Data Processing – Challenges and Tools

Mr Joseph Bonello
University of Malta
Email: joseph.bonello@um.edu.mt

Prof Ernest Cachia
University of Malta
Email: ernest.cachia@um.edu.mt

***Abstract*—Data has grown at incredible rates these last few year, especially with the increasing popularity of social media and video streaming services such as YouTube. This paper looks at some of the challenges associated with the acquisition and processing of data. These challenges defer from the opportunities that can be exploited from the acquired data and are areas that can benefit from scientific research. In particular, the need to process large amounts of data in real-time is becoming a critical need in many areas such as social network trends, website statistics and intrusion detection in large data centres.**

I. Introduction

Data has become a very important aspect of a modern, information-driven society. A brief look at statistics provides enough empirical evidence of the growth in the amount of data that is generated. YouTube, for instance, claims that users generate around 300 hours of video every minute of every day[1]. Craig Smith reports that an estimated number of days worth of YouTube videos that are watched on Facebook every minute is of 323[2]. Twitter reports around 500 million tweets every day, with around 80% of active users using the service from a mobile device[3]. Data from 2014 shows that over 20 billion photos have been uploaded to Instagram.

The data that is being generated is mostly unstructured generated mostly through social media[4]. It consists of free text, images, videos and other forms of artistic expressions (such as images, sound or a combination of different modalities) which are not easily interpreted by electronic means. Data is not limited to just the consumer-generated domain. Meaningful data is also generated by computer systems in the form of extensive logs that help diagnose shortcomings and debug while the system is running.

The scientific community also generates a large amount of data for analysis. Vidal and Cid state that the Large Hadron Collider (LHC) project at CERN is capable of generating data at around 1 petabyte per second (1 PB/s)[5]. Given the massive scale of this data, it is sifted in realtime to retain only the data that is relevant. Many other datasets are also available freely for researchers, such as those provided at the Open Science Data Cloud[6]. The data here consists of a large variety of datasets that range over a number of subject areas, including biology, genomics, social science and music.

Given the massive scale and diversity of data, it is tempting to solely focus on the size of the data and ignore the processes that are associated with the acquired abilities to represent many aspects of the physical world as tangible and searchable data. Cuckier and Mayer-Schoenberger call this process "datafication" and provide examples of how location is represented through GPS coordinates or how friendships are represented through one's circles in social media[7]. The collection, sharing and processing of data, however, present a number of challenges in order to transform the data into valuable information that helps its consumer in a meaningful way. In this paper, we will look at some of the challenges in data collection and tools that facilitate its storage and processing towards the goal of storing value against data.

II. The challenges of data processing

The ability to manipulate large datasets provides new opportunities in terms of the competitive advantage gained through the analysis and predictive strengths that are available as a result of the ability to harness and store the available data. As Cuckier and Mayer-Schoenberger put it, it is a move from causation to correlation[7]. This means that instead of attempting to always understand what is going on with the data, it is possible to understand the relationship between different phenomena and use that understanding to find a solution.

A community whitepaper by Purdue University list a number of challenges that affect the data processing pipeline at many stages[8]. The authors highlight the difficulties that start even at the data acquisition stage, where a number of decisions are required (often using impromptu methods) about what data to keep and what to discard.

The challenges to data processing are the result of different factors. They can be legal commitments and personal concerns as in the case of privacy, business-driven as in the case of security, technical as in the case of storage and speed and analytical challenges that arise from the data processing itself. The following sections describe some of the challenges involved in the data processing pipeline that prevent the efficient use of the available data.

A. Privacy

Privacy is defined as the relationship between the collection of sensitive data and its dissemination[9]. The challenge arises whenever personally-identifiable or other sensitive information is collected and stored. Data related to a person's health care records, financial records, location and residence, ethnicity and biological traits are primary subjects of privacy concerns.

Privacy is an ongoing concern in data collection as people are more acutely aware of how personal data is collected, stored and processed. Privacy laws are enacted to regulate how data is gathered and processed. Companies provide publicly-available privacy policy statements that inform potential users how their data will be stored and appoint a data protection officer whose responsibility is to ensure that data is handled prudently[10]. Government and Non-Governmental Agencies

regularly provide campaigns to promote awareness of privacy issues such as the Privacy Awareness Week that is organised by the Asia Pacific Privacy Authorities forum[11].

Data privacy concerns extend beyond the confines of private company or agency usage. Online activity has given rise to a number of important court rulings that have a profound significance on how data is stored and treated. An example of one such case is the court ruling by the European Court of Justice against Google that was delivered in May 2014[12]. This ruling, later called the "Right to be forgotten" ruling determined that search engines are controllers of data and therefore are subject to data protection laws (i.e. applicability of jurisdiction on data) and that individuals can, under some restrictions, have a right to demand that data about the individual is removed if the information is "*inaccurate, inadequate, irrelevant or excessive*"[12].

Besides storage, companies need to ensure that data access is also secure. This includes the transmission of data over a network, which must be duly encrypted to render snooping difficult over network connections. Access to sensitive data should be limited and audited, to ensure that only authorised personnel are privy to the information but also to have a reliable audit trail should a data breach be found or authenticity be sought. Moreover, should a third-party be used to acquire data (such as, for example, using an online survey tool), the company's data controller should thoroughly check how data is stored and what the retention policy is.

B. Security

Data security is defined as the measures that are applied to prevent unauthorised access to computers and databases. Furthermore, it also refers to measures taken to protect data from corruption and loss. Unauthorised access to data leads to not only privacy concerns but also to financial and legal concerns[9]. Industrial theft is a serious threat to an organisation as it threatens its competitive edge. The loss of trade secrets can be catastrophic for an organisation due to the financial implications. PriceWaterhouseCoopers estimate Research and Development (R&D) to be around 1.8% of the global GDP [13]. The same report, quoting a 2012 Association of Certified Fraud Examiners (ACFE) report, states that companies worldwide lose as much as $3.5 trillion (5% of the GDP) to occupational fraud.

The "Celebgate" incident of August 2014, where private photos of some celebrities were leaked after being reported stolen from Apple's cloud services, highlights the risks associated with securing data online and highlights the weaknesses of computer systems[14]. Although vulnerabilities in computer systems exist and are detected on a daily basis, regular updates are not always applied in time leading to security compromises that facilitate illicit activities. A more standard approach to data security is critical and should include processed data (e.g. data extracts and reports).

Data security is also related to privacy. Security breaches where personal and financial data is stolen can lead to loss of trust and financial loss for the company or agency that has been entrusted with the data and reduces customers' trust in using online services. Two such examples include the hacking of Target in 2013[15] and Ebay in March 2014[16]. Ebay's shares are reported to have lost as much as 3.2% as a result of the news, showing how markets react adversely to news of security breaches.

C. Storage

To cope with protection against loss of data and remote accessibility, many companies opt to store data on cloud services that offer large storage facilities. Companies such as Dropbox offer Enterprise solutions that afford businesses a large storage space and governed by strict Service Level Agreements on availability and security.

Users of these organisations share the same corporate account which allows them to collaborate together by sharing copies of the same file. Cloud storage solutions ensure that data is replicated across a number of nodes and that it is properly segregated. These solutions are able to detect conflicts, and merge changes made by different users in the same document.

de Borja discusses some important features of cloud computing especially on how cloud systems offer great scalability features[17]. The services can offer increased bandwidth, speed and data storage that is driven by the particular requirements of the business. Cloud systems are designed to be resilient due to the in-built replication mechanisms often at geographically separate data centres that aims to provide high availability to the end users. Users have the possibility to sync locally the part of they are working on. This allows users to continue working offline even in case of failures (such as limited network availability).

With storage, there is an associated issue with the Total Cost of Ownership of the using cloud based resources. Cloud based resources facilitate the scaling requirements of a server, such that a system can scale up if it requires more resources or down if the requirements change[18]. The cost of owning a server or a group of servers compared to renting cloud facilities is much higher due to the operational costs involved (such as administrators, licences, backups and maintenance) which are often ignored[18].

D. Speed

Data does not deal with static data only. As hinted earlier in the paper, massive amounts of data can be generated by systems during their execution. Nick Clayton quoting Nick Halstead of DataSift claims that the New York Stock Exchange generates one terabyte of data on a daily basis[19]. The Stock Exchange is a good example of the importance of the data's speed, as it directly affects the choices performed by high-speed trading algorithms.

The issue with speed is that traditional models of data storage are not adequate to support an environment that requires decisions to be taken in real-time[20]. Traditionally, database systems store structured data that is then queried at a later stage. This is not possible with streaming data, as it can be too large to query and traditional models are too slow to react[20].

Some examples given by IBM are the Internet of Things and fraud detection in high volume transaction systems[21]. In the former example, numerous sensors monitor real-world objects for availability, performance, capacity and resource

utilisation. These sensors generate a large amount of data that needs to be analysed in realtime and corrective action taken as quickly as possible. In this case, "normal" data can be discarded after it has been used.

In the second example, detecting a fraudulent pattern as quickly as possible is essential to prevent fraud and to be able to gather enough concrete data to enable possible legal action to be taken against the perpetrator. In this case, once the analytical process is complete, it is stored permanently for safe keeping.

E. Modeling and using data

Data modeling is the process that is used to define, analyse and learn about data[22]. Wambler describes it as the act exploring data-oriented structures[23]. The aim is to transform the collected raw data into a usable data product that can be used to make informed decisions. The process involves cleaning data processing and cleaning the data coupled with a method of systematically going through the data to understand the nature of the data.

One of the main problems with large datasets is that as data grows, it often becomes difficult to ascertain the validity and correctness of data. This means that data may be incomplete, ambiguous and may be inaccurate due to model approximations. In particular, unstructured data contains significant amounts of uncertain and imprecise data[7].

As the volume of data grows, organisations cannot continue to invest in data cleanup and preparation and must instead embrace the fact that data may be incomplete. In practice, as data grows, one can infer a conclusion about the population from a subset of the data, all within a certain margin of error. Whereas previously the samples collected were required to necessarily be correct, it is now possible to benefit from the analysis of bigger datasets provided they are not fundamentally incorrect[7].

A discipline that is used in data modelling is statistical inference. Statistical inference is concerned with developing methods, theorems and procedures for extracting meaning and information from data that has been generated by a stochastic process[22]. This technique is useful to draw conclusions based on the data while allowing for the magnitude of errors to be estimated.

III. Tools for data processing

The previous section of this paper analysed some challenges that are encountered in data processing. This section provides an overview of tools that are used to acquire, model and process data. Hadoop and Storm are two tools that are used in storing and processing large datasets. The Hadoop and Storm projects are two projects that focus on acquisition, processing and storage of data. R is a tool that is used in data analysis and statistical model building. It has strong statistical and visualisation toolset that is continuously updated.

A. Apache Hadoop

Apache Hadoop is a technology that was originally devised by Google as a way of indexing the content that they were collecting from the web. Yahoo has contributed a lot of effort for developing Hadoop for Enterprise applications. Mike Olson, CEO of Cloudera explains the benefits and architecture of Hadoop in an interview with Mike Turner[24].

Hadoop address the problems associated with massive amounts of data which can be both complex and structured and that do not always fit into a table structure. Hadoop is useful where computationally extensive analytic analysis and sophisticated modelling is required.

The architecture of Hadoop is built so that computations can be distributed across a large number of independent machines, that is they do not share memory and disks. This means that Hadoop can run on many cheap machines rather than expecting a high-end expensive server. This is possible because Hadoop can break the data down into small chunks that can be processed on different servers. Data is replicated across several nodes making the platform reliable.

Hadoop uses the concept of MapReduce to break the processing task into smaller chunks which are then reassembled back into a single whole once the processing is complete. Since the computations are farmed out on multiple servers, it is possible to request complex computations as each processor works on its own copy in parallel with other processors.

B. Apache Storm

Apache Storm is a computation system that processes large volumes of high-velocity data in real-time. Like Hadoop, it uses YARN for clustering and managing multiple data processing engines. Storm is optimised for that require real-time analytics, machine learning and continuous monitoring of operations.

Storm's architecture is driven by a special node called Nimbus, which is a master node analogous to Hadoop's JobTracker. It is responsible for uploading computations for executions, distributing code across nodes in the cluster, launching workers in the cluster and for monitoring computation and reallocation workers as needed.

One or more ZooKeeper nodes are responsible for coordinating the Storm cluster. ZooKeeper is a centralised service that is used to maintain configuration information and for provided group services and distributed synchronisation. The last type of node in a Storm cluster is the Supervisor node. Supervisor nodes communicate with Nimbus by means of the ZooKeeper and are responsible for starting and stopping workers as directed by the Nimbus node.

Storm processes are organised as a topology of Spouts and Bolts. Storm uses the abstraction of Streams to denote unbounded sequences of data, which in Storm terms are ordered lists of elements called Tuples. A Sprout is a source of tuples that is used in computations whereas a Bolt is a processor that manipulates input streams and optionally produces an output stream. A Topology represents the overall Storm network of Spouts and Bolts and how they are connected together.

C. R

R is an implementation of the S language, that was initially developed at AT&T Bell Laboratories. It is an open source language that is provided free of charge and available through the

Comprehensive R Archive Network (CRAN). R has extensive and powerful graphics abilities and is known for its strong analytical abilities that are continuously being updated.

R's strength lies in its wide range of tools that can be used for data analysis. It includes tools for running basic statistics and supports probability distributions that allows an analyst to quick generate datasets for basic exploratory data analysis.

It also supports machine learning algorithms for cluster analysis, neural networks and trees/recursive partitioning. R has libraries to support optimisation and mathematical programming, signal processing and for simulation.

IV. Conclusion

As data takes more centre stage in the "information society", new opportunities for research will arise on how data can be processed to allow stakeholders to reap the maximum benefit. The challenges highlighted in this paper present some of the areas where additional research is required to facilitate the aim of acquiring data and to exploit the benefits and opportunities that are associated with it.

New models of using data are providing the ability to perform analysis on data as it is being generated. The value that is derived can be used to make informed decisions on the fly. For example, in the context of an advertising company (such as Google Ads), displaying outdated advertisements based on stale data can result in a loss in revenue.

References

[1] Youtube, "Youtube statistics," Retrieved from https://www.youtube.com/yt/press/statistics.html. Last accessed 01 April 2015., 10 2014.

[2] C. Smith, "By the numbers: 80+ amazing youtube statistics," Retrieved from http://expandedramblings.com/index.php/youtube-statistics. Last accessed 01 April 2015., 03 2015.

[3] Twitter, "Twitter statistics," Retrieved from https://about.twitter.com/company. Last accessed 01 April 2015., 2015.

[4] D. Stewart, "Big content: The unstructured side of big data," Retrieved from http://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-big-data. Last accessed 05 April 2015., 2013.

[5] C. V. Vidal and R. Cid, "Lhc data," Retrieved from http://www.lhc-closer.es/1/3/12/0. Last accessed 01 April 2015., 2012.

[6] OSDC, "Public data commons," Retrieved from https://www.opensciencedatacloud.org/publicdata. Last Accessed 30 March 2015., 2015.

[7] K. N. Cuckier and V. Mayer-Schoenberger, "The rise of big data," *Foreign Affairs*, 2013.

[8] Purdue University, "Challenges and opportunities with big data," Purdue University, Tech. Rep., 2012.

[9] Y. Sun, J. Zhang, Y. Xiong, and G. Zhu, "Data security and privacy in cloud computing," *Internation Journal of Distributed Sensor Networks*, 2014.

[10] Governor Technology Limited, "Data protection officer," Retrieved from http://www.eudataprotectionlaw.com/data-protection-officer. Last accessed 05 April 2015., 2013.

[11] Asia Pacific Privacy Authorities forum, "Privacy awareness week 2015," Retrieved from http://www.privacyawarenessweek.org. Last accessed 02 April 2015., 2015.

[12] European Commission, "Factsheet on the "right to be forgotten" ruling (c-131/12)," European Commission, Tech. Rep., 2014.

[13] P. Passman, S. Subramanian, and G. Prokop, "Economic impact of trade secret theft: A framework for companies to safeguard trade secrets and mitigate potential threats," PricewaterhouseCoopers LLP (PwC), Tech. Rep., 2014.

[14] O. Williams, "This could be the icloud flaw that led to celebrity photos being leaked." Retrieved from http://thenextweb.com/apple/2014/09/01/this-could-be-the-apple-icloud-flaw-that-led-to-celebrity-photos-being-leaked/. Lasct accessed on 30 March 2015., 2014.

[15] T. Berman, "Credit card, personal data stolen from 40 million target customers," Retrieved from http://gawker.com/credit-card-personal-data-stolen-from-40-million-targe-1486354625. Last accessed 05 April 2015., 2013.

[16] Reuters, "Ebay: Hackers stole users' personal data," *New York Post*, 2014.

[17] F. de Borja, "Important features of cloud computing," Retrieved from http://cloudtweaks.com/2012/03/important-features-of-cloud-computing. Last accessed 02 April 2015., 03 2012.

[18] B. Golden, "The case against cloud computing, part four," Retrieved from http://www.cio.com/article/2430760/cloud-computing/the-case-against-cloud-computing–part-four.html. Last accessed 02 April 2015., 2009.

[19] N. Clayton, "Filtering profits," *Wall Street Journal*, 2011.

[20] MongoDB Inc. (2015) Big data explained. Retrieved from http://www.mongodb.com/big-data-explained. Last accessed 06 April 2015.

[21] IBM, "What is stream computing?" Retrieved from http://www-01.ibm.com/software/data/infosphere/stream-computing. Last acessed 03 April 2015., 2015.

[22] C. O'Neill and R. Schutt, *Doing Data Science*. O'Reilly Media, 2013.

[23] S. Wambler, "Data modeling 101," Retrieved from http://www.agiledata.org/essays/dataModeling101.html. Last accessed 04 April 2015., 2013.

[24] J. Turner, "Hadoop: What it is, how it works, and what it can do," Retrieved from http://radar.oreilly.com/2011/01/what-is-hadoop.html. Last accessed 04 April 2015., 01 2011.