

Entornos de computación distribuida para grandes volúmenes de datos

El **BIG DATA** permite obtener información útil a partir de volúmenes de datos grandes. El objetivo es resolver problemas nuevos y mejorar la solución de problemas viejos.

El big data ha surgido debido al incremento en la cantidad de datos generados disponibles y debido a la aparición de tecnologías de bajo coste que permiten el almacenamiento de todos estos datos.

Las 5 “V” del Big data son, **Volume**, **Velocity**, **Variety**, **Veracity** y **Value**.

El problema del big data es su volumen de datos. Para solucionar este problema se aborda el tema de la **escalabilidad**, esto significa que los sistemas deben tener mayor capacidad de almacenamiento y procesamiento. Hay dos tipos de escalabilidad:

- **Vertical**: Se aumenta la potencia de la máquina. Si el software ya está preparado para esta mejora aumenta la simplicidad, pero tarde o temprano habrá que modificar el software.
- **Horizontal**: Se distribuye la carga de trabajo entre más ordenadores. Los límites ahora son mucho más altos. Comparada con la escalabilidad principal, cuanto más capacidad, menor es el aumento con respecto a la mejora de la escalabilidad vertical.

La escalabilidad horizontal implica cierta probabilidad de que falle uno de los nodos, por lo tanto, necesitamos tener 3 propiedades deseables en cada nodo:

- **Consistencia**
- **Disponibilidad**
- **Tolerancia al particionado**