**Nadia Arellano – Ana Karina Muñoz**
**Ingenieras en Informática**

# Optimización del Transporte Público con Big Data y GCP

## Descripción General

Este proyecto consiste en el diseño e implementación de una arquitectura Big Data sobre Google Cloud Platform para abordar desafíos críticos del transporte público en Santiago, tales como congestión, puntualidad y eficiencia operativa. Mediante la recopilación, procesamiento y visualización de grandes volúmenes de datos en tiempo real y batch, se generaron insights valiosos para mejorar la toma de decisiones, planificación de rutas y experiencia del usuario. El sistema integra múltiples servicios de GCP para automatizar el ciclo de vida de los datos desde su captura hasta su análisis y visualización, permitiendo una gestión eficiente y estratégica de la información pública de transporte.

## Arquitectura y Procesos

La solución desarrollada se basó en una arquitectura cloud sobre Google Cloud Platform, compuesta por procesos batch y streaming para el tratamiento integral de datos del transporte público de Santiago.

1. **Ingesta de datos históricos (.csv)**
   Se descargaron archivos desde fuentes públicas y se almacenaron en Cloud Storage. Estos archivos representan información estructurada sobre servicios de transporte urbano.

2. **Procesamiento por lotes (batch)**
   Utilizando Apache Beam sobre Dataflow, se construyó un pipeline llamado dataflow_servicios.py, el cual realiza:

   - ✓ Limpieza de datos
   - ✓ Transformación de campos
   - ✓ Carga estructurada hacia BigQuery, en el dataset lake y tabla servicios_procesados

3. **Ingesta de datos en tiempo real (streaming)**
   Se desarrolló el script mensaje-test.py que obtiene datos JSON desde la API pública https://www.red.cl/restservice_v2/rest/conocerecorrido, los publica en Pub/Sub y los guarda temporalmente en local para ser subidos a Cloud Storage.

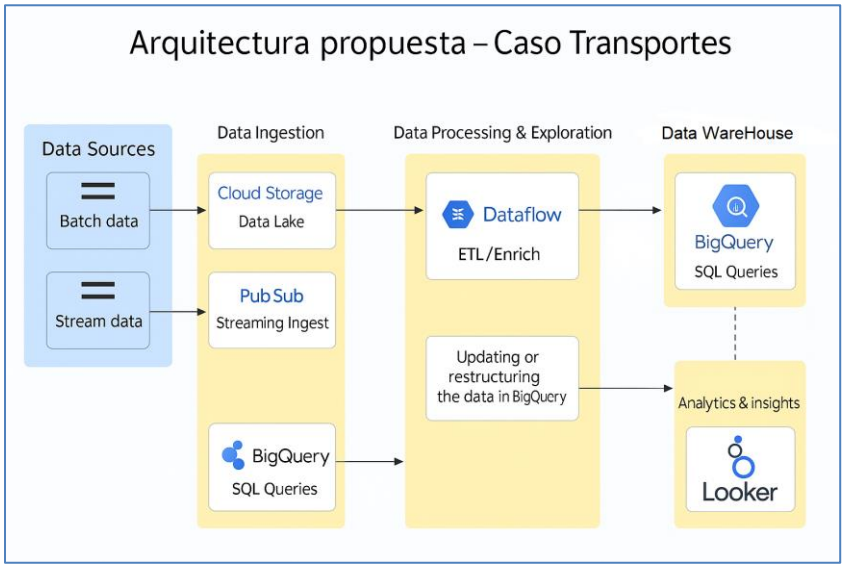
4. **Pipeline de transformación en tiempo real**
   Un segundo pipeline en Dataflow se conecta al tópico Pub/Sub para:

   - ✓ Decodificar los mensajes JSON
   - ✓ Validar la estructura
   - ✓ Cargar los datos procesados en tiempo real en BigQuery, en el dataset lake_real_time

5. **Visualización con Looker Studio**
   Los datos almacenados en BigQuery fueron visualizados mediante paneles dinámicos que incluyen filtros por comuna, empresa operadora, accesibilidad y frecuencia de servicios.

Cada componente de esta arquitectura está automatizado, es escalable, y permite extender el análisis hacia otras fuentes de datos o aplicar modelos predictivos en futuras fases.

La siguiente imagen, muestra el flujo de datos desde fuentes batch y en tiempo real, su almacenamiento en Cloud Storage, ingesta vía Pub/Sub, procesamiento con Dataflow y almacenamiento estructurado en BigQuery. Finalmente, los datos se visualizan mediante dashboards interactivos en Looker Studio para generar insights y apoyar la toma de decisiones estratégicas.



## Gobierno de Datos y Calidad de la Información

Se definió un flujo end-to-end que asegura la trazabilidad del dato desde su captura hasta su visualización. Incluye almacenamiento seguro en buckets configurados, control de errores en ingesta y transformación, registro de logs, y condiciones para reintento y eliminación de archivos locales temporales.

## Casos de Uso

¿Cuáles son las agencias que operan servicios de transporte público en Santiago?
¿Qué paraderos son accesibles para personas en silla de ruedas?
¿Cuántos viajes se realizan por cada ruta en el sistema de transporte?
¿Qué comunas concentran la mayor cantidad de servicios activos?
¿Qué paraderos tienen el mayor número de servicios registrados?
¿Qué empresas operan la mayor cantidad de recorridos?

## Herramientas Tecnológicas

| Tecnología | Aplicación en el proyecto |
|---|---|
| **Google Cloud Storage** | Almacenó tanto los archivos históricos en formato .csv descargados desde fuentes públicas, como los archivos JSON generados por los scripts de ingesta en tiempo real. Se utilizaron buckets separados para procesos batch y streaming. |
| **Google Cloud Pub/Sub** | Ingestó datos en tiempo real desde una API externa del sistema de transporte público (Red.cl). Los mensajes se publicaron en un tópico personalizado y sirvieron como entrada al pipeline de Dataflow. |
| **Google Cloud Dataflow** | Ejecutó pipelines construidos en Python con Apache Beam para procesar los datos en ambos escenarios (batch y streaming). Se aplicaron transformaciones, limpieza de campos, formatos de fecha y estructuras para carga eficiente en BigQuery. |
| **Apache Beam** | Framework utilizado dentro de Dataflow para definir la lógica de los pipelines ETL. Se desarrollaron scripts específicos como dataflow_servicios.py y mensaje-test.py para manejar múltiples estructuras de datos. |
| **BigQuery** | Almacenó los datos procesados desde los pipelines. Se crearon dos datasets: uno para datos por lotes (lake) y otro para streaming (lake_real_time). Fue la base para todas las consultas analíticas. |

| | |
|---|---|
| **Looker Studio** | Se conectó directamente a BigQuery para crear dashboards con filtros por comuna, empresa, y accesibilidad. Permitió visualizar los indicadores clave y responder preguntas del negocio en tiempo real. |
| **Python 3.8** | Lenguaje usado para automatizar el flujo completo de datos: consumo de API, transformación de JSON, publicación en Pub/Sub, y ejecución de pipelines. También se empleó para manejar logs y reintentos. |

## Resultados e Impacto Profesional

La solución desarrollada permitió integrar datos históricos y en tiempo real del sistema de transporte público de Santiago dentro de una arquitectura escalable y automatizada en Google Cloud Platform (GCP).

Entre los resultados más relevantes:

- Se diseñaron y ejecutaron pipelines de transformación en Dataflow para datos batch (.csv) y streaming (JSON en Pub/Sub), garantizando limpieza, estandarización y calidad en la carga a BigQuery.
- Se automatizó el flujo de ingesta desde APIs abiertas con control de errores, reintentos, y logs, asegurando trazabilidad y disponibilidad de los datos.
- Se implementaron consultas analíticas en BigQuery que permitieron responder preguntas estratégicas como: paraderos más utilizados, cobertura por comuna, y accesibilidad universal.
- Se crearon dashboards interactivos en Looker Studio, listos para toma de decisiones operativas y visualización pública.

Desde una perspectiva profesional, este proyecto demuestra competencias clave en:

- Arquitectura cloud en GCP para soluciones de análisis de datos.
- Desarrollo y ejecución de pipelines ETL en Apache Beam.
- Automatización de flujos de datos con Python.
- Aplicación de buenas prácticas en gobierno de datos, almacenamiento y visualización.

Esta solución es totalmente escalable y adaptable, preparada para incorporar modelos de predicción o extenderse a otros contextos como movilidad urbana, logística o servicios públicos inteligentes.