I. Propón una arquitectura que considere los siguientes aspectos:
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?

R: Con ayuda de Dataflow, se extraerían sólo los campos que el área operativa requiera mismos que se almacenarían en Cloud Storage al terminar el pipeline

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

R: Como las tres fuentes de datos utilizan diferentes tecnologías, esto implica crear 3 instancias para cada una de las bases de datos, lo cual puede incrementar los costos mensuales para la operación de esta arquitectura y también incrementaría en cierta medida la dificultad para la configuración y administración de las mismas.

Se escogió Cloud SQL ya que es un servicio totalmente administrado y que proporciona algunas ventajas por ejemplo, Cloud SQL admite bases de datos como MySQL y Postgrsql, se puede escalar vertical u horizontalmente cada instancia, también brinda seguridad ya que los datos se encriptan mientras están en reposo y en tránsito, también cuenta con alta disponibilidad entre otras.

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

R: Debido a que las 3 fuentes fueron diseñadas de manera diferente, conlleva un incremento de esfuerzo para diseñar un pipeline así como de su almacenamiento.

Para resolverlo, esta arquitectura sugiere crear un pipeline diferente para cada fuente y tambíen se recomienda guardar el resultado de los procesos en 3 buckets diferentes para evitar conflictos con los esquemas de las mismas.

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

R: Se podrían tener ambientes diferentes uno de desarrollo, otro ambiente Beta productivo y por último un ambiente de producción. Se podrían probar los pipelines de desarrollo en el ambiente más bajo para no impactar a la información productiva.

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

R: Es la tapa intermedia entre la fuente de datos y el storage del Datalake. Principalmente se realizarían operaciones de limpieza de información así como la

aplicación de reglas de negocio y estandarización para que finalmente puedan ser consumidas en un Datawarehouse.

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

R: Para esta arquitectura se escogió utilizar Dataflow el cual funciona sobre Apache Beam y los pipelines se pueden diseñar tanto con Python como con Java y permite el procesamiento de grandes volúmenes de información de manera eficiente.

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

R: Para la fuente, se utiliza Cloud SQL ya que es totalmente administrado y permite diferentes tecnologías como MySQL, PostgrSQL entre otras.

Para el Datalake se utiliza Cloud Storage ya que cuenta con escalabilidad ilimitada, también proporciona durabilidad y confiabilidad, acceso rápido y eficiente y se pueden guardar los datos en una estructura jerárquica con directorios lo que facilita la organización de la información.

Para el DatawareHouse se escogió BigQuery porque proporciona una plataforma eficiente, escalable y fácil de usar para el almacenamiento y análisis de grandes conjuntos de datos. Además cuenta con controles de acceso granulares lo cual permite que a cierto grupo de usuarios se les otorgue únicamente los permisos que necesitan lo cual es un requerimiento solicitado para esta arquitectura de datos.

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?

R: La opción más viable es a traves de Cloud Composer (Apache AirFlow) creando un Dag para cada pipeline. La ventaja de crear un pipeline para cada fuente es que se pueden ejecutar en paralelo, sin esperar a que termine uno después del otro. Se configuraría su ejecución diaria buscando una hora óptima para no saturar los recursos y no afectar la operación.

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.

En GCP se utiliza la encriptación para proteger datos en reposo y en tránsito. Se pueden gestionar claves de encriptación proporcionadas por google o por el mismo usuario. También se puede controlar el acceso a los datos a través de Cloud IAM en

donde se pueden configurar políticas para restringir acceso a recursos sensibles.
Así mismo también existe otro servicio de nombre Data Loss Prevention que identifica y protege los datos sensibles como información PII.

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?

Para el control de metadatos existe Cloud Catalog que proporciona una interfaz centralizada para acceder a metadatos de datos de diferentes fuentes, incluidos servicios de Google Cloud, fuentes de datos locales y fuentes de datos personalizadas.

Igualmente cuando está corriendo un job de Dataflow se puede monitorear desde GCP en el apartado de DataFlow. Todos los logs se van almacenando en Cloud Logging que es un servicio que permite recopilar, buscar, analizar y monitorear registros de recursos en Google Cloud Platform.