LA TRANSFORMACIÓN DIGITAL DE LA EMPRES

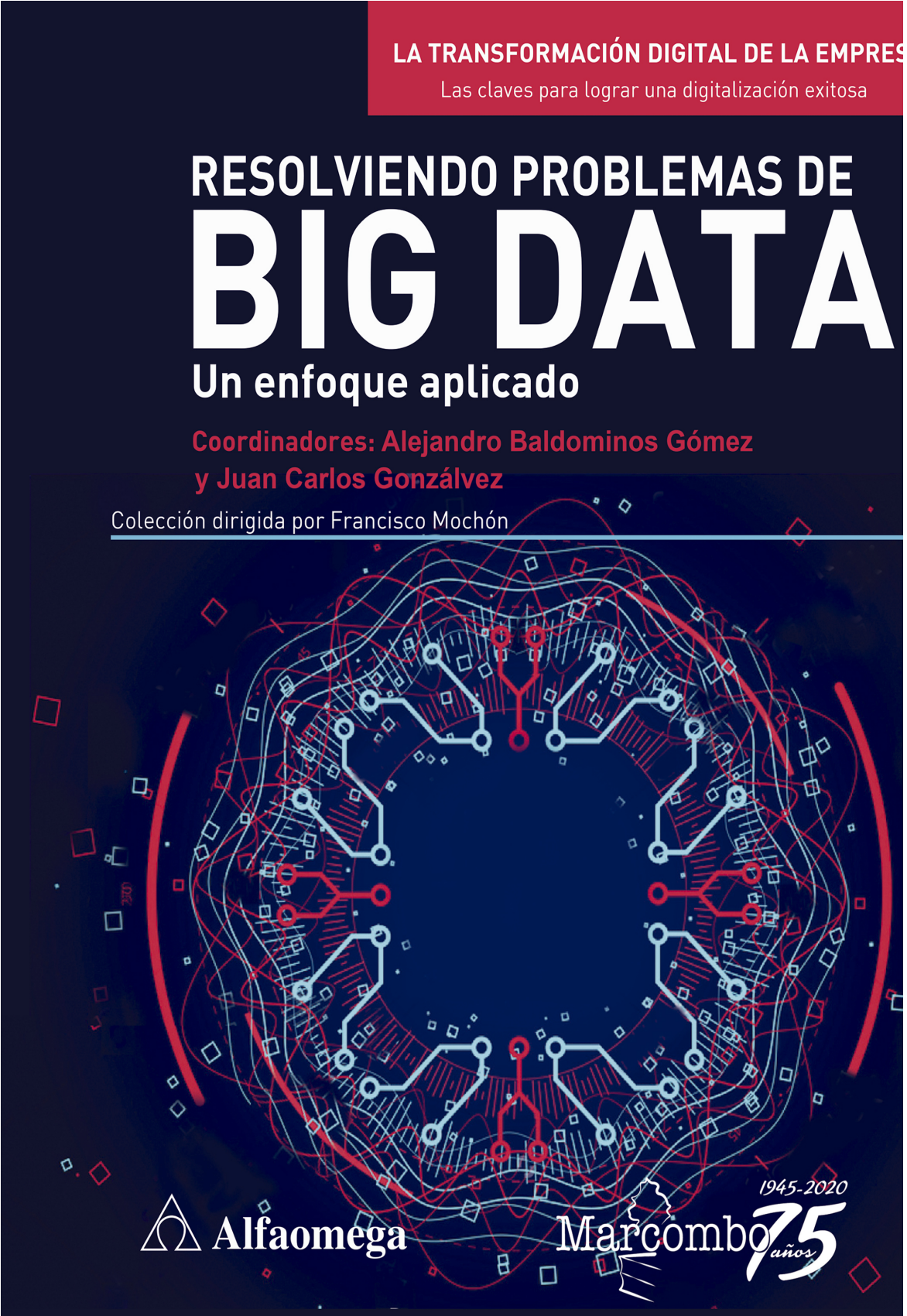
Las claves para lograr una digitalización exitosa

# RESOLVIENDO PROBLEMAS DE BIG DATA

## Un enfoque aplicado

Coordinadores: Alejandro Baldominos Gómez
y Juan Carlos Gonzálvez

Colección dirigida por Francisco Mochón



Alfaomega

Marcombo

1945-2020 75 años

# Resolviendo problemas de Big Data

Un enfoque aplicado

**Coordinadores:**
Alejandro Baldominos Gómez y Juan Carlos Gonzálvez

**Autores:**
Alejandro Baldominos Gómez
Juan Carlos Gonzálvez
Francisco Mochón Morcillo
Teófilo Redondo Pastor
Argelia Berenice Urbina Nájera
Rosa María Cantón Croda
Damián Emilio Gibaja Romero
Clara Ramón Lozano

# Resolviendo problemas de Big Data

Un enfoque aplicado

**Coordinadores:**
Alejandro Baldominos Gómez y Juan Carlos Gonzálvez

**Autores:**
Alejandro Baldominos Gómez
Juan Carlos Gonzálvez
Francisco Mochón Morcillo
Teófilo Redondo Pastor
Argelia Berenice Urbina Nájera
Rosa María Cantón Croda
Damián Emilio Gibaja Romero
Clara Ramón Lozano

Alfaomega

Marcombo 1945-2020 75 años

*Resolviendo problemas de Big Data*

Baldominos G· mez, Alejandro; Gonz§lvez, Juan Carlos;
Moch· n Morcillo, Francisco; Redondo Pastor, Te· filo;
Urbina N§jera, Argelia Berenice; Cant· n Croda, Rosa Mar²a;
Gibaja Romero, Dami§n Emilio; Ram· n Lozano, Clara


Primera edición: 2020
ISBN: 978-607-538-480-1

Primera edición: MARCOMBO, S.L. 2020



ISBN: 978-84-267-3213-2
D.L.: B 17881-2020

Impreso en Servicepoint
*Printed in Spain*

*Los autores dedicamos este libro a sus lectores, confiando en que los casos de uso que en él se presentan sirvan de guía para poder dar los primeros pasos en el mundo del Big Data, y animándoles a que una vez terminado, no dejen de seguir "caminando" por estos terrenos tan apasionantes. El mundo está lleno de datos y usted, lector, no puede dejar pasar la oportunidad de sacarles partido.*

# Acerca de la serie “La transformación digital de la empresa”

## ¿Qué se entiende por transformación digital?

Internet y las nuevas tecnologías están cambiando la forma en que nos relacionamos, trabajamos y aprendemos. Algunos identifican la transformación digital con el nuevo diseño de los negocios como fruto de integrar lo digital y lo físico. De hecho, la transformación digital conlleva la reorientación de toda la organización hacia un modelo eficaz de relación digital en lo referente a los procesos, los clientes y el modelo de negocio.

La transformación digital supone reinventar, mejorar la conectividad, optimizar los procesos de negocio, redefinir la experiencia del cliente y asumir la esencia del cambio. El nuevo contexto se caracteriza por el gran poder otorgado a los datos, la integración de los procesos, las aplicaciones y los sistemas en un todo armónico y por la prevalencia del talento. El talento y el conocimiento experto son un bien escaso y su papel es clave para llevar a cabo la transformación digital de los negocios.

## ¿Por qué es algo de lo que todo el mundo habla?

La transformación digital está de plena actualidad por las razones siguientes:

- La profundidad de los cambios asociados a la transformación digital es tan intensa que es comparable a la revolución industrial que arrancó en Inglaterra en la segunda mitad del siglo XVIII. De hecho, a la revolución digital se le denomina la cuarta revolución industrial.
- La velocidad del cambio asociado a la transformación digital es exponencial, mientras que las organizaciones y la sociedad tienden a cambiar a una velocidad lineal, lo que supone un importante reto a superar.

- La transformación digital se trata de un fenómeno mucho más amplio que algo ligado a las empresas de Internet; a afecta a la sociedad en general, a los distintos sectores y a prácticamente todas las empresas, si bien la magnitud del impacto y el plazo es distinto.

## ¿Qué interés tiene la serie "La transformación digital de los negocios" para Ud.?

Es una serie que aborda el análisis de la transformación digital de forma integral y con un enfoque eminentemente aplicado. En todos los libros de forma sistemática se recurre a ejemplos y casos de uso para ilustrar los contenidos.

## Títulos de la Serie

- eCommerce 360: Casos de éxito latinoamericanos.
- Emprendimiento en el entorno digital. El lanzamiento de una startup
- La transformación digital de la empresa
- Marketing Digital: Casos Latinoamericanos.
- Las claves del entorno mobile
- Resolviendo Problemas de Big Data. Un Enfoque Aplicado
- La Inteligencia Artificial y los Negocios: Casos de uso

# Acerca de los autores

**Alejandro Baldominos** es Doctor en Ciencia y Tecnología Informática por la Universidad Carlos III de Madrid (España), donde trabaja como investigador en el grupo de computación evolutiva, redes de neuronas e inteligencia artificial (EVANNAI). Su trayectoria investigadora está centrada en las aplicaciones de inteligencia artificial, habiendo publicado diversos artículos en revistas de reconocido prestigio y conferencias internacionales. En el año 2016, completó una estancia de investigación en el Computer Science and Artificial Intelligence Lab del MIT. Además, ha impartido docencia en diferentes asignaturas relacionadas con la programación de ordenadores, la ciencia de datos y la inteligencia artificial

**Juan Carlos Gonzálvez Cabañas** es Licenciado en Ciencias Químicas por la Universidad Autónoma de Madrid (UAM) y posee un MBA por el Instituto de Empresa (IE) Business School. Ha desempeñado distintos cargos directivos en empresas de ámbito internacional, en los sectores TMT (tecnología, media y telecomunicaciones) y entretenimiento (Internet, Juegos y Animación). Ha liderado y participado en la elaboración, diseño, desarrollo e implantación de productos y servicios de base tecnológica en los que uno de sus principales activos son los datos y su tratamiento inteligente para el negocio.

**Francisco Mochón**. Doctor en Economía por la Universidad Autónoma de Madrid y PhD en Economía por Indiana University (Becado Fulbright). Actualmente es catedrático de Análisis Económico de la UNED. Ha sido asesor del Ministerio de Economía y Hacienda de España, Director General de Política Financiera de la Junta de Andalucía, CEO de la empresa de investigación ESECA y Director General de Finanzas (CFO) del Grupo Telefónica. Ha sido Presidente del Consejo Social de la Universidad de Málaga y miembro del comité asesor de la U-TAD. Es patrono de la Fundación de software libre (FIDESOL). Ha publicado numerosos artículos de investigación y es autor de más de cincuenta libros sobre economía, finanzas y negocios. Desde hace unos años su investigación se ha centrado en dos campos: la economía digital y la economía de la felicidad en el entorno empresarial; habiendo publicado diverso artículos y libros en ambos campos. Ha sido el director del curso MOOC "Felicidad y Práctica empresarial".

**Teófilo Redondo** (Master en Lingüística Computacional – 1986, por la Universidad Complutense de Madrid - UCM) es consultor especializado en gestión de la Innovación en Ayming Consulting. Anteriormente fue Coordinador Tecnológico de Proyectos de Investigación en Zed Worldwide; Arquitecto de Tecnología y de Proyectos de Innovación en la Universidad Internacional de La Rioja (UNIR), donde también desarrolló actividad docente en el Master eLearning y Redes Sociales. Previamente desarrolló su carrera profesional en IBM cubriendo áreas como Computación en la Nube y Arquitecturas para Big Data, Soluciones de Arquitectura Empresarial (SAP, Oracle Solutions, Dassault Systemes) y Arquitecto SOA. Comenzó en la División de Investigación de IBM, en el Centro Científico UAM-IBM, con varios proyectos de Traducción Automática. Fue Visiting Scholar en la Universidad de Stanford (1987 – Linguistic Institute).

**Argelia Berenice Urbina Nájera**. Tiene el grado de Doctora en Planeación Estratégica y Dirección de Tecnología por la Universidad Popular Autónoma del Estado de Puebla (UPAEP), el grado de Maestra en Ciencias en Ingeniería de la Computación por la Universidad Autónoma de Tlaxcala, el grado de Maestra en Ciencias de la Educación por el Instituto de Estudios Universitarios y la Licenciatura en Ciencias de la Computación por la Benemérita Universidad Autónoma de Puebla. Pertenece al Sistema Nacional de Investigadores en el nivel Candidato. Sus líneas de investigación se enfocan en la tecnología educativa y la aplicación de aprendizaje automático en el ámbito educativo, salud y negocios. Actualmente es Profesora de Tiempo Completo de la Maestría en Ciencia de Datos e Inteligencia de Negocios ofertada por UPAEP.

**Rosa María Cantón Croda**. Doctora en Ciencias Computacionales por el Tecnológico de Monterrey, Ciudad de México, Maestra en Tecnologías de Información y Licenciada en Sistemas Computacionales Administrativos, ambas por el Tecnológico de Monterrey, Campus Monterrey. Ha desempeñado puesto como Controladora de Gestión de Procesos y Gerente de Sistemas en empresas de diversos giros. Cuenta con una amplia experiencia en el ámbito académico tanto en el Tecnológico de Monterrey, la Universidad Tecmilenio y actualmente, Decana de Posgrados de Ingenierías y Negocios en UPAEP. Ha escrito para congresos nacionales e internacionales, así como artículos en revistas arbitradas, capítulos de libro, reportes técnicos para diversas empresas en el área de Sistemas de Información y ha sido parte del comité de revisores de congresos. Tiene una amplia experiencia docente en el área de Bases de Datos, Administración de Proyectos y Sistemas de Información, labor por la que ha sido reconocida en varias ocasiones. Está certifica en PMBoK por Itera, en Psicología Positiva por la Universidad Tecmilenio, en Aprendizaje Basado en Proyectos por la Universidad de Aalborg en Dinamarca y en Ciencia de Datos por la Universidad de California en San Diego. Actualmente su línea de investigación es Ciencia de Datos e Inteligencia de Negocios.

**Damián Emilio Gibaja Romero**, es licenciado en Matemáticas por la Universidad Autónoma del Estado de México, Maestro y Doctor en Economía por El Colegio de México. Realizó estancias de investigación en la Universidad de Glasgow y en la Escuela de Economía de París. Ha participado en congresos nacionales e internacionales como la 25th Summer School in Economic Theory en la Universidad Hebrea de Jerusalén. La International Federation Operational Research Conference organizado por la Universidad de Quebec, y las Jornadas Latinoamericanas de Teoría de Económica en el Centro de Investigación en Matemáticas. Ha impartido clases en El Colegio de México y el Centro de Investigación y Docencia Económicas. Su línea de investigación es en Teoría Microeconómica, Diseño de Mecanismos (matching), Economía Matemática y Teoría de Juegos. Forma parte del claustro de profesores de los posgrados en Ingenierías y Negocios, donde actualmente se desempeña como director académico de matemáticas. Ha publicado artículos sobre la existencia y unicidad de soluciones en problemas estratégicos, y su aplicación en el diseño de mecanismos para la búsqueda de soluciones eficientes y justas.

**Clara Ramón Lozano** Es Ingeniera Biomédica por la Universidad Carlos III de Madrid (España) y tiene un máster en Ingeniería Biomecánica por l'École Polytechnique (Francia). Actualmente está haciendo un doctorado en el laboratorio de hidrodinámica de l'École Polytechnique (LadHyX) en el desarrollo de organs-on-chip.

Es fundadora de CEEIBIS, el Consejo Estatal de Estudiantes de Ingeniería Biomédica e Ingeniería de la Salud, y ha publicado artículos en revistas científicas de prestigio y conferencias internacionales, algunos de ellos orientados a la aplicación de inteligencia artificial al dominio biomédico. En 2017 fue premiada en el Certamen Universitario Arquímedes por su trayectoria de joven investigadora.

# Contenido

# Prólogo

Cuando hablamos de Big Data nos podemos referir a muchos conceptos distintos. Algunas personas lo emplean, por ejemplo, para referirse a los procesos que permiten extraer información de los datos en un entorno corporativo. Es el uso que se le da, por ejemplo, en la frase "En mi empresa hacemos Big Data". También se emplea, aunque con menos frecuencia, para referirse a las herramientas que se pueden emplear para ejecutar los mencionados procesos. Lo cierto es que a pesar de estos usos, en el sentido más purista del término hablar de Big Data es hablar de información. Y no de cualquier información: se trata de una cantidad de información tan voluminosa que el simple hecho de almacenarla plantea retos, y tratar de procesarla y estudiarla podría suponer un enorme desafío.

Por suerte, desde hace más de una década, científicos e ingenieros de compañías como Google o Yahoo!, que empezaban a darse cuenta de los problemas que planteaba el disponer de tales volúmenes de datos, comenzaron a desarrollar técnicas y herramientas para poder almacenar los datos y procesarlos de forma eficiente. En los años sucesivos, cada vez más empresas descubrieron el valor de extraer la mayor cantidad de datos posibles, y estos datos pueden surgir de muchas fuentes distintas: navegación de los usuarios con una plataforma de compras, interacción entre usuarios en una red social, etc. La principal ventaja es que durante estos años han ido apareciendo herramientas que han simplificado enormemente el trabajo con grandes cantidades de datos, reduciendo la brecha entre Big Data y datos en un sentido más tradicional. Algunas de las herramientas importantes en este sentido son aquellas que están disponibles en la nube, pues evitan a los usuarios tener que adquirir equipamiento especializado para lidiar con los datos, aumentando la productividad al poder comenzar el estudio de los datos desde el minuto uno.

En este libro, se introduce al lector en el uso de algunas de estas herramientas. No se trabajará con volúmenes de datos especialmente grandes, ya que se emplearán herramientas que disponen de una capa

gratuita, es decir, que el lector no deberá incurrir en costes, y generalmente esta capa gratuita impone una limitación en el tamaño de los datos a usar o en el tiempo de procesamiento.

No es el objetivo de este libro proporcionar un profundo conocimiento sobre Big Data, sino por el contrario dotar al lector de un cierto bagaje que le permita identificar algunos problemas frecuentes en Big Data, y proceder a su resolución. Es por tanto, un recurso ideal para aquellas personas que no deseen ejercer una profesión técnica especializada en el manejo de datos, sino adquirir cierta capacitación para comunicarse con un equipo técnico y así prototipar soluciones simplificadas a problemas reales.

Para lograr este objetivo, el libro parte de un enfoque absolutamente práctico. La única excepción la constituye el capítulo 1, donde se presentan a modo introductorio algunas de las claves del Big Data en entornos organizativos. El resto del libro está constituido por casos de uso, que ilustran un problema de tratamiento de datos y su resolución empleando herramientas accesibles. En algunos casos, se han introducido conceptos teóricos en estos capítulos para que el lector pueda entender mejor los pasos que debe seguir.

En particular, los tres primeros casos de uso plantean problemas de análisis de datos de diversa índole. El caso contenido en el capítulo 2 proporciona herramientas para realizar un análisis descriptivo de un conjunto de datos, para entender mejor su estructura y resumir la información importante allí escondida.

Por el contrario, los capítulos 3 y 4 se centran en un análisis más inteligente de los datos. Así, el capítulo 3 proporciona un ejemplo de problema de análisis predictivo, donde dado un histórico de datos queremos ser capaces de construir un modelo de forma automática que nos permita inferir información sobre datos futuros. Este tipo de problemas surge de forma natural y con una elevada frecuencia. Por ejemplo, en un entorno bancario, cabría preguntarse si un cliente de un préstamo hipotecario va a tener dificultades para pagarlo con intereses antes de concederlo. Alternativamente, en un entorno médico podría ser interesante conocer la probabilidad de que un paciente sufra en un futuro cercano un paro cardíaco, examinando sus constantes vitales y sus hábitos. Como se ha afirmado anteriormente, estos problemas resultan de gran interés y su resolución involucra en algunos casos la fusión de técnicas de Big Data con otras de una disciplina distinta pero interrelacionada: la Inteligencia Artificial.

El capítulo 4 se centra en un problema distinto, el de la detección de anomalías, donde lo que se pretende es localizar aquellos datos que

puedan constituir una anomalía dentro de un conjunto de datos. De nuevo, este tipo de problema puede resultar interesante en numerosos escenarios. Así, en un contexto financiero, podría ser clave detectar patrones anómalos en los hábitos de consumo y gasto de un usuario de una tarjeta de crédito, pues podría involucrar un uso no autorizado o robo. Este tipo de problemas (como el anterior) también se resuelve mediante el cruce de técnicas de Big Data con Inteligencia Artificial.

A continuación, los tres siguientes casos de uso se centran en otro tipo de cuestiones que cobran una especial relevancia: la visualización. La visualización es crítica, porque permite comunicar información de un modo gráfico, visual y entendible. En muchos casos, esta comunicación visual puede realizarse a terceras personas (un potencial cliente, un público en una conferencia, etc.), pero igualmente puede servir para que nosotros mismos seamos capaces de entender la estructura de los datos de forma mucho más clara que si nos limitáramos a ver su resumen estadístico.

En este sentido, el capítulo 5 plantea un contenido introductorio a la visualización de datos. Sin presentar un problema concreto que atacar, como ocurre en los otros casos de uso, el capítulo fusiona contenido teórico y práctico, describiendo las principales tendencias y técnicas en visualización de datos.

A continuación, el capítulo 6 presenta un caso de visualización de datos en la transformación de una industria. El objetivo de este caso práctico es que el lector sea capaz de diseñar un cuadro de mandos integral que permita comunicar informes sobre tendencias, complementado además con una visualización geográfica.

El último caso, contenido en el capítulo 7, presenta una herramienta para la construcción de gráficos de forma sencilla, buscando una aplicación mucho más ligera que la del caso anterior. En esta ocasión, lo que se pretende es que el lector conozca recursos para construir distintos tipos de visualizaciones, especialmente aquellas con contenido geográfico, de forma rápida, y así prototipar una visualización que comunique ideas sencillas sin la necesidad de un realizar un arduo trabajo.

Finalmente, el libro presenta algunas líneas conclusivas para que el lector pueda asimilar todo lo que ha venido estudiando a lo largo de los casos.

Los autores confiamos en que este libro sea útil para despertar el interés y la curiosidad de los lectores por el Big Data, y que sirva como