CS4103 Distributed Systems
Coursework Part 1: Big Data

Student ID: 080010830

March 16, 2012

Word Count: 3887

Abstract

Big data is one of the most vibrant topics among multiple industries, thus in this paper we have covered examples as well as current research that is being conducted in the field. This was done based on real applications that have to deal with big data on a daily basis together with a clear focus on their achievements and challenges. The results are very convincing that big data is a critical subject that will continue to receive further study.

1 Introduction

Big data – in information technology – refers to the extremely large volume of data that needs to be captured [1], stored [2], searched [3, 1], shared [4, 1], analysed [2] and visualised [5, 1]. The exponential growth of these datasets can result in exabytes[1] or even zettabytes[2] of information. For example, telecommunications networks have seen their capacity to exchange information grow from 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 and predictions say that it will reach 667 exabytes annually by 2013 [6]. Furthermore, to put these numbers into perspective, 5 exabytes of information is equal to "all words ever spoken by human beings" [7, 8, 9] and if we add all the combined capacity of all the computer hard drives that were available in the world in 2006 the total amount of free space would be approximately 160 exabytes [10]. However, this storage capacity is increasing at an astonishing rate and a proof of that is Seagate's report that during the 2011 fiscal year alone, they have sold hard drives of a combined capacity of 330 exabytes [11].

These impressive statistics and the fact that more people than ever before interact directly with data [6] makes the analysis of big data very relevant, if not crucial.

[1] 1 EB = 10^{18} bytes = 1 000 000 000 gigabytes = 1 000 000 terabytes
[2] 1 ZB = 10^{21} bytes = 1 000 000 000 000 gigabytes = 1 000 000 000 terabytes

2 Examples

Data went from scarce to abundant in the last few decades, bringing on one hand extensive benefits but on the other hand a number of difficulties. Furthermore, data is continuously being gathered at an ever increasing rate due to the ubiquity of "information-sensing mobile devices, aerial sensory technologies also known as remote sensing, software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks" [12]. We will expand on the benefits and disadvantages of big data in a few key scientific and industrial applications that are currently facing them.

2.1 Scientific Applications

The main scientific applications where scientists work in a regular basis with tremendous amounts of information include meteorology, genomics, connectomics, complex physics simulations, biological, and environmental research [12]. It is important to analyse each of these sciences in detail because in each and every case there are unique benefits and difficulties being faced.

2.1.1 Meteorology

Meteorology is "the science dealing with the atmosphere and its phenomena, including weather and climate" [13]. Although, it might give the impression that it is trivial, meteorology can have life-shattering consequences, particularly in the case of hurricanes and tornados. Thus, it is vital that data is examined and understood thoroughly.

After Hibbard [14] concluded that all the most important weather modeling centers had charts and printed maps on the walls, were trying to build 3D Plexiglas, and were generally not indifferent to the great amount of data they had to deal with, it became obvious that they needed a tool to integrate all these different kinds of data into an unified 3D picture. Furthermore, this new tool would allow scientists to look at their large data sets in a much more accessible and interactive way, making it less difficult to comprehend. The program used to combine all this data into an unified visualisation was *Vis5D*, which is an "open source system for interactive visualization of large 5-D gridded data sets" [15]. This was an excellent solution because *Vis5D* was designed to work with data sets produced by numerical weather models. Furthermore, the fact that 3D computer games started to gain popularity resulted in direct improvements on large data sets visualization systems. One of the possible approaches to make big data easier to analyse in this case is to combine a large number of smaller screens into a wall display that contains millions of pixels. However this system would use unnecessary large resources due to the fact that the human mind could not grasp the extensive

amount of detail being displayed. Thus, a better alternative would be to simply use zoom capabilities by allowing the user to zoom in and out if they want to have more detail, or an overview of the image, respectively. There are nonetheless, a number of problems [14] that need to be addressed in this case; *response time*, because organizing the large amounts of meteorological data into a visualisation has the main purpose of allowing the user to think in a clearer way, thus any delay on the response time will invariably cause interruptions in the flow of thought; *abstraction*, the system has to be abstract enough to be general and flexible, thus making room for new and unpredicted requirements; *user accessibility*, even though the system is developed for users, they are in fact part of the problem because it is very difficult to systematize large amounts of data in a concise manner that is also accessible for the average user.

Meteorology also suffers from "large spatial data fields" [16], which can be analysed using the principle component analysis (PCA) method introduced by Pearson [17] in 1901. This method was initially developed to "find the correlated patterns in a set of variables" [16], and it is now used as a prefilter for compressing large spatial data fields.

Another important area where meteorology can provide a dramatic insight into climate change. Xue et al. in the paper "Quantitative retrieval of geophysical parameters using satellite data" discuss the remote sensing information service grid node (rSiN) that is a tool based on the high-throughput computing grid to deal with climate change. It uses satellite remote sensing to monitor aerosol properties which Xue et al. believe to be very effective in understanding the aerosol radiation that leads to climate change, and their conclusions are very satisfactory in that respect.

2.1.2 Genomics

Genomics is "the branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes" [18]. The human genome, for example, is a complex string of approximately 3 billion [19] As, Ts, Gs and Cs which was considered "absurd" and "impossible" to be decoded when first proposed in the 80s [20, 21]. In fact, originally deciphering took ten years [22] and $3 billion [23] to be completed. Presently, the entire human genome can be decoded in a day for less than $1000 by the Ion Proton^{TM} Sequencer [24]. Genomics, as meteorology (section 2.1.1), uses principal-component analysis (PCA) and partial least-squares (PLS) to analyse the large-volume high-density data structures which are obtained [25]. Eriksson et al. [25] used the example of the cell-cycle-regulating genes of *Saccharomyces cerevisia* and concluded that "analytical bioinformatics data can be visualized using PCA" as it "gives an overview of the data and highlights experimental variations, trends, and outliers". This is a really interesting study and proves that big data can be

displayed graphically in a variety of ways if the right tools are used. However, genomics is a very new science and only recently the technoloy allows scientists to analyse and make full use of the data gathered.

2.1.3 Connectomics

As genomics is the science concerned with the study of genomes, connectomics is the science concerned with the study of connectomes. As first described by Sporns et al. [26] in 2005; a "connectome is a comprehensive map of neural connections in the brain" [27]. It is important to note that in the human brain there exist hundreds of billions of single neurons and possible connections in the the realm of 10^{15}. Thus, there is a need for "powerful tools for handling the vast amount of information given by advanced imaging techniques" [28].

One of the obstacles when comparing large-scale medical images is that intra or interindividual variations have to be removed first [28]. Neurological diseases, for instance schizophrenia [29], multiple sclerosis [30] and Alzheimers [31] cause disruption in the small-world topology of the brain making data harder to analyse. In order to overcome this and other obstacles principal-component analysis (PCA) and linear discriminant analysis (LDA) were used by Robinson et al. [32] when trying to identify population differences in whole-brain structural networks.

The fact that connectomics was only described for the first time in 2005 makes it an exceptionally young science, and the technology has only recently evolved to manage the phenomenal amount of data crucial to its understanding. Furthermore, the research is not very extensive and there are a limited amount of practical applications in this area.

2.1.4 Complex physics simulations

The rising popularity of quantum mechanics, particle physics, biophysics and atomic, molecular, and optical physics brought also with it the challenge of finding a suitable method of dealing with such large amounts of information.

In spite of some difficulties we have had in finding papers regarding complex physics simulations and big data it is, however, clear that simulation software is extremely complex as a result of the need for interaction between several physical models. These difficulties usually lead to bottle-necks in both testing and validation. Moreover, network, and fluid or particles simulations require extremely large amounts of data that needs to be processed. There are solutions being developed, such as i) *OpenFOAM* [33], the computational fluid dynamics (CFD) toolbox; *VORPAL*, designed for modern parallel computing platforms with capabilities such as "accelerator component modeling, charge-particle simulation, fluid simulation, and plasma modeling" [34]; and *Partitioned Global Address Space (PGAS)* [35] which

supports complex coupled multi-physics simulations. However, further research is required to consolidate ideas on how complex physics simulations can benefit from big data tools.

2.1.5 Biological and environmental research

In terms of biological and environmental research there are extensive studies regarding big data. For example in 2002, *GenMAPP* [36] a tool to view and analyse microarray data on biological pathways was developed. Furthermore, there has been research on the "comparative assessment of large-scale data sets of proteinprotein interaction" [37], "literature mining for the biologist" [38], *VisANT*;"an online visualization and analysis tool for biological interaction data" [39], and many other studies [40, 41, 42, 43] that show the motivation that biologists have in using big data tools to improve data management and analysis.

2.2 Industrial Applications

2.2.1 Twitter

Twitter [44] has over 200 million accounts generating a staggering 230 million tweets a day [45]. With this amount of data being generated every single day it came as no surprise that Twitter was interested in using this data to predict trends. Thus, Twitter acquired the company BackType [46] that had a software called *Storm* which parsed live data streams in a "fault-tolerant and scalable" [47] way. Nathan Marz [45], stated that using *Storm* to calculate how links are shared across Twitter users in real-time "is a really intense [continuous] computation, which could involve thousands of database calls and millions of follower records". Furthermore, *Storm* is a distributed RPC (DRPC) with the objective of being able to instantaneously "parallelize the computation of really intense functions" [48]. In spite of the high volume of data, *Storm* is capable of guaranteeing message processing; it is a robust process management tool, capable of managing processes around the cluster; it has fault detection and automatic reassignment of tasks that have timed out; it is an efficient message passing tool that does not use any intermediate queuing; it has a local mode and distributed mode. Above all, "*Storm* is easy to use, configure, and operate" [47] which are some of the key principles in a big data managing tool. In the end it turned out to be what Twitter really needed to analyse and manage big data in order to detect and explore trends. Another problem that Twitter was facing was understanding the connections between users, in order to suggest users to follow or to know how many times a link was shared. Thus *Cassovary*, a "Big Graph-Processing Library" [49] was developed. It was writen in Scala for the JVM with the intention of handling graphs containing "billions of

edges" [49] with ease. *Cassovary* was recently[3] released as an open source program and it can be used to do "large-scale graph mining and analysis" [49].

2.2.2 Search Engines

Search engines, such as Google [50], Yahoo [51] and Bing [52] have to be on the cutting edge of big data analysis. Google for example has to process "more than a billion searches" [53] every single day and has more than "10 billion images indexed"[54]. Big data tools are critical to manage and analyse this data, thus in order to overcome these challenges, Hadoop [55] – which was developed mainly by Google – was created as a distributed file system designed to run on commodity hardware. Also, the need to manage structured data in a scalable manner gave rise to an indexing system called Bigtable [56] that is used with Hadoop. Bigtable is a distributed storage system that handles petabytes of information on thousands of commodity servers. It is used by more than 60 Google products including Google Earth, Orkut and Google Analytics as a flexible and high-performance solution to store the extremely large amounts of data. For example, *HBase* [57] a non-relational, distributed database was modelled on Googles BigTable and its used in extremely large databases. Although Google intended to keep Bigtable an internal technology, engineer Doug Cutting [58] created an open source version that was rapidly adopted by Yahoo. Furthermore, Yahoo started to dedicate considerable resources to improve Hadoop in 2006 and it is now one of the most extensive[4] users of the technology [59]. The numerous uses Yahoo gives to Hadoop include advertisement activity, listings with all the articles and content published, and large log files that store in which sections and stories users click on.

In conclusion, Hadoop is a complete tool to deal "with large amounts of data by splitting it into subsets and sending the data to multiple computer processor unit cores" [60]. One of the few disadvantages, however, it is that users have difficulties adapting to the system, especially if they were used to relational databases supporting general-purpose transactions. Nonetheless, the proof that Hadoop is a reliable system comes precisely from the fact that Google uses it in more than 60 of their products; and a big advantage is that in order to scale it, the only required is that users add more machines to their cluster [56].

2.2.3 Amazon

Amazon [61], is not only the world's largest online retailer but it also has one of the largest databases in the world with more than 42 terabytes of

[3] March 8, 2012

[4] Yahoo uses Hadoop in more than 40,000 servers

data [62]. Furthermore, Amazon has some of world's best analytics and an infrastructure capable of an incredibly big data service. Products such as *Amazon Retail Analytics* which supplies business intelligence for product development and sales optimization [63] are a proof that Amazon is taking big data analysis very seriously and has plans to offer it as a service themselves.

Cassandra [64] a "highly available key-value storage system" [65] was developed by Jonathan Ellis and is based on Dynamo, which was initially developed by Amazon with the intent to store what users were adding to their shopping cart. Moreover, in its design each and every node is capable of accepting information from the entire system making it possible to replicate data across different hosts.

Amazon is thus a very important player in the big data arena, the proof is products such as i) *Amazon Elastic MapReduce* that "enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data" [66]; ii) *Amazon Relational Database Service* [67] which is still in beta stage but has the goal to "makes it easy to set up, operate, and scale a relational database in the cloud"; iii) *Amazon Simple Storage Service* [68], is an interface that stores and retrieves large amount of information; among other products that belong to *Amazon Web Services* [69].

2.2.4 Facebook

Facebook [70] is colossal; it has currently more than 845 million active users [71] which share more than 100 billion connections between each other. Moreover, every single day 250 million photos are uploaded, there are 2.7 billion likes and an average 20 minutes spent on the site per visit. Thus, it is crucial that Facebook has the tools to manage and analyse such large amounts of data.

Apache Cassandra, an open souce distributed database was originally developed by Facebook engineers [72]. Cassandra database can store 2 million columns in a single row which makes it possible to append more information to user accounts without knowing first hand how to format that data. Another advantage is that data can be spread across different servers which aids databases scalability beyond a single server.

Google's Bigtable architecture was also considered by Facebook, however because the entire operation depended on a single node coordinating all the activities across every other node it would be a very risky alternative. Ellis stated that Google's Bigtable "is not the best design. You want one where if one machine goes down, the others keep going" [45].

Facebook is also using Hadoop, where its migration problems were described in an excellent blog post by Paul Yang [73]. Its cluster has grown to more than 30 petabytes of data, 3000 times the amount of information in the Library of Congress. Moreover, they use Hive [74] for analytics and

HDFS [75] to store data.

2.2.5 Finance

As well as other industries, the financial sector has seen a major increase in the amount of quality data accessible. Moreover if this data is analysed adequately it can dramatically improve organisational performance. For example, sentiment analysis, also known as opinion mining, is the analysis of others opinions. This might be very relevant for the finance sector, as big data tools can be used to harvest user-generated content to find opinions about certain stock. The main problem with this type of analysis is that it is a natural language processing task [76]. Research on natural language analysis is very limited because we are relying on machine learning algorithms [76] that are not capable of understanding human meaning. Nonetheless, the need for such analysis is real, due to the fact that each and every company would like to know what consumers think about their services and products. Furthermore, it is also true that consumers would like to know what other consumers think about a company, product, or service. It is believed by Amir Halfon [77] that "sentiment analysis is becoming so popular that some hedge funds are basing their entire strategies on trading signals generated by Twitter analytics". This is extraordinary, and although it might be an extreme-case scenario it is further proof that sentiment analysis is becoming a very important tool for financial institutions.

Considering the quantity of historical market data and the exponential speed at which new data is being created, makes it difficult for finantial firms to deal with big data. Moreover, these firms can not afford to delay the analysis of these great amounts of data, thus they are moving from proprietary tools to open source frameworks such as Hadoop and R [78]. Hadoop was discussed in section 2.2.2 and R is a framework for microarray analysis which allows us "to analyze 100-1,000s of arrays of various chip types" [78] such as SNP chips, exon arrays and expression arrays.

We are moving towards a continuous risk management solution; where correlating data from different sources can be the key to stop fraudulent activites. The continuous risk management solution includes computing the aggregation of counter party exposure or VAR, among other situations [77] which are also related to big data. These facts again raise the difficulty of dealing with massive amounts of data that current systems simply can not perform and the need for tools that allow their analysis.

2.2.6 Retail

Big retail companies such as Walmart [79] which handles over "1 million customer transactions every hour" [12] also have the need for big data tools in order to process such large amounts of data. The estimated 2.5 petabytes

of data that Walmart has in its database is comparable to 167 times of all the data contained on all the books in the US Library of Congress [80]. Although still far from Facebook's 30 petabytes it is an impressive amount of data that needs to be managed and analysed efficiently.

3 Current Research

Current research was largely covered in the examples above, which makes it clear that research in the field of big data is extremely vibrant "with significant activity in both industry and academia" [81]. This comes as no surprise, not only because of all of the immediate practical applications that we have covered in this paper but also because there is much more to be learnt and discovered. The research is dominated by studies to try and solve one or more of the three-dimensional data growth challenges – increasing volume, variety, and speed – that were defined by Doug Laney in 2001 [82]. Furthermore, we have encountered numerous papers describing the engineering [2, 83, 43] and semantic [84, 85, 86, 87] obstacles that make big data a "use case for multidisciplinary problem solving" [88]. Nonetheless, all the research boils down to a few key factors i) define the problem to be solved; ii) search the big data; iii) extract, transform, and load (ETL) what is relevant; iv) confirm that the information is comprehensive, relevant and unique; v) find the solution to the initial problem. However Huberman has raised some concerns regarding the fact that most big data comes from private sources and is inaccessible to the majority of researchers, he further argues that if big data does not become ubiquitous a small group of scientists will have privileged access to private data repositories restricted its access to equally talented researchers that do not have the same "connections" [89].

4 Conclusion

Big data is a very broad topic and impossible to clearly define because it depends upon the resources of each and every organization. Moreover, it is a moving target, for what is regarded as big data today might not be tomorrow. Nonetheless, the research in the area is vast and the numerous papers published thus far are a proof of that. There are already a great number of technologies being applied to big data, such as datamining grids [49], distributed file systems [55], distributed databases [57, 64, 67], interactive visualization of large gridded data sets [15], cloud computing platforms [69], scalable storage systems [66, 68], computational fluid dynamics [33], parallel computing platforms [34], among many others [39, 74, 75, 78] previously described in this paper. Moreover, there is a real need for professionals that can work with these technologies, the proof of which is in the steep growth

of the number of positions available in this field as the graphic below demonstrates:

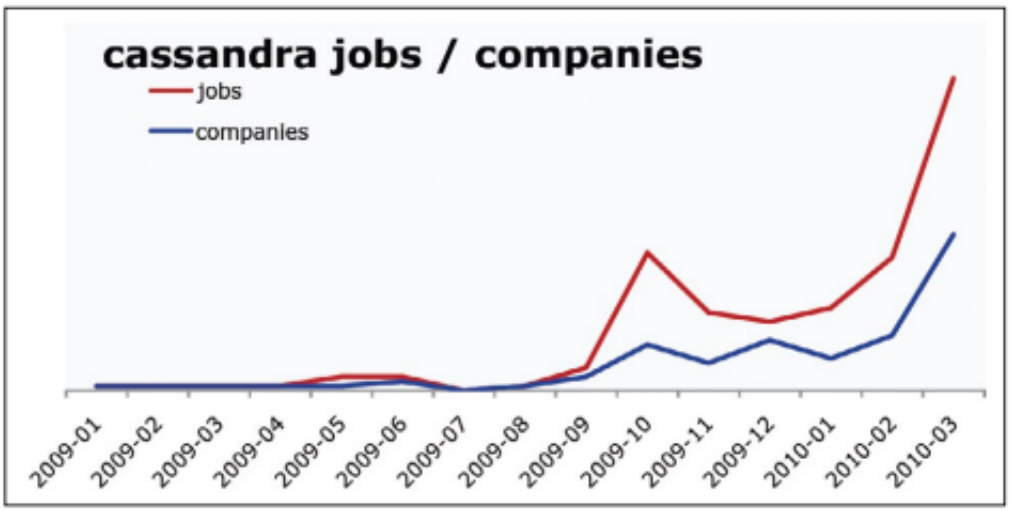

Figure 1: Data from O'Reilly Research [90] which shows the increase over the years for Hadoop [59] and Cassandra [64] related jobs as well as a growth in the number of companies listing jobs related to this technology.

Forrester Research analyst James Kobielus goes even further and argues that technology is not the main obstacle for big data, it is rather, "finding the right talent to analyze the data" [45]. It is also believed that this will give rise to a new breed of professionals – the data scientist – that will have a vast understanding of mathematics and statistics.

We conclude this paper with a quote from Hal Varian, Google's Chief Economist whose words are very relevant: "I keep saying the sexy job in the next ten years will be statistician... The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids" [91].

Bibliography

[1] M. Waldrop. Big data: wikiomics. *Nature*, 455(7209):22–25, 2008.

[2] J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, and C. Welton. Mad skills: New analysis practices for big data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492, 2009.

[3] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S. St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.

[4] D. Field, S.A. Sansone, A. Collis, T. Booth, P. Dukes, S.K. Gregurick, K. Kennedy, P. Kolar, E. Kolker, M. Maxon, et al. 'omics data sharing. *Science*, 326(5950):234, 2009.

[5] J. Ahrens, K. Brislawn, K. Martin, B. Geveci, C.C. Law, and M. Papka. Large-scale data visualization using parallel data streaming. *Computer Graphics and Applications, IEEE*, 21(4):34–41, 2001.

[6] Cconomist. Data, data everywhere, 2010. [Online; accessed 11-March-2012].

[7] V. Klinkenborg. Trying to measure the amount of information that humans create. *New York Times, p. A*, 20, 2003.

[8] H. Kobayashi, F. Dolivo, and E. Eleftheriou. 35 years of progress in digital magnetic recording. In *Proc. 11th Intl. Symp. Problems of Redundancy in Info. and Control Sys*, pages 2–6.

[9] N.K. Agarwal and D.C.C. Poo. Making sense of an electronic document-visualization strategies for concept presentation. In *Enterprise Distributed Object Computing Conference Workshops, 2006. EDOCW'06. 10th IEEE International*, pages 56–56. IEEE, 2006.

[10] J.F. Gantz, J. Mcarthur, and S. Minton. The expanding digital universe. *Director*, 285(6), 2007.

[11] PC Advisor. Microsoft windows server 8 review, 2012. [Online; accessed 11-March-2012].

[12] Wikipedia. Big data, 2012. [Online; accessed 12-March-2012].

[13] LLC. Dictionary.com. Meteorology — define meteorology at dictionary.com, 2012. [Online; accessed 12-March-2012].

[14] B. Hibbard. The top five problems that motivated my work [data visualisation]. *Computer Graphics and Applications, IEEE*, 24(6):9–13, 2004.

[15] Space Science and Engineering Center. Vis5d, 1998. [Online; accessed 13-March-2012].

[16] W.W. Hsieh and B. Tang. Applying neural network models to prediction and data analysis in meteorology and oceanography. 1998.

[17] I.T. Jolliffe and MyiLibrary. *Principal component analysis*, volume 2. Wiley Online Library, 2002.

[18] LLC. Dictionary.com. Genomics — define genomics at dictionary.com, year =.

[19] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. *science*, 291(5507):1304, 2001.

[20] L. Roberts. Controversial from the start. *Science*, 291(5507):1182, 2001.

[21] PO JWG423. Decoding the human genome. *Biochemical Society Transactions*, 28(Part 5):A97, 2000.

[22] H. Varmus. Ten years onthe human genome and medicine. *New England Journal of Medicine*, 362(21):2028–2029, 2010.

[23] F.S. Collins. Medical and societal consequences of the human genome project. *New England Journal of Medicine*, 341(1):28–37, 1999.

[24] Life Technologies Corporation. Ion proton^{tm} sequencer, 2012. [Online; accessed 13-March-2012].

[25] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, and S. Wold. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and bioanalytical chemistry*, 380(3):419–429, 2004.

[26] O. Sporns, G. Tononi, and R. Kötter. The human connectome: a structural description of the human brain. *PLoS Computational Biology*, 1(4):e42, 2005.

[27] Wikipedia. Connectome, 2012. [Online; accessed 13-March-2012].

[28] P.T. Yap, G. Wu, and D. Shen. Human brain connectomics: Networks, techniques, and applications [life sciences]. *Signal Processing Magazine, IEEE*, 27(4):131–134, 2010.

[29] Y. Liu, M. Liang, Y. Zhou, Y. He, Y. Hao, M. Song, C. Yu, H. Liu, Z. Liu, and T. Jiang. Disrupted small-world networks in schizophrenia. *Brain*, 131(4):945, 2008.

[30] Y. He, A. Dagher, Z. Chen, A. Charil, A. Zijdenbos, K. Worsley, and A. Evans. Impaired small-world efficiency in structural cortical networks in multiple sclerosis associated with white matter lesion load. *Brain*, 132(12):3366, 2009.

[31] K. Supekar, V. Menon, D. Rubin, M. Musen, and M.D. Greicius. Network analysis of intrinsic functional brain connectivity in alzheimer's disease. *PLoS computational biology*, 4(6):e1000100, 2008.

[32] E.C. Robinson, A. Hammers, A. Ericsson, A.D. Edwards, and D. Rueckert. Identifying population differences in whole-brain structural networks: A machine learning approach. *Neuroimage*, 50(3):910–919, 2010.

[33] H. Jasak, A. Jemcov, and Z. Tukovic. Openfoam: A c++ library for complex physics simulations. In *International Workshop on Coupled Methods in Numerical Dynamics, IUC, Dubrovnik, Croatia*, 2007.

[34] D. Smithe, P. Stoltz, M.C. Lin, D. Karipides, H. Wang, K. Tian, and G. Cheng. 19.4: Multi-physics simulations with vorpal. In *Vacuum Electronics Conference (IVEC), 2010 IEEE International*, pages 445–446. IEEE, 2010.

[35] F. Zhang, C. Docan, M. Parashar, and S. Klasky. Enabling multi-physics coupled simulations within the pgas programming framework. In *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*, pages 84–93. IEEE, 2011.

[36] K.D. Dahlquist, N. Salomonis, K. Vranizan, S.C. Lawlor, and B.R. Conklin. Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nature genetics*, 31(1):19–20, 2002.

[37] C. Von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *NATURE-LONDON-*, pages 399–404, 2002.

[38] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7(2):119–129, 2006.

[39] Z. Hu, J. Mellor, J. Wu, and C. DeLisi. Visant: an online visualization and analysis tool for biological interaction data. *BMC bioinformatics*, 5(1):17, 2004.

[40] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl 1):S145–S154, 2002.

[41] C. Bartels, T. Xia, M. Billeter, P. Güntert, and K. Wüthrich. The program xeasy for computer-supported nmr spectral analysis of biological macromolecules. *Journal of Biomolecular NMR*, 6(1):1–10, 1995.

[42] B.F. Cravatt, G.M. Simon, and J.R. Yates Iii. The biological impact of mass-spectrometry-based proteomics. *Nature*, 450(7172):991–1000, 2007.

[43] E. Aronova, K.S. Baker, and N. Oreskes. Big science and big data in biology: From the international geophysical year through the international biological program to the long term ecological research (lter) network, 1957-present. *Historical Studies in the Natural Sciences*, 40(2):183–224, 2010.

[44] Twitter Inc. Twitter, 2012. [Online; accessed 14-March-2012].

[45] Joab Jackson. The big promise of big data, 2012.

[46] Backtype. Backtype has been acquired by twitter, 2012.

[47] Nathan Marz. A storm is coming: more details and plans for release, 2012. [Online; accessed 14-March-2012].

[48] Nathan Marz. Distributed rpc, 2012. [Online; accessed 14-March-2012].

[49] Pankaj Gupta. Cassovary: A big graph-processing library, 2012. [Online; accessed 14-March-2012].

[50] Google Inc. Google, 2012. [Online; accessed 14-March-2012].

[51] Yahoo Inc. Yahoo, 2012. [Online; accessed 14-March-2012].

[52] Microsoft Inc. Bing, 2012. [Online; accessed 14-March-2012].

[53] Google Inc. Google internet stats, 2010. [Online; accessed 14-March-2012].

[54] MG Siegler. Google image search: Over 10 billion images, 1 billion pageviews a day, 2010. [Online; accessed 14-March-2012].

[55] D. Borthakur. The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 2007.

[56] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R.E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):4, 2008.

[57] R.C. Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics*, 11(Suppl 12):S1, 2010.

[58] D. Cutting and E. Baldeschwieler. Meet hadoop. *OSCON, Portland, OR, USA*, 2007.

[59] T. White. *Hadoop: The definitive guide*. Yahoo Press, 2010.

[60] J.N. Burney, G. Koch, and J.B.C. Hall. A study on the viability of hadoop usage on the umfort cluster for the processing and storage of cresis polar data.

[61] Amazon Inc. Amazon, 2012. [Online; accessed 14-March-2012].

[62] Focus Inc. Top 10 largest databases in the world, 2010. [Online; accessed 16-March-2012].

[63] R. Bordawekar, B. Blainey, C. Apte, and M. McRoberts. Analyzing analytics part (1): A survey of business analytics models and algorithms.

[64] E. Hewitt. *Cassandra: the definitive guide*. O'Reilly Media, Inc., 2010.

[65] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: amazon's highly available key-value store. *ACM SIGOPS Operating Systems Review*, 41(6):205–220, 2007.

[66] Amazon Web Services LLC. Amazon elastic mapreduce (amazon emr), 2012. [Online; accessed 16-March-2012].

[67] Amazon Web Services LLC. Amazon relational database service (amazon rds), 2012. [Online; accessed 16-March-2012].

[68] Amazon Web Services LLC. Amazon simple storage service (amazon s3), 2012. [Online; accessed 16-March-2012].

[69] Amazon Web Services LLC. Amazon web services, 2012. [Online; accessed 16-March-2012].

[70] Facebook Inc. Facebook, 2012. [Online; accessed 14-March-2012].

[71] Anson Alexander. Facebook user statistics 2012, 2012. [Online; accessed 14-March-2012].

[72] D. Borthakur, J. Gray, J.S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, et al. Apache hadoop goes realtime at facebook. In *ACM SIGMOD Conf*, pages 1071–1080, 2011.

[73] Paul Yang. Moving an elephant: Large scale hadoop data migration at facebook, 2011. [Online; accessed 16-March-2012].

[74] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment*, 2(2):1626–1629, 2009.

[75] D. Borthakur. Hdfs architecture guide, 2008.

[76] B. Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 978–1420085921, 2010.

[77] Amir Halfon. Big data use cases, 2012. [Online; accessed 14-March-2012].

[78] Henrik Bengtsson. An open-source r framework for your microarray analysis, 2012. [Online; accessed 15-March-2012].

[79] Inc. Wal-Mart Stores. Walmart.com: Save money. live better., 2012. [Online; accessed 15-March-2012].

[80] K. Cukier. A special report on managing information: Data, data everywhere. *The Economist*, 12(8), 2010.

[81] J. Lin and C. Dyer. Data-intensive text processing with mapreduce. *Synthesis Lectures on Human Language Technologies*, 3(1):1–177, 2010.

[82] L. Douglas. 3d data management: Controlling data volume, velocity and variety. *Gartner*, 2001.

[83] J. Bughin, M. Chui, and J. Manyika. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, 56, 2010.

[84] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

[85] R. Agrawal, A. Ailamaki, P.A. Bernstein, E.A. Brewer, M.J. Carey, S. Chaudhuri, A.H. Doan, D. Florescu, M.J. Franklin, H. Garcia-Molina, et al. The claremont report on database research. *ACM SIGMOD Record*, 37(3):9–19, 2008.

[86] D. Agrawal, S. Das, and A. El Abbadi. Big data and cloud computing: new wine or just new bottles? *Proceedings of the VLDB Endowment*, 3(1-2):1647–1648, 2010.

[87] W. Tantisiriroj, S. Patil, and G. Gibson. Data-intensive file systems for internet services: A rose by any other name. *Parallel Data Laboratory, Tech. Rep. UCB/EECS-2008-99*, 2008.

[88] M. Greaves M.L. Brodie and J.A. Hendler. Databases and ai: The twain just met. *STI Semantic Summit, Riga, Latvia, July 6-8, 2011*, 2011.

[89] B.A. Huberman. Sociology of science: Big data deserve a bigger audience. *Nature*, 482(7385):308–308, 2012.

[90] M. Loukides. What is data science? *The future belongs to the companies and people that turn data into products*, 2010.

[91] M. Townsend and A. Ridgeway. Making official data relevant to students: Statistics canada's education outreach program.