A. De cada fuente de datos se tienen identificados que campos requiere el área operativa. ¿Para cumplir con los dos objetivos que subconjunto de cada fuente de datos extraerías?

uente F1 (CRM):

- Campos a extraer:
    - ID de cliente
    - Nombre
    - Apellidos
    - curp
    - rfc
    - Fecha de nacimiento
    - Correo electrónico
    - Teléfono
    - Dirección

Fuente F2 (SQL Server):

- Campos a extraer:
    - ID de cliente
    - ID de transacción
    - Fecha de transacción
    - Producto
    - Cantidad
    - Monto total

Fuente F3 (PostgreSQL):

- Campos a extraer:
    - ID de cliente
    - ID de transacción
    - Fecha de transacción
    - Producto
    - Cantidad
    - Monto total

B. ¿Qué posibles retos implica la extracción de cada una de las
fuentes de datos por separado y qué herramientas utilizas ?

Una de las principales problemas para la fuente F1 (CRM) es validar si existe alguna herramienta o api que nos permita la extracción de los datos y también influyen en F1,F2 y F3 los principales retos para todas las fuentes son si existen reglas de firewall , necesidades de conexion de vpn y formas especificas de conexiones

C. ¿Qué posibles retos implica la independencia en el modelo de
datos de las tres fuentes y cómo los resolverías?
Al tener diferentes fuentes nos podemos enfrentar a inconsistencias en los nombres de los campos , asi como diferencias en tipos de datos entras las distintas fuentes , estructuras distintas, ademas generar un proceso de etl por fuente para obtener un modelo unificado y que nos permita unir la informacion de las distintas fuentes

D. ¿Aparte de un proceso batch en la hora de menor uso, cómo
podrías mitigar el impacto de tu pipeline sobre las fuentes originales ?

Para mitigar el impacto tambien se podria optar por implementar mecanismos de CDC o hacer alguna replica de base de datos para no impactar a la base productiva

E. ¿Cuáles etapas considerarías en tu proceso de transformación de
- datos y qué uso les darías?
- ingesta de informacion
- limpieza de datos
- desnormalizacion de información
- data productiva

F. ¿Qué herramientas utilizas para las etapas de transformación?

Las opciones podemos utilizar pandas si los cantidad de datos no es demasiada o pyspark si se requiere un proceso distribuidos de información

G. ¿Qué storage usarías para cada propósito y por qué ?

utilización de una base de datos como snowflake para la gestión de la información o uns sistema SMTP para el almacenamiento de datos semi y no estructurados

H. Recuerda que al menos a diario tendrás que llevar data nueva a tu
etapa de transformación final, ¿Como orquestarias tu pipeline y con
qué herramienta?

generando workflow generado en apache airflow para orquestar cada una de las etapas

II. Seguridad (manteniendo tu rol de ingeniero de datos).
A. ¿Cómo mantendrías la seguridad de tu flujo de datos end-to-end? Es

decir disminuir riesgos de posibles fugas o intrusiones no deseadas al entorno de ejecución que estás construyendo.

el uso de un de SMTP para la ingesta de la información asi como un sistema de encriptación asincrono para mayor seguridad en el transporte de la información

III. Gobernanza de datos
A. ¿Cómo llevarías control de la metadata y sus cambios al igual que los procesos de tu pipeline y cómo almacenamos estos datos?

Documentar origen, transformaciones y destino de cada dato, versionado de esquemas ,registrar los cambios en un sistema de versionamiento , generación y centralización de mata data , generar campos de auditorías y registro de logs