**3. Diseño Conceptual (20 minutos)  
Camilo Del Rio Florez**

**Escenario:**

"La compañía necesita un pipeline diario que consolide datos de 3 fuentes: (1) CSV en un SFTP, (2) API REST de un proveedor externo, (3) MongoDB. Los datos deben ser validados, unidos y cargados en un warehouse para análisis."

**Preguntas**:

1. **Describe los pasos del pipeline y las herramientas que usarías**

**Paso 1 – Extracción de los datos:** como se da a entender en el escenario, se tienen 3 fuentes de datos, el CSV mediante el protocolo SFTP se extraen los datos, sea cual sea la estación en la que se encuentre, para la fuete API Rest, se requiere coordinar con el proveedor externo los accesos correspondientes, y posterior hacer los request necesarios para extrar la data, y finalmente para la fuente de MongoDB, se usaría ***Python*** para hacer las consultas pertinententes, es aclarar que en este punto se debe conocer a cabalidad la necesidad de todos los datos solicitados a extraer, esto con el propósito de no generar sobrecarga en las consultas de MongoDB ni en los ***request*** de API Rest

**Paso 2 – Validación, limpieza y transformación de datos:** en este paso es importante y crucial el uso de ***Python*** y sus librerías como ***Pandas***, con el propósito de poder validar formatos en los datos, revisar valores nulos, sus posibles transformación con métodos estadísticos y/o eliminación de datos que no generan valor para el ejercicio, como duplicados o datos incompletos, y de acuerdo a la identifiaccion de variables clave entre los dataframes generados se deben hacer las correspondientes uniones atendiendo a la necesidad del negocio

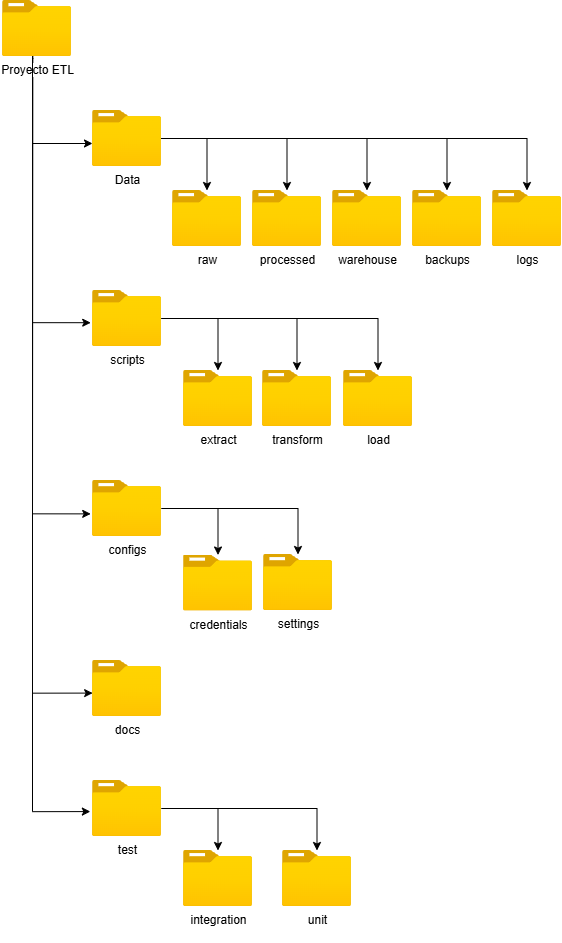
**Paso 3 – Cargue al Warehouse:** Si bien en síntesis los data warehouse son bases de datos para el manejo de grandes volumnes de datos, se puede utilizar sentencias ***SQL*** para cargar, o utilizar ***PySpark*** con para pipelines as automatizados, finalmente tambien se puede contar con plataformas dedidcadas a este proceso como ***Airflow***, en caso de ser datos que se deban procesar en tiempo real se puede ustilziar S3 de ***AWS*** o ***Kafla***

**Paso 4 – Monitoreo:** En este punto los datos fueron extraídos, procesados y de acuerdo con el requerimiento se están utilizando en BI o Ciencia de Datos para la toma de decisiones, sin embargo, es pertinente poder tener alertas sobre fallas en las plataformas o en la ingesta de datos en el warehouse, para este caso usaría plataformas de monitoreo como ***Grafana***

1. **¿Cómo manejarías errores (ej: ¿API caída) y garantizarías idempotencia**

En el caso de que la API este caída, optaría por siempre tener intentos automáticos controlados con tenancy de los request manejando expeciones con requests.exceptions, en caso de no ser posible, “desconectar el enchufe” evitando sobre cargas  
  
En el caso del CSV con SFTP, verificaría que el archivo exista, hacer intentos controlados de solicitud, y tener algún tipo de log para ver el histórico, adicional de poder generar un respaldo fuera de la estación con políticas de autoguardado y eliminado para no tener sobrecarga de almacenamiento, ya sea en nube de Azure, AWS o Onpremise  
  
En el caso de la conexión con MongoDB, revisar si las consultas están siendo muy densas, y optimizarlas, y revisar si no hay conexión para comunicarse con el responsable  
  
Y finalmente para no tener duplicados y fgarantizar la idempotencia, haría que los cargues de los datos siempre tengan valores únicos en sus claves, antes de cargar revisar que no sean datos ya existentes y categorizar los archivos en el caso de los CSV, adicional siempre tener un backup en caso de que un rollback no funcionara

1. **Propón una estructura de carpetas para el proyecto (ej: `/data/raw`, `/scripts`, `/tests`)**

****