# Semana 3 - Actividad 1

## Limpieza de datos

Lázaro Lara Martínez. Matricula A01793198

José Mtanous Treviño. Matricula A00169781

Ciencia y Analítica de datos.

Profesor Titular. Jobish Vallikavungal Devassia

Profesor Tutor. Mtro. Mario Alberto Solano Saldaña

01/Octubre/2022

## Fundamentos de bases de datos y para ciencia de datos.

¿Qué es una base de datos?:

Es un sistema capaz de almacenar gran cantidad de datos que tienen relación entre sí y están estructurados. Sus datos pueden ser consultados rápidamente utilizando un lenguaje estructurado de consultas llamado SQL. En una base de datos existen diferentes tipos de objetos como son tablas las cuales contienen columnas, filas y celdas. Los registros en una tabla son los renglones los cuales contienen diferentes campos o columnas, las celdas contienen un dato de dicha columna y renglón. También existen vistas las cuales pueden unir lógicamente diferentes tablas, índices para hacer más rápidas las consultas, triggers para ejecutar acciones antes o después de insertar, actualizar o borrar datos y constraints para restringir el contenido de los datos.

Las bases de datos utilizan el paradigma de ETL, en donde los datos se depuran y organizan antes de cargarlos, este proceso puede ser costoso para fuentes de datos muy grandes.

Existen distintas aplicaciones para las bases de datos como bancos, aerolíneas, universidades, ventas, tiendas en línea, RH etc. Las Bases de datos más utilizadas son, Oracle MySQL, PostgreSQL y MS SQL Server.

¿Qué es un Data Warehouse?

Es un sistema de almacenamiento de datos que permite a las empresas comprender y utilizar sus datos para tomar decisiones estratégicas, generalmente los Data Warehouse contienen datos no estructurados o con estructura mínima (particiones), a diferencia de las bases de datos tradicionales, esta arquitectura de almacenamiento utiliza el paradigma ELT, en donde los datos se cargan sin depurarse, incluso muchas veces en un formato 'crudo' raw. El consumidor de los datos es el encargado de depurarlos y transformarlos en algo que haga sentido para el contexto en que se usarán.

Estos sistemas de Data Warehouse generalmente se implementan en hardware de propósito general y están diseñados para ser redundantes y ejecutar tareas en paralelo (MapReduce), la estructura de datos que almacenan están optimizadas para lectura y es muy costoso actualizar de manera parcial los datos, son sistemas optimizados para lectura y procesamiento.

## Selección y limpieza de los Datos en Python

## Se importan las librerías Pandas y numpy

In [ ]:
```
import pandas as pd
import numpy as np
```

Primero obtenemos los datos frescos desde el archivo csv Y mostramos los primeros 5 renglones.

In [ ]:
```
myDataSetUrl = 'https://raw.githubusercontent.com/PosgradoMNA/Actividades_Aprendizaje-/main/default%20of%20cred
ndf = pd.read_csv(myDataSetUrl, index_col=0)
ndf.head(5)
```

In [ ]:
```
# Describimos como vienen los datos.
ndf.describe()
```

In [ ]:
```
# Copiamos el dataframe original para no sobre escribirlo.

df = ndf.copy()
```

In [ ]:
```
# Obtenemos las llaves
df.keys()
```

In [ ]:
```
# Varias formas de checar los valores NaN, NA y NULL
#df.isnull()
df.isnull().sum()
# df.isnull().values.any()
# df.isnull().values.sum()
# df.isnull().any()
```

In [ ]:
```
#Con esta instrucción obtenemos la información del dataset y vemos que todos los datos son numéricos.
df.info()
```

In [ ]:
```
#Existen datos nulos?

df.isnull().values.any()
```

In [ ]:
```
#Que campos tienen datos nulos?
df.isnull().any()
```

In [ ]:
```
df.count()
```

In [ ]:
```
df.dropna(inplace = True)
```

In [ ]:
```
df.count()
```

In [ ]:
```
#Localizamos los valores nulos del Dadaframe original y comparamos que en el nuevo ya se borraron.
null_columns=ndf.columns[ndf.isnull().any()]
ndf[null_columns].isnull().sum()
print(ndf[ndf.isnull().any(axis=1)][null_columns].head())
```

In [ ]:
```
null_columns=df.columns[df.isnull().any()]
df[null_columns].isnull().sum()
print(df[df.isnull().any(axis=1)][null_columns].head())
```

In [ ]:
```
#De nuevo mostramos los primeros 5 datos.
df.head(5)
```

In [ ]:
```
# y la info resultante.
df.info()
```

## Decidimos borrar los datos nulos porque no tiene sentido desde el punto de vista del negocio asignar valores de media, mediana o moda diferentes clientes además, se borraron solo 42 registros del total de 30,000.

In [ ]:
```
## Obtener los Valores diferentes en los datos de Género para corroborar que sea consistente.
# X2: Gender (1 = male; 2 = female). Todo se ve bien.

df.X2.unique()
```

In [ ]:
```
## Obtener los Valores diferentes en los datos de Género para corroborar que sea consistente.
# X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
# Podemos ver que existen valores fuera de lo esperado. 0, 5 y 6. El cero puede considerarse vacío o nulo debe

df.X3.unique()
```

In [ ]:
```
## Obtener los Valores diferentes en los datos de Género para corroborar que sea consistente.
# X4: Marital status (1 = married; 2 = single; 3 = others). Existe el valor 0 que no se encuentra en la descrip

df.X4.unique()
```

## Sección 3

1. ¿Qué datos considero mas importantes?

Para este tipo de análisis todos los datos del archivo son necesarios, sin embargo, creemos que es primordial saber X1 el monto del crédito, X6 – X11 el historial de pagos, X12 – X17 monto del estado de cuenta, y X18 – X23 cantidad de los últimos pagos.

1. ¿Se eliminaron o reemplazaron datos nulos? ¿Qué se hizo y por qué?

Decidimos borrar los datos nulos porque no tiene sentido desde el punto de vista del negocio asignar valores de media, mediana o moda diferentes clientes. Además, se borraron solo 42 registros del total de 30,000, aproximadamente el 0.14%

1. ¿Es necesario ordenar los datos para el análisis? Sí / No / ¿Por qué?

Desde el punto de vista estadístico no consideramos necesario ordenarlos, debido a que cada registro es independiente del siguiente, no es necesario comparara un registro contra el siguiente. Desde el punto de vista de procesamiento tampoco consideramos que es necesario ordenarlos ya que el tamaño del conjunto de datos es muy pequeño y cabe en memoria, si fuera un conjunto enorme en donde tuviéramos que paralelizar el procesamiento podríamos pensar en ordenar los datos o particionarlos, Sin embargo, para este conjunto no es necesario.

1. ¿Existen problemas de formato que deban solucionar antes del proceso de modelado? Sí / No / Por qué.

No existen problemas de formato, los datos categóricos X2, X3 y X4 contienen datos numéricos como se indica en la información de los atributos. El único problema que veo en los datos categóricos es unos cuantos ceros que pueden considerarse nulos, pero no puede asignarse un valor arbitrario, ya que es el nivel de educación del cliente, o su estatus marital. En Educación existen valores no esperados, 5 y 6 que no vienen descritos en la información del atributo. Sería cuestión de investigar cuál es el significado de dichos valores.

Todos los demás datos hablan de montos y son numéricos, los datos negativos tienen su significado en el negocio como por ejemplo X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

1. ¿Qué ajustes se realizaron en el proceso de limpieza de datos (agregar, integrar, eliminar, modificar registros (filas), cambiar atributos (columnas)?

Nosotros decidimos eliminamos los datos que tenían nulos, debido a que pensamos que no es correcto asignar valores como media a datos financieros de diferentes personas, no fue necesario hacer otro ajuste, y solo es un pequeño porcentaje del total de registros, si fuera un porcentaje mayor entonces si sería necesario buscar otra estrategia.