Tarea 1

Bases de Datos a gran escala

mail: claudio.torresf@usm.cl

1. Objetivo

El objetivo de esta tarea es aplicar tus conocimientos de bases de datos y programación en Python para transformar un conjunto de datos en formato CSV a formatos Avro y Parquet, y luego analizar el tamaño de los archivos resultantes.

2. Descripción

Debe convertir el dataset `Earthquakes Chile 2000-2024` [1] en formato CSV a Avro y Parquet. Se entregan 5 archivos CSV con el 100, 50, 25, 10 y 1 % de los datos. Su código debe guardar los datos con y sin compresión.

El dataset posee el siguiente esquema:

- UTC_Date: Timestamp de 64 bytes
- Profundity: String
- Magnitude: String
- Date: Date de 32 bytes
- Hour: Time de 32 bytes
- Location: String
- Latitude: Decimal(precision=5, scale=3)
- Longitude: Decimal(precision=6, scale=3)

Junto con este enunciado se le entrega un workspace de trabajo para Visual Studio code basado en devcontainers. El espacio de trabajo incluye una aplicación base en python junto con las librerías recomendadas (fastavro, pyarrow, pandas) para leer el archivo CSV y escribir archivos Avro y Parquet.

3. Actividades

Debe realizar las siguientes actividades (100 puntos):

1. Agregar el codigo Python necesario para leer los archivos CSV y convertirlo a Avro y Parquet utilizando los algoritmos de compresión (50 puntos):

[1] Extraído de `https://www.kaggle.com/datasets/javierquinterosm/earthquakes-in-chile-2000-2024/data`

- Avro: sin compresión, deflate y snappy
- Parquet: sin compresión, snappy, gzip y lz4

Los archivos resultantes deben mantener el esquema de los datos.

2. Completar la siguiente tabla y graficar los resultados: (10 puntos)

	Avro			Parquet			
	sin comprimir	deflate	snappy	sin comprimir	snappy	gzip	lz4
1 %							
10 %							
20 %							
50 %							
100 %							

3. Responda las siguientes preguntas:

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

   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)

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

4. Consideraciones

- La tarea puede ser desarrollada en grupo de máximo 3 estudiantes.
- Puede modificar el proyecto base como estime pertinente. Las librerías incluidas son solo una sugerencia.
- La conversión de los archivos se debe realizar usando el comando: **python app.py** Cualquier paso extra necesario para ejecutar su aplicación debe ser detallado en su informe y en el archivo Readme.md de su proyecto.
- Debe entregar el código fuente.
- Se descontarán 4 puntos por cada campo de los datos que no cumplán el esquema solicitado. Solo está permitido cambiar la cantidad de bytes de los campos.
- Está permitido utilizar Inteligencias Artificiales para el desarrollo de la tarea, pero debe agregar el prompt utilizado en el informe.

Fecha entrega: 27 Abril 2025