Letra Obligatorio Certificado de Big Data.

Para el obligatorio deberán utilizar las herramientas utilizadas en el curso. Deberá seleccionar un conjunto de datos tabulares con más de 4 tablas que el docente propondrá y deberá seleccionar 8 preguntas relativas a los datos para contestarlas.

Los pasos a seguir son:

- Tomar los datos que fueron seleccionados junto al docente.
- Luego se deberá realizar un análisis exploratorio de los datos vía pandas, identificando el tipo de datos que hay en cada columna y que significado tienen dentro del dominio de los datos. Dentro de un Jupyter notebook se mostrará, una vista previa de las primeras filas, cantidad de columnas de cada tabla, nombre de cada columna, descripción de los datos de cada tabla, cómo está compuesto el esquema de los datos, revisar valores nulos o faltantes y limpiarlos si es necesario. Revisar registros duplicados. Claves primarias únicas.
- Los archivos resultantes se deberán almacenar en otra carpeta.
- A partir de estos nuevos archivos, se deben crear visualizaciones dentro de otro notebook con las herramientas dadas en clase u otras de elección del equipo, que ayuden a responder las preguntas seleccionadas.

Parte 2

El mismo análisis realizado en la parte 1 realizarlo vía spark, ya sea dentro de la máquina virtual si se tienen créditos si no dentro de Google Collab.

Parte 3

Se píde desarrollar un dashboard que responda algunas de las preguntas planteadas, implementado en Tableau Public o superset.

Parte 4

Una vez que termine con la exploración y limpieza de datos, deberá elegir una forma de modelarlos, esta puede ser, Normalizada, Diagrama Estrella, Data Vault, o OBT. Describir en

Hive, como lo modelaría, que tablas crearía y de que tipo (externas, internas).

La entrega final consiste en un informe donde se detalle todo el proceso realizado y todo lo aprendido durante la realización del obligatorio. Se deben entregar los notebooks de análisis via pandas, de preguntas y respuestas con las visualizaciones correspondientes, así como el notebook de análisis utilizando spark. Para el caso del dashboard se pide entregar el link a Tableau Public o Superset, o alguna captura que demuestre su funcionamiento.