# Modulo 2. Características y escalabilidad

## Unidad 1. Características de Big Data y cómo hacerlo escalable

Hablar de Big Data implica tratar distintas cuestiones al mismo tiempo. A diferencia del paradigma tradicional, que solo pone la atención en el volumen y tamaño de los datos recopilados, trabajar con grandes cúmulos de información supone distintos tipos de elementos, útiles y vitales, para realizar una incorporación escalable y sostenible en el tiempo, que genere valor para quien lo está utilizando, manteniendo y financiando.

Por eso, es necesario conocer y entender las características de Big Data y cómo influyen sobre el proceso natural de un proyecto o programa de este tipo.

Existen modelos clásicos, como el de las 3 V (volumen, velocidad y variedad), y también críticas y planteamientos de otras alternativas para solucionar problemas prácticos de Big Data, enfocados en casos reales.

Este capítulo está destinado a hablar de los distintos modos enfocados en crear proyectos sostenibles y escalables que generen valor y retorno de inversión en el desarrollo de proyectos de Big Data.

### 2.1.1 Características principales de Big Data

Los procesos de Big Data suelen estar basados en tres características que rigen la lógica con la que se tratará el análisis de datos de gran envergadura. En este sentido, se puede hablar del modelo clásico de las 3 V:

- Volumen
- Variedad
- Velocidad

Según el autor que se consulte, a estos tres elementos se le podrá sumar uno que influye principalmente sobre la calidad del análisis resultante: la veracidad. Esta cuarta característica se ha vuelto cada vez más popular, ante la necesidad de aumentar consistentemente la calidad de las conclusiones obtenidas del análisis de datos.

Para explicar cada una en detalle podemos decir lo siguiente:

- El volumen es la cantidad de datos almacenados por cada *dataset* y, en conjunto, entre los distintos recursos utilizados para este fin. El tamaño de los datos determina el valor y el potencial de generar conclusiones, según sea o no Big Data. Si bien en la actualidad se considera que hablamos de Big Data cuando el *dataset* utilizado para analizar tiene un tamaño superior a 1terabyte, no existe un criterio único de tamaño para determinar dicha frontera. Esto se da, principalmente, porque la evolución del tamaño de la información generada hace que sea exponencial la cantidad de datos que se procesan. Hace 10 años podría haberse considerado como una gran cantidad de información a unos pocos gigabytes y, seguramente, en un futuro cercano será considerado común el uso de varios petabytes. Estas escalas varían, ya que los volúmenes crecen año tras año.
- La variedad de la tipología o naturaleza de los datos también influye en la forma de trabajar sobre Big Data. Existe una gran cantidad de formatos que pueden ser utilizados para este fin; los más comunes son los textos, imágenes, videos, audios, cifras e incluso la imputación de datos para determinar elementos perdidos o datos faltantes. Esta variedad puede ser dividida en dos grandes áreas: datos estructurados y datos no estructurados. Esto no solo afecta a cómo se relacionan y clasifican los datos, sino también a la forma en que se almacena, se manipula y se le da utilidad a cada byte analizado. La multiplicidad de fuentes y formatos de datos afecta su complejidad y,

por lo tanto, aproxima su necesidad a tomar técnicas de Big Data para descubrir patrones, tendencias y hacer segmentaciones.

- La velocidad a la que los datos son requeridos o analizados influye en la utilización de Big Data para su manipulación. Un volumen importante de datos, aunque no supere el parámetro de 1 terabyte de información, puede considerarse Big Data si la necesidad de responder rápidamente requiere un procesamiento importante. En gran medida, se puede utilizar Big Data en períodos cortos de tiempo, e incluso necesitar una respuesta, reacción o acción concreta en tiempo real, las cuales deben ser ejecutadas en milisegundos, por lo que un procesamiento lento no sería suficiente.

- La veracidad de los datos plantea una verificación que garantice que la información obtenida de esos datos pueda ser útil y pertinente. Tener datos que no sean certeros, completos o verídicos afecta completamente la calidad del análisis, porque implicaría realizar esfuerzos sobre elementos no verdaderos. ("Las 7 V del Big Data: Características más importantes", 2016).

  A estas características se le agregan otras mínimas que debería tener el equipo de analistas o científicos de datos que los analizarán para crear alguna conclusión de valor.

  Para tener un equipo apropiado que trabaje sobre proyectos de Big Data es necesario reunir al menos 3 características esenciales:

1. Conocimiento sobre el tipo de negocio o actividad que se pretende analizar. Si no se sabe qué buscar, los patrones y tendencias no serán advertidos con la misma firmeza que una persona que entiende cómo puede generar valor con ello y darle un accionable concreto.

2. Habilidades técnicas y de programación. Muchas de las técnicas de análisis relacionadas con Big Data y Analytics refieren a la ejecución de algoritmos programados o a la ejecución sobre una infraestructura tecnológica. Para esto, es necesario entender los requerimientos para elevar la *performance* de las acciones ejecutadas por capacidad de cómputo.

3. Habilidades estadísticas y matemáticas para poder generar modelos que predigan, segmenten y estimen comportamientos, interacciones y cualquier característica común a los datos que se están investigando.

En el módulo 4 hablaremos más en detalle sobre las habilidades y *skills* necesarios que debería tener un equipo de Big Data.

En distintas charlas y conferencias que constantemente llevo a cabo sobre temas de Big Data, suelo incluir usos comunes y realistas en distintos rubros que buscan cambiar al mundo o, al menos, sus realidades particulares con el análisis de datos. Algunos de estos casos son:

- **A nivel del Gobierno,** se realizan múltiples iniciativas en todo el mundo, desde predecir dónde habrá un delito y distribuir a las fuerzas de seguridad para minimizar su probabilidad de ocurrencia hasta la simulación de distintos caminos para resolver un conflicto entre países, evaluación de variables para la disminución del déficit fiscal, productividad e innovación en trámites y procesos de atención a la ciudadanía, entre otros.
- **Con respecto al desarrollo internacional**, se pueden realizar mapeos de distribución de recursos para poder mitigar la pobreza, combatir plagas y mejorar la salud de las personas, por ejemplo, evitando brotes de epidemias y mejorando la calidad de vida a través de la automatización de procesos monótonos que solían ser ineficientes con la intervención humana.
- **Con respecto a la fabricación industrial y a la producción de bienes**, la estimación de *forecasting* y la demanda de inventario y de productos puede implicar la vida o muerte de una compañía. Por ello, es necesario aprender a realizar mejores estimaciones o, al menos, obtener una menor desviación de la realidad contra el valor predicho.
- **En el ámbito de las instituciones relacionadas con la salud**, es posible desde hacer más eficiente la detección de enfermedades y sus respectivos tratamientos para salvar la vida de un paciente hasta el correcto manejo de la institución desde su gerenciamiento para maximizar sus recursos y beneficios.
- **Con respecto a la estimación de la demanda**, en casi cualquier rubro se genera incertidumbre cuando se pretende predecir un suceso, más aún cuando hay que estimar cuánta demanda habrá para un producto, servicio o evento. En cada compañía hay distintos tamaños de equipos y recursos para esta tarea, pero las dispersiones entre lo pronosticado y lo que efectivamente sucede pueden mejorar notablemente.

- **En el ámbito de la educación**, se generan nuevas formas de aprendizaje a través del seguimiento personalizado de cada alumno, en tiempo real, a medida que va aprendiendo. Así, se empodera a cada persona que toma una asignatura para que logre no solo aprenderla, sino también contribuir al conocimiento colectivo, a la investigación y al desarrollo de dicha profesión.

- **Con respecto al *marketing* digital**, optimizar los esfuerzos es una tarea cada vez más estratégica. Se delega paulatinamente la ejecución a robots o *softwares* de automatización, que son los que hacen el trabajo fuerte operativo, y los humanos pasan a tener un rol de planeamiento, definición de criterios y configuración de herramientas.

- **Con respecto a la automatización de trabajos rutinarios y a la creación de nuevos empleos**, el ser humano debe adaptarse, puesto que la automatización y la inteligencia artificial están reemplazando a los trabajadores humanos en tareas rutinarias, como líneas de fabricación, conducción de un auto o camión, o atención al cliente. Posteriormente, estas actividades incluirán labores de toma de decisiones con base en parámetros estadísticos que permitan calcular la probabilidad de ocurrencia de un suceso. Para esto es necesario no solo saber que los empleos con disciplinas automatizables, que actualmente tienen humanos, tienen los días contados, sino también que esta realidad genera empleos que hasta hace algunos años no existían y que renuevan la base laboral de las empresas y naciones a partir de trabajos más calificados y tecnificados.

  Estos son solo algunos ejemplos de los posibles casos de aplicación de técnicas de Big Data. En cada uno de ellos, juega un rol preponderante cada factor característico, y su manipulación, ejecución, mantenimiento y almacenamiento tienen particularidades específicas para cada caso. Por esto, enmarcar todo en una teoría única es complejo cuando se desea dar una bajada práctica sobre la forma de aplicar estos conocimientos a distintas industrias, porque depende de la forma con la que, en cada caso, se deseen resolver sus problemas.

### 2.1.2 La escalabilidad en Big Data

El uso de Big Data ha desarrollado un trastorno constante en las capacidades de almacenamiento y procesamiento de datos, ya que las capacidades del *hardware* físico son limitadas y paulatinamente se llega al punto de la saturación. Es necesario ampliar la capacidad instalada, lo cual puede ser bastante desafiante por el consumo de recursos requerido.

Al hablar de escalabilidad no solo se hace referencia a un rápido crecimiento en volumen de almacenamiento y procesamiento de datos, sino también a la flexibilidad que permite que las capacidades se adapten cuando la demanda decrece.

Escalabilidad eficiente implica pensar tanto en picos altos como bajos y aprovechar ambos extremos para brindar experiencias de calidad a la hora de realizar análisis de datos, así como también ser costo-eficientes y tener la menor capacidad ociosa posible.

Para lograr escalabilidad es necesario identificar dónde se generan los cuellos de botella, y para ello hay tres puntos en los que son bastante comunes:

- **Alto uso del CPU**: suele ser el cuello de botella más común y visible, dado que disminuye el desempeño de los servidores de forma contundente y limita la capacidad y velocidad con que los equipos trabajan.
- **Memoria con poca disponibilidad**: normalmente, al correr procesos complejos que saturan la memoria disponible en los equipos, se produce una eficiencia muy baja y los servidores o equipos con poca memoria no pueden correr todas las aplicaciones o procesos de análisis necesarios para lograr extraer el conocimiento requerido de los sets de datos. Este caso puede tener dos soluciones principales: añadir mayor memoria RAM o identificar si existe pérdida de memoria, dónde se encuentra y repararla.
- **Alto uso del disco**: es un gran indicador de la necesidad de escalabilidad, puesto que está directamente relacionado con el almacenamiento de datos. Si se llena el disco significa que se está almacenando mayor cantidad de información que la que el disco puede manejar.

Para escalar en la base de su infraestructura, se puede aumentar la capacidad de los equipos físicos propios o implementar espacio en la nube, y pagar solo por lo utilizado.

Mantener infraestructura eficiente y escalable permite mejorar las capacidades y productividad de cualquier proyecto de Big Data.

### 2.1.3 Críticas al Modelo de las V

Las características descritas en este capítulo hablan sobre volumen, variabilidad, velocidad y, en algunos casos, de veracidad. Este modelo no es definitivo puesto que tiene muchos adeptos, pero también muchos detractores.

Como, por ejemplo, Davenport (2014) en su libro *Big Data at Work* afirma que, si bien Big Data tiene mucho que ver con el tamaño de los datos, el desafío principal pasa por su estructura. Esto debido a que es simple guardar toda la información que se genera, pero para analizarla hay que darle estructura. Las bases de datos tradicionales requieren mínimos esfuerzos para estructurarlos, pero las bases de datos no estructuradas requieren grandes esfuerzos.

El Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid plantea la existencia de 7 V, no solo de 3. ("Las 7 V del Big Data: Características más importantes", 2016). Las principales críticas radican en que se dan por sentados elementos que no son comprobables hasta que se demuestren empíricamente. Así mismo, todo depende del tipo de uso o finalidad que se pretenda dar al proyecto de Big Data y su alcance real.

Las principales críticas que recibe este modelo radican en los siguientes factores:

- **Completitud de los datos:** Se le critica que no siempre se entiende lo que no es obvio, por lo que para encontrar patrones sobre esto quizás no sea necesario un proyecto de Big Data complejo.
- **Correlación de los datos:** Encontrar correlaciones entre variables no quiere decir que exista causalidad entre ellas. Si no se cuenta con el manejo estadístico apropiado, pueden llegar a cometerse varios errores en este sentido. Esto afecta la capacidad predictiva del Big Data, puesto que, si un elemento no es causante del otro, por más correlación que posean no se podrá crear un modelo válido.

- **Decisiones relegadas a la automatización:** Se establece que una máquina no tiene la capacidad cognitiva de los humanos ni su evaluación ética y moral, por lo que dejar que decidan por sí mismas basadas en algoritmos tiene implicancias técnicas y éticas respecto de la privacidad y singularidad de las personas.

**Conceptos clave**

- Las principales características de Big Data en el modelo de las V son volumen, variedad, velocidad y veracidad.
- Lograr escalabilidad es vital para proyectos de Big Data, tanto para picos altos de consumo de datos y procesamiento como cuando la demanda de recursos disminuye. Es necesario considerar una infraestructura flexible que sea fácilmente adaptable.
- Big data tiene enorme aplicación para resolver problemas, aunque también es criticada por tener falencias técnicas, éticas y de conceptualización práctica. De igual manera se siguen desarrollando nuevas y mejores metodologías que superan constantemente a las formas anteriores de trabajo sobre esta disciplina.

# Referencias

**Las 7 V del Big Data: Características más importantes.** (28 de junio del 2016). *Instituto de Ingeniería del Conocimiento*. Recuperado de https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

**Davenport, T. J.** (2014). *Big Data at Work. Dispelling the Myths, Uncovering the Opportunities*. Recuperado de https://books.google.es/books?hl=es&lr=&id=apjBAgAAQBAJ&oi=fnd&pg=PR5&dq=%22Big+Data+at+Work:+Dispelling+the+Myths,+Uncovering+the+Opportunities%22&ots=fmlfku8uLy&sig=Q8AWJFqdZ3FPRD10x2-9WoN4uoA#v=snippet&q=2014&f=false