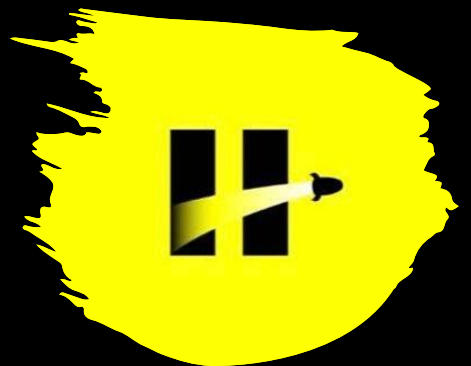
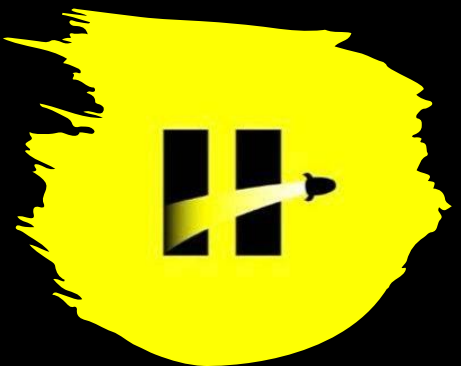
M4
BIG DATA

Pre Modulo

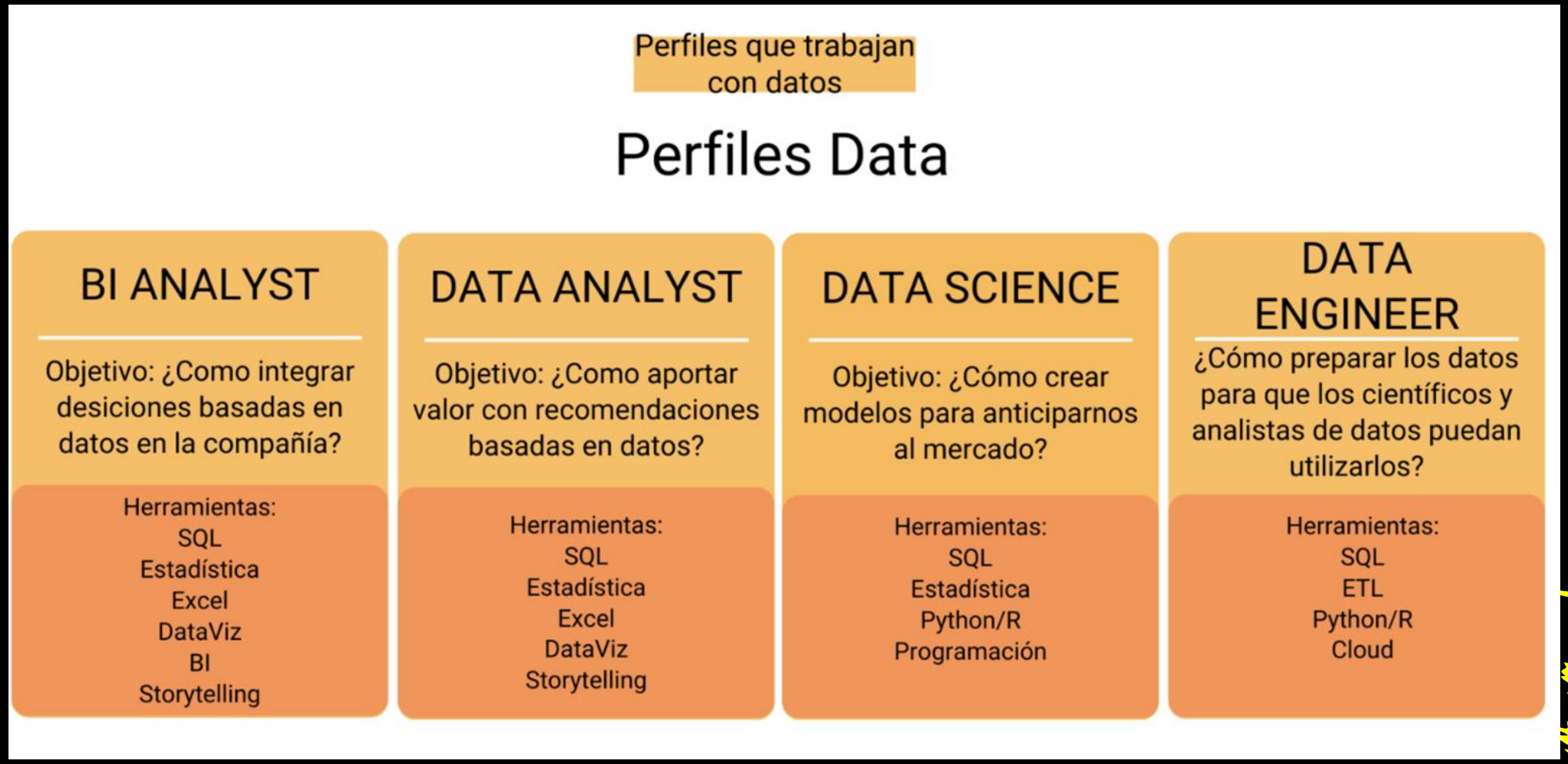


Objetivos del Modulo

- Definición de Big Data
- Comprender conceptos de virtualización y contenedores
- DataLake Vs Datawarehouse
- Casos de Uso
- Conocimiento acerca del ecosistema HADOOP
- Poder determinar en que situaciones es conveniente un ecosistema de Big Data y cuando no.
- Poder comprender las diferentes herramientas que conforman el ecosistema de Apache Hadoop y el lugar que ocupan en el mismo.
- Entiender la arquitectura de HDFS y desde esa comprensión valoren el impacto de los comandos de manipulación de archivos.
- Bases de datos SQL VS No SQL
- Determinar la importancia de las herramientas de Big Data en un proyecto



Perfiles Data



Donde nos encontramos?

M4

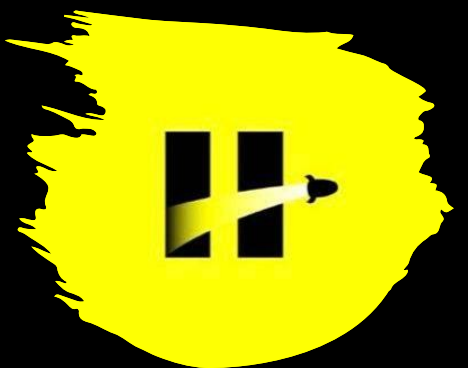
- En este modulo conoceras las herramientas que nos brindan la infraestructura para trabajar en Big Data

M5

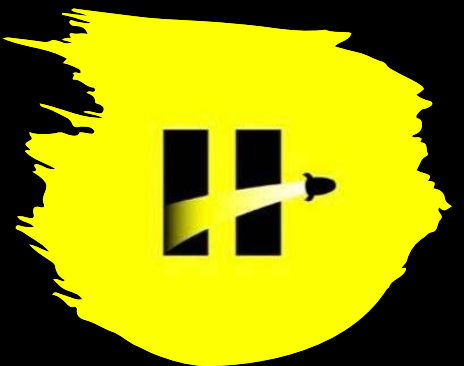
- Durante el M5 conoceras mucho mas acerca del analisis de datos, herramientas de visualización y como extraer valor a los datos e impactar el negocio

M6

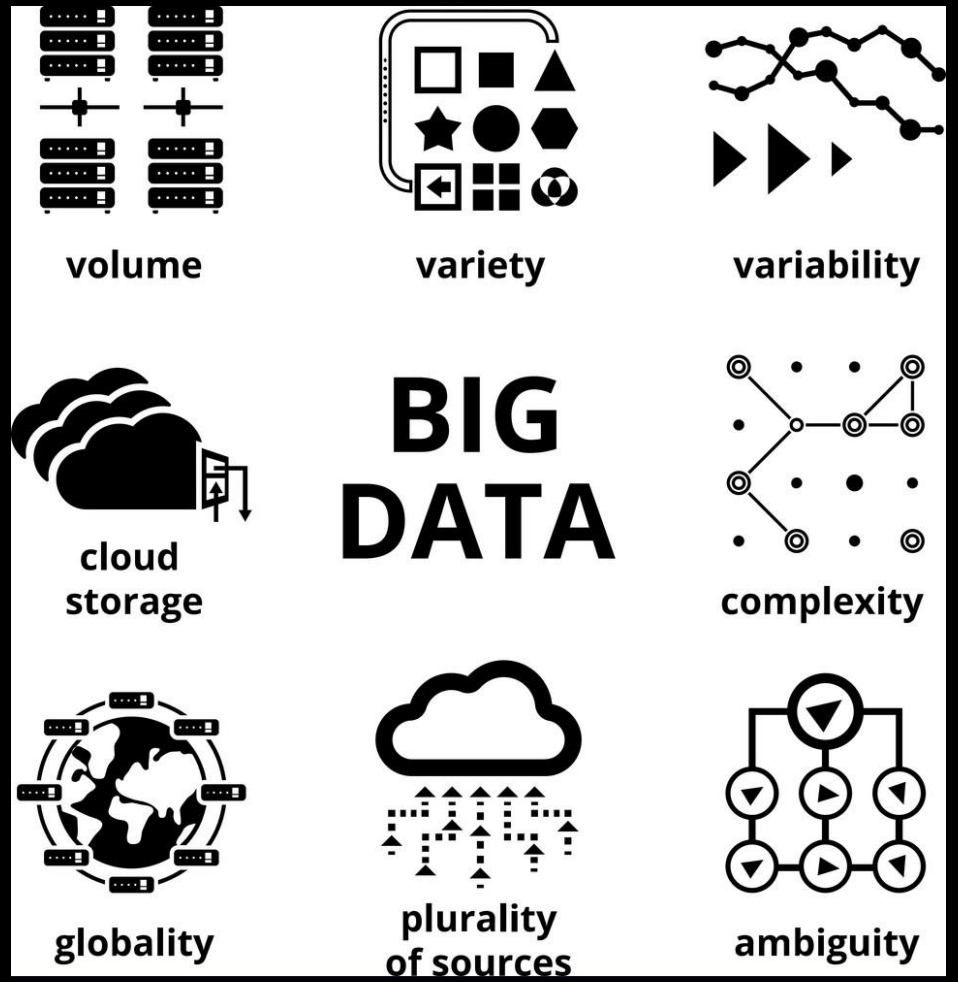
- En el M6 podras obtener todo el conocimiento acerca de Machine Learning y modelos predictivos.



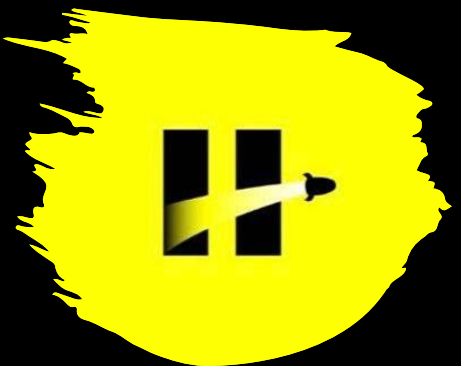
Te dejamos algunos contenidos para que vayas introduciéndote en el M4

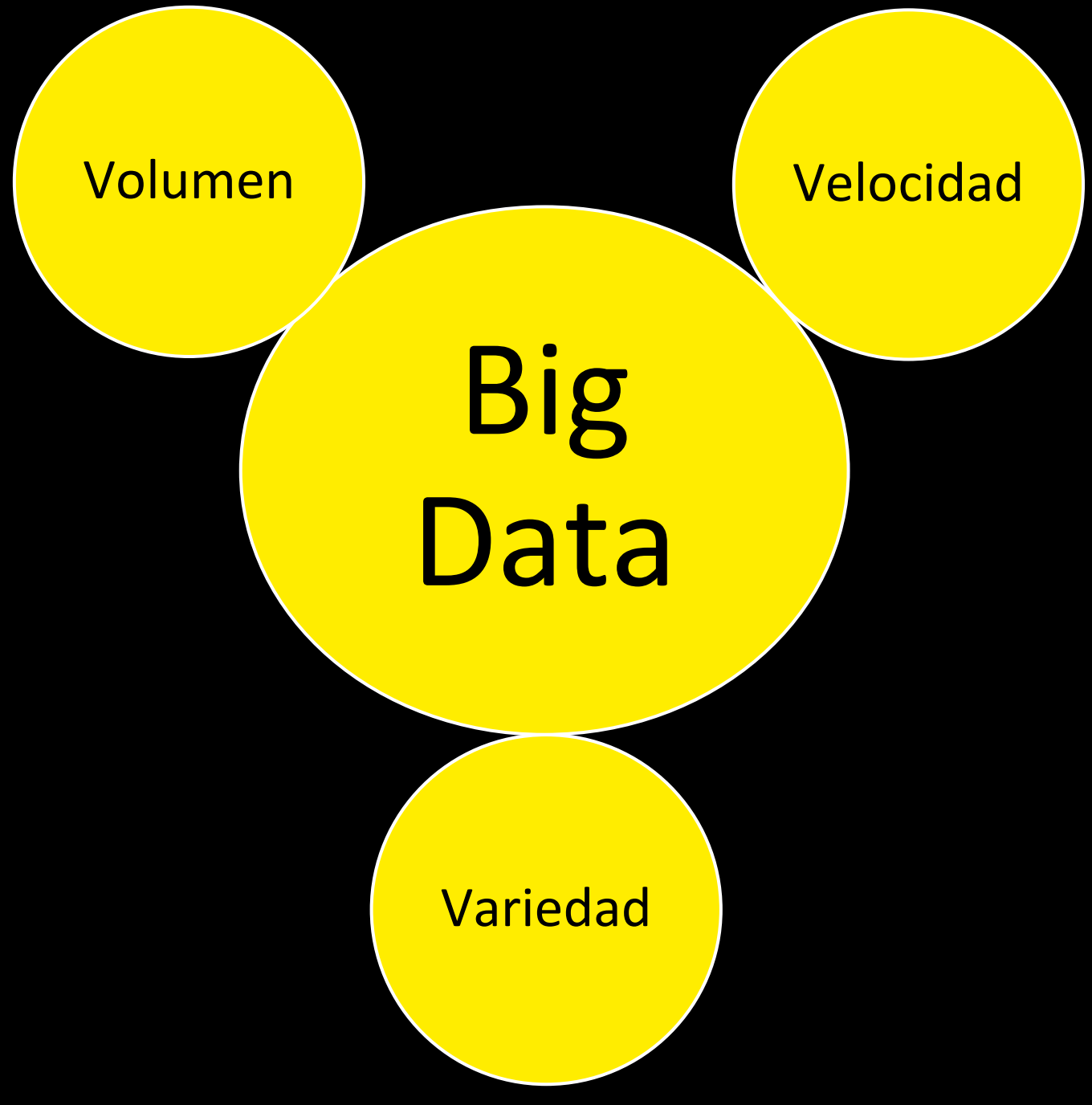


** El Modulo M4 se dará un primer acercamiento a muchas herramientas del mundo de Big Data, será introductorio y podras reconocer la importancia de dichas herramientas.*



¿A que llamamos Big Data?




Volumen
Velocidad
Big Data
Variedad

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo

Volumen

Tamaño

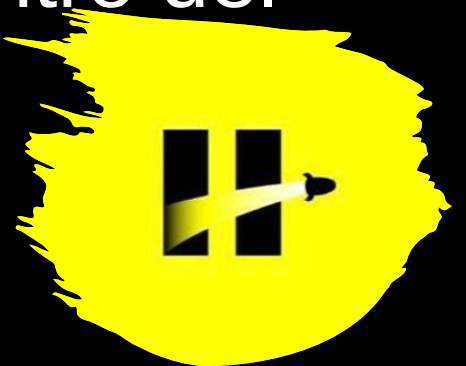
Velocidad

Velocidad de Crecimiento

Variedad

Complejidad de tipos de datos

Dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.



Big Data

Velocidad

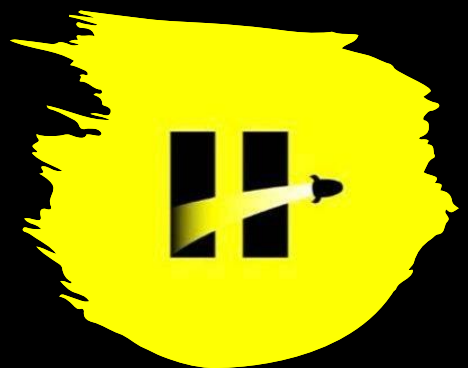
Refiere a la **rapidez con que los datos son creados, accedidos, almacenados y procesados**.

Volumen

Refiere a la almacenamiento procesamiento de una **grandísima cantidad de datos** que supera ampliamente las capacidades de los equipos y sistemas

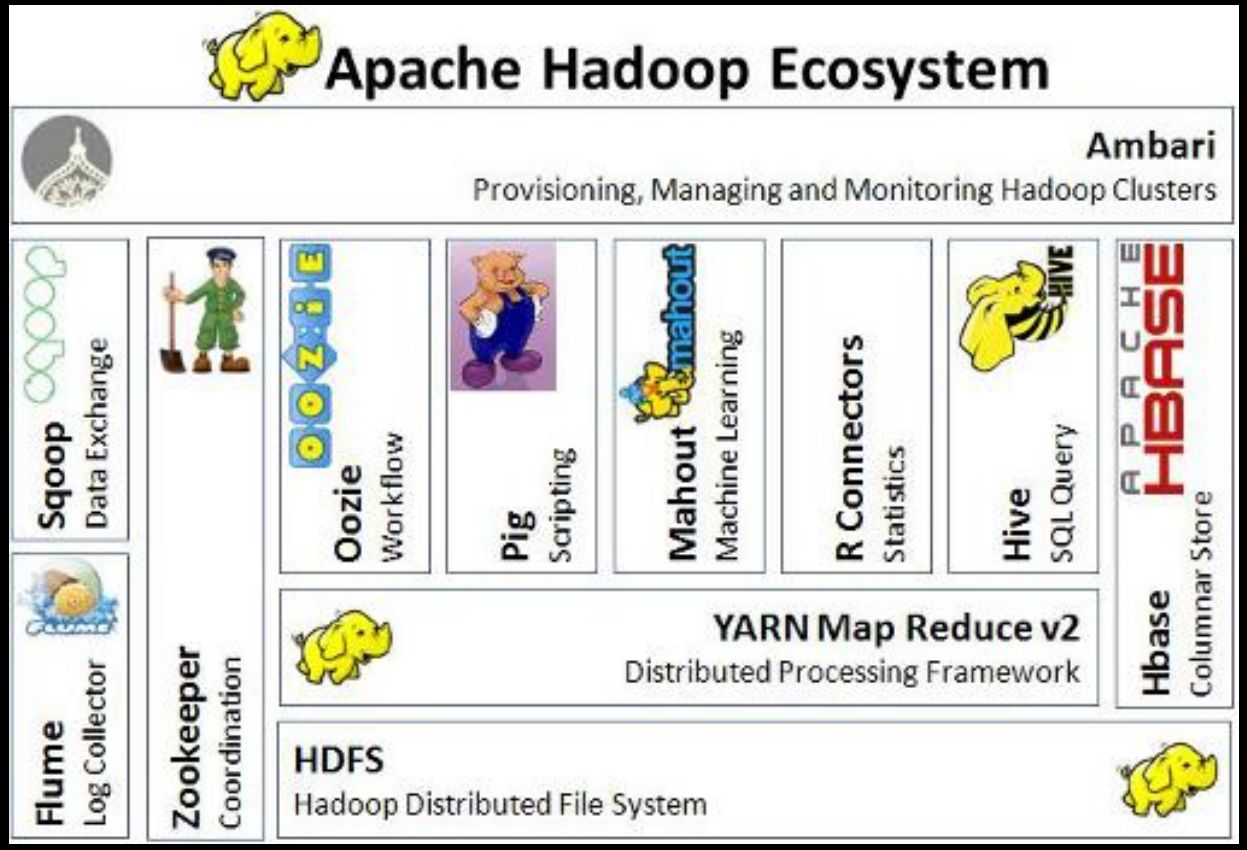
Variedad

Refiere a que deben **ser capaces de procesar datos de diversas formas, tipos y fuentes**. Datos estructurados *(SQL)*, semiestructurados *(XML,JSON)* y no estructurados *(WORD, PDFs, IMAGENES)*



Ecosistema Hadoop

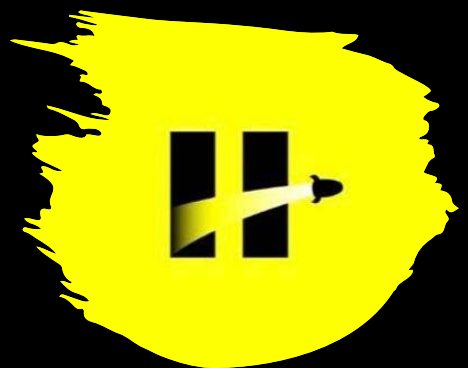
- Durante este modulo iremos navegando por los diferentes frameworks del ecosistema Hadoop.



Te dejamos algunos links para que visites

- Previo al primer encuentro del M4 te dejamos algunos links para que vayas descargando y asi tener todo listo para dicho modulo.

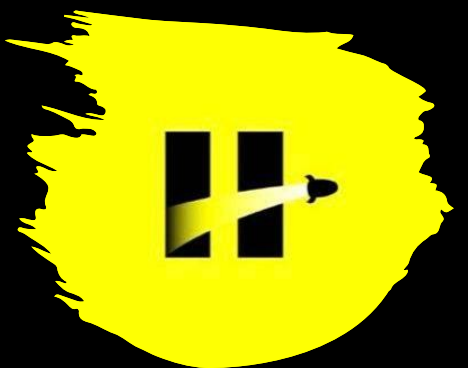
1.Realizar la instalación de VirtualBox: https://www.virtualbox.org/wiki/Downloads
2.Realizar la instalación de Putty: https://www.putty.org/ - https://www.compuhoy.com/como-descargo-putty-en-linux/
3.Realizar la instalación de WinSCP: https://winscp.net/eng/download.php (FileZilla es una alternativa si no usas sistema operativo Windows)
4.Imagen UBUNTU: https://drive.google.com/file/d/1oh-GJdQKCzi75l9b4TVoIEejs94Ke_V0/view?usp=share_link



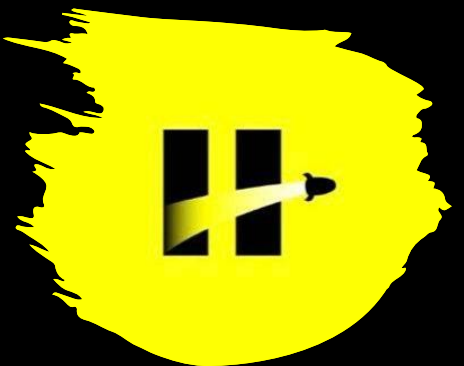
- Aunque el tamaño utilizado para determinar si un conjunto de datos determinado se considera Big Data no está firmemente definido y sigue cambiando con el tiempo, la mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van **desde 30-50 Terabytes a varios Petabytes.**

Que tipos de datos manejamos en Big Data

La naturaleza compleja del Big Data se debe principalmente a la naturaleza **no estructurada** de gran parte de los datos generados por las tecnologías modernas, como los logs de servidores web, los sensores incorporados en dispositivos, las búsquedas en Internet, las redes sociales, computadoras portátiles, teléfonos inteligentes y otros teléfonos móviles, dispositivos GPS, registros de centros de llamadas y en general todo lo que refiere a tecnología multimedia, es decir, imágenes, audio y video.

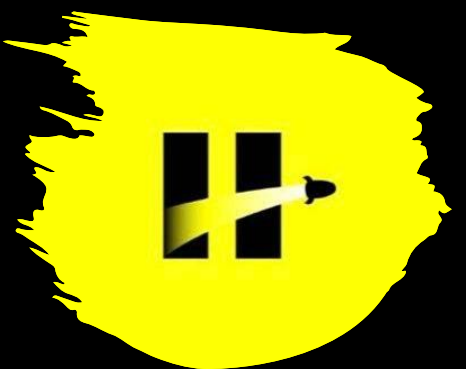


¿Que Ventajas Brinda?

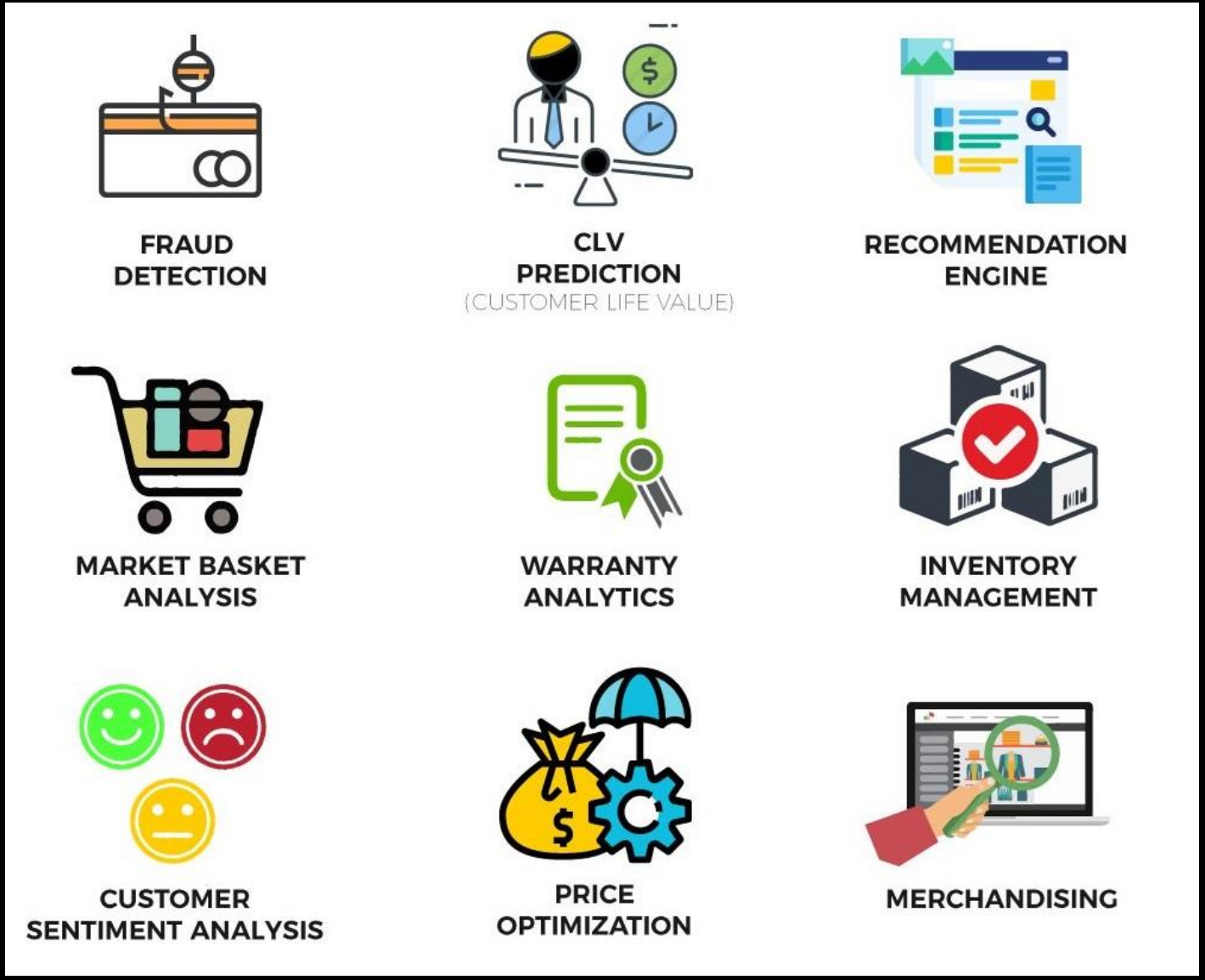


Optimizando nuestros negocios con Big Data

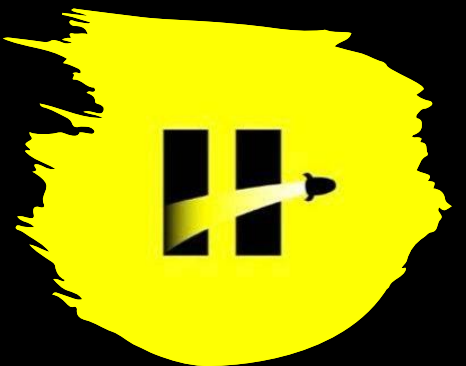
- El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para **identificar nuevas oportunidades**. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:
  - Reducción de coste: Optimizando Procesos
  - Mas Rápido y mejores decisiones
  - Nuevos Productos y servicios



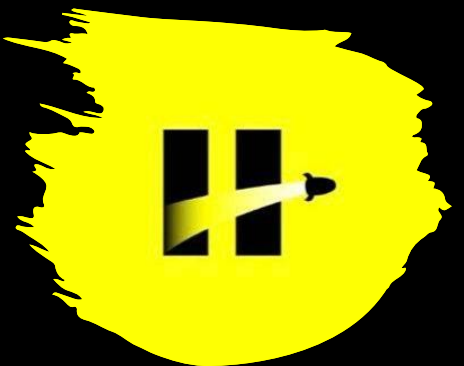
Casos de Uso


FRAUD DETECTION
CLV PREDICTION
(CUSTOMER LIFE VALUE)
RECOMMENDATION ENGINE
MARKET BASKET ANALYSIS
WARRANTY ANALYTICS
INVENTORY MANAGEMENT
CUSTOMER SENTIMENT ANALYSIS
PRICE OPTIMIZATION
MERCHANDISING

DATAWAREHOUSE
VS
DATALAKE

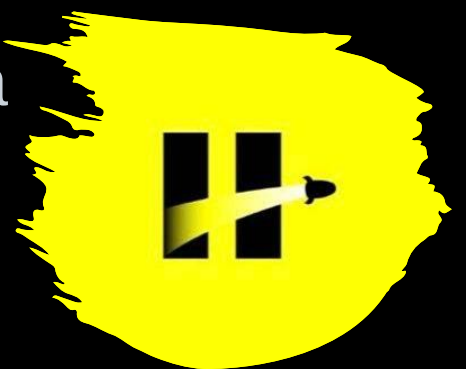


DATALAKE



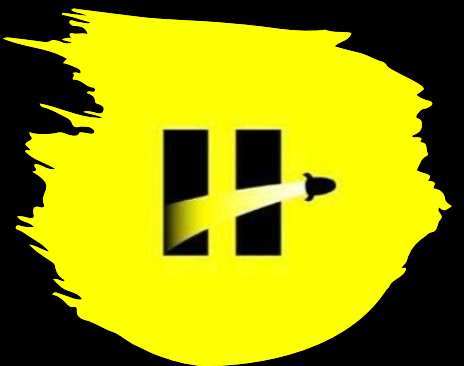
Datalake

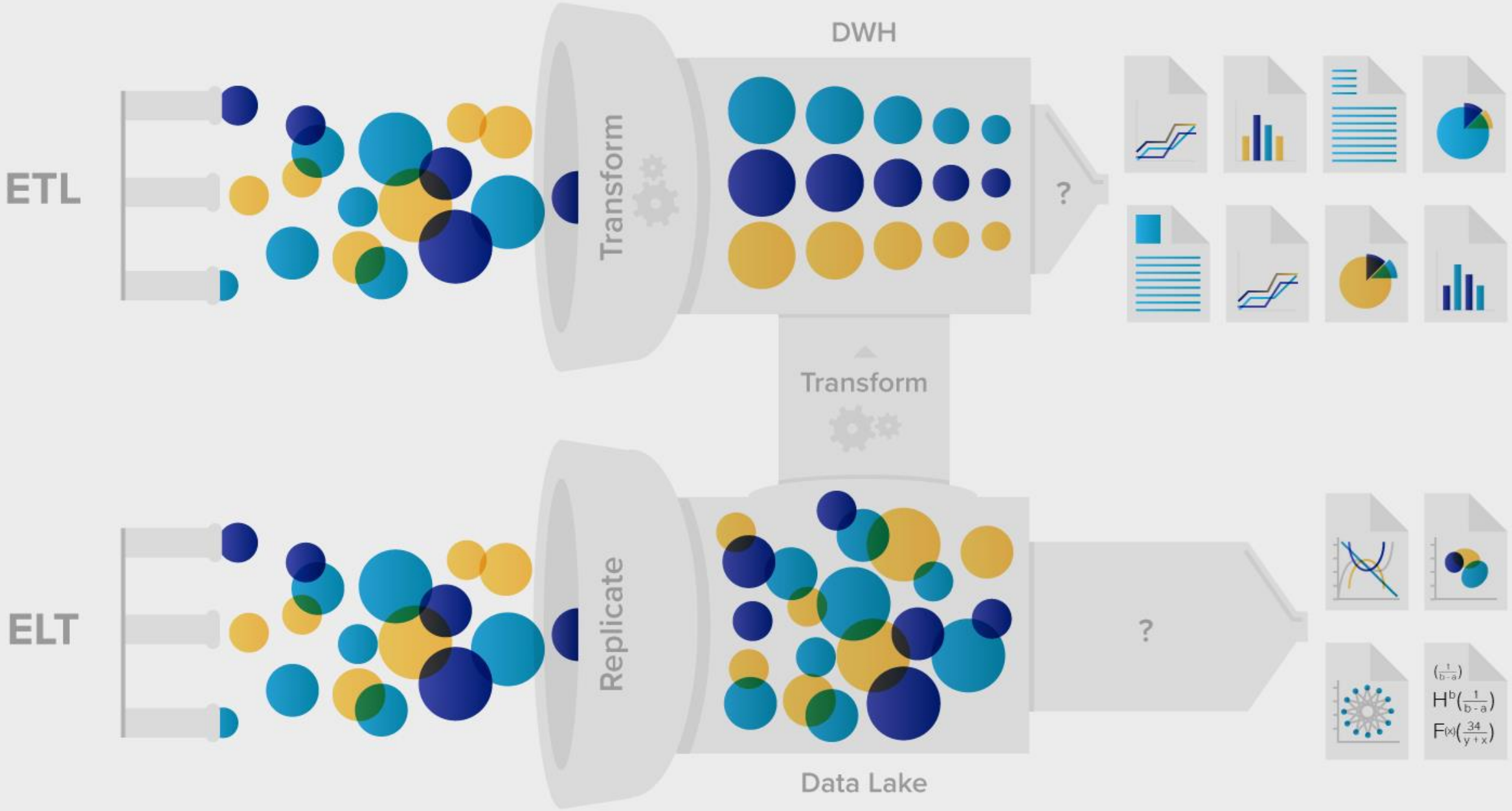
- Dada la afluencia en los últimos tiempos, del uso de Internet y la tecnología de redes en general, como por ejemplo sensores o APIs, concretamente el desarrollo de lo que se conoce como IoT (Internet de las Cosas), se comenzó a trabajar datos que **no necesariamente son llevados a una estructura tabular**, dando lugar al desarrollo de una serie de herramientas conocidas como motores de bases de datos No-SQL y al desarrollo de una arquitectura conocida como Data Lake, la cuál contempla el almacenamiento y disponibilización de todo tipo de datos, estructurados y no estructurados, manejando esa variedad y también soportando grandes volúmenes de datos, que se generan a gran velocidad.



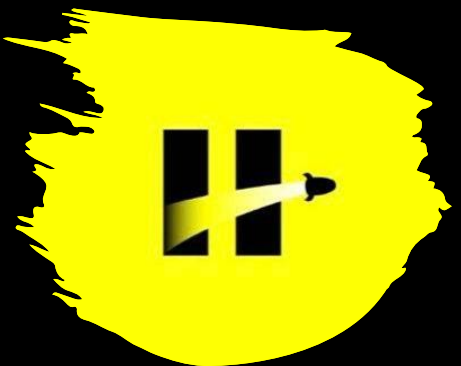
Características

- Es un repositorio unificado de datos, estructurados y no estructurados.
- Está diseñado soportar las cargas de trabajo de Big Data y Machine Learning.
- Prioriza el almacenamiento de los datos en su formato original para luego ser procesados de acuerdo a la demanda.
-



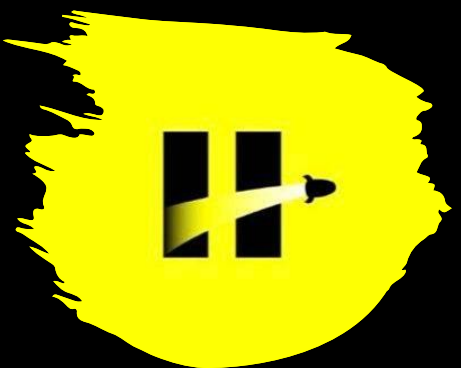

ETL
Transform
DWH
?
Transform
ELT
Replicate
?
Data Lake

DATAWAREHOUSE

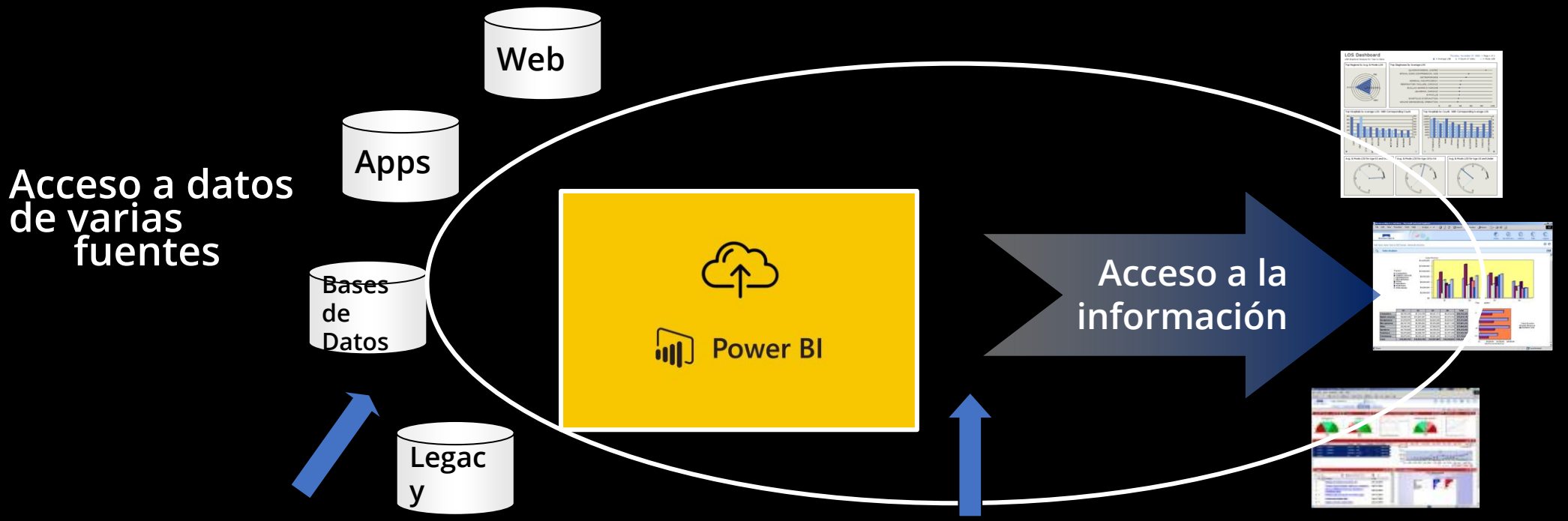


Datawarehouse

Es un **repositorio de datos** de fácil acceso, alimentado de numerosas fuentes, transformadas en **grupos de información sobre temas específicos de negocios**, para permitir nuevas consultas, análisis, reportes y decisiones.



Datawarehouse



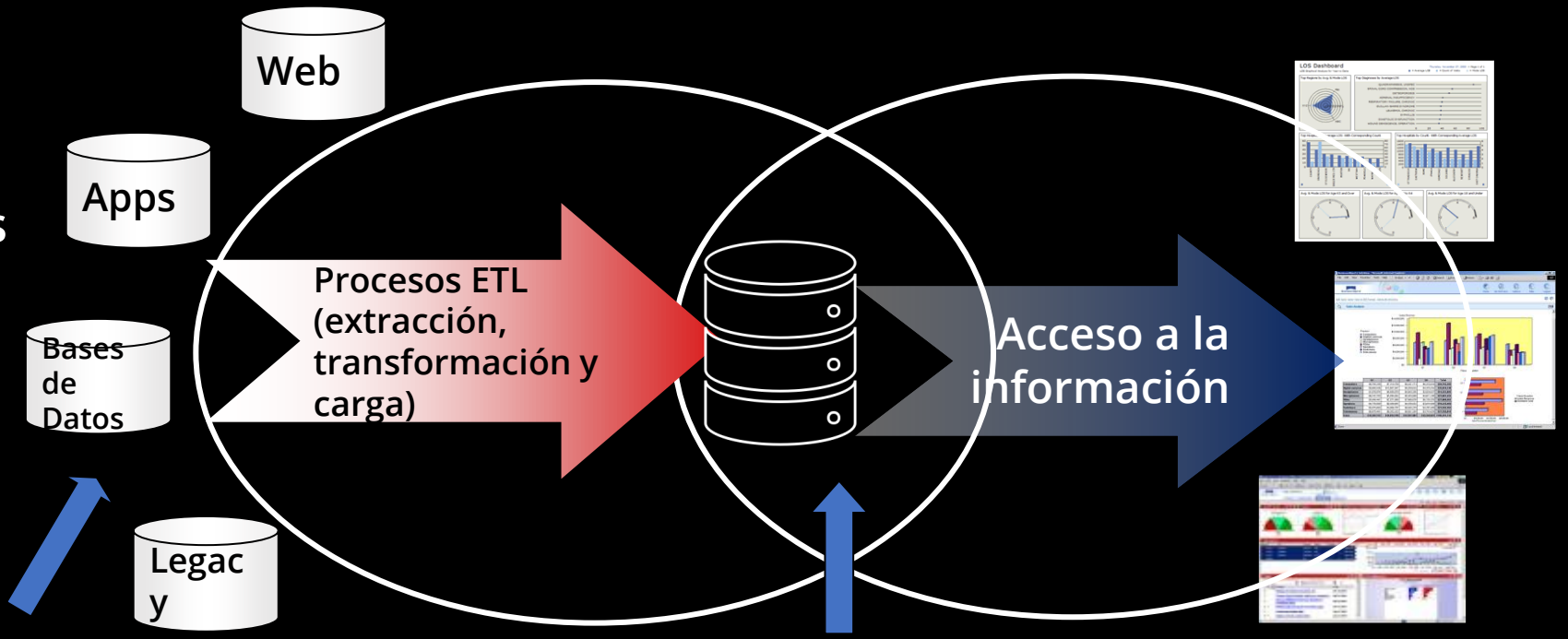
Acceso a información:

Reportes, Dashboard, Consultas, Aplicaciones

OLTP: Estructuras orientadas a ser eficientes en las transacciones

Datawarehouse

Acceso a datos de varias fuentes



Acceso a información:

Reportes, Dashboard, Consultas, Aplicaciones

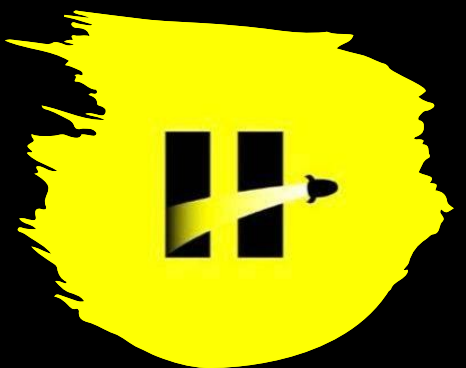
OLTP: Estructuras orientadas a ser eficientes en las transacciones

Datawarehouse: (OLAP) Estructuras orientadas a ser eficientes en las consultas

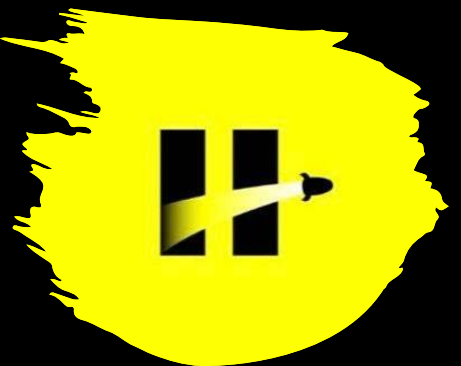
DOCKER MONGO
KAFKA
ZOOKEPER
M4
STREAMING DE DATOS DE TWITTER

NOTEBOOK JUPYTER
M5 M3
DESCRIPTIVO DE LOS TWEETS STREMEADOS
STREAMLIT CON VISUALIZAIOCNES DE SEABORN Y MATPLOT

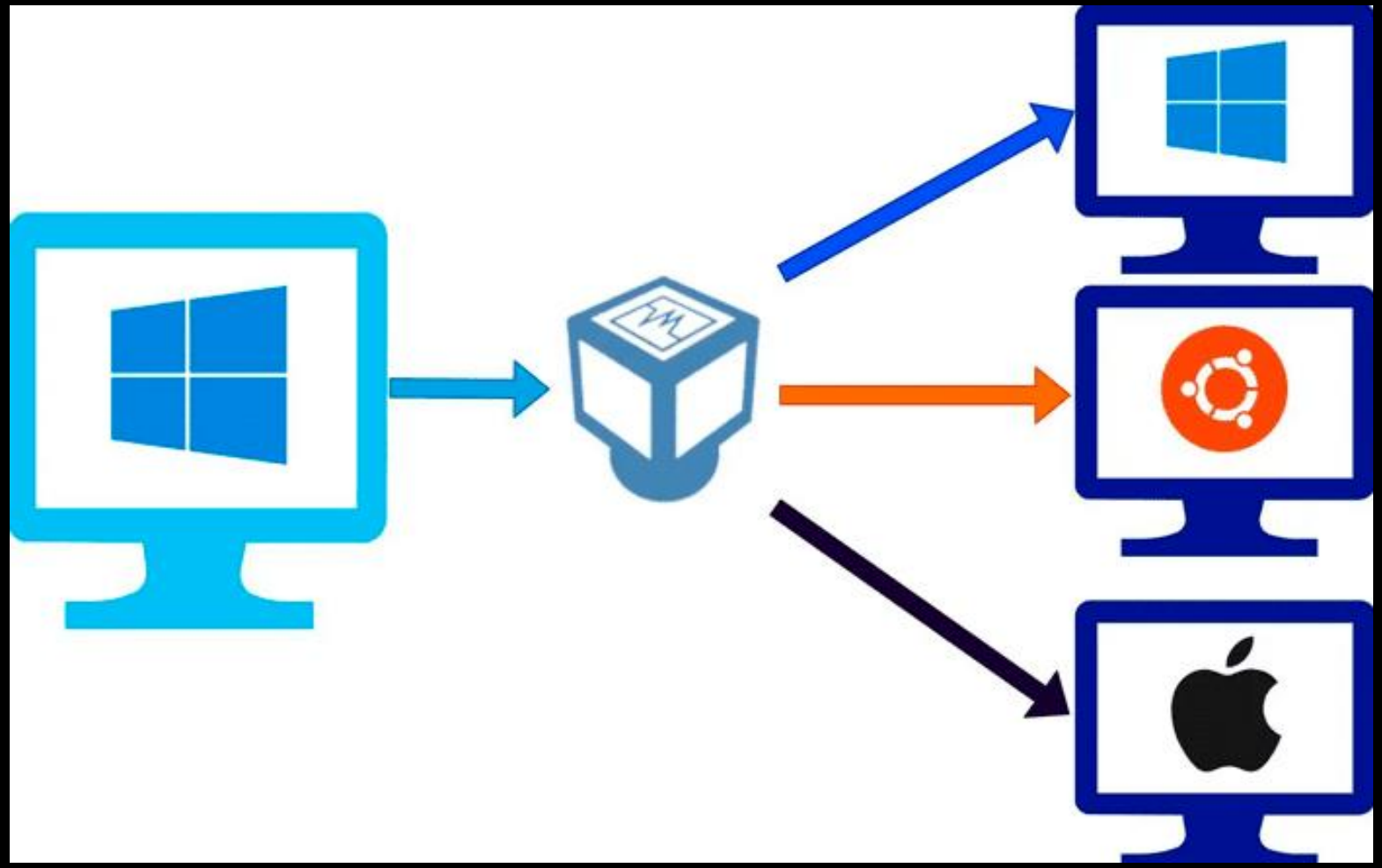
NOTEBOOK JUPYTER
M6
NLP SENTIMIENTOS DE LOS TWEETS



Algunos conceptos previos

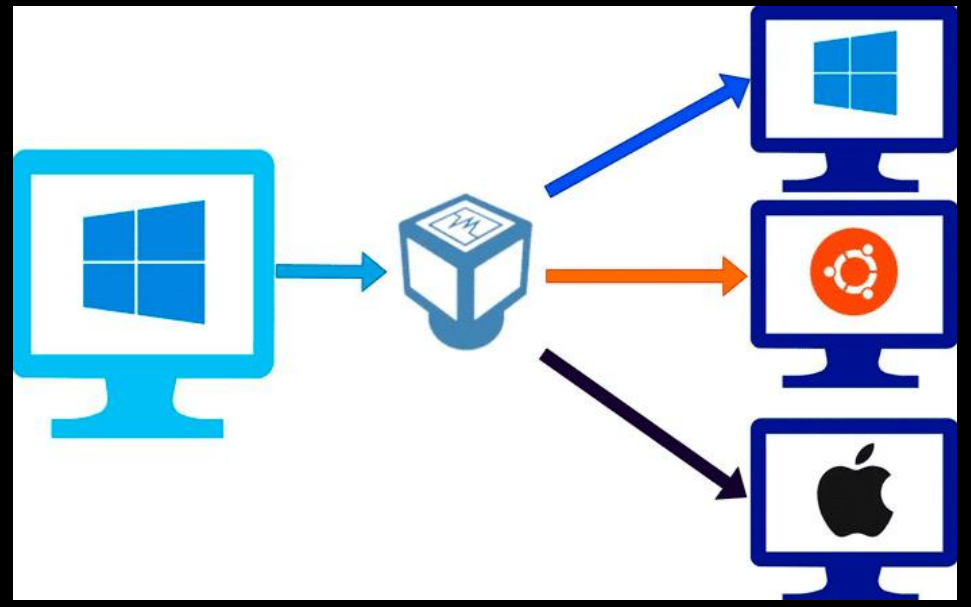


Virtualización



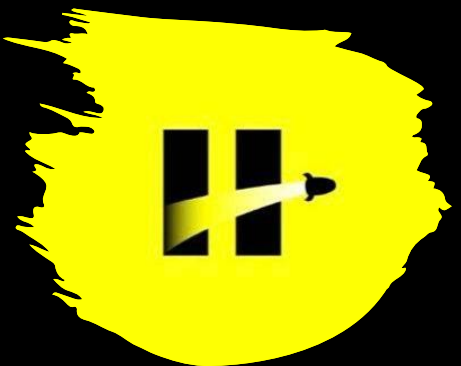
Virtualización

Una máquina virtual no es más que un software **capaz de cargar en su interior otro sistema operativo** haciéndole creer que es un PC de verdad. Tal y como su nombre indica, el concepto es tan sencillo como crear una máquina (PC, consola, móvil o lo que sea) que en vez de ser física es virtual o emulada.



Putty

- En general, Putty no es más que una terminal de simulación *open source* que fue desarrollado para actuar como cliente de conexiones seguras a través de protocolos raw TCP, Telnet, rlogin y portal serial.
- Por lo tanto, este software se indica para establecer conexiones seguras de acceso remoto a servidores a través de Shell Seguro (SSH y para construir canales encriptados entre servidores.

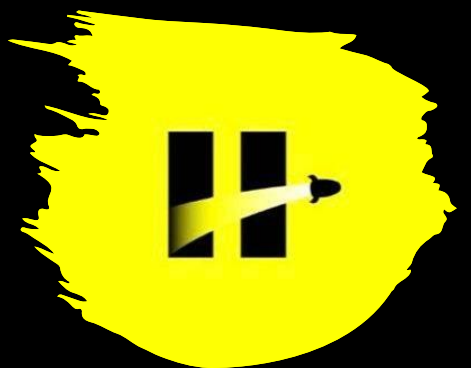


Hadoop

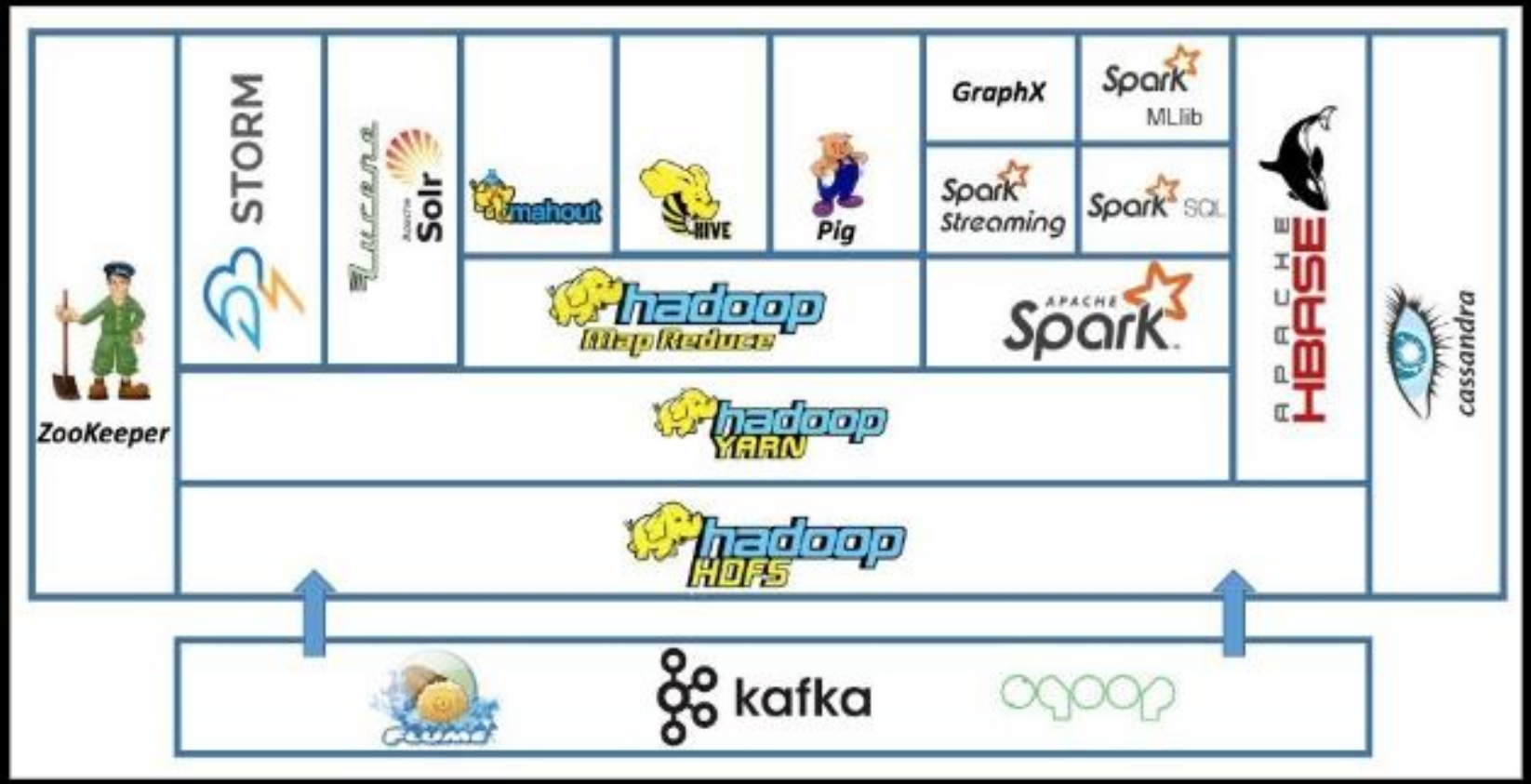
Es un sistema open-source diseñado para almacenar y procesar Big Data de forma distribuida utilizando un cluster de servidores.

Características:

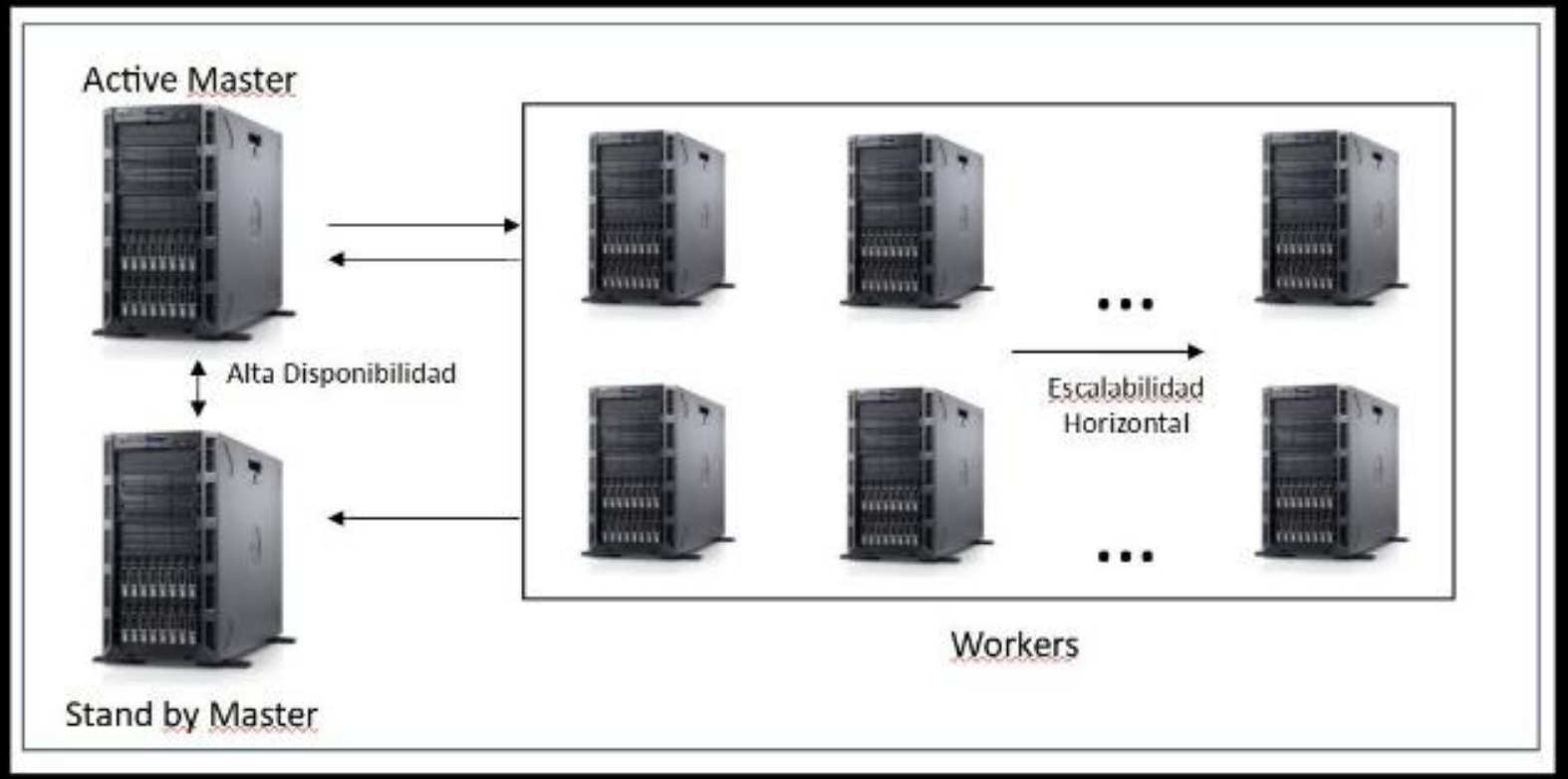
- Tolerancia a Fallos.
- Escalabilidad Horizontal.
- Utiliza "Commodity Hardware".
- Desarrollado en lenguaje Java.
- Procesamiento en paralelo.



Ecosistema Hadoop



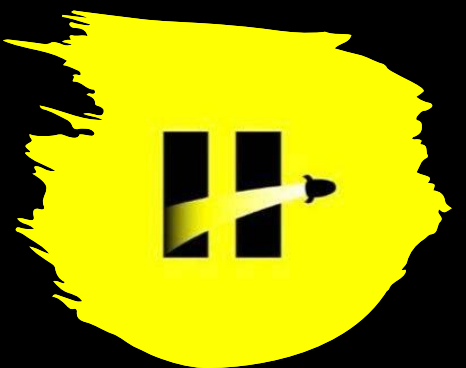
Cluster Hadoop



Cluster Hadoop

- Hadoop permite organizar computadoras en una relación maestro-esclavo que contribuye a conseguir una gran escalabilidad para el procesamiento.

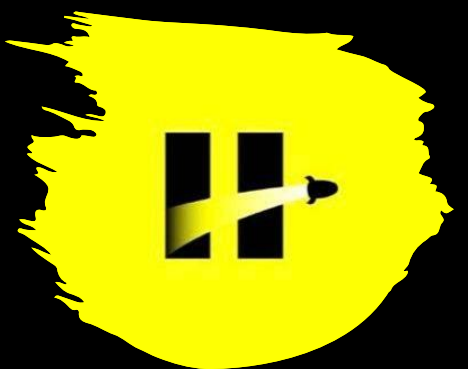
- Un Cluster Hadoop tiene dos tipos de nodo: 1) Un “Master Node” llamado NameNode y 2) “Workers Nodes” llamados DataNodes



Master Node

Almacenan toda la información relevante acerca de los DataNodes y de los archivos almacenados en los DataNodes:

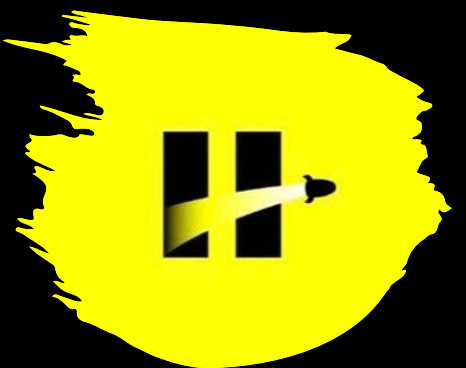
- Para cada DataNode: nombre, rack, capacidad y estado
- Para cada archivo: nombre, réplicas, tipo, tamaño, ubicación y estado.



DataNode

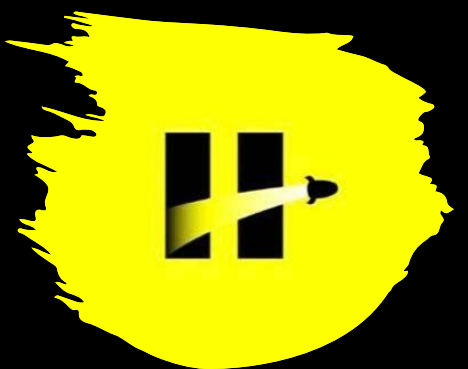
- Si un DataNode Falla, se puede recrear la información automáticamente en otra computadora mediante los bloques de archivos.

- Los DataNodes se comunican entre si por medio de mensajes “heartbeats” para conocer el estado de los otros nodos.



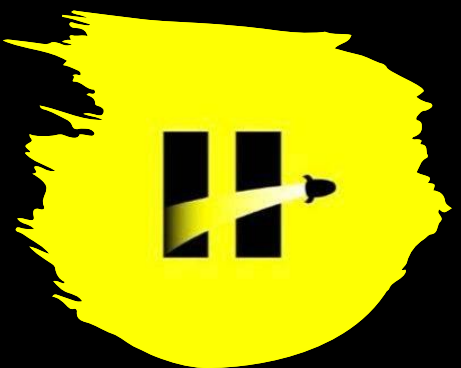
Bloques de Archivos

- En el Sistema de Bloques, un bloque es la unidad fundamental de almacenamiento en HDFS.
- Se almacenan información de grandes archivos distribuidos en segmentos llamados bloques para ser almacenados en diferentes computadoras. (64 o 128mb)
- Cada archivo de datos utiliza un determinado numero de bloques organizados en bloques consecutivos para facilidad y velocidad de acceso.
- La cantidad de bloques se puede configurar.



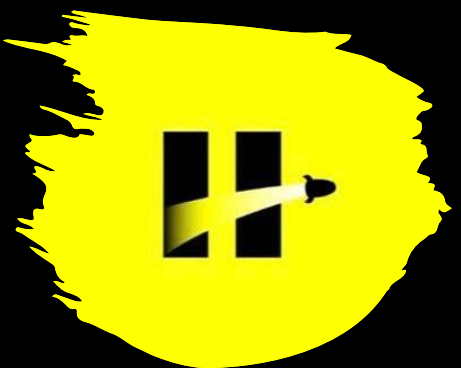
Componentes Core

- HDFS
- YARN
- MapReduce



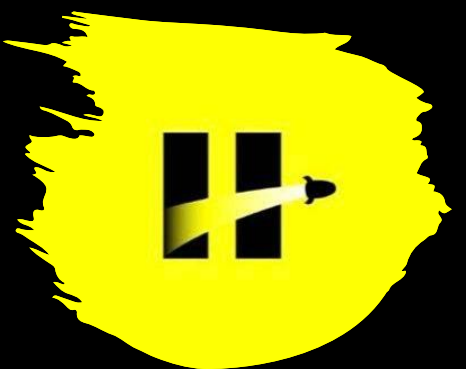
HDFS

- Hadoop Distributed File System
- Es el componente principal del ecosistema Hadoop
- Permite almacenar data-sets masivos con distintos tipos de datos estructurados, semi-estructurados y no estructurados.



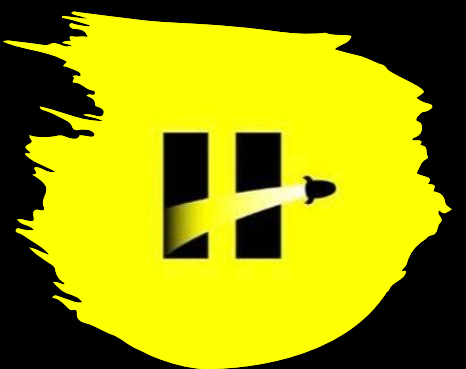
YARN

- Yet Another Resource Negotiator
- Administra los recursos y los flujos de trabajo en un entorno seguro y asegura la alta disponibilidad de los multiples clusters Hadoop
- Brinda flexibilidad para ejecutar múltiples aplicaciones y herramientas como ser consultas interactivas (Hive), procesos de flujo en tiempo real (Spark) y procesamiento por lotes (MapReduce)
- Brinda escalabilidad a los nodos



Map Reduce

- Permite procesar enormes cantidades de datos utilizando los servicios de gran cantidad de computadoras para trabajar en diferentes partes del trabajo (Jobs)
- La tarea de procesamiento se divide en muchas partes y cada una se procesa de forma independiente de las otras y luego los resultados intermedios se combinan en el resultado final.
- Es un framework de procesamiento paralelo



Componentes Core

MapReduce