



BIG DATA IN PRACTICE

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

James Kobielski, Big Data Evangelist at IBM

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

BERNARD
MARR

WILEY

BIG DATA IN PRACTICE

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

BERNARD MARR

WILEY

This edition first published 2016

© 2016 Bernard Marr

Registered office

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

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book and on its cover are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher and the book are not associated with any product or vendor mentioned in this book. None of the companies referenced within the book have endorsed the book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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	1
1	Walmart: How Big Data Is Used To Drive Supermarket Performance	5
2	CERN: Unravelling The Secrets Of The Universe With Big Data	11
3	Netflix: How Netflix Used Big Data To Give Us The Programmes We Want	17
4	Rolls-Royce: How Big Data Is Used To Drive Success In Manufacturing	25
5	Shell: How Big Oil Uses Big Data	31
6	Apixio: How Big Data Is Transforming Healthcare	37
7	Lotus F1 Team: How Big Data Is Essential To The Success Of Motorsport Teams	45
8	Pendleton & Son Butchers: Big Data For Small Business	51
9	US Olympic Women's Cycling Team: How Big Data Analytics Is Used To Optimize Athletes' Performance	57
10	ZSL: Big Data In The Zoo And To Protect Animals	63
11	Facebook: How Facebook Use Big Data To Understand Customers	69
12	John Deere: How Big Data Can Be Applied On Farms	75
13	Royal Bank of Scotland: Using Big Data To Make Customer Service More Personal	81
14	LinkedIn: How Big Data Is Used To Fuel Social Media Success	87
15	Microsoft: Bringing Big Data To The Masses	95
16	Acxiom: Fuelling Marketing With Big Data	103

CONTENTS

17	US Immigration And Customs: How Big Data Is Used To Keep Passengers Safe And Prevent Terrorism	111
18	Nest: Bringing The Internet of Things Into The Home	117
19	GE: How Big Data Is Fuelling The Industrial Internet	125
20	Etsy: How Big Data Is Used In A Crafty Way	131
21	Narrative Science: How Big Data Is Used To Tell Stories	137
22	BBC: How Big Data Is Used In The Media	143
23	Milton Keynes: How Big Data Is Used To Create Smarter Cities	149
24	Palantir: How Big Data Is Used To Help The CIA And To Detect Bombs In Afghanistan	157
25	Airbnb: How Big Data Is Used To Disrupt The Hospitality Industry	163
26	Sprint: Profiling Audiences Using Mobile Network Data	169
27	Dickey's Barbecue Pit: How Big Data Is Used To Gain Performance Insights Into One Of America's Most Successful Restaurant Chains	175
28	Caesars: Big Data At The Casino	181
29	Fitbit: Big Data In The Personal Fitness Arena	189
30	Ralph Lauren: Big Data In The Fashion Industry	195
31	Zynga: Big Data In The Gaming Industry	199
32	Autodesk: How Big Data Is Transforming The Software Industry	205
33	Walt Disney Parks and Resorts: How Big Data Is Transforming Our Family Holidays	211
34	Experian: Using Big Data To Make Lending Decisions And To Crack Down On Identity Fraud	217
35	Transport for London: How Big Data Is Used To Improve And Manage Public Transport In London	223
36	The US Government: Using Big Data To Run A Country	229
37	IBM Watson: Teaching Computers To Understand And Learn	237
38	Google: How Big Data Is At The Heart Of Google's Business Model	243

CONTENTS

39	Terra Seismic: Using Big Data To Predict Earthquakes	251
40	Apple: How Big Data Is At The Centre Of Their Business	255
41	Twitter: How Twitter And IBM Deliver Customer Insights From Big Data	261
42	Uber: How Big Data Is At The Centre Of Uber's Transportation Business	267
43	Electronic Arts: Big Data In Video Gaming	273
44	Kaggle: Crowdsourcing Your Data Scientist	281
45	Amazon: How Predictive Analytics Are Used To Get A 360-Degree View Of Consumers	287
	<i>Final Thoughts</i>	293
	<i>About the Author</i>	297
	<i>Acknowledgements</i>	299
	<i>Index</i>	301

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.

Peddemail 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.