UCAB UNIVERSIDAD CATÓLICA ANDRÉS BELLO

# CADENA DE VALOR CON EL ECOSISTEMA HADOOP

# Tabla de contenido

# Objetivo

- Conocer las fases de la cadena de valor de *big data* e instanciarla a una plataforma tecnológica con herramientas del ecosistema Hadoop.

# Introducción

El concepto de cadena de valor surge en 1985, gracias a Michel Porter.

Se trata de una herramienta de gestión que permite analizar las actividades que aportan valor a una empresa, distribuyéndolas en actividades principales o primarias (las dedicadas al desarrollo del producto o servicio que genera valor a la empresa) y actividades secundarias o de soporte (aquellas necesarias para el correcto funcionamiento de la empresa).

El objetivo de la cadena de valor es identificar cuáles son las fuentes de generación de valor de la empresa en el proceso de producción de sus bienes o servicios, pero también se han utilizado como una herramienta de apoyo a la toma de decisiones.

La cadena de valor de datos CVD es un modelo que ve los datos como materia prima y como un recurso importante en el negocio. Esta describe una serie de procesos necesarios para generar valor, paso a paso, e información útil. Sin embargo, con la llegada del *big data* la CVD ha evolucionado para abarcar los elementos necesarios que permitan generar valor en grandes volúmenes de datos, como veremos en este tema.

# Cadena de valor del dato

Uno de los conceptos que está evolucionando más rápidamente en la industria de la información es el de "cadena de valor" y, por ende, el de "ciclo de vida". La idea de la cadena de valor es que desde que una información se crea hasta que se usa pasa por una serie de procesos cuya función, típicamente, es añadir valor para que resulte de mayor utilidad a quien ha de emplearla (Cornella, 1997).

En 1988, R. L. Ackoff especificó por primera vez una jerarquía basada en la filtración, la reducción y la transformación, que mostraba cómo los datos conducen a la información, al conocimiento y finalmente a la sabiduría, representada como la pirámide del conocimiento (Ackoff, 1989).


Fuente: Pirámide del conocimiento. Adaptao de The Valley (s.f.)

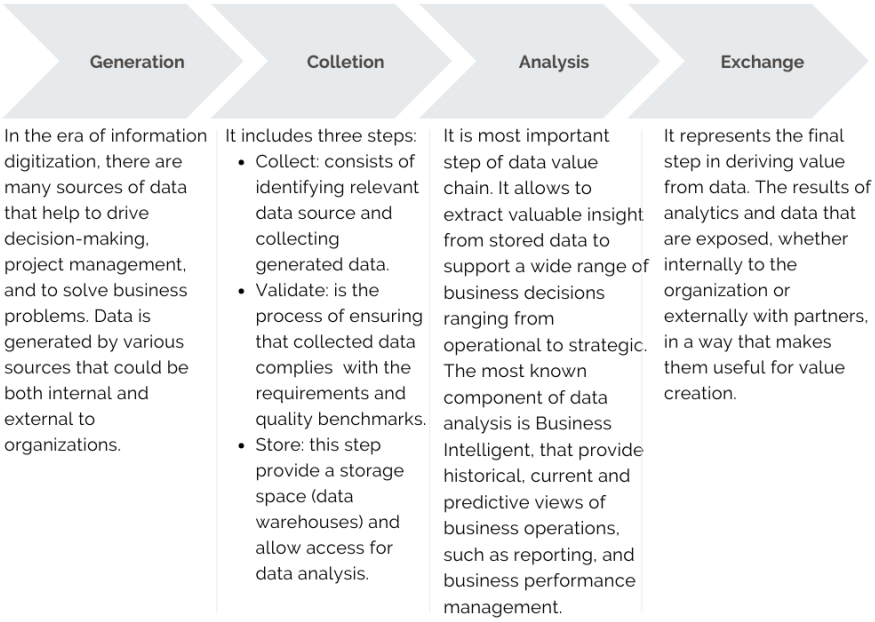
Donde:

- Los datos son hechos o eventos en crudo, sin significado por sí mismo: Ejemplo: rojo, María, 18, Nilo, Mérida, 45.23.
- La información son datos a los que se les ha dado significado a través de una conexión relacional. Puede responder preguntas como: "¿Qué, ¿quién, ¿dónde y cuándo?" Ejemplo: consultas en una base de datos, análisis estadístico.
- El conocimiento es información categorizada, analizada y clasificada. Sirve como marco para la incorporación de experiencias e información; también

puede ser el proceso cognitivo simulado por computadora como la percepción, el aprendizaje y el razonamiento y responde a preguntas del tipo "¿Cómo?".

- La sabiduría es la capacidad de ver las consecuencias a largo plazo de cualquier acto, y evaluarlas en relación con la idea de control total. Es un proceso no determinista y no probabilístico que responde a preguntas como "¿qué se debe hacer y por qué?" (Hey, 2004; Figuerola, 2013).

La cadena de valor del dato describe una serie de procesos necesarios para generar, paso a paso, valor e ideas útiles. Consta de cuatro fases principales, descritas en la imagen anterior, a saber: generación de datos, recopilación de datos, análisis de datos e intercambio de datos.




Fuente: Procesos de la cadena de valor. Adaptado de Faroukhi et al. (2020)

Pero, en síntesis, la cadena de valor es un modelo que se basa en un conjunto de actividades que consideran los datos como una materia prima y como un recurso importante en el negocio. En esta, el flujo de información se describe como una serie de pasos necesarios para generar valor y conocimientos útiles a partir de los datos. Permite el paso de los datos al conocimiento, pasando por varios pasos como el

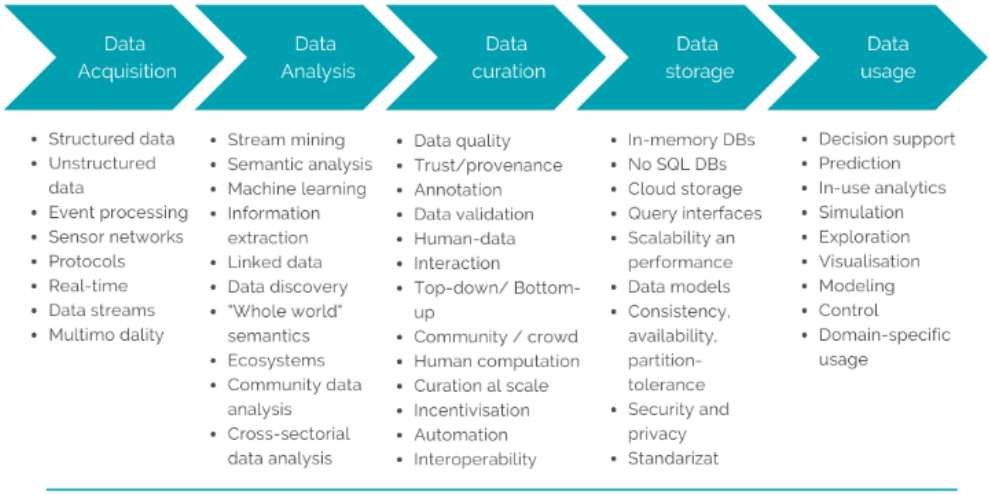
descubrimiento, el procesamiento y la explotación, facilitando así la toma de decisiones (Faroukhi, Alaoui, Gahi y Amine, 2020).

Con la llegada de *big data*, las operaciones se han vuelto cada vez más basadas en datos, enfrentando nuevos desafíos relacionados con el volumen, la variedad y la velocidad, y dando origen a otro tipo de cadena de valor llamada "*Big Data Value Chain*" (BDVC). Las organizaciones se han interesado cada vez más en este tipo de cadena de valor para extraer conocimiento confinado y monetizar sus activos de datos de manera eficiente.

## Cadena de valor de *big data*

En palabras de Rado Kotorov, director ejecutivo de Trendalyze, "…dar el salto de la recopilación y el análisis de datos al uso de datos para marcar una diferencia real en los resultados de la empresa es un desafío para muchas organizaciones" (Dhawal, 2022, s.p.)

La clave es tener una "cadena de valor de datos", que puede permitir a las empresas aprovechar los datos como una nueva fuente de ingresos o para reducir costos. En *big data* en específico, la cadena de valor fue definida en 2014 por Edward Curry para modelar las actividades de alto nivel que componen un sistema de información (Curry, 2015).

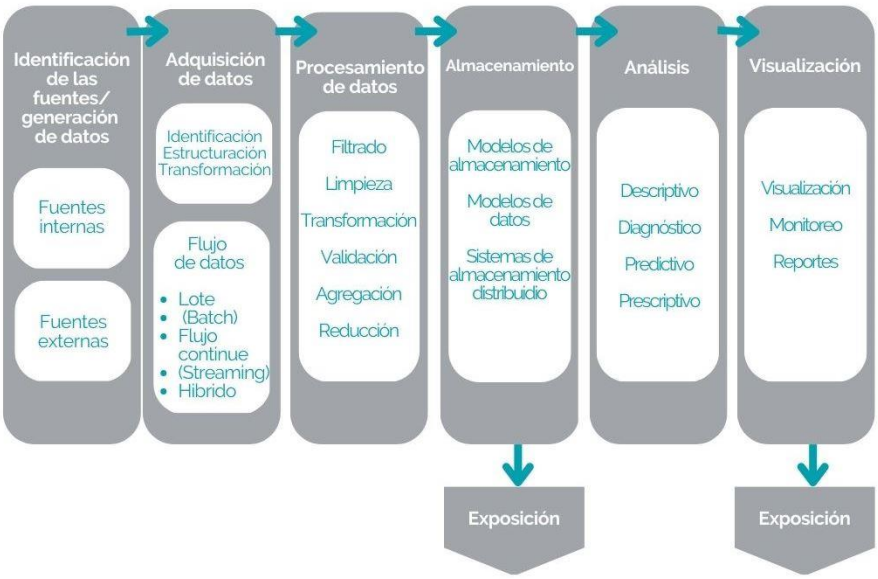



Fuente: Technical Working Groups. Adaptado.

Faroukhi, Alaoui, Gahi y Amine (2020) proponen un marco de trabajo para la monetización en la cadena de valor de *big data*, con el fin de hacer que los procesos

de las organizaciones estén completamente basados en datos, respaldar la toma de decisiones y facilitar la creación conjunta de valor.

Para esta formación vamos a utilizar el siguiente esquema de la cadena de valor en *big data*, basado en la literatura consultada:




Fuente: Cadena de valor en *big data.* Adaptado, basado en Faroukhi, Alaoui, Gahi y Amine (2020), Curry (2015) y Aguilar (2019)

## Fases de la cadena de valor de *big data*

1. **Identificación de las fuentes/generación de datos:** en esta etapa la disponibilidad, cantidad y accesibilidad definen el valor de las fuentes de datos, lo que significa que, si los datos de las fuentes son fácilmente accesibles, tienen un valor más alto. Las actividades en esta etapa abarcan:

   - Identificar las fuentes internas y externas
   - Calcular la cantidad de datos detectada (por ingerir) de cada fuente de datos
   - Identificar los mecanismos de obtención de datos (*push* o *pull*)
   - Determinar el tipo de fuente de datos (generadas por máquinas o por personas, archivos, bases de datos de la empresa, datos web)

– Determinar el tipo de datos: estructurado, no estructurado o semiestructurado.

Ejemplos de fuentes de datos externas son: web y redes sociales, datos provenientes de dispositivos IoT, datos de transacciones online, datos biométricos y datos generados por personas como llamadas o correos.


Fuente: Fuentes de datos externos en *big data*. Extraída de One to Market (s.f.)

Y los datos internos de una organización están constituidos por todos aquellos datos que recoge y pertenecen a ella, y cuyo control está gestionado por ella misma.

2. La adquisición (ingesta) de datos: se refiere a la forma en que se pueden recibir y recopilar los datos y se ha convertido en una etapa de gran interés en el proceso de *big data*.

- Identificación de datos: se refiere a determinar el contenido que se debe considerar.
- Estructuración de los datos: se refiere a la identificación de estructuras de datos preliminares que deben seguirse para adaptarse a la estrategia de gestión de datos.
- Transformación: se refiere al procesamiento previo de los datos para que sean consistentes con los demás datos almacenados. Por ejemplo:

formatos y rangos de fecha, estandarización de uso de mayúsculas, siglas, etc.

- **Flujo de datos:** se refiere a la transferencia de los datos sin procesar recopilados a una infraestructura de almacenamiento de datos específica. La mayoría de las veces un lago de datos.

En esta fase se identifica el modo de flujo de datos y cómo serán tratados al conectarse a plataformas de generación. Este flujo de datos podría ser:

- Carga por lotes (*Batch*)
- Carga de flujo continuo (*Streamng*)
- Microlotes
- Sus combinaciones: Lambda, Kappa.


Fuente: Adquisición de datos. Extraída de One to Market (s.f.)

3. **Preprocesamiento de datos:** los datos recopilados de varias fuentes heterogéneas contienen mucho ruido, redundancia y anomalías. Los métodos analíticos requieren un cierto nivel de calidad de los datos. Esta actividad se ocupa de hacer que los datos adquiridos sean aptos para su uso en la toma de decisiones, así como en el uso específico del dominio. Las tareas son:

- **Filtrado:** eliminación de datos considerados corruptos, de acuerdo con los requisitos de la estrategia de datos de la organización.
- **Transformación:** modificación, adaptación y empaquetado de datos en formas apropiadas a la estandarización de escalado de atributos para mejorar los procesos de análisis de datos.
- **Validación:** establecimiento de reglas de validación y borrado para eliminar datos no válidos y desconocidos.
- **Agregación y reducción:** tratamiento en conjuntos de datos que pertenecen al mismo campo. Esta agregación nos permite tratar con datos voluminosos.
- **Limpieza:** identificación y procesamiento de datos incompletos, inexactos e irrazonables para eliminarlos o completarlos.


Fuente: Procesamiento de datos. Extraída de One to Market (s.f.)

4. **Almacenamiento:** refiere a la persistencia de una gran cantidad de datos recopilados y preprocesados. Las estrategias de los sistemas de almacenamiento tienen un impacto significativo en la escalabilidad y el rendimiento de BDVC en términos de acceso y exposición de datos.

- **Modelos de almacenamiento:** desarrollado principalmente en torno a tres modelos de almacenamiento, definidos por el sistema de archivo:
    - Almacenamiento por bloques
    - Almacenamiento por archivo

- Almacenamiento por objeto.

- **Modelos de datos:** tales como clave-valor, orientado a columnas, orientado a gráficos u orientado a documentos de las bases de datos NoSQL.
- **Sistemas de almacenamiento distribuido:** se verifica el cumplimiento del teorema de CAP, por lo que surgen:
  - Sistemas CA (consistentes y de alta disponibilidad)
  - Sistemas CP (consistentes y tolerantes a particiones)
  - Sistemas AP (altamente disponibles y tolerantes a particiones).


Fuente: Almacenamiento. Extraída de One to Market (s.f.)

5. **Análisis de datos:** manipulación de datos masivos para identificar patrones, encontrar correlaciones y descubrir nuevos modelos de conocimiento emergentes. Los tipos de análisis pueden ser:

- **Análisis descriptivo:** se refiere a la descripción y síntesis de modelos de conocimiento utilizando métodos estadísticos que describen una situación, como informes estándar, cuadros de mando y análisis detallados. Describe

los eventos que ocurrieron en cualquier momento del pasado y proporciona una idea de lo que realmente sucedió.

- **Análisis diagnóstico:** permite a los usuarios comprender lo que está sucediendo y por qué sucedió, para que se pueda tomar una acción correctiva si algo salió mal.
- **Análisis predictivo:** se refiere a las probabilidades de predicción empleadas para definir tendencias futuras. Utiliza modelos de aprendizaje supervisados, no supervisados y semisupervisados, para proporcionar modelos analíticos predictivos.
- **Análisis prescriptivo:** se aplica para predecir eventos futuros e impulsar decisiones proactivas, fuera de los límites de la interacción humana.


Fuente: Análisis. Extraída de One to Market (s.f.)

6. **Visualización de datos:** se refiere a ilustrar las relaciones de los datos con una representación visual artística, como gráficos, mapas, cuadrículas de datos y alertas, que ayudan a la toma de decisiones rápida y eficiente.

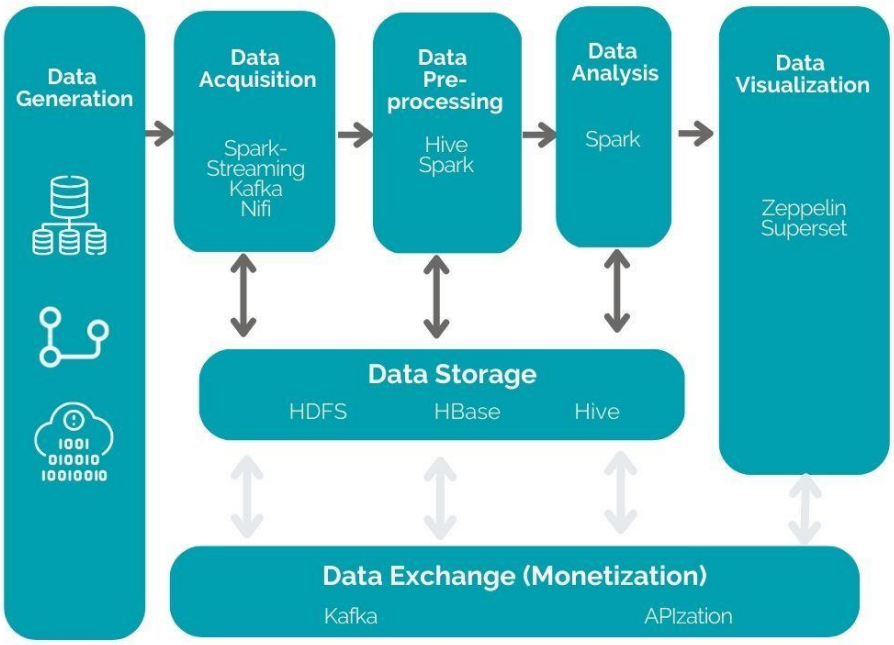
Utiliza herramientas adecuadas con capacidades ampliadas que permiten a los usuarios comerciales encontrar nuevas tendencias o descubrir respuestas a preguntas no formuladas.

7. **Exposición de los datos:** se refiere a la puesta a disposición de los datos para el consumo. Consiste en:
   - Configurar muchas API (interfaces de programación de aplicaciones)

- Respetar las políticas de seguridad y confidencialidad y permitir el acceso a los datos en diferentes estados: analizados, preprocesados, transformados o incluso tan crudos como se recopilan
- Generalmente sirve para muchas aplicaciones internas, como CRM (*Customer Relationship Management*), para promover productos específicos, pero también podría extenderse para servir a los socios.

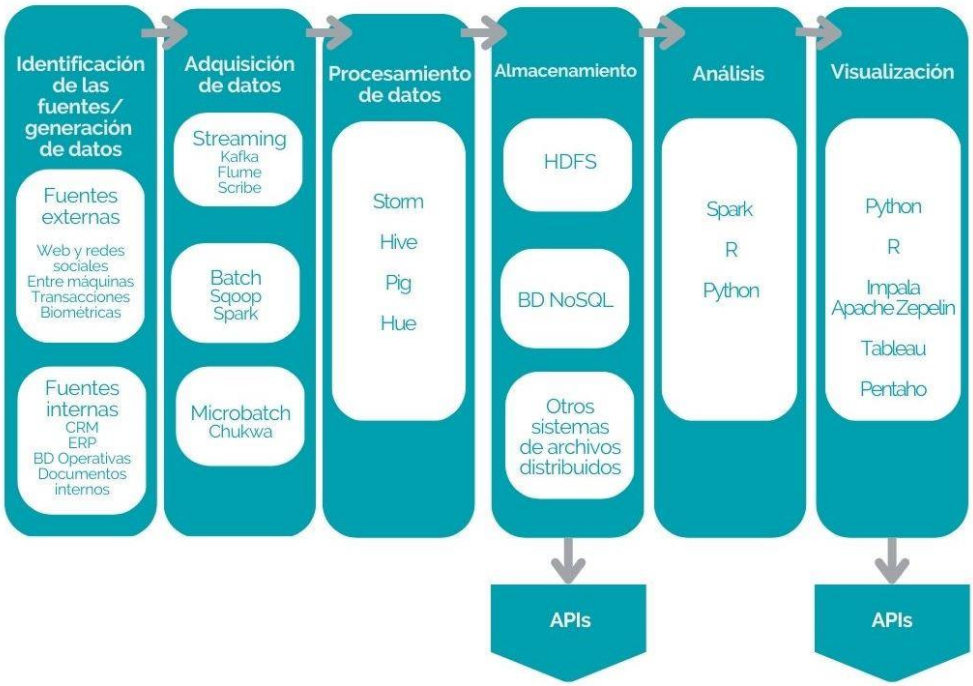
## Tecnologías asociadas a la cadena de valor de *big data*

A continuación, se muestra una plataforma tecnológica que instancia cada una de estas etapas con herramientas del ecosistema Hadoop:




Fuente: Plataforma *big data* para la cadena de valor. Adaptado de Faroukhi, Alaoui, Gahi y Amine (2020)

Estas y otras tecnologías las podemos agrupar en la cadena de valor de *big data,* tal como se muestra en la siguiente imagen:




Fuente: Cadena de valor del dato en una arquitectura de *big data* con herramientas del ecosistema Hadoop. Adaptado de One to Market (s.f.)

# Cierre

La cadena de valor de *big data* consiste en una serie de etapas que permiten analizar grandes volúmenes de información en formatos variados para organizar procesos que contribuyan directa o indirectamente a generar valor agregado para la organización.

El análisis de *big data* nos ha permitido descubrir patrones ocultos y definir nuevas tendencias mediante el análisis y la predicción de comportamientos. Dado que los datos se vuelven más valiosos que nunca, las organizaciones tienen que adaptar todos los procesos que se ocupan de las especificidades de *big data* para hacerlos más rentables

La revisión de los diferentes modelos propuestos nos lleva a trabajar con un marco global y genérico que admite la mayoría de las fases necesarias para obtener el valor de los datos.

Con el presente tema ya eres capaz de identificar, de manera integrada, las diferentes fases de la cadena de valor de *big data* y las actividades que se llevan a cabo, las cuales en su mayoría han sido ampliamente cubiertas en el curso, dando a su vez una guía para llevar adelante proyectos de *big data*.

# Referencias

Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, pp. 3-9.

Aguilar, L. (2019). *Inteligencia de negocios y analítica de datos. Una visión global de Business Intelligence & Analytics.* Alfaomega Grupo Editor.

Cornella, A. (1997). *Ciclo de vida y cadena de valor en información*. El profesional de la Información. http://profesionaldelainformacion.com/contenidos/1997/enero/ciclo_de_vida_y_cadena_de_valor_en_informacin.html

Curry, E. (2015). The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. En *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe.* Heidelberg: Springer.

Dhawal, K (2022). *Data Value Chain: Analysis-Enable Data Products*. Tdan. https://tdan.com/data-value-chain-analysis-enable-data-products/29012

Faroukhi, Z., Alaoui, I. E., Gahi, Y. y Amine, A. (2020). An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization. *Big Data and Cognitive Computing, 4*(4), p. 34.

Faroukhi, Z., Alaoui, I. E., Gahi, Y. y Amine, A. (2020). Big data monetization throughout Big Data Value Chain: a comprehensive review. Journal of Big Data, https://doi.org/10.1186/s40537-019-0281-5

Figuerola, N. (2013). *Gestión del Conocimiento (Knowledge Management).* Artículos pm. https://articulospm.files.wordpress.com/2013/08/gestic3b3n-de-conocimiento-dikw.pdf

Hey, J. (2004). *The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link.* Jonohey. https://www.jonohey.com/files/DIKW-chain-Hey-2004.pdf

## Referencia de las imágenes

Faroukhi, Z., Alaoui, I. E., Gahi, Y. y Amine, A. (2020). Plataforma *big data* para la cadena valor [Imagen]. Disponible en: An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization. *Big Data and Cognitive Computing, 4*(4), p. 34.

Faroukhi, Z., Alaoui, I. E., Gahi, Y. y Amine, A. (2020). Procesos de la cadena de valor [Imagen]. Disponible en: An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization. *Big Data and Cognitive Computing, 4*(4), p. 34.

One to Market (s.f.). Adquisición de datos [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

One to Market (s.f.). Almacenamiento [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

One to Market (s.f.). Análisis [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

One to Market (s.f.). Cadena de valor del dato en una arquitectura de *big data* con herramientas del ecosistema Hadoop [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

One to Market (s.f.). Fuentes de datos externos en *big data* [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

One to Market (s.f.). Procesamiento de datos [Imagen]. Disponible en: https://onetomarket.es/wp-content/uploads/2022/02/tipos-de-fuentes-de-datos-big-data.png

The Valley (s.f.). Pirámide del conocimiento [Imagen]. Disponible en: https://thevalley.es/wp-content/uploads/2017/03/pyramid.jpg