

MarketLine Case Study

Big Data

A road map for business
intelligence

Reference Code: ML00019-004

Publication Date: March 2015

WWW.MARKETLINE.COM

MARKETLINE. THIS PROFILE IS A LICENSED PRODUCT AND IS NOT TO BE PHOTOCOPIED

MarketLine

OVERVIEW

Catalyst

For a number of years, Big Data has been intensively promoted as a groundbreaking technological development that will revolutionize how businesses operate. Major corporations have therefore been fairly quick to understand and exploit the opportunities available. However, for most people the technology involved is not especially accessible and since the industry is awash with companies vying for position and market share, it can be difficult to obtain a coherent picture of precisely what the fuss is all about. In order to make sense of the nature of the industry and the opportunities it offers, this report provides an understanding of the core technologies and key players. In addition, this study analyzes the uses of Big Data in terms of business intelligence and provides a structure to those uses for ease of comprehension. Finally, this report seeks to identify potential gaps in the market and provides an outline of a conceivable future for the industry.

Summary

- Big Data technologies can be viewed simply as a contemporary way to store, query and model information. The essential reason that businesses have been quick to adopt Big Data is to do with the level of insight that can be obtained and the types of problems that can be solved through Big Data technologies.
- The majority of Big Data providers have centered on software developed by the non-profit Apache Software Foundation, particularly through the Hadoop framework.
- Hadoop is made up of a freely available software library that includes Hadoop Common, Hadoop YARN, the Hadoop Distributed File System and Hadoop MapReduce. This library combines to allow the rapid processing of large datasets, by distributing files across clusters of computers, with the capability of scaling up to thousands of machines. Vast amounts of multi-format information can consequently be stored and analyzed.
- The key players in the industry include: major technology companies such as Google, IBM, Oracle and Microsoft; services companies such as PwC and Accenture; and rapidly rising upstarts such as Splunk and Cloudera.
- A wide range of industries are using Big Data technologies but the dominant operational feature in recent years has been a restriction to internally-focused data, either from information about the company itself and its processes, or through the interaction of customers with that company.
- Two ways of looking at the potential future of the industry are in terms of externally generated information and the feasible limits of Big Data analytics. The result may well be a broader business intelligence perspective, forming what might be termed Big Data 2.0, or holistic intelligence.
- In terms of competitive advantages gained from Big Data technologies, the future belongs to those who are able ask the right questions, rather than those who can capture the most information.

TABLE OF CONTENTS

Overview	2
Catalyst.....	2
Summary	2
The core technologies and the key players	6
The core technologies are open & free but complex to operate	6
Open source software has been central to the development of Big Data.....	6
How can a business make sense of Big Data and why does it matter?	6
How does Big Data really work?	7
Hadoop Common.....	7
The Hadoop Distributed File System (HDFS)	7
Hadoop YARN	7
Hadoop MapReduce	8
The key players include software majors and rising upstarts.....	9
Technology powerhouses are likely to reap the most rewards in the long-term.....	9
The new and the hungry have seen exponential growth but not all can be winners	10
Early adopters have been introspective	12
Customer-generated information has provided many new strategic insights.....	12
Customer analytics gives businesses greater predictive power	12
Sentiment analysis can be used to tailor public interactions and optimize production processes	13
Analysis of the general public can lead to improved cost and time efficiencies	14
Company processes and operational information become clearer with Big Data	14
Previously impenetrable research and development problems have become solvable	14
Industrial usage points to huge cost savings through optimization techniques	14
Big Data 2.0 – holistic intelligence.....	16
Where do we go from here?	16
What happens when everyone is using Big Data?	16
Sourcing external data is a logical next step in the short-term	17
Conclusions.....	18
The future belongs to those who ask the right questions.....	18
Appendix	19
Definitions	19
Sources	19
Further Reading.....	22

Ask the analyst	23
About MarketLine	23
Disclaimer	23

LIST OF TABLES

Table 1: The core Big Data platforms of the major players in the industry 9

Table 2: The core Big Data platforms of the smaller players in the industry..... 10

TABLE OF FIGURES

Figure 1: The Hadoop Big Data framework 8

Figure 2: The Gnip approach to social data 13

THE CORE TECHNOLOGIES AND THE KEY PLAYERS

The Big Data industry is truly one for the 21st century; it uses new and advanced technology, many innovations are shared through a non-profit organization and it has spread quickly all over the world. The core technologies stem from research by Google and developments by the Apache Software Foundation. In essence, these technologies can be viewed simply as a contemporary way to store, query and model information, much like a new Microsoft Excel for the Internet age. The key players include: major technology companies such as Google, IBM, Oracle and Microsoft; services companies such as PwC and Accenture; and rapidly rising upstarts such as Splunk and Cloudera.

The core technologies are open & free but complex to operate

Open source software has been central to the development of Big Data technologies; without this system of innovation the industry may not have developed as rapidly as it has done. From a business perspective, Big Data can be thought of as a new tool to study and model large volumes of different types of information, opening up the possibility of capturing a wider range of data to inform a company's business intelligence process. From a technological perspective, Big Data is a major advance in computing power, combining new types of storage systems, cloud computing capabilities, and the implementation of a wide range of mathematical functions, including machine-learning algorithms.

Open source software has been central to the development of Big Data

The technology behind Big Data has broadly been developed through an open system of innovation centered on the Apache Software Foundation, which developed the Hadoop framework offered by most Big Data providers. These providers then contribute their own developments back to the Apache Software Foundation to improve the Hadoop framework for everyone. Due to this open system, input into the development of these technologies has come from a wide range of players. This means that whilst the Big Data industry has advanced very quickly, those who understand how to fully operate and customize these systems generally need to have a level of expertise comparable to a research scientist in computing or mathematics. Nevertheless, companies have been quick to employ data scientists and computing experts to take advantage of the Big Data phenomenon. For more details about open innovation systems, see MarketLine Case Study, ['The open source and graph database revolutions: How software companies manage knowledge as a strategic asset'](#) (ML00017-059, November 2014).

How can a business make sense of Big Data and why does it matter?

Ignoring the technological and structural differences for a moment, from a business point of view it is useful to consider whether Big Data is just 'the new Excel'. This question is helpful in order to frame what Big Data is actually offering beyond the rhetoric of 'data at the speed of thought' (Oracle) or 'data at the speed of business' (IBM). In its most distilled form, Big Data is just another way of organizing information so that it can be made useful, just as spreadsheets allow the majority of businesses to store and model data. The ultimate difference with Big Data is that the information that is being dealt with is so large and is made up of so many different formats that a traditional spreadsheet or database cannot cope with the volume or variety. Big Data technologies are therefore a direct response to the amount and type of information that can be collected through the Internet and, more recently, through physical devices. This puts Big Data at the heart of how it is possible to gain some sort of strategic advantage from the mountains of digital information being created in the modern world (worldwide data reached four zettabytes in 2013 and some commentators expect there to be 44 zettabytes of global data by 2020 (one zettabyte = one trillion gigabytes)).

So, from a business perspective, Big Data is a lot like 'the new Excel'. Just as many contemporary businesses could not operate as efficiently without spreadsheets, in five to ten years, many businesses are unlikely to be able to compete as effectively without Big Data technologies purely because of the amount of digital information that can be collected and analyzed today.

Moreover, organizations have been quick to adopt Big Data due to the level of insight that can be obtained and the types of problems that can be solved through Big Data technologies. For example, in 2012 NASA held 32 petabytes of information on the earth's climate and expects to hold nearly 350 petabytes by 2030 (five petabytes roughly matches the level of information contained in the total number of letters delivered by the US Postal Service in one year – around 160

billion letters). The maximum Excel spreadsheet size is 1,048,576 rows by 16,384 columns. Although this enables 17.2 trillion points of data, one cannot imagine that any process run on that data would be particularly fast or that the data itself could be more than just numbers. It is therefore only with new technologies that problems such as modeling satellite images of the earth's climate can be solved. Whilst this is an extreme example of the amount and variety of information that can be collected on a given subject, businesses also face many challenges that they have been unable to get past until now. For example, there is now the potential for airlines to collate enough machine-generated information to determine when a plane is likely to have an engine failure up to 30-50 flights before the issue occurs, therefore allowing engineers to fix small problems before they become catastrophic.

As such, companies like IBM have invested heavily in Big Data and analytics technologies in order to stay ahead of the competition. In its 2013 annual report for example, IBM claims to have invested \$24bn to date in order to build Big Data and analytics capabilities, which includes \$7bn of organic investment. Globally, IBM obtained revenues of \$15.7bn in 2013 alone in relation to analytics, making the investment appear well spent.

How does Big Data really work?

The majority of Big Data providers have centered on software developed by the non-profit Apache Software Foundation, particularly the Hadoop framework. Hadoop is made up of a freely available software library that includes the following modules: Hadoop Common, Hadoop YARN, the Hadoop Distributed File System and Hadoop MapReduce. This library combines to allow the analysis of large datasets, which have been distributed across clusters of computers or servers to maximize storage space, with the capability of scaling up to thousands of machines. The system is designed so that computations occur at the point of information (rather than the information being in front of you like with Excel) and so that failures in individual pieces of hardware (on which the data is stored) can be overcome without a breakdown in the overall system. The system is therefore fast and can be run on relatively cheap hardware that has a propensity to fail. The Hadoop framework functions as follows and is shown in Figure 1 further below:

Hadoop Common

Hadoop Common contains the common utilities (files) that support the other Hadoop modules, such as the files required to start up the whole framework. A full understanding of how Hadoop Common actually works is therefore not essential for an understanding of the overall system.

The Hadoop Distributed File System (HDFS)

This is the central part of the Hadoop framework, determining where and how all the information is stored. It is based on the Google File System. The HDFS stores vast volumes of large files by distributing them across multiple servers, replicating files in multiple locations so that individual hardware failures do not impact on the operation of the system. Amongst other functions, the files can be read, edited and scanned by applications for analysis; Hadoop YARN (detailed below) is used to manage these functions in terms of scheduling and allocation. Combined with cloud storage or on-premises servers, the HDFS allows multiple users to store massive amounts of data outside of their own computers, yet be able to access and use that data quickly and easily.

Whilst this may not sound all that different to a standard network system of folders and files, it is important to remember that the volume and variety of information being held is potentially enormous due to the distributed nature of the system. The functions that can be used on the files are also far more sophisticated due to the parallel nature of Hadoop MapReduce (detailed below) and the way those functions are managed by Hadoop YARN.

Hadoop YARN

Hadoop YARN is like a project manager for the whole framework, acting as a job scheduler and cluster management resource (i.e. it manages the use of the data that is distributed across different servers through the HDFS). This allows multiple acts of data processing on the information stored within Hadoop (the processing is actually done by Hadoop MapReduce, detailed below). Hadoop YARN project management therefore means that real-time analytics, data science (statistics and machine-learning algorithms) as well as longer-term batch processing jobs can all be performed in a single platform, often at the same time.

Hadoop MapReduce

This is where the action happens in terms of gaining actual business intelligence from the information being stored in the HDFS. It is the part of the Hadoop framework that delivers parallel processing capabilities and therefore allows computations to be carried out at speed. It is based on Google MapReduce. In order to understand the process, imagine you have a HDFS filled with files about Facebook users and you want to know the average age of all the users. Facebook has around 1.5 billion users so it is worth remembering that this cannot be done with a traditional spreadsheet since the rows are limited to just over one million entries. Instead, with the information held in a large file system, one way to do this would be to go through each file one at a time, find the age value, add them all up and then find the average. Going through this process consecutively 1.5 billion times would take massive amounts of time. With Hadoop MapReduce and its parallel processing capability however, the Map function is able to split the task out across multiple processors at the location of the distributed files all at the same time, and report the ages back much more quickly. In simplified terms, the Reduce function then carries out the averaging operation on the information returned by the Map function. MapReduce is therefore used to carry out a variety of analytical processes on the information being stored and can be used to provide valuable insights. The disadvantage with MapReduce is that it is designed for batch processing and is generally run overnight before results are returned. Big Data providers have therefore sought to supplement MapReduce in order to provide real-time analytics and further computational capabilities.

Additional software can be added to the core Hadoop framework

In addition to the central framework of Hadoop, an ever increasing and improving variety of further software libraries can be added alongside or on top of the traditional Hadoop framework to maximize functionality. For example, HBase (based on Google Bigtable) can be run on top of HDFS. It is a distributed non-relational data store that works through a system of very large tables, supporting billions of rows and millions of columns. It is designed to augment the HDFS by allowing for the fast lookup of records within the file system. Effectively, HBase puts the data held in HDFS in indexed store files and acts like a map for the system so that information can be found quickly.

Apache Spark is relatively new and will potentially replace Hadoop MapReduce. As mentioned above, MapReduce was designed for batch orientated (overnight) processing, whereas many business applications require real-time analytics and the use of a wider variety of algorithms in order to form predictive business intelligence capabilities. Apache Spark is far faster than MapReduce and includes the ability to operate faster machine learning and graph computations (to examine how information is connected together). The addition of Apache Spark to the Hadoop framework is likely to create an extremely powerful business intelligence platform.

Figure 1: The Hadoop Big Data framework



The key players include software majors and rising upstarts

The Big Data industry is almost overrun with businesses grappling for market share, each one either claiming to be the most powerful, fastest, and easiest to use or that they operate in some sort of niche that cannot be found elsewhere. The industry is currently shifting from an exciting early stage, within which new entrants have been able to gain a foothold and obtain relatively high levels of engagement, to a structure in which only the major players and some of the more well-known upstarts are likely to survive on a global scale.

Technology powerhouses are likely to reap the most rewards in the long-term

As might be imagined from the description of the Hadoop framework above, although the software libraries are freely available and Hadoop is designed to run on relatively cheap hardware, do-it-yourself (DIY) Big Data is not easy. As such, household name technology companies offer private 'enterprise ready' distributions of the Hadoop system (or their own Big Data technologies, particularly in terms of analytics) together with expert guidance and assistance.

Table 1: The core Big Data platforms of the major players in the industry

Company	Core Big Data Distribution	Additional functionality
IBM	IBM BigInsight (based on Hadoop)	IBM BigInsight
Hewlett Packard	HAVEn (based on Hadoop)	HAVEn predictive analytics, R statistical programming
Oracle	Hadoop via Cloudera	R statistical programming
Microsoft	HDInsight (via Hortonworks' Hadoop framework)	SQL programming and Microsoft Excel integration
Amazon	Amazon Elastic Cloud Compute (do-it-yourself) or through Amazon Elastic MapReduce	Amazon Kinesis, Cloudera Impala, Splunk Hunk, plus others
Google	Google Cloud Dataflow	Google Compute Engine, Google Cloud Storage, Google BigQuery

SOURCE: Company websites

MARKETLINE

IBM

IBM offers a range of hardware, software and services in terms of Big Data. Its key product is known as an 'IBM Watson Foundation', which involves the full gamut of Big Data analytics services including data management and warehousing, content management, information governance, and their own Hadoop distribution (BigInsight). IBM claims that BigInsight runs four times faster than the standard Hadoop distribution as well as providing data scientists with the capability of running more machine-learning and mathematical algorithms than can be used with Hadoop MapReduce.

Hewlett Packard (HP)

HP also offers a range of hardware, software and services in terms of Big Data. The central product is the HAVEn Big Data platform, which is available either with private servers or in the cloud. HAVEn stands for 'Hadoop, Autonomy, Vertica, Enterprise Security' with the 'n' standing for the large number of offerings that can be tied into the service. To augment the HP Hadoop distribution, HAVEn Predictive Analytics offers large-scale machine learning and statistical analysis through a HP distribution of the open source R statistical programming language, which is popular with many data scientists due to its flexibility and large online community.

Oracle

The Oracle Big Data Appliance offering includes both hardware and software. The Appliance uses a Hadoop distribution via Cloudera CDH (a smaller private company detailed further below), an Oracle database, an Oracle data integrator application adaptor for Hadoop, an Oracle loader for Hadoop and an open source distribution of the R statistical programming language. Oracle appears to market its Big Data Appliance as being a comprehensive and secure offering rather than a full analytics package, with most of the analytics being run through the R language, in a similar way to HP.

Microsoft

Microsoft's Big Data offering, HDInsight, is based on the Hortonworks Data Platform, which itself is built through the Apache Software Foundation libraries. The platform has been designed for use on Microsoft Azure cloud services and with an SQL database system, which will be attractive to many programmers who already understand the widely-used SQL programming language. A huge selling point for Microsoft is that it has developed the platform so that it can integrate with Microsoft Excel, meaning business staff with more widely-used IT skills could soon begin to embrace Big Data analytics.

Amazon

Amazon Web Services includes a wide range of Big Data services available in the cloud. These include: Amazon Redshift, which acts as a data warehouse; Amazon Kinesis, which is a platform for analyzing data in real-time; Amazon Elastic MapReduce, which runs the data processing aspect of the offering and uses Hadoop to distribute the data as well as being able to work alongside Apache software libraries including Spark and HBase, the Cloudera Impala analytics platform, and Splunk's Hunk software for machine-generated data analytics (Cloudera and Splunk are detailed further below); and Amazon Dynamo DB, which is a form of database. Amazon Elastic Cloud Compute is also available, which allows you to run a self-managed DIY Big Data analytics platform on the Amazon cloud.

Google

The Google Cloud Dataflow service is new (summer 2014) and, as of time of writing, is currently in 'Alpha release'. It is a combination of Google Compute Engine, Google Cloud Storage and Google BigQuery which, tied together, can be used for data processing tasks 'of any size'. Given that Google literally wrote the book on the technology that inspired the Hadoop framework, these are likely to form an advanced offering. In fact, the company has noted that the technology behind Cloud Dataflow evolved from Google MapReduce and then from successor technologies used internally such as Flume (which enables the use of parallel data pipelines rather than singular parallel data processing) and MillWheel (which is a framework for rapid data processing). As such, it may well be that Cloud Dataflow quickly becomes the leading player.

Professional services firms

Companies such as PwC, Deloitte and Accenture also offer Big Data services. However, these are in relation to assisting with the operation of the analytical system and providing insights into how to apply the framework to a particular business or industry rather than the actual distribution of a Big Data package.

The new and the hungry have seen exponential growth but not all can be winners

There is almost a never-ending list of smaller private companies offering Big Data technologies, services and analytics. These include Palantir, Pivotal, Splunk and Cloudera. Whilst these are some of the larger private companies operating in the market, the hundreds of other smaller companies are unlikely to be able to compete on a global scale now that Big Data is becoming more established and accepted. Even a significant advance in technology from a smaller player would likely have to be made open source and freely available in order to gain enough attention.

Table 2: The core Big Data platforms of the smaller players in the industry

Company	Core Big Data Distribution	Additional functionality
Palantir	Palantir Gotham and Palantir Metropolis	In-house analytics and data management through Phoenix, Raptor, Horizon and RevDB
Pivotal	Pivotal HD (based on Hadoop)	Pivotal Analytics and the use of the data-lake concept
Splunk	Splunk Enterprise	Splunk Hunk analytics, specialized for machine-generated data
Cloudera	Cloudera CDH (based on Hadoop)	Cloudera Impala analytics

SOURCE: Company websites

MARKETLINE

Palantir

Palantir offers Palantir Gotham (all data types) and Palantir Metropolis (quantitative data), known as 'data fusion platforms', which are layered over with machine-assisted analytics software. Palantir Gotham is very much based on the data lake concept, whereby every aspect of an organization's data is integrated into a single model. Palantir uses a variety of in-house software to make up the data fusion platforms including Phoenix, Raptor, Horizon and RevDB. Based on various rounds of funding, Palantir has been valued at somewhere between \$6bn and \$9bn, and revenues have been reported at \$250m for 2013. Palantir has been used by the CIA and the FBI, providing an almost Minority Report style intuition of criminal activities, verging on an Orwellian nightmare depending on your point of view.

Pivotal

Pivotal distributes Pivotal HD (a Hadoop distribution) and runs the Pivotal Analytics platform. The company runs a subscription based storage service, with costs based on volume. The company is set up to maximize the amount of insight that can be obtained through the entirety of an organization's data, rather than a specific section of an organization's information, i.e. Pivotal, like Palantir, follows the data lake concept. Despite the initial success of the company, Pivotal laid off 60 employees on the Big Data side of its business in 2014, which demonstrates the fragile nature of some of the smaller companies in the industry.

Splunk

Although able to monitor and analyze a full range of data, Splunk is largely a specialized offering in relation to the Internet of Things. For example, Splunk provides the analytics involved in the Nest smart thermostat. It has therefore historically been aimed at businesses that create a lot of in-house data through their own physical devices. Splunk has over 8,400 customers in 100 countries.

Cloudera

Cloudera was created by former employees of Facebook, Google, Oracle and Yahoo. It offers the Hadoop framework under its Cloudera CDH platform. In addition to the standard Apache Software Foundation libraries, Cloudera also offers a user interface and runs the Impala analytics software to supplement the operations of Hadoop MapReduce. Cloudera claims that its distribution of Hadoop is the world's most popular in terms of download volumes.

EARLY ADOPTERS HAVE BEEN INTROSPECTIVE

So, which companies are using all these technologies and what are they using them for? A wide range of industries are applying Big Data technologies; at present, the dominant feature has been a restriction to internally-focused data rather than a particular industry vertical. Major corporations are therefore beginning to have a very clear picture of their own internal operations and customer bases, which has given rise to business analytics at a whole new level of detail.

The majority of use cases documented by Big Data providers highlight a trend that most companies using Big Data technologies are staying well within their comfort zones. For instance, many of the current applications of Big Data are simply extensions of what companies already carry out in terms of understanding their operations and customers; the difference is in the level of detail. As such, instead of conducting market surveys or focus groups to understand the public perception of a particular product, or carrying out regular manual checks on machinery, Big Data has stepped in to monitor consumer trends and check production facilities. In terms of the types of internally-focused information that can be created, these can broadly be split between customer generated information and company process or operational information.

In light of this trend, professional services companies marketing themselves as being involved in Big Data therefore currently offer services to help companies analyze their own data. This is in contrast to what one might expect, i.e. professional services firms are not using Big Data themselves to provide a separate form of business or competitive intelligence.

Customer-generated information has provided many new strategic insights

A central feature of current Big Data usage has been the analysis of 'customer'-generated information, which might include customers interacting with retail websites, banking customers' use of credit facilities, or the general public's use of government programs. Customer analytics, sentiment analysis and governmental analysis of the general public have therefore been some of the early uses of Big Data.

Customer analytics gives businesses greater predictive power

A famous example of customer analytics involves Target's "pregnancy-prediction" model. Research shows that previously ingrained shopping habits fall apart when customers become parents due to a sense of overwhelming exhaustion, leading to them becoming more willing to purchase everything in one place rather than from separate outlets. New parents are therefore worth a lot of money to retailers if they become loyal customers.

As birth records are public in the US, new parents are therefore inundated with advertisements, offers and incentives from retailers after the birth of their child. This means that retailers such as Target face fierce competition from rivals for potential new customers, with new parents often overlooking or disposing of much of the marketing without even looking at it. To overcome this issue, Target used Big Data technologies to determine when a potential customer was pregnant, even if that customer had never told them.

Using its baby-shower registry and REDcard customer loyalty information, the company was able to analyze the data and discover specific buying habits that emerged at various stages of pregnancy. As a result, Target was able to assign each shopper a pregnancy prediction score and even estimate a due date for customers based on the purchasing habits of around 25 specific products.

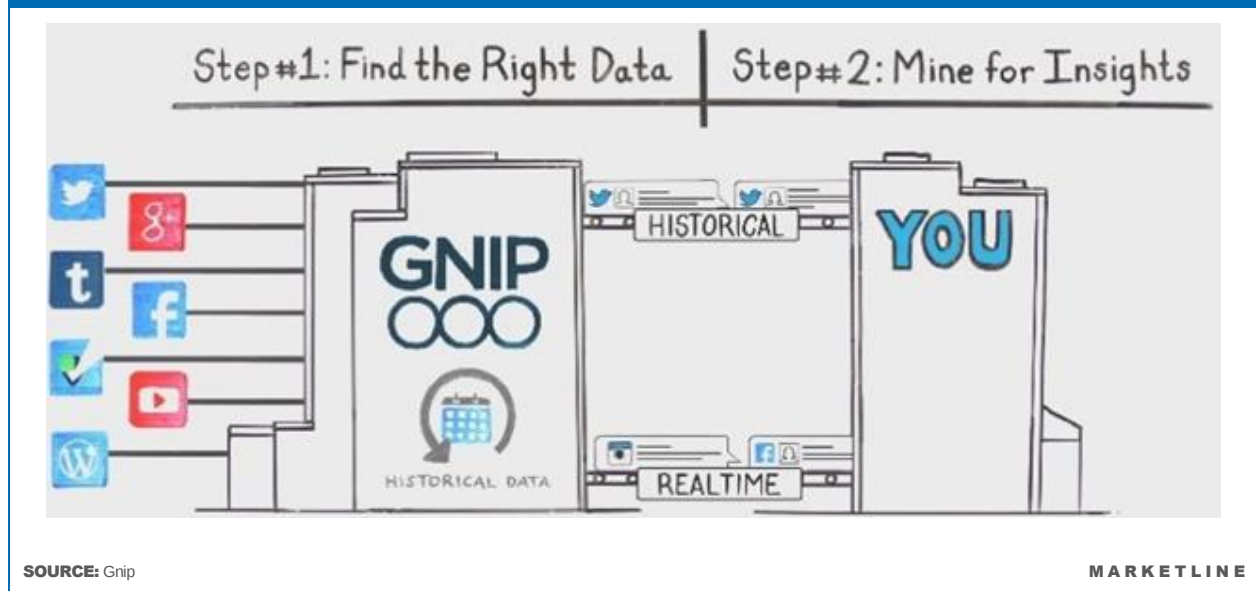
With this data, Target was able to send coupons to customers it believed to be pregnant, timed to specific stages of their pregnancy, therefore beating the competition. There was even a recorded incident of a disgruntled father demanding to see a Target store manager and asking why the company was sending his high school daughter coupons and marketing for baby products. The father also questioned if the company was encouraging his daughter to get pregnant. It turned out the daughter was pregnant unbeknownst to the father, showing the power of Big Data analytics. For a detailed study of customer analytics with Big Data, see MarketLine case study, [*'From token loyalty to meaningful relationships: How loyalty programs and Big Data Analytics are facilitating CRM in the retail sector'*](#) (ML00017-011, April 2014).

Sentiment analysis can be used to tailor public interactions and optimize production processes

In order to understand what customers and media outlets are thinking about a product, sentiment analysis can be used alongside swathes of data generated through social media and the press. For example, NASCAR has around 75 million fans in the US and uses HP HAVEn technologies to help it understand the data that can be collected about the public perception of NASCAR through social media, broadcast media, and traditional press outlets. This therefore includes text, image and video content. Tying all of this information together through an analytics and visualization package, NASCAR is able to understand public sentiment and react to fan and press impressions of the sport extremely quickly. NASCAR is therefore able to tailor its public interactions based on a real-time understanding of public perceptions.

Whilst large organizations such as NASCAR are able to capture all the information about their own industries themselves, companies such as Gnip also act as repositories of a wide array of social information for others to access. For example, Gnip includes data that has been mined from websites such as Twitter, Tumblr, Wordpress, Facebook, Instagram and YouTube.

Figure 2: The Gnip approach to social data



Although companies do not divulge how they are actually analyzing public perception, the only real limit to sentiment analysis is the skill and imagination of the data scientists utilizing the Big Data frameworks. For instance, understanding public sentiment could be a key driver of strategic advantage in terms of stocking the right products at the right time, vastly improving 'just in time' production processes.

One can imagine mining the data from the public Twitter stream in order to predict 'the next big thing' before it really takes off. In order to do this computationally, a company might apply an algorithm in order to understand how a particular flow of public information impacts on product sales. Social network analysis and the analysis of complex adaptive systems are only in the early stages of academic development but certain models already stand out as having the potential to be utilized in mapping public sentiment in relation to product sales.

For example, mathematical models have been created to mirror the actions of ants as they randomly search for and find a good source of food; the ants give off signals to each other as they move between the nest and the food source meaning that eventually all of the ants can find the food. The point at which a significant number of ants move to the new food source can therefore be predicted. It is not much of a step to imagine business intelligence analysts applying similar models to people randomly talking about products on social media outlets and therefore being able to predict which products might be successful in real-time. Such systems could then be used to provide real-time demand forecasts and therefore the ability to optimize production processes. This example shows the importance of having links between researchers and businesses in order to fully exploit Big Data technologies.

Analysis of the general public can lead to improved cost and time efficiencies

Big Data has been used in terms of crime prediction and prevention through Palantir technologies. For example, through Palantir, the Salt Lake City Police Department relies on dozens of large datasets to investigate and prevent crime. In 2010, investigators had access to more than 988,000 documents, 40,000 mug shots, 117,000 arrest records, and 520,000 case reports. Searches on these files through Big Data technologies can be completed in minutes rather than days or weeks. Following the implementation of these technologies, the Police Department estimates that Palantir has reduced the time it takes to perform a complex investigation by 95%.

A further area in which Big Data has been heralded as being able to provide major insights is population health. IBM has worked with Premier, a large US healthcare alliance, which includes over 2,700 member hospitals and health systems, 90,000 non-acute care facilities and 400,000 physicians. Through Big Data technologies, Premier has compiled an enormous comparative database containing clinical, financial, supply chain and operational information, which actually holds data on a quarter of all discharged hospital patients in the US. The database is able to provide users with detailed clinical outcome measures, amongst other details, so that Premier can improve internal processes and health outcomes.

Whilst IBM does not provide details of the specific analytics they use with Premier, it claims that through one performance improvement combining information from 330 hospitals, more than 29,000 lives have been saved and spending on healthcare has been reduced by almost \$7bn. Furthermore, Premier is in the process of extending its database by integrating information from physician offices, nursing homes, long-term care facilities, other non-acute care settings and home monitoring devices. Once this information has been fed into the database, Premier will have an eerily complete picture of the health of the US population, in real time.

Company processes and operational information become clearer with Big Data

A second aspect of the introspective use of Big Data has been through the analysis of company processes and operational information. These uses can be seen in research and development programs as well as through industrial use in terms of equipment monitoring and process tracking.

Previously impenetrable research and development problems have become solvable

Big Data has been useful for many organizations in terms of assisting with complex research and development tasks. A prime example of this is through Big Data analytics being applied to genomic sequences in order to find better treatments for patients. For example, Intel has assisted the University of North Carolina in this regard, which currently holds 200 to 300 terabytes of genomic data on Intel servers. The processes involved in utilizing this dataset are extremely complex because taking a tissue sample from a single patient can generate hundreds of millions of short DNA sequences that need to be reassembled through parallelized computations, into the patient's specific genome. Once a patient's genome has been reassembled in a way that can be read by the Big Data system, it can be compared to the 200 to 300 terabytes of genomic data that has been captured from a larger patient population, in order to detect variations. Linking that information with public domain and other private databases helps the University to understand the impact of genomic variations and whether they are associated with clinical conditions relevant to a particular case. A level of relevancy can then be assigned to each variation and reported back to the relevant physician in order to improve diagnoses and subsequent treatments. The Hadoop MapReduce parallel processing system allows these comparisons to happen at much greater speed than could possibly have been achieved before, with the distributed file system allowing for the analysis of enormous amounts of comparative genomic data.

Industrial usage points to huge cost savings through optimization techniques

The impact of Big Data technologies is wide and complex, with many new uses being found in industry. Again however, many of these types of implementations are ways of monitoring aspects of a business that a company already deals with in some shape or form, relying on internally generated information to inform business intelligence led decisions.

In terms of manufacturing, Oracle claims that Big Data has responded to the fact that sensors and data-capture equipment have multiplied substantially in industrial use. For instance, Oracle gives the example that modern car manufacturers are now able to do a lot more to prevent expensive vehicle recalls due to the analysis of machine-generated information. This is because modern cars contain onboard sensors that continuously monitor the workings of critical aspects of each car's functionality, from tires to brakes to fuel-injection systems. This information is stored in the car's computer, which is then captured by the manufacturer when the car is serviced. The data can then be analyzed to provide insights into any engineering and maintenance issues that might not be picked up in a standard service due to the subtle nature of the results. When combined together with the results from millions of cars, these subtle results take on more significance. As a result, by studying the data streams, manufacturers can use the information to improve products and project more accurate service cycles. Small defects that might turn into damaging product recalls can therefore be found early, enabling production processes to be updated as soon as possible, thus minimizing these expensive scenarios.

Splunk has used its Big Data technologies to assist the New York Air Brake Company (NYAB) with fuel efficiencies for its customers. NYAB is a leading supplier of braking systems to the railroad industry. Fuel efficiency is extremely important in the rail freight business as it acts as a competitive advantage for the industry in comparison to other forms of freight transport. NYAB therefore uses its 'Train Dynamic Systems' technology to optimize fuel efficiency for its customers but, prior to the use of Big Data technologies, it had issues with the speed at which it could provide freight companies with information on a given train's performance. Splunk has been used to augment this fuel optimization system by collating multiple formats of machine-generated data so that real-time information can be given on train performance. As such, NYAB can now help companies model a whole host of real-time problems and scenarios, as well as their likely impact on costs, such as whether a train loaded with coal arriving half an hour later than expected might actually be beneficial if it saves 2% on fuel costs.

IBM also notes that Big Data can be used to help with industrial drilling checks, production surveillance and optimization, telecoms network analytics, distribution load forecasting and scheduling for energy outputs, condition based maintenance systems, smart meters, the observation of weather correlation effects, capacity and pricing optimization, and predictive maintenance optimization.

BIG DATA 2.0 – HOLISTIC INTELLIGENCE

In light of the historic trend towards analyzing internally-focused data and the recent use cases of Big Data technologies, the question arises as to where Big Data might go next once all internal operations have been fully optimized. Two ways of looking at this are in terms of externally generated information outside of a company's standard parameters, and the feasible limits of Big Data analytics. The result may well be a broader business intelligence perspective, including non-computational insights and the use of wider sources of information, forming what might be termed Big Data 2.0 - Holistic Intelligence.

Where do we go from here?

Although it might not seem like it, there must be limitations to what Big Data analytics can achieve. These limitations can be found to some extent, and speculatively, in the limits of computers themselves. Furthermore, as with any period of introspection, some things are going to be missed. As an analogy, the current trends in Big Data have the appearance of knowing everything about what a car is doing, its likelihood of engine failure, everything about the person who is driving it and their capabilities, how well the car is gripping to the road, and many other points of information, but not knowing that the car is about to drive off a cliff or that another car is about to crash into it. Data about wider markets, economies and competitors are therefore ripe for Big Data analytics.

What happens when everyone is using Big Data?

At the point when the majority of companies are using Big Data analytics, it will be as if every company has its own private Google repository, storing all of the information relevant to its business in a way that can be modeled and analyzed. These 'data lakes' will become assets in themselves and companies will have a tight grip on the knowledge required to exploit their particular customer bases. The proportion of global data that is no longer 'open' or publicly available will become substantial, leading to an exponential increase in the value of that data. As each company gains control of its own data lake, significant value will also be held in the skill and imagination of the data scientists employed to maximize the functionality of that information.

In a Big Data laden world, there is therefore likely to be a greater proliferation (and increase in complexity) of advanced analytics. Mirroring this process, there might also be a parallel shift away from computational intelligence in order to gain a wider understanding of a company's business, though how this will manifest is difficult, if not impossible to quantify.

It is going to get harder before it gets easier

Given that Google already claims that its Cloud Dataflow service can handle volumes of data 'of any size', the real arena for competition is going to be in the power of the analytical techniques being used to gain greater levels of insight. Those Big Data providers offering easy graphical user interfaces as part of their sales pitch are therefore going to need continual updates in order to keep pace with the development of new and more complex algorithms being used to analyze data. Those providers instead offering the flexibility of something like the R statistical programming language, which allows skilled programmers to design their own algorithms and create tailored mathematical models depending on the data being used, may therefore have made the right choice for now.

At present, most business intelligence operations use various forms of statistical analysis, from finding averages to creating predictive models. Some other areas of mathematics from which Big Data analytics might pick and choose algorithms include complexity theory (the analysis of non-linear relationships and understanding large effects from small causes), social network analysis (the analysis of individual and group interactions), complex adaptive systems (the analysis of evolving individuals and groups), and game theory (the analysis of strategy). Big Data business intelligence analysts are therefore likely to need a whole host of complex skills in the short term. Eventually, however, many of these algorithms are likely to be integrated into a market-leading business analytics platform, much like Excel has been.

Machine learning and the use of artificial intelligence will also be areas in which analysts are likely to operate. At present, these areas are particularly important for the analysis of visual content by using algorithms to classify images; the more images the algorithm is applied to, the more accurate the classifications become. Such techniques are also being used to update predictive models and determine the next best actions for a particular company objective.

The Chinese room argument and the limits of computers

One way of looking at Big Data is to see it as another step on the road to realizing artificial intelligence. In a similar way to how the human brain memorizes vast amounts of varied information and can answer most questions quickly, but may need to 'sleep on it' for more difficult issues, servers are now able to store formats of digital data that cover most types of visual and sonic information. This information can then be queried quickly with relatively simple issues (known as low latency processing), or queried over long periods for more complex problems (overnight batch processing). As such, the absolute limits of Big Data analytics are closely tied to the limits of what can be done computationally.

John Searle first published his Chinese room argument in 1980, which is a useful way to understand the potential limits of Big Data technologies. The argument goes that if a human mimicked the operations of an advanced computer program precisely, it would demonstrate the lack of cognition that seemingly intelligent computers have at a higher level. For example, Searle, who has no knowledge whatsoever of the Chinese language, imagines being sat in a room following a computer program designed for translation, such as Google translate. He is passed Chinese characters under the door and follows the logical steps of the program to produce appropriate mixtures of Chinese symbols to convince someone outside that there is a fluent Chinese speaker in the room. As such, even though computer programs may appear to attain a level of understanding that can be used to pass something like the Turing Test (the ability of a computer program to answer questions and convince the investigator that it is human), computers do not have any real understanding of meaning and therefore cannot provide comprehensive insights.

This argument provides a tentative indication that once Big Data analysts master a wealth of algorithms that can be applied to large datasets, companies will begin to seek non-computational insights and find those areas of perception that cannot be performed through a computer program. In other words, human understanding and communication will gain more value in terms of the ability to comprehend Big Data output and the ability to understand what questions need to be asked about a business. As such, data visualization platforms such as Tableau may exploit this because they already provide a non-computational level of communication through the use of easy to understand info-graphics.

Sourcing external data is a logical next step in the short-term

Aside from increasing complexity and non-computational insights, a logical next step for Big Data business intelligence is to make use of a wider variety of information. In short, Big Data is likely to get even bigger, shifting from marketing and internal optimization to a broader understanding of the business climate and therefore the ability to inform what long-term strategies to undertake.

As noted earlier, in 2012, NASA held 32 petabytes of information on the earth's climate and expects to hold nearly 350 petabytes by 2030, and it is only with new technologies that problems such as modeling satellite images of the earth's climate can be solved. Equally, problems such as understanding the impact of macroeconomic interactions, social and political changes, and competitive landscapes are likely to benefit from Big Data technologies.

For example, Big Data technologies could be used to model competitive and supply chain networks that could be informed through text mining company announcements and reports. In this way, companies could obtain a near real-time understanding of competitor strategies and how they might impact their own business, particularly through the development of game theory based algorithms. As another example, the ever-increasing level of digital information could also allow for the analysis of competitive environments on a city-by-city level, which would involve an inordinate amount of data collection to build a national view. Although these are only possibilities, it is clear that a much broader understanding of a company's environment could be achieved through Big Data by leveraging existing information with these technologies.

Collecting external data on its own is unlikely to be enough to match the expectations of companies that have used Big Data analytics for their own internal purposes. The transformation of externally generated information into valuable insights, such as broader predictions, complex analytics and recommendations, could therefore be a second string to the Big Data industry. Being able to understand the broader business environment through the use of Big Data technologies, in a manner that can be easily communicated might thus form the type of wide-ranging, holistic intelligence that is currently lacking from the Big Data industry.

CONCLUSIONS

The future belongs to those who ask the right questions

Big Data technologies can be viewed simply as a contemporary way to store, query and model information. The essential reason that businesses have been quick to adopt Big Data is to do with the level of insight that can be obtained and the types of problems that can be solved through Big Data technologies.

Whilst a wide range of industries are using Big Data technologies, the dominant operational feature in recent years has been a restriction to internally-focused data, either from information about the company itself and its processes, or through the interaction of customers with that company. This has led to strategic advantages being gained by analyzing company information at a whole new level of detail. Due to the level of internal analysis being done with Big Data technologies, Big Data providers are in a unique position to become business intelligence advisors in their own right in terms of exploiting the potential of a company's specific level and type of data. Computer and data scientists are therefore way ahead of the curve in terms of business intelligence insights as they are able to harness the computational and storage power of Big Data.

Two ways of looking at the potential future of the industry are in terms of externally generated information and the feasible limits of Big Data analytics. Looking at the industry in this way leads to the conclusion that the Big Data industry will advance through the sourcing of macro-level information and utilization of ever more complex analytics until a point is reached whereby non-computational insights will become more and more valuable. The combination of such processes has the potential to form something that might be termed holistic intelligence. In terms of competitive advantages to be gained from Big Data technologies, the future therefore belongs to those who are able ask the right questions, rather than those who can capture the most information.

APPENDIX

Definitions

Computer algorithm: This is effectively a logical sequence of commands designed to tell a computer which actions to carry out, through a step-by-step process, to achieve a desired result. Algorithms can be very simple or very complicated. For example, a simple algorithm might average a set of numbers by first adding the numbers together to find a total, secondly counting the amount of numbers in the set, thirdly dividing the total by the amount of numbers in the set, and finally displaying the result. The actual algorithm would be written in mathematical form. A far more complicated algorithm might be used with the goal of improving the accuracy of the final result if more and more information is input into the algorithm (machine-learning). For example, this type of algorithm might be used to carry out a step-by-step process that gets better and better at classifying different images.

Big Data: In one sentence, Big Data is a common term used to describe sets of data that are so large and multi-faceted that wholly new technologies have had to be developed in order to deal with them in a useful manner.

Parallel processing: This is a particular way that advanced computers can function to improve performance. Specifically, computer processes, such as the carrying out of algorithms, can be run in parallel i.e. they can be run simultaneously by using multiple processing units. In terms of data processing, with software such as Hadoop MapReduce, singular data processes (such as finding and remembering a piece of information from millions of files) can be carried out in parallel rather than consecutively, i.e. more than one piece of information can be found at any one time, thus increasing the speed of results.

Programming language: This is the specific language used to communicate with a computer and thus make it function. There are thousands of different languages, each one with a different purpose and syntax. For example, the R statistical programming language is used to write statistical algorithms for a computer to perform on a given data set. The SQL (Structured Query Language) on the other hand is used to determine the functions of a particular type of database.

Sources

All Analytics: NASA takes Big Data to new heights for you & me, September 2012

http://www.allanalytics.com/author.asp?section_id=1411&doc_id=250894

Amazon: Big Data Analytics Options on AWS, December 2014

http://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf

Amazon Web Services: Big Data, accessed February 2015

http://aws.amazon.com/big-data/?nc1=f_dr

Apache Software Foundation: Hadoop, accessed February 2015

<http://hadoop.apache.org>

Apache Software Foundation: Hadoop Distributed File System, accessed February 2015

<http://wiki.apache.org/hadoop/HDFS>

Apache Software Foundations: Spark, accessed February 2015

<https://spark.apache.org>

Bloomberg: Latest round of funding values Palantir, the CIA's favorite startup at \$9bn, December 2013

<http://www.bloomberg.com/bw/articles/2013-12-05/palantir-the-cias-favorite-startup-is-now-valued-at-9-billion>

Cloudera: CDH, accessed February 2015

<http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html>

Cloud Times: Internet of Things will multiply the digital universe data to 44 trillion GBs by 2020, April 2014
<http://cloudtimes.org/2014/04/17/internet-of-things-will-multiply-the-digital-universe-data-to-44-trillion-gbs-by-2020/>

CRN: Pivotal laying off 60 employees mostly on Big Data side of business, Kevin McLaughlin, November 2014
<http://www.crn.com/news/cloud/300074934/sources-pivotal-laying-off-60-employees-mostly-on-big-data-side-of-business.htm>

CSC: Climate change simulation – NASA's weather supercomputer, accessed February 2015
http://www.csc.com/cscworld/publications/81769/81773-supercomputing_the_climate_nasa_s_big_data_mission

eWeek: Trends to Expect in Business Intelligence, Big Data in 2015, Chris Priemesberger, December 2014
<http://www.eweek.com/innovation/trends-to-expect-in-business-intelligence-big-data-in-2015.html>

Google: Bigtable - A Distributed Storage System for Structured Data, 2006
<http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf>

Google: Cloud Platform - Data Flow, accessed February 2015
<https://cloud.google.com/dataflow/>

Google: The Google file system, October 2003
<http://research.google.com/archive/gfs-sosp2003.pdf>

Google: MapReduce – simplified data processing on large clusters, 2004
<http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf>

Hortonworks: Data Platform, accessed February 2015
<http://hortonworks.com/hdp/>

IBM: 2013 Annual Report, accessed February 2015
http://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf

IBM: Big Data, accessed February 2015
<http://www-01.ibm.com/software/data/bigdata/industry.html>

IBM: Data-driven healthcare organizations use big data analytics for big gains, accessed February 2015
<http://www-01.ibm.com/software/data/bigdata/industry-healthcare.html>

Intel: Big Data genomics sequencing, accessed February 2015
<http://www.intel.com/content/www/us/en/big-data/renci-peer-story.html>

Microsoft: Microsoft Azure HD Insight, accessed February 2015
<http://azure.microsoft.com/en-gb/services/hdinsight/>

MongoDB: Big Data – Examples and Guidelines for the Enterprise Decision Maker, February 2015
<http://www.mongodb.com/lp/white-paper/big-data-nosql>

MongoDB: Big Data explained, accessed February 2015
<http://www.mongodb.com/big-data-explained>

NASA: What is NASA doing with Big Data today? October 2012
<http://open.nasa.gov/blog/2012/10/04/what-is-nasa-doing-with-big-data-today/>

Oracle: Big Data Appliance, accessed February 2015

<https://www.oracle.com/engineered-systems/big-data-appliance/index.html>

Oracle: Business Analytics – Big Data meets business intelligence, September 2014

<http://www.oracle.com/technetwork/issue-archive/2014/14-sep/o54ba-2279189.html>

Palantir: Fighting crime in Salt Lake City, accessed February 2015

https://www.palantir.com/wp-assets/wp-content/uploads/2012/06/ImpactStudy_SLCPD.pdf

Pivotal: Big Data, accessed February 2015

<http://pivotal.io/big-data/pivotal-big-data-suite>

Splunk: Splunk at New York Air Brake, 2014

<http://www.splunk.com/content/dam/splunk2/pdfs/customer-success-stories/splunk-at-new-york-air-brake.pdf>

Splunk Enterprise, accessed February 2015

http://www.splunk.com/en_us/products/splunk-enterprise.html

United States Postal Service: Postal Facts, 2014

<https://about.usps.com/who-we-are/postal-facts/postalfacts2014.pdf>

University of Berkeley: MAD Skills – new analysis practices for Big Data, March 2009

<http://db.cs.berkeley.edu/jmh/papers/madskills-032009.pdf>

University of Cambridge, Turing Gateway to Mathematics: Coping with Big Data, accessed February 2015

http://www.turing-gateway.cam.ac.uk/cbd_jan2015-programme

UCL: The theory of Big Data, accessed February 2015

<http://www.ucl.ac.uk/bigdata-theory/accepted-abstracts/>

University of Stanford: The Chinese room argument, accessed February 2015

<http://plato.stanford.edu/entries/chinese-room/>

University of Texas: Center for Big Data Analytics, accessed February 2015

<http://bigdata.ices.utexas.edu>

VSAT: In 2013 the amount of data generated worldwide will reach four zettabytes, June 2013

<https://vsatglobalseriesblog.wordpress.com/2013/06/21/in-2013-the-amount-of-data-generated-worldwide-will-reach-four-zettabytes/>

Wikibon: Big Data vendor revenue and market forecast, Jeff Kelly, February 2014

http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

World Economic Forum: What are the top data analytics platforms of 2015, Bernard Marr, February 2015

https://agenda.weforum.org/2015/02/what-are-the-top-data-analytics-platforms-of-2015/?utm_content=bufferd28e9&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Further Reading

[The open source and graph database revolutions: How software companies manage knowledge as a strategic asset](#) - MarketLine Case Study

[From token loyalty to meaningful relationships: How loyalty programs and Big Data Analytics are facilitating CRM in the retail sector](#) - MarketLine Case Study

[Google Inc.](#) - MarketLine Company Profile

[Microsoft Corporation](#) - MarketLine Company Profile

[Oracle Corporation](#) - MarketLine Company Profile

[Global Software & Services](#) - MarketLine Industry Profile

Ask the analyst

We hope that the data and analysis in this brief will help you make informed and imaginative business decisions. If you have any questions or further requirements, MarketLine's research team may be able to help you. The MarketLine Research team can be contacted at ReachUs@MarketLine.com.

About MarketLine

At MarketLine, we deliver accurate, up-to-date insights on over 30,000 companies, 300 industries, and 215 countries, as well as the latest news and financial deal information from within your market and across the globe.

Established in 1997 when the Internet was in its infancy, we recognized the need for a convenient and reliable data service to help our clients understand local and global markets and the companies operating within them.

In today's information-rich world, sifting fact from fiction to pick out what's relevant and what's up to date has become the new 'holy grail' in business information provision.

Our 170 dedicated research professionals aggregate, analyze, and cross-check facts in line with our strict research methodology, ensuring a constant stream of new and accurate information is added to MarketLine every day.

Disclaimer

All Rights Reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher, MarketLine.

The facts of this report are believed to be correct at the time of publication but cannot be guaranteed. Please note that the findings, conclusions and recommendations that MarketLine delivers will be based on information gathered in good faith from both primary and secondary sources, whose accuracy we are not always in a position to guarantee. As such MarketLine can accept no liability whatever for actions taken based on any information that may subsequently prove to be incorrect.

MARKETLINE | CHRISTCHURCH COURT | 10-15 NEWGATE
STREET | LONDON EC1A 7AZ
T: +44 20 3377 3045 | F: +44 161 238 4141
E: REACHUS@MARKETLINE.COM
WWW.MARKETLINE.COM

MarketLine