INTRODUCCIÓN AL PROCESAMIENTO DE DATOS CON PANDAS

Martín David Llauce Santos
@davidllaucesantos

AGENDA

1. Introducción

- Introducción sobre la data.
- Tendencias de crecimiento de la data en el mundo
- ¿ Qué es python?
- Un poco de historia de pandas
- ¿Qué es Pandas?

2. Característica y funcionalidades

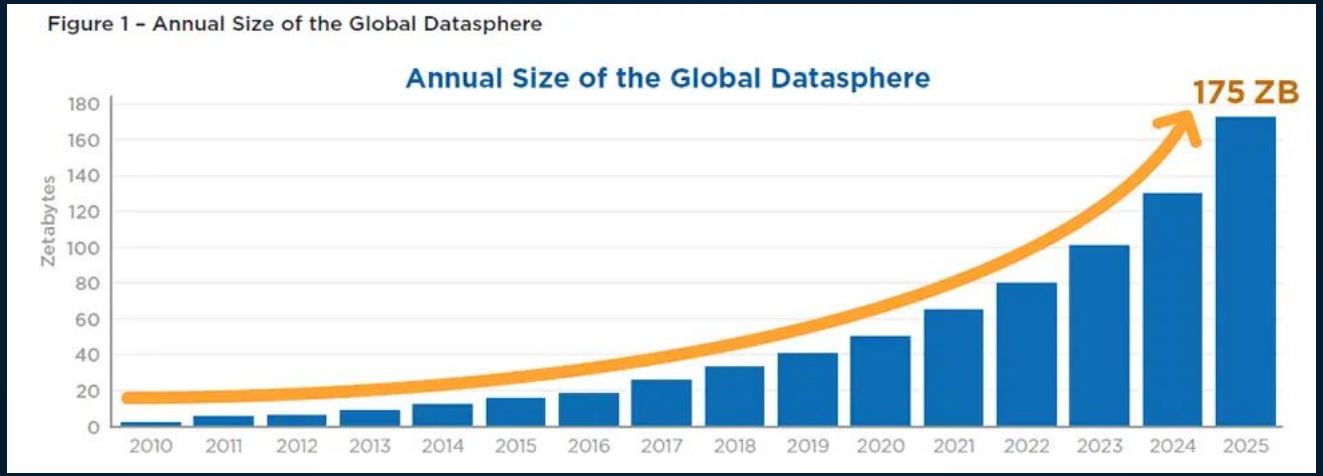
- Características interesantes sobre Pandas
- Comprender el significado y uso de los DataFrames en Pandas
- Ejecutar operaciones básicas con Pandas
- Graficar la data con Matplotlib
- Leer y exportar datos en diferentes formatos de archivos
- Limitaciones sobre Pandas

3. Demo

- Demo: Analizando un archivo con Dataframes.
- Recomendaciones/conclusiones

TENDENCIAS DE CRECIMIENTO DE LA DATA EN EL MUNDO

La data va cada vez en crecimiento lo que hace que sea casi imposible procesarlos en máquinas de escritorio.



Source: IDC Data Age 2025

TENDENCIAS DE CRECIMIENTO DE LA DATA EN EL MUNDO

- Instagram suben una media de más **100 millones de fotos** todos los días(1,000 fotos cada segundo).
- En instagran hacen clic en "Me gusta" **4 mil millones** de publicaciones todos los días.

- Un vehículo autónomo genera **11 Terabytes** de data al día.
- En Twitter se postean **3000 tweets por segundo**

- **167 millones de videos** de Tik Tok vistos en un minuto de Internet
- Los datos son fácil de almacenar pero difícil de analizar

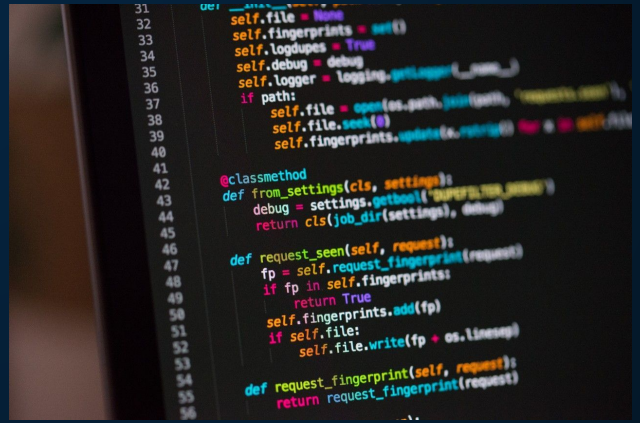
Source: 40 + estadísticas y datos de Instagram para 2022

¿QUÉ ES PYTHON?

Es un lenguaje de programación multiplataforma y multi paradigma que se destaca por su código legible y limpio.

Plataformas como Google, Instagram, YouTube, Spotify, Uber, SpaceX, Tesla y hasta la plataforma que utiliza la CIA usan python.

Creador: Guido Van Rossum



¿QUÉ ES PANDAS?

- Su principal creador, **Wes McKinney**, empezó a desarrollar Pandas en el año 2008 mientras trabajaba en AQR Capital
- Nació por la necesidad que tenía de una herramienta flexible de alto rendimiento para realizar **análisis cuantitativo en datos financieros**.
- Pandas es una biblioteca de software escrita como extensión de NumPy(operaciones matemáticas) y matplotlib(visualización de datos), para manipulación y análisis de datos para el lenguaje de programación Python

```
In [1]: # Import the pandas package
        import pandas as pd

In [2]: # Read airbnb dataset
        airbnb = pd.read_csv('airbnb.csv')

In [3]: # Print pandas DataFrame
        airbnb
```

Out[3]:

	listing_id	5_stars	availability_365	borough	coordinates	host_id	host_name	is_rated	last_review	name	neighbourhood	number_of
0	3831	0.757366	194	Brooklyn	(40.68514, -73.95976)	4869	LisaRoxanne	1	2019-07-05	Cozy Entire Floor of Brownstone	Clinton Hill	
1	6848	0.789743	46	Brooklyn	(40.70837, -73.95352)	15991	Allen & Irina	1	2019-06-29	Only 2 stops to Manhattan studio	Williamsburg	
2	7322	0.669873	12	Manhattan	(40.74192, -73.99501)	18946	Doti	1	2019-07-01	Chelsea Perfect	Chelsea	
3	7726	0.640251	21	Brooklyn	(40.67592, -73.94694)	20950	Adam And Charity	1	2019-06-22	Hip Historic Brownstone Apartment with Backyard	Crown Heights	
4	12303	0.918593	311	Brooklyn	(40.69673, -73.97584)	47618	Yolande	1	2018-09-30	1bdr w private bath. in lofty apt	Fort Greene	
...	...	...	...	...	...	...	...	...	...	...	...	...

CARACTERÍSTICA Y FUNCIONALIDADES

- **Estructuras**:
  - **Series**, son estructuras de una dimensión.
  - **DataFrame**, estructuras de dos dimensiones (tablas).
  - **Panel**, estructuras de tres dimensiones (cubos)

- **Indexación integrada:** Leer y escribir datos entre estructuras de dato en memoria y formatos de archivos variados(excel, sql, csv, otros).
- **Cadenas de operaciones:** Dividir, aplicar y combinar sobre conjuntos de datos, la mezcla y unión de datos.
- **Performance optimizado:** Escrito en Cython o C
- **Operaciones Matemáticas.**
- **Visualización de datos**
- **Integración con Jupyter.**

OPERACIONES CON DATAFRAMES

```
dict_data = {'Name': ["John", "Anna", "Lunna", "Christian"],
             'Location': ["Lima", "Chiclayo", "Arequipa", "Lambayeque"]}

df = pd.DataFrame(data=dict_data)
```

```
# Creamos dataframe con un diccionario
df
```

	Name	Location
0	John	Lima
1	Anna	Chiclayo
2	Lunna	Arequipa
3	Christian	Lambayeque

Operaciones: **https://programmerclick.com/article/84081520515/**

DEMO

El cliente entrega dos archivos(venta e inventario). Se requiere crear un proceso que sanitice los datos y análisis de venta en dinero, venta unidades y stock .
La salida requerida es un solo archivo(csv) que contenga el cruce de venta unidades, venta dinero y stock de acuerdo a las columnas sku producto, código local y fecha.

Columnas salida del archivo:

- Clave: Sku Producto o código producto
- Id Sucursal: Código del local
- Producto: Descripción del producto
- Fecha Pedido: Fecha en formato '%Y-%m-%d'
- Cantidad': Unidades vendidas
- Importe s/impuesto: Venta en dinero
- Existencia: Stock

LIMITACIONES SOBRE PANDAS

- La sintaxis es poco intuitiva de la programación habitual.
- Uso de gran cantidad de memoria RAM, no recomendado para archivos gigantes(mayor 10 millones de registros).
- Reporte de fallos de carga de csv a partir de 18 GB o superación de memoria RAM de 32 GB

CONCLUSIONES/ RECOMENDACIONES

CONCLUSIONES

- Es posible leer y guardar archivos como csv, excel, parquet, hdf5 de disco duro o desde servicios en la nube por ejemplo cloud storage, S3 y Bigquery.
- Pandas puede ser muy útil para manipular, transformar, limpiar data (ETL) en entornos cloud cuando el tamaño de la data en muy variable.
- Pandas permite visualizar fácilmente la data(matplotlib) para detectar outliers con pocos recursos y líneas de código.

CONCLUSIONES

- No es recomendable para archivos mayores a 80 millones de registros.
- Una alternativa es el uso de vaex o dask con sintaxis muy parecida a pandas que utiliza almacenamiento(disco duro) para todo el proceso de transformación(1 billón de registros).
- Utilizar la computación distribuida para archivos o datos en streaming, existen muchos frameworks como Dataflow, Spark, Hadoop, etc. Esto implica un overhead ya que es necesario configurar y administrar los clusters.

RECURSOS

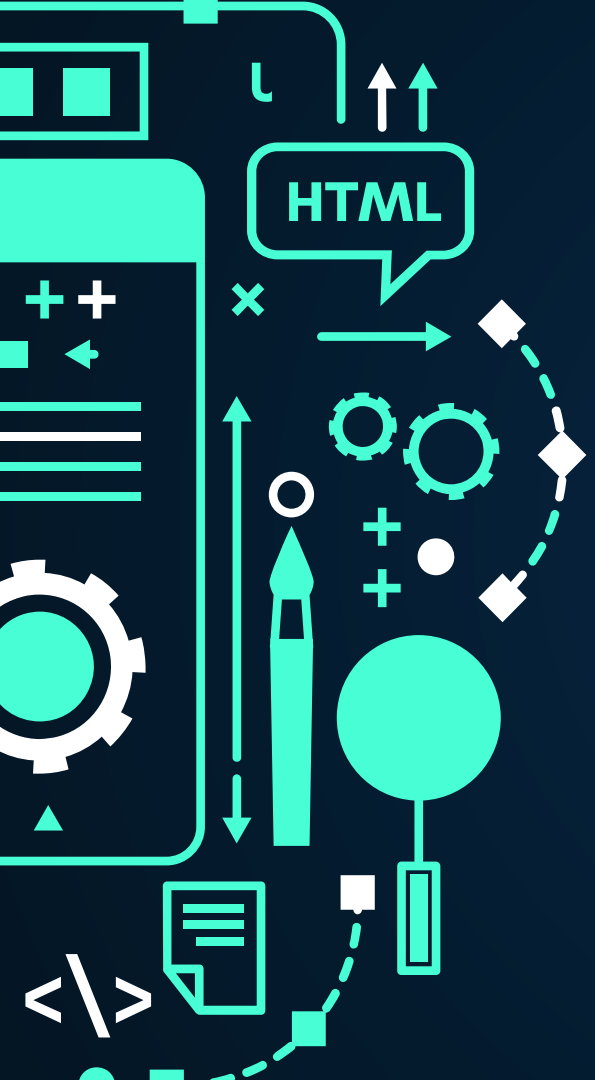
- **Pandas**

  https://pandas.pydata.org/docs/getting_started/index.html
- **Python**

  https://www.python.org/
- **Matplotlib**

  https://matplotlib.org/
- **Slides & Notebook :**

  https://github.com/davidllauce/introduccion_pandas



MUCHAS GRACIAS!

¿Preguntas?

David Llauce
davidllaucesantos@gmail.com
@davidllauce