceso

5. Identificación, análisis y corrección de registros con problemas en la tabla "operaciones". Tip: tiene que cargar la tabla operaciones primero (utilice script)

Problema: Algunos registros de la tabla 'operaciones' no presentan el formato de fecha correcto (AAAA-MM-DD). Se observan inconsistencias como fechas con guiones sin ceros a la izquierda (2024-8-21), barras (21/08/2024), o guiones en orden día-mes-año (21-08-2024).

Para normalizar todas las fechas se construyó un pipeline de actualizaciones en SQL, ejecutado en el siguiente orden:

- Normalizar ceros faltantes (YYYY-M-D → YYYY-MM-DD)

```
UPDATE operaciones

SET fecha = TO_CHAR(TO_DATE(fecha, 'YYYY-M-D'), 'YYYY-MM-DD')

WHERE fecha ~ '^[0-9]{4}-[0-9]{1,2}-[0-9]{1,2}$';
```

- Transformar años de dos dígitos (YY-MM-DD → YYYY-MM-DD)

```
UPDATE operaciones

SET fecha = TO_CHAR(TO_DATE(fecha, 'YY-MM-DD'), 'YYYY-MM-DD')

WHERE fecha ~ '^[0-9]{2}-[0-9]{2}-[0-9]{2}$';
```

- Reinterpretar meses imposibles (>12) como YY-DD-MM.

```
UPDATE operaciones

SET fecha = TO_CHAR(TO_DATE(fecha, 'YY-DD-MM'), 'YYYY-MM-DD')

WHERE fecha ~ '^[0-9]{2}-[0-9]{2}-[0-9]{2}$'

  AND SPLIT_PART(fecha,'-',2)::int > 12;
```

- Corregir años de tres dígitos anteponiendo 2

```
UPDATE operaciones

SET fecha = '2' || fecha

WHERE fecha ~ '^[0-9]{3}-[0-9]{2}-[0-9]{2}$';
```

- Convertir fechas en formato DD-MM-YYYY a YYYY-MM-DD

```
UPDATE operaciones
SET fecha = TO_CHAR(TO_DATE(fecha, 'DD-MM-YYYY'), 'YYYY-MM-DD')
WHERE fecha ~ '^[0-9]{2}-[0-9]{2}-[0-9]{4}$'
  AND SPLIT_PART(fecha,'-',2)::int <= 12;
```

- Convertir fechas en formato MM-DD-YYYY a YYYY-MM-DD

```
UPDATE operaciones
SET fecha = TO_CHAR(TO_DATE(fecha, 'MM-DD-YYYY'), 'YYYY-MM-DD')
WHERE fecha ~ '^[0-9]{2}-[0-9]{2}-[0-9]{4}$'
  AND SPLIT_PART(fecha,'-',2)::int > 12;
```

De esta manera se logró que todos los registros cumplan con el formato estándar requerido.

6. Estrategia y solución de “limpieza” de los registros que no son válidos

Problema: Algunos registros de la tabla 'operaciones' no contienen información completa. Se identificaron los siguientes casos: cantidades en cero, cantidades negativas e identificadores de producto incompletos.

Propuesta de solución: Aplicar técnicas de imputación y reglas de negocio para recuperar la información, evitando la eliminación de registros.

Tipo de problema	Número Registro	Solución
Fecha mal formateada	152	Convertidas con TO_DATE a YYYY-MM-DD
Cantidad negativa	35	Transformadas con ABS(cantidad)
Cantidad = 0	120	Imputadas con promedio histórico por municipio/producto
id_producto=0	12	Imputados como “NARANJITA” (id=4)

Cantidades = 0:

```
UPDATE operaciones o
SET cantidad = sub.promedio
FROM (
   SELECT id_producto, id_municipio, ROUND(AVG(cantidad)) AS promedio
   FROM operaciones
   WHERE cantidad > 0
   GROUP BY id_producto, id_municipio
) sub
WHERE o.cantidad = 0
  AND o.id_producto = sub.id_producto
  AND o.id_municipio = sub.id_municipio;
```

Cantidades negativas:

```
UPDATE operaciones
SET cantidad = ABS(cantidad)
WHERE cantidad < 0;
```

Productos faltantes (`id_producto=0`):

```
UPDATE operaciones
SET id_producto = 4
WHERE id_producto = 0
  AND id_municipio = 5705108;
```

7. Tabla con las consultas SQL solicitadas.

#	Descripción	Consulta SQL
7.1	**Seleccionar los 8 departamentos con mayor *volumen* de ventas (monto) de productos ordenados de mayor a menor.** Datos solicitados: nombre de departamento y monto total por departamento de todos los productos. Nota: Recuerde que tiene agrupar por departamento	***SELECT departamento, SUM(venta) AS monto_total FROM vista_operaciones GROUP BY departamento ORDER BY monto_total DESC LIMIT 8;***
7.2	**Seleccionar los 15 municipios con mayor *cantidad* de productos vendidos en el departamento de Antioquia ordenados de mayor a menor.** Datos solicitados: nombre municipio y cantidad total por municipio. Nota: Recuerde que tiene agrupar por municipio	***SELECT municipio, SUM(cantidad) AS total_cantidad FROM vista_operaciones WHERE departamento = 'Antioquia' GROUP BY municipio ORDER BY total_cantidad DESC LIMIT 15;***
7.3	**Seleccionar los 5 departamentos con mayor cantidad de gaseosas vendidas del producto "MANZALOCA" ordenados de mayor a menor.** Datos solicitados: nombre de departamento y cantidad total por departamento. Nota: Recuerde que tiene agrupar por departamento y filtrar por el producto.	***SELECT departamento, SUM(cantidad) AS total_cantidad FROM vista_operaciones WHERE producto = 'MANZALOCA' GROUP BY departamento ORDER BY total_cantidad DESC LIMIT 5;***
7.4	**Seleccione los 5 municipios con el menor monto de ventas de gaseosas ordenados de menor a mayor.** Datos solicitados: nombre del departamento al que pertenece, nombre municipio y monto total de ventas por municipio. Nota: Recuerde que tiene agrupar por municipio	***SELECT departamento, municipio, SUM(venta) AS monto_total FROM vista_operaciones GROUP BY departamento, municipio ORDER BY monto_total ASC LIMIT 5;***
7.5	**Consultar la cantidad de gaseosas vendidas de cada producto por cada región ordenados de mayor a menor.**	***SELECT d.codigo_region AS region, producto, SUM(cantidad) AS total_cantidad FROM vista_operaciones vo JOIN departamentos d ON vo.id_departamento = d.id_departamento GROUP BY d.codigo_region, producto ORDER BY total_cantidad DESC;***
7.6	**Consultar el total del monto de ventas de cada producto en Antioquia de mayor a menor.**	***SELECT producto, SUM(venta) AS monto_total FROM vista_operaciones***

		WHERE departamento = 'Antioquia' *GROUP BY producto* *ORDER BY monto_total DESC;*

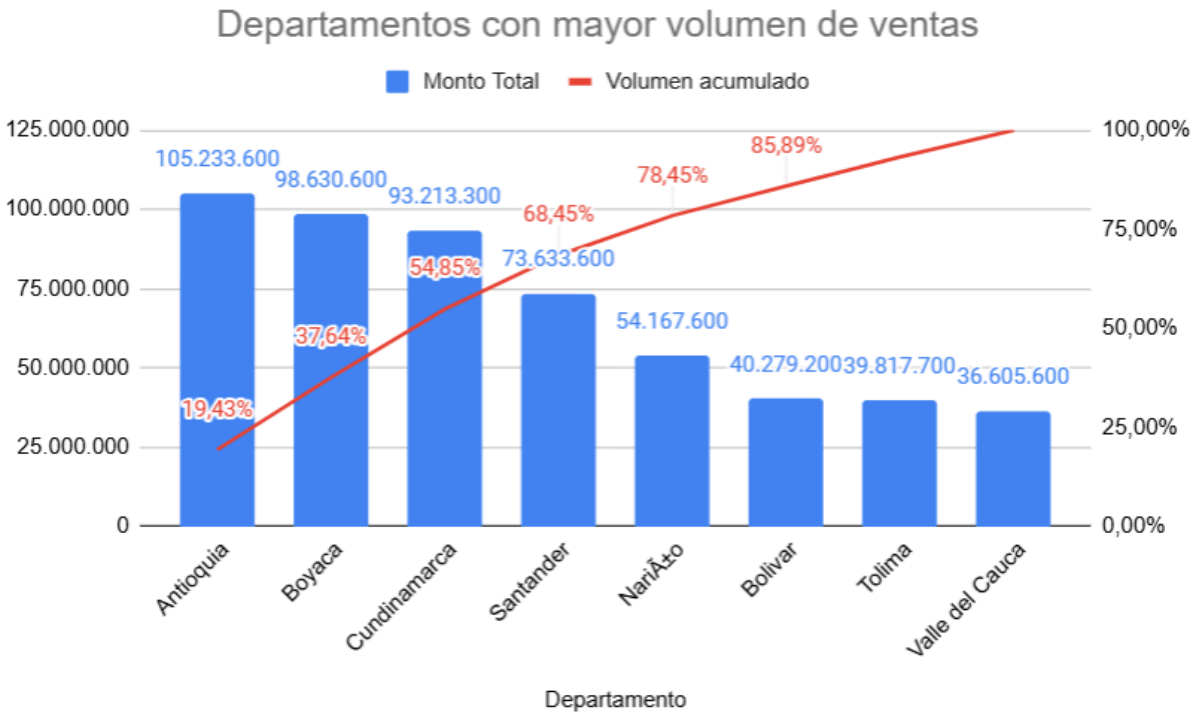
8.- Gráficos

8.1.- Gráfico de Pareto que muestra los resultados de la consulta #7.1.

8.1.1.- Resultados de la consulta.

departamento character varying (70)	monto_total bigint
Antioquia	105233600
Boyaca	98630600
Cundinamarca	93213300
Santander	73633600
NariÃ±o	54167600
Bolivar	40279200
Tolima	39817700
Valle del Cauca	36605600

8.1.2.- Gráfico.

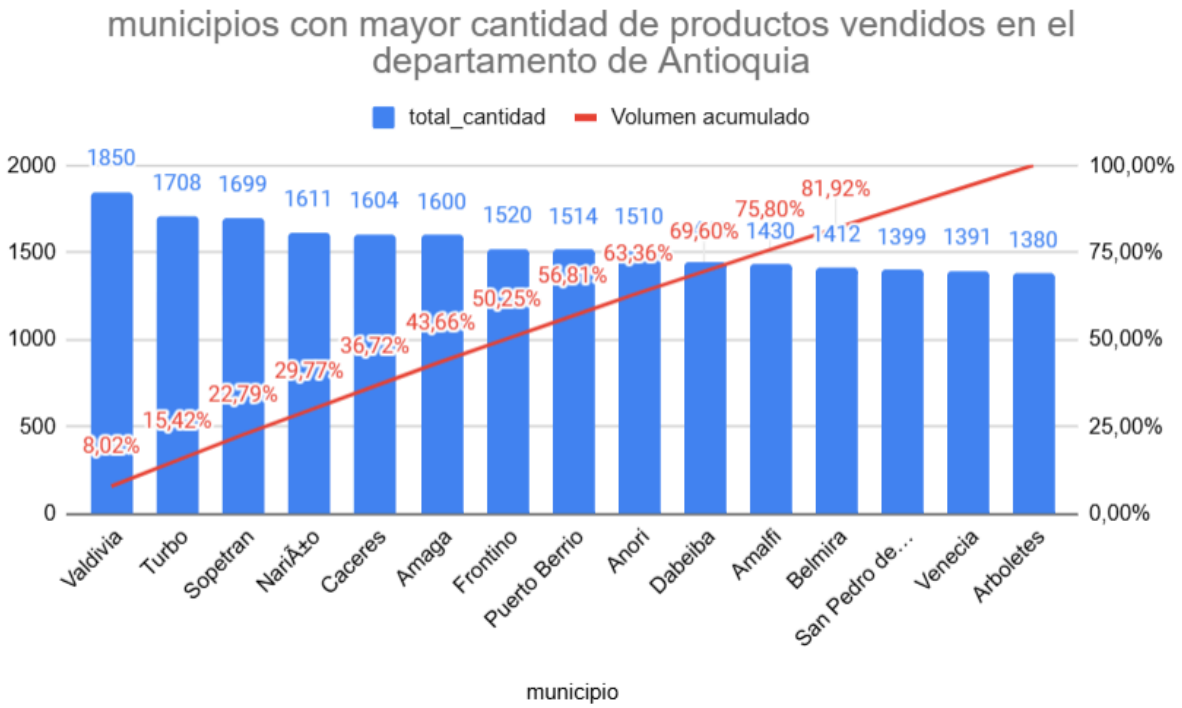


8.2.- Gráfico de Pareto que muestra los resultados de la consulta #7.2.

8.2.1.- Resultados de la consulta.

municipio character varying (70)	total_cantidad bigint
Valdivia	1850
Turbo	1708
Sopetran	1699
NariÃ±o	1611
Caceres	1604
Amaga	1600
Frontino	1520
Puerto Berrio	1514

8.2.2.- Gráfico.

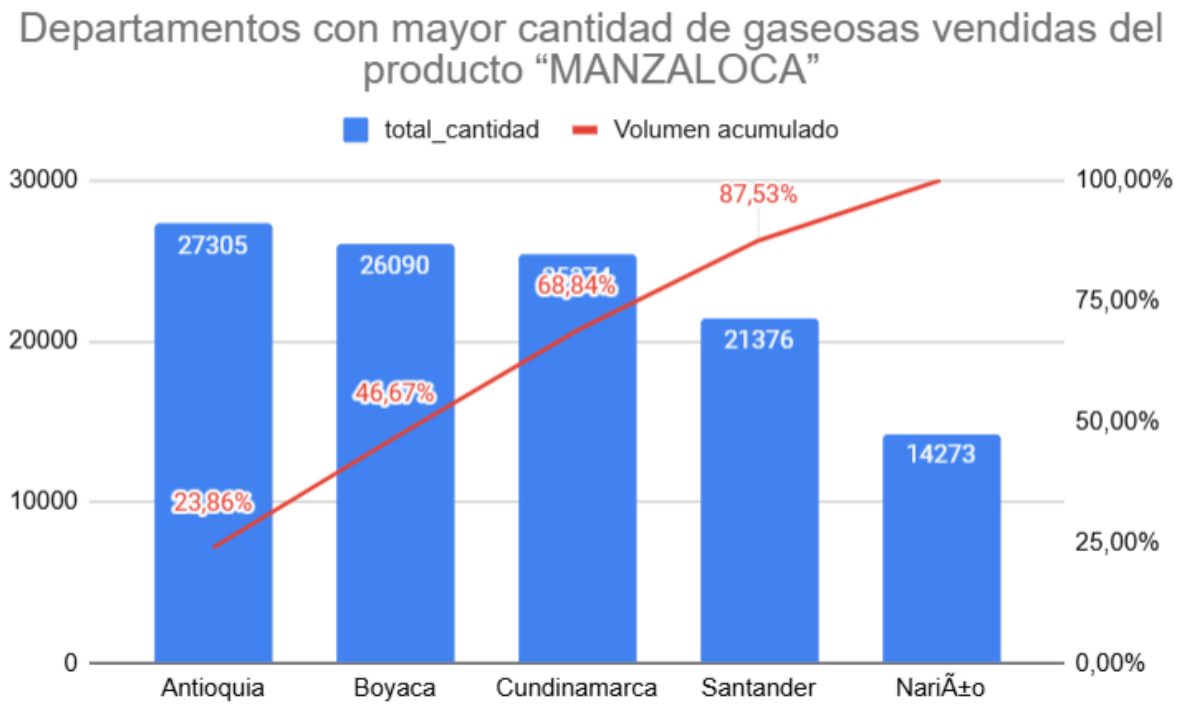


8.3.- Gráfico de Pareto que muestra los resultados de la consulta #7.3.

8.3.1.- Resultados de la consulta.

departamento character varying (70)	total_cantidad bigint
Antioquia	27305
Boyaca	26090
Cundinamarca	25374
Santander	21376
NariÃ±o	14273

8.3.2.- Gráfico.

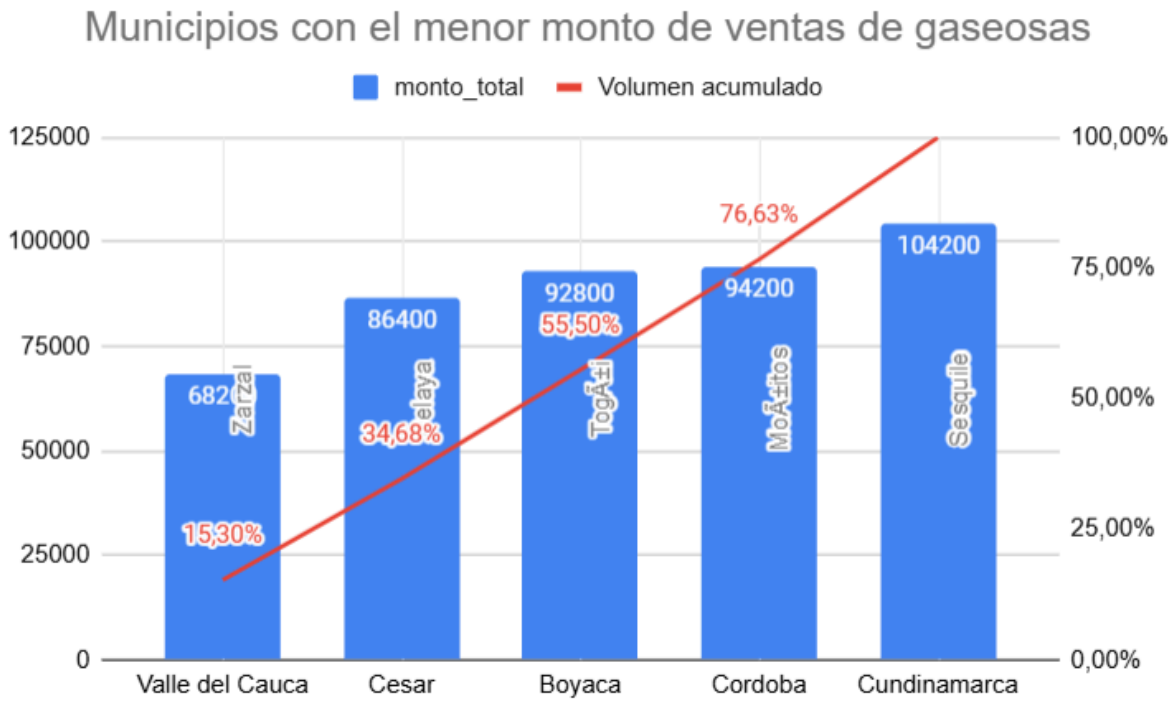


8.4.- Gráfico de Pareto que muestra los resultados de la consulta #7.4.

8.4.1.- Resultados de la consulta.

departamento character varying (70)	municipio character varying (70)	monto_total bigint
Valle del Cauca	Zarzal	68200
Cesar	Pelaya	86400
Boyaca	TogÃ±i	92800
Cordoba	MoÃ±itos	94200
Cundinamarca	Sesquile	104200

8.4.2.- Gráfico.

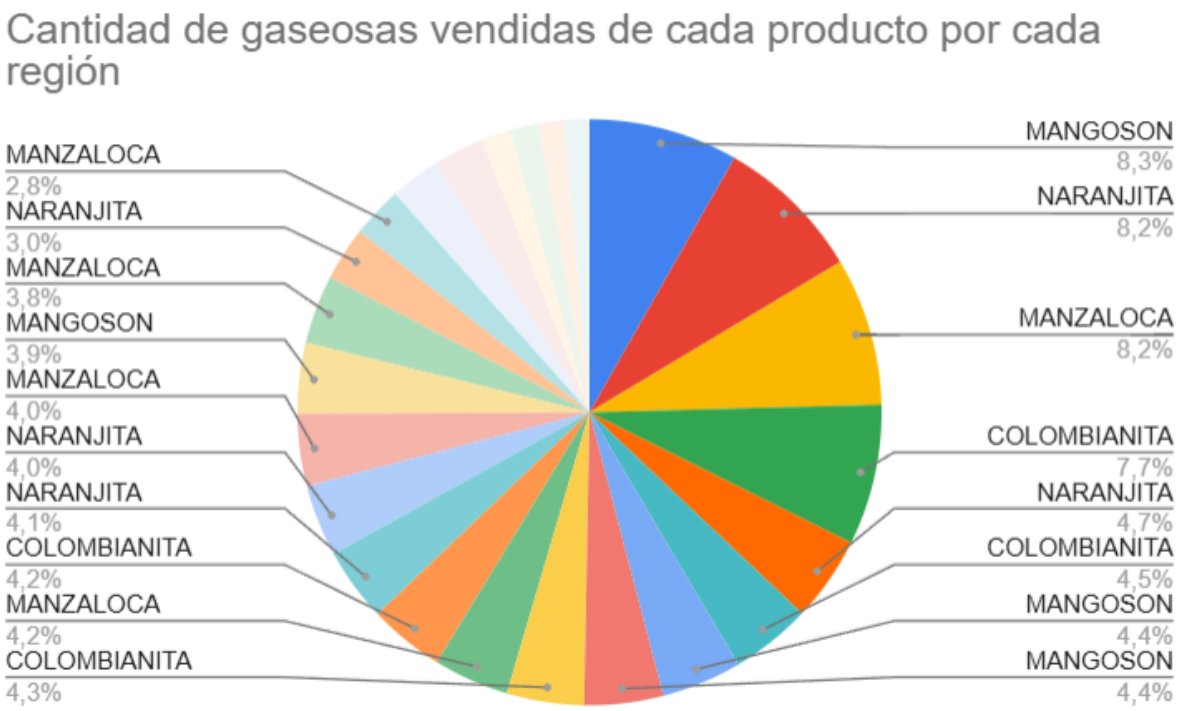


8.5.- Gráfico de Tortas que muestra los resultados de la consulta #7.5.

8.5.1.- Resultados de la consulta.

	region integer	producto character varying (20)	total_cantidad bigint
1	2	MANGOSON	83000
2	2	NARANJITA	82011
3	2	MANZALOCA	81967
4	2	COLOMBIANITA	77835
5	4	NARANJITA	47274
6	4	COLOMBIANITA	44702
7	4	MANGOSON	44596
8	6	MANGOSON	43780

8.5.2.- Gráfico.



8.6.- Gráfico de Torta que muestra los resultados de la consulta #7.6.

8.6.1.- Resultados de la consulta.

producto character varying (20)	monto_total bigint
COLOMBIANITA	35554800
MANZALOCA	27305000
MANGOSON	24820200
NARANJITA	17553600

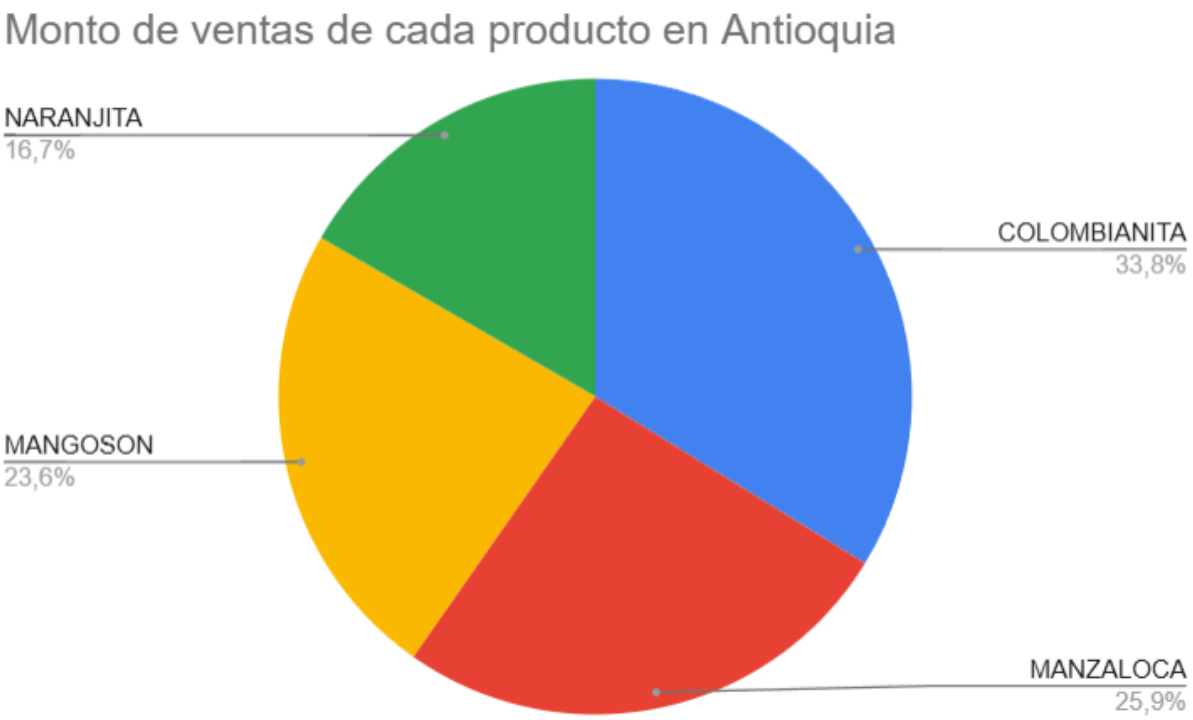
8.6.2.- Gráfico.

I.U. PASCUAL BRAVO

ET-0155 – Fundamentos de Big Data – Grupo 0100

Periodo 2024-2

Profesor: Jaime E Soto U



10. Cálculo de tiempos de procesamiento según cantidad de registros.

Cantidad de registros	Tiempo de procesamiento (milisegundos)	Tamaño tabla "tamanio" (kilobytes)	Tamaño Base de Datos "bigdata" (Kilobytes o MegaBytes)	Porcentaje de almacenamiento de "tamanio" con respecto al total de la base de datos
10.000	23552.04	696 kB	9707 kB	7.17%
100.000	256058.13	6704 kB	15 MB	44,69%
1.000.000	2923171.73	65 MB	74 MB	87,84%
10.000.000	35127751.94	650 MB	659 MB	98.63%

11. Link del video explicativo de todo el proceso

https://www.youtube.com/watch?v=eXl6eMXcSYo

12. Análisis de los resultados

Durante el análisis de los registros de ventas se evidenció que existen algunas inconsistencias en la calidad de los datos. Se encontraron fechas en distintos

formatos que dificultan la comparación y el ordenamiento de las operaciones. También aparecen registros con cantidades iguales a cero, lo cual sugiere errores de captura o información incompleta, y algunos registros con cantidades negativas que pueden corresponder a devoluciones mal clasificadas o errores de digitación. Además, se identificaron registros en los que el código del producto no estaba asignado, lo que genera problemas de trazabilidad y análisis.

A pesar de estas inconsistencias, los datos muestran patrones relevantes. Los ocho departamentos con mayores ventas concentran aproximadamente el 58% del total, lo que refleja un comportamiento de tipo Pareto donde una pequeña parte de los territorios explica la mayor parte de los ingresos. Este hallazgo señala la importancia de concentrar esfuerzos en los departamentos con mayor contribución, ya que allí se define buena parte del desempeño de la empresa. Entre ellos, Antioquia ocupa una posición destacada, con la proporción más alta de ventas. Esto puede explicarse porque allí se encuentra la sede central, lo que facilita la logística, el control de inventarios y las campañas comerciales.

Por otra parte, se observaron departamentos grandes en población y relevancia económica que reportaron un nivel de ventas menor al esperado. Esto puede responder a una menor presencia de sucursales de la franquicia, dificultades logísticas en la distribución, una competencia local más fuerte o simplemente a problemas de registro de la información. Estos resultados sugieren la necesidad de revisar la cobertura comercial y las estrategias de mercado en dichas zonas para no perder oportunidades de crecimiento.

En cuanto al análisis por municipios, algunos mostraron niveles de ventas muy bajos en comparación con su tamaño poblacional. En estos casos es posible que existan pocas o ninguna sucursal activa, limitaciones en la capacidad de distribución o una baja preferencia de los consumidores por los productos de la marca. También es probable que algunos de estos valores tan bajos estén asociados a registros incompletos o a la falta de reporte por parte de ciertas tiendas.

Al revisar los productos, se encontró que las ventas están lideradas por la gaseosa COLOMBIANITA, seguida por MANZALOCA y MANGOSON, mientras que NARANJITA ocupa la última posición en participación. Esto refleja las preferencias de los consumidores y permite orientar la planeación de la producción y las campañas de mercadeo. El hecho de que COLOMBIANITA concentre casi un tercio de las ventas confirma su papel como producto

estrella y plantea la necesidad de asegurar su disponibilidad permanente en todos los puntos de venta.

En algunos casos puntuales, como el municipio de Támesis, se encontraron registros de ventas sin producto asignado. Estos registros representan una cantidad significativa y requieren imputación para evitar la pérdida de información. Dado que en Támesis se comercializa principalmente NARANJITA, se propone asignar estas ventas a dicho producto para mantener la coherencia de los datos.

En síntesis, el análisis evidencia que las ventas de la franquicia presentan una alta concentración en unos pocos departamentos, lo que plantea oportunidades de crecimiento en otras regiones aún no explotadas. También revela la importancia de mejorar la calidad de los datos mediante procesos de limpieza, imputación y validación en la captura, con el fin de obtener una base sólida que permita construir tableros de control confiables y útiles para la toma de decisiones.

13. Conclusiones. Debe incluir la importancia del trabajo en el desarrollo académico y profesional.

Maria Rodriguez:
Este trabajo me ayudó a comprender más acerca del proceso ETL, este es un proceso fundamental que se puede aplicar en el campo laboral, permite mi crecimiento en este mundo tan grande que es el software, las bases de datos y los análisis de datos.

Con este trabajo pude ver el potencial de las bases de datos en SQL y al permitirme trabajar con una amplia cantidad de herramientas reforcé mi conocimiento técnico con respecto a estos.

Harlan Enciso:
El trabajo realizado impacta mi desarrollo académico permitiendo desarrollar mis habilidades para mi desarrollo profesional ya que se realiza un tema que en verdad si es utilizado en un ambiente laboral, donde los datos están alcanzando un gran valor para cada tipo de empresa lo que me permite poder desempeñarme en cargos con alto valor y demanda en el mercado laboral.

Por la parte técnica el usar herramientas como Sql o Python refuerza mis conocimientos técnicos en esta tecnologías permitiendo seguir creciendo

profesionalmente, además de mostrarme la utilidad de estas herramientas para el procesamiento de datos

Miguel Rojas Pabon
En el desarrollo de las actividades entendí lo importante que es la fase de transformación y limpieza de datos en un proceso ETL. Al trabajar con la tabla operaciones me encontré con muchos problemas en los formatos de fecha y con datos incompletos que al principio parecían difíciles de solucionar. Sin embargo, con práctica y pruebas fui aplicando diferentes reglas en SQL hasta lograr que todas las fechas quedarán en el formato correcto AAAA-MM-DD. Incluso aprendí a rescatar casos que parecían imposibles, como cuando los años venían con dos o tres dígitos o cuando el mes aparecía mayor a 12.

En la limpieza de datos también comprendí que no siempre la solución es borrar registros, sino que muchas veces se pueden corregir aplicando reglas lógicas o usando promedios para imputar valores faltantes. Eso me ayudó a valorar la importancia de mantener la mayor cantidad de información posible para que los análisis sean más confiables.

Este trabajo me sirvió para mejorar mis conocimientos en SQL, para entender cómo funcionan las expresiones regulares en las consultas y para darme cuenta de que el orden en el que se aplican las transformaciones es clave para no cometer errores. Al final, siento que logré dejar la base de datos en mejores condiciones, lista para análisis y consultas posteriores.

14. Bono (opcional). Diagrama de flujo del algoritmo ETL

Diagrama de flujo:
https://miro.com/app/board/uXjVJKS2J0M=/?share_link_id=108253361429

15. Github
https://github.com/harlanenciso112/BigData