Big data, machine learning y data science en python

Big data, machine learning y data science en python

José Manuel Ortega

Ra-Ma®

La ley prohíbe
fotocopiar este libro

Big data, machine learning y data science en python
Materia: GPH - Ciencia y análisis de datos

Editado por:
RA-MA Editorial
Calle Jarama, 3A, Polígono Industrial Igarsa
28860 PARACUELLOS DE JARAMA, Madrid
Teléfono: 91 658 42 80
Fax: 91 662 81 39
Correo electrónico: *editorial@ra-ma.com*
Internet: *www.ra-ma.es* y *www.ra-ma.com*
ISBN impreso: 978-84-1944-458-5
ISBN ePub: 978-84-19444-59-2
Depósito legal: M-319-2023
Maquetación: Antonio García Tomé
Diseño de portada: Antonio García Tomé
Filmación e impresión: Safekat
Impreso en España en enero de 2023

A mi familia.

ÍNDICE

OBJETIVOS

El libro está dirigido aquellos lectores que estén trabajando en proyecto relacionados con big data y busquen identificar las características de una solución de Big Data, los datos asociados a estas soluciones, la infraestructura requerida, y las técnicas de procesamiento de esos datos. Entre los principales **objetivos** podemos destacar:

- Introducir los conceptos de ciencias de datos y machine learning.
- Introducir las principales librerías que podemos encontrar en Python para aplicar técnicas de machine learning a los datos.
- Dar a conocer los pasos para construir un modelo de machine learning, desde la adquisición de datos, pasando por la generación de funciones, hasta la selección de modelos.
- Dar a conocer los principales algoritmos para resolver problemas de machine learning.
- Introducir scikit-learn como herramienta para resolver problemas de machine learning.
- Introducir pyspark como herramienta para aplicar técnicas de big data y map-reduce.
- Introducir los sistemas de recomendación basados en contenidos.

El libro trata de seguir un enfoque teórico-práctico con el objetivo de afianzar los conocimientos mediante la creación y ejecución de scripts desde la consola de Python. Además, se provee un repositorio donde se pueden encontrar los ejemplos que se analizan a lo largo del libro para facilitar al lector las pruebas y asimilación de los contenidos teóricos.

1

INTRODUCCIÓN A BIG DATA

1.1 INTRODUCCIÓN

En el presente capítulo se va a detallar las diferentes arquitecturas utilizadas en un ecosistema Big Data y las capas más importantes como seguridad, gestión, generación, adquisición, almacenamiento y análisis, los actores, tecnologías y herramientas que forman parte de la arquitectura. Estas tecnologías y herramientas son implementadas según las características del proyecto o tipo de investigación a realizar, es por eso que se van a definir los componentes funcionales de una arquitectura Big Data, resaltando en qué casos suelen ser más útiles o cómo en combinación con otras pueden aportar mejores resultados.

Si hablamos de Big Data, esta no es una sola tecnología, sino una combinación de viejas y nuevas tecnologías que se integran para poder abordar las nuevas características de los datos como velocidad, variedad y volumen. Por lo tanto, Big Data es la capacidad de manejar un gran volumen de datos de diversas fuentes, a la velocidad correcta, y dentro del marco de tiempo adecuado para permitir el análisis ya sea posterior a la recolección de los datos o en tiempo real. Big Data está típicamente dividido en tres características que son las 5Vs.

El volumen que es la cantidad de datos, la velocidad que hace referencia la tasa de flujo de los datos en la creación, almacenamiento, análisis y visualización, y variedad que son las distintas fuentes de datos. Aunque se tiende a simplificar Big Data en 5Vs existen propuestas que hacen referencia a otras como la variabilidad que se refiere a cualquier cambio de los datos en el tiempo como puede ser la tasa de transferencia o el formato, la veracidad la cual indica la exactitud o precisión de los datos.

Por lo que no debe entenderse la definición de Big Data limitada a solo 5Vs; por ejemplo, puede darse el caso de una cantidad relativamente pequeña de datos muy diversos y complejos o es posible que se procese un gran volumen de datos muy simples. Esos datos simples pueden ser estructurados, semiestructurados o no estructurados. Es por eso que se suele incluir la V de valor que hace referencia al aporte de valor a la organización de parte del análisis de los datos a través del procesamiento Big Data. Esto nos indica, por ejemplo, cuán precisos son los datos elegidos para predecir el valor del negocio o si en realidad tiene sentido los resultados del análisis de Big Data.

1.2 DEFINICIÓN DE BIG DATA

Big Data o datos a gran escala hace referencia a un conjunto de datos tan grande que las aplicaciones informáticas tradicionales de procesamiento de datos no son capaces de tratar con ellos ni de encontrar patrones repetitivos. Se encuentra dentro del sector de las tecnologías de la información y la comunicación (TIC) y se ocupa de la manipulación y procesamiento de grandes volúmenes de datos.

Big Data es la agrupación de múltiples tendencias tecnológicas, maduradas a partir del año 2000. Dichas tecnologías se han consolidado entre los últimos años, momento en el que la sociedad se encuentra generando información alrededor de las redes sociales, un mayor ancho de banda, reducción de los costes de conexión a internet, telefonía móvil, internet de las cosas y computación en la nube.

La popularización de Big Data ha venido explicada inicialmente por 3 Vs: el procesamiento de grandes **volúmenes** de datos que llegan a grandes **velocidades** y con una **variedad** de fuentes de información nunca vista hasta ahora. En el modelo en V de Big Data se proponen 5 grupos de procesos:

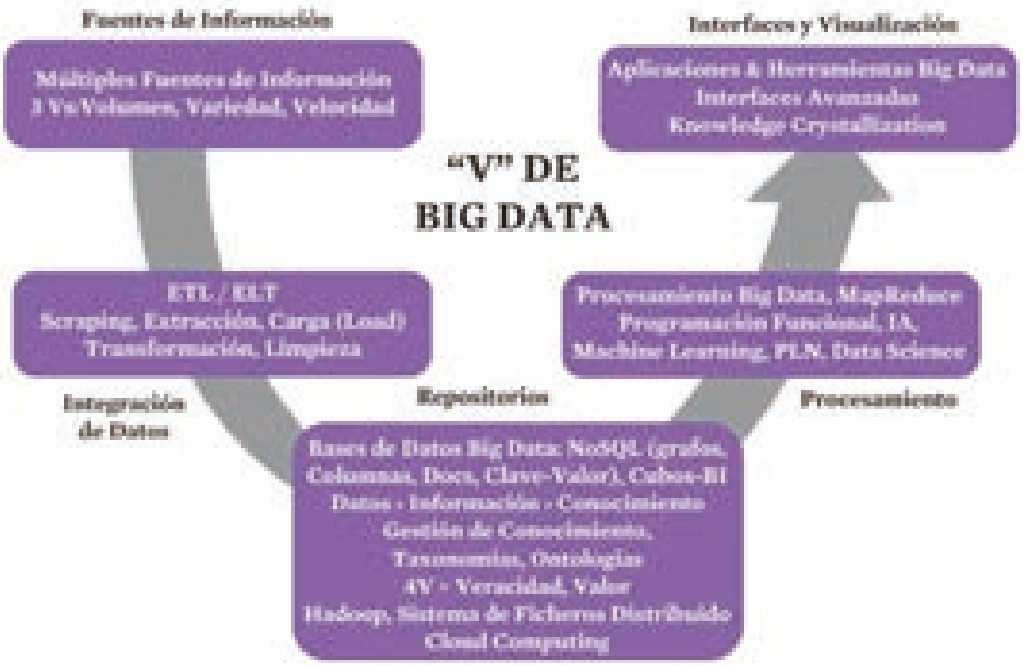

Figura 1.1. Modelo de proceso en Big Data

- **Fuentes de Información Big Data:** enriquecemos nuestras fuentes de datos con nuevas fuentes disponibles de forma abierta en internet. Toda esta variedad de fuentes de información genera grandes volúmenes de datos que llegan a gran velocidad. Las taxonomías que clasifican esas fuentes son relevantes.
- **Integración de datos Big Data**: extraemos los datos y los cargamos en Repositorios de Información especialmente diseñados para tratar Big Data. Frente a la posibilidad de transformar y limpiar los datos antes de cargarlos la tendencia es cargar todos los datos para poder explotarlos a posteriori para otros fines. Cobra asimismo importancia el proceso de Scraping de información, de lectura de datos directamente de la web mediante aplicaciones software que llamamos Bots.
- **Sistema y Repositorios Big Data**: nuevos tipos de Bases de Datos, que llamamos NoSQL son los nuevos contenedores de información, especialmente preparados para los tipos de procesamiento necesarios. Además de datos e información gestionamos el

conocimiento en Ontologías, que son reflejo de una 4a V, la Veracidad. El Sistema de Ficheros Distribuido y el Cloud Computing son la base de este Sistema Big Data.

- **Procesamiento Big Data:** tecnologías tradicionales como la programación funcional, el machine learning, el procesamiento de lenguaje natural, y un grupo de áreas de conocimiento que agrupamos bajo los paraguas de la "Data Science" y la Inteligencia Artificial se aprovechan de nuevas capacidades de procesamiento distribuido y masivo de datos para ser el 4o eslabón de la "V" de Big Data. En torno a este grupo de procesos aparece para algunas empresas una 5a "V", la Viscosidad, referenciando con ese concepto la mayor o menor facilidad para correlacionar los datos.
- **Interfaces y Visualización Big Data:** los usuarios necesitan nuevos sistemas de visualización, interacción y análisis para interactuar con el Big Data, diferentes a los tradicionales provenientes del mundo del Business Intelligence. Aparecen situaciones en las que, por ejemplo, una misma pregunta cristaliza en diferentes respuestas para diferentes usuarios según su contexto.

La consultora Gartner lo describe como "Big Data son los grandes conjuntos de datos que tiene tres características principales: **volumen** (cantidad), **velocidad** (velocidad de creación y utilización) y **variedad** (tipos de fuentes de datos no estructurados, tales como interacción social, video, audio, cualquier cosa que se pueda clasificar en una base de datos)".

El ritmo actual de generación de datos está sobrepasando las capacidades de procesamiento de los sistemas actuales en compañías y organismos públicos. Las redes sociales, el Internet de las Cosas y la industria 4.0 son algunos de los nuevos escenarios con presencia de datos masivos.

La necesidad de procesar y extraer conocimiento valioso de tal inmensidad de datos se ha convertido en un desafío considerable para científicos de datos y expertos en la materia. El valor del conocimiento extraído es uno de los aspectos esenciales del Big Data.

Con el objetivo de cubrir la problemática existente del almacenamiento, tratamiento y aprovechamiento de los grandes volúmenes de datos que se producen en la actualidad por factores como son: la elevada y creciente cantidad de fuentes de datos (sensores y redes sociales, por ejemplo) y la generalización de las redes de telecomunicaciones, en muchos casos inalámbricas. El conjunto de estos elementos, junto con las mayores capacidades de almacenamiento, ha hecho crecer de una manera enorme la cantidad de datos disponibles en los últimos años, tendencia que se sigue manteniendo en la actualidad.

Otra posible definición es la que describe Big Data a través de tres características:

- **Volumen**: gran cantidad de datos.
- **Velocidad**: procesamiento cercano a tiempo real.
- **Variedad**: distintas fuentes de información y formato.

La primera de las características más importantes de este concepto hace referencia a la circunstancia de que la cantidad de datos que se manejan supera actualmente el desproporcionado rango de los Exabytes de información. Obviamente, toda esta gran cantidad de datos puede obtenerse de diversas fuentes o ser presentados en infinidad de formas (variedad).

El volumen se incrementa en órdenes de magnitud no vistos anteriormente en los almacenes de información tradicionales, hablándose incluso de Zetabytes. Por otra parte los datos empiezan a llegar a los sistemas en tiempo real (Velocity) y hay que ser capaz de tratar esa información para que no se pierda nada.

Por último, empiezan a llegar fuentes de datos eminentemente desestructuradas (básicamente texto procedente de Internet) que siguen conviviendo con las fuentes estructuradas clásicas, aquí

estamos hablando de variedad (Variety) en las fuentes de información que será necesario integrar para tener una visión global de cada escenario.

Todas las aplicaciones que hacen uso de estos datos necesitan obtener unos tiempos de respuesta mínimos que permitan lograr la obtención de la información correcta en el momento preciso. Esta información debe ser lo más veraz posible; es decir, las fuentes de las cuáles se obtiene deben ser lo más fiable posible para así poder generar el valor tan ansiado que haga que nuestros datos sirvan para un fin concreto, como puede ser la toma de decisiones críticas en organizaciones o la comprobación de la evolución del tráfico en un portal de Internet, por ejemplo.

Debido a esto, en el mundo en el que nos encontramos es necesario determinar qué información queremos obtener, para que el volumen de los datos no nos desborde. Para tal fin, se utilizarán un conjunto de herramientas que permitan el almacenamiento, procesamiento, recuperación y análisis de una cantidad inmensa de datos.

Big Data se suele definir como "conjunto de técnicas que permiten analizar, procesar y gestionar conjuntos de datos extremadamente grandes que pueden ser analizados informáticamente para revelar patrones, tendencias y asociaciones". Además, el volumen no tiene definido un tamaño mínimo que divida, lo que es Big Data y lo que no. Según un estudio, no existe una cantidad de datos específica, aunque afirma que usualmente se habla en términos de petabytes y exabytes de datos.

- **Gigabyte**: equivale aproximadamente a 256 canciones si el tamaño promedio de cada canción son 4 MB.
- **Terabyte**: : cantidad equivalente a 4 portátiles de 256 GB, teniendo en cuenta que el S.O. ocupa parte de ese espacio.
- **Petabyte**: todas las fotos que posee Facebook equivalen a 1.5 PB.
- **Exabyte**: Empresas como Google, Amazon o Facebook suelen manejar tales cantidades de datos.

La capacidad de cómputo del hardware y el software crece exponencialmente. Hoy en día tenemos en nuestro bolsillo, concretamente en nuestros modernos teléfonos móviles, más capacidad de cómputo que los ordenadores de la NASA que llevaron al hombre a la luna. Los ordenadores personales de los que disponíamos a finales de los años 90 son hoy tristes antiguallas, apenas útiles más que en exposiciones de juegos retro.

En los últimos años han evolucionado tanto las técnicas como las nuevas capacidades del hardware y del software que nos hacen posible usar ahora paradigmas informáticos de altas capacidades que hasta hace pocos años eran computacionalmente inviables.

Estas nuevas tecnologías pueden habilitar nuevas capacidades para las organizaciones fundamentadas en el término paraguas Big Data, materializadas en servicios, funciones u operaciones nuevas o muy mejoradas. La implementación de estas nuevas capacidades puede conseguir como resultado importantes beneficios.

Big Data como paradigma también nos ha aportado Sistemas de Archivos Distribuidos y escalables y nuevos sistemas de gestión de bases de datos preparados para dar respuesta a la necesidad de manejar grandes volúmenes de información de forma distribuida. Ejemplos hoy de rabiosa actualidad son las **Bases de Datos NoSQL**, entre las que destacan las orientadas a columnas, las de clave-valor, las orientadas a la gestión de documentos, objetos o grafos.

Los otros enfoques emergentes son los del Aprendizaje Automático, popularmente conocido por su denominación en inglés, "**Machine Learning**", y los Métodos Probabilísticos

y Estadísticos. Estos dos enfoques, aplicados tanto a textos desestructurados como a datos masivos, proporcionan resultados novedosos aplicados a los procesos analíticos, prospectivos y predictivos.

En **Machine Learning** utilizamos conjuntos de información y un algoritmo para entrenar a una aplicación. Una vez entrenada, cada vez que necesitemos analizar una nueva información dicha aplicación clasificará la nueva información a partir del entrenamiento recibido. En el algoritmo de entrenamiento podemos estar utilizando tanto los métodos probabilísticos y estadísticos mencionados anteriormente como otras técnicas de inteligencia artificial como redes neuronales, árboles de decisión, etc.

Los métodos probabilísticos y estadísticos nos van a ofrecer un modelo de referencia para un conjunto de datos, gracias al cual podamos clasificar una nueva información ofreciendo una predicción a partir de dicho modelo. Estos modelos se aplican tanto a datos numéricos como a conjuntos de palabras dentro de documentos. Son aplicados actualmente, por ejemplo, por los grandes buscadores de Internet para determinar qué documentos son más relevantes para una búsqueda dada.

Para agrupar todo este conocimiento que se está concentrando en torno al término de Big Data ha emergido el concepto de Data Science. Las implementaciones Big Data serían imposibles sin las nuevas capacidades de los ordenadores actuales, que han evolucionado enormemente tanto en el hardware como en el software. Además de la capacidad de procesamiento, el almacenamiento es el otro punto en el que el hardware ha evolucionado: el coste de un dispositivo de 1Gb de capacidad ha disminuido de 300.000 € en 1980, a unos 10 € en el año 2000 y apenas unos céntimos en la actualidad.

En cuanto al software las claves están en la evolución y mejora de los sistemas operativos y en la virtualización, encarnada en las máquinas virtuales, un software capaz de emular a una computadora, pudiendo ejecutarse en un mismo ordenador varias máquinas virtuales. Ambas evoluciones, de hardware y software, han habilitado una paralelización potente y fiable, haciendo posible poner a funcionar en paralelo cientos o miles de estos ordenadores que, aplicando el viejo lema de Julio César "divide et vinces", divide y vencerás, separamos los problemas en multitud de pequeños problemas fáciles de solucionar y luego integran todas esas pequeñas soluciones en la solución final del problema planteado, todo ello realizado en un intervalo de tiempo pequeño. A este tipo de sistemas lo llamamos **sistemas distribuidos**.

Gracias a todo esto se ha habilitado la posibilidad de que en grandes centros de datos se implementen todas estas nuevas capacidades de cómputo y se le ofrezcan nuevos servicios al mercado. A este otro paradigma lo llamamos "**Cloud Computing**", computación remota, en definitiva.

Por último, la aparición de proyectos de software libre, entre los que destaca el **Apache Hadoop**, ha hecho posible esta revolución. Las grandes empresas de internet han promovido un uso masivo de software libre principalmente por su capacidad de adaptación rápida a sus nuevas necesidades, pero también hay que mencionar que el reducido o inexistente coste de licencias del mismo ha posibilitado la viabilidad económica de estas empresas.

Big Data contempla las nuevas herramientas, tecnologías y (nuevos) los conceptos relacionados con la adquisición de (mucha) data (volumen), de distinto tipo (variedad) que a su vez podría estar no estructurada, con unos aspectos opcionales pero que también puede marcar la diferencia para definirlo como "really Big Data" como la movilidad (por ejemplo la adquisición de información mediante IoT o dispositivos móviles) y el tiempo real. De hecho al trabajar con Big Data se podrían considerar las siguientes vertientes que pueden o no trabajar en conjunto:

- **Ingeniería:** Enfocado en el uso de las herramientas por ejemplo al tratar verdaderamente mucha data con poco o nada de análisis, un rol de esta vertiente sería el Arquitecto de Datos, esa persona encargada de manipular estructurar los datos, manipularlos, masticarlos y dejarlos bien preparados para aquellos encargados de hacer análisis sobre los datos, esta persona trabajaría con Hadoop, Pig, Spark.
- **Científica:** Donde sin que estrictamente se tenga que trabajar con muchísima data (podría ser tanto small Data como Big Data) se lleva a cabo análisis mayormente de tipo estadístico como análisis predictivos, construyendo modelos, un rol de esta vertiente sería la del Data scientist, esa persona encargada de hacer data mining, machine learning, etc.

1.3 TIPOS DE DATOS

Una vez hemos fijado con mayor precisión el concepto de Big Data, vamos a proceder a analizar los tipos de datos existentes, además de aclarar la diferencia entre lo que es Big Data y lo que son datos desde el punto de vista tradicional. Cuando las empresas deciden llevar a cabo un proyecto de Big Data deben dar solución a una serie de cuestiones tales como: el origen de los datos, el volumen de información necesario para tomar una decisión, la información que aporta cada dato a mi negocio... Por tanto, es importante que la empresa reconozca las fuentes de datos existentes y el tratamiento que necesita cada dato.

En Big Data los datos son diferentes a los datos tradicionales es decir los datos estructurados almacenados en bases de datos relacionales. Los datos se consideran en dos tipos, los estructurados y los no estructurados como podemos ver en la siguiente imagen:

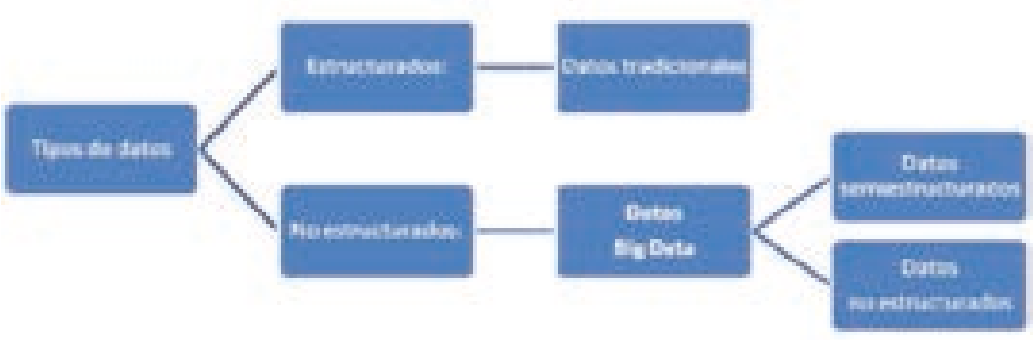

Figura 1.2. Tipos de datos en Big Data

- **Datos estructurados:** son aquellos datos con formato y campos fijos, en el que el formato es anticipadamente definido, para ser almacenados en bases de datos relacionales; este tipo de datos guardan un orden específico lo que facilita trabajar con ellos.
- **Datos semi estructurados**: son aquellos datos que no tienen formatos fijos, pero que contienen etiquetas, marcadores o separadores que permiten entenderlos; se procesan a base de reglas para extraer la información en piezas. Por ejemplo, los lenguajes XML y HTML son ejemplos de texto con etiquetas. no siguen un patrón claramente comprensible (como sí hacen los datos estructurados), a pesar de que, presentan un flujo claro y un formato definible. No existen formatos fijos como en los estructurados, pero sí marcadores para separar los datos. En esta categoría destacamos registros de logs procedentes de conexiones a internet.

- **Datos no estructurados:** son aquellos datos que no tienen formatos predefinidos, es decir no tienen estructura uniforme. Generalmente son datos binarios que no tienen estructura interna identificable. Es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se identifican y almacenan de manera organizada. Por ejemplo los correos electrónicos, mensajes instantáneos SMS, WhatsApp, Viber, fotos, audios, videos entre otros. Su almacenamiento se da sin estructura uniforme y no existe capacidad para controlar estos datos. Los ejemplos más claros son los videos, audios, fotos o datos de texto (SMS, WhatsApp, Correos electrónicos...) Estos datos suponen el 80% de los datos que poseen las empresas, siendo con diferencia aquellos que presentan una mayor dificultad en su análisis, por tanto, han dado lugar al nacimiento de herramientas como MapReduce, Hadoop o bases NoSQL que analizaremos más adelante.

1.4 CARACTERÍSTICAS DE BIG DATA

Los últimos diez años han visto un aumento extraordinario del interés de empresas y organizaciones por el uso de herramientas que les permitan manejar la ingente cantidad de datos que recogen diariamente a través de sus sistemas de información, de sus canales de ventas y compras, de la información recogida a través de su presencia en la Web (anuncios, páginas de acceso a información, a servicios, etc.) o incluso cada vez más de comentarios y mensajes que se puedan generar en las redes sociales.

Este fenómeno ha incrementado enormemente la demanda de aplicación de procedimientos de análisis de datos para detectar la presencia de patrones o de tendencias que no resultan obvias, aportan información muy valiosa para mejorar significativamente su actividad: sus operaciones, sus ventas o sus resultados. Por otra parte, y asociado a este interés, se ha iniciado un proceso de revisión y mejora de las técnicas cuantitativas existentes para el tratamiento de datos y la extracción de la información relevante.

Uno de los aspectos más significativos asociado a este nuevo interés, y uno que resulta especialmente relevante por los cambios que implica tanto en la formación básica necesaria como en las aplicaciones para los profesionales interesados en el tratamiento de datos, es el aumento extraordinario en el volumen de los datos disponibles.

Cada vez es más habitual que las organizaciones y empresas dispongan de cantidades de datos medibles en peta- o exabytes (miles de billones o trillones de bytes).Se ha popularizado el uso del término “Big Data” para referirse a estas cantidades de información y a las técnicas adecuadas para su tratamiento. Un problema asociado a estos volúmenes de datos es que las técnicas tradicionales no resultan aplicables por ineficientes; es necesario utilizar nuevos métodos, adaptados especialmente a estas situaciones, creando una demanda y ofreciendo una oportunidad de formación de profesionales muy relevante en el futuro inmediato.

Tecnologías como Internet generan datos a un ritmo exponencial gracias al abaratamiento y gran desarrollo del almacenamiento y los recursos de red. El volumen actual de datos ha superado las capacidades de procesamiento de los sistemas clásicos de minería de datos. Hemos entrado en la era del Big Data o datos masivos, que es definida con la presencia de gran volumen, velocidad y variedad en los datos, tres características que fueron introducidas por D. Laney en el año 2001, con el requerimiento de nuevos sistemas de procesamiento de alto rendimiento, nuevos algoritmos escalables, etc.

IBM y Gartner plantean tres dimensiones para el entendimiento de la naturaleza de los Big Data, conocido como el modelo de las 3V; inclusive IBM considera una cuarta V correspondiente a la veracidad, y otras fuentes añaden una más, el valor. Sin embargo, las distintas fuentes tienen en común el modelo de las 3V que corresponde a volumen, velocidad y variedad. Otros dos

aspectos importantes que caracterizan los datos masivos son la veracidad de los datos y el valor intrínseco del conocimiento extraído. La figura muestra estas cinco características.

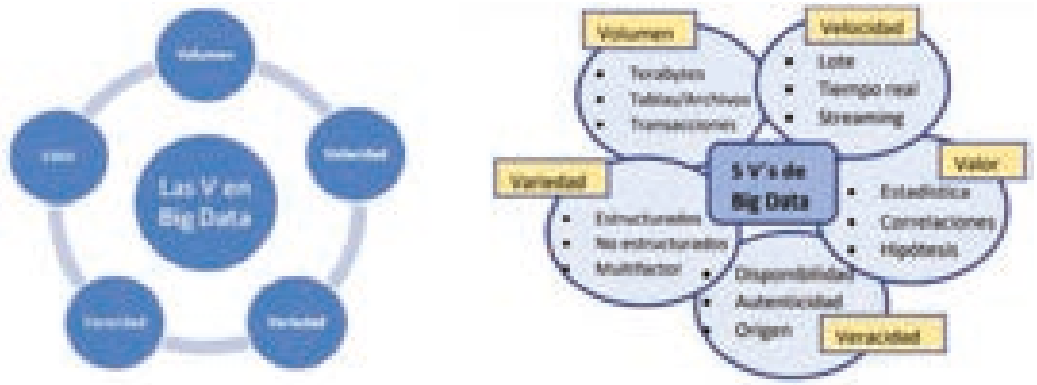

Figura 1.3. Características en Big Data

Dichos volúmenes de datos poseen cuatro características principales que vienen definidas como las cinco Vs:

- **Volumen de información.** Cantidad de datos que son generados a lo largo del tiempo. Es una de las principales características de Big Data, ya que hace referencia a las cantidades masivas de datos que se almacenan para ser procesados. Se espera que entre el presente año y el 2020, la humanidad entre en la era de zettabyte; en el año 2000 se almacenaron en el mundo 800.000 petabytes y se proyecta que para el año 2020 se alcancen 35 zettabytes (ZB); en 2012 Twitter generó más de 9 terabytes (TB) de datos al día. La gran cantidad de datos que las organizaciones generan es mayor día tras día, es por eso que hablar de un gran tamaño de datos en estos momentos es dar lugar a que solo en unos semestres ese valor no sea tan grande para ese momento. El volumen en Big Data hace referencia al presente de la cantidad de información que genera y almacena una organización, es decir a su volumen masivo de datos. Para IBM el volumen de datos disponible que tiene una organización está en ascenso en contraste con la disminución en el análisis que se hace de ellos.
- **Velocidad de los datos.** Rapidez con la que los datos son creados, almacenados y procesados en tiempo real. En muchas ocasiones es necesario hacer un estudio en tiempo real. En Big Data este tema merece consideración ya que el aumento de los flujos de datos en las organizaciones aumenta la velocidad en que se deben almacenar datos y sugiere últimas versiones de los gestores de grandes bases de datos. Este aumento en la velocidad de los datos requiere que el procesamiento de ellos se haga en tiempo real para mejorar la toma de decisiones.
- **Variedad de los datos.** Formas, tipos y fuentes en las que los datos son registrados. Los datos pueden ser estructurados y fáciles de gestionar como son las bases de datos, o no estructurados, como son los documentos de texto, correos electrónicos, datos de sensores, etc. Los datos no estructurados requieren un tratamiento diferente de los datos no estructurados, ya que es necesario un procesamiento de los datos recogidos de múltiples fuentes de información con la herramienta adecuada. representa todos los tipos de datos, entre datos estructurados y no estructurados. Las fuentes de datos son de cualquier tipo y ello aumenta la complejidad; por ejemplo los videos no pueden ser tratados en bases de datos relacionales como si se puede hacer con el registro de ingreso de los empleados

a una empresa. Considerar toda esta variedad de tipos de archivos supone un reto en la clasificación de los datos para que el análisis de ellos entregue resultados de valor a quien los investiga.

- **Veracidad de los datos**. Grado de falibilidad de los datos recibidos. Es necesario tener la certeza de que los datos obtenidos son de calidad, aplicando soluciones y métodos que puedan eliminar datos imprevisibles. En Big Data este término se relaciona a la fiabilidad de las fuentes de datos, debido al aumento de fuentes de ellos, y además a la variedad en los tipos de datos.
- **Valor**: Es la característica más importante de los datos. Como hemos visto anteriormente, el potencial de los datos es espectacular, pero de nada sirve tener acceso a una gran cantidad de datos si no somos capaces de convertirlos en algo con valor. Es decir, la información no sirve de nada a las organizaciones si esta no les otorga una fuente de valor, por tanto, para que las empresas realicen la inversión en almacenes de datos y sistemas de procesamiento y análisis debe existir un retorno claro de esta inversión. Como consecuencia de esto nace Hadoop cuya función es el análisis de grandes volúmenes de datos.

Es importante resaltar que, al pasar de la administración de una simple base de datos a adoptar el uso de Big Data, se necesita implementar una determinada arquitectura. Ésta viene marcada por el ciclo de vida del procesamiento de datos: capturar, organizar, integrar, analizar, actuar. En la siguiente imagen vemos los principales elementos cuando trabajamos con Big Data.

Figura 1.4. Elementos en Big Data

- **Collection (recogida):** una de las mayores dificultades a la hora de disponer los datos es cómo conseguirlos.
- **Storage (almacenamiento):** una vez han sido obtenidos, hay que determinar cómo almacenarlos de la manera más óptima para su gestión y posterior consulta.
- **Research (investigación):** la información que se pretende extraer de los datos debe ser parte de un proceso de investigación y de mejora continua para el descubrimiento de nuevas capacidades.
- **Analysis (análisis):** para que de los datos se pueda extraer una información valiosa, deben ser analizados.
- **Volume (volumen):** hablamos de Big Data y no de otras variaciones cuando se incluye un componente de volumen y complejidad.
- **Visualization (visualización):** para su mejor comprensión y sobre todo, de cara a poder orientar y convencer a los actores decisivos de una empresa, es imprescindible una visualización amigable del resultado del análisis.
- **Cloud technology (tecnología en la nube):** los datos deben estar disponibles para su consulta por distintos agentes en cualquier momento y desde distintas ubicaciones, además del hecho de que tener externalizados servicios en la nube tiene ventajas adicionales para una empresa, como se verá más adelante.
- **Network (red):** se trata de la infraestructura física que sustenta el punto anterior.

1.5 DESAFÍOS DE BIG DATA

Como toda tecnología en desarrollo, Big Data presenta desafíos relacionados a distintos factores, desde el hecho de hacer cambiar las infraestructuras y formas de pensar de los desarrolladores que hoy están acostumbrados a tecnologías como information retrieval y data mining, utilizando estilos tradicionales de desarrollo, hasta saber qué tipo de datos son los adecuados para buscar información para estas implementaciones. Entre los desafíos más comunes podemos citar los siguientes:

- **Skills:** Este problema trata básicamente la capacidad de las personas a cargo del manejo de la información recolectada. Al ser una tecnología en desarrollo, la cantidad de personas que tengan el "know how" o conocimiento para poder procesar de manera correcta el volumen de información es relativamente poco, lo que dificulta el desarrollo de proyectos.
- **Estructura de datos:** Otro gran desafío es la forma en la que se guardan los datos. La forma misma en que tenemos concebida la idea de cómo guardar los datos en la actualidad presenta un desafío enorme para Big Data. El desafío de hoy es que la mayoría de los almacenes de datos empresariales ven un cliente o una entidad que la empresa trabaja con una fila de datos en lugar de una columna. Esa fila se rellena y se actualiza quizás a diario con la instantánea o al agregado de la situación actual del cliente. Al realizar esta actualización, estamos perdiendo la información recolectada, lo que conlleva a menor capacidad de predicción o información a procesar.
- **La tecnología:** Lo interesante es que Hadoop es ideal para el procesamiento por lotes a gran escala, que es como las operaciones de agregación o cómputo. El problema es que Hadoop no es una tecnología en tiempo real o muy dinámica en absoluto. La ejecución de consultas en un clúster Hadoop suele tener una gran latencia ya que hay que distribuir cada consulta individual, luego, hacer su etapa de reducción, que está trayendo todos los datos de nuevo juntos. Así que es una tecnología de alto rendimiento, pero de alta latencia.

- **Privacidad:** Junto con la obtención de volúmenes de datos incalculables, viene una cantidad de datos que podríamos considerar intrusiva, podría darse ejemplos como Facebook, Twitter, Google que manejan grandes volúmenes de datos de clientes, con esta capacidad de Big Data de intentar analizar absolutamente todo, podría darse una examinación inapropiada de los datos de usuarios, conllevando rupturas en la privacidad de los datos de los usuarios. (Si bien esta problemática no es nueva, podría agravarse con la capacidad avanzada de procesamiento que se obtiene con Big Data).
- **Volumen, Variedad, Velocidad:** La capacidad de encontrar un equilibrio entre todas ellas depende de la capacidad de plantear un desarrollo sustentable y un plan acorde a las posibilidades tecnológicas de la empresa que desarrolla con esta tecnología.

A nivel técnico, la adopción de tecnologías big data supone una serie de desafíos entre los que podemos destacar:

- El análisis de datos estructurados es necesario para comprender los métodos de análisis de Big Data, incluso existen métodos que se comparten con el análisis convencional, pero con muchos más datos.
- La administración de bases de datos es un fundamento para el análisis de datos y para manejar datos operacionales. En Big Data, las bases de datos son una fuente importante que alimenta el núcleo de procesamiento.
- La programación orientada a objetos es el pilar para desarrollar cualquier tipo de aplicación, incluso para manejar bases de datos. El Big Data se utiliza para manejar y procesar distintos tipos de datos.
- La administración de servidores es necesaria para aprovechar al máximo las tecnologías de la información. En Big Data son primordiales pues son el soporte de toda la infraestructura de aprovechamiento de los datos masivos.

Figura 1.5. Desafíos en Big Data

1.6 TECNOLOGÍAS PARA BIG DATA

Las tecnologías y algoritmos sofisticados y novedosos son necesarios para procesar eficientemente lo que se conoce como Big Data. Estos nuevos esquemas de procesamiento han de ser diseñados para procesar conjuntos de datos grandes, datos masivos, dentro de tiempo de cómputo razonable y en un rango de precisión adecuado.

Desde el punto de vista del aprendizaje automático, esta problemática ha causado que muchos algoritmos estándar se conviertan en obsoletos en el paradigma Big Data. Como resultado surge la necesidad de diseñar nuevos métodos escalables capaces de manejar grandes volúmenes de datos, manteniendo a su vez su comportamiento en términos de efectividad.

Google diseñó MapReduce en 2003 la que es considerada como la plataforma pionera para el procesamiento de datos masivos, así como un paradigma para el procesamiento de datos mediante el particionamiento de ficheros de datos. MapReduce es capaz de procesar grandes conjuntos de datos, a la vez que proporciona al usuario un manejo fácil y transparente de los recursos del clúster subyacente.

En el paradigma **MapReduce**, existen dos fases: Map y Reduce. En la fase Map, el sistema procesa parejas clave-valor, leídas directamente del sistema de ficheros distribuido, y transforma estos pares en otros intermedios usando una función definida por el usuario. Cada nodo se encarga de leer y transformar los pares de una o más particiones. En la fase Reduce, los pares con claves coincidentes son enviadas al mismo nodo y finalmente fusionados usando otra función definida por el usuario. La siguiente figura muestra un esquema general del proceso completo MapReduce:

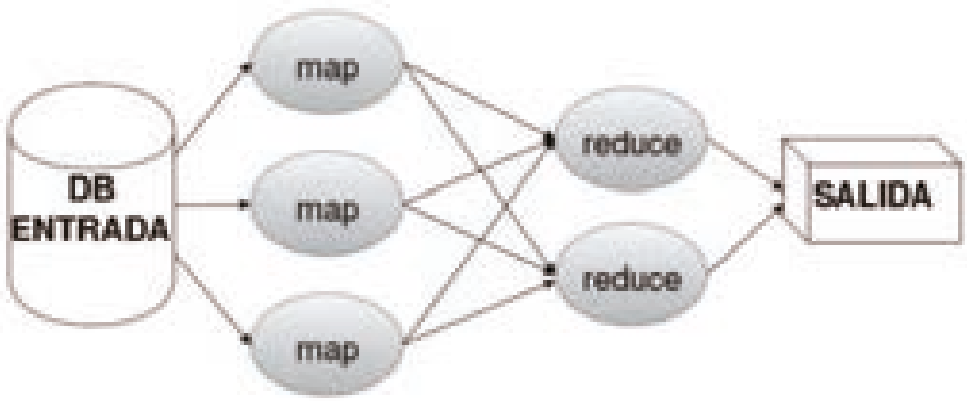

Figura 1.6. Modelo Mapreduce en Big Data

Este modelo consiste en dos funciones primitivas “Map” y “Reduce”. La entrada de “Map” es un conjunto de pares clave-valor (k1, v1) a los cuales se les aplica una función “Map” que devuelve como resultado un conjunto intermedio de pares clave-valor (k2 ,v2). Este conjunto intermedio se agrupa según claves iguales, las cuales sirven de entrada para la función “Reduce”, la cual trabaja sobre toda la lista de valores asociados a la misma clave y produce cero o más resultados agregados en forma de lista (list v3). Destacar que los conjuntos de pares clave-valor pueden pertenecer a dominios diferentes.

```
Map
Map(k1,v1) -> list(k2,v2)

Reduce
Reduce(k2, list (v2)) -> list(v3)
```

La función Map tiene como entrada una serie de pares <clave, valor> y produce una lista de pares intermedios como salida. La función Map, que internamente procesa los datos en cada proceso, es definida por el usuario siguiendo el esquema clave-valor. El esquema general para dicha función es el siguiente:

```
Map(<clave1, valor1>) -> lista(<clave2, valor2>)
```

En la segunda fase, el nodo maestro agrupa pares por clave y distribuye los resultados combinados a los procesos Reduce en cada nodo. La función de reducción es aplicada a la lista de valores asociada a cada clave y genera un valor de salida. Dicho proceso es esquematizado a continuación:

```
Reduce(< clave2, lista(valor2) >) →< clave3, valor3>
```

1.7 PERFILES BIG DATA

Un especialista en Big Data es un profesional que cuenta con amplios conocimientos en una serie de tareas involucradas en el ciclo de vida de la gestión de los datos tales como: identificar diversos orígenes de información, almacenar y extraer grandes volúmenes de datos, diseñar la arquitectura del ecosistema empresarial donde se procesa y consumirá los datos para su exploración, modelado, análisis, visualización y monitorización en tiempo real. Dependiendo de sus funciones, un especialista en Big Data debe poseer habilidades empresariales, técnicas y analíticas para obtener el mayor provecho de la información.

La constante y creciente generación de datos en todas las actividades humanas, y la consecuente necesidad de procesar y analizar un volumen cada vez mayor de información, implica una enorme oportunidad laboral. Un experto en Big Data forma parte de uno de los sectores profesionales con mayor oferta de empleos.

La clave para poder obtener, procesar, analizar y darles un aprovechamiento efectivo a los datos, pasa por la implementación de tecnologías adecuadas y contar con expertos en big data que sean capaces de gestionarlas e interpretar la información con foco en el negocio.

Dado que el uso de plataformas de Big Data aumenta cada vez más para dar paso a la transformación digital, es común que las empresas desarrollen sus propios sistemas con componentes legacy, en la nube o en ambos, por lo que los expertos de Big Data deben tener dominio en diferentes lenguajes de programación, aplicaciones tecnológicas, pero además de herramientas en entornos cloud.

Big Data con el panorama actual catapulta a los científicos de datos como otra muy buena opción de carrera profesional y sobre todo bien remunerada. Ya que el Big Data es una herramienta clave para las empresas para ganar competitividad, tomar decisiones basadas en datos.

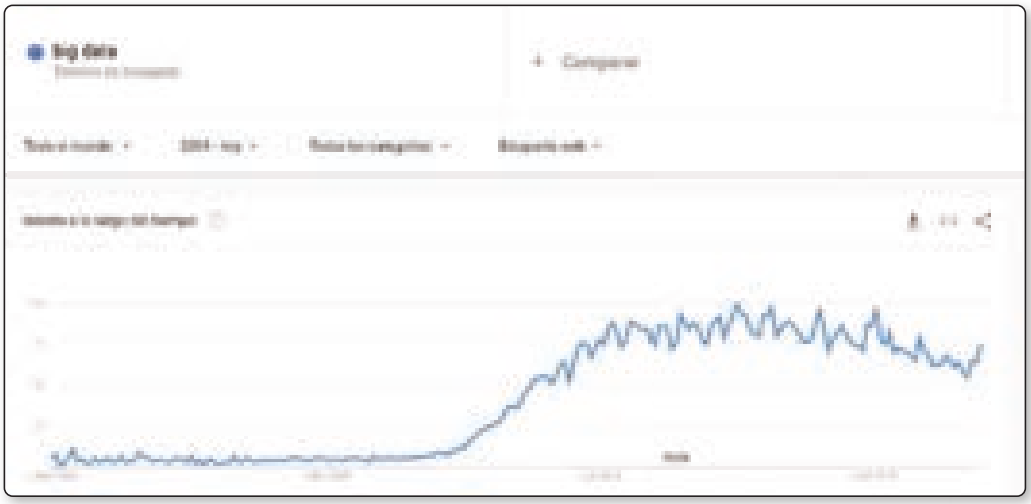

Figura 1.7. Evolución del término a lo largo del tiempo

Un aspecto muy importante es que los científicos de datos, no sólo se desarrollan como personas técnicas, es decir no están aislados en el área de sistemas y de allí no tienen interacción con el resto de la empresa a la que pertenecen, sino todo lo contrario, los científicos de datos van de la mano de la toma de decisiones de las empresas e interactúan con la mayoría de las áreas para obtener datos valiosos y saber cómo interpretarlos, es decir los científicos de datos están tomando decisiones o están al lado de los tomadores de decisiones.

Pero no solo eso se necesita para convertirse en un profesional de Big Data, además de tener algún máster o doctorado, se necesitan tener habilidades de comunicación ya que como se mencionó los científicos de datos tienen que estar en contacto con la mayoría de áreas de las empresas y por ende saber comunicarse con conocedores del dominio a tratar para sacar el mayor valor a los datos, se necesita un alto grado de curiosidad y tener una comprensión de lo que son negocios reales, deben de saber que una mala decisión tiene consecuencias reales en las empresas.

1.7.1 DIRECCIÓN DE DATOS(CHIEF DATA OFFICER-CDO)

Es el responsable de todos los equipos especializados en Big Data de la organización. Su función combina la rendición de cuentas y responsabilidad en cuanto a privacidad y protección de la información, calidad y gestión de los datos. Se trata del director digital de la empresa. Es una figura clave, ya que este profesional es el director digital de la empresa.

Se trata del líder de la gestión de datos y analítica asociada por el negocio, quien debe dirigir los equipos especializados en dato, definir políticas de seguridad para gestionar y almacenar datos, mantenerse actualizado en las regulaciones vigentes en cada país, decidir qué datos se utilizarán, incluyendo cómo y para qué, validar las tecnologías que se utilicen y ayudar a democratizar el acceso a los datos a todos los empleados y empleadas. Para aspirar a este puesto se requieren las siguientes **competencias**:

- Varios años de experiencia en el sector de la tecnología y trayectoria en el campo de la analítica aplicada al negocio.
- Formación en estadística y graduación en carreras como ingeniería, informática o telecomunicaciones. Se valoran los Másteres en Big Data, MBA o gestión de negocios.

- Habilidades de comunicación, planificación y gestión integral de proyectos, trabajo en equipo y marcación de objetivos.
- Capacidad analítica y orientación al cliente.

Este profesional es el encargado de coordinar los esfuerzos de todos los profesionales dedicados al Big Data en una organización. Debe establecer la metodología de trabajo, y asegurarse de que esta se encuentre enfocada en obtener los datos que la empresa necesita.

La formación profesional requerida para el cargo es la misma que requiere un experto en Big Data, pero generalmente para llegar a un puesto de CDO se requieren años de experiencia en el área. También se puede alcanzar este perfil combinando experiencia de Big Data con experiencia a nivel de gestión.

1.7.2 CIENTÍFICO DE DATOS(SCIENTIST)

El científico de datos analiza, interpreta y comunica las nuevas tendencias en el área y las traduce a la empresa para que puedan hacer uso de ellas y así adaptar sus productos y servicios y crear nuevas oportunidades de negocio. Es el encargado de traducir la información para que los analistas puedan tomar decisiones.

Para el perfil de científico de datos se precisan conocimientos estadísticos que un programador no suele tener y conocimientos informáticos que un estadístico no suele manejar. Dentro de este perfil diferenciamos entre los profesionales orientados al campo de las matemáticas y las estadísticas y los que proceden del ámbito de la inteligencia artificial y el machine learning. Este perfil debe unir conocimientos de matemáticas, estadística y programación, y conocer también muy al detalle el sector de actividad de la compañía para la que trabaja, además de ser buen comunicador para trasladar los datos que interpreta.

La principal función del científico de datos es la de traducir los grandes volúmenes de datos y convertirlos en información útil para la empresa. Tiene conocimientos matemáticos, estadísticos y de programación. También cuenta con una visión de negocio y habilidades comunicativas, para dar a conocer el resultado de su trabajo al resto de la organización.

Permiten extraer conocimiento e información valiosa de los datos. Tienen visión general del proceso de extremo a extremo y pueden resolver problemas de ciencias datos, la construcción de modelos analíticos y algoritmos. Combinan diversas habilidades relacionadas con las matemáticas, la estadística, la programación y visualización, pero también deben tener habilidades comunicativas, para explicar los resultados obtenidos en la organización. Estas disciplinas están en línea con las habilidades que se demandan hoy en día de un **data scientist**:

- **Programación**: Para la limpieza, tratamiento, filtrado, etc. de los datos es necesario conocimientos de Programación.
- **Informática**: Nos dará la infraestructura y herramientas necesarias para almacenar los datos, procesarlos, etc., especialmente cuando nos movemos en el mundo Big Data.
- **Estadística**: Para la obtención y visualización de insights, responder las cuestiones planteadas, representar la información que obtengamos, ... saber qué modelos, algoritmos, etc. podemos utilizar, cómo validar los resultados, …
- **Matemáticas**: Para entender los fundamentos de los modelos y técnicas estadísticas que empleemos.

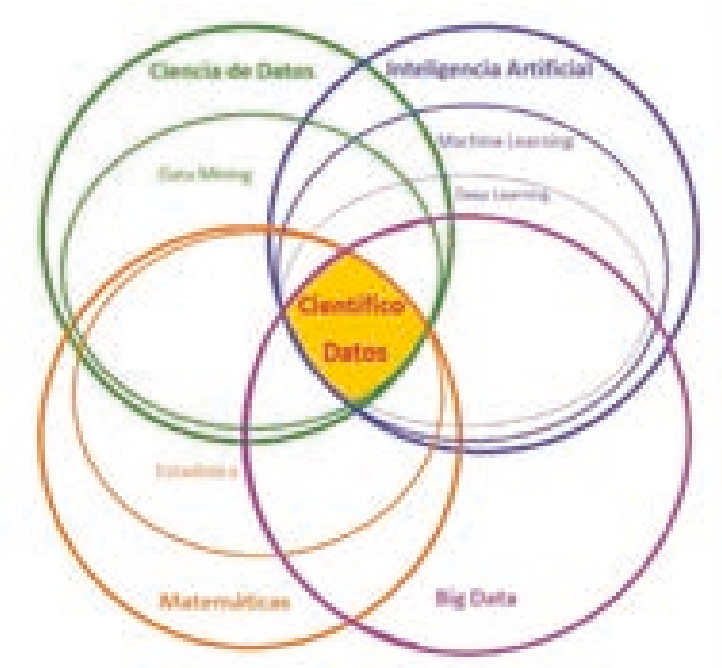

Figura 1.8. Áreas de conocimiento de un científico de datos

Se trata de un perfil muy buscado por empresas vinculadas a motores de búsqueda, servicios financieros y e-commerce, ya que su aporte reside en la extracción de información valiosa de los datos generados en el marco de la operación, con una visión general del proceso. Para aspirar a este puesto se requieren las siguientes **competencias**:

- Conocimientos de ingeniería de software en sistemas distribuidos, algorítmica y estructuras de datos.
- Ser experto en matemáticas, estadística, informática, etc.
- Saber de Machine Learning, lenguajes de programación como R o Python, y el uso de notebooks y ecosistemas Big Data.
- Poseer gran capacidad para la resolución de problemas.
- Capacidad para analizar, resolver y explicar en forma entendible evitando conceptos científicos, y predecir comportamientos futuros.
- Conocimientos en programación y aptitud para expresarse en lenguajes informáticos.
- Facilidad en álgebra lineal, cálculo y probabilidad.
- Comprensión y manejo de las técnicas de machine learning.
- Pensamiento lógico, análisis, predicciones y capacidad de detección de patrones.
- Capacidad para usar librerías como Tensor Flow para técnicas de Deep Learning basadas en redes neuronales.

Básicamente, su trabajo consiste en analizar y gestionar la información generada por los usuarios, y transformarla en datos comprensibles para las empresas. Mediante la creación de gráficos y estadísticas, este profesional será capaz de arrojar luz sobre miles de millones de datos en bruto.

Para alcanzar este perfil, lo recomendable es optar por carreras de Informática o Tecnología y luego realizar una especialización. Aunque existen instituciones privadas que ofrecen cursos breves con los que se puede adquirir este perfil.

Este perfil presenta un nivel de conocimientos superior al del analista de datos, pudiendo realizar sus mismas tareas, pero con un nivel de profundidad superior a nivel de conocimientos matemáticos y de programación, que les permiten conocer los detalles de implementación de los métodos y algoritmos de minería de datos y machine learning.

Los científicos de datos se dedican a resolver problemas con casuísticas complejas, muchas veces de problemas ad hoc que requieren un análisis y dedicación profunda. Deben de ser capaces de hacer investigación y conocer el estado del arte en los temas de minería de datos y machine learning, ya que la optimización de los algoritmos mediante la parametrización (fine-tuning) es una de sus responsabilidades.

1.7.3 ANALISTA DE DATOS(DATA ANALYST)

El perfil de analista de datos se encuentra en la intersección de otras disciplinas como Informática y Programación, Estadística y Matemáticas. Como su cargo indica, se encarga de participar en el análisis de los datos y recoge las necesidades de los clientes para presentarlas al Data Scientist. También se encarga de extraer, procesar y agrupar datos, analizar esas agrupaciones de datos y generar informes.

Es uno de los perfiles profesionales más demandados actualmente por las empresas ya que se encarga de procesar la información y obtener conclusiones que ayuden a mejorar resultados. Estos profesionales son los que saben extraer patrones de conducta de los usuarios y saben por qué actúan de una manera.

Tiene la responsabilidad de descubrir cómo extraer datos, procesarlos y sintetizarlos para obtener conclusiones y resolver aquellos problemas surgen en una organización, a través de modelos computarizados avanzados, y modelos analíticos y de visualización de datos sintonizados con los requerimientos de una compañía. Para aspirar a este puesto se requieren las siguientes competencias:

- Estudios de grado en Estadística, Matemáticas o Ingenierías (técnica y/o superior).
- Dominio de lenguajes de programación como Python, y programas estadísticos.
- Capacidad para extraer, limpiar, analizar, modelar e interpretar datos.
- Habilidades de comunicación, planificación y trabajo en equipo.
- Además de los conceptos de Machine Learning, deben destacarse por el conocimiento del entorno Big Data en el que trabajan, como Spark o Hadoop.
- Son valorados los conocimientos de Bases de Datos SQL y Business Intelligence.

1.7.4 INGENIERIO DE DATOS(DATA ENGINEER)

La principal tarea de un ingeniero de datos es la de distribuir datos de manera accesible a los Data Scientist. Su perfil es más especializado en gestión de bases de datos y en sistemas de procesamiento y de programación. Podríamos definir un Data Engineer como aquel profesional

enfocado en el diseño, desarrollo y mantenimiento de los sistemas de procesamiento de datos dentro de un proyecto de Big Data. Entre las principales que desempeña podemos destacar:

- Proporcionar los datos de una manera accesible y apropiada a los usuarios y Data scientists.
- Desarrollar y explotar técnicas, procesos, herramientas y métodos que deben servir para el desarrollo de aplicaciones Big Data.
- Tiene un gran conocimiento en gestión de bases de datos, arquitecturas de clusters, lenguajes de programación y sistemas de procesamiento de datos.
- Construir y mantener las estructuras y las arquitecturas tecnológicas necesarias para el procesamiento, ingestión e implementación a gran escala de aplicaciones que usan datos de forma intensiva.
- Se especializa en infraestructura Big Data, creando e implementando técnicas, procesos, herramientas y métodos para el desarrollo de aplicaciones Big Data.
- Juega un papel clave a la hora de convertir una prueba de concepto de Big Data en un proyecto real y palpable.

Para aspirar a este puesto se requieren las siguientes **competencias**:

- Conocimiento en gestión de bases de datos, arquitecturas de clusters, lenguajes de programación y sistemas de procesamiento de datos.
- Trabajar con Linux y Git, y también con Hadoop y Spark a nivel de entornos, Mapreduce a nivel de modelos computacionales, y HDFS, MongoDB y Cassandra a nivel de tecnologías NoSQL.
- Se suelen requerir los siguientes lenguajes: Python para el procesado de datos con librerías como **PySpark** y Scala como lenguaje nativo de Spark y Java, en muchos casos.

1.7.5 ARQUITECTO DE DATOS(DATA ARCHITECT)

El Arquitecto de Big Data es quien define la infraestructura de la plataforma de Big Data. Tiene una visión global tanto de las necesidades de las empresas u organizaciones como de las soluciones de tratamiento de datos recomendadas para cada caso.

Cuando el ingeniero de datos se dedica a diseñar, implementar y mantener infraestructura relacionada con sistemas de Big Data, se llama arquitecto de Big Data. En este caso, su trabajo se centra en el diseño, creación y mantenimiento de clusters de procesamiento distribuido, como por ejemplo Apache Hadoop, Apache Spark o Apache Flink, y sistemas de almacenamiento distribuido de datos, como por ejemplo el sistema de ficheros distribuido HDFS o las bases de datos NoSQL.

Este perfil tiene como objetivo velar por el buen funcionamiento y la seguridad de las plataformas y el hardware que contienen los datos, debiendo prever los nuevos escenarios de volumen de datos que se puedan presentar en un futuro. Para aspirar a este puesto se requieren las siguientes **competencias**:

- Formación en Informática y/o matemáticas.
- Experiencia para manejar tecnologías de datos no estructurados como Hadoop, Spark o Cassandra.
- Conocimientos de lenguajes de programación distribuida y funcional como Java/scala, SQL y Python.
- Conocimientos en bases de datos como Oracle y PostgreSQL.

Estos perfiles tienen una formación en ingeniería informática, matemática, física o telecomunicaciones. Algunas de las herramientas que deben manejar son Hadoop, MapReduce, Hive, Pig, Spark, Flink, experiencia en SQL (MySQL, PostgreSQL) y NoSQL (Hive, Couchbase, Redis, ElasticSearch, Solr), conocimientos avanzados en Java, Scala o Python o conocimientos avanzados de estructuras de datos, data mining aplicado y machine learning.

En los casos de Data Scientist provenientes del mundo de la estadística suelen trabajar con R como herramienta principal y tratan de realizar las tareas de manipulación y agregación de datos mediante R, aunque a veces en el mundo Big Data no sea la solución ideal.

Sin embargo, en los casos de Data Scientists que vienen del mundo del desarrollo de software, estos profesionales se sienten más cómodos con lenguajes más formales, y aquí es donde Python juega un papel fundamental, las distintas contribuciones, módulos y librerías de análisis, de machine learning y librerías específicas (series temporales, procesado de lenguaje natural, ...) junto con su fácil integración con las plataformas Big Data hacen de Python la tecnología ideal para análisis de datos.

1.7.6 GESTOR DE DATOS(DATA MANAGER)

El objetivo principal de los gestores de datos o Data Manager es supervisar los diferentes sistemas de datos de una empresa. Son los encargados de organizar, almacenar y analizar los datos de la forma más eficiente posible.

Los Data Manager tienen conocimientos relacionados con la informática y de uno a cuatro años de experiencia. Destacan en el mundo de los números, los registros y los datos en bruto. Pero también tiene que estar familiarizado con todo el sistema de datos y tener una mente lógica y analítica con buena capacidad para resolver problemas.

1.7.7 CIUDADANO CIENTÍFICO DE DATOS(CITIZEN DATA SCIENTIST)

Es el profesional que no tiene una formación específica en Data Scientist, pero que con su experiencia puede aportar valor. Por ejemplo, realizando tareas analíticas y de gestión de datos a través de herramientas más sencillas. Se define como una persona que crea o genera modelos que utilizan análisis de diagnóstico avanzado o capacidades predictivas, pero cuya función principal de trabajo está fuera del campo de la estadística y el análisis.

En resumen, son personas no técnicas que pueden usar herramientas de ciencia de datos para resolver problemas relacionados con big data. Su experiencia y su conocimiento de las prioridades de la organización les permiten integrar de forma eficaz la ciencia de datos y el desarrollo de machine learning en los procesos.

1.7.8 ADMINISTRADOR DE DATOS(DATA STEWARD)

Este especialista es el responsable de velar por la calidad, la seguridad y la disponibilidad de los datos. Su función se resume en saber utilizar los datos dentro del proceso de negocio y presentarlos a toda la organización.

Este perfil trata la gestión y supervisión de los activos de datos de una organización para ayudar a proporcionar a los comerciales datos de alta calidad a los que poder acceder fácilmente. Data Steward se enfoca en la coordinación e implementación de tácticas. Es responsable de llevar a cabo el uso de datos y las políticas de seguridad según lo determinado a través de iniciativas empresariales, actuando como enlace entre el departamento de IT y el comercial.

1.7.9 TABLA COMPARATIVA

Como se puede observar en la siguiente tabla la diferencia entre algunos de los roles es cuestión de matices.

Ingeniero de datos	Perfil orientado al desarrollo de software y con experiencia en el tratamiento de datos desde la extracción y depuración hasta el procesamiento y el almacenamiento.
Arquitecto Big Data	Encargado de definir la arquitectura de los sistemas Big Data, eligiendo las alternativas más óptimas desde el punto de vista de seguridad, gobierno del dato y rendimiento. También es el responsable de mantener las plataformas actualizadas tecnológicamente y proponer nuevas alternativas que mejoren lo existente cuando van apareciendo.
Científico de datos	Perfil que cuenta con background de investigación en ámbitos de ingeniería, física y estadística. Experto en tratar problemas complejos, extrapola el conocimiento adquirido en el contexto académico para resolver problemas planteados en el entorno empresarial.
Analista de negocio	Profesional orientado al negocio con capacidad para comprender los resultados derivados del análisis avanzado de datos. Crea propuestas de valor para el negocio con el fin de generar beneficios para la empresa.
Ingeniero de visualización de datos	Perfil diferencial en visualización de datos y storytelling con capacidad para explotar el valor de los datos y hacerlos entendibles. Aplica herramientas de programación, de Data Discovery y de visualización.

El rol del Científico de Datos es el más importante en cuanto a la interpretación de los datos, diseño de algoritmos y análisis predictivos, es el que aplica métodos matemáticos y estadísticos a los datos para obtener valor de ellos, adicionalmente aplica conocimientos y metodologías de distintas áreas a los datos como machine learning, deep learning, inteligencia artificial.

En cuanto al Ingeniero de Datos es quien diseña e implementa la solución de Big Data para almacenar, consumir, analizar, visualizar los datos. También es el encargado de decidir qué tecnologías se adaptan mejor a la situación que se está tratando para obtener el mayor beneficio y valor de los datos. Al estar relacionado con el desarrollo, suele tener conocimiento sobre lenguajes de programación orientados a análisis de datos como R y Python.

1.8 BIG DATA ANALYTICS

En la actualidad los Big Data pueden ayudar a responder cuestiones clave acerca de cómo se comportan los clientes, cómo van a funcionar los nuevos lanzamientos, las futuras campañas o las posibles promociones. Esto está contribuyendo a realizar mejoras en los negocios mediante el marketing personalizado (one-to-one), las estrategias de competencia monopolística en precios, el análisis de atribución para estímulos comerciales, etc.

Por este motivo, se suele considerar que Big Data, más que tratar sobre datos, trata "sobre la transformación empresarial, sobre pasar del planteamiento retrospectivo de la monitorización y el procesamiento de datos por lotes a la obtención de conocimientos empresariales en tiempo

real". La Era del Big Data, por tanto, produce una creciente competencia en la comprensión de las necesidades del cliente en todo momento.

La posibilidad de aplicar técnicas de Big Data Analytics está haciendo que las empresas introduzcan paulatinamente una mayor "cultura de los datos" ("Data-driven culture") dentro de su operativa empresarial habitual, recurriendo tanto a nuevas tecnologías de almacenamiento y gestión de datos, como a herramientas de visualización y monitorización de métricas acerca del funcionamiento de la empresa (Key Performance Indicators - KPI) insertas en cuadros de mando ("dashboards").

Los indicadores clave de desempeño o KPI son valores que indican el rendimiento de un proceso de acuerdo con un objetivo predeterminado. Toda organización debe ser capaz de identificar sus propios KPI, por lo que deben tener:

- Definido completamente y acotado su proceso de negocio.
- Objetivos claros o el rendimiento del proceso de negocio.
- Una medida cuantitativa o cualitativa de los resultados con relación a los objetivos.
- Información sobre las variaciones entre los resultados y los objetivos planteados para ajustar procesos o recursos y alcanzar metas a corto plazo.

De este modo, la era del Big Data está haciendo que las empresas evolucionen por los estados de madurez siguientes: en primer lugar, la analítica descriptiva, en la que únicamente se dispone del dashboard en estado inicial; en segundo lugar la analítica de diagnóstico, enfocada a una comprensión avanzada y continua de la situación empresarial; en tercer lugar, la analítica predictiva, enfocada en la anticipación de riesgos y oportunidades; en cuarto lugar, la analítica prescriptiva, enfocada a la recomendación de acciones; y, por último, la analítica cognitiva, hoy en día emergente.

La analítica de Big Data es el proceso de examinar con gran velocidad, conjuntos de grandes volúmenes de datos entre una amplia variedad de tipos y descubrir patrones ocultos, nuevas correlaciones y más información útil, en un tiempo razonable en el que la oportunidad de la información proporcione ventajas competitivas al investigador.

Los grandes volúmenes de información pueden proceder de fuentes de datos no estructurados como los que generan smartphones, medios de comunicación, información suministrada por sensores, actividades sociales, entre otros; pero, además pueden proceder de datos estructurados almacenados en bases de datos relacionales.

El análisis de grandes datos (analítica de Big Data o Big Data analytics) corresponde a datos estructurados, no estructurados y semiestructurados. El análisis de grandes datos relacionales se puede realizar con herramientas de software tradicionales con técnicas sencillas o avanzadas como minería de datos, análisis predictivo y análisis estadísticos.

En cuanto a las fuentes de datos no estructuradas, pueden no encajar dentro de los esquemas de los almacenes de datos tradicionales o EDW (Enterprise Data Warehouse) o no estar en capacidad de atender la demanda de procesamiento de datos requerido.

Para atender la demanda de procesamiento de grandes datos han surgido tecnologías de bases de datos distintas a las relacionales llamadas bases de datos NoSQL, bases de datos en memoria y MapReduce. Este sistema se integra a través de un cluster, y bien puede ser por medio de software de código abierto o propietario. Para el tratamiento de los grandes volúmenes de datos se requieren las siguientes etapas:

- **Adquisición de datos:** los datos proceden de fuentes de datos diversas, es decir, de fuentes de datos tradicionales como almacenes de datos, bases de datos relacionales, entre otros; y, de fuentes de datos no estructurados o semi estructurados. Los datos procedentes de ambos tipos de fuentes de datos, pueden ser almacenados en bases de datos NoSQL o en bases de datos "en memoria".
- **Organización de los datos**: el origen distinto en las fuentes de datos requiere que luego de que se adquiera la información, deba prepararse, siendo tal vez necesario eliminar datos o parte de ellos para dejar lo más relevante de estos.
- **Análisis de información**: es una etapa muy importante dentro del tratamiento de los grandes volúmenes de datos. Consiste en analizar todos los datos por medio de herramientas estadísticas avanzadas como minería de información, minería social, herramientas desarrolladas para diseño de estadística avanzada como el lenguaje de programación R.
- **Decisión**: es en esta etapa en donde con los resultados obtenidos del análisis de información se obtiene conocimiento, preferiblemente en tiempo real; para que se incluya en los tableros de control, cuadros de mando y herramientas de visualización, y así predecir el comportamiento que va a tener el objeto de estudio.

El **preprocesamiento de datos** es una etapa fundamental en el proceso de extracción de conocimiento, cuyo objetivo principal es obtener un conjunto de datos final que sea de calidad y útil para la fase de extracción de conocimiento. El preprocesamiento de datos se vislumbra como una herramienta muy importante en el paso de Big Data a Smart Data, esencial para convertir los datos almacenados (material en bruto) en datos de calidad (valga el símil del paso de un diamante en bruto sin pulir y sin tallar a la piedra preciosa tras su procesado).

Para la mayoría de problemas actuales con datos masivos es necesario el uso de una solución distribuida escalable porque las soluciones secuenciales no son capaces de abordar tales magnitudes. Varias plataformas para el procesamiento a gran escala (como Spark o Hadoop) han intentado afrontar la problemática del Big Data en los últimos años. Estas plataformas requieren algoritmos escalables que den soporte a las tareas más relevantes de la analítica de datos masivos.

Los algoritmos de preprocesamiento también están afectados por el problema de la escalabilidad, por lo tanto deben ser rediseñados para su uso con tecnologías Big Data si queremos preprocesar conjuntos de datos masivos en los diferentes escenarios de aplicación, aprendizaje supervisado y no supervisado, procesamiento en tiempo real (flujo masivo de datos), etc.

Los que se llaman modelos no supervisados, en los que se incluyen las Reglas de Asociación, Patrones Secuenciales y Clustering. Los modelos supervisados, que necesitan un conjunto de entrenamiento del que aprender y que se suele etiquetar manualmente. Aquí se incluyen los modelos que pretenden adivinar un valor numérico para la variable objetivo que se quiere adivinar (predicción) o una etiqueta (clasificación) que puede tener únicamente dos valores posibles (binaria) o más de dos (multiclase).

La preparación de datos está formada por una serie de técnicas que tienen el objetivo de inicializar correctamente los datos que servirán de entrada para los algoritmos de minería de datos. Este tipo de técnicas pueden clasificarse como de uso obligado, ya que sin ellas los algoritmos de extracción de conocimiento no podrían ejecutarse u ofrecerían resultados erróneos. En esta área se incluye la transformación de datos y normalización, integración, limpieza de ruido e imputación de valores perdidos.

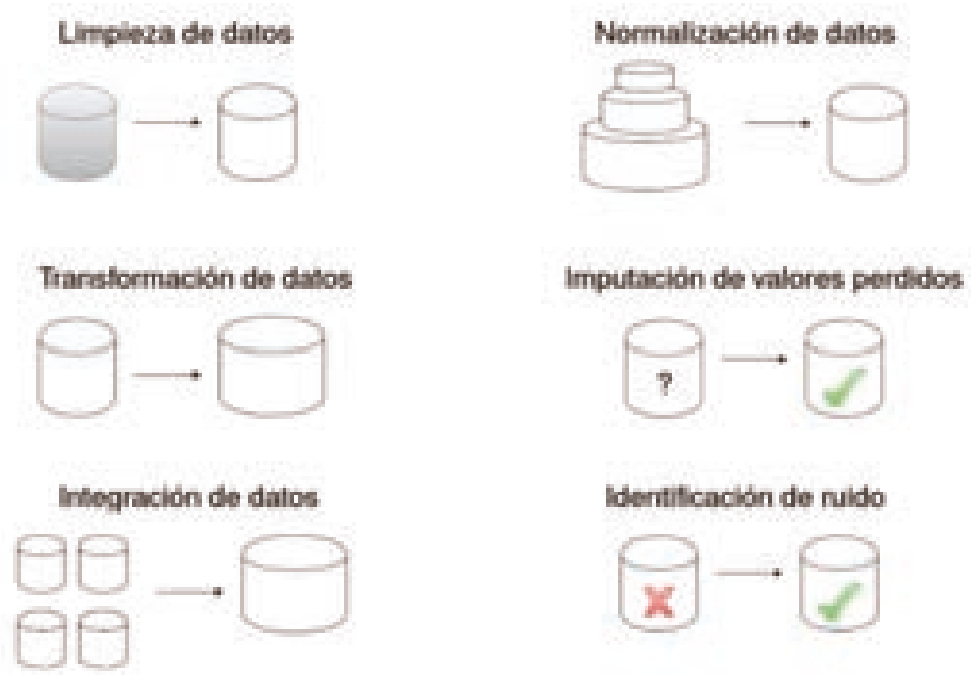

Figura 1.9. Etapas relacionadas con la preparación de los datos

Las técnicas de reducción de datos se orientan a obtener una representación reducida de los datos originales, manteniendo en la mayor medida posible la integridad y la información existente en los datos. Cuando el tiempo de ejecución de un algoritmo o el tamaño de los datos comienza a ser bastante elevado, para los algoritmos de extracción, estas técnicas deben ser aplicadas para obtener conjuntos de datos más pequeños y de calidad. En esta área las técnicas de reducción más relevantes son:

- **Selección de atributos (Feature Selection)**. El objetivo es reducir el número de atributos iniciales para reducir la complejidad a la hora de realizar el análisis.

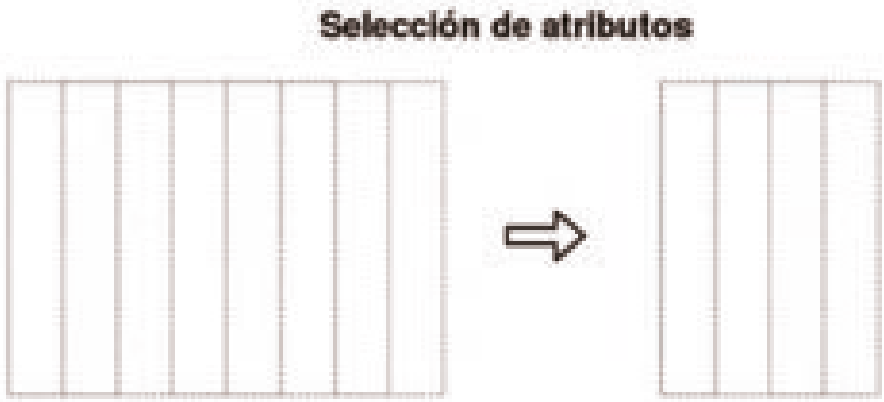

Figura 1.10. Selección de atributos

- **Selección de instancias (Instance Selection)**. El objetivo es reducir el número de filas que contiene nuestro dataset inicial.

Figura 1.11. Selección de instancias

- **Discretización**. El objetivo es convertir variables numéricas en variables categóricas que nos permitan realizar una clasificación de los posibles valores que puede tomar esa variable.

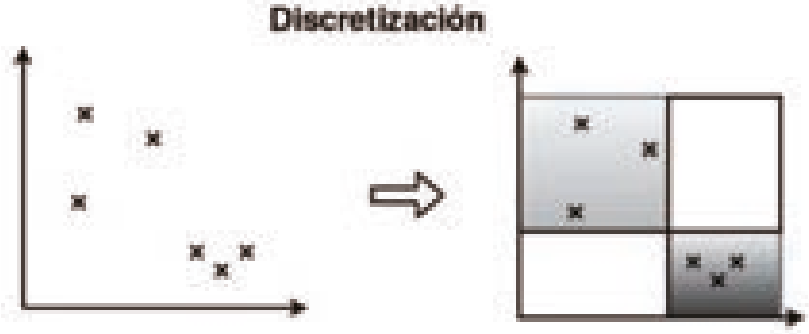

Figura 1.12. Discretización