A Cloud-based Exploration of Open Data: Promoting Transparency and Accountability of the Federal Government of Australia

Edwin Salvador
Department of Computing and Information Systems
University of Melbourne
Melbourne, Australia
edwin.salvador@epn.edu.au

Richard Sinnott
Department of Computing and Information Systems
University of Melbourne
Melbourne, Australia
rsinnott@unimelb.edu.au

Abstract

The Open Data movement has become more popular since governments such as USA, UK, Australia and New Zealand decided to open up much of their public information. Data is open if anyone is free to use, reuse and redistribute it. The main benefits that a government can obtain from Open Data include transparency, participation and collaboration. The aim of this research is to promote transparency and accountability of the Federal Government of Australia by using Cloud-related technologies to transform a set of publicly available data into human-friendly visualizations in order to facilitate its analysis. The datasets include details of politicians, parties, political opinions and government contracts among others. This paper describes the stages involved in transforming an extensive and diverse collection of data to support effective visualization that helps to highlight patterns in the datasets that would otherwise be difficult or impossible to identify.

1 Introduction

In recent years, the Open Data movement has become increasingly popular since the governments of various countries such as USA, UK, Australia, New Zealand, Ghana amongst many others decided to open up (some of) their data sets. In order to consider data as open, it should ideally be available preferably online in formats that are easy to read by computers and anyone must be allowed to use, reuse and redistribute it without any restriction (Dietrich et al., 2012). Furthermore, in the Open Data Handbook (Dietrich et al., 2012), the authors state that most of the data generated by governments are public data by law, and therefore they should be made available for others to use where privacy of citizens and national security issues are not challenged. According to the Open Government Data definition ("Welcome to Open Government Data," 2014), there are three main benefits that governments can obtain by opening up their data: transparency, participation and collaboration.

Acquiring and processing the amount of data generated by Governments may lead to workloads that are beyond the capacity of a single computer. Fortunately, the emergence of new technologies, such as Cloud Computing, makes it easier to scale the data processing demands in a seamless and scalable manner (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009). Whilst for some disciplines and domains where finer grained security is an impediment to adoption of Cloud computing, e.g. medicine, open data has by its very nature, no such impediments. Cloud computing also encourages the creation of more innovative services including those based on processing and analyzing datasets made available by governments. The sharing of technologies as open source solutions also goes hand in hand with open data initiatives.

The aim of this paper is to describe an approach taken to leverage the benefits provided by Open Data from the Australian government using Cloud-related technologies through the Australian national cloud facility: National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) and specifically the NeCTAR Research Cloud. The paper begins with a brief introduction to Open Data, providing its definition, its benefits and also its potential disadvantages. We then describe the advantages of using Cloud Computing to deal with Open Data. The details of the approach taken to harvest, clean and store publicly available data from Australian government resources followed by their analyses and visualizations of these datasets is given. Finally, the paper concludes by

pointing out the importance of Open Government Data and the role of Cloud Computing to leverage the benefits offered by Open Data. It is emphasized that there is no causality implied in this paper regarding the analysis of the data offered. However we strongly believe that open discussions about causality are an essential element in the transparency of Government more generally.

2 Open Data

2.1 Definition

The Open Knowledge Foundation defines Open Data as 'any data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement of attribute and/or share-alike' (Doctorow et al., 2014). We emphasize two important conditions that are not clearly mentioned in this short definition. First, data can be considered as open if it is easily accessible which means that data should be available on the Internet and in formats that are machine readable. Second, the terms reuse and redistribute include the possibility of intermixing two or more datasets in order to discover relations that would not be visible when having the datasets separated. The full definition provided by the Open Knowledge Foundation (Doctorow et al., 2014) gives further details of the conditions that should be satisfied by data to be considered as open. The final purpose of all these conditions required by Open Data is to ensure the potential interoperability of datasets, i.e. it is possible to combine any number of these datasets and subsequently identify their inter-relationships. Ideally this should be part of a larger system as opposed say to having many individual data sets (e.g. spreadsheets). The true power of Open Data is derived from the analytical tools and capabilities used to identify patterns that would otherwise remain hidden across multiple, diverse data sets.

2.2 Open Government Data

Governments are constantly gathering data from many types of sources: the population, taxes, quality of life indicators and indeed anything that could help the government to monitor and improve the management and governance of their country. Historically, only governmental entities (departments) have had access to process and analyze these data. However, according to (Davies, 2010; Dietrich et al., 2012; Lathrop & Ruma, 2010; Robinson, Yu, Zeller, & Felten, 2008), most of the data collected by government is public by law and therefore, it should be made open and available for everyone to use. In some cases, when governments have failed to make data easily accessible, citizens have had to find alternative ways to harvest and process these data to give it a meaningful use. A well-known case is the portal GovTrack.us which was launched in 2004 by a student who harvested a set of government data and published it in more accessible formats. This kind of initiatives have influenced in governments' decisions to make government data publicly available (Brito, 2007; Hogge, 2010; Lathrop & Ruma, 2010). It should be noted also that government does not always share data effectively across its own departments – here the data includes both open and non-open data. The government departments of immigration, employment, education, health, transport, etc. all have subsets of the total "government" data, but the use of this data in integrated frameworks by government is currently lacking.

Since 2009, various countries have started Open Data initiatives by launching portals in which they publish government datasets to be downloaded by anyone. Among these countries are the USA (data.gov), the UK (data.gov.uk), Australia (data.gov.au), Ghana (data.gov.gh) and New Zealand (data.govt.nz). These sources of data are useful but do not include the tools to compare all of the data sets in any meaningful manner. Instead they are typically large collections of documents and individual (distinct) data sets. Often they are available as spreadsheets, CSV files with no means for direct comparison or analysis across the data sets.

2.3 Benefits

Many authors (Brito, 2007; Davies, 2010; Dietrich et al., 2012; Hogge, 2010; Lathrop & Ruma, 2010; Robinson et al., 2008) agree about the benefits that can be obtained by governments when they decide to open up their data, namely: transparency, participation and collaboration. These benefits are directly derived from the corresponding Open Data requirements: freedom of use, reuse and redistribution. In this context, the fact that anyone is free to use government

data leads to an increment in government transparency. Hogge (2010), in her study mentions that transparency is not only about citizens trying to find irregularities in government actions, it is also about citizens constantly monitoring their governments' activities and providing feedback to improve processes and public services, and according to the Open Government Data definition ("Welcome to Open Government Data," 2014), this is what defines a well-functioning democratic society.

Open Data not only requires data to be accessible, but it requires the freedom to reuse these data for different purposes. This allows citizens to combine two or more datasets to create mash-ups and highlight potentially hidden relations between different datasets (Brito, 2007; Davies, 2010; Lathrop & Ruma, 2010). This improves the participation of citizens from different fields such as developers, scientists and indeed journalists. This is particularly important to governments since citizens can experiment in the creation of new services based on government data and the government is subsequently able to evaluate the most useful services and where appropriate shape future policy based on new knowledge. This has the added value of encouraging the participation of more citizens in government activities and increases the number of new services that could benefit the government.

The third key benefit of Open Data is collaboration which is directly derived from the freedom of users to redistribute government data, e.g. combining two or more datasets for a specific purpose and making the resulting dataset available for others to use. In this way, citizens are collaborating with each other while they are contributing to the government by creating services and solving problems. In some cases, this model of combining data sets to develop new, targeted solutions has spurned a range of start-ups and industries, e.g. San Francisco and the Civic Innovation activities (http://innovatesf.com/category/open-data/)

Although the process of making data publicly available can be seen as laborious and cost intensive to the government agencies involved, it brings further economic benefits to governments since it will improve the participation of people in the creation of innovative services (Hogge, 2010).

2.4 Barriers

According to (Davies, 2010; Lathrop & Ruma, 2010), transparency should not be focused only on the accountability and transparency of government. In fact, this could generate an excessive attention to government's mistakes and consequently, create an image of government as corrupt. This is clearly a reason why governments might not want to open up their data. However, the authors state that instead of reducing transparency, this problem could be addressed by creating a culture of transparency that not only judges when public entities behave badly, but a culture that is also capable to register approval when governments successfully solve public problems or deliver services in a cost effective manner.

Furthermore, many governments and indeed individuals are concerned about the privacy of citizens. Although, it is possible to anonymize datasets before they are made publicly available, it requires considerable time, effort and expense of public workers and sometimes it is not possible to guarantee that the data will be fully anonymized (Lathrop & Ruma, 2010). For this reason, some governments prefer to keep the data private. However it is the case that often terms such as protecting national security or citizen privacy are used as a blanket to deny access to many other data sets that are not contentious.

Additional barriers that stop governments making data publicly available is the fact that many data sets are stored on older forms of data storage media such as paper files and proprietary databases which do not allow for easy extraction and publication. Furthermore open data also requires appropriate (rich) metadata to describe it: the context in which it was collected, by whom and when. In some cases, this additional information is not directly available.

2.5 Disadvantages

Data can be open to misinterpretation, which can subsequently generate civic confusion and extra problems for governments. For instance, (Lathrop & Ruma, 2010) mentions a case where people correlated locations of crimes in a city with the number of immigrants in that location and make conclusions like "This is a high crime neighborhood because many immigrants live

here". Something which is not necessarily true as many other aspects must be taken into account to determine the reasons of high levels of crimes in a location.

Another disadvantage of publicly available data is for the potential for it to be manipulated with the intention of satisfying personal interests. This is difficult for a government to control and could be problematic since people often do not always verify data before making conclusions. Key to tackling this is the spirit of open data: it should be possible to verify or refute the conclusions that are drawn by access to the original data sets. Thus for any data that is accessed (harvested) it should always be possible to go back to the original (definitive) sources of the data (since it is open).

3 Cloud Computing

Open data benefits greatly by access to open data processing platforms. Cloud computing offers one approach that is directly suited to the processing of open data. The National Institute of Standards and Technology (NIST) (Mell & Grance, 2011), points out five essential characteristics that define the Cloud Computing model: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. In order to adapt to different types of users, Cloud providers offer three levels of abstraction: Software as a Service (SaaS) with examples being Salesforce's CRM and Google Docs; Platform as a Service (PaaS) with examples being Microsoft Azure and Google App Engine, and Infrastructure as a Service (IaaS) with examples being Amazon EC2, Rackspace and the Australian NeCTAR Research Cloud. There are also many different Cloud deployment models: Private Clouds, Public Clouds, Community Clouds, and Hybrid Clouds (Mell & Grance, 2011; Sriram & Khajeh-Hosseini, 2010; Velte, Velte, & Elsenpeter, 2009; Zhang, Cheng, & Boutaba, 2010). Ideally open data should be processed on open Clouds and the applications and interpretation of the data utilizing open sources data models for complete transparency of the data and the associated data processing pipelines.

One of the main reasons for the success of Cloud computing is the capacity to rapidly scale up or scale down on demand, at an affordable cost and ideally in an automated fashion. This is particularly important when working with government data as they can become quite voluminous, they can change over time, they require veracity of information to be checks, and when comparisons and analyses are made between data sets these can result in computationally expensive requirements. Cloud Computing is especially suited to this environment since it is possible to scale out resources to satisfy needs and (in principle) pay for those extra resources only for the time that are actually being used. This is convenient specially for people considered 'civil hackers' who create services based on government data and most often without financial reward (Davies, 2010; Hogge, 2010). This contributes to the emergence of new questions and reduces the time needed to answer these questions, which encourages people to collect more data and create more innovative services.

The Cloud provider utilized here is the NeCTAR Research Cloud, which is an Australian government funded project that offers an IaaS platform with free access to Australian academics, or more precisely members of organizations subscribed to the Australian Access Federation (AAF – www.aaf.edu.au) such as the University of Melbourne. This work utilised two virtual machines (VMs) each with 2 cores, 8GB RAM and 100GB of storage. While the VMs were located in different zones, both have the same architecture (Figure 1). This allowed them to act as master at any time providing high availability to the system.

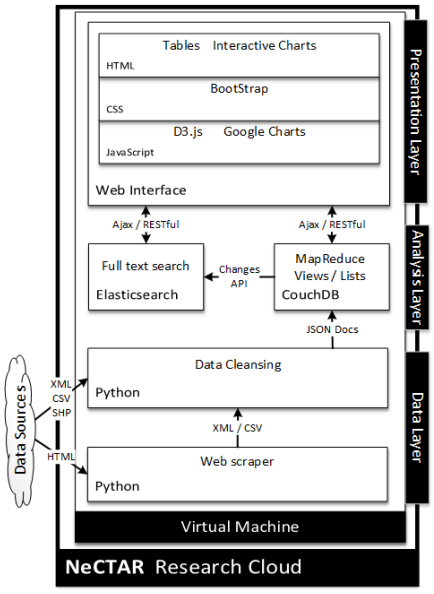

Figure 1. Architecture.

4 Implementation of the Open Government Data Access, Process and Visualise Platform

4.1 Data Layer

The key focus of the work is on access to and use of open government data. A *Data Layer* that harvested and processed these data was key to this. This layer was responsible for dealing with raw data coming from external sources. The data sets that were harvested and used as the basis for the work described here included:

- Australian Electoral Commission (www.aec.gov.au)
  - Annual Returns (2003 - 2013) (includes: party returns, political donations, Associated Entities, Political Expenditure)
  - Election Returns (2003 - 2013) (includes: donations to candidates, donors details, senate groups)
  - Election Results (2004 - 2010) (includes: Details of Federal election results divided in general, house and senate)
  - Federal electoral boundary GIS data (Census 2011)
- Portal data.gov.au
  - Historical Australian Government Contract Data (1999 - 2013)
  - Members of Parliament webpages and social networks
  - Portfolio of Responsibilities
- Parliament of Australia (www.aph.gov.au/Parliamentary_Business/Hansard)
  - House Hansard
  - Senate Hansard
- Lobbyists Details
  - Australian Government (www.lobbyists.pmc.gov.au)
  - Victoria (www.lobbyistsregister.vic.gov.au)
  - Queensland (www.lobbyists.integrity.qld.gov.au)
  - Western Australia (www.lobbyists.wa.gov.au)
  - Tasmania (www.lobbyists.dpac.tas.gov.au)
  - New South Wales (www.dpc.nsw.gov.au)
  - South Australia (www.dpc.sa.gov.au)

The analyses and visualizations of these data that drove and shaped the work were based on: political donations, election results, historical contracts data and political speeches. These data were selected following with researchers at the Centre for Advanced Data Journalism at the University of Melbourne. The Data Layer itself was divided into three stages: data harvesting, data cleansing and data storage which are described here.

Data Harvesting

It should be noted that most of the datasets that were harvested satisfy the requirements of Open Data, i.e. they are downloadable and are provided in machine-readable formats such as CSV and XML. It is also noted that there are other data that do not satisfy all of these requirements. For instance, the lobbyist registers for all the Australian States are available only in HTML (via web pages). In this case, it was necessary to implement web scrapers for webpages to extract the data and then store it in the database. This technique is inefficient and has several disadvantages for how data can be released as open data and subsequently used and interpreted. Firstly, it is error prone because a scraper may assume that a webpage follows a standard but there is the possibility of mistakes in the scraped HTML, which would cause the scraper to obtain erroneous data. Furthermore, it is a tedious task since it is almost impossible to build a scraper that works with many webpages as different sites use completely different designs. Lastly, the design of a webpage can change without any notice, which would render a given scraper totally useless and require a new scraper to be produced. Nevertheless, it is an effective technique when used carefully and after ensuring that all data obtained is verified before performing further analyses and interpretations. The information should also include metadata on when the data was accessed and scraped.

Additionally, even when data is made available in a more accessible (downloadable) format, further work is often required. For example, despite the fact that the Hansard political speeches of Australia are provided as downloadable XML files, there is no way to download the whole collection of speeches or the possibility of selecting a range of speech dates that could be downloaded. Consequently, it is often necessary to download one file at a time,

which makes the process inefficient taking into account that there are thousands of files. As a result, whilst the data is open, the way in which it is made available is not really conducive to further processing without computational approaches to overcome these limitations, e.g. implementing processes to download all of the XML files.

It is recognized that the difficulties faced in harvesting public data are understandable since many governments (including the Australian government) are still in the process of opening their datasets and learning about how best to do this. These lessons are often spotlighted through important initiatives such as organized events used to receive feedback from data enthusiasts about how to improve the available datasets or which new datasets could be made available. For instance, GovHack is an annual event which brings together people from government, industry, academia and general public to experiment with government data and encourage open government and open data in Australia. Additionally, there exist various open data portals around Australia including the national portal data.gov.au, portals for every State such as the http://data.nsw.gov.au and even some cities like Melbourne have launched their own open data portals, e.g. https://data.melbourne.vic.gov.au/.

Data Cleansing

Every dataset collected will typically contain some extra and/or useless data that needs to be removed in order to improve the quality of data and increase the consistency between different datasets allowing them to be combined and interpreted more easily. To aid in data consistency, datasets from different formats such as CSV or XML were converted to JavaScript Object Notation (JSON) objects. Although, this process was simple, there were some difficulties to overcome in specific datasets. For instance, the XML files of the Hansard political speeches have different structures over different time periods, which made the process of parsing the whole collection more complex. However, it was possible to find certain levels of consistency in most of the datasets, which allowed use of Python scripts to convert hundreds of datasets and then store them in the database.

Data storage

Due to the variety of sources and the lack of a well-defined schema, CouchDB was selected as an appropriate database to store all the harvested data. CouchDB is a schema-free NoSQL and document-oriented database (Anderson, Lehnardt, & Slater, 2010). It stores its documents as JSON objects. In this model, each row of each dataset was stored as an independent JSON document adding an extra field “type”, and in some cases “subtype”, in order to facilitate the exploration of different datasets in the database.

Although both VMs were set up to act as a master at any given time, in this stage one VM could be considered as master and the other one as slave because only one of them could harvest data from external sources at a time while it replicated all the new data to the other. CouchDB provides strong replication processes that allow setting up a bi-directional replication between the databases in each VM. This allowed having both databases up to date while only one of them was harvesting data.

4.2 Analysis Layer

In addition to the flexibility provided by CouchDB to store schema-free documents, one of the main reasons to choose this database was its support for MapReduce based views. MapReduce is one of most effective approaches to deal with large-scale data problems and allows to separate what computations are performed and how those computations are performed (Buyya et al., 2009; Dean & Ghemawat, 2008; Ekanayake, Pallickara, & Fox, 2008; Lin & Dyer, 2010; Segaran & Hammerbacher, 2009; White, 2012). Therefore, to analyze the data the developer only needs to focus on the first part which consists on writing two functions: a map function and a reduce function. The run-time system handles how those computations are performed by managing failures, schedules and intercommunication. The complexity of map and reduce functions can be diverse and depends on the type of analysis to be performed on the data.

Furthermore, CouchDB documents and views are indexed using a B-Trees data structures, which are very efficient for storing large amounts of data (Anderson et al., 2010; Bayer, 1997). The index for a view is created only the first time that the view is queried and allows to retrieve large amount of data very quickly. In

order reflect the current state of the database, the index of a view only needs to introduce the documents that have changed. Although this process is very efficient, it can introduce high latency to queries when a large amount of documents have changed (Anderson et al., 2010). This is a common problem faced by applications where documents in the database tend to be updated frequently. However, since the type of data used in this project is largely historical and not changing dynamically, CouchDB views were used successfully.

Most of the data analyses where performed using CouchDB views, these analyses included political donations over time, data aggregation of donations such as retrieving the largest donation in certain election period and correlation between different datasets, for instance, donations *vs* votes received by a political party. However, there were some cases where it was not possible to perform more complex analyses using only CouchDB views. For example, despite the fact that CouchDB owes many of its advantages to B-Trees, it also inherits one of its main drawbacks which is the inability to perform multi-dimensional queries (Bayer, 1997). In other words, CouchDB views are excellent to process queries such as the sum of donations received in the year 2004 (point queries) or the sum of donations received between 2004 and 2010 (range queries). However, for multi-dimensional queries such as the sum of donations received by a candidate from a specific donor in 2004 (4-dimensional), there were challenges that required support for other data processing capabilities. For this kind of query it was required to provide a visualization that showed an overview of the political donations. This visualization was required to group, color and filter donations in multiple ways and shows a summary for every group of donations. The summary includes the total sum of donations, the number of donations in that group, details of the largest donation and the top 3 donations received by candidates, parties and States. In order to solve this multi-dimensional query limitation, CouchDB functionalities were extended with ElasticSearch.

ElasticSearch is a distributed search engine built on top of Apache Lucene, which among other features provides full text search capabilities whilst hiding the complexity of Lucene behind a simple and coherent API. In spite of the document storage capabilities of ElasticSearch, it is mainly used as an extension for NoSQL databases thanks to a range of available plugins. For instance, it offers the possibility of indexing any CouchDB database through a plugin that listens to the changes API of CouchDB making the database searchable and allowing to perform more complex queries and more complete analyses of the data (Gormley & Tong, 2014). Using CouchDB in conjunction with ElasticSearch allows taking advantage of the most important features provided by each technology, namely durability and advanced search capabilities respectively. The number of features offered by ElasticSearch is vast and more details can be found in (Gormley & Tong, 2014).

ElasticSearch was also useful to build visualizations that correlate the political donations and the government contracts by searching all the occurrences of the donors' names in the dataset of government contracts. This was done through the Python API client provided by ElasticSearch and a Python script which returned a list of donor names that appeared in the government contracts dataset indicating the field where it was found, this helped to show the correlation of both datasets in a timeline.

4.3 Presentation Layer

Visualisation is essential when dealing with large-scale heterogeneous data sets. Indeed all of the data analyses would have limited value if it were not possible to visualize them in a human-friendly way. This is especially important in open government data initiatives where the focus is less on the detailed scientific model of discovery and more on the big picture questions that can be illustrated through the data itself. The presentation layer was based mainly in JavaScript using the D3.js library, Google Charts API and jQuery. In this section we illustrate some of the visualizations for the analyses mentioned previously.

Figure 2 shows an example of the multiple ways of visualizing political donations through one of the visualizations that were built. Each bubble in the figure represents a donation received by a candidate, the size of the bubble represents the amount of money donated, the color in this case, represents a political party and

each group of bubbles is an election period. This is an interactive visualization so, donations could be grouped, colored and filtered by all the features contained in the dataset which include election period, candidate, party, electorate, electorate state, donor, donor state, donor suburb, and nil return. Furthermore, the labels for each group (including the main title) are clickable and they contain the summary for every group of donations and the main title contains the summary for all the four groups.

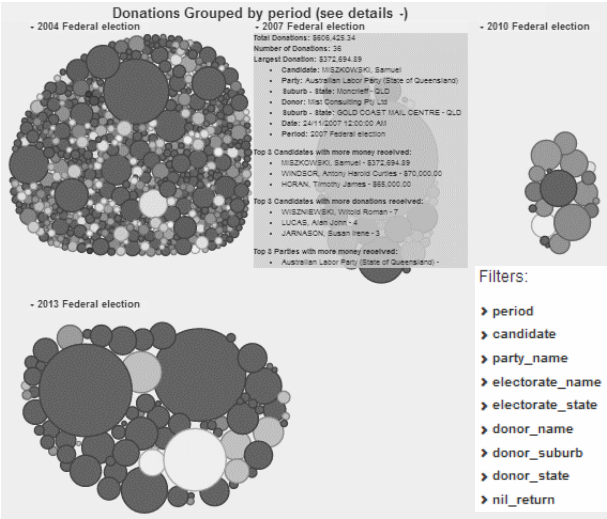

Figure 2. Overview of Donations.

This visualization facilitates a detailed exploration of the donations dataset and could be considered as a starting point for further analyses.

Another way of visualizing the donations is on a timeline as exposed in Figure 3. This shows the total number of donations received by date. Something interesting to point out here is how we can see that most of the peaks are in the years 2004, 2007 and 2010, years in which federal elections have taken place. This pattern of donations increasing in election years is also visible when looking at donations made by individual entities. Figure 4 illustrates all the donations made by an entity over time and highlights the tendency of making more donations in election years.

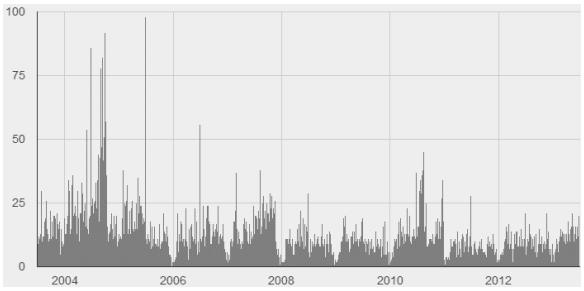

Figure 3. Summation of Political Donations Visualised over Timeline.

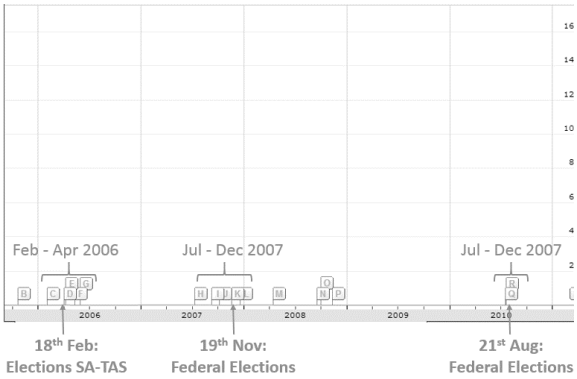

Figure 4. Individual Donations made by a given entity over time.

An additional scenario of interest is the correlation between political donations and government contracts, i.e. grants/award made to entities (most often companies). With the results obtained from the ElasticSearch analysis described in the previous section, donations and contracts were displayed in a timeline to explore whether the number of donations or the amount of money donated by an entity influenced (was correlated with) the number and the value of contracts that they subsequently obtained. Figure 5 shows this scenario for a specific entity.

It can be seen that there are several cases where contracts are obtained right before or after

some donations have been made. In addition to the graph showed in Figure 5, this visualization also provides the details of the donations and contracts related with the entity being analyzed. Thus one can see the persons involved as well as political parties and governmental agencies and try to find more patterns to perform further investigations. For instance, a next step might be to investigate who is on the board of the companies making donations and to see if there exists a direct or indirect relation with the governmental agency that is offering the contract. It is emphasized that this is only one of the many scenarios that can be visualized with this analysis and there did not exist a clear correlation between the two datasets in many of the cases. However, this specific scenario helps us to demonstrate how mash-ups highlight hidden relations between apparently unrelated datasets. For transparency of government it is important to ensure that where major grants are awarded, independent review of political donations prior to the award can be scrutinized to ensure independence of government.

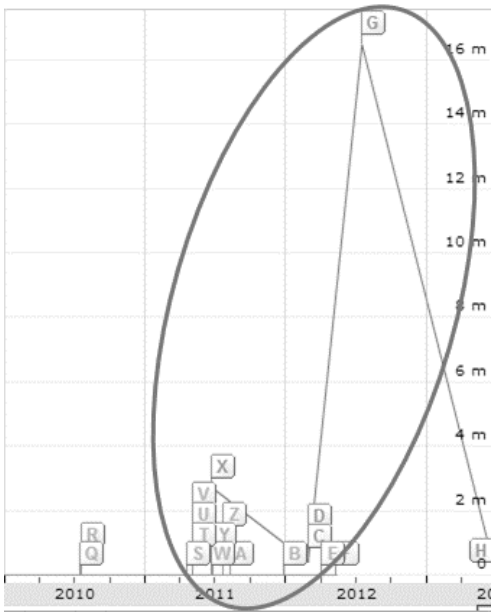

Figure 5. Colour-coded Correlation of Donations (A, B, E, F, Q-W, Y, Z) vs Government Contracts/Awards (C, D, G, H, X).

A further visualization is illustrated in Figure 6, which shows the correlation of terms used in political speeches over time. The figure demonstrate the correlation between the terms “boats” and “immigration” and it indicates how both terms tend to be used in the same dates. This visualization is useful to get an idea of what topics are being discussed by the members of the parliament in different periods.

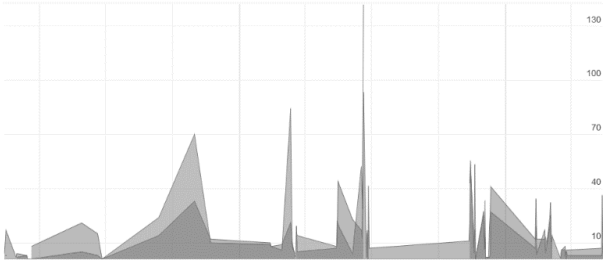

Figure 6. Political Speeches: correlation of words used over time.

An additional visualization using word clouds (Figure 7) was implemented to explore the most popular terms used by politicians in their speeches. This visualization allows to see a general word cloud for all the politicians in the collection of speeches and provides the possibility of filtering by year, month and day as well as selecting up to three politicians to show a comparison of the terms used by each of them over time. These word clouds provide a simple overview of the topics discussed by each politician. For instance, in Figure 7 the word cloud on the left belongs to the Shadow Minister for Transport and Infrastructure and so we can see that the most popular words are highly related to this charge such as “infrastructure”, “transport”, “highway”, “safety”, and “airport”. The word cloud on the right shows the words used by the Prime Minister in his speeches in May 2014 which is the month when the federal budget was presented to the parliament. In this case, we can see that the words “budget”, “deficit”, “spending”, and “tax” are amongst the most popular ones. This demonstrates that word clouds give us an idea of the topics that are dealt in parliament in different periods of time by different politicians. The combination of terms used in speeches and decisions made in award of contracts are also essential to correlate, e.g. speeches about the important of the Australian car industry should not be correlated/associated with political donations from car manufacturers for example if government is to be truly transparent and ultimately accountable for the independence of the decisions it makes.

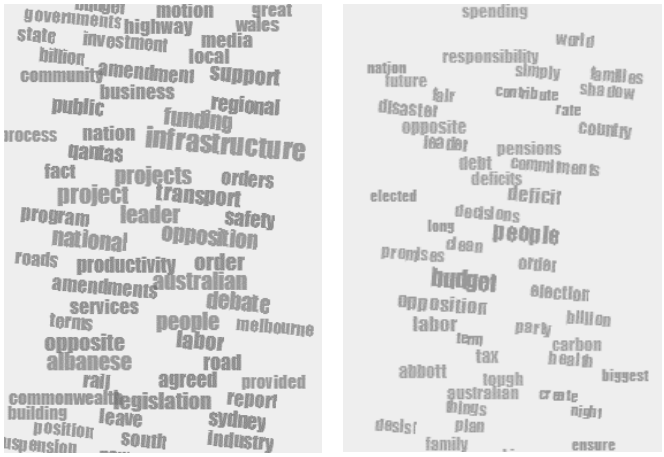

Figure 7. Word clouds comparing terms used by two politicians.

5 Conclusions

This paper presents an introduction to Open Data and points out how it could help governments to improve transparency and accountability. Furthermore, it describes some reasons why governments refuse to engage in Open Data initiatives as well as the existing disadvantages encountered if they are not managed correctly. The work described how and why Cloud Computing provide an appropriate environment for working with Open Data and identified and presented one of the many approaches that can be taken to set up this environment and the associated technologies involved. It also identified some of the common challenges faced by projects that deal with publicly available data and the methods used to overcome these challenges. Moreover, it showed multiple ways of visualizing data and how different datasets could be correlated to explore a portfolio of government data that is openly available on the web.

This work has many refinements that are currently ongoing. Incorporation of further data, e.g. membership of companies by politicians/their families/associates, as well as exploring social media use. The use of Twitter in particular offers a rich source of Open Data that can be accessed and used to help promote the overall information of government. Who is following whom on Twitter; who tweets on what topics; what is their sentiment on particular topics and how does this change over time are all on-going activities that are being pursued.

In all of this, it is emphasized that the purpose of this work is not to draw conclusions on any given government activity – this is the responsibility of others, e.g. investigative journalists. However for truly democratic and transparent governments it is essential that the data can be reviewed and analysed and stand up to public scrutiny. We strongly encourage this endeavor. All of the software and systems used in this work are also available. The existing prototype system is available at http://130.56.249.15/proj/.

Acknowledgments

The authors are thankful to the feedback and discussions that have shaped this work. Specifically we would like to thank Prof. Margaret Simons (Director of the Centre for Advanced Journalism at the University of Melbourne).

References

Anderson, J. C., Lehnardt, J., & Slater, N. (2010). *CouchDB: the definitive guide*: O'Reilly Media, Inc.

Bayer, R. (1997). The universal B-tree for multidimensional indexing: General concepts *Worldwide Computing and Its Applications* (pp. 198-209): Springer.

Brito, J. (2007). Hack, mash & peer: Crowdsourcing government transparency.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems, 25*(6), 599-616.

Davies, T. (2010). Open data, democracy and public sector reform. *A look at open government data use from data. gov. uk. Über: http://practicalparticipation.co.uk/odi/report/wp-content/uploads/2010/08/How-is-open-governmentdata-being-used-in-practice. pdf*.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107-113.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., et al. (2012). The Open Data Handbook. 2014, from http://opendatahandbook.org/en/

Doctorow, C., Suber, P., Hubbard, T., Murray-Rust, P., Walsh, J.,

Tsiavos, P., et al. (2014). The Open Definition. 2014, from http://opendefinition.org/

Ekanayake, J., Pallickara, S., & Fox, G. (2008). *Mapreduce for data intensive scientific analyses*. Paper presented at the eScience, 2008. eScience'08. IEEE Fourth International Conference on.

Gormley, C., & Tong, Z. (2014). *Elasticsearch: The Definitive Guide*: O'Reilly Media, Inc.

Hogge, B. (2010). Open data study. *a report commissioned by the Transparency and Accountability Initiative, available for download at: http://www.soros.org/initiatives/information/focus/communication/articles_publications/publications/open-data-study-20100519*.

Lathrop, D., & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice*: O'Reilly Media, Inc.

Lin, J., & Dyer, C. (2010). *Data-Intensive Text Processing with MapReduce*: Morgan and Claypool Publishers.

Mell, P., & Grance, T. (2011). The NIST Definition of Cloud Computing.

Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2008). Government data and the invisible hand. *Yale JL & Tech., 11*, 159.

Segaran, T., & Hammerbacher, J. (2009). *Beautiful data: the stories behind elegant data solutions*: O'Reilly Media, Inc.

Sriram, I., & Khajeh-Hosseini, A. (2010). Research agenda in cloud technologies. *arXiv preprint arXiv:1001.3259*.

Velte, T., Velte, A., & Elsenpeter, R. (2009). *Cloud computing, a practical approach*: McGraw-Hill, Inc.

Welcome to Open Government Data. (2014). 2014, from http://opengovernmentdata.org/

White, T. (2012). *Hadoop: The definitive guide*: " O'Reilly Media, Inc.".

Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications, 1*(1), 7-18.