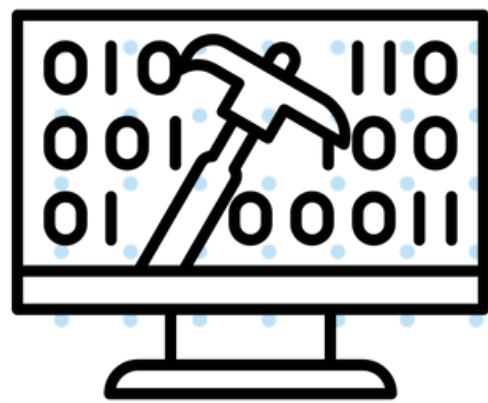


Ecosistemas de ingeniería de datos




× Herramientas de procesamiento de datos

Existen muchas herramientas para analizar y procesar datos, cada una ofrece ventajas y desventajas respecto a las demás. En esta lección se estudiarán algunas de las herramientas disponibles en el mercado.

· · Hojas de cálculo (Excel , Google sheets, calc, entre otras)

Las hojas de cálculo son los programas más comunes para el tratamiento inicial de datos y el almacenamiento de la información. Son populares debido a que los usuarios no necesitan un conocimiento avanzado en lenguajes de programación y permiten realizar tareas de forma intuitiva con su interfaz gráfica. Dentro de sus herramientas internas permite analizar datos, crear tablas para análisis descriptivo, crear gráficos de diversos tipos, limpiar y ver datos de forma manual e incluso permite agregar funcionalidades complejas con las macros que son segmentos de código que se pueden adjuntar a los libros de cálculo.

Es una herramienta completa, sin embargo, las hojas de cálculo suelen estar muy limitadas en cuanto a cantidad de datos que pueden albergar. Esto se debe a que tienen límites para procesar más de 16.384 columnas o 1.048.576 filas. En los escenarios modernos, un dataset puede sobrepasar fácilmente esos límites con lo que las hojas de cálculo pueden quedarse cortas. Adicionalmente, para sus cálculos requieren una capacidad grande de procesamiento (uso de CPUs) haciendo que, a mayor cantidad de datos por procesar, más lento puede tornarse el manejo de un conjunto  de datos.

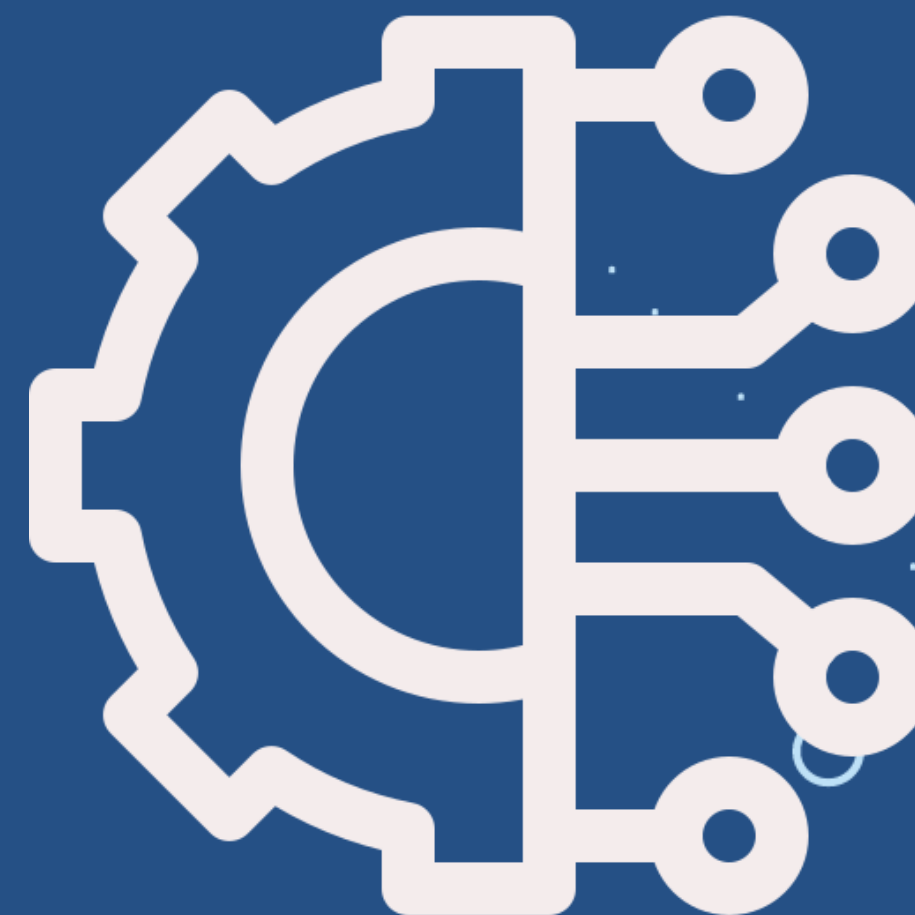
R

R es un lenguaje de programación empleado para estadística computacional y generación de gráficos. Es un lenguaje fácil de aprender, compatible con formatos modernos de almacenamiento de datos.

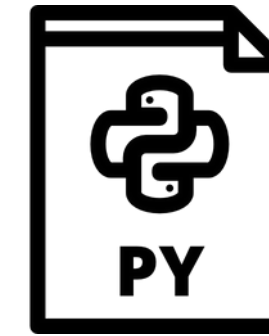


Es de código abierto y de uso libre, por lo que no se requieren licencias (como en la mayoría de las hojas de cálculo) y es modular, por lo que se puede extender su funcionalidad instalando librerías. Cualquier persona puede contribuir a la creación de librerías y estas pueden ser inspeccionadas por cualquiera para saber como funcionan (código abierto). Entre las ventajas de R se tiene que funciona en cualquier equipo de computo, como computadores Mac, Windows, Linux e incluso en maquinas no tan convencionales, como las Raspberry pi o las playstation 3.

Para usar R se suele emplear un entorno de desarrollo integrado (IDE) llamado R studio, que incluye una consola, editor de texto y herramientas para trabajar de forma automatizada con R.



Python



TIC

Python es un lenguaje de programación muy popular, con cientos de herramientas de alto nivel para el manejo de datos, la creación de aplicaciones y la programación en general. Se basa en el intérprete de Python que es un programa capaz de leer instrucciones y ejecutarlas. Su popularidad se debe a la facilidad de su sintaxis (escritura) y que es fácil de aprender.

También permite crear prototipos y segmentos de código de forma rápida. Como es un lenguaje de programación interpretado, no se necesita compilar un programa para obtener resultados. Actualmente es el de mayor crecimiento en la encuesta de Stack overflow y tiene muchas herramientas para el análisis de datos, la visualización, el aprendizaje de máquina, el procesamiento distribuido y muchas posibilidades adicionales que ofrece.

Apache spark



Spark es un entorno de trabajo (framework) de código abierto para el procesamiento de datos en clústeres computacionales. Proporciona una interfaz unificada para analizar datos distribuidos en memoria con lo que es más rápido que sistemas tradicionales de almacenamiento en discos locales. Una de sus ventajas es que permite crear programas empleando Python, R o Scala. Esta herramienta soporta la creación de pipelines de datos, consultas de datos en streaming, herramientas interactivas de visualización y herramientas de aprendizaje automático.

Google cloud autoML



Es un conjunto de herramientas orientadas a datos y machine learning proporcionadas por Google para que el trabajo con datos sea de alto nivel, es decir, no hay que programar para crear rutinas de procesamiento. También permite la creación de modelos de aprendizaje de máquina de forma fácil conectando los datos desde orígenes y visualizando fácilmente la información de lo que sucede durante las etapas de entrenamiento y validación.



Power BI

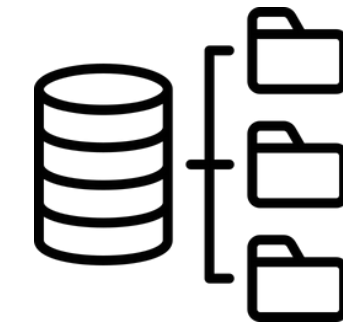


Es una herramienta de procesamiento y visualización interactiva de datos. Está orientado al bussines intelligence (BI) y diseñado para asociar datos desde orígenes (bases de datos, archivos en la nube entre otros) cargar los datos, procesarlos y visualizarlos desde su interfaz gráfica. Su principal ventaja es la creación de visualizaciones dinámicas, desde las que se pueden consultar datos manipularlos en alto nivel para cambiar una vista dependiendo del contenido de los datos.



Es muy popular y empleada como solución para publicar el análisis de resultados en la web y facilitar a cualquier persona que incluso no sepa de programación, crear vistas y cambiar los contenidos de los gráficos a unos cuantos clics.

Databricks



TIC

Es una plataforma unificada de análisis de datos basada en apache Spark. Proporciona una interfaz fácil de usar para los usuarios y colaborativa para el trabajo en equipo. Se enfoca en la creación de pipelines de datos y las aplicaciones de aprendizaje automático. Con databricks es fácil escalar soluciones construidas sobre apache spark en la nube, sin preocuparse por la infraestructura ya que se escala de forma automatizada. Adicionalmente databricks ofrece la integración con herramientas y servicios en la nube, como son los sistemas de AWS, Azure, Delta Lake y tensorflow.



Snowflake



TIC

Es una plataforma de almacenamiento de datos en la nube con servicio de data warehousing, permite a las organizaciones almacenar grandes volúmenes de datos de manera eficiente y escalable. Está diseñada para operar en la nube, ofreciendo alto rendimiento y facilidad de uso. En el caso de snowflake Se pueden separar las cargas de trabajo del almacenamiento, con lo que los datos pueden almacenarse de forma independiente al sistema de procesamiento. También incluye sistemas de modelado compartido de datos, para que, si múltiples aplicaciones intentan acceder a los datos lo puedan hacer sin problemas.

Informática



TIC

Conjunto de herramientas que permiten realizar tareas de procesamiento, análisis de datos así como la manipulación y presentación de datos. Sus ventajas son que permite la automatización de tareas repetitivas de gestión de datos, permite realizar cálculos complejos y el desarrollo de software que procesa datos de forma eficiente.