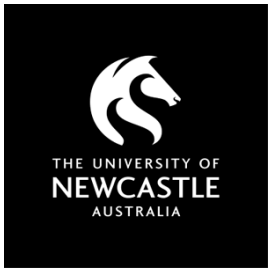
THE UNIVERSITY OF NEWCASTLE

SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING



WORK INTEGRATED LEARNING

COMP3851A – SEMESTER 1, 2019

Interim Report

Authors:

Kepler MANU-LONG

Supervisor:

Kathryn LAMBKIN

June 7, 2019

Contents

Data Engineering Transformation and Business Intelligence Visualisations 3
Background 3
Aims 7
Ultimate Goal 7
Change of Aim 7
Getting Access 7
Report Creation 7
Requirements Gathering 7
Data Quality 7
Report Publishing 8
Methods 8
Project Management 8
Requirements Gathering 8
Getting Access 9
Data Quality 9
Report Creation 10
Stakeholder Liaison 12
Results 13
Data Quality 13
Report Creation 14
Stakeholder Liaison 15
Self-reflection, project management and ethics 16
References 17

Data Engineering Transformation and Business Intelligence Visualisations

Background

This semester, I have been undertaking a Big Data [1] related project at the international company, Komatsu [2]. The branch that I have been working under is Komatsu Mining Corporation's (KMC) Hunter Valley Smart Solutions Team (SST). This team aims to optimise the performance of surface and underground mining, increase the safety of these operations, as well as help reduce overhead costs. SST accomplishes this by "leveraging the Internet of Things (IoT) [3] to rapidly increase onsite insights, powering data-based decision making". [4]

One of the datasets that I am utilising in this project is directly related to the data that is being recorded by sensors installed in trucks that belong to certain mining sites where Komatsu's electric shovels (Figure 1) are installed, specifically the measurements of truck payloads (the dirt in the back of the truck) recorded by weight sensors. Electric shovels are machines that dig up large amounts of dirt and earth, which then dump these materials into a truck via a dipper (a bucket or scooper that holds dirt and releases the contents that are held into trucks) which also has a sensor attached to it to calculate the weight of the amount of dirt that was in the electrical shovel before it was unloaded. These trucks usually have a target payload weight that they want to attain before leaving the shovel's vicinity. This target payload and other statistics are all included in the dataset previously mentioned. (Figure 2) All the data that I'm working with comes from Modular [5] software. Modular is a company that was acquired by Komatsu that operates independently. Therefore, Komatsu needs to liaise with Modular if they want access to their data.

Figure 1: A shovel loading a payload into truck

asset_id	load start time	load end time	truck	target	dip00tons	dip01tons	dip02tons	dip03tons	actual_payload	operator name	bucket count
93	2018-07-16 08:00:22	2018-07-15 22:02:31	T799	320	135	226	346	NULL	359	ADR	3
93	2018-07-16 08:10:44	2018-07-15 22:12:31	T793	320	93	195	284	NULL	290	ADR	3
93	2018-07-16 08:43:33	2018-07-15 22:47:19	T795	320	86	179	275	309	324	ADR	4
93	2018-07-16 09:04:48	2018-07-15 23:06:39	T775	320	106	184	281	NULL	302	ADR	3

Figure 2: Modular Truck Data Dataset. Includes shovel serial no., relevant payload times, truck id, truck target payload, cumulative number of tons after each load, actual number of tons of payload once finished, shovel operator name, and amount of times the truck was loaded (with much more hidden information)

When an operator is operating a machine, they are required to log data about the behaviour of that machine. Any status change that occurs throughout the shift must be manually reported by the operator. This data is recorded in a different dataset to the truck data (Figure 3) and can be used to gauge operational performance of a mining site, e.g. minor incidents and efficiency of a certain machine or all relevant information about down time during a shift.

asset_id	duration(secs)	status	event	operator name	comments
93	1028	Delay	4301 - BOARDING	DAV	TRAVEL TO/FROM PIT
93	140	Delay	4160 - CLEAN UP	WAY	
93	1629	Ready	7000 - OPERATING	WAY	
93	1064	Delay	5100 - SHIFT CHANGE	DAV	
93	6349	Delay	5280 - SAFETY SHUTDOWN	DAN	INCIDENT

Figure 3: Operator Data. Includes shovel serial no., duration of status, status and event, operator name, and event explanations

I am currently working under the supervision of the Business Intelligence analyst at KMC whose main responsibilities include; transforming data into usable formats for reporting, as well as building graphs, dashboards, and reports to answer business questions of stakeholders.[6] What we are doing at SST is the next step of "leveraging the internet of things" after all the data is transmitted to databases. I am currently editing and engineering data to give to end users in a form where it is easy to understand the end results of the large amounts of data that is being recorded during mining operations. Komatsu's large mining equipment are fitted with sensors that record a datapoint every $1/10^{th}$ of a second over 2000 metrics across at least 600 machines (more than 60,000 data points per second).[3] This data doesn't mean anything to a stakeholder if they look at the raw data in a database. Therefore, it's our job to transform this data into easy to read and straightforward to understand reports that will give these stakeholders meaningful insights into the operations that are being undertaken at their mining sites based on the stakeholders' business questions. The datasets that I have been working on have never been viewed before, which means that there are a multitude of potential insights that could be used to increase the efficiency of procedures at the sites where the Modular dispatch software has been implemented.

Power BI [7] (Figure 4) is the business intelligence tool that has been chosen to create the report for this project. Up until now, SST's reporting service of choice has been SQL Server Reporting Services (SSRS) [8]. However, recently there has been a change in SST's business intelligence schema, and Power BI is now under consideration as a contender for their future reporting undertakings. This is because SST

wants to empower their end users to have access to the self-service, easy to use tools that Power BI provides.[9] SSRS also has a more waterfall design style approach when it comes to report building, whereas Power BI has a more dynamic style of report building with its easy to understand graphical components, which suits SST's agile development approach when building reports.[10] For example, when a stakeholder asks a SST employee that they want to see a graph spliced by a certain field, it's easy to drag and drop that field from the sidebar of Power BI and add it as a splicer, where in SSRS, you'd have to edit each data source to splice a graph using that specific field. This leads to greater satisfaction for the stakeholder as they can receive results faster and this also brings about faster iterations of a report's prototype. In these ways, SST is hoping that Power BI will be a viable tool for them moving forward.

Figure 4: An example of a Power BI dashboard in Power BI Mobile

SST uses Apache Impala [11] to query Kudu [12] which communicates with Hadoop Distributed File System (HDFS) [13] where all their data is stored. SST use HDFS as opposed to SQL Server as the time series data that SST receives is an excessive amount for a SQL server to contain and maintain, so they must use a NoSQL type of database.[14] SST stores up to 12 months' worth of data in Kudu, and 3 months' worth of data in OpenTSDB.[15] OpenTSDB is fast to access and has high availability so it's been chosen as the data source that drives the data visualisation dashboard named Grafana.[16] Grafana is another reporting system that SST uses which is incredibly useful to look at datapoints in a time series but cannot splice than more by one set specified field at a time and is also not suitable towards showing aggregated data as legible information. To edit and format the data that can be viewed in Impala, the open-source SQL cloud editor of the same Cloudera product range, Apache Hue [17] is employed. SQL has been the standard for a long period of time for transforming data into a format that's suitable for reporting. SST employs the Microsoft Application Stack [18] and Cloudera Stack [19] (Figure 5), which is why the smart solutions team chooses applications that fit the requirements of their needs which fall under these stacks' range.

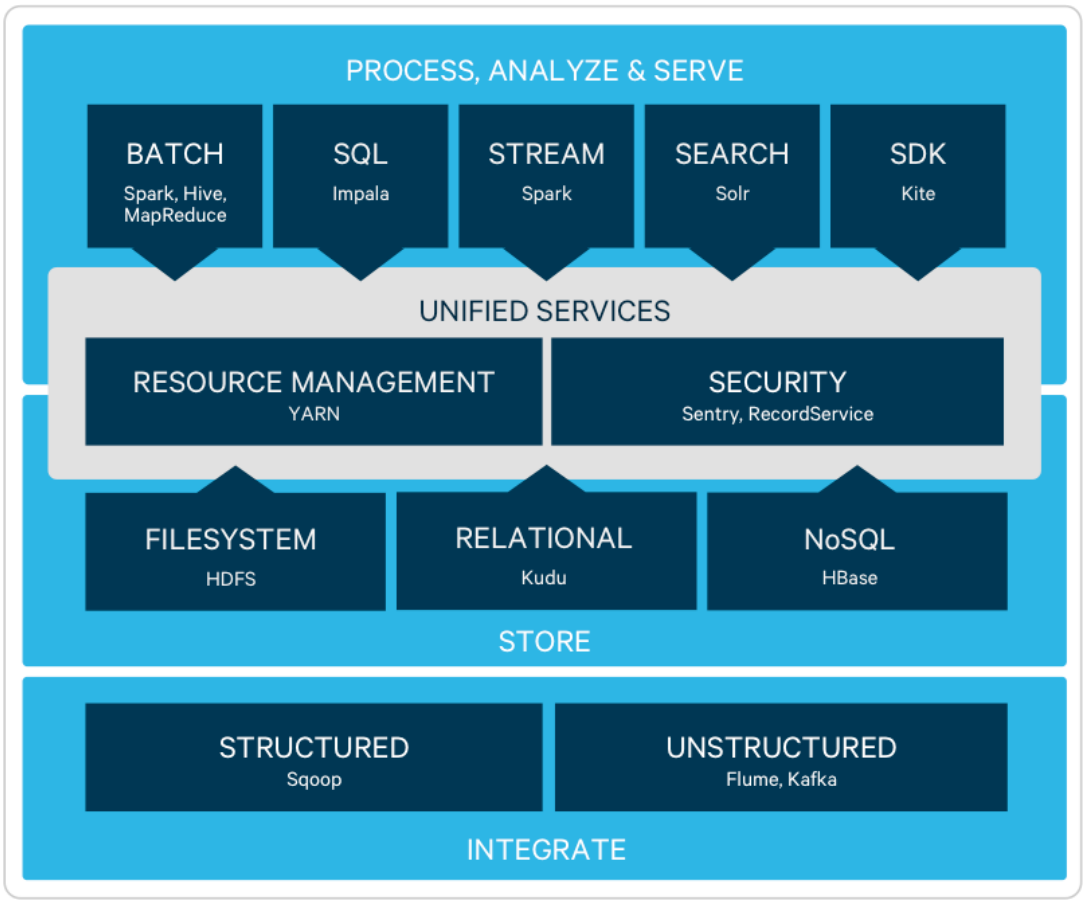


Figure 5: Cloudera Product Stack

Aims

Ultimate Goal

Throughout my time at SST I will be undergoing work experience on what was originally intended to be multiple projects through the form of an internship. The aim and outcome I intend on achieving from this experience is to **deepen and specialise my knowledge in the fields of Big Data and Business Intelligence**. This means to not only to **increase my software skills to a higher level**, but to also **gain sufficient business and communication skills**, as technical skills alone are not enough to satisfy many job requirements.

Change of Aim

Originally, I was planning on undertaking two projects this semester, considering the projects chosen were foreseen to take approximately one month each. However, scope creep of the current project made it unnecessary to move onto the next project. My project supervisor realised that getting access to the data for the KPAR NoComms project that we were initially planning to do first would take longer than doing the Modular Truck Data project first, so I am now currently working on just the Modular Truck Data project.

Getting Access

Upon arrival at SST, it became clear that their data is under tight security and that I would need credentials to access anything on their system to begin working on the project. **Getting access to the company's network and installing the necessary tools and applications** required was essential to start my project.

Report Creation

The main objective of these projects is to **visualise data in an easily understandable and efficient format** with the use of Business Intelligence (BI) and visualisation programs. SSRS was considered as one of tools for this project, but the data visualisation program of choice was changed to Power BI considering that it has been deemed more suitable for SST's needs for the reasons previously stated above (in the Background section),[9][10] provided that its functionality is up to par with SSRS. SSRS has been the go-to product for SST's reporting services up until now, so the functionalities of Power BI are unknown to the Smart Solutions team. Therefore, one of the main aims of these projects is to **assess the compatibility and viability of the use of Power BI in creating mining reports compared to SSRS.**

Requirements Gathering

The idea was to **transform data to meet the requirements and visions of the stakeholders' business questions and demands**, and then afterwards, **create an easy to read report that would provide meaningful insights** for the stakeholder. This stakeholder could potentially be a subject matter expert in mining machines but not necessarily knowledgeable in the fields of business intelligence and programming. Therefore, I had to learn how to **perform accurate requirements gathering so that there is no misunderstanding between myself and the customer** and **clarify my queries and their requirements in a manner that the customer's needs are effectively communicated, captured, and validated**.

Data Quality

Upon further inspection of the data within one of the datasets, it became clear that the data was 'dirty' (incorrect/faulty data). This meant that it would be necessary to **conduct several data quality tests to**

get the dataset to a level of standard where there wouldn't be any dirty data left within the dataset that would heavily skew the information within the diagrams of a report (Figure 6).[20] A common problem in businesses when cleaning datasets is accidentally deleting outliers in place of legitimately faulty records. Therefore, I had **to identify which records were unquestionably inaccurate records and come up with an appropriate way to address them**.

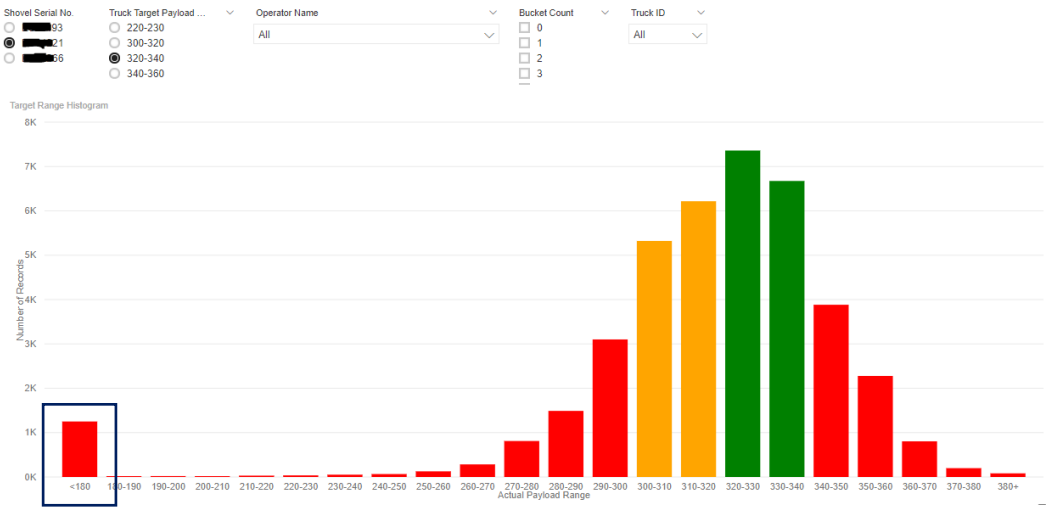


Figure 6: An example of a graph with skewed data due to faulty data

Report Publishing

Once the report is complete, I will **collaborate with senior software engineers to push the report and visualisations created in Power BI into production by following company procedure and getting permissions** by filling out company forms.

Methods

Project Management

SST utilises the Agile Methodology when it comes to developing their reports, code and solutions, and operate on fortnightly sprints. [21] However, I only come in a maximum of 3 days a week and am working on a project independent to the rest of the team, so I don't participate in the Smart Solution Team's sprints. However, I still work in an **iterative style** where, **when I accomplish a milestone, I review the progress of my project with the major stakeholder and refine the requirements of the project further.**

Requirements Gathering

Projects with a goal that seemed attainable within a year were chosen from the Business Development Manager's project backlog. To gain a further understanding of the projects and to gather the requirements needed, we **held a meeting with the Business Development Manager to get a grasp of the what he wanted from us and his vision of the final product**. **I asked him in detail about what each section meant** so there was no misunderstanding between his desired requirements and my understanding of what he wanted me to create.

Getting Access

As when starting at any company, I was unable to access any of the software and data needed to undergo the suggested project. The process of **communicating with my supervisor and other officials to get everything set up** took approximately around a month. Once credentials were assigned to me, I was able to **install Barracuda VPN [22] that is required to access all the data** on Komatsu's network and begin working on my project.

Data Quality

Upon observing that the data within the Modular truck data dataset was often incorrect or faulty, my supervisor suggested I run data quality tests on this dataset so we can flag which data is not reliable and write up a document so that we can report our findings, and ultimately **decide whether creating a report is viable or if there is too much unreliable data to do so**.

When I first got the data, there were more nulls than expected, so I decided to investigate it. When I got full access to the dataset, I did **data quality investigation for records where there were nulls that should have an integer value in its place and searched for values that were too low or too high to be possible.** While using the data I observed multiple other figures that seemed implausible, such as the cumulative number of tons of a payload in a truck decreasing with each load, or trucks driving away empty.

target	loadstart	loadend	dumpend	dip00tons	dip01tons	dip02tons
325	2018-09-06 17:37:34	1536255540	1536256026	113	230	328
326	2018-09-06 19:24:22	1536261921	1536262951	NULL	NULL	NULL
335	2018-09-06 19:53:23	1536263651	1536264815	211	304	NULL
335	2018-09-06 20:19:29	1536265231	1536265714	119	204	266
335	2018-09-06 20:35:49	1536266149	1536266623	NULL	NULL	NULL
323	2018-09-06 20:54:06	1536267251	1536267990	312	NULL	NULL
335	2018-09-06 22:46:23	1536274056	1536274594	94	213	324
323	2018-09-06 23:27:52	1536276547	1536277058	133	226	317
326	2018-09-06 23:31:00	1536276707	1536277389	108	213	291
340	2018-09-07 00:11:25	1536279174	1536280465	NULL	NULL	NULL
355	2018-09-07 01:02:48	1536282281	1536282876	124	216	330
324	2018-09-07 01:11:55	1536282799	1536283323	113	325	NULL
326	2018-09-07 02:57:27	1536289143	1536289874	116	229	320
335	2018-09-07 03:33:17	1536291310	1536292519	115	231	327

Figure 7: Dip00tons should always have an integer value

dip00tons	dip01tons	dip02tons	dip03tons	actual_payload
222	203	295	NULL	315
332	177	264	309	334
336	207	309	NULL	329
134	238	314	240	318
118	19	250	364	356
225	314	125	NULL	317
124	241	342	178	353
271	146	NULL	NULL	249
119	191	42	232	231
97	71	201	NULL	191
118	48	193	NULL	196

Figure 8: Implausible for the cumulative number of dips to go down in number

The **data from the Modular truck data dataset was cross checked against an alternative data source, OpenTSBD**, to see if the Modular data was accurate (Figure 10 & 11). I also **conducted several data quality tests to calculate how many faulty records there were within the dataset**. I **then combined all these exclusions into a single SQL WHERE Clause to filter the data from my SELECT statement** so that when I build a report in Power BI, the graphs wouldn't be skewed by incorrect and false data. [23][24]

value	day	hours	minutes	seconds
114.500000000	24	0	2	51
108.800000000	24	0	3	35
126.700000000	24	0	4	17

Crosscheck

Figure 9: Open TSBD data that should roughly match up with Figure 11(Modular Truck Data dataset)

loadstart	loadend	dumpend	dip00tons	dip01tons	dip02tons
2019-04-24 00:02:44	1556064261	1556064764	127	234	372

Figure 10: Modular Truck data that should roughly match up with Figure 10(OpenTSBD)

Report Creation

Once it seemed like I could create a report with the new clean version of the dataset, I started to begin **working in Power BI to create reports based on the main project stakeholders requirements** as well as business questions that stakeholders have asked and what I think they would ask. Power BI can get access to data from Impala or ODBC (and many other data sources if so desired).[25][26]

The datasets used for this project were not available in the test environment of Impala, so therefore any changes or views that I would have wanted to make to import into Power BI would have to be made in the production environment. However, development in the production environment is considered poor practice and is advised against. When importing data from Impala, Power BI does not allow SQL to be passed to filter the imported data whereas ODBC does. Therefore, **I have chosen to use ODBC for the data source in my Power BI report**.

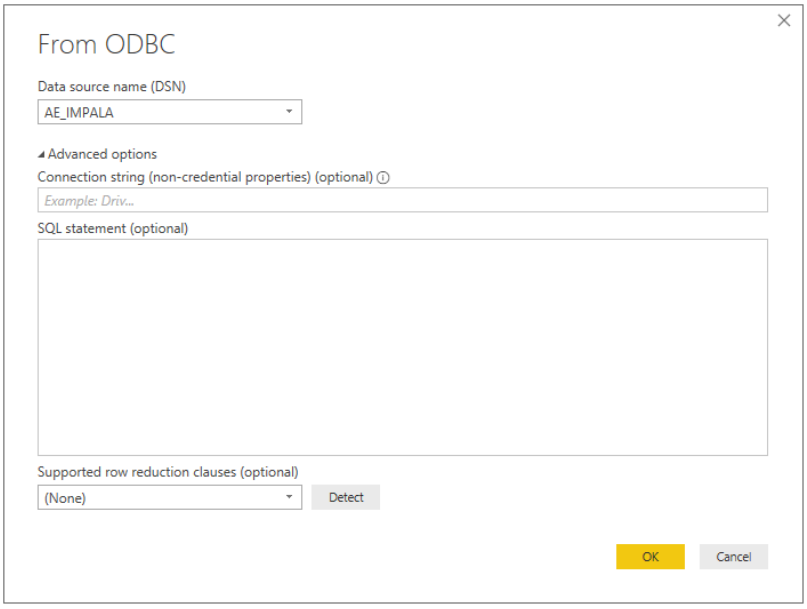


Figure 11: Option to insert SQL statements when importing ODBC data into Power BI

My stakeholder required a histogram, so I needed to research how best to structure the data to drive this. I found there were 2 options, wide and long format data. [28] Different software has different suitable formats of data for each diagram. In Power BI, to make an easy to read histogram graph, through trial and error, I realised that long data is more suited to this need as opposed to wide data, so I **tested how to change the wide data query (Figure 13) to long data (Figure 14) that I created with SQL queries**.

asset_id	<180	180-190	190-200	200-210	210-220	220-230	230-240	240-250	250-260	260-270	270-280	280-290	290-300
21	325	106	106	115	85	76	80	84	121	297	905	1855	4096
22	3443	4358	4358	6368	6346	4315	2148	1336	836	449	465	916	1957
66	186	34	34	60	65	101	92	114	188	293	804	1651	3882
93	182	34	34	49	48	39	47	46	94	275	868	1901	4221

Figure 12: Wide Data - Sum of records within a certain payload range for each shovel

asset_id	bucketrange
93	310-320
93	280-290
93	370-380
93	260-270
93	360-370
93	320-330
93	350-360
93	300-310
93	320-330
93	290-300
93	310-320
93	360-370
93	330-340
93	340-350
93	330-340
93	310-320
93	300-310
93	310-320
93	270-280
93	310-320

Figure13: Long Data - A new column is created, and each record is assigned a certain payload range

In Power BI, there is a limitation where you cannot customise the order of your data and it is automatically ordered either numerically or alphabetically, which was a problem when creating column labels that contain both numbers and text (Figure 15). **I planned to use Data Analysis Expressions (DAX) [29]** for retrieving data from a lookup table and inserting it into a dataset so I could sort by this inserted column to display columns in the correct order. [30] I also realised I needed to add another column to meet the requirements of one of my tables, so I planned on creating DAX measures to derive a new column that changes dynamically depending on the applied filter. [31]

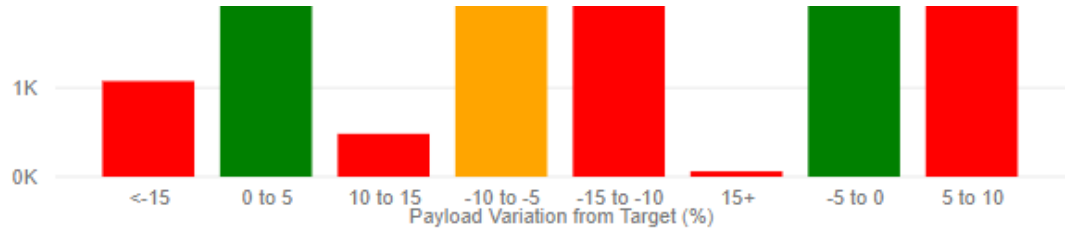


Figure 14: Columns ordered incorrectly, automatically ordered in alphabetical order

Stakeholder Liaison

During this period of cleaning data and creating the report, I had multiple meetings with the main project stakeholder, who when I presented after each iteration of my project, provided me with information about the datasets and suggested new ideas and feedback on the report which I encouraged and welcomed. The stakeholder decided to extend this project rather than having me start work on his alternative projects next semester. Recently, I prepared to demonstrate the current iteration and progress of my project in a meeting **with a wider stakeholder group via a WebEx video conference call [32].** My supervisor and I **presented my project so I could receive feedback on its progress and exhibit the viability of Power BI when building reports**.

Results

Data Quality

I wrote a report articulating the 8 data quality problems that I identified within the dataset and forwarded the document to Modular so that they could solve these issues at the source of the problem.

Section 1: Problems within the data 2
Null entries 2
Repetitive dipNtons entries 3
dipNtons is lower than the previous dipNtons 4
Weight differences between dipNtons too large or too small 5
Too many dips 6
Inaccurate bucket count 7
Actual payload is inaccurate 8
Load start time is occurring after load end time 9
Summary 9

Figure 15: List of faulty record categories from Modular Data Quality document

When I was checking data against Grafana, I noticed that Grafana had errors in its data as well. I gave this feedback to a stakeholder, who indicated it would be caused by broken sensors. He then informed the correct people to fix the sensors. Fixed sensors will lead towards better data quality for this project before completion.

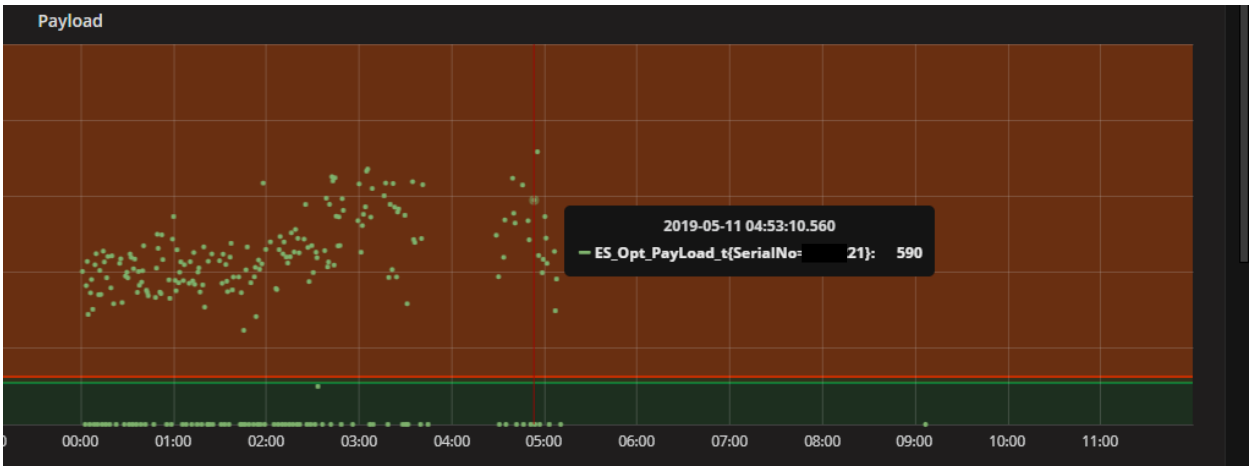


Figure 16: Faulty Grafana data (this integer should never be above 150)

I have built a library of SQL code snippets that can be reused not only on this project, but by other team members on their future work. E.g. I assisted my supervisor in rounding timestamps to the lowest 10-minute period which I then implemented into my own SQL query from this library of SQL code snippets when creating my own report. [33]

Report Creation

I chose ODBC connection which imports the data into the file. Therefore, the file can be distributed to stakeholders without requiring VPN access, any configuration, or even internet access.

I created a histogram and a data table that showed information related to the histogram directly underneath it. I used a long dataset to drive the histogram and a separate wide dataset to drive the data table. There was a limitation where the slicers didn't work as intended (wouldn't filter other slicers). I reworked the solution by combining not only the wide and long datasets, but also the lookup table data that was previously used for the slicers, as I found a constraint of Power BI is that all the slicers work best if all the slicers are within the one dataset.

I used DAX to create new columns with derived metrics inside of them and for arranging data in the desired way. I was able to sort the percentage variance histogram via this use of DAX (Figure 18),[30] as well as insert a percentage of total records column in the dipper averages table based on the selected electric shovel serial number (Figure 19). [34] The format of these DAX codes can be reused for future reference and projects.

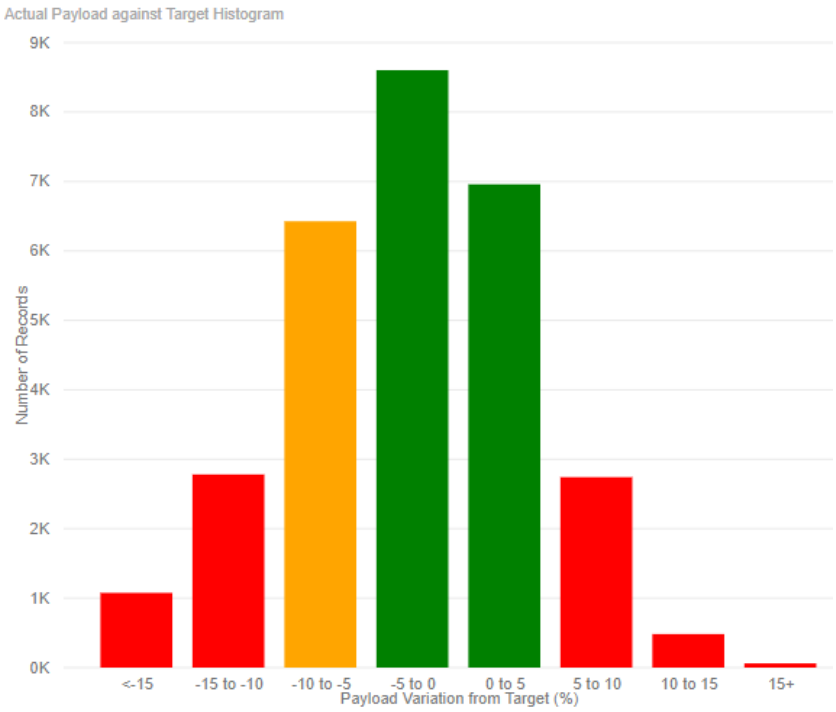


Figure 17: Correct order of figure 15 because of DAX used

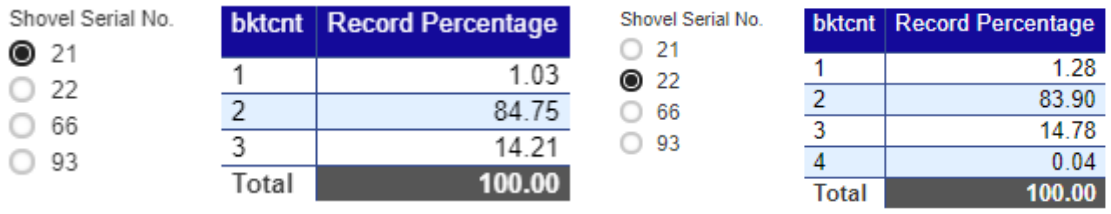


Figure 18: Dynamic Record Percentage column based on selected filter

Stakeholder Liaison

The final meeting with the project stakeholders was held on the 31st of May. This meeting acted as a presentation of the progress of my report, and also as a showcase of what can be accomplished with Power BI. The stakeholders were very satisfied with the results of my project, and they also requested extra features, and I advised them that I will need to investigate the feasibility of these features within the constraints of Power BI.

The stakeholders attending the meeting saw great value in the results and requested additional features. As a result, my project sponsor has decided that I should continue this piece of work for the rest of the year rather than starting a new smaller project. A stakeholder indicated that they can use this information to support further part sales with a significant financial benefit to the company. I was then invited to present my project to a larger stakeholder group because of this success.

Self-reflection, project management and ethics

SST has a wide and well set-up workspace with a relaxed workplace environment and friendly staff. I mainly interact with my supervisor and the main project stakeholder, sometimes emailing other personnel when reporting problems about electric shovels or when in need of administration help. I have a small meeting with my supervisor at the end of each day to show her my progress and get feedback and suggestions, which I proceed to work on the next day I'm at the office. Microsoft Teams [35] and Outlook [36] is used for communication within the Smart Solution's Team. WebEx video conference calls are used for large meetings with many participants. Any task that needs to be officially kept track of are sent via email and meetings are scheduled in Outlook, and for immediate problems and when asking for general advice, communication is undergone via Microsoft Teams.

It wasn't necessary to make a project requirements document and all notes were taken down either in my laptop in a notes word file, or my notebook. Notes were taken frequently throughout each meeting, which always widened the scope of the project, so I always had something to work on. I always made sure to comment my code as a method of documentation, so my code is maintainable and reusable. Any time that I encountered a problem that I was uncapable of doing with my own knowledge, I would search for the solution through Google.

There are no time constraints on this project, provided I finish the project within the time allowed for COMP3851A/B. My supervisor and project sponsor are satisfied, provided I regularly produce deliverables.

Like with any Big Data project there are big ethics considerations that need to be considered. One of the biggest ethics considerations when handling data is the European Union General Data Protection Regulation (GDPR). This regulations states that data related to a subject should not be able to be related back to that subject from merely observing said data without the use of additionally stored information.[37] i.e. Looking at data based on an Employee, just by looking at this data, it should not be able to be traced back to the employee, so names should be either pseudonymized or fully anonymized. The data should at all means not be publicly available without explicit and informed consent from said subject.

All work has been conducted on the computer provided to me by SST on the premises of the company. It is not possible for me to access the database to work on my project from my personal computer. It's advised to not put company file on a personal USB, as the files on the USB can be easily accessed by people not working for SST if lost outside of the company's premises. Data and files are sent to stakeholders via emails. All measures possible are taken so that the company doesn't leak confidential information to the public or to customers. Sensitive data,[38] for example, data that shows an operator's performance, should not ever be publicly revealed to people besides the management of a mining site, but non-sensitive data, such as data about the performance of an electrical shovel, may be viewed by even the employees of that mining site because it doesn't involve information about other personnel.

Considering these ethics considerations, I ensured no names of employees or shovel id numbers were fully displayed in my presentation and this report by shortening every name down to the first three letters, and also use the last two digits of electrical shovel serial numbers so they can't be identified and traced back to their allocated mining site.

References

[1] Oracle.com. (2019). *What is Big Data?*. [online] Available at: https://www.oracle.com/au/big-data/guide/what-is-big-data.html [Accessed 4 Jun. 2019].

[2] Komatsu.com.au. (2019). *Komatsu Australia*. [online] Available at: https://www.komatsu.com.au/ [Accessed 4 Jun. 2019].

[3] Minerals, W. (2019). *How Industrial Internet of Things is changing the mining industry*. [online] MINING.com. Available at: http://www.mining.com/web/industrial-internet-things-changing-mining-industry/ [Accessed 4 Jun. 2019].

[4] Mining.komatsu. (2019). *Smart Solutions*. [online] Available at: https://mining.komatsu/docs/default-source/non-product-documents/services/joysmart-solutions/smart-solutions-brochure.pdf?sfvrsn=4aa50c6b_65 [Accessed 5 Apr. 2019].

[5] Modular Mining. (2019). *Modular Mining*. [online] Available at: https://www.modularmining.com/ [Accessed 4 Jun. 2019].

[6] Snagajob.com. (2019). *Business Intelligence Analyst Job Description*. [online] Available at: https://www.snagajob.com/job-descriptions/business-intelligence-analyst/ [Accessed 4 Jun. 2019]. [7] https://powerbi.microsoft.com/en-us/

[8] Sparkman, M., Guyer, C., Schmidtke, R., Saxton, A., Hassler, M., Kumar, S., Milener, G., Ghanayem, M. and Howell, J. (2019). *What is SQL Server Reporting Services (SSRS)?*. [online] Docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/sql/reporting-services/create-deploy-and-manage-mobile-and-paginated-reports?view=sql-server-2017 [Accessed 4 Jun. 2019].

[9] Ulag, A. (2018). *Power BI expands self-service prep for big data, unifies modern and enterprise BI*. [online] Powerbi.microsoft.com. Available at: https://powerbi.microsoft.com/en-us/blog/power-bi-expands-self-service-prep-for-big-data-unifies-modern-and-enterprise-bi/ [Accessed 4 Jun. 2019].

[10] Team, D. (2018). *SSRS Vs Power BI - 11 Major Difference Between Power BI Vs SSRS - DataFlair*. [online] DataFlair. Available at: https://data-flair.training/blogs/ssrs-vs-power-bi/ [Accessed 4 Jun. 2019].

[11] Impala.apache.org. (n.d.). *Impala - Overview*. [online] Available at: https://impala.apache.org/overview.html [Accessed 4 Jun. 2019].

[12] Kudu.apache.org. (2016). *Apache Kudu - Overview*. [online] Available at: https://kudu.apache.org/overview.html [Accessed 5 Jun. 2019].

[13] Borthakur, D. (2018). *HDFS Architecture Guide*. [online] Hadoop.apache.org. Available at: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction [Accessed 4 Jun. 2019].

[14] Kim, D. (2015). *NoSQL and Hadoop for Time Series Data*. [online] Mapr.com. Available at: https://mapr.com/blog/nosql-and-hadoop-time-series-data/ [Accessed 5 Jun. 2019].

[15] Opentsdb.net. (n.d.). *How does OpenTSDB work?*. [online] Available at: http://opentsdb.net/overview.html [Accessed 4 Jun. 2019].

[16] Grafana Labs. (2019). *Grafana*. [online] Available at: https://grafana.com/ [Accessed 4 Jun. 2019].

[17] Gethue.com. (2019). *Hue*. [online] Available at: http://gethue.com/ [Accessed 4 Jun. 2019].

[18] Acharyya, P. (2015). *What is the Microsoft technology stack all about?*. [online] www.quora.com. Available at: https://www.quora.com/What-is-the-Microsoft-technology-stack-all-about [Accessed 5 Jun. 2019].

[19] Cloudera.com. (2019). *CDH Overview*. [online] Available at: https://www.cloudera.com/documentation/enterprise/5-9-x/topics/cdh_intro.html [Accessed 4 Jun. 2019]. [20] https://www.marklogic.com/blog/the-staggering-impact-of-dirty-data/

[21] Linchpin SEO Team (2019). *A Beginners Guide To The Agile Method & Scrums*. [online] Linchpin SEO. Available at: https://linchpinseo.com/the-agile-method/ [Accessed 4 Jun. 2019].

[22] Barraguard.com.au. (n.d.). *Barracuda SSL VPN*. [online] Available at: https://www.barraguard.com.au/Network-SSL-VPN.asp [Accessed 4 Jun. 2019].

[23] W3schools.com. (n.d.). *SQL WHERE Clause*. [online] Available at: https://www.w3schools.com/sql/sql_where.asp [Accessed 4 Jun. 2019].

[24] W3schools.com. (n.d.). *SQL SELECT Clause*. [online] Available at: https://www.w3schools.com/sql/sql_select.asp [Accessed 4 Jun. 2019].

[25] Saxton, A., Iseminger, D., Blythe, M., Agiewich, R., Peterson, T., Hu, J., Sparkman, M. and Brown, S. (2019). *Connect to an Impala database in Power BI Desktop*. [online] Docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/power-bi/desktop-connect-impala [Accessed 4 Jun. 2019].

[26] Iseminger, D., Harvey, B., Schonning, N., Blythe, M., Sparkman, M., Saxton, A., Petersen, T., Hu, J. and Sebolt, M. (2019). *Connect to data using generic interfaces in Power BI Desktop*. [online] Docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/power-bi/desktop-connect-using-generic-interfaces#data-sources-accessible-through-odbc [Accessed 4 Jun. 2019].

[27] DS, B. and W, B. (2019). *How to use SQL with Power BI in Direct Query Mode for Impala data sources?*. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/47817530/how-to-use-sql-with-power-bi-in-direct-query-mode-for-impala-data-sources [Accessed 4 Jun. 2019].

[28] Grace-Martin, K. (n.d.). *The Wide and Long Data Format for Repeated Measures Data*. [online] The Analysis Factor. Available at: https://www.theanalysisfactor.com/wide-and-long-data/ [Accessed 4 Jun. 2019].

[29] Iseminger, D., Petersen, T., Blythe, M., Duncan, O., Saxton, A., Sebolt, M., Hu, J., Schonning, N. and Koudelka, M. (2019). *DAX basics in Power BI Desktop*. [online] Docs.microsoft.com. Available at: https://docs.microsoft.com/en-us/power-bi/desktop-quickstart-learn-dax-basics [Accessed 4 Jun. 2019].

[30] Community.powerbi.com. (2018). *change column based on filter*. [online] Available at: https://community.powerbi.com/t5/Desktop/change-column-based-on-filter/m-p/498138#M232332 [Accessed 4 Jun. 2019].

[31] Blake, C. (2018). *How to Reorder the Legend in Power BI*. [online] Seer Interactive. Available at: https://www.seerinteractive.com/blog/reorder-powerbi-legend/ [Accessed 4 Jun. 2019].

[32] Webex. (2019). *Cisco Webex*. [online] Available at: https://www.webex.com/ [Accessed 4 Jun. 2019].

[33] Lang, P. (2010). *Round date to 10 minutes interval*. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/2192424/round-date-to-10-minutes-interval [Accessed 4 Jun. 2019].

[34] Kasper On BI. (2015). *Dynamic format using DAX*. [online] Available at: http://www.kasperonbi.com/dynamic-format-using-dax [Accessed 5 Jun. 2019].

[35] Products.office.com. (2019). *Microsoft Teams*. [online] Available at: https://products.office.com/en-us/microsoft-teams/group-chat-software [Accessed 4 Jun. 2019].

[36] Outlook.live.com. (2019). *Microsoft free personal email*. [online] Available at: https://outlook.live.com/owa/ [Accessed 4 Jun. 2019].

[37] Oaic.gov.au. (2019). *Australian businesses and the EU General Data Protection Regulation*. [online] Available at: https://www.oaic.gov.au/resources/agencies-andorganisations/business-resources/privacy-business-resource-21-australianbusinesses-and-the-eu-general-data-protection-regulation.pdf [Accessed 5 Apr. 2019].

[38] Alrc.gov.au. (n.d.). *Sensitive information | ALRC*. [online] Available at: https://www.alrc.gov.au/publications/6.%20The%20Privacy%20Act%3A%20Some%20Important%20Definitions/sensitive-information [Accessed 4 Jun. 2019].