BIG DATA IN PRACTICE

"A comprehensive compendium of why, how, and to what effects Big Data analytics are used in today's world"
James Kobielus, Big Data Evangelist at IBM

HOW 45 SUCCESSFUL COMPANIES USED BIG DATA ANALYTICS TO DELIVER EXTRAORDINARY RESULTS

BERNARD MARR

WILEY

"Amazing. That was my first word, when I started reading this book. Fascinating was the next. Amazing, because once again, Bernard masterfully takes a complex subject, and translates it into something anyone can understand. Fascinating because the detailed real-life customer examples immediately inspired me to think about my own customers and partners, and how they could emulate the success of these companies. Bernard's book is a must have for all Big Data practitioners and Big Data hopefuls!"

Shawn Ahmed, Senior Director, Business Analytics and IoT at Splunk

"Finally a book that stops talking theory and starts talking facts. Providing real-life and tangible insights for practices, processes, technology and *teams* that support Big Data, across a portfolio of organizations and industries. We often think Big Data is big business and big cost, however some of the most interesting examples show how small businesses can use smart data to make a real difference. The businesses in the book illustrate how Big Data is fundamentally about the customer, and generating a data-driven customer strategy that influences both staff and customers at every touch point of the customer journey."

Adrian Clowes, Head of Data and Analytics at Center Parcs UK

"*Big Data in Practice* by Bernard Marr is the most complete book on the Big Data and analytics ecosystem. The many real-life examples make it equally relevant for the novice as well as experienced data scientists."

Fouad Bendris, Business Technologist, Big Data Lead at Hewlett Packard Enterprise

"Bernard Marr is one of the leading authors in the domain of Big Data. Throughout *Big Data in Practice* Marr generously shares some of his keen insights into the practical value delivered to a huge range of different businesses from their Big Data initiatives. This fascinating book provides excellent clues as to the secret sauce required in order to successfully deliver competitive advantage through Big Data analytics. The logical structure of the book means that it is as easy to consume in one sitting as it is to pick up from time to time. This is a must-read for any Big Data sceptics or business leaders looking for inspiration."

Will Cashman, Head of Customer Analytics at AIB

"The business of business is now data! Bernard Marr's book delivers concrete, valuable, and diverse insights on Big Data use cases, success stories, and lessons learned from numerous business domains. After diving into this book, you will have all the knowledge you need to crush the Big Data hype machine, to soar to new heights of data analytics ROI, and to gain competitive advantage from the data within your organization."

Kirk Borne, Principal Data Scientist at Booz Allen Hamilton, USA

"Big Data is disrupting every aspect of business. You're holding a book that provides powerful examples of how companies strive to defy outmoded business models and design new ones with Big Data in mind."

Henrik von Scheel, Google Advisory Board Member

"Bernard Marr provides a comprehensive overview of how far Big Data has come in past years. With inspiring examples he clearly shows how large, and small, organizations can benefit from Big Data. This book is a must-read for any organization that wants to be a data-driven business."

Mark van Rijmenam, Author *Think Bigger* and Founder of Datafloq

"This is one of those unique business books that is as useful as it is interesting. Bernard has provided us with a unique, inside look at how leading organizations are leveraging new technology to deliver real value out of data and completely transforming the way we think, work, and live."

Stuart Frankel, CEO at Narrative Science Inc.

"Big Data can be a confusing subject for even sophisticated data analysts. Bernard has done a fantastic job of illustrating the true business benefits of Big Data. In this book you find out succinctly how leading companies are getting real value from Big Data – highly recommended read!'

Arthur Lee, Vice President of Qlik Analytics at Qlik

"If you are searching for the missing link between Big Data technology and achieving business value – look no further! From the world of science to entertainment, Bernard Marr delivers it – and, importantly, shares with us the recipes for success."

Achim Granzen, Chief Technologist Analytics at Hewlett Packard Enterprise

"A comprehensive compendium of why, how, and to what effects Big Data analytics are used in today's world."

James Kobielus, Big Data Evangelist at IBM

"A treasure chest of Big Data use cases."

Stefan Groschupf, CEO at Datameer, Inc.

BIG DATA IN PRACTICE

BIG DATA IN PRACTICE

HOW 45 SUCCESSFUL COMPANIES USED BIG DATA ANALYTICS TO DELIVER EXTRAORDINARY RESULTS

BERNARD MARR

WILEY

This edition first published 2016

Registered office
John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data is available

A catalogue record for this book is available from the British Library.

ISBN 978-1-119-23138-7 (hbk) ISBN 978-1-119-23139-4 (ebk)
ISBN 978-1-119-23141-7 (ebk) ISBN 978-1-119-27882-5 (ebk)

Cover Design: Wiley
Cover Image: © vs148/Shutterstock

Set in 11/14pt MinionPro Light by Aptara Inc., New Delhi, India
Printed in Great Britain by TJ International Ltd, Padstow, Cornwall, UK

This book is dedicated to the people who mean most to me: My wife Claire and our three children Sophia, James and Oliver.

CONTENTS

INTRODUCTION

We are witnessing a movement that will completely transform any part of business and society. The word we have given to this movement is Big Data and it will change everything, from the way banks and shops operate to the way we treat cancer and protect our world from terrorism. No matter what job you are in and no matter what industry you work in, Big Data will transform it.

Some people believe that Big Data is just a big fad that will go away if they ignore it for long enough. It won't! The hype around Big Data and the name may disappear (which wouldn't be a great loss), but the phenomenon will stay and only gather momentum. What we call Big Data today will simply become the new normal in a few years' time, when all businesses and government organizations use large volumes of data to improve what they do and how they do it.

I work every day with companies and government organizations on Big Data projects and thought it would be a good idea to share how Big Data is used today, across lots of different industries, among big and small companies, to deliver real value. But first things first, let's just look at what Big Data actually means.

What Is Big Data?

Big Data basically refers to the fact that we can now collect and analyse data in ways that was simply impossible even a few years ago. There

are two things that are fuelling this Big Data movement: the fact we have more data on anything and our improved ability to store and analyse any data.

More Data On Everything

Everything we do in our increasingly digitized world leaves a data trail. This means the amount of data available is literally exploding. We have created more data in the past two years than in the entire previous history of mankind. By 2020, it is predicted that about 1.7 megabytes of new data will be created every second, for every human being on the planet. This data is coming not just from the tens of millions of messages and emails we send each other every second via email, WhatsApp, Facebook, Twitter, etc. but also from the one trillion digital photos we take each year and the increasing amounts of video data we generate (every single minute we currently upload about 300 hours of new video to YouTube and we share almost three million videos on Facebook). On top of that, we have data from all the sensors we are now surrounded by. The latest smartphones have sensors to tell where we are (GPS), how fast we are moving (accelerometer), what the weather is like around us (barometer), what force we are using to press the touch screen (touch sensor) and much more. By 2020, we will have over six billion smartphones in the world – all full of sensors that collect data. But not only our phones are getting smart, we now have smart TVs, smart watches, smart meters, smart kettles, fridges, tennis rackets and even smart light bulbs. In fact, by 2020, we will have over 50 billion devices that are connected to the Internet. All this means that the amount of data and the variety of data (from sensor data, to text and video) in the world will grow to unimaginable levels.

Ability To Analyse Everything

All this Big Data is worth very little unless we are able to turn it into insights. In order to do that we need to capture and analyse the data.

In the past, there were limitations to the amount of data that could be stored in databases – the more data there was, the slower the system became. This can now be overcome with new techniques that allow us to store and analyse data across different databases, in distributed locations, connected via networks. So-called distributed computing means huge amounts of data can be stored (in little bits across lots of databases) and analysed by sharing the analysis between different servers (each performing a small part of the analysis).

Google were instrumental in developing distributed computing technology, enabling them to search the Internet. Today, about 1000 computers are involved in answering a single search query, which takes no more than 0.2 seconds to complete. We currently search 3.5 billion times a day on Google alone.

Distributed computing tools such as Hadoop manage the storage and analysis of Big Data across connected databases and servers. What's more, Big Data storage and analysis technology is now available to rent in a software-as-a-service (SAAS) model, which makes Big Data analytics accessible to anyone, even those with low budgets and limited IT support.

Finally, we are seeing amazing advancements in the way we can analyse data. Algorithms can now look at photos, identify who is on them and then search the Internet for other pictures of that person. Algorithms can now understand spoken words, translate them into written text and analyse this text for content, meaning and sentiment (e.g. are we saying nice things or not-so-nice things?). More and more advanced algorithms emerge every day to help us understand our world and predict the future. Couple all this with machine learning and artificial intelligence (the ability of algorithms to learn and make decisions independently) and you can hopefully see that the developments and opportunities here are very exciting and evolving very quickly.

Big Data Opportunities

With this book I wanted to showcase the current state of the art in Big Data and provide an overview of how companies and organizations across all different industries are using Big Data to deliver value in diverse areas. You will see I have covered areas including how retailers (both traditional bricks 'n' mortar companies as well as online ones) use Big Data to predict trends and consumer behaviours, how governments are using Big Data to foil terrorist plots, even how a tiny family butcher or a zoo use Big Data to improve performance, as well as the use of Big Data in cities, telecoms, sports, gambling, fashion, manufacturing, research, motor racing, video gaming and everything in between.

Instead of putting their heads in the sand or getting lost in this startling new world of Big Data, the companies I have featured here have figured out smart ways to use data in order to deliver strategic value. In my previous book, *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance* (also published by Wiley), I go into more detail on how any company can figure out how to use Big Data to deliver value.

I am convinced that Big Data, unlike any other trend at the moment, will affect everyone and everything we do. You can read this book cover to cover for a complete overview of current Big Data use cases or you can use it as a reference book and dive in and out of the areas you find most interesting or are relevant to you or your clients. I hope you enjoy it!

1
WALMART

How Big Data Is Used To Drive Supermarket Performance

Background

Walmart are the largest retailer in the world and the world's largest company by revenue, with over two million employees and 20,000 stores in 28 countries.

With operations on this scale it's no surprise that they have long seen the value in data analytics. In 2004, when Hurricane Sandy hit the US, they found that unexpected insights could come to light when data was studied as a whole, rather than as isolated individual sets. Attempting to forecast demand for emergency supplies in the face of the approaching Hurricane Sandy, CIO Linda Dillman turned up some surprising statistics. As well as flashlights and emergency equipment, expected bad weather had led to an upsurge in sales of strawberry Pop Tarts in several other locations. Extra supplies of these were dispatched to stores in Hurricane Frances's path in 2012, and sold extremely well.

Walmart have grown their Big Data and analytics department considerably since then, continuously staying on the cutting edge. In 2015, the company announced they were in the process of creating

the world's largest private data cloud, to enable the processing of 2.5 petabytes of information every hour.

What Problem Is Big Data Helping To Solve?

Supermarkets sell millions of products to millions of people every day. It's a fiercely competitive industry which a large proportion of people living in the developed world count on to provide them with day-to-day essentials. Supermarkets compete not just on price but also on customer service and, vitally, convenience. Having the right products in the right place at the right time, so the right people can buy them, presents huge logistical problems. Products have to be efficiently priced to the cent, to stay competitive. And if customers find they can't get everything they need under one roof, they will look elsewhere for somewhere to shop that is a better fit for their busy schedule.

How Is Big Data Used In Practice?

In 2011, with a growing awareness of how data could be used to understand their customers' needs and provide them with the products they wanted to buy, Walmart established @WalmartLabs and their Fast Big Data Team to research and deploy new data-led initiatives across the business.

The culmination of this strategy was referred to as the Data Café – a state-of-the-art analytics hub at their Bentonville, Arkansas headquarters. At the Café, the analytics team can monitor 200 streams of internal and external data in real time, including a 40-petabyte database of all the sales transactions in the previous weeks.

Timely analysis of real-time data is seen as key to driving business performance – as Walmart Senior Statistical Analyst Naveen Peddamail tells me: "If you can't get insights until you've analysed your sales for a week or a month, then you've lost sales within that time.

"Our goal is always to get information to our business partners as fast as we can, so they can take action and cut down the turnaround time. It is proactive and reactive analytics."

Teams from any part of the business are invited to visit the Café with their data problems, and work with the analysts to devise a solution. There is also a system which monitors performance indicators across the company and triggers automated alerts when they hit a certain level – inviting the teams responsible for them to talk to the data team about possible solutions.

Peddamail gives an example of a grocery team struggling to understand why sales of a particular produce were unexpectedly declining. Once their data was in the hands of the Café analysts, it was established very quickly that the decline was directly attributable to a pricing error. The error was immediately rectified and sales recovered within days.

Sales across different stores in different geographical areas can also be monitored in real-time. One Halloween, Peddamail recalls, sales figures of novelty cookies were being monitored, when analysts saw that there were several locations where they weren't selling at all. This enabled them to trigger an alert to the merchandizing teams responsible for those stores, who quickly realized that the products hadn't even been put on the shelves. Not exactly a complex algorithm, but it wouldn't have been possible without real-time analytics.

Another initiative is Walmart's Social Genome Project, which monitors public social media conversations and attempts to predict what products people will buy based on their conversations. They also have the Shopycat service, which predicts how people's shopping habits are influenced by their friends (using social media data again) and have developed their own search engine, named Polaris, to allow them to analyse search terms entered by customers on their websites.

What Were The Results?

Walmart tell me that the Data Café system has led to a reduction in the time it takes from a problem being spotted in the numbers to a solution being proposed from an average of two to three weeks down to around 20 minutes.

What Data Was Used?

The Data Café uses a constantly refreshed database consisting of 200 billion rows of transactional data – and that only represents the most recent few weeks of business!

On top of that it pulls in data from 200 other sources, including meteorological data, economic data, telecoms data, social media data, gas prices and a database of events taking place in the vicinity of Walmart stores.

What Are The Technical Details?

Walmart's real-time transactional database consists of 40 petabytes of data. Huge though this volume of transactional data is, it only includes from the most recent weeks' data, as this is where the value, as far as real-time analysis goes, is to be found. Data from across the chain's stores, online divisions and corporate units are stored centrally on Hadoop (a distributed data storage and data management system).

CTO Jeremy King has described the approach as "data democracy" as the aim is to make it available to anyone in the business who can make use of it. At some point after the adoption of distributed Hadoop framework in 2011, analysts became concerned that the volume was growing at a rate that could hamper their ability to analyse it. As a result, a policy of "intelligently managing" data collection was adopted which involved setting up several systems designed to refine and categorize the data before it was stored. Other technologies in use

include Spark and Cassandra, and languages including R and SAS are used to develop analytical applications.

Any Challenges That Had To Be Overcome?

With an analytics operation as ambitious as the one planned by Walmart, the rapid expansion required a large intake of new staff, and finding the right people with the right skills proved difficult. This problem is far from restricted to Walmart: a recent survey by researchers Gartner found that more than half of businesses feel their ability to carry out Big Data analytics is hampered by difficulty in hiring the appropriate talent.

One of the approaches Walmart took to solving this was to turn to crowdsourced data science competition website Kaggle – which I profile in Chapter 44.[1]

Kaggle set users of the website a challenge involving predicting how promotional and seasonal events such as stock-clearance sales and holidays would influence sales of a number of different products. Those who came up with models that most closely matched the real-life data gathered by Walmart were invited to apply for positions on the data science team. In fact, one of those who found himself working for Walmart after taking part in the competition was Naveen Peddamail, whose thoughts I have included in this chapter.

Once a new analyst starts at Walmart, they are put through their Analytics Rotation Program. This sees them moved through each different team with responsibility for analytical work, to allow them to gain a broad overview of how analytics is used across the business.

Walmart's senior recruiter for its Information Systems Operation, Mandar Thakur, told me: "The Kaggle competition created a buzz about Walmart and our analytics organization. People always knew

that Walmart generates and has a lot of data, but the best part was that this let people see how we are using it strategically."

What Are The Key Learning Points And Takeaways?

Supermarkets are big, fast, constantly changing businesses that are complex organisms consisting of many individual subsystems. This makes them an ideal business in which to apply Big Data analytics.

Success in business is driven by competition. Walmart have always taken a lead in data-driven initiatives, such as loyalty and reward programmes, and by wholeheartedly committing themselves to the latest advances in real-time, responsive analytics they have shown they plan to remain competitive.

Bricks 'n' mortar retail may be seen as "low tech" – almost Stone Age, in fact – compared to their flashy, online rivals but Walmart have shown that cutting-edge Big Data is just as relevant to them as it is to Amazon or Alibaba.[2] Despite the seemingly more convenient options on offer, it appears that customers, whether through habit or preference, are still willing to get in their cars and travel to shops to buy things in person. This means there is still a huge market out there for the taking, and businesses that make best use of analytics in order to drive efficiency and improve their customers' experience are set to prosper.

REFERENCES AND FURTHER READING

1. Kaggle (2015) Predict how sales of weather-sensitive products are affected by snow and rain, https://www.kaggle.com/c/walmart-recruiting-sales-in-stormy-weather, accessed 5 January 2016.
2. Walmart (2015) When data met retail: A #lovedata story, http://careersblog.walmart.com/when-data-met-retail-a-lovedata-story/, accessed 5 January 2016.

2
CERN

Unravelling The Secrets Of The Universe With Big Data

Background

CERN are the international scientific research organization that operate the Large Hadron Collider (LHC), humanity's biggest and most advanced physics experiment. The colliders, encased in 17 miles of tunnels buried 600 feet below the surface of Switzerland and France, aim to simulate conditions in the universe milliseconds following the Big Bang. This allows physicists to search for elusive theoretical particles, such as the Higgs boson, which could give us unprecedented insight into the composition of the universe.

CERN's projects, such as the LHC, would not be possible if it weren't for the Internet and Big Data – in fact, the Internet was originally created at CERN in the 1990s. Tim Berners-Lee, the man often referred to as the "father of the Internet", developed the hypertext protocol which holds together the World Wide Web while at CERN. Its original purpose was to facilitate communication between researchers around the globe.

The LHC alone generates around 30 petabytes of information per year – 15 trillion pages of printed text, enough to fill 600 million filling cabinets – clearly Big Data by anyone's standards!

In 2013, CERN announced that the Higgs boson had been found. Many scientists have taken this as proof that the standard model of particle physics is correct. This confirms that much of what we think we know about the workings of the universe on a subatomic level is essentially right, although there are still many mysteries remaining, particularly involving gravity and dark matter.

What Problem Is Big Data Helping To Solve?

The collisions monitored in the LHC happen very quickly, and the resulting subatomic "debris" containing the elusive, sought-after particles exists for only a few millionths of a second before they decay. The exact conditions that cause the release of the particles which CERN are looking for only occur under very precise conditions, and as a result many hundreds of millions of collisions have to be monitored and recorded every second in the hope that the sensors will pick them up.

The LHC's sensors record hundreds of millions of collisions between particles, some of which achieve speeds of just a fraction under the speed of light as they are accelerated around the collider. This generates a massive amount of data and requires very sensitive and precise equipment to measure and record the results.

How Is Big Data Used In Practice?

The LHC is used in four main experiments, involving around 8000 analysts across the globe. They use the data to search for elusive theoretical particles and probe for the answers to questions involving antimatter, dark matter and extra dimensions in time and space.

Data is collected by sensors inside the collider that monitor hundreds of millions of particle collisions every second. The sensors pick up light, so they are essentially cameras, with a 100-megapixel resolution capable of capturing images at incredibly high speeds.

This data is then analysed by algorithms that are tuned to pick up the telltale energy signatures left behind by the appearance and disappearance of the exotic particles CERN are searching for.

The algorithms compare the resulting images with theoretical data explaining how we believe the target particles, such as the Higgs boson, will act. If the results match, it is evidence the sensors have found the target particles.

What Were The Results?

In 2013, CERN scientists announced that they believed they had observed and recorded the existence of the Higgs boson. This was a huge leap forward for science as the existence of the particle had been theorized for decades but could not be proven until technology was developed on this scale.

The discovery has given scientists unprecedented insight into the fundamental structure of the universe and the complex relationships between the fundamental particles that everything we see, experience and interact with is built from.

Apart from the LHC, CERN has existed since the 1950s and has been responsible for a great many scientific breakthroughs with earlier experiments, and many world-leading scientists have made their name through their work with the organization.

What Data Was Used?

Primarily, the LHC gathers data using light sensors to record the collision, and fallout, from protons accelerated to 99.9% of the speed of light. Sensors inside the colliders pick up light energy emitted during the collisions and from the decay of the resulting particles, and convert it into data which can be analysed by computer algorithms.

Much of this data, being essentially photographs, is unstructured. Algorithms transform light patterns recorded by the sensors into mathematical data. Theoretical data – ideas about how we think the particles being hunted will act – is matched against the sensor data to determine what has been captured on camera.

What Are The Technical Details?

The Worldwide LHC Computing Grid is the world's largest distributed computing network, spanning 170 computing centres in 35 different countries. To develop distributed systems capable of analysing 30 petabytes of information per year, CERN instigated the openlab project, in collaboration with data experts at companies including Oracle, Intel and Siemens. The network consists of over 200,000 cores and 15 petabytes of disk space.

The 300 gigabytes per second of data provided by the seven CERN sensors is eventually whittled down to 300 megabytes per second of "useful" data, which constitutes the product's raw output. This data is made available as a real-time stream to academic institutions partnered with CERN.

CERN have developed methods of adding extra computing power on the fly to increase the processing output of the grid without taking it offline, in times of spikes in demand for computational power.

Any Challenges That Had To Be Overcome?

The LHC gathers incredibly vast amounts of data, very quickly. No organization on earth has the computing power and resources necessary to analyse that data in a timely fashion. To deal with this, CERN turned to distributed computing.

They had already been using distributed computed for some time. In fact, the Internet as we know it today was initially built to save

scientists from having to travel to Geneva whenever they wanted to analyse results of CERN's earlier experiments.

For the LHC, CERN created the LHC Distributed Computing Grid, which comprises 170 computer centres in 35 countries. Many of these are private computing centres operated by the academic and commercial organizations partnered with CERN.

This parallel, distributed use of computer processing power means far more calculations per second can be carried out than even the world's most powerful supercomputers could manage alone.

What Are The Key Learning Points And Takeaways?

The groundbreaking work carried out by CERN, which has greatly improved our knowledge of how the universe works, would not be possible without Big Data and analytics.

CERN and Big Data have evolved together: CERN was one of the primary catalysts in the development of the Internet which brought about the Big Data age we live in today.

Distributed computing makes it possible to carry out tasks that are far beyond the capabilities of any one organization to complete alone.

REFERENCES AND FURTHER READING

Purcell, A. (2013) CERN on preparing for tomorrow's big data, http://home.web.cern.ch/about/updates/2013/10/preparing-tomorrows-big-data

Darrow, B. (2013) Attacking CERN's big data problem, https://gigaom.com/2013/09/18/attacking-cerns-big-data-problem/

O'Luanaigh, C. (2013) Exploration on the big data frontier, http://home.web.cern.ch/students-educators/updates/2013/05/exploration-big-data-frontier

Smith, T. (2015) Video on CERN's big data, https://www.youtube.com/watch?v=j-0cUmUyb-Y