Actividad 2: Ciclo de vida de ingeniería de datos

Hugo Porras

1. Contexto de la actividad

Esta semana trabajaremos con los datos **TLC Trip Data**. El **New York City Taxi & Limousine Commission (TLC)** publica datos abiertos sobre los servicios de transporte en la ciudad de Nueva York. La información disponible corresponde a los siguientes tipos de servicio:

- Taxi amarillo (Yellow Taxi)
- Taxi verde (Green Taxi)
- For-Hire Vehicles (FHV)

La **data descargable**, correspondiente a los registros de viajes, se encuentra disponible en el siguiente enlace oficial:

`https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

Adicionalmente, información general sobre el organismo, el contexto regulatorio y la descripción de los datos puede encontrarse en:

`https://www.nyc.gov/site/tlc/about/about-tlc.page`

El objetivo de esta actividad es que el estudiante analice estos datos desde la perspectiva de la **ingeniería de datos**, aplicando los conceptos del ciclo de vida de la ingeniería de datos vistos en clase.

2. Actividades

2.1. Comprensión del sistema fuente (Generación de datos)

Revisa la página del TLC y la descripción de los datasets disponibles. Seleccionar **uno** de los siguientes tipos de datos:

- Taxi amarillo
- Taxi verde
- FHV

Luego responde:

- ¿Qué sistema o conjunto de sistemas genera estos datos?
- ¿Qué evento real representa cada registro del dataset?
- ¿Los datos se generan de forma continua o periódica?
- ¿El ingeniero de datos tiene control directo sobre el sistema fuente? ¿Por qué?

2.2. Diseño conceptual del almacenamiento

Suponga que trabaja como ingeniero de datos para una organización que desea explotar estos datos.

Responde las siguientes preguntas:

- ¿Estos datos serían *hot*, *warm* o *cold*?
- ¿Puede cambiar esta clasificación según el caso de uso?
- ¿Qué tipo de almacenamiento conceptual utilizaría (data lake, data warehouse o ambos)?
- ¿Qué datos conservaría en crudo y cuáles transformaría?

2.3. Estrategia de ingesta

Para el dataset seleccionado, responde:

- ¿La ingesta sería *batch* o *streaming*?
- ¿Se utilizaría un modelo *push* o *pull*?

Justifica tu respuesta considerando impacto sobre el sistema fuente, requerimientos de latencia y complejidad operativa.

2.4. Disponibilización de los datos

Para el dataset seleccionado, responde cómo cambian los requisitos de latencia, acceso y disponibilidad para los siguientes casos:

- Análisis y reportería del área de finanzas de la comisión de transporte
- Departamento de logística de cada compañía de taxis

2.5. Condiciones transversales

Elige dos condiciones transversales vistas en clase (seguridad, calidad, gobernanza, metadatos u operación).

Para cada una, indica:

- Un riesgo concreto asociado a estos datos.
- Una acción de mitigación desde la ingeniería de datos.

3. Entrega

Debes entregar tus respuestas en un solo documento de máximo dos páginas con letra de tamaño 11 e interlineado simple.

Si utilizas IA generativa:

- Se permite el uso de herramientas como **ChatGPT únicamente para revisión de sintaxis y ortografía**.
- No está permitido utilizar estas herramientas para generar o responder el contenido de la actividad.
- El uso indebido de IA para responder la actividad podrá ser detectado y conlleva la calificación de **cero**.