IET Sectors
The Institution of Engineering and Technology

Transport

Big Data in Transport

Data is increasingly significant in the management and use of transport systems. This Insight will explore big data management, best practice and consider the challenges and developments ahead for those responsible for big data in a transport environment.

1. Introduction

Data sets can be so large and complex that they become difficult to process using traditional data processing applications and existing data management tools. As a result, capturing, storing, searching, sharing, transferring and analysing the data sets can be a significant challenge.

The trend to larger data sets is a result of the plethora of information that can now be derived from the analysis of a single large set of related data. Increasingly, data is gathered by information-sensing mobile devices, remote sensing, software logs, cameras, microphones and wireless sensor networks. Global technological information per-capita capacity has approximately doubled every 40 months since the 1980s. Some predictions show that data production will be 44 times greater in 2020 than it was in 2009.

In transport, the increase in data is manifest in the availability of traffic information, particularly through sat nav applications. Similar passenger information applications provide departure information for public transport users. Payment for transport (ticketing and tolling) is increasingly reliant on data-dependent technology, applications and services.

In all modes of transport, there is now a considerable quantity and diversity of data available for operators to improve performance, efficiency, service provision, safety and security. Data also enables operators to manage demand conflicts, customer service, environmental impacts and innovation. This can be seen in traffic signal co-ordination, trains reporting track defects, on-line flight check-ins and cargo tracking.

So data is increasingly significant in the management and use of transport systems. It is important to identify the needs for data and its capabilities and constraints. This will help to determine the impact of big data on transport and the innovations that are expected or necessary (such as the internet of things).

William Perugini / Shutterstock.com

1.1. Definition

Big data uses data sets with sizes beyond the capability of traditionally-used software applications to capture, store, manage and process data within acceptable time frames. A single big data set size can range from a few dozen terabytes to several petabytes. Gartner research defined big data challenges and opportunities as being three-dimensional (3Vs model):

- Volume (increasing amount of data)
- Velocity (speed of data in and out)
- Variety (range of data types and their sources).

New forms of processing are required to enable enhanced insight, decision making and process optimisation to address the characteristics of 3Vs. In transport, the volume of data has increased because of growth in the amount of traffic (all modes) and detectors. Also, travellers, goods and vehicles generate more data from mobile devices and tracking transponders (including trains, ships and aircraft). Infrastructure, environmental and meteorological monitoring also produces data that is related to transport operations and users.

The velocity of data has increased in transport due to improved communications technology and media (particularly fibre optic cabling) and increased processing power and speed for monitoring and processing. Some applications have experienced a step change in data velocity as technology has changed. For example, ticketing and tolling transactions that use smart cards or tags are now immediately reported, whereas paper-based ticketing depends on human processing to acquire data from the transactions.

The variety of transport-related data has increased significantly. Modern trains and aircraft report internal system telemetry in real time from anywhere in the world and it is possible to acquire information about all crew members and passengers. In past years, only static data might have been available about the rolling stock, aircraft and crew and once outside human monitoring (e.g. signalman observations or radar range), its progress or position was unknown.

2. Big Data Management

2.1. General

The big data process includes data acquisition, processing, aggregation and delivery. Data acquisition in transport relates to the collection of a high volume of data from specific data sources e.g. presence detection data, tolling and passenger transaction data. Data acquisition in a big data environment is characterised by a high volume of semi/unstructured raw data ready for processing (e.g. traffic speed).

Data processing involves cleansing (e.g. anonymization), the application of unique IDs to records and identification of errors. Clean data from multiple data sources is then made available for aggregation.

Big data aggregation is achieved by organising and processing data from an unstructured to a structured state. For example, vehicle presence detections are used to establish characteristics of traffic, such as flow or occupancy, which is used to establish congestion or delay data. Or train departure data is used to predict delays. Aggregated data may or may not be moved from its original location. Data sets may be aggregated into one big data set, which can then be processed using intensive analytics to identify relations, trends and insight. This is then available for analysis and dissemination.

Information delivery uses advanced statistics and data modelling to join disparate data so that it is organised for presentation to end users. For a rail freight operator, this might be a dashboard of progress, performance and predictions. For a bus passenger, it might be an expected arrival time for a service at a stop.

2.2 Data Acquisition

The infrastructure that is required to support big data acquisition needs to handle high volumes of transactions and to deliver low, predictable latency capture and processing. Mobile network providers depend on significant numbers of base stations to meet the service levels demanded by their customers. Transport network managers use mobile networks to acquire and deliver data. Transport operators also use bespoke communications services such as fibre optic networks to ensure they can have immediate access to traffic and travel data and so they can regulate and control traffic flows for normal and incident operations. This applies to all modes.

The infrastructure also needs to support flexible and dynamic data structures. This means that all data that is available relating to a data item needs to be transmitted and stored, even if it is not used by its primary application. For example, all elements of a ticketing transaction need to be transmitted and stored so that it can be used effectively, even if its initial ticketing application only needs to associate one barrier passage with a user ID on a database.

Data reliability is dependent on accuracy and precision (for measured quantities and associated metadata, such as time stamping). This means that the resolution of measured quantities needs to be as high as possible and comms error rates need to be low. Operational demands for safety drive precision and accuracy in monitoring and control in all modes. This encourages the use of redundancy and multiple data sources. Safety is not compromised in the event of failures, but service levels often suffer. So, failures to detect train movements or control railway signals result in service disruptions, rather than higher accident rates. These requirements demand high data rates and fast processing, which necessitates ever greater investment in detection, communications and processing infrastructure.

pisaphotography / Shutterstock.com

It is notable that as well as increasing the demands on infrastructure provision directly, the need to monitor, transmit, process, and store all data elements also increases the need to manage data privacy and security, which has an impact on data infrastructure provision.

This means that if an in-house storage solution is adopted by a transport service provider, it requires significant capital expenditure on data storage. Cloud storage provides an alternative, with the capital and operational risk transferred to a third party, but with higher operational costs manifest in service charges. Similarly, communications services that support transport operators used to be implemented and delivered by the operators themselves. This is less common now, with collaborative and third party provision becoming more popular.

An Australian analysis of sensor options provides an indicator of how to match functions to detectors:

- Inductive loops (presence, count and speed)
- Piezo-electric strips (counts, pressure, speed)
- Pneumatic tubes (counts, speed)
- Cameras (counts, classification, speed, presence)
- Infrared sensors (counts, speed, classification)
- Passive acoustic (counts, speed)
- Microwave (counts, speed, presence)
- RFID (presence, counts, classification).

In the maritime arena, big data and analytics has been identified in a recent report addressing its application in commercial shipping and naval applications. It recognises the proliferation of big data solutions enabled by wireless communications, novel sensor technologies and the creation of ad hoc networks, with widespread applications including, meteorological oceanographic, traffic data, material and machinery performance data, cargo data and accident data.

Mobile-sourced data also provides data acquisition opportunities, but with a different set of performance challenges to traditional detectors:

- GPS/mobile (speed, presence, count)
- Bluetooth (speed, presence, count).

As part of a Technology Strategy Board study with Deloitte, Imperial College London and INRIX, Transport for London (TfL) also compared three datasets from existing detector sources (off-call mobile phone data provided by INRIX, ANPR2 journey time data provided by the TfL LCAP3 system and TfL iBus bus journey time data, based on GPS vehicle tracking). It was found that mobile phone-sourced data quality depends on the context (e.g. time of day, type of user and speed).

The exploitation of mobile data to infer traffic flows in urban environments is limited by its lack of flexibility in measuring flows on-demand on a specific path, but it might be aggregated with other sources to improve its performance. Also, mobile technology is subject to potential biases (for example, some age or social groups might have a greater tendency to use bicycles or trains instead of road-based motor vehicles). Unknown vehicle occupancy also increases the level of uncertainty when sourcing traffic flow data from mobile technology.

The *Global Marine Technology Trends (GMTT) 2030* report highlights the need to integrate a range of emerging technologies as a critical factor in developing robust, reliable and efficient solutions to exploit data from a wide range of sources in varying and constantly changing structures and architectures. This includes the need for robust, high bandwidth, secure communications supported by sophisticated analytics to augment highly skilled operators.

Connected vehicles developments are expected to promote changes in the way data is acquired for highway applications. The timeliness, availability and accuracy of data will be much higher than for existing techniques and it can be expected to contain more telemetry, rather than alerts. However, its value is dependent on penetration rates, which are currently low, but expected to increase.

This illustrates the challenges in using a wide variety of data sources. It is essential to know and understand the quality of the data available. It highlights that big data processing and aggregation needs to apply a data quality assessment, otherwise outputs will be inaccurate, so users' decisions will carry a higher risk that their needs will not be met and data providers' reputations will suffer accordingly.

Transport network and service operators need to manage data to ensure it is available, reliable, accurate and true. This means that relevant data standards should be established and incoming data needs to be monitored, controlled and refined. For example, specific data quality requirements for a transport network operator might include:

- Spatial granularity (line, station, urban road network, link, junction etc)
- Temporal granularity (minutes, hours days, annual, etc)
- Direction discrimination
- Modal discrimination
- Sample size within provided spatial and temporal quantities
- Bias (free or bias).

Other sources of data that are expected to contribute to big data in transport include live feeds from social media (e.g. Twitter – particularly for public transport), traffic data and weather. Many historic data sets are becoming available. These can be analysed to identify performance of particular train services, for example, over the preceding 6 months, which can be used to increase the confidence of customers or the speed of operational decision making.

Big data is expected to play an increasing role for transport infrastructure owners and operators in managing their assets. BIM (Buildings Information Modelling) generates asset information as soon as it is designed. Maintenance and service functions add to that data so that infrastructure owners are able to develop a clear picture about the state of their assets, how they need to be managed and what resources might be needed to preserve their capability.

2.3 Data Processing

The value of big data increases as latency decreases, i.e. the faster data is delivered, the more value it provides to users. Performance improvements lead to qualitatively better analysis outputs (e.g. closer to real-time). The challenge, then, is to deliver data as fast as possible.

This is supported by standardisation and DATEX II provides a set of specifications for exchange of traffic information in a standard format between separate systems. DATEXII is a structured data model that utilises UML, is platform independent and seeks to harmonise the exchange of traffic and travel information across the EU. Processing data within these standards is the challenge.

Data standards are also developing for maritime application. The Automatic Identification System (AIS) is an automatic tracking system used on ships and by Vessel Traffic Services (VTS) for identifying and locating vessels by electronically exchanging data with other nearby ships, AIS base stations and satellites. The National Marine Electronics Association (NMEA) standard uses two primary sentences for AIS data to receive data from other vessels and for own vessel's information.

Transport operators need to consider if current server storage has sufficient capacity to handle data within desired time parameters. Requirements for time parameters, storage solutions etc. will need to be considered and assessed. Cloud computing solutions may be a viable option, subject to careful feasibility analysis.

2.4 Aggregation

The infrastructure required for organising big data must be able to manipulate and process data at the original storage location and manage high throughput as part of the big data processing step in addition to being able to handle data of varying types and converting unstructured data to structured data.

Transport service operators cannot always extrapolate meaningful outputs from original source data (e.g. mobile phone data) because of lack of expertise or investment in systems. Third party intervention is available to process data into a meaningful and usable format. For example, sample bias can inhibit analysis and mobile data might not be representative of the travelling population and additional analysis and aggregation with other data sets might be necessary to create useful inputs for operators to use. As a result, transport operators will become more reliant upon third parties to process and aggregate the data necessary for their own analysis and delivery.

Careful management is needed, therefore, to ensure data quality is maintained. This is likely to benefit from a collaborative approach with reciprocal arrangements in place for the two way exchange of data. Third party service providers can process data to a required quality performance level in exchange for free access to input data so that they can add value for subscription customers.

The accuracy of data is likely to be an increasingly significant factor in data quality so that the value of information in a big data solution can be enhanced. This is particularly relevant in predictive applications and it can be difficult to achieve. The challenge for transport service providers is that when travellers are presented with forecasts about their journey, they expect them to be fulfilled. However, the only certainties that transport operators can offer are records of past events.

For example, journey time postings on motorways are based on the measured journey times of recent travellers. If this is perceived as travel time prediction, it will be correct as long as the traffic and highway conditions do not change. Similar challenges exist in all public transport operations. Big data can provide more realistic predictions by comparing current conditions with historic data and by assigning confidence levels and tolerances to predictions.

The presentation of predictive information to travellers raises more challenges. For example, travellers might rely on variable message signs, real time passenger information or platform displays when there are no disruptions, but a mobile solution might be more appropriate for dissemination of disruption information.

If appropriate standards are used for data exchange (such as DATEXII), meta data will be available that can be used to speed up searches. For example, timestamping can be used to filter historic data according to day, date or age, which enhances the quality of predictions for traffic information. It also speeds up batch processing by narrowing down the data set for analysis. This is important for some operational applications, such as incident detection, where confidence levels can be raised by rapid analysis of multiple data sets using narrow time and location parameters.

2.5 Information Delivery

The value of big data needs to be challenged because big data analysis (e.g. fusion and mining) might not produce the 'truth'. Analysis could identify patterns where none exist because they might emerge if data is analysed for long enough. Conversely, trends can be lost when data is combined. This indicates that skills and expertise are likely to be important in big data processing.

If there is no understanding of context, it can be lost within a big data set. Diverting motorists to switch modes and catch a train to avoid congestion will be fruitless if there are no parking spaces at the station. Finding ways to convert information into simple messages can be a significant challenge, particularly if the output media have constraints (for example, Variable Message Signs - VMS).

Also, it is important to ensure a single source of truth and that third party users do not corrupt data and misuse it.

The challenge of maintaining control of the data can be achieved with appropriate agreements, which underlines the need to work collaboratively with third parties (such as mobile apps and traffic information service providers) in what is essentially an un-regulated market.

Organisations such as TfL, Network Rail and Highways England have pursued an open data approach, making data and reports freely available to third parties and the public. This commonly involves removal of restrictions on commercial usage of data in a bid to increase information availability and dissemination. TfL provides Application Programme Interface (API) for web and app developers for journey planning, live travel disruptions and underground and bus service information.

Transport operators need to ensure that a single source of truth can still be maintained if big data is made freely available. They also need to ensure that third parties do not reduce the quality of this data and that it is used in pursuance of its transport obligations.

3. Best Practice

3.1 Data Acquisition

SQL databases cannot be used for collection and storage due to the variety of data formats stored within big data sets. This means that big data needs to be processed by converting data from an unstructured to a structured form, or by using approaches that do not rely on relational databases, such as NoSQL .

Using a systems engineering approach, data can be seen in the context of business requirements driving operational requirements. User needs can be viewed in the context of an architecture that is governed by standards. The choice of data sources should be driven by business needs. For example, Highways England's business objectives are driven by safety, performance and customer service. Data from traffic and environmental sensors and CCTV cameras is used to meet operational objectives for the levels, severity and resolution of accident, congestion, journey time reliability and traveller information.

The Digital Railway concept is building on the principles of systems engineering to improve performance and increase efficiency, with expectations that it will support future integration to achieve multimodal travel, bringing together airlines, buses, trams and taxis.

For urban highways, Urban Traffic Management and Control (UTMC) is likely to be a useful way of obtaining data from a range of compatible systems and equipment (e.g. VMS and ANPR). Highways England is implementing a new control system for motorways (CHARM), which is primarily intended for operational control but it will use significant data input from traffic detection, which will be used in parallel by the National Traffic Information Service (NTIS). Developments in mobile data processing provide new opportunities for data acquisition (for example, TomTom®).

Automotive suppliers provide services that acquire data from the vehicles they have supplied whilst their customers are using them. This is subsequently processed to provide useful information for all of their customers.

For example, some Jaguar Landrover (JLR) vehicles collect data about problems in the highway surface (e.g. potholes), which is subsequently used to create warnings for other JLR drivers. Ideally, this data should be available to highway authorities for asset management purposes and to other drivers as part of traffic information and JLR is working to that end.

3.2 Data Processing

Building a legacy big data environment should be avoided because of the risk of potential disruptive changes such as new data types, hardware and programming approaches. This means that standards and commercial off the shelf (COTS) solutions should be used wherever possible.

DATEX II is a European data interoperability standard that includes a range of data types and delivery mechanisms associated with traffic and travel information delivery and exchange.

It is now adopted by Highways England as its primary information protocol. DATEXII is a structured data model that utilises UML, is platform independent and seeks to harmonise the exchange of traffic and travel information across the EU. Highways England believes adoption of DATEX II will improve data quality and timeliness.

3.3 Aggregation

Packages are available that provide an open-source software framework designed for large-scale processing and storage of data sets on clusters of commodity hardware (e.g. Hadoop, Apache Spark). This allows big data to be processed and organised whilst data is stored on an original database.

Once data has been processed and aggregated, the aggregated data set can provide multiple reporting processes or reports with a source of data i.e. one big data set can be reused for multiple activities.

TomTom® uses data aggregated from its subscribers and inputs from network operators to provide traffic information. Also, Twitter® feeds can be monitored to identify hot spots of activity, which might indicate public transport delays. The growth of non-relational database processing is likely to continue because it brings unstructured data sources into big data solutions.

3.4 Information Delivery

UTMC enables traffic management applications to share and communicate information amongst themselves e.g. VMS with ANPR.

DATEX II is a structured data model that utilises UML, is platform independent and seeks to harmonise the exchange of traffic and travel information across the EU. Transxchange is a Department for Transport (DfT)-sponsored national standard for bus information exchange with other systems and SIRI ((Service Interface for Real Time Information) is an EU standard for exchange of current, planned and predicted real time public transport information between systems).

4. Big Data Challenges and Developments

4.1 Data Access

Network operators want travel information to be distributed effectively so that travellers can make effective journey decisions. This is commonly achieved by making data freely available to third parties for processing and onward dissemination. The commercial models that support the third party providers involve app purchases, subscriptions or advertising revenue.

4.2 Data Quality

Whilst travellers perceive a benefit, the business model is sustainable, but any disruptions or changes in perceived benefits might challenge revenue streams. For example, if all motorists are advised to take an alternative route to avoid an incident, significant delays might ensue. Similarly, if motorists are advised to change modes and take the train, the credibility of the advice will be undermined if there is no parking availability at the station and all the trains are late or full.

This means that even without external business disruptions, it is important to maintain and improve the quality of data and associated advice.

4.3 Privacy

Perceptions of privacy are also likely to influence the value of big data, particularly where it relates to the use of personal data derived from mobile or ANPR sources. This can be anonymised so that there is no risk to privacy, but the perception that authorities are tracking individuals is likely to continue. This puts at risk the ability to acquire data because individuals will not trust authorities or app suppliers.

Geographic location is associated with a device through a relevant identification (e.g. GPS coordinates, Internet Protocol (IP) address, RFID, or Wi-Fi positioning system). For safety applications (such as eCall), regulations have been developed to protect privacy.

It is notable that individuals are prepared to share their location data if they perceive a clear benefit in doing so, which explains why, for example, fitness apps are popular. The challenge of this approach for big data is that inputs to mobile-sourced data might not be representative because subscribers will have particular motivations for permitting access to their location data. Demographic attitudes towards sharing personal data might also favour younger generations.

Principles of proportionality and minimisation, with transparent processes, policies and strategies are likely to be important in retaining user confidence. Ideally, this should be publicised alongside the benefits in sharing data.

4.4 Interoperability

Big data solutions are more likely to succeed if the data is interoperable, enabling systems to process data from any source. The key to integration is standardisation and an open architecture. Regulation is needed to deliver this when the market cannot.

Standardisation through regulation can be seen in the application of Telematics Applications for Passenger Services Technical Specifications for Interoperability (TAP TSI). This defines European-wide procedures and interfaces between all types of railway stakeholders. TAP TSI supports interoperable and cost-efficient information exchange for high quality journey information and ticketing. Similarly, specifications have been adopted to improve interoperability of real-time highway status and traffic data to be made accessible in a standardised format (DATEX II) as part of the ITS Directive .

4.5 Business Case

Mobile data provides significant opportunities for big data deployment. However, rapid development in technology and services creates uncertainty for big data investment. Business cases could be undermined if solutions are obsolete before they reach maturity, so flexibility needs to be built into the delivery.

For example, cloud-based deployment provides some protection against obsolescence by partially offsetting capital risk against ongoing service costs.

4.6 Skills

Big data is likely to affect skill sets in the transport industry in the future. As operations become more complex, the drive for improvements in services and efficiencies can be expected to increase the dependence on systems and data. Over time, system processes will develop to perform better than their human counterparts in such a scenario, which will reduce network manager's reliance on operators' skills and knowledge. However, this trend is likely to increase the dependency on data specialist skills to manage performance.

Application program interfaces (APIs) are also expected to create dependencies on skills for big data deployment. APIs provide the building blocks of protocols, tools and routines for the interaction of software components in order to create applications, particularly when developing graphical user interfaces (GUIs). The challenge for suppliers, authorities and managers will be to ensure that skills are available at the right level and in sufficient quantity to support big data solutions.

4.7 Internet of Things

The internet of things is expected to create significant data content as more and more devices become connected. This is likely to provide opportunities for big data application developers and cloud service providers to innovate.

This IET *Transport Sector Insight* was written by Matthew Clarke, ATKINS Transportation.

Image of driverless pod courtesy of the Transport Systems Catapult.