Big Data

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

Abstract—In modern day computing, the term Big Data has become increasingly popular and equally significant. In a nutshell, it's a term that's used to describe large amounts of data. This Data is being created and stored on a massive scale globally. The purpose of this literary review is to delve into the world of Big Data and ascertain where it's used, why organisations want it and where it fits into our modern day society.

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 worlds data created[2]. This data is being created by a number of different devices and sources, for example mobile phones that are so much more than call making devices. The University of cambridge suggest that by 2020, 80% of the worlds population will own a mobile phone[3]. With these predictions and statistics, it's fair to say that data creation will continue to increase. Social media has seen a dramatic increase in usage, reports suggest that Facebook are 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 are 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 / production, there will inevitably be an increase in Data related positions and careers. The UK government have reported that they predict an increase in demand for Big Data staff of between 13 and 23% between now and 2017[5]. To add further to this domino effect, it's important to mention that data needs to be stored somewhere if it's going to be of any use and i'm not just talking about excel spreadsheets or traditional databases methods, but something that works on a much larger scale, 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 operate them.

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 look in more depth at what data actually is we can gain a more insightful perspective, which in turn will help to understand how data works. Essentially if we break any type of data down to it's 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 and as we put letters and sentences together it's worth noting that there will always be a binary representation at the lowest level. Data can be structured or unstructured, with structured there are specific dataypes ie integers, strings and Floats. 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 Pettabytes, Exabytes, Zettabytes and even Yottabytes are becoming common place. To help put this into perspective, consider the following information taken from the School of Information Management and Systems, Berkeley University in 2003[7];

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 in mind, we can build a picture as to how big big data can actually get, bearing in mind the fact that this paper was written thirteen years ago. 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) to

explain this. 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 and his colleagues 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[8]. Furthermore if we fast forward to 2009, a survey from the TDWI(transforming data into intelligence), 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[9]. Based on this, it seems that Fayyad and his colleagues were spot on the mark and while it's important to acknowledge there foresight, it's also equally important to ask the question why this happened? 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[10].

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."[11].

Let's explore this a bit further. Fig. 1 outlines what most Data mining algorithms contain, initialisation, data input, data scan, rules construction and also rules update operators[12].

Figure 1: 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;

Lets 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 Fig 2. we see another method used called Clustering. With this method data can be seperated into different labelled groups using *k-means*[13].

Figure 2: *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 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;

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

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 jth datum in the ith cluster; ci is the mean of the ith cluster; [14].

With Data Mining, the most common method to measure distance is the Euclidean distance, defined as

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

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

1.2.2 Data Warehousing

Hadoop, Casandra, Map Reduce.....

1.3 The Importance of Big Data

1.4 Big Data, the next steps

Mention about "Spark", the new Paradigm [A survey on platforms for big data analytics]

2 Conclusion

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] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag.1996
[9] TDWI Best Practices Report Next Generation Data Warehouse Platforms (Q4 2009), available on tdwi.org.
[10] Big Data Analytics by Philip Russom 2011 http://www.sas.com/content/dam/SAS/en_us/doc/research2/big-data-analytics-105425.pdf
[11] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm
[12] 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
[13] SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS by J Macqueen, University of California https://pdfs.semanticscholar.org/a718/b85520bea702533ca9a5954c33576fd162b0.pdf
[14] Krishna K, Murty MN. Genetic k-means algorithm. IEEE Trans Syst Man Cyber Part B Cyber. 1999;