PUNTO 3: PREGUNTAS ABIERTAS DE BUSINESS INTELLIGENCE

1. Modelado de Datos y Diseño de Soluciones

Una empresa maneja múltiples fuentes de datos desconectadas (CRM, ERP, Google Analytics, bases de datos internas, etc.).

* ¿Cómo diseñarías un modelo de datos que integre esta información para generar reportes confiables?
* Podemos utilizar procesos de integración de datos como (ETL), que nos permita trabajar con fuentes heterogéneas como (**SQL, NoSQL, APIs, archivos planos)** y **(**ELT) que luego nos permita cargar la información y trabajarla en tiempo real usando (**Fivetran, Apache NiFi, dbt o Airflow)**, a su vez necesitamos almacenamiento centralizado (DataWareHouse) que permite almacenar datos estructurados y previamente procesados con el fin de automatizar procesos y generar reportes interactivos para toma de decisiones.
* ¿Qué desafíos podrían surgir y cómo los abordarías?
* Podemos tener problemas como:
* **Datos inconsistentes entre fuentes**: se pueden presentar diferentes formatos e irregularidades entre los datos, por ende, podríamos utilizar un ETL/ELT robusto que estandarice formatos como (Fechas, monedas, unidades).
* **Integración de Fuentes con Diferentes Tecnologías:** podemos presentar problemas e incompatibilidades respecto al tipo de tecnología que utilice la fuente de información, para ello podemos implementar APIs personalizadas o middleware para normalizar datos antes de cargarlos en el Data Warehouse, con esto estamos seguros de tener una vista unificada de los datos presentados
* **Volumen y Escalabilidad:** Almacenamiento y procesamiento de grandes volúmenes de datos puede afectar el rendimiento. podríamos implementar almacenamiento en capas (Data Lake para datos en bruto o Data Ware optimizado para reportes.
* **Datos en Tiempo Real vs. Batch:** Algunas fuentes requieren análisis en tiempo real, mientras que otras funcionan con cargas batch, consultas en tiempo real, monitoreo en vivo 24/7, baja latencia de datos.

1. Procesamiento de Datos: Tiempo Real vs Batch

Un startup de e-commerce recopila datos de navegación, compras y soporte al cliente.

* ¿En qué escenarios utilizarías procesamiento en tiempo real y en cuáles batch?

**Procesamiento en Tiempo Real:**

* Analizar transacciones en tiempo real para identificar patrones sospechosos y prevenir fraudes.
* Reflejar las ventas inmediatamente y evitar la venta de productos fuera de stock.
* Identificar y resolver problemas rápidamente.
* Comportamientos de los datos en vivo

**Procesamiento por Batch:**

* Procesamiento de grandes volúmenes de datos como compras masivas
* Crear informes y paneles de control periódicos para el análisis del rendimiento y la toma de decisiones estratégicas.
* ¿Cómo equilibrarías la necesidad de datos frescos con la eficiencia del sistema?
* **Identificar casos de uso críticos:** No todos los datos requieren la misma capacidad de carga. Es importante identificar qué datos necesitan ser actualizados en tiempo real y cuáles pueden subirse por **Batch** .

1. Optimización de Consultas y Rendimiento

Manejas una base de datos con cientos de millones de registros y las consultas para reportes tardan demasiado.

* ¿Cómo diagnosticarias el problema de rendimiento?
* Identificar posibles consultas complejas, consultas innecesarias o uso inadecuado de funciones.
* Diagnóstico de Indexación y Particionamiento, utilizamos índices para realizar consultas en tablas grandes.
* podemos utilizar particiones para acelerar el proceso ejemplo (rangos de fechas).
* Medición del Uso de Recursos (CPU, RAM, IOPS) para evitar desbordamiento del sistema.
* Estrategias de Almacenamiento y Cacheo ejemplo (Si una consulta se ejecuta con frecuencia, almacenar los resultados evita computación).
* ¿Qué estrategias aplicarías para optimizar las consultas sin comprometer la integridad de los datos?
* Evaluar cuidadosamente la necesidad de cada índice y eliminar los redundantes o poco utilizados.
* Utilizar índices compuestos para consultas que involucran múltiples columnas en filtros o uniones.
* Clustering y Ordenación de Datos Optimizando el acceso secuencial a los datos.
* Evitar subconsultas innecesarias y usar joins eficientes.
* Selección Eficiente de Datos, Especificar sólo las columnas necesarias reduce el volumen de datos transferidos.
* Evitar traer datos innecesarios optimiza la memoria y CPU
* Uso de Particionamiento para Tablas Grandes (Particionar por fecha, región o categoría)
* Uso de Caching para Evitar Repetir Consultas Costosas, Reducimos carga en la base de datos y mejora el tiempo de respuesta.

1. Calidad y Gobierno de Datos

En una empresa de retail, los reportes de ventas presentan inconsistencias debido a fuentes con distintos formatos y criterios.

* ¿Qué medidas implementarías para garantizar la calidad de los datos?
* Filtros y prohibiciones en la carga de datos para evitar errores humanos.
* Aplicar reglas de validación en la carga (datos negativos).
* Clasificación de datos duplicados o nulos.
* Estandarizar formatos para evitar inconsistencias.
* Recopilar comentarios y realizar ajustes según sea necesario.
* Completitud de los datos (validar espacios vacíos).
* Unicidad (registros duplicados o reiterativos)
* ¿Cómo estructurarías la gobernanza de datos en la organización?
* La Gobernanza de Datos es clave para asegurar que los datos sean confiables, seguros y accesibles en toda la organización. Un buen marco de gobernanza evita errores, duplicidades, incumplimientos normativos y silos de información (sistemas de almacenamiento de datos que no pueden comunicarse entre sí). por consiguiente, se definen roles al interior de, y se garantiza la calidad, acceso, seguridad y confiabilidad de los datos. Así mismo protegerlos basado en cumplimiento normativo (Reglamento General de Protección de Datos (GDPR), ISO/IEC 27001)

1. Casos de Uso en BI

Describe un proyecto real o hipotético de BI o Data Engineering en el que hayas trabajado.

* ¿Qué problema intentan resolver?

Identificar si las condiciones sociodemográficas de un estudiante infieren en su puntaje final en las pruebas de estado, teniendo en cuenta variables como (estrato, ubicación de vivienda, nivel educativo de padres, acceso a internet, dispositivos de uso, etc).

* ¿Qué tecnologías y metodologías utilizaste?

-**Tecnologías**: Como principal herramienta se usó visual studio code con un entorno de python para realizar el proyecto junto con una extensión llamada Jupyter Notebook para mejor distribución del código, aparte de ciertas librerías como pandas, numpy, sklearn y matplotlib, las cuales permiten realizar funciones específicas en función de recolección, modificación, análisis predictivo y gráficos de información.

**Metodologías:**

**Definición del problema:** El objetivo principal fue analizar si las condiciones sociodemográficas de los estudiantes afectan al puntaje final de la prueba de estado.

**Recolección y preparación de los datos:** Se recolectaron y limpiaron los datos. En esta etapa se manejaron los valores faltantes, se transformaron las variables y se estructuraron los datos de manera adecuada para el análisis.

**Análisis exploratorio de datos:** Se realizó un análisis para examinar la distribución de las variables, las correlaciones entre ellas y se generaron visualizaciones que ayudaron a identificar patrones y tendencias en los datos.

**Modelado predictivo:** En este paso, se utilizó el enfoque de **machine learning** para construir modelos predictivos que pudieran predecir el puntaje de los estudiantes. Se emplearon algoritmos como regresión lineal. Se entrenaron varios modelos y se evaluó su desempeño utilizando métricas apropiadas, como el **R²** para regresión.

**Validación y evaluación del modelo:** Se dividió el conjunto de datos en dos partes: entrenamiento y prueba, para evitar el sobreajuste y garantizar que el modelo sea adecuado en su predicción

* ¿Cuál fue el mayor desafío y cómo lo superaste?
* La relación de las variables para determinar el puntaje, ya que el Dataset presentaba mucha información que se relaciona entre si con una finalidad (Puntaje Total), fue complejo determinar que variables si y que variables no, dentro de la Data existían datos atípicos que minorizaban el porcentaje de precisión del modelo, lo cual hacia que se extendiera mas el proceso de encontrar las variables precisas.
* **Solución**: Se realizó un gráfico de temperatura en donde se mostraban todas las columnas seleccionadas y su importancia en la variable a predecir, de esta manera se logro identificar cuales ayudaban a mejorar la precisión del modelo y continuamos entrenando el modelo con las nuevas variables hasta encontrar un mejor margen de precisión