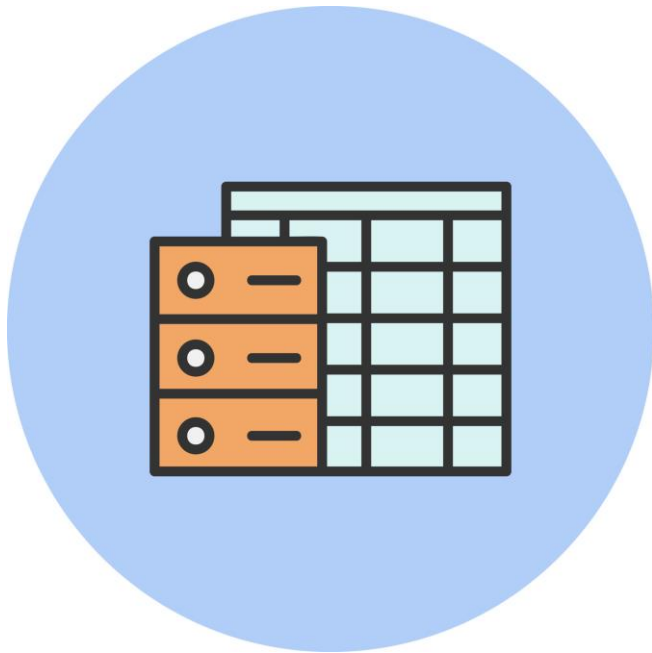


# Ciencia de Datos

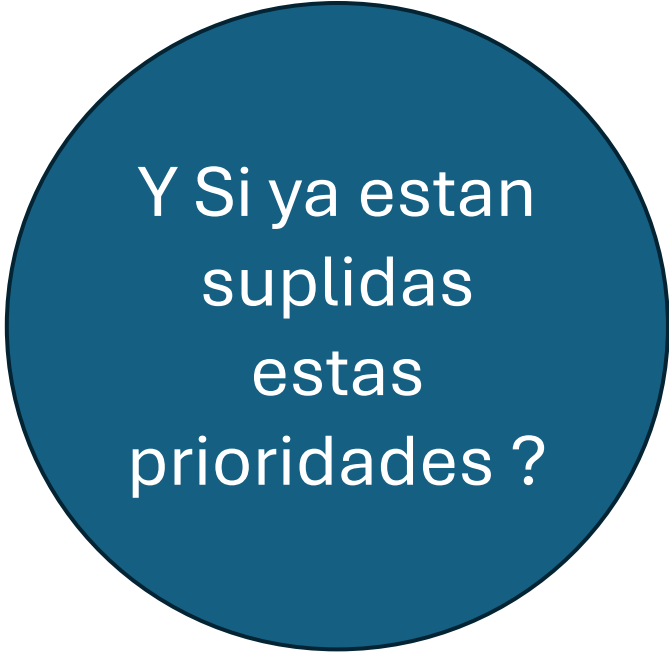
## DATASET Y HOJAS DE CALCULO



# El problema de Almacenamiento

## Prioridades últimos 10 años:

- **-Captura de informacion**
  - En archivos de texto
  - Hojas de calculo
  - Sistemas de informacion
  - formularios
  - CRM – ERP -Apps
  - Sensores
- **Seguridad**
- **Conectividad- ubicuidad**
- **Costos**

A large blue circle containing the text "Y Si ya estan suplidas estas prioridades ?" in white, sans-serif font.

Y Si ya estan  
suplidas  
estas  
prioridades ?

# El problema de Almacenamiento



## Lo Nuevo:

- Problemas
- Prioridades
- Oportunidades

1. Crecimiento Exponencial de Datos
2. Seguridad y Privacidad de los Datos
3. Gestión del Ciclo de Vida de los Datos
4. Costos de Almacenamiento
5. Almacenamiento en la Nube
6. Rendimiento y Accesibilidad
7. Escalabilidad – Nuevas arquitecturas
8. Integridad y Recuperación de Datos
9. Compatibilidad y Migración de Datos
- 10. Big Data y Analítica / IA –**
  - 1. Datos no estructurados**
  - 2. semiestructurados**

**Control  
Gobierno del  
Dato**

# Que hacer con alto volumen de Datos?

 Identificación de fuentes

 Selección de datos

 Calidad de los datos

 Limpieza

 Procesamiento

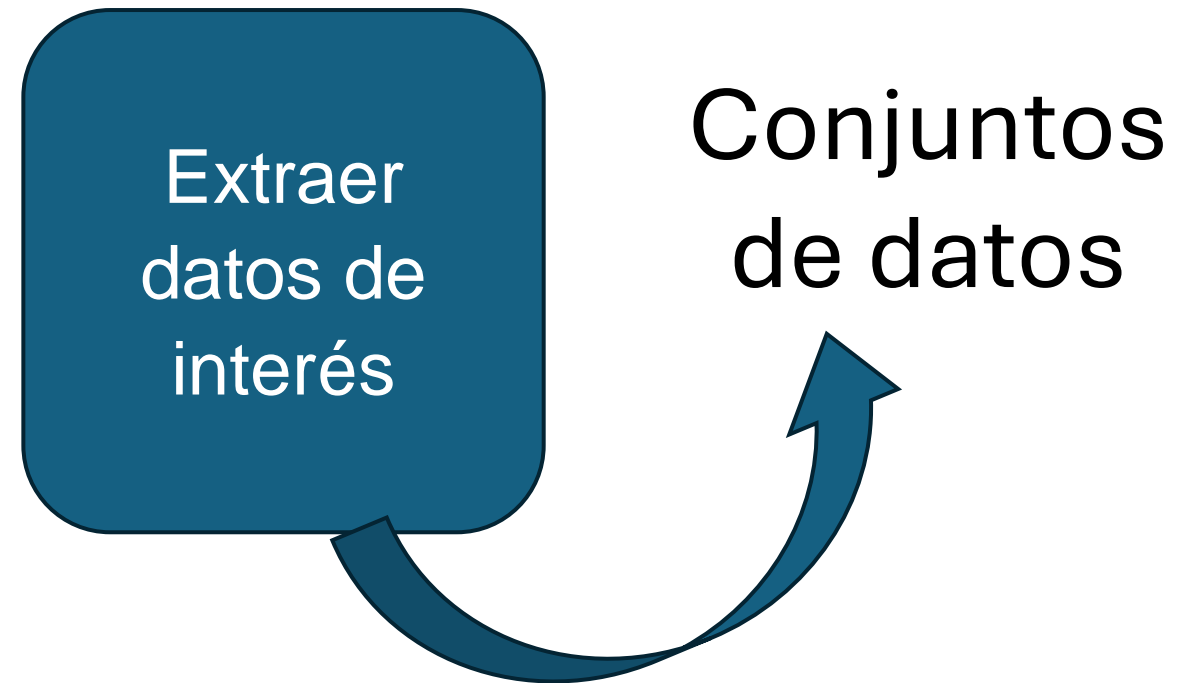
 Resultados

 Cuantitativos

 Cualitativos

 Visualización





 Toma desiciones



## Dataset –(Conjuntos de datos)

Un **Dataset** es una colección de datos estructurados




.Usos frecuentes:.

-  Entrenar modelos en ciencia de datos.
-  Análisis y resultados
-  Visualización –Dashboard
-  Tomar decisiones



| Title                            | Budget      | Genre     | Director     | Cast                                   | Box Office  | Rating | Year |
|----------------------------------|-------------|-----------|--------------|--|-------------|--------|------|
| Star Wars: The Force Awakens     | 249,000,000 | Adventure | J.J. Abrams  | Daisy Ridley, John Boyega, Oscar Isaac | 206,971,101 | 7.8    | 2015 |
| Star Wars: The Last Jedi         | 275,000,000 | Adventure | Rian Johnson | Solo, Kelly Marie Tran, John Boyega    | 220,531,101 | 6.9    | 2017 |
| Star Wars: The Rise of Skywalker | 270,000,000 | Adventure | J.J. Abrams  | Daisy Ridley, John Boyega, Adam Driver | 220,531,101 | 6.5    | 2019 |

De diversos tipos




-  tablas,
-  Imágenes
-  Textos, etc.









# Dataset –(Conjuntos de datos)

- Fuentes:**
- Privados
  - Públicos

**Consideraciones:**

-  Con objetivos claros
-  Estructura definida
-  Con Metadata

**Como se Crean:**

-  Por captación directa
-  Consultas en Bases de datos (SQL, NoSql)
-  Apis
-  Web scraping
-  Etiquetado
-  OCR...



# Dataset –Composición

Dependiendo del tipo de datos la composición generalmente esta relacionada a una estructura:

## **Headers:**

Encabezados identificadores de cada variable del dataset.

## **Fila:**

Representa cada registro relacionado a datos de una misma entidad

## **Columna**

Cada variable del Dataset

## **Metadata**

Descripción del dataset, informacion, contenido, estructura, tipos de dato, fecha de creación, fuente etc.



**Metadata:** o metadatos son, datos que proporcionan información sobre otros datos.

## Es como una etiqueta que proporciona contexto y detalles sobre:

- [illegible]



# Dataset –Metadata

**La Metadata se usa para:** organizar y recuperar datos de manera eficiente, especialmente en grandes conjuntos de información.

| Metadata example |                    |
|------------------|--------------------|
| Structural       | Descriptive        |
| Song Title:      | Better Man         |
| Artist Name:     | Pearl Jam          |
| Album Title:     | Vitalogy           |
| Genre:           | Rock               |
| Release Year:    | 1994               |
| Track Number:    | 11 of 14           |
| Composer:        | Eddie Vedder       |
| Copyright:       | © Pearl Jam        |
| Administrative   |                    |
| Added By:        | Robert Godino      |
| Date Added:      | 26/11/2016 8:19 pm |
| Encoded With:    | iTunes v7.6.1      |
| Media Kind:      | MPEG audio file    |

↑ Elements      ↑ Values

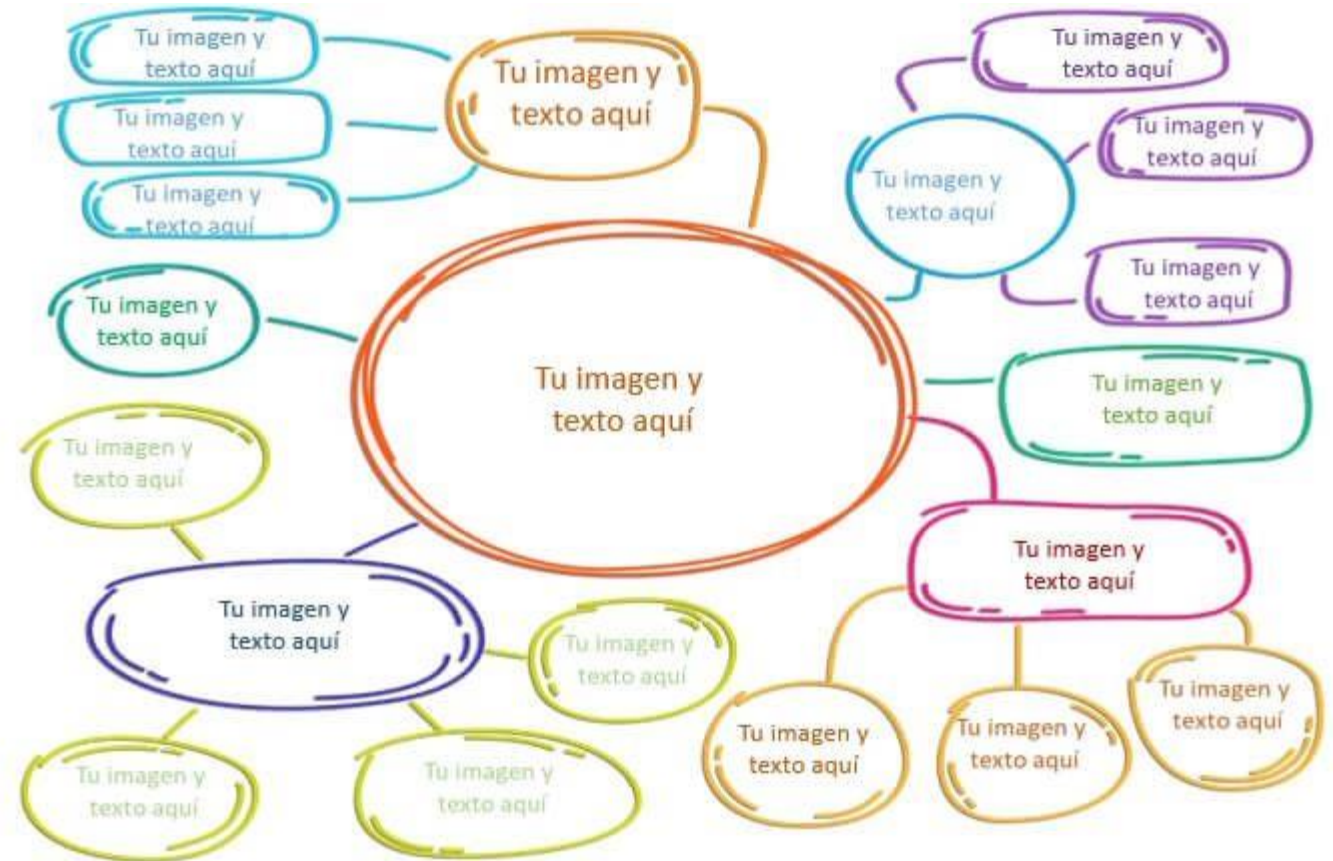
## Algunos Metadatos:

- Título
- Autor
- Información de contacto
- Fecha de Creación
- Fuente
- Palabras clave
- Resumen o descripción
- idioma
- formato de archivos
- Tipo de contenido
- Categorías o temas
- Cantidad de registros
- Estructura – Variables
- Licencias de uso /restricciones
- Etc...

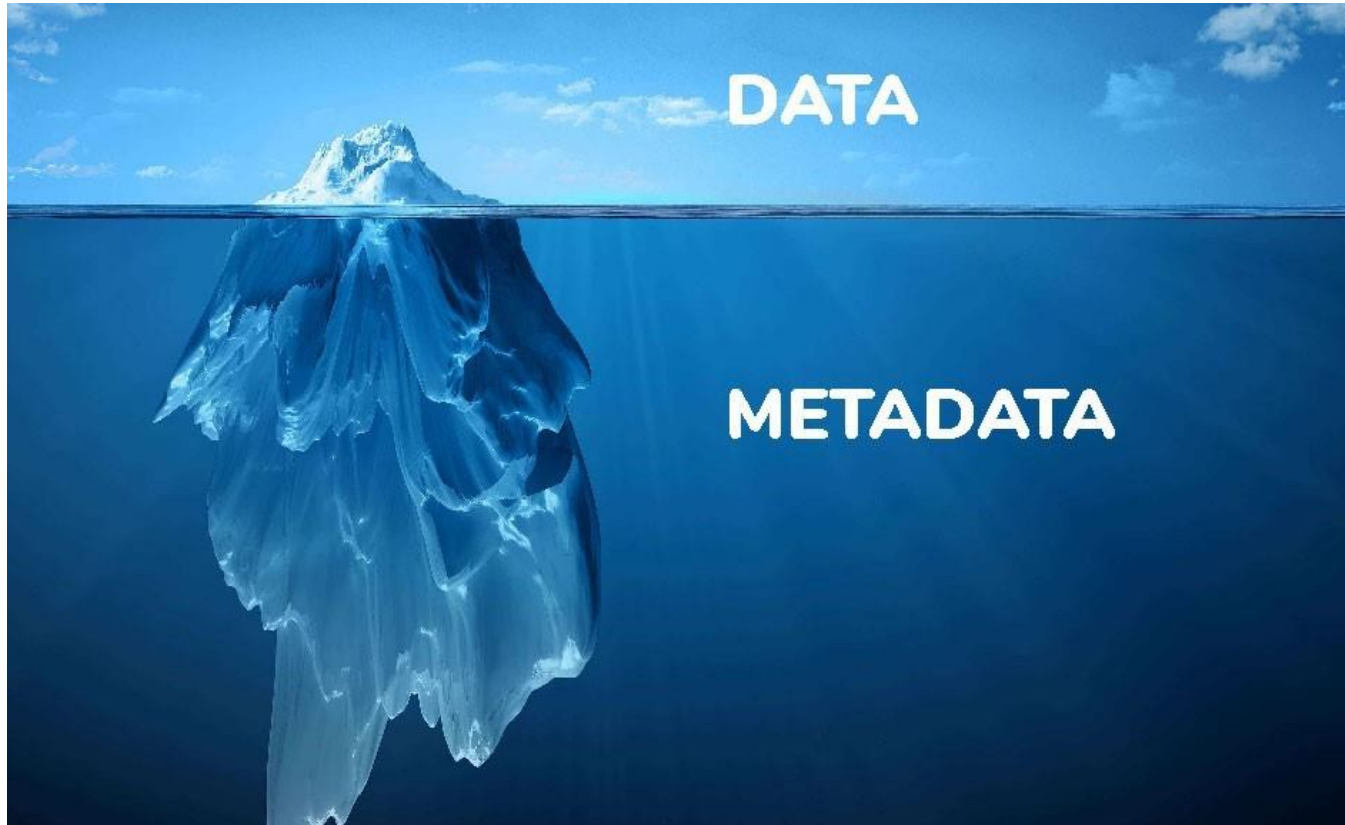
# Dataset –Metadata

La Metadata se pueden agrupar según el interés:

- ✓ Descriptivos
- ✓ Administrativos
- ✓ Estructurales
- ✓ de Proceso
- ✓ de uso
- ✓ de Localización
- ✓ Geográficos
- ✓ Temporales
- ✓ Sociales
- ✓ de Seguridad



# Dataset –Metadata



**Los metadatos desempeñan un rol crítico en la gestión de los datos**

- Facilitan la búsqueda eficiente
- permiten la interpretación y comprensión de los datos
- Tiene un papel crucial en la seguridad y el cumplimiento normativo.
- **son esenciales para las políticas de data governance. –(Gobierno del Dato)**

HOJAS DE CALCULO



| Evaluación Recuperativa (3) - Excel (Error de activación de productos)  |       |                                     |                   |          |        |        |        |        |          |        |               |
|---|-------|-------------------------------------|-------------------|----------|--------|--------|--------|--------|----------|--------|---------------|
| Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista ¿Qué desea hacer?   |       |                                     |                   |          |        |        |        |        |          |        |               |
| Inicio sesión Compartir   |       |                                     |                   |          |        |        |        |        |          |        |               |
| <div><div><div>Cortar Copiar Copiar formato</div><div>Portapapeles Fuente</div></div><div><div>Calibri 12</div><div>N K S</div><div>Ajustar texto</div><div>Alineación</div></div><div><div>General</div><div>\$ % 000 0.00 4,9</div><div>Número</div></div><div><div>Formato condicional Dar formato como tabla</div><div>Estilos</div></div><div><div>Normal Bueno Incorrecto Neutral</div><div>Estilos</div></div><div><div>Insertar Eliminar Formato</div><div>Celdas</div></div><div><div>Σ Autosuma Rellenar Borrar</div><div>Modificar</div></div><div><div>Ordenar y filtrar Buscar y seleccionar</div><div>Modificar</div></div></div> |       |                                     |                   |          |        |        |        |        |          |        |               |
| J7  |       |                                     |                   |          |        |        |        |        |          |        |               |
| A   | B     | C                                   | D                 | E        | F      | G      | H      | I      | J        | K      | L             |
|   | Nº    | NOMBRE                              | CARRERA           | BECA     | NOTA 1 | NOTA 2 | NOTA 3 | NOTA 4 | PROMEDIO | ESTADO | CERTIFICACION |
| 6   | 1     | ABARCA ZUÑIGA SUSANA MARLEN         | LOGISTICA         | SIN BECA | 36     | 21     | 70     | 64     |          |        |               |
| 7   | 2     | ARRIAGADA ROJAS LORENA ALEJANDRA    | LOGISTICA         | SIN BECA | 68     | 68     | 50     | 52     |          |        |               |
| 8   | 3     | AVENDAÑO ANDIA CATHERINE NINOSKA    | COMERCIO EXTERIOR | SIN BECA | 18     | 26     | 20     | 57     |          |        |               |
| 9   | 4     | BUSTOS GARCÍA MARCO ANTONIO         | RECURSOS HUMANOS  | SIN BECA | 70     | 69     | 70     | 31     |          |        |               |
| 10  | 5     | CAMPOS FUENTEALBA JAVIERA DEL PILAR | COMERCIO EXTERIOR | SIN BECA | 66     | 21     | 66     | 67     |          |        |               |
| 11  | 6     | CORTÉS MARQUEZ FERNANDO ESTEBAN     | ADMINISTRACIÓN    | SIN BECA | 42     | 70     | 15     | 63     |          |        |               |
| 12  | 7     | DÍAZ MARTÍNEZ RODRIGO GIOVANNI      | RECURSOS HUMANOS  | SIN BECA | 15     | 62     | 34     | 38     |          |        |               |
| 13  | 8     | DONOSO SANCHEZ DAVID ISAAC          | COMERCIO EXTERIOR | BECA     | 66     | 11     | 56     | 60     |          |        |               |
| 14  | 9     | ERICES GÓMEZ NICOLÁS IGNACIO        | RECURSOS HUMANOS  | SIN BECA | 19     | 70     | 67     | 29     |          |        |               |
| 15  | 10    | ESPIÑOZA SOPOLA PABLA MARCELA       | COMERCIO EXTERIOR | SIN BECA | 39     | 65     | 30     | 55     |          |        |               |
| 16  | 11    | FIGUEROA CUEVA FRANCISCA YOLANDA    | RECURSOS HUMANOS  | SIN BECA | 65     | 51     | 31     | 38     |          |        |               |
| 17  | 12    | GARCÉS MARQUESSI YEISON ALEJANDRO   | LOGISTICA         | SIN BECA | 29     | 37     | 35     | 41     |          |        |               |
| 18  | 13    | GONZÁLEZ GÓMEZ JUAN LEANDRO         | COMERCIO EXTERIOR | BECA     | 52     | 51     | 70     | 66     |          |        |               |
| 19  | 14    | GONZÁLEZ SALGADO JAVIERA IGNACIA    | RECURSOS HUMANOS  | SIN BECA | 46     | 35     | 24     | 35     |          |        |               |
| 20  | 15    | LÓPEZ LARA ALICIA CRISTINA          | ADMINISTRACIÓN    | SIN BECA | 70     | 54     | 58     | 39     |          |        |               |
| 21  | 16    | MARQUEZ ÁLVAREZ CARLA ELIZABETH     | LOGISTICA         | SIN BECA | 14     | 45     | 53     | 64     |          |        |               |
| 22  | 17    | MEJÍAS GREZ VÍCTOR IGNACIO          | RECURSOS HUMANOS  | BECA     | 60     | 50     | 57     | 53     |          |        |               |
| 23  | 18    | MOLINA CONTRERAS EMILIO ANDRÉS      | ADMINISTRACIÓN    | SIN BECA | 70     | 31     | 64     | 34     |          |        |               |
| 24  | 19    | MORAGA AMADO NICOL ALEJANDRA        | ADMINISTRACIÓN    | SIN BECA | 65     | 69     | 17     | 23     |          |        |               |
| 25  | 20    | MORENO ROSSI IGNACIO ANDRÉS         | LOGISTICA         | BECA     | 50     | 52     | 58     | 62     |          |        |               |
| 26  | 21    | MUÑOZ MANDUJANO LORENA ALEJANDRA    | ADMINISTRACIÓN    | SIN BECA | 25     | 30     | 29     | 60     |          |        |               |
| 27  | 22    | NAVARRETE ADASME FABIÁN ANDRÉS      | LOGISTICA         | SIN BECA | 59     | 56     | 56     | 38     |          |        |               |
| 28  | 23    | ORMEÑO CONTRERAS VICTORIA FERNANDA  | COMERCIO EXTERIOR | SIN BECA | 65     | 70     | 55     | 30     |          |        |               |
| 29  | 24    | RIVAS VIVAS ÍTALO ANTONIO           | RECURSOS HUMANOS  | SIN BECA | 30     | 60     | 46     | 37     |          |        |               |
| 30  | 25    | ROJAS LILLO JAIR TORIBIO            | LOGISTICA         | SIN BECA | 36     | 58     | 35     | 62     |          |        |               |
| 31  | 26    | SALAZAR ARANGUIZ PABLO ANDRÉS       | COMERCIO EXTERIOR | SIN BECA | 24     | 62     | 42     | 22     |          |        |               |
| 32  | NOTAS |                                     |                   |          |        |        |        |        |          |        |               |

0:23

14-12-2019



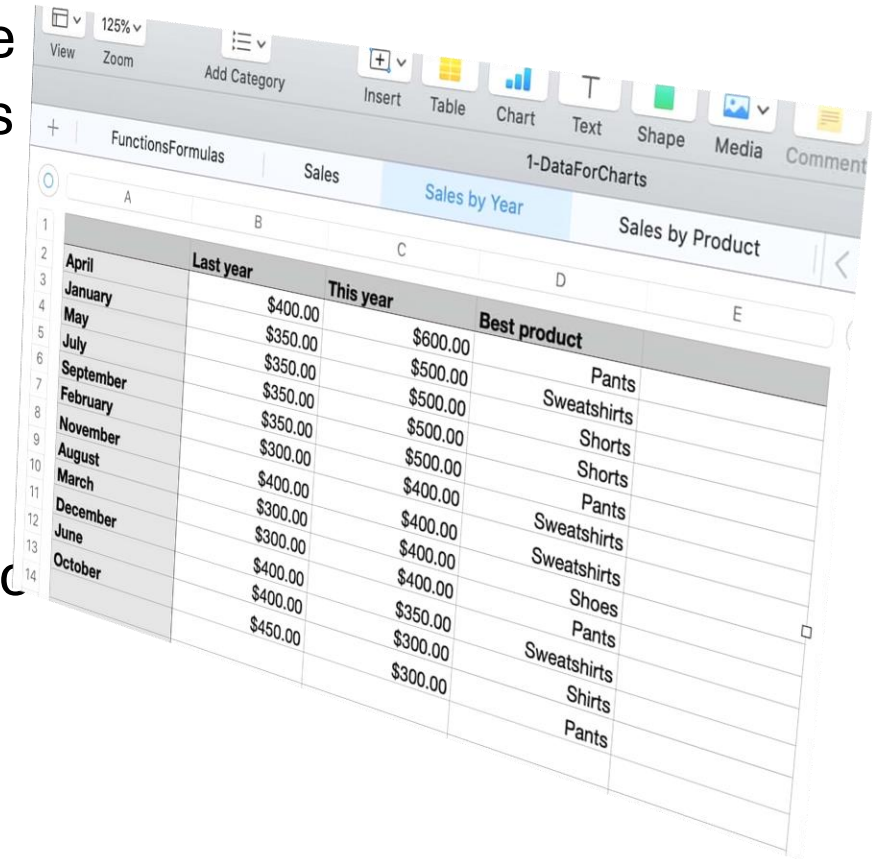


## HOJAS DE CALCULO

Una hoja de cálculo es una aplicación de software que permite organizar, analizar y almacenar datos en formato de tabla.

Consiste en una cuadrícula (**Matriz**) de celdas dispuestas en filas y columnas,

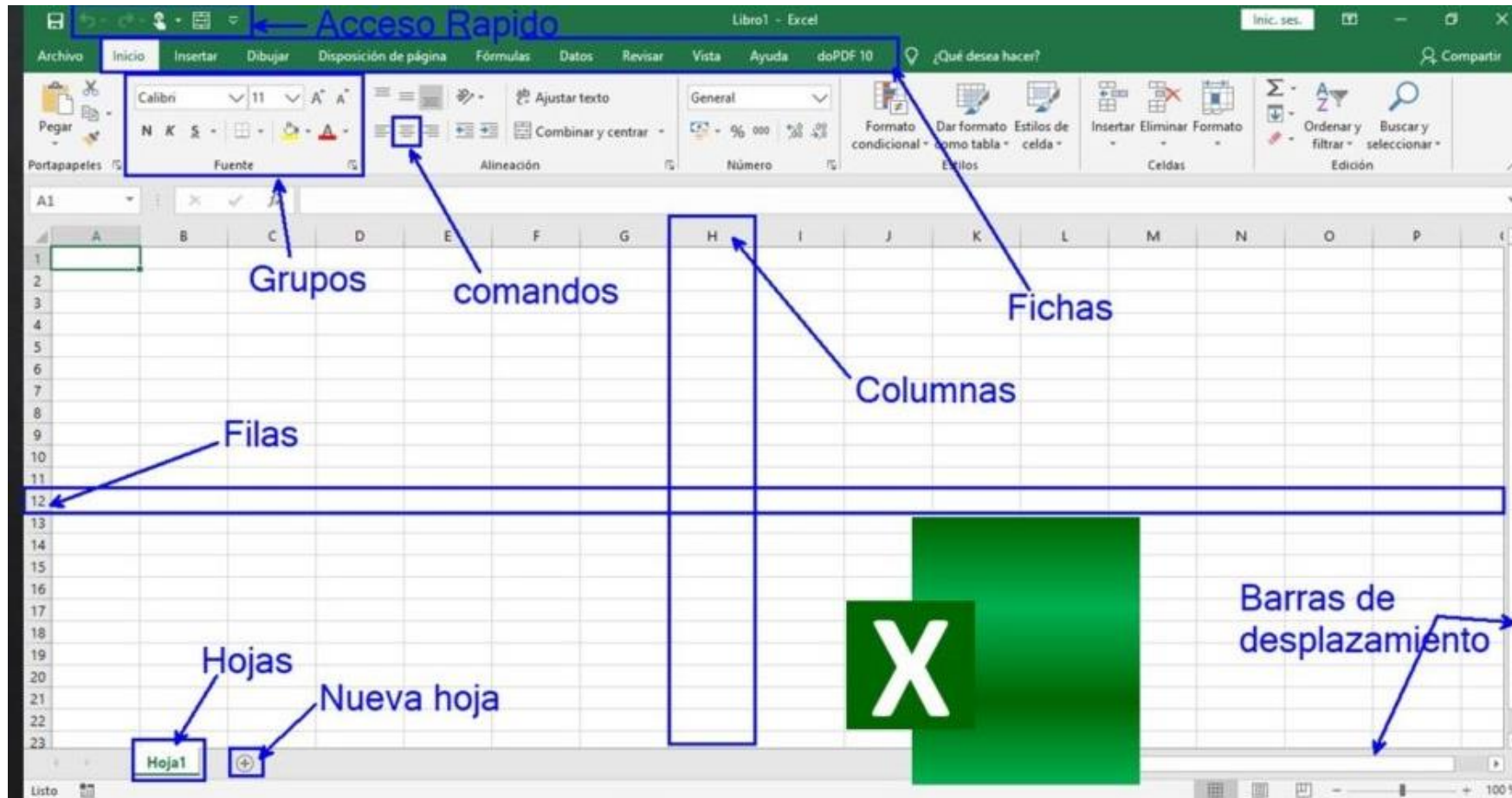
Cada celda puede contener datos de diferente tipo: numéricos, texto o fórmulas que realizan cálculos automáticos sobre los datos.



The screenshot shows a spreadsheet application interface. The top menu bar includes options like View, Zoom, Add Category, Insert, Table, Chart, Text, Shape, Media, and Comment. Below the menu, there are tabs for FunctionsFormulas, Sales, and 1-DataForCharts. The main data area is a table with columns A, B, C, D, and E. The table has a header row with 'Last year', 'This year', and 'Best product'. The rows list months from April to October. The 'Best product' column lists various items like Pants, Sweatshirts, Shorts, Shoes, and Shirts. The numerical values represent sales figures.







|           | Last year | This year | Best product |
|-----------|-----------|-----------|--------------|
| April     |           |           |              |
| January   | \$400.00  | \$600.00  | Pants        |
| May       | \$350.00  | \$500.00  | Sweatshirts  |
| July      | \$350.00  | \$500.00  | Shorts       |
| September | \$350.00  | \$500.00  | Shorts       |
| February  | \$350.00  | \$500.00  | Pants        |
| November  | \$300.00  | \$400.00  | Sweatshirts  |
| August    | \$400.00  | \$400.00  | Sweatshirts  |
| March     | \$300.00  | \$400.00  | Shoes        |
| December  | \$300.00  | \$400.00  | Pants        |
| June      | \$400.00  | \$350.00  | Sweatshirts  |
| October   | \$450.00  | \$300.00  | Shirts       |
|           |           | \$300.00  | Pants        |

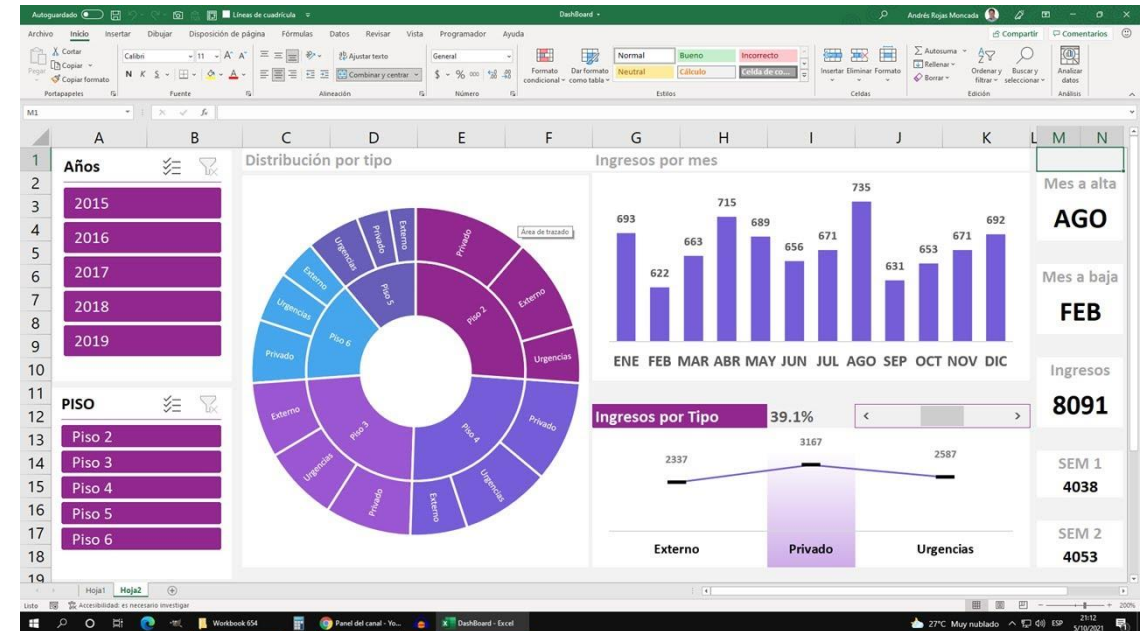
# Hojas de cálculo – componentes básicos



## HOJAS DE CALCULO

### En ciencia de datos:

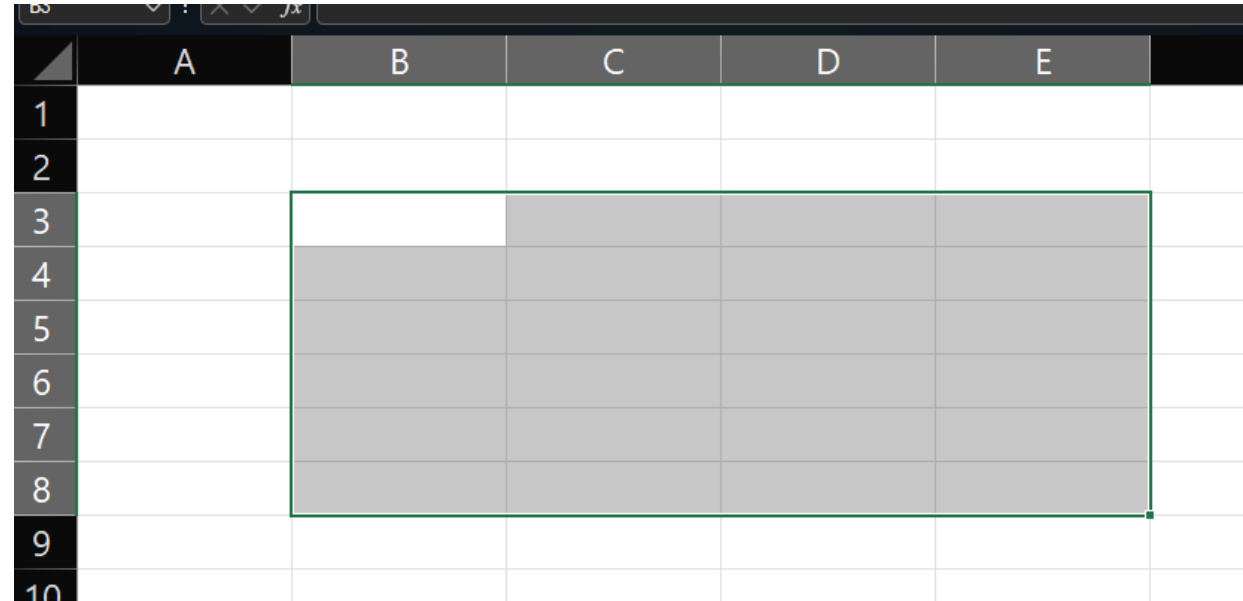
-  Análisis descriptivo y exploratorio de datos
-  Manipulación y limpieza de datos
-  Modelado y análisis predictivo (Trend, Forecast)
-  Automatización y scripting (Macros, script)
-  Visualización
-  Programacion



## Hojas de cálculo

### Organización matricial

Filas >> Numeros  
Columnas >> Letras



The image shows a portion of an Excel spreadsheet. The columns are labeled A, B, C, D, and E. The rows are numbered 1 through 10. A range of cells from B3 to E8 is highlighted in gray, indicating a selected range.

|    | A | B | C | D | E |
|----|---|---|---|---|---|
| 1  |   |   |   |   |   |
| 2  |   |   |   |   |   |
| 3  |   |   |   |   |   |
| 4  |   |   |   |   |   |
| 5  |   |   |   |   |   |
| 6  |   |   |   |   |   |
| 7  |   |   |   |   |   |
| 8  |   |   |   |   |   |
| 9  |   |   |   |   |   |
| 10 |   |   |   |   |   |

Índices de Celda  
**Columna-Fila**

**A1, B3, E8, R200, K80**

RANGO: bloque contiguo, se indica con celdas extremos de la diagonal

**B3:E8**



# Hojas de cálculo

## Desplazamientos

- Moverse una pantalla a la derecha en la hoja.

**Alt** + **Avpág**

- Moverse una pantalla a la izquierda en la hoja.

**Alt** + **Repág**

- Moverse a la hoja siguiente.

**Ctrl** + **Avpág**

- Moverse a la hoja anterior.

**Ctrl** + **Repág**

**Otros atajos: Ctrl, Shift, Alt .**

- Moverse a la siguiente esquina de un rango seleccionado.

**Ctrl** + **.**

- Moverse a la celda A1 o a la celda superior izquierda visible en la hoja.

**Ctrl** + **Inicio**

- Moverse a la última celda utilizada del rango actual.

**Ctrl** + **Fin**

- Moverse al siguiente libro abierto.

**Ctrl** + **Tab**


- Moverse al extremo de la fila o columna actual de acuerdo a la tecla de dirección pulsada.


**Ctrl** + **Tecla dirección**




# Hojas de cálculo

## Funciones $f(x)$

 **Una función en hojas de cálculo es una fórmula que ya se está predefinida**

 Ejecuta los cálculos de diferente tipo: numéricas, textuales, de búsqueda, de fecha, estadísticas, de ingeniería etc.

 Inician con el símbolo “=”

 Relacionan datos constantes, celdas, y rangos (de la misma hoja, otra hoja u otro archivo)

## Grupos de Funciones

- De Búsqueda Y Referencia
- De Texto
- Lógicas
- De Fecha Y Hora
- De Base De Datos
- Matemáticas Y Trigonométricas
- Financieras
- Estadísticas
- De Información
- De Ingeniería
- De Cubo
- Web

# Hojas de cálculo

## Funciones $f(x)$

- Empiezan con un signo de igual (=)
- Seguido por el nombre de la función.
- Un conjunto de argumentos entre paréntesis.

Por ejemplo, la función SUMA

**=SUMA (A1:A5)**

suma los valores en el rango de A1 a A5.

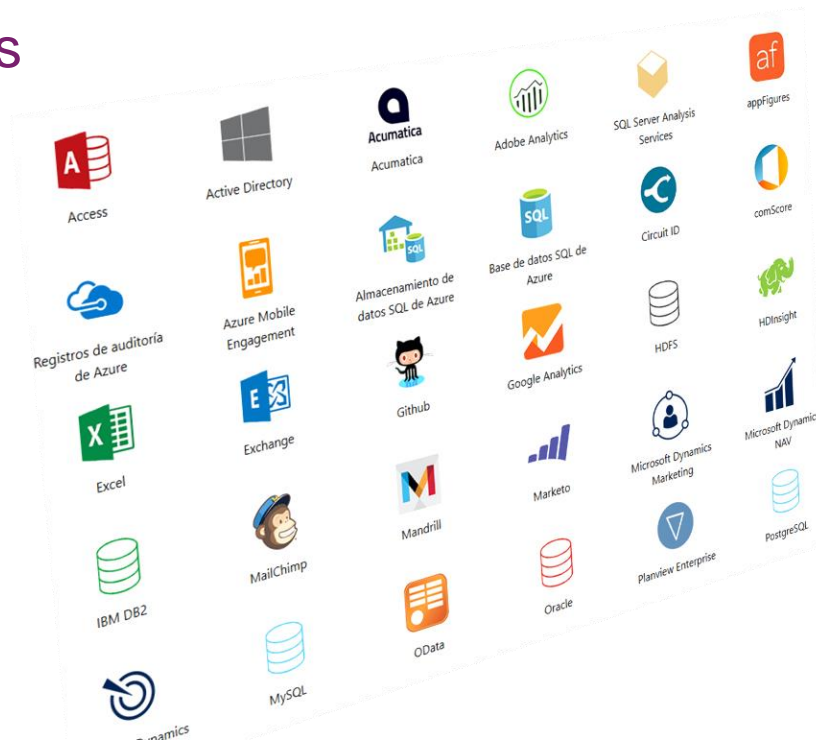
### Algunas funciones

```
=PROMEDIO (A1:A20)  
=BUSCARV("10.";A1:E21;4;FALSO)  
=MAX(A2:A6;30)  
=CONTAR(A2:A7)  
=MAYUSC(O4:Q4)  
=MEDIANA(O3;P3;Q3)
```

## Hojas de cálculo

### Fuentes de datos

- Dataset
- Archivos en Diferente formato:
  - CSV, XLS, TXT, Json...
- Bases de Datos externas
- Apis



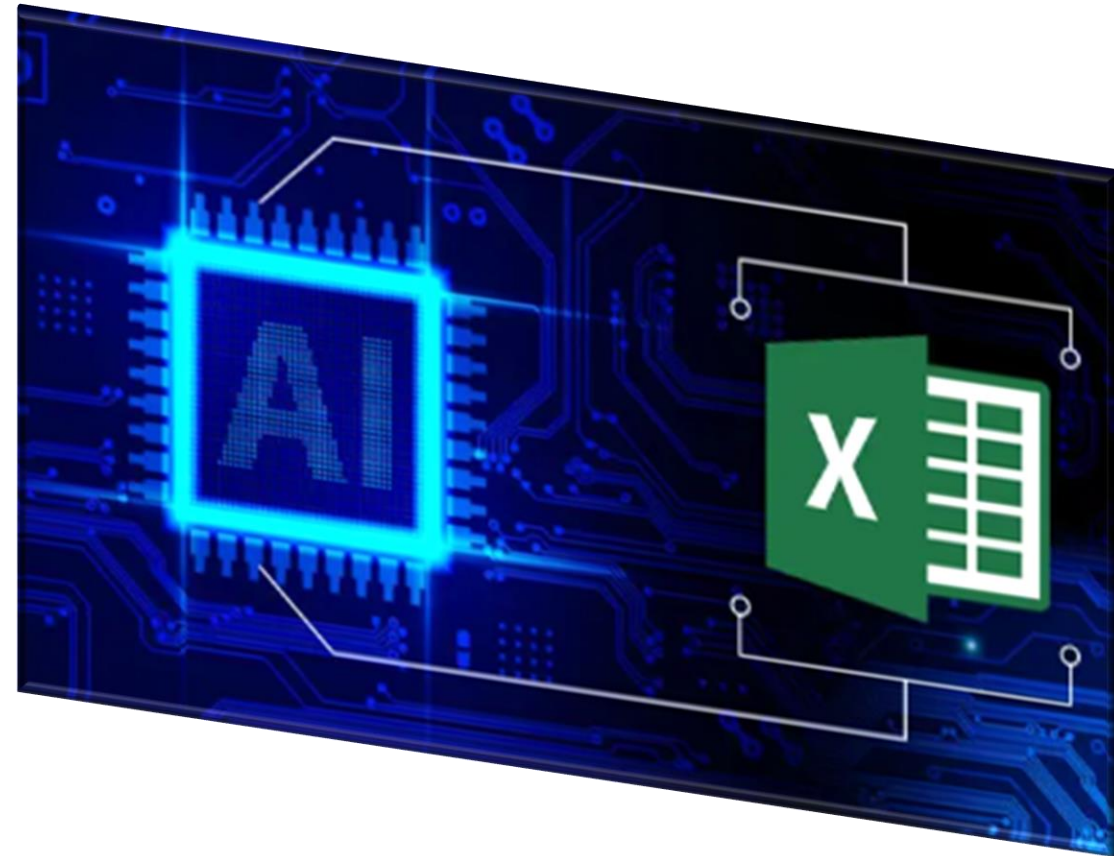
### Formatos

Todos los archivos  
Todos los archivos de Excel  
Archivos de Excel  
Todas las páginas web  
Archivos XML  
Archivos de texto  
Todos los orígenes de datos  
Bases de datos de Access  
Archivos de Query  
Archivos dBase  
Macros de Microsoft Excel 4.0  
Libros de Microsoft Excel 4.0  
Hojas de cálculo  
Áreas de trabajo  
Plantillas  
Complementos  
Barras de herramientas  
Archivos SYLK  
Formato para intercambio de datos  
Copias de seguridad  
Hoja de cálculo de OpenDocument

# Hojas de cálculo

## Utilidad de hojas de cálculo en la Ciencia de datos

- Facilidad de uso y accesibilidad
- Herramientas de análisis y visualización
- Manipulación y limpieza de datos
- Integración con otras herramientas y plataformas
- Automatización y scripting



# Hojas de cálculo

## Desde Python acceder a archivos de excel

### Modulos:

- **xlrd**
- **Pandas**

```
from xlrd import open_workbook

wb = open_workbook("sample.xls")
sheet = wb.sheet_by_index(0)
sheet.cell_value(0, 0)
columns = []
print("Columns")

for i in range(sheet.ncols):
    columns.append(sheet.cell_value(0, i))

print(columns)
```

```
import pandas

df = pandas.read_excel("sample.xls")
print("Columns")
print(df.columns)
```

# Taller





**Gracias**