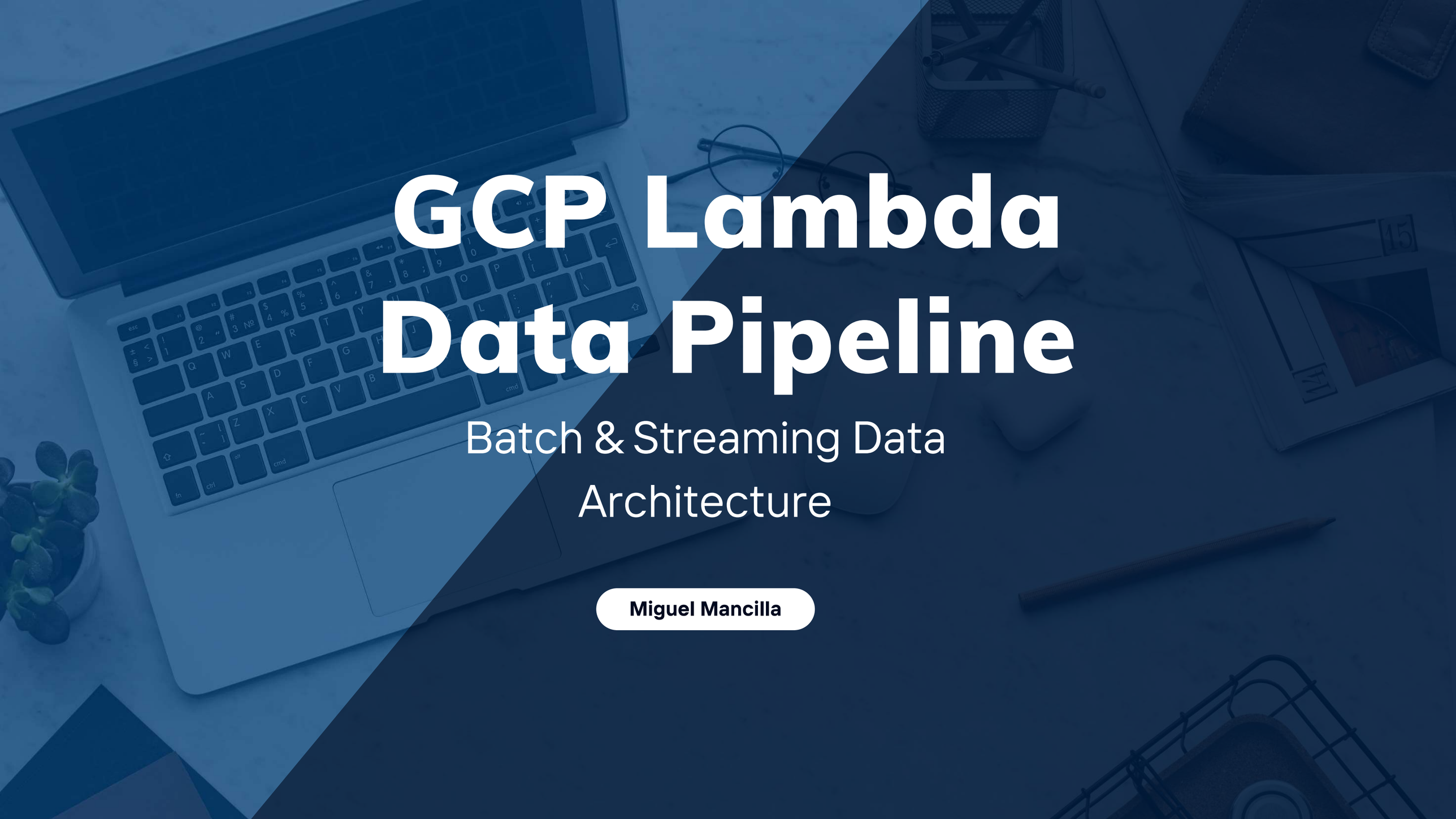


GCP Lambda Data Pipeline

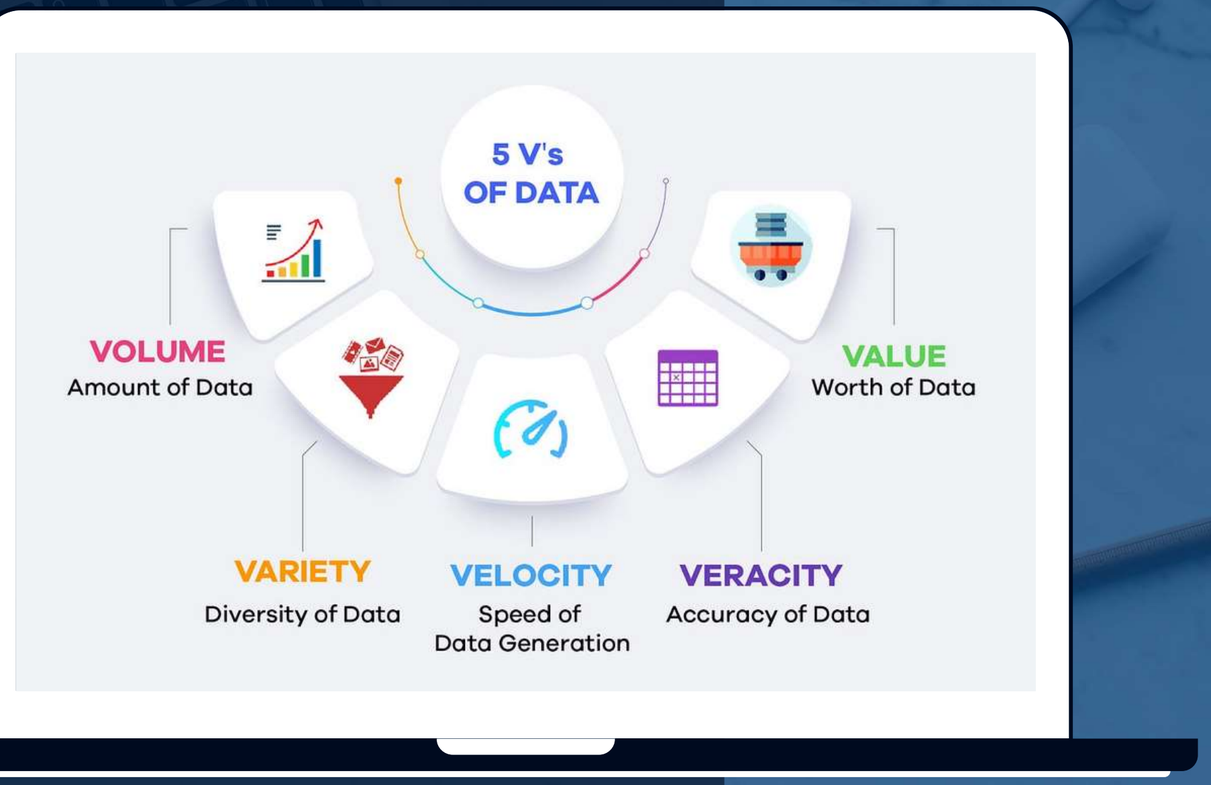
Batch & Streaming Data Architecture

Miguel Mancilla

INTRODUCCIÓN

El proyecto integró datos históricos (taxis de Nueva York) y en tiempo real (subastas electrónicas) en Google Cloud Platform, utilizando una arquitectura Lambda con procesamiento batch y streaming en BigQuery. Se aplicaron análisis exploratorios, modelos predictivos y visualizaciones en Looker Studio. Además, se abordaron las 5Vs, herramientas, gobernanza y soluciones que transformaron los datos en valor para la toma de decisiones.

JUSTIFICACIÓN 5V´S



GOBIERNO DE DATOS

Políticas de calidad de datos

Se establecieron y aplicaron reglas de validación y transformación dentro del proceso en Dataflow, tales como la eliminación de registros duplicados y la validación de campos obligatorios. Estas políticas garantizaron la integridad y coherencia tanto de los datos históricos (procesamiento por lotes) como de los datos en tiempo real (streaming). Como resultado, los datos almacenados en BigQuery se mantuvieron preparados para su análisis y visualización en Looker Studio, minimizando errores e inconsistencias.

Políticas de seguridad y acceso

El control de acceso fue gestionado a través de Cloud IAM, estableciendo roles y permisos específicos según el tipo de usuario y la función desempeñada. Se asignaron permisos de solo lectura para la visualización de datos en Looker Studio y permisos de edición para los responsables de los procesos de transformación en Dataflow. Esta configuración permitió proteger la integridad de la información almacenada en BigQuery, asegurando que solo usuarios y servicios autorizados pudieran acceder o modificar los datos.

CICLO DE VIDA DEL DATO

Obtención

Cloud Storage , Cloud Run Y Pub/sub

Procesamiento

Dataflow

Almacenamiento

Bigquery

Explotación

Loocker Studio

ARQUITECTURA LAMBDA

INGESTA EN TIEMPO REAL

- Utiliza servicios como Cloud Pub/Sub para capturar datos en tiempo real.

PROCESAMIENTO EN TIEMPO REAL

- Se apoya en Dataflow para transformaciones y análisis de datos a medida que llegan.

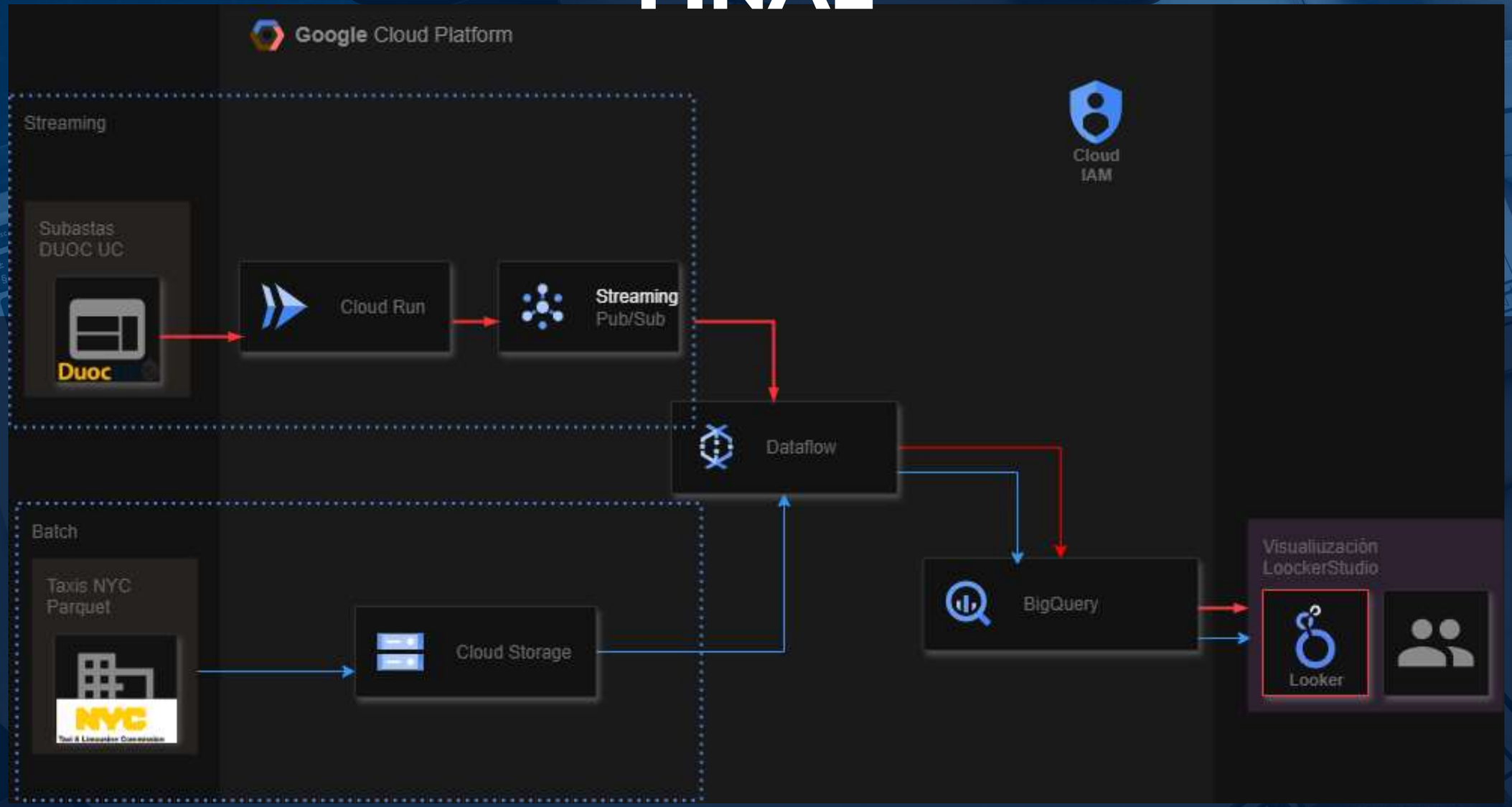
PROCESAMIENTO BATCH

- Implementa herramientas como Dataproc para analizar grandes volúmenes de datos históricos.

ALMACENAMIENTO

- Datos en tiempo real se almacenan en BigQuery, mientras que los datos procesados por lotes se guardan en Cloud Storage o Bigtable.

ARQUITECTURA FINAL



CONTEXTO BASH

FUENTE DE DATOS

Dataset historio de viajes de Nueva York, almacenado en formato Parquet.

REQUERIMIENTO

Procesar y almacenar más de tres años de datos, organizados en 36 archivos mensuales, para análisis histórico

FLUJO DE DATOS

Los parquet se cargan y almacenan en Cloud Storage, se procesan mediante Dataflow y se cargan a BigQuery para su explotación analítica.

IMPLEMENTACIÓN BATCH

Ingesta

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Fermentum leo.

Procesamiento

Control de nulos , etc...

Almacenamiento

119 Millones de datos y uso 15.78 GB

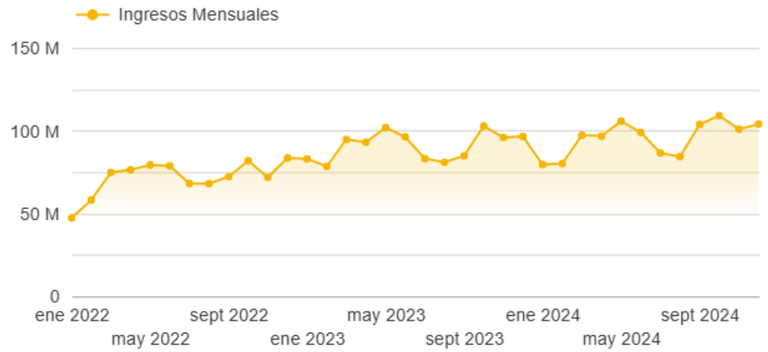
Visualización

Dashboard con grafica para el analisis del negocio

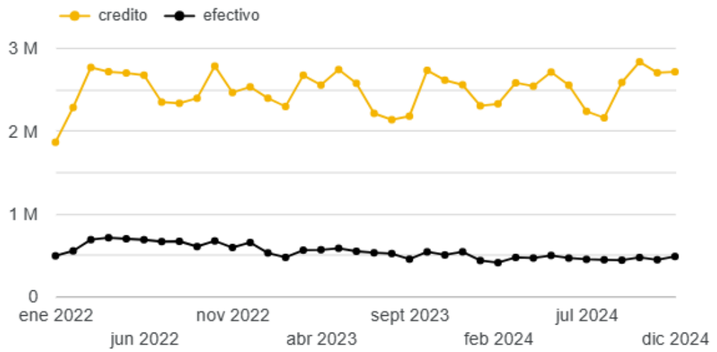
DASHBOARD BATCH

Analisis de Taxis Amarillos a travez del tiempo

Ingresos Mensuales a lo largo del tiempo



Comparativa de credito y efectivo mensual



Comparativa de cantidad de viajes por día de la semana (2022-2024)

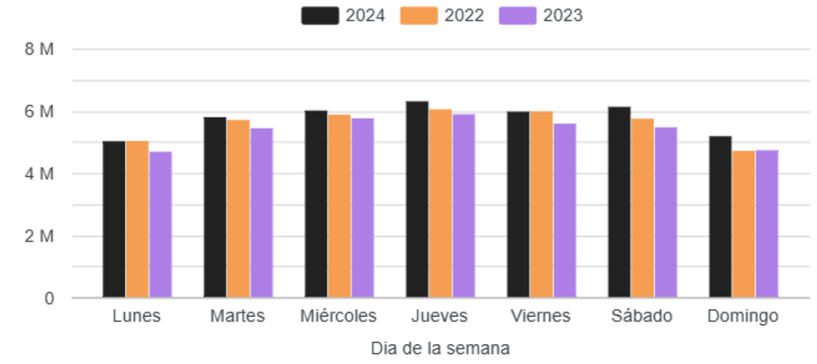
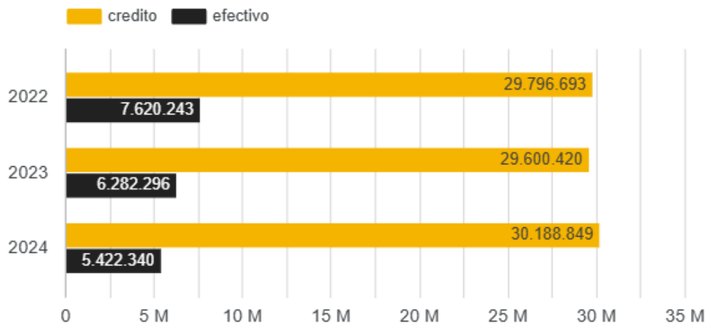


Grafico Comparativo de Metodos de pago cada año



VALOR

COMPORTAMIENTO

Los ingresos mensuales han crecido de forma sostenida desde 2022, superando los 100 millones en 2024. Se observa un uso predominante del pago con crédito sobre efectivo durante todo el periodo.

TOMA DE DECISIONES

La preferencia clara por el pago con crédito permite priorizar este método para promociones o mejoras tecnológicas. Además, los días jueves y viernes registran mayor volumen de viajes, por lo que se podrían asignar más recursos o conductores en esas jornadas.

IDENTIFICACIÓN

Los datos evidencian que el método de pago en efectivo ha disminuido progresivamente cada año. Asimismo, se mantiene una mayor actividad entre semana que durante los fines de semana, especialmente en 2024.

EVOLUCION

Desde 2022 a 2024, los ingresos han aumentado de manera consistente. A la vez, se mantiene una tendencia estable en la cantidad de viajes por día, aunque con una leve alza los días jueves y viernes en los últimos años.

CONTEXTO STREAMING

FUENTE DE DATOS

Subasta electronica en vivo de DUOC UC, enviadas cada dos minutos como registro en JSON

REQUERIMIENTO

Procesamiento y visualiacion en tiempo real los datos generados por el sistema de subastas .

Para monitorear y analizar

FLUJO DE DATOS

Los datos ingresan a travez de un servicio desplegado en CloudRun, se publican en pub/sub , se transforman con dataflow y se alamancena en bigquery

IMPLEMENTACIÓN STREAMING

Ingesta

Topic Pub/sub , configurando cloud run , atravez de un codigo pyton otorgando permiso , y el registro de la url en duoc uc

Procesamiento

Datos de subastas capturadop via pub/sub se procesan con dataflow y se alamacena en la tabla de datosTR en bigquery

Almacenamiento

Cantidad de 39.107 registros ocupando de almacenamiento
2.78 MB desde 29-06-2025
hasta 12-07-2025

Visualización

Dashboard con grafica para el analisis del negocio

DASHBOARD STREAMING

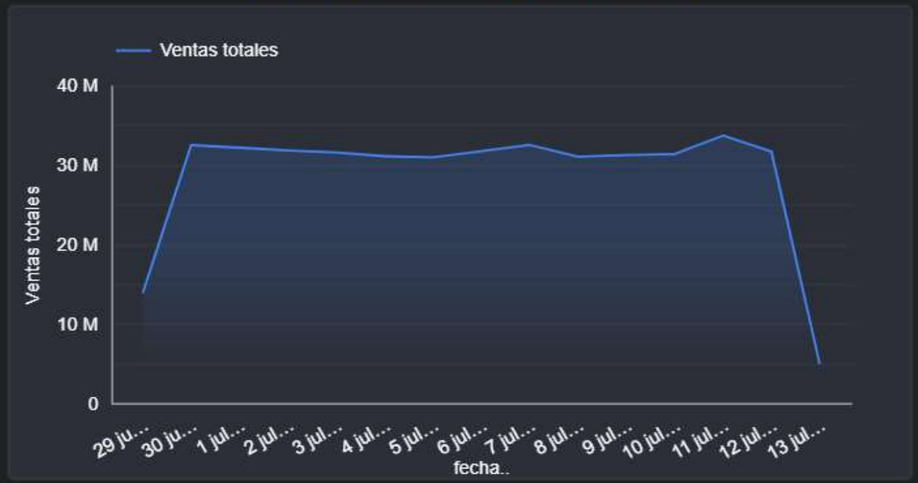
Analisis de subastas

DuocUC

Distribución de transacciones por metodo de pago

Distribución de ventas por Producto y forma de pago

Ventas totales generadas por dia



Distribución por ingreso total generado por cliente

VALOR

1

Actividad

Las ventas diarias se mantienen estables, con montos sobre los 30 millones, lo que refleja un buen nivel de actividad constante en el sistema.

2

Clientes a considerar

Mauricio Correa y Elizabeth Cordero destacan por su alto monto total y número de transacciones, siendo clientes clave para retención o beneficios exclusivos.

3

Preferencias

No hay una diferencia marcada entre métodos de pago, pero el crédito domina en productos de mayor valor como Microsoft y Nvidia, lo que indica una tendencia en compras de alto monto.

CONCLUSIÓN

Este proyecto consistió en implementar una solución utilizando herramientas de Google Cloud Platform, integrando datos históricos (batch) y en tiempo real (streaming) mediante el modelo Lambda. De este modo, los datos fueron transformados en información valiosa para apoyar la toma de decisiones.

GRACIAS

Miguel Mancilla