Technical research priorities for big data

DOI:
10.1007/978-3-030-68176-0_5

Document status and date:
Published: 01/01/2021

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Technical Research Priorities for Big Data

Edward Curry, Sonja Zillner, Andreas Metzger, Arne J. Berre, Sören Auer, Ray Walshe, Marija Despenic, Milan Petkovic, Dumitru Roman, Walter Waterfeld, Robert Seidl, Souleiman Hasan, Umair ul Hassan, and Adegboyega Ojo

Abstract To drive innovation and competitiveness, organisations need to foster the development and broad adoption of data technologies, value-adding use cases and sustainable business models. Enabling an effective data ecosystem requires overcoming several technical challenges associated with the cost and complexity of management, processing, analysis and utilisation of data. This chapter details a community-driven initiative to identify and characterise the key technical research priorities for research and development in data technologies. The chapter examines the systemic and structured methodology used to gather inputs from over 200 stakeholder organisations. The result of the process identified five key technical research priorities in the areas of *data management*, *data processing*, *data analytics*, *data*

E. Curry (✉) · S. Hasan · U. ul Hassan · A. Ojo
Insight SFI Research Centre for Data Analytics, NUI Galway, Galway, Ireland
e-mail: edward.curry@nuigalway.ie

S. Zillner
Siemens AG, Munich, Germany

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

A. J. Berre · D. Roman
SINTEF Digital, Oslo, Norway

S. Auer
Leibniz Universität Hannover, Hannover, Germany

R. Walshe
ADAPT SFI Centre for Digital Content, Dublin City University, Dublin, Ireland

M. Despenic
ABN AMRO Bank, Amsterdam, the Netherlands

M. Petkovic
Philips and Eindhoven University of Technology, Eindhoven, the Netherlands

W. Waterfeld
Saarbrücken, Germany

R. Seidl
Nokia Bell Labs, Munich, Germany

E. Curry et al. (eds.), *The Elements of Big Data Value*,
https://doi.org/10.1007/978-3-030-68176-0_5

visualisation and user interactions, and *data protection,* together with 28 sub-level challenges. The process also highlighted the important role of data standardisation, data engineering and DevOps for Big Data.

Keywords Research challenges · Data management · Data processing · Data analytics · Data visualisation · User interactions · Data protection · Data standardisation · Data ecosystem

1 Introduction

The expectations in refining data as the new oil of the twenty-first century are currently so high that virtually no business can afford not to have a big data project that 'unlocks' the value in their data (Chen et al. 2012). There is a noticeable increase in the adoption of data-driven business scenarios in sectors other than the web-based 'traditional' big data companies such as Google, Yahoo, Facebook and Twitter (Lavalle et al. 2011). However, many sectors still struggle with the adoption of data technologies, often due to a lack of expertise, regulatory barriers and unclear business value. This is especially true in non-IT-focused sectors, such as the energy sector that struggles with the adoption of data technologies (Rusitschka and Curry 2016). The benefits of sharing and linking data across domains and industry are apparent. Initiatives such as Smart Cities are showing how different sectors (i.e. energy and transport) can collaborate to maximise the potential for optimisation and value return (*Communication: A European strategy for data* 2020). The cross-fertilisation of stakeholders and datasets from different sectors is a key element for advancing the data economy.

To support the emergence of a data ecosystem, it was important that the different actors within the ecosystem 'define a shared vision and jointly identify gaps in the current data landscape' (DG Connect 2013). Data ecosystems face several problems such as data discovery, curation, linking, synchronisation, distribution, business modelling, sales and marketing (José María Cavanillas et al. 2016). To address these issues, the Big Data Value contractual Public-Private Partnership (BDV PPP) between the European Commission and the Big Data Value Association aimed to strengthen the data value chain (Curry 2016), foster cooperation in data research and innovation, enhance community building around data and set the groundwork for a thriving data-driven economy in Europe. The BDV PPP was driven by the conviction that research and innovation focusing on a combination of business and usage needs is the best long-term strategy to deliver value from big data and create jobs and prosperity. An essential requirement was to identify and characterise the key technical research challenges that need to be tackled to enable a data ecosystem.

This chapter identifies the key technical research priorities for research and development in data technologies. It presents the results of an investigation and consultation process that was conducted to capture the priorities for big data in public and private organisations across Europe. The chapter starts with an introduction to the methodology for the identification and prioritisation of the technical challenges for the adoption of data technologies. The chapter details the key challenges and outcomes needed in terms of data management, data processing, data analytics, data visualisation and user interaction, and data protection. It highlights the role of standardisation to further the development of data technology and the key role of data standards. Challenges with data engineering and DevOps for big data systems ensure productivity and quality are detailed. Finally, the chapter presents a scenario from the healthcare sector to emphasise the importance of adopting better big data strategies.

2 Methodology

In order to correctly identify the technical research priorities a systemic and structured methodology was needed to gather inputs from over 200 stakeholder organisations. The methodology built on and extended an established roadmapping methodology to gather consensus from a range of stakeholders (Curry et al. 2016). The key phases in the methodology, as illustrated in Fig. 1, are (a) technology state of the art and sector analyses, (b) subject matter expert interviews, (c) stakeholder workshops, (d) requirements consolidation and (e) community survey.

2.1 Technology State of the Art and Sector Analysis

The goal of the first phase was to identify the sectorial needs and requirements gathered from different stakeholders and the state of the art of data technologies, as well as identifying research challenges. As part of the investigation, application sectors expressed their need for the technology as well as possible limitations and expectations regarding its current and future deployment. The first step was to perform a systematic literature review based on the following activities:

- Identification of relevant type and sources of information
- Analysis of key information in each source
- Identification of key topics for each technical working group
- Identification of the key subject matter experts for each topic as potential interview candidates

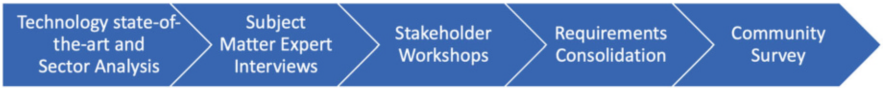

Technology state-of-the-art and Sector Analysis
Subject Matter Expert Interviews
Stakeholder Workshops
Requirements Consolidation
Community Survey

Fig. 1 The workflow of research methodology

- Synthesisation of the key message of each data source into state-of-the-art descriptions for each identified topic

The following types of data sources were used: scientific papers published in workshops, symposia, conferences, journals and magazines, company white papers, technology vendor websites, open-source projects, online magazines, analysts' data, web blogs other online sources and interviews. The groups focused on sources that mention concrete technologies and analysed them concerning their values and benefits. The synthesis step compared the key messages and extracted agreed views. Topics were prioritised based on the degree to which they can address business needs.

2.2 *Subject Matter Expert Interviews*

The literature survey was complemented by a series of interviews with subject matter experts for relevant topic areas. Subject matter expert interviews are a technique well suited to data collection and particularly for exploratory research because it allows extensive discussions that illuminate factors of importance (Oppenheim 1992; Yin 2013). The information gathered is likely to be more accurate than information collected by other methods since the interviewer can avoid inaccurate or incomplete answers by explaining the questions to the interviewee (Oppenheim 1992). The interviews followed a semi-structured protocol. The topics of the interview covered different aspects of big data:

- Goals of big data technology
- Beneficiaries of big data technology
- Drivers and barriers for data technologies
- Technology and standards for data technologies

Interviewees were selected to be representative of the different stakeholders within the data ecosystem. The selection of interviewees covered (1) established providers of big data technology (typically MNCs), (2) innovative sectorial players who are successful at leveraging big data, (3) new and emerging SMEs in the big data space and (4) world-leading academic authorities in technical areas related to the big data value chain.

The data collection and the analysis strategy were inspired by the triangulation approach (Flick 2004). Reviewing and quantitatively assessing the high-level application scenarios derived a reliable analysis of user needs. Examinations of the likely constraints of big data applications helped to identify the relevant requirements that needed to be addressed.

2.3 Stakeholder Workshops

The third step involved a cross-check and validation of the initial results of the first two steps by involving stakeholders from multiple domains in dedicated workshops and webinars to discuss and review the outcomes. Multiple workshops and consultations took place to ensure the most comprehensive representation of views and positions, including the full range of public and private sector entities not only from technology provision but also technology adoption. Sectoral workshops were conducted in various fields: geospatial/environment, energy, media, mobility, manufacturing, retail, health and the public sector. The purpose was to identify the main priorities with approximately 200 organisations and other relevant stakeholders physically participating and contributing. A wide range of stakeholders contributed to the process with inputs and analysis from SMEs and large enterprises, public organisations, and research and academic institutions. They included suppliers and service providers, data owners and early adopters of big data in many sectors. Extensive analysis reports were then produced, which helped both formulate and reformulate the identified requirements. From the analysis of the results, it was clear that addressing the technical needs of these vertical application markets required a set of cross-sector technologies.

2.4 Requirement Consolidation

Comparison among the different sectors enabled the identification of commonalities and differences at multiple levels. The analysis was used to define integrated cross-sectorial priorities that provide a coherent, holistic view of the big data domain and establish a common understanding of requirements, as well as technology descriptions and terms used across domains. A consolidated description was established to align the sector-specific labelling of requirements. In doing so, each sector provided its requirements with the associated user needs. Thus, the initial list of 13 high-level requirements and 28 sub-level requirements could be reduced to 5 high-level requirements and 20 sub-level requirements.

2.5 Community Survey

The objective of the community survey was to engage with the broader community to ensure a comprehensive perspective concerning the technical and business impact of the identified technical priorities, as well as to identify emerging priorities with high impact for the European big data economy. An inclusive approach was taken to ensure stakeholder engagement, with inputs actively solicited from the wider community composed of experts in technical domains as well as in business sectors. The

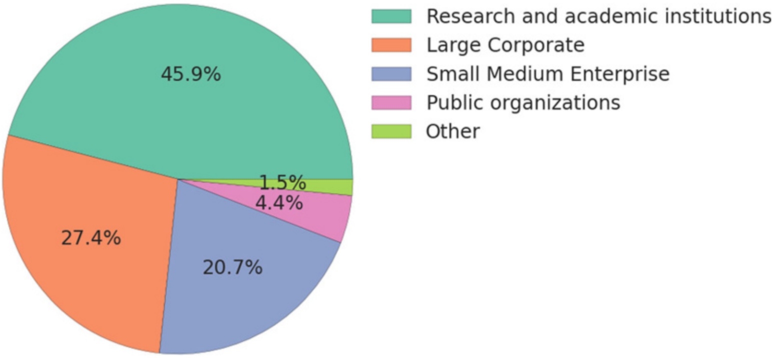


Fig. 2 Distribution of participants in terms of the type of organisation

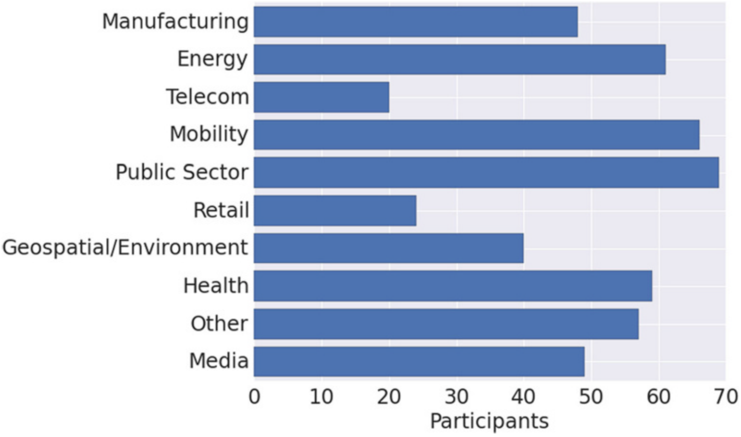


Fig. 3 Number of organisations associated with different sectors

survey received participation from a wide range of organisations. In total, 135 organisations responded to the survey through their representatives.

Figure 2 shows the distribution of participants in terms of the type of organisation. The majority of participant organisations (almost 95%) were either private companies or research and academic institutions. The response indicates a broader interest and contribution from stakeholders in shaping the future of the European big data community.

Figure 3 shows the number of organisations working in various sectors. In general, the organisations identified themselves as being active in multiple sectors, which underlines the cross-sectoral perspectives on the technical and non-technical priorities of big data as identified by the survey. Figure 4 shows that more than 70% of the participants chose two or more sectors. On average, more than three different sectors were chosen by participants to indicate the diversity of their portfolio. This

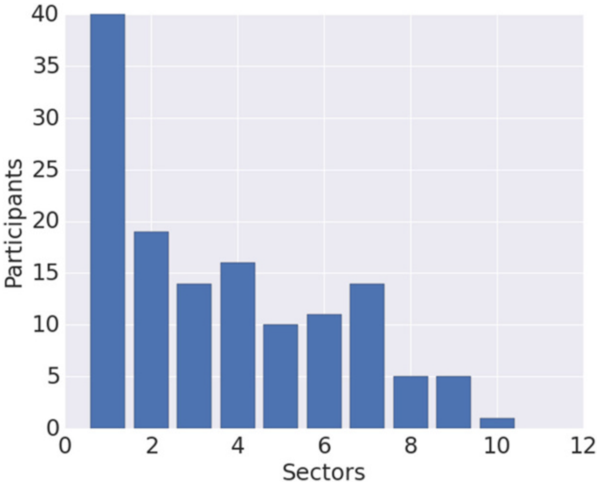


Fig. 4 Histogram of the number of sectors per organisation

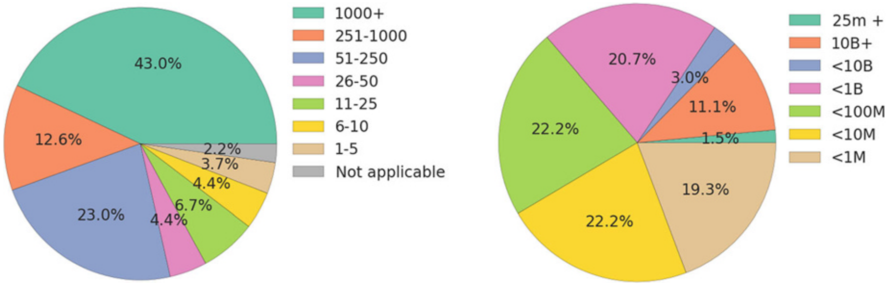


Fig. 5 Composition of participating organisations in terms of number of employees (left) and annual revenue (right)

also highlights the need to consider the multidisciplinary nature of the big data economy.

To quantify the size of the organisation, the survey participants were asked to indicate the number of employees (full-time equivalent) and annual revenue. Figure 5 summarises the composition of participating organisations in terms of employees and revenue. Primarily due to participation from the public sector and large corporates, the majority of organisations have more than 200 employees and revenue higher than 10 million. It should be noted that big data challenges for companies with more than 1000 employees are not only limited to their specific sectors but also in their day-to-day operations, such as human resource management and finance. The following section discusses the technical priorities for data technologies, in addition to their ranking based on the community survey.

3 Research Priorities for Big Data Value

The first three steps of the methodology produced a set of consolidated cross-sectorial technical research requirements. The result of this process was the identification of five key technical research priorities as illustrated in Fig. 6 (data management, data processing architectures, deep analytics, data protection and pseudonymisation, advanced visualisation and user experience), together with 28 sub-level challenges to delivering big data value. In this section, we report on the results of the survey to identify a prioritisation of the cross-sectorial requirements. As far as possible, the roadmaps were quantified using the results of the survey to allow for well-founded prioritisation and action plans, as illustrated in Fig. 7. The remainder of this chapter summaries the technical priorities as defined in the Strategic Research and Innovation Agenda (SRIA) of the BDVA (Zillner et al. 2017).

3.1 Priority 'Data Management'

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,

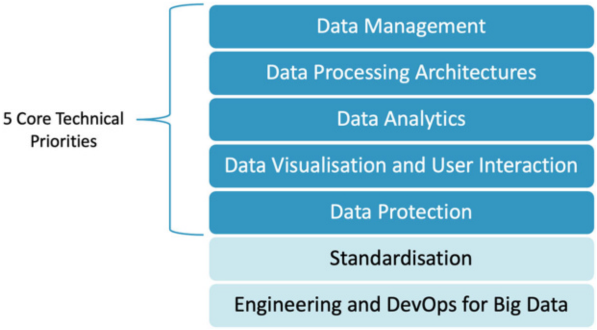

5 Core Technical Priorities
Data Management
Data Processing Architectures
Data Analytics
Data Visualisation and User Interaction
Data Protection
Standardisation
Engineering and DevOps for Big Data

Fig. 6 High-level technical priorities for data technologies

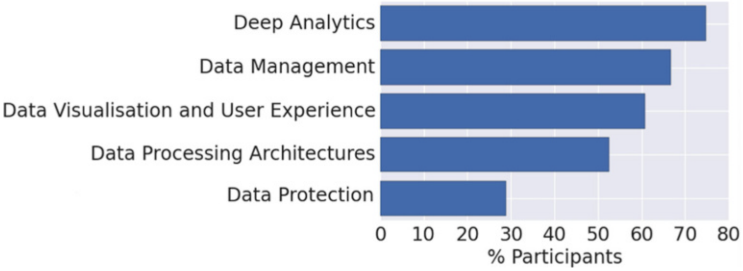


Fig. 7 Distribution of high-level technical priorities across participants

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 formats, such as reports, Web 2.0 data, images, sensor data, mobile data, geospatial data and multimedia data. For instance, 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 a 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.

To ensure a valuable big data analytics outcome, the incoming data has to be 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, or DaaS).

3.1.1 Challenges

As of today, collected data is rapidly increasing; however, the methods and tools for data management are not evolving at the same pace. From this perspective, it becomes crucial to have – at a minimum – good metadata, Natural Language Processing (NLP), and semantic techniques to structure the datasets and content, annotate them, document the associated processes, and deliver or syndicate information to recipients. The following research challenges have been identified:

- **Semantic annotation of unstructured and semi-structured data:** Data needs to be semantically annotated in digital formats, without imposing extra effort on data producers. In particular, unstructured data, such as videos, images or text in a natural language (including multilingual text), or specific domain data, such as Earth observation data, have to be pre-processed and enhanced with semantic annotation.
- **Semantic interoperability:** Data silos have to be unlocked by creating interoperability standards and efficient technologies for the storage and exchange of semantic data and tools to allow efficient user-driven or automated annotations and transformations.
- **Data quality:** Methods for improving and assessing data quality have to be created, together with curation frameworks and workflows. Data curation methods might include general-purpose data curation pipelines, online and offline data filtering techniques, improved human–data interaction, and standardised data curation models and vocabularies, as well as ensuring improved integration between data curation tools.
- **Data lifecycle management and data governance:** With the tremendous increase in data, integrated data lifecycle management is facing new challenges in handling the sheer size of data, as well as enforcing consistent quality as the data grows in volume, velocity and variability, including providing support for real-time data management and efficiency in data centres. Furthermore, as part of the data lifecycle, data protection and management must be aligned. Control, auditability and lifecycle management are key for governance, cross-sector applications and the General Data Protection Regulation (GDPR).
- **Integration of data and business processes:** This relates to a conceptual and technically sound integration of results from the two ‘worlds’ of analytics. Integrating data processes, such as data mining or business intelligence, on the one side, with business processes, such as process analysis in the area of Business Process Management (BPM), on the other side, is needed.
- **Data-as-a-Service:** The issue here is how to bundle both the data and the software and data analytics needed to interpret and process them into a single package that can be provided as an (intermediate) offering to the customer.
- **Distributed trust infrastructures for data management:** Mechanisms are required to enforce consistency in transactions and data management, for example, based on distributed ledger/blockchain technologies. Flexible data

management structures are based on microservices with the possibility of integrating data transformations, data analysis and data anonymisation, in a decentralised manner.

3.1.2 Outcomes

The main expected advances in data management are as follows:

- Languages, techniques and tools for measuring and ensuring data quality (such as novel data management processing algorithms and data quality governance approaches that support the specifics of big data) and for assessing data provenance, control and IPRs.
- Principles for a clear Data-as-a-Service (DaaS) model and paradigm fostering the harmonisation of tools and techniques with the ability to easily reuse, interconnect, syndicate, auto/crowd annotate and bring to life data management use cases and services across sectors, borders and citizens by decreasing the costs of developing new solutions. Furthermore, trusted and flexible infrastructures need to be developed for the DaaS paradigm, potentially based on technologies such as distributed ledgers, blockchains and microservices.
- Methods and tools for a complete data management lifecycle, ranging from data curation and cleaning (including pre-processing veracity, velocity integrity and quality of data) and using scalable big data transformations approaches (including aspects of automatic, interactive, sharable and repeatable transformations), to long-term data storage and access. New models and tools to check integrity and veracity of data, through both machine-based and human-based (crowdsourcing) techniques. Furthermore, mechanisms need to be developed for the alignment of data protection and management, addressing aspects such as control, auditability and lifecycle management of data.
- Methods and tools for the sound integration of analytics results from data and business processes. This relies on languages and techniques for semantic interoperability such as standardised data models and interoperable architectures for different sectors enriched through semantic terminologies. Particularly important are standards and multilingual knowledge repositories/sources that allow industries and citizens to seamlessly link their data with others. Mechanisms to deal with semantic data lakes and industrial data spaces and the development of enterprise knowledge graphs are of high relevance in this context.
- Techniques and tools for handling unstructured and semi-structured data. This includes natural language processing for different languages and algorithms for the automatic detection of normal and abnormal structures (including automatic measuring, tools for pre-processing and analysing sensor, social, geospatial, genomics, proteomics and other domain-orientated data), as well as standardised annotation frameworks for different sectors supporting the technical integration of different annotation technologies and data formats.

3.2 Priority 'Data Processing Architectures'

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 affect all layers of future big data infrastructures, ranging from the specifications of low-level data, to 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 leverages their true potential. Optimised frameworks and toolboxes to enable the best use of both data-in-motion (e.g. data streams from sensors) and data-at-rest 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 the predictive models based on data-at-rest 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 can focus on the development of business models.

The capability of existing systems to process such data-in-motion and answer queries in real time and for thousands of concurrent users is 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.

3.2.1 Challenges

There have been several advances in big data analytics to support the dimension of big data volume. In a separate development, stream processing has been enhanced in terms of analytics on the fly to cover the velocity aspect of big data. This is especially important as business needs to know what is happening now. The main challenges to be addressed are:

- **Heterogeneity:** Big data processing architectures form places to gather and process various pieces of relevant data together. Such data can vary in several aspects, including different syntactic formats, heterogeneous semantic representations and various levels of granularity. In addition, data can be structured, semi-structured or unstructured, or multimedia, audio-visual or textual. Hardware can also be heterogeneous (CPUs, GPUs and FPGAs). Having the ability to handle big data's variety and uncertainty over several dimensions is a challenge for big data processing architectures.
- **Scalability:** Being able to apply storage and complex analytics techniques at scale is crucial to extract knowledge out of the data and develop decision-support applications. For instance, predictive systems such as recommendation engines must be able to provide real-time predictions while enriching historical databases to continuously train more complex and refined statistical models. The analytics must be scalable, with low latency adjusting to the increase of both the streams and volume of big datasets.
- **Processing of data-in-motion and data-at-rest:** Real-time analytics through event processing and stream processing, spanning inductive reasoning (machine learning), deductive reasoning (inference), high-performance computing (data centre optimisation, efficient resource allocation, quality of service provisioning) and statistical analysis, has to be adapted to allow continuous querying over streams (i.e. online processing). The scenarios for big data processing also require a greater ability to cope with systems which inherently contain dynamics in their daily operation, alongside their proper management, to increase operational effectiveness and competitiveness. Most of these processing techniques have only been applied to data-at-rest and in some cases to data-in-motion. A challenge here is to have suitable techniques for data-in-motion and also integrated processing for both types of data at the same time.
- **Decentralisation:** Big data producers and consumers can be distributed and loosely coupled as in the Internet of Things. Architectures have to consider the effect of distribution on the assumptions underlying them, such as loose data agreements and missing contextual data. The distribution of big data processing nodes poses the need for new big data-specific parallelisation techniques, and (at least partially) the automated distribution of tasks over clusters is a crucial

element for effective stream processing. Especially important is efficient distribution of the processing to the Edge (i.e. local data Edge processing and analytics), as a part of the ever-increasing trend of Fog computing.

- **Performance:** The performance of algorithms has to scale up by several orders of magnitude while reducing energy consumption compatible with the best efforts in the integration between hardware and software. It should be possible to utilise existing and emerging high-performance-computing and hardware-oriented developments, such as main memory technology, with different types of caches, such as Cloud and Fog computing, and software-defined storage with built-in functionality for computation near the data (e.g. Storlets). Also to be utilised are data availability guarantees to avoid unnecessary data downloading and archiving, and data reduction to support storing, sharing and efficient in-place processing of the data.
- **Novel architectures for enabling new types of big data workloads (hybrid big data and HPC architecture):** Some selected domains have shown a considerable increase in the complexity of big data applications, usually driven by computation-intensive simulations, which are based on complex models and generate enormous amounts of output data. On the other hand, users need to apply advanced and highly complex analytics and processing to this data to generate insights, which usually means that data analytics needs to take place in situ, using complex workflows and in synchrony with computing platforms. This requires novel big data architectures which exploit the advantages of HPC infrastructure and distributed processing, and includes the challenges of maintaining efficient distributed data access (enabling the scaling of deep learning applications) and efficient energy consumption models in such architectures.
- **The introduction of new hardware capabilities:** Computing capacity has become available to train larger and more complex models more quickly. Graphics processing units (GPUs) have been repurposed to execute the data and algorithm crunching required for machine learning at speeds many times faster than traditional processor chips. In addition, Field Programmable Gate Arrays (FPGAs) and dedicated deep learning processors are influencing big data architectures.

3.2.2 Outcomes

The main expected advances in data processing architectures are:

- **Techniques and tools for processing real-time heterogeneous data sources:** The heterogeneity of data sources for both data-at-rest and data-in-motion requires efficient and powerful techniques for transformation and migration. This includes data reduction and mechanisms to attach and link to arbitrary data. Standardisation also plays a key role in addressing heterogeneity.
- **Scalable and dynamic data approaches:** The capabilities for processing very large amounts of data in a very short time (in real-time applications and/or

reacting to dynamic data) and analysing sizable amounts of data to update the analysis results as the information content changes. It is important to access only relevant and suitable data, thereby avoiding accessing and processing irrelevant data. Research should provide new techniques that can speed up training on large amounts of data, for example by exploiting parallelisation, distribution and flexible Cloud computing platforms, and by moving computation to Edge computing.

- **Real-time architectures for data-in-motion:** Architectures, frameworks and tools for real-time and on-the-fly processing of data-in-motion (e.g. IoT sensor data) and integrating it with data-at-rest. Furthermore, there is a need to dynamically reconfigure such architectures and dynamic data processing capabilities on the fly to cope with, for example, different contexts, changing requirements and optimisation in various dimensions (e.g. performance, energy consumption and security).
- **Decentralised architectures:** Architectures that can deal with the big data produced and consumed by highly decentralised and loosely coupled parties such as in the Internet of Things, with secure traceability such as blockchain. Additionally, architectures with parallelisation and distributed placement of processing for data-in-motion and its integration with data-at-rest.
- **Efficient mechanisms for storage and processing:** Real-time algorithms and techniques are needed for requirements demanding low latency when handling data-in-motion. Developing hardware and software together for Cloud and high-performance data platforms will, in turn, enable applications to run agnostically with outstanding reliability and energy efficiency.
- **Hybrid big data and high-performance computing architecture:** Efficient hybrid architectures that optimise the mixture of big data (i.e. Edge) and HPC (i.e. central) resources – combining local and global processing – to serve the needs of the most extreme and/or challenging data analytics at scale, called high-performance data analytics (HPDA).

3.3 Priority 'Data Analytics'

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 have a positive influence on all parts of the data value chain, 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 substantial 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 recent 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 production (Lund et al. 2013).

The next generation of analytics needs 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 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, aside from 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 massive datasets. These can be used for iterative refinements of the models, but also for providing new data analytics services which can process massive datasets.

3.3.1 Challenges

Understanding data, whether it is numbers, text or multimedia content, has always been one of the most significant challenges for data analytics. Entering the era of big data, this challenge has expanded to a degree that makes the development of new methods necessary. The following list details the research areas identified for data analytics:

- **Semantic and knowledge-based analysis:** Improvements in the analysis of data to provide a near-real-time interpretation of the data (i.e. sentiment, semantics, etc.). Also, ontology engineering for big data sources, interactive visualisation and exploration, real-time interlinking and annotation of data sources, scalable and incremental reasoning, linked data mining and cognitive computing.
- **Content validation:** Implementation of veracity (source reliability/information credibility) models for validating content and exploiting content recommendations from unknown users.
- **Analytics frameworks and processing:** New frameworks and open APIs for the quality-aware distribution of batch and stream processing analytics, with minimal development effort from application developers and domain experts. Improvement in the scalability and processing speed of the algorithms mentioned above to tackle linearisation and computational optimisation issues.
- **Advanced business analytics and intelligence:** All of the above items enable the realisation of real and static business analytics, as well as business intelligence

empowering enterprises and other organisations to make accurate and instant decisions to shape their markets. The simplification and automation of these techniques are necessary, especially for SMEs.

- **Predictive and prescriptive analytics:** Machine learning, clustering, pattern mining, network analysis and hypothesis testing techniques applied on extremely large graphs containing sparse, uncertain and incomplete data. Areas that need to be addressed are building on the results of related research activities within the current EU work programme, sector-specific challenges and contextualisation combining heterogeneous data and data streams via graphs to improve the quality of mining processes, classifiers and event discovery. These capabilities open up novel opportunities for predictive analytics in terms of predicting future situations, and even prescriptive analytics providing actionable insights based on forecasts.
- **High-performance data analytics:** Applying high-performance computing techniques to the processing of extremely large amounts of data. Taking advantage of a high-performance infrastructure that powers different workloads and starting to support workflows that accelerate insights and lead to improved business results for enterprises. The goal is to develop new data analytics services with workloads typically characterised as follows: insights derived from analysis or simulations that are extremely valuable; the time-to-insight must be extremely fast; models and datasets are exceptionally complex.
- **Data analytics and Artificial Intelligence:** Machine-learning algorithms have progressed in recent years, primarily through the development of deep learning and reinforcement-learning techniques based on neural networks. The challenge is to make use of this progress in efficient and reliable data analytics processes for advanced business applications. This includes the intelligent distribution of the processing steps, from very close to data sources to Cloud (e.g. distributed deep learning). In addition, different techniques from AI can be used to enable better reasoning about data analytics' processes and outcomes.

3.3.2 Outcomes

The main expected advanced analytics innovations are as follows:

- **Improved models and simulations:** Improving the accuracy of statistical models by enabling fast non-linear approximations in very large datasets. Moving beyond the limited samples used so far in statistical analytics to samples covering the whole or the largest part of an event space/dataset.
- **Semantic analysis:** Deep learning, contextualisation based on AI, machine learning, natural language and semantic analysis in near real time. Providing canonical paths so that data can be aggregated and shared easily without dependency on technicians or domain experts. Enabling smart analysis of data across and within domains.

- **Event and pattern discovery:** Discovering and predicting rare real-time events that are hard to identify since they have a small probability of occurrence, but a great significance (such as physical disasters, a few costly claims in an insurance portfolio, rare diseases and treatments).
- **Multimedia (unstructured) data mining:** The processing of unstructured data (multimedia, text) Linking and cross-analysis algorithms to deliver cross-domain and cross-sector intelligence.
- **Deep learning techniques for business intelligence:** Coupled with the priorities on visualisation and engineering, providing user-friendly tools which connect to open and other datasets and streams (including a citizen's data), offering intelligent data interconnection for business- and citizen-orientated analytics, and allowing visualisation (e.g. diagnostic, descriptive and prescriptive analytics).
- **HPDA reference applications**: Well-defined processes for realising HPDA scenarios. Through enabling the combination of models (so-called Digital Twins) with the real-time operation of complex products/systems to more speedily project the inferences from (Big-Data-based) real-time massive data streams into (HPC-based) models and simulations (processing terabytes per minute/hour to petabytes of data per instance), the temporal delta between as-designed and as-operated can be reduced considerably.

3.4 Priority 'Data Visualisation and User Interaction'

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

3.4.1 Challenges

In the data visualisation and user interaction domain, the tools that are currently used to communicate information need to be improved due to the significant changes brought about by the expanding volume and variety of big data. Advanced visualisation techniques must therefore consider the range of data available from diverse domains (e.g. graphs or geospatial, sensor and mobile data). Tools need to support user interaction for the exploration of unknown and unpredictable data within the visualisation layer. The following list briefly outlines the research areas identified for visualisation and user interaction:

- **Visual data discovery:** Access to information is at present based on a user-driven paradigm: the user knows what they need, and the only issue is to define the right criteria. With the advent of big data, this user-driven paradigm is no longer the most efficient. Data-driven paradigms are needed in which information is proactively extracted through data discovery techniques, and systems anticipate the user's information needs.
- **Interactive visual analytics of multiple-scale data:** There are significant challenges in visual analytics in the area of multiple-scale data. Appropriate scales of analysis are not always clear in advance, and single optimal solutions are unlikely to exist. Interactive visual interfaces have great potential for facilitating the empirical search for acceptable scales of analysis and the verification of results by modifying the scale and the means of any aggregation.
- **Collaborative, intuitive and interactive visual interfaces:** What is needed is an evolution of visual interfaces towards their becoming more intuitive and exploiting the advanced discovery aspects of big data analytics. This is required to foster effective exploitation of the information and knowledge that big data can deliver. In addition, there are significant challenges for effective communication and visualisation of big data insights to enable collaborative decision-making processes in organisations.
- **Interactive visual data exploration and querying in a multi-device context:** A key challenge is the provisioning of cross-platform mechanisms for data exploration, discovery and querying. Some difficult problems are how best to deal with uniform data visualisation on a range of devices and how to ensure access to functionalities for data exploration, discovery and querying in multi-device settings, requiring the exploration and development of new approaches and paradigms.

3.4.2 Outcomes

The main expected advances in visualisation and user experience are as follows:

- **Scalable data visualisation approaches and tools:** To handle extremely large volumes of data, the interaction must focus on aggregated data at different scales of abstraction rather than on individual objects. Techniques for summarising data in different contexts are highly relevant. There is a need to develop novel interaction techniques that can enable easy transitions from one scale or form of aggregation to another (e.g. from neighbourhood level to city level) while supporting aggregation and comparisons between different scales. It is necessary to address the uncertainty of the data and its propagation through aggregation and analysis operations.

- **Collaborative, 3D and cross-platform data visualisation frameworks:** Novel ways to visualise large amounts of possibly real-time data on different kinds of devices are required, including the augmented reality visualisation of data on mobile devices (e.g. smart glasses), as well as real-time and collaborative 3D visualisation techniques and tools.
- **New paradigms for visual data exploration, discovery and querying:** End-users need simplified mechanisms for the visual exploration of data, intuitive support for visual query formulation at different levels of abstraction, and tool-supported mechanisms for the visual discovery of data.
- **Personalised end-user-centric reusable data visualisation components:** Also useful are plug-and-play visualisation components that support the combination of any visualisation asset in real time and can be adapted and personalised to the needs of end-users. These also include advanced search capabilities rather than pre-defined visualisations and analytics. User feedback should be as simple as possible.
- **Domain-specific data visualisation approaches:** Techniques and approaches are required that support particular domains in exploring domain-specific data, for example innovative ways to visualise data in the geospatial domain, such as geo-locations, distances and space/time correlations (i.e. sensor data, event data). Another example is time-based data visualisation (it is necessary to take into account the specifics of time) – in contrast to common data dimensions which are usually ‘flat’. Finally, the visualisation of interrelated/linked data that exploits graph visualisation techniques to allow easy exploration of network structures.

3.5 Priority ‘Data Protection’

Data protection and anonymisation is a significant 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, to support 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.

There is a need for a truly modern and harmonised legal framework on data protection which has teeth and can be enforced appropriately to 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.

3.5.1 Challenges

In this perspective, the following main challenges have been identified:

- A more generic, easy-to-use and **enforceable data protection** approach suitable for large-scale commercial processing is needed. Data usage should conform to current legislation and policies. On the technical side, mechanisms are needed to provide data owners with the means to define the purpose of information gathering and sharing and to control the granularity at which their data will be shared with authorised third parties throughout the lifecycle of the data (data-in-motion and data-at-rest). Moreover, citizens should be able, for example, to have a say over the destruction of their personal data (the right to be forgotten). Data protection mechanisms also need to be ‘easy’, or at least capable of being used and understood with a reasonable level of effort by the various stakeholders, especially the end-users. Technical measures are also needed to enable and enforce the auditability of the principle that the data is only used for the defined purpose and nothing else – in particular, in relation to controlling the usage of personal information. In distributed settings such as supply chains, distributed trust technologies such as blockchains can be part of the solution.

- Maintaining robust **data privacy with utility guarantees** is a significant challenge and one which also implies sub-challenges, such as the need for state-of-the-art data analytics to cope with encrypted or anonymised data. The scalability of the solutions is also a critical feature. Anonymisation schemes may expose weaknesses exploitable by opportunistic or malicious opponents, and thus new and more robust techniques must be developed to tackle these adversarial models. Thus, ensuring the irreversibility of the anonymisation of big data assets is a key big data issue. On the other hand, encrypted data processing techniques, such as multiparty computation or homomorphic encryption, provide stronger privacy guarantees, but can currently only be applied to small parts of computation due to their large performance penalty. Also important are data privacy methods that can handle different data types as well as co-existing data types (e.g. datasets containing relational data together with sequential data about users), and methods that are designed to support analytic applications in different sectors (e.g. telecommunications, energy, and healthcare). Finally, preserving anonymity often implies removing the links between data assets. However, the approach to preserving anonymity also has to be reconciled with the needs for data quality, on which link removal has a very negative impact. This choice can be located on the side of the end-user, who has to balance the service benefits and possible loss of privacy, or on the side of the service provider, who has to offer a variety of added-value services according to the privacy acceptance of their customers. Measures to quantify privacy loss and data utility can be used to allow end-users to make informed decisions.
- Risk-based approaches calibrating information controllers’ obligations regarding **privacy and personal data protection** must be considered, especially when dealing with the combined processing of multiple datasets. It has indeed been shown that when processing combinations of anonymised, pseudonymised and even public datasets, there is a risk that personally identifiable information can be retrieved. Thus, providing tools to assess or prevent the risks associated with such data processing is an issue of significant importance.

3.5.2 Outcomes

The main expected advances in data protection are as follows:

- **Complete data protection framework:** A good mechanism for data protection includes protecting the Cloud infrastructure, analytics applications and the data from leakage and threats, but also provides easy-to-use privacy mechanisms. Apart from the specification of the intended use of data, usage control mechanisms should also be covered.

- **Mining algorithms:** Developed privacy-preserving data mining algorithms.
- **Robust anonymisation algorithms:** Scalable algorithms that guarantee anonymity even when other external or publicly available data is integrated. In addition, algorithms that allow the generation of reliable insights by cross-referring data from a particular user in multiple databases, while protecting the identity of the user. Moreover, anonymisation methods that can guarantee a level of data utility to support intended types of analyses. Lastly, algorithms that can anonymise datasets of co-existing data types or generate synthetic data, which are commonly encountered in many business sectors, such as energy, healthcare and telecommunications.
- **Protection against reversibility:** Methods to analyse datasets to discover privacy vulnerabilities, evaluate the privacy risk of sharing the data and decide on the level of data protection that is necessary to guarantee privacy. Risk assessment tools to evaluate the reversibility of the anonymisation mechanisms.
- **Multiparty mining/pattern hiding:** Secure multiparty mining mechanisms over distributed datasets, so that data on which mining is to be performed can be partitioned, horizontally or vertically, and distributed among several parties. The partitioned data cannot be shared and must remain private, but the results of mining on the ‘union’ of the data are shared among the participants. The design of mechanisms for pattern hiding so that data is transformed in such a way that certain patterns cannot be derived (via mining) while others can.

4 Big Data Standardisation

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. Further development of technology and data standards for big data is needed by:

- Leveraging existing common standards as the basis for an open and thriving 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 BDVA Big Data 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 major 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. A pragmatic approach to standardisation is needed by influencing, in addition to NoSQL databases, the standardisation of technologies such as complex event processing for real-time 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. Collaborative efforts are needed to support, where possible and pragmatic, the definition of semantic standardised data representation, ranging from 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.

5 Engineering and DevOps for Big Data

Big data technologies have gained significant momentum in research and innovation. However, mature, proven and empirically sound engineering methodologies for building next-generation big data value systems are not yet available. Also, 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.

5.1 Challenges

Engineering and DevOps toolchains for big data value systems need to look at and systematically integrate a diverse set of aspects for: (1) system/software engineering, (2) development and operations and (3) quality assurance.

The main challenges to be addressed include:

- **Big data value engineering:** The engineering of big data value systems needs to be supported by targeted methodologies and tooling. Particularly important is significantly extending from online analytical processing (OLAP) systems to fully fledged frameworks which integrate data management, data analytics and data protection by bringing these data technologies into a unified systems perspective.
- **DevOps:** Integrated development and operations (DevOps) approaches need to be tailored to data systems. In particular, these approaches should align the work of data scientists (who develop data analytics solutions) and data engineers (who manage and curate data for and during operations).
- **Quality assurance:** Novel methods of quality assurance are required to deliver trustworthy and reliable big data value systems. Proven quality assurance techniques from software engineering, for example, can only be a starting point, as these techniques have to be significantly extended to cope with the values of big data. This may include generating (e.g. using simulation) sufficient and representative test data (e.g. incorporating extreme cases) to cover the volume and variety of big data. As testing may not scale to the ever-increasing size, velocity and variety of data, complementary (formal) verification techniques may be required to deliver confidence in the systems' quality. Also, to cope with velocity, existing monitoring techniques need to be extended to ensure the quality of big data value systems during their operation.
- **Considering multiple dimensions of big data value:** The design and advancement of methodologies, tooling and platforms should carefully consider the multifaceted issues of big data, such as real-time processing and analytics, as well as data veracity and variety.

5.2 *Outcomes*

The expected primary outcomes for engineering and DevOps are:

- Engineering principles, as well as fully integrated toolchains and frameworks, that significantly increase productivity in terms of developing and deploying big data value systems
- Testing, monitoring and verification tools and methodologies to significantly increase reliability, security, energy efficiency and quality of big data value systems
- Enhancing real-time capabilities of data systems and platforms to handle high-intensity and highly distributed data and event streams

6 Illustrative Scenario in Healthcare

This section illustrates how the technical priorities may help in delivering big data solutions for specific industry sectors. To this end, we present a scenario from the healthcare sector. A BDVA white paper collected and analysed the needs, opportunities and challenges for big data technologies in healthcare (TF7 Healthcare subgroup 2016).

There is a clear opportunity to transform healthcare by applying data technologies. To improve the productivity of the healthcare sector, it is necessary to reduce costs while maintaining or improving the quality of the care provided. The fastest, least costly and most effective way to achieve this is to use the knowledge that is hiding within the already existing large amounts of generated medical data. According to current estimates, medical data is already at the zettabyte scale and will soon reach the yottabyte (e.g. 1000 zettabytes, a billion petabytes) scale. While most of this data was previously stored in hard copy format, the current trend is towards digitisation of these large amounts of information, thus making them amenable to analysis, resulting in what is known as big data.

The challenges and needs for research and innovation in this illustrative scenario are quite evident for each of the technical priorities listed above. Let's consider them one by one, starting with data management.

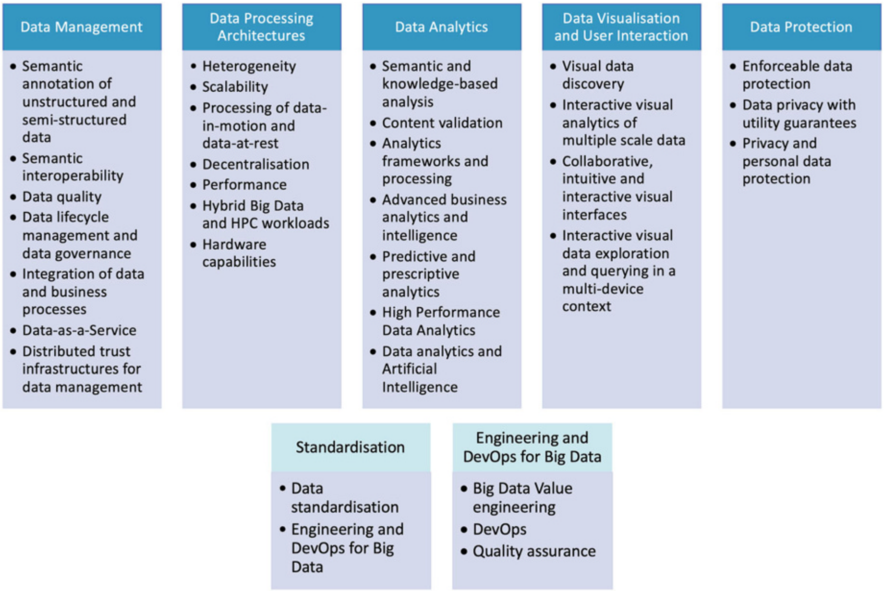
- **Data management:** Access to high-quality, large healthcare datasets to optimise care processes, disease diagnosis, personalised care and the healthcare system in general. Furthermore, a real transformation of the healthcare sector can only be achieved if all stakeholders and verticals in the healthcare sector (the HealthTech industry, healthcare providers, pharma, and insurance) share data and allow free data flow. Topics such as data quality, semantic interoperability and data management lifecycles are of the utmost importance in breaking down data silos in healthcare.

- **Data processing:** Consequently, the data processing architecture needs to be able to deal with heterogeneous health data (medical records, medical images and lab results), ensuring scalability (e.g. to process millions of patient records to find a similar patient) and performance (e.g. for smart alarms in intensive care units).
- **Data analytics:** The main challenges arise in the field of data analytics. The core of healthcare transformation is expected to come from AI-based propositions to enable personalised medicine, clinical decision support, workflow optimisation, clinical research and, finally, better diagnosis and medical treatment for patients.
- **Data visualisation and user interaction:** An area closely related to analytics and data interpretation is data visualisation and user interaction. Visualising models obtained by machine learning, as well as effective and clear user interaction technologies, is of utmost importance for the acceptance of AI technologies in the healthcare sector.
- **Data protection:** The developing focus on data protection is especially important in the healthcare sector, which deals with sensitive health data. Robust data privacy and anonymisation techniques, privacy-preserving data mining, end-to-end security and consent management are significant challenges to be addressed.
- **Standards:** Finally, in the healthcare sector data is often fragmented or generated by different systems with incompatible formats. Therefore, interoperability and standardisation are key to deploying the full potential of data held.
- **Engineering and DevOps:** Linked to this are the engineering methodologies for building next-generation big data value systems in healthcare, which need to be correctly validated by clinical trials and regulatory approval. An interesting challenge is to create methodologies to regulate AI-based propositions more quickly and also address the liability and regulatory aspects of techniques such as continuous learning.

7 Summary

Enabling an effective data ecosystem requires overcoming several technical challenges associated with the cost and complexity of extracting value from data. This chapter identifies and characterises the key research areas. A systemic and structured methodology was used to gather inputs from over 200 stakeholder organisations. The results of this process, as illustrated in Fig. 8, identify the five technical research priorities together with 28 sub-challenges of big data. The requirement analysis was done in consultation with a community of stakeholders that included organisations for industry, research and government.

The results presented in this chapter provide a prioritised list of cross-sectorial business needs of data technologies and their impact in industry, research and government. These findings serve as a guide for directing the research and development efforts towards fostering a data ecosystem. The findings indicate that deep analytics and data management are viewed as the top two technical challenges for big data, with more than 60% of organisations prioritising them as having a high



Data Management

- Semantic annotation of unstructured and semi-structured data
- Semantic interoperability
- Data quality
- Data lifecycle management and data governance
- Integration of data and business processes
- Data-as-a-Service
- Distributed trust infrastructures for data management

Data Processing Architectures

- Heterogeneity
- Scalability
- Processing of data-in-motion and data-at-rest
- Decentralisation
- Performance
- Hybrid Big Data and HPC workloads
- Hardware capabilities

Data Analytics

- Semantic and knowledge-based analysis
- Content validation
- Analytics frameworks and processing
- Advanced business analytics and intelligence
- Predictive and prescriptive analytics
- High Performance Data Analytics
- Data analytics and Artificial Intelligence

Data Visualisation and User Interaction

- Visual data discovery
- Interactive visual analytics of multiple scale data
- Collaborative, intuitive and interactive visual interfaces
- Interactive visual data exploration and querying in a multi-device context

Data Protection

- Enforceable data protection
- Data privacy with utility guarantees
- Privacy and personal data protection

Standardisation

- Data standardisation
- Engineering and DevOps for Big Data

Engineering and DevOps for Big Data

- Big Data Value engineering
- DevOps
- Quality assurance

Fig. 8 High-level technical priorities and sub-challenges for big data value

impact on the data ecosystem. Although data privacy was considered a significant challenge, it was ranked lowest compared to other key challenges. This may be because not all data applications and domains have privacy implications and may focus on industrial/machine data.

Finally, these data research priorities have laid the foundations for a joint Strategic Research, Innovation and Deployment Agenda for an AI, Data and Robotics Partnership in Europe (Zillner et al. 2020) with the goal to unify the strategic focus of each of the three disciplines engaged in creating the Partnership.

Acknowledgements We greatly acknowledge the collective effort of the SRIA teams: Carlos A. Iglesias, Antonio Alfaro, Jesus Angel, Sören Auer, Paolo Bellavista, Arne Berre, Freek Bomhof, Stuart Campbell, Geraud Canet, Giuseppa Caruso, Edward Curry, Paul Czech, Davide Dalle Carbonare, Nuria de Lama, Stefano de Panfilis, Thomas Delavallade, Marija Despenic, Ana Garcia Robles, Wolfgang Gerteis, Aris Gkoulalas-Divanis, Nuria Gomez, Paolo Gonzales, Thomas Hahn, Souleiman Hasan, Jim Keneally, Bjarne Kjær Ersbøll, Bas Kotterink, Yannick Legré, Yves Mabiala, Julie Marguerite, Dirk Mayer, Ernestina Menasalves, Andreas Metzger, Elisa Molino, Thierry Nagellen, Dalit Naor, Maria Perez, Milan Petkovic, Roberta Piscitelli, Klaus-Dieter Platte, Pierre Pleven, Dumitru Roman, Titi Roman, Alexandra Rosén, Nikos Sarris, Stefano Scamuzzo, Simon Scerri, Corinna Schulze, Robert Seidl, Bjørn Skjellaug, Caj Södergård, Claire Tonna, Francois Troussier, Colin Upstill, Josef Urban, Meilof Veeningen, Tonny Velin, Ray Walshe, Walter Waterfeld, Stefan Wrobel, and Sonja Zillner.

References

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* (pp. 1–303). New York: Springer. https://doi.org/10.1007/978-3-319-21569-3

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165. https://doi.org/10.2307/41703503

Communication: A European strategy for data. (2020). Retrieved from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

Curry, E. (2016). The big data value chain: Definitions, concepts, and theoretical approaches. In J. M. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*. New York: Springer. https://doi.org/10.1007/978-3-319-21569-3_3

Curry, E., Becker, T., Munné, R., De Lama, N., & Zillner, S. (2016). The BIG project. In J. M. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*. New York: Springer. https://doi.org/10.1007/978-3-319-21569-3_2

DG Connect. (2013). *A European strategy on the data value chain*. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=3488

Flick, U. (2004). Triangulation in qualitative research. In *A companion to qualitative research* (p. 432).

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21–32.

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

Oppenheim, A. N. (1992). Questionnaire design, interviewing and attitude measurement. *Journal of Marketing Research, 30*. Retrieved from http://www.amazon.com/Questionnaire-Design-Interviewing-Attitude-Measurement/dp/1855670445

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

Rusitschka, S., & Curry, E. (2016). Big data in the energy and transport sectors. In J. M. Cavanillas, E. Curry, & W. Wahlster (Eds.), *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe*. New York: Springer. https://doi.org/10.1007/978-3-319-21569-3_13

TF7 Healthcare subgroup. (2016). *Big Data Technologies in Healthcare*.

Yin, R. K. (2013). Case study research: Design and methods. *SAGE Publications, 26*(1), 93–96. https://doi.org/10.1017/CBO9781107415324.004

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

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.