Big Data

Will Hogan
4th Year Software Development
Research Methodologies in Computing and I.T.

Abstract—In modern day computing, the term Big Data is used to describe large amounts of data, this data is being created and stored by human made devices on a massive scale globally. The purpose of this review is to explore the world of Big Data and look at it from the ground up in terms of what data actually is, how it's structured and why it's used by organisations. We'll take an in-depth look at Big Data Analytics and how Data Mining can be extremely useful for uncovering information relating to current trends or customer spending. We'll explore the importance of why organisations store and access Big Data and subsequently see that this yields key benefits such as Cost Reduction and improved decision making. Finally, we'll take a look at the direction Big Data could take in the immediate future and see how new technologies like Apache Hadoop and Apache Spark may contribute to this by using clustering and MapReduce to condense and refine large data sets.

1 Introduction

We live in an ever-changing world and technology is one of the things that's moving quickly. There are millions of devices globally that are storing, sending and receiving large amounts of data and there are multiple reports suggesting that the rate of data creation will continue to grow at a rate between 40 and 60% a year [1]. IBM have stated that there are 2.5 quintillion bytes of data created every day and that the last two years has seen 90% of the world's data created [2]. This data is being created by a number of different devices and sources, for example, mobile phones, that have become much more than devices that make phone calls. The University of Cambridge suggests that by 2020, 80% of the worlds population will own a mobile phone [3]. With these statistics in mind, it's fair to say that data creation will continue to increase. Social media has seen a dramatic increase in usage, with reports suggesting that Facebook is dealing with a billion content information queries per day [4]. But it's not just Social Media that's creating large amounts of data, Netflix is accumulating billions of viewer ratings, with members searching and adding millions of items each day [4]. It's also worth highlighting that with these increases in data creation, comes the inevitable increase in Data related positions and careers as a by-product. The UK government have reported that they predict an increase in demand for Big Data staff of between 13 and 23% [5] between 2016 and 2017. To add further to this domino effect, it's vitally important to highlight that data needs to be stored somewhere, especially if it's going to be of any use in the grand scheme of things. With this fresh in our minds, we need to think on a much larger scale that surpasses spreadsheet storage or smaller traditional databases methods and start looking at something that works on a much larger scale, something that can deal with the vast amounts of data and information being circulated globally. This increase creates a need for better software to handle the data, bigger and better servers to store it and more staff to deal with it.

1.1 A Comprehensive look at Data

The term Big Data has been around for a number of years, but it's really become more relevant with the increase in Social Media usage, contributed to by big names like Facebook, Twitter and Instagram. But if we take an in-depth look into what data actually is at a low level, we can gain a much more insightful perspective into how it operates on a larger scale. Essentially if we break any type of data down to its most raw component, data in technological terms is simply just a collection of 1's and 0's that form binary code. Humans have mapped binary code into the more human readable form, which is known as the ASCII (American Standard Code for Information Interchange) character encoding standard [6]. This standard contains all the letters of the alphabet and their equivalent binary values. As we start to put words and sentences together it becomes clear that there will always be a binary representation at the lowest level, no matter how large the body of text is. The same concept holds true for other forms of data also, for example, a digital image that may be comprised of different pixels will essentially have binary at its heart as each pixel must have a binary equivalent.
Data can be structured or unstructured and to explain what this means, we might say that structured data has a specific datatype associated with it, like an integer, string or Float. From a Relational database model the structured data may also be normalised. On the flipside, unstructured data is pure raw data and doesn't necessarily comply with any format or type. As time has moved on and technology has advanced, it's hard to think of things in terms of bits and bytes, however, in the world of Big Data, the words Petabytes, Exabytes, Zettabytes and even Yottabytes are becoming common place. To help put this into perspective, consider the following information, in Figure 1 which is taken from the School of Information Management and Systems, Berkeley University [7] in 2003;

Figure 1: Data Size and Examples

Data Size	Example
100 Kilobytes	A low resolution photo.
5 Megabytes	The complete works of Shakespeare.
100 Gigabytes	A library floor of academic journals.
10 Terabytes	Print collections of the U.S. Lib. of Congress.
200 Petabytes	All printed material.
2 Exabytes	Total volume of information generated in 1999

With these statistics, we can build a picture as to how big, big data can actually get, bearing in mind the fact that this paper was written in 2003. As the years have elapsed, data has grown exponentially and with that comes new terminologies and buzz words that fit specific circumstances. As detailed by Doug Laney [8], who outlines a well-known definition (also called 3Vs) Volume, velocity, and variety. The definition of 3Vs implies that the data size is large, the data will be created rapidly, and the data will exist as multiple types and captured from different sources, respectively.

1.2 Big Data Analytics

Big Data analytics is a combination of Big Data and Analytics. But firstly just for clarification purposes, let us define what exactly Data Analytics is. Data analytics is the science of scrutinising raw data with the hope of discovering trends or habits in specific business areas. It's important to note that Fayyad et al [9], mentioned back in 1996, that due to the emerging field of Knowledge Discovery in Databases(KDD), that there was an urgent need for new computational tools and theories to aid humans in discovering new and useful information. Furthermore, if we fast forward to 2009, a survey from the TDWI (transforming data into intelligence) [10], revealed that 35% of Organisations have reported practicing some form of advanced analytics, whereas 85% explained that they would be practicing it within the next 3 years. Based on this, it seems that Fayyad and his colleagues were right on the mark and while it's important to acknowledge their foresight, it's also equally important to ask the question why this happened? The answer is simply because it needed to accommodate the growing data and as I have mentioned earlier, this in turn creates a snowball effect that requires other sectors to adapt accordingly.
Change is something that happens in many different environments, business being just one. But analytics itself hasn't just helped to assess situations, it has helped us discover what has changed, it's then with these discoveries that we can react accordingly and make the right decisions [10]. To add to this, it's worth noting that companies and organisations opted to use Data Analytics as a chance to beat the worldwide recession and help build a path to recovery.

1.2.1 Data Mining

One of the methods that organisations can use in order to benefit from analysing their data, is Data Mining. To quote Oracle;

"Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis." [12].

Let's explore this a bit further. Figure 2 outlines what most Data mining algorithms contain, initialisation, data input, data scan, rules construction and also rules update operators [11].

Figure 2: Data Mining Algorithm

Input data D
Initialize candidate solutions r
while *the termination criterion is not met* **do**
  d = *Scan(D);*
  v = *Construct(d, r, o);*
  r = *Update(v);*
end
Output rules r;

Let's break this down piece by piece in order to understand it better. D is the actual raw data, d is the data entered in by the scan operator, r are the rules, o are the predefined measurement and v the candidate rules. The scan, construct and update operators will continue to loop until the search criteria have been met.

There are different methods used in Data Mining, in Figure 3. we see another method used called Clustering. With this method data can be separated into different labeled groups using *k-means* [13].

Figure 3: *k-means* Algorithm

Input data D
Randomly create a set of centroids c
while *the termination criterion is not met* **do**
  v = *Assign(D, c);*
  c = *Update(v);*
end
Output centroids c;

To detail what's happening in the above, firstly we create a random set of Centroids, these centroids represent the patterns created by user input and are divided into specific groups[14]. After this, the assignment operator checks the distance between the centroids and the patterns in order to ascertain which group each pattern belongs too. The formula used to calculate this(1) can be written as so;

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} D(x_{ij} - c_i) \quad (1)$$

Where SSE is the squared sum of errors, which is used to measure cohesion of the data mining results. In the above formula k is entered by the user; ni the number of data in

the ith cluster; *xij* the *j*th datum in the *i*th cluster; *ci* is the mean of the *i*th cluster [14].

With Data Mining, the most common method to measure distance is the Euclidean distance(2), defined as;

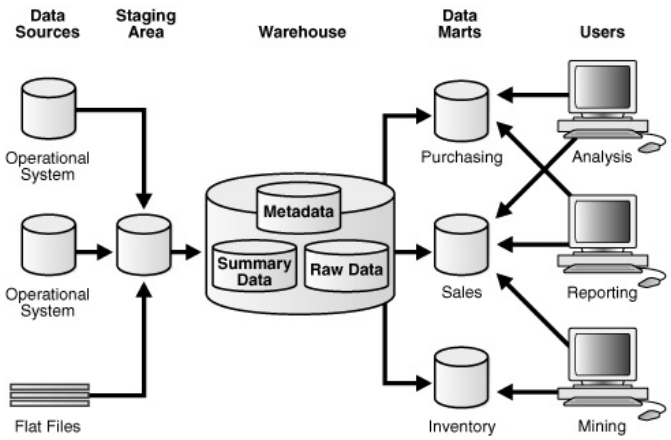
$$\mathbf{D}(\mathbf{p_i}, \mathbf{p_j}) = \left(\sum_{l=1}^{d} |p_{il}, p_{jl}|^2 \right)^{1/2} \quad (2)$$

Where *pi* and *pj* are the positions of two different pieces of data.

1.2.2 Data Warehousing

Like the buzz words Big Data and Data Mining, Data Warehousing is another term that's being widely used in the world of Big Data & Analytics. Essentially a Data Warehouse is a large-scale database that stores data. But it's what it stores that defines it and sets it apart from other databases. Oracle have stated that it's a Relational Database that's not designed for transaction purposes, but more for querying and Analysis [15]. To put this into perspective, you might be looking to see trends, spikes and customer spending as opposed to searching a database for when a particular employee started in a given organisation. To enable this sort of broad scale searching and analysis, several databases from different sections of an organisation, feed their data into the Data Warehousing which essentially stores everything that's being sent to it. From here, the data can be split and separated into department specific databases known as Data Marts. The method used to get to this point is called ETL (Extraction, Transformation, and Loading). Extraction means to retrieve the data from the various Database sources, Transformation is turning that data into something useful and Loading refers to the process of saving the data to the Data Warehouse, illustrated in Figure 4 [15].

Figure 4: Data Warehousing Example



This type of concept may seem convoluted to the average client, but what will make sense is how it will affect their organisations financially. Inmon states [16], that Data Warehousing significantly reduces the cost of information, with the result that organisations that use a Data Warehouse, can access a specific piece of information for $100, as opposed to an organisation that doesn't have a Data Warehouse, who will access the same piece of information for $10,000.

1.3 Benefits of Big Data & Analytics

If we analyse the information detailed in this review, it becomes clear that there are benefits to organisations who invest time and money in Big Data and Data Analysis. Here are some of the key benefits;

1) **Cost Reduction:**
As outlined in the last section, Data Warehousing and Data Analysis, in general, can yield substantial financial savings. But it's not just the implementation that's cost effective. According to Thomas Davenport [17], most companies and organisations are using Big Data and Warehousing to accompany their existing databases, as opposed to replacing the traditional architecture that's currently in-situ. This is a huge cost reducing benefit from a maintenance perspective alone.

2) **Improved Decision Making:**
To enable organisations to make proper decisions, they need information. The more information they have the better insight they have into what's been happening and be able to plan effectively for what could happen. To gain a visual insight into this and thanks to advancements in technology, snapshots can be taken of a specific period in time that display a visual perspective that can highlight peaks and troughs in particular areas of business.

3) **New Products and Services**
With the explosion of Data in the last number of years, comes new products and services that can be tailored to customers needs. These services can help customers plan new advertising and marketing campaigns based on information they are receiving from Big Data & Analysis.

1.4 Big Data, the next steps

The creation of Data is unlikely to stop, so the next steps could involve capturing relevant data in as many different ways as possible that will of course benefit companies and organisations. Apache Hadoop [18] is an open source framework for storing and processing large datasets using clusters. They use what's called a MapReduce method to condense or reduce large amounts of data into smaller parts. Although Hadoop has been around for a number of years, it's worth highlighting that they'll be involved in Big Data for the foreseeable future due to the fact that they are highly involved in Big Data Warehousing, with many companies globally using their services [19].

Apache Spark is worth mentioning as a contributor to the current and possible future world of Big Data. It's a fast, general purpose engine used for large-scale data processing [20]. It focuses on Speed, that allows for greater MapReduce performance in memory, than its counterpart Hadoop. It also allows for a general ease of use that enables applications to be written quickly in languages such as Java, Scala, Python and R. They claim that this allows you to combine

libraries such as SQL and machine learning programs into one application [20]. Apache Spark is currently being used by companies such as Amazon, Yahoo, Nokia, NASA JPL (Jet Propulsion Lab) and Hitachi Solutions [21].

2 Conclusion

In this paper, we've had an in-depth look into the world of Big Data and Data Analytics. Firstly, we looked at how Big Data is used by companies and organisations, where we gained some knowledge about how data works with the help of some interesting statistical analysis. We looked at Data in its most basic form and looked at some size comparisons. We've discussed in great detail about Big Data, Data Analysis, Data Mining and how data operates on a much larger scale with the help of Data Warehousing. We outlined that reduction in cost and improved decision making are key benefits as a result of using Big Data & Analytics. Finally, we looked at some key new technologies that are working with large data sets and data warehousing to produce more efficient results.

Upon reflection and based on the topics covered in this review, one might see that certain outcomes are a by-product or result of another action with regard to Big Data. Essentially what this means is that as data creation accelerates, it creates a need in other areas. As we create more data, well need to store it somewhere and as this data grows it will eventually contain useful information, information that can be searched for something of benefit to an organisation. Finally, this creates a need for employers to fill gaps by employing and training more people as their business needs change and evolve. Based on the rate at which technology is advancing, it's fair to estimate that data will continue to grow and as it grows, humans will find a way to evolve with it in order to benefit from it.

References

[1] OECD (2013) New sources of growth- knowledge based capital. OECD, Paris. http://www.oecd.org/sti/inno/knowledge-based-capital-synthesis.pdf.

[2] Harness the Power of Big Data The IBM Big Data Platform https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[3] http://www.cam.ac.uk/research/discussion/talkin-bout-a-revolution-how-to-make-the-digital-world-work-for-us

[4] Amatriain X (2013) Beyond Data: from user information to business value through personalized recommendations and consumer science, CIKM13. San Francisco, CA, USA

[5] Seizing the data opportunity: A strategy for UK data capability bis-13-1250-strategy-for-uk-data-capability-v4.pdf

[6] http://www.asciitable.com

[7] How Much data.pdf Release date: October 27, 2003. 2003 Regents of the University of California

[8] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling- Data-Volume-Velocity-and-Variety.pdf.

[9] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag.1996

[10] TDWI Best Practices Report Next Generation Data Warehouse Platforms (Q4 2009), available on tdwi.org.

[11] Big Data Analytics by Philip Russom 2011 http://www.sas.com/content/dam/SAS/en_us/doc/research2/big-data-analytics-105425.pdf

[12] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm

[13] Tsai C-W, Lai C-F, Chiang M-C, Yang L. Data mining for internet of things: a survey. IEEE Commun Surveys Tutor. 2014;16

[14] SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS by J Macqueen, University of California https://pdfs.semanticscholar.org/a718/b85520bea702533ca9a5954c33576fd162b0

[15] https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm

[16] Building the Data Warehouse Third Edition, W.H.Inmon, 2002

[17] Competing on Analytics by Thomas H. Davenport http://www.milwaukeespin.com/Files/competing-on-analytics.pdf

[18] Hadoop. http://hadoop.apache.org/

[19] http://wiki.apache.org/hadoop/PoweredBy

[20] http://spark.apache.org/

[21] https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark