SUPUESTOS:

- Las tres fuentes de datos F1, F3 Y F3 fueron diseñadas de manera independiente tanto a nivel de software como de modelo de datos.
- Como las fuentes de datos son operativas diariamente hay datos nuevos.

Tu objetivo es diseñar una arquitectura que consolide la información contenida en las fuentes de datos de F1, F2 y F3 que cumpla con dos propósitos en primer lugar que habilite a un grupo de usuarios del área operativa a la extracción de consultas por medio de SQL y de manera secundaria que permita al equipo de ciencia de datos la aplicación de algoritmos de detección de patrones como clustering o búsquedas en grafos.

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

De la fuente F1 se tendría que proporcionar la información de los clientes, datos principales como nombre o usuario, de la F2 los campos relacionados con los productos que ofrece la compañía y de F3 la información faltante sobre los productos.

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

El reto más grande es unir la información de las 3 fuentes, crear lógicas para la nueva creación de campos porque posiblemente con los campos que trae cada una no sean llaves de la otra fuente, entonces se tendrían que crear para poder hacer los cruces, además de homologar tipos de datos, etc.

Con respecto al CRM- F1 ya que es propietario y es un servidor físico de la empresa, hay que realizar el mantenimiento de éste, actualización constante de la base de datos, costo al unificar la BD.

C. ¿Qué posibles retos implica la independencia en el modelo de datos de las tres fuentes y cómo lo resolverías?

El que estén separados no se podría realizar un análisis completo, se tendría que hacer cada análisis por separado, además de que no se tendría la información necesario posiblemente para realizar el modelo de datos, es necesario unificar la información.

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

E. ¿Cuáles etapas considerarías en tu proceso de transformación de datos y qué uso les darías?

Etapas:

1. Creación de documentos funcionales para saber las lógicas, realizar un análisis exploratorio de los datos para conocer los tipos de datos, la información que viene con nulos, la información sobre outliers, tener conocimiento de la información que tiene cada campo, etc.
2. Aplicación de lógicas en caso de crear nuevos campos si es necesario.
3. Realización del código
4. Hacer pruebas de funciones y pruebas generales.
5. Realizar pruebas en un ambiente de pruebas similar al ambiente productivo para ver cómo es la salida de información.
6. Realizar pruebas de calidad.

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

Ambiente de pruebas vbox, herramients de GCP, AWS, IDE, utilización de spark-scala, pyspark, python

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

Dropbox – Google Drive para toda la documentación, diagramas, etc
Amazon Web Services o GCP para almacenamiento en la nube

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

Debido a que la información se actualiza diariamente se tendría que realizar hacer una ejecución diaria para tener los registros nuevos y actualizar si es necesario registros antiguos.

1. Ejecuciones de cada fuente
2. Ejecución de la fuente nueva donde estará almacenada la información unificada.
3. Ejecución de las fuentes que requiere cada área

Airflow, Control-M

I. Proporciona un diagrama de tu propuesta de arquitectura.

II. A. ¿Cómo mantendrías la seguridad de tu flujo de datos end-to-end? Es decir, disminuir riesgos de tu flujo o intrusiones no deseadas al entorno de ejecución que estás construyendo.

Al ingestar los datos se tokenizarían – cifrarían, no tener acceso en ambientes de prueba, sólo se tendrían acceso a nivel productivo teniendo el mayor cuidado posible y teniendo en cuanta sólo tendrían acceso ciertas personas, no todo el quipo para mantener el cuidado con el acceso o la fuga de datos.

III. A. ¿Cómo llevarías el control de la metada y sus cambios al igual que los procesos de tu pipeline y cómo almacenarías estos datos?

Con un control de versiones, para poder controlar los cambios y si son varias personas las que intervienen podrían trabajar en conjunto, git, bitbucket, etc