Big Data (Análisis de grandes volúmenes de datos)

DIPLOMADO, MÓDULO III

Dr. José Luis Cendejas Valdez, UTM

Dr. Heberto Ferreira Medina, IIES-UNAM

Dr. Sergio Rogelio Tinoco Martínez, ENES-UNAM

Actividades sincrónicas:

Classroom: https://classroom.google.com/c/NzM0NTk3Mzc3MzIz?cjc=5wc4s6l

ZOOM: https://cuaieed-unam.zoom.us/j/87963800329?pwd=il8rD7Ai14yqCN4nSKxijXRvxCkwjC.1

DIPLOMADO EN APRENDIZAJE MÁQUINA Y PROFUNDO APLICADO A GRANDES VOLÚMENES DE DATOS (Big Data)
PROYECTO PAPIME PE103124

Objetivos

- **Objetivo del módulo:**
  - Desarrollar aplicaciones para la búsqueda, extracción, procesamiento, interpretación y visualización de grandes volúmenes de datos.

- **Específicos:**
  - Aportar la capacidad para procesar, interpretar y graficar datos previamente obtenidos.
  - Aplicar sus conocimientos y habilidades para análisis de grandes cantidades de datos.

Calendario:

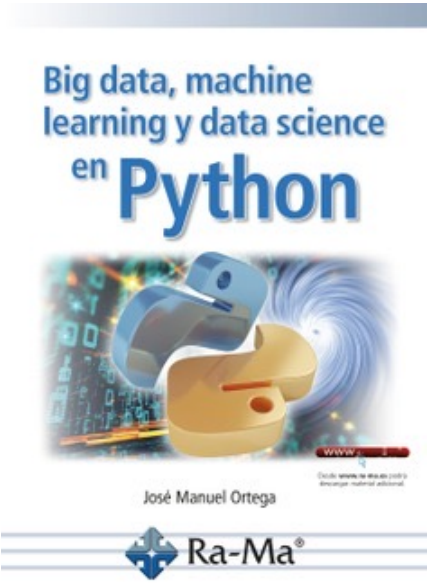
Sesiones síncronas, martes y jueves de 17:00 a 20:00 hrs.

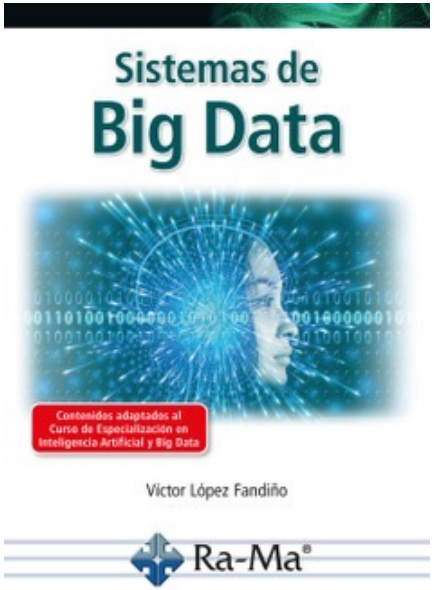
Sesión	Fechas	Actividades	Entregas	Puntos
1	Martes 26 de noviembre	El uso del Big Data en la Ciencia de Datos.		
2	Jueves 28 de noviembre	Capítulo libro, Web Scraping (Análisis del Big Data con ML).	Tarea 1	3
3	Martes 3 de diciembre	Uso de almacenamiento en la Red (NAS).	Tarea 2	2
4	Jueves 5 de diciembre	Uso de HDFS, DASK.	Tarea 3	3
5	Martes 10 de diciembre	Análisis de grandes volúmenes de datos para el DL (Aplicación). Práctica 4.	Tarea 4	3
6	Martes 17 de diciembre	Uso de Bases de Datos (DB) no estructuradas.	Tarea 5	2
7	Jueves 19 de diciembre	Despliegue de un proyecto de Big Data con servicios Web.	Tarea 6	2
8	Jueves 19 de diciembre	Cierre de sesiones síncronas	Capítulo de libro	3
9	Semana 6 al 10 de enero 2025.	Sesiones asíncronas de apoyo **Cierre 17** de enero de 2025	Puntos	13

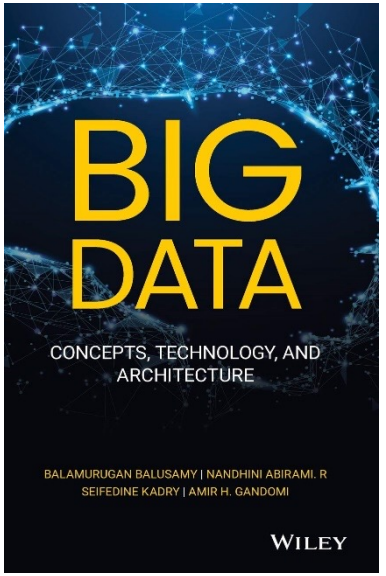
Actividades sesión 1.

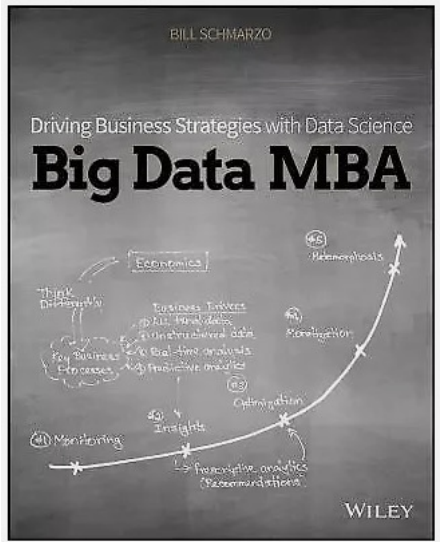
- Evaluación diagnóstica.
- Plataforma de curso en línea o a distancia.
- Herramientas de software (Anaconda Navigator):
  1) Python; instalar
  2) Jupyter-Lab; instalar
  3) Matplotlib; instalar
  4) Numpy; instalar
  5) Pandas; instalar
  6) Scikit-Learn; instalar
  7) Keras y TensorFlow-GPU
  8) DASK
- Equipo de cómputo, conexión a Internet y GPU reciente.
- Módulo II: https://classroom.google.com/c/NzAxNDM0OTY3MTk2?cjc=lhqzs2h
- Módulo I: https://classroom.google.com/c/NjY2Njg3OTA1NzQ0?cjc=ka5qs3c

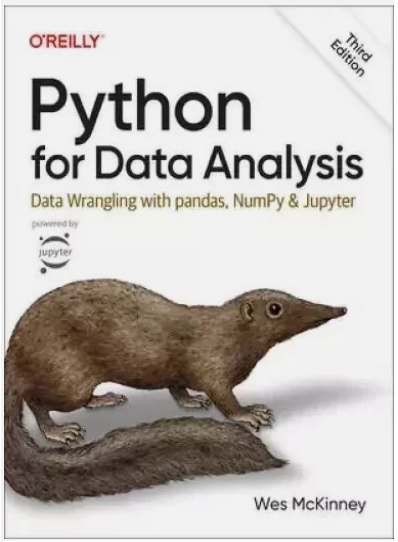
Bibliografía:











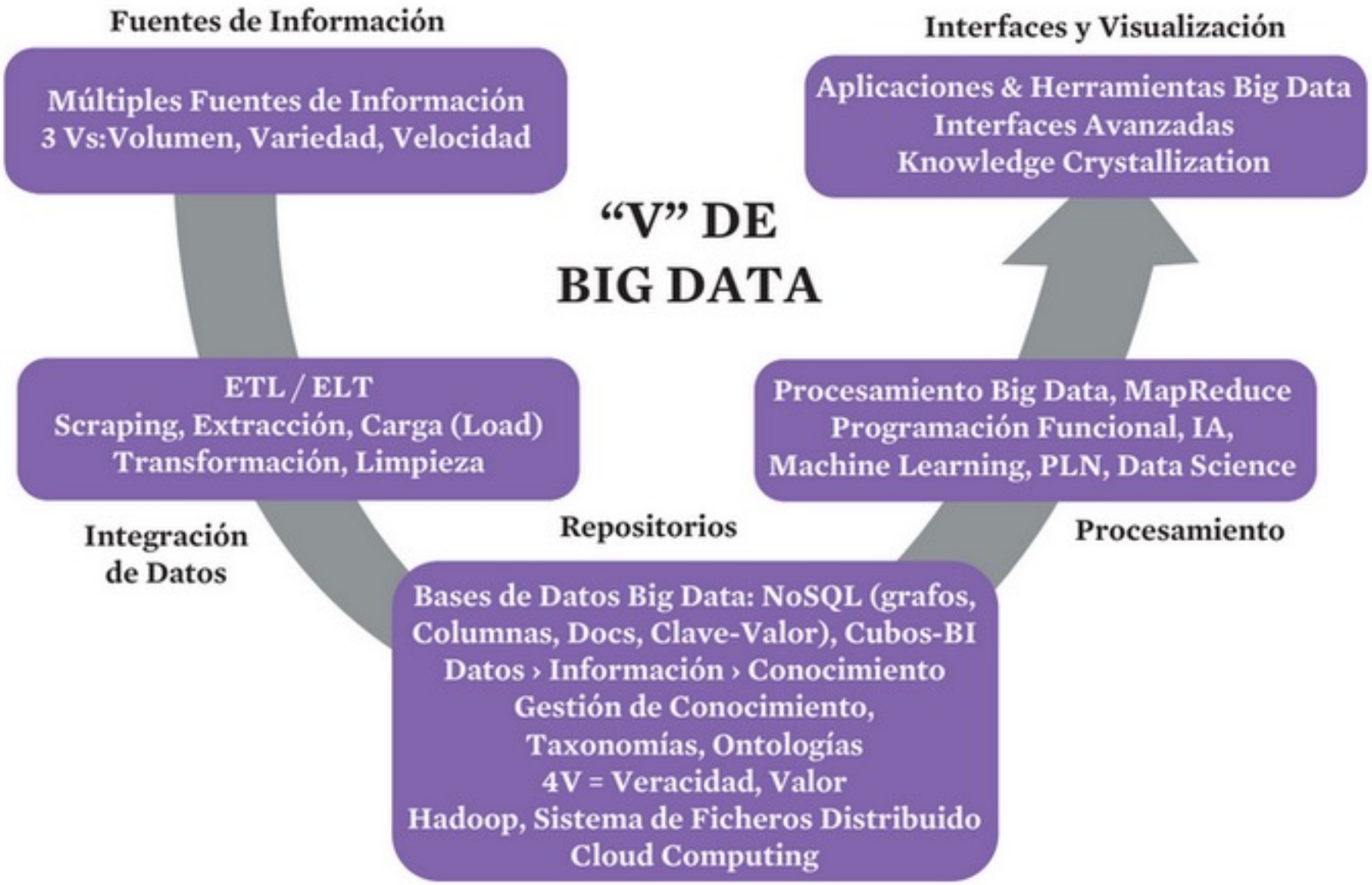
Big Data:

- Los “macrodatos”, también llamados **datos masivos**, **inteligencia de datos**, **datos a gran escala, grandes volúmenes de datos** es un término que hace referencia a conjuntos de datos tan grandes y complejos que precisan de aplicaciones informáticas no tradicionales de procesamiento de datos para tratarlos adecuadamente (Wikipedia)
- Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho (RAE)

Temas de Big Data en libros:

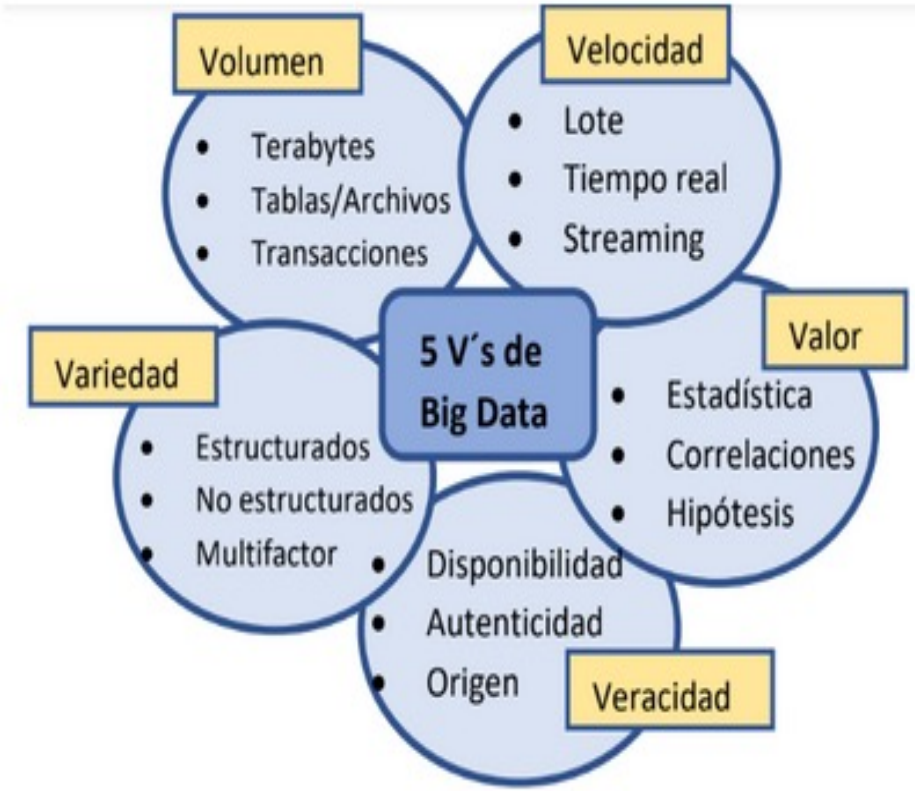
Potencial del Big Data en negocios
Ciencia de datos
Ciencia de datos para negocios
Construcción del Big Data
Modelo de madurez del Big Data
Estrategias de implementación
Lago de datos
Técnicas de análisis
Limpieza de datos
Minería de datos

Big Data


Fuentes de Información
Múltiples Fuentes de Información
3 Vs:Volumen, Variedad, Velocidad
Interfaces y Visualización
Aplicaciones & Herramientas Big Data
Interfaces Avanzadas
Knowledge Crystallization
"V" DE
BIG DATA
ETL / ELT
Scraping, Extracción, Carga (Load)
Transformación, Limpieza
Procesamiento Big Data, MapReduce
Programación Funcional, IA,
Machine Learning, PLN, Data Science
Integración
de Datos
Repositorios
Procesamiento
Bases de Datos Big Data: NoSQL (grafos,
Columnas, Docs, Clave-Valor), Cubos-BI
Datos › Información › Conocimiento
Gestión de Conocimiento,
Taxonomías, Ontologías
4V = Veracidad, Valor
Hadoop, Sistema de Ficheros Distribuido
Cloud Computing

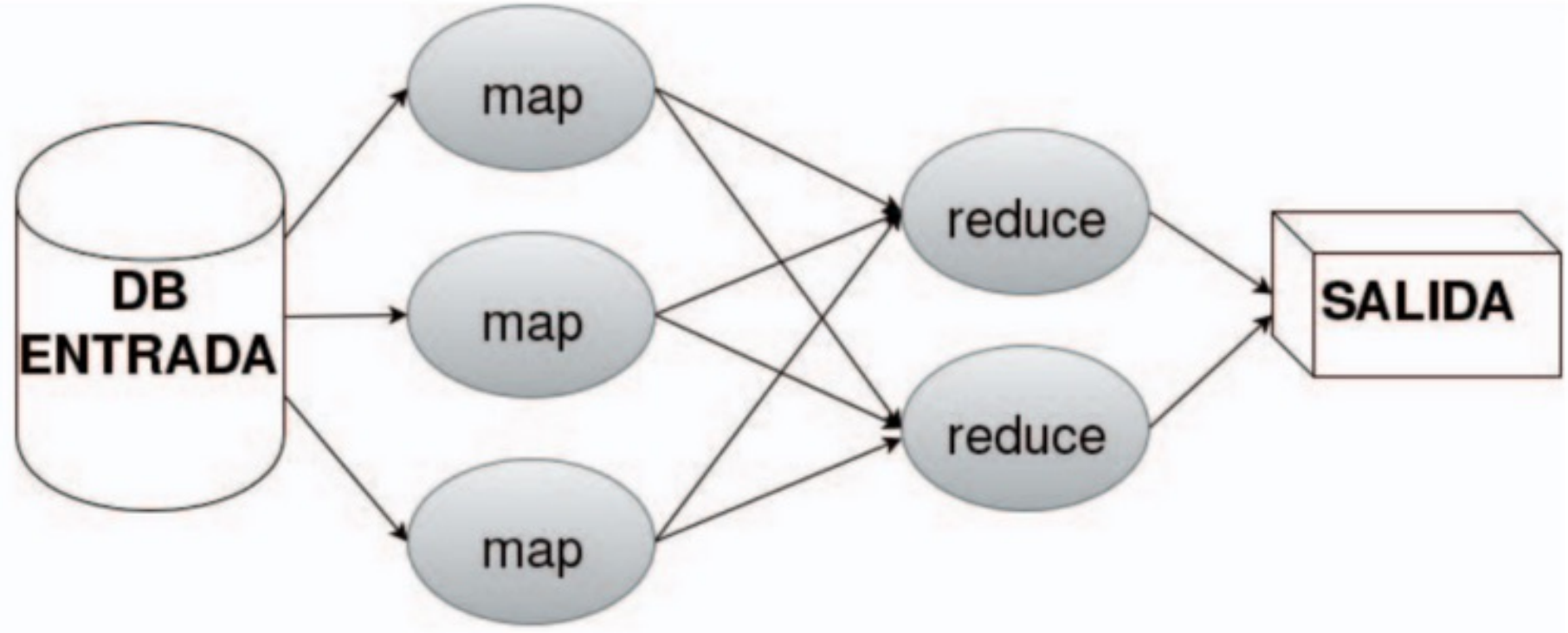
Fuentes de información

V´s del Big Data

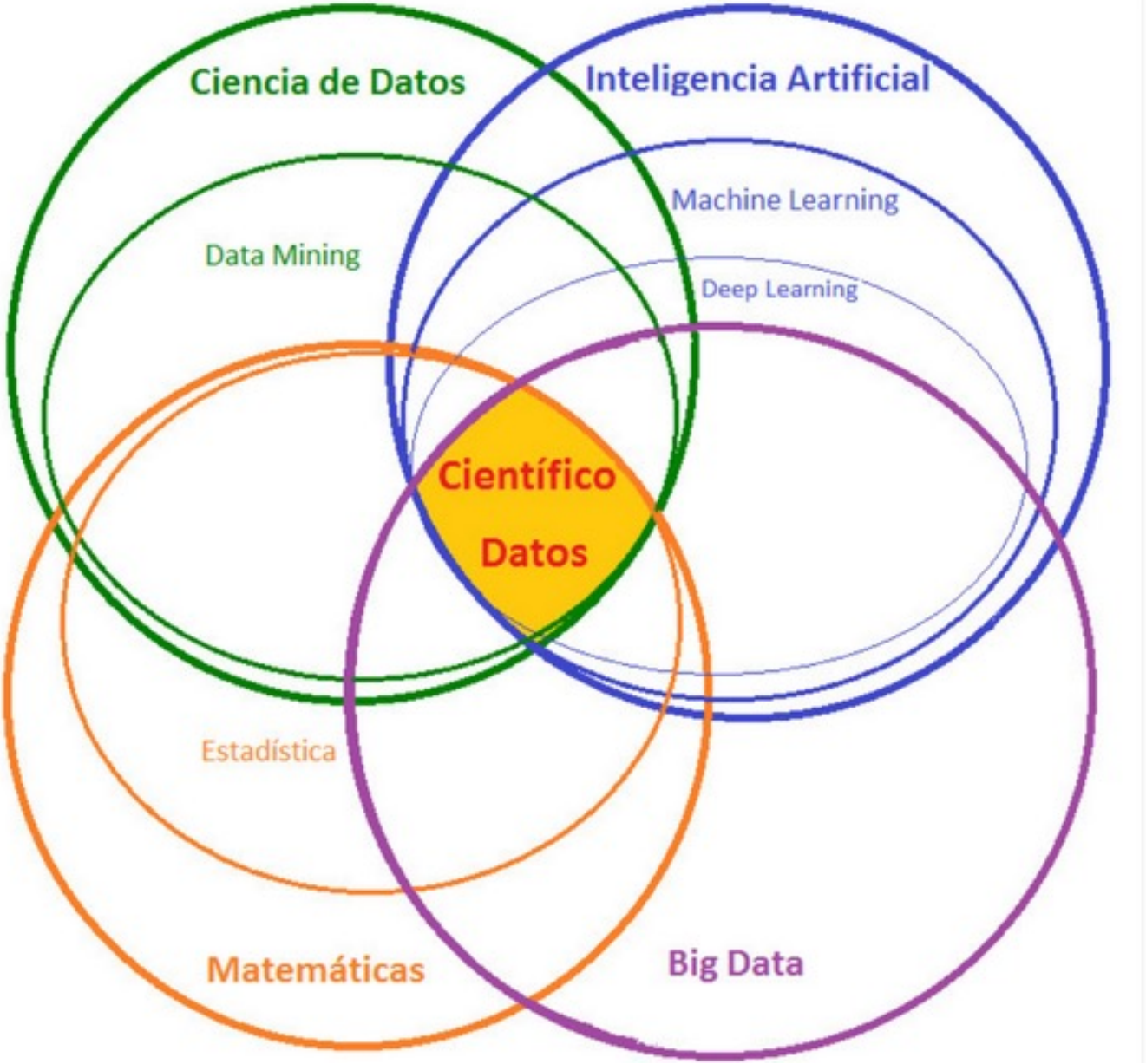


Integración de datos

Map - Reduce



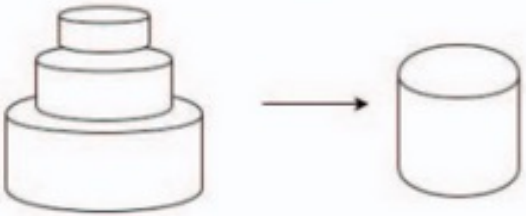
Ciencia de datos



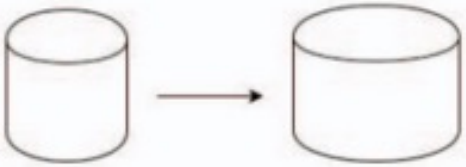
Procesamiento de datos

Limpieza de datos

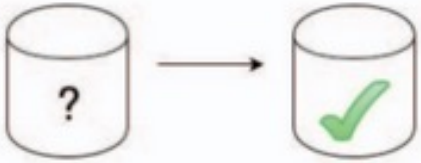
Normalización de datos



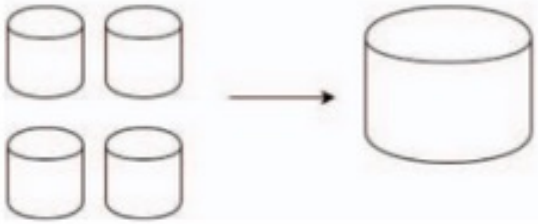
Transformación de datos



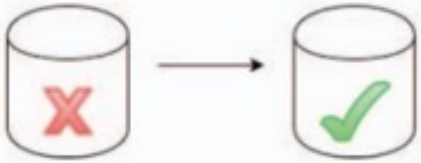
Imputación de valores perdidos



Integración de datos

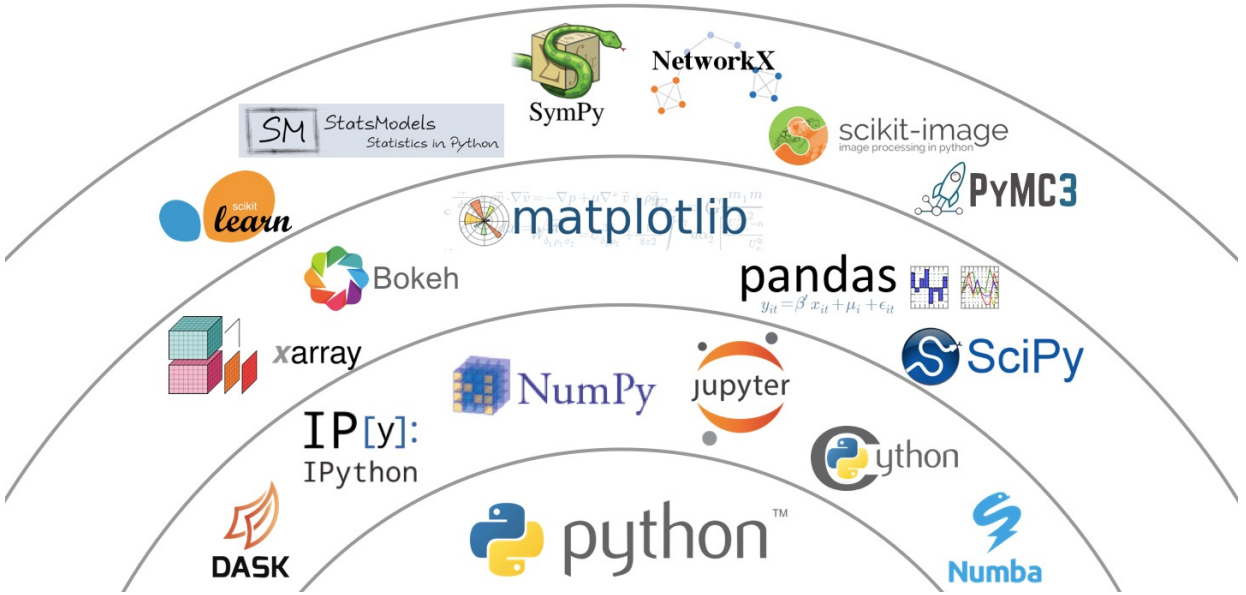


Identificación de ruido

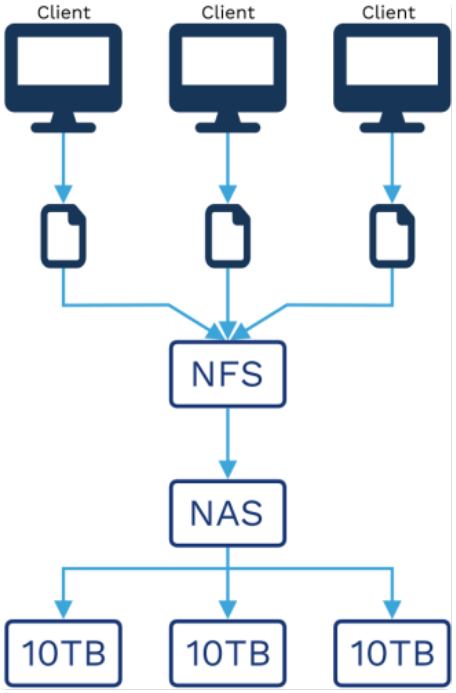


Herramientas

Tecnologías de BigData Python



Network Atachment Storage (NAS)



	Abstraction	Protocols	Random IO support	Typical applications
SAN	"virtual disks" (LUNs, block storage)	Fibre Channel, iSCSI, NVMe-over-Fabric	yes	Virtual machines, some databases
NAS	files and directories	NFS, CIFS, native file system clients	yes	Virtually any, including virtual machines, databases, analytics, machine learning
S3 / Object	write-once objects	S3 protocol, SWIFT	no, only full object overwrites	Limited, mostly archival

Nodos del Cluster IIES (132.248.172.201/202/203)

Windows: openssh-keygen –t RSA
id_rsa -> key private
id_rsa.pub – key pública
ssh heberto@132.248.172.201 –p 10022 -> GPU nvidia 1060
scp -P 10022 archivo-copiar.ext heberto@132.248.172.201:. Win->nodo
scp -P 10022 heberto@132.248.172.201:archivo-copiar.ext . Nodo->Win
exit ->salir del nodo
Versión del Sistema operativo Ubuntu server 22.04
cd ..
ls
pwd
df
nvidia-smi
htop
rm –rf archivo.ext
cp
 mv

Búsqueda y recolección de información.

- **Búsqueda de información:**

  Proceso iterativo en el que se realiza una búsqueda, se analizan los resultados y se va modificando la estrategia de búsqueda hasta identificar los términos y las fuentes de información que nos proporcionan los resultados más pertinentes. (google, 2022)
- https://zety.com/blog/google-search-hacks
- https://pentest-tools.com/information-gathering/google-hacking
- Herramientas de búsqueda en bibliotecas:
  - Schoolar google
  - Bibitex
  - Proquest, etc ..

Recolección de información

- ¿Cuáles son los métodos de recolección de datos?
- Algunos métodos de recolección de información incluyen: encuestas, entrevistas, pruebas, evaluaciones fisiológicas, observaciones, revisión de registros existentes y muestras biológicas.
- Los métodos para recabar datos cualitativos son la observación, la entrevista, los grupos de enfoque, la recolección de documentos y materiales, y las historias de vida. (Sampieri, 2017)

- https://www.pcmag.com/how-to/google-search-tips-youll-want-to-learn

Motores de búsqueda

- Top ten de buscadores 2024
- Web Crawlers: Son la razón por la que motores de búsqueda como Google, Bing, Yahoo y DuckDuckGo siempre proporcionan resultados de búsqueda nuevos y actualizados. Estos bots deambulan por la web como arañas, recopilando información y organizándola en índices. ¿En qué ámbitos se utilizan los rastreadores web y qué tipos existen en Internet?
- Web Scraper: Técnica en la cual una API extrae datos del resultado generado por otro programa. El scraping de datos se manifiesta normalmente en el scraping web, el proceso de utilizar una aplicación para extraer información valiosa de un sitio web.
- https://www.woorank.com/en/edu/seo-guides/search-engine-crawlers
- https://popupsmart.com/blog/web-scraping-tools

Extracción de datos usando API´s

- API web for data extraction:

  Aplicaciones que consisten en extraer datos de la Web. Específicamente, un Web scraper es una herramienta que realiza la extracción y generalmente está representado por un script. Python es uno de los lenguajes de programación más sencillos y fiables que existen. Además, viene con una amplia variedad de bibliotecas. Esto hace que Python sea el lenguaje de programación perfecto para el scraping Web. En detalle, raspado web con Python requiere sólo unas pocas líneas de código

- **10 Best Web Scraping API in 2023**
- https://realpython.com/python-web-scraping-practical-introduction/
- https://docs.scrapy.org/en/latest/intro/tutorial.html

Bibliotecas de Python para Web scraping

- Scrapy. https://docs.scrapy.org/en/latest/intro/tutorial.html
- Requests. https://www.tutorialspoint.com/requests/requests_web_scraping_using_requests.htm
- Beautifulsoup4: https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/
- Selenium: https://selenium-python.readthedocs.io/index.html

- https://brightdata.es/blog/procedimientos/web-scraping-with-python#libraries

Bibliografía

- Big data, machine learning y data science en python. José Manuel Ortega Candel. Ed. Ra-Ma. 2023
- Sistemas de Big Data. Victor Manuel López Fandiño. Ed. Ra-Ma. 2023
- Big Data: Concepts, Technology, and Architecture. Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H Gandomi. Ed. Wiley. 2021
- Big Data For Dummies. Judith S Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman, Ed. A Wiley Brand. 2013
- Big Data MBA: Driving Business Strategies with Data Science, Schmarzo, Bill. Ed. Wiley
- Hands-On Machine Learning, with Scikit-Learn, Keras, and TensorFlow
- Concepts, Tools, and Techniques to Build Intelligent Systems. Aurélien Géron. O'Reylli. 2023
- Python for Data Analysis. 3a. Edición. Data Wrangling with pandas, NumPy, and Jupyter. Wes McKinney. Ed. O'Reylli. 2022
- Scaling Python with Dask. From Data Science to Machine Learning. Holden Karau and Mika Kimmins. Ed. O'Reylli. 2023
- Documentos oficiales de Anaconda Python. https://docs.anaconda.com/
- Documentos oficiales de Flask.