

# ***Capítulo 1***

## ***Introducción al Big Data***

**Al finalizar el capítulo, el alumno podrá:**

- Identificar el valor agregado del análisis de datos para las empresas.
- Conocer las tendencias del mercado en Big Data.


### **Temas**

1. ¿Qué es Big Data?
2. Características del Big Data
3. Las V's del Big Data
4. El impacto del Big Data en los negocios
5. Ejemplos de Big Data

## 1. ¿Qué es Big Data?

**¿Qué es Big Data?**

Big Data es un conjunto de conocimientos, métodos y tecnologías orientados a facilitar la adquisición, gestión y uso de los datos generados en la actualidad.



1 - 8

Copyright © Todos los Derechos Reservados - Cibertec Perú S.A.C.

**Wikipedia** indica que: Big data o macrodatos es un término que hace referencia a una cantidad de datos tal que supera la capacidad del software convencional para ser capturados, administrados y procesados en un tiempo razonable.

**IBM:** el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

**Google:** Los datos grandes se refieren a datos que, por lo general, serían demasiado caros de almacenar, gestionar y analizar utilizando sistemas de bases de datos tradicionales (relacionales y / o monolíticos). Por lo general, estos sistemas son ineficientes en función de los costos debido a su inflexibilidad para almacenar datos no estructurados (como imágenes, texto y video), acomodar datos de "alta velocidad" (en tiempo real) o escalar a soporte muy grande (escala de petabytes ) volúmenes de datos.

**Gartner:** : Big Data son activos de información de gran volumen, alta velocidad y / o gran variedad que demandan formas rentables e innovadoras de procesamiento de la información que permiten un mejor conocimiento, toma de decisiones y automatización de procesos.

Big Data es un conjunto de conocimientos, métodos y tecnologías orientados a facilitar la adquisición, gestión y uso de los datos generados en la actualidad.

## 2. Características del Big Data

### Características del Big Data

- N = ALL
  - Fin del muestreo
  - Todo el conjunto de datos es válido, no se descartan casos
  - Se puede eliminar el problema del sesgo
- La inexactitud de los datos ya no es un problema
  - El error de la muestra se minimiza, asumimos datos menos exactos.
  - Dicho de otra forma, cuando dependemos de una muestra, queremos que los datos sean exactos.
- Obtenemos el qué, pero no el porqué
  - Las técnicas del Big Data no explican la causalidad (correlaciones).

1 - 9 Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.

Entre las principales características del Big Data tenemos:

- Ya no se utiliza muestreo para el análisis de datos, ahora se obtiene toda la información. Al disponer de ella se pueden analizar todos los casos de ser necesarios.
- La inexactitud de los datos ya no es un problema, debido a que se dispone de toda la fuente de datos la cual brinda todos los datos exactos e inexactos.
- Obtenemos el qué, pero no el porqué, esto es debido a que las técnicas de Big Data no explican la causalidad (correlaciones).
- Utilización de la datificación con el fin de convertir cualquier acción o evento susceptible de ser medido en datos digitales, para que posteriormente sea tabulado y analizado. Dicho análisis permite detectar patrones de comportamiento.

### 3. Las V's del Big Data



En este capítulo vamos a proporcionar una definición de Big Data, que inevitablemente está asociada a lo que se conoce como las Vs del Big Data.

#### a) Volumen

Volumen de los datos disponibles para el análisis, los cuales exceden la capacidad de los RDBMS tradicionales y que son gestionados por diversos sistemas (correo, Facebook, otros).

- Actualmente las empresas que dan soporte al Big Data manejan volúmenes que van desde unos pocos terabytes ( $10^3$  Gb) hasta los petabytes ( $10^6$  Gb)
- En la actualidad muchas ya superan los 10 terabytes y se espera que en tres años la norma sean 100 terabytes.

Pero, el volumen, no es lo más importante. Por ejemplo, no supone el mismo esfuerzo analizar y extraer conocimiento de 1 terabyte de texto que de 1 gigabyte de imágenes médicas. Se ha de valorar el tamaño en relación a los recursos disponibles, las preguntas a las que buscamos dar respuesta y el tipo de datos analizados.

**b) Velocidad**

- Incremento de la velocidad a la que se genera (y se distribuye) los datos.
- Datos en “streaming”: Nuevas fuentes que se generan y distribuyen en tiempo real. Ejemplo: datos generados por sensores, servidores web o las redes sociales.
- El aumento de la velocidad de generación es una de las razones del incremento del volumen de datos, pues los pequeños mensajes o eventos generados en cada instante, por dispositivos y aplicaciones como los anteriores, dan lugar a enormes conjuntos de datos.
- La velocidad también hace referencia a la necesidad para extraer conocimiento de los datos en el momento oportuno. Ejemplo: Datos de carácter financiero pueden reducir su valor en cuestión de segundos.

**c) Variedad**

- La variedad se refiere al importante aumento en la heterogeneidad en las fuentes de datos debido a diversos factores como:
  - Incremento en el número de fuentes disponibles.
  - Posibilidad de procesar fuentes con distinto nivel de estructura.
  - Diversidad de formatos en que se distribuyen las fuentes.
- Tradicionalmente los datos de una organización se han almacenado en formatos altamente estructurados
  - Ejemplo: Sistemas de Gestión de Bases de Datos Relacionales (SGBDR)
- En el nuevo escenario actual, a estas fuentes se unen otras, como las semiestructuradas (ejemplo. XML, JSON) o las que carecen de estructura alguna (ejemplo. texto, imagen o video).

**d) Veracidad**

- Aumento de la incertidumbre respecto a la veracidad o calidad de los datos disponibles
  - Incertidumbre de datos → incertidumbre del conocimiento extraído
- Es uno los retos principales del nuevo contexto:
  - Su comprobación sobrepasa las capacidades del ser humano.
  - Las fuentes de datos, incluso en el mismo dominio, difieren ampliamente en su calidad en lo que respecta a la cobertura, precisión y oportunidad de los datos proporcionados.
  - Uso de datos incorrectos supone grandes pérdidas (varios billones al año).

**e) Valor**

- Además de las características anteriores, es necesario tener en cuenta el valor, este se define como una medida de la utilidad de los datos, para la toma de decisiones en la organización, poniendo de manifiesto la dificultad para conocer y evaluar dicha utilidad a priori.

## 4. El impacto del Big Data en los negocios


### El impacto del Big Data en los negocios

#### El valor de los datos

- El uso intensivo de los datos ha pasado a ser el petróleo de muchas compañías.
- El nuevo enfoque es almacenar cualquier tipo de dato, por irrelevante que pueda parecer, para su posterior análisis.
  - Clics de ratón en la página web de mi negocio.
  - Vibración del motor del coche.
  - Movimiento del acelerómetro del Smartphone.
- Permite crear modelos para responder preguntas complejas, mostrar percepciones y brindar resultados únicos.

1 - 18

Copyright © Todos los Derechos Reservados - Cibertec Perú SAC.



El valor de los datos se ha convertido en el petróleo de muchas compañías. En Big Data se va almacenar cualquier tipo de dato, por irrelevante que pueda parecer para su posterior análisis.

Algunos ejemplos:

- Clic en la página web de mi negocio.
- Vibración de motor de un auto

Al disponer de una gran cantidad de datos, vamos a poder crear modelos para responder preguntas complejas, mostrar percepciones contra intuitivas y aprender de resultados únicos.

El Big Data abarca todos los sectores como el sector automovilístico (recorrido del auto), salud (saber qué tipo de medicamento toma y como afecto su tratamiento), o registrar el movimiento de un atleta en una competencia.

Sin embargo, para aprovechar el Big Data, se ha creado un nuevo perfil llamado Data Scientist, el cual tiene una alta demanda en el sector. Este tiene un tercio de las habilidades de las ciencias de la computación, un tercio de las habilidades de matemáticas y estadísticas y finalmente es un experto del negocio.



## 5. Ejemplos de Big Data

**Ejemplo de Big Data - Netflix**

Utiliza la información de sus suscriptores para predecir que contenidos tienen más probabilidades de triunfar.

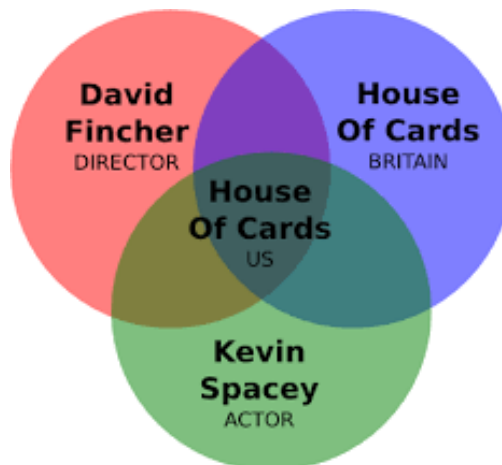
- ¿Qué búsquedas realizan?
- ¿Qué dispositivos usan?
- ¿Cuál es su día preferido?
- ¿Cuánto tiempo emplean en el servicio y en cada uno de los contenidos?
- Si ven los capítulos enteros o parcialmente, incluso, ¿qué fragmentos vuelven a visionar?
- ¿En qué momento abandonan el visionado y si lo recuperan o abandonan?
- Las valoraciones de los consumidores.
- ¿Qué preferencias tienen en común con sus amigos o con la audiencia de su misma zona geográfica?
- La información de sus perfiles en redes sociales.




1 - 22 Copyright © Todos los Derechos Reservados - Cibertec Perú S.A.C.

Uno de los casos del uso de Big Data es Netflix el cual utiliza al máximo su capacidad de Big Data para producir una serie que se adaptará a las preferencias de su audiencia. Así, analizando sus datos, Netflix observó que tenía una gran cantidad de usuarios que tenían tres factores en común:

- Muchos habían visto “Red Social”, película dirigida por David Fincher, de principio a fin.
- La versión británica de “House of Cards” había sido bien valorada.
- Quienes vieron la versión británica de “House of Cards”, también habían visto películas protagonizadas por Kevin Spacey y/o dirigidas por David Fincher.





La combinación de estos factores, sumados a la popularidad de los thrillers políticos, parecía cerrar una fórmula perfecta para Netflix. Además, optaron por estrategias innovadoras de distribución, como estrenar todos los capítulos de una misma temporada en simultáneo.

Netflix invirtió más de 100 millones de dólares en la producción de las dos primeras temporadas de la serie y se encargó de promoverla. Con los datos disponibles, hicieron un “tráiler personalizado” para cada tipo de miembro de Netflix. La empresa hizo 10 cortes diferentes del tráiler de “House of Cards”, cada una dirigida hacia diferentes públicos, con base en su comportamiento de visualización anterior. Si habían visto muchas películas de Kevin Spacey, verían un tráiler donde él apareciera; aquellos que vieron una gran cantidad de películas con protagonistas mujeres, vieron un tráiler con mujeres; y los fans de David Fincher vieron un tráiler con su toque.

Ya fuera una estrategia para ganar suscriptores o para diferenciarse de las cadenas de televisión tradicionales, a Netflix le fue muy bien con su nuevo producto. Solo en el primer trimestre del 2013, atrajo dos millones de nuevos suscriptores de Estados Unidos, lo cual fue un aumento del 7% respecto del trimestre anterior. También trajo un millón de nuevos suscriptores de otras partes del mundo.

En 2015, la plataforma de streaming produjo 16 ficciones propias –series, películas, documentales y otros– y este año tiene previsto producir un total de 31 series.

¿Qué lección nos deja Netflix? La hipercomercialización e hipersegmentación del cliente que se logra con una estrategia eficaz del Big Data, el cual permite a las empresas dar respuestas precisas a las nuevas demandas de consumo. A su vez, posibilita generar estrategias exitosas para retener a los clientes existentes, maximizando beneficios y, por tanto, brindando un producto o servicio superior.