1. **Big Data features, scales and data sources**

When discussing Big Data, three main characteristics are mentioned: volume, velocity, and variety, derived from Laney's 3 V model [12]. These elements are critical in how data is ingested, analysed, managed, stored, distributed, and processed. The following list and Figure 2.1.1 describes in detail the three most important indicators that distinguish Big Data from "normal" data [13] :

1. **Volume:** The total amount of data captured, managed, and analysed is volume. Nowadays, it is not uncommon to find experimental setups dealing with volumes up to the yottabyte scale (Y.B., 1024 bytes), with petabytes (P.B., 1015 bytes) as the minimum data volume for a scenario to be recognised as Big Data.
2. **Velocity:** this feature can refer to two concepts: (1) the rate at which data instances are generated and (2) the rate at which such data samples are received and processed. Currently, investigated ITS scenarios necessitate real-time processing latencies in nanoseconds, a stringent constraint requiring critical design decisions in the computing infrastructure and the developed data analysis models.
3. **Variety:** the heterogeneity of the captured data is due to the diversity of digitalised domains related to the application or service at hand. The requirement for jointly processing and analysing them has significant implications for system design, including data integration and fusion functionalities.

Although the above description of Big Data features is not limited to any particular field of interest, as seen in Kwan M. P's study [14], it is a good starting point for narrowing them down to the domains addressed in this paper: transportation big data.

In terms of volume and size, Big Data refers to datasets with sizes that make traditional approaches to collection, management, processing, and analysis impractical. Datasets have grown in size over the years, with single datasets ranging from a few dozen terabytes to many petabytes. When it comes to transportation and mobility, an excellent example of new geo-localised data sources with increasing relevance in these fields is social media. For example, Twitter has announced that its users send over 500 million tweets per day, whereas Instagram claims that its users shared about 60 million photos per day in 2014. Some of this media content is geotagged with GPS data, allowing us to infer people's behavioural trends over time.

Similarly, the volume and coverage of sensed transportation data are growing due to the widespread use of sensors in vehicles, wearables, cell phones, and other devices. Sensors are deployed on roadways in major metropolitan areas that were not previously instrumented or even visually inspected [15]. In addition, smaller communities with limited or no data collection are gradually deploying sensors and implementing data collection mechanisms, such as crowdsourcing. Overall, the increased number of sensors results in a significant increase in data volume in all of these cases.

According to the same logic, velocity is how new data is generated and transferred from its source to its destination(s). Big Data technology enables data analysis while it is being generated rather than after it has been stored in databases. Real-time data stream processing for traffic control is one example. In terms of data variety, many opportunities now arise from capturing massive amounts of data from various sources and the ability to connect and exploit them with people, governments, and businesses. Traditional legacy systems, relational databases, sensor networks, Open Data, email, video, blogs, call centre conversations, and social media are just a few examples. Because of the diversity of data sources, the type and form in which data are represented vary greatly; indeed, even for a fixed nature of the data (e.g., high-definition radar signals for vehicle speed and count detection), differences between the data delivered by sensors capturing them may appear simply because different companies manufacture them.

Big Data represents a significant increase in the quantity, variety, and accessibility of data. There are three primary Big Data sources in transportation and mobility: unstructured social media, sensor data (unstructured or semi-structured), and open data (unstructured – raw text, semi-structured – JSON/XML, or semantically structured – Linked Open Data).

1. **Social Media:** Nowadays, Social Media platforms store enormous amounts of data that could reveal critical information upon their eventual analysis. The underlying reason is that the user role has shifted from being a mere consumer to providing content assets. Social Media is defined by Kaplan et al. [16] as a set of Internet-based applications that build on Web 2.0's ideological and technological foundations and allow the creation and exchange of user-generated content. Therefore, social media can be considered a context-rich source of Big Data, usually referred to as Social Big Data. Furthermore, it is an important data source in transportation and mobility because the gathered information is typically geotagged, allowing for the inference of geo-localised knowledge about the mobility of the people who produce data traces. Other data sources in this category include collaborative applications and initiatives such as crowdsourcing, which is the practice of gathering information or input for a task or project by enlisting the help of a large number of people. Logistics is another area where this type of data is widely used nowadays, giving rise to novel concepts and collaborative services. Among them, crowdsourcing delivery refers to the scenario in which citizens deliver goods to each other along their route to reduce the carbon footprint of the delivered product while keeping in mind that deliveries must be done with a minimal detour from the deliverer's original path.
2. **In terms of sensor data**, the proliferation of physical devices embedded with electronics, software, sensors, and actuators fueled by the IoT paradigm is dramatically accelerating the pace of innovation in the transportation industry, particularly in concepts such as the connected car (also known as automobile as a sensor), which allows for the exchange of sensed information not only between vehicles but also between vehicles and infrastructure. Furthermore, new modes of last-mile delivery using new connected means, such as drones or autonomous vehicles, are expected to benefit supply chains and logistics operations significantly. Because of the criticality of an accident in an automotive scenario, transportation safety is another topic of increasing relevance for the IoT domain. To summarise, IoT is expected to be a critical technology in current and future transportation and mobility management solutions, as it can improve safety and efficiency, reduce environmental impact, and foster new services and capabilities. Similar claims have been made in research related to urban computing, which refers to the acquisition and integration of data generated by various sources in urban areas. Ubiquitous sensing technologies (e.g., IoT) play an essential role in urban computing, mainly when dealing with transportation issues throughout the city.
3. **Open Data:** The term "open" refers to the fact that anyone can freely access, use, modify, and share data for any purpose. Both citizens and the public sector stand to benefit from open transportation data. Therefore, governments should publish open transportation information in machine-readable and easily consumable formats. Many countries have recently embraced this potential by creating data repositories for national public transportation services rather than transit systems.
4. **Data acquisition and recording.**

People are increasingly leaving a digital footprint wherever they go (both voluntarily and involuntarily). Each text message, phone call, credit card purchase, social media post, email, online search, and many other electronic transactions use technology to report the user's location at a given point in time. This data is then routed to the service providers' central servers, which enable these actions. When cross-referenced with the geographical terrain, data at this scale provides a means of understanding and responding to the city's urban dynamics in real-time. However, making sense of this data, particularly for policy purposes, necessitates familiarity with technical aspects of data production methods and knowledge of how or from whom the data is sourced [17].

Data can be "born digital" or "born analogue." Data that has been "born digital" is created by users of a computing device specifically for use in a machine processing environment. Data that was "born digital" includes:

* Commercial transaction data (transaction records and credit card use, RFID tag reading and bar-code etc.).
* GPS (Global Positioning System) or other geo-located spatial data stamps.
* Devices, vehicles and networked objects produce data.
* Mobile devices use device identity, status, and location metadata to connect to various networks (GSM, Wi-Fi, etc.).
* SMS and Emails, metadata concerning phone calls
* Public transport card swipes and other data associated with portal access (key cards, RFID tags, badge)
* Timestamps and process logs.

Data "born digital" is created to meet one or more specific needs. Efficiency considerations have meant that only the specific data required for a process was generated and retained to avoid overburdening storage and computational capacity or inflating costs. However, the decrease in processing and storage costs effectively means that data over-collection (data collection outside the stated actual purpose for data collection) is easy and has a near-zero actual cost. Data "born analogue" is data created by imprinting a physical phenomenon (light, motion, sound, biological compound or presence of a chemical, magnetic impedance, etc.) on a sensing device and then converted into a digital signal. Cameras, microphones, magnetic field detectors, heart rate monitors, accelerometers, thermal sensors, and other sensors are examples of sensors. As a result, sensor costs have dropped dramatically, contributing to a rapidly pervasive sensing environment. Here are some examples of "born analogue" data:

* Electromagnetic or light (laser) reflectance of objects (laser-based LIDAR systems or synthetic aperture radar –SAR).
* The audio content of voice phone calls, ambient audio from video cameras or microphone networks.
* Data relating to heart rate, respiration, gait, and other physical and health parameters.
* Inertia/motion (ultrasonic sensors, accelerometers), temperature, electromagnetic fields, infrared radiation, air pressure etc.
* Surveillance, in-vehicle, roadside, and other cameras' video streams.

With the proliferation of sensors, some may argue that "born analog" data is rapidly expanding to include all potential observations now and in the future. However, two filtering mechanisms operate at the start of data collection, limiting the scope of data collected. The first filter is any sensor or recording device's design specifications (and thus limitations). A heat sensor, for example, will not record audio signals due to design specifications, whereas an accelerometer will not record geographic coordinates. Nonetheless, the ability to infer one phenomenon from another is constantly evolving. For example, studies on electromagnetic interference (which road authorities have attempted to shield roadside data transmission cables from) have revealed that the data produced by the interference can be used to infer vehicle movements and provide traffic counts.

The second operational filter focuses on the rate of observation of sensed or monitored events versus the rate of data retention and transmission. Engine sensors in a commercial aircraft, for example, may process up to 10 terabytes of data per 30 minutes of flight time, but the majority of this data is discarded as soon as it is used. Data retention rates are orders of magnitude lower, and the amount of data transmitted during flight is even lower. Therefore, much of the data sensed or generated is quickly discarded. What is left can then be filtered and compressed before being treated and used. However, the potential value of sensed data may not be recognised during the system design phase. Therefore, it is necessary to ensure that filters or compression algorithms do not discard potentially valuable data.

Filtering and compression-related data loss are becoming less an issue as low-cost data storage and transmission technologies combine with cutting-edge database platforms like Hadoop. These changes, combined with the increasing sophistication and lower cost of dual sensor-computing devices, increase the relevance of retained data. Sensor cost and size reductions, improved sensor performance, lower data storage costs, and increased relevance of retained data all contribute to a large-scale increase in exploitable data.

**Big Data core technologies**

Big Data draws on various technologies and system architectures designed to extract social or economic value from high velocity, large scale and highly diverse and heterogeneous data streams. At the heart of them are four interconnected technological developments:

1. **Pervasive data logging and sensor platforms**

Advanced software event logging (and storage) and the implementation of millions of sensing devices enable the global production of petabytes of real-time data.

1. **Real-time in-stream data analysis**

Sophisticated algorithms and distributed computing capacity (often hard-wired to sensor platforms) enable real-time data parsing and analysis. For example, an in-memory analysis helps extract relevant information from unstructured analog video or audio streams.

1. **Advances in data storage**

Reduced data storage costs have increased the retained-to-generated data ratio. This data includes previously dismissed information as insignificant or trivial (e.g., "digital dust"). When "digital dust" is analysed by sophisticated algorithms or combined with other contextual data sources, it can provide new insights. This data is increasingly being stored remotely in data centres in other countries. The emergence of "cloud" computing capacity, which can be used to analyse large and real-time data sets, is related to the development of remote data centres. The term "cloud" refers to remote data storage facilities and data transfer and networking protocols that enable access to and analysis of distributed data as if it were on a single server. Not only does "cloud computing" provide economies of scale in terms of data storage, management, and support costs, but it also opens up new avenues for ad hoc and customisable access to computing capacity on public cloud-based platforms (e.g., Amazon Web Services, Google Cloud Platform, etc.)

1. **New analytic frameworks**

New techniques for efficiently processing massive data sets within runtime computing capacity constraints have emerged. Many of these techniques have been made available under non-commercial open-source licenses. These techniques have significantly accelerated their adoption. For example, map-reduce work processes (such as Hadoop or its derivatives) take advantage of parallel processing by dividing large, complex semi-structured and unstructured data sets into more manageable subsets. They then distribute coordinated processing tasks to several distributed servers. These algorithms are entirely scalable and are not constrained by the need to formalise database relationships prior to storage and analysis. They can be applied directly to data, regardless of size, format, or complexity. Nonetheless, they may not be reactive enough to be used in instream data analysis. Other approaches specifically geared to analysing real-time streaming data and involving some forms of in-memory processing have emerged (analysis occurs without data storage).

1. **Real-time data analytics**

In terms of transportation, current ITS infrastructure and platforms are not designed for real-time data processing, nor are they capable of analysing captured data at the rates required by critical applications (e.g. safety). This situation creates a challenging paradox because, for most transportation and mobility problems, a quick response time is critical to ensure that information and decision making are tightly coupled in time and thus valuable for practice. Three broad categories of Big Data systems can be identified based on the required latency level [18]:

* **In Batch systems**, the data processing cycle is subject to very loose (hours to days) or no latency constraints. These systems collect data, process it, and generate results in stages. This is the case of delay-insensitive applications, which are exemplified in the context of ITS by control panels for medium-term (e.g., weekly) freight transportation monitoring and planning.
* **Soft real-time systems** have weak real-time constraints ranging from seconds to minutes. Online systems are also included in this category, as the response time to the user should be quick but not critical. A query service on the occupation of parking lots in a Smart City is an example. While a more responsive user interface with the system improves the user experience, compliance with real-time latency constraints is not required for the application at hand.
* **Hard real-time systems**, in which meeting ultra-low-latency constraints is a hard constraint, with thresholds ranging from milliseconds to nanoseconds. If the designed data processing life cycle does not meet this timing requirement, the developed system fails and cannot support the application for which it was designed. Examples of latency-critical applications include incident prediction and inference of the state of the road for autonomous vehicles

Real-time processing is thus deemed critical for providing new forms of on-demand mobility for people and things, necessitating real-time data sharing and interpretation, and necessitating legislative/public attitude shifts to implement intelligent mobility use cases utilising real-time network capacity data for people, vehicles, and goods, public transportation schedules, vehicle location data, dynamic fare and pricing data, sentiment data from service users and non-users, third-party usage data, and third-party usage. Real-time service automation and optimisation will be critical enablers for future intelligent mobility and transportation solutions. The gap to be bridged is not the technological transformation of this sector but rather the transformation of people toward processes and methodologies for using and profiting from future ITS. Real-time systems must be quick. If they are not fast enough to process data as it comes in, the system will be overloaded, and the system will no longer work in real-time.

Nothing is true in real-time if you think about it. There will always be some lag. Even the most advanced real-time systems will have a delay of 10 to the power of something seconds. So, when we say "real-time," we mean processing that takes less time than a predefined benchmark. This metric varies depending on the system. For example, a bank ATM's data processing can be considered real-time if it responds in less than one-tenth of a second. In a supercomputer, however, the same speed would be considered slow.

Real-time processing is used when you need the output immediately, and it has advantages and disadvantages.

Advantages of Real-Time Processing:

* The information is continually updated and ready to use.
* You have increased uptime.
* The delay in data processing is minimal.
* System synchronisation would require fewer resources.
* It aids in identifying problems, allowing you to take immediate action.

Disadvantages of Real-Time Processing:

* It adds a data overload in the event of a system failure.
* It is not easy to implement with simple systems.
* It requires high-performance hardware and is expensive.

1. **Security and privacy**

Security and privacy are long-standing and widely researched issues in transportation. Because of the personal geo-localised data and content shared by users via their devices and computers, privacy is likely the most stringent security requirement imposed by transportation applications leveraging Big Data. As a result, the transportation industry continues to face unresolved challenges in terms of collecting private data. Security flaws and gaps range from identity theft to unwanted location and data tracking. Furthermore, trends like using payment data (in all its forms and scales) as an additional source to infer macroscopic mobility patterns significantly increase the privacy risks associated with the Big Data paradigm. These privacy concerns jeopardise the widespread adoption of Big Data technologies for ITS and mobility; surprisingly, we find a scarcity of references and contributions with effective ITS technologies that ensure confidentiality, integrity, and secure data handling, including financial and personal information. Without a doubt, tackling this challenge with specific, viable technical approaches would cast a thick blanket of certainty over the use of Big Data in transportation and mobility, assuring users that their rights are protected following international laws, directives, and recommendations.

Much of the regulatory debate over collecting and using personal data focuses on the design of data collection mechanisms and practices. Most of the time, it is assumed that little will change in terms of how data and data systems embed and transmit bits of personally identifiable information. However, this viewpoint is challenged, particularly by proponents of the "Privacy by Design" approach. According to this viewpoint, data collection systems and practices should be designed (or redesigned) from the ground up to include solid and irreversible pro-privacy measures for data collection and handling systems – even in the design of machine logging protocols and sensors.

Ann Cavoukian, Executive Director of the Institute for Privacy and Big Data at Ryerson University in Canada and former Information and Privacy Commissioner for the Province of Ontario, developed the "Privacy by Design" approach to design data collection mechanisms and practices. "Privacy by Design" is founded on the tenet that pro-privacy concrete measures should be addressed during the design stage for data collection and analysis rather than after data has been collected and developed analytic systems. The approach is based on seven fundamental principles [19].

* 1. **Proactive, not reactive – preventative, not remedial:** The "Privacy by Design" (PbD) approach is proactive rather than reactive. It anticipates and prevents privacy-invading events from occurring. PbD does not wait for privacy concerns to arise, nor does it provide remedies for resolving privacy violations after they have occurred – it seeks to prevent them from occurring. "Privacy by Design" occurs before, not after the fact.
  2. **Setting privacy as the default:** "Privacy by Design" ensures that personal data are automatically protected in any given I.T. system or organisational practice, providing the highest level of privacy. Even if a person does nothing, their privacy is still protected. Individuals are not required to take any action to protect their privacy because it is built into the system by default.
  3. **Privacy embedded into design:** "Privacy by Design" is incorporated into the design and architecture of information technology systems, organisational structures, and business practices. It is not tacked on as an afterthought. As a result, privacy becomes an essential component of the core functionality provided. Privacy is essential to the system without sacrificing functionality.
  4. **Full functionality: positive-sum, not zero-sum:** "Privacy by Design" seeks to accommodate all legitimate interests and objectives in a positive-sum, win-win fashion, rather than a zero-sum approach that results in unnecessary trade-offs. "Privacy by Design" avoids the appearance of false dichotomies, such as privacy vs security, by demonstrating that both can be had.
  5. **End-to-end security: complete lifecycle protection:** "Privacy by Design" extends securely throughout the entire lifecycle of the personal data involved, having been embedded into the system prior to the first element of information being collected – strong security measures are essential to privacy, from start to finish. This ensures that all personal data is securely retained and then securely destroyed promptly. Thus, "Privacy by Design" ensures end-to-end, cradle-to-grave, secure lifecycle management of personal information.
  6. **Visibility and transparency:** keep it open "Privacy by Design" seeks to reassure all stakeholders that it operates following the stated promises and objectives, subject to independent verification regardless of organisation, business practice, or technology. Furthermore, its constituent parts and operations are visible and transparent to users and providers.
  7. **Respect for User Privacy: keep it user-centric:** Above all, "Privacy by Design" requires architects and operators to prioritise the individual's interests by implementing safeguards such as substantial privacy defaults, appropriate notice, and empowering user-friendly options. "Privacy by Design" is centred on the user.