**Trabajo final de Big Data**

Pasos

1. Leer y entender trabajo.final.analitca.grandes.datos.pdf

**Empresa**: UNALWater

**Actividad económica**: venta de botellas de agua

**Sede**: Medellín

**Objetivo**: potenciar sus capacidades analíticas, mediante un área especializada de Big Data

**Problema**: Grandes volúmenes de datos

**Requerimiento**: Simular datos de transacciones diarias

**Métricas** que se pueden generar y visualizar en Power BI

* + **Comunas con mayores ventas de botellas de agua**
  + **Ventas por vendedor**: Analiza el rendimiento de cada vendedor, evaluando el número de ventas realizadas, el valor total de las ventas y la comisión generada.
  + **Comisión por vendedor**. Esto puede ayudar a identificar a los vendedores más eficientes y a proporcionarles incentivos adecuados
  + **Ventas por cliente**: Examina el comportamiento de compra de los clientes, incluyendo la frecuencia de compra, el valor total de las compras y la lealtad del cliente. Identificar a los clientes más rentables y entender sus preferencias puede ayudar a adaptar las estrategias de marketing y retención de clientes, como promociones y creación de nuevos productos
  + **Análisis temporal de ventas**: Observa las tendencias temporales en las ventas, como los picos de ventas diarios, semanales o mensuales, para identificar patrones de comportamiento de compra y ajustar las estrategias de marketing en consecuencia. También se puede analizar la relación entre la hora de la venta y el valor de la transacción
  + **Análisis de retención de clientes:** Examina la frecuencia con la que los clientes regresan para realizar compras adicionales y la duración promedio de tiempo entre las compras sucesivas. Esto puede proporcionar información sobre la lealtad del cliente y la efectividad de las estrategias de retención

1. Análisis de datos

Nos entregaron 4 fuentes de datos en formato parquet

* Un archivo con información de contacto de 500 clientes (customers): nombre, teléfono, correo electrónico, identificación y dirección
* Un archivo con información de 20 vendedores (employees): nombre, teléfono, correo electrónico, identificación, dirección y comision
* Un archivo con las 16 comunas de Medellín y los 5 corregimientos cada uno con su ubicación geoespacial
* Un archivo con la ubicación geoespacial de la cuidad de Medellín en el departamento de Antioquia







Para leer los archivos customers.parquet y employees.parquet.parquet, se utilizó pandas, donde examinamos los datos para entender su estructura y contenido.

Se utilizó geopandas para leer los archivos medellin\_neighborhoods.parquet y 50001.parquet, donde verificamos la información geoespacial y asegurándonos que esté alineada con las comunas de Medellín.

Esto nos dio una idea inicial de la estructura y el contenido de los datos, así como una visualización de la información geoespacial de las comunas y corregimientos de Medellín

1. Procesamiento de datos

* Crear una función para simular eventos de transacciones de botellas de agua cada 30 segundos.

Generar datos aleatorios para cada campo dentro de rangos específicos y con distribuciones que reflejen situaciones realistas.

Para la latitud y longitud, generar valores dentro del rango correspondiente a Medellín.

Para los IDs de empleados y clientes, muestrear aleatoriamente de los IDs disponibles en los conjuntos de datos de empleados y clientes que han sido cargados previamente.

La cantidad de productos y el ID de orden también pueden ser generados aleatoriamente.

La fecha puede ser generada aleatoriamente dentro de un rango de fechas específico

* Utilizar pandas para generar datos simulados basados en reglas definidas para demostrar valor en la PoC.

Utilizar distribuciones específicas para ciertos campos, por ejemplo, una distribución normal para la cantidad de productos.

Aplicar reglas específicas para garantizar que los datos generados sean coherentes y realistas, ¿que reglas pudiéramos utilizar?

1. Almacenamiento de datos

Decidir sobre el formato de almacenamiento adecuado para cada zona (bronze, silver, gold)

Estos serían los campos que propongo para cada zona de almacenamiento, en negrita están los que debemos adicionar y en rojo me parece redundante comuna y barrio porque en el primero es el código y en el segundo es el nombre. Me parece que en la capa gold se requiere ver los nombres de los vendedores y compradores para poder hacer la visualización en Power Bi

|  |  |  |
| --- | --- | --- |
| **Bronze** | **Silver** | **Gold** |
| partition\_date | partition\_date | partition\_date |
| order\_id | order\_id | order\_id |
| commune | **commune** | **commune** |
| customer\_id | customer\_id | customer\_id |
| employee\_id | employee\_id | employee\_id |
| event\_date | event\_date | event\_date |
| event\_day | event\_day | event\_day |
| event\_hour | event\_hour | event\_hour |
| event\_minute | event\_minute | event\_minute |
| event\_month | event\_month | event\_month |
| event\_second | event\_second | event\_second |
| event\_ year | event\_ year | event\_ year |
| latitude | latitude | latitude |
| longitude | longitude | longitude |
| neighborhood | **neighborhood** | **neighborhood** |
| quantity\_products | quantity\_products | quantity\_products |
|  | **price** | **price** |
|  | **comission** | **comission** |
|  |  | **name\_customer** |
|  |  | **name\_employee** |

1. Diseño de arquitectura

Crear un diagrama de arquitectura que muestre cómo se diseñó la solución, incluyendo los procesos orquestados para la simulación de eventos y el procesamiento de los datos en diferentes zonas de almacenamiento

1. Pruebas y ajustes