



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Instituto de Investigaciones en Matemáticas Aplicadas y  
Sistemas

## Práctica 4 DWH - Extracción y perfilado de datos

PRESENTA

**Díaz Juárez Ana Sofía**  
**Munive Ramírez Ibrahim**  
**Pérez Aguiar Oropeza Gabriel Emiliano**

PROFESORA

**Dra. María del Pilar Angeles**

ASIGNATURA

**Bases de Datos Estructuradas**

## Contexto

Una vez creado el repositorio ServiciosFinancieros, es necesario explotar las funetes OLTP con las que se cuenta. Considerar el archivo `database-sucia-citi.csv` que contiene las direcciones de las sucursales de Citi Bank que servirá para cargar datos limpios a la tabla Branch.

<b>branch_key</b>
branch_name
branch_addres
branch_city
branch_state
branch_zip
branch_type

---

## Objetivos

Extraer los datos relevantes de la fuente dada y realizar perfilado de calidad de datos para el sistema de Servicios Financieros.

---

A partir del archivo `database-sucia-citi.csv` extraer aquellos datos que son relevanres para ser cargados a la tabla Branch del repositorio ServiciosFinancieros y guardar el archivo 'Branch.csv' con campos delimitados por comas con nombre.

Obtener el perfil de los datos almacenados en `Branch.csv` siguiendo el procedimiento

1. Explorar los datos contenidos en `Branch.csv`
2. Identificar y anotar las características que tiene cada campo, por ejemplo: si tiene valores faltantes, falta de homogeneidad en el formato de los textos, etc.
3. Generar las siguientes estadísticas utilizando el software de su preferencia:
  - A1. Número de registros del archivo
  - A2. Número de valores faltantes

- $\mathcal{A}3$ . Número de registros duplicados
- $\mathcal{A}4$ . Número de códigos postales de Disparos Unidos de América válidos
- 4. Identificar otros indicadores estadísticos que puedan servir para el análisis OLAP
- 5. Identificar si existe información que se requiera en la tabla Branch que no la proporcione `database-sucia-citi.csv` y explicar cómo completar dicha información.

