A Reference Model for Big Data Technologies

Edward Curry, Andreas Metzger, Arne J. Berre, Andrés Monzón, and Alessandra Boggio-Marzet

Abstract The Big Data Value (BDV) Reference Model has been developed with input from technical experts and stakeholders along the whole big data value chain. The BDV Reference Model may serve as a common reference framework to locate big data technologies on the overall IT stack. It addresses the main technical concerns and aspects to be considered for big data value systems. The BDV Reference Model enables the mapping of existing and future data technologies within a common framework. Within this chapter, we detail the reference model in more detail and show how it can be used to manage a portfolio of research and innovation projects.

Keywords Reference model · Big data technologies · Data management · Data processing · Data analysis · Data visualisation · Data protection

1 Introduction

The Big Data Value (BDV) Reference Model has been developed with input from technical experts and stakeholders along the whole big data value chain. The BDV Reference Model may serve as a common reference framework to locate big data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for big data value systems. Within this chapter, we detail the reference

E. Curry (✉)
Insight SFI Research Centre for Data Analytics, NUI Galway, Galway, Ireland
e-mail: edward.curry@nuigalway.ie

A. Metzger
paluno, University of Duisburg-Essen, Duisburg, Germany

A. J. Berre
SINTEF Digital, Oslo, Norway

A. Monzón · A. Boggio-Marzet
Universidad Politécnica de Madrid, Madrid, Spain

E. Curry et al. (eds.), *The Elements of Big Data Value*,
https://doi.org/10.1007/978-3-030-68176-0_6

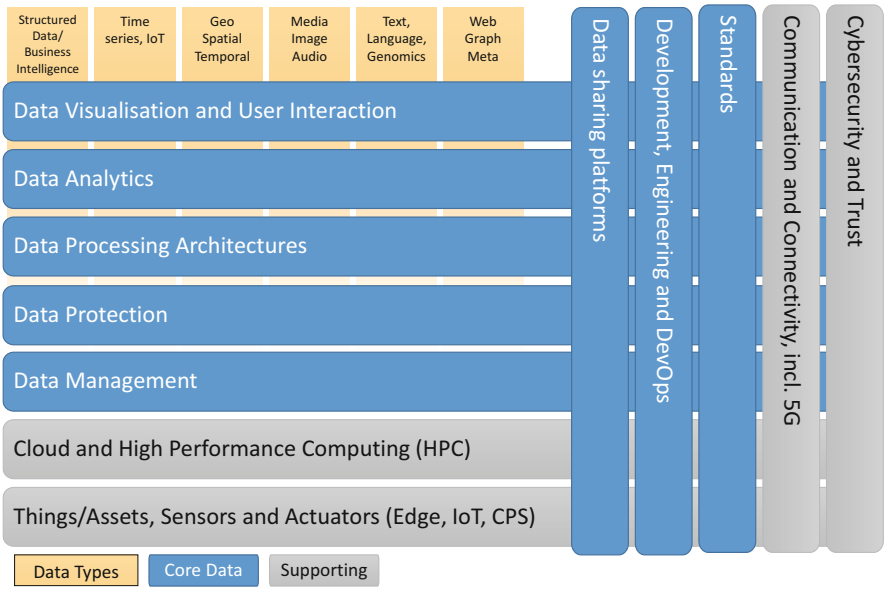

Fig. 1 Big Data Value Reference Model

model in more detail and show how it can be used to manage a portfolio of research and innovation projects. Section 2 details the Reference Model with its horizontal and concerns. Section 3 describes the use of the Reference Model within large-scale data projects to map projects' technical outcomes. Finally, Sect. 4 concludes the chapter.

2 Reference Model

An overview of the BDV Reference Model is shown in Fig. 1. It distinguishes between two different elements. On the one hand, it describes the elements that are at the core of the BDVA (also see Chap. "The European Big Data Value Ecosystem"); on the other, it outlines the features that are developed in strong collaboration with related European activities.

The BDV Reference Model has been developed by the Big Data Value Association (BDVA), taking into account input from technical experts and stakeholders along the whole big data value chain, as well as interactions with other related public-private partnerships (PPPs) (Zillner et al. 2017). The BDV Reference Model may serve as a common reference framework to locate big data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for big data value systems.

The BDV Reference Model is structured into horizontal and vertical concerns.

- Horizontal concerns cover specific aspects along the data processing chain, starting with data collection and ingestion, and extending to data visualisation. It should be noted that the horizontal concerns do not imply a layered architecture. As an example, data visualisation may be applied directly to collected data (the data management aspect) without the need for data processing and analytics.
- Vertical concerns address cross-cutting issues, which may affect all the horizontal concerns. In addition, vertical concerns may also involve non-technical aspects.

It should be noted that the BDV Reference Model has no ambition to serve as a technical reference architecture. However, it is compatible with such reference architectures, most notably the emerging ISO JTC1 WG9 Big Data Reference Architecture.

The following elements as expressed in the BDV Reference Model are elaborated in the remainder of this section.

2.1 Horizontal Concerns

Horizontal concerns cover specific aspects of a big data system. On the one hand, they cover the different elements of the data processing chain, starting from data collection and ingestion up to data visualisation and user interaction. On the other hand, they cover elements that facilitate deploying and operating big data systems, including Cloud and HPC, as well as Edge and IoT.

2.1.1 Data Visualisation and User Interaction

This concern covers advanced visualisation approaches for improved user experience. Data visualisation plays a key role in effectively exploring and understanding big data. Visual analytics is the science of analytical reasoning assisted by interactive user interfaces. Data generated from data analytics processes need to be presented to end-users via (traditional or innovative) multi-device reports and dashboards which contain varying forms of media for the end-user, ranging from text and charts to dynamic, 3D and possibly augmented-reality visualisations. In order for users to quickly and correctly interpret data in multi-device reports and dashboards, carefully designed presentations and digital visualisations are required. Interaction techniques fuse user input and output to provide a better way for a user to perform a task. Common tasks that allow users to gain a better understanding of big data include scalable zooms, dynamic filtering and annotation.

When representing complex information on multi-device screens, the design issues multiply rapidly. Complex information interfaces need to be responsive to human needs and capacity (Raskin 2000). Knowledge workers need to be supplied with relevant information according to the just-in-time approach. Too much information, which cannot be efficiently searched and explored, can obscure the

information that is most relevant. In fast-moving time-constrained environments, knowledge workers need to be able to quickly understand the relevance and relatedness of information.

2.1.2 Data Analytics

This concern covers data analytics, which ranges from descriptive analytics ("What happened and why?") through predictive analytics ("What will happen and when?") to prescriptive analytics ("What is the best course of action to take?"). The progress of data analytics is key not only for turning big data into value but also for making it accessible to the wider public. Data analytics will have a positive influence on all parts of the data value chain (Cavanillas et al. 2016) and increase business opportunities through business intelligence and analytics while bringing benefits to both society and citizens.

Data analytics is an open, emerging field, in which Europe has strong competitive advantages and a promising business development potential. It has been estimated that governments in Europe could save $149 billion (Manyika et al. 2011) by using big data analytics to improve operational efficiency. Big data analytics can provide additional value in every sector where it is applied, leading to more efficient and accurate processes. A study by the McKinsey Global Institute placed a strong emphasis on analytics, ranking it as the main future driver for US economic growth, ahead of shale oil and gas productions (Lund et al. 2013).

The next generation of analytics will be required to deal with a vast amount of information from different types of sources, with differentiated characteristics, levels of trust and frequency of updating. Data analytics will have to provide insights into the data in a cost-effective and economically sustainable way. On the one hand, there is a need to create complex and fine-grained predictive models for heterogeneous and massive datasets such as time series or graph data. On the other hand, such models must be applied in real time to large amounts of streaming data. This ranges from structured to unstructured data, from numerical data to micro-blogs and streams of data. The latter is exceptionally challenging because data streams, in addition to their volume, are very heterogeneous and highly dynamic, which also calls for scalability and high throughput. For instance, data collection related to a disaster area can easily occupy terabytes in binary GIS formats, and real-time data streams can show bursts of gigabytes per minute.

In addition, an increasing number of big data applications are based on complex models of real-world objects and systems, which are used in computation-intensive simulations to generate new huge datasets. These can be used for iterative refinements of the models, but also for providing new data analytics services which can process extremely large datasets.

2.1.3 Data Processing Architectures

This concern covers optimised and scalable architectures for analytics of both data-at-rest and data-in-motion, thereby delivering low-latency real-time analytics.

The Internet of Things (IoT) is one of the key drivers of the big data phenomenon. Initially, this phenomenon started by applying the existing architectures and technologies of big data that we categorise as data-at-rest, which is data kept in persistent storage. In the meantime, the need for processing immense amounts of sensor data streams has increased. This type of data-in-motion (i.e. non-persistent data processed on the fly) has extreme requirements for low-latency and real-time processing. What has hardly been addressed is the concept of complete processing for the combination of data-in-motion and data-at-rest.

For the IoT domain, these capabilities are essential. They are also required for other domains like social networks or manufacturing, where huge amounts of streaming data are produced in addition to the available big datasets of actual and historical data.

These capabilities will affect all layers of future big data infrastructures, ranging from the specifications of low-level data flows with the continuous processing of micro-messages, to sophisticated analytics algorithms. The parallel need for real-time and large data volume capabilities is a key challenge for big data processing architectures. Architectures to handle streams of data such as the lambda and kappa architectures will be considered as a baseline for achieving a tighter integration of data-in-motion with data-at-rest.

Developing the integrated processing of data-at-rest and data-in-motion in an ad hoc fashion is of course possible, but only the design of generic, decentralised and scalable architectural solutions will leverage their true potential. Optimised frameworks and toolboxes allowing the best use of both data-in-motion (e.g. data streams from sensors) and data-at-rest will leverage the dissemination of reference solutions which are ready and easy to deploy in any economic sector. For example, proper integration of data-in-motion with predictive models based on data-at-rest will enable efficient, proactive processing (detection ahead of time). Architectures that can handle heterogeneous and unstructured data are also important. When such solutions become available to service providers, in a straightforward manner, they will then be free to focus on the development of business models.

The capabilities of existing systems to process such data-in-motion and answer queries in real time and for thousands of concurrent users are limited. Special-purpose approaches based on solutions like Complex Event Processing (CEP) are not sufficient for the challenges posed by the IoT in big data scenarios. The problem of achieving effective and efficient processing of data streams (data-in-motion) in a big data context is far from being solved, especially when considering the integration with data-at-rest and breakthroughs in NoSQL databases and parallel processing (e.g. Hadoop, Apache Spark, Apache Flink, Apache Kafka). Applications, for instance of Artificial Intelligence, are also required to fully exploit all the capabilities

of modern and heterogeneous hardware, including parallelism and distribution to boost performance.

To achieve the agility demanded by real-time business and next-generation applications, a new set of interconnected data management capabilities is required.

2.1.4 Data Protection

This concern covers privacy and anonymisation mechanisms to facilitate data protection. This is shown related to data management and processing as there is a strong link here, but it can also be associated with the area of cybersecurity.

Data protection and anonymisation is a major issue in the areas of big data and data analytics. With more than 90% of today's data having been produced in the last 2 years, a huge amount of person-specific and sensitive information from disparate data sources, such as social networking sites, mobile phone applications and electronic medical record systems, is increasingly being collected. Analysing this wealth and volume of data offers remarkable opportunities for data owners, but, at the same time, requires the use of state-of-the-art data privacy solutions, as well as the application of legal privacy regulations, to guarantee the confidentiality of individuals who are represented in the data. Data protection, while essential in the development of any modern information system, becomes crucial in the context of large-scale sensitive data processing.

Recent studies on mechanisms for protecting privacy have demonstrated that simple approaches, such as the removal or masking of the direct identifiers in a dataset (e.g. names, social security numbers), are insufficient to guarantee privacy. Indeed, such simple protection strategies can be easily circumvented by attackers who possess little background knowledge about specific data subjects. Due to the critical importance of addressing privacy issues in many business domains, the employment of privacy-protection techniques that offer formal privacy guarantees has become a necessity. This has paved the way for the development of privacy models and techniques such as differential privacy, private information retrieval, syntactic anonymity, homomorphic encryption, secure search encryption and secure multiparty computation, among others. The maturity of these technologies varies, with some, such as k-anonymity, more established than others. However, none of these technologies has so far been applied to large-scale commercial data processing tasks involving big data.

In addition to the privacy guarantees that can be offered by state-of-the-art privacy-enhancing technologies, another important consideration concerns the ability of the data protection approaches to maintain the utility of the datasets to which they are applied, with the goal of supporting different types of data analysis. Privacy solutions that offer guarantees while maintaining high data utility will make privacy technology a key enabler for the application of analytics to proprietary and potentially sensitive data.

A truly modern and harmonised legal framework on data protection which has teeth and can be enforced appropriately will ensure that stakeholders pay attention to

the importance of data protection. At the same time, it should enable the uptake of big data and incentivise privacy-enhancing technologies, which could be an asset for Europe as this is currently an underdeveloped market. In addition, users are beginning to pay more attention to how their data are processed. Hence, firms operating in the digital economy may realise that investing in privacy-enhancing technologies could give them a competitive advantage.

2.1.5 Data Management

This concern covers principles and techniques for data management, including data ingestion, sharing, integration, cleansing and storage. More and more data are becoming available. This data explosion, often called a "data tsunami", has been triggered by the growing volumes of sensor data and social data, born out of Cyber-Physical Systems (CPS) and Internet of Things (IoT) applications. Traditional means for data storage and data management are no longer able to cope with the size and speed of data delivered in heterogeneous formats and at distributed locations.

Large amounts of data are being made available in a variety of formats – ranging from unstructured to semi-structured to structured – such as reports, Web 2.0 data, images, sensor data, mobile data, geospatial data and multimedia data. Important data types include numeric types, arrays and matrices, geospatial data, multimedia data and text. A great deal of this data is created or converted and further processed as text. Algorithms or machines are not able to process the data sources due to the lack of explicit semantics. In Europe, text-based data resources occur in many different languages, since customers and citizens create content in their local language. This multilingualism of data sources means that it is often impossible to align them using existing tools because they are generally available only in the English language. Thus, the seamless aligning of data sources for data analysis or business intelligence applications is hindered by the lack of language support and gaps in the availability of appropriate resources.

Isolated and fragmented data pools are found in almost all industrial sectors. Due to the prevalence of data silos, it is challenging to accomplish seamless integration with and smart access to the various heterogeneous data sources. And still today, data producers and consumers, even in the same sector, rely on different storage, communication and thus different access mechanisms for their data. Due to the lack of commonly agreed standards and frameworks, the migration and federation of data between pools impose high levels of additional costs. Without a semantic interoperability layer being imposed upon all these different systems, the seamless alignment of data sources cannot be realised.

In order to ensure a valuable big data analytics outcome, the incoming data has to be of high quality, or, at least, the quality of the data should be known to enable appropriate judgements to be made. This requires differentiating between noise and valuable data, and thereby being able to decide which data sources to include and which to exclude to achieve the desired results.

Over many years, several different application sectors have tried to develop vertical processes for data management, including specific data format standards and domain models. However, consistent data lifecycle management – that is, the ability to clearly define, interoperate, openly share, access, transform, link, syndicate and manage data – is still missing. In addition, data, information and content need to be syndicated from data providers to data consumers while maintaining provenance, control and source information, including IPR considerations (data provenance). Moreover, to ensure transparent and flexible data usage, the aggregation and management of respective datasets enhanced by a controlled access mechanism through APIs should be enabled (Data-as-a-Service).

2.1.6 Cloud and High-Performance Computing (HPC)

Efficient big data processing, data analytics and data management require the effective use of Cloud and High-Performance Computing infrastructures to address the computational resource and storage needs of big data systems.

Cloud Data ecosystems, promoted by the BDVA, should include strong links to scientific research that is becoming predominantly data driven. The BDVA is in a strong position to nurture such links as it has established strong relationships with European big data academia. However, a lack of access, trust and reusability prevents European researchers in academia and industry from gaining the full benefits of data-driven science. Most datasets from publicly funded research are still inaccessible to the majority of scientists in the same discipline, not to mention other potential users of the data, such as company R&D departments. Approximately 80% of research data is not in a trusted repository. However, even if the data openly appears in repositories, this is not always enough. As a current example, only 18% of the data in open repositories is reusable.[1] This leads to inefficiencies and delays; in recent surveys, the time reportedly spent by data scientists in collecting and cleaning data sources made up 80% of their work (G. Press 2016).

In response to these challenges, the Commission has launched a large effort to create "a European Open Science Cloud to make science more efficient and productive and let millions of researchers share and analyse research data in a trusted environment across technologies, disciplines and borders"[1]. The initial outline for the European Open Science Cloud (EOSC) was laid out in the report from the High-Level Expert Group.[2] The report advised the Commission on several measures needed to implement the governance and the financial scheme of the European Open Science Cloud, such as being based on a federated system of existing and emerging research (e-)infrastructures operating under light international governance with well-defined Rules of Engagement for participation. Machine understanding of

[1] "Are FAIR data principles FAIR?" LIBER Webinar by Alastair Dunning, 10.03.2017.

[2] Realising the European Open Science Cloud, 2016, https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

data – based on common or widely used data standards – is required to handle the exponential growth in publications. Attractive career paths for data experts should be created through proper training and by applying modern reward and recognition practices. This should help to satisfy the growing demand for data scientists working together with substance scientists. Turning science into innovation is emphasised, and alongside this there is a need for industry, especially SMEs and start-ups, to be able to access the appropriate data resources.

A first phase aims at establishing a governance and business model that sets the rules for the use of the EOSC, creating a cross-border and multi-disciplinary open innovation environment for research data, knowledge and services, and ultimately establishing global standards for the interoperability of scientific data.

The EU has already initiated and will go on to launch several more infrastructure projects, such as EOSC-hub, within H2020 for implementing and piloting the EOSC. In addition to these projects, Germany and the Netherlands, among other countries, are promoting the GO FAIR initiative (Germany and the Netherlands 2017). The FAIR principles aim to ensure that Data and Digital Research Objects are Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al. 2016). As science becomes increasingly data driven, making data FAIR will create real added value since it allows for combining datasets across disciplines and across borders to address pressing societal challenges that are mostly interdisciplinary.

The GO FAIR initiative is a bottom-up, open-to-all, cross-border and cross-disciplinary approach aiming to contribute to a broad involvement of the European science community as a whole, including the "long tail" of science.

The EOSC initiative is aligned with the BDVA agenda, as both promote data accessibility, trustworthiness and reproducibility over domains and borders. In the BDVA, this mainly applies to the i-Spaces and Lighthouse instruments, where the interoperability of datasets is central. Data standardisation is a self-evident topic for cooperation, but there are also common concerns in non-technical priorities – most notably skills development (relating to data-intensive engineers and data scientists). Both industry and academia benefit from findable, accessible, interoperable and reproducible data.

High-Performance Computing In some sectors, big data applications are expected to move towards more computation-intensive algorithms to reap deeper insights across descriptive (explaining what is happening), diagnostic (exploring why it happens), prognostic (predicting what can happen) and prescriptive (proactive handling) analysis. The adoption of specific HPC-type capabilities by the big data analytics stack is likely to be of assistance where big data insights will be of the utmost value. Faster decision-making is crucial and extremely complex datasets are involved – i.e. extreme data analytics.

The Big Data and HPC communities (through BDVA and ETP4HPC collaboration[1]) have recognised their shared interests in strengthening Europe's position regarding extreme data analytics. Recent engagements between PPPs have focused on the relevant issues of looking at how HPC and Big Data platforms are implemented, understanding the platform requirements for HPC and Big Data

workloads, and exploring how the cross-transfer of certain technical capabilities belonging to either HPC or big data could benefit each other. For example, the application of deep learning is one such workload that readily stands to benefit from certain HPC-type capabilities regarding optimising and parallelising difficult optimisation problems.

Major technical requirements include highly scalable performance, high memory bandwidth, low power consumption and excellent short arithmetic performance. Additionally, more flexible end-user education paths, utilisation and business models will be required to capitalise on the rapidly evolving technologies underpinning extreme data analytics, as well as continued support for collaboration across the communities of both big data and HPC to jointly define the way forward for Europe.

2.1.7 IoT, CPS, Edge and Fog Computing

The main source of big data is sensor data from an IoT context and actuator interaction in Cyber-Physical Systems. To meet real-time needs, it will often be necessary to handle big data aspects at the edge of the system. This area is separately elaborated further in collaboration with the IoT (Alliance for Internet of Things Innovation (AIOTI)) and CPS communities.

Internet of Things (IoT) technology, which enables the connection of any type of smart device or object, will have a profound impact on many sectors in the European economy. Fostering this future market growth requires the seamless integration of IoT technology (such as sensor integration, field data collection, Cloud, Edge and Fog computing) and big data technology (such as data management, analytics, deep analytics, edge analytics and processing architectures).

The mission of the Alliance of Internet of Things Innovation (AIOTI) is to foster the European IoT market uptake and position by developing ecosystems across vertical silos, contributing to the direction of H2020 large-scale pilots, gathering evidence on market obstacles for IoT deployment in the Digital Single Market context, championing the EU in spearheading IoT initiatives, and mapping and bridging global, EU and Members States' IoT innovation and standardisation activities. AIOTI working groups cover various vertical markets from smart farming to smart manufacturing and smart cities, and specific horizontal topics on standardisation, policy, research and innovation ecosystems. The AIOTI was launched by the European Commission in 2015 as an informal group and established as a legal entity in 2016. It is a major cross-domain European IoT innovation activity.

Close cooperation between the AIOTI and the BDVA is seen as being very beneficial for the BDVA. The following areas of collaboration are of particular interest to the BDVA:

- Alignment of high-level reference architectures: A common understanding of how the AIOTI High-Level Architecture (HLA) and the BDVA Reference Model are related to each other enables well-grounded decisions and prioritisations related to the future impact of technologies.

- Deepening the understanding about sectorial needs: Through the mutual exchange of roadmaps, accompanied by insights about sectorial needs in the various domains, the BDVA will receive additional input about drivers for and constraints on the adoption of big data in the various sectors. In particular, insights about sector-specific user requirements as well as topics related to the BDV strategic research and innovation roadmap will be fed back into our ongoing updating process.
- Standardisation activities: To foster the seamless integration of IoT and big data technologies, the standardisation activities of both communities should be aligned whenever technically required. In addition, the BDVA can benefit from the already established partnerships between the AIOTI and standardisation bodies to communicate big-data-related standardisation requirements.

Aligning Security Efforts The efforts to strengthen security in the IoT domain will have a huge impact on the integrity of data in the big data domain. When IoT security is compromised, so too is the generated data. By developing a mutual understanding on security issues in both domains, trust in both technologies and their applications will be increased.

2.2 Vertical Concerns

Vertical concerns address cross-cutting issues, which are relevant and may affect more than one of the horizontal concerns. They may not be purely technical and also involve some non-technical aspects.

2.2.1 Big Data Types and Semantics

One specific vertical concern defined by the BDV Reference Model is data types. Different data types may require the use of different techniques and mechanisms in the horizontal concerns, for instance for data analytics and data storage.

The following six big data types have been identified as the main relevant data types used in big data systems: (1) structured data, (2) time series data, (3) geospatial data, (4) media data (image, video, audio, etc.), (5) text data (including natural language data and genomics representations) and (6) graph or network data. In addition, it is important to support both the syntactical and semantic aspects of data for all big data types, in particular, considering metadata.

2.2.2 Standards

This concern covers the standardisation of big data technology areas to facilitate data integration, sharing and interoperability.

Standardisation is a fundamental pillar in the construction of a Digital Single Market and Data Economy. It is only through the use of standards that the requirements of interconnectivity and interoperability can be ensured in an ICT-centric economy. The PPP will continue to lead the way in the development of technology and data standards for big data by:

- Leveraging existing common standards as the basis for an open and successful big data market
- Supporting standards development organisations (SDOs), such as ETSI, CEN-CENELEC, ISO, IEC, W3C, ITU-T and IEEE, by making experts available for all aspects of big data in the standardisation process
- Aligning the BDV Reference Model with existing and evolving compatible architectures
- Liaising and collaborating with international consortia and SDOs through the TF6SG6 Standards Group and Workshops
- Integrating national efforts on an international (European) level as early as possible
- Providing education and educational material to promote developing standards

Standards are the essential building blocks for product and service development as they define clear protocols that can be easily understood and adopted internationally. They are a prime source of compatibility and interoperability and simplify product and service development as well as speeding the time-to-market. Standards are globally adopted; they make it easier to understand and compare competing products, and thus drive international trade.

In the data ecosystem, standardisation applies to both the technology and the data.

Technology Standardisation Most technology standards for big data processing are de facto standards that are not prescribed (but are at best described after the fact) by a standards organisation. However, the lack of standards is a significant obstacle. One example is the NoSQL databases. The history of NoSQL is based on solving specific technology challenges that lead to a range of different storage technologies. The broad range of choices, coupled with the lack of standards for querying the data, makes it harder to exchange data stores, as this may tie application-specific code to a specific storage solution. The PPP is likely to take a pragmatic approach to standardisation and look to influence, in addition to NoSQL databases, the standardisation of technologies such as complex event processing for real-time big data applications, languages to encode the extracted knowledge bases, Artificial Intelligence, computation infrastructure, data curation infrastructure, query interfaces and data storage technologies.

Data Standardisation The "variety" of big data makes it very difficult to standardise. Nevertheless, there is a great deal of potential for data standardisation in the areas of data exchange and data interoperability. The exchange and use of data assets are essential for functioning ecosystems and the data economy. Enabling the seamless flow of data between participants (i.e. companies, institutions and individuals) is a necessary cornerstone of the ecosystem.

To this end, the PPP is likely to undertake collaborative efforts to support, where possible and pragmatic, the definition of semantic standardised data representation, ranging from the domain (industry sector)-specific solutions, like domain ontologies, to general concepts, such as Linked Open Data, to simplify and reduce the costs of data exchange.

In line with JTC1 Directives Clause 3.3.4.2, the Big Data Value Association (BDVA) requested the establishment of a Category C liaison with the ISO/IEC JTC1/WG9 Big Data Reference Architecture. This request was processed at the August Plenary meeting of ISO IEC JTC1 WG9, and the recommendation was unanimously approved by the working group. This liaison moves the BDVA work forward from a technology standardisation viewpoint, and now the BDVA Big Data Reference Model is closely aligned with the ISO Big Data Reference Architecture, as described in ISO IEC JTC1 WG9 20547-3. The BDVA TF6SG6 Standardisation Group is now also in the process of using the WG9 Use Case Template to extract data from the PPP Projects to extend the European use case influence on the ISO big data standards.

As the data ecosystem overlaps with many other ecosystems, such as Cloud computing, IoT, smart cities and Artificial Intelligence, the PPP will continue to be a forum for bringing together industry stakeholders from across these other domains to collaborate. These fora will continue to drive interoperability within the big data domain but will also extend this activity across the other technological ecosystems.

2.2.3 Communication and Connectivity

This concern covers effective communication and connectivity mechanisms, which are necessary for providing support for big data. This area is separately further elaborated, along with various communication communities, such as the 5G community.

The 5G PPP will deliver solutions, architectures, technologies and standards for the ubiquitous next generation of communication infrastructures in the coming decade. It will provide 1000 times higher wireless area capacity by facilitating very dense deployments of wireless communication links to connect over 7 trillion wireless devices serving over 7 billion people. This guarantees access to a wider panel of services and applications for everyone, everywhere.

5G provides the opportunity to collect and process big data from the network in real time. The exploitation of Data Analytics and big data techniques supports Network Management and Automation. This will pave the way to monitoring users' Quality of Experience (QoE) and Quality of Service (QoS) through new metrics combining network and behavioural data while guaranteeing privacy. 5G is also based on flexible network function orchestration, where machine learning techniques and approaches from big data handling will become necessary to optimise the network.

Turning to the IoT arena, the per-bit value of IoT is relatively low, while the value generated by holistic orchestration and big data analytics is enormous. Combinations

of 5G infrastructure capabilities, big data assets and IoT development may help to create more value, increased sector knowledge and ultimately more ground for new sector applications and services.

On the agenda of 5G PPP is the realisation of prototypes, technology demos, and pilots of network management and operation, Cloud-based distributed computing, edge computing and big data for network operation – as is the extension of pilots and trials to non-ICT stakeholders to evaluate the technical solutions and their impact on the real economy.

The aims of 5G PPP are closely related to the agenda of the BDVA. Collaborative interactions involving both ecosystems (e.g. joint events, workshops and conferences) could provide opportunities for the BDVA and 5G PPP to advance understanding and definition in their respective areas. The 5G PPP and BDVA ecosystems need to increase their collaboration with each other, and in so doing could develop joint recommendations related to big data.

2.2.4 Cybersecurity

This concern covers security and trust elements that go beyond privacy and anonymisation. The aspect of trust frequently has links to trust mechanisms such as blockchain technologies, smart contracts and various forms of encryption.

Cybersecurity and big data naturally complement each other and are closely related, for instance in using cybersecurity algorithms to secure a data repository, or reciprocally, using big data technologies to build dynamic and smart responses and protection from attacks (web crawling to gather information and learning techniques to extract relevant information).

By its nature, any data manipulation presents a cybersecurity challenge. The issue of Data Sovereignty perfectly illustrates the way in which both technologies can be intertwined. Data Sovereignty consists in merging personal data from several sources, always allowing the data owner to retain control over their data, be it by partial anonymisation, secure protocols, smart contracts or other methods. The problem as a whole cannot be solved by considering each of these technologies separately, especially those relevant to cybersecurity and big data. The problem has to be solved globally, taking a functionally complete and secure-by-design approach.

In the case of personal data space, both security and privacy should be considered. For industrial dataspaces, the challenges relate more to the protection of IPRs, the protection of data at large and the secure processing of sensitive data in the Cloud.

In terms of research and innovation, several topics have to be considered, for example homomorphic encryption, threat intelligence and how to test a learning process, assurance in gaining trust, differential privacy techniques for privacy-aware big data analytics and the protection of data algorithms.

Artificial Intelligence could be used and could even be more efficient in attacking a system rather than protecting it. The impact of falsified data, and trust in data,

should also be considered. It is essential to define the concepts of measurable trust and evidence-based trust. Data should be secured at rest and in motion.

The European Cyber Security Organisation (ECSO) represents the contractual counterpart to the European Commission for the implementation of the Cybersecurity contractual Public-Private Partnership (PPP)[1]. A collaboration with ECSO, supporting the Cybersecurity PPP, has been initiated and further steps planned.

2.2.5 Engineering and DevOps for Building Big Data Value Systems

This concern covers methodologies for developing and operating big data systems.

While big data technologies gain significant momentum in research and innovation, mature, proven and empirically sound engineering methodologies for building next-generation big data value systems are not yet available. Moreover, we lack proven approaches for continuous development and operations (DevOps) of big data value systems. The availability of engineering methodologies and DevOps approaches – combined with adequate toolchains and big data platforms – will be essential for fostering productivity and quality. As a result, these methodologies and approaches will empower the new wave of data professionals to deliver high-quality next-generation big data value systems.

2.2.6 Marketplaces, Industrial Data Platforms and Personal Data Platforms (IDPs/PDPs), Ecosystems for Data Sharing and Innovation Support

This concern covers data platforms for data sharing, which include, in particular, IDPs and PDPs, but also other data sharing platforms such as Research Data Platforms (RDPs), Data Platforms for Smart Environments (Curry 2020) and Urban/City Data Platforms (UDPs). These platforms facilitate the efficient usage of a number of the horizontal and vertical big data areas, most notably data management, data processing, data protection and cybersecurity.

Data sharing and trading are seen as essential ecosystem enablers in the data economy, although closed and personal data present particular challenges for the free flow of data (Curry and Ojo 2020). The following two conceptual solutions – Industrial Data Platforms (IDPs) and Personal Data Platforms (PDPs) – introduce new approaches to addressing this particular need to regulate closed proprietary and personal data.

3 Transforming Transport Case Study

This section illustrates the use of the BDV Reference Model within the large-scale European big data project TransformingTransport (http://www.transformingtransport.eu). The model was used to structure systematically, map, coordinate and align the project's technical outcomes, thereby also serving to distil lessons learned for the different technical concerns.

The TransformingTransport project demonstrated in a realistic, measurable and replicable way the transformations that big data can bring to the mobility and logistics market (Castiñeira and Metzger 2018; Metzger et al. 2019a). Structured into 13 different pilots, which cover areas of major importance for the mobility and logistics sector in Europe, TransformingTransport validated the technical and economic viability of big data for reshaping transport processes and services. To this end, TransformingTransport exploited access to industrial data sets from over 160 data sources, totalling 410,000 GB.

TransformingTransport ran from January 2017 to July 2019 and brought together knowledge, solutions and impact potential of major European ICT and big data technology providers with the competence and experience of key European industry players and public bodies in the mobility and logistics domain. TransformingTransport was one of the first two Lighthouse projects of the European Big Data Value Public-Private Partnership (http://www.big-data-value.eu/) funded by the European Commission within the framework of the Horizon 2020 programme.

TransformingTransport addresses 13 pilots in seven highly relevant pilot domains within mobility and transport that will benefit from big data solutions and the increased availability of data. The seven pilot domains and 13 pilots are shown in Fig. 2. For each pilot, TransformingTransport explored innovative use cases and engaged key players in the sector to demonstrate the transformative nature that big data technologies can bring about.

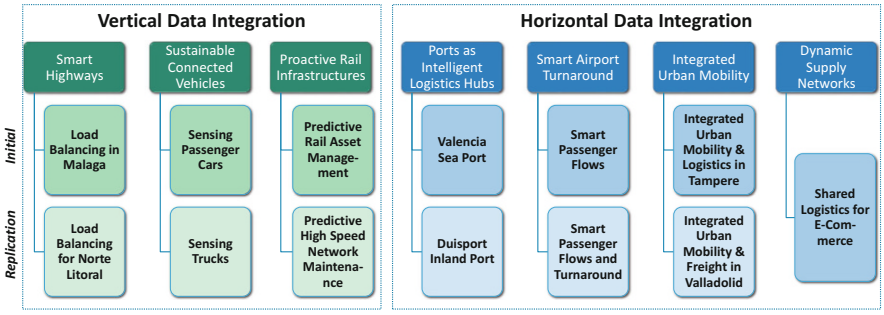

Fig. 2 Thirteen pilots in seven pilot domains

	Smart Highways		Sustainable Connected Vehicles		Proactive Rail Infrastructures		Ports as Intelligent Logistics Hubs		Smart Airport Turnaround		Integrated Urban Mobility		Dynamic Supply Networks
Data Management													
Semantic Annotation of unstructured and semi-structured data	2	2	3	3	3	3	3	3	3	3	4	4	3
Semantic interoperability	3	3	3	4	3	3	3	3	3	3	4	4	3
Data quality	3	3	4	4	2	2	4	4	4	4	4	4	4
Data lifecycle management and data governance	4	4	4	4	4	4	4	4	3	3	3	3	3
Integration of data and business processes	3	2	3	4	4	4	4	4	4	4	4	4	4
Data-as-a service	4	4	4	4	4	4	4	4	4	4	4	4	3
Distributed trust infrastructures for data management	4	4	4	4	4	4	4	4	4	4	4	4	4
Data Processing Architectures													
Heterogeneity	4	4	4	4	4	4	4	4	4	4	3	3	4
Scalability	3	3	3	3	3	3	3	3	3	3	3	3	3
Processing of data-in-motion and data-at-rest	4	4	4	4	4	4	4	4	4	4	4	4	4
Decentralizatrion	4	4	4	4	4	4	4	4	4	4	4	4	4
Performance	4	4	4	4	4	4	4	4	4	4	4	4	4
Novel architectures for enabling new types of big data workloads	3	3	4	4	4	4	4	4	4	4	4	4	4
Introduction of new hardware capabilities	4	4	4	3	4	4	4	4	4	4	4	4	3
Data Analytics													
Semantic and knowledge-based analysis	3	2	3	3	2	2	3	3	2	2	2	2	2
Content validation	4	4	4	4	3	3	4	4	3	3	4	4	4
Analytics frameworks & processing	2	3	3	3	3	3	3	3	3	3	3	3	3
Advanced business analytics and intelligence	3	2	2	1	1	1	2	2	3	3	2	2	2
Predictive and prescriptive analytics	1	1	1	2	1	1	1	1	1	1	1	1	1
High Performance Data Analytics (HPDA)	2	2	2	2	1	1	2	2	2	2	3	3	2
Data analytics and Artificial Intelligence	4	4	4	3	4	4	4	4	4	4	4	4	3
Data Protection													
Generic and easy to use data protection approaches	4	4	4	4	4	4	4	4	4	4	4	4	4
Robust Data privacy (incl. multi-party computation)	4	4	4	4	4	4	4	4	4	4	4	4	4
Risk based approaches	4	4	4	4	4	4	4	4	4	4	4	4	4
Data Visualisation and User Interaction													
Visual data discovery	3	3	3	2	2	2	3	3	3	3	3	3	3
Interactive visual analytics of multiple scale data	2	2	3	2	2	2	2	2	3	3	2	2	2
Collaborative, intuitive and interactive visual interfaces	2	2	2	2	2	2	2	2	3	3	2	2	2
Interactive visual data exploration and querying in a multi-device context	2	2	2	2	2	2	2	2	3	3	2	2	2

Fig. 3 Coverage of Big Data Value Reference Model (1 = Main focus; 2 = Topic addressed, but not main focus; 3 = Topic marginally addressed; 4 = Topic not addressed)

Figure 3 shows how the different pilots contributed to the different horizontal concerns of the Big Data Value Reference Model (as introduced in Sect. 2), breaking down their contributions to different technical priorities per concern. The numbers indicate the focus of the pilots on the respective technical priorities.

As can be seen, the most relevant horizontal concerns of TransformingTransport were (1) Data Analytics, (2) Data Visualisation and (3) Data Management, which we elaborate below together with lessons learned from the project. We then elaborate on how the impact of big data solutions on key business outcomes can be measured to assess the usefulness of these techniques, and then conclude the use case with some final observations.

3.1 Data Analytics

The key enabling analytics technology employed by TransformingTransport is predictive data analytics. Predictive analytics is a significant next step from descriptive analytics. While descriptive analytics answers the question "What happened and why?", predictive analytics attempts to answer the question "What will happen and when?" (see Sect. 2.1.2). For example, predictive analytics may help predict whether there may be a delay in a transport process, helping transport operators to be proactive and take action to decrease or prevent delays (Metzger et al. 2019a).

A case in point is the Smart Passenger Flows pilot at Athens Airport. With passenger demand increasing annually, the challenge for Athens Airport has been to identify intelligent ways to improve and streamline the flow of people through the airport, i.e., increase throughput, while at the same time ensuring the safety and the experience of passengers (Feltus et al. 2018). Increasing throughput requires sophisticated data analysis to build powerful big data models that can segment passengers and identify patterns and trends that will lead to actionable strategies on behalf of the airport.

Lessons learned in data analytics include:

- **Data quality**: Among the most universally accepted principles of analytics is "**Garbage in – Garbage out**", which refers to the quality of the data in the training models. It means that if poor-quality data enter the system, no matter how trendy the software for the analysis, the output value is expected to be of low quality too. To overcome this, checking and coping with missing data, data accuracy, data timelines, different time-zones (clocks), etc., is a must; so is assigning "data owners" that understand data and its field (domain) being able to be in the care of data quality.
- Using **Deep Learning** and Neural Networks helps to create more efficient development and engineering. They have been proven to work well even without extensive hyper-parametrisation, provided that enough good-quality data is available. This means that the time- and resource-consuming step of extensive experimentation with hyper-parameters may be skipped, leading to a more efficient development and deployment process of big data applications (Palm et al. 2020).
- **Data accuracy**: Operators benefit from information about data accuracy. This results in improved decision-making and helps to determine when to trust a prediction. Augmenting the quality of data (live or predicted) with confidence intervals, error ranges or reliability estimates allows operators to acquaint themselves with the most realistic situation.
- **Time series models** can be successfully approached by traditional **machine learning techniques**. It has been verified that machine learning techniques and Arima models are quite similar in short-term predictions, while the former tend to be more accurate as the time to be predicted increases. Not only are predictive models useful to improve a process, but it is also necessary to have teams with *enough experience* to select the most suitable alternative (descriptive or predictive). Another lesson learned is that external variables are easily included in the modelisation.

- **Historical data**: Regarding data analytics, pilots found it useful to **keep historical non-reproducible data** and, when possible, in **raw format**. Several reasons support this method, such as possible errors or improvements in the code that do not allow rebuilding of processed data if the original data is deleted. If one substitutes raw data with processed data, and there are no possible mechanisms to reverse the process, important information can be missed in ulterior processing stages. A drawback in maintaining unprocessed raw data could be the need for increased storage capacity. Raw historical data can also be used for training in machine learning algorithms. The main idea is to keep the complete historical data since some bits of previously untreated information can be very important for future analyses.

3.2 Data Visualisation

As the project concluded, one of the most useful and profitable visualisation techniques that was considered as a "*key success factor*" was **cockpit** for data visualisation and real-time control. Cockpit is a flexible human-machine interface (HMI) designed to help operators in day-to-day monitoring, where pilots have shared their knowledge to gain the most valuable insights from these tools.

A case in point is developed as part of Dusiport inland port pilot. This cockpit exploits advanced data processing, predictive analytics capabilities and interactive visualisation to support terminal operators in proactive decision-making and process adaptation (Metzger et al. n.d.). In addition to raising alarms in the case of a predicted delay, the terminal productivity cockpit also shows a reliability estimate for the predicted delay. The reliability estimate gives the probability (in %) of whether the alarm is indeed a true alarm. Reliability estimates facilitate distinguishing between more and less reliable predictions on a case-by-case basis (Metzger et al. 2019b).

Lessons learned in data visualisation include the following:

- Despite being an excellent tool to see what is happening around the pilot, a cockpit should not be exhaustive in relation to the amount of information displayed, which can lead to cognitive overload due to information overflow. There are three main requisites. First, the **information** must be shown **hierarchically** from top to bottom interface, enabling making summaries with the most relevant details. Second, widgets must be **intuitive, simple and "clean"** for the user and allow for quick handling to easily grasp the information shown. Third, cockpit should **only display critical and sufficiently well-validated events**, in order to avoid overloading the interface with superfluous warnings and focus the attention on the most important ones.
- Static user interfaces (UI) may be limiting. Providing **dynamic customisation** of UI from simple multi-option dropdowns to more complex interchangeable

requests could boost the efficiency of the analysis, adapting itself to specific user and operator needs.

- Visualisation helps to take decisions, with synthetic and clear results. Implications of the **human factors** team were found to be useful in understanding these aspects. Moreover, involving them in the early stages of the project also helped to gain a better perspective of the demonstrator.
- It is relevant to address the **right customer or user** who is going to work with the visualised data. In day-to-day business, there is often not enough time to only look at visualisations without an explicit added value. Yet, if the cockpit also serves as a decision-making tool, e.g. to plan routes, or has other technical implementations, it provides more added value. Another group to be approached could be **decision-makers** who can use these cockpits for strategic planning purposes.
- The goal of data visualisation is to make the data easily understandable and usable by the operators. To accomplish this, visualisations beyond just showing the quantitative data in big tables must be developed, thereby enabling the users to make a qualitative assessment of quantitative data intuitively. The terminal operators must be sure that the **data** is current. However, only knowing the current state is not sufficient for the operator. In addition, the **date and time** of the last critical event were perceived as important, to allow the operator to visualise/search for anomalies around the fault in historical data, and not only rely on the prediction algorithm. To enable the user to recognise critical trends more easily, it is recommended that spaces above and below certain thresholds be **colour-coded**.
- As it turns out, cockpits are an excellent means to gain a clear perception of the current status of activities. Nevertheless, **excessive overload in the presentation of the results can be risky** for a good understanding of the actual and relevant situation.

3.3 Data Management

Data collection, integration and quality requires significant effort and time in TransformingTransport. It has been estimated at around 80% by some pilots. Access to the data sources has turned out to be much more complicated than expected due to the following reasons: first, the number of **different sources** and data production and storage systems; secondly, the **access characteristics of data sources** – from a technical point of view, some of these sources and systems did not have the optimal flexibility. Using domain-specific data platforms (such as the BDV data platform project DataPorts: http://dataports-project.eu/) together with domain-specific machine learning components could significantly increase productivity in developing and deploying data analytics solutions.

Further lessons learned for data management are as follows:

- Concerning real-time analysis, tools have in many cases been implemented not as pure real-time but as **near-real-time** systems adapting the reaction time of the tools to the more lagged data-producing process. This is an important lesson because expecting pure real-time systems is nowadays far from easy due to ageing systems in several cases. This technology should be updated for further replications, mainly concerning big data projects, to take advantage of the new technologies.
- In order to provide services in real time, **extra storage** is required (which should be considered in the dimensioning phase of the system). This means that special care must be taken in defining optimised structures derived from the raw data that allows lower latency to process data. Additionally, in the case of databases, it is important to define appropriate indexes, reaching a compromise between the speed of writing in the database and reading from it. It has also been found that non-relational databases are more appropriate than traditional relational databases for evolving systems. Relational databases are more restrictive in their structure and do not allow rapid changes, offering advantages such as flexible schemas and better scaling (e.g. when new datasets are added and more fields in a table – or collection – are necessary, the addition is much easier in a non-relational database).
- One of the research goals was to identify **valuable data sources** that support the understanding of the different transport domains. Therefore, many different types of data and data sources were part of the pilots. These data sources differed in terms of format, timely availability and geographical spread (for pilots with large areas of action). One of the first things that many pilots learned was to abandon the idea of a holistic technical integration of all data sources. Data can also provide valuable insights when considered separately to some extent. Concerning visualisation, it was important to develop good use cases and to define the right data for them. Therefore, only useful data were used and further processed, which finally reduced complexity and increased understandability.
- The management of data required, in many cases, two approaches depending on whether processes required the use of raw datasets or processed datasets. Raw data were stored in file structures which were accessible to all workers. The parallelisation of computations was then organised such that each process task would use a different file from other processes, resulting in a mitigation of file access conflicts. Results were then stored in a variety of data structures that were capable of receiving data very quickly from multiple sources and enabled very fast search and retrieval times for records. Key was to have access to people who really know the data, because standardisation has not always been completed.
- **Data availability and fit for purpose:** Having data available on day 1 of the project does not mean it is fit for purpose (enough to answer the addressed business or operational needs) since technical access (interfaces) and organisational access (ownership) may require time to resolve. Because of this, first data analytics and visualisation goals must be defined and then it must be determined which data needs to be accessed and how, or vice versa.

3.4 Assessing the Impact of Big Data Technologies

As reported above, different lessons learned were collected for different technical concerns. However, such lessons learned were mostly qualitative. In order to complement these qualitative insights with quantitative measurements, TransformingTransport followed a stringent KPI measurement regime to demonstrate the transformative effects that big data could have on the transport sector through pilot projects in different countries, locations, transport modes and operating conditions. It applied big data for reshaping transport processes and services, increasing operational efficiency, improving customer experience and fostering new business models. As previously mentioned, data collection, integration and quality require significant effort and time, estimated at around 80% by some pilots mainly due to difficulties to be faced such as different data sources and storage characteristics. In this context, good and consistent data management is essential to improve operations.

A multi-criteria analysis (MCA) was designed specifically to assess the multiple impact levels of big data technologies implemented in the 13 different pilot cases of the project. The use of MCA appears to be an adequate option for simultaneously evaluating a certain number of both quantitative and qualitative criteria, some incommensurable, that ultimately need to be aggregated. MCA arose in the context of operations research (Charnes and Cooper 1977) and assessed alternatives on a set of criteria reflecting the decision-makers objectives, ranked based on an aggregation procedure. The scores achieved do not need to be translated into monetary terms but can simply be expressed in physical units or in qualitative terms (de Brucker et al. 2011). To make this method possible, a set of "Key Performance Indicators" (KPIs) were selected, defined as measurable figures able to shed light on how effective a certain application is. Applying the groundings of MCA, which enables the combination of both qualitative and quantitative aspects, TransformingTransport developed a methodology of assessing a high number of indicators pertaining to entirely different transport sectors (Velazquez et al. 2018) and Assessment Categories of major relevance, i.e. operational efficiency, asset management, environmental quality, energy consumption and safety. These categories have been used to perform a complete assessment of the different pilots and manage data collected through pilot-only evaluation and then – in a transversal way across pilots – a comparison between them.

The large differences among pilots and domains have led to the creation of a specific methodology out of which the analysis of results showed the impacts of the tested technological improvements. Throughout several consciously selected KPIs, it has been possible to assess the benefits of big data implementation on the transportation sector. Then, a four-level assessment was carried out. The first level consists of the evaluation of each pilot individually for each of the Assessment Categories, after an aggregation process. The second level goes through the analysis of the aggregated achievements within the same pilot domain, comparing the performance of the pilots within the domain. Therefore, the effects of big data in

the same mode in different settings and conditions are analysed. The third level of the evaluation is the transversal assessment of the pilots for each category; the goal was to perform a comparative analysis through the different pilots on each of the aspects, e.g. how operational efficiency or energy savings vary among them. The fourth assessment is the strategic level, for which only the most relevant KPIs for each pilot are considered (Vázquez et al. 2020).

The evaluation procedure analyses the impact of the big data implementation over different transport sectors, by comparing KPI final measurements with the original ones. There is thus a four-level assessment comparison between two scenarios: the reference scenario before leveraging the big data technology (baseline or ex ante scenario) and the scenario once the technologies have been introduced (big data technology scenario) (Velazquez et al. 2018). The results of this assessment reveal improvements of around 40-60% regarding the operative cost, energy consumption, environmental quality and enhancement of the predictive maintenance of assets, among others. Big data technologies have demonstrated their usefulness when it comes to gaining deeper insights from the huge quantity of data to boost the different transport processes.

Effective and consistent data management is essential to improve transport operations. A further lesson learned from TransformingTransport is that due to the huge volume and variety of data and data sources, a coherent, in-depth and integrated approach for data management and analysis is necessary.

3.5 Use Case Conclusion

As can be concluded from the use case presented above, big data technologies promise to deliver profound economic and societal impact in mobility and logistics. TransformingTransport pursued big data use cases in all areas of major importance for the mobility and logistics sector in Europe, demonstrating the technical and economic viability of big data for reshaping transport processes and services. TransformingTransport employed predictive data analytics and predictive maintenance as the key enabling big data technologies to bring about this transformation.

The significant growth of transport data volumes and the rates at which such data is generated will be an important driver for the next level of technology innovation in transport: data-driven Artificial Intelligence (AI). Data-driven AI has a tremendous potential to benefit European citizens, economy and society (Sonja Zillner et al. 2018; Zillner et al. 2020). From an industrial point of view, AI means algorithm-based and data-driven computer systems that enable machines and people with digital capabilities such as perception, reasoning, learning and even autonomous decision-making. AI will facilitate software to draw conclusions, learn, adapt and adjust parameters accordingly. With recent advances in computing power, connectivity and algorithms, AI is making great strides. With today's promising results in using AI technology, we can expect the next level of efficiency and operational improvements in the mobility and transport sectors in Europe.

4 Summary

The Big Data Value Reference Model has been developed with input from technical experts and stakeholders along the whole big data value chain. The BDV Reference Model may serve as a common reference framework to locate big data technologies on the overall IT stack. This chapter elaborated the various elements (both horizontal and vertical) of the framework and illustrated how it might be used to map technical elements stemming from research and innovation projects. Complementing this application of the reference model, it has also been used to systematically monitor the technical progress of the Big Data Value PPP. To determine how well the technical priorities and challenges are covered by ongoing research and innovation activities, the BDVA performed a systematic collection of data, where the BDV Reference Model provided the structure for a common data collection template and frame for data analysis.

Acknowledgements Research leading to these results received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement nos. 732630 (BDVe), 731932 (TransformingTransport) and 871493 (DataPorts). This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under grant no. SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

References

Castiñeira, R., & Metzger, A. (2018, April). *The transforming transport project – Mobility meets big data*. https://doi.org/10.5281/zenodo.1484954

Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.). (2016). *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*. https://doi.org/10.1007/978-3-319-21569-3

Charnes, A., & Cooper, W. W. (1977). Goal programming and multiple objective optimisations: Part 1. *European Journal of Operational Research*. https://doi.org/10.1016/S0377-2217(77)81007-2

Curry, E. (2020). *Real-time linked dataspaces*. https://doi.org/10.1007/978-3-030-29665-0

Curry, E., & Ojo, A. (2020). Enabling knowledge flows in an intelligent systems data ecosystem. In *Real-time Linked Dataspaces* (pp. 15–43). https://doi.org/10.1007/978-3-030-29665-0_2

de Brucker, K., Macharis, C., & Verbeke, A. (2011). Multi-criteria analysis in transport project evaluation: An institutional approach. *European Transport - Trasporti Europei*.

Feltus, C., Proper, H., Metzger, A., Lopez, J., & Castineira, R. (2018). *Value CoCreation (VCC) Language design in the frame of a smart airport network case study* (pp. 858–865). https://doi.org/10.1109/AINA.2018.00127

G. Press. (2016). *Cleaning big data: Most time-consuming, least enjoyable data science task, survey says*.

Germany and the Netherlands. (2017). *Joint position paper on the European open science cloud*.

Lund, S., Manyika, J., Nyquist, S., Mendonca, L., & Ramaswamy, S. (2013). *Game changers: Five opportunities for US growth and renewal*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011*). Big data: The next frontier for innovation, competition, and productivity*. Retrieved from McKinsey

Global Institute website http://scholar.google.com/scholar.bib?q=info:kkCtazs1Q6wJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,47&ct=citation&cd=0

Metzger, A., Thornton, J., Valverde, F., Lopez, J. F. G., & Rublova, D. (2019a). *TransformingTransport, Predictive analytics and predictive maintenance innovation via big da-ta: The case of TransformingTransport*. In 13th Intelligent Transport Systems - European Congress (ITS Europe), Brainport-Eindhoven, The Netherlands, June 3–6.

Metzger, A., Neubauer, A., Bohn, P., & Pohl, K. (2019b). Proactive process adaptation using deep learning ensembles. In P. Giorgini & B. Weber (Eds.), *Advanced information systems engineering* (pp. 547–562). Cham: Springer.

Metzger, A., Franke, J., & Jansen, T. (n.d.). Data-driven deep learning for proactive terminal process management. In J. V. Brocke, J. Mendling, & M. Rosemann (Eds.), *Business process management cases* (Vol. 2). New York: Springer.

Palm, A., Metzger, A., & Pohl, K. (2020). Online reinforcement learning for self-adaptive information systems. In S. Dustdar, E. Yu, C. Salinesi, D. Rieu, & V. Pant (Eds.), *Advanced Information systems engineering* (pp. 169–184). Cham: Springer.

Raskin, J. (2000). *Humane interface, the: New directions for designing interactive systems*. Addison-Wesley Professional.

Vázquez, P., Monzon, A., Boggio-Marzet, A., & Corral, V. (2020). Assessing the impact of big data for improving transport efficiency: A cross-modal approach. In *8th Transport Research Arena TRA 2020*. Helsinki, Finland.

Velazquez, G., Monzon, A., & Roman, A. (2018). Big Data value for improving transport performance in all modes, an assessment methodology. In *7th Transport Research Arena TRA 2018*. Vienna, Austria.

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Zillner, S, Curry, E., Metzger, A., & Auer, S. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Retrieved from Big Data Value Association website www.bdva.eu

Zillner, S, Gómez, J. A., Robles, A. G., & Curry, E. (2018). *Data for artificial intelligence for European economic competitiveness and societal progress – BDVA position statement*.

Zillner, S., Bisset, D., Milano, M., Curry, E., Hahn, T., Lafrenz, R., et al. (2020). *Strategic research, innovation and deployment agenda - AI, data and robotics partnership. Third Release (3rd)*. Brussels: BDVA, euRobotics, ELLIS, EurAI and CLAIRE.