# Data Fabric

## Objetivos del taller

* Mediante un script Python, simular el acceso a datos a través de la arquitectura Data Fabric, demostrando cuatro de sus características.
* Visualizar los datos obtenidos mediante la librería streamlit.

Prerrequisitos.

* Python => 3.9
* pandas\*
* streamlit\*

\*Instalación mediante pip (gestor de paquetes de Python).

## Material

* Clonar el repositorio <https://github.com/karengutierrezl/DataFabric/> . (Se puede descargar como zip y descomprimir los archivos).

## Background

Los archivos .json contenidos en el archivo zip, establecen las reglas para la validación de usuarios, asignación de roles y configuración de los diccionarios de datos a los que pueden tener acceso los diferentes usuarios por medio de los roles asignados.

Los apartados de la arquitectura *Data Fabric* que se probarán con este taller son los siguientes.

### Data Policies

Mediante la aplicación de roles descritos en el archivo roles, su asociación con el archivo *users* y su relación con el archivo *DD*, se establece que es a lo que un usuario puede o no tener acceso.

### Data Lineage

Usando los metadatos, descritos en el archivo *DD*, se pueden acceder a las bases de datos y columnas de acuerdo con las reglas ahí establecidas.

### Data Ingestion

Consumir los datos marcados por las políticas, y descritos por los metadatos, para para su utilización en el script, en este caso son los archivos *IMDb Top TV Series.csv* y *netflix.csv*.

### Data Virtualization

Las fuentes de datos definidas en los metadatos del archivo *DD*, en la columna golden\_name, son en esta ocasión un archivo *csv*, pero de igual manera se pueden establecer conexiones a otras fuentes de datos como bases de datos, otros archivos, etc.

## Metodología

Este ejercicio a la par de la simulación de la arquitectura Data Fabric, intenta recrear condiciones similares a las encontradas en la industria, donde los tiempos de entrega están marcados estrictamente y se deben mostrar avances para las actividades en los tiempos determinados por las organizaciones. Es por eso por lo que cada apartado tiene indicado el tiempo disponible para cada actividad.

1. Leer los archivos json. (30 minutos)
   1. users.json
   2. roles.json
   3. DD.json

Revisar las rutas desde donde se cargarán dichos archivos.

1. Validar que un usuario determinado tiene o no acceso. (15 minutos)
2. Establecer relaciones entre las claves de cada uno de los archivos. (45 minutos)

El archivo users.json contine una clave referencia a roles.json, el cual contiene un campo que es referencia del archivo DD.json. En el archivo users .json, se establece también una clave referencia a cuáles de los diccionarios de datos, un usuario tiene acceso.

1. Determinar a cuáles columnas y de cual dataset tiene acceso un usuario y cargar la información a listas. (30 minutos)
2. Leer los archivos .csv con la librería pandas (import pandas as pd) mediante la instrucción pd.read\_csv('ruta') y filtrar las columnas a las cuales tiene acceso el usuario enviando los resultados del paso anterior como parámetro para el filtrado del dataframe. (30 minutos)
3. Utilizar la librería streamlit (import streamlit as st). (45 minutos)

Streamlit es una herramienta de python, que se utiliza para crear visualización de datos mediante la generación de una aplicación web, la cual se ejecuta mediante la ejecución de un servidor web que la misma librería contiene, pero para ejecutar una aplicación streamlit ya no basta dar clic en el botón ejecutar del IDE o el comando en la terminal python3 *<nombre\_archivo.py>*, sino se debe ejecutar a través del comando en la terminal:

streamlit run *<nombre\_archivo.py>*

* 1. st.title() agrega titulo a la página generada.
  2. st.write(dataframe) escribe el contenido del dataframe en una tabla.
  3. with st.sidebar: para agregar una barra lateral.
  4. st.text\_input() agrega una caja de texto para entrada de datos.
     1. variable = st.text\_input() Asigna el texto a una variable string
  5. st.button() agrega un boton.
     1. if st.button() puede escuchar el evento de clic de un boton.

\*\*Cada función requiere y acepta diferentes parametros. Estas son solo pistas para su utilización.