INSTITUTO PROVINCIAL DE LA ADMINISTRACIÓN PÚBLICA (IPAP)

2022

USO DE NUEVAS TECNOLOGÍAS PARA EL ANÁLISIS DE DATOS EN EL ESTADO

Tomás Barbieri
2021

Clase 1

Objetivos

Comprender el concepto dato, información, tipos de datos, elementos del Big Data y cómo evoluciona la información en el mundo. Etapas del análisis de datos.

Identificar dónde localizar la información, sus características y cómo la organizamos para su procesamiento.

Conocer nuestro ambiente de trabajo, tecnologías y herramientas que vamos a utilizar.

Datos e información

- En el mundo existen un montón de elementos que pueden ser registrables e informatizados. Por ejemplo, listado de personas, lugares turísticos, cantidad de de calles asfaltadas o locales comerciales y su horario de apertura y cierre.
- Un **dato** podemos definirlo entonces como un valor específico de algo de la realidad que decidimos analizar.
- Cuando logramos conectar muchos datos, con un objetivo de estudio y un criterio de análisis, podemos decir que muchos datos se convierten en **información**.
- Intentaremos ver a lo largo de este curso, cómo lograr convertir esos datos que parecen inicialmente inconexos en información valiosa para la toma de decisiones.

Tipos de datos

- **No Estructurados**
  Comúnmente encontramos los datos distribuidos en diferentes formatos y no cuentan con una estructura unificada: por ejemplo archivos Excel, PDF, Word, contenido de emails, datos en teléfonos móviles, audios, videos, redes sociales, etc.

- **Estructurados**
  Se refiere a todos los datos que si se encuentran bajo una estructura, es decir, bases de datos relacionales.
  Se utilizan herramientas de consulta para dichas bases de datos.

  Otro objetivo de este proceso es utilizar datos no estructurados para generar herramientas estructuradas de análisis

Big Data ¿Qué es?

- Podemos definir Big Data como el análisis y gestión de grandes volúmenes de datos y toda la infraestructura física y de software que se necesita para que esto suceda.
- El principal objetivo de este análisis es convertir los datos dispersos en información. Entendiendo con información como una herramienta útil para tomar decisiones en diferentes ámbitos, público, privado, personal, etc.
- Se suelen utilizar varias V para definir qué significa big data: Volumen, Variedad, Velocidad, Veracidad y Valor (entre algunas de ellas)

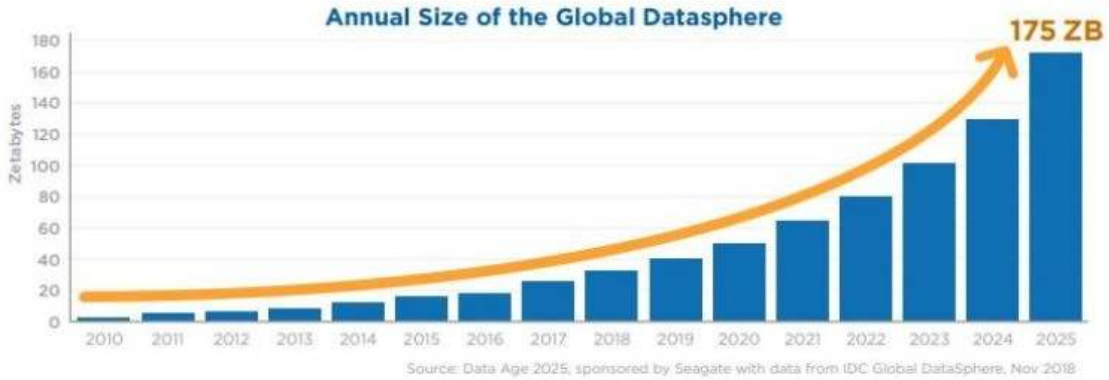
Big Data ¿Qué es?

- **Volumen**: hace referencia a la enorme cantidad de datos generados al correr del tiempo.
- **Velocidad**: se refiere a la enorme velocidad en la que se van generando esos datos y cada vez más.
- **Variedad**: Los datos se encuentran en diferentes formatos, estructurados, no estructurados.
- **Valor**: el valor que se le puede añadir a toda esa información si se conecta, organiza y analiza con un objetivo concreto.
- **Veracidad**: Dentro de todos estos datos, debemos encontrar aquellos que sean reales, confiables y verdaderos.

Para dimensionar

- 1 Terabyte (TB) = 1000 Gigabytes (GB).
- 1 Petabyte (PB) = 1000 Terabytes.
- 1 Exabyte (EB) = 1000 Petabytes.
- 1 Zettabyte (ZB) = 1000 Exabytes = 1 millón de petabytes = 1000 millones de terabytes.

Se estima que para el 2025 la atmósfera de datos mundial sea de 175 ZettaBytes



Etapas del análisis de datos

- **La pregunta:** ¿Qué queremos hacer? ¿Qué información necesitamos? No siempre nos sale la primera pregunta...

- **Recopilación de datos**: obtener de diferentes fuentes de datos la información que necesitamos. ¿Qué fuentes existen? Muchísimas... excel, pdf, documentos, encuestas, datos internet, informes, mucha información NO ESTRUCTURADA.

- **Procesamiento de datos:** Principalmente estructurar los datos en un formato común para su posterior análisis. Por ejemplo, armar un buen excel (o varios) con la información ordenada. Esto puede ser un proceso manual o automatizado dependiendo las fuentes de datos y el volumen.

Etapas del análisis de datos

- **Limpieza de datos**: Esta etapa es clave, ya que los datos previamente organizados y sistematizados pueden estar incompletos, contener duplicados o contener errores.

- **Análisis de los datos**: podemos definirla como una fase fundamental, donde vamos a ordenar, filtrar, segmentar y estudiar los datos. También se hacen "pequeñas visualizaciones" para ir entendiendo el set de datos.

- **Visualización/Comunicación**: Esta parte completa el proceso, es cómo mostrar los resultados en diferentes formatos, gráficos, mapas, etc. Debe pensarse también en clave de comunicación, es decir, cómo generar un relato con la información obtenida.

Tecnologías utilizadas

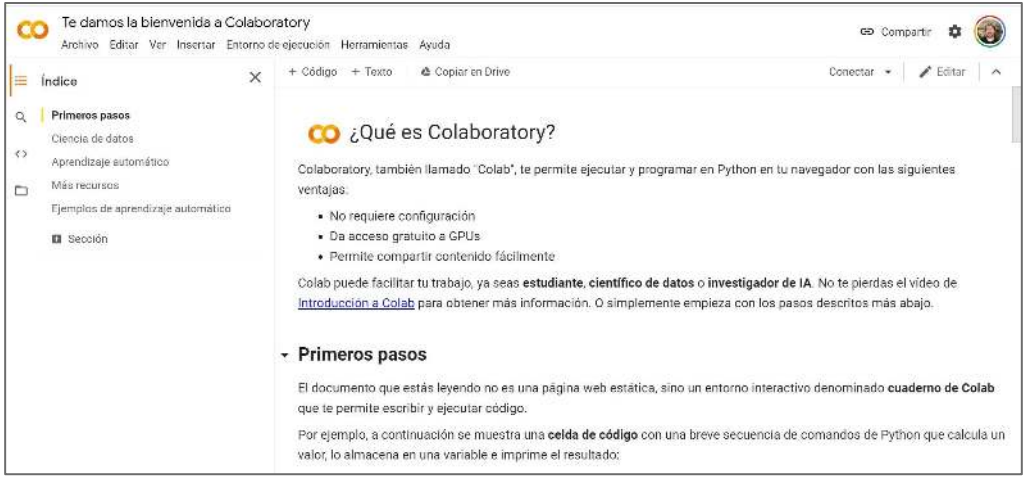
- **Python**: Lenguaje de programación de base para poder ejecutar los comandos con los que van a permitirnos la exploración y armado de la información.
- **Pandas**: Librería de python especial para la gestión de datos, análisis y segmentación de la información.
- **MatPlotLib**: Librería de python que se utiliza para la visualización de datos con gráficos y herramientas visuales.
- **NumPy**: Se usa para algunas operaciones matemáticas dentro del código, trabajo con matrices y con funciones.

Nuestro ambiente de trabajo

- **Google colab:** es una herramienta de Google que está asociada a Google Drive que nos permite crear documentos en donde combinemos análisis de datos y texto que pueda ir explicando lo que vamos haciendo. Todo esto de forma remota.
- Nos permite ejecutar código Python
- Armar textos con Markdown
- Podemos armar una bitácora de lo que estamos realizando
- Visualización en el mismo ámbito de ejecución y escritura del código.

Nuestro ambiente de trabajo

- **¿Como lo instalamos?** Hay un sobre como instalar google collaboratory en nuestro google drive. “Instalar Google Colaboratory”



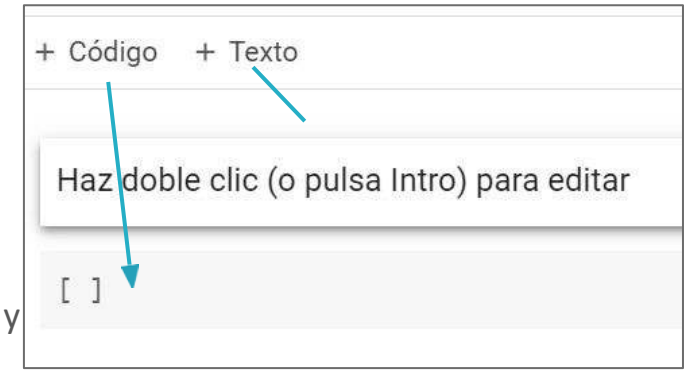
https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=2fhs6GZ4qFMx

¿Cómo funciona Colab?

Un documento en google colab es básicamente una serie de bloques que pueden ser de **Texto** o de **Código**.

Los **bloques de texto** nos permiten agregar palabras en el documento que vayan explicando lo que vamos haciendo, anticipando comandos o simplemente agregando comentarios.

Los **bloques de código** (en Python) son las instrucciones que vamos a ir usando para ir procesando los datos, cargarlos o mostrarlos en gráficos.



Es probable que para quienes no tengan experiencia en programación sea un poco complejo de entender, pero en principio nos conformamos con entender qué hace el código y copiar instrucciones de otros lugares.

Bloques de texto

Los **bloques de texto** pueden tener varios formatos, nos permiten añadir títulos, letra negrita, cursiva, listados, enlaces e incluso imágenes.

Dicho formato lo genera con el lenguaje de marcado Markdown (en el enlace hay detalles de cómo se usa). Del lado izquierdo del bloque añadimos el contenido, y del lado derecho nos va mostrando como queda.

Bloques de código

Los **bloques de código** pueden ser un poco complejos de entender, pero nos vamos a detener únicamente en “usar” el código y no tanto en entender cómo funciona internamente.

Como se mencionó previamente, el código es en Python y nos permiten ejecutar diferentes instrucciones (que iremos viendo a lo largo del curso).

Importar librerías, abrir los archivos de datos, realizar observación de datos, observar las columnas, realizar filtrados en la información, imprimir gráficos, etc.

Para ejecutar el código

Funcionamiento del colab

En la clase 2 veremos como interactuar con nuestro google colab con más profundidad y con el formato de video para la comprensión de su funcionamiento.

Links de interés

- Python
- Pandas
- Matplotlib
- Anaconda navigator
- Big data
- Big data (global datasphere)
- Video Big data
- Datos nacionales
- Proceso de análisis de datos
- Exploración de datos
- Data science roadmap

Links de interés

- Dato
- Datos estructurados vs no estructurados
- Definiciones big data
- Catálogo de datos de la provincia de Buenos Aires

ipap.gba.gob.ar