Tarea 1

Base Datos a gran escala

Diego Moyano 202004509-7
Luis Zegarra 202073628-6
Nicolas Cancino 202004680-8

Objetivo

Transformar un conjunto de datos en formato CSV a formatos Avro y Parquet, y luego analizar el tama˜no de los archivos resultantes.

Tabla

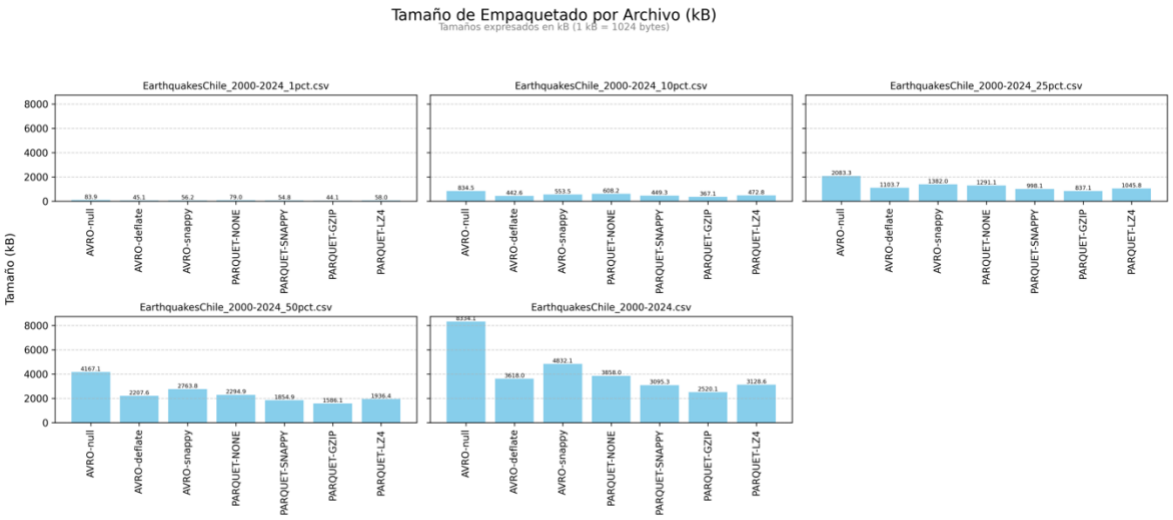
Tabla creada con los valores de archivo `empaquetado_size.json`, el cual contiene los registros de los tamaños de los archivos.

	Avro			**Parquet**			
	sin comprimir	**deflate**	**snappy**	**sin comprimir**	**snappy**	**gzip**	**lz4**
1%	83.88	45.07	56.17	79.00	54.81	44.08	58.00
10%	834.53	442.57	553.48	608.25	449.27	367.08	472.81
25%	2083.26	1103.66	1381.99	1291.08	998.14	837.09	1045.83
50%	4167.06	2207.64	2763.80	2294.88	1854.86	1586.12	1936.37
100%	8334.09	3618.00	4832.11	3858.03	3095.30	2520.06	3128.59

Tabla 1: Tamaño archivo por tipo compresión (kB) vs cantidad de datos archivo origen

Gráfica

Este gráfico es el obtenido con los valores almacenados en `empaquetado_size.json`, se encuentra en la ruta workspace/images/Figura1.png



Respuestas

Responda las siguientes preguntas:

a) ¿Qué conclusiones puede obtener de los resultados anteriores? (15 puntos)

Los resultados indican que el formato Parquet es más eficiente que Avro en términos de almacenamiento. En el caso de Avro, el método más eficiente es 'deflate', mientras que para Parquet, el método más eficiente es 'GZIP', siendo estos los que generan los archivos de menor tamaño en la mayoría de los casos, con el método Parquet dominando sobre Avro. De forma predeterminada (ambos sin métodos), se observa que Parquet logra una mayor compresión que Avro. En el último caso, donde se manejan mayores volúmenes de datos, Avro sin compresión es el que genera el archivo de mayor tamaño, mientras que Parquet sin compresión supera a Avro con 'snappy' y se aproxima al tamaño de Avro con 'deflate'.

b) Basado en los resultados: ¿Qué combinación (formato/compresión) elegiría para almacenar el dataset en un data lake en la nube? Justifique su respuesta.(15 puntos)

Dado que estamos operando en un entorno de nube, el objetivo principal es optimizar el uso de recursos. En este sentido, métodos de compresión como 'Deflate' para Avro y 'Gzip' para Parquet son altamente recomendables, ya que permiten una mayor compresión, reduciendo el tamaño de los archivos y, por lo tanto, los costos de almacenamiento. Además, la elección del formato de almacenamiento (filas o columnas) también influye en la elección del método de compresión. En conclusión, como se tiene pregunta por un Data Lake, este contiene muchos archivos, por lo que lograr mayor compresión sería una ventaja para optimizar espacio y recursos.

c) ¿Cuál fue el principal desafío para desarrollar la presente tarea? (10 puntos)

El principal desafío al desarrollar esta tarea fue definir los esquemas adecuados para el empaquetado de los datos, ya que ninguno de los integrantes del grupo tenía experiencia previa trabajando con los formatos Parquet ni Avro. Esto implicó una etapa inicial de aprendizaje para comprender su estructura. Establecer correctamente estos esquemas fue fundamental para garantizar una correcta serialización de los datos y obtener resultados confiables.

Prompt utilizado

1. "dame un ejemplo de cómo se utiliza la parquet y avro en Python"
2. "Entre Parquet y Avro para compresión de archivos, cuál debería ser más eficiente teoricamente?"
3. "Corrígeme este texto en ortografía y formalidad: {texto}"