FIT5145 ASSESSMENT 2: BUSINESS AND DATA CASE STUDY

Nhung Seidensticker

Monash ID: 29395968

INTRODUCTION

Organisation

Public Transport Victoria (PTV) is a statutory authority that "*aims to improve public transport in Victoria by ensuring better coordination between mode, facilitating expansions to the network, auditing public transport assets, promoting public transport as an alternative to the car*" (PTV, 2017).

The transport network comprises of train, tram, bus and coach services, spanning across metropolitan and regional Victoria. The public transport ticketing system (PTTS) used by PTV is myki. The benefits of having a PTTS is to provide an interface for customers and the usage data collected can be used to assist in strategic planning.

Project

PTV collects data on commuters from the myki cards. Surrounding touch on and touch off by location for their transports networks. The myki data is not currently being utilised to inform business objectives. PTV would like to use the data to derive insights into their projects. PTV has provided sample data that is temporal and spatial for touch on and touch off points in the form of text (*.txt) files.

Aim

The project aims to provide PTV with insights on how data collected through their myki systems can be utilised to achieve the organisational goals of:

- making sure there's better coordination between trains, trams, and buses
- facilitating network improvements
- auditing public transport assets
- promoting public transport as an alternative to the car

Purpose

The purpose of the project is to set a strategic direction on how the data is managed and utilised throughout the organisation.

The strategic direction will consider the business objectives, curation and management of data, the business model, data landscape, data characteristics, analysis methods, and the potential resources needed.

ORGANISATIONAL STRUCTURE

Within the Department of Economic Development, Jobs, Transport, and Resources, PTV sits under Transport of Victoria. PTV's organisational structure as at June 2017, consists of 6 divisions:

1. Customer Service
2. Network Service Delivery
3. Network Integrity and Project Assurance
4. Franchise Operator Management
5. Communications
6. Corporate Services

The myki ticketing system is outsourced to the company NTT Data (PTV, 2016). The myki ticketing system has over 12 million active cards with more than 600,00 myki cards used daily (PTV, 2016).

BUSINESS MODEL

Using Schroder (2016) big data business model typology set out in Table 1, PTV's business model type is **data user**. The business model is classified as such because PTV has expressed a business need to use the data to make informed business decisions with the outcome of well-planned transport networks and improved services to ensure that Victorian's have access to reliable public transport.

This business model can be further broken down into "*Informing business decisions*" and "*Data analytics as a service.*" **Informing business decisions** is a business that uses data internally to inform strategic decisions and refine business processes, whereas, **data analytics as a service**, is about creating action items based on the analyse of the data (Schroeder, 2016).

Table 1. Big data business model typology

TYPE	EXAMPLE FUNCTIONS	DEPENDENCIES
DATA USERS	Using data to inform strategic decisions; building data into products	Depend on suppliers for raw data, and on facilitators for infrastructure and skills
DATA SUPPLIERS	Gathering primary data; aggregating and packaging data for sale	Depend on facilitators for infrastructure and skills, and on users both as customers and as sources of data
DATA FACILITATORS	Supplying infrastructure; consultancy; outsourced analysis	Depend on users and on suppliers as customers

Note. Reprinted [adapted] from "Big data business models: Challenges and opportunities," by Schroeder, R., 2016, Cogent Social Sciences, 2, 9.

DATA LANDSCAPE

Identifying the data landscape is essential in the strategic planning to adequately identify the potential data science roles required and other inputs from the organisation. The data landscape proposed for this PTV project is illustrated in Figure 1.

Figure 1. Data landscape adapted from Turck's Big Data Landscape and Crowdflower ecosystem.

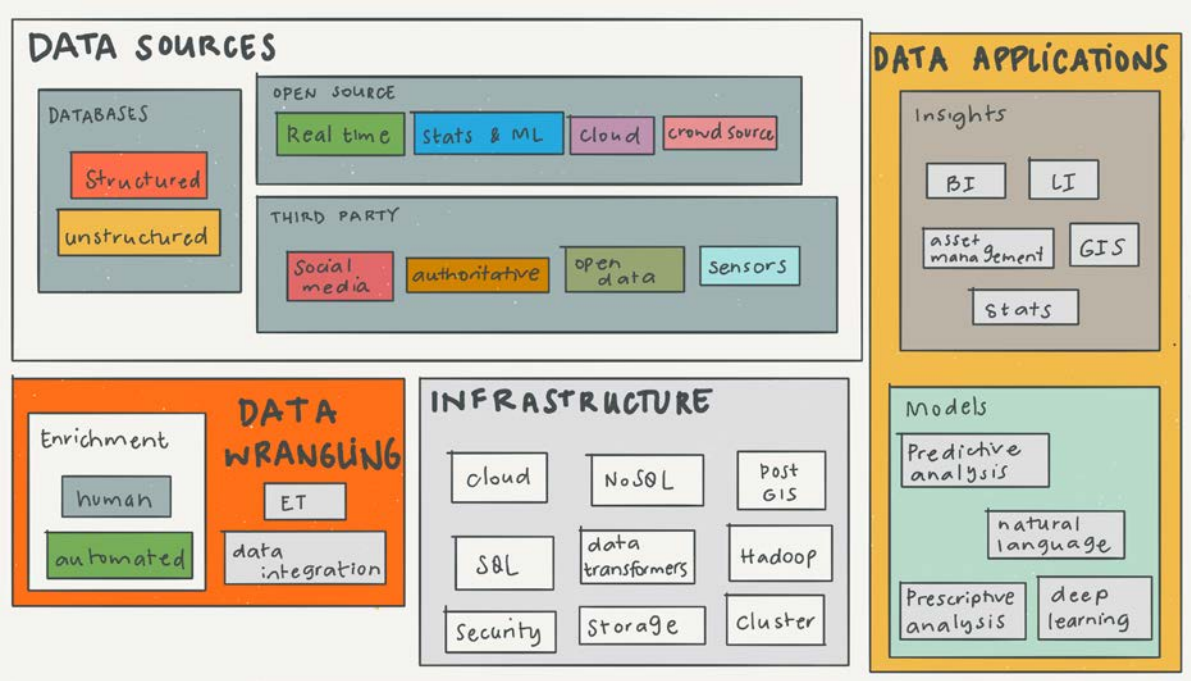

Figure 1. Data landscape for PTV project. Adapted [reprinted] from "Data models and curation in organisations (online)," by Li, M., 2017.

Data science roles

There are many different roles defined within the data science industry. For this project, the roles defined by Theuwissen (2015) will be used. The **data architect** is responsible for creating the "blueprint" for data management systems that ensure data sources are centralised, integrated, protected, and maintained. The **data engineer** will take these blueprints and create an infrastructure that possesses scalable processing systems that includes but not limited to: data warehousing solutions, database systems (SQL & No SQL), data APIs, data modeling tools, and extract, transform, load (ETL) tools. The **business analyst** (BA), being an expert in business processes will link data insights into actional business insights, bridging the gap between **IT** and the business. The **database administrator** (DA) role enables the business users readily have access to data. The DA makes sure that the data is properly secured and backed up as well as having recovery systems in place and are abreast of the different and new technology needed. The **data statistician** models the data for predictive analysis and deep learning. The **data analyst** will work with the statistician and will integrate the various datasets to derive insights and display the information in data applications like Power BI or a GIS system. Overseeing the project and ensuring the success is the **data and analytics manager** that manages the team and setting the strategic direction and prioritising projects and tasks.

DATA CURATION

Data curation is the process of turning data sources (structure and semi-structured) into unified data sets that are ready to be analysed (Freitas & Curry, 2016). The purpose of setting data curation is to turn raw data into information, that information is advanced by adding more content to provide knowledge that is extracted to improve business insights, a visual interpretation of this in illustrated in Figure 2 (Khine & Shun, 2017).

At a high level, data curation for the PTV project is to:

1. Identify the various data sources (internal and external) that are needed.
2. Verify the data to understand the composition.
3. Wrangling the data ensuring that there is no duplication.
4. Transform so that it can be stored and access readily in the database systems.
5. Integrate the data with other sources to provide deeper insights.
6. Export the data to business intelligence (BI) or location intelligence (LI) for users to gain insights.

Figure 2. Turning PTV data into insights.

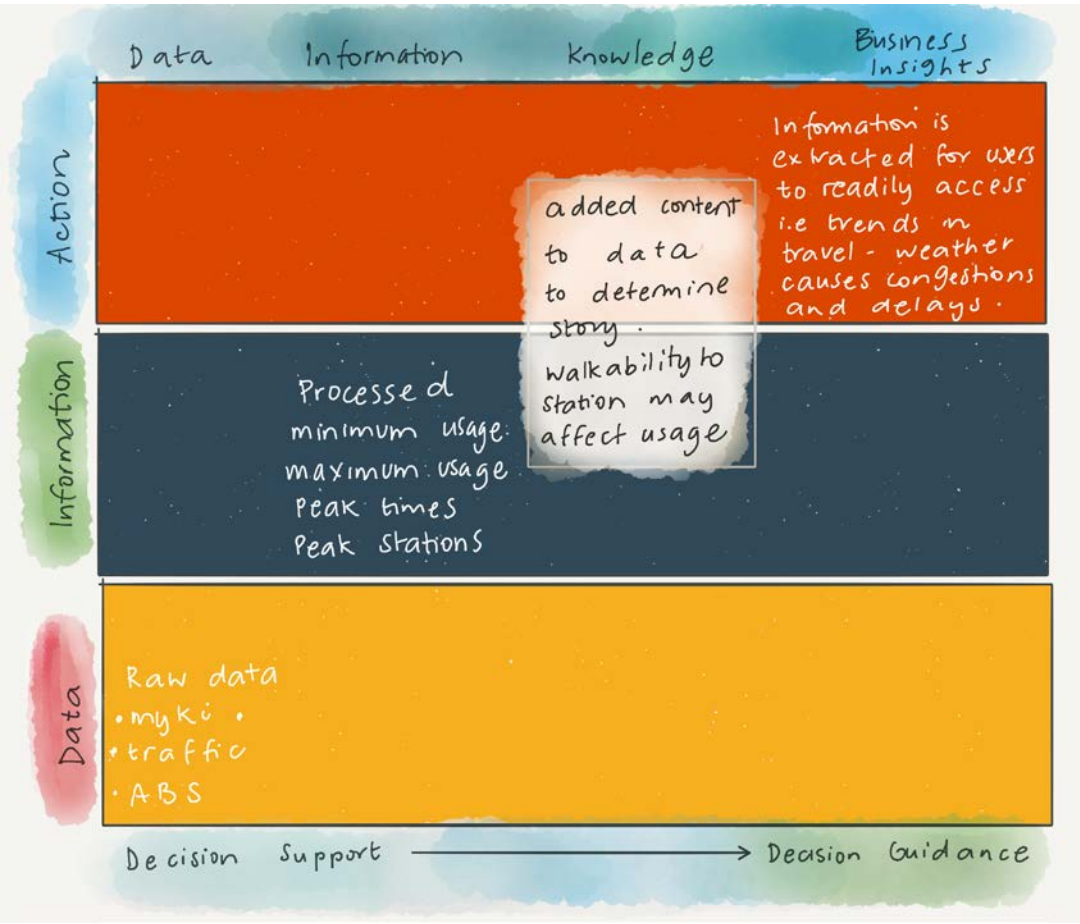

Figure 2. Turning data into insights. Visual interpretation of Khine & Shun (2017) Hierarchy of data.

POLICIES

Before starting the project is governed by PTVs data policy. This policy provides sets broad, high-level principles in which the project will operate (ALGA & ANZLIC, 2005).

The policy ensures that the correct data is acquired and is suitable and fit for purpose, the custodian is identified to determine ownership and assign accountabilities, data security is adhered to, and the use of the data is monitored to establish user patterns and the archival of data not being used.

VALUE

Utilising myki data with other data sets such as, but not limited to, traffic, demographics, social media, open data sources to extract business knowledge and insights adding value and insights to various projects.

Table 2. Summary of the operations that PTV carried out in 2016/17 financial year (PTV, 2017).

HIGH LEVEL OUTCOMES		PROJECTS
EVERY VICTORIAN CAN CONNECT AND PARTICIPATE	1	Extra V/Line services
	2	Better bus networks
	3	Woodend's new flexible bus service
	4	New train and tram network maps
	5	Transporting passengers to special events
	6	Helping passengers plan their journey
	7	Exploring Melbourne by bus
	8	myki improvements
	9	Reducing the number of escalated complaints
VICTORIA IS PROSPEROUS AND SUSTAINABLE	10	Flinders Street Station restoration
	11	New train terminus in Acland Street
	12	New station at Caroline Springs
	13	New Bus depots
	14	Lilydale station's second entrance
	15	Intelligent Transport Systems Congress
	16	Transition to new ticketing system services
	17	Extra trains

	18	Extra E-Class trams ordered
	19	E-Class trams on Route 86
	20	Making trams greener
EVERY VICTORIAN TRAVELS SAFELY	21	Improving level-crossing safety
	22	Try Before You Ride
	23	Continuing to focus on public transport accessibility
	24	Integrated safety and environmental management systems
	25	City Loop safety upgrade
	26	Managing incident responses

Note: Reprinted [adapted] from *Annual Report: 2016/17*, by PTV, 2017, Docklands, Public Transport Development Authority. Copyright 2017 by "Public Transport Victoria".

Having access to real-time commuter data through myki touch on/off locations provides a live picture of the transit network, highlighting blockages and delays. This real-time information is communicated to commuters to allow the user to choose how to get to their destination. This has benefits in reducing complaints and assisting planners in rerouting services and adding extra services. The day-to-day operations are already benefiting from this data set. The focus for the data can move into advanced analytics, combining the real-time commuter data with a wide range of data sources (traffic, demographics, events, social data) to understand future commuter behaviour and plan networks that match and direct movement more efficiently. Comprehensively understanding commuter behaviour will ensure the success of projects outlined in table 2. In addition, data can be used to monitor the success or highlight improvements within these projects.

DATA

The data that PTV would like to utilise and integrate into their business decision making is found in Table 3.

Table 3. Overview of data provided by PTV.

DATA	WHAT	VALUES
SAMPLE 0	1.09GB of Scan off transactions from 2015 – 2018 1.55GB of Scan on transactions from 2015 - 2018	Card type, date (y:m:d) and time (h:m:s), mode, stopID
SAMPLE 1	As above	As above
STOP LOCATION	*.txt of all stop locations with geospatial information	219 metro train stations; 1,679 tram stops, 18,011 bus stops, 89 regional train stations, and 6,590 regional bus stops.
CARD TYPE	*.txt of different types of card holders	72 different types of myki passes.
TRAFFIC	*.csv with geospatial information	Mean and standard speed every hr for 24 hrs.

Note: anon (2018). *MelbourneDatathon*. Retrieved from datathon conference.

CHARACTERISTICS OF DATA

In 2001 Gartner analyst Doug Laney introduced the 3Vs concept of data in the MetaGroup research publication. As illustrated in Figure 3, the 3Vs of data coined by Laney (2001) are volume, variety, and velocity. Since 2001, other "Vs" have been used to define the characteristics of data, these are: value (Tole, 2013) (Gadnomi & Haider, 2015) (Koseleva & Ropaite, 2017), veracity (Tole, 2013) (Gadnomi & Haider, 2015), viability (Khine & Shun, 2017), and, variability (Gadnomi & Haider, 2015).

Figure 3. Infographic of the 3Vs of big data.

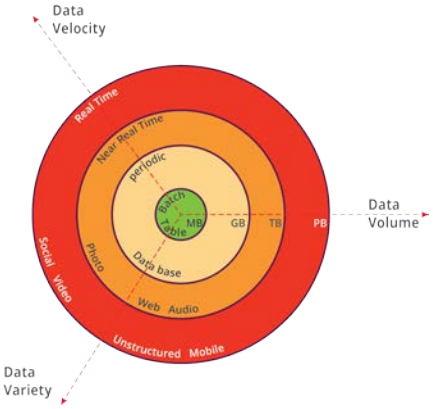

Figure 3. Infographic of the 3Vs. Adapted from "The 3Vs that define Big Data," by Soubra, D. (2012) retrieved from https://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data

The data collected by the myki system falls in the realm of spatio-temporal, meaning the data has geospatial and temporal information. For instance, the data collected has the where (touch on/off location) and when (year, month, day, minutes, secs) the myki card was used. This type of data is commonly referred to as geospatial data.

Geospatial data has always been considered big data (Lee & Kang, 2015). The United Nations Initiative on Global Geospatial Information Management (UN-GGIM), reported an estimation of approximately 2.5 quintillion bytes of data being created every day, with a significant amount having some form of location references (Carpenter & Snell, 2013).

Using the Vs of data, this dataset is large in volume and is collected throughout the day which gives it the characteristics of having velocity and variety. The data could also be characterised by variability as the data is coming from different networks which would lend itself to being characterised as veracity as the data may be unreliable at times.

CHALLENGES AND TECHNOLOGICAL SOLUTIONS

Throughout each step of the data lifecycle (acquisition, processing, and management) there are challenges to overcome, this has been illustrated in Figure 4.

Figure 4. Challenges with data throughout the data lifecycle

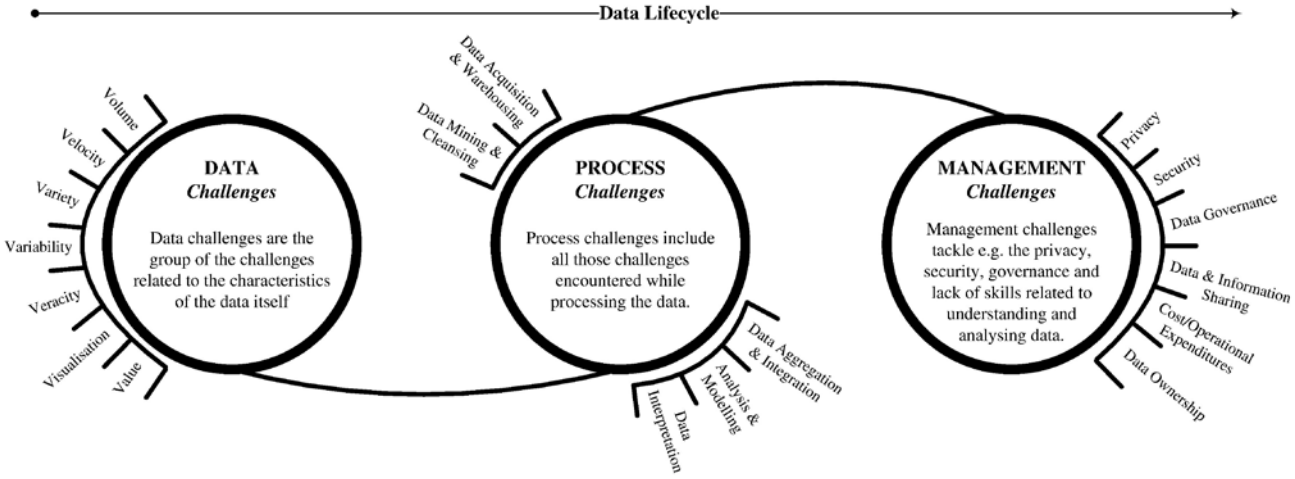

Figure 4. Challenges throughout the data lifecycle. Adapted [reprinted] from "Critical analysis of Big Data Challenges and analytical methods," by Sivarajah et *al.*, 2017, *Journal of Business Research*, *70*, 263-286.

The data supplied by PTV can be unlimited as new data is collected every minute, while historical data is also retrieved for analysis. Compressing this data takes time, and this leads to a delay in analysing the data, a solution is to use the method of Online Analytical Processing (OLAP) (Tole, 2013). As the PTV data is geospatial data, spatial OLAP provides a visual platform for analysing spatio-temporal data quickly with the added value of being able to explore the data in a multi-dimensional way with aggregation levels (Lee & Kang, 2015). A product called **GeoMondrian** is a spatial OLAP system that uses **PostGIS** as the data warehouse. However, it has limitations, and **Spatial Hadoop** has been suggested as a base platform (Lee & Kang, 2015).

The challenge of Big Data **volume** provides challenges when wanting to deal with the data fast (Tole, 2013). Spending time cleaning the data causes a delay in the information utilised. Therefore an automated system to "clean" the data is needed (Tole, 2013). **MapReduce** addresses the issues of cleaning as it provides solutions by "looping," "divide et impera," and "filtering" data and addresses data loss that can occur from a hardware malfunction (Tole, 2013). Other technologies that could handle volume are **Amazon EC2**, **Hive**, **DB2**, **Oracle**, and **MongoDB** (Lee & Kang, 2015).

Lee & Kang (2015) provided an illustrated view of the technologies that could be used to overcome the data challenges outlined by Sivarajah et *al.* (2015) as illustrated in Figure 5.

Figure 5. Example of Geospatial project architecture.

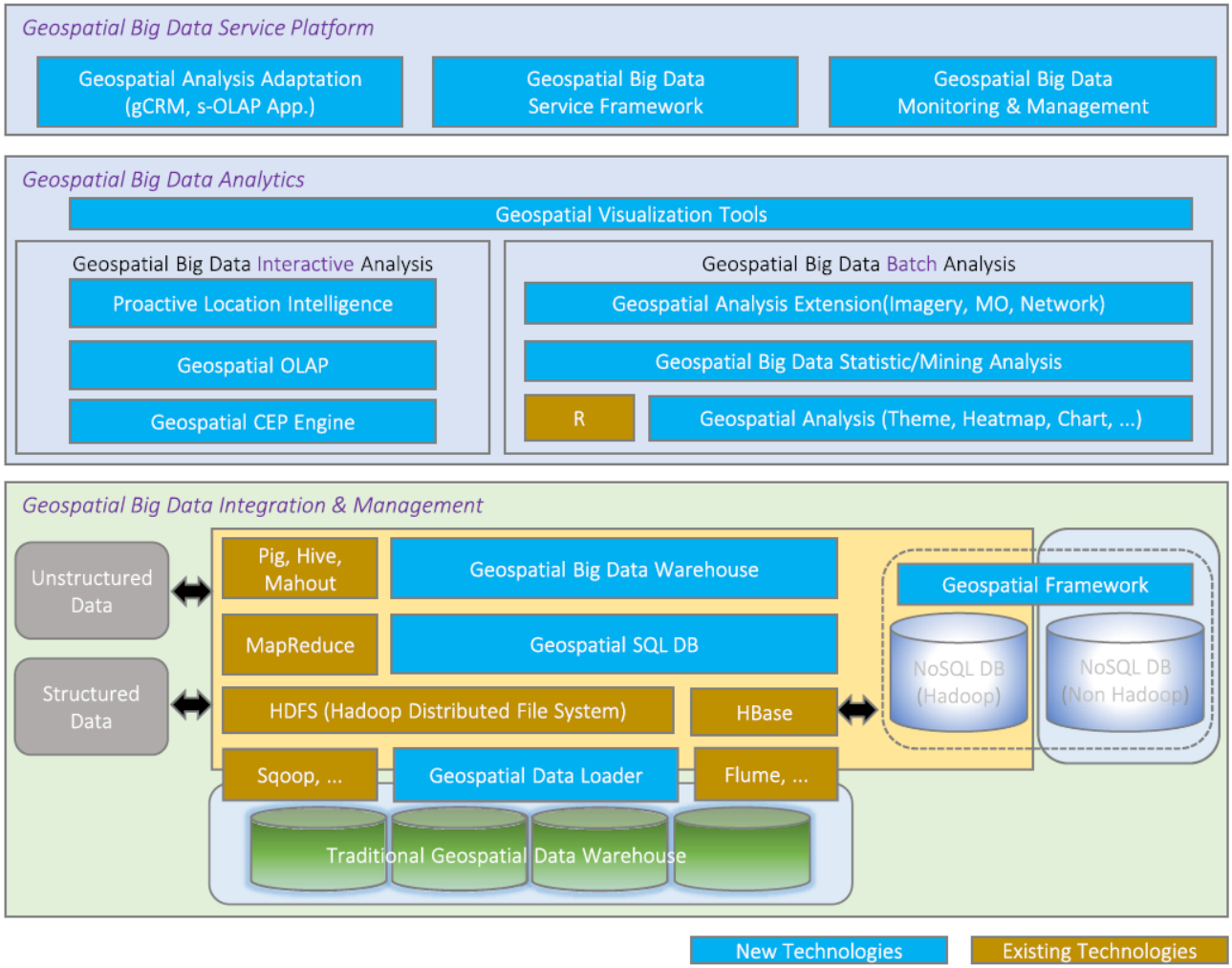

Figure 5. Example of Geospatial project architecture. Adapted [reprinted] from "Geospatial Big Data: Challenges and Opportunities," by Lee, J., & Kang, M., 2015, *Big Data Research*, 2, 74-81.

There are many technological solutions that the role of the data engineer and data architect would explore.

ANALYSIS

As discussed earlier in section Data Curation, data analytics aims to extract as much information as possible to turn data into business insights; this is further illustrated in Figure 6.

Figure 6. Types of data analytics utilised to turn data into information that is actionable.

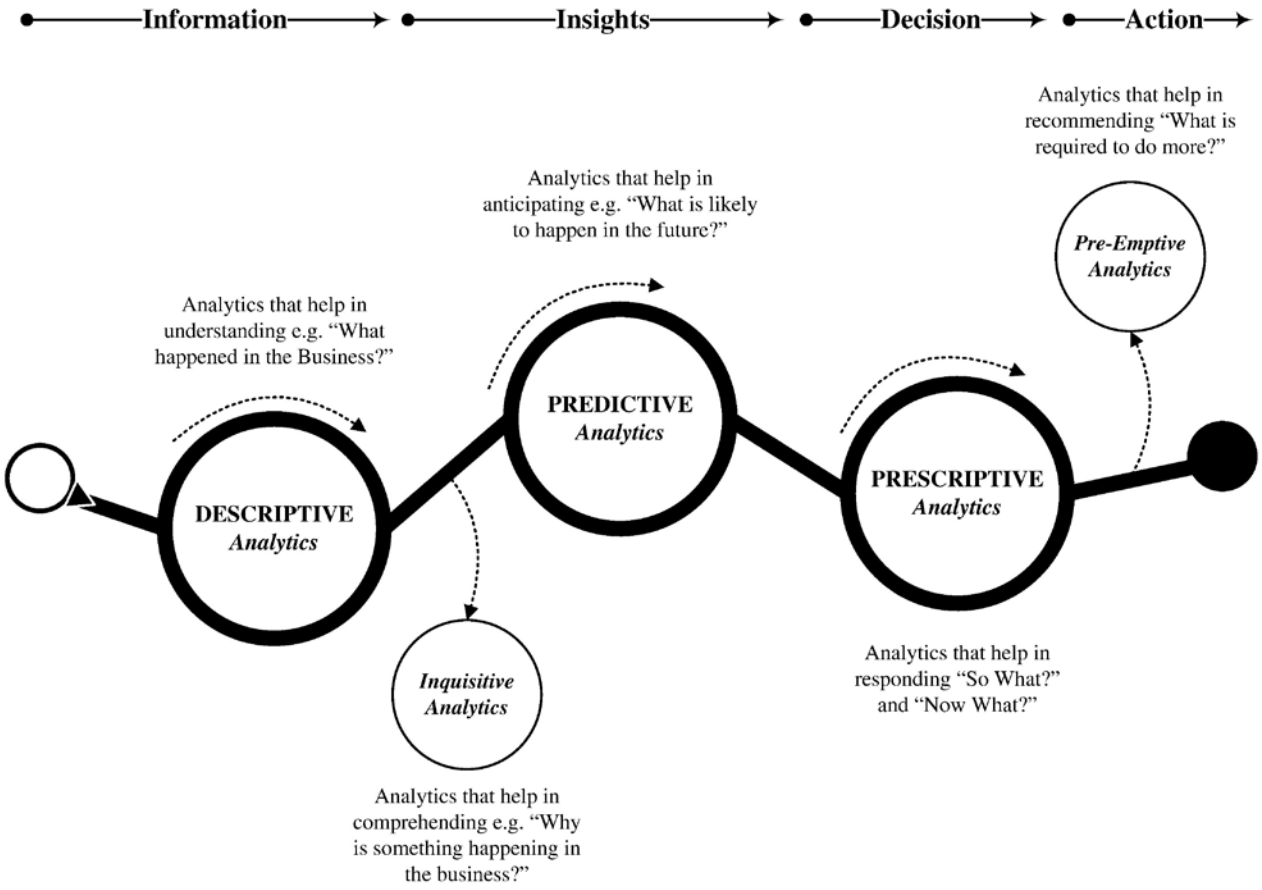

Figure 4. Turning data analytics into action items to meet business goals. Adapted [reprinted] from "Critical analysis of Big Data Challenges and analytical methods," by Sivarajah et *al.*, 2017, *Journal of Business Research*, *70*, 263-286.

The **descriptive analytics** on the data collected about the number of passengers using the various transport services can describe what had occurred; an example is to use regression to find simple trends (Sivarajah et *al.*, 2017). With inquisitive analytics, other data sources such as the Australian Bureau of Statistics (ABS) is integrated with passenger data to understand why some stations frequented more than others. For example, the size of the population and age group, together with looking at walkability to the train station can provide insights into why some stations are utilised more than others. This data can be taken further if the technology discussed earlier is implemented, this dataset can be consumed in real-time leading to the ability of **predictive analytics**. For example, a train is delayed due to a traffic accident; machine learning can predict how long the delays would be and how other transportation could be affected. The information is pushed to the commuter who can decide what to do next. Also, PTV can use the predictions to deploy extra services to assist the number of passengers in getting home. PTV could also predict the demands

in certain areas based on population growth and new development and analyse gaps within their network allowing to better plan new stations or stops. Applying **prescriptive analytics** allows PTV to find optimal solutions given the known constraints. The techniques used in prescriptive analytics are:

1. moving averages: this statistical technique attempts to discover historical patterns in the outcomes and extrapolates it to the future (Gandomi & Haider, 2015).
2. linear regression: this captures interdependencies between outcome and explanatory variables and exploits them to make predictions (Gandomi & Haider, 2015).

With **pre-emptive analytics,** PTV can predict the disruption to commuters due to the removal of a level crossing (Sivarajah et al., 2017). Using pre-emptive analytics, PTV can determine where traffic choke points are and where added bus services are needed to supplement the loss in train services.

RESOURCES

Data science isn't a new field; however, it is constantly evolving and enhancing with new technology being developed (Gandomi & Haider, 2015). Table 4 provides a summary of resources that are useful when embarking on a big data project.

Table 4. Summary of resources to assist in big data project.

TYPE		
RESEARCH ARTICLES	Gartner Journal of Big Data IEEE Journals & Magazine	https://www.gartner.com/en/about
OPEN DATA	Open data	https://www.data.vic.gov.au/ https://search.data.gov.au/ http://www.abs.gov.au/
GIS SOFTWARE	QGIS (open source) Leaflet (open source) Aurin (free for education and government)	https://qgis.org/en/site/ https://leafletjs.com/ https://aurin.org.au/
DATABASES	Cubrid (open source) PostGIS (open source) PostgreSQL (open source)	https://www.cubrid.org/ https://postgis.net/ https://www.postgresql.org/

SOFTWARE LIBRARIES	Hadoop	http://hadoop.apache.org/
APIS	Vicmap	https://www2.delwp.vic.gov.au/maps/maps-and-services/vicmap-api
	VicRoads	http://api.vicroads.vic.gov.au/
	PTV	https://www.ptv.vic.gov.au/about-ptv/data-and-reports/datasets/ptv-timetable-api/
	ABS	http://www.abs.gov.au/websitedbs/D3310114.nsf/home/absstat
DATA STANDARDS	Open data standards	https://www.dta.gov.au/standard/design-guides/open-data/
	ANZLIC	http://anzlic.gov.au/
	IEEE Big Data	https://bigdata.ieee.org/standards
	United Nations Initiative on Global Geospatial Information Management	http://ggim.un.org/
VISUALISATION	Power BI	https://powerbi.microsoft.com/en-us/
	Yellowfin	https://www.yellowfinbi.com/
	SAS	https://www.sas.com/en_au/home.html
	ArcGIS Online	https://www.arcgis.com/home/index.html
	Tableau	https://www.tableau.com/
	Alteryx	https://www.alteryx.com/
ETL	FME	https://www.safe.com/how-it-works/
	Confluent	https://www.confluent.io/
	Alooma (real time ETL)	https://www.alooma.com/solutions/real-time-data-ingestion

Note: Seidensticker, information taken from various resources on the internet

PRESENTATION

Link to presentation:

Total word count: 3279

REFERENCES

Australian Local Government Association (ALGA), ANZLIC (2007), *Local Government Spatial Information Management*, Toolkit v2.0. Retrieved from https://alga.asn.au/?ID=136.

Carpenter, J., Snell, J. (2013). *United Nations Initiative on Global Geospatial Information Management (UN-GGIM): Future trends in geospatial information management: the five to ten year vision*. Retrieved from https://unstats.un.org/unsd/ggim/index.html

Curry, E. (2016). The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. *New Horizons For A Data-Driven Economy*, 29-37. doi: 10.1007/978-3-319-21569-3_3

Freitas, A., & Curry, E. (2016). Big Data Curation. *New Horizons For A Data-Driven Economy*, 87-118. doi: 10.1007/978-3-319-21569-3_6

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 35, 137-144. Retrieved from https://www.sciencedirect.com/science/article/pii/S0268401214001066

Hu, H., Wen, Y., Chua, T-S., & Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, 2, 652-687. doi: 10.1109/access.2014.2332453

Khine, P., & Shun, W. (2017). Big Data for Organizations: A Review. *Journal Of Computer And Communications*, 05(03), 40-48. doi: 10.4236/jcc.2017.53005

Koseleva, N., & Ropaite, G. (2017). Big data in building enery efficiency: understanding of big data and main challenges. *Procedia Engineering, 172*, 544-549. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877705817305702

Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity, and Variety. *Meta Group Res Note 6.6*. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lee, J., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Elsevier*, 2(2), 74-81. Retrieved from https://www.sciencedirect.com/science/article/pii/S2214579615000040.

Li, M. (2017). *FIT5145: Introduction to Data Science (Online)* [Alexandria notes]. Retrieved from https://www.alexandriarepository.org/syllabus/fit5145-introduction-to-data-science-online/116092/

Mousannif, M., Sabah, H., Sayad, Y. O., & Douiji, Y. (2017). *From Big Data to Big Projects: A Step-by-Step Roadmap*. Paper presented at the *Conference: Future Internet of Things and Cloud (FiCloud)*, Barcelona, Spain. Retrieved from https://www.researchgate.net/publication/277657883_From_Big_Data_to_Big_Projects_A_Step-by-Step_Roadmap

Oussous, A., Benjelloun, F., Lahcen, A. A., & Belfkih, S. (2017). Big Data technologies: A survey. *Journal of King Saud University – Computer and Information Sciences, inpress*. Retrieved from http://dx.doi.org/10.1016/jksuci.2017.06.001.

Public Transport Victoria (PTV) (2017). *Annual report 2016-17*. Retrieved from https://www.ptv.vic.gov.au/about-ptv/public-transport-victoria/annual-report/

Public Transport Victoria (PTV) (2016, July 5). *New myki contract awarded* [Press release]. Retrieved from https://www.ptv.vic.gov.au/news-and-events/news/new-myki-contract-awarded/

Schroeder, R. (2016). Big data business models: Challenges and opportunities. *Cogent Social Sciences*, 2(1). doi: 10.1080/23311886.2016.1166924

Sivarajah, U., Kamal, M., Irani, Z., & Weerakkody, V. (2018). Critical analysis of Big Data challenges and analytical methods. *Elsevier*, 70, 263-286. Retrieved from https://www.sciencedirect.com/science/article/pii/S014829631630488X?via%3Dihub

Theuwissen, M. (2015). The different data science roles in the industry. Retrieved from https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html

Tole, A. (2013). Big Data Challenges. *Database Systems Journal*, 4(3), 31-40. Retrieved from http://www.dbjournal.ro/13.html