# INTRODUCCIÓN A BIG DATA

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) 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.

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.

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 web logs, la identificación por radiofrecuencia (RFID), los sensores incorporados en dispositivos, la maquinaria, los vehículos, las búsquedas en Internet, las redes sociales como Facebook, computadoras portátiles, teléfonos

inteligentes y otros teléfonos móviles, dispositivos GPS y registros de centros de llamadas.

En la mayoría de los casos, con el fin de utilizar eficazmente el Big Data, debe combinarse con datos estructurados (normalmente de una base de datos relacional) de una aplicación comercial más convencional, como un ERP (Enterprise Resource Planning) o un CRM (Customer Relationship Management).

**Nota**

Twitter, Facebook, Amazon, Verizon, Macy's y Whole Foods son todas las empresas que dirigen su negocio utilizando análisis de datos y basan muchas de las decisiones en él. Piense en qué tipo de datos están recopilando, cuántos datos podrían estar recopilando y, a continuación, cómo podrían estar utilizando los datos.

Echemos un vistazo a nuestro ejemplo de tienda de comestibles visto anteriormente. **¿Qué pasa si la tienda comienza a expandir su negocio para establecer 100s de tiendas?** Naturalmente, las transacciones de ventas tendrán que ser recogidas y almacenadas en una escala que es 100s de veces más que la tienda única. Pero entonces, ya no hay negocios que trabajen de forma independiente. Hay mucha información por ahí a partir de noticias locales, tweets, reseñas de yelp, quejas de clientes, actividades de encuestas, competencia de otras tiendas, cambios demográficos, o la economía del área local, y así sucesivamente. Todos estos datos adicionales pueden ayudar a comprender mejor el comportamiento del cliente y los modelos de ingresos.

Por ejemplo, si vemos un sentimiento negativo creciente con respecto a la instalación de estacionamiento de la tienda, entonces podríamos analizar esto y tomar medidas correctivas como estacionamiento validado o negociar con el departamento de transporte público de la ciudad para proporcionar trenes o autobuses más frecuentes para un mejor alcance.

Esta cantidad creciente y una variedad de datos, al tiempo que proporciona un mejor análisis, también plantea desafíos a la organización de TI empresarial que intenta almacenar, procesar y analizar todos los datos. De hecho, no es raro ver TBs de datos.

**Nota**

Cada día, creamos más de 2 quintillones de bytes de datos (2 Exa Bytes), y se estima que más del 90% de los datos se han generado solo en los últimos años. 1

KB a 1024 Bytes1 MB a 1024 KB1 GB a 1024 MB1 TB , 1024 GB, 1.000.000 MB1 PB, 1024 TB, 1.000.000 GB, 1.000 MB, 1.000 MB 000,000,000 MB1 EB a 1024 PB a 1.000.000 DE TB a 1.000.000 de 000 GB a 1.000.000.000 de MB

Tales grandes cantidades de datos desde la década de 1990, y la necesidad de entender y dar sentido a los datos, dio lugar al término big data.

El término big data, que abarca la informática y las estadísticas/econometría, probablemente se originó en las conversaciones de la mesa de almuerzo en Silicon Graphics a mediados de la década de 1990, en las que John Mashey figuraba prominentemente.

En 2001, Doug Laney, entonces analista de la consultora Meta Group Inc (que fue adquirida por Gartner) introdujo la idea de 3Vs (variedad, velocidad y volumen). Ahora, nos referimos a 4 Vs en lugar de 3Vs con la adición de la veracidad de los datos a los 3V.

## 4 “V” de Big Data

Los siguientes son las 4 “V” de los datos utilizados para describir las propiedades del Big Data.

## Variedad de datos

Los datos pueden ser de sensores meteorológicos, sensores de automóviles, datos censales, actualizaciones de Facebook, tweets, transacciones, ventas y marketing. El formato de datos también está estructurado y no estructurado. Los tipos de datos también pueden ser diferentes; binario, texto, JSON y XML.

## Velocidad de los datos

Los datos se pueden obtener en un almacén de datos, archivos de archivos en modo por lotes, actualizaciones casi en tiempo real o actualizaciones instantáneas en tiempo real del viaje de Uber que acabas de reservar.

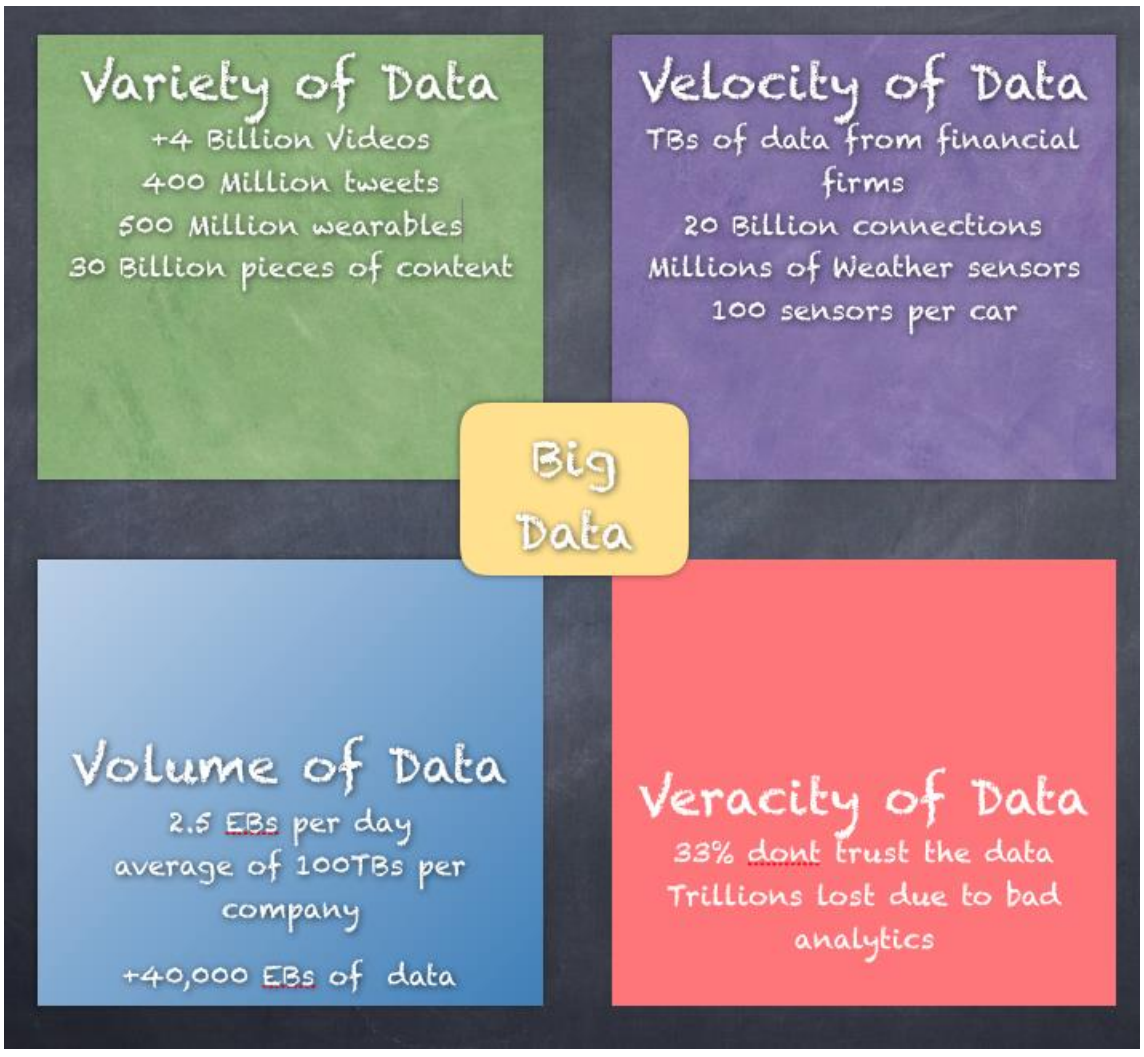
## Volumen de datos

Los datos se pueden recopilar y almacenar durante una hora, un día, un mes, un año o 10 años. El tamaño de está creciendo a 100s de TBs para muchas empresas.

## Veracidad de los datos

Los datos se pueden analizar para obtener información procesable, pero con los datos de todos los tipos que se analizan desde todas las fuentes de datos, es muy difícil garantizar la corrección y la prueba de precisión.

**Los siguientes son los 4 Vs del big data:**




Para dar sentido a todos los datos y aplicar el análisis de datos al big data, necesitamos ampliar el concepto de análisis de datos para que funcione a una escala mucho mayor que se ocupe del 4 Vs del big data. Esto cambia no sólo las herramientas, tecnologías y metodologías utilizadas en el análisis de datos, sino también la forma en que incluso abordamos el problema. Si una base de datos SQL se usó para datos en una empresa en 1999, ahora para manejar los datos

para el mismo negocio, necesitaremos una base de datos SQL distribuida escalable y adaptable a los matices del espacio de **big data.**

Las aplicaciones de análisis de **Big Data** a menudo incluyen datos de sistemas internos y fuentes externas, como datos meteorológicos o datos demográficos sobre consumidores compilados por proveedores de servicios de información de terceros. Además, las aplicaciones de análisis de streaming se están volviendo comunes en entornos de big data, ya que los usuarios buscan realizar análisis en tiempo real de datos alimentados en sistemas Hadoop a través del módulo de streaming Spark de Spark u otros motores de procesamiento de flujos de código abierto, como Flink y Storm.

Los primeros sistemas de big data se implementaron principalmente en las instalaciones, especialmente en grandes organizaciones que recopilaban, organizaban y analizaban cantidades masivas de datos. Sin embargo, los proveedores de plataformas en la nube, como Amazon Web Services (AWS) y Microsoft, han facilitar la configuración y administración de clústeres de Hadoop en la nube, al igual que los proveedores de Hadoop, como Cloudera y Hortonworks, que admiten sus distribuciones del marco de big data en las nubes de AWS y Microsoft Azure. Los usuarios ahora pueden crear clústeres en la nube, ejecutarlos durante el tiempo que sea necesario y, a continuación, desconectarlos, con precios basados en el uso que no requieren licencias de software en curso.

Los posibles escollos que pueden hacer frente a las organizaciones en iniciativas de análisis de big data incluyen la falta de habilidades de análisis interno y el alto costo de contratar científicos de datos e ingenieros de datos experimentados para llenar las lagunas.

La cantidad de datos que suelen implicarse y su variedad pueden causar problemas de administración de datos en áreas como la calidad de los datos, la coherencia y la gobernanza; también, los silos de datos pueden resultar del uso de diferentes plataformas y almacenes de datos en una arquitectura de big data. Además, la integración de Hadoop, Spark y otras herramientas de big data en una arquitectura cohesiva que satisfaga las necesidades de análisis de big data de una organización es una propuesta desafiante para muchos equipos de TI y análisis, que tienen que identificar la combinación correcta de tecnologías y luego juntar las piezas.