Ejercicio 2.

Se proponen dos capas de almacenamiento para el procesamiento de datos

1. Capa de Almacén de datos crudos

Se plantea realizar réplicas de las bases sql haciendo uso del CDC para extraer los eventos de cambio de la base de datos, transformarlos y cargarlos de forma tabular en una capa que almacene la información cruda y de esta manera lograr tener en tiempo real todos los cambios realizados por los sistemas transaccionales.

El objetivo de crear estas réplicas es para no interferir o degradar el performance de los servidores transaccionales.

F2 PostgresSQL - Se haría uso del servicio DataStream para extraer el CDC y depositarlo en un bucket de storage.Un pipeline en Apache beam usado Dataflow como runner, se encargaría de leer, transformar y cargar los eventos de cambio de la base de datos en la capa de crudos. La ejecución del proceso sería desencadenada cada que se detecte un cambio en el bucket que almacena el CDC
En caso de usar infraestructura onpremise se haría uso de spark como runner para el pipeline de Apache beam

F3 MS sql server - Se puede ir al leer el CDC con un pipeline en Apache beam usado Dataflow como runner, se encargaría de leer, transformar y cargar los eventos de cambio de la base de datos en la capa de crudos. La ejecución sería programada por medio de un scheduler dentro de GCP.
En caso de usar infraestructura onpremise se haría uso de spark como runner para el pipeline de Apache beam y sería orquestado en Air Flow

CRM - Serán almacenados en esta misma capa por medio de proceso batch. El proceso batch sería un pipeline en Apache beam usado Dataflow como runner y sería orquestado en Air Flow desplegado en el servicio Cloud Composer de GCP.

En caso de usar infraestructura onpremise se haría uso de spark como runner para el pipeline de Apache beam y sería orquestado en Air Flow

2. Capa de consumo

En esta capa se almacenarán las tablas finales listas para su consumo.

El procesamiento y carga de los datos se realizará por medio de proceso batch sql haciendo uso de hadoop y spark y serían orquestados por Air Flow desplegado en el servicio Cloud Composer de GCP.

La Capa de consumo cumpliría dos propósitos.

1. Se habilitará el acceso a los objetos procesados para los usuarios operativos y serían capaces de realizar consultas sql a través de bigquery.

2. El equipo de ciencia de datos podría realizar peticiones de datos por medio de la API de Bigquery y usarlos dentro de sus modelos de ML.

Se opta por Bigquery como medio de almacenamiento por el rendimiento y costos, además que los usuarios operativos pueden hacer consultas sql desde la consola de bigquery y el equipo de científicos puede acceder por medio de una API.