I. Propón una arquitectura (solo proponla no la construyas) que considere los siguientes aspectos y preguntas:

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?
Todos los insumos de las 3 fuentes de origen para satisfacer los requerimientos del negocio.

B. ¿Qué posibles retos implica la extracción de cada una de las fuentes de datos por separado y qué herramientas utilizas ?
Al tener diferentes insumos se tiene que realizar diferentes scripts para extraer la información hacia el DataLake. Se puede utilizar Spark y los conectores de Postgres y SQL Server, para el CRM se tendría que revisar si se tiene que consumir un API para obtener la información.

C. ¿Qué posibles retos implica la independencia en el modelo de datos de las tres fuentes y cómo los resolverías?
Haría un diagrama de datos para revisar la compatibilidad de la información entre las 3 fuentes y saber cuantos campos se tendríamos como resultado de esa comparación.

D. ¿Aparte de un proceso batch en la hora de menor uso, cómo podrías mitigar el impacto de tu pipeline sobre las fuentes originales ?
Cargas incrementales por día (para no traerme la historia diariamente), Optimización de joins (ya que es la transformación que más recursos consume) y Optimización de la configuración del cluster.

E. ¿Cuáles etapas considerarías en tu proceso de transformación de datos y qué uso les darías?
3, Proceso de limpieza, proceso de cálculos (matemáticos, agregaciones, string, etc) y proceso de carga final.

F. ¿Qué herramientas utilizas para las etapas de transformación?
Spark debido a que es el motor más eficiente para procesos Big Data.

G. ¿Qué storage usarías para cada propósito y por qué ?
Al ser un DataLake su storage es un FileSystem y usaría 3 capas bronce (datos de origen sin modificación), plata (datos de origen que pasaron un proceso de limpieza y reglas de negocio) y oro (datos finales que van a ser explotados).

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?
Se puede hacer con Control -M o Airflow ejecutando los scripts de extracción de información y después los scripts de transformación de la información en un proceso diario ambos scripts.

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.
Habilitando roles para que los usuarios solo tengan acceso a los flujos que van a ejecutar y no tengan acceso a recursos que no les corresponden.

III. Gobernanza de datos

A. ¿Cómo llevarías control de la metadata y sus cambios al igual que los procesos de tu pipeline y cómo almacenarías estos datos?
Se tendría que solicitar los permisos a un supervisor que tenga la autoridad de dar acceso a las diferentes capas del Datalake y a los insumos de cada área operativa.