

Contexto General.

El transporte público es fundamental para la movilización de las personas en una ciudad, es por ello que conocer los trayectos, paradas, horarios y duración de los trayectos es indispensable para planificar correctamente los viajes.

Se necesita generar una plataforma de datos que contenga la información histórica de los viajes, de tal forma de identificar la cantidad de transportes disponibles por cada uno de los medios disponibles (buses, metro), ver qué horarios tiene mayor disponibilidad de transporte en una zona determinada, y cuáles de esas zonas han tenido la mayor variabilidad en los recorridos (señalando si han agregado o quitado recorridos). Por otro lado, es importante tener siempre la información lo más actualizada posible, es por ello que también será requerido en este control la obtención diaria de los recorridos disponibles en Santiago.

Información.

1. **Datos Históricos:** Estos datos se pueden obtener desde la plataforma de datos abiertos del Gobierno de Chile y el link es el siguiente:

<https://datos.gob.cl/dataset/33245>.

Este set de datos contiene la información mensual de la planificación de los distintos medios de transporte en Santiago, y está disponible cerca del día 15 de cada mes. Se requiere que la descarga de datos sea automatizada, para ello deberá utilizar la siguiente API:

https://datos.gob.cl/api/action/package_show?id=33245

Esta API devuelve los recursos disponibles, debe identificar la ULR en la respuesta que le permitirá descargar los archivos.

2. **Datos Diarios:** Para obtener la información de transportes diarios, primero deberá consultar una API que le devolverá todos los recorridos disponibles:

https://www.red.cl/restservice_v2/rest/getservicios/all

Luego, por cada uno de esos recorridos deberá obtener la información de su trayecto, horarios y paradas desde la siguiente API:

https://www.red.cl/restservice_v2/rest/conocerecorrido?codsint=101

(donde “101” es un código de servicio devuelto por la API anterior)

Encargo

Etapas 1:

- Seleccionar herramientas de procesamiento, transformación y visualización, justificando su aplicación.
- Incorporar ciclo de vida del dato en el proceso end to end.
- Analizar problemática planteada utilizando alguna arquitectura de referencia.
- Definir los procesos, flujos de información y orquestación de datos para iniciar la construcción de la solución.

Etapas 2:

- Construir procesos de carga hacia storage, considerando disponibilidad de la información desde de la fuente.
- Construir procesos de transformación, limpieza de datos Batch.
- Construir procesos orquestados considerando disponibilidad de información y dependencias de grandes volúmenes de datos en formato Batch.

Entregables

- Informe:
(Todo lo que lleva un informe, al principio y al final)
Desarrollo con etapas de gestión de volúmenes de datos.
Información para implementación.
- Presentación:
Debe incluir visualizaciones interactivas hacia lo construido.

Notas

- Si necesita utilizar API's de terceros no indicadas en este contexto, utilícelas e infórmelas.
- Si necesita trabajar sobre supuestos, hágalo e infórmelo.
- No es necesario que explique, por ejemplo, que es Big Data, que es GCP, que es SQL, etc.
- Las conclusiones deben referirse única y exclusivamente a las temáticas abordadas e investigadas.
- La presentación debe ser equilibrada entre todos los integrantes del grupo. De no ser así, la nota grupal se verá afectada.
- No se permite leer el contenido que presentará de manera íntegra, aunque si está permitido utilizar algún apunte para apoyar ciertas ideas (nunca he esperado que se aprenda una presentación de memoria, pero si espero que sepa de lo que está exponiendo)
- La fecha/hora límite para cargar el informe y la presentación está en AVA.

Éxito.