Using big data technologies to support Bus Punctuality Improvement in Scotland.

Tobi Tonner, Steven Revill, UrbanTide

1. Introduction

This paper will set out the award winning work delivered by UrbanTide on the use of big data analytics on bus prediction data and to indicate how this work could be applied in Scotland to provide real benefits to operators and citizens.

1.1 Background and Context

UrbanTide partnered with Datalab to take advantage of the Datalab MSc option where skilled Masters student's are utilised by Companies to develop their technology in return for which they use their work to create their masters dissertation.

This masters dissertation project used our USMART technology stack which is built on a technology stack that has the capability to handle big data in a scalable way and that can provide the analytics that generates additional value from the data.

1.2 Objectives

The business goal of this project was to demonstrate the big data analytics capabilities of the USMART platform and the technology stack that it is built upon. UrbanTide have business reasons to maintain a transport theme and had acquired access to the Transport for London Countdown API which provides a live data feed of the predicted arrival times of the buses in the London bus network. The company was keen to do some analysis of this data feed with the goal of producing evidence of the big data analytics capabilities of USMART and new mobility intelligence. This evidence would be used to support future sales, funding and investment opportunities.

1.3 Scope

1.3.1 In Scope

- Explore and assess usefulness of open data sources for the project.
- Gather data from various data sources for use in the analysis.

- Design and implement a suitable storage pattern for the data that is gathered from the various data sources to be stored within the USMART platform and technology stack.
- Analyse data and assess the adherence of buses to the timetables.
- Analyse data and assess the accuracy of the predicted bus arrival times.
- Analyse Twitter data and assess the bus users' satisfaction of bus services.
- Analyse Twitter data and assess the relationship between the satisfaction of bus users and the adherence to timetables and accuracy of predicted arrival times.
- Design and implement an analytics application within Apache Spark to generate the analysis described above on real-time data feeds.
- Design and implement a suitable data structure for storing the output of the real-time analysis described above.
- Design and implement a publicly accessible API for the presentation of the resulting analysis as a machine-readable open dataset.
- Design and implement a method of visualizing the results of the real-time analysis within the USMART web front-end.

1.3.2 Out of Scope

- USMART application development
- USMART platform deployment

1.4 Achievements

The project set out to demonstrate the big data analytics capabilities of the USMART smart cities platform developed by UrbanTide and it achieved this by employing Apache Spark and the Spark Streaming module to do real-time data analytics on the TfL Countdown API data feed which publishes a live data feed of the predicted arrival times of the buses in the London bus network. This data feed generates about 5 Gbytes of data every day containing 12 to 18 million messages about prediction updates and this qualifies as big data in terms of quantity and velocity.

The project also demonstrated the capability of the USMART platform technology to do batch analytics using Apache Spark and an Apache Cassandra NoSQL database for data storage.

2 Research and Planning

The definition of the project scope as outlined in section 1.3 was established after the research and planning phase of the project described in this section. For that reason some date sources were explored that may not seem completely relevant to the project and that is because we were keeping an open mind about what direction the project could go in and this would depend a lot on what data was available.

2.1 Open and Proprietary Datasets

UrbanTide had no datasets of their own and one of the goals of the project was to gather suitable data from available data sources.

2.1.1 Transport for London

Transport for London (TfL) decided on a strategy to make their data open starting in 2010 with the expectation that it will encourage innovation and with the understanding that building data applications was not their core business **Error! Reference source not found.**.

There are currently two main APIs provided by TfL with different access restrictions.

2.1.1.1 TfL Countdown API

The TfL Countdown API provides a live data feed of the predicted arrival times of the buses in the London bus network and is described in more detail in section **Error! Reference source not found.**. It is free to access but requires pre-authorization with business justification. Prior to the project starting, UrbanTide had obtained authorization and credentials to access the TfL Countdown data feed. This API is the primary data source used in this project.

2.1.1.2 TfL Unified API

The TfL Unified API provides unrestricted access to a large amount of transport-related data for the London area via a single RESTful API. The API includes many datasets including ones related to journey planning, arrival predictions (for a specified stop), bus route data, timetables, underground and road data, and much more **Error! Reference source not found.**.

2.1.2 Twitter

Twitter provide access to a REST API for historical queries and a streaming API for a live data feed. Both of these are freely accessible with API keys that can be generated upon registration. The streaming API could provide data useful for sentiment analysis of the bus network performance in real-time. However the data available is a sample of the full data and

may not include every tweet, some studies show that the percentage of tweets delivered could be as low as 1% **Error! Reference source not found.**. Twitter provide a paid for service operated by GNIP and DataSift called Firehose that does guarantee to include every tweet in the search criteria but due to the cost, this could not be used in the project. However, even if only a sample of tweets was available, the free APIs could still provide useful insights in this project.

2.1.3 Lothian Bus Tracker

Lothian buses provide an API called 'my bus tracker' (www.mybustracker.co.uk) which is also freely available upon request of an API key. This service provides, among other things, estimated bus arrival times on a query by query basis but it does not provide a continuous stream of prediction updates like the TfL Countdown API **Error! Reference source not found.**.

2.1.4 CKAN

CKAN is an open-source open data portal that is being used by a number of cities in the UK as a platform to publish their open data (e.g. www.edinburghopendata.info, data.glasgow.gov.uk). CKAN can be seen as a potential competitor to USMART but it is focused purely on open data publishing whereas USMART has a more specific focus on smart cities and generating value from data.

These city data portals contain a number of open datasets but a large portion of them are simple spreadsheets of aggregated quarterly statistics which are not very useful for the project. Some datasets are links to services such as Edinburgh's 'my bus tracker' service described above.

2.1.5 Traffic Scotland DATEX II

The DATEX II platform is an initiative delivering the European Transport Policy and aims to support the development of information exchange in road traffic management **Error! Reference source not found.**. Traffic Scotland provide access to their DATEX II infrastructure and during the project, a request was submitted and granted for UrbanTide to access the API. The API provides a feed of the 'Variable Message Signs' which are the dot-matrix message boards displayed on roads around the country that display messages such as the journey times to major destinations (e.g. airports), warnings of bad weather, road works and safety messages.

2.1.6 TransportAPI.com

TransportAPI.com is a transport data platform for the UK that provides a variety of transport-related information via a paid-for API service such as live departure and arrival times, journey planning and performance indicators **Error! Reference source not found.**. A particularly interesting service that could be of use in this project is 'Tweet Mapping' which provides sentiment analysis and traffic-related content tweeted by commuters.

2.1.7 Met Office DataPoint.

The Met Office offer a new free API accessible with an API key upon registration that can provide weather-related data that could be of interest in this project (http://www.metoffice.gov.uk/datapoint). This service is new and is in beta stages. It is understood that weather conditions have a significant influence on traffic.

2.1.8 Google APIs

Google make it their business to gather data and they offer a wide range of APIs to share this data. Some APIs are free for small usage in applications and websites that offer free public services, and they also offer a tiered price structure for medium to large usage **Error! Reference source not found.**. One service that would be very useful but doesn't seem to be available would be one that provides the origin and destination of people travelling, such as commuters.

2.2 Potential Issues with Twitter Analysis

Around the start of the project, UrbanTide had made contact with Jorge Gonzalez, a PhD research student at the University of Glasgow who was doing a project at the Urban Big Data Centre researching the use of Twitter to identify traffic-related incidents. As there were potential synergies with this project, we setup a meeting to discuss some ideas and gained the following insights:

- Identifying the topic of the tweet as bus-network related would be challenging and one approach could be linear discriminant analysis. Another simpler approach would be the bag-of-words method and building a model would be an iterative process requiring refinement.
- Because no dataset was available, we would be required to capture Twitter feeds and the fact that the Twitter data feed is only a sample of data could cause issues as no indication is given for the method used to create the sample and this could change over time without warning affecting the model.

- The difficulties could make this more than a complete dissertation project in itself and there wouldn't be enough time to gather data and do analysis in the limited time that was available.

The outcome of this meeting was a decision to drop the Twitter sentiment analysis from the scope of the project due to time constraints.

2.3 Project Selection

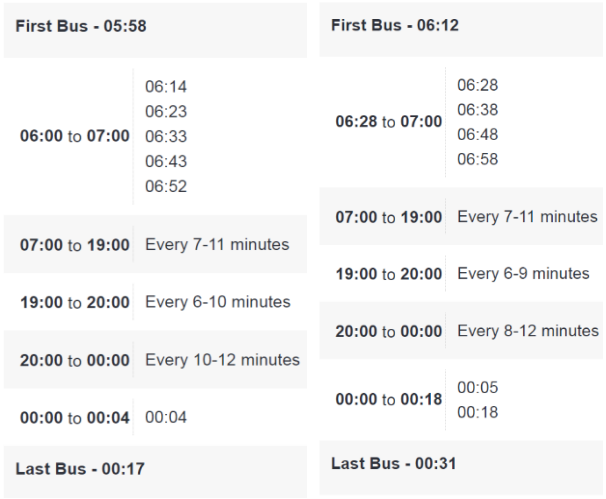
After the research phase and exploring alternative ideas in a company-wide brainstorming session, we decided to revert to the original plan to perform analysis of the TfL Countdown API data feed. UrbanTide had a business reason to maintain a transport theme for the project and wished to make use of their access to the TfL Countdown API. The data feed provided a large volume of data (about 5 Gbytes of data per day with 12 to 18 million messages over the API stream) and would be a good dataset to prove the capabilities of the system to cope with big data in terms of volume and velocity. The ability to effectively analyse this type of data would be useful for the business because it is expected that there will be many project opportunities in the future doing analysis of streaming data from IoT sensors in cities and this project would provide the proof-of-concept, experience and lessons learned to make these future projects successful and this would assist greatly in tender and bid processes.

2.4 Data Analysis

This section describes the analysis performed to answer the question: *how well do the buses adhere to their timetables?*

2.4.1 Background on London Bus Timetables

The bus timetables in London are quite intricate. The timetables don't state specific times when the bus will arrive at a bus stop, instead they indicate a wait time (i.e. a time gap) between the buses. Moreover, the wait time varies through the time of day and is not necessarily the same for every stop on the same bus route. In addition to that, the times of day when the wait times change are not the same for every bus stop on the same bus route. There are three different timetables for 1: Mondays to Fridays, 2: Saturdays and 3: Sundays. Wait times are provided as a minimum and maximum expected wait time, e.g. 10 to 15 minutes.



First Bus - 05:58	
06:00 to **07:00**	06:14 06:23 06:33 06:43 06:52
07:00 to **19:00**	Every 7-11 minutes
19:00 to **20:00**	Every 6-10 minutes
20:00 to **00:00**	Every 10-12 minutes
00:00 to **00:04**	00:04
Last Bus - 00:17	

First Bus - 06:12	
06:28 to **07:00**	06:28 06:38 06:48 06:58
07:00 to **19:00**	Every 7-11 minutes
19:00 to **20:00**	Every 6-9 minutes
20:00 to **00:00**	Every 8-12 minutes
00:00 to **00:18**	00:05 00:18
Last Bus - 00:31	

Timetable for Bus Route 1 towards 'Canada Water Bus Station' on Mon-Fri (as of 17th Aug 2016). Left is timetable for bus stop 'Elephant & Castle / New Kent Road' and right is timetable for bus stop 'Surrey Quays Station'

It is expected that this style of timetable with different time gaps for different stops has evolved from historical experience and from the practicalities of buses travelling through levels of congestion that vary with the time of day and with week days versus weekend days. It may also be necessary to increase the quantity and frequency of buses during periods of high demand to meet capacity.

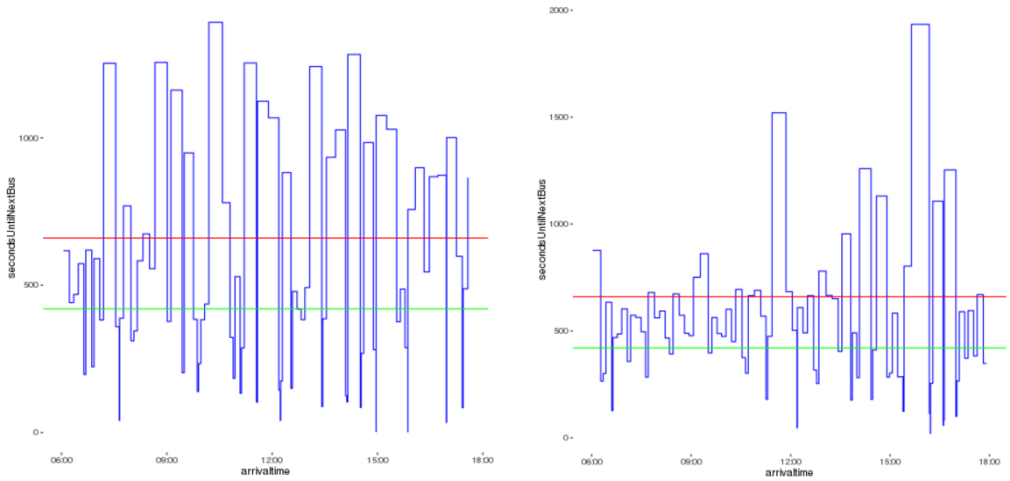
2.4.2 A Measure for Bus Performance

Given the way that the timetables are structured, the most relevant way to measure how well the buses adhere to their timetables is to analyse the gaps between the buses to see if they were within the timetabled wait time. Because the timetables are so intricate, this is difficult to achieve over the bus system as a whole but it is relatively straight forward to measure the performance of a particular bus route at a particular stop on that route if the time range is restricted to a period where the timetabled wait time is consistent.

2.4.3 Analysing Gaps between Buses and Comparing with Timetable

A single bus stop on a route was selected at random from the data to explore the length of the gaps between the buses. The stop chosen is the one described in section 2.4.1 (bus route 1 towards 'Canada Water Bus Station' on Mon-Fri for bus stop 'Elephant & Castle / New Kent Road') during peak time 7 am to 7 pm BST (i.e. 6 am to 6 pm GMT) on 27th July 2016. During this time, the buses on that bus route are scheduled to arrive at that bus stop every 7 to 11 minutes.

To analyse the gaps between the buses over the day, the data was plotted on a stepped line chart where the x-axis represents the time of day (increasing from left to right) and the y-axis represents the time gap between the bus arriving immediately before and the bus arriving immediately after that time of day. A green horizontal bar depicts the point where the gap is 7 minutes and a red bar where the gap is 11 minutes. See figures below.



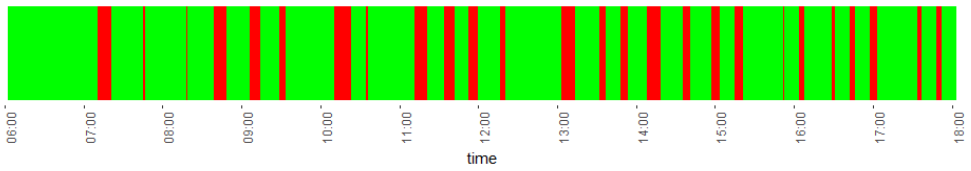
Plot of bus gaps at a single bus stop of a bus route. Left 27th July 2017, Right 3rd August 2017

The plot shows that there are a large number of gaps between buses that are above the red line and longer than the timetabled wait time for that bus route at that stop. There are also a large number of gaps that are very short showing the 'bus bunching' problem where a person has to wait for a bus for a long time and then several arrive in quick succession. It is somewhat surprising how few of the bus gaps are within the timetabled wait time of 7 to 11 minutes (only 20% on 27th July and 34% on 3rd August).

2.4.4 Visualization for Bus Gap Analysis

Plotting the time gaps between buses on the stepped line graph shows the time gaps clearly for a single stop. To visualize the gaps for all of the stops on a bus route, an array of these graphs can be created (see Appendix 3). However this doesn't work very well to quickly see the patterns at a glance. Therefore a different type of plot was made by creating a result set that contains a data point for every minute of the day and a discrete value to indicate whether the next will arrive within a specified wait time or not. This data is then plotted on a bar chart with the vertical bars coloured according to the discrete value just described with red indicating long wait times and green indicating short wait times. So if a person arrives at the bus stop during a red slice, they will be waiting a long time for the next bus. The result is a plot that

looks somewhat like a red and green barcode. This allows the data to be easily visualized at a glance for multiple bus stops (see Appendix 4). A single bus stop is shown in the figure below.



A 'barcode-style' bar chart showing times of day when a passenger will need to wait a long time for a bus (red) for the 27th July 2016.

There may be a name for this type of graph and it might already exist as a concept but it wasn't known of nor was it found in research so it might be an innovation created in this project.

It should be noted that this visualization does not show when the buses are arriving too close together, where the wait time between buses is shorter than the timetabled minimum wait time. From the perspective of passenger experience, these occasions are not so relevant. It should also be noted that, when plotting the data for all of the buses on a bus route, because the timetabled wait time varies on a per bus stop basis, the visualization doesn't show how the wait times compare directly with the timetables. Instead the visualization shows how the wait time between the buses compares with a static, standard wait time.

2.5 Implementation of Data Product – Real-Time Performance Monitoring of Arrival Time Predictions

The primary business goal for the project is to demonstrate that the USMART application has the capability to do analytics with big data. The real-time nature of the TfL Countdown data feed that we had access to provided an opportunity to do real-time analytics, moreover the Apache Spark technology provides the real-time processing capability via the Spark Streaming module, so doing real-time analytics was a good opportunity to create something that would help the USMART product to stand out against its competitors during demonstrations to potential clients. With velocity being one of the V's of big data and the added value that real-time analytics brings by generating insights immediately instead of during overnight batch processing, this was something that got a lot of focus during the project.

To make this component effective as a good component to demonstrate, it was important to have a good visual impact and perhaps a level of interactivity with the users. Because the data source was the output of the prediction system for bus arrival times, it was decided that the real-time monitoring would display in a visual format some analysis of the performance of the prediction system itself.

2.5.1 How to Quantify the Performance of the Prediction System?

It is reasonable to propose that a system that predicts the arrival times of buses performs well if it accurately predicts the arrival times of the buses. Moreover, the predictions should be consistent.

From a passenger's perspective, when the message system displays a message that the bus will arrive at a specified time, that insight should be actionable for the passenger. For example the passenger might decide to get a different bus to get to their destination sooner, when taking into account the journey times of each bus route and the time remaining until each of the buses arrives; or perhaps the passenger might decide that they have enough time to do something like go to a shop and return before their bus will arrive.

If the prediction system works well and the predicted arrival times are accurate then passengers will get a good experience and can trust the predictions and make decision based on them. If on the other hand the prediction system is not consistent, i.e. one minute it predicts that the bus will arrive in 5 minutes and then the next minute it predicts that the bus will arrive in 15 minutes then these predictions are less useful to the passenger and the passenger cannot depend on them so much to make their decisions.

For the reasons described above, the method used to measure the performance of the prediction system is to measure the spread of arrival time predictions for each particular bus arrival at each bus stop on its route. Specifically this is measured by taking the latest (not necessarily the most recent) predicted arrival time and subtracting the earliest predicted arrival time to find the number of seconds between them. The smaller the spread in the predictions, the better the performance of the prediction system (from the passenger's perspective).

It should be noted that this method of measuring the prediction system is purely from the passenger's perspective. It does not measure the quality of the prediction system itself. It is expected that there will be many factors that affect the bus arrival times and that these factors will change over time as the bus travels on its route – it is not expected that the bus arrival times are in reality 100% predictable. If the prediction system did not update its predictions according to real-time information about the bus locations and traffic patterns, then it would be an inferior system.

2.5.2 Real-Time Prediction Performance Visualization

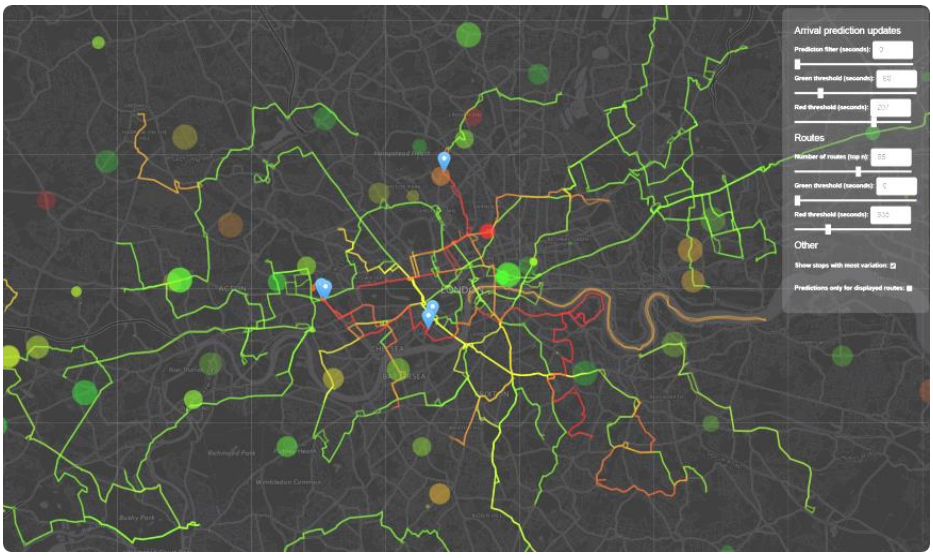
The real-time system that was developed in this project displayed animated overlays on a map of London. When an update to a predicted arrival time was issued by the TfL Countdown data feed, the difference between the new predicted arrival time and the previous predicted arrival time (if there was one) was calculated – this was called the delta. For the visualization output, a coloured circle was briefly overlaid on the map geo-located to the position of the bus stop that the prediction was issued for. The colour of the circle is a gradient between green and

red and this indicated the delta. The values mapping the delta to the colour are configurable by the user using sliders on the user interface and the concept is that a green circle shows a small change to the predicted arrival time of a bus at that bus stop and a red circle depicts a large change to the predicted arrival time. This gives the user a feel for how steady or fluctuating the predictions are in general and where the larger fluctuations are located on the map.

In addition to the coloured circles indicating the deltas in predicted arrival times of buses at the bus stops, the spread of each predicted bus arrival time was calculated and the spreads on each route were averaged over a 15 minute moving window to generate a performance metric for the bus route as a whole. The user can select the number of routes to display on the map and the top *n* worst performing routes are displayed with their paths overlaid on the map, again coloured from green to red according to their performance in a user-configurable manner.

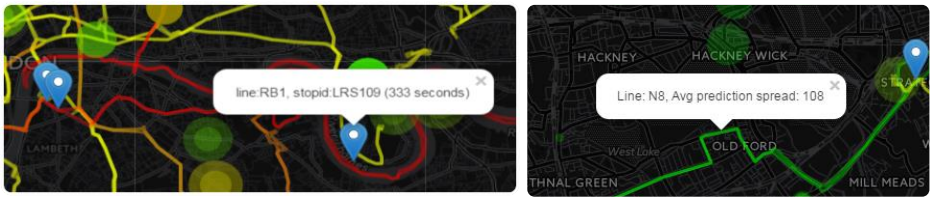
The last addition to the map is to display markers for the worst 5 bus stops (worst being the one with the largest spread of predicted arrival time).

See figure below for a screenshot of the user interface of real-time analytics of the performance of the TfL Countdown bus arrival prediction system.



Screenshot of user interface of real-time analytics on bus arrival prediction performance.

To provide further information, a user can click on a bus route or a bus stop marker to get more details about it such as the route name and the value of the performance metric.



Balloon popups shows details of bus stop (left) or bus route (right) when clicked.

3 Project Conclusion

3.1 Summary

The project set out to demonstrate that the USMART platform and the technology stack that it is built upon has the capability to deliver big data analytics in the smart cities domain. Using the live data feed of bus arrival-time predictions in the London Bus network, and processing this data in Apache Spark in batch, and in real-time with the Spark Streaming module, we demonstrated that the solution is capable of both batch and real-time analytics of big data. We did some analysis of the gaps between buses measuring the performance of the buses against their timetables and measuring user experience of buses in terms of the wait time. We also created an interactive map dashboard to visualize in near-time the variations and aggregated statistics of the arrival-time prediction system in the London bus network.

The analysis of the bus gaps reveals that there are many instances through the day when a passenger will have to wait for longer than the timetabled wait time for their bus and that the problem of buses 'bunching up' where several buses arrive in quick succession with long gaps in between is prevalent in the system. However it is important to understand how these findings can be positively used to improve mobility in cities:

- Despite bus operation managers deliberately slowing or speeding up buses to attempt to maintain even gaps it is clear that the manual process is not achieving the desired results
- It is felt that there is opportunity to bring in other big data sources to potentially improve the real time operation of the buses to maintain gaps. This could include
    - Live traffic data on conditions of the road network on both bus lanes and normal traffic
    - Real time road works information as one line in particular suffered significant impacts to the bus gaps due to road works
    - Improvements to bus arrival information
    - Improved customer experience due to the accuracy of timetables
    - Increased patronage as a result of improved customer experience
- The absence of recorded arrival times, all analysis using the assumed arrival times imputed from the last received predicted arrival time could be dubious and this is an important point to acknowledge, should also be provided as open data to help improve accuracy.

The real-time map dashboard visualization showing analytics of the arrival-time predictions can be demonstrated to potential clients of, and investors in, USMART to demonstrate, in a

visually appealing way, that any sensor data gathered in cities can be analysed in real-time and visualized, revealing patterns and insights to users – this can be particularly valuable with the expected proliferation of IoT devices that gather and stream many types of real-time data from around a city (such as air pollution levels).

The opportunity for Scotland

It is the belief of UrbanTide that the findings from this research project can be applied in a number of ways to improve mobility and mobility services in Scotland.

Transport Scotland sets the national policy framework on buses which is delivered by bus operators, local authorities, Regional Transport Partnerships and the regulatory authorities. They also published a suite of best practice guidance documents to assist in the provision of bus services. The bus policy aims most relevant to this project are:

- To enable bus to provide an effective alternative to the car by improving reliability, average bus speed and encouraging improvements to the quality of services and infrastructure
- To link communities, people, places of business and employment and essential services through encouraging the maintenance and development of the bus network in Scotland

Specifically, Transport Scotland sets out Guidance for local authorities and bus operators on the creation of Bus Punctuality Improvement Partnerships (BPIPs) outlining considerations in bus monitoring and data sharing, use and publication. The document goes into detail on the opportunity of data sharing and data publication in general but makes no mention of the opportunity of big data technologies as such or indeed open data.

This paper argues that new big data technologies could significantly improve understanding of bus performance and lead to improvements for passengers. And although the BPIP notes that reliability is one of the principal factors cited by non bus users as being necessary to achieve modal shift UrbanTide believe there are further opportunities not being realised. If these were explored it could arguably not only sustain but enhance passenger levels and the ongoing delivery of a quality public transport system for Scotland. Punctual and reliable bus services are essential elements that should not be underestimated when considering how the overall service can be improved.

This paper believes that properly analysing all the available bus data provides opportunity for improvements in reliability and therefore in helping to support modal shift.

The role of Open Data

Over and above the impact detailed analytics may have on improving services UrbanTide recognises that the value in the “open” provision of data using an open licence in itself has the potential to provide value and improve services.

UrbanTide propose that bus operating conditions could be altered in the future to include open data as a component to support future development and tie in with the Scottish Government’s

Pioneer Status in Open Governance. These findings could also be used in informing updating of bus policy in Scotland.