"Mining the Datasphere: Big Data, Technologies, and Transportation"

CONGESTION MANAGEMENT

Version 3 (January 2017)

Project Steering Group

Dr Ken Michael AO (Chairperson), SBEnrc

Professor Keith Hampson, CEO SBEnrc

Professor Peter Newman, Curtin University

Dr Charlie Hargroves, Curtin University

Professor Bela Stantic, Griffith University

Kamal Weeratunga, MRWA

Kim Thomas, Aurecon

Jannatun Haque, NSW RMS

University Research Team

Program Leader:

Professor Peter Newman, Sustainability Policy Institute, Curtin University

Project Leader/Chief Investigator:

Dr. Karlson 'Charlie' Hargroves, Curtin University Sustainability Policy Institute, Curtin University

Researchers:

Associate Professor Bela Stantic, Institute for Integrated and Intelligent Systems, Griffith University

Daena Ho, Sustainability Policy Institute, Curtin University

Research Assistants:

Daniel Conley, Andrew Hojem, Oliver Pyke, Josh Wood, Rohan Aird, Khiem Nguyen, Georgia Grant, and Harry Carpenter, Sustainability Policy Institute, Curtin University.

CONTENTS

Executive Summary **4**

1 Introduction **7**

1.1 What is 'Big Data'? 7

1.2 Data Superhighways: Big Data in Transport Systems 9

1.2.1 What value can 'Big Data' create? *9*

1.2.2 Collecting Big Data *10*

1.2.3 Data Analysis: 'Big Analytics' *11*

1.2.4 Data Visualisation and Communication *11*

2 Harnessing Data to Better Manage Congestion **12**

2.1 Role of Big Data in Congestion Management 12

2.1.1 Real-Time Congestion Management *12*

2.1.2 Predictive Congestion Management *13*

2.1.3 Public Transport Planning and Deployment *14*

2.2 The Future of Technology Enabled Transport 15

2.3 Privacy Concerns 17

3 Platforms and Technologies for Big Data Analysis and Analytics **19**

3.1 Emerging Digital Platforms for Big Data 19

3.1.1 Overview *19*

3.1.2 Hadoop *20*

3.1.3 Spark *21*

3.1.4 SAP HANA *22*

4 Big Data Case Studies and Applications **24**

4.1 International Examples: Real-Time Congestion Management 24

4.1.1 Zenryoku Annai!, Tokyo *24*

4.1.2 Copenhagen Connecting, Copenhagen *24*

4.1.3 Midtown in Motion, New York *25*

4.1.4 Concert, Seattle *25*

4.1.5 Dublin Road Congestion System, Dublin *26*

4.1.6 Transport Information System, Hong Kong *26*

4.1.7 Traffic Management, Zhejiang *27*

4.1.8 Twitraffic, United Kingdom *27*

4.1.9 Transportation Decision-Making, California *27*

4.2 International Examples – Predictive Congestion Management using Historic Big Data 29

4.2.1 JamBayes, Seattle *29*

4.2.2 Stockholm, Sweden *29*

4.2.3 Intelligent Transport System, Singapore *30*

4.2.4 The Hague, Netherlands *31*

4.3 An Integrated International Example: Transport for London 32

4.3.1 Traffic Management and Transport Planning *32*

4.3.2 Emergency Response and Management *33*

4.4 Australian Examples 34

4.4.1 Congestion Management, Western Australia *34*

4.4.2 Traffic Management and Emergency Response, New South Wales *34*

4.4.3 Sydney Coordinated Adaptive Traffic System (SCATS) *35*

4.4.4 STREAMS, Queensland *36*

4.4.5 Traffic Management and Emergency Response, Queensland *37*

5 Conclusion 38

5.1 Existing Challenges 38

5.2 Next Steps 39

5.2.1 Better Integration *39*

5.2.2 Smarter Technology: Vehicle-to-vehicle (V2V) Systems *39*

5.2.3 Specific Privacy Policies for Big Data *40*

5.2.4 Better Strategy *40*

5.2.5 Big Data Software Selection *40*

EXECUTIVE SUMMARY

Traffic congestion is a key issue facing transport planners and managers with many now asking if there are any promising technologies offering new solutions? In 2015 alone Australia's capital cities were estimated to have a combined congestion cost of $16 billion, expected increase to $37 billion by 2030.[1] With the rapidly growing availability of data and the ability to analyse large data sets this report seeks to answer the question "*What value can 'Big Data' provide to assist with congestion?*". There is great interest and hype around 'Big Data' and this report provides a summary of an assessment of its value to assist in relieving congestion informed by world's best practice. In basic terms, 'Big Data' refers to a very large and rapidly expanding pool of data that is collected from multiple sources and platforms often at high speed and in real time. In recent times the term has become a popular topic for discussions and research on smart urban transportation and mobility. This report seeks to investigate the hype around the term and identify some tangible value for road agencies in Australia that has yet to be captured. It is important to distinguish between 'small data' and 'Big Data', with the difference being that the tools used to assess small data are not sufficient to assess big data and new approaches are needed.

The harnessing of ever expanding streams of urban data has huge potential to inform approaches to alleviate severely congested transport corridors via two main mechanisms:

1. **Real Time Data Monitoring**: This approach allows transport management officials to manage traffic in real-time by harnessing data from a range of sources including those currently available an those anticipated to be available in the near future:

   - *Currently available data*: There are multiple examples of this form of congestion management in Australia and internationally with most using small data from traffic sensors rather than harnessing new data sources such as personal devices, social media, vehicles, weather conditions, etc. The research has found that some cities are beginning to integrate multiple 'small data' streams along with other datasets such as mobile location data to create larger data sets. Tokyo is one of the forefront users of multiple data streams that are starting to look like a 'Big Data' approach to real-time management (See Section 4.1.1). Copenhagen, New York, Seattle and Dublin are also at the leading edge in this field (4.1.2 to 4.1.5). Currently, however, many more cities use specialised, for-purpose 'small data' to manage real-time congestion. Hong Kong and Zhejiang (4.1.6 and 4.1.7) have installed for-purpose sensors to manage congestion and use smaller datasets, but the results and observations that can be made from these cities still have important implications for traffic management in Australian cities. Further, UK's novel 'Twitraffic' application shows that social media can be used effectively as a way to diagnose and identify traffic congestion (4.1.8).

   - *Anticipated near future data:* As will be explored in the upcoming SBEnrc project "*Tech-Enabled Transport: Informing the Transition to Technology Enabled Transportation Vehicles and Infrastructure*", it is anticipated that numerous additional data streams will

[1] Bureau of Infrastructure, Transport and Regional Economics (BITRE) 2015, *Traffic and congestion cost trends for Australian capital cities*, Commonwealth of Australia, Canberra. (ISBN 978-1-925216-99-8)

be available for traffic management in the near future, generated by connected vehicles and civil infrastructure. Vehicles (including private vehicles, public transport, and freight vehicles) are increasingly acting as mobile computers and sensors which can both produce and receive data in order to inform travel options. Imagine your car knowing that another vehicle that you cannot see yet is racing towards your intersection and will very likely run the red light and hit you but your car tells you stop and let them through. This data is termed '*Vehicle-to-Vehicle*' (V2V) and will revolutionise safety and route selection in the coming decades. Likewise infrastructure like motorways, bridges, traffic signals, bus stops, train stations etc. can be sending and receiving data from various vehicles, termed '*Vehicle-to-Infrastructure*' (V2I) which can then allow sophisticated real time transport monitoring and management, while enhancing predictive capacities. Then there is what is called '*Vehicle-to-Everything*' (V2X) that combines V2V and V2I and also communicates with mobile phones in the pockets of people on the bus next to you to tell you how long a bus trip would have taken you, taps into wireless networks to find out occupancy rates at car parks you typically park in (and even how safe they are), and checks the driving history of the person in front of you to see if they are a safe driver. This then begs the question '*What is the best way to collect data by transport agencies to aggregate it for travellers and use for management and prediction?*' which will be part of the focus on the SBEnrc Project 1.52 on 'Tech-Enabled Transport'.

2. **Historic Data Collection and Analysis:** This approach informs preventative measures to prevent congestion through two main mechanisms, namely:
    - *Predictions of future demand*: Historic data can be used to inform predictions of future transport demand based on past commuter behaviour, which is an effective alternative to costly, time-consuming and sometimes inaccurate travel preference surveys. While research into longer-term transport demand patterns is still in its early stages there are computer programs for short-term predictions such as Microsoft's 'JamBayes' program which uses datasets to predict congestion 30 minutes in advance of an unexpected incident, based on current conditions (See Section 4.2.1). Preliminary studies have also been conducted in this area to optimise and predict travel times for the bus fleet in Stockholm (4.2.2).
    - *Simulation of planning options*: Historic data can also be used for simulation of various transport planning options in order to quantify the improvements delivered by suggested upgrades. This can allow city planners to optimise future transport corridors to best meet the needs of commuters. Research in this field is relatively new, but is promising, with Singapore (4.2.3) now using large historical datasets to inform plans for transport corridors and public transport routes. Furthermore, researchers have modelled the impact of changes in a public transport system on the overall transport network in the Netherlands (4.2.4).

These real-time and predictive congestion management techniques can act to mitigate and prevent key bottlenecks in transport systems, allowing transport infrastructure investment to be deferred by opening up the bottlenecks and avoiding the need for additional infrastructure in order to reduce pressure on transport systems. Congestion management will also reduce existing negative impacts of congestion on both the economy and the environment, especially in dense metropolitan areas. For instance, in the US, the cost of congestion in 2012 was estimated to be in the order of $121 billion, the equivalent of $818 per commuter per year, and some additional 25 million tonnes of CO_2 per year.[2]

Transport for London (TfL) provides insight into how truly 'Big' data can be harnessed for congestion management. Using data from multiple sources across the bus network, trains, roads, taxis, ferries and cycle paths, TfL produces a real-time view of traffic conditions and sets variable speed limits across its transport network accordingly. Citizens of London also receive real-time information covering various subjects such as weather, air pollution, delays in public transport, availability of public bikes and real time camera feeds.[3] During an unexpected event, real-time and historical data is analysed to quantify the affected passengers and alter the transport systems to meet their new predicted travel needs. TfL's use of 'Big Data' has allowed better management of higher numbers of passengers: operational performance across 2011-2016 shows improvement in the quality of passenger journeys, including a reduction in wait times and traffic flow, increased customer satisfaction, and an increased number of kilometres serviced by public transport.[4] (See Section 4.3 for more detail)

Essential in the harnessing of Big Data is the use of analytics technology which sifts through the various data streams to identify key information and trends. This report identifies three of the main platforms in Big Data analysis and management: 1) "Hadoop", an open source data collection and analysis platform; 2) "Spark", a Hadoop extension with improved processing performance; and 3) "SAP HANA", a licensed, customisable software proven effective in collecting and analysing Big Data for traffic congestion (3.1). Researchers have also developed novel platforms specific to congestion management (3.5).

With data exploding across a growing array of platforms, and the promise of high-resolution information, many cities across the world are beginning the shift to Big Data. This report details international efforts to provide city planners and transport managers with ideas and avenues to harness Big Data to improve transport systems in Australian cities.

[2] Mullich, J. 2013, 'Drivers avoid traffic jams with Big Data and Analytics', *Bloomberg L.P.,* New York.
[3] Kitchin, R. 2014, 'The real-time city? big data and smart urbanism', *GeoJournal*, 79(1):1–14. ISSN 1572-9893. DOI: 10.1007/s10708-013-9516-8
[4] Transport for London 2015, 'Annual Report and Statement of Accounts', *Greater London.*

1 INTRODUCTION

1.1 What is ‘Big Data’?

There are multiple definitions of Big Data. Most commonly, the term is used to broadly characterise data sets so large they cannot be stored and analysed by traditional data storage and processing methods. The research firm McKinsey describes ‘Big Data’ as a large pool of structured and unstructured data which can be analysed, aggregated, and communicated.[5] A large volume of data is now available for a growing number of sources; however, this is only one dimension of its complexity. The velocity at which data is received and the variety of information available adds to the challenge of creating value. Further, data is now produced in multiple formats, languages and software configurations depending on where the data is sourced. These formats include relational, textual, multimedia and document mark-up languages such as XML.

It is these three characteristics (referred to as the three V’s – Volume, Velocity and Variety) that distinguish ‘Big Data’ from other forms of data. The emergence of such large and complex data sets has primarily been the result of a decrease in the cost of sensory and observational technologies in conjunction with mass digitisation of systems and processes around the globe. Combined with large-scale sensor networks and computer simulations, there is now a vast amount of information stored in a worldwide network of distributed archives. Such changes have allowed for a proliferation of platforms that have enabled the transformation of the analogue world into one made of vast realms of digital information.[6]

Not only can data be used to observe direct phenomenon (such as streamlining traffic signal timings) but the interrogation of data streams to identify commonalities in response to perturbations provides a unique potential to identify linkages between previously seemingly unrelated data. Because of the extremely large volumes of information produced, ‘Big Data’ requires analysis in order to produce meaningful results. The term ‘Big Analytics’ is used to describe the processing of multiple massive data sets to extract useful algorithms and information.

When considering data related to congestion management streams such as traffic counts, average velocity, temperature conditions, traffic light signal durations etc. would be classified as ‘Small Data’ however as the word cloud below presents there are literally hundreds of data sources that stand to inform congestion management efforts.

[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H. ‘Big data: the next frontier for innovation, competition, and productivity’, *McKinsey Global Inst.*, 2011.

[6] Hassanien, A.E., Azar, A.T., Snasael, V., Kacprzyk, J., Abawajy, J.H. (ed) 2015, *Big Data in Complex Systems: Challenges and Opportunities*, Springer International Publishing, Switzerland. (ISBN: 978-3-319-11056-1)

WIDTH OF BIKE LANE
No. of passengers sitting
ROAD INCLINATION
Size of footpath
GPS IDLE TIME
Vehicle acceleration rate
TEMPERATURE
Vehicle break usage
GPS location
TIME TO FIND CARPARK
TRAFFIC FLOW
No. of buses
STANDARD HEAVY VEHICLE ROUTES
No. of bike accidents
Average vehicle velocity
VEHICLE IDLE TIME
SIZE OF TRAN/TRAIN CARRIAGES (PASSENGER CAPACITY)
DRIVER AGE
No. of passengers getting on/off each stop
FREQUENCY OF RAIN
NO. OF CARS THROUGH SCHOOL ZONE
VEHICLE OIL TEMPERATURE
Vehicle size
NO. OF AVAILABLE CARPARK SPACES
FREQUENCY OF ACCIDENTS ON ROAD/INTERSECTION
% OF CARS TRAVELLING AT SPEED LIMIT
Average vehicle waiting time at pedestrian crossing
Average overtaking distance between vehicle and bike
Road wetness
Vehicle age
Distance to carpark
VEHICLE WEIGHT
Average vehicle wait time at an intersection
Frequency of roadworks on road/intersection
Social media event location/size/attendance/time
No. of road potholes per unit distance
HEAVY VEHICLE LENGTH
NO. OF LANE CHANGES PER VEHICLE
Traffic volume
WIFI availability and usage on public transport
NO. OF BREAKDOWNS PER YEAR PER VEHICLE
Phone location
Average vehicle acceleration time to reach speed limit
NO. OF PHONE CALLS
CARPARK EARLY BIRD USAGE
DIRECTION OF TRAFFIC
No. of passengers getting on/off at bus stop
No. of passengers standing
No. of lanes
Average distance before intersection vehicle breaks
HEAVY VEHICLES AS % OF ALL TRAFFIC USAGE ON ROAD
AVERAGE BIKE SPEED
Average time bus is stationary at bus stop
AVERAGE VEHICLE MERGE TIME
Average heavy vehicle spee
% OF CARS TRAVELLING OVER SPEED LIMIT
FREQUENCY OF HEAVY VEHICLES ON ROAD
AVERAGE VEHICLE SPEED THROUGH AN INTERSECTION
CARPARK EARLY BIRD COST
Vehicle value
% OF TIME HAND BREAK IS ON
No. of stops on train/tram route
Carpark cost
No. of tram/train carriages per train/tram
Nature of heavy vehicle g
Carpark size
Average heavy vehicle acceleration
AVERAGE TIME VEHICLE STOPPED BEHIND STATIONARY BUS
FREQUENCY OF BIKE ACCIDENTS
FREQUENCY OF EMERGENCY PHONE
NO. OF PASSENGERS IN VEHICLE
Average wait time to overtake car
GPS speed
Fuel mileage
TOLL ROAD COST
Average bus size
Road ove
% of cars travelling under speed limit
No. of toll roads
No. of bi
ROAD SURFACE TYPE
SOCIAL MEDIA KEY WORDS: ACCIDENT/CONGESTION ETC.
Average wait time to overtake heavy vehicle
NO. OF HEAVY VEHICLES ON ROAD
Intersection size
Heavy vehicle lane usage on multilane roads
Road size
NO. OF CHILDREN BEING PICKED UP/DROPPED OFF
Rain intensity
AVERAGE VALUE OF VEHICLE INSURANCE
AVERAGE TIME OF PASSENGER PICK-UP/DROP-OFF
Frequency of motorbike lane filte
Sun strength
Vehicle type
AVERAGE VEHICLE WAIT TIME TO OVERTAKE BIKES
NO. OF PEDESTRIANS USING CROSSING
Cloud coverage
No. of seats available on train/
VEHICLE USAGE (% TIME)
NO. OF CARS STOPPING AT SCHOOL ZONE
AVERAGE DISTANCE BETWEEN VEHICLES
No. of passengers getting on/off at train/tram stop
Instantaneous vehicle velocity
No. of stopped delivery vehic
No. of road cracks per unit distance
Carpark height restrictions
No. of past tyre changes per road
% tread wear of vehicle tyres
STREETLIGHT COVERAGE
Instantaneous vehicle emissions
SPEED OF CARS THROUGH SCHOOL ZONE
No. of parked car break-ins
NO. OF STOPS ON BUS ROUTES
No. of accidents on road/intersection
% of oxygen in vehicle emissions
No. of double parked cars
Carpark usage
STRENGTH OF SUNLIGHT ON CARPARK
Last service date of vehicle
FREQUENCY OF ROAD UPGRADES
% USAGE OF INDICATORS
BUS WEIGHT

meaning custom hardware does not need to

1.2 Data Superhighways: Big Data in Transport Systems

1.2.1 What value can 'Big Data' create?

The link between 'Big Data' and transportation is not a new phenomenon; traffic systems have long produced streams of observational and sensory information both directly and indirectly. In 2014, a study by the Australian Government Bureau of Infrastructure, Transport, and Regional Economics (BITRE) identified a number of available data collection technologies and concluded: '*Recent and emerging technologies offer significant opportunities for collecting more information, more cost effectively, about personal travel activity and road use, that can better inform day-to-day network management, long-term infrastructure planning and road user travel choices*'.[7]

Smart transport systems using Big Data can achieve a higher level of efficiency, which leads to cost savings, reduced energy demand, better delivery of services, improving life quality and reducing environmental impacts. In this report, transportation refers to systems of mobility including vehicles, roads, railways, subways, buses, taxis, bicycles, ferries and share-rides. Each transportation mode plays an essential role in mobility of a city and if properly harnessed can move people and products to their destination safely and efficiently at a reasonable cost.

One of the key 'push' factors for the integration of Big Data into transportation systems is growing levels of congestion, which puts greater demands on current systems and calls for additional investment in transport infrastructure. Many major cities experience severe congestion and while the cost to business is increasing so too is the impact on the environment, such as exhaust pollution and greenhouse gas emissions from vehicle engines spending additional time in road systems due to congestion. Efforts to reduce congestion through mode shifts to public transport and better management of the road network lead to a number of benefits, such as:

- Enhanced liveability of cities due to less lost time in congestion,
- Faster, cheaper journeys that reduce wear and tear on vehicles and the road network,
- Attracting businesses to cities by providing better and more efficient mobility,
- Reduced environmental impact such as air pollution and greenhouse gas emissions,
- Easing the stress on the city transport budget and maximising the benefit of expensive transportation assets.

According to the Australian Bureau of Transport and Resource Economics, in Australia '*the avoidable social costs of traffic congestion will rise to about $20 billion by 2020'*.[8] In order to create a data system that has a meaningful impact on congestion the system should:

- Cover the whole city using multi-service communications,
- Access integrated accounts through multiple channels,

[7] BITRE 2014, 'New traffic data sources – An overview', *Australian Department of Infrastructure and Regional Development,* Canberra.
[8] Transmax 2015, 'Streams ITS', Transmax Pty Ltd 2015.

- Optimise and integrate all transport modes,
- Employ demand-based and dynamic pricing,
- Analyse and respond to full situational data,
- Optimise relevant operations, and
- Target predictive analysis.

1.2.2 Collecting Big Data

The rapid rise in the capacity of data storage options (both in-house and remotely) along with the increase in computational ability means that there is great potential to harness additional value from data that can inform the working of a modern city, especially its transportation systems. Data from transport systems is highly varied and comes in three broad categories:

i. *Highly structured datasets* that originate from technology implemented to address well-defined problems (e.g. data from automatic toll road payment transponders for the use in processing toll road payments or data from intersection sensors on traffic flows and time of day usage of the road network) which can more aptly be defined as 'Small Data',

ii. *Unstructured datasets* that are produced from any interaction between road users and digital infrastructure. Given the explosion of mobile phones, personal computers, sensors, cameras, and devices, there is huge (yet largely untapped) potential to harness these data streams to inform congestion and disaster response, which is now moving into the realm of 'Big Data', and

iii. *Data from seemingly unrelated sources* that stand to provide insights into the behaviour and functioning of the transport system, such as the price of parking at particular public carparks, the level of fines for illegal parking, the amount of people walking more than 1 kilometre to public transport, weather conditions etc., which now makes the overall data set unmanageable using small data techniques.

Currently access to data is not a concern as there are a multitude of data sources available which produce a wealth of information (however it may be the case that additional data sets that are currently not available may be more valuable to congestion management than those that are currently openly available). The challenge is to harness the data by processing and interpreting it both at the higher levels of trends and scenarios and at the lower levels related to the day to day management of transportation infrastructure. High data volumes mean that it takes time to process and advanced computing technologies are required to improve response times.[9] Currently, data is used to inform trip times and route selection; however, Big Data can be used to inform predictive analyses and the development of advanced user information platforms.

This analysis requires programs and technologies that extract value from what our research team refer to as the 'Datasphere', which contains data that may seemingly be disconnected from transportation but when assessed shows correlations that would otherwise be hidden. It

[9] Mullich, J. 2013, 'Drivers avoid traffic jams with Big Data and Analytics', *Bloomberg L.P.,* New York.

is the combined 'Big Analytics' processing of all available data streams within which the true potential value of Big Data exists.[10] Effectively harnessing such data can provide significant benefits due to the development of temporal, spatial, and historical correlations.[11]

1.2.3 Data Analysis: 'Big Analytics'

Because of the volume and complexity of the data produced, there are inherent difficulties in data analytics and challenges exist in the analysis and harnessing of this information. In particular, the different data formats and languages in which data is stored may lead to difficulties in processing using data mining algorithms.[12] However, the potential rewards are impressive. The availability of 'Big Data' provides insight into actual passenger and road use behaviour, as opposed to reported behaviours and preferences which may not present the whole picture.[13]

The multilayered nature of Big Data also allows data mining programs to find correlations and convergent traveller preferences across multiple platforms such as surveillance cameras, smartphone and metro card (smart card) use, and sensors.[14] This can aid in the development of transport demand projections that take into account multiple modes of transport (private vehicles, public transport, cycling etc.), forming a big picture overview of expected travel patterns.

1.2.4 Data Visualisation and Communication

Communicating value from 'Big Data' to road users, planners and operators is crucial for the improvement of transport networks. To curtail road congestion before the congestion becomes severe, Big Analytics algorithms must be able to communicate with traffic lights and other traffic control systems when real-time congestion pre-cursors match historical information on severe congestion events.[15] Demand projections can inform transport infrastructure investment for planners and ensure that implemented projects are catered specifically to customer demand. In the case of public transport, the provision of real-time traffic conditions and accurate wait time estimations greatly improves customer perceptions of service effectiveness.[16] As such, the benefits of using Big Data must be weighed against the effectiveness and efficiency of Big Analytics technologies, as well as the costs incurred in using these analytic procedures.

[10] International Transport Forum 2015, 'Big Data and Transport', *International Transport Forum.*

[11] International Transport Forum 2015, 'Big Data and Transport', op. cit.

[12] Hassanien, A.E., Azar, A.T., Snasael, V., Kacprzyk, J., Abawajy, J.H. (ed) 2015, op. cit. ISBN: 978-3-319-11056-1

[13] van Oort, N. & Cats, O. 2015, 'Improving public transport decision making, planning and operations by using Big Data: Cases from Sweden and the Netherlands', *IEEE 18th International Conference on Intelligent Transportation Systems*. DOI 10.1109/ITSC.2015.1

[14] Carter, K.B. 2014, Actionable Intelligence: A guide to delivering business results with Big Data fast!, John Wiley & Sons, Singapore. (ISBN: 1118920651)

[15] Sawyers, P. 2015, 'How Microsoft's using big data to predict traffic jams up to an hour in advance, Venturebest, April.

[16] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

2 HARNESSING DATA TO BETTER MANAGE CONGESTION

2.1 Role of Big Data in Congestion Management

Congestion reduction has economic, environmental and health benefits. Firstly, reducing peak-time congestion defers required capital investment: an additional road or highway does not have to be built if peak-time traffic is no longer an influential factor. In addition, road congestion has a financial cost. In the US, the cost of congestion was $121 billion in 2012, which equates to $818 per commuter per year.[17] According to a study by BITRE, Australia's capital cities were estimated to have a combined congestion cost of $16 billion AUD in 2015, with an expected increase to $37 billion AUD by 2030.[18] In particular, Perth's congestion costs are expected to increase the most drastically, tripling from $2 billion in 2015 to $5.7 billion in 2030.[19] Further, reduced vehicle wait times in traffic jams reduces vehicle exhaust, thus reducing carbon emissions and air pollution. In the US alone, 25 million tonnes of CO_2 per year was emitted from vehicles stuck on congested roads.[20] In addition, inhaling vehicle exhaust for extended periods has also been linked to human health problems such as brain-cell damage.[21] These negative externalities all point to the increased need to manage road congestion, and the growing availability of data might provide part of the solution.

The potential to harness Big Data in congestion management is two pronged:

1. *Mitigation of existing traffic jams through real-time data use*: Real-time information from a variety of sources (such as traffic signals, vehicle counters, social media streams, CCTV streams etc) is being used in many cities across the world as a form of congestion management, however this is considered an application of 'small data' (Sections 4.1 and 4.2.1)

2. *Avoidance of traffic jams through predictive strategies*: Big Data allows for going beyond real time data analysis of small data streams to allow predictive algorithms to blend real time data with historical data sets on commuter habits and preferred routes to allow for predictive traffic management (See Section 3.2)

3. *Create sophisticated public transport routings*: Big Data produces high-resolution information which can be used to build public transport demand maps which result in the better allocation of public transport resources. Research in this area is still relatively new, but several authors have conducted studies in this field (4.2.2 to 4.2.4)

The following sections investigate each application in detail.

2.1.1 Real-Time Congestion Management

Many well-established congestion management strategies use real-time data. Multiple types of software exist which respond quickly to real-time changes in traffic volume, traffic movement

[17] Texas A&M Transportation Institute 2013, 'As traffic jams worsen, commuters allow extra time for urgent trips', *Texas A&M University*, February 5.

[18] Bureau of Infrastructure, Transport and Regional Economics (BITRE) 2015, *Traffic and congestion cost trends for Australian capital cities*, Commonwealth of Australia, Canberra. (ISBN 978-1-925216-99-8)

[19] Bureau of Infrastructure, Transport and Regional Economics (BITRE) 2015, *Traffic and congestion cost trends for Australian capital cities*, op. cit.

[20] Mullich, J. 2013, 'Drivers avoid traffic jams with Big Data and Analytics', *Bloomberg L.P.*, New York.

[21] Hotz, R.L. 2011, 'The hidden toll of traffic jams', *The Wall Street Journal*, November 8.

demands, and direction of travel. In Australia, three main types of software are currently used to inform traffic control systems, namely SCATS, STREAMS and InSync.[22] SCATS stands for 'Sydney Co-ordinated Adaptive Traffic System', which monitors real-time traffic signals and vehicle volumes to coordinate adjacent traffic signals to reduce traffic congestion and optimise traffic flow. The use of SCATS has been shown to correspond to a reduction in overall travel times, vehicle stops, fuel consumption and waiting times at red traffic signals.[23] As of November 2011, more than 3,700 traffic lights were connected, monitored and controlled using the SCATS network within New South Wales. STREAMS, a similar type of program, has also been implemented in Queensland, with promising results (4.4.3 and 4.4.4).

InSync is another adaptive traffic control system that uses cameras installed at traffic intersections to detect and manage traffic conditions. InSync differs to SCATS in that it does not use the concept of cycle lengths, splits, and offsets, but rather uses the concept of a finite state machine which consists of all possible states within the intersection. This means that at any given moment, a specific state can be identified which will lead to an appropriate signal transition.

Main Roads Western Australia has recently developed a tool called NetPReS (Network Performance Report System), which integrates data from multiple sources and reports road network performance in terms of multiple indicators. The tool is currently limited to historical performance but is expected to be expanded in to real-time performance analysis.

In terms of data visualisation and reporting, real-time congestion reporting is so commonplace that it has become both ubiquitous and expected. The best example of this is Google Maps' Directions application, which incorporates real-time congestion information in a readily accessible form known to almost all first-world road users.

While real-time congestion mitigation techniques have been implemented extensively across multiple cities and countries, any real-time strategy only has a limited scope to improve traffic conditions. This is primarily because it is already too late to avoid the congestion once it has been observed. Real-time mitigation strategies are often based around deterring additional traffic from moving towards this area through traffic signals or responsive road tolls, but neither strategy eliminates the existing congestion. As such, great interest is now focused on predictive strategies which seek to curtail a traffic jam before it even begins.[24]

2.1.2 Predictive Congestion Management

The question is can 'Big Data' really predict traffic conditions, and the short answer is "yes, but not perfectly, and not right now". The long answer is that while high-resolution data collected from millions of sources contains the necessary information to predict travel patterns and

[22] Fernando, B., Gray, E., Kellner, J. 2013, 'A review of current traffic congestion management in the City of Sydney', *Infrastructure Australia*, Canberra.

[23] Main Roads WA 2014, 'SCATS', *Government of Western Australia,* Perth.

[24] Lu, H.P., Sun, Z.Y., Qu, W.C. 2014, 'Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction', *Discrete Dynamics in Nature and Society*, vol 1. (DOI:10.1155/2015/284906)

identify problem areas, the challenge is to process this information to extract the useful information and correlations to be compared to historical data.[25]

In addition, no prediction is perfect and there is always a margin of error. In particular, traffic conditions are also influenced by factors such as road accidents caused by traffic dynamics and human responses and are hence near impossible to predict. Even a perfectly safe vehicle under perfect geometric and environmental conditions may still crash due to sudden changes in road dynamics, disruptions in the normal flow of traffic, or a disruption inside the vehicle. In order to better improve congestion prediction models, aggregated traffic flow variables (e.g. assuming vehicle speed to be equal to the speed limit) can be replaced with real-time data in order to create a more realistic model.[26]

There are currently a number of 'Big Analytics' traffic prediction systems in development. An early mover in this space is the global company HERE that processes information collected from over 2 billion traffic probes per day and compares it to historical data since 2011 using algorithms to generate predictions of road traffic congestion issues.[27] Microsoft has also developed various software to predict traffic conditions, with some software platforms taking into account unexpected traffic conditions, and have achieved promising results (See Section 4.2.1).[28]

2.1.3 Public Transport Planning and Deployment

Public transport systems are increasingly equipped with automated data collection systems, which can be harnessed along with other data streams to provide insight into passenger demand and identify optimal public transport networks, routes and connections.[29] Analysis of such data can provide information on passenger needs and behaviour, as well as provide an assessment of system performance and real-time conditions. Furthermore, such data analysis can allow road and transport organisations to quantify the costs of service deficiencies. Crucially, quantification also allows an even-handed and simulation based evaluation of possible solutions (eg. timetable synchronization), allowing each solution to be ranked based on its cost-effectiveness and user experience benefit.[30]

Two key Big Data sources are of interest for public transport networks: 1) Automated Vehicle Location (AVL) data, provided by mobile phones, and 2) Automated Passenger Counting (APC) data, provided by smart cards (metro cards), surveillance systems (ie. video cameras), Wi-Fi and Bluetooth trackers, and sensors connected to assets, signals and switches. Currently, both AVL and APC data are used for system performance evaluation. However, neither data source has been used extensively in system planning and development, making AVL and APC data a largely underused resource. If harnessed, these forms of data can inform projections of passenger

[25] Jie Xu et al. 2015, 'Mining the Situation: Spatiotemporal Traffic Prediction with Big Data', *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4

[26] Hossain, M. 2012, 'A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways', *Accident Analysis & Prevention,* vol. 45. (http://dx.doi.org/10.1016/j.aap.2011.08.004)

[27] Highway Engineering Australia 2015, 'Big data: The key to unlocking the future of traffic, transport and infrastructure', *HEA* , vol. 47, no. 2, pp. 40-41.

[28] Horvitz, E. 2011, 'Predictive Analytics for Traffic', *Microsoft Research*, Sept 26,.

[29] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

[30] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

volumes, which are essential in the effective prediction of future demand and can act to enable the design and optimisation of transport networks.[31]

Furthermore, the analysis of data sources such as AVL and APC data can replace large, costly surveys on travel habit and stated preferences. Algorithms can directly construct travel demand based on observed travel patterns and provide a basis for public transport planning such as tactical planning (a mid-term plan that involves service frequency, timetabling, and vehicle and crew scheduling) and strategic planning (a long-term plan and concerned with overall network and service design such as stop positioning and line topology and capacity). [32,33] Not only does using Big Data reduce the costs of transport surveys, it also provides more detailed, high-resolution information such as seasonal effects and within-day and day-to-day demand variations which are essential in timetabling.[34]

There is great interest in ways to mine the topology of a public transport network, and this has led to the development of tools to support this such as the 'Density Consensus Clustering' approach.[35] This approach seeks to deduce static knowledge of a public transport network by means of a GPS time series data. The proposed method is able to be developed to generate static data, manage data changes and to check on sudden detours in real time. The creators of the approach suggest that the infrastructure required to collect data using this approach is small and low-cost, comprising of one main server and on-board units for each vehicle.

Using Big Data to inform public transport systems requires the collecting and processing of multiple data streams to input into prediction algorithms. Information is generated by analysing traffic data and public transport vehicle data through the application of machine learning techniques, which can utilise large amounts of data to reveal complex patterns. Service disruptions can also be mitigated by offline analysis of passenger behaviour during severe disruptions, which allow the adjustment of transport lines to high demand areas.[36] Ultimately, prediction algorithms built based on Big Data can be used to optimise future transport networks as well as monitor, schedule and manage disruptions in real-time (See Sections 4.2.2 to 4.2.4).

2.2 The Future of Technology Enabled Transport

As will be explored in the upcoming SBEnrc project "*Tech-Enabled Transport: Informing the Transition to Technology Enabled Transportation Vehicles and Infrastructure*", it is anticipated that in the near future vehicles will increasingly act as mobile computers which produce, process and react to a constant stream of road data, including the locations of other cars and objects, real and expected traffic conditions, and optimal routes to a given destination. Vehicles will be able to harness data from various sources, such as other vehicles, road conditions, and road infrastructure. Vehicles that have these capabilities are being referred to as 'connected

[31] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

[32] Ma, X., Wu, Y.J., Wang, Y., Chen, F. and Liu, J. 2013, 'Mining smart card data for transit riders' travel patterns', *Transportation Research Part C: Emerging Technologies*, Vol. 36, pp. 1-12, http://dx.doi.org/10.1016/j.trc.2013.07.010

[33] Zhao, J., Rahbee A. and Wilson, N.H.M. 2007, 'Estimating rail passenger trip origin-destination matrix using automatic data collection systems', *Computer Aided Civil and Infra. Eng.*, Vol. 22, pp. 376-387. (DOI: 10.1111/j.1467-8667.2007.00494.x)

[34] Berkow, M., El-Geneidy, A.M., Bertini, R.L. and Crout, D. 2009, 'Beyond generating transit performance measures', *Transportation Research Record*, Vol. 2111, pp. 158-168. (DOI: 10.3141/2111-18)

[35] Fiori, A., Mignone, A., Rospo, G. 2016, 'DeCoClu: Density consensus clustering approach for public transport data', *Information Sciences*, vol. 328, no. 1, pp. 378-388.

[36] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

vehicles' and extensive access to multiple communications services is required for effective operation of such vehicles.[37] Vehicle manufacturers around the world are in a race to embed greater levels of technology into vehicles with the goal of eventually providing what is referred to as an 'autonomous vehicle', which is a vehicle that does not require a driver. Weather this goal is achieved or not, or in fact even preferable with concerns that it may actually increase congestion and reduced ridership on public transport, given such vehicles would need to be connected to other vehicles and infrastructure significant benefits can be reaped weather the driver takes their hands off the wheel or not.

Connected vehicles provide the potential to harness high-velocity vehicle-generated information streams and data from associated infrastructure in real time. To clarify, there are three types of vehicle related data transfers: Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I) and Vehicle to Everything (V2X).

– *Vehicle-to-Vehicle (V2V)*: Vehicle-to-Vehicle communication allows transmission of information between vehicles, creating the potential for organisation and cooperation to prevent accidents and map vehicle information across networks to improve safety and reduce congestion. It is anticipated that such a connected vehicle could intervene to prevent an accident, for example braking before a collision occurs, either caused by the driver or due to the behaviour of other vehicles the driver cannot see till it's too late. A number of car manufacturers are now testing V2V prototypes with Toyota announcing in 2016 it will increase its V2V enabled fleet test size to 5,000 vehicles in Ann Abor, Michigan.[38] However data compatibility is proving to be a challenge for the industry with the Mercedes E-Class only capable of communicating with other E-Class models. In order to streamline efforts the United States Department of Transport propose to require that all new light vehicles from 2021 have V2V technologies with a standard V2V frequency to allow cross make and mode communication.[39]

– *Vehicle-to-Infrastructure*: In the near future vehicles themselves will send data streams that include multiple variables directly related to transport infrastructure to be used for congestion management, emergency response, and predictive analysis. Visa versa infrastructure can communicate directly to vehicles to provide alerts, nominate optimal speeds to reduce congestion and travel time, and create open corridors around emergency vehicles and public transport vehicles. For instance, Audi's Traffic light information system released on select 2017 Audi Q7 and A4 models will inform drivers of the timings until traffic lights change to green.[40] Third party applications such as the 'EnLighten' application provide similar information to motorists, harnessing traffic signal timings to provide travel speed recommendations, and is now being installed in BMW vehicles.[41]

[37] Weeratunga, K. and Somers, A. 2015, Connected vehicles: Are we ready?, Main Roads Western Australia internal report, Perth.

[38] Asian Development Bank 2016, 'Safety and Intelligent Transport Systems Development in the People's Republic of China', *Asian Development Bank,* Section B.8.4.

[39] National Highway Traffic Safety Administration 2016, 'US DOT advances deployment of connected vehicle technology to prevent hundreds of thousands of crashes', *United States Department of Transportation*, Washington DC.

[40] Herndon, Virginia. Audi USA Homepage, Press Release (August 15, 2016) "Audi announces the first vehicle to infrastructure (V2I) service – the new Traffic light information system"

[41] Zurschmeide, J. 2015, 'Stop wasting gas in the city with the app that knows when traffic lights will change', *Digital Trends*.

– *Vehicle-to-Everything:* Vehicle-to-Everything (V2X) combines V2V and V2I while also communicating with pedestrians, devices and networks, essentially allowing a vehicle to communicate with all surrounding elements on a road network that may affect it. The application of V2X approaches 'Big Data' proportions and harnesses a wide array of data streams and inputs in order to provide road users with information to create safer and more efficient road networks.

One of the crucial points of connected systems is interoperability as connected vehicles should have compatible technology so that vehicles on the road can communicate with infrastructure and each other regardless of their make and model. In this vein, progress on harmonisation of technology standards internationally has delayed implementation and adoption of connected vehicles in Australia.[42] The Australian Communications and Media Authority is in the process of consultation with industry to develop a regime for the authorisation of 'Cooperative Intelligent Transport Systems (C-ITS).[43] Transport Certification Australia and Austroads are also working on a system to ensure the security, robustness, and credibility of C-ITS systems in Australia.[44] Testing of connected vehicles has also commenced in NSW, which has implemented a C-ITS testbed in Illawarra for 60 participating heavy vehicles fitted with V2V and V2I technology which broadcasts on the 5.9GHz radio spectrum.[45] In an internal report, Main Roads WA also recommends the staged installation of road-side units which have connection capability, especially for planned road developments.[46]

2.3 Privacy Concerns

The exponential growth of mobility-related data will trigger significant changes in the transport industry accompanied by rising concerns relating to the adequacy of regulations ensuring privacy.[47] Even data that is said to be 'anonymous' may still be able to be linked to specific individual sources if cross-referenced with other sources of related data, especially as much of the data is currently shared with private companies with no accountability. Not only do traffic management centres have to tackle this issue, they also have to decide on whether the data is reliable enough. Much of this data right now has to be verified with other data sources such as sensors and camera footage or still-shots. In addition, companies may need to migrate to non-relational (NoSQL) databases to accommodate and process large unstructured data sets. These NoSQL databases usually use external security enforcing mechanisms; hence to reduce security breaches, companies have to use additional security software, reviewing security policies for the 'middleware' between the operating system and the NoSQL database, while also toughening the NoSQL database itself to match its counterpart relational databases.[48]

The multi-tiered nature of Big Data means that transaction logs are stored in multi-tiered media. In smaller datasets, IT managers can manually move data between tiers, giving them a measure

[42] ACMA 2016, 'Proposed regulatory measures for the introduction of C-ITS in Australia', Australian Communications and Media Authority, Australian Government, accessed 25 Jan 2017.
[43] ACMA 2016, 'Proposed regulatory measures for the introduction of C-ITS in Australia', op. cit.
[44] TCA 2017, 'Cooperative Intelligent Transport Systems (C-ITS)', Transport Certification Australia, Australian Government, accessed 25 Jan 2017.
[45] Centre for Road Safety 2016, 'Cooperative Intelligent Transport Initiative', *NSW Government*, accessed 25 Jan 2015.
[46] Weeratunga, K. and Somers, A. 2015, Connected vehicles: Are we ready?, op. cit.
[47] International Transport Forum 2015, 'Big Data and Transport', op. cit.
[48] CSA 2012, *Top ten Big Data Security and Privacy Challenges*, Cloud Security Alliance, Rolling Meadows, Illinois.

of control; however, as the dataset grows exponentially, auto-tiering is likely to become increasingly necessary for big data storage management. As auto-tiering does not keep track of where the data is stored, unauthorised access to data stores is less easy to detect and data breaches may occur. Thus, new mechanisms must be developed to prevent data theft and maintain the 24/7 availability.[49] In Australia, the Privacy Act regulates and protects personal information, including the Australian Privacy Principles (APPs) which define the standards, rights and obligations in relation to handling and assessing personal information. Big Data changes how key privacy principles—which include data collection, minimisation of data retention and use limitation—are applied. However, as the APPs are technologically neutral, corporations and other organisations can adapt their Big Data handling policies to protect personal information while also retaining the maximum use from the information derived from Big Data analysis.[50]

According to the Privacy Act, organisations must take reasonable steps to implement practices, procedures and systems that protect personal information. These organisations must also be able to deal with privacy related complaints from individuals. A systematic risk management approach must be used to identify reasonable steps according to the size of the entity, its resources and the complexity of its operations. Organisations dealing directly with Big Data must take more rigorous and detailed privacy protection procedures than an entity handling the results of Big Data analytics. [51] One possible technique to ensure privacy is to use de-identification to remove the personal identifiers such as addresses and date of birth, as well as any other unique individual characteristics. This means that the Privacy Act no longer applies.[52] However, this technique is not foolproof and if de-identified datasets are matched to other datasets or other information, it is possible that individuals can be re-identified.[53]

[49] CSA 2012, *Top ten Big Data Security and Privacy Challenges*, op. cit.
[50] OAIC 2015, *Consultation draft: Guide to Big Data and the Australian Privacy Principles*, Office of the Australian Information Commissioner, Canberra.
[51] OAIC 2015, *Consultation draft: Guide to Big Data and the Australian Privacy Principles*, op. cit.
[52] OAIC 2015, *Consultation draft: Guide to Big Data and the Australian Privacy Principles*, op. cit.
[53] OAIC 2016, *Privacy business resource 4: De-identification of data and information*, Office of the Australian Information Commissioner, Canberra.

3 PLATFORMS AND TECHNOLOGIES FOR BIG DATA ANALYSIS AND ANALYTICS

3.1 Emerging Digital Platforms for Big Data

3.1.1 Overview

There are a number of promising emerging software platforms that can potentially be applied to further harness the potential of 'Big Data'. Some of these platforms both collect and analyse data, while others require the input of raw data collected by the user. For transportation applications, the input datasets come from a range of sources including GPS data from mobile phones, taxis and buses, as well as information taken from traffic light sensors, fixed sensors, and other datasets. The data analysis platforms then analyse and interpret the information to respond to user queries. While previous platforms have mainly focused on small data mining, there are now a robust set of digital platforms that are leaders in Big Data processing when users have numerous flexible requirements. The use of cloud based or in house processing platforms each have their own advantages. Cloud based services provide a less expensive startup cost, however the user is generally charged for the amount of data they wish to process. On the other hand, in house systems incur larger initial costing and maintenance fees, but offer an unlimited amount of data processing.

Hadoop distributes data collections across multiple nodes within a cluster of servers, meaning custom hardware does not need to be bought or maintained. Spark is a data-processing tool which operates on the data distributed using Hadoop. Although Hadoop can process the data, Spark does so at a significantly faster rate. Both programs are open-source, free and can be used in conjunction to increase processing speed; however, modifications have to be made to the program in order to customise it to a user-specific application. Because Hadoop and Spark are open source, many companies build their own data analytics software based on these frameworks. These companies can be hired to develop a more specialised system and establish a support network. Many of these companies are also start-ups, which means that while complex, high level services may be obtained from such companies at a lower price (compared to more experienced competitors), there is a level of risk that the company may collapse or fail to perform to the expected standard.

Alternatively, SAP HANA provides an all in one platform which has been proven to be effective in handling the data required to analyse traffic congestion. Being run through cloud based or in house servers makes SAP HANA versatile and the most cost effective SAP HANA setup can be chosen based on how the system must perform. However, this platform is not open source and licensing must be purchased or rented. This is not necessarily a bad thing, as SAP HANA provides large amounts of support, especially with hardware, and can help set up a Big Data system quicker than using an open source platform. Table 1 compares the three technologies and tools often used to analyse Big Data.

Table 1: Comparison of the three Big Analytics platforms

Hadoop with MapReduce	Hadoop with Spark	SAP HANA
Open-source	Open-source	Closed-source
Disk memory, which uses batch processing where data is stored and then processed at specific intervals	In memory, allowing real-time continuous processing of incoming data	
No support	No support	Hardware support
Machine learning capability has to be specifically programmed into the platform		In-built machine learning capability for better predictive ability

3.1.2 Hadoop

Hadoop is a framework platform that can be used for the processing of large data sets across clusters of computers. It is designed to be highly scalable, meaning it can be scaled from a single machine to many thousands, with each offering local computation and storage. Rather than rely on hardware, the library itself is designed to detect and handle failures at the application layer, making it a highly robust and efficient tool. Scalability also gives it a large range of uses and users can modify it for specific needs. Hadoop has multiple systems that can be used in conjunction with each other to enhance specific properties. These include:

- *Hadoop Common*: The common utilities that support the other Hadoop modules.
- *Hadoop Distributed File System (HDFS)*: A distributed file system that provides high-throughput access to application data.
- *Hadoop YARN*: A framework for job scheduling and cluster resource management.
- *Hadoop MapReduce*: A YARN-based system for parallel processing of large data sets.[54]

Figure 1 shows how the applications can be used for both single use and multi-Purpose Platforms.

[54] Apache Hadoop 2016, 'What Is Apache Hadoop', The Apache Software Foundation.

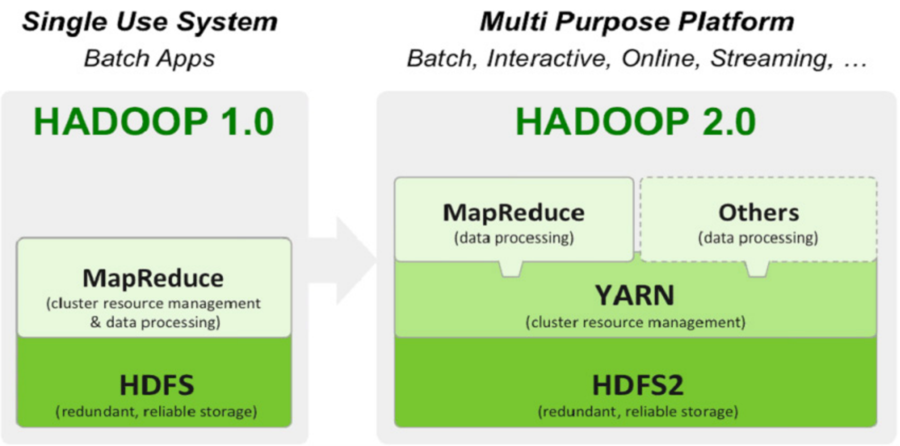

Single Use System
Batch Apps
HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
Multi Purpose Platform
Batch, Interactive, Online, Streaming, …
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS2
(redundant, reliable storage)

Figure 1: Data Analytics News 2014[55]

Hangzhou Trustway Technology Co. Ltd. has significantly improved its transportation management capability using Apache Hadoop on Intel® Xeon® processors (specifically using MapReduce, Hive and optimised HDFS). Hadoop's software is connected with traffic monitoring equipment such as the city's checkpoints, video monitoring, traffic flow detection, signal systems, and devices to provide the city with a big data storage system with high throughput and fault tolerance. This allows for a fast and efficient dynamic monitoring system that enables vehicle track analysis and searching, fake plate number analysis, vehicle control and traffic violation data storage.

> *Testing has shown the system is capable of carrying out collision analysis on 2.4 billion plates in only ten seconds*.[56]

From members of the Hadoop framework, different software can be produced for different needs, as highlighted in the following section with Apache Spark. Spark was implemented through Hadoop YARN, HDFS and MapReduce and is discussed in the following section.[57]

3.1.3 Spark

Apache Spark is an open source framework designed to perform real-time processing. As a whole, Spark cannot distribute files throughout the system, however it is used in conjunction with Hadoop as an alternative to MapReduce when data sets are in the processing stage. Both Spark and MapReduce rely on resilient distributed data (RDD) sets, which are representative of the incoming data for which the analysis takes place. RDDs have the capability to recompute their own data if a systematic failure were to occur due to their long lineage.[58]

The development of Spark is to be implemented alongside Hadoop due to the various limitations of MapReduce, which essentially processes the data as a batch, and stores the processed data on a disk. Conversely, Spark allows for analysis to be undertaken on a single

[55] Bigdata 2014,'Hadoop 2.0 and YARN Architecture', BDAN.
[56] Intel 2013, 'Improving traffic management with Big Data analytics', Intel Xeon.
[57] Hortonworks 2015, 'Apache Spark'.
[58] Anoop Daware 2015, 'Apache Spark vs. MapReduce', MapR.

cluster of both analytical and operational data using real-time processing. Additionally, the storage of the data in memory creates a significantly quicker run-time.[59]

Both programs differ in the way they process the data, however since MapReduce was designed solely to perform batch processing, its performance can plateau. As a result, Spark has been known to process the same dataset 100 times faster than Hadoop's MapReduce using memory, or 10 times faster using its disk.[60] Figure 2 shows the structure of the Spark platform.

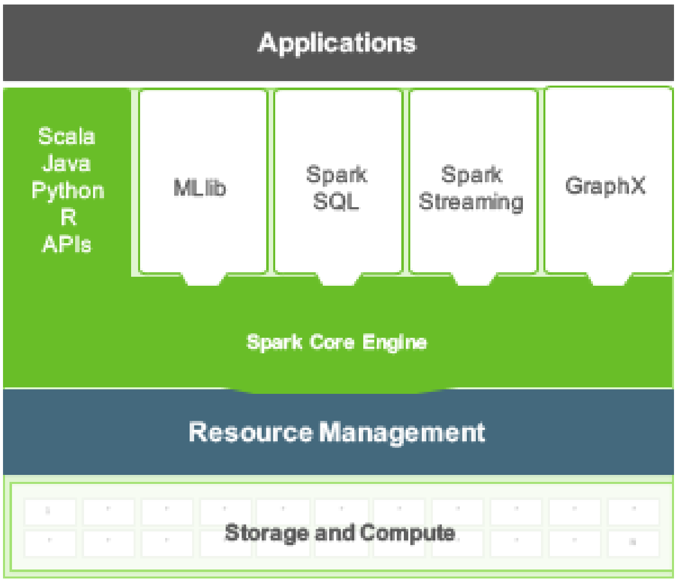


Figure 2: Structure of the Spark platform[61]

A case study undertaken in India illustrates the power of the Spark platform. Using gathered sensor values containing vehicle speed, count of the vehicle and the time taken for the vehicle to pass by the sensor area, a significant amount of data points is converted into a comma separated value (CSV) file. The CSV file is processed in Spark in order to predict the severity of traffic congestion, and the study found that …

> *…where the existing system took a significant amount of time, the implementation of Spark predicted traffic conditions in half a second.*[62]

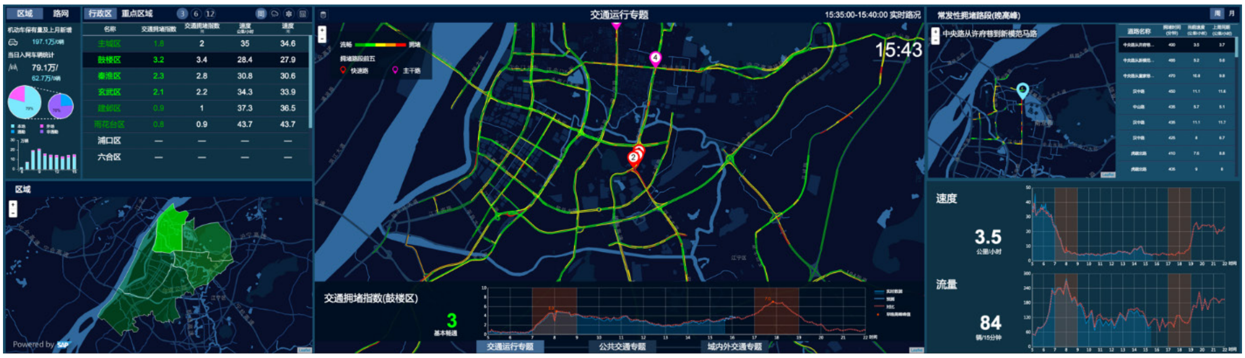
3.1.4 SAP HANA

SAP HANA is a single in-memory platform which combines application services, high-speed analytics and data acquisition tools to deliver an all in one package. SAP HANA can be integrated as either an in house system or through a cloud based database service such as SAP HANA Cloud Platform, HP Helion, Microsoft Azure and many others. The platform lends itself perfectly to be applied as a Big Data analytics tool and is being used around the world for this purpose. SAP HANA can handle multiple data inputs and provide predictive analytics as well as spatial and graphical data processing.[63] The platform also excels at delivering deeper insight from Big Data and the Internet of Things due to its strong machine learning capabilities. Due to the software's

[59] Vardhan 2015, 'Apache Spark vs Hadoop MapReduce', Edureka.
[60] Xin 2016, '$1.44 per terabyte: setting a new world record with Apache Spark', databricks.
[61] Hortonworks 2015, 'Apache Spark'
[62] Prathilothamai, M. Lakshmi, A.M. and Viswanthan, D. 2016, 'Cost Effective Road Traffic Prediction Model using Apache Spark', *Indian Journal of Science and Technology*, vol. 9, no. 17. DOI: 10.17485/ijst/2016/v9i17/87334
[63] SAP HANA n.d., What is SAP HANA?

ability to perform analytics in-memory on a single data copy, outputs from the raw data can be obtained in real-time allowing for instantaneous reactions in dynamic systems,[64] such as traffic management.

SAP HANA is being used around the world for many different purposes such as business analytics, bank processing and more recently traffic management. Both Japan and China have used SAP HANA to reduce congestion on city roads. The Nomura Research Institute (NRI) conducted an experiment using GPS data from 18,000 taxis located in Tokyo using Intel®Xeon® processor E7's to run SAP HANA.[65] Using the platform, they managed to process approximately 360-million items of data in a single second, as discussed in Section 4.1.1. In Nanjing, China, SAP HANA has been used in a similar way to provide real time traffic congestion analyses. Nanjing also used Intel®Xeon® processor E7 as their main processor and were able to process 100 million points of data from both floating car data (GPS) and fixed device data (eg. traffic sensors) each day.[66] The software then gives a graphical "temperature" rating to the roads in the city which is used by city management and accessed by over 800,000 members of the public.[67] Figure 3 is an example of the graphical output SAP HANA is able to provide.



名称	交通拥堵指数	交通拥堵指数	速度	速度
主城区	1.8	2	35	34.6
鼓楼区	3.2	3.4	28.4	27.9
秦淮区	2.3	2.8	30.8	30.6
玄武区	2.1	2.2	34.3	33.9
建邺区	0.9	1	37.3	36.5
雨花台区	0.8	0.9	43.7	43.7
浦口区	—	—	—	—
六合区	—	—	—	—

Figure 3: Output from SAP HANA in a traffic management application[68]

[64] SAP HANA n.d., What is SAP HANA?, op. cit.

[65] Intel 2011, Not limited to ERP applications alone, the Intel® Xeon® processor E7 family also provides new business opportunities via in-memory technology, Intel.

[66] Chen, 2016, Smart Traffic: an IOT solution, SAP HANA Innovation Award

[67] Chen, 2016, Smart Traffic: an IOT solution, op. cit.

[68] Chen, 2016, Smart Traffic: an IOT solution, op. cit.

4 BIG DATA CASE STUDIES AND APPLICATIONS

Harnessing Big Data and associated technologies across the transportation system has huge potential to reduce traffic congestion. The challenge is to find cost effective, efficient ways to use Big Data to inform transport management policy and decision making on both planning and active real-time levels. Multiple different techniques, methods and programs have been implemented to apply Big Data to transport issues and this section provides snapshots of international and Australian examples of Big Data applications.

4.1 International Examples: Real-Time Congestion Management

Real-time use of both for-purpose 'small data' and Big Data for congestion control and mitigation is the most well-developed of the Big Data transport applications. Tokyo, Copenhagen, New York, Seattle and Dublin (Sections 4.1.1 to 4.1.5) have begun the move towards Big Data by combining multiple types of data from multiple for-purpose small data streams, then supplementing these datasets with unstructured data from other sources. On the other hand, many large cities, such as Hong Kong and Zhejiang (4.1.6 and 4.1.7), use 'small data' from sensors installed specifically for congestion control to implement real-time congestion management systems. The UK's Twitraffic is an interesting application which uses Twitter to identify congestion and hence allows commuters to decide on alternative routes (4.1.8)

4.1.1 Zenryoku Annai!, Tokyo

This Japan-based application uses in-memory computing technology to analyse traffic jams. The data streams come from traffic sensors, satellite navigation systems, mobile location data and taxi GPS location data. The software analyses traffic data through statistical analysis on position and speed information from subscribers, moving vehicles and pedestrians.[69] Due to the high density within Tokyo, the Zenryoku Annai! application receives approximately 360 million data packages every second. The software platform is built by Nomura Research Institute (NRI), one of Japan's leading IT firms, and uses SAP HANA's processing system (See Section 3.3). One of the key aspects of this case study is the use of advanced IT processing technology to provide near instantaneous response times. NRI's in-memory computing technology can process millions of different data types in one second, far outstripping ordinary relational databases which take a several minutes to process the same volume of data.[70]

4.1.2 Copenhagen Connecting, Copenhagen

In Denmark, Copenhagen Connecting (CC) spent an estimated $9 million to implement an ITS to control traffic digitally, adapt to weather and real time traffic conditions to improve traffic within the city. This system provides real time traffic data and combines GPS data from cars, personal navigation devices, mobile apps and fleet vehicles. This provides the transport agency (Danish Road Directorate) with information of the current road conditions and may provide individuals with information of congestion spots and potential re-routing options. The use of GPS probe data to monitor traffic and congestion allows for a significant improvement in traffic

[69] Nomura Research Institute 2011, 'Not limited to ERP applications alone, the Intel Xeon processor E7 family also provides new business opportunities via in-memory technology', *Intel*.

[70] Mullich, J. 2013, 'Drivers avoid traffic jams with Big Data and Analytics', *Bloomberg L.P.,* New York.

efficiency. Implementation of CC has resulted in drivers spending an average of 7.5 minutes in traffic a day.[71]

4.1.3 Midtown in Motion, New York

In 2011, the New York City Department of Transport enacted the first phase of the 'Midtown in Motion' (MiM) project. Designed to reduce crippling traffic congestion, the system allows traffic engineers to identify and respond to traffic conditions in real time across the Midtown area in New York City.[72] The initial phase of the project saw the installation of:

- 100 microwave sensors for vehicle flow and occupancy measurement
- 23 ETC readers at intersections to sense the metro cards of pedestrians and commuters as they pass by
- 32 traffic cameras for verification and system monitoring.[73]

Using a previously commissioned wireless communication network, real time data streams from the system are centralised and processed at a traffic management centre in Long Island City. All MiM data is transmitted wirelessly and engineers at the traffic management centre can immediately analyse the data, identify congestion issues and adjust network traffic signals accordingly through network connected advanced solid state traffic controllers. Before full implementation of the MiM system, data from installed sensors was used to create a traffic database of Big Data. Once live, real-time traffic conditions are compared to the database and using an 'algorithm of adaptive control', the system recognises the build-up of traffic and automatically adjusts green-time phasing of traffic lights. In addition to this, the system allows for management centre operators to manually respond to isolated incidents through the use of pre-determined signalling. Since its implementation, Midtown average vehicular travel speed has seen a 10 per cent improvement from pre MiM baseline levels.[74] The Department of Transport has invested $300 million in traffic management tools and advanced technology across New York which have facilitated the implementation of MiM.

4.1.4 Concert, Seattle

Siemens's integrated traffic management system 'Concert' is to be implemented in 2016 to link previously separated traffic planning and control centres and improve traffic management and congestion. The Concert platform will integrate the traffic control system, existing variable message signs, freeway systems provided by the Washington Department of Transportation, historical traffic data and real time weather and road conditions.

Real-time data provided to Concert gives a comprehensive view of the traffic, allowing congested zones to be quickly identified so that system wide mitigating action can be implemented (eg. through dynamic traffic signs). Additionally, information on the location of congested zones will be shared with travellers through dynamic mapping and messages, who

[71] Clinton, N. 2015, 'Denmark takes lead in smart traffic control with procurement innovation', *Public Spend Forum.*

[72] Solomonow S., Mosquera, N. 2012, 'NYC DOT announces expansion of Midtown congestion management system, receives National Transportation Award', *The City of New York,* June 5.

[73]Siemens AG 2014, *Traffic Management Transformed*, Siemens.

[74] Solomonow S., Mosquera, N. 2012, 'NYC DOT announces expansion of Midtown congestion management system, receives National Transportation Award', op. cit.

can then change their routes accordingly. Traffic-related incidents can be quickly identified and the best management response employed.[75]

4.1.5 Dublin Road Congestion System, Dublin

Policies were introduced in Dublin during the 1980s and 90s to restrict the construction of new roads in order to preserve Dublin's historic fabric. As a result the growing city suffered from traffic congestion. Dublin City Council started a partnership with IBM in 2010 to make Dublin a smart city test bed which involved a proposal to collect and analyse Big Data throughout the city to combat congestion. The project involves the capture and monitoring of a range of data sources, including road sensors and GPS updates from the city's 1,000 buses to produce a digital map of the city displaying the real-time positions of the buses. By updating the journey information every minute, city council made it possible for 1.2 million residents to find the fastest route for their travel, moving more efficiently through Dublin's extensive network of roads, tramways and bus lanes. Using reporting facilities in the traffic centre, the optimal traffic-calming measures were identified to manage congestion.[76]

Operational data from Dublin's four local authorities are also provided in open data format as Dublinked,[77] which is followed by many other municipal governments around the world. While being used by passengers for trip planning and scheduling, this open data can also be used for app development, which can provide information on 2700 points of interest such as public buses, taxis, bikes, beaches, parks and gardens, along access to traffic cameras, weather reports and bike-rentals services.

4.1.6 Transport Information System, Hong Kong

To improve transport safety, efficiency and implement a smarter transport system, Hong Kong reviewed its existing system and established the Transport Information System (TIS). This system allows for the collection and processing of transport information to support real time traffic. Information from small scale for-purpose traffic control systems, as well as traffic control surveillance systems, is integrated into a Traffic Management and Information Centre (TMIC). While the system currently only uses two types of data, hence constituting a 'small data' platform, two facts are of key importance in this case study:

- The plethora of services that can be provided using just these few datasets. The TMIC provides real-time traffic information for the TIS, which provides 4 services:
  - Road Traffic Information Services, which integrates unexpected incident warning and real-time traffic speeds
  - Hong Kong eRouting, which enables commuters to plan their driving routes to avoid delays
  - Hong Kong eTransport, which enables commuters to plan public transport routes across different modes (eg. bus, train)
  - Intelligent Road Network, which provides up-to-date traffic information for private service provides

[75] SCC 2016, 'Lessons from a CTO: Seattle's path to becoming smarter', *Smart Cities Council*.
[76] Berst, J. 2013, 'Smart mobility: Dublin uses real-time data to reduce congestion', *Smart Cities Council*, May 18.
[77] Dublin City Council 2016, 'Smart Dublin Challenge', *Dublin City Council*.

- Rapid consumer uptake of its services, indicating commuter interest in these features. Commuters are increasingly using these trip planning systems, suggesting that it has been useful in improving travel times: from 2010 to 2013, the number of eRouting downloads has increased five-fold from 200 to 1000 downloads per day, reflecting increased use of the transport planning systems.[78]

4.1.7 Traffic Management, Zhejiang

Rapid development in Zhejiang, China, has led to heavier traffic and an increase in accidents and traffic offences. Key city checkpoints were equipped with 1,000 monitoring devices to capture image and video data reaching to a terabyte in size monthly. The next challenges were enabling centralised management of traffic data, optimised use of this huge amount of data and improving the traffic flow city-wide. The city uses 22 servers running on Intel Xeon processor E5 series, a unified data centre with 198-terabyte storage space. The Hadoop Distributed File System (HDFS) and Apache HBase* provide 24 month storage and the Trustway system is deployed for massive-scale data mining and analysis. Using the databank of Big Data collected and stored, traffic police departments can identify traffic offenders and extract relevant historical data including the behaviour, routing and car related information. In addition, traffic conditions can be analysed in real-time[79].

Although the Zhejiang example uses only one set of data (that of video imaging), the city is noted in this report because of its large and technologically advanced data collection and mining system, which sets it apart from many other cities and gives an indication of the extent of data load that needs to be analysed in Big Data applications.

4.1.8 Twitraffic, United Kingdom

Another interesting 'small data' case study with Big Data implications is in the UK, where an application has been developed based on social media. Twitraffic[80] is a novel smartphone application that provides real-time traffic information and journey-planning to its users according to Twitter posts.[81] It trawls through Twitter searching for keywords such as 'traffic', 'accident', 'congestion' and 'road-works', then applying a fast and effective sentiment analysis software to extract and validate the useful data for real-time traffic flow. Twitraffic reportedly flags traffic incidents around 7.1 minutes faster (on average) than the UK Government Highways Agency.[82] While the application only uses data from one source and hence cannot be classed as Big Data, it does demonstrate the importance and effectiveness of social media in diagnosing and identifying congestion. As such, social media can in fact be a viable component of a Big Data congestion management system.

4.1.9 Transportation Decision-Making, California

While the technologies discussed in Section 3.1 to 3.3 are broad-based applications that can be adapted to suit transport applications, a novel platform that uses multi-source data and applies

[78] Siemens AG 2014, 'Traffic Management Transformed', *Siemens*.
[79] Cohen, B. 2012, *The top 10 smart cities on the planet*, *Fast Company & Inc*, New York.
[80] Twitraffic 2016.
[81] GeoConnexion 2016, 'Twitraffic: the power of Twitter for real-time traffic information', *GeoConnexion*.
[82] Fankhauser, D. 2012, 'Get there faster with these 4 traffic apps', August 23.

it specifically to congestion management also exists. TransDec[83] is developed by the Integrated Media Systems Centre in the University of Southern California. It includes real-time large-scale traffic sensor data collection, efficient real-time and historical spatio-temporal data processing and data analysis and visualisation. The system has a three-tier architecture: data tier, presentation tier and query interface. On the data tier, TransDec employs various real-time traffic data such as NAVTEQ and RIITS (Regional Integration of Intelligent Transportation Systems). The RIITS dataset is made of historical and minute-to-minute real-time data provided by various organisations such as Caltrans D7, Metro, LADOT, and CHP, which includes CCTV snapshots, events, bus locations and arterial congestion. TransDec also uses data mined from the US transportation network map and points of interest. Next, the presentation tier and query interface provides a web-based map to the user, enabling spatio-temporal queries to be made interactively. The TransDec intelligent transportation system provides monitoring, decision-making and management services. However, it is not exempt from the processing challenges faced by many Big Data and Big Analytics platforms, which are discussed in Section 5.1.

[83] Integrated Media Systems Center 2016, *Transdec: Big Data for Transportation*, University of Southern California, Columbia.

4.2 International Examples – Predictive Congestion Management using Historic Big Data

Unlike real-time congestion management, historical Big Data mining for use in congestion prediction and prevention is still in the early stages of development and hence the scope for further development and application is extensive. There are two ways in which predictive congestion management can be applied:

- Debottlenecking of key routes during peak periods to defer infrastructure investment
- Transport corridor planning to ensure that any implemented infrastructure is optimal and best suits the requirements of commuters

This section looks at four case studies using Big Data for transport demand projections in order to highlight the various directions for use of Big Data.

4.2.1 JamBayes, Seattle

This Microsoft-based research project collects historical Big Data, including traffic data such as sensed highway data and accident reports, as well as contextual data such as weather reports and news reports of major regional events. Using the collected data, the software identifies key bottlenecks (22 were identified in Seattle) and uses a large model of all the bottlenecks in the area, taking into account interdependencies between bottlenecks as well as other factors such as the time of day and weather conditions. It then compares real-time events to the case library built from historical data in order to predict traffic conditions. The final output is a predictive program which produces an estimate of the location and duration of traffic jams.[84]

One of the key drawbacks of predictive software is its inability to consider system shocks and unprecedented events such as accidents or natural disasters. The JamBayes software attempts to consider such unexpected events by storing Big Data during surprising events in a 'case library'. The software then takes a snapshot of traffic and data conditions 30 minutes before the surprising state occurred. If current road conditions are found to match circumstances that previously occurred 30 minutes before a surprising event, the software will predict a surprising events 30 minutes into the future. A surprising state which leads to increased congestion on particular routes and surprisingly low traffic volumes on others. The effectiveness of these predictions has been found to be a 0.05 false positive rate, which translates to a 95% chance that the software makes the right prediction.[85]

By identifying the key bottlenecks in Seattle's transport system, and predicting when these bottlenecks will occur, transport planners can devise methods to prevent congestion in these areas; hence, infrastructure investment to reduce pressure on these areas can be deferred.

4.2.2 Stockholm, Sweden

Big Data provided by Stockholm's bus fleet, which uses automated vehicle location and automated passenger counting sensors which circulate from bus to bus, enables the

[84] Horvitz, E., Apacible, J., Sarin, R., Liao, L. 2005, 'Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service', *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, Edinburgh, Scotland. Arxiv ID: 1207.1352

[85] Horvitz, E., Apacible, J., Sarin, R., Liao, L. 2005, op. cit., Arxiv ID: 1207.1352

development of real-time performance algorithms. van Oort and Cats[86] analyse Big Data to diagnose system limitations and estimate passenger travel time, including waiting times between services (termed 'headways'). Their analysis found that unpredictable wait times between services caused poor capacity utilisation and congestion on the buses. Implementing a headway-based control system resulted in reduced travel times and improved transport system performance, allowing buses to run 'on schedule' more often.

In addition, Big Data was used to inform real-time control systems and provide reliable information to passengers about bus service availability. Big Data enabled prediction schemes that took into account current traffic, bus fleet and travel demand conditions. When historical data is integrated with information about downstream traffic conditions provided by Big Data, better travel time prediction is achieved.[87]

4.2.3 Intelligent Transport System, Singapore

The Intelligent Transport System (ITS) manages traffic and mitigates congestion in Singapore's central business district using ITS infrastructure and data collection to ensure free flowing and safely moving traffic. Singapore has multiple ITS infrastructure technologies, including the Electronic Road Pricing system (ERP), a physical gantry which detects congestion level and deducts time of day based charges from smart cards inserted within all vehicles), activated green wave system installed on all traffic signals,[88] parking detection systems, an expressway advisory and monitoring system and a taxi GPS system which monitors traffic conditions within the city. Big Data produced from these sources is processed at the ITS Operations Control Centre, which collates the data and provides real time updates with the aim to divert traffic from congested areas in real-time.[89]

While the above is another example of real-time congestion management using Big Data, Singapore's key point of difference is that its transport authorities pre-empt traffic congestion by improving public transport. To provide networks which meet the needs of commuters, historical and real-time Big Data from metro card, bus, taxi and Wi-Fi data was collected and analysed. Ridership trends and passenger routes are processed to identify areas where demand outstripped supply, and public transport resources and vehicles will then be redeployed according to passenger demand. These public transport improvements aim to reduce wait time, improve transport efficiency, and encourage increased uptake of public transport, hence reducing the number of private vehicles and improving traffic conditions.[90] The transport authorities have already used historical Big Data to find preferred routes during specific time periods throughout the day and are considering running shuttle bus services during those periods to remove the need for individual private vehicles travelling the same route.[91]

In 2020, Singapore will update its ERP system to one based on Global Navigation Satellite System (GNSS) Technology. This updated satellite-based system has great potential to be a

[86] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

[87] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

[88] Land Transport Authority of Singapore 2010, *Intelligent Transport Systems Centre*, Singapore Government.

[89] Hong Kong Transport Development 2006, 'Electronic Road System', Hong Kong Government, Hong Kong.

[90] Lui, T.Y. 2014, 'Singapore Urban Mobility', *International Transport Forum*, Leipzig, Germany.

[91] Tan, W.Z. 2015, 'Big Data could power on-demand public transport: IDA', *Channel NewsAsia*, 23 April.

store of Big Data, as it will track the movement of individual vehicles (in order to charge drivers per distance travelled on congested roads) and hence will provide detailed information on preferred routes and driver behaviour.[92] This has the potential to inform planned transport infrastructure developments, but it is not yet known if the transport authorities plan to use the information in this manner.

Big Data can also be used for a variety of novel applications which act to manage congestion on the level of individual travellers. Recently, MIT's SENSEable City Lab39 led a 5-year research project, named LIVE Singapore40, which collected, fused and shared a huge volume of city-wide Big Data. By sharing this data with the public, platforms were developed which predicted rain 10 minutes in advance and directed taxi-drivers to the part of the city where rain was expected. Rain 'spots' can be easily identified on a real-time map, as shown in Figure 4. Other models show the estimated temperature throughout the city and the resulting energy consumption; the level of mobile phone network usage; the global reach of the ports and airports in Singapore; the amount of text messaging as a result of the Formula One racing and where they originates form; and a map which deforms to represent the extra travel time as a result of congestion[93]. Commuters can also choose to receive real-time traffic flow updates collected using GPS and speed data from cars.[94]

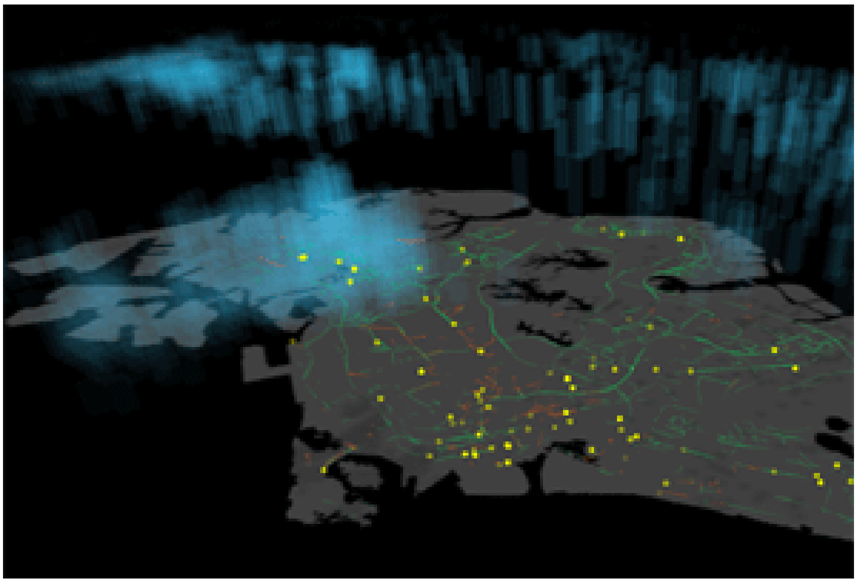


Figure 4: A rain forecast from LIVE Singapore which provides information to taxi drivers to move to the affected areas[95]

4.2.4 The Hague, Netherlands

Another study by van Oort and Cats[96] analysed smart (metro) card data in The Hague to develop a profile of actual passenger behaviour and create a prediction for future network demand. They distribute passenger flows over the network by superimposing passengers' routes, then investigate which model parameters best predict passenger loads. Subsequently, they construct

[92] Land Transport Authority 2016, 'Tender awarded to develop next-generation electronic road pricing system', *Singapore Government*, Singapore.

[93] Senseable City Lab, 2011, 'LIVE Singapore', MIT

[94] Copenhagen Cleantech Cluster 2012, *Danish smart cities: sustainable living in an urban world*, Copenhagen, vol. 41, no. 2.

[95] Senseable City Lab, 2011, 'LIVE Singapore', op. cit.

[96] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

a prediction model by converting Big Data smart card information into passenger journeys and place them on an origin-destination matrix. This matrix is used to reproduce measured passenger flows, enabling a simple analysis to be conducted.

Using their model, the authors analyse the results of a specific timetable adjustment such as a frequency increase of two public transport lines. The model produces a detailed projection of the impact of such a change on the entire network (both the affected lines and the other lines), including expected changes in ridership patterns. The algorithms generated using Big Data were found to be valuable in quantifying the impacts of changes in tactical planning such as vehicle frequency, timetabling and maintenance schedules. It allowed for an accurate prediction of how the frequency of public transport use can be increased based on timetable changes and how nearby lines are affected. However, the authors note that the model is only reliable for short term predictions and suggest further research to develop more accurate models.

4.3 An Integrated International Example: Transport for London

In 2015, London was estimated to have the worst congestion in the world, with drivers wasting an average of 101 hours (more than four days) in gridlock. Big Data collection and analysis is integrated extensively into the Transport for London (TfL) system, which handles bus network, trains, roads, taxis, ferries and cycle paths. These connected networks provide huge amount of information through smart cards, ticketing systems, vehicle sensors, traffic signals, social media, GPS location and mobile networks. TfL collects Big Data for both traffic management and emergency responses.[97] Currently, big data collected from TfL services is shared by TfL through third-party app developers which provide tailored solutions. A number of Microsoft and Oracle platforms are in use for running the TfL systems. Transport planners also plan to integrate open source solutions so the system can cope with future data demand, including increasing the capacity for real-time analytics, integrating a wider range of data sources and improved plan services.[98]

4.3.1 Traffic Management and Transport Planning

TfL utilises Big Data from FVD and CCTV throughout the transport network to control traffic management and produce a real-time view of traffic conditions. Using the collected data, variable speed limits are set to improve road network performance. The road management data is also shared across private sectors to enable freight companies to calculate optimal travel routes based on traffic conditions. Using Big Data produced from its extensive transport network, the citizens of London receive real-time information covering various subjects such as weather, air pollution, delays in public transport, availability of public bikes, level of the river, twitter trends in the city, and traffic camera feeds. These sites produce visualisations of complex Big Data, increasing ease of interpretation and analysis for non-experts.[99]

Big Data is also used for asset maintenance. Floating vehicle data, smartphone and GPS locations data are analysed to locate track defects to an accuracy of less than 5 metres. Using

[97] Marr, B. 2015, 'How big data and the internet of things improve public transport in London', *Forbes,* May 27.
[98] Cohen, B. 2012, *The top 10 smart cities on the planet*.
[99] Kitchin, R. 2014, 'The real-time city? big data and smart urbanism', *GeoJournal*, 79(1):1–14. ISSN 1572-9893. DOI: 10.1007/s10708-013-9516-8

big data analysis techniques overcomes the limitations of fixed sensors and reduces the cost and time taken to fix infrastructure. Crucially, transport planners also use Big Data to gain a deeper understanding of how people from different groups use public transport (eg. students compared to seniors) and hence improve network planning. One of the key sources of transport data is the Oyster prepaid travel card, which provides access to buses and trains. Combined with bus positioning transport data, a huge amount of data regarding journeys taken is collated and used to create a map of where and when passengers are travelling in the network, and how they are using more than one method of transport. The system can be used to plan interchanges so other services can benefit and the walking time can be minimised.

The use of Big Data in real-time and future (planned) traffic management has led to direct economic benefits, with increased transport efficiencies. TfL's data was valued at £15-58 million per year and resulted in 20 travel apps developed by private companies. Public transport usage has also increased to record highs, with the number of customers on public transport during the 2012 Olympics becoming a daily occurrence. The expansion in accumulated Big Data has allowed the adequate management of increased passenger numbers: operational performance across 2011-2016 shows a continued, and in some areas improved, level of passenger journeys. These include a reduction in wait times and traffic flow, and increased customer satisfaction and kilometres operated.[100] These statistics highlight the success of TfL's Big Data incorporation for transport network applications.

4.3.2 Emergency Response and Management

Emergency repairs on Putney Bridge, which is crossed by 870,000 people every day, led to an unexpected transport bottleneck in surrounding areas. Using Big Data analysis from historical and real-time Big Databases, TfL addressed this disruption by quantifying the affected passengers. Since half of the journeys started or ended very close to the bridge, TfL concluded that this group of passengers would be fine and they needed to manage the other half. A transport interchange was set up and bus services on alternate routes were increased. The affected passengers were also informed by personalised messages.

As such, the UK government has strategized the use of Big Data on a broad scale, maximising the opportunities of using such datasets. The fundamental focus behind the Government's strategy is 'opening up' data in transport by making it more widely available across various sectors to improve transparency and encourage economic growth. The Government established the Transport Systems Catapult,[101] overseen by the Technology Strategy Board (TSB), and is set to receive £46.6 million from TSB and £16.9 million from the Department for Transport (DFT), with the specific objective to encourage analysis of Big Data.[102]

[100] Transport for London 2015, 'Annual Report and Statement of Accounts', *Greater London.*

[101] Yianni, S., Zanelli, P. 2016, *Technology Strategy 2016 for Intelligent Mobility*, *Catapult Transport Systems*, London.

[102] The Parliamentary Office of Science & Technology 2014, *Big and Open Data in Transport*, UK Government, London.

4.4 Australian Examples

Currently, Australia has not yet tapped into the enormous potential of Big Data. Australia's cities use 'small data' from sensors installed specifically for congestion control to implement congestion management systems, much like Hong Kong and Zhejiang (Sections 4.1.6 and 4.1.74.1.7), but on a smaller scale due to the smaller volumes of data received. The systems currently implemented in cities across Australia are discussed in further detail in the following section.

4.4.1 Congestion Management, Western Australia

According to the *Intelligent Transport Systems Master Plan* released by Main Roads Western Australia (MRWA) in 2014, '*Big Data can help Main Roads better understand how the transport system is currently being used, match available capacity to changing demand and manage its assets through data from the vehicles themselves*.' The plan focuses on intelligent infrastructure, smart vehicles and information services, and estimates that managed freeways could deliver a 27 per cent efficiency during peak times.[103] Western Australia's Traffic Congestion Management Program plans to harness multiple data intensive strategies in the coming five years in order to reduce congestion.

Information on the location of congestion is collected using for-purpose sensors such as vehicle detectors and CCTV. Congestions are managed using Variable Message Signs (VMS), variable speed limits and lane-use management signals. VMS are installed on the Kwinana Freeway, Mitchell Freeway, and Graham Farmer Freeway as well as major roads servicing these vital network links. These signs show the current traffic speed at nominated freeway points, as well as arrows indicating whether the speed is increasing or decreasing. VMS can also provide warnings of unusual traffic congestions due to car crashes, roadworks or other related hazards. VMS utilisation is being expanded with the integration of travel time information,[104] allowing road users to make informed decisions to plan their journey with greater knowledge of current travelling conditions.[105]

In 2015, traffic signal timings at key metropolitan traffic locations were optimised according to the amount of traffic demand. Trip times decreased significantly, with some corridors experiencing up to 24.4 per cent in travel time reduction during the morning peak period, and even larger 28.4 per cent during the evening peak hour period.[106]

4.4.2 Traffic Management and Emergency Response, New South Wales

The New South Wales government aims to increase the use of open data as '*sharing data allows government agencies to focus on delivering core public services. It encourages innovative solutions to our citizen's problems*'.[107] The successes so far in creating innovative solutions include 'TripView' and 'Fires near me', where for-purpose data has been used to assist in reducing peak congestion and enhancing disaster resilience and response. The TripView

[103] Main Roads WA 2015, 'Intelligent Transport System', *The Government of Western Australia,* Perth.
[104] Main Roads WA 2015, 'Traffic congestion management program', *The Government of Western Australia*, Perth.
[105] Main Roads WA 2015, 'Facts & Figures', *The Government of Western Australia,* Perth.
[106] Main Roads WA 2015, 'Traffic signal timing improvement project', *The Government of Western Australia*, Perth.
[107] NSW Government 2016, 'Premier's Innovation Initiative – Open Data', *New South Wales Government*, Sydney.

website and application provides the user with real-time public transport updates, which assists in reducing congestion and decreasing public transport transit times.[108] Other initiatives include using social media to help motorists in deciding a route, and utilising satellite technology to identify buses behind schedule so as to provide them with traffic signal priority where possible.[109]

The M4 Smart Motorway project planned for Sydney West has been designed to reduce congestion through intelligent traffic management drawing on real-time information and communication. The key features of the project include Variable Message Signs (VMS), ramp signalling, variable lane usage, and variable speed signalling.[110] The installation of CCTV and In-Road Sensors also provide monitoring of the network, and by installing these sensors every 500 meters along the motorway real-time traffic flow changes can be detected.

Predicted benefits of the upgrade include: [111]

- 30 per cent reduction in motorway crash rates
- Entry ramp signalling enabling smooth merging of traffic
- A reduction in journey times on the motorway of up to 15 minutes
- Real-time updates on congestion events via VMS while drivers are approaching the incident or before they enter the motorway
- Adjustable speed limits and lane closures to lessen the detrimental impact of any incident
- Ease of access to accidents and disasters for emergency vehicles through the use of VMS and lane signalling

Simulations conducted by National Information Communication Technology Australia (NICTA) suggest that travel times can be up to 40 per cent during peak hour, which increases the capacity of the highway equal to adding an extra lane. This equates to a $22 million saving per year from this project alone, ultimately providing the opportunity to save over $500 million per year across the entire Sydney system.[112]

4.4.3 Sydney Coordinated Adaptive Traffic System (SCATS)

First developed in Sydney, SCATS links a combination of traffic signals for road management coordination. These adaptive traffic systems assist in complex traffic management strategies implemented to synchronise traffic signals and optimise traffic flow in real time.[113] SCATS essentially analyses real-time traffic data to primarily control signal timings suitable for current traffic conditions. This system has been successfully implemented in several cities within the US, such as Bellevue Washington. Through the use of SCATS on Factoria Boulevard, travel times during peak hours have reduced by 36% since the adaptive traffic systems were installed.[114] It

[108] NSW Government 2016, 'Premier's Innovation Initiative – Congestion', *New South Wales Government*, Sydney.
[109] NSW Government 2016, 'Premier's Innovation Initiative – Congestion', op.cit.
[110] Service NSW 2016, 'Key features of the proposed M4 Smart Motorway', *New South Wales Government*, Sydney.
[111] Service NSW 2016, 'Key benefits of the proposed M4 Smart Motorway', op. cit.
[112] CSIRO 2015, 'Advanced data analytics in transport – Machine learning perspective', CSIRO, Sydney.
[113] NSW Transport Roads and Maritime Services 2015, 'How SCATS works', *New South Wales Government*, Sydney.
[114] Sanburn, J. 2015, 'How smart traffic lights could transform your commute', *Time*, May 5.

is estimated that the $5.5 million system saves drivers $9-12 million annually, if a driver's time is costed at $15/hr.

When comparing the operational aspects of the control system, InSync outperformed both SCATS and ACS-Lite in the three categories shown in Figure 5.[115] InSync has the lowest cost per intersection, averaging at $28,700 USD, whereas the cost of SCATS was up to $60,000 USD per intersection. While InSync have the lower cost and operational benefits, SCATS controllers have a higher proliferation across a large portion of NSW. The additional benefit of the use of IP cameras with the InSync control systems provides visual monitoring of all intersections across urban environments, resulting in more rapid and efficient detection of traffic delays, accidents or poor weather conditions. With the rollout of the National Broadband Network (NBN) in Australia hopefully predicted to be complete by 2021, if the InSync Ethernet communication systems were to be used in conjunction with this network, instantaneous and reliable transmission of data from both the controllers and IP cameras to control authorities could be achieved in order to provide real time information of conditions.[116] However, it must be noted that the study of each technology was examined at different intersections and not the same one therefore, there may be a slight variance in results.

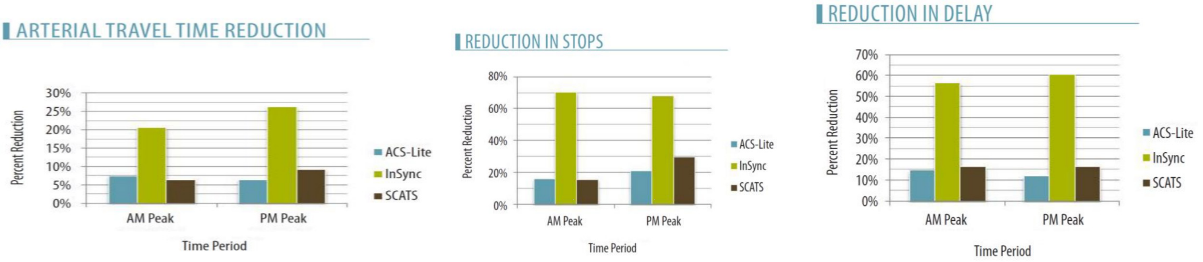


Figure 5: Comparative operational performance of ACS-Lite, InSync and SCATS

4.4.4 STREAMS, Queensland

Brisbane based company Transmax has developed the software platform STREAMS which uses SCATS real-time transport data and SCATS ITS ports. On top of the existing SCATS platform, STREAMS integrates CCTV, variable message signs and vehicle detectors to produce a map-based, browser-style ITS interface for transport network management. The combination of this integrated data set allows STREAMS to build a Geographic Information System (GIS) that models transport network infrastructure in real-time under a single software interface, with the ability to send SCATS control requests and manage ITS devices under the existing infrastructure. Currently operating in all Australian states (Northern Territory excluded), managing 1955 intersections with 48,500+ STREAMS connected devices[117], the Big Data ITS system delivers efficiency and performance across the entire road network. STREAMS provides the necessary integration of information to make large scale increased performance possible.

STREAMS has proved to be an effective platform in managing transport, reducing crashes and improving congestion. VicRoads reported a number of positive outcomes after implementing STREAMS for motorway ramp metering in 2007; economic benefits of $94,000 per day, travel

[115] Fernando, B., Gray, E., Kellner, J. 2013, 'A review of current traffic congestion management in the City of Sydney', op. cit.
[116] Fernando, B., Gray, E., Kellner, J. 2013, 'A review of current traffic congestion management in the City of Sydney', op. cit.
[117] Transmax 2016, 'STREAMS', *Transmax Pty Ltd*, Perth.

time savings of 42% during peak periods, a 30% reduction in motorway accidents, and an 11% reduction in greenhouse gas emissions.[118] Transmax has recently partnered with Parsons Brinckerhoff to introduce the emerging technologies of Big Data in transport management to the Colorado Department of Transport[119] to assist the U.S. in reliability, capacity and safety of its freeways.

4.4.5 Traffic Management and Emergency Response, Queensland

Managed Motorways, an initiative of Queensland's Department of Transport and Main Roads, draws upon state-of-the-art smart technology to manage the South East Queensland Road network, reducing congestion and improving safety.[120] The Intelligent Transport Systems (ITS) manages the flow of traffic through utilisation of variable message, speed limit and lane signs. Merging traffic will be effectively controlled using ramp signalling in order to minimise the interruption of flow. Electronic message signs display real-time travel time information to drivers, while roadside data systems such as sensors and CCTV monitor the conditions and enable rapid response to incidents. The Department aims to reduce stop-start travel, increase the ability to predict journey times, increase the capacity of the roadway while reducing crashes and emissions.[121]

In addition, the Queensland government and Transmax have developed an Emergency Vehicle Priority (EVP) solution which allows emergency vehicles to travel more quickly and safely. The EVP system uses the location of the emergency vehicle and the flow of surrounding traffic to estimate time of arrival at intersections. Using this information it is possible to pre-emptively change traffic lights to green before arrival to ensure the emergency vehicle can continue through without having to slow down. This system has been found to improve the travel time for emergency vehicles by 10-18%.[122] Not only is there little impact to the surrounding traffic, the stress of navigating red traffic signals by the drivers is greatly reduced while their safety is increased.

[118] Victorian Auditor-General's Office 2010, 'Using ICT to Improve Traffic Management', *Victorian Government*, Melbourne.

[119] Mena Report 2016, *Australia: Queensland congestion-busting traffic technology hits American highways*, GALE|A461442569.

[120] Department of Transport and Main Roads 2016, 'QLDTraffic', *Queensland Government*, Brisbane.

[121] DTMR 2016, op. cit., 'QLDTraffic'.

[122] Transmax 2016, op. cit., 'STREAMS'.

5 CONCLUSION

As huge amounts of data continue to explode across a huge array of platforms, Big Data is increasingly relevant in providing high-resolution information to optimise transport systems. The exciting implications of Big Data have yet to be fully realised, especially in Australia, which currently only draws on real-time, for-purpose 'small data'. As the case studies in this report illustrate, Big Data has a high potential to prevent congestion and has achieved substantial returns internationally. The work conducted and successes achieved internationally from the use of Big Data provides Australia's city planners and transport organisations with ideas and avenues to harness Big Data to improve transport systems in the nation. This task, however, is not without its challenges.

5.1 Existing Challenges

Although Big Data can provide key information to evaluate, plan and improve transport systems, the key challenge in the utilisation of Big Data is the fact that the extensive volume of information requires multiple modes of data analysis and processing. Because so much information is available, software and programs must be developed which can sift out irrelevant information and focus on key features of the data which will provide necessary inputs into transport prediction patterns.[123]

However, due to the scale of data, data variety and rapid frequent changes, it is a challenging task to integrate, visualise, analyse and respond to queries. Current data analytics systems provide limited analysis capabilities with long response times of several minutes, which is an impediment for real-time data analytics. Recently, in-memory computing techniques have been found to achieve significantly higher efficiencies, with processing speeds of approximately one second (Sections 4.1.1 and 3.3). Multiple IT firms are actively working in this field, with researchers currently investigating new methods to improve processor speed and responsiveness.

Further, when using Big Data for future transport volume projections, highly specialised and accurately calibrated data mining programs must be used in order to develop accurate and robust projections, because the sheer volume of information available makes analysis difficult.[124] The algorithms and projections developed using Big Data must also be properly calibrated against real-life transport volume scenarios in order to ensure that the projected system performance is sufficiently accurate.

These challenges must be overcome in the future in order for Big Data to be accurately, effectively and efficiently harnessed for congestion management and emergency response. Yet as a whole, what transport planners stand to win is far greater than what they could lose:

- Better run transport systems which best meet the needs of commuters, and

[123] Minelli, M, Chambers, M, & Dhiraj, A 2012, *Wiley CIO: Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley, Somerset, US.

[124] van Oort, N. and Cats, O. 2015, op. cit. DOI 10.1109/ITSC.2015.1

– Removal of key bottlenecks in transport systems: Prediction of when these bottlenecks will occur allows transport planners to devise methods to prevent congestion in these areas; hence, infrastructure investment can be deferred as it is no longer needed so urgently.

5.2 Next Steps

In order to reap the benefits of effective, efficient transport systems, Australia should join the field of Big Data integration, analysis and application. The software is available and there are already existing 'small data' streams from for-purpose sensors (as highlighted in Section 4.4) which can be combined with other data streams (eg. mobile phone location data, Twitter feeds, taxi and bus GPS) in order to reap significant transport organisation benefits. This section outlines several key steps for the future as Australia moves forward towards the Big Data space.

5.2.1 Better Integration

Existing systems: Multiple separate GPS tracking systems, segmented according to the operator's organisation and disseminated by the operator. An Australian example of this is TransPerth's newly released 'real time arrivals' capability, which tracks arrival times of trains and buses[125]; taxi GPS tracking systems also exist but are often shared only with the taxi operator.

Next steps: Integrating each of these disparate interfaces so as to better visualise the location of vehicles and people. This requires gaining access to existing public transport and taxi GPS data. In addition, sensors can be installed to track pedestrian traffic by detecting the presence of nearby smart cards at intersections.

By combining the multiple interfaces, transport authorities can gain a more in depth understanding of existing demand and build better forecasts of future demand, which will greatly aid in transport planning.

5.2.2 Smarter Technology: Vehicle-to-vehicle (V2V) Systems

Existing systems: There is no current legislation in Australia around the integration of V2V on roads.

Next steps: Considering the mandatory implementation of V2V systems in Australia to make use of the 5.9 GHz frequency that has been reserved for ITS. The US is leading the way in this field, with a proposed rule that requires all new light vehicles to have V2V technology from 2021. Currently, the V2V interface sees only cars from the same manufacturer (eg. Mercedes-Benz V2V technologies will only perceive other Mercedes-Benz cars on the road).

[125] Adelaide Metro 2016, 'Real-time passenger information', Government of South Australia, Adelaide.

5.2.3 Specific Privacy Policies for Big Data

Existing systems: Australia’s Privacy Act and its constituent Australian Privacy Principles define the standards and rights in handling and assessing personal information, but are technologically neutral.

Next steps: Because Big Data changes the game in terms of the collection, storage and application of data, specific policies should be implemented (with many such policies under consideration) to handle privacy concerns associated with Big Data, in particular personal information such as mobile phone and private vehicle GPS tracking systems. One possible technique to minimise privacy risks is to use in-memory computing: this system continuously processes raw data and produces analysis results. Hence, only derived trends and patterns are stored, hence reducing the likelihood of sensitive personal information being accessed illegally.

5.2.4 Better Strategy

Existing systems: Transport planning is often based on costly surveys about travel habit and stated preference.

Next steps: Big Data from integrated systems (Section 5.2.1) allow critical transport corridors to be easily identified. Subsequently, the following steps can be undertaken:

- A strategic roll out of more sensors in these areas and surrounding roads, in order to pre-emptively identify congestion before it occurs.
- The development of better, usage-based transport planning by:
    - I. Using systems built from Big Data to simulate changes in public transport arrangements (eg. changing a bus route or changing the frequency of a particular train line) and quantify the effectiveness of the change in reducing congestion across the transport network (see Section 4.2.2).
    - II. Changing public transport routes and/or frequency to meet public demand and reduce pressure on critical transport corridors, hence deferring the need for infrastructure investment.
    - III. Ensuring that any infrastructure investment that is conducted will effectively reduce congestion on critical corridors, meet the needs of commuters and be well-catered to commuters’ demonstrated travel preferences.

5.2.5 Big Data Software Selection

Hadoop/Spark and SAP HANA are both versatile Big Data systems which can be applied to congestion management, each with different strengths and weaknesses (Section 3). Systems like Hadoop must be customised to suit client needs, which may require significant development. There is also a choice between cloud-based or in-house platforms. While cloud-based systems do not have large upfront costs, they mean that hardware and analysis functions are outsourced overseas and privacy issues may occur for sensitive information. On the other hand, in-house platforms can be stored in a secure location in Australia, hence reducing privacy risks; however, these platforms come with high upfront fees and running costs for the electricity required to run the servers.

Hence, a cost analysis must be conducted in order to decide the most applicable software. A thorough analysis requires the consideration of several key factors:

- Length of usage of the software platform (eg. 4 year trial period, 10 year period).
- Amount of data to be processed and how much these data streams will grow over the usage period.
    a. Number of structured and unstructured data sources.
    b. Frequency of sensor readings and the resolution of data processed (eg. GPS location data at 5 minute intervals vs 1 minute intervals).
- Intended purpose of the data (eg. real-time congestion management, demand prediction for transport planning).

Australia’s transport systems have some catching up to do if the nation wants to join the world stage in building effective and efficient transport systems. Harnessing Big Data brings great potential rewards: Big Data provides a wealth of information that can be mined for high-resolution correlations, relations and predictions and can play a pivotal role in the shaping and development of truly smart transport systems catered to the needs of its commuters.