



# **Curso de Big Data con Hadoop y Spark**

## **La Era de los datos**

### **Evolución de la Tecnología de Hardware e Infraestructura de Redes**

- Nuevas formas de generar dato: Internet, API's, IoT
- El dato generado es heterogéneo y de gran volumen
- Se genera dato más rápidamente
- Las grandes empresas comienzan a invertir en desarrollar herramientas que permitan gestionar y analizar esos datos

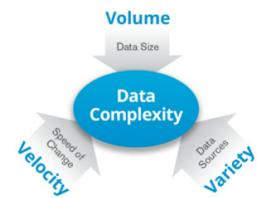


# Big Data

## Volúmen, Variadad y Velocidad

Son las 3 V de big data y su característica principal en cuanto a como se intenga gestionar el dato que se considera complejo.

Estos datos no pueden ser gestionados por sistemas tradicionales.



- Surgen necesidades específicas para canalizar, gestionar y procesar cierto tipo de datos
- Por lo tanto, se desarrollan herramientas específicas para cada caso.
- Las infraestructuras se arman en clusters y se trabaja sobre nuevas características que saquen el máximo provecho del Hardware.
- Ambientes distribuidos, escalabilidad horizontal y disponibilidad continua.

## Casos de Uso

- Detección de fraude.

Detección en todo el canal de operaciones, estas se analizan y se aplican técnicas estadísticas para detectar casos anómalos.

- Motores de recomendación.

Implica para lo que esté recomendando al usuario, análisis de perfiles de usuario tanto como características de los productos, se plantea una clusterización para poder brindar opciones a los usuarios.

- Predicción de comportamientos de clientes.

Acciones de machine learning, capacidad de computo elevado, almacenamientos de grandes volúmenes de datos.

Se analiza el historial de los clientes para poder determinar a futuro el comportamiento de los mismos.

- Análisis de sentimiento.

Se plantea un proceso que esté escuchando la generación de información continuamente, para analizar sobre si se está hablando de determinado producto o de una empresa de forma positiva o negativa.

- Mercadotecnia.

Se analiza y examina grandes volúmenes de datos de distintos tipos de cuales podrían ser los mejores próximos pasos para las campañas de marketing.

- Administración de inventarios.

Se aplican algoritmos basados en estadística que permitan que la operatoria diaria del negocio sea lo más eficiente posible.

Para todos los casos es necesario y contar con gran capacidad de cómputo y hardware para el almacenamiento de los datos.



---

# Data Lake - Data Warehouse

Las necesidades que nos llevan a construir cada uno de estos son:

## Data Lake

- Se requiere visualizar y analizar datos que no son estructurados necesariamente.
- Se requiere soportar las cargas de los procesos de Big Data y Machine Learning.
- Los datos se almacenan y se disponibilizan, aún sin conocer si van a ser utilizados para su análisis.
- Los datos no tienen que ser transformados ni limpiados antes de ser almacenados
- Acceso a los datos en tiempo real tanto en crudo como refinado.
- Escalable y puede manejar grandes cantidades de datos

## Data Warehouse

- Tener un repositorio unificado de todos los datos de la organización.
- Se evalúan las necesidades de análisis y visualización de datos y se integra lo necesario.
- La prioridad es reducir costos de almacenamientos y optimización del proceso ETL.
- Los datos deben ser transformados y limpiados antes de ser almacenados
- Acceso a los datos con una latencia mayor
- Menos escalable y puede tener dificultades para manejar grandes cantidades de datos



## Esquema de procesamiento

### Data Lake

- Esquema “On read”
- Obtener los datos.
- Guardar los datos.
- Transformar los datos.
- Analizar los datos.

### Data Warehouse

- Esquema “On Write”.
- Obtener datos.
- Transformar datos.
- Guardar los datos.
- Analizar los datos.



## Arquitectura Data Lake

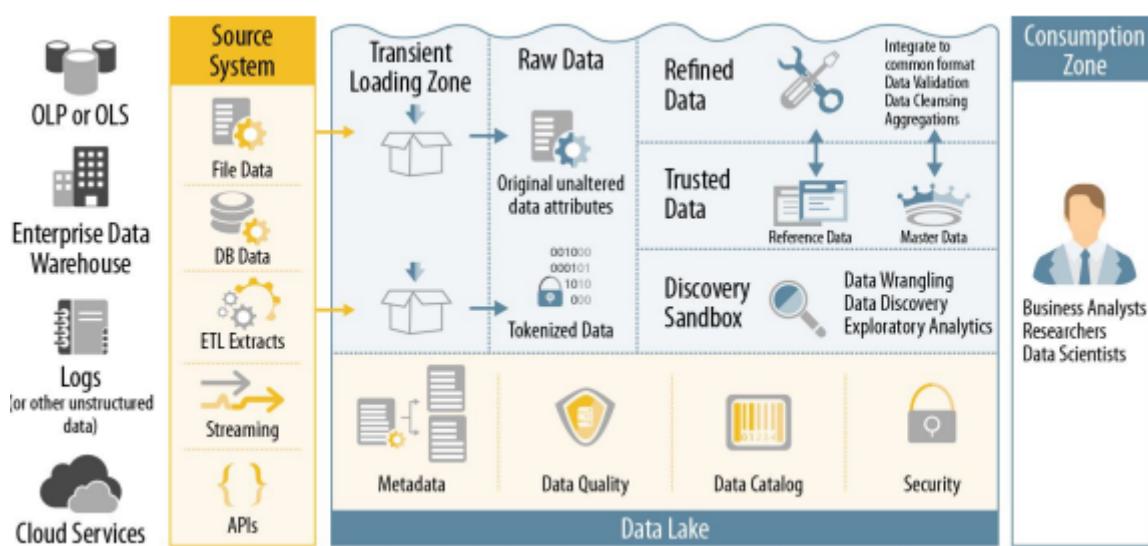
El Data Lake es un repositorio unificado de datos, estructurados y no estructurados.

Prioriza el almacenamiento de los datos en su formato original para luego ser procesados de acuerdo a la demanda.

- Zona de carga y transición.
- Zona de datos en crudo.

Solamente tenemos la información con el formato en el que fue creado.

- Zona de datos refinados.  
Se realizan pasos de transformación de los datos.
- Zona de datos validados.
- Zona de análisis exploración y descubrimiento.  
Se disponibilizan para las personas asignadas a esa tarea.



## Hadoop

Hadoop es una plataforma de software de código abierto utilizada para el procesamiento y almacenamiento de grandes cantidades de datos.

Sus características más importantes son:

- Clúster de servidores

Una red de servidores, conformada por varios nodos, donde cada servidor representa un nodo. El clúster está conformado por nodos maestros y nodos esclavos.

- Tolerancia a fallos.

El clúster es tolerante a fallos, lo que significa que si uno de los nodos presenta un problema y se cae, el clúster puede seguir funcionando.

- Escalabilidad horizontal.

Si se desea mejorar la capacidad de cómputo y almacenamiento del clúster, se pueden agregar más nodos.

- Comodity Hardware.

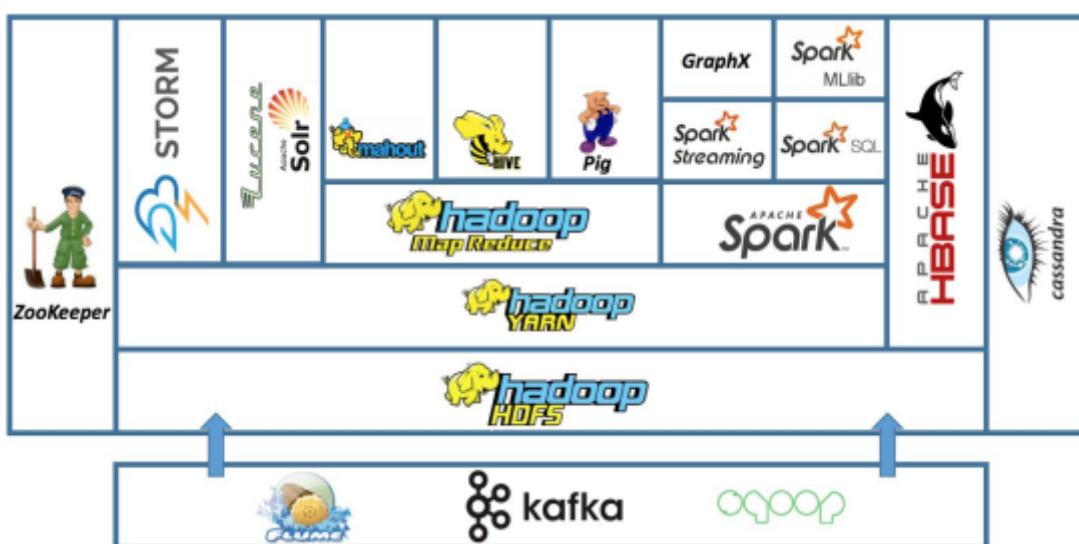
El esquema previo permite que el costo de hardware sea menor.

- Procesamiento en paralelo.

Cada nodo puede subdividir una gran tarea para que cada uno procese una parte y llegar a un resultado final.

- Desarrollado en Java.

Hadoop está desarrollado en Java, lo que lo hace compatible con una amplia gama de sistemas operativos.



## Ecosistema Hadoop

Hadoop es un conjunto de herramientas que se montan sobre un sistema distribuido cuyos componentes fundamentales:

- HDFS (Hadoop Distributed File System)

Es un sistema de archivos distribuido que permite almacenar grandes cantidades de datos de manera distribuida a través de varios nodos en un clúster.

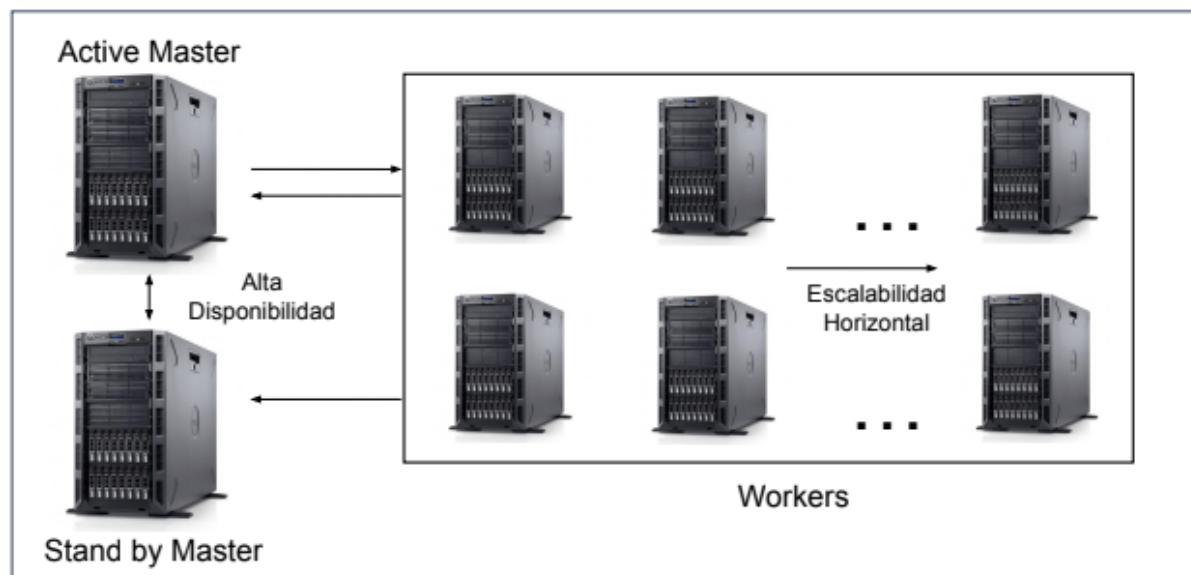
- YARN (Yet Another Resource Negotiator)

Es un gestor de recursos que permite asignar y gestionar los recursos del clúster para el procesamiento de datos. YARN se utiliza junto con MapReduce, que es un sistema de procesamiento de datos en paralelo.

Sobre estos componentes principales, se montan otras herramientas como motores de bases de datos (como Cassandra o HBase), herramientas para la comunicación con otros sistemas (como Kafka y Sqoop), herramientas de bases de datos relacionales (como Hive) y Spark con sus diversos módulos.

## Cluster Hadoop

Un cluster Hadoop es una red de servidores o nodos que están diseñados para el almacenamiento y procesamiento distribuidos de grandes volúmenes de datos, asegurando así la disponibilidad de la información. En un cluster Hadoop, hay un nodo maestro y varios nodos worker.

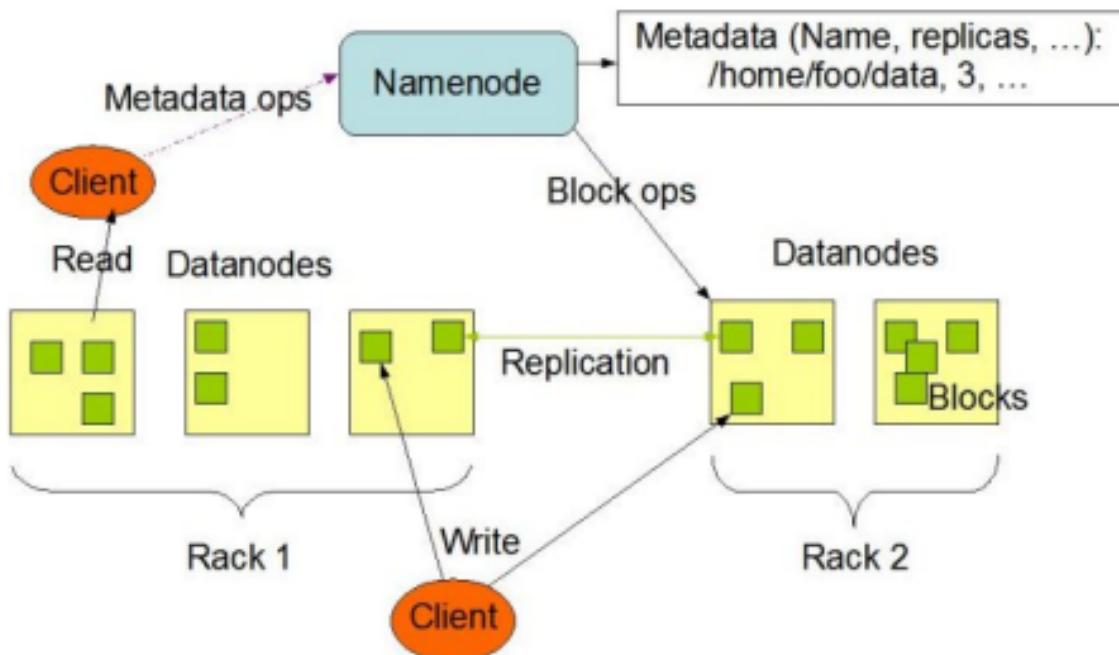


## HDFS (Hadoop Distributed File System)

HDFS (Hadoop Distributed File System) es el sistema de archivos distribuido utilizado por Hadoop. Algunas de sus características más importantes son:

- Permite organizar computadoras en una relación maestro-esclavo dentro del cluster.
- Ofrece escalabilidad para el almacenamiento y el procesamiento.

- Existen dos tipos de nodos:
  - Nodo maestro (NameNode): administra el sistema de archivos, controla el acceso y conoce cómo se distribuyen los archivos en los nodos worker (DataNodes).  
Utiliza un log de transacciones para registrar cada cambio en el sistema de archivos y trata de asegurar una distribución equilibrada de los archivos en el clúster, optimizando también el ancho de banda de la red y el balance de carga de procesamiento y almacenamiento.
  - Nodos worker (DataNodes): almacenan los bloques de datos que conforman los archivos del sistema de archivos.  
Cada DataNode típicamente contiene muchos discos para maximizar la capacidad de almacenamiento y la velocidad de acceso, y tienen su propio sistema de archivos local. Almacenan y distribuyen bloques de datos sobre la red usando un protocolo de bloques.

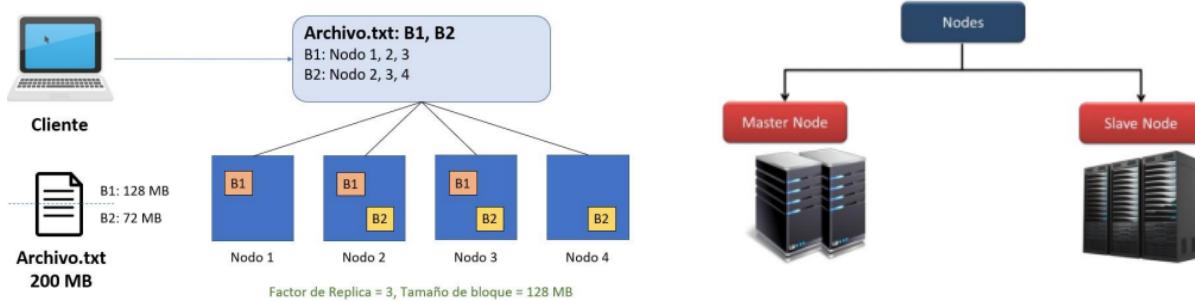


Los NameNodes almacenan toda la información relevante acerca de todos los DataNodes, y los archivos almacenados en los DataNodes:

- Para cada DataNode, su nombre, rack, capacidad y estado.
- Para cada archivo, su nombre, réplicas, tipo, tamaño, "timeStamp", ubicación, estado.
- El sistema de bloques es la unidad fundamental de almacenamiento y un archivo ocupa un determinado número de bloques. El tamaño predeterminado de los bloques es de 64 o 128 MB dependiendo de la distribución:

```
hdfs getconf -confKey dfs.blocksize
```

- Cada archivo es replicado según un factor de réplica. Los datos suelen ser almacenados en tres nodos diferentes, dos en el mismo rack y uno en un rack diferente, para permitir la recreación automática en caso de fallo escribiéndose todos los bloques de archivos desde réplicas.
- Los archivos secuenciales son una estructura de datos especializada para manejar pequeños archivos en registros pequeños.
- MapReduce está diseñado para gestionar archivos de gran tamaño, de manera de “empaquetarlos” en archivos secuenciales pequeños, para facilitar su almacenamiento y procesamiento.

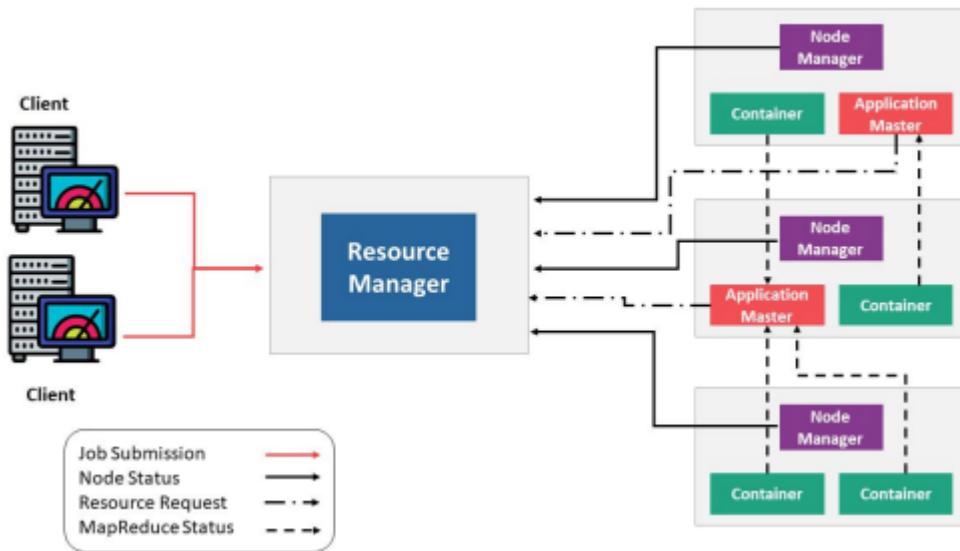


## YARN (Yet Another Resource Negotiator)

YARN es un gestor de recursos que administra y garantiza la disponibilidad de los mismos en un clúster Hadoop. Algunas de sus características más importantes son:

- YARN brinda flexibilidad como una plataforma común para ejecutar múltiples aplicaciones y herramientas de consultas interactivas SQL (Hive), de proceso de flujos en tiempo real (Spark), y procesamiento por lotes (MapReduce) para trabajar con los datos almacenados en una plataforma HDFS.
- Brinda gran escalabilidad para expandirse más allá de 1000 nodos y provee ubicación dinámica de recursos del clúster.

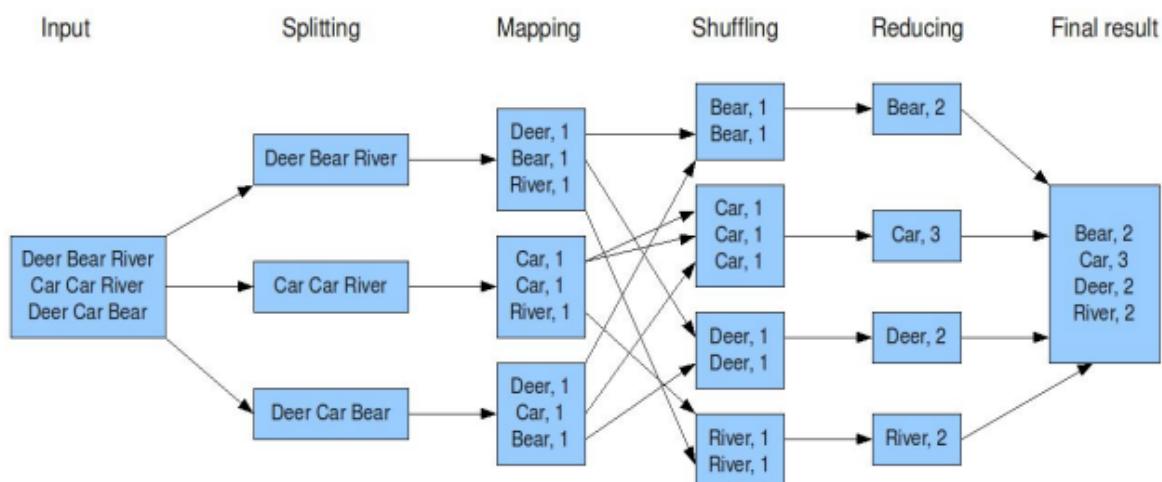
El cluster es un conjunto de recursos, tales como capacidad de almacenamiento, memoria RAM, capacidad de procesamiento, capacidad de infraestructura de red que son administrados por un gestor o módulo encargado de distribuir correctamente la carga sobre todo el cluster.



## MapReduce

MapReduce es un sistema de procesamiento de datos en paralelo que se utiliza en Hadoop para mejorar la capacidad de procesamiento y tolerancia a fallos. Algunas de sus características más importantes son:

- Permite procesar grandes cantidades de datos utilizando la capacidad de procesamiento de varias computadoras que trabajan de forma simultánea en diferentes partes del trabajo.
- El procesamiento se divide en varias partes que se procesan de forma independiente, y luego se combinan los resultados.
- MapReduce acelera el procesamiento de datos a gran escala, con un mínimo movimiento de los datos en el sistema de archivos del clúster, y ofrece resultados cercanos al tiempo real.



---

## Vendors

Son proveedores de servicio de Big Data en la nube.

- Cloudera
- Amazon EMR
- Azure HDInsight
- IBM Analytics Engine
- Google Dataproc
- MapR

## Frameworks

- Hive.  
Consultas SQL sobre Hadoop.
- Sqoop.  
Transferencia entre bases de datos relacionales y Hadoop.
- Spark.  
Procesamiento en memoria de ETL's, Streaming, Machine Learning y Grafos.
- Kafka.  
Sistema de colas de mensajería. Patrón productor/consumidor.
- HBase.  
Base de datos NoSQL columnar que ejecuta sobre HDFS.
- Ranger.  
Administración de políticas de seguridad sobre componentes Hadoop.
- Atlas.  
Funcionalidades de Data Governance en Hadoop.
- Nifi.

Orquestador de flujos de datos, desde y hacia Hadoop.

## Otras tecnologías



## Máquinas Virtuales

Las máquinas virtuales son versiones virtuales de ciertos recursos tecnológicos, como hardware, sistema operativo, dispositivos de almacenamiento o redes.

La virtualización es un proceso en el que se ejecuta un sistema huésped sobre otro anfitrión, pero que tiene su propio sistema de archivos y puede tener varios formatos, como VDI, VMDK, VHD o RAW, entre otros.

## Docker

Docker es una plataforma de virtualización que utiliza contenedores para reutilizar el kernel, es decir, la parte más profunda del sistema operativo de la máquina anfitriona, y optimizar el uso de los recursos disponibles.

La containerización es una forma más ligera, portable, de bajo acoplamiento, escalable y segura de realizar la virtualización.

Característica	Máquinas Virtuales	Docker
¿Qué es?	Una máquina virtual es un software que simula un sistema	Docker es una plataforma de contenedores que permite ejecutar

	informático completo con su propio sistema operativo, hardware y aplicaciones.	aplicaciones en entornos aislados.
¿Cómo funciona?	Una máquina virtual se ejecuta en un sistema operativo huésped y emula hardware para permitir que otro sistema operativo (invitado) se ejecute en ella. Esto significa que se necesita una máquina virtual por cada sistema operativo que se quiera ejecutar.	Docker utiliza contenedores, que son una forma ligera de empaquetar y distribuir aplicaciones. Los contenedores comparten el kernel del sistema operativo del host y pueden ejecutarse en cualquier máquina que tenga Docker instalado, sin necesidad de un sistema operativo invitado.
Recursos necesarios	Una máquina virtual necesita su propio sistema operativo, lo que significa que necesita recursos para ejecutarlo. Además, cada máquina virtual necesita su propio espacio de almacenamiento y memoria RAM.	Los contenedores comparten el kernel del sistema operativo del host y, por lo tanto, son más ligeros y necesitan menos recursos que las máquinas virtuales.
Facilidad de uso	Configurar y utilizar una máquina virtual puede ser complicado, ya que se necesita instalar un sistema operativo y configurarlo adecuadamente.	Docker es más fácil de usar, ya que los contenedores ya están configurados y listos para usar, por lo que es más rápido implementarlos y ejecutarlos.
Facilidad de distribución	Las máquinas virtuales son más difíciles de distribuir debido a su tamaño y a la cantidad de recursos que necesitan.	Los contenedores son más fáciles de distribuir, ya que son más ligeros y pueden ejecutarse en cualquier máquina que tenga Docker instalado.
Aislamiento	Las máquinas virtuales proporcionan un aislamiento completo, ya que cada una tiene su propio sistema operativo y hardware.	Los contenedores proporcionan aislamiento

## Docker Engine

Docker Engine es la plataforma principal de Docker que permite ejecutar contenedores en un sistema operativo. Se ejecuta de forma nativa en Linux, pero para otros sistemas operativos requiere levantar una máquina virtual. Algunos de sus componentes más importantes son:

- Docker Daemon

Es el centro de Docker que permite la comunicación con los servicios de Docker.

- REST API

Es una interfaz de programación de aplicaciones que permite visualizar Docker de forma gráfica.

- Cliente de Docker

Es una herramienta que permite la comunicación con el centro de Docker (Docker Daemon), y que por defecto se utiliza a través de la línea de comandos.