**Qué es Minería de Datos y Big Data**

Miguel Molina.

Agosto 2020.

Fundación Universitaria Unipanamericana.

Ingeniería de Sistemas.

Extracción, Transformación y Carga de Datos.

**Tabla de Contenidos**

[Qué es Mineria de Datos **¡Error! Marcador no definido.**](#_Toc410628920)

[Qué es Big Data 5](#_Toc410628925)

[Cuadro Comparativo. 12](#_Toc410628929)

[List of References 14](#_Toc410628930)

**Minería de Datos**

## Significado y características

La **minería de datos (*data mining*),** es un proceso para detectar información de conjuntos grandes de datos, de la manera más automáticamente posible. Su objetivo es encontrar patrones, tendencias o reglas que expliquen el comportamiento de los datos en un contexto específico.

La característica que distingue a la minería de datos es que puedes obtener información sin tener que hacer preguntas específicas, ya que su tarea es resolver este ámbito:

Tiene como rol ser predictiva, en relación a la posibilidad de decir qué es lo que puede pasar utilizando estadísticas y probabilidades de información que está oculta en datos almacenados.

También funciona para conocer, además de eventos futuros, datos importantes que conviene saber, pero que no están a la vista. Por ejemplo, una relación inusual entre las ventas de dos productos diferentes y cómo esa relación varía de acuerdo a cómo se posicionan estos productos en los locales de ventas al por menor.

## Beneficios de la minería de datos

La minería de datos brinda innumerables beneficios. Algunos de ellos son:

* Se puede aplicar a un montón de escenarios empresariales diferentes.
* Contribuye a gestionar y organizar la información de tu empresa de forma más eficiente.
* Colabora en el ahorro de costos y tiempo en los procesos de tu organización.
* Te ayuda a prevenir situaciones adversas futuras basándose en datos verídicos.
* Al mostrar información clave, contribuye en la toma de decisiones estratégicas.
* Facilita el conocimiento de los usuarios de tu empresa, sus gustos, preferencias, comportamientos.
* Te ayuda a optimizar tus productos o servicios en base a la información que te brinda sobre el comportamiento online de tus clientes. Esto a su vez te lleva a aumentar tus ventas.
* Brinda la posibilidad de desarrollar estrategias para encontrar y atraer nuevos clientes.
* Te posibilita mejorar la atención al cliente de tu empresa a partir de la información obtenida.

## El proceso de la minería de datos

Un proceso común de Minería de Datos suele tener cuatro partes principales:

1. **Determinación de los objetivos:** A partir de las necesidades del cliente y con orientación del especialista en data mining.
2. **Pre-procesamiento de datos:** Es la selección y transformación de las bases de datos.
3. **Determinación del modelo:** Comienza con los análisis estadísticos de los datos, y después se implementa un primer boceto del diseño de la visualización. En esta etapa pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. **Análisis de los resultados:** Se verifica la coherencia de los resultados y se los compara con los obtenidos por estadística y visualización gráfica. También aquí, el cliente opina si se trata de resultados novedosos, útiles y si sirve a sus objetivos iniciales.

La minería de datos puede ser útil en tipos específicos de misiones comerciales. Por ejemplo: determinando la efectividad de un programa de Marketing, manejo del ciclo de vida de un cliente, mostrando escenarios posibles en un proceso de negocio y procesando datos para luego combinarlos en partes para mejor visualización.

**Big Data**

**¿Qué es Big Data y para qué sirve?**

En definitiva, cuando hablamos de Big Data no nos referimos únicamente a los datos, sino sobre todo a **la capacidad de poderlos explotar para extraer información y conocimiento de valor para nuestro negocio**. La finalidad del Big Data es poder diseñar nuevos productos y servicios basándose en los nuevos insights que adquirimos sobre nuestros clientes, sobre nuestra competencia o el mercado en general.

Una vez recogida y almacenada la información, se deben extraer indicadores que puedan ser útiles para tomar decisiones, incluso en tiempo real. Por lo tanto, **la verdad sobre qué es Big Data y para qué sirve va mucho más allá de solo pensar en «datos masivos».**

## Las cinco “Vs” del Big Data

**La primera pregunta que nos viene a la mente al considerar qué es Big Data y para qué sirve**, está relacionada con cuánto de “grande” tienen que ser los datos para ser considerados “Big”. Finalmente, el enfoque correcto está en **no establecer un tamaño en absoluto, sino relativo**. Lo que ahora puede parecernos un gran tamaño de datos, en dos o tres años puede ser normal o hasta irrelevante. La mayoría de expertos **definen el Big Data en términos de las cinco “Vs”:**

* **Volumen**: como hemos visto, la cantidad de datos se define “Big” no cuando supera un tamaño definido, sino cuando su almacenamiento, procesamiento y explotación empieza a ser un reto para una organización.
* **Velocidad**: la segunda característica del Big Data está relacionada con el ritmo a los cuales los datos se están generando, que suele aumentar constantemente y que necesita una respuesta en tiempo real por parte de las empresas.
* **Variedad**: sin embargo, el reto principal del Big Data reside en la gran diferencia de formatos distintos en los cuales encontramos los datos y que pueden ir desde texto sencillo, a imágenes, videos, hojas de cálculos y enteras bases de datos.
* **Veracidad**: además, los datos tienen que ser confiables y han que ser mantenidos limpios. Una gran cantidad de datos no tiene valor si son incorrectos y puede ser altamente perjudicial, sobre todo en la toma de decisión automatizada.
* **Valor**: finalmente, los datos y su análisis tienen que generar un beneficio para las empresas.

## Tipos de Big Data

* **Web y Redes Sociales**: información disponible en Internet como contenido Web, generada por los usuarios en su actividad en las redes sociales o información de búsquedas en buscadores.
* **Machine-to-Machine (M2M)**: datos generados a partir de la comunicación entre sensores inteligentes integrados en objetos de uso cotidiano.
* **Transacciones**: incluye registros de facturación, llamadas o transacciones entre cuentas.
* **Biométricos**: datos generados por tecnología de identificación de personas mediante reconocimiento facial, de huellas dactilares o mediante información genética.
* **Generados por personas**: a través de correos electrónicos, servicios de mensajería o grabaciones de llamadas.
* **Generados por organizaciones** tanto públicas como privadas: datos relacionados con el medioambiente, estadísticas gubernamentales sobre población y economía, historiales clínicos electrónicos, etc.

Por otro lado, según su estructura, los datos pueden ser:

* **Estructurados**: datos que tienen definidos su formato, tamaño y longitud, como las bases de datos relacionales o Data Warehouse.
* **Semiestructurados**: datos almacenados según una cierta estructura flexible y con metadatos definidos, como XML y HTML, JSON, y las hojas de cálculo (CSV, Excel).
* **No estructurados**: datos sin formato específico, como ficheros de texto (Word, PDF, correos electrónicos) o contenido multimedia (audio, vídeo, o imágenes).

**Herramientas en el Mercado**

**Minería de Datos:**

* **Clementine / SPSS:** Herramienta de data mining que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales, de manera que no es preciso ser un experto en data mining.
* **SAS Enterprise Miner / SAS:** Solución de minería de datos que proporciona gran cantidad de modelos y de alternativas. Permite determinar pautas y tendencias, explica resultados conocidos e identifica factores que permiten asegurar efectos deseados. Además, compara los resultados de las distintas técnicas de modelado, tanto en términos estadísticos como de negocio, dentro de un marco sencillo y fácil de interpretar.
* **Keel:** Es un software para evaluar la evolución de los algoritmos de minería de datos y problemas de regresión, entre ellos: clasificación, agrupamiento y patrón de la minería. Contiene una gran colección de algoritmos clásicos de extracción de conocimientos, técnicas de pre procesamiento y modelos híbridos como sistemas difusos genéticos, redes neuronales evolutivas, etc. Nos permite realizar un análisis completo de cualquier modelo de aprendizaje en comparación con los existentes, incluido un módulo de prueba estadística para la comparación entre ellos.
* **SAS Analytics / SAS:** Suite de soluciones analíticas que permiten transformar todos los datos de la organización en conocimiento, reduciendo la incertidumbre, realizando predicciones fiables y optimizando el desempeño.
* **RapidMiner / Yale:** Es el líder mundial de código abierto para la minería de datos debido a su combinación de su tecnología de primera calidad y su rango de funcionalidad. Esta aplicación de RapidMiner cubre un amplio rango de minería de datos. Además de ser una herramienta flexible para aprender y explorar la minería de datos, la interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área.
* **Microsoft SQL Server 2005 / Microsoft:** Solución que ofrece un entorno integrado para crear modelos de minería de datos y trabajar con ellos. La solución SQL Server Data Mining permite el acceso a la información necesaria para tomar decisiones inteligentes sobre problemas empresariales complejos.
* **Dlife / Apara:** Plataforma bioinformática para la toma de decisiones clínicas. Proporciona un soporte computacional a la toma de decisiones médicas en los procesos de diagnóstico, tratamiento y seguimiento de la evolución de los pacientes que permite a los profesionales clínicos incrementar su precisión y la eficiencia de la prestación sanitaria en una media del 20%.
* **KNIME:** Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold.

**Big Data:**

* [**Hadoop**](http://hadoop.apache.org/): es una herramienta de código abierto que nos permite tanto gestionar los grandes volúmenes de datos, como analizarlos y procesarlos. Hadoop implementa **MapReduce**, un modelo de programación que da soporte a la computación paralela sobre grandes colecciones de datos.
* **NoSQL**: se trata de sistemas que no utilizan SQL como lenguaje de consultas, lo que, a pesar de no poder garantizar la integridad de los datos (principios ACID: atomicidad, consistencia, integridad y durabilidad), les permite obtener ganancias significativas en escalabilidad y rendimiento a la hora de trabajar con Big Data. Una de las bases de datos NoSQL más populares es [MongoDB](https://www.mongodb.com/).
* [**Spark**](https://blog.mdcloud.es/del-dato-a-los-insights-donde-queda-hadoop-spark/): es un framework de computación en clúster de código abierto que permite procesar los datos de forma rápida. Permite escribir aplicaciones en Java, Scala, Python, R y SQL y funciona tanto sobre Hadoop, Apache Mesos, Kubernetes, como de forma independiente o en la nube. Puede acceder a centenares de fuentes de datos.
* [**Storm**](http://storm.apache.org/): es un sistema de computación en tiempo real distribuido de código libre. Storm permite procesar flujos ilimitados de datos en tiempo real de manera sencilla, pudiendo usarse con cualquier lenguaje de programación.
* [**Hive**](https://hive.apache.org/): es una infraestructura de [Data Warehouse](https://blog.mdcloud.es/dwh-ejemplos-arquitectura-data-warehouse/) construida sobre Hadoop. Facilita la lectura, escritura y administración de grandes conjuntos de datos que residen en almacenamiento distribuido mediante SQL.
* [**R**](https://www.r-project.org/): es uno de los lenguajes de programación más utilizados en análisis estadísticos y en minería de datos. Puede integrarse con distintas bases de datos y permite generar gráficos con alta calidad.
* [**D3.js**](https://d3js.org/): es una biblioteca de JavaScript para producir visualizaciones dinámicas e interactivas de datos en navegadores web, usando HTML, SVG y CSS.

**Cuadro Comparativo**

|  |  |  |
| --- | --- | --- |
|  | Big Data | Minería de Datos |
| ¿Qué es? | Solución de análisis de grandes cantidades de estructurados o semi estructurados. | Es el proceso de exploración y análisis de manera automática o semiautomática de los datos para obtener patrones significativos y reglas de negocio. |
| ¿Para qué sirve? | Ayuda a descubrir patrones, tendencias que plantean preguntas que no conocemos de antemano. | Obtener información sin tener que hacer preguntas específicas, ya que su tarea es resolver este ámbito. |
| Características | * Se refiere a una gran cantidad de datos. * Se basa en variedad, volumen, asi como almacenamiento y procesamiento de datos. * Depende de la minería de datos para darle utilidad a la información. * Es la información que se almacena. | * No necesita una gran cantidad de datos para usarse. * Realiza un análisis de los patrones y las relaciones de los datos presentados. * No se requiere de Big Data, ya que puede manejar una pequeña cantidad de datos. * Es el administrador de la información, es decir, el encargado de otorgarle un uso |

**¿Qué puede ayudar la minería de datos y big data en la ETL?**

Dentro de la minería de datos, su primera fase consiste en extraer los datos de una o varias fuentes transformarlos en un formato estandarizado y cargarlos en el data warehouse. El software analiza y ordena los datos en función de las consultas del usuario final. Los datos se presentan al usuario final en un formato comprensible, por ejemplo, un informe o una gráfica.

La transformación y carga de datos se da debido a la utilización de información que es suministrada por varias fuentes, el Big data al recolectar masivas cantidades de información estás serán procesadas, al hacer la carga de datos el Big data será fundamental debido a la cantidad de datos se podrá cargar todo en niveles jerárquicos.

# Lista de referencias

A, Garassino. (2015). ¿Para qué sirve la minería de datos o data mining?.URL: <https://www.syloper.com/blog/recursos/para-que-sirve-la-mineria-de-datos/>

E, Carisio. (2018). ¿Qué es Big Data y para qué sirve? Ejemplos de uso. URL: <https://blog.mdcloud.es/que-es-big-data-y-para-que-sirve/>

ChannelBeatz Redaction. (2015). ETL el gran coste del Big Data. URL: <https://www.channelbiz.es/2020/08/05/competencia-autoriza-la-compra-de-gti-por-esprinet/>

J, Lopez. (2020). Minería de Datos. URL:<https://economipedia.com/definiciones/mineria-dedatos.html#:~:text=Minería%20de%20datos.%20La%20minería%20de%20datos%20es,de%20comportamiento%20dentro%20de%20la%20base%20de%20datos>

Bolger, A., & Giorgi, F. Trimmomatic: A Flexible Read Trimming Tool for Illumina NGS Data. URL http://www. usadellab. org/cms/index. php.