

HANSARD DATA MINING PROJECT PLAN

Aaron Butler

Bipin Karki

Katherine Noack

Mahmoud Yousefi

August 16, 2019

Executive Summary

The Hansard data mining project is focused on enabling the Auditor-General's Department (AGD) staff to better understand, interrogate and summarise key topics discussed between parliament members whilst formally sitting in the South Australian Government Legislative Council and House of Assembly. These discussions are documented in the official Hansard¹.

This project uses all Hansard documents as the primary data source that is stored in the publicly available Hansard website (<http://hansardpublic.parliament.sa.gov.au>)

A minimum of one year of data will be extracted and analysed using different text mining and text analytics techniques. However, advanced text analytic techniques will not be used due to time constraints.

A dashboard, potentially developed using Tableau Desktop, will be provided to the AGD to visualise trending discussions topics in parliament and issues where AGD and their clients are mentioned.

At the conclusion of the project, AGD will also be provided with instructions on how to deploy and run the automated processes, code, and dashboard(s) that have been developed.

This project will help AGD employees save time and better interrogate text compared to the current manual search used by various AGD audit teams. This project will cost approximately \$60,284 for a team of four to complete the deliverables in 13 weeks starting from 2nd August 2019.

¹ Hansard is the complete record of Parliamentary debates and questions.

Project Background and Description

The AGD contributes to public sector accountability in South Australia by providing independent assurance to the Parliament that government activities are conducted and accounted for properly and in accordance with the law (AGD, 2019). The Auditor-General is appointed by Parliament according to the *Public Finance and Audit Act 1987*. This Act establishes the Auditor-Generals's mandate and specifies the financial reporting obligations of the Treasurer and public sector agencies (AGD, 2019).

According to AGD (2019) the Auditor-General's responsibilities are to:

- conduct and report on the financial report and controls audits of the Treasurer and public sector agencies
- conduct and report on special audits relating to accountability and probity
- examine publicly funded bodies at the request of Parliament, the Treasurer, a Minister or the Independent Commissioner Against Corruption
- undertake examinations of the local government sector
- examine issues referred by whistleblowers and members of the community
- review and report on summaries of confidential government contracts

Hansard is a division of South Australia (SA) parliament that keeps the official record of the debates of the parliament (Hansard, 2008). Currently, auditors at AGD manually read through the Hansard records from the SA parliament and find the relevant topics to assist with their audit planning. To save staff time and to reduce manual overhead the AGD Data Analytics team and the University of South Australia have created the Hansard data mining project.

The main objective of this project is to develop a proof-of-concept dashboard for AGD to analyse unstructured text data sources, focusing on Hansard records. The dashboard would allow staff to better interrogate and summarise Parliament discussions in Hansard to identify relevant information for their audits. It would ingest Hansard extracts from Parliament's website and provide a dashboard for AGD auditors to review. It is anticipated that this dashboard will assist in speeding up AGD's audit process.

Project Scope

Inclusions

The following aspects of the project are considered within the scope:

- Identification and analysis of customer requirements
- Build an automatic web scraping process to extract Parliament discussions from Hansard web page
- Processing of text data scraped from Hansard web page into a format suitable for storage in a database and later analysis.
- Manage the storage of the scraped data in a database
- Data analysed will be at least one year of Hansard Parliament Discussions
- Exploratory approaches for initial data analysis and visualisations
- Background research on text analysis (such as topic modelling and Latent Dirichlet Allocation), dashboard design and visualisation methods
- Text analysis of the collected data
- Search functionality for clients and topics over collected data
- Building the visualisation dashboard using a third-party application (such as Tableau Desktop)
- The project will be developed in R or Python, making use of existing packages and libraries
- Version control for the project will be maintained in GitHub

- Documentation of deliverables, including instructions on the deployment of the process to scrape, store and visualise Hansard data
- Process and products produced will be deployable by Auditor-General's Department

Exclusions

The following aspects of the project are explicitly out of scope:

- Advanced text analytics techniques, such as sentiment analysis. Due to time constraints, advanced techniques cannot be adequately developed or tested.
- Data sources outside of those listed in scope inclusions such as Parliament Minutes of the Proceedings
- Predictions or forecasting from scraped data
- Web application development
- Evaluation of third-party software such as Tableau
- Evaluation of Hansard data accuracy. Due to time constraints this project will not investigate the accuracy of Hansard transcripts compared to the original discussions.

Risks

The table below outlines the risks to this project, their probability, impact and the mitigation strategy for each risk. Some risk factors that are specific to this project are due to several team members being unavailable for in-person meetings due to other commitments and the use of Hansard Parliament discussions as the primary data source.

Using the Hansard Parliament discussions as the main data source has some risks involved surrounding data quality, and potential scope creep due to the project team being unfamiliar with this data source. There is also the risk of the project being delayed due to different parts of the project taking longer to develop than expected. There are several discrete parts to the project involving data scraping, data processing, data storage, text analysis and development of one or more dashboards. A delay in one or more of these parts will delay subsequent parts of the project.

Risk	Probability (out of 1.0)	Impact	Mitigation Strategy
Team members unable to work on-site	0.9	Will affect the distribution of work amongst team members and in-person attendance at client meetings	Use of Microsoft Teams and conference calls to communicate with clients. The project team will identify what parts of the project can be completed off-site by external team members.
Changes in Government department names can cause challenges in conducting analysis. For example, the Department for Health and Wellbeing was formerly Department for Health and Aging (University of Adelaide, 2018).	0.8	Government departments could be identified as separate entities when they are the same department. Thereby decreasing the quality and usefulness of the analysis in AGD audits.	AGD will provide the project team with a list of departments that have changed names. The project team will conduct a data quality check to ensure that all department name changes have been captured.
Delays in receiving timely responses and feedback from key stakeholders due to operational commitments	0.5	Deliverables do not meet business requirements	AGD will establish a project reference group to coordinate and manage stakeholder feedback. Andrew Corrigan to escalate any

Risk	Probability (out of 1.0)	Impact	Mitigation Strategy
			<p>concerns raised by the project team.</p> <p>In the event that the above two mitigation strategies are not working, agreed communication timeframes will be set between AGD staff and the project team.</p>
Unable to find an appropriate platform to complete the project that meets confidentiality restrictions	0.2	Unable to complete the project to AGD requirements	Identify possible platforms early in the project and gain approval from AGD on their use in the project. For example, the use of a platform such as Microsoft Azure.
Lack of team member contribution	0.1	Impact on deliverables and schedule	Team members are to update team regularly on their progress, so that lack of contribution/progress is detected early and can be completed by another team member if required to meet project schedule and deliverable deadlines.
Scope creep	0.5	Variation in scope may occur	Update clients with project progress regularly to ensure that it meets their requirements and seek clarification of requirements at the early stages of the project. Clients will be updated on progress at least once a week in person at client meetings, by email or Microsoft Teams.
Team member leaving the group	0.1	Impact on deliverables and schedule	If a team member leaves the project, dashboard functionality may be reduced so that the deliverable deadline can be met
Very few debates on the Hansard dataset related to Auditor-General's Department and its clients.	0.4	Relevance of the project to AGD audits may be reduced	Clients are to be updated regularly so that they can be made aware of any issues that could impact on the usefulness of the project.
The project team consists of only four team members to conduct the project in a short time frame which may result in some functionality being unable to be delivered	1.0	Not all desired functionality for the dashboard and analysis may be completed by the project deadline	<p>Employ an Agile approach to develop highest priority features first.</p> <p>Use of planning software to keep track of tasks to be completed and deadlines.</p>
Data scraping of Hansard web site takes longer than the time allocated	0.4	The completed dashboard may not contain the full set of data.	<p>Employ an Agile approach by building a proof-of-concept process using a smaller set of data.</p> <p>Use of planning software to keep track of tasks to be completed and deadlines. Additional time will be allocated to the early stages of the project that is crucial to the progression of the project.</p>
Text analysis development takes longer than the time allocated	0.4	Dashboards do not have the desired functionality	Advanced text analytics will not be completed as part of this project. This will give more time for requested dashboard functionality such as searching, interactivity and the ability to drill-down into data.

Budget

The potential costs associated with this project are described in the table below. A national police check was required to complete this project and is an actual rather than hypothetical budget item. It has not been determined what software will be used for developing the project dashboards. Therefore, Tableau and Elasticsearch are both included in the estimated budget.

Budget Item	Estimated Cost	Justification
Salary for four data scientists	\$52,000	Cost calculated at \$50/hr, 20 hours a week for 13 weeks. Based on the median hourly rate of approximately \$56 for data scientists in Australia (PayScale, 2019)
Team members travel to client meetings	\$2,880	Cost calculated 80\$/day, once a week for 12 weeks for three project members. At least one team member will be communicating with clients via conference call.
Police Checks for the project team	\$141.40	Four police checks costing \$30.35 each through National Crimes Check (National Crime Check, 2019). This includes a 1.5% credit card processing fee.
Azure	\$1,125.64	For Azure virtual machine of 28 GB RAM and 50 GB temporary memory for 730 hours. The cost was estimated using Azure price calculator (Azure, 2019).
Two licenses of Tableau Creator	\$420	The suite of products supports end-to-end analytics workflow. Every dashboard deployment needs at least one Tableau Creator license. Two licenses required for two team members to work on the dashboard(s) concurrently for three months. Tableau Creator license costs \$70 user/month (Tableau, 2019).
Elasticsearch	Free	The free version of Elasticsearch contains the required functionality such as dashboards, visualisation, full-text search and data transformation (Elasticsearch, 2019). Security features are limited in the free version.
SQL Server License	\$3,717	A free developer license can be used during project development. On completion of the project, an SQL Server License will be required to host the database on AGD servers. Estimated cost for the deployment of SQL Server on AGD servers is for 1 core of standard edition license (Microsoft, 2019).
Total estimated cost:	\$60,284	

Roles and Responsibilities

The table below identifies the roles and responsibilities of the four project team members and project mentor.

Name	Role	Justification
Aaron Butler	Data Science Engineer	Experience in software development developed during University courses
Bipin Karki	Data Scientist, Lead Client Contact	Experience in data analysis developed during University courses
Katherine Noack	Product Owner, Developer, Editor	Industry experience in software development, data analysis and dashboard development in Tableau
Mahmoud Yousefi	Data Analyst, Dashboard Designer	Experience in data analysis developed during University courses
Eric Lam	Project Mentor	Assigned by University of South Australia

Communications Plan

General

Individual communication amongst team members, and with the mentor, will be primarily through email and WhatsApp. WhatsApp is a secure messaging service accessible on mobile and personal computers. An email group has been created for this project, which can contain shared documents editable by all team members. A one-hour meeting with team members and the mentor occurs weekly on Monday evenings for the duration of the project.

Client

Team members will meet with the client in-person at the SA Auditor-Generals Department weekly on Thursday 4pm for the duration of the 13-week project. This may change to fortnightly meetings as required. This meeting can also occur by a conference call for team members who are unable to attend in person. Other communication with the client will be through Microsoft Teams. Updates on progress and draft copies of reports will be communicated using Microsoft Teams to the Project Coordinator James Baker at AGD.

Reports

The project plan and final deliverable report will be submitted to the client for approval one week before the submission due dates required by the University. Reports will be submitted to the University according to the project schedule.

Project Stakeholders

The table below outlines the relevant stakeholders and their interest in the project. These stakeholders have an interest in the project because they will use the developed dashboard in their audits or make use of the audit results.

Name	Role	Interest in the project
Dr Gaye Deegan Dr Ben Martini	Course Coordinator	<ul style="list-style-type: none">• Evaluating and make decisions regarding the project documents.• Establishing the relationship with the AGD.• Building the project team and assigning the mentor.• Providing ongoing support to the project team.• Monitoring the project progress
Data Analytics team at AGD - James Baker & Aaron Steicke	Project Owner/ Coordinator	<ul style="list-style-type: none">• Providing an overview of the AGD• Explaining the project objectives in high level• Providing the permits required to start the project
Project reference group	End User	<ul style="list-style-type: none">• Reference group would meet fortnightly for one hour from August to November 2019 to review draft deliverables with the project team and provide feedback. Working iteratively, the project team will refine their approach and dashboard based on this feedback.• It is expected that the reference group will contain the following AGD members:<ul style="list-style-type: none">○ 1-2 members from the Policy, Planning and Standards Directorate○ 1-2 members from the Performance Audit team○ 1-2 members from Financial Audit teams○ Assistant Auditor-General at AGD- Andrew Corrigan representing the IT Audit and Local Government teams.
Assistant Auditor-General at AGD- Andrew Corrigan	Project Sponsor	<ul style="list-style-type: none">• Explaining the project objectives, scope and requirements in more details• Setting up meetings with the joint AGD Project reference group and project team meetings• Providing regular feedback to the team• Facilitating the project at AGD

Deliverables and Critical Success Factors

The items that will be delivered to the client at the conclusion of this project are:

- Automated process for scraping data from the Hansard website
- Process for preparing data into a format suitable for analysis and visualising in dashboards
- Database with one year of past Hansard Parliament discussions that has been prepared for further analysis and visualisation in dashboards
- Dashboards for use in AGD audits
- Instructions on how to deploy and run the automated processes, code, and dashboards that have been developed
- Project closure documents to handover the project to the client

The critical success factors for this project are:

- Dashboards will be interactive and allow the searching of text
- Ability to be able to identify in documents when the Auditor-General's Department of South Australia and its clients are mentioned
- Ability to be able to identify hot topics in parliament discussions
- Software and automated processes developed will be developed and documented in a way that allows them to be deployed by AGD on their own systems

Implementation Plan

The project team will be using an agile work process, as shown in the figure below.

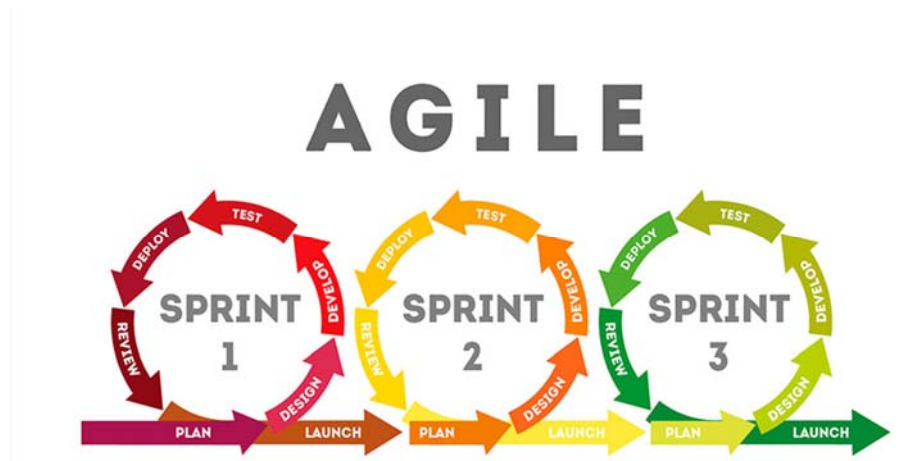


Figure 1: Agile Methodology (300 librarians, 2019)

Agile is a team-based workflow based on an iterative and incremental approach that is often used in software development (Scaled Agile, 2019). Using agile methodology, developers are not told how to build, solve or deliver the product, only “what to deliver”. Agile methodology functions in sprints, where a sprint contains a planned amount of work to be finished within a specified time period. For example, development of code to obtain data from the Hansard web site will be planned for, designed, developed, tested and reviewed in its own sprint. A sprint typically lasts one to four weeks. At the end of a sprint, the results are shown to what’s known as a Product Owner who compares the work to previously established criteria of success. Once a sprint has finished, no more work will be completed on that phase of the project, and the remaining phases of the project will continue as planned.

Why Agile

Agile methodology will be used in this project because it is the preferred working style of the AGD and many other organisations. Agile will assist with tracking tasks as well as the individuals that will be responsible for delivering them. An advantage of this approach is that it focuses on the rapid delivery of software and allows changing requirements to be included in the final product. This is important because of the short duration of the project. Iterative planning and feedback involving developers and the stakeholders also allow the development of a product that closely meets the needs of the client.

Agile methodology will be used in this project to segment the work into manageable parts that can be more easily planned, developed, tested and evaluated with stakeholder input. The following Project Schedule will describe how the project has been divided into phases, sprints and milestones according to the agile methodology.

Project Schedule

The project schedule is shown in a Gantt chart on the next page. This chart, and the table below, outlines the project's planned stages and milestones throughout the duration of the 13-week project.

Week Beginning	Task Name	Duration	Start	Finish
Week 1	Project Initiation Phase	3 days	29/07/2019	31/07/2019
Week 1	Project kick-off and Client Meeting	2 days	29/07/2019	30/07/2019
Week 1	Client Meeting	1 day	31/07/2019	31/07/2019
Week 1	Planning Phase	14 days	1/08/2019	16/08/2019
Week 1	Project Plan Draft Completed	6 days	1/08/2019	8/08/2019
Week 2	Project Plan submitted to the client	1 day	9/08/2019	9/08/2019
Week 3	Project Plan approved by the client	4 days	12/08/2019	15/08/2019
Week 3	Project Plan submitted to University	2 days	16/08/19	16/08/19
Week 4	Project Plan Presentation	1 day	22/09/19	22/09/19
Week 4	Development Phase	70 days	19/08/2019	27/10/2019
Week 4	Sprint 1	14 days	19/08/19	01/09/19
Week 6	Sprint 2	14 days	02/09/19	15/09/19
Week 8	Sprint 3	14 days	16/09/19	29/09/19
Week 10	Sprint 4	14 days	30/09/19	13/10/19
Week 12	Sprint 5	14 days	14/10/19	27/10/19
Week 14	Closing phase in UniSA	6 days	28/10/2019	4/11/2019
Week 14	Making the changes on final documents as per UniSA requirements	4 days	28/10/2019	31/10/2019
Week 14	Submitting the final deliverables on the course webpage	1 day	1/11/2019	1/11/2019
Week 15	Final Presentation	1 day	4/11/2019	4/11/2019

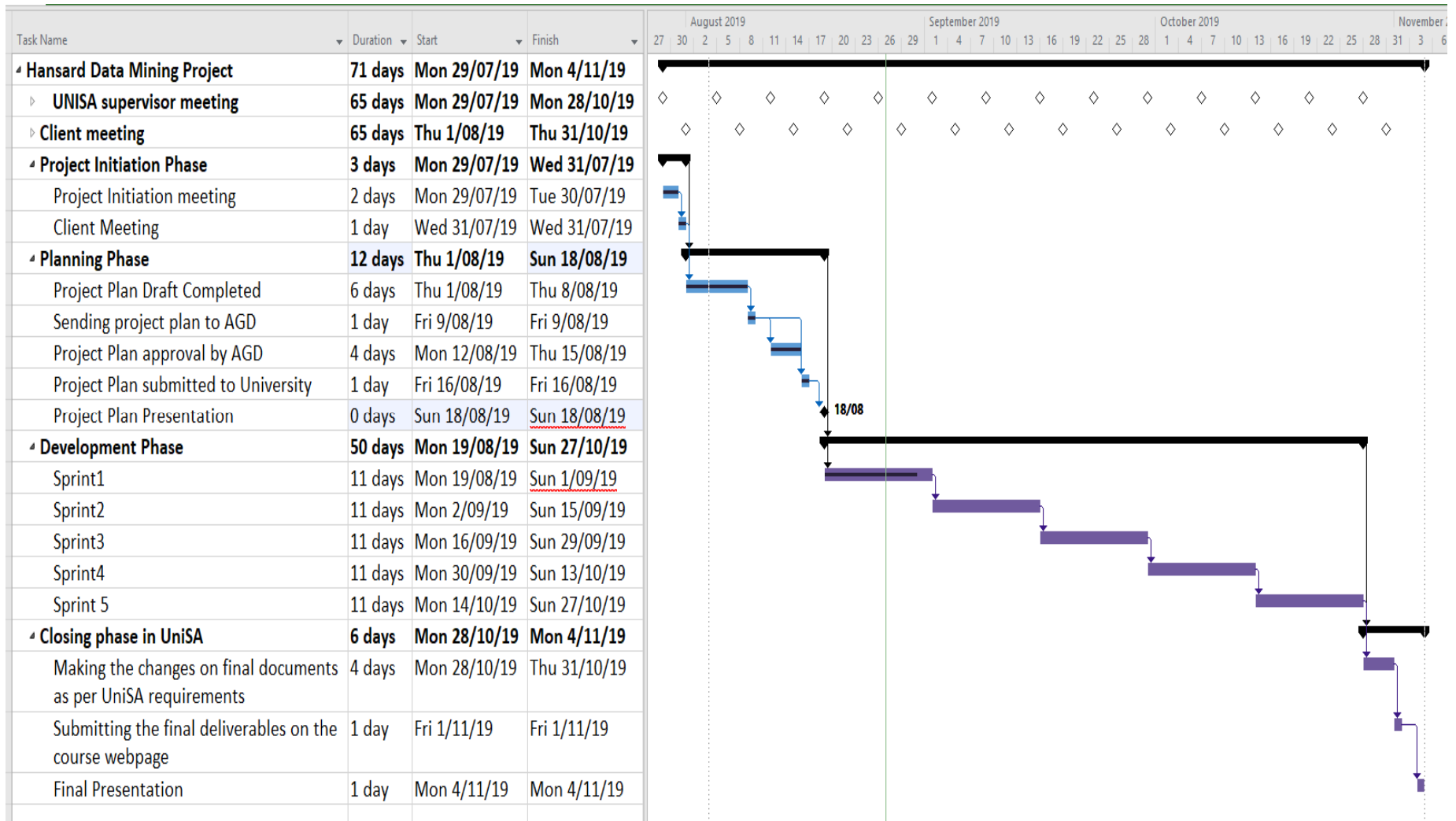

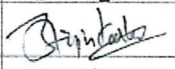



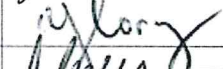




Figure 2: Project Schedule

APPROVAL

We approve the project as described above.

Name	Role (e.g. student, client, mentor)	Signature	Date
Aaron Butler	Student		15/8/19
Bipin Karki	Student		15/8/2019
Katherine Noack	Student		16/08/2019
Mahmoud Yousefi	Student		15/8/19
James Baker	Client		30/08/19
Andrew Corrigan	Client		15/8/2019
Aaron Steicke	Client		15/8/2019
Eric Lam	Mentor		15/8/2019

REFERENCES

300 librarians 2019, *The concept of rapid product development*, Shutterstock, Vector illustration Eps10 file, <<https://www.shutterstock.com/image-vector/concept-rapid-product-development-sprint-diagram-709541407>>

Auditor-General's Department 2019, *Auditor-General's Department*, Viewed 2nd August 2019, <<https://www.audit.sa.gov.au/about-us/the-auditor-general>>

Azure 2019, *Azure Pricing for Virtual Machine*, Viewed 7th August 2019, <<https://azure.microsoft.com/en-gb/pricing/calculator/#virtual-machines69ecc6e4-6817-4661-aae0-e6616c236091>>

Hansard, S. A. I. a. 2008, *Parliament of South Australia*, Viewed 2nd August 2019, <<http://www.parliament.sa.gov.au/Hansard/Pages/GeneralHansardInformation.aspx>>

Elasticsearch 2019, *Elastic Stack Subscriptions*, Viewed 5th August 2019, <<https://www.elastic.co/subscriptions>>

Microsoft 2019, *SQL Server pricing*, Viewed 12th August 2019, <<https://www.microsoft.com/en-au/sql-server/sql-server-2017-pricing>>

National Crime Check 2019, *National Police Check*, Viewed 9th August 2019, <<https://www.nationalcrimecheck.com.au/>>

PayScale 2019, *Average Data Scientist IT Salary in Australia*, Viewed 4th August 2019, <https://www.payscale.com/research/AU/Job=Data_Scientist%2C_IT/Salary>

Scaled Agile 2019, *Scaled Agile Framework*, Viewed 15th August 2019, <<https://www.scaledagileframework.com/>>

Tableau 2019, *Tableau Pricing for Teams and Organisations*, Viewed 5th August 2019, <<https://www.tableau.com/en-au/pricing/teams-orgs>>