BIG DATA CASE STUDY COLLECTION

'Data and analytics power everything that we do. This book is the go-to-guide on data for 2015.'
– Henrik von Scheel, Advisory Board Member at Google

7 AMAZING COMPANIES THAT REALLY GET BIG DATA

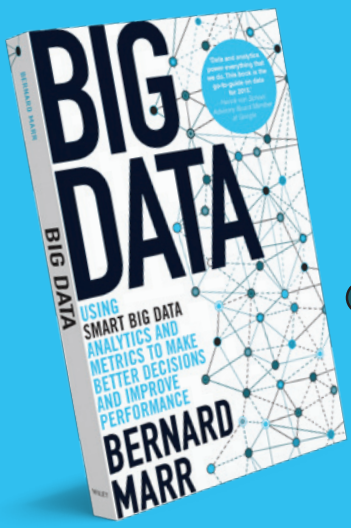
BERNARD MARR

WILEY

Big Data is a big thing and this case study collection will give you a good overview of how some companies really leverage big data to drive business performance. They range from industry giants like Google, Amazon, Facebook, GE, and Microsoft, to smaller businesses which have put big data at the centre of their business model, like Kaggle and Cornerstone.

This case study collection is based on articles published by Bernard Marr on his LinkedIn Influencer blog.

BROUGHT TO YOU BY THE BESTSELLING AUTHOR OF...



1

Google

Big data and big business go hand in hand – this is the first in a series where I will examine the different uses that the world's leading corporations are making of the endless amount of digital information the world is producing every day.

Google has not only significantly influenced the way we can now analyse big data (think MapReduce, BigQuery, etc.) – but they are probably more responsible than anyone else for making it part of our everyday lives. I believe that many of the innovative things Google is doing today, most companies will do in years to come.

Many people, particularly those who didn't get online until this century had started, will have had their first direct experience of manipulating big data through Google. Although these days Google's big data innovation goes well beyond basic search, it's still their core business. They process 3.5 billion requests per day, and each request queries a database of 20 billion web pages.

This is refreshed daily, as Google's bots crawl the web, copying down what they see and taking it back to be stored in Google's index database. What pushed Google in front of other search engines has been its ability to analyse wider data sets for their search.

Initially it was PageRank which included information about sites that linked to a particular site in the index, to help take a measure of that site's importance in the grand scheme of things. Previously leading search engines worked almost entirely on the principle of matching relevant keywords in the search query to sites containing those words. PageRank revolutionized search by incorporating other elements alongside keyword analysis.

Their aim has always been to make as much of the world's information available to as many people as possible (and get rich trying, of course…) and the way Google search works has been constantly revised and updated to keep up with this mission.

Moving further away from keyword-based search and towards semantic search is the current aim. This involves analysing not just the "objects" (words) in the query, but the connection between them, to determine what it means as accurately as possible.

To this end, Google throws a whole heap of other information into the mix. Starting in 2007 it launched Universal Search, which pulls in data from hundreds of sources including language databases, weather forecasts and historical data, financial data, travel information, currency exchange rates, sports statistics and a database of mathematical functions.

It continued to evolve in 2012 into the Knowledge Graph, which

displays information on the subject of the search from a wide range of resources directly into the search results.

It then mixes what it knows about you from your previous search history (if you are signed in), which can include information about your location, as well as data from your Google+ profile and Gmail messages, to come up with its best guess at what you are looking for.

The ultimate aim is undoubtedly to build the kind of machine we have become used to seeing in science fiction for decades – a computer which you can have a conversation with in your native tongue, and which will answer you with precisely the information you want.

Search is by no means all of what Google does, though. After all, it's free, right? And Google is one of the most profitable businesses on the planet. That profit comes from what it gets in return for its searches – information about you.

Google builds up vast amounts of data about the people using it. Essentially it then matches up companies with potential customers, through its Adsense algorithm. The companies pay handsomely for these introductions, which appear as adverts in the customers' browsers.

In 2010 it launched BigQuery, its commercial service for allowing companies to store and analyse big data sets on its cloud platforms. Companies pay for the storage space and computer time taken in running the queries.

Another big data project Google is working on is the self-driving car. Using and generating massive amounts of data from sensors,

cameras, tracking devices and coupling this with on-board and real-time data analysis from Google Maps, Streetview and other sources allows the Google car to safely drive on the roads without any input from a human driver.

Perhaps the most astounding use Google have found for their enormous data though, is predicting the future.

In 2008 the company published a paper in the science journal *Nature* claiming that their technology had the capability to detect outbreaks of flu with more accuracy than current medical techniques for detecting the spread of epidemics.

The results were controversial – debate continues over the accuracy of the predictions. But the incident unveiled the possibility of "crowd prediction", which in my opinion is likely to be a reality in the future as analytics becomes more sophisticated.

Google may not quite yet be ready to predict the future – but its position as a main player and innovator in the big data space seems like a safe bet.

2

GE

General Electric – a literal powerhouse of a corporation involved in virtually every area of industry, has been laying the foundations of what it grandly calls the Industrial Internet for some time now.

But what exactly is it? Here's a basic overview of the ideas which they are hoping will transform industry, and how it's all built around big data.

If you've heard about the Internet of Things which I've written about previously <click here>, a simple way to think of the industrial internet is as a subset of that, which includes all the data-gathering, communicating and analysis done in industry.

In essence, the idea is that all the separate machines and tools which make an industry possible will be "smart" – connected, data-enabled and constantly reporting their status to each other in ways as creative as their engineers and data scientists can devise.

This will increase efficiency by allowing every aspect of an industrial operation to be monitored and tweaked for optimal performance, and reduce down-time – machinery will break down less often if we know exactly the best time to replace a worn part.

Data is behind this transformation, specifically the new tools that technology is giving us to record and analyse every aspect of a machine's operation. And GE is certainly not data poor – according to Wikipedia, its 2005 tax return extended across 24,000 pages when printed out.

And pioneering is deeply engrained in its corporate culture – being established by Thomas Edison, as well as being the first private company in the world to own its own computer system, in the 1960s.

So of all the industrial giants of the pre-online world, it isn't surprising that they are blazing a trail into the brave new world of big data.

GE generates power at its plants which is used to drive the manufacturing that goes on in its factories, and its financial divisions enable the multi-million transactions involved when they are bought and sold. With fingers in this many pies, it's clearly in the position to generate, analyse and act on a great deal of data.

Sensors embedded in their power turbines, jet engines and hospital scanners will collect the data – it's estimated that one typical gas turbine will generate 500Gb of data every day. And if that data can be used to improve efficiency by just 1% across five of their key sectors that they sell to, those sectors stand to make combined savings of $300 billion.

With those kinds of savings within sight, it isn't surprising that GE

is investing heavily. In 2012 they announced $1 billion was being invested over four years in their state-of-the-art analytics centre in San Ramon, California, in order to attract pioneering data talent to lay the software foundations of the Industrial Internet.

In aviation, they are aiming to improve fuel economy, maintenance costs, reduction in delays and cancellations and optimize flight scheduling – while also improving safety.

Abu Dhabi-based Etihad Airways was the first to deploy their Taleris Intelligent Operations technology, developed in partnership with Accenture.

Huge amounts of data are recorded from every aircraft and every aspect of ground operations, which is reported in real-time and targeted specifically to recovering from disruption, and returning to regular schedule.

And last year it launched its Hadoop <click here> based database system to allow its industrial customers to move its data to the cloud. It claims it has built the first infrastructure which is solid enough to meet the demands of big industry, and works with its GE Predictivity service to allow real-time automated analysis. This means machines can order new parts for themselves and expensive downtime minimized – GE estimates that its contractors lose an average of $8 million per year due to unplanned downtime.

Green industries are benefitting too – its 22,000 wind turbines across the globe are rigged with sensors which stream constant data to the cloud, which operators can use to remotely fine-tune the pitch, speed, and direction the blades are facing, to capture as much of the energy from the wind as possible.

Each turbine will speak to others around it, too – allowing automated responses such as adapting their behaviour to mimic more efficient neighbours, and pooling of resources (i.e wind speed monitors) if the device on one turbine should fail.

Their data gathering extends into homes too – millions are fitted with their smart meters which record data on power consumption, which is analysed together with weather and even social media data to predict when power cuts or shortages will occur.

GE has come further and faster into the world of big data than most of its old-school tech competitors. It's clear they believe the financial incentive is there – chairman and CEO Jeff Immelt estimates that they could add $10 trillion to $15 trillion to the world's economy over the next two decades. In industry, where everything including resources is finite, efficiency is of utmost importance – and GE are demonstrating with the Industrial Internet that they believe big data is the key to unlocking its potential.

3

Cornerstone

Employees are a both a business's greatest asset and its greatest expense. So hitting on the right formula for selecting them, and keeping them in place, is absolutely essential. One company offering unique solutions to help others tackle this challenge is Cornerstone. I will give a brief overview of what they do, and why it's an important – but controversial – example of big data analysis driving business growth.

Cornerstone is a software tool which helps assess and understand employees and candidates by crunching half a billion data points on everything from gas prices, unemployment rates and social media use.

Clients such as Xerox use it to predict, for example, how long an employee is likely to stay in his or her job, and remarkable insights gleaned include the fact that in some careers, such as call centre work, employees with criminal records perform better than those without.

Its prowess has made Cornerstone into a huge success, with sales growing by 150% from 2012 to 2013 and the software being put to use by 20 of the Fortune 100 companies.

The "data points" are measurements taken from employees working across 18 industries in 13 different countries, providing information on everything from how long they take to travel to work, to how often they speak to their managers. Data collection methods include the controversial "smart badges" that monitor employee movements and track which employees interact with each other.

Cornerstone has certainly caused positive change in companies using it – Bank of America reportedly improved performance metrics by 23% and decreased stress levels (measured by analysing worker's voices) by 19%, simply by allowing more staff to take their breaks together.

And Xerox reduced call centre turnover by 20% by applying analytics to prospective candidates – finding among other things that creative people were more likely to remain with the company for the 6 months necessary to recoup the $6,000 cost of their training than inquisitive people.

So far data gathering and analysis has focused mainly on customer-facing members of staff, who in larger organizations will tend to be those with less responsibility and decision-making power. Could even greater benefits be taken by applying the same principles to the movers and shakers in the boardroom, who hold the keys to wider-reaching business change? Certainly some companies are starting to think that way.

The director of research and strategy at one firm that uses the software – David Lathrop of Steelcase – told the *Financial Times* this year that improving the performance of top executives has a "disproportionate effect on the company". Although he did not disclose precise details of methods or results, much research is being carried out in the name of finding exactly what it is that makes high-fliers tick. This will inevitably find its way into analytical projects at big companies which spend millions hiring executives.

Crunching employee data at this level plainly has the opportunity to bring huge benefits, but it could also prove disastrous if a company gets it wrong.

Failing to take proper consideration of individuals' rights to privacy in some jurisdictions (eg Europe) can lead to severe legal penalties. In my opinion, any company thinking about carrying out data-gathering and analysis for these purposes needs to take great care.

In workplaces where morale is low or relationships between workers and managers are not good, it could very easily be seen as a case of taking snooping too far.

Interestingly, Cornerstone's privacy policy makes it clear that information on applicants is provided to them by their clients, including names, work history and contact details. How many people know that simply by applying for a job with one of these clients, their personal data will be made available for analysis? It appears that Cornerstone absolves itself of responsibility here by declaring itself a "mere data processor" – putting the onus on the client businesses to gain permission to distribute their applicants' and employees' data.

It is vitally important that staff are made aware of precisely what data is being gathered from them, and what it is being used for. Everyone (and certainly those running the operation) needs to be aware that the purpose is to increase overall company efficiency, rather than assess or monitor individual members of staff.

With more than half of human resources departments reporting an increase in data analytics since 2010, according to a report by the Economist Intelligence Unit, it's obvious that like it or not, it's here to stay. Companies that use it well, with respect for their employees' privacy and an understanding of the vital principle mentioned above, are likely to prosper. Those who don't – be warned!

4

Microsoft

Since it was founded in 1975 by Bill Gates and Paul Allen, Microsoft has been a key player in just about every major advance in the use of computers, at home and in business.

Just as it anticipated the rise of the personal computer, the graphical operating system and the internet, it wasn't taken by surprise by the dawn of the big data era. It might not always be the principle source of innovation, but it has always excelled at bringing innovation to the masses, and packaging it into a user-friendly product (even though many would argue against this).

It has caused controversy along the way, though, and at one time was called an "abusive monopoly" by the US Department of Justice, over its packaging of Internet Explorer with Windows operating systems. And in 2004 it was fined over $600m by the European Union following anti-trust action.

The company's fortunes have wavered in recent years – notably, they were slow to come up with a solid plan for capturing a significant share of the booming mobile market, causing them to lose ground (and brand recognition) to competitors Apple and Google.

However it remains a market leader in business and home computer operating systems, office productivity software, web browsers, games consoles and search – Bing having overtaken Yahoo as the second most-used search engine.

It is now angling to become a key player in big data, too – offering a suite of services and tools including data hosting and analytics services based on Hadoop to businesses.

But Microsoft had a substantial head-start over the competition – in fact their first forays into the world of big data started way before even the first version of MS-DOS. Gates and Allen's first business venture, two years before Microsoft, a service providing real-time reports for traffic engineers using data from roadside traffic counters. It's clear that the founders of what would grow into the world's biggest software company knew how important information (specifically, getting the right information to the right people, at the right time) would become in the digital age.

Microsoft competed in the search engine wars from the beginning, rebranding its engine along the way from MSN Search, to Windows Live Search and Live Search before finally arriving at Bing in 2009. Although most of the changes it brought in appeared designed to ape the undisputed champion of search Google (such as incorporating various indexes, public records and relevant paid advertising into its results) there are differences. Bing places more importance on how well-shared information is on social networks when ranking it, as well as geographical locations associated with the data.

Microsoft's Kinect device for the Xbox aims to capture more data than ever from our own living rooms. It uses an array of sensors to capture minute movements and is already able to monitor and record the heart rate of users, as well as activity levels. Patent applications suggest there are plans for much wider use, including monitoring the behaviour of television viewers, to provide a more interactive watching experience. The move fits in with Microsoft's strategy of rebranding the Xbox – generally thought of as a games console – into an intelligent living room activity hub which monitors, records and adapts to users' behaviour. No, you are not the only person who finds that idea a little bit scary!

In the business-to-business market, where Microsoft made its first fortunes with its OS and office software, it is now throwing all of its considerable weight into big data-related services for enterprise.

Like Google with its Adwords, Bing Ads provides pay-per-click advertising services which are targeted at a precise audience segment, identified through data collected about our browsing habits.

And like competitors Google and Amazon it offers its own "big data in a box" solutions, combining open-source with proprietary software to offer large-scale data analytics operations to businesses of all sizes.

Its Analytics Platform System marries Hadoop with its industry-standard SQL Server database management technology, while its ubiquitous Office 365 will soon make data analytics available to an even wider audience, with the inclusion of PowerBI – adding basic analytics functions to the world's most widely used office productivity software.

It is also looking to stake its claim on the Internet of Things with Azure Intelligent Systems Service. This is a cloud-based framework built to handle streaming information from the growing number of online-enabled industrial and domestic devices, from manufacturing machinery to bathroom scales.

It may have missed a trick with mobile – prompting many premature declarations that Microsoft was falling behind the competition – but its keen embrace of data and analytics services show that it is still a key player.

When CEO Satya Nadella took up his post at the start of this year he emailed all employees letting them know he expected huge change in the industry, and the wider world, very soon, prompted by "an ever-growing network of connected devices, incredible computing capacity from the cloud, insights from big data and intelligence from machine learning."

So it's clear that Microsoft aims to put big data at the heart of its business activities for the foreseeable future, and provide (relatively) simple software solutions to help the rest of us do the same.

5

Kaggle

If you're looking for a company which seems to embody all the principles of big data entrepreneurship under one roof, then look no further than Kaggle.

Crowd sourcing, predictive modelling, gamification – Kaggle has it all - and has worked out how to turn a profit from them.

The San Francisco-based business awards cash prizes to its teams of "citizen scientists" who compete to untangle big data challenges of all shapes and sizes.

And it isn't just businesses which are benefitting – by applying the concept of crowd-sourcing to data analytics, they are helping to further scientific and medical research. Their projects include looking deep into the cosmos for traces of dark matter, and furthering research into HIV treatment.

Chief scientist at Google (which has itself benefitted from Kaggle's research) and Kaggle investor, Hal Varian, describes it as "a way to organize the brainpower of the world's most talented data scientists and make it accessible to organizations of every size."

And that's certainly an intriguing aim – as well as a highly profitable one – in a world where businesses of all sizes are beginning to cotton on to the benefits of big data. Even if every company could afford to set up its own data analytics department, there aren't nearly enough people trained to do the job to go around!

As with all emerging sciences, there is a shortage of trained data scientists at the moment – but Kaggle has 150,000 of them, ready to farm out to the highest bidder.

As well as charging companies they work with (including Amazon, Facebook, Microsoft and Wikipedia) up to $300 per hour for consultancy work, the company organizes competitions – which is where the gamification comes in.

I've written about gamification before – and Kaggle works along the same lines, with the theory being that it is easier to get people to take part in something if it is presented to them as a challenge or competition of some sort.

Current challenges include assisting with schizophrenia diagnosis by identifying the condition from MRA neuroimaging data, and finding the Higgs Boson amidst the mountains of data collected by CERN's Atlas particle physics experiments.

They are open to anybody to take part in, and all the information (as well as the necessary data sets can be found at Kaggle's website.

Although it is frequently reported that they have "over 100,000 data scientists", these are actually registered users and competitors rather than employees. There are no qualification or experience barriers to registering as a Kaggle data scientist, previous winners have ranged from data science academics and professionals to enthusiastic, knowledgeable amateurs. However certain competitions are occasionally reserved for "masters" – those who have shown they have the right stuff through their previous work with Kaggle.

The company also recruit its own staff to work on internal projects. In fact they are advertising for recruits now – and although no requirements are listed, other than that applicants be "experienced", two questions on the application form ask for the mean and standard deviation of two sets of numbers.

The concept is undoubtedly inspired by earlier pioneering work in crowd-sourcing data analysis, such as the Search For Extra-terrestrial Intelligence at Home (SETI @home) project, and a competition organized by Netflix in 2009 offering £1 million to the person who came up with a better algorithm for providing movie recommendations.

Kaggle has taken those idea and expanded on them, basically – it acts as the middle man, with companies or organizations bringing their problems, and Kaggle packaging them into competitions, gathering the contestants and sharing out the rewards.

The data itself is often simulated – and contestants are challenged to come up with methods or algorithms which are more efficient than existing methods at solving the problem in hand. Using simulated data means that issues surrounding access to sensitive data can be

sidestepped. Once that is done, the reward – currently up to $30,000, although occasionally much larger for the top projects – is paid.

One of its best known success stories was the Heritage Health Prize, which awarded $3 million last year to the winning entrant, whose algorithm most accurately predicted which patients would be admitted to hospital in the coming 12 months, from a set of medical data.

They also offer the Kaggle In Class service – an academic spin-off of the main brand which offers free data processing tools and simulated challenges. It is intended for use in schools and colleges struggling to meet the challenges of training the first generations of professional data scientists.

Of course like anything new it isn't without its critics. In particular, questions have been asked about how valuable the research it leads to actually is – often, they say, the biggest challenges in data analysis revolve around what data is needed, and what questions should be asked. Kaggle's pre-packaged competitions take this element out of the equation. The crowdsourced data scientists might be working on the solution to a particular problem – but is it the correct one? And might there be more relevant data elsewhere, other than that supplied in the competition package?

This might be a fundamental limitation to the competition model, until data collection and distribution evolves to the point where it can be made available to contestants in real-time, and then of course there will be serious privacy and data protection issues to hurdle.
But as it stands today, Kaggle is one of the more forward-thinking innovations in big data, and has done much to raise awareness of the power that crowd sourcing data analysis can bring to businesses and organizations of all sizes.

6

Facebook

Facebook – it's the world's biggest social network by a huge margin, and most of us are used to using it to share details of our everyday lives with our friends and families. It's no secret now that we're also sharing it with their advertisers, but that hasn't put most of us off using it! So here's a brief rundown of how Facebook has been one of the most successful companies in the world at gathering our data and turning it into profit – and why some think its business practices sometimes overstep the mark.

Recently, Facebook has been causing a stir amongst those interested in online privacy and data protection. The latest accusations are that is has been carrying out unethical psychological research – effectively experimenting on its users without their permission. Critics have said that by attempting to alter people's moods by showing them specific posts with either a positive or negative vibe,

and then measuring their response, several ethical guidelines have been broken.

The truth though, is that Facebook (and the internet at large) is making its own rules as it goes along. Putting 1.25 billion people – that's getting on for one fifth of the world's population, if we pretend for a second that none of the accounts are duplicates – within a mouse click of each other was always going to have far reaching consequences. And with hindsight it was a bit silly to have ever expected it to be manageable within established social and legal boundaries.

Of course those of us who love social media believe the potential benefits far outweigh the hazards. Putting aside how much easier it makes keeping in touch with our friends and family, there's clearly a lot to be learned from studying the data generated during that communication. And gathering data from us is the foundation of Facebook's business model.

Don't forget though - although it now seems to be dipping its toes into psychological experiments, Facebook's main motivation for collecting and analysing our data has always been to sell us adverts.

Advertisers benefit from highly detailed profiles users build up over time as they use the site – meaning their messages can be targeted precisely at "women over 40 who love books" or "men under 25 living in the UK who love football".

The huge and speedy success of Facebook was prompted by its simple interface and, somewhat ironically given how things have developed, emphasis on user privacy. This helped it quickly become more popular than other early social networks such as Myspace and

Bebo. But with hindsight, it's clear to see it was always gunning for bigger targets.

A big difference between Google and Facebook is that Google's information on who we are is often a "best guess" based on what sites we are visiting. From the start, Facebook explicitly asks us who we are, where we live and what we are interested in. Yes, Google eventually started to do the same with Google+, but by then, they were simply playing catch-up. Advertisers clearly value this direct approach – ad revenues at Facebook grew by 129% from 2011 to 2013, compared to 49% at Google during the same period.

Like Google and all of the other big tech firms, buying up smaller firms to make use of their IP and, crucially, the data from their user base, is a core business strategy. Notable acquisitions have included Instagram and Whatsapp, both of which came with existing communities of millions of users to add to Facebook's own. Interestingly, their highest profile recent purchase was the makers of the upcoming Oculus Rift virtual reality headset. They are clearly thinking ahead to a time when we may be looking for more convenient methods than existing screens offer to view our data.

Facebook has always said that the privacy worries this causes are addressed by the fact that all information is shared with our permission and anonymized when sold on for marketing purposes. That hasn't stopped a lot of critics taking issue with their practices though. For example, many say that the privacy settings are too complex or not clearly explained, meaning it is too easy for people to share things they didn't mean to. Facebook have tried to fix this several times over the years – often confusing people who had got used to the way they were!

Another feature which caused concern when it was introduced was facial recognition. When you upload a picture, you might see suggestions for people you could tag on it. This is based on analysis of the picture data, which is compared against pictures of people in your Friends list, and prompted an investigation by EU privacy regulators in 2011.

More recently, changes to the way its users' habits are monitored have caused concerns. Its latest monitoring tools record everything from how long a user "hovers" their cursor over certain parts of the page to what websites they visit outside of Facebook. Last month it announced that this information is being used in their algorithms that determine which adverts to show us.

But, at least it is now possible to delete your data permanently, if you don't want Facebook to have access to it at all, any more. Before 2010 you couldn't even really delete your account – although you could remove your profile, everything was kept on their servers for an unspecified amount of time, for unspecified purposes (it would hardly be any use to target adverts at you, if you no longer had an account with which to view the site.) Outcry when this was discovered prompted Facebook to add the ability to erase yourself completely.

Facebook's data strategy is led by its Data Science team – who have their own page, of course, which you can see here. They regularly post updates on insights they have gleaned from analysing the habits of the millions who browse the site.

Overall I think that a lot of the problems caused by Facebook are a symptom of its enormous success. Regulators and lawmakers have shown themselves to be slow to get to grips with the revolution it

(and other social media) have brought to the way we communicate with each other, day to day. And, as with Google, it seems like there are more than enough people who think the problems are worth putting up with, for the convenience it brings.

7

Amazon

Amazon is a big data giant, which is why I want to look at the company in my second post of my series on how specific organizations use big data.

We all know that Amazon pioneered e-commerce in many ways, but possibly one of its greatest innovations was the personalized recommendation system – which, of course, is built on the big data it gathers from its millions of customer transactions.

Psychologists speak about the power of suggestion – put something that someone might like in front of them and they may well be overcome by a burning desire to buy it – regardless of whether or not it will fulfil any real need.

This is of course how impulse advertising has always worked – but instead of a scattergun approach, Amazon leveraged their customer data and honed its system into a high powered, laser-sighted sniper

rifle. Or at least that is the plan – they don't seem to get it completely right yet. I have had some very strange recommendations from Amazon.

Anyway, their systems are getting better and it looks like what we have seen so far is only the beginning – as I've previously mentioned, Amazon has recently obtained a patent on a system designed to ship goods to us before we have even decided to buy it – predictive despatch – you can read more about that here. This is a strong indicator that their confidence in reliable predictive analytics is increasing.

An important factor to consider when looking at Amazon is how commercial its big data is, compared to those of other companies that deal with data on a comparable scale. Unlike, say, Facebook – which might know an awful lot about which movies you like or who your friends are – the vast majority of Amazon's data on us relates to how we spend hard cash.

And having worked out how to use it to get more money out of our pockets, it is now setting out on a mission to help other global corporations do the same – by making that data, as well as its own tools for analysing it, available to buy.

This means that, as with Google, we have started to see adverts driven by Amazon's platform and based on its data appearing on other sites over the past few years. As noted by MIT Technology Review last year, this makes the company now a head-on competitor to Google – with both online giants fighting for a chunk of marketers' budgets.

However, ad sales is not the only arena in which Amazon is taking on Google – its Amazon Web Services offers cloud-based computing

and big data analysis on an enterprise scale. This allows companies which need to run highly processor-intensive procedures to rent the computing time far more cheaply than setting up their own data processing centres – just like Google's BigQuery.

These services include datawarehousing (Redshift), hosted Hadoop solution (Elastic Map Reduce), S3 – the database service it uses to run its own physical warehousing operations and Glacier, an archival service. Recently added to this list is Kinesis, which is a real-time "stream processing" service designed to aid analysis of high volume, real-time data streams.

Amazon has also incorporated big data analysis into its customer service operations. Its purchase of shoe retailer Zappos is often cited as a key element in this. Since its founding, Zappos had earned a fantastic reputation for its customer service and was often held up as a world leader in this respect. Much of this was due to their sophisticated relationship management systems which made extensive use of their own customer data. These procedures were melded together with Amazon's own, following the 2009 acquisition.

Finally, it is worth mentioning the public data sets that Amazon hosts, and allows analysis of, through Amazon Web Services. Fancy digging around in the data unearthed through the Human Genome Project, NASA's Earth science datasets or US census data? Amazon hosts all of this and much more, and makes it available for anyone to browse for free.

Amazon has grown far beyond its original inception as an online bookshop, and much of this is due to its enthusiastic adoption of big data principles. It looks set to continue breaking new ground in this field, for the foreseeable future.

ABOUT THE AUTHOR

Bernard Marr is a leading global authority and best-selling author on organizational performance and business success. He is a LinkedIn Influencer, he writes the Big Data Guru blog and is the world's #1 expert on big data. He regularly advises leading companies, organizations and governments across the globe, and is an acclaimed and award-winning keynote speaker, researcher, consultant and teacher.

Bernard is the founder and CEO of the Advanced Performance Institute. Prior to this he held influential positions at the University of Cambridge and at Cranfield School of Management. Bernard Marr's expert comments on organizational performance have been published in *The Times*, *The Financial Times*, *The Sunday Times*, *Financial Management*, the *CFO Magazine* and *The Wall Street Journal*.

Connect with Bernard

GET SMART AND LEARN TO CONVERT THE PROMISE OF BIG DATA INTO REAL–WORLD RESULTS

There is so much buzz around big data. We all need to know what it is and how it works. But what will set you apart from the rest is actually knowing how to use big data to get solid, real-world business results and putting that in place to improve performance.

If you like what you've read here get yourself a copy of *Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance* and learn how to devise your own big data strategy.

- **Understand why you need to clearly define what it is you need to know from your data**
- **Learn how you can collect relevant data and measure the metrics that will help you answer your most important business questions**
- **Find out how the results of big data analytics can be visualized and communicated to ensure key decisions-makers understand them**

BUY TODAY FROM YOUR FAVOURITE BOOKSTORE!