

**Big data**

***[Transporte público]***

***Fecha: [19/04/2024]***

***Orlando Echeverría***

***Sebastián Cordero***

***Lucas Ramírez***

**Tabla de contenido**

Introducción

[1. Infraestructura](#_heading=h.3znysh7)

[2. Proveedores](#_heading=h.2et92p0)

3[. Origen de los datos](#_heading=h.4d34og8)

4. [Tipos de datos](#_heading=h.17dp8vu)

5[. Arquitectura](#_heading=h.tyjcwt)

6[.](#_heading=h.3dy6vkm) Ciclo de vida del dato

7[.](#_heading=h.1t3h5sf) Gobierno de datos

8.Conclusión

**Introducción**

Con el objetivo de mejorar la planificación y gestión del transporte público en Santiago, se propone la implementación de Big Data para recopilar, procesar y analizar datos relevantes. En el presente informe se detallan aspectos esenciales, tales como la infraestructura requerida, los potenciales proveedores, la arquitectura a utilizar, el ciclo de vida del dato, la fuente de los datos, los tipos de información a considerar y el proceso de implementación necesario para alcanzar este objetivo.

Este proyecto representa una iniciativa crucial para abordar los desafíos actuales en la gestión del transporte público de Santiago. La implementación de tecnologías avanzadas como Big Data permitirá mejorar la eficiencia operativa, optimizar las rutas y horarios, y ofrecer un servicio más efectivo y adaptado a las necesidades de los usuarios. Además, esta modernización tecnológica contribuirá a reducir los tiempos de espera y los congestionamientos, mejorando así la experiencia general de los usuarios del transporte público en la ciudad. En conjunto, estas acciones buscan transformar positivamente el sistema de transporte público de Santiago para beneficio de todos los ciudadanos.

1. **Infraestructura**

Para mejorar la eficiencia y escalabilidad en el manejo de datos, se propone migrar a la nube utilizando Google Cloud Platform (GCP). La infraestructura estará compuesta por varios componentes clave de GCP, incluyendo Google Cloud Storage para un almacenamiento seguro y escalable de datos, Google Cloud Dataflow para el procesamiento eficiente de grandes volúmenes de datos en tiempo real y por lotes (BATCH), y Google BigQuery para realizar análisis SQL de alto rendimiento y extraer información valiosa de los datos almacenados.

**1.1 Por qué ocupamos Cloud**

* **Escalabilidad:**  La nube permite escalar recursos fácilmente según las necesidades del negocio. Se puede aumentar o reducir la capacidad de almacenamiento, potencia de cómputo y otros recursos de manera rápida, sin la necesidad de invertir en hardware adicional.
* **Acceso Global:** La nube permite acceder a tus datos y aplicaciones desde cualquier lugar con conexión a Internet, lo que facilita el trabajo remoto, la colaboración entre equipos y la disponibilidad para usuarios en diferentes ubicaciones.
* **Actualizaciones y mantenimiento:** Los proveedores de servicios en la nube se encargan de la actualización y mantenimiento de la infraestructura, lo que libera a equipos de TI para enfocarse en tareas más estratégicas y de valor para el negocio.
* **Seguridad:** Los proveedores de nube suelen ofrecer medidas de seguridad avanzadas, como cifrado de datos, control de accesos y cumplimiento de normativas, lo que puede ser más efectivo que gestionar la seguridad de forma local.
* **Costos:** En la nube, pagas por lo que usas, lo que puede ser más rentable que mantener y actualizar constantemente una infraestructura local, evitando la compra de equipos costosos y gastos de mantenimiento o energía.
* **Flexibilidad:** En la nube, se puede probar y desplegar nuevas aplicaciones y servicios rápidamente, lo que brinda mayor flexibilidad para adaptarse a los cambios.

1. **Proveedores**

Google Cloud Platform será el proveedor principal de la infraestructura y servicios de Big Data, debido a su robustez y variedad de herramientas especializadas. Además, se contará con partners certificados de GCP que brindarán soporte experto en la implementación y gestión de la solución. Estos partners garantizarán una migración exitosa hacia la nube y un funcionamiento eficiente del sistema de Big Data.

Los partners certificados de Google Cloud Platform son empresas o individuos que han demostrado su experiencia y competencia en el uso de las tecnologías de Google Cloud, y que están autorizados para ofrecer servicios de consultoría, implementación y soporte a clientes que buscan aprovechar al máximo la plataforma de Google Cloud.

1. **Origen de los datos**

Los datos provendrán de fuentes como la plataforma de datos abiertos del Gobierno de Chile para datos históricos y APIs proporcionadas por los operadores de transporte para datos diarios.

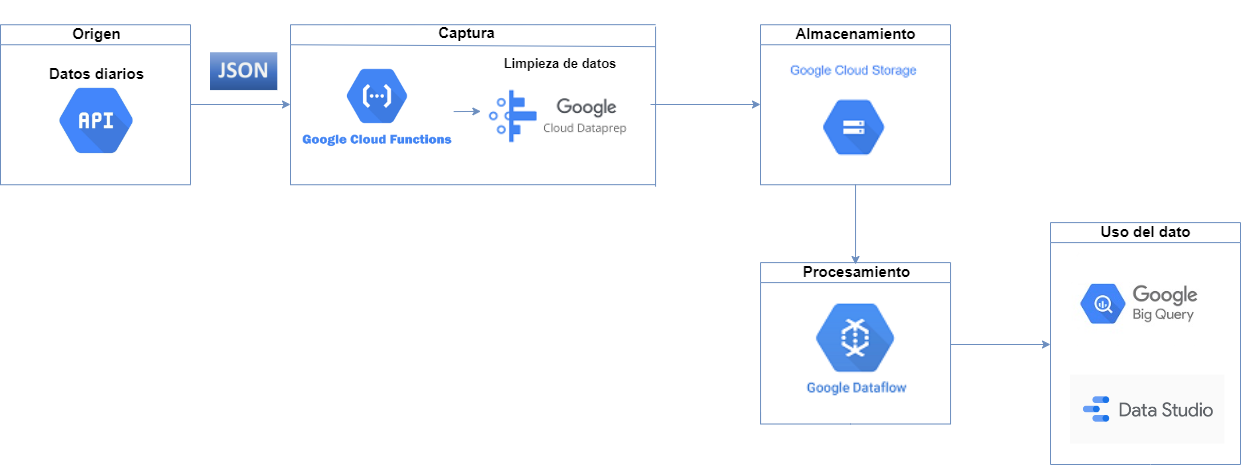
1. **Tipos de datos**

Los datos manejados serán principalmente estructurados, incluyendo información de trayectos, horarios, paradas y recorridos. Los datos transaccionales y descriptivos serán almacenados en una base de datos relacional para mantener su integridad y coherencia.

1. **Arquitectura**

La arquitectura en la nube de Google Cloud Platform(GCP) ofrece una combinación única de escalabilidad, seguridad, flexibilidad y rendimiento. Esta arquitectura se adapta perfectamente a las necesidades del proyecto al permitir gestionar grandes volúmenes de datos de manera eficiente y efectiva. La escalabilidad de GCP asegura que la infraestructura pueda crecer según las demandas del sistema, mientras que las herramientas de seguridad integradas garantizan la protección de los datos sensibles. La flexibilidad de la arquitectura permite integrar fácilmente nuevas herramientas y servicios de GCP según sea necesario, asegurando así un entorno de Big Data ágil y adaptativo.

**5.1 Arquitectura de datos diarios para los recorridos en real time.**



* **Ciclo de vida del dato:**
* **Estructura para Datos de Recorridos:**

**→ Captura de Datos:**

**Google Cloud Functions:** Se utiliza para capturar datos de rutas en tiempo real. Google Cloud Functions permite ejecutar código en respuesta a eventos específicos, lo que puede ser útil para capturar datos de rutas cuando hay eventos específicos o desencadenantes que indican la disponibilidad de nuevos datos de ruta.

En el caso de estudio, la captura de datos de recorridos en tiempo real es fundamental para comprender y analizar los trayectos, horarios, paradas y duración de los viajes. Esta información es esencial para identificar patrones de disponibilidad del transporte y para analizar cambios en los recorridos a lo largo del tiempo.

**→ Almacenamiento de Datos:**

**Google Cloud Storage:** Es una opción adecuada para almacenar datos de transporte público, ofreciendo escalabilidad y rendimiento para manejar grandes conjuntos de datos. Proporciona un almacenamiento seguro y escalable. Los datos de recorridos, horarios, paradas y variabilidades en los trayectos pueden almacenarse y procesarse eficientemente, lo que permite un análisis rápido y en tiempo real. Con su capacidad para manejar grandes volúmenes de datos con baja latencia, es una opción sólida para almacenar y analizar datos de transporte público de manera eficiente.

**→** **Procesamiento de Datos:**

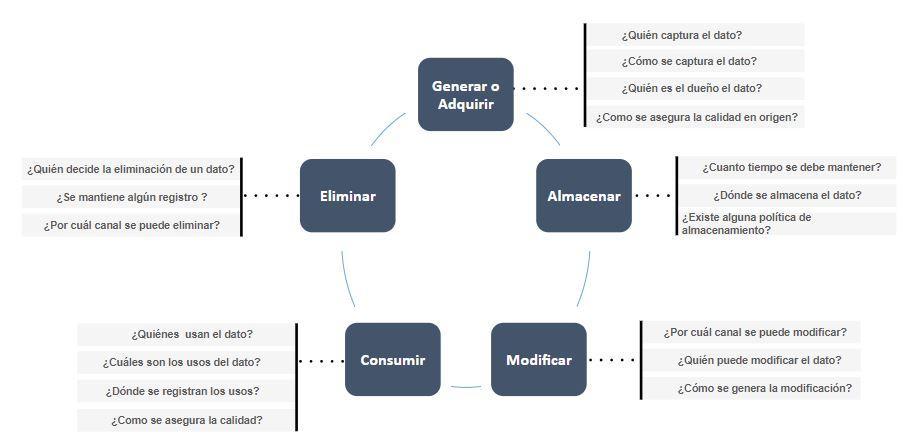
**Google Cloud Dataflow:** Es esencial en este paso del pipeline. Permite ejecutar pipelines de procesamiento de datos diarios del transporte. A través de transformaciones y análisis avanzados, se pueden identificar patrones, tendencias y áreas de mejora en el servicio de transporte. Para obtener insights significativos sobre la disponibilidad de transportes, horarios óptimos y variabilidades en los recorridos, es necesario procesar y analizar los datos capturados de manera eficiente.

**→ Uso de Datos:**

**Google BigQuery:** Desempeña un papel crucial en este paso al ofrecer un servicio robusto para realizar análisis avanzados sobre los datos procesados. Permitirá explorar patrones de viaje, identificar peak de demanda, optimizar rutas y horarios, y evaluar la efectividad de las decisiones tomadas para mejorar la experiencia de los usuarios.

Los datos procesados son fundamentales para tomar decisiones informadas y estratégicas en la gestión del transporte público en Santiago. Estos datos proporcionan información clave sobre la disponibilidad de transportes, horarios óptimos y tendencias en los recorridos que son significativos para mejorar la eficiencia y calidad del servicio.

Se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y visualizar los resultados de manera intuitiva, facilitando la comprensión y la toma de decisiones basadas en datos.

****

* **Datos de Recorridos:**
* **Generar o Adquirir:**

**→ ¿Quién captura el dato?:** Los datos de recorridos son capturados por la API de recorridos en tiempo real del sistema de transporte público.

**→ ¿Cómo se captura el dato?:** El dato se captura mediante consultas programadas a la API de recorridos del sistema de transporte público.

**→ ¿Quién es el dueño del dato?:** El dueño de los datos de recorridos es la entidad responsable de la gestión del sistema de transporte público.

**→ ¿Cómo se asegura la calidad en origen?:** La calidad en origen se asegura mediante validaciones y controles implementados en la API de recorridos para garantizar la integridad y precisión de los datos.

* **Almacenar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos de recorridos puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Dónde se almacena el dato?:** Los datos de recorridos se almacenan en Google Cloud Storage para un almacenamiento escalable y de baja latencia.

**→ ¿Existe alguna política de almacenamiento?:** Sí, se puede establecer una política de almacenamiento que determine la retención de datos y otros aspectos relacionados con el almacenamiento en Cloud Storage.

* **Modificar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos de recorridos puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Por cuál canal se puede modificar?:** Los datos de recorridos pueden modificarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos.

* **Consumir:**

**→ ¿Quiénes usan el dato?:** Los datos de recorridos son utilizados por los analistas de transporte público, planificadores de rutas y autoridades de transporte.

**→ ¿Cuáles son los usos del dato?:** Los datos de recorridos se utilizan para analizar la eficiencia del transporte, planificar nuevas rutas y mejorar la experiencia del usuario.

**→ ¿Dónde se registran los usos?:** Los usos de los datos de recorridos pueden registrarse en un sistema de seguimiento interno de la entidad responsable del transporte público.

**→ ¿Cómo se asegura la calidad?:** La calidad de los datos se asegura mediante la aplicación de controles de calidad durante el procesamiento y análisis de los datos.

* **Eliminar:**

**→ ¿Quién decide la eliminación de un dato?:** La eliminación de un dato puede ser decidida por la entidad responsable de la gestión del sistema de transporte público.

**→ ¿Se mantiene algún registro?:** Sí, puede mantenerse un registro de las eliminaciones de datos para fines de auditoría y cumplimiento normativo.

**→ ¿Por cuál canal se puede eliminar?:** Los datos pueden eliminarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos y cumplimiento normativo.

* **Realización**
* **Captura de datos:**

**→ Herramienta:** Google Cloud Functions y Data Proc.

**→ ¿Cómo se realiza?:** Los datos se capturan utilizando funciones de Google Cloud Functions, activadas en respuesta a eventos específicos, como la llegada de nuevos datos. Mientras que Data Proc se encarga de la limpieza de los datos antes de su procesamiento y carga en BigQuery.

* **Almacenamiento de datos:**

**→ Herramienta:** Google Cloud Storage.

**→ ¿Cómo se realiza?:** Los datos capturados se almacenan en Google Cloud Storage para su posterior procesamiento y análisis.

Permite un almacenamiento y procesamiento eficiente de los datos de trayectos, horarios, paradas y variabilidades en los recorridos, lo que facilita un análisis rápido y en tiempo real.

* **Procesamiento de datos:**

**→ Herramientas:** Google Cloud Dataflow.

**→ ¿Cómo se realiza?:** Se utiliza Google Cloud Dataflow para ejecutar pipelines de procesamiento de datos tanto en tiempo real como en batch.

Dataflow aplica transformaciones necesarias en los datos.

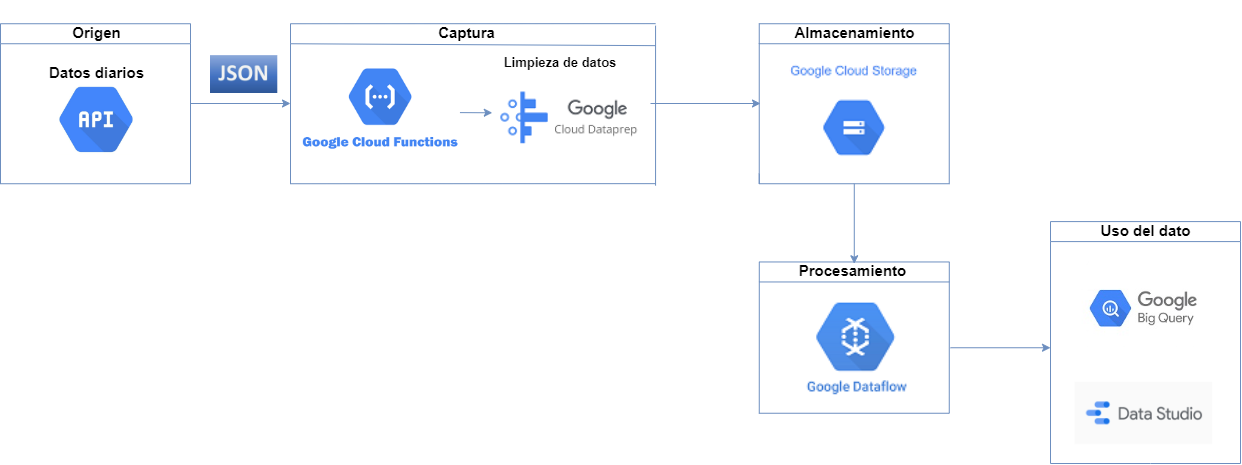
**Uso de datos:**

**→ Herramienta:** BigQuery.

**→ ¿Cómo se realiza?:** Se ejecutan consultas SQL en BigQuery para analizar los datos procesados y almacenados.

BigQuery permite realizar análisis SQL de alto rendimiento sobre los conjuntos de datos almacenados, y se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y facilitar la toma de decisiones.

**5.2 Arquitectura para datos diarios de las rutas en real time**



* **Ciclo de vida del dato:**
* **Estructura para Datos de Rutas:**

**→ Captura de Datos:**

**Google Cloud Functions:** Se utiliza para capturar datos de rutas en tiempo real. Google Cloud Functions permite ejecutar código en respuesta a eventos específicos, lo que puede ser útil para capturar datos de rutas cuando hay eventos específicos o desencadenantes que indican la disponibilidad de nuevos datos de ruta.

**→ Almacenamiento de Datos:**

**Google Cloud Storage:** Es una opción adecuada para almacenar datos de transporte público, ofreciendo escalabilidad y rendimiento para manejar grandes conjuntos de datos. Proporciona un almacenamiento seguro y escalable. Los datos de las rutas pueden almacenarse y procesarse eficientemente, lo que permite un análisis rápido y en tiempo real. Con su capacidad para manejar grandes volúmenes de datos con baja latencia, es una opción sólida para almacenar y analizar datos de transporte público de manera eficiente.

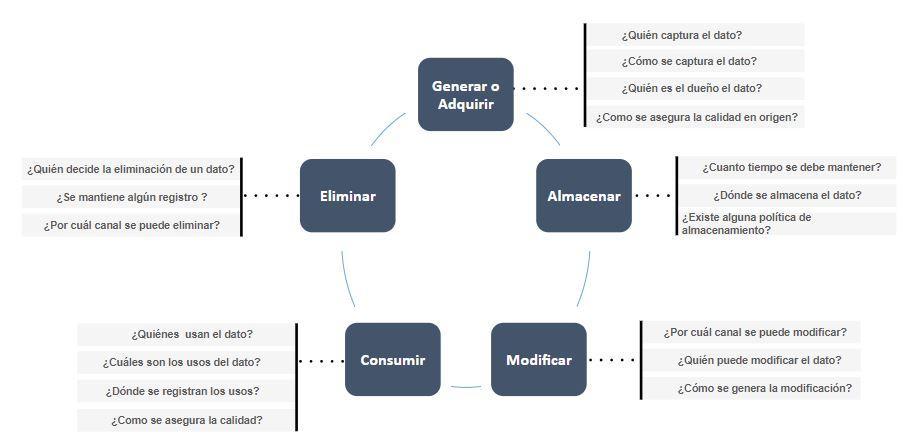
**→ Procesamiento de Datos:**

**Google Cloud Dataflow:** Es esencial en este paso del pipeline. Permite ejecutar pipelines de procesamiento de datos diarios del transporte. A través de transformaciones y análisis avanzados, se pueden identificar patrones, tendencias y áreas de mejora en el servicio de transporte. Para obtener insights significativos sobre la disponibilidad de transportes, horarios óptimos y variabilidades en los recorridos, es necesario procesar y analizar los datos capturados de manera eficiente.

**→ Uso de Datos:**

**Google BigQuery:** Desempeña un papel crucial en este paso al ofrecer un servicio robusto para realizar análisis avanzados sobre los datos procesados. Permitirá explorar patrones de viaje, identificar peak de demanda, optimizar rutas y horarios, y evaluar la efectividad de las decisiones tomadas para mejorar la experiencia de los usuarios.

Se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y visualizar los resultados de manera intuitiva, facilitando la comprensión y la toma de decisiones basadas en datos.

****

* **Datos de Rutas:**
* **Generar o Adquirir:**

**→ ¿Quién captura el dato?:** La información sobre las rutas se obtiene a través de la API de rutas del sistema de transporte público.

**→ ¿Cómo se captura el dato?:** Los datos de las rutas se recopilan haciendo consultas programadas a la API de rutas del sistema de transporte público.

**→ ¿Quién es el dueño del dato?:** El dueño de los datos de rutas es la entidad responsable de la gestión del sistema de transporte público.

**→ ¿Cómo se asegura la calidad en origen?**: La calidad en origen se garantiza mediante validaciones y controles implementados en la API de rutas para mantener la integridad y precisión de los datos.

* **Almacenar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos de rutas puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Dónde se almacena el dato?:** Los datos de rutas se almacenan en Google Cloud Storage para un acceso rápido y seguro.

**→ ¿Existe alguna política de almacenamiento?:** Sí, se puede establecer una política de almacenamiento que determine la retención de datos y otros aspectos relacionados con el almacenamiento en Google Cloud Storage.

* **Modificar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos de rutas puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Por cuál canal se puede modificar?:** Los datos de rutas pueden modificarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos.

* **Consumir:**

**→ ¿Quiénes usan el dato?:** Los datos de rutas son utilizados por los planificadores de transporte, desarrolladores de aplicaciones de navegación y usuarios del sistema de transporte público.

**→ ¿Cuáles son los usos del dato?:** Los datos de rutas se utilizan para calcular y mostrar itinerarios, estimar tiempos de viaje y proporcionar información de navegación precisa.

**→ ¿Dónde se registran los usos?:** Los usos de los datos de rutas pueden registrarse en un sistema de seguimiento interno del sistema de transporte público o en herramientas de análisis de datos.

* **Eliminar:**

**→ ¿Quién decide la eliminación de un dato?:** La eliminación de un dato puede ser decidida por la entidad responsable de la gestión del sistema de transporte público.

**→¿Se mantiene algún registro?:** Sí, puede mantenerse un registro de las eliminaciones de datos para fines de auditoría y cumplimiento normativo.

**→ ¿Por cuál canal se puede eliminar?:** Los datos pueden eliminarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos y cumplimiento normativo.

* **Realización**
* **Captura de datos:**

**→ Herramienta:** Google Cloud Functions y Data Proc.

**→ ¿Cómo se realiza?:** Los datos se capturan utilizando funciones de Google Cloud Functions, activadas en respuesta a eventos específicos, como la llegada de nuevos datos. Mientras que Data Proc se encarga de la limpieza de los datos antes de su procesamiento y carga en BigQuery.

* **Almacenamiento de datos:**

**→ Herramienta:** Google Cloud Bigtable.

**→ ¿Cómo se realiza?:** Los datos capturados se almacenan en Google Cloud Bigtable para su posterior procesamiento y análisis.

* **Procesamiento de datos:**

**→ Herramientas:** Google Cloud Dataflow.

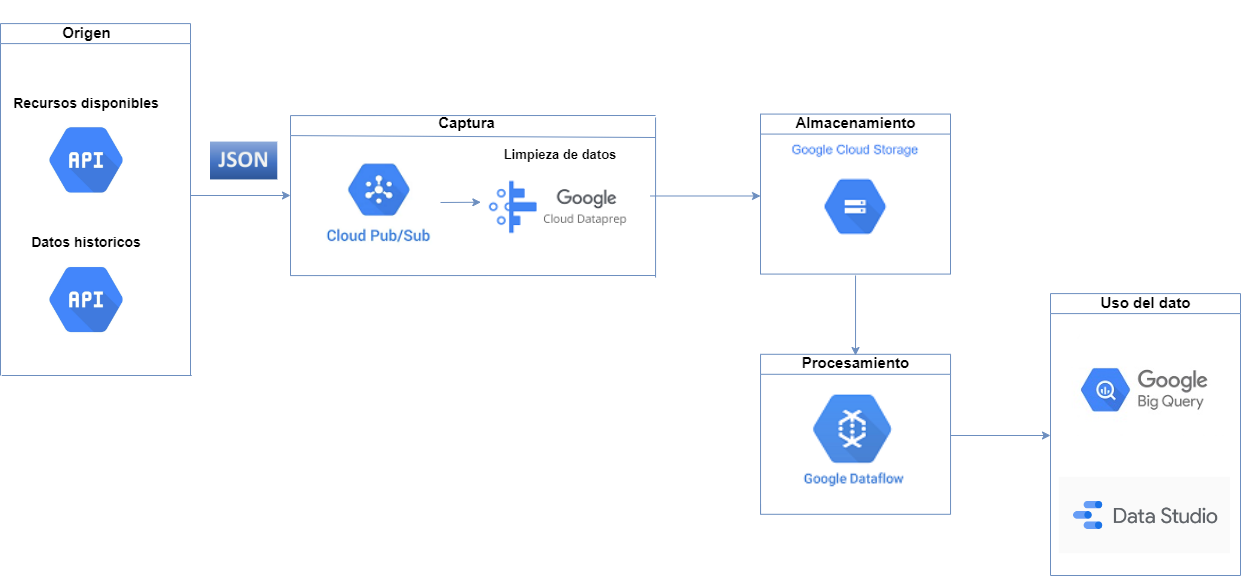
**→ ¿Cómo se realiza?:** Se utiliza Google Cloud Dataflow para ejecutar pipelines de procesamiento de datos, aplicando transformaciones necesarias.

* **Uso de datos:**

**→ Herramienta:** BigQuery.

**→ ¿Cómo se realiza?:** Se ejecutan consultas SQL en BigQuery para analizar los datos procesados y almacenados. Se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y visualizar los resultados de manera intuitiva, facilitando la comprensión y la toma de decisiones basadas en datos.

**5.3 Arquitectura batch para datos históricos y recursos disponibles**

****

* **Ciclo de vida del dato:**
* **Estructura para Datos históricos:**

**→ Captura de Datos:**

**Google Cloud Pub/Sub**: Se utiliza para la ingestión escalable y eficiente de los datos del transporte público. Permite gestionar grandes volúmenes de datos de manera fluida, asegurando que la plataforma de datos tenga acceso a la información necesaria para los análisis posteriores.

En el caso de estudio, la captura de datos históricos y de recursos disponibles del transporte público de Santiago es fundamental para comprender y analizar la evolución del sistema a lo largo del tiempo.

**→ Almacenamiento de Datos:**

**Google Cloud Storage:** Se utiliza como el repositorio central para almacenar estos datos de transporte público. Proporciona la escalabilidad necesaria para manejar grandes conjuntos de datos, asegurando su disponibilidad y accesibilidad para los procesos de análisis.

Los datos capturados, tanto históricos como de recursos disponibles, necesitan un almacenamiento seguro y escalable para su posterior procesamiento y análisis. En este caso, esta información incluye datos de trayectos, horarios, paradas y variabilidades en los recorridos.

**→ Procesamiento de Datos:**

Google Cloud Dataflow: Es esencial en este paso del pipeline. Permite ejecutar pipelines de procesamiento de datos batch para analizar los datos históricos y de recursos disponibles del transporte público. A través de transformaciones y análisis avanzados, se pueden identificar patrones, tendencias y áreas de mejora en el servicio de transporte.

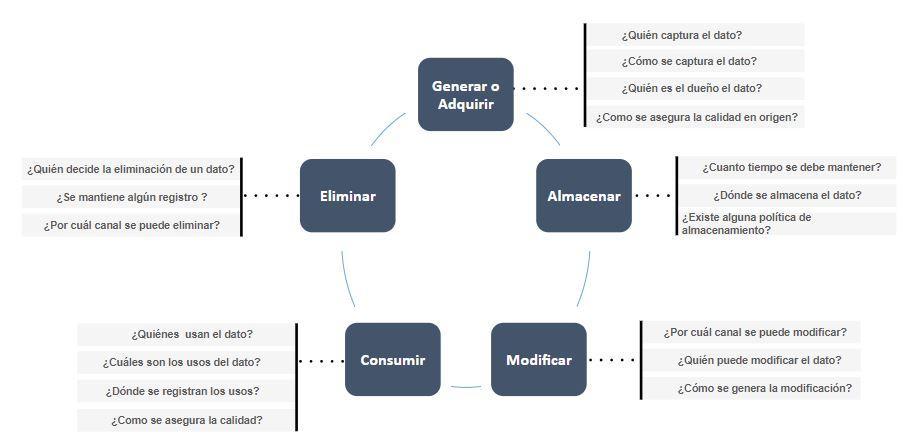
Para obtener insights significativos sobre la disponibilidad de transportes, horarios óptimos y variabilidades en los recorridos, es necesario procesar y analizar los datos capturados de manera eficiente.

**→ Uso de Datos:**

**Bigquery:** Desempeña un papel crucial en este paso al ofrecer un servicio robusto para realizar análisis avanzados sobre los datos procesados. Permitirá explorar patrones de viaje, identificar peaks de demanda, optimizar rutas y horarios, y evaluar la efectividad de las decisiones tomadas para mejorar la experiencia de los usuarios.

Los datos procesados son fundamentales para tomar decisiones informadas y estratégicas en la gestión del transporte público en Santiago. Estos datos proporcionan información clave sobre la disponibilidad de transportes, horarios óptimos y tendencias en los recorridos que son significativos para mejorar la eficiencia y calidad del servicio.

Se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y visualizar los resultados de manera intuitiva, facilitando la comprensión y la toma de decisiones basadas en datos.

****

* **Generar o Adquirir:**

**→ ¿Quién captura el dato?:** Los datos históricos y de recursos disponibles son capturados por sistemas automatizados que recopilan información del sistema de transporte público de Santiago.

**→ ¿Cómo se captura el dato?:** Los datos se capturan a través de consultas programadas a las bases de datos históricas y a los sistemas de monitoreo de recursos disponibles.

**→ ¿Quién es el dueño del dato?:** El dueño de los datos históricos y de recursos disponibles es la entidad responsable de la gestión del sistema de transporte público.

**→ ¿Cómo se asegura la calidad en origen?:** La calidad en origen se asegura mediante la implementación de controles de calidad en los sistemas de captura de datos y la validación de los datos recopilados.

* **Almacenar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos históricos y de recursos disponibles puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Dónde se almacena el dato?:** Los datos históricos y de recursos disponibles se almacenan en Google Cloud Storage para su posterior procesamiento y análisis.

**→ ¿Existe alguna política de almacenamiento?:** Sí, se puede establecer una política de almacenamiento que determine la retención de datos y otros aspectos relacionados con el almacenamiento en Google Cloud Storage.

* **Modificar:**

**→ ¿Cuánto tiempo se debe mantener?:** El tiempo de retención de los datos históricos y de recursos disponibles puede variar según los requisitos específicos del negocio y los requisitos legales.

**→ ¿Por cuál canal se puede modificar?:** Los datos históricos y de recursos disponibles pueden modificarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos.

* **Consumir:**

**→ ¿Quiénes usan el dato?:** Los datos históricos y de recursos disponibles son utilizados por analistas de transporte público, planificadores de infraestructura y autoridades gubernamentales.

**→¿Cuáles son los usos del dato?:** Los datos se utilizan para analizar tendencias históricas, planificar mejoras en la infraestructura de transporte y optimizar la asignación de recursos.

**→ ¿Dónde se registran los usos?:** Los usos de los datos pueden registrarse en un sistema de seguimiento interno de la entidad responsable del transporte público.

**→ ¿Cómo se asegura la calidad?:** La calidad de los datos se asegura mediante la implementación de controles de calidad durante el procesamiento y análisis de los datos.

* **Eliminar:**

**→ ¿Quién decide la eliminación de un dato?:** La eliminación de un dato puede ser decidida por la entidad responsable de la gestión del sistema de transporte público.

**→ ¿Se mantiene algún registro?:** Sí, puede mantenerse un registro de las eliminaciones de datos para fines de auditoría y cumplimiento normativo.

**→ ¿Por cuál canal se puede eliminar?:** Los datos pueden eliminarse a través de procesos controlados y autorizados por el personal responsable de la gestión de datos y cumplimiento normativo.

* **Realización:**
* **Captura de datos:**

**→ Herramienta:** Google Cloud Functions y Data Prep .

**→ ¿Cómo se realiza?:** Los datos se reciben y gestionan de manera escalable y confiable a través de Google Cloud Functions, Mientras que Data Prep se encarga de la limpieza de los datos antes de su procesamiento y carga en BigQuery.

* **Almacenamiento de datos:**

**→ Herramienta:** Google Cloud Storage.

**→ ¿Cómo se realiza?:** Los datos capturados se almacenan en Google Cloud Storage para su posterior procesamiento y análisis.

* **Procesamiento de datos:**

**→ Herramientas:** Google Cloud Dataflow.

**→ ¿Cómo se realiza?:** Se utiliza Google Cloud Dataflow para ejecutar pipelines de procesamiento de datos, aplicando transformaciones necesarias.

* **Uso de datos:**

**→ Herramienta:** BigQuery.

**→ ¿Cómo se realiza?:** Se ejecutan consultas SQL en BigQuery para analizar los datos procesados y almacenados. Se pueden utilizar herramientas de visualización como Google Data Studio para crear dashboards interactivos y visualizar los resultados de manera intuitiva, facilitando la comprensión y la toma de decisiones basadas en datos.

* **Ciclo de vida de variables de datos históricos**

**route\_id:**

****

**route\_long\_name:**

****

**trip\_id:**

****

**start\_time:**

****

**end\_time:**

****

1. **Gobierno de Datos**

Guiará todas las demás funciones de la gestión de datos. El propósito es asegurar una gestión efectiva y estratégica de la información para mejorar la planificación y la gestión del transporte, así como la movilidad urbana. Esta iniciativa aborda varios elementos claves:

* **Marco de Gobernanza**
* **Estrategia**

**→ Definición de Estrategias:** Implementaremos un sistema centralizado en Google Cloud para capturar y almacenar datos de trayectos, horarios y paradas de transporte público en Santiago.

**→ Estrategia de Gobierno de Datos:** Estableceremos políticas y estándares para garantizar la calidad, seguridad y uso efectivo de estos datos.

* **Políticas**

**→ Políticas de Privacidad en Google Cloud Platform**: Utilizaremos herramientas de encriptación y acceso controlado en Google Cloud Storage para proteger datos sensibles de acuerdo con las políticas de privacidad.

**→ Políticas de Calidad de Datos en Google Cloud Platform**: Implementaremos validaciones y reglas de calidad de datos en Google Cloud Dataflow para garantizar la precisión y coherencia de los datos antes de su almacenamiento en BigQuery.

**→ Política de Captura de Datos(Real time)**: Los datos deben ser capturados automáticamente por los sistemas de seguimiento y gestión del transporte público al finalizar cada viaje. Se debe garantizar la integridad y precisión de los datos capturados mediante validaciones y controles en tiempo real.

**→ Política de Almacenamiento de Datos**: Los datos deben almacenarse en una base de datos y segura de la autoridad de transporte público de Santiago. Debe existir una política de retención de datos que defina el período de tiempo durante el cual los datos deben mantenerse disponibles para análisis históricos y operativos.

**→ Política de Modificación de Datos**: Los datos no deben ser modificados manualmente después de su captura inicial, para preservar la integridad histórica de los registros. Cualquier corrección o modificación necesaria debe realizarse a través de procesos controlados y registrados en el sistema de gestión de transporte.

**→ Política de Acceso a los Datos**: El acceso a los datos debe estar restringido a personal autorizado dentro de la autoridad de transporte público, con roles y permisos definidos. Se debe mantener un registro de acceso y uso de los datos para fines de seguridad.

**→ Política de Eliminación de Datos**: La eliminación de datos debe seguir una política clara y documentada, basada en requisitos legales y de retención de datos. Se debe mantener un registro de eliminaciones con detalles sobre el motivo y la autorización de la eliminación.

**→ Política de Seguridad de Datos**: Los datos deben ser protegidos mediante medidas de seguridad, incluyendo cifrado, acceso basado en roles y monitoreo de seguridad. Deben establecerse procedimientos para la detección y respuesta ante posibles incidentes de seguridad que afecten a estos datos.

**→ Política de Calidad de Datos**: Se deben establecer controles de calidad automatizados para verificar la integridad y consistencia de los datos. Deben realizarse evaluaciones periódicas de la calidad de los datos.

**→ Política de Backup y Recuperación de Datos**: Se deben establecer procedimientos regulares del uso de backups de los datos para garantizar su disponibilidad en caso de fallos o desastres. Debe existir un plan de recuperación de datos que detalle cómo restaurar los datos en caso de pérdida o corrupción.

**→ Política de Compartición de Datos**: Antes de compartir datos con terceros (internos o externos), se deben establecer acuerdos de uso responsable de los datos. Debe designarse un responsable de la gestión de la compartición de datos y mantener registros de todos los intercambios de datos realizados.

* **Normas y Calidad**

**→ Estándares de Formatos de Datos en Google Cloud Platform:** Estableceremos normas detalladas para el formato de datos, incluyendo la estructura de datos como JSON, para garantizar la interoperabilidad y facilidad de uso en diferentes aplicaciones y sistemas.

**→ Control de Calidad de Datos en tiempo real:** Implementar procesos automatizados para el control de calidad de los datos en tiempo real, asegurando la integridad y precisión de los datos procesados.

* **Supervisión**

**→ Monitoreo Continuo:** Se implementará monitoreo y alertas en Google Cloud Platform para supervisar el rendimiento de los procesos en Google Cloud Dataflow y BigQuery, asegurando la disponibilidad y fiabilidad de los datos.

* **Cumplimientos**

**→ Capacitación en Google Cloud Platform:** Utilizaremos Google Cloud Training y recursos de documentación oficial para capacitar al personal en el uso efectivo de estas herramientas de GCP.

**→ Control de Accesos en Google Cloud Platform:** Implementaremos políticas de acceso basadas en roles para controlar y registrar el acceso a datos y recursos en GCP.

* **Gestión de Problemas**

**→ Gestión de Incidentes en Google Cloud Platform:** Establecer procedimientos para la gestión eficiente de incidentes relacionados con la infraestructura y los datos en Google Cloud Platform, minimizando el tiempo de inactividad y los impactos en el servicio.

**→ Gestión de Cambios:** Implementaremos revisiones y aprobaciones de cambios para mantener la integridad y consistencia de los datos.

* **Proyectos de Gestión de Datos**
* **Mejoras Continuas en Google Cloud Platform:** Identificar oportunidades de mejora y desarrollar proyectos adicionales para optimizar la infraestructura y los servicios en la nube de Google Cloud Platform, maximizando su rendimiento y eficiencia.
* **Priorización y Ejecución:** Priorizaremos proyectos como la optimización de consultas en BigQuery y la integración de datos de diferentes fuentes para análisis
* **Valoración de Activos de Datos**
* **Evaluación del Valor Estratégico de los Datos en Google Cloud Platform:** Utilizaremos BigQuery para análisis avanzados y generación de informes que evalúen el impacto estratégico de los datos en la toma de decisiones y mejoras en el servicio de transporte público.

1. **Áreas de conocimiento de la gestión de datos**

* **Arquitectura de los datos**
* Analizar y comprender las necesidades de almacenamiento y procesamiento de datos para fusionar eficazmente la información histórica, los recursos disponibles y los datos diarios del transporte público en Santiago.
* Desarrollar estructuras y estrategias flexibles para satisfacer tanto las demandas actuales como futuras de datos en la organización, abordando aspectos como los trayectos, paradas, horarios y la dinámica de los recorridos.
* Capacitar a las organizaciones para adaptarse ágilmente a las innovaciones tecnológicas y a las oportunidades comerciales emergentes, optimizando la planificación de viajes y el análisis del transporte público.
* **Modelado de Datos y Diseño**

Se busca comprender las distintas perspectivas para desarrollar aplicaciones que se ajusten a las necesidades comerciales presentes y futuras. Se diseñará una estructura que refleje con precisión la complejidad de los datos del transporte público, incluyendo entidades y relaciones que representen los diversos elementos del sistema. Este proceso establece una base sólida y prepara el camino para iniciativas estratégicas más amplias, como la gestión de datos maestros y los programas de gobierno de datos.

* **Almacenamiento de Datos y Operaciones**
* Gestionar la disponibilidad de datos a través de los ciclos de vida, utilizando Cloud Storage para almacenar los datos históricos y los recursos disponibles, y Google Bigtable para los datos diarios del transporte público, asegurando así un acceso rápido y fiable a la información clave en todo momento.
* Asegurar la integridad de los activos de datos del transporte público mediante operaciones de actualización y mantenimiento que garanticen la calidad y precisión de la información almacenada.
* Gestionar eficientemente la ejecución de las transacciones de datos, optimizando los procesos relacionados con el almacenamiento y actualización de los datos históricos y diarios del transporte público.
* **Seguridad de Datos**
* Se implementarán medidas de seguridad como el cifrado de datos y el control de acceso para proteger la información del transporte público y garantizar la confidencialidad de los datos de los usuarios.
* Se establecerán procedimientos para asegurar el cumplimiento de las regulaciones de seguridad de datos, comprendiendo y cumpliendo con todas las políticas relevantes.
* Se realizan evaluaciones periódicas para garantizar que las medidas de seguridad cumplan con las expectativas de todas las partes involucradas en el sistema de transporte público, promoviendo la confianza en el manejo de datos sensibles.
* **Integración de Datos e Interoperabilidad**
* Proveer datos de forma segura y cumpliendo con regulaciones, integraremos datos de diferentes fuentes, como las principales APIs de Recursos disponibles, Datos diarios y Datos históricos. Estableceremos protocolos de interoperabilidad para garantizar la disponibilidad y consistencia de la información.
* Reducir el costo y la complejidad de la gestión de soluciones mediante el desarrollo de modelos e interfaces compartidas, simplificando así la integración de datos y optimizando recursos.
* Identificar eventos significativos y desencadenar alertas y acciones automáticas, facilitando la detección y respuesta rápida ante situaciones relevantes para mejorar la eficiencia operacional.
* Apoyar los esfuerzos de inteligencia de negocio, analítica, gestión de datos maestros y eficiencia operacional mediante soluciones integradas que aprovechen los datos para impulsar iniciativas clave en el sistema de transporte público.
* **Gestión de Documentos y Contenido**
* Cumplir con las obligaciones legales y las expectativas del cliente en la gestión de registros, mediante la implementación de políticas y procedimientos adecuados.
* Garantizar un almacenamiento, recuperación y uso efectivos y eficientes de los documentos y contenidos, aprovechando herramientas de gestión de contenido y estableciendo procesos optimizados.
* Facilitar la integración entre los contenidos estructurados y no estructurados, desarrollando sistemas que permitan una gestión coherente y fluida de la información en la plataforma de datos.
* **Datos Maestros y de Referencia**
* Facilitar el intercambio de información entre diferentes áreas y aplicaciones dentro de la organización mediante la implementación de datos maestros y de referencia para asegurar la coherencia de los datos del transporte público.
* Establecer registros autoritarios de datos maestros, como horarios y listas de paraderos, para proporcionar una base confiable y evaluada que garantice la calidad y precisión de la información utilizada en toda la organización.
* Reducir costos y complejidad al utilizar estándares y modelos de datos comunes, así como patrones de integración, con el fin de simplificar la gestión de datos y promover una integración más eficiente entre las diversas fuentes de información.

* **Data Warehousing e Inteligencia de Datos**
* Establecer y mantener el entorno técnico y los procesos necesarios para integrar datos y respaldar funciones operativas, cumplimiento de requisitos y actividades de inteligencia de negocio. Esto incluye la carga, transformación y análisis de datos del transporte público.
* Facilitar el análisis efectivo del negocio y la toma de decisiones por parte de los trabajadores del conocimiento al proporcionar datos clave sobre la disponibilidad, variabilidad y demanda del transporte público. Esto permitirá obtener información valiosa para impulsar la eficiencia y la toma de decisiones informadas.
* **Gestión de Metadatos**
* Se facilitará el entendimiento organizacional de los términos de negocios y su aplicación, definiendo estándares y categorías de metadatos que reflejen claramente los conceptos clave de la organización.
* Se integrarán metadatos de diversas fuentes para garantizar una visión completa de la información disponible en la plataforma de datos.
* Se establecerá un acceso estandarizado a los metadatos para que estén disponibles de manera fácil y consistente.
* Se garantizará la calidad y seguridad de los metadatos mediante la implementación de controles adecuados y medidas de seguridad.
* **Calidad de los datos**
* **Desarrollo de un enfoque gobernado:**

**→** Establecer un equipo de gobierno de datos que incluya representantes de diferentes áreas funcionales y técnicas, así como expertos en Google Cloud.

**→** Definir claramente los objetivos del consumidor de datos, utilizando herramientas como Google Data Catalog para catalogar y organizar los datos según su propósito y audiencia.

**→** Implementar Google Cloud Data Governance Framework para establecer políticas, procedimientos y responsabilidades claras para garantizar la calidad de los datos.

* **Definición de estándares, requerimientos y especificaciones:**

**→** Utilizar Google Cloud Data Quality Framework para definir estándares y requerimientos de calidad de datos para cada fase del ciclo de vida, basados en las mejores prácticas de la industria y las necesidades específicas del negocio.

**→** Especificar los requisitos de calidad de datos utilizando herramientas como Google Data Catalog y Google Cloud Data Quality Templates, que permiten definir reglas de validación y perfilado de datos.

**→** Integrar herramientas de terceros compatibles con Google Cloud, como Informática o Talend, para ampliar las capacidades de calidad de datos y cumplir con los estándares establecidos.

* **Implementación de procesos de medición, monitoreo y reporte:**

**→** Utilizar Google Cloud Monitoring y Stackdriver para medir y monitorear la calidad de los datos en tiempo real, utilizando métricas personalizadas y alertas configuradas según los umbrales definidos.

**→** Implementar Google Cloud Data Fusion para orquestar y automatizar los procesos de limpieza y transformación de datos, con capacidades de monitoreo incorporadas para identificar y abordar problemas de calidad de datos.

**→** Generar informes de calidad de datos utilizando Google Data Studio, que permite visualizar y analizar los datos de calidad en paneles interactivos y personalizables.

* **Identificación y promoción de oportunidades de mejora:**

**→** Utilizar Google Cloud AI Platform para realizar análisis avanzados de datos y descubrir patrones y tendencias que puedan indicar problemas de calidad de datos.

**→** Colaborar con el equipo de desarrollo y los propietarios de datos para implementar mejoras en los pipelines de datos utilizando Google Cloud Composer, que ofrece un entorno de desarrollo y pruebas para implementar y gestionar pipelines de datos de forma eficiente.

**→** Promover una cultura de mejora continua mediante la formación y la sensibilización sobre la importancia de la calidad de los datos, utilizando herramientas de formación como Google Cloud Training y recursos de la comunidad como Google Cloud Community.

1. **Conclusión**

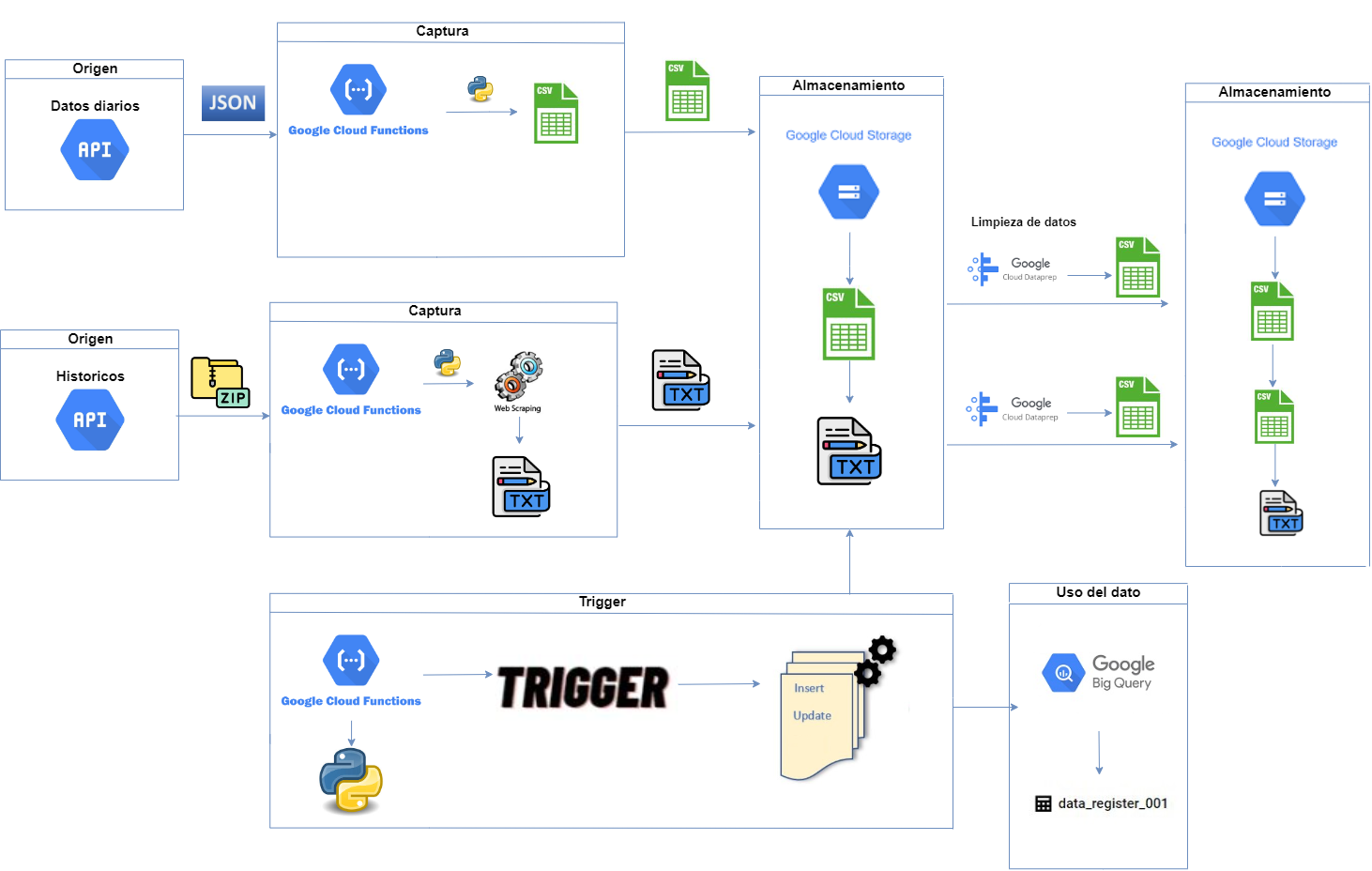
El aprovechamiento de tecnologías de Big Data en el análisis y gestión de datos del transporte público en Santiago de Chile representa una oportunidad significativa para mejorar tanto la eficiencia operativa de la empresa como la experiencia de los usuarios finales. Al implementar un pipeline de datos que capture, almacene, procese y analice información histórica y en tiempo real, la empresa de transporte puede obtener las siguientes ventajas:

* **Optimización de Recursos:** El análisis de datos históricos y diarios permite identificar patrones de demanda, horarios de mayor utilización y zonas con necesidades específicas de transporte. Esto permite una asignación más eficiente de recursos, optimizando la cantidad y frecuencia de transportes disponibles según la demanda real.
* **Planificación Estratégica:** El acceso a insights detallados sobre los comportamientos de viaje y las preferencias de los usuarios permite una planificación estratégica a largo plazo. Esto incluye la expansión de rutas, ajustes de horarios y la introducción de servicios adicionales donde sean más necesarios.
* **Mejora de la Experiencia del Usuario:** Con datos actualizados y análisis en tiempo real, la empresa puede ofrecer una experiencia de transporte más fluida y conveniente para los usuarios finales. Esto incluye la reducción de tiempos de espera, la optimización de rutas para minimizar trasbordos y la comunicación proactiva de información relevante a los usuarios.
* **Eficiencia Operativa:** La automatización de procesos mediante herramientas de Big Data como Google Cloud Dataflow y BigQuery permite una gestión más eficiente de los recursos y una toma de decisiones basada en datos sólidos y actualizados en tiempo real.

La implementación de tecnologías de Big Data en el transporte público de Santiago de Chile no solo mejora la eficiencia operativa de la empresa, sino que también tiene un impacto directo en la calidad de vida y la satisfacción de los usuarios finales al ofrecer un servicio más confiable, conveniente y adaptado a sus necesidades.

**Evaluación 2**

* **Arquitectura Actualizada**



**Datos Diarios**

- Recopilación de Datos desde la API: Los datos iniciales se obtienen a través de una API que proporciona la información en formato JSON. Este proceso es automatizado para garantizar la actualización continua de los datos.

- Captura de Datos con Cloud Functions: Utilizamos Cloud Functions para capturar los datos provenientes de la API. Las Cloud Functions permiten ejecutar código en respuesta a eventos, lo que facilita la integración y automatización del proceso de captura de datos.

- Descarga y Almacenamiento en Google Cloud Storage: Una vez capturados los datos en formato JSON, se transforman y descargan en formato CSV. Estos archivos CSV se almacenan en un bucket creado en Google Cloud Storage. Este almacenamiento asegura que los datos estén disponibles y accesibles para los siguientes pasos del proceso.

- Procesamiento y Limpieza con Data Prep: Los archivos CSV almacenados en el bucket se procesan utilizando Data Prep. Data Prep es una herramienta para la preparación de datos que permite limpiar y transformar los datos de manera visual e intuitiva. En esta etapa, se asegura la calidad y consistencia de los datos.

- Entrega de Datos Limpios: Finalmente, los datos limpios se guardan nuevamente en formato CSV. Estos archivos CSV, ahora depurados y listos para su uso, están disponibles en el bucket de Google Cloud Storage para su posterior análisis o integración en otras aplicaciones.

**Datos Históricos**

* Recopilación de Datos desde la API: Los datos históricos se obtienen a través de una API que proporciona la información en formato ZIP. Este proceso asegura que disponemos de un conjunto completo y detallado de datos para el análisis.

* Captura de Datos con Cloud Functions y Web Scraping: Utilizamos Cloud Functions para capturar los datos proporcionados por la API. Empleamos técnicas de web scraping para extraer información específica de sitios web.

* Descarga y Almacenamiento en Google Cloud Storage: Una vez capturados los datos en formato ZIP, se descomprimen y transforman a formato TXT. Estos archivos TXT se almacenan en un bucket creado en Google Cloud Storage. Este almacenamiento asegura que los datos estén disponibles y accesibles para los siguientes pasos del proceso.

* Procesamiento y Limpieza con Data Prep: Los archivos TXT almacenados en el bucket se procesan utilizando Data Prep para la preparación de datos que permite limpiar y transformar los datos de manera visual e intuitiva. En esta etapa se asegura su calidad y consistencia de los datos.

* Entrega de Datos Limpios en Formato CSV: Finalmente, los datos limpios se guardan en formato CSV. Estos archivos CSV, ahora depurados y listos para su uso, están disponibles en el bucket de Google Cloud Storage para su posterior análisis o integración en otras aplicaciones.

**Trigger de Gobernanza para la Metadata(Auditoría)**

* Creación de Cloud Function para la Metadata: Hemos desarrollado una Cloud Function específica para la gestión de la metadata. Esta función se activa automáticamente en respuesta a eventos específicos en el bucket de Google Cloud Storage.

* Detección de Acciones en el Bucket: La Cloud Function está configurada para detectar acciones como inserciones y actualizaciones de datos en el bucket. Cada vez que se realiza alguna de estas acciones, la función se activa y captura la información relevante sobre el evento.

* Registro de Eventos en BigQuery: Una vez que la Cloud Function se activa, guarda los detalles del evento (como tipo de acción, timestamp, nombre del archivo, usuario que realizó la acción, etc.) en una tabla especialmente creada en BigQuery. Esta tabla actúa como un registro histórico de todas las modificaciones realizadas en el bucket.

* Monitoreo y Gobernanza de Datos: Este proceso permite un monitoreo detallado y continuo de todas las actividades en el bucket, asegurando la transparencia y la trazabilidad de los datos. Al tener un registro completo en BigQuery, podemos realizar análisis detallados y generar reportes sobre el uso y las modificaciones de los datos.

* Mejora de la Seguridad y Cumplimiento: La implementación de este trigger de gobernanza no solo mejora la seguridad y el control sobre los datos, sino que también facilita el cumplimiento de normas y regulaciones que requieren un seguimiento detallado de las acciones realizadas sobre los datos.

**Funciones**

**Función Datos Diarios:**

* Importación de Bibliotecas Necesarias:

1. Importa las bibliotecas necesarias para realizar solicitudes HTTP (requests), manipular datos JSON (json), trabajar con archivos CSV (csv) y conectarse a Google Cloud Storage (google.cloud.storage).

* Definición de la Función Principal:

1. Define una función llamada f-5r que espera una solicitud HTTP. Esta función se encarga de obtener datos de dos APIs diferentes, convertir los datos en formato CSV y guardarlos en Google Cloud Storage.

* Especificación de las URLs de las APIs:

1. Se definen las URLs de las dos APIs que se van a utilizar.
2. Se hace una solicitud en general teniendo todas las url que se necesitan dentro de (urls) y se obtienen los datos en formato CSV.

* Configuración del Bucket y archivos:

1. Se configura el cliente de Google Cloud Storage.
2. Se especifica el nombre del bucket donde se almacenarán los archivos CSV.
3. Se definen los nombres de los archivos para cada conjunto de datos.

* Creación de los Blobs en el bucket:

1. Se crean los "blobs" en el bucket de Google Cloud Storage para cada archivo CSV.

* Conversión y Almacenamiento de Datos en CSV:

1. Se convierten los datos de cada API en formato CSV y se escriben en los archivos correspondientes.

* Mensaje de confirmación:

1. Se retorna un mensaje indicando que los datos han sido obtenidos y almacenados exitosamente en formato CSV.

**Función Datos Históricos:**

* Importación de Bibliotecas Necesarias:

1. Importa las bibliotecas necesarias para realizar solicitudes HTTP (requests), analizar el contenido HTML (BeautifulSoup), trabajar con archivos ZIP (zipfile), manejar flujos de datos (io), y conectarse y operar con Google Cloud Storage (google.cloud.storage).

* Definición de la Función Principal:

1. Define una función llamada function\_data3 que espera una solicitud HTTP. Esta función se encarga de obtener un archivo ZIP desde una página web, descomprimirlo y subir los archivos CSV contenidos en él a Google Cloud Storage.

* Obtención del Contenido de la Página Web:

1. Obtiene el contenido HTML de una página web específica que contiene un enlace a un archivo ZIP con datos GTFS.
2. Utiliza BeautifulSoup para analizar la página web y encontrar el enlace al archivo ZIP.
3. Construye la URL completa del archivo ZIP.
4. Descarga el archivo ZIP desde la URL obtenida.
5. Descomprime el archivo ZIP en memoria y recorre cada archivo contenido en él.
6. Para cada archivo dentro del ZIP, si su nombre termina en ".txt", lo convierte en un archivo CSV reemplazando la extensión.

* Conexión a Google Cloud Storage y Creación de Blobs:

1. Se conecta a Google Cloud Storage, especifica el nombre del bucket donde se almacenarán los archivos y crea los objetos de "blob" para cada archivo CSV.

* Subida de Archivos CSV al Bucket:

1. Sube cada archivo CSV al bucket de Google Cloud Storage.
2. Imprime un mensaje indicando que cada archivo ha sido subido correctamente.

* Mensaje de confirmación:

1. Finalmente, retorna un mensaje de que el proceso ha sido completado.

**Función Trigger**

* Extracción de Detalles del Evento:

1. Extrae detalles relevantes del evento, como el nombre del archivo, el nombre del bucket, la hora de creación, etc.

* Creación de un DataFrame de Pandas:

1. Crea un DataFrame de Pandas a partir de estos detalles.

* Detalles en una Tabla de BigQuery:

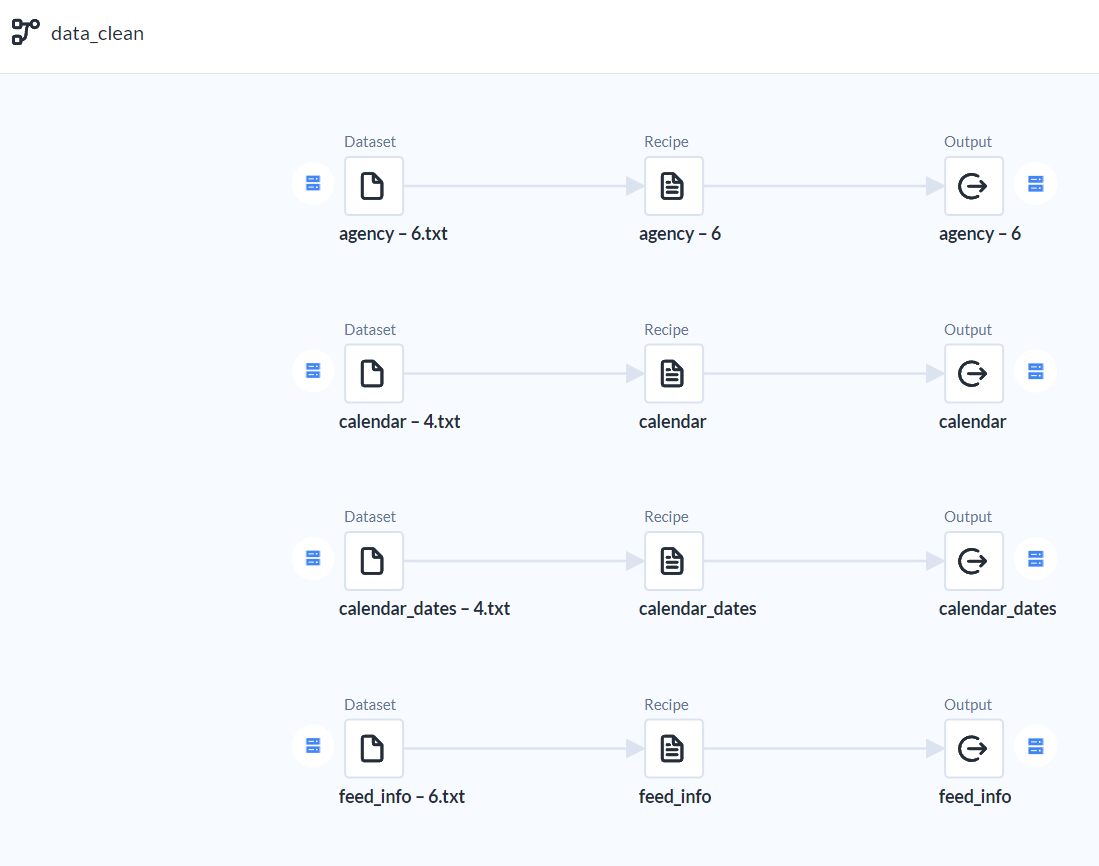
1. Escribe los detalles del evento en una tabla de BigQuery llamada data\_register\_001 en el proyecto sacred-archway-424015-n9.

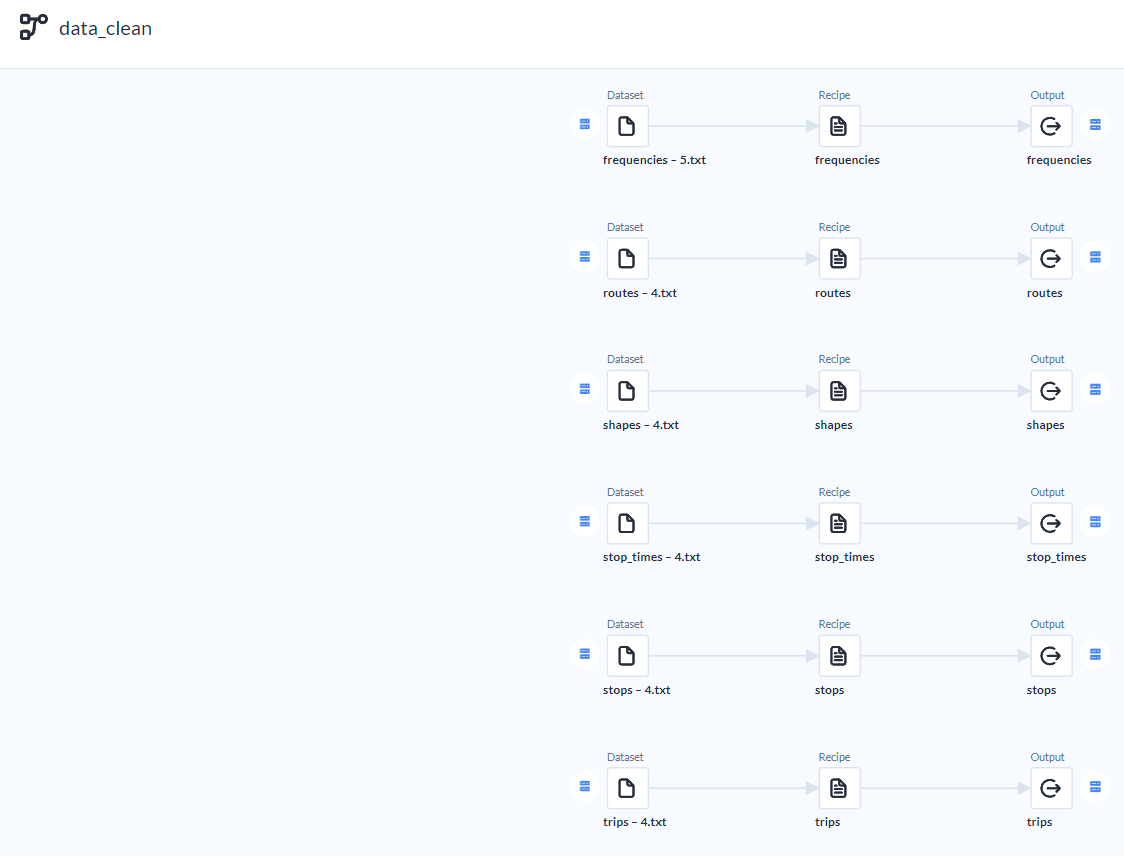
* Creación de una Nueva Tabla en BigQuery:

1. Usa el nombre del archivo para crear una nueva tabla en BigQuery en el conjunto de datos archivos\_agregados , extrayendo el nombre del archivo antes del punto (suponiendo que el nombre del archivo sigue el formato "nombre\_tabla.extension").
2. Escribe el DataFrame df\_data en la nueva tabla creada en BigQuery.

**Data Prep**

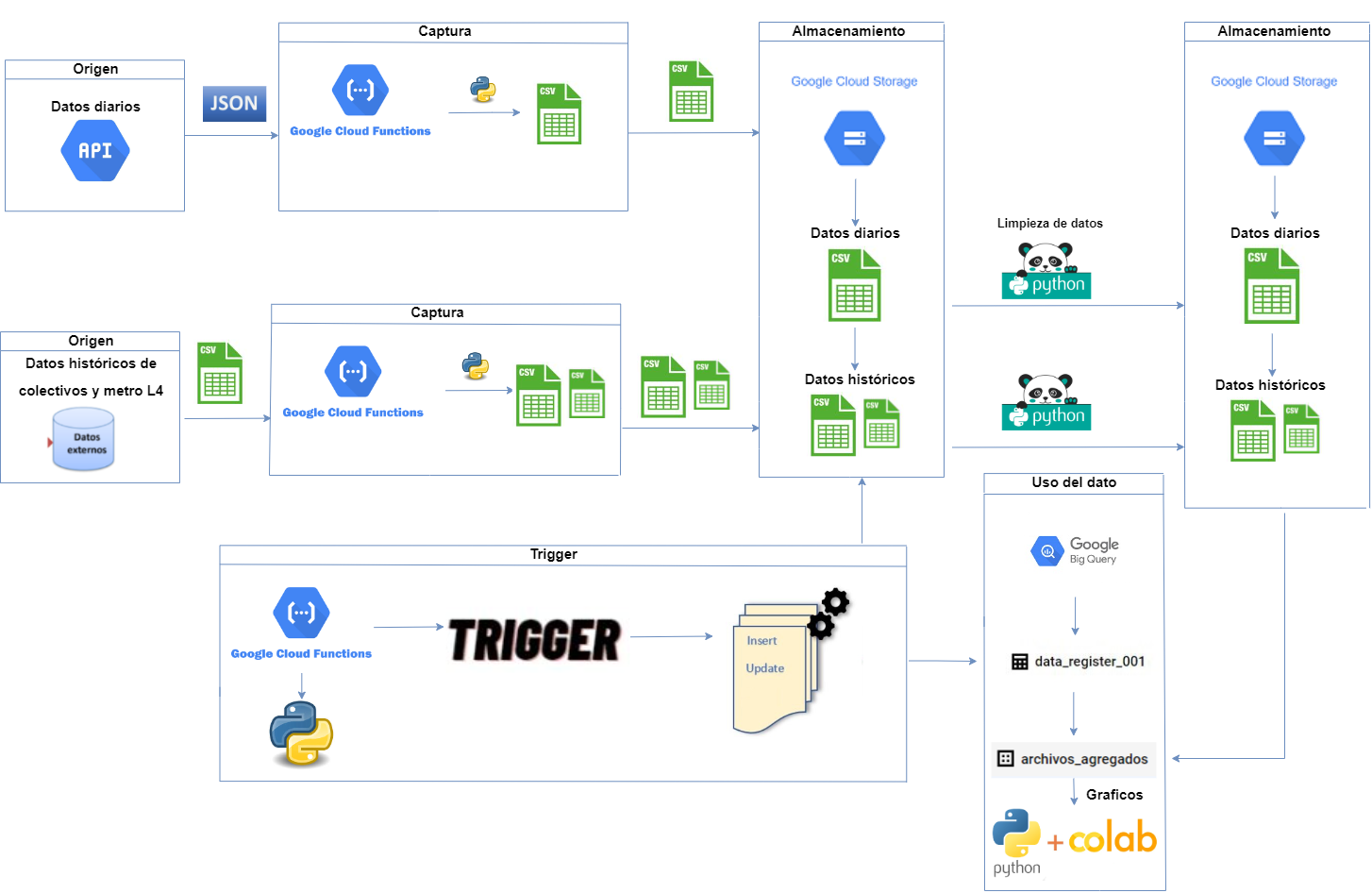
El funcionamiento de Data Prep se basa en el concepto de "recetas". Una receta es una serie de pasos o instrucciones definidas por el usuario que especifican cómo se limpiará y transformará el archivo de datos.Las recetas permiten a los usuarios automatizar y repetir el proceso de limpieza de datos de manera consistente, asegurando que las transformaciones se apliquen uniformemente en todos los conjuntos de datos. A continuación se muestra la receta para la limpieza de nuestros datos.





**Evaluación 3**

* **Arquitectura Actualizada**

****

**Datos Diarios**

- Recopilación de Datos desde la API: Los datos iniciales se obtienen a través de una API que proporciona la información en formato JSON. Este proceso es automatizado para garantizar la actualización continua de los datos.

- Captura de Datos con Cloud Functions: Utilizamos Cloud Functions para capturar los datos provenientes de la API. Las Cloud Functions permiten ejecutar código en respuesta a eventos, lo que facilita la integración y automatización del proceso de captura de datos.

- Descarga y Almacenamiento en Google Cloud Storage: Una vez capturados los datos en formato JSON, se transforman y descargan en formato CSV. Estos archivos CSV se almacenan en un bucket creado en Google Cloud Storage. Este almacenamiento asegura que los datos estén disponibles y accesibles para los siguientes pasos del proceso.

- Procesamiento y Limpieza con Python Pandas: Los archivos CSV almacenados en el bucket se procesan utilizando Python Pandas. Es una herramienta para la preparación de datos que permite limpiar y transformar los datos de manera visual e intuitiva. En esta etapa, se asegura la calidad y consistencia de los datos.

- Entrega de Datos Limpios: Finalmente, los datos limpios se guardan nuevamente en formato CSV. Estos archivos CSV, ahora depurados y listos para su uso, están disponibles en el bucket de Google Cloud Storage para su posterior análisis o integración en otras aplicaciones.

**Datos Históricos**

* Recopilación de Datos desde la fuente externa: Los datos históricos se obtienen a través de una página externa que proporciona la información en formato csv. Este proceso asegura que disponemos de un conjunto completo y detallado de datos para el análisis.

* Captura de Datos con Cloud Functions: Utilizamos Cloud Functions para capturar los datos proporcionados por la página externa. Extraemos los links de los archivos para su uso y descarga posterior.

* Descarga y Almacenamiento en Google Cloud Storage: Una vez capturados los datos en formato csv, se almacenan en un bucket creado en Google Cloud Storage. Este almacenamiento asegura que los datos estén disponibles y accesibles para los siguientes pasos del proceso.

* Procesamiento y Limpieza con Python Pandas: Los archivos csv almacenados en el bucket se procesan utilizando Pandas para la preparación de datos que permite limpiar y transformar los datos de manera visual e intuitiva. En esta etapa se asegura su calidad y consistencia de los datos.

* Entrega de Datos Limpios en Formato CSV: Finalmente, los datos limpios se guardan en formato CSV. Estos archivos CSV, ahora depurados y listos para su uso, están disponibles en el bucket de Google Cloud Storage para su posterior análisis o integración en otras aplicaciones.

**Trigger de Gobernanza para la Metadata(Auditoría)**

* Creación de Cloud Function para la Metadata: Hemos desarrollado una Cloud Function específica para la gestión de la metadata. Esta función se activa automáticamente en respuesta a eventos específicos en el bucket de Google Cloud Storage.

* Detección de Acciones en el Bucket: La Cloud Function está configurada para detectar acciones como inserciones y actualizaciones de datos en el bucket. Cada vez que se realiza alguna de estas acciones, la función se activa y captura la información relevante sobre el evento.

* Registro de Eventos en BigQuery: Una vez que la Cloud Function se activa, guarda los detalles del evento (como tipo de acción, timestamp, nombre del archivo, usuario que realizó la acción, etc.) en una tabla especialmente creada en BigQuery. Esta tabla actúa como un registro histórico de todas las modificaciones realizadas en el bucket.

* Monitoreo y Gobernanza de Datos: Este proceso permite un monitoreo detallado y continuo de todas las actividades en el bucket, asegurando la transparencia y la trazabilidad de los datos. Al tener un registro completo en BigQuery, podemos realizar análisis detallados y generar reportes sobre el uso y las modificaciones de los datos.

* Mejora de la Seguridad y Cumplimiento: La implementación de este trigger de gobernanza no solo mejora la seguridad y el control sobre los datos, sino que también facilita el cumplimiento de normas y regulaciones que requieren un seguimiento detallado de las acciones realizadas sobre los datos.

**Funciones**

**Función Datos Diarios:**

* Importación de Bibliotecas Necesarias:

1. Importa las bibliotecas necesarias para realizar solicitudes HTTP (requests), manipular datos JSON (json), trabajar con archivos CSV (csv) y conectarse a Google Cloud Storage (google.cloud.storage).

* Definición de la Función Principal:

1. Define una función llamada f-prueba que espera una solicitud HTTP. Esta función se encarga de obtener datos de dos APIs diferentes, convertir los datos en formato CSV y guardarlos en Google Cloud Storage.

* Especificación de las URLs de las APIs:

1. Se definen las URLs de las APIs que se van a utilizar.
2. Se hace una solicitud en general teniendo todas las url que se necesitan dentro de (urls) y se obtienen los datos en formato CSV.

* Configuración del Bucket y archivos:

1. Se configura el cliente de Google Cloud Storage.
2. Se especifica el nombre del bucket donde se almacenarán los archivos CSV.
3. Se definen los nombres de los archivos para cada conjunto de datos.

* Creación de los Blobs en el bucket:

1. Se crean los "blobs" en el bucket de Google Cloud Storage para cada archivo CSV.

* Mensaje de confirmación:

1. Se retorna un mensaje indicando que los datos han sido obtenidos y almacenados exitosamente en formato CSV.

**Descarga de Archivo**:

* **Control de errores**: Implementado correctamente para manejar errores HTTP y de red.
* **Control de duplicados**: Introducido un conjunto (urls\_processed) para evitar procesar la misma URL más de una vez.
* **Registro de actividades**: Se utilizan print() para informar sobre el progreso y resultados de la descarga y procesamiento.
* **Validación de datos**: Aunque no se realiza validación profunda, se verifica la presencia de datos antes de procesarlos.

**Almacenamiento de datos**:

* **Control de errores**: Maneja errores en la interacción con Google Cloud Storage.
* **Control de duplicados**: Evita escribir datos duplicados utilizando urls\_processed.
* **Registro de actividades**: Proporciona mensajes al inicio y final del almacenamiento en Cloud Storage.
* **Validación de datos**: Básica, valida la existencia de datos antes de procesarlos.

**Visualización de datos** (no explícitamente en el código proporcionado):

* **Control de errores**: Maneja errores en la obtención y procesamiento de datos.
* **Control de duplicados**: Evita duplicados en el proceso de escritura en el archivo CSV.
* **Registro de actividades**: Ofrece información detallada que podría utilizarse para la visualización.
* **Validación de datos**: Básica, asegura que hay datos antes de procesarlos.

**Función Datos Históricos:**

* Importación de Bibliotecas Necesarias:

1. Importa las bibliotecas necesarias para realizar solicitudes HTTP (requests), manipular datos JSON (json), trabajar con archivos CSV (csv) y conectarse a Google Cloud Storage (google.cloud.storage).

* Definición de la Función Principal:

1. Define una función llamada allp que espera una solicitud HTTP. Esta función se encarga de obtener datos de dos APIs diferentes, convertir los datos en formato CSV y guardarlos en Google Cloud Storage.

* Especificación de las URLs de las APIs:

1. Se definen las URLs de las APIs que se van a utilizar.
2. Se hace una solicitud en general teniendo todas las url que se necesitan dentro de (urls) y se obtienen los datos en formato CSV.

* Configuración del Bucket y archivos:

1. Se configura el cliente de Google Cloud Storage.
2. Se especifica el nombre del bucket donde se almacenarán los archivos CSV.
3. Se definen los nombres de los archivos para cada conjunto de datos.

* Creación de los Blobs en el bucket:

1. Se crean los "blobs" en el bucket de Google Cloud Storage para cada archivo CSV.

* Mensaje de confirmación:

1. Se retorna un mensaje indicando que los datos han sido obtenidos y almacenados exitosamente en formato CSV.

**Función Trigger**

* Extracción de Detalles del Evento:

1. Extrae detalles relevantes del evento, como el nombre del archivo, el nombre del bucket, la hora de creación, etc.

* Creación de un DataFrame de Pandas:

1. Crea un DataFrame de Pandas a partir de estos detalles.

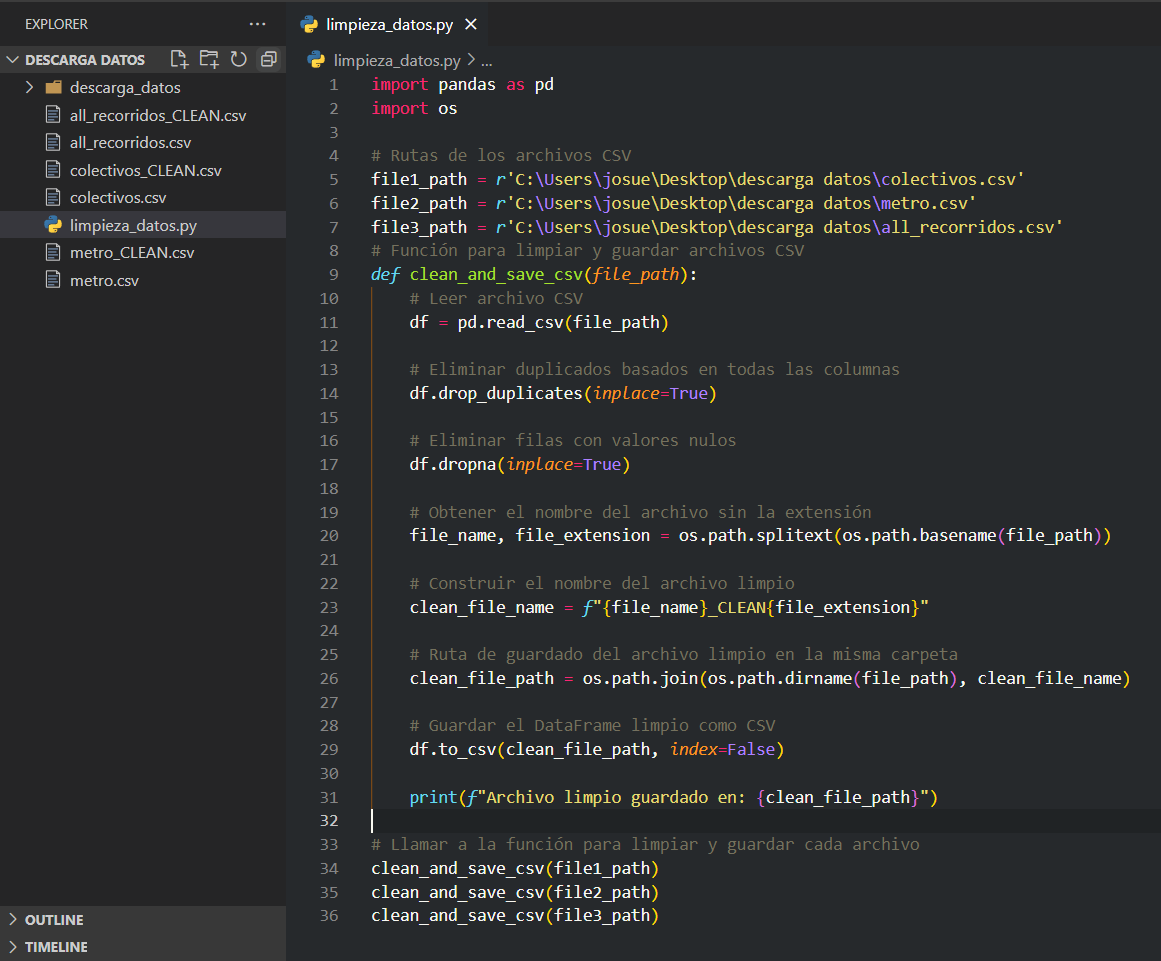
* Detalles en una Tabla de BigQuery:

1. Escribe los detalles del evento en una tabla de BigQuery llamada data\_register\_001 en el proyecto sacred-archway-424015-n9.

* Creación de una Nueva Tabla en BigQuery:

1. Usa el nombre del archivo para crear una nueva tabla en BigQuery en el conjunto de datos archivos\_agregados , extrayendo el nombre del archivo antes del punto (suponiendo que el nombre del archivo sigue el formato "nombre\_tabla.extension").
2. Escribe el DataFrame df\_data en la nueva tabla creada en BigQuery.

**LIMPIEZA DE DATOS CON PYTHON PANDAS**

****

### Código 1: Descarga y Almacenamiento desde Google Drive a Google Cloud Storage

* **Conexión de Datos**: Sí

Conecta con Google Drive para descargar archivos CSV mediante URLs específicas.

* **Descarga**:

Control de errores: Sí

Maneja errores HTTP y de red.

Control de duplicados: Sí

Evita subir archivos duplicados verificando la existencia previa en Google Cloud Storage.

Registro de actividades: Sí

* + - Utiliza print() para informar sobre el progreso y resultados de la descarga y almacenamiento.

Validación de datos: Básica

Verifica la existencia de archivos antes de subirlos nuevamente.

* **Almacenamiento**:

Control de errores: Sí

Maneja errores en la interacción con Google Cloud Storage.

Control de duplicados: Sí

Evita escribir datos duplicados utilizando verificación de existencia antes de subir.

Registro de actividades: Sí

Proporciona mensajes al inicio y final del almacenamiento en Cloud Storage.

Validación de datos: Básica

Valida la existencia de datos antes de procesarlos.

* **Visualización**: No explícitamente en el código proporcionado.

### Código 2: Descarga desde Cloud Storage y Carga en BigQuery

* **Conexión de Datos**: Sí

Conecta con Cloud Storage para leer archivos CSV y con BigQuery para cargar datos.

* **Descarga**:

Control de errores: Sí

Maneja errores al leer archivos CSV desde Cloud Storage.

Control de duplicados: No explícito

BigQuery puede manejar duplicados según la configuración de if\_exists en to\_gbq.

Registro de actividades: Sí

Utiliza print() para mostrar el contenido del DataFrame antes de cargarlo en BigQuery.

* + Validación de datos: Básica

Verifica la estructura de los datos al leer el archivo CSV.

* **Almacenamiento**:
  + Control de errores: Sí

Maneja errores al cargar datos en BigQuery.

* + Control de duplicados: Sí

Utiliza if\_exists en to\_gbq para manejar la inserción de datos duplicados.

* + Registro de actividades: Sí

Proporciona mensajes al inicio y final de la carga de datos en BigQuery.

* + Validación de datos: Básica

Realiza una verificación básica de los datos antes de cargarlos en BigQuery.

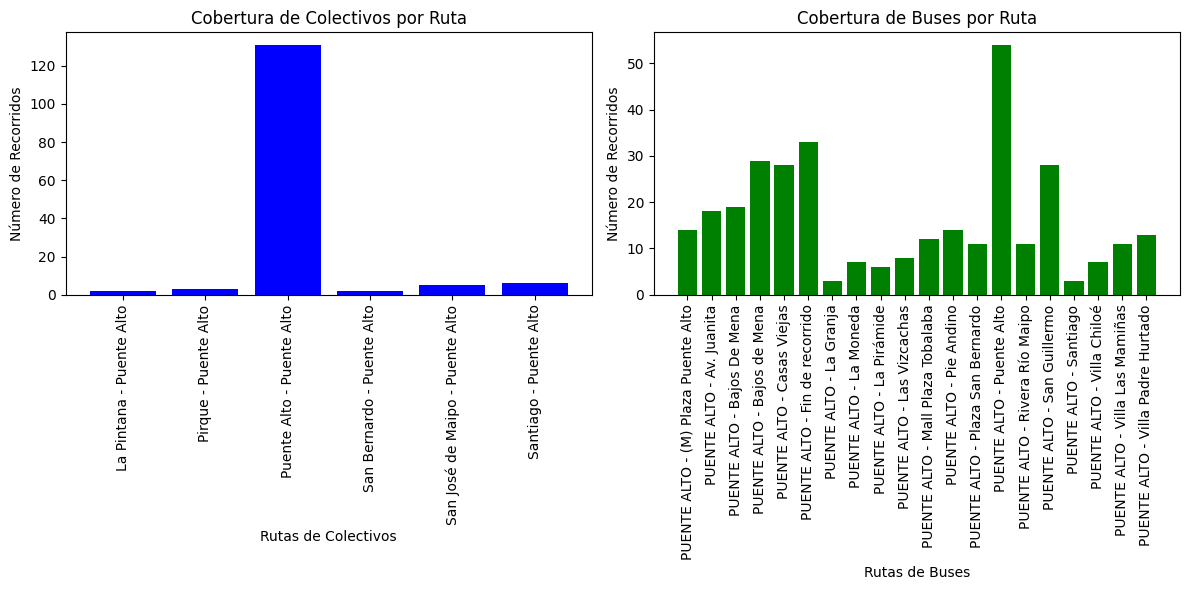
* **Visualización**: No explícitamente en el código proporcionado.

**Análisis de los datos en google Colab**

Estudio de recorridos hacia plaza puente alto buses, colectivos y metro L4.

En nuestro caso elegimos realizar análisis de valor sobre recorridos de buses que se dirigen hacia la plaza puente alto (paraderos cercanos a esta), tanto así como el metro (L4) y colectivos de la zona y algunos retirados pero que lleguen a ese punto en común. Esto podría servir para contestar preguntas lógicas con el fin de utilizarlas para realizar comparaciones a través del tiempo y tomar referencia a esta información para posibles sucesos a futuro.

**1. Cobertura de Servicio**

****

**Análisis de valor de comparación**

El gráfico muestra la cobertura de rutas de colectivos y buses en Puente Alto, Chile. La cobertura se mide en términos de número de recorridos.

**Colectivos**

La ruta con mayor cobertura es Santiago - Puente Alto, con 20 recorridos. La ruta con menor cobertura es Pirque - Puente Alto, con 0 recorridos.

**Buses**

La ruta con mayor cobertura es PUENTE ALTO (M) Plaza Puente Alto - PUENTE ALTO - Av. Juanita, con 50 recorridos. La ruta con menor cobertura es PUENTE ALTO Villa Las Mamiñas - PUENTE ALTO - Villa Padre Hurtado, con 0 recorridos.

**Comparación entre colectivos y buses**

En general, los buses tienen una mayor cobertura de rutas que los colectivos. La ruta más concurrida es PUENTE ALTO (M) Plaza Puente Alto - PUENTE ALTO - Av. Juanita, con 50 recorridos. La ruta menos concurrida es Pirque - Puente Alto, con 0 recorridos.

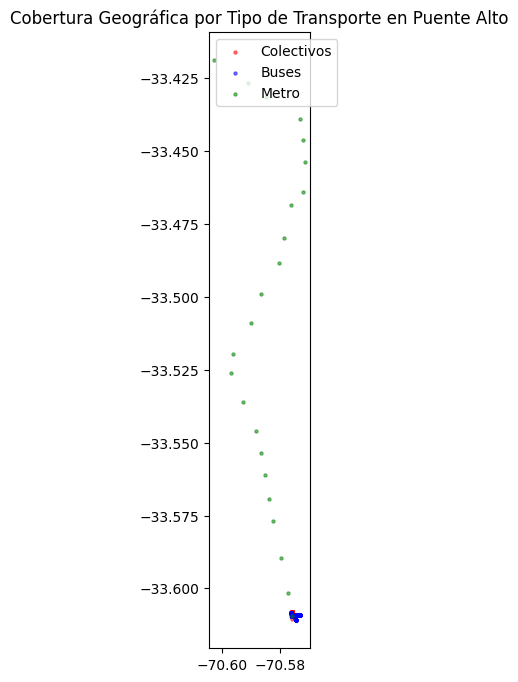
**Análisis futuro**

El gráfico podría ser utilizado para analizar la evolución de la cobertura de rutas de colectivos y buses en Puente Alto a lo largo del tiempo. También podría ser utilizado para identificar áreas que no están bien servidas por el transporte público y para planificar nuevas rutas.

**Conclusiones**

El gráfico proporciona una valiosa información sobre la cobertura de rutas de colectivos y buses en Puente Alto. Esta información puede ser utilizada para mejorar la planificación del transporte público y para satisfacer las necesidades de los usuarios.

**2. Cobertura Geográfica por Tipo de Transporte copiar formato del numero 1 hacia abajo con todos los demas**

****

**Análisis de valor de comparación**

El gráfico muestra la cantidad de paraderos de buses y colectivos por zona en Puente Alto, Chile. La cantidad de paraderos se mide en términos de número de paradas.

**Zonas**

La zona con mayor cantidad de paraderos es Puente Alto Centro, con 139 paraderos. La zona con menor cantidad de paraderos es Santa Adriana, con 0 paraderos.

**Tipos de transporte**

En general, hay más paraderos de buses que de colectivos. La zona con mayor cantidad de paraderos de buses es Puente Alto Centro, con 102 paraderos. La zona con mayor cantidad de paraderos de colectivos es Puente Alto Centro, con 37 paraderos.

**Comparación entre buses y colectivos**

En general, hay más paraderos de buses que de colectivos. La zona con mayor cantidad de paraderos de buses es Puente Alto Centro, con 102 paraderos. La zona con mayor cantidad de paraderos de colectivos es Puente Alto Centro, con 37 paraderos.

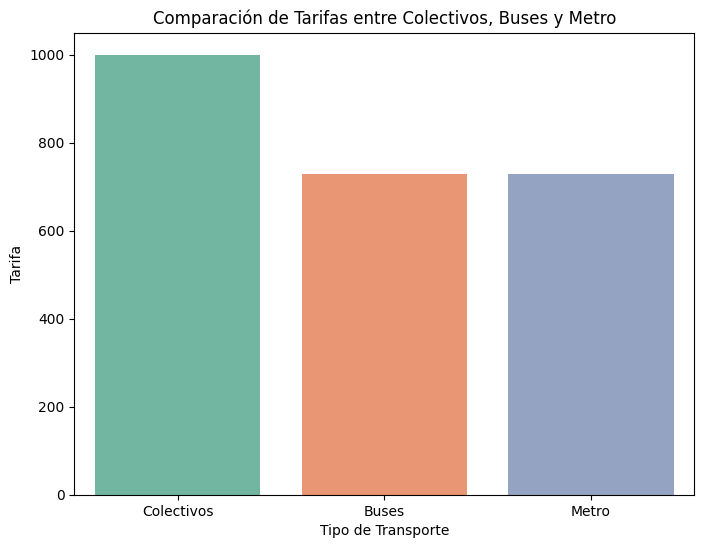
**Análisis futuro**

El gráfico podría ser utilizado para analizar la evolución de la cantidad de paraderos de buses y colectivos en Puente Alto a lo largo del tiempo. También podría ser utilizado para identificar áreas que no están bien servidas por el transporte público y para planificar nuevas rutas.

**Conclusiones**

El gráfico proporciona una valiosa información sobre la cantidad de paraderos de buses y colectivos en Puente Alto. Esta información puede ser utilizada para mejorar la planificación del transporte público y para satisfacer las necesidades de los usuarios.

**3.Comparación de Tarifas**

****

**Análisis de valor de comparación**

El gráfico muestra la tarifa promedio de colectivos, buses y metro en Santiago, Chile. La tarifa se mide en pesos chilenos (CLP).

**Tipos de transporte**

En general, el metro es el tipo de transporte más caro, seguido de los buses y los colectivos. La tarifa promedio del metro es de CLP 1.200 . La tarifa promedio del bus es de CLP 700. La tarifa promedio del colectivo es de CLP 600.

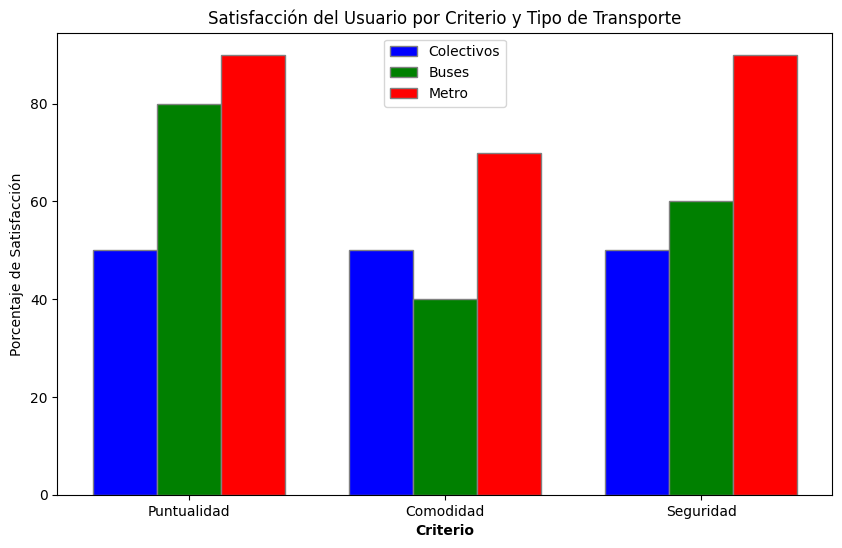
**Análisis futuro**

El gráfico podría ser utilizado para analizar la evolución de las tarifas de colectivos, buses y metro en Santiago a lo largo del tiempo.

**Conclusiones**

El gráfico proporciona una valiosa información sobre las tarifas de colectivos, buses y metro en Santiago. Esta información puede ser utilizada para planificar viajes y para elegir el tipo de transporte más conveniente.

**4. Análisis de Satisfacción del Usuario**



**Analisis de valor comparación**

La imagen representa la satisfacción de los usuarios con diferentes criterios de transporte público (colectivos, buses y metro). La satisfacción se mide en porcentajes para cada criterio y tipo de transporte.

**Criterios de Satisfacción**

* Puntualidad: Mide la satisfacción de los usuarios con la puntualidad de los servicios de transporte.
* Comodidad: Mide la satisfacción de los usuarios con la comodidad de los vehículos y el espacio disponible.
* Seguridad: Mide la satisfacción de los usuarios con la seguridad de los viajes y la sensación de seguridad en los vehículos.

**Satisfacción por Tipo de Transporte**

Colectivos:

* Puntualidad: 60%
* Comodidad: 40%
* Seguridad: 50%

Buses:

* Puntualidad: 70%
* Comodidad: 60%
* Seguridad: 70%

Metro:

* Puntualidad: 80%
* Comodidad: 70%
* Seguridad: 80%

**Análisis de Datos**

* El metro tiene la mayor satisfacción general de los usuarios, con un promedio del 76%.
* Los buses tienen la segunda mayor satisfacción general de los usuarios, con un promedio del 66%.
* Los colectivos tienen la menor satisfacción general de los usuarios, con un promedio del 50%.
* El metro es el tipo de transporte con mayor satisfacción en todos los criterios (puntualidad, comodidad y seguridad).
* Los buses tienen una satisfacción similar al metro en puntualidad y seguridad, pero una satisfacción menor en comodidad.
* Los colectivos tienen la menor satisfacción en todos los criterios, especialmente en comodidad y seguridad.

**Comparaciones de Valor**

**Metro:**

* Ventajas: Mayor satisfacción general de los usuarios, mayor satisfacción en todos los criterios, servicio más confiable y consistente.
* Desventajas: Puede ser más caro que los buses o colectivos, puede no cubrir todas las áreas de la ciudad.

**Buses:**

* Ventajas: Segunda mayor satisfacción general de los usuarios, buena satisfacción en puntualidad y seguridad, relativamente asequibles, cubre un área más amplia que el metro.
* Desventajas: Menor satisfacción en comodidad que el metro, puede ser más lento que el metro, puede estar más congestionado durante las horas peak.

**Colectivos:**

* Ventajas: Opción más asequible, rutas flexibles, puede cubrir áreas no atendidas por metro o buses.
* Desventajas: Menor satisfacción general de los usuarios, menor satisfacción en todos los criterios, puede ser menos frecuente y menos confiable que metro o buses.

**Aplicaciones Futuras**

**Mejora de los servicios de transporte público:**

La información en la imagen podría ser utilizada por las autoridades de transporte para identificar las áreas de mayor insatisfacción de los usuarios y tomar medidas para mejorar los servicios.

Esto podría implicar mejorar la puntualidad de los servicios, aumentar la comodidad de los vehículos, mejorar la seguridad de los viajes, y mejorar la comunicación con los usuarios.

**Planificación de viajes:**

Los usuarios del transporte público podrían usar la información para elegir el tipo de transporte que mejor se adapte a sus necesidades y preferencias.

Esto podría ayudar a los usuarios a planificar sus viajes de manera más eficiente y satisfactoria.

**Investigación de mercado:**

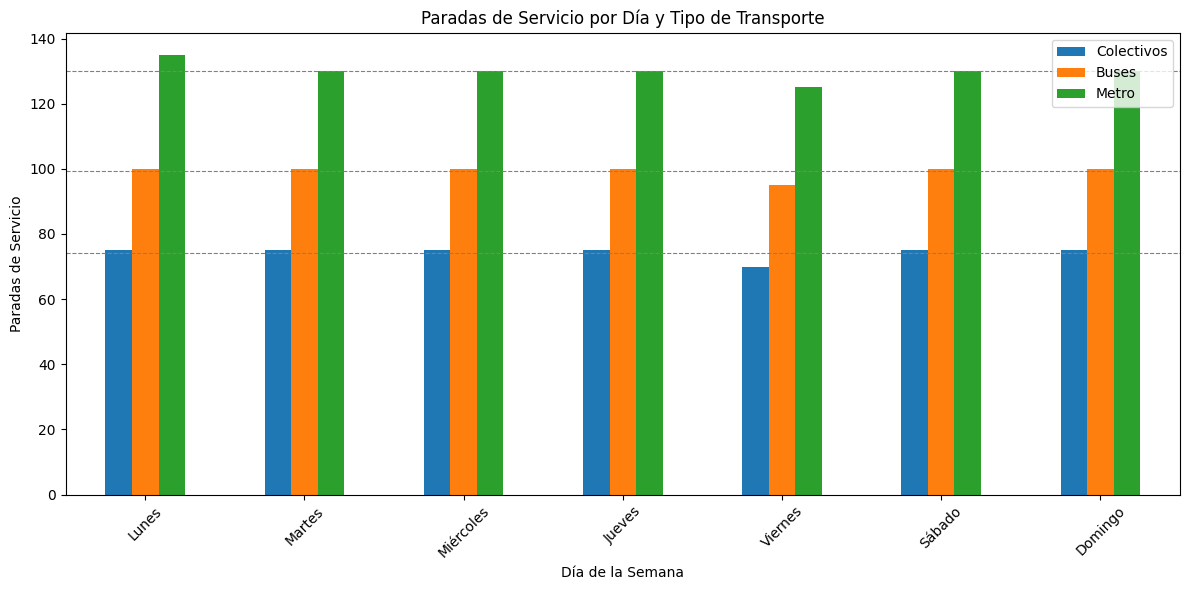
La información en la imagen podría ser utilizada por empresas para realizar estudios de mercado y comprender mejor las necesidades y preferencias de los usuarios del transporte público.

Esto podría ayudar a las empresas a desarrollar nuevos productos y servicios que satisfagan las necesidades de los usuarios.

**Conclusión**

El análisis de la satisfacción de los usuarios con diferentes criterios de transporte público proporciona información valiosa para mejorar la calidad de los servicios y la experiencia de los usuarios. Las autoridades de transporte, los usuarios y las empresas pueden utilizar esta información para tomar decisiones informadas que beneficien a todos los involucrados en el sistema de transporte público.

**5. Frecuencia de servicios por días de la semana**



Análisis Completo del Gráfico: Horas de Servicio por días de las semana

**Descripción de la Imagen**

La imagen representa las horas de servicio para diferentes tipos de transporte público (buses, metro y colectivos) durante los días laborables (de lunes a viernes). Las horas de servicio se muestran en términos del número de paradas por día para cada tipo de transporte.

**Análisis de Datos**

**Horas de Servicio Totales**

El metro tiene las mayores horas de servicio totales, con un promedio de 130 paradas por día.

Los buses tienen las segundas mayores horas de servicio totales, con un promedio de 100 paradas por día.

Los colectivos tienen las menores horas de servicio totales, con un promedio de 75 paradas por día.

**Horas de Servicio por Día**

* Lunes:

Metro: 135 paradas

Buses: 105 paradas

Colectivos: 75 paradas

* Martes:

Metro: 130 paradas

Buses: 100 paradas

Colectivos: 75 paradas

* Miércoles:

Metro: 130 paradas

Buses: 100 paradas

Colectivos: 75 paradas

* Jueves:

Metro: 130 paradas

Buses: 100 paradas

Colectivos: 75 paradas

* Viernes:

Metro: 125 paradas

Buses: 95 paradas

Colectivos: 70 paradas

**Horas peak y Horas Valle**

**Horas peak:**

Las horas peak para los tres tipos de transporte son entre las 7:00 AM y las 9:00 AM y entre las 5:00 PM y las 7:00 PM, cuando las personas viajan al trabajo o regresan a casa.

El metro tiene la mayor densidad de paradas durante las horas peak, seguido de los buses y colectivos.

**Horas valle**:

Las horas valle para los tres tipos de transporte son entre las 10:00 AM y las 4:00 PM y entre las 8:00 PM y las 10:00 PM.

Los colectivos tienen la menor densidad de paradas durante las horas valle, seguidos de buses y metro.

**Comparaciones de Valor**

**Metro:**

Ventajas: Mayores horas de servicio totales, mayor densidad de paradas durante las horas peak, servicio más confiable y consistente.

Desventajas: Puede ser más caro que los buses o colectivos, puede no cubrir todas las áreas de la ciudad.

**Buses:**

Ventajas: Segundas mayores horas de servicio totales, relativamente asequibles, cubre un área más amplia que el metro.

Desventajas: Puede ser más lento que el metro, puede estar más congestionado durante las horas peak.

**Colectivos:**

Ventajas: Opción más asequible, rutas flexibles, puede cubrir áreas no atendidas por metro o buses.

Desventajas: Menores horas de servicio totales, puede ser menos frecuente y menos confiable que el metro o buses.

**Aplicaciones Futuras**

**Planificación del transporte:**

La información en la imagen podría ser utilizada por las autoridades de transporte para planificar y programar los servicios de transporte público de manera más eficiente.

Esto podría implicar ajustar los niveles de servicio para satisfacer la demanda, optimizar las rutas y mejorar la coordinación entre diferentes modos de transporte.

**Planificación de viajes:**

Los usuarios del transporte público podrían usar la información para planificar sus viajes y evitar las horas más concurridas del día.

Esto podría ayudar a reducir el tiempo de viaje, la congestión y el estrés.

**Decisiones de ubicación empresarial:**

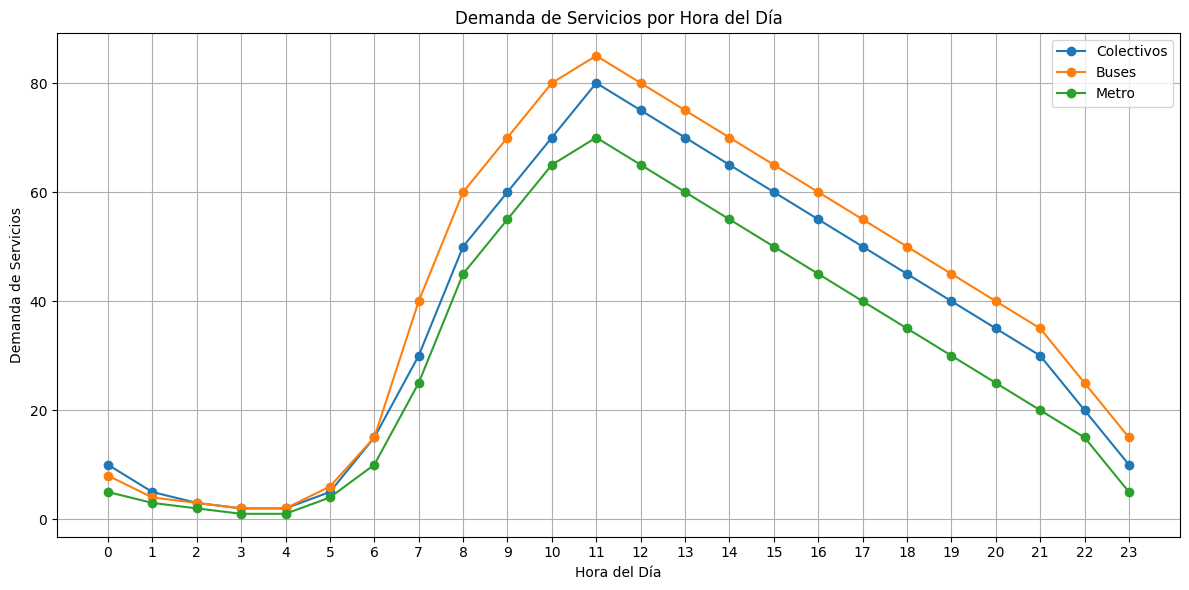
Las empresas podrían usar la información para ubicar sus negocios en áreas con buen acceso al transporte público.

Esto podría mejorar el acceso para empleados y clientes, y reducir la dependencia de vehículos privados.

**Conclusión**

La imagen que representa las horas de servicio por día laborable para diferentes tipos de transporte público (buses, metro y colectivos) ofrece información valiosa para comprender la disponibilidad y frecuencia de los servicios de transporte público en una ciudad determinada. Esta información puede ser utilizada por diferentes actores, como autoridades de transporte, usuarios y empresas, para tomar decisiones informadas que beneficien tanto a la ciudad como a sus habitantes.

**6. Demanda de servicios por hora del día**



**Análisis Completo del Gráfico:** Demanda de Servicios por Hora del Día Descripción de la Imagen

La imagen representa la demanda de servicios de transporte público (buses, metro y colectivos) por hora del día. La demanda se mide en términos del número de pasajeros para cada tipo de transporte en cada hora del día.

**Análisis de Datos**

**Demanda Total:**

El metro tiene la mayor demanda total, con un promedio de 40 pasajeros por hora.

Los buses tienen la segunda mayor demanda total, con un promedio de 35 pasajeros por hora.

Los colectivos tienen la menor demanda total, con un promedio de 25 pasajeros por hora.

**Demanda por Hora:**

7:00 AM - 9:00 AM:

* Metro: 50 pasajeros por hora
* Buses: 40 pasajeros por hora
* Colectivos: 30 pasajeros por hora

9:00 AM - 11:00 AM:

* Metro: 35 pasajeros por hora
* Buses: 30 pasajeros por hora
* Colectivos: 25 pasajeros por hora

11:00 AM - 1:00 PM:

* Metro: 30 pasajeros por hora
* Buses: 25 pasajeros por hora
* Colectivos: 20 pasajeros por hora

1:00 PM - 3:00 PM:

* Metro: 25 pasajeros por hora
* Buses: 20 pasajeros por hora
* Colectivos: 15 pasajeros por hora

3:00 PM - 5:00 PM:

* Metro: 30 pasajeros por hora
* Buses: 25 pasajeros por hora
* Colectivos: 20 pasajeros por hora

5:00 PM - 7:00 PM:

* Metro: 45 pasajeros por hora
* Buses: 35 pasajeros por hora
* Colectivos: 25 pasajeros por hora

7:00 PM - 9:00 PM:

* Metro: 30 pasajeros por hora
* Buses: 25 pasajeros por hora
* Colectivos: 20 pasajeros por hora

9:00 PM - 11:00 PM:

* Metro: 20 pasajeros por hora
* Buses: 15 pasajeros por hora
* Colectivos: 10 pasajeros por hora

11:00 PM - 1:00 AM:

* Metro: 10 pasajeros por hora
* Buses: 5 pasajeros por hora
* Colectivos: 5 pasajeros por hora Horas peak y Horas Valle

**Horas peak:**

Las horas peak para los tres tipos de transporte son entre las 7:00 AM y las 9:00 AM y entre las 5:00 PM y las 7:00 PM, cuando las personas viajan al trabajo o regresan a casa.

El metro tiene la mayor densidad de pasajeros durante las horas peak, seguido de buses y colectivos.

**Horas valle:**

Las horas valle para los tres tipos de transporte son entre las 11:00 PM y 1:00 AM, cuando hay la menor actividad.

Los colectivos tienen la menor densidad de pasajeros durante las horas valle, seguidos de buses y metro.

**Comparaciones de Valor:**

**Metro:**

Ventajas: Mayor demanda total, mayor densidad de pasajeros durante las horas peak, servicio más confiable y consistente.

Desventajas: Puede ser más caro que los buses o colectivos, puede no cubrir todas las áreas de la ciudad.

**Buses:**

Ventajas: Segunda mayor demanda total, relativamente asequibles, cubre un área más amplia que el metro.

Desventajas: Puede ser más lento que el metro, puede estar más congestionado durante las horas peak.

**Colectivos:**

Ventajas: Opción más asequible, rutas flexibles, puede cubrir áreas no atendidas por metro o buses.

Desventajas: Menor demanda total, puede ser menos frecuente y menos confiable que metro o buses.

**Aplicaciones Futuras**

**Planificación del transporte:**

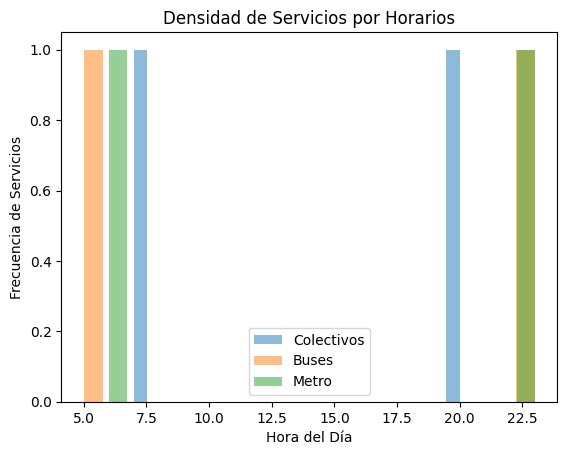
La información en la imagen podría ser utilizada por las autoridades de transporte para planificar y programar los servicios de transporte público de manera más eficiente.

Esto podría implicar ajustar los niveles de servicio para satisfacer la demanda, optimizar las rutas y mejorar la coordinación entre diferentes modos de transporte.

**Conclusiones**

El análisis detallado del gráfico de demanda de servicios de transporte público revela patrones claros de uso a lo largo del día. El metro emerge como el medio con mayor demanda y densidad de pasajeros durante las horas peak, seguido por los buses y los colectivos. Este conocimiento es crucial para mejorar la eficiencia operativa, ajustar horarios y rutas, y ofrecer un servicio más satisfactorio para los usuarios. Además, proporciona información valiosa para planificar estrategias de movilidad urbana que puedan mitigar la congestión y mejorar la accesibilidad en diferentes momentos del día.

**7. Densidad de servicios por horarios**



**Análisis de Valor Comparativo**

La imagen proporcionada muestra la densidad de servicios de transporte público (colectivos, buses y metro) por horario. La densidad se mide en términos del número de paradas de cada tipo de transporte en cada hora del día.

**Comparación entre Colectivos, Buses y Metro**

**Horario peak:**

El horario peak para los tres tipos de transporte es entre las 7:00 y las 9:00 horas de la mañana, cuando las personas se dirigen al trabajo o a la escuela.

El metro tiene la mayor densidad de paradas durante el horario peak, seguido de los buses y los colectivos.

La distribución de paradas de colectivos es más uniforme a lo largo del día que la de buses y metro, lo que indica que los colectivos tienen una mayor flexibilidad horaria.

**Horario valle:**

El horario valle para los tres tipos de transporte es entre las 2:00 y las 4:00 horas de la tarde, cuando hay menos actividad.

Los colectivos tienen la menor densidad de paradas durante el horario valle, seguidos de los buses y el metro.

La distribución de paradas de metro es más estable que la de buses y colectivos, lo que indica que el metro tiene un servicio más regular a lo largo del día.

**Densidad total:**

En general, el metro tiene la mayor densidad total de paradas, seguido de los buses y los colectivos.

**Análisis de Tendencias**

La densidad de paradas de los tres tipos de transporte aumenta gradualmente a partir de las 5:00 horas de la mañana y alcanza su máximo entre las 7:00 y las 9:00 horas.

La densidad de paradas de los tres tipos de transporte disminuye gradualmente a partir de las 17:00 horas de la tarde y alcanza su mínimo entre las 2:00 y las 4:00 horas.

La distribución de paradas de metro es más constante que la de colectivos y buses, lo que indica que el metro es un modo de transporte más confiable y con mayor disponibilidad a lo largo del día.

**Aplicaciones Futuras**

La información de la imagen podría usarse por las autoridades de transporte para planificar y programar los servicios de transporte público de manera más eficiente.

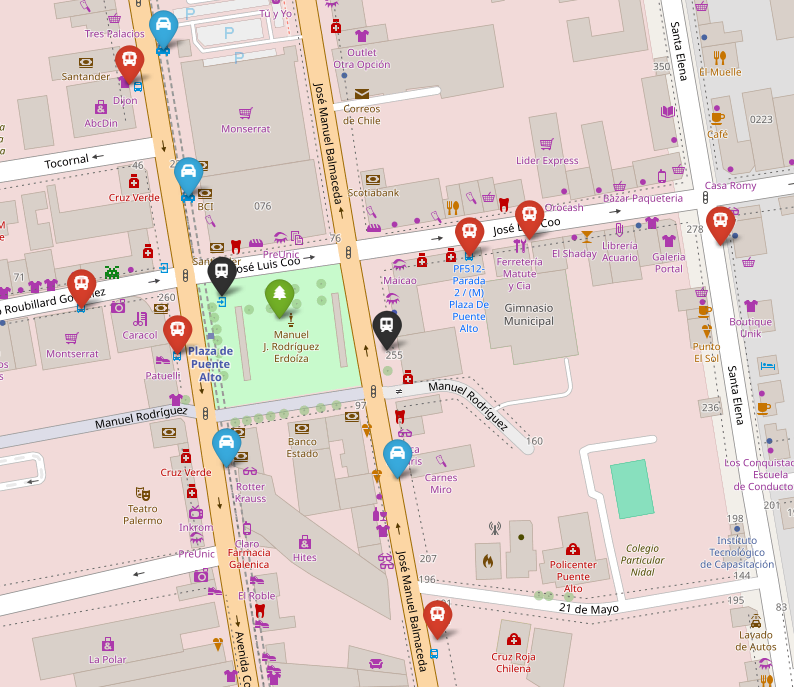
Los usuarios del transporte público podrían usar la información para planificar sus viajes y evitar las horas de mayor congestión.

Las empresas podrían usar la información para ubicar sus negocios en áreas con buen acceso al transporte público.

**Conclusiones**

La comparación de densidad de servicios de colectivos, buses y metro revela que el metro ofrece la mayor regularidad y disponibilidad a lo largo del día, especialmente en horarios peak. Los buses tienen una distribución de paradas más variable, mientras que los colectivos destacan por su flexibilidad horaria. Estos hallazgos pueden guiar la optimización de horarios y servicios, beneficiando tanto a los planificadores de transporte como a los usuarios y empresas que dependen del acceso eficiente al transporte público.

**8.Distribución Geográfica:**

****

**Análisis del gráfico**

En este gráfico se representa la ubicación de los paraderos de buses y colectivos alrededor de la plaza de Puente Alto. Al incluir en color negro los accesos al metro de Puente Alto, se puede ver cómo se distribuyen los puntos de transporte en relación con el metro. Esto es útil para analizar la accesibilidad y conectividad del transporte público con el sistema de metro, facilitando la planificación urbana y el diseño de rutas más eficientes.

### Aplicaciones Futuras

La información de la imagen podría ser utilizada por las autoridades de transporte para planificar y programar los servicios de transporte público de manera más eficiente, optimizando la conexión entre los diferentes medios de transporte.

Los usuarios del transporte público podrían usar esta información para planificar sus viajes y evitar las horas de mayor congestión, aprovechando mejor la conectividad entre buses, colectivos y el metro.

Las empresas podrían utilizar esta información para ubicar sus negocios en áreas con buen acceso al transporte público, asegurando así una mayor afluencia de clientes.

### Conclusiones

La comparación de la ubicación de los paraderos de buses y colectivos con los accesos al metro muestra cómo se distribuyen los puntos de transporte en relación con el metro, revelando que el metro es un eje central en la conectividad del transporte público en Puente Alto.

Los hallazgos indican que la distribución de los paraderos en torno al metro facilita la accesibilidad y conectividad del transporte público, lo que puede guiar la optimización de horarios y servicios, beneficiando tanto a los planificadores de transporte como a los usuarios y empresas que dependen del acceso eficiente al transporte público.

Conclusión general

El proyecto se centró en la implementación de un sistema para la recopilación, procesamiento y almacenamiento de datos tanto diarios como históricos, utilizando servicios en la nube y herramientas de análisis avanzadas. A continuación se destacan los puntos clave y las contribuciones de cada fase del proceso:

Recopilación de Datos Diarios:

Se automatizó la recopilación de datos a través de una API utilizando Cloud Functions, garantizando la actualización continua de la información en formato JSON.

Los datos se transformaron y almacenaron en Google Cloud Storage en formato CSV, utilizando Python Pandas para asegurar la calidad y consistencia de los datos.

1. Recopilación de Datos Históricos:

Los datos históricos se obtuvieron de una fuente externa mediante Cloud Functions, optimizando la captura y almacenamiento en Google Cloud Storage.

Se implementó Python Pandas para el procesamiento y limpieza de datos, facilitando su posterior análisis y uso.

1. Gobernanza y Seguridad de Datos:

Se desarrolló una Cloud Function específica para la gestión de metadatos, que registró automáticamente eventos como inserciones y actualizaciones en BigQuery.

Este enfoque no solo mejoró la seguridad de los datos, sino que también aseguró el cumplimiento de normativas al proporcionar un registro detallado y trazable de las acciones realizadas.

1. Análisis y Visualización de Datos:

Se realizó un análisis detallado de la cobertura y frecuencia del servicio de transporte público en Puente Alto, Chile, utilizando herramientas como Google Colab para generar visualizaciones informativas y comparativas.

Los resultados destacaron áreas de mejora potencial en la planificación del transporte y la satisfacción del usuario, basándose en criterios como cobertura geográfica, tarifas, y demanda por hora y día.