

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/NzM0NTk3Mzc3Mzlz?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.

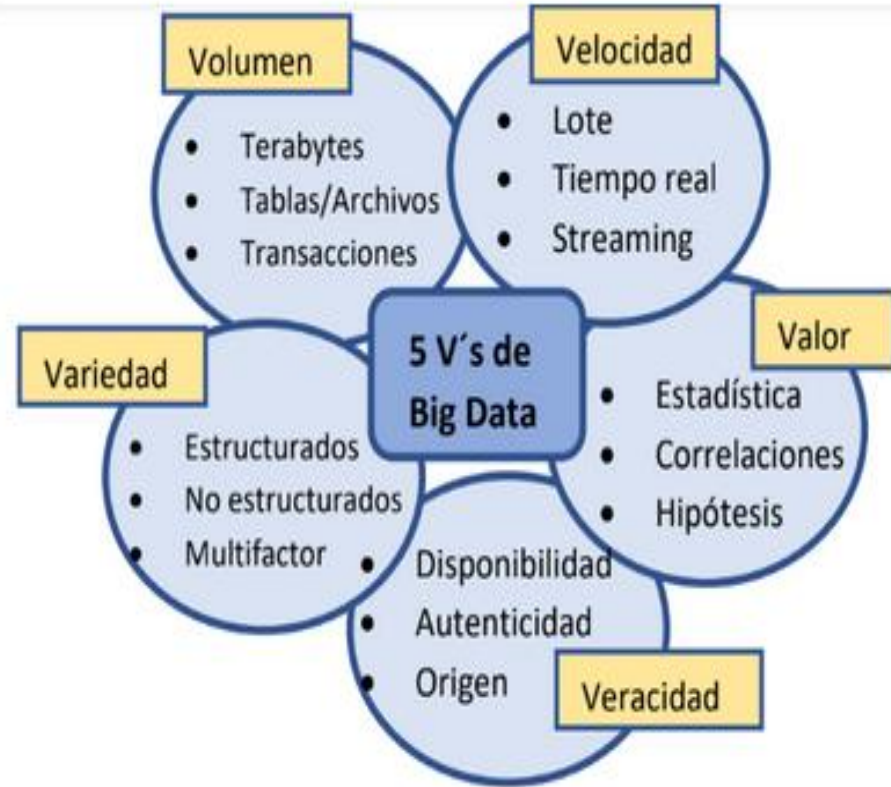
Sesión	Fechas	Actividades	Entregas	Puntos
1	Martes 26 de noviembre	El uso del Big Data en la Ciencia de Datos.		
2	Jueves 28 de noviembre	Capítulo libro, Web Scraping (Análisis del Big Data con ML).	Tarea 1	3
3	Martes 3 de diciembre	Uso de almacenamiento en la Red (NAS).	Tarea 2	2
4	Jueves 5 de diciembre	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 3.

- Evaluación diagnóstica.
- Descargar los sistemas:
 - 1.Vbox (VMWare) para virtualización
 - 2.TrueNAS community (versión 13 o 24)
- Instalar el TrueNAS
- <https://www.truenas.com/docs/core/13.0/gettingstarted/install/>



V's del Big Data



Integración de datos



NAS:

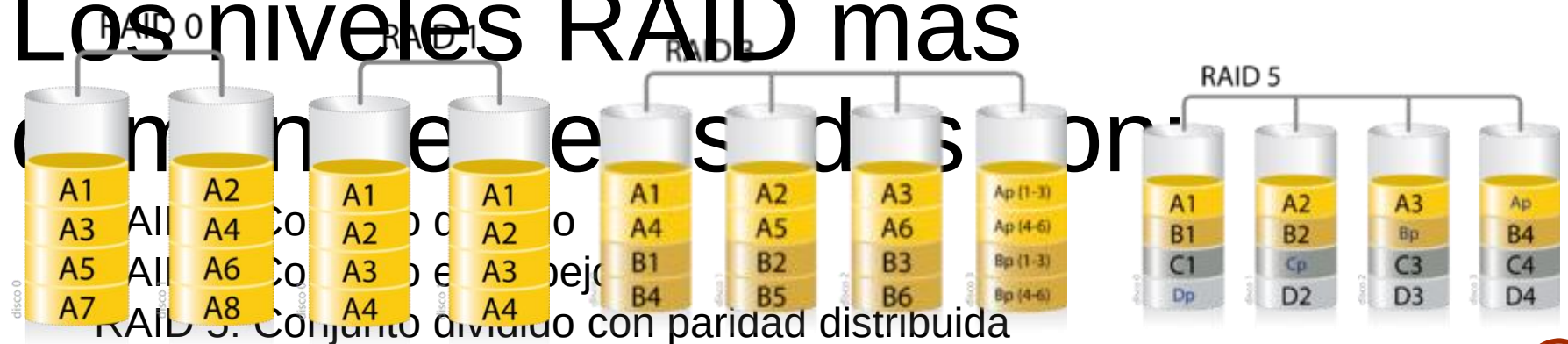
- Un servidor NAS es un dispositivo dedicado al **almacenamiento de datos** que se conecta directamente a una red
- Su función principal es facilitar el almacenamiento centralizado y el intercambio de archivos
- Los servicios que ofrece son:
 - ✓ Conexión para usuarios (login/password)
 - ✓ AFP: archivos de Apple, y compartirlos en red.
 - ✓ iSCSI: Almacenamiento por bloques (LUN).
 - ✓ NFS: archivos en Linux.
 - ✓ WebDAV: Compartidos en web.
 - ✓ CIFS/SMB (Samba): Para sistemas windows.
 - ✓ FTP (Proftpd): Transferencia de archivos en red.
 - ✓ S3: Para compartir con la nube.
 - ✓ Instalación de un Pool de discos



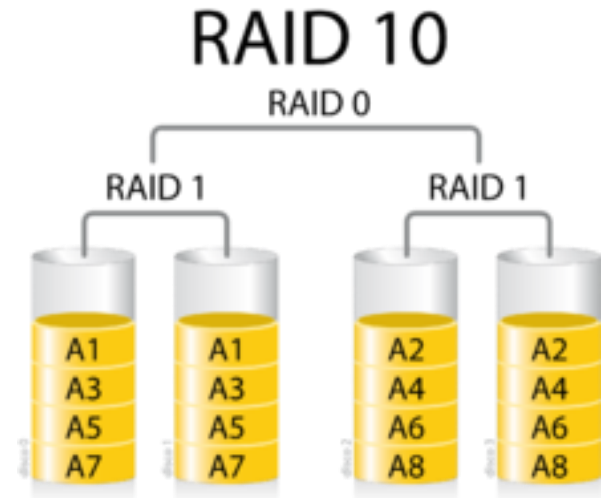
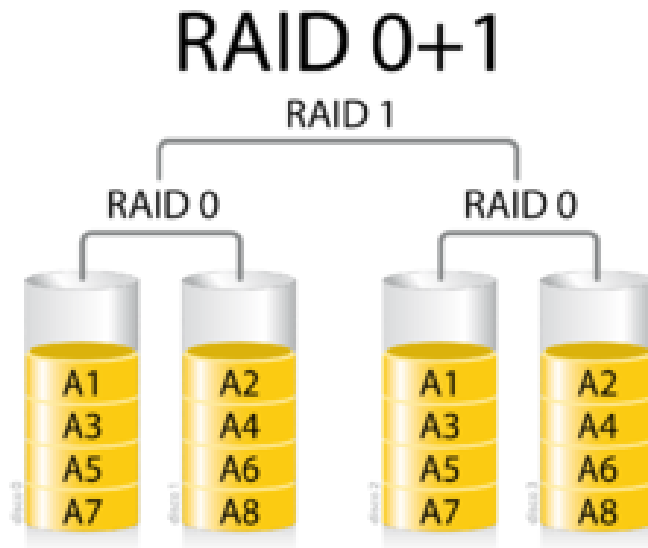
Discos (Redundancia):

Se instalan y configuran como una arreglo redundante de discos independientes (RAID, redundant array of independent disks)

Los niveles RAID más comunes



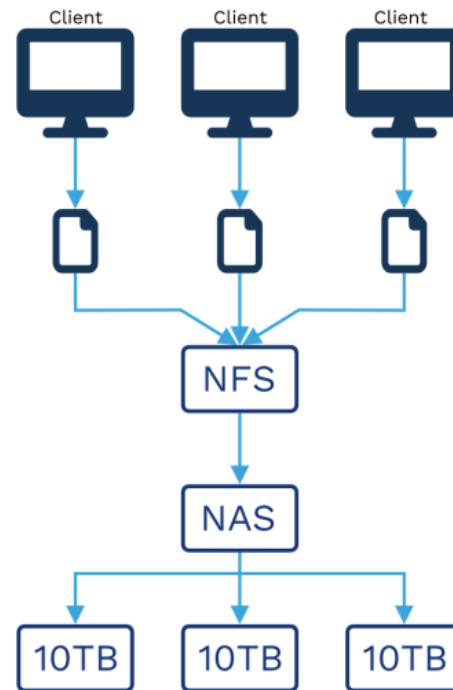
RAID:



Best SSDs for Synology NAS



Network Attachment 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



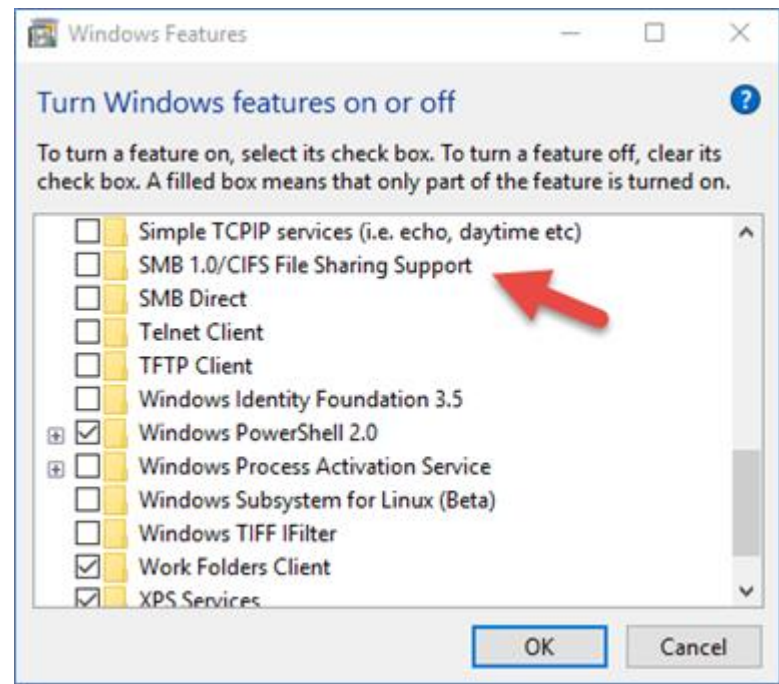
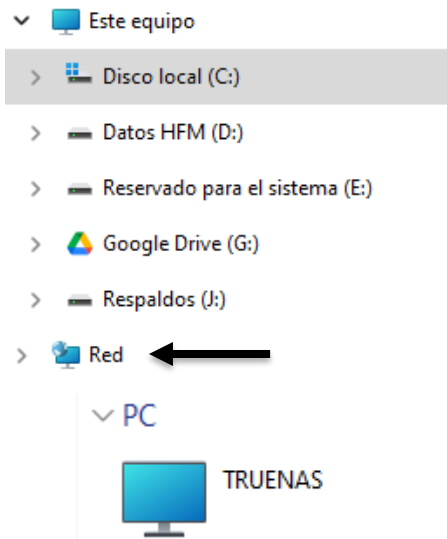
Instalación de TrueNAS:

Se requiere una máquina virtual
con 2 procesadores
30 Gb de Disco para instalación
4 GB de RAM
1 Tarjeta de red
2 o tres discos virtuales
Utilizar la imagen más
actualizada de TrueNAS desde
el sitio Web.



Uso de un SMB/CIFS:

- Instalar cliente SMB/CIFS versión 2
- Después acceder al TrueNAS



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.

