Justificación

Se establece la infraestructura de acuerdo al documento .svg debido a que mantener dos fases en la información permite tener un registro historico, respaldo para auditorias.
Usar un orquestador, permite gestionar los procesos por horarios y condicionarlos de diferentes maneras (Permite integrar reglas de calidad que condiciones la continuación del flujo).
Se consolidan en un mismo punto las tres fuentes de información y permite la reproducibilidad del flujo para diferentes origenes o diferentes insumos dentro del mismo origen.

Preguntas

A. Decada 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?

Para los archivos avro mantendría el insumo con el 100% de sus campos.
Para la fase master, delimitaría unicamente los campos que serán usados para eficientar costos de almacenamiento y proceso.

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

Es necesario configurar una conexión de fuentes diferentes hacia el dataleake de fase raw.
Apache airflow, control M permiten gestionar (una vez establecida la conexión) el proceso de carga periodica de la información.

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

Al se cada configuración diferente, es necesario conocer lo escencial para el mantenimiento de las conexiones, el formato de llegada será diferente por lo que la normalización tendrá caracteristicas diferente para cada insumo dependiendo de su origen.

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

Generar vistar en los origenes SQL que permitan realizar la consulta sin apuntar a la información original, evitando accidentes y modificando la información.

Si la técnología lo permite, implementar condiciones que actualicen la información solo cuando haya una modificacion (insert, update, delete) como es el caso de snowflake.

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

Etapa raw y etapa master.

Etapa raw brindaría un respaldo de la información original dentro del ambiente en el que se esta trabajando, de forma ordenada y estructurada.

La etapa master, en un DWH en donde la información contenida esta limpia y lista para su consumo con información vital, sin campos innecesarios y transformaciones en caso de ser requeridas.

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

Spark, python, SQL, Pandas.

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

Etapa raw
S3 glaciar para la etapa raw ya que es un ambiente que no suele generar consultas constantes y esto generaría ahorros.

Etapa Master
Amazon Redshift debido a que se especializa en permitir consultar SQL
S3 para requerimientos de ML ya que es posible conectarlo con sagemaker y spark.

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?

Control M o Apache airflow.
Orientado a airflow ya que es open source, más conocido y tiene una interfaz más amigable

¿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.
Utilizando el principio de minimo acceso, no otorgando acceso general al personas, si no por medio de un IAM teniendo control sobre unicamente el personal que requiere el acceso. Tambien con una posible configuración de logs por accesos realizados y 2FA.

¿Cómo llevarías control de la metadata y sus cambios al igual que los procesos de tu pipeline y cómo almacenarías estos datos?

Usando un aplicativo que permite guardar metadata es decir, el schema con el detalle de columnas y tipo de datos, mismo que sería usado como validación en elproceso de carga de información para mantener consistencia y las particiones.