BDV BIG DATA VALUE ASSOCIATION

BDV SRIA

European Big Data Value Strategic Research and Innovation Agenda

Version 4.0 October 2017

Accelerating Data-Driven Innovation in Europe

www.bdva.eu

Executive Summary

This Strategic Research and Innovation Agenda (SRIA) defines the overall goals, main technical and non-technical priorities, and a research and innovation roadmap for the European Public Private Partnership (PPP) on Big Data Value. The SRIA has been developed by the Big Data Value Association (BDVA), an industry-led organisation representing large businesses, small and medium-sized enterprises (SMEs), and research organisations in Europe.

The SRIA explains the strategic importance of Big Data, describes the data value chain and the central role of ecosystems, details a vision for Big Data Value in Europe in 2020, and sets out the objectives and goals to be accomplished by the PPP within the European research and innovation landscape of Horizon 2020 (H2020) and at both national and regional levels.

The multiple dimensions of Big Data Value are described and the overarching strategic objectives for the PPP are set out. These embrace data, skills, legal and policy issues, technology leadership through research and innovation, transforming applications into new business opportunities, the acceleration of business ecosystems and business models, with a particular focus on SMEs, and successful solutions for the major societal challenges Europe is facing in areas such as health, energy, transport and the environment. The objectives of the SRIA are broken down into specific areas, such as competitiveness, innovation and technology, and societal and operational objectives.

The implementation strategy for addressing the goals of the SRIA involves four mechanisms: i-Spaces; Lighthouse projects; technical projects; and cooperation and coordination projects. I-Spaces are cross-organisation, cross-sector, interdisciplinary Innovation Spaces intended to anchor targeted research and innovation projects. They offer secure accelerator-style environments for running experiments in both private data and open data, bringing technology and application development together. I-Spaces will act as incubators for new businesses in the development of skills, competences and best practices. Lighthouse projects are large-scale, data-driven innovation and demonstration projects that will create superior visibility, awareness and impact. The four mechanisms together will foster the development of the European data ecosystem in three distinct phases by establishing an innovation ecosystem, pioneering disruptive Big Data value solutions, and setting long-term ecosystem enablers. Moreover, the PPP will drive the development of the European Data Value Ecosystem to promote synergies and cooperation among members and with other PPPs such as ETP4HPC, 5G, ECSO, AIOTI, and others.

The strategic and specific goals, which together will ensure Europe's leading role in the data-driven world, are supported by key specific technical and non-technical priorities. Five technical priority areas have been identified for research and innovation: data analytics to improve the understanding of data; optimised architectures for analytics of data-at-rest and data-in-motion; mechanisms ensuring data protection and anonymisation, to enable the vast amounts of data which are not (and never can be) open data to be incorporated into the data value chain; advanced visualisation and user

experience; and, underpinning these, data management engineering accompanied by aspects of big data standardisation. The complementary non-technical priorities are: skills development, business models and ecosystems; regulation and policy; and social perceptions and societal implications.

Finally, the expected impact of the objectives is summarised, together with KPIs to frame and assess that impact. The activities set out in this SRIA will deliver solutions, architectures, technologies and standards for the data value chain over the next decade, leading to a comprehensive ecosystem for achieving and sustaining Europe's role, delivering economic and societal benefits, and enabling a future in which Europe is the world leader in the creation of Big Data Value.

Significant updates of content between SRIA Version 3 and this SRIA, Version 4, are indicated by

BDV BIG DATA VALUE ASSOCIATION

Contents

1. INTRODUCTION

The recent developments of the European policy and data market have been reflected in this section.

1.1. Strategic Importance of Big Data Value

The continuous and significant growth of data together with improved access to data and the availability of powerful Information and Communication Technologies (ICT) systems have led to intensified activities around Big Data Value. **Powerful data techniques and tools** allow collecting, storing, analysing, processing and visualising vast amounts of data. **Open data initiatives** are gaining momentum, providing broad access to data from the public sector, business and science.

The European data market measured by the value of the data products and services bought by European businesses and consumers is a rapidly growing multibillion Euro business. According to the International Data Corporation (IDC)[1] , the compound annual growth rate (CAGR) of the EU data market over the period 2016–2020 may be as high as 15.7% under the most favourable scenario. This would mean that the size of the data market in Europe is expected to more than double in the coming years, boosted by sustained economic recovery and the swift adoption of data-driven technologies, thus reaching a value of around 107 billion EUR by 2020.

The exploitation of Big Data in various sectors has a potential socio-economic impact far beyond the specific Big Data market. Therefore, it is essential to embrace new technology, applications, use cases and business models within and across various sectors and domains. This will ensure the rapid adoption of Big Data by organisations and individuals, and provide major returns in terms of growth and competitiveness. In particular, the efficiency gains made possible by Big Data will also have a profound **societal impact**. As an example, the OECD[2] reports that 380 megatonnes of CO2 emissions may be saved worldwide in transport and logistics, while the utility sector may see a CO2 reduction of over 2 gigatonnes.

The volume of data is rapidly growing. **By 2020, there will be more than 16 zettabytes** of useful data (16 trillion GB)[3], which implies growth of 236% per year from 2013 to 2020. This data explosion is a reality that Europe must both face and exploit in a

[1] IDC et al., European Data Market, SMART 2013/0063, D9 – Final Report, 1 February 2017, http://datalandscape.eu/study-reports
[2] OECD, Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by 'Big Data', OECD, Paris, 2013.
[3] Vernon Turner, John F. Gantz, David Reinsel and Stephen Minton, The digital universe of opportunities: rich data and the increasing value of the Internet of Things, Report from IDC for EMC April 2014.

structured, aggressive and ambitious way to create value for its citizens, businesses in all sectors and society as a whole.

From the European policies perspective, the mid-term review on the implementation of the Digital Single Market Strategy[4] (released on 10 May 2017) provides a good overview of the strategic importance and positioning of Big Data Value. Among the current legislative priorities and commitments to implement a connected Digital Single Market (DSM), the European Commission is working closely with Member States, the independent Data Protection Supervisory Authorities, and with businesses and civil society to prepare for the application of the General Data Protection Regulation (GDPR)[5] from 25 May 2018, the implementation of which is essential to 'safeguard individuals' fundamental right to the protection of personal data in the digital age'.

'**Developing the European Data Economy**' is one of the new pillars of the extended DSM strategy designed to keep up with emerging trends and challenges. It focuses on defining and implementing the framework conditions for a European Data Economy ensuring a fair, open and secure digital environment. To foster common approaches, the Commission has undertaken a public consultation as well as detailed exchanges with Member States over a European framework for the free flow of data within the Digital Single Market. The main focus is on ensuring the effective and reliable cross-border flow of non-personal data, and access to and reuse of such data, as well as looking at the challenges to safety and liabilities posed by the Internet of Things (IoT).

To manage the digital transformation of our society and economy, the Digital Single Market Strategy focuses on four different areas:

1. **Digital skills**, addressing the impact of information and communication technologies on the transformation of jobs and skills.
2. Start-ups and the digitisation of all industry and service sectors, with a special focus on the role of ICT standards.
3. **Digital innovation for modernising public services**, to allow public authorities to deliver services more quickly, precisely and efficiently.
4. **Stepping up investments in digital technologies and infrastructures**, in particular: (i) in developing a European Open Science Cloud, High Performance Computing and a European Data Infrastructure; and (ii) building Artificial Intelligence capacities.

Large companies and SMEs in Europe are clearly seeing the fundamental potential of Big Data Value for causing disruptive change in markets and business models, and are beginning to explore the opportunities which are now appearing. IDC confirms that Big Data adoption in Europe is accelerating[7]. According to IDC findings[8], in 2016 the European data market was second in value only to the US (with around half the market size), and was growing almost as fast. Companies intending to build and rely on data-driven solutions appear to have begun fruitfully addressing challenges that extend

[4]https://ec.europa.eu/digital-single-market/en/news/digital-single-market-commission-calls-swift-adoption-key-proposals-and-maps-out-challenges

[5]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, which entered into force on 24 May 2016 and shall apply from 25 May 2018. In this context, the Radio Equipment Directive 2014/53/EU, Article 3(3)(e), empowers the Commission to adopt delegated acts inter alia on safeguards in radio equipment to protect the personal data and privacy of users.

[6]https://ec.europa.eu/digital-single-market/en/news/public-consultation-building-european-data-economy

[7]Gabriella Cattaneo, 'The European Data Market', IDC, presentation at the European Data Forum in Luxembourg, November 2015, http://2015.data-forum.eu/sites/default/files/1140-1155_Gabriela%20Cattaneo_SEC.pdf

[8]IDC et al., European Data Market, SMART 2013/0063, D9 – Final Report, 1 February 2017, http://datalandscape.eu/study-reports

well beyond technology usage. The successful adoption of Big Data requires changes in business orientation and strategy, processes, procedures and organisational set-up. European enterprises are creating new knowledge and are starting to hire new experts, enhancing a new ecosystem.

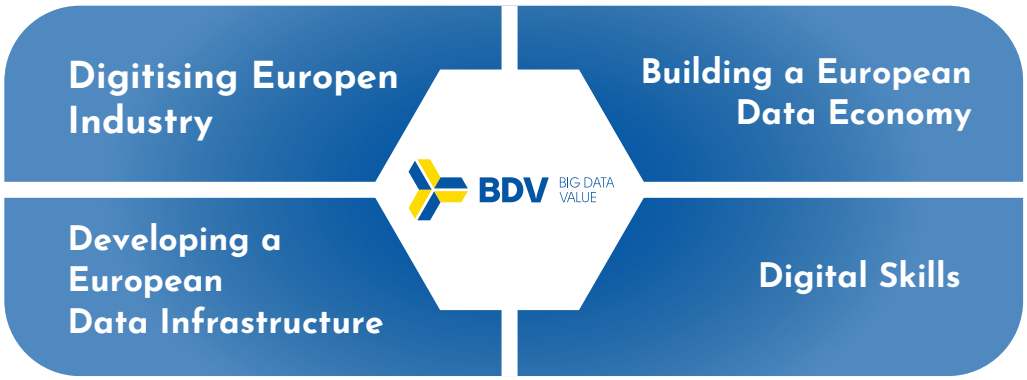
Economic and social activities have long relied on data. But the increased volume, velocity, variety, and social and economic value of data signal **a paradigm shift towards a data-driven socio-economic model**. The significance of data will only grow in importance beyond 2020 as it is used to make critical decisions in our everyday lives, from the course of treatment for a critical illness to safely driving a car. The challenges beyond 2020 will be multifaceted: How can we trust the large-scale, data-driven decision-making provided by data-powered **Artificial Intelligence (AI) platforms**? How will decision-making processes evolve between humans and AI-based systems? And what are the legal and ethical issues associated with making data-driven critical decisions?

1.2 The Big Data Value PPP (BDV PPP)

Europe must aim high and mobilise stakeholders throughout society, industry, academia and research to enable the creation of a European Big Data Value economy, supporting and boosting agile business actors, and delivering products, services and technology, while providing highly skilled data engineers, scientists and practitioners along the entire Big Data Value chain. This will result in an innovation ecosystem in which value creation from Big Data flourishes.

To achieve these goals, the **European contractual Public Private Partnership** on **Big Data Value (BDV PPP)** was signed on 13 October 2014. This marked the commitment of the European Commission, industry and partners from academia to build a data-driven economy across Europe, mastering the generation of value from Big Data and creating a significant competitive advantage for European industry, thus boosting economic growth and jobs. **The Big Data Value Association (BDVA)** is the private counterpart to the EU Commission in implementing the BDV PPP programme. The BDVA has a well-balanced composition of large, small and medium-sized industries and enterprises, as well as research organisations to support the development and deployment of the PPP work programme and to achieve the Key Performance Indicators (KPI) set out in the PPP contract. The BDV PPP commenced in 2015 and was operationalised with the launch of the Leadership in Enabling and Industrial Technologies (LEIT) work programme 2016/2017 of Horizon 2020. The BDV PPP activities will address the development of technology and applications, business model discovery, ecosystem validation, skills profiling, regulatory and IPR environments and a number of social aspects. The BDV PPP will lead to a comprehensive innovation ecosystem for achieving and sustaining European leadership in Big Data, and delivering the maximum economic and societal benefits to Europe – its businesses and citizens. Finally, the value generated by applying intelligence on Big Data will empower Artificial Intelligence to foster linking, cross-cutting and vertical dimensions of value creation at the technical, business and societal levels across many different sectors.
The Big Data Value PPP plays a central role in the implementation of the revised DSM strategy, contributing to its different pillars, as visualised in Figure 1.

Figure 1 BDV PPP contribution to the European DSM strategy



1.3 The Role of Big Data Value in Digitising European Industry (DEI)

In April 2016 the European Commission adopted a comprehensive strategy on Digitising European Industry (DEI)[9]. This strategy highlights the importance of stimulating private investment in digital innovations in all industrial sectors across the EU. The DEI strategy calls for: (1) the development of digital innovation hubs all across Europe, so that 'access to [the] latest technologies will be possible for any industry in Europe with the aim of spurring a wave of bottom-up innovations across sectors'; and (2) the reinforcement of 'public private partnerships on innovation and strategic R&D to ensure EU-wide industry academia collaboration involving stakeholders across value chains'. The Big Data Value PPP is aimed at supporting the implementation of this strategy of developing the Data Platforms to support the growth of innovative data-driven businesses in Europe and the exploitation of the potential value of data across sectors. Big Data technologies and Big Data Value ecosystems play a crucial role as the main enablers for the Digitising European Industry strategy. The BDV PPP provides unique means to pool the resources needed to achieve ground-breaking developments in Big Data technologies and platforms, including experimentation spaces for science and innovation as well as large-scale test-beds to accelerate standards setting.

Reports from DEI Working Group 1[10] (Digital Innovation Hubs) and DEI Working Group 2[11] (Digital Industrial Platforms) refer to the BDV PPP, and in particular to the value and contribution of the i-Spaces and Lighthouse implementation instruments.

The future competitive position of Europe will rest on the capability of regions, public administrations and organisations (in particular SMEs) to extract data insights from next-generation infrastructure. In 2018 we are at the beginning of a

[9]COM(2016) 180.
[10]https://ec.europa.eu/futurium/en/system/files/ged/dei_working_group1_report_june2017_0.pdf

great synergy in enabling digital technologies, from IoT to 5G, Cloud and High Performance Computing (HPC), Edge Computing and Big Data. We believe that the common dominator of next-generation digital infrastructure is the knowledge about technologies that is necessary to extract insights from the Big Data collected and generated by them. These Big Data insights will form the foundations on which the transformation of industry and society will be built.

The Digitising European Industry (DEI) initiative[12] recognises that all sectors of the economy need to be digitised for the EU to reinforce its competitiveness, build a strong industrial base, and manage the transition to a smart economy. In particular, this requires strengthening leadership in digital technologies and in digital industrial platforms across value chains in all sectors of the economy. As a result, the concept of Digital Innovation Hubs (DIHs) and their regional networks of competence centres was born. DIHs aim to allow any business to gain access to knowledge, expertise and technology for the purpose of conducting tests and experiments in digital innovations relevant to its products, processes or business models. They do this by developing an ecosystem comprising stakeholders from technology, business, and finance and funding, as well as policy makers and industry partners.[13]

When referring to digital industrial platforms across value chains, one key requirement is the Big Data Value chain, as depicted Figure 2. Europe needs strong players along this Big Data Value chain, in areas ranging from data generation and acquisition, through data processing and analysis, to curation, usage, service creation and provisioning. Each link in the value chain has to be strong so that a vibrant Big Data Value ecosystem can evolve.

Figure 2: The Big Data Value chain [14]

[11]https://ec.europa.eu/futurium/en/implementing-digitising-european-industry-actions/report-wg2-digital-industrial-platforms-final
[12]EC Communication (COM(2016) 180).
[13]http://s3platform.jrc.ec.europa.eu/digital-innovation-hubs

There are a number of major companies in Europe that provide services and solutions along the Big Data Value chain. Some of them generate, and provide access to, huge amounts of data, including structured and unstructured data. They acquire or combine real-time data streams from different sources, or add value by pre-processing, validating or augmenting data and ensuring data integrity. There are companies that specialise in analysing data and recognising correlations and patterns. Furthermore, some companies use these insights to make predictions and decisions in various application domains.

However, despite the growing number of companies active in the data business, strengthening an economic Big Data Value ecosystem by bringing organisations together along the Big Data Value chain at the European level is required. Data usage is growing, but it is treated and handled in a fragmented way in the fields of business and science. To ensure a coherent use of data, a wide range of stakeholders along the Data Value chain need to be brought together to create a basis for cooperation.

The stakeholders that will form the foundation for interoperable data-driven ecosystems as a resource for new businesses and innovations using Big Data are:

- **Vendors of the ICT industry** (both large concerns and SMEs), that provide access to innovative ICT in dedicated explorative settings and can benefit from the feedback of test users in experimental settings, thus gaining valuable guidance for the optimisation of their technology and influencing standards.

- **Users across different industrial sectors** (private and public), who:
  - will make use of Big Data solutions for advanced decision making or automation;
  - can provide valuable insights into user needs and the roles and/or interests of important user groups, as well as promising application scenarios;
  - can benefit from advanced Big Data technology by generating value in the context of their business;
  - can inform the ecosystem about industrial requirements and challenges leading to new research questions.

- **Data Entrepreneurs**, who build innovative data businesses and data services based on Big Data on the demand and supply side;

- **Researchers and academics**, who provide access to state-of-the-art research in Big Data technology;

- **Policy makers**, responsible for establishing policy framework conditions that foster the adoption of Big Data technology in various sectors.

[14]CSee Michael E. Porter, Competitive advantage: creating and sustaining superior performance, The Free Press, New York, 1998, and M. Cavanillas, E. Curry and W. Wahlster: New horizons for a data-driven economy: a roadmap for big data in Europe, Berlin, Springer International Publishing, 2016

1.4 The Multiple Dimensions of Big Data Value

In order to sustain the growth of Big Data and remain competitive with other countries and regions, Europe needs to foster, strengthen and support the development and wide adoption of Big Data Value technologies, successful use cases and data-driven business models. At the same time, it is necessary to deal with many different aspects of an increasingly complex landscape. The main issues that Europe must tackle in creating and sustaining a strong Big Data ecosystem relate to the following dimensions:

- **Data**: The availability of data and access to data sources are paramount concerns. There is a broad range of data types and data sources: structured and unstructured data; multilingual data sources; data generated from machines and sensors; data-at-rest and data-in-motion. Value is created by acquiring data, combining data from different sources, and providing access to data with low latency while ensuring data integrity and preserving privacy. Pre-processing, validating and augmenting data, as well as ensuring their integrity and accuracy, add value.

- **Skills**: In order to leverage the potential of Big Data Value, a key challenge for Europe is to ensure the availability of highly and relevantly skilled people who have an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and solutions. There will be a need for data scientists and engineers who have expertise in analytics, statistics, machine learning, data mining and data management. These specialists should be combined with other experts who have strong domain knowledge and the ability to apply this know-how within organisations to create value.

- **Legal**: The increased importance of data will intensify the debate on data ownership and usage, data protection and privacy, security, liability, cybercrime and Intellectual Property Rights (IPRs). These issues need to be resolved in order to remove the barriers to adoption. Favourable European regulatory environments are required to facilitate the development of a genuine pan-European Big Data market.

- **Technical**: Key aspects such as real-time analytics, low latency and scalability in processing data, new and rich user interfaces, interacting with and linking data, information and content, all have to be developed in order to open up new opportunities and to sustain or develop competitive advantages. As well as having agreed approaches, the interoperability of datasets and data-driven solutions is essential to ensure wide adoption within and across sectors.

- **Application:** Business and market-ready applications should be the target. Novel applications and solutions must be developed and validated in ecosystems, providing the basis for Europe to become the world leader in the creation of Big Data Value.

- **Business**: Making more efficient use of Big Data, and understanding data as an economic asset, carries great potential for the EU economy and society. The establishment of Big Data Value ecosystems and the development of appropriate business models on top of a strong Big Data Value chain must be supported in order to generate the desired impact on the economy and employment.

- **Societal: Big Data** will provide solutions for major societal challenges in Europe, such as improved efficiency in healthcare information processing or reduced CO2 emissions through climate impact analysis. In parallel, an accelerated adoption of Big Data will critically increase awareness of the benefits and value that Big Data can create for business, the public sector and the individual citizen.

Creating a favourable **ecosystem** for Big Data and promoting its accelerated adoption requires an interdisciplinary approach that addresses all of the aforementioned dimensions of Big Data Value.

1.5 BDV PPP Vision and Objectives for European Big Data Value

This section was extended by a mission and vision statement and merged with a section documenting the general objectives of the BDV PPP.

The Big Data Value Association and the launch of the **BDV PPP** pursue a common shared vision of positioning **Europe as the world leader in the creation of Big Data Value**.

Structured along the dimensions introduced in Section 1.4, the BDV PPP vision for Europe in 2020 concerns the following aspects:

- **Data**: Zettabytes of useful public and private data will be widely and openly available. By 2020, smart applications such as smart grids, smart logistics, smart factories and smart cities will be widely deployed across the continent and beyond. Ubiquitous broadband access, mobile technology, social media, services, and the IoT on billions of devices will have contributed to the explosion of generated data to a global total of 40 zettabytes. Much of this data will yield valuable information. Extracting this information and using it in intelligent ways will revolutionise decision making in business, science and society, enhancing companies' competitiveness and leading to the creation of new industries, jobs and services.

- **Skills**: Millions of jobs will become established for data engineers and scientists, and the Big Data discipline will be integrated into technical and business degrees. The European workforce is increasingly data-savvy, regarding data as an asset.

- **Legal**: Privacy and security can be guaranteed along the Big Data Value chain. Data sharing and data privacy can be fully managed by citizens in a trusted data ecosystem.

- **Technology**: Real-time integration and interoperability among different multilingual, sensorial and non-structured datasets will be accomplished, and content automatically managed and visualised in real time. By 2020, European research and innovation efforts will have led to advanced technologies that make it significantly easier to use Big Data across sectors, borders and languages.

- **Application**: Applications using the BDV technologies can be built which will allow anyone to create, use, exploit and benefit from Big Data. By 2020, thousands of specific applications and solutions will address data-in-motion and data-at-rest. There will be a highly secure and traceable environment supporting organisations and citizens, with the capacity to sustain various monetisation models.

- **Business**: One true EU single data market will be established, thus allowing EU companies to increase their competitiveness and become world leaders. By 2020 value creation from Big Data will have a disruptive influence on many sectors. From manufacturing to tourism, from healthcare to education, from energy to telecommunications services, and from entertainment to mobility, Big Data Value will be a key success factor in fuelling innovation, driving new business models, and supporting increased productivity and competitiveness.

- **Societal**: Societal challenges will be addressed through BDV systems, focusing on areas such as the high volume, mobility and variety of data.

These issues will impact the European Union's priority areas as follows:

- **Economy**: The competitiveness of European enterprises will be significantly higher compared to their worldwide competitors, due to improved products and services and greater efficiency based on the value of Big Data. One true EU single data market will be established, allowing EU companies to increase their competitiveness and become world leaders.

- **Growth**: A flourishing sector of expanding new small and large businesses will result in a significant number of new jobs focusing on creating value out of data.

- **Society**: Citizens will benefit from better and more economical services in a stable economy where data can be shared with confidence. Privacy and security will be guaranteed throughout the lifecycle of BDV exploitation.

These three factors will support the major **EU pillars as stated in Rome Declaration** of March 2017[15]: a safe and secure Europe; a prosperous and sustainable Europe; a social Europe; and a stronger Europe on the world stage.

The **mission** of the Big Data Value Association is to develop the Innovation Ecosystem that will enable the data-driven digital transformation in Europe, delivering maximum economic and societal benefit, and achieving and sustaining Europe's leadership in the fields of Big Data Value creation and Artificial Intelligence.

To achieve this mission, in 2017 the BDVA defined four strategic priorities (Figure 3):

- **Develop Data Innovation Recommendations:** Providing guidelines and recommendations on data innovation to the industry, researchers, markets and policy makers.

- **Develop Ecosystem:** Developing and strengthening the European Big Data Value Ecosystem.

[15]http://www.consilium.europa.eu/en/press/press-releases/2017/03/25-rome-declaration

- **Guiding Standards:** Driving Big Data standardisation and interoperability priorities, and influencing standardisation bodies and industrial alliances.
- **Know-How and Skills:** Improving the adoption of Big Data through the exchange of knowledge, skills and best practices.

Figure 3: BDVA strategic priorities

1.6 BDV PPP Objectives

As laid out in the Contractual Arrangement (CA) of the BDV PPP[16], the overarching **general objectives** are:

- To foster European Big Data technology leadership in terms of job creation and prosperity by creating a Europe-wide technology and application base, and building up the competence and number of European data companies, including start-ups;
- To reinforce Europe's industrial leadership and ability to compete successfully in the global data value solution market by advancing applications which can be converted into new opportunities, so that European businesses secure a 30% market share by 2020;
- To enable research and innovation work, including activities related to interoperability and standardisation, and secure the future basis of Big Data Value creation in Europe;
- To facilitate the acceleration of business ecosystems and appropriate business models, with a particular focus on SMEs, enforced by a Europe-wide benchmarking of usage, efficiency and benefits;
- To provide and support successful solutions for major societal challenges in Europe, for example in the fields of health, energy, transport and the environment, and agriculture, etc.;
- To demonstrate the value of Big Data for businesses and the public sector and to increase citizens' acceptance levels by involving them as 'prosumers' and accelerating take-up;
- To support the application of EU data protection legislation and provide effective secure mechanisms to ensure its enforcement in the Cloud and for Big Data, thus facilitating its adoption.

The more specific objectives of the BDV PPP are documented in the Contractual Agreement as well as reflected in the discussion of the KPIs in Section 5.

[16]http://ok-bdva.iais.fraunhofer.de/sites/default/files/BDVPPP_Contractual_Arrangement_.pdf

1.7 BDV SRIA Document History

To establish a contractual counterpart to the European Commission for the implementation of the PPP, the Big Data Value Association, a fully self-financed not-for-profit organisation under Belgian law, was founded by 24 organisations including large businesses, SMEs and research organisations. As of October 2017 the BDVA has over 180 members, representing Big Data Value stakeholders from across the European Union.

This **Strategic Research and Innovation Agenda (SRIA)** defines the main technical and non-technical priorities to achieve the BDV PPP objectives (see Section 1.6), and describes a research and innovation roadmap for the BDV PPP. The BDV SRIA was prepared using an extensive process that has heavily engaged with the wider Big Data Value community. A wide range of stakeholders has contributed to the SRIA in different forms of engagement (see Annex 6.3). The BDV SRIA is constructed from inputs and analyses from SMEs and large businesses, public organisations, and research and academic institutions. Stakeholders include suppliers and service providers, data owners and early adopters of Big Data in many sectors. The process included multiple workshops and consultations to ensure the widest representation of views and positions, including the full range of public and private sector entities. The aim was to identify the main priorities of the stakeholders, with approximately 200 organisations and other relevant parties physically participating and contributing. Extensive analysis reports were then produced which helped to both formulate and construct this SRIA.

The BDVA is responsible for providing regular (yearly) updates of the SRIA to ensure it remains relevant to the priorities of the community. In this update, the main focus is to highlight the ongoing development of the European Data Value Ecosystem as well as to publish important BDV initiatives, such as the BDV reference model or the Big Data standardisation initiatives. In addition, SRIA version 4 incorporates important strategic directions within the BDV PPP that are reflected in the Work Programme 2018-2019 and provide guidance for the way forward to 2020 and beyond.

2. IMPLEMENTATION STRATEGY

Given the broad range of objectives around the many aspects of Big Data Value, a complete implementation strategy is needed. In this section, we set out such a strategy, the formulation of which is the result of a very broad discussion process involving a large number of relevant European BDV stakeholders.

The result is an interdisciplinary approach that integrates expertise from the different fields necessary to tackle both the strategic and specific objectives. To this end, European cross-organisational and cross-sector environments have to be developed in such a way that large enterprises and SMEs alike will find it easy to discover economic opportunities based on data integration and analysis, and then develop working prototypes to test the viability of actual business deployments. Such environments will add value to other existing hubs (e.g. DIHs) by providing services or forming a technological nucleus for individual hubs.

The growing number and complexity of valuable data assets will drive existing and new research challenges. Cross-sectorial and cross-organisational environments will enable research and innovations in new and existing technologies. Business applications that need to be evaluated for usability and fitness for purpose can be deployed within these environments, so ensuring their practical applicability. This, in turn, will require validations, trials and large-scale experiments in existing or emerging business fields, the public sector, industry, and, jointly, with end-users and individual consumers.

To support such validations, trials and large-scale experiments, access to valuable data assets needs to be provided with minimal obstacles in environments that simultaneously support legitimate ownership, privacy and security for data owners and their customers. These environments will facilitate experimentation for researchers, entrepreneurs, SMEs and large ICT providers.

2.1 Four Kinds of Mechanism

In order implement the research and innovation strategy, and to align technical issues with aspects of cooperation and coordination aspects, four major types of mechanism are recommended:

- **Innovation Spaces (i-Spaces):** These are cross-organisational and cross-sectorial environments that allow challenges to be addressed in an interdisciplinary way and will serve as a hub for other research and innovation activities.
- **Lighthouse projects:** These will help raise awareness of the opportunities offered

by Big Data and the value of data-driven applications for different sectors, acting as incubators for data-driven ecosystems.

- **Technical projects:** These will take up specific Big Data issues, addressing targeted aspects of the technical priorities as defined in Section 3.

- **Cooperation and coordination projects:** These projects will foster international cooperation for efficient information exchange and coordination of activities.

Further information about the ongoing PPP projects in alignment with these four implementation mechanisms is available from: http://www.big-data-value.eu/projects.

2.1.1 European Innovation Spaces (i-Spaces)

Extensive consultation with many stakeholders from areas related to Big Data Value (BDV) has confirmed that besides technology and applications, a number of key issues require consideration. First, infrastructural, economic, social and legal issues have to be addressed. Second, the private and the public sector need to be made aware of the benefits that BDV can provide, thereby motivating them to be innovative and to adopt BDV solutions.

To address all these aspects, European cross-organisational and cross-sectorial environments, which rely and build upon existing national and European initiatives, play a central role in a European Big Data ecosystem. These so-called **European Innovation Spaces** (or **i-Spaces** for short) are the main elements to ensure that research on BDV technologies and novel BDV applications can be quickly tested, piloted and thus exploited in a context with the maximum involvement of all the stakeholders of BDV ecosystems. As such, i-Spaces enable stakeholders to develop new businesses facilitated by advanced BDV technologies, applications and business models. They contribute to the building of a community, providing a catalyst for community engagement and acting as incubators and accelerators of data-driven innovation.

In this sense, i-Spaces are hubs for uniting technical and non-technical activities, for instance by bringing technology and application development together and fostering skills, competence, and best practices. To this end, i-Spaces offer both state-of-the-art and emerging technologies and tools from industry, as well as open source software initiatives; they also provide access to data assets. In this way, i-Spaces foster community building and an interdisciplinary approach to solving BDV challenges along the core dimensions of technology, applications, legal, social and business issues, data assets and skills.

The creation of i-Spaces is driven by the needs of large and small companies alike to ensure they can easily access the economic opportunities offered by BDV and develop working prototypes to test the viability of actual business deployments. This does not necessarily require moving data assets across borders; rather, data analytic tools and computation activities are brought to the data. In this way, valuable data assets are made available in environments that simultaneously support the legitimate ownership, privacy and security policies of corporate data owners and their customers, while

facilitating ease of experimentation for researchers, entrepreneurs and small and large IT providers.

Concerning the discovery of value creation, i-Spaces support various models: at one end, corporate entities with valuable data assets are able to specify business-relevant data challenges for researchers or software developers to tackle; at the other end, entrepreneurs and companies with business ideas to be evaluated are able to solicit the addition and integration of desired data assets from corporate or public sources. I-Spaces also contribute to filling the skills gap Europe is facing in providing (controlled) access to real use cases and data assets for education and skills improvement initiatives. I-Spaces themselves are data driven, both at the planning and the reporting stages. At the planning stage, they prioritise the inclusion of data assets that, in conjunction with existing assets, present the greatest promise for European economic development (while taking full account of the international competitive landscape); at the reporting stage, they provide methodologically sound quantitative evidence on important issues such as increases in performance for core technologies or reductions in costs for business processes. These reports foster learning and continuous improvement for the next cycle of technology and applications.

The particular European added value of i-Spaces is that they federate, complement and leverage activities of similar national incubators/environments, existing PPPs and other national or European initiatives. With the aim of not duplicating existing efforts, complementary activities considered for inclusion have to stand the test of expected economic development: new data assets and technologies are considered for inclusion to the extent that they can be expected to open new economic opportunities when added to and interfaced with the assets maintained by regional or national data incubators or existing PPPs.

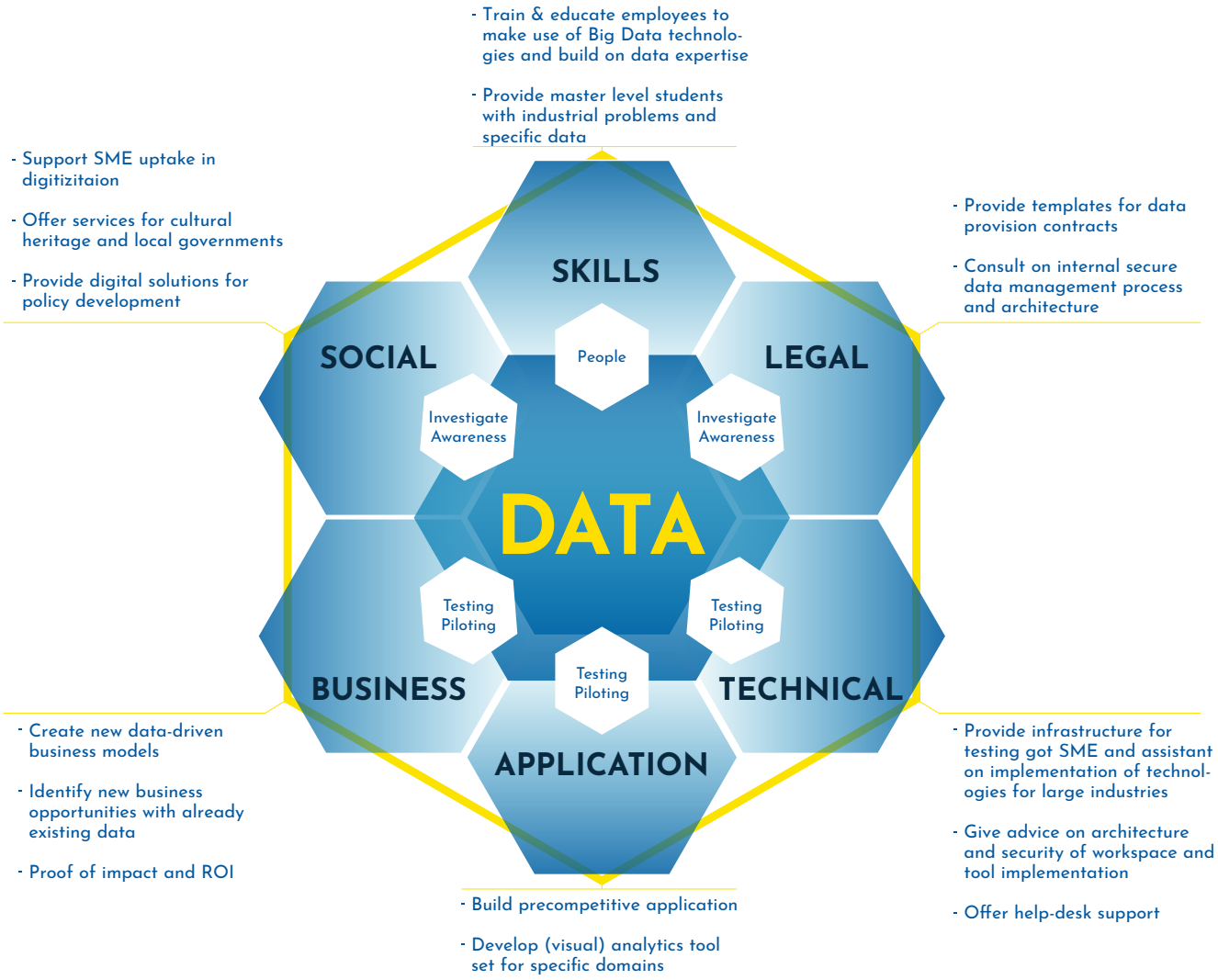
The successive inclusion of data assets into i-Spaces, in turn, drives and prioritises the agenda for addressing data integration or data processing technologies. One example is the existence of data assets with homogenous qualities (e.g. geospatial factors, time series, graphs and imagery), which calls for optimising the performance of existing core technology (e.g. querying, indexing, feature extraction, predictive analytics and visualisation). This requires methodologically sound benchmarking practices to be carried out in appropriate facilities. Similarly, business applications exploiting BDV technologies are evaluated for usability and fitness for purpose, thereby leading to the continuous improvement of these applications.

Due to the richness of data that i-Spaces offer, as well as the access they afford to a large variety of integrated software tools and expert community interactions, the data environments provide the perfect setting for the effective training of data scientists and domain practitioners. They encourage a broader group of interested parties to engage in data activities. These activities are designed to complement the educational offerings of established European institutions.

While economic development is the principal objective of BDV, this cannot happen without taking into proper account the legislative requirements pertaining to the treatment of data, as well as ethical considerations. In addition, BDV creates value for society as a whole by systematically supporting the transfer of sophisticated data management practices to domains of societal interest such as health, environment or sustainable development, among others. Especially when it comes to SMEs, the issues of skills and training, reliable legal frameworks, reference applications and access to an ecosystem become central for a rapid take-up of the opportunities offered by BDV.

In this holistic interdisciplinary approach, i-Spaces are a key mechanism that targets BDV challenges along the relevant dimensions, as depicted in Figure 4. I-Spaces are instrumental in the testing, showcasing and validation of new technology, applications and business models. The central need for the availability of open and industrial data assets is catered for, as well as the requirements for skills development, best practices identification, and favourable legal, policy and infrastructural frameworks and tools across sectors and borders.

Figure 4: Interconnected challenges of the PPP within i-Spaces



All i-Spaces provide a set of basic services to support Lighthouse projects and technical projects, as well as collaboration and coordination projects. These **basic services** include:

- **Community Building**: Contributing to the identification and management of stakeholder ecosystem communities along thematic and/or regional dimensions. This activity capitalises on existing thematic and/or regional initiatives.
- **Asset Support**: Supporting data providers in integrating datasets in a quality-secured way while maintaining a catalogue of available data assets.
- **ICT Support**: Providing basic ICT assistance as well as focused support from Big Data scientists and data specialists, and business development during research and innovation projects. This includes assistance in benchmarking datasets, technologies, applications, services and business models.
- **On-boarding**: Running an induction process for new project teams.
- **Resourcing**: Allocating the resources (computing, storage, networking, tools and applications) to individual research and innovation projects, and scheduling these resources among different projects.
- **Protection**: Data protection, including ensuring compliance with laws and regulations, and the deployment of cutting-edge, state-of-the-art security technologies in protecting data and controlling data access.
- **Privacy**: Data privacy and anonymisation in terms of handling and deleting personally identifiable information (PII) in compliance with laws and regulations such as the EU GDPR (General Data Protection Regulation), and the deployment of anonymisation technologies for preventing the processing of PII when necessary.
- **Data Governance**: Taking into account privacy and protection issues, defining the rules for accessing and sharing data. This includes the standardisation of procedures for sharing metadata, defining the (smart) contract between stakeholders, assessing technologies such as encryption and blockchain, and formulating the necessary solutions to orchestrate the agreed governance.
- **Federation**: Supporting linkages to other innovation spaces and facilitating experiments across multiple innovation spaces. An effective federation will help to support research and innovation activities through accessing and processing data assets across national borders (data spaces).
- **Business Support**: Facilitating start-ups and SME inclusion in the value creation process by leveraging community engagement.
- **Incubation and Acceleration**: Delivering all forms of suitable support to data-driven value creation projects by liaising with existing thematic, national or regional initiatives.

I-Spaces are understood as incubator environments, where the outcomes of research into novel technologies and applications can be quickly tested and piloted in a context that incorporates the maximum involvement of stakeholders, including business innovators and end-users.

To summarise, the main characteristics of i-Spaces are:

- Forming the **hubs for bringing technology and application developments together** while catering for the development of skills, competences and best practices. These environments offer new and existing technologies and tools from industry and open source software initiatives as a basic service to tackle the Big Data Value challenges.

- Ensuring that **data is at the centre** of Big Data Value activities. The i-Spaces make data assets based on industrial, private and open data sources accessible. I-Spaces are secure and safe environments that ensure the availability, integrity and confidentiality of data sources.

- Serving as **incubators for the testing and benchmarking** of technologies, applications and business models. This provides early insights into potential issues and helps to avoid failure in the later stages of commercial deployments. In addition, it is expected that this activity will provide **input for standardisation and regulation**.

- **Developing skills and sharing best practices** is an important task of i-Spaces and their federation. They will also link with other existing initiatives at both the European and national levels.

- New **business models and ecosystems** will emerge from exposing new technologies and tools to industrial and open data. The i-Spaces are a playground for testing new business model concepts and the emerging ecosystems of existing and new BDV 'players'.

- Gaining early insights into the **social impact** of new technologies and data-driven applications and how they will change the behaviour of individuals and the characteristics of data ecosystems.

- Acting as a catalyst to foster data-driven **communities** in the ecosystem and accelerate value creation.

Set-up of i-Spaces

To ensure that i-Spaces can achieve their ambitious objectives, the following design considerations are taken into account when setting up i-Spaces:

- A strong relationship with **Data Innovation ecosystems**, in particular to industrial and institutional data owners;

- The availability of a team providing basic **IT assistance**, as well as focused **support** by Big Data experts;

- **Business Development** resources to initiate and materialise projects, taking into account value, legal and technical dimensions.

- A well-managed **IT infrastructure**, including remote access capabilities.

- **Secure and trustworthy** data hosting and access.
- Project management resources to ensure delivery.

Key elements for the implementation of i-Spaces include at least the following:

- **A multidisciplinary team** able to manage the value creation process from community building to the final delivery of projects. In particular, this may include steering the progress of SMEs participating in i-Spaces activities and identifying and guiding the potential technology transfer from research. This should be performed in full coordination with existing thematic or regional initiatives.
- **Secure access to data storage** that provides the necessary security mechanisms required by industry, and other data asset owners, thus promoting trust in sharing data assets with scientists and data specialists for experimentation. At the same time, open data will be made available. Support will also be provided for running experiments on site or remotely, as well as data governance mechanisms for leveraging the different rights and duties (roles) within a data space.
- **Offering hybrid computing models**. The Cloud paradigm will be one important computing model for Big Data Value technology and thus i-Spaces, yet it will not be the only one. For instance, due to the volume and velocity of data, transferring this information from data sources (such as IoT sensors) to the Cloud providers might not be feasible. This means that i-Spaces infrastructures also consider other computing models, such as 'distributed computing' and 'high-performance computing', as well as 'computing at the edge'.
- **Delivering platforms and tools** from different sources, including open source and proprietary models, to enable data scientists and data engineers to develop and run new technology and applications. I-Spaces start from a 'state-of-the-art' base and will continuously evolve, incorporating new technology as it becomes available. They will contribute to the establishment of data lifecycle management and data organisation in order to develop methods for data preservation and curation as well as data sharing.
- **Providing a tool for continuous benchmarking**, so that businesses and, in particular, start-ups and SMEs, can evaluate whether their products and services will work in a real-world context.
- **Establishing tools for anonymisation** in order to prevent the identification of sensitive information which may interfere with the right to secrecy (e.g. industrial trade information) and privacy.
- **Business (model) benchmarking**. This form of benchmarking may, among other aspects, focus on process, financial or investor perspectives.
- **Technical benchmarking**. This type of benchmarking is about determining how the performance or operational cost of a product or service compares with existing products or services.
- **User experience benchmarking**. Besides performance and cost, the quality of the customer's experience of a product or service is key to its success. A user-centric approach is thus vital in providing products and services.

- **Dataset benchmarking**. The datasets are at the core of i-Spaces. Measuring and ensuring data quality, not only for existing datasets but also in live data streams, is the main concern here.

BDVA i-Space label

To ensure the quality of the European Innovation Spaces and connect existing initiatives under one umbrella, the BDVA set up the BDVA i-Space label. On a yearly basis, candidates from all over Europe are invited to apply for this label. The previously mentioned criteria are tested by way of a survey to collect the relevant information. A committee appointed by the BDVA Board of Directors (BoD) carefully examines individual candidates and recommends the appropriate quality ranking (bronze, silver, gold) to the BoD, which ultimately grants the label. In the future these activities will lead to the establishment of a European network of i-Spaces that will ensure easy access for industrial partners to foster the development and testing of precompetitive solutions on high-quality platforms offering certified services and training.

Further information about the ongoing PPP i-Spaces is available from http://www.big-data-value.eu/projects/

2.1.2 Lighthouse projects

Minor updates have been included reflecting the positioning of BDVA Lighthouses in the context of the new work programme.

Lighthouse projects[17] are projects with a high degree of innovation that **run large-scale data-driven demonstrations** whose main objectives are to create high-level impact and to promote visibility and awareness, leading to a faster uptake of Big Data Value applications and solutions.

They are the major mechanism to demonstrate Big Data Value ecosystems and sustainable data marketplaces, and thus promote increased competitiveness of established sectors as well as the creation of new sectors in Europe. Furthermore, they propose replicable solutions by using existing technologies or very near-to-market technologies that show evidence of data value and could be integrated in an innovative way.
Lighthouse projects will lead to explicit business growth and job creation, and so all projects are required to define clear indicators and success factors that can be measured and assessed in both qualitative and quantitative terms against these goals.

Increased competitiveness is not only a result of the application of advanced technologies; it also stems from a combination of changes that expand the technological

[17] Sometimes also labelled as large-scale demonstrations or pilots.

level, as well as political and legal decisions, among others. Thus, Lighthouse projects are expected to involve a combination of decisions centred on data, including the use of advanced Big-Data-related technologies, but also other dimensions. Their main purpose is to render results visible to a widespread and high-level audience in order to accelerate change, thus making the impact of Big Data in a specific sector, and/or a particular economic or societal ecosystem, explicit.

Lighthouse projects are defined through a set of well-specified goals that materialise through large-scale demonstrations deploying existing and near-to-market technologies. Projects may include a limited set of research activities, if that is needed to achieve their goals, but it is expected that the major focus will be on data integration and solution deployment.

Lighthouse projects are different to Proof of Concepts (which are more related to technology or process) or pilots (which are usually an intermediate step on the way to full production): they should pave the way for a faster market roll-out of technologies (Big Data with Cloud and HPC or the IoT); they should be conducted on a large scale; and they should use their successes to rapidly transform the way an organisation thinks or processes are run.

Sectors or environments to be included are not pre-determined but should be in line with the aforementioned goal of creating a high-level impact. For example, the deployment of new e-health services in the EU through a large-scale implementation of Electronic Health Records (EHRs) is aligned with the expected degree of impact; however, this would require not only the application of some Big Data technologies but also advocating the making of political and regulatory decisions in the fields of data privacy and interoperability.

The first call for Lighthouse projects made by the BDV PPP resulted in two actions in the domains of bio-economy (including agriculture, fisheries and forestry) and transport and logistics. The second call resulted in two actions for health and smart manufacturing. Therefore, even though additional use cases in those domains could be valuable, our aim is to diversify the sectorial approach of the PPP and ensure that the benefits of Big Data technologies expand over different industries.

Lighthouse projects operate primarily in a single domain, where a meaningful (as evidenced by total market share) group of EU industries from the same sector will jointly provide a safe environment in which they make available a proportion of their data (or data streams) and demonstrate, on a large scale, the impact of Big Data technologies. It is expected that projects will use data sources other than those of the specific sector addressed, thereby contributing to breaking silos. In all cases, projects are intended to have access to appropriately large, complex and realistic datasets.

One of the expected outcomes of this approach is data interoperability. Solutions at the EU level (i.e. going beyond national boundaries) and which avoid vendor lock-in are especially desired in an attempt to achieve economies of scale.

Projects are asked to show sustainable impact beyond the specific large-scale demonstrator/s running through a project's duration. This should be done whenever possible through solutions that can be replicable by other companies in the sector or by other application domains.

All Lighthouse projects have to involve, as appropriate, the relevant stakeholders to reach their goals. As a result, it is expected that complete data ecosystems will be developed. When needed, Lighthouse projects may use the infrastructure and ecosystems facilitated by one or more i-Spaces.

Some of the indicators that are used to assess the impact of Lighthouse projects are the number and size of datasets processed (integrated), the number of data sources made available for use and analysis by third parties, or the number of services provided for integrating data across sectors. Market indicators are obviously of utmost importance. Lighthouse projects are expected to contribute to:

- a 20% increase in market share in the corresponding sector through selling integrated data and/or data integration services;

- the establishment of cross-sectorial standards for data sharing at the EU level, when applicable.

The nature of these contributions should be made clear.

Key elements for the implementation of Lighthouse projects include at least the following areas.

The use of existing or close-to-market technologies: Lighthouses are not expected to develop completely new solutions; instead, they should make use of existing or close-to-market technologies and services by adding and/or adapting current relevant technologies, as well as accelerating the roll-out of Big Data value solutions using the Cloud and the IoT or HPC. Solutions should provide answers for real needs and requirements, showing an explicit knowledge of the demand side. Even though projects should concentrate on solving concrete problems and this may lead to specific deployments, the replicability of concepts should be a priority in order to ensure impact beyond the particular deployments of the project. Lighthouse projects should address frameworks and tools from a holistic perspective, considering, for example, not only analytics but also the complete data value chain (data generation, the extension of data storing, analysis, etc.).

Interoperability and openness: Projects should take advantage of both closed and open data; they can also determine if open source or proprietary solutions are the most suitable to address their challenges. However, they should promote the interoperability of solutions in order to avoid locking in customers.

Proposals are encouraged to include working with the respective communities and initiatives funded within the HPC PPP, and relevant Cloud and IoT projects. In particular, open specifications that allow different stakeholders to make their solutions interoperable and compatible with the proposed solutions are welcome. The focus should be on the creation of markets in line with the European Digital Single Market Strategy and on projects that take advantage of economies of scale. As mentioned previously, in many sectors pan-European solutions may require going beyond purely technological decisions.

The involvement of smaller actors (for example, through opportunities for start-ups and entrepreneurs) who can compete in the same ecosystem in a fair way should be a must. Open Application Programming Interfaces (APIs) could play an important role here

(e.g. third party innovation through data sharing). In addition, projects should focus on usability and reduce possible barriers or gaps resulting from Big Data methods impacting on end-users (break the 'Big Data for data analysts only' paradigm).

Performance: Proposals should contribute to common data collection systems and have a measurement methodology in place. Performance monitoring should last for at least two-thirds of the duration of the project. However, a longer-term commitment will give value to the proposal. For further information on quantitative impact check Work Programme 2018-20[18].

The setting-up of ecosystems: Lighthouse projects should have a transformational power; that is, they should not be restricted to a very narrow-minded experiment with a limited impact. They should demonstrate that they are able to improve (sometimes changing associated processes) the competitiveness of the selected industrial sector in a relevant way. This requires the active involvement of different stakeholders, and therefore attention should be paid to the ecosystem that will enable such changes. Lighthouse projects should be connected to communities of stakeholders from the design phase. Ecosystems should evolve, extend or connect existing networks of stakeholders and hubs.

As is well known, European industry is characterised by a huge number of small and medium-sized enterprises; a lack of consideration of this factor would lead to a less than healthy environment. Thus, adequate consideration of SME integration in the projects is required.

Even though proposals should focus on a particular sector, the use of data from different sources and industrial fields should be encouraged, with priority given to avoiding the 'silo' effect. The proposals should align and work with their communities and projects funded within the HPC PPP, and relevant Cloud and IoT projects.

Long-term commitment and sustainability: The budgets assigned to the projects should act as seeds for more widely implemented plans. It is expected that the proposed activities will be integrated into a more ambitious strategy which will involve additional stakeholders and further funding (preferably private but also possibly a combination of public and private).

As was mentioned before, four Lighthouse projects have already been selected and recently launched, leading to an evolution of the concept. That is why this updated version of the SRIA suggests more concrete requirements for the upcoming large-scale pilots, in some cases further specifying aspects that were already worked out. The following should be regarded as guidance rather than a complete list:

- It is important to reuse technologies and frameworks by combining and/or adapting existing relevant technologies (Big Data with the Cloud and HPC or IoT) that are already in the market or close to it (i.e. those with a high technology readiness level (TRL)) to avoid the development of new platforms where a good basis already exists (for example, as part of the Open Source community). Projects are especially encouraged to build on the technologies created by the ongoing projects of the Big Data PPP that fit their requirements (e.g. in the area of privacy-preserving technologies)

- Special attention should be paid to interoperability. This applies to all layers of the solution, including data (here, some of the results of the projects funded under the

[18] http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-leit-ict_en.pdf

Big Data PPP with a focus on data integration could be particularly useful), and to relevant efforts within the HPC, Cloud and IoT communities.

- It is expected that projects will combine the use of open and closed data. While it is understandable that some closed data will remain as such, we also expect these projects to contribute to increasing the availability of datasets that could be used by other stakeholders, such as SMEs and start-ups. This could happen under different regimes (not necessarily for free). Projects should declare how they will contribute to this objective by quantifying and qualifying datasets (when possible) and by including potential contributions to the ongoing data incubators/accelerators and Innovation Spaces.

- Lighthouse projects have to contribute to the horizontal activities of the Big Data PPP as a way of helping in the assessment of the PPP implementation and increasing its potential impact. Some of the targeted activities include contributing to the standardisation of activities, the measurement of KPIs, and coordination with the PPP branding, or active participation in training and educational activities proposed by the PPP.

2.1.3 Technical projects

Technical projects focus on addressing one issue or a few specific aspects identified as part of the BDV technical priorities (also see Section 3). In this way, technical projects provide the technology foundation for Lighthouse projects and i-Spaces. Technical projects may be implemented as Research and Innovation Actions (RIAs) or Innovation Actions (IAs), depending on the amount of research work required to address the respective technical priorities.

2.1.4 Cooperation and coordination projects

The aim of cooperation and coordination projects[19] is to work on detailed activities that ensure coordination and coherence in the PPP implementation and provide support to activities that fall under the skills, business, policy, regulatory, legal and social domains.

[19]For instance Collaboration and Support Actions (CSAs).

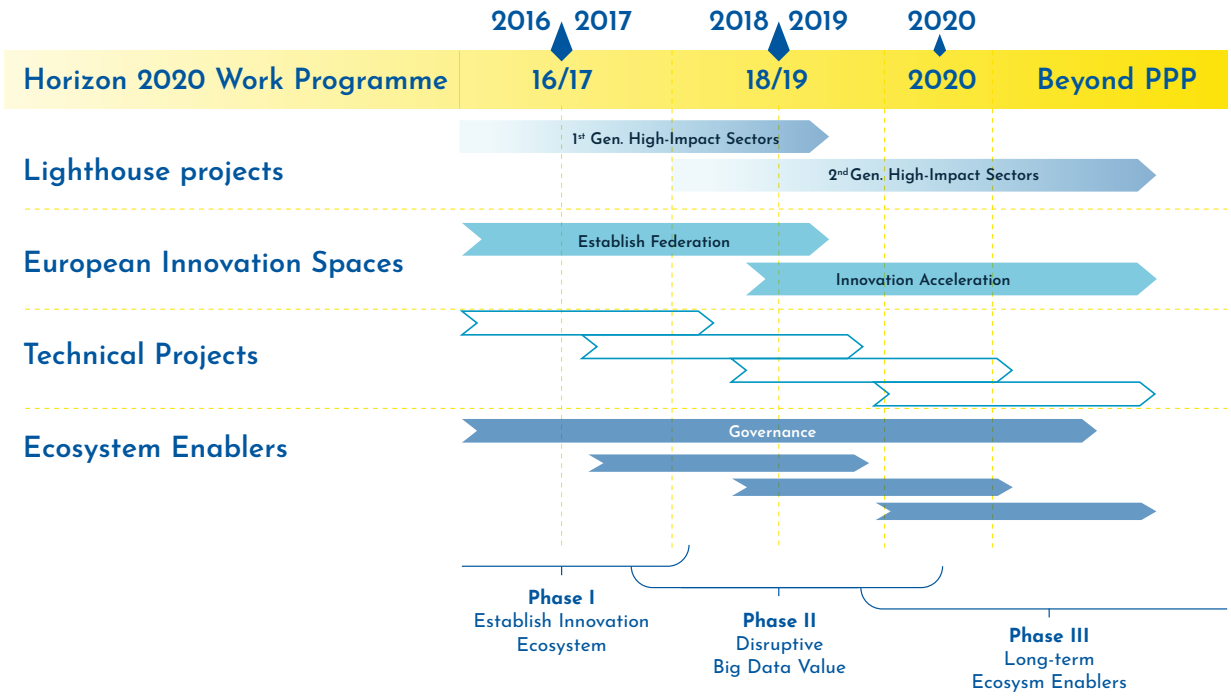
2.2 BDV Methodology

This section contains minor updates reflecting the positioning of the BDV PPP in the context of the new Work Programme.

The programme will develop the ecosystem in distinct phases of development, each with a primary theme. The three phases, as depicted in Figure 5, are:

- **Phase I**: Establish the ecosystem (governance, i-Space, education, enablers) and demonstrate the value of existing technology in high-impact sectors (Lighthouses, technical projects) (Work Programme WP 16–17).
- **Phase II**: Pioneer disruptive new forms of Big Data Value solutions (Lighthouses, technical projects) in high-impact domains of importance for EU industry, addressing emerging challenges of the data economy (WP 18–19).
- **Phase III**: Develop long-term ecosystem enablers to maximise sustainability for economic and societal benefit (WP 19–20).

Figure 5: Three-phase timeline of the BDV PPP



Phase I: Establish Innovation Ecosystem (2016–2017)

The first phase of the programme will focus on laying the foundations necessary to establish a sustainable European data innovation ecosystem. The key activities of Phase I include:

- Establish a European network of i-Spaces for cross-sectorial and cross-lingual data integration, experimentation and incubation (ICT14 – 2016–17).
- Demonstrate Big Data Value solutions via large-scale pilot projects in domains of strategic importance for EU industry, using existing technologies or very near-to-market technologies (ICT15 – 2016–17).
- Tackle the main technology challenges of the data economy by improving the technology, methods, standards and processes for Big Data Value (ICT16 – 2017).
- Advance the state of the art in privacy-preserving Big Data technologies and explore the societal and ethical implications of this (ICT18 – 2016).
- Establish key ecosystem enablers including programme support and coordination structures for industry skills and benchmarking (ICT17 – 2016-17).

Phase II: Disruptive Big Data Value (2018–2019)

Building on the foundations established in Phase I, the second phase will have a primary focus on Research and Innovation (R&I) activities to deliver the next generation of Big Data Value solutions. The key activities of Phase II include:

- Supporting the emergence of the data economy with a particular focus on accelerating the progress of SMEs, start-ups and entrepreneurs, as well as best practices and standardisation (ICT-13-c).
- Pioneering disruptive new forms of Big Data Value solutions with the Cloud and HPC or the IoT via large-scale pilot projects in emerging domains of importance for EU industry using advanced platforms, tools and test-beds (ICT-11, DT-ICT-11-2019).
- Tackling the next generation of Big Data research and innovation challenges for extreme-scale analytics (ICT-12-a).
- Addressing ecosystem roadblocks and inhibitors to the take up of Big Data Value platforms for data ecosystem viability, including platforms for personal and industrial data (ICT-13).
- Providing programme support (continuing), facilitating networking and cooperation among ecosystem actors and projects, and promoting community building between BDV, Cloud, HPC and IoT activities (ICT-12-b).

Phase III: Long-term Ecosystem Enablers (2019–2020)

While the sustainability of the ecosystem has been considered from the start of the PPP, the third phase will have a specific focus on activities that can ensure long-term self-sustainability. The key activates of Phase III include:

- Sowing the seeds for long-term ecosystems enablers to ensure self-sustainability beyond 2020 (ICT-13).

- Creating innovation projects within a federation of i-Spaces (European Digital Innovation Hubs for Big Data) to validate and incubate innovative Big Data Value solutions and business models (DT-ICT-05-2020).

- Ensuring continued support for technology outputs of PPP (Lighthouses, R&I, CSA), including non-technical aspects (training) beyond 2020 (i.e. Open Source Community, Technology Foundation).

- Establishing a Foundation for European Innovation Spaces with a charter to continue collaborative innovation activity beyond 2020, in line with the concept of European Digital Innovation Hub for Big Data.

- Liaising with private funding (including Venture Capital) to accelerate entry into the market and socio-economic impacts, including the provision of ancillary services to develop investment-ready proposals and support scaling for BDV PPP start-ups and SMEs to reach the market.

- Tackling the necessary strategy and planning for the BDV Ecosystem until 2030, including the identification of new stakeholders, emerging usage domains, technology, business and policy road-mapping activity (ICT-13).

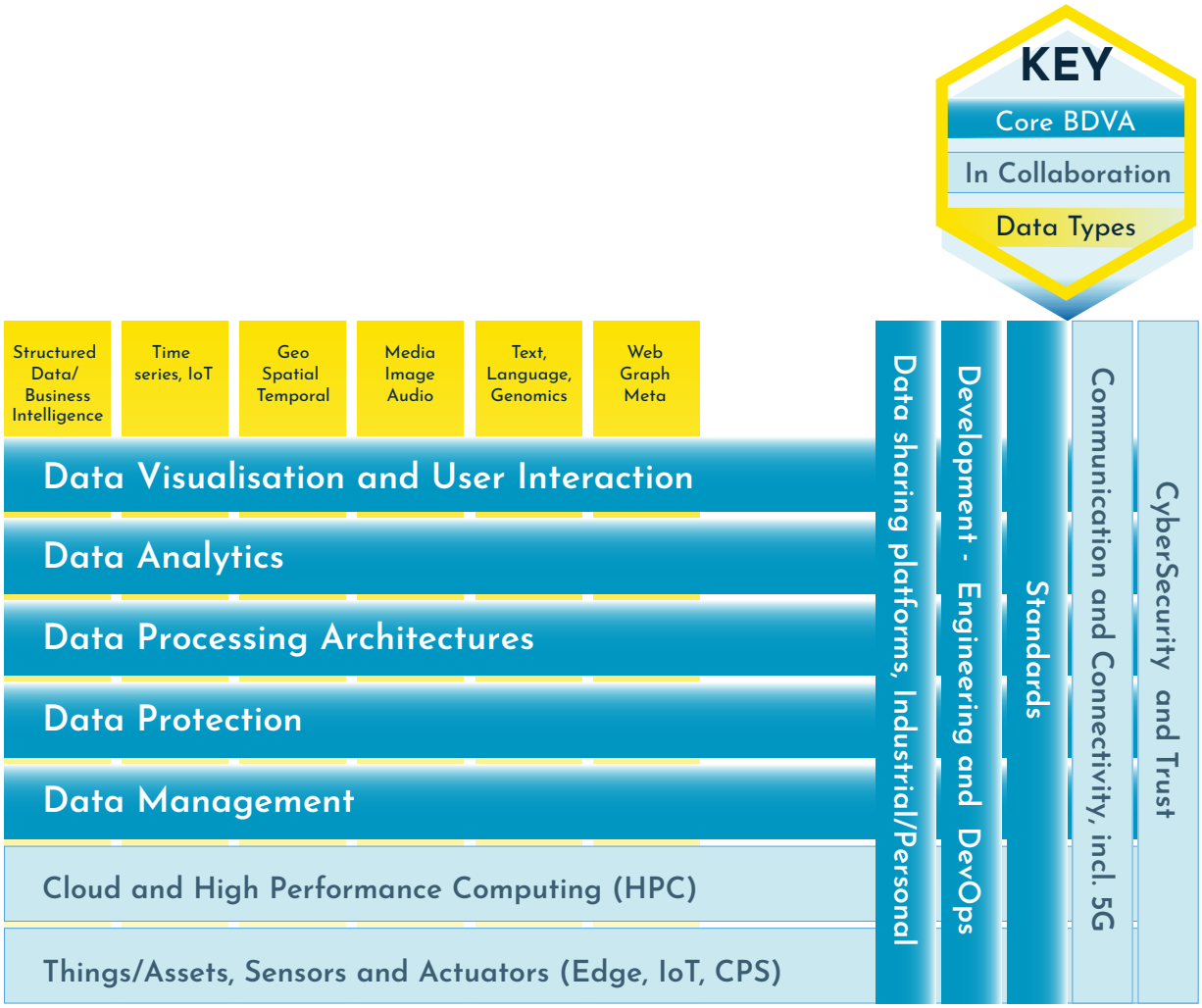
2.3 BDV Reference Model

This section was added to SRIA version 4.

Overview

In order to structure the discussions in the remainder of this SRIA, we now describe the Big Data Value Reference Model, as shown in Figure 6.

Figure 6: Big Data Value Reference Model



The BDV Reference Model has been developed by the BDVA, taking into account input from technical experts and stakeholders along the whole Big Data Value chain, as well as interactions with other related PPPs. The BDV Reference Model may serve as 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 distinguishes between two different elements. On the one hand, it describes the elements that are at the core of the BDVA; on the other, it outlines the features that are developed in strong collaboration with related European activities.

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 structure. However, the BDV Reference Model 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:

Horizontal concerns

- *Data Visualisation and User Interaction:* Advanced visualisation approaches for improved user experience. This is described further in Section 3.4.

- *Data Analytics:* Data analytics to improve data understanding, deep learning and the meaningfulness of data. This is described further in Section 3.3.

- *Data Processing Architectures:* Optimised and scalable architectures for analytics of both data-at-rest and data-in-motion, with low latency delivering real-time analytics. This is described further in Section 3.2.

- *Data Protection:* 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. This is described further in Section 3.5.

- *Data Management:* Principles and techniques for data management. This is described further in Section 3.1.

- *The Cloud and High Performance Computing (HPC):* Effective Big Data processing and data management might imply the effective usage of Cloud and High Performance Computing infrastructures. This area is separately elaborated further in collaboration with the Cloud and High Performance Computing (ETP4HPC) communities. This is covered in Sections 2.5.2 and 2.5.1.

- *IoT, CPS, Edge and Fog Computing:* A main source of Big Data is sensor data from an IoT context and actuator interaction in Cyber Physical Systems. In order 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. This is covered in Section 2.5.4.

Vertical concerns

- *Big Data Types and Semantics:* The following 6 Big Data types have been identified, based on the fact that they often lead to the use of different techniques and mechanisms in the horizontal concerns, which should be considered, for instance, for data analytics and data storage: (1) Structured data; (2) Time series data; (3) Geospatial data; (4) Media, Image, Video and Audio data; (5) Text data, including Natural Language Processing data and Genomics representations; and (6) Graph data, Network/Web data and Metadata. In addition, it is important to support both the syntactical and semantic aspects of data for all Big Data types.

- *Standards:* Standardisation of Big Data technology areas to facilitate data integration, sharing and interoperability. This is described further in Section 3.6.

- *Communication and Connectivity:* Effective communication and connectivity mechanisms are necessary in providing support for Big Data. This area is separately further elaborated, along with various communication communities, such as the 5G community. This is covered in Section 2.5.5.

- *Cybersecurity:* Big Data often need support to maintain security and trust 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. This is covered in Section 2.5.3.

- *Engineering and DevOps for building Big Data Value systems:* This topic will be elaborated in greater detail along with the NESSI Software and Service community. This is described further in Section 3.7.

- Marketplaces, Industrial Data Platforms and Personal Data Platforms (IDPs/PDPs), Ecosystems for Data Sharing and Innovation Support: Data platforms for data sharing include, in particular, IDPs and PDPs, but also other data sharing platforms like Research Data Platforms (RDPs) 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. This is described in detail in Section 2.4.

2.4 Platforms for Data Sharing

This section was added to SRIA version 4.

Data sharing and trading are seen as important ecosystem enablers in the data economy, although closed and personal data present particular challenges for the free flow of data. 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.

2.4.1 Industrial Data Platforms (IDPs)

IDPs have increasingly been touted as potential catalysts for advancing the European data economy. Recent versions of the EC's Work Programme have called for innovative actions to establish or evolve IDPs as solutions for emerging data markets, focusing on the need to offer secure and trusted data sharing to interested parties, primarily from the private sector (industrial implementations). The Digitising European Industry initiative has also identified IDPs as one of two horizontal topics, alongside the IoT, which can be seen as drivers for next-generation digital platforms[20]. The focus of activities within the Digital Industrial Platforms working group (DEI-WG2) is on 'strengthening Europe's position in digital technologies and digital industrial platforms across value chains in industrial sectors'. IDPs are also highlighted by the Digital Single Market strategy as offering solutions to some of the identified legal issues hampering the establishment of a European Data Economy[21].

The IDP conceptual solution is oriented towards proprietary (or closed) data, and its realisation should guarantee a trusted, secure environment within which participants can safely, and within a clear legal framework, monetise and exchange their data assets. A functional realisation of a continent-wide IDP promises to significantly reduce the existing barriers to a free flow of data within an advanced European Data Economy. The establishment of a trusted data-sharing environment will have a substantial impact on the data economy by incentivising the marketing and sharing of proprietary data assets (currently widely considered by the private sector as out of bounds) through guarantees for fair and safe financial compensations set out in black-and-white legal terms and obligations for both data owners and users. The 'opening up' of previously guarded private data can thus vastly increase its value by several orders of magnitude, boosting the data economy and enabling cross-sectorial applications that were previously unattainable or only possible following one-off bilateral agreements between parties over specific data assets.

[20] Report of WG2: Digital Industrial Platforms, final version, August 2017, p. 53, https://ec.europa.eu/futurium/en/implementing-digitising-european-industry-actions/report-wg2-digital-industrial-platforms-final

[21] Commission staff working document on the free flow of data and emerging issues of the European data economy, January 2017, p. 18, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41247

The IDP conceptual solution complements the drive to establish BDVA i-Spaces by offering existing infrastructure and functional technical solutions that can better regulate data sharing within the innovation spaces. This includes better support for the secure sharing of proprietary or 'closed' data within the trusted i-Space environment. Moreover, i-Spaces offer a perfect testbed for validating existing implementations of conceptual solutions such as the IDP.

The identified possibilities for action can be categorised into two branches:

- ***Standardisation:*** Addressing the lack of an existing standard platform (technical solution) that limits stakeholders from participating in the European Digital Single Market, and the availability of clear governance models (reference models, guidelines and best practices) regulating the secure and trusted exchange of proprietary data.

- ***Implementation:*** Establishing, developing or aligning existing IDP implementations to provide a functional European-wide infrastructure within which industrial participants can safely, and within a clear legal framework, monetise and exchange data assets.

Standardisation activities outlined by the SRIA (Section 3.6) have taken into account the need to accommodate activities related to the evolving IDP solutions. The opportunity to drive forward emerging standards also covers the harmonisation of reference architectures and governance models put forward by the community. Notable advanced contributions in this direction include the highly relevant white paper and the reference architecture[22] provided by the Industrial Data Space (IDS) Association. The Layered Databus, introduced by the Industrial Internet Consortium[23], is another emerging standard advocating the need for data-centric information-sharing technology that enables data market players to exchange data within a virtual and global data space.

The implementation of IDPs needs to be approached on a European level, and existing and planned EU-wide, national, and/or regional platform development activities could contribute to these efforts. The industries behind existing IDP implementations, including the IDS reference architecture and other examples such as the MindSphere Open Industrial Cloud Platform[24], can be approached to move towards a functional European Industrial Data Platform. The technical priorities outlined by the SRIA, particularly the Data Management priority (see Section 3.1), need to address data management across a data ecosystem comprising both open and closed data. The broadening of the scope of data management is also reflected in the latest BDVA reference model, which includes an allusion to the establishment of a digital platform whereby marketplaces regulate the exchange of proprietary data.

The Digitising European Industry WG2 identifies three vertical industrial sectors that could immediately benefit from more advanced IDP implementations: Connected Smart Factories; Smart Agriculture; and the Digital Transformation of Health and Care. Starting from these sectors, other communities could be approached to participate in validations of key Europe-wide IDP implementations.

[22] Reference Architecture Model for the Industrial Data Space, April 2017, https://www.fraunhofer.de/content/dam/zv/de/Forschungsfelder/industrial-data-space/Industrial-Data-Space_Reference-Architecture-Model-2017.pdf
[23]The Industrial Internet of Things, Volume G1: Reference Architecture, January 2017, https://www.iiconsortium.org/IIC_PUB_G1_V1.80_2017-01-31.pdf
[24]MindSphere: The cloud-based, open IoT operating system for digital transformation, Siemens, 2017, https://www.plm.automation.siemens.com/media/global/en/Siemens_MindSphere_Whitepaper_tcm27-9395.pdf

2.4.2 Personal Data Platforms (PDPs)

So far consumers have trusted, including companies like Google, Amazon, Facebook, Apple and Microsoft, to aggregate and use their personal data in return for free services. While EU legislation, through directives such as the Data Protection Directive (1995) and the ePrivacy Directive (1998), has ensured that personal data can only be processed lawfully and for legitimate use, the limited user control offered by such companies and their abuse of a lack of transparency have undermined the consumer's trust. In particular, consumers experience everyday leakage of their data, traded by large aggregators in the marketing networks for a value only returned to consumers in the form of often unwanted digital advertisements. This has recently led to a growth in the number of consumers adopting ad blockers to protect their digital life[25], while at the same time they are becoming more conscious of and suspicious about their personal data trail.

In order to address this growing distrust, the concept of Personal Data Platforms (PDP) has emerged as a possible solution that could allow data subjects and data owners to remain in control of their data and its subsequent use[26]. PDPs leverage 'the concept of user-controlled cloud-based technologies for storage and use of personal data ("personal data spaces")'[27]. However, so far consumers have only been able to store and control access to a limited set of personal data, mainly by connecting their social media profiles to a variety of emerging Personal Information Management Systems (PIMS). More successful (but limited in number) uses of PDPs have involved the support of large organisations in agreeing to their customers accumulating data in their own self-controlled spaces. The expectation here is the reduction of their liability in securing such data and the opportunity to access and combine them with other data that individuals will import and accumulate from other aggregators. However, a degree of friction and the lack of a successful business model is still hindering the potential of the PDP approach.

A new driver behind such a self-managed personal data economy has recently started to appear. As a result of consumers' growing distrust, measures such as the General Data Protection Regulation (GDPR), which will be in force from May 2018, have emerged. The GDPR will constitute the single pan-European law on data protection, and, among other provisions and backed by the risk of incurring high fines, it will force all the companies dealing with European consumers to (1) increase transparency, (2) provide users with granular control for data access and sharing, and (3) guarantee consumers a set of fundamental individual digital rights (including the right to rectification, erasure, data portability and to restrict processing). In particular, by representing a threat to the multibillion Euro advertising business, we expect individuals' data portability right, as enshrined in the GDPR, to be the driver for large data aggregators to explore new

[25]Used by 615 million devices at the end of 2016, http://uk.businessinsider.com/pagefair-2017-ad-blocking-report-2017-1?r=US&IR=T

[26]See a Commission paper on 'Personal information management services – current state of service offers and challenges' analysing feedback from public consultation: https://ec.europa.eu/digital-single-market/en/news/emerging-offer-personal-information-management-services-current-state-service-offers-and

[27]A Personal Data Space is a concept, framework and architectural implementation that enables individuals to gather, store, update, correct, analyse and/or share personal data. This is also a marked deviation from the existing environment where distributed data is stored throughout organisations and companies internally, with limited to no access or control from the user that the information concerns. This is a move away from the B2B (business to business) and B2C (business to consumer) models, with a move towards Me2B – when individuals start collecting and using data for their own purposes, and sharing data with other parties (including companies) under their control (https://www.ctrl-shift.co.uk/news/2016/09/19/shifting-from-b2c-to-me2b/).

business models for personal data access. As result, this will create new opportunities for PDPs to emerge. The rise of PDPs and the creation of more decentralised personal datasets will also open up new opportunities for SMEs that might benefit from and investigate new secondary uses of such data, by gaining access to them from user-controlled personal data stores – a privilege so far available only to large data aggregators. However, further debate is required to reach an understanding on the best business models (for demand and supply) to develop a marketplace for personal data donors and what mechanisms are required to demonstrate transparency and distribute rewards to personal data donors. Furthermore, the challenges organisations face in accessing expensive data storage, and the difficulties in sharing data with commercial and international partners due to the existence of data platforms which are considered to be unsafe, need to be taken into account. Last but not least, questions around data portability and interoperability also have to be addressed.

The benefits of adopting PDPs will align neatly with BDVA values, by increasing community engagement and participation through the creation of new societal value from big personal-data-based services and the development of new business models for SMEs. A number of task forces have committed to integrating activities on personal data and user control through the organisation of stakeholder meetings and workshops at the BDVA General Assembly meetings, as well as through joint sessions with external events. The BDVA's collaboration with other PPPs, such as the European Cyber Security Organisation (ECSO), will ensure that both privacy and security in the context of personal data platforms will be discussed.

2.5 European Data Value Ecosystem Development

This section, covering all ongoing engagements with BDV stakeholders, is new to this SRIA version.

Developing the European Data Value Ecosystem is at the core of the mission and strategic priorities of the Big Data Value Association and the Big Data Value PPP. The European Data Value Ecosystem brings together **communities** (all the different stakeholders who are involved, affected or stand to benefit), **technology, solutions** and **data platforms**, experimentation, incubation and know-how **resources**, and the **business models** and **framework conditions** for the data economy. In this section we refer to the 'community' and stakeholder's aspect of the European Big Data Value Ecosystem.

A dimension to emphasise in the European Data Value Ecosystem is its two-fold nature of vertical versus horizontal in respect to the different sector or application domains (e.g. transport health, energy, etc.). While specific data value ecosystems are needed per sector (in relation to particular targeted markets, stakeholders, regulations, type of users, data types, challenges, etc.), one of the main values identified for the Big Data Value Association and the PPP is its horizontal nature, allowing cross-sector value creation, considering both the reuse of value from one sector to another, and the creation of new innovations based on cross-sector solutions and consequently new value chains.
The Big Data Value PPP and the Big Data Value Association have defined instruments

and actions to serve both the horizontal and vertical nature of the Data Value (e.g. Lighthouse projects, a dedicated Task Force and targeted collaborations for sector-specific ecosystems, or I-Spaces, data skills and most of the BDVA collaborations at the horizontal level).

The **Big Data Value ecosystem project** (BDVe project) has a crucial role in the development of this ecosystem and it is contributing in a number of areas: by delivering a European **Big Data Landscape**; through its **engagement with National/Regional initiatives**; by developing a **Marketplace for Big Data services/solutions**; and through the establishment of an **Investment Forum** to promote European data-related start-ups. In performing the coordinating and support actions of the PPP the BDVe project also has a central position in clustering all the PPP projects, stakeholders and outcomes together under the umbrella of this European BDV Ecosystem. The BDVe project also has resources to deliver outreach to the so-called user communities (mainly industrial players from the different sectors).

Of particular importance is the **Big Data Landscape**, which will geographically represent organisations involved in the Big Data Value ecosystem in Europe. The map will represent different tiers of information and make it possible for users to conduct an intelligent search and filtering of these layers. The map will represent categories of information such as, existing national initiatives in Big Data, the distribution of use cases for all PPP projects, SMEs/start-ups working in Big Data, organisations specialised in Big Data solutions for vertical applications, among others. The map will also be connected to the marketplace. One of the main missions of The **Big Data Value Association** is to enable and accelerate European data-driven innovation through the development of an interoperable data-driven ecosystem. Its **membership strategy** pursues balance at the geographical level, for each type of stakeholder and membership, and in terms of the quality and level of contributions. At the time of delivering the SRIA version 4 the BDVA had reached almost 200 members, with representation in 27 different countries (mainly EU Member States), and a good balance between industry and research interests. Over 30% of its membership is made up of SMEs and approximately 20% constitutes large corporate bodies. Contributions from members are organised around task forces that help the community to deliver Data Innovation Guidelines, support Big Data Standards priorities, and provide access to know-how, infrastructure and business to business (B2B) or research to business cross-border collaboration. In addition to its members, the BDVA offers easy access to individuals and organisations who wish to become part of the **BDVA extended community** and thus gain faster access to news and know-how.

Establishing **collaboration** with other European, international and local organisations is crucial for the development of the Ecosystem, to generate synergies between communities and impact in the fields of Research and Innovation, Standards, Regulations, Markets and Society.

Collaborations, in particular with other PPPs, European and international standardisation bodies, industrial technology platforms, data-driven research and innovation initiatives, user organisations and policy makers have been identified and developed at national, European and international levels since the launch of the PPP and the creation of the Association (linked to SRIA v1), influencing the level of maturity of these collaborations.

This section delves more deeply into developed collaborations with an impact on technology integration and the digitisation of industry challenges, and in particular into collaborations with the ETP4HPC (European Technology Platform for HPC) (for HPC), ECSO (for cybersecurity), AIOTI (for IoT), 5G (through 5G PPP), the European Open Science Cloud (EOSC) (for the Cloud), and the European Factories of the Future

Research Association (EFFRA) (for factories of the future). Other sections of this document and a tailored paper developed by the BDV Association provide further details on other collaborations[28].

This ecosystem is further enriched by the collaboration established with sectorial user and data communities (verticals), such as those developed by the Lighthouse projects (on bio-economy and mobility, transport and logistics) and the application domains under BDVA Task Force 7 (Health, Media, Energy, Smart Cities, Geospatial Issues, Finances, Smart Manufacturing, etc.)

2.5.1 High Performance Computing with ETP4HPC

In some sectors, Big Data applications are expected to move towards more compute-intensive algorithms to reap deeper insights across descriptive (explaining what is happening), diagnostic (exploring why it happens), prognostics (predicting what can happen) and prescriptive (proactive handling) analysis. The adoption of certain HPC-type capabilities by the Big Data analytics' stack is likely to be of assistance where Big Data insights will be of the upmost 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[29]) have recognised their common 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. Specific technical challenges are detailed in the relevant technical sections of this document – under the sub-heading 'High Performance Data Analytics'. 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.

2.5.2 European Cloud Initiative with EOSC

Big 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

[28] E.g. collaboration with standardisation bodies is described in Section 3.7. Sector- and collaboration-specific papers can be found at http://www.bdva.eu/
[29] European Technology Platform for High Performance Computing: http://www.etp4hpc.eu/

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. About 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[30]. 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[31].

In response to these challenges, the Commission has launched a large effort with the objective of creating '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 …'[32]. The initial outline for the **European Open Science Cloud (EOSC)** was laid out in the report from the High Level Expert Group[33]. The report advised the Commission on a number of 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 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 for scientific data[34]. The recently approved **EOSC-hub** project[35]will create an integration and management system (the Hub) for the EOSC, the aim of which is to deliver a catalogue of services, software and data from major European research e-infrastructures.

The EU has already initiated and will go on to launch several more infrastructure projects, like EOSC-hub, within H2020 for implementing and piloting the EOSC. In addition to these projects, Germany and the Netherlands, among others, are promoting the **GO FAIR initiative**[36]. The FAIR principles aim to ensure that Data and Digital Research Objects are **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR)[37]. 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 in nature.

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

[30]'Are FAIR data principles FAIR?' LIBER Webinar by Alastair Dunning, 10.03.2017.
[31]G. Press, 'Cleaning Big Data: most time-consuming, least enjoyable data science task, survey says', Forbes [Internet], 2016 March 23. Available at http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3643728a7f75
[32]C. Moedas, Commissioner for Research, Science and Innovation, 19.4.2016.
[33]Realising the European Open Science Cloud, 2016, https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf
[34]EOSC pilot project: http://www.eoscpilot.eu/
[35]EOSC-hub project: http://go.egi.eu/eosc-hub
[36]Joint Position Paper on the European Open Science Cloud, 30.05.2017, Germany and the Netherlands.
[37]M. Wilkinson et al., 'The FAIR Guiding Principles for scientific data management and stewardship', http://www.nature.com/articles/sdata201618.pdf

community as a whole, including the 'long tail' of science.

The EOSC initiative is clearly aligned with the BDVA agenda, as both promote data accessibility, trustworthiness and reproducibility over domains and borders. In the BDVA, this particularly 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.

2.5.3 Cybersecurity with ECSO

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/or 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 that both technologies can be intertwined. Data Sovereignty consists in merging personal data from several sources, always allowing the data owner to keep control over their own 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 data spaces 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, a number of 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; 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 important to define the concepts of measurable trust and evidence-based trust. Data should be secured at rest and in motion.

ECSO represents the contractual counterpart to the European Commission for the implementation of the Cybersecurity contractual Public-Private Partnership (cPPP)[38]. A collaboration with ECSO, supporting the Cybersecurity PPP, has been initiated and further steps planned.

[38] The European Cyber Security Organisation (ECSO) is a fully self-financed not-for-profit organisation established under Belgian law in June 2016. ECSO members include a wide variety of stakeholders such as large companies, SMEs and start-ups, research centres, universities, end-users, operators, clusters and associations, as well as European Member States' local, regional and national administrations, countries that are part of the European Economic Area (EEA) and the European Free Trade Association (EFTA), and H2020 associated countries (https://www.ecs-org.eu).

2.5.4 Internet of Things with AIOTI

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. This will trigger significant growth in the amount of data. IDC estimates that machine-generated data will grow from 1.5ZB in 2013 to 18ZB in 2018[39].

This growth in data will lead to a future market expansion in the IoT business; for instance, IDC's prognosis in May 2014 was that the revenue forecast for the IoT will increase from $1.9 trillion in 2013 to $7.1 trillion in 2020, while in November 2015 Gartner predicted a market growth of connected objects from 6.4 billion in 2016 to 20.8 billion in 2020.

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, 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**: Using 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 to be made.

- **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.

[39] IDC CEO Summit 2015.

- **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 in their applications will be increased.

2.5.5 Connectivity and data access with 5G PPP

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 1 000 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 guaranties 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 monitor 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 rather 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.

2.5.6 Factories of the Future with EFFRA

The materialisation of the EFFRA 2020 Roadmap[40] in the last three years of H2020 (2018–2020) is driven and coordinated by a Consultation Document called 'Factories 4.0 and Beyond' released by EFFRA in September 2016. 'Factories 4.0 and Beyond' updates 'Factories of the Future 2020 Roadmap', duly considering the increasing impact on manufacturing by advanced ICT technologies when adopted in synergy with advanced material processing technologies (e.g. additive manufacturing) and mechatronics systems (Cyber Physical Systems). 'Factories 4.0 and Beyond' identifies five key priority areas and targets which EFFRA proposes for the 'Factories of the Future' work programme 2018-19-20:

- **Agile value networks:** Lot-size one – distributed manufacturing;
- **Excellence in manufacturing:** Advanced manufacturing processes and services for zero-defect and innovative processes and products;
- **The human factor:** Developing human competences in synergy with technological progress;
- **Sustainable value networks:** Manufacturing driving the circular economy;
- **Inter-operable digital manufacturing platforms:** Supporting an ecosystem of manufacturing services.

At the recent BDVA Valencia Summit (November 2016), EFFRA detailed the five key priorities above in seven main Research Headlines concerned with the Big Data and Industrial Analytics domains:

- HL16 Digital Factory Modelling and Simulation, including Real–Digital World Synchronisation (the Digital Twin);
- HL17 Multiple Source (Big) Data Mining and Real Time Advanced Analytics in Product and Production Lifecycle Ecosystems;
- HL19 Digitisation of the Supply Chain – Manage complex customer-driven value networks;
- HL22 Manufacturing as a Service (MaaS) – Servitisation of autonomous and reconfigurable production systems;
- HL25 Digital Platforms Interoperability and Open Standards development;
- HL26 Security, Privacy and Liability – Cybersecurity and Industrial Safety;
- HL28 European Circular Economy Open Platform.

[40] The European Factories of the Future Research Association (EFFRA) is the private section of the Factories of the Future FoF cPPP. Its 2020 roadmap includes six domains: (i) advanced manufacturing processes; (ii) adaptive and smart manufacturing systems; (iii) digital/virtual and resource-efficient factories; (iv) collaborative and mobile enterprises; (v) human-centred manufacturing; (vi) customer-focused manufacturing.

The BDVA subgroup on Smart Manufacturing Industry (SMI) aims to match the BDVA SRIA 5 Technical Priorities with three main Manufacturing Industry Grand Challenges: Smart Factories; Smart Supply Chains; and Smart Product Lifecycles. Industrial scenarios, use cases and requirements are being analysed in order to identify research and innovation challenges to be addressed in the next few years.

The relation between the three SMI Grand Challenges and the seven Research Headlines identified by EFFRA will be further expanded in the upcoming activities of the SMI subgroup, and especially in the planned Position Paper. This paper elaborates on that relationship, including topics such as Digital Twins, Predictive Maintenance, Data and Protocols semantics and Open Standards; in addition it reports on the concept of Industrial Data Platforms in the context of the manufacturing domain.

3. TECHNICAL ASPECTS

The identification of emerging technical priorities and the updating of existing technical priorities was based on a three-way analysis. First, the BDVA Task Force TF6 'Technical' gathered key stakeholders and experts to discuss technical trends and open issues, resulting in a consolidated view among the BDVA members. In particular, this led to a structuring of the technical priorities into the BDV Reference Model (explained in section 2.3). Second, ongoing BDV PPP projects were asked to identify which of the technical priorities are being addressed by the running projects and where there may still be gaps that require future research and innovation. Third, feedback was solicited from partners of the BDV strategic collaborations (see Section 2.5).

Based on this analysis, the overall, updated strategic technical goal may be summarised as:

Deliver Big Data technology empowered by deep analytics for data-at-rest and data-in-motion, while providing data protection guarantees and optimised user experience, through sound engineering principles and tools for data-intensive systems.

Section 2.3 provides an introduction to the BDV Reference Model, which structures the technical priorities identified during the needs analysis. The individual technical priorities are presented in Sections 3.1 to 3.4. Finally, Section 3.8 gives an illustrative example for how the elements in the BDV reference model help to address the Big Data concerns of a specific industry sector.

3.1 Priority 'Data Management'

Background

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 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 use existing tools and to align available resource, because they are generally provided 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.

In almost all industrial sectors, isolated and fragmented data pools are found. Due to the prevalence of **data silos**, it is difficult 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/or federation of data between pools imposes high levels of additional cost. 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 a 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, a 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 needs to be syndicated from data providers to data consumers while maintaining provenance, control and source information including IPR considerations (**data provenance**). Moreover, in order to ensure transparent and flexible data usage, the aggregating and managing of respective datasets enhanced by a controlled access mechanism through APIs should be enabled (**Data-as-a-Service**).

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, on-line and off-line 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 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 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, data anonymisation, etc., in a decentralised manner.

Outcomes

The main expected advances in data management are the following:

- Languages, techniques and tools for measuring and assuring **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 diminishing 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/or 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 the data) and using scalable big data transformations approaches (including aspects of automatic, interactive, sharable and repeatable transformations), to long-term storage and data access. New models and tools to check integrity and veracity of data, through both machine-based and human-based (crowd-sourcing) 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'

Minor updates are included, reflecting the feedback from the BDVA meetings with ETP4HPC.

Background

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, a proper integration of data-in-motion with the 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.

Challenges

There have been a number of 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,** various levels of granularity, etc. Besides, data can be structured, semi-structured or unstructured, multimedia, audio-visual or textual data. Hardware can be heterogeneous too (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 in order 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 have to be adapted to allow continuous querying over streams (i.e., on-line processing). The scenarios for Big data processing also require a greater ability to cope with the systems which inherently contain dynamics in their daily operation, alongside their proper management, in order 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, missing contextual data, etc. The distribution of Big Data processing nodes pose 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 an efficient distribution of the processing

to the Edge (i.e. local data edge processing and analytics), as a part of the ever-increasing trend for 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 like main memory technology, with different type 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 huge increase in the complexity of Big Data applications, usually driven by the 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 will 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 will influence Big Data architectures in the future.

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 with the aim of updating the analysis results as the information content changes. It is important to access only the relevant and suitable data, thereby avoiding the access and processing of 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 the 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, named High Performance Data Analytics (HPDA).

3.3 Priority 'Data Analytics'

Minor updates are included, reflecting the feedback from the BDVA meeting with ETP4HPC.

Background

The progress of data analytics is not only key 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 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[41] 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 future main driver for US economic growth, ahead of shale oil and gas production[42].

[41] 'Big Data: the next frontier for innovation, competition and productivity', McKinsey Global Institute, June 2011.

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 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 extremely challenging because data streams, besides 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 extreme huge datasets.

Challenges

Understanding data, whether it is numbers, text, or multimedia content, has always been one of the greatest challenges for data analytics. Entering into 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 of the scalability and processing speed for the aforementioned algorithms in order to tackle linearisation and computational optimisation issues;

- **Advanced business analytics and intelligence:** All 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 is necessary, especially for SMEs.

- **Predictive and prescriptive analytics:** Machine learning, clustering, pattern mining, network analysis and hypothesis testing techniques applied on extremely

[42] 'Game changers: five opportunities for US growth and renewal', McKinsey Global Institute, 2013.

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 will open up novel opportunities for predictive analytics in terms of predicting future situations, and even prescriptive analytics with regard to providing actionable insights based on forecasts.

- **High Performance Data Analytics (HPDA):** Applying high performance computing techniques to the processing of extremely huge amounts of data. Taking advantage of a high-performance infrastructure that powers different workloads, and starting to support workflows that actually 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; and models and datasets are exceptionally complex.

- **Data analytics and Artificial Intelligence:** Machine-learning algorithms have progressed in recent years, especially through the development of deep learning and reinforcement-learning techniques based on neural networks. The challenge is to make use of the progress in efficient and reliable data analytics processes for advanced business applications. This includes the intelligent distribution of the processing steps, especially those close to data sources (e.g. distributed deep learning). In addition, different techniques from AI can be used in order to enable better reasoning about data analytics' processes and outcomes.

Outcomes

The main expected advanced analytics innovations are the following:

- **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 the 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 of 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 min/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'

Background

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 together 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[43]. 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.

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, etc.). Tools need to support user interaction for the exploration of unknown and unpredictable data within the

[43] J. Raskin, The humane interface: new directions for designing interactive systems, Addison-Wesley, Reading, MA, 2000.

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 will emerge 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 in order 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. How best to deal with uniform data visualisation on a range of devises and to ensure access to functionalities for data exploration, discovery and querying in multi-device settings are difficult problems that require new approaches and paradigms to be explored and developed.

Outcomes

The main expected advances in visualisation and user experience are the following:

- **Scalable data visualisation approaches and tools:** In order to handle extremely large volumes of data, 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, also which 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). Other examples are 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', time has an inherent semantic structure and a hierarchical system of granularities which must be addressed), and the visualisation of interrelated/linked data, exploiting graph visualisation techniques to allow easy exploration of network structures.

3.5 Priority 'Data Protection'

Background

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 two 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 important 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, etc.), 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 to 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 properly enforced 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 clearly 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.

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 in order 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 whole 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 an important 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 a computation due to their great 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, healthcare, etc.). 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, even public, datasets, there is a risk that personal 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.

Outcomes

The main expected advances in data protection are the following:

- **Complete data protection framework:** A mechanism for data protection within innovation spaces. This 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, 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.

3.6 Big Data Standardisation

This section was added to SRIA version 4.

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 assured 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 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. This is a prime source of compatibility and interoperability, and simplifies 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 Big Data ecosystem, standardisation applies to both the **technology** and to 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 large 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 certain 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 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 Big Data ecosystem overlaps with many other ecosystems, like Cloud computing, IoT, Smart Cities and Artificial Intelligence, the PPP will continue to be a forum for bringing industry stakeholders from across these other domains together 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.

3.7 Engineering and DevOps for Big Data

This section was added to SRIA version 4.

Background

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 tool chains 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.

Challenges

Engineering and DevOps tool chains 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 on-line analytical processing (OLAP) systems to fully fledged frameworks which integrate data management, data analytics and data protection by bringing these Big Data technologies into a unified systems perspective.

- **DevOps:** Integrated development and operations (DevOps) approaches need to be tailored to Big 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 ways for 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 (for instance by means of 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.

Outcomes

The main expected outcomes for engineering and DevOps are:

- Engineering principles, as well as fully integrated tool chains 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 Big Data systems and platforms to handle high-intensity and highly distributed data and event streams.

3.8 Illustrative Scenario in Healthcare

This section was added to SRIA version 4.

This section illustrates how the technical priorities arranged in the BDV Reference Model may help in delivering Big Data solutions for specific industry sectors. To this end, we present a scenario from the healthcare sector. A recent BDVA white paper has collected and analysed the needs, opportunities and challenges for Big Data technologies in healthcare[44].

There is a clear opportunity to transform healthcare through the application of Big Data. 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. While most of this data was previously stored in a 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 will optimise care processes, disease diagnosis, personalised care and the healthcare system in general. Furthermore, a true transformation of the healthcare sector can only be achieved if all stakeholders and verticals in the healthcare sector (the HealthTech industry, healthcare providers, Pharma, Insurance, etc.) share Big 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, data processing architecture needs to be able to deal with heterogeneous health data (medical records, medical images, lab results, etc.), 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:** Still the main challenges will arise in the field of data analytics. The core of healthcare transformation is expected to come from AI-based propositions that will enable personalised medicine, clinical decision support, workflow optimisation, clinical research and, finally, better diagnosis and treatment of patients.

- **Data visualisation and user interaction:** An area closely related to analytics and data interpretation is data visualisation and user interaction. Visualising models

[44] Big Data Technologies in Healthcare – Needs, Opportunities and Challenges, BDVA, TF7 Healthcare subgroup, Dec 2016; http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20Healthcare.pdf

obtained by machine learning, as well as effective and clear user interaction technologies, are of the 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 important 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 properly 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.

4. NON-TECHNICAL ASPECTS

The portfolio of activities of the Big Data Value SRIA needs to comprise support actions that address complementary, non-technical issues alongside the European Innovation Spaces, Lighthouse projects, and the research and innovation activities. In addition to the activities addressing the governance of the PPP[45], the non-technical activities will focus on:

- Skills development;
- Business models and ecosystems;
- Policy and regulation;
- Social perceptions and societal implications.

4.1 Skills Development

Related activities of the CSA BDVe were incorporated into this section.

In order to leverage the potential of Big Data Value, a key challenge for Europe is to ensure the availability of highly and appropriately skilled people with an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and with regard to solutions. In addition to meeting the technical, innovation, and business challenges as laid out in this document, Europe needs to systematically address the need for educating people so that they are equipped with the right skills and are able to leverage Big Data Value, thus enabling best practices to flourish. Education and training will play a pivotal role in creating and capitalising on EU-based Big Data Value technologies and solutions. Promoting the 'transparency and recognition of skills and qualifications' is particularly relevant to the task of recognising data science skills, and consequently the challenge will be to provide a framework in order to validate these skills.

Over the past few years several European initiatives have begun to fill the gap of data professional profiles and to identify the competences that will be required. In particular, the European Data Science Academy (EDSA – http://edsa-project.eu/) is a project led by The Open University in the UK that aims to **analyse the required sector-specific skill-sets for data analysts** across the main industrial sectors in Europe and

[45] Which are described in detail in the Big Data Value PPP proposal.

has developed a **data science curricula** to meet these needs. In the same period the EIT-Digital Master on Innovation launched a major on Data Science as a joint initiative of six European Universities (Universidad Politecnica de Madrid (UPM), Eindhoven University of Technology (TU/e), and Université Nice Sophia Antipolis (UNS), KTH Royal Institute of Technology (KTH), Technical University of Berlin (TUB). More recently EDISON (http://edison-project.eu/), an e-infrastructures project led by University of Amsterdam, has devoted the last two years to analysing the industrial and academic needs in terms of competences and skills in data literacy and the data profession. As result of this extensive work it developed a comprehensive Data Science Framework (EDSF), based on existing standards and preliminary accepted works; the framework includes a Competence Framework for Data Science, the Data Science Professional Profiles, the Data Science Model Curricula, and a Data Science Body of Knowledge. The project also supported the creation of an on-line Community Portal (https://www.datasciencepro.eu/) and a network of Champion Universities, whose representatives meet at six-monthly conferences to share experiences, lessons learned and issues common to academies promoting data literacy and supporting the creation of the Data Science profession.

Following on from these activities and their respective legacies, the Big Data Value ecosystem (BDVe) CSA will target activities to improve skills, education and Centres of Excellence around Big Data. It will facilitate coordination between Member States, help to align curricula with industry needs, and accelerate skills development to increase the number of European data scientists by 2020. The BDVe will address these challenges through building on the work already done and in cooperation with existing initiatives. Specific activities include:

- establishing a Network of National BDV Centres of Excellence to foster collaboration and share best practices between existing centres and support the setting up of new ones;
- exchanging knowledge on data science educational programmes across all Member States by delivering a Big Data Value Education Hub as a platform and living repository for knowledge;
- the certification of curricula and training programmes for BDV professionals to ensure their alignment with industry needs;
- stimulating and promoting the mobility of students, data professionals and domain experts, while creating mobility opportunities beyond the BDV PPP such as industrial internships.

Understanding the skills challenge calls for a clear definition of the appropriate profiles required to cover the complete data value chain. The first SRIA of the BDVA identified data scientists and data engineers. In the second version of SRIA requirements and needs related to different-sized companies and sectors were analysed. In that version the SRIA distinguished three different profiles, covering: (i) the hardware and software infrastructure area; (ii) the analytical field; and (iii) business expertise.

The educational support available for data science engineers is, however, far too limited to meet industry's requirements, mainly due the spectrum of skills and technologies involved. By transforming the current knowledge-driven approach into an experience-driven one, we can fulfil industry's needs for individuals capable of shaping the data-driven enterprise. The next generation of data professionals needs this wider view in order to deliver the data-driven organisations of the future.

The EDISON and EDSA projects have been rigorously working on defining data science profiles and related competences; in particular EDISON contributed to the identification of 23 data-related professions based on the Data Science Competence Framework, an extension of the European e-Competence Framework (eCF), an EU-promoted standard (in the next version of eCF (v.4.0) two new data-related profiles will be added based on the EDISON Data Science Framework).

Extensive experience and skills acquired by working on projects in the specified technical priority areas of the SRIA, together with the domain-specific knowledge obtained from the development of Lighthouse projects will guide the identification of skill development requirements that can be addressed by collaborating with higher education institutes and education providers to support the establishment of:

- new educational programmes based on interdisciplinary curricula with a clear focus on high-impact application domains;
- professional courses to educate and re-skill/up-skill the current workforce with the specialised skill-sets needed to be Data-intensive Engineers, Data Scientists and Data-intensive Business Experts. These course will stimulate life-long learning in the domain of data and the adoption of new data-related skills;
- foundational modules in data science, statistical techniques and data management within related disciplines such as law and the humanities;
- a network connecting scientists (academia) and industry that leverages Innovation Spaces to foster the exchange of ideas and challenges;
- datasets and infrastructure resources, provided by industry, that enhances the industrial relevance of courses.

The regularly updated strategic challenge areas will provide the orientation for the development of the required data skills to support building extensive know-how and skills (e.g. through European curricula and the sharing of best practices) for future systems within both the industrial field and the research community.

The New Skills agenda emphasises 'the strategic importance of skills for sustaining jobs, growth and competitiveness', and is centred around two key points:

- improving the quality and relevance of skills formation; and
- making skills and qualifications more visible and comparable.

Skills recognition in data science should:

- require renewal after a set period of time;
- provide a framework which can quickly adapt to changes in skill requirements;
- measure skills on a highly granular and individual basis.

The BDVe project is currently working on the different workflows for awarding badges for skills in Big Data and is working with BDVA and other stakeholders to define a business model for the effective implementation of the badge system for data science skills recognition in Europe.

4.2 Ecosystems and Business Models

The Big Data Value ecosystem will comprise many new stakeholders. New concepts for data collection, processing, storing, analysing, handling, visualisation and, most importantly, usage will emerge and business models will be created around them.

There are three key ways to generate value for companies along the value chain, regardless of sector or domain: optimising and improving the core business; selling data services; and, perhaps most importantly, creating entirely new business models and business development.

Identifying sustainable business models and ecosystems in and across sectors and platforms will be an important challenge. In particular, many SMEs that are now involved in highly specific or niche roles will need support to help them align and adapt to new value chain opportunities.

Dedicated projects and activities investigating and evaluating existing and emerging business propositions and models will in part be linked to the innovation spaces where suppliers and users will meet. Those projects will:

- establish a map of technology or platform providers and their value contribution;
- identify mechanisms by which the value of data can be adequately determined;
- provide a platform for entrepreneurs and financial actors to gain adequate levels of understanding about the value chain of Big Data;
- scope, describe and validate business propositions and models that might be successful and sustainable in the data economy of the future.

The outcomes of these projects will contribute to the creation of a more stable business environment that will enable companies, particularly web entrepreneurs and SMEs, to access Big Data markets and ecosystems. Europe needs to foster more and stronger players to make the whole Big Data Value ecosystem strong, vibrant and valuable, such that it will lift the entire Europe's economy. The following **key stakeholders** are seen as the main actors along the Big Data Value chain:

- **User enterprises**, e.g. enterprises in all sectors and of all size that want to improve their services and products using Big Data technology, data products and services;
- **Data generators and providers** who create, collect, aggregate, transform and model raw data from various public and non-public sources and offer them to customers.
- **Technology providers** who provide tools and platforms which offer data management and analytics tools to extract knowledge from data, curate and visualise it;
- **Service providers** who develop Big Data applications on top of the tools and platforms to provide services to user enterprises.

In addition, the following organisations and communities will have an impact on data-driven ecosystems building on the Big Data Value chain:

- Regulatory bodies that define privacy and legacy issues related to data usage;
- **International/national de jure and de facto standardisation bodies** that promote new concepts, systems and solutions for global adoption in international standards;
- **Collaborative networks** where different players in the value chain collaborate to offer value services to their customers based on data value creation.

The current stakeholders in H2020 who operate along the phases of research, innovation, exploitation and usage will also play an important role in leveraging the Big Data Value chain.

To engage the full value chain of stakeholders, the strategic intention of the BDVA is to move forward with analysing and defining new business propositions and models. It will do so by investigating the emergence and evolution of the Big Data ecosystem in two key ways: first by addressing the SME, start-ups and entrepreneurship aspect, and, second, by investigating how the value of Big Data can be leveraged effectively in (transforming) traditional business and industries. Taken together, these tracks ensure a comprehensive analysis of new business propositions.

4.3 Policy and Regulation

The PPP has no mandate or competence to be involved directly in policy making for legal or regulatory framework conditions. However, the PPP needs to contribute to the policy and regulatory debate about the non-technical aspects of future Big Data Value creation as part of the data-driven economy. Dedicated projects have to address the realities of data governance and usage, data protection and privacy, security, liability, cybercrime and Intellectual Property Rights (IPRs), among other issues.

These projects will initiate activities that are envisaged to promote exchange between stakeholders from industry, end-users, citizens and society to develop input into on-going policy debates where appropriate. Equally such exchanges will identify the concrete legal problems faced by actors in the value chain, particularly SMEs who have no legal resources. This will establish a body of knowledge on legal issues with a helpdesk for the project participants and ultimately for the wider community. The outlined projects will:

- establish an inventory of roadblocks inhibiting a flourishing data-driven economy (e.g. by materialising the value of Big Data collections);
- make and collect observations about the discovery of new legal and regulatory challenges along with the implementation of state-of-the-art technology and the introduction of new technology.

By doing so, these projects will contribute from the perspective of developing novel technology and solutions, by promoting direct contact with the actors to help legislators and regulators undertake exhaustive consideration of the framework conditions. Furthermore, these projects will support the BDV actors, particularly SMEs, to navigate the legal barriers to integrating into new ecosystems.

4.4 Social Perceptions and Societal Implications

Big Data will provide solutions for major societal challenges in Europe. For an accelerated adoption of Big Data it is critical to increase awareness of the benefits and the value that Big Data offers, and to understand the obstacles to building solutions and putting them into practice. End-users' lack of trust in Big Data technology is an important barrier that may hinder adoption, affecting aspects such as privacy, transparency (the ability to understand and interpret), perceived efficacy (the expected benefits), manageability (ease of use and level of control that the user can exert), and acceptability (related to ethical issues that arise when new technology creates new questions, for instance in the case of profiling customers for insurance companies; but also willingness to share data: in many cases end-users are expected to contribute to the service by providing data themselves). In addition, collaboration and co-innovation between organisations, the public sector and private individuals should be enhanced to support value creation from big data solutions.

Societal challenges cover a wide range of questions:

- How to establish and increase trust in Big Data innovations, addressing transparency, efficacy, manageability and acceptability?
- How to incorporate privacy-by-design principles and create a common understanding among the technical community leading to an operational and validated method that is applicable to data-driven innovations development?
- How to develop a better understanding of inclusion and the collective awareness aspects of Big Data innovations? How to enable a clear profile of the social benefits Big Data Value technology can provide?
- How to identify the ethical issues created by Big Data innovations, leading to a clear formulation of these issues and finding the path that leads to solutions?

By addressing the listed topics, citizens' views and perception will be taken into account so that technology and applications are not developed without the chance of their being widely accepted. The above actions will be based on and related to work that addresses the bridge between ICT and society, for instance, the BYTE and Big Data Europe projects, as well as the activities of other NGOs, such as the Digital Enlightenment forum and national organisations.

5. EXPECTED IMPACT

5.1 Expected Impact of Strategic Objectives

The expected impact of the PPP should be recognised in the great enhancement that Big Data analysis techniques will provide to all decision-making processes. From this point of view every sector, private or public, industrial or academic, will be affected, as will society. The PPP will show that Big Data Value is not just a new buzzword, but a shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.

The general impact of the PPP is anticipated along the following lines:

- **Effective service provision** from public and private organisations will be achieved by developing and making available to industry and the public sector technology, applications and solutions for the creation of value from Big Data leading to increased productivity, optimised production and more efficient logistics (both inbound and outbound).

- **Extensive experience and skills will be acquired** and a base of Intellectual Property Rights established to support the building of extensive know-how and skills in Europe (e.g. by European curricula and the sharing of best practices) for future systems in the industrial sector and research community.

- **New business models and the optimisation of existing industries** will drive the integration of the BDV services with private and public decision-making systems such as Enterprise Resource Planning and marketing systems.

A significant impact is expected on society, with opportunities for a wide range of applications:

- **Big Data Value technologies** will be a key contributing factor to solutions for major societal challenges in areas such as health, demographic change, climate change, transport, energy and cities. Innovative Big Data technologies will provide insight into the different aspects of societal challenges and unlock new potential solutions to address them. Similarly, BDV is associated other areas such as the Future Internet (FI) and the Internet of Things (IoT). In these emerging markets the integration of huge volumes of data needs to be supported by solid data-orientated technologies. All these solutions will lead to a transformation of our everyday lives, with direct impacts on individuals' behaviour and habits. In the future, citizens can expect benefits from a more personalised healthcare system, novel decision-support systems for their everyday life, or new ways to interact with companies and administrations, based on Big Data Value solutions.

- **The availability of public government information and open data** will influence educational and cultural services. Large databases containing information on cultural heritage such as digitised books and manuscripts, photographs and paintings, television and film, sculpture and crafts, diaries and maps, sheet music and recordings will be made available and allow for innovative ways of educating people and new forms of interaction between people across cultural borders.

- **Big Data technology will improve societal insight** into individual and collective behaviour. Such technologies may allow for greater fact-based decision making in the fields of politics and the economy. Fundamental research will be deeply impacted by the availability of Big Data resources and analysis, providing fresh insights and new developments in many areas such as biology, physics, mathematics, materials and energy. These developments themselves will produce new Big Data and further enhance societal developments.

- **Collaboration:** Big Data Value will help to improve collaboration by providing access to various data sources such as media content, traffic flows, etc. Better services and collaboration will be possible, for instance in emergency and crisis situations. Individuals will be empowered by their new roles as co-creators or co-innovators as well as the generators and providers of personal data.

Industry surveys show that the gains from Big Data Value are expected across all sectors, from manufacturing and production to services and retail. The following are examples of sectors that are especially promising with regard to Big Data Value.

- **Environment:** A better understanding and management of environmental and geospatial data is of crucial importance. Environmental data helps us to understand how our planet and its climate are changing and also addresses the role humans play in these changes. For example, the European Earth observation programme, Copernicus, aims to provide reliable and up-to-date information on how our planet's climate is changing in order to provide a foundation for the creation of sustainable environmental policies. In addition, the EU project Galileo will offer a global network of satellites providing precise timing and location information to users on the ground and in the air. The overall intention is to improve the accuracy and availability of location data for the benefit of sectors including transport and industry as well as Europe's new air-traffic control system.

- **Energy:** The digitisation of the energy system, from production and distribution to smart meters monitored by the consumer, enables the acquisition of real-time, high-resolution data. This coupled with the addition of other data sources, such as weather data, usage patterns and market data, and accompanied by advanced analytics, means that efficiency levels can be increased immensely. Existing grid capacities could be better utilised, and renewable energy resources could be better integrated.

- **Mobility, transport and logistics:** Urban multimodal transportation is one of the most complex and rewarding Big Data settings in the logistics sector. In addition to sensor data from infrastructure, vast amounts of mobility and social data are generated by smart phones, C2x technology (communication among and between vehicles), and end-users with location-based services and maps. Big Data will open up opportunities for innovative ways of monitoring, controlling and managing logistical business processes. Deliveries could be adapted based on predictive monitoring, using data from stores, semantic product memories, internet forums, and weather forecasts, leading to both economic and environmental savings.

- **Manufacturing and production:** As a result of industry's growing investments in smart factories with sensor-equipped machinery that is both intelligent and networked (Internet of Things, Cyber-Physical Systems), in 2020 the production sectors will be one of the major producers of (real-time) data. The application of Big Data into this sector will bring efficiency gains and predictive maintenance. Entirely new business models are expected as the mass production of individualised products becomes possible where consumers can have direct access to influence and control in product supply.

- **Public sector:** Big Data Value will contribute to increased efficiency in public administration processes. The continuous collection and exploitation of real-time data from people, devices and objects will be the basis for smart cities, where people, places and administrations are connected through innovative ICT services and networks. In the physical and cyber domains, security will be significantly enhanced by Big Data techniques; visual analytics approaches will be used to allow algorithms and humans to cooperate. From financial fraud to public security, Big Data will contribute to establishing a framework that enables a safe and secure digital economy.

- **Healthcare:** Applications range from comparative effectiveness research to the next generation of clinical decision support systems, which make use of comprehensive heterogeneous health datasets as well as advanced analytics of clinical operations. Of particular importance are aspects such as patient involvement, privacy and ethics.

- **Media and content:** By employing Big Data analysis and visualisation techniques, it will be possible to allow users to interact with the data, and have dynamic access to new data as they appear in the relevant repositories. Users will be able to register and provide their data or annotations to existing data. The environment will move from a few state-orientated broadcasters to a prosumer approach, where data and content are linked together, blurring the lines between data sources and modes of viewing. Content and information will find organisations and consumers, rather than vice versa, with a seamless content experience.

- **Financial services:** Huge amounts of data are processed to detect issues such as fraud and risk and to analyse customer behaviour, segmentation, trading, etc. Big Data analysis and visualisation will open up new use cases and permit new techniques to be put into practice. Possibilities include managing regulation, reporting, audits and compliance, and the automatic detection of behaviour patterns and cyber-attacks. Open sources of information can be combined with proprietary knowledge to analyse competitive positions, and recommendation engines will be able to identify potential customers for products.

- **Telecommunications services**: Big Data enables improved competitiveness by transforming data into customer knowledge. Possible use cases could include the improvement of service levels, churn reduction, services based on combining location with data about personal context, and better analysis of product and service demand.

- **Retail:** Digital services for customers provided by smart systems will be essential for the success of future retail businesses. The retail domain will be especially focused on highly efficient and personalised customer assistance services. Retailers are currently confronted with the challenge to meet the demand of a new generation

of customers who expect information to be available anytime and anywhere. New intelligent services that make use of Big Data will allow a new level of personalised and high-quality Efficient Consumer Response (ECR).

- **Tourism**: Personalised services for tourists are essential for creating real experiences within a powerful European market. The analysis of real-time and context-aware data with the help of historic data will provide customised information to each tourist and contribute to a better and more efficient management of the whole tourism value chain. The application of Big Data in this sector will enable the development of new business models, services and tourism experiences.

5.2 Monitoring of Objectives

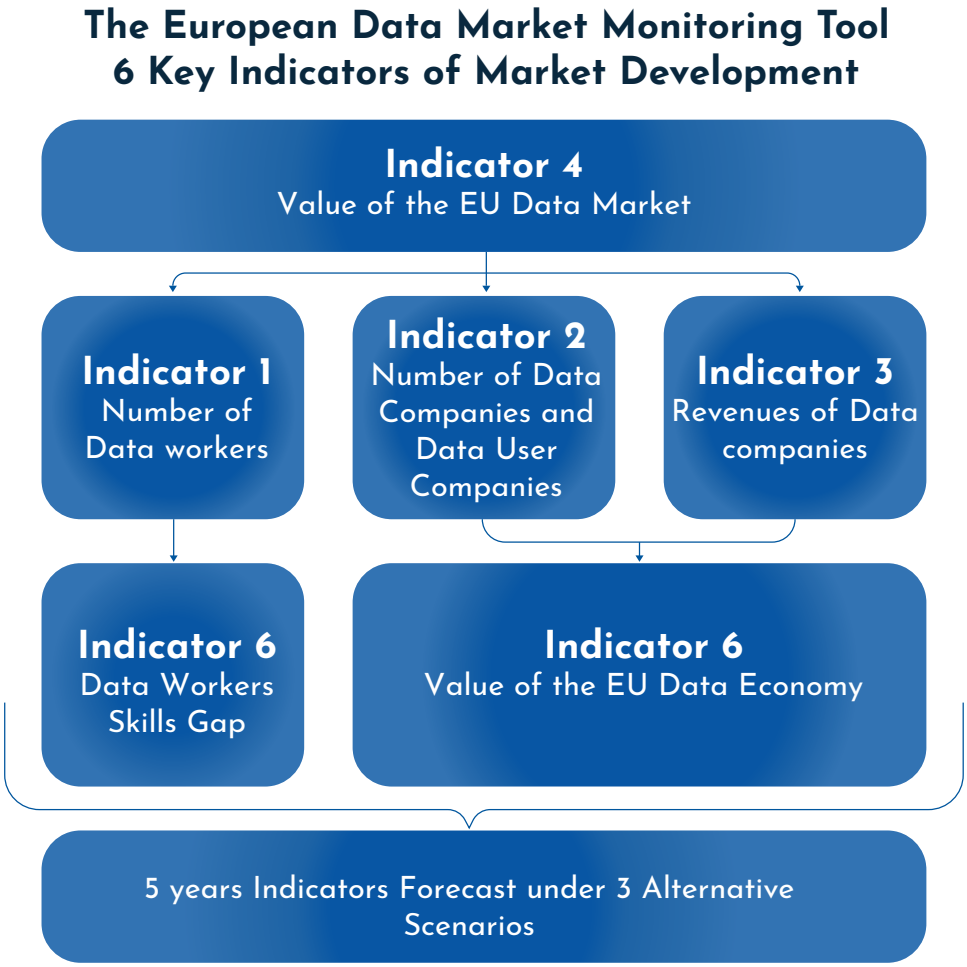
Big Data Value generation and the technology behind it will have a tremendous impact on industry and the economy as a whole. In terms of assessing this impact there are two basic types of measurements, with related indicators:

- **Indirect monitoring:** The monitoring is done using indicators that cannot be directly influenced or monitored by activities resulting from this Big Data Value SRIA. Typically, the monitoring is based on tracking the progress of particular developments and uses comparison rather than specific numbers or targets. For example, the proposed Big Data Value activities can provide a research and innovation ecosystem, but ultimately jobs, sales information and business progress will be under the control of individual organisations. Indirect indicators include, for instance, economic and usage information. The success of the SRIA strategy will be mainly assessed based on indirect indicators.

- **Direct KPIs**: The second type of measurement is by means of Key Performance Indicators (KPIs) that are directly related to the performance of the SRIA activities themselves and are clearly measurable; for example, providing solutions to the technical priorities or the stimulation of SME participation in research and innovation activities and i-Spaces.

According to the strategic and specific objectives of the Big Data Value SRIA described in Section 1.6 an interdisciplinary and holistic approach will be followed. Consequently, the indicators to be used for assessing the impact of the SRIA have to address strategic, social, competitiveness and innovation aspects.

Figure 7: The European Data Market Indicators (IDC, 2016)



Indicators to measure the achievement of the strategic objectives

The development of the BDV market will be driven by new and innovative products. However, success in this area depends heavily on various market conditions and the overall economic climate. IDC[46] has developed a European Data Market Monitoring tool with European Data Market Indicators (see Figure 7) that are in close alignment with certain strategic objectives and KPIs specified in the Contractual Arrangement (CA) of the PPP. Therefore, the SRIA proposes to use these kinds of market metrics as indirect indicators for monitoring its strategic impact. The KPIs and their monitoring will need to be adapted in alignment with the on-going development of the European Data Market Monitoring Tool, based on the assumption that, using this tool, the development of the market can be observed throughout the lifetime of the PPP. The following table lists the relevant KPIs of the CA (the numberings are derived from the Arrangement and the IDC report, and are therefore non-consecutive, as the table focuses on the KPIs of the Contractual Arrangement that are relevant in this area). Furthermore, it shows the measurements available from 2013 and 2016.

[46] IDC et al., European Data Market, SMART 2013/0063, D9 – Final Report, 1 February 2017, http://datalandscape.eu/study-reports

Strategic Indicators

CA-KPI[47]	IDC-KPI	Description and initial measurements	Societal	Competitiveness	Innovation	Operational
KPI.CA.1 (II-1)		Market share of the European suppliers of the global Big Data market in 2020 The IDC report does not provide any figures to measure this KPI; however, it presents a comparison on the assessed indicators for the EU, the US, Japan and Brazil. For details, see the IDC report on the European Data Market[48].		■		
KPI.CA.5 (II-2)	2.1	Number of European companies offering data technology, applications and services, including start-ups. According to IDC the total number of data companies in the EU, measured as legal entities based in one EU country, increased from 239 845 in 2013 to 243 600 in 2014, and to 249 100 in 2015 and 254 850 in 2016, representing a total increase of 6.3% since 2013.		■	■	
KPI.CA.6 (II-3)	3.1	Number of European companies offering data technology, applications and services, including start-ups. According to IDC the total number of data companies in the EU, measured as legal entities based in one EU country, increased from 239 845 in 2013 to 243 600 in 2014, and to 249 100 in 2015 and 254 850 in 2016, representing a total increase of 6.3% since 2013.		■		
KPI.CA.8 (II-5)	1.1	Increased number of European data workers. IDC estimates an overall number of 5.7 million data workers in 2013 and 6.1 million data workers in 2016, representing an increase of almost 6.7% since 2013.	■	■	■	

Direct KPIs to measure the achievement of the specific objectives

The SRIA activities will deliver solutions, architectures, technologies and standards for the data value chain over the next decade. The following KPIs are proposed to frame and assess the impact of those SRIA activities.

[47] Numbers in parentheses refer to the numbering scheme used in the Monitoring Report of the PPP.
[48] IDC et al., European Data Market, SMART 2013/0063, D9 – Final Report, 1 February 2017, http://datalandscape.eu/study-reports, pp.171.

Direct KPIs

			Societal	Competitiveness	Innovation	Operational
Business	KPI.D.1 (II-11)	**At least 50 large-scale experiments are conducted in i-Spaces involving closed data.** Multiple SMEs should be encouraged to perform experiments by using i-Spaces. This will foster their growth from small companies into larger ones and/or their expansion from national markets into the EU (or even global) market. The i-Space and the hosted experiments will provide a unique opportunity for exploitation.		■	■	
	KPI.D.2 (II-12)	**30% year-on-year increase in Big Data Value use cases supported in i-Spaces.** The number of use cases within the large-scale experiments will be an indicator of acceptance and will also prove the innovative capacity of the BDV partnership. An ever-expanding increase will guarantee a continuous value creation out of Big Data and will speed up the innovation process, thus also addressing the issue of time to market. It will also support market development in existing industries and potentially the establishment of entirely new business models.		■	■	
Skills	KPI.D.3 (II-8)	**At least 50 training programmes are established with the participation of at least 100 participants per training session arising from the PPP.** Continuous development of skills and competences on the basis of the Big Data Value PPP will be supported by training and education activities. An appropriate environment (e.g. e-learning platform, contribution to university curricula) should be created to attract potential participants. This expands the number of skilled people and serves as a unique opportunity to create new jobs and start-ups as a result of the PPP activities.	■	■		
	KPI.D.4 (II-8)	**At least 10 European training programmes involving three different disciplines with the participation of at least 100 participants.** These interdisciplinary programmes will contribute to developing the knowledge and skills needed to deal with the complexity of Big Data. To expand the number of students, Massive Open Online Courses (MOOCs) would be proposed, building on the diversity of skills available and European multiculturalism	■	■		

Applications	KPI.D.5 (II-13)	**At least 10 major sectors and major domains are supported by Big Data technologies and applications developed in the PPP.** The usage of BDV technologies and applications developed in the PPP in different sectors will lead to increased value generation and finally to job growth in all the addressed sectors. The broad take-up of those technologies and applications across a number of sectors is also an indicator of the efficient sharing of best practices and expertise leading to a build-up of a broad skills base. Furthermore, cross-sector activities should prove domain independent, while cross-domain deployment will lead to the setting of standards.				
Data	KPI.D.6 (II-14)	**Total amount of data made available to i-Spaces – including closed data – is in the 10x Exabyte range.** Experiments conducted in i-Spaces benefit from their scale, the amount of different but integrated data sources, and, especially, the value of the data. This is key to accelerating Data Driven Innovation in Europe in liaison with Research and Education, with major advances expected in data management techniques, semantics, analytics, data learning and visualisation. This includes both economic and societal objectives. These experiments will also contribute to the advances in data governance practices, such as encryption, pseudonymisation and anonymisation to ensure better Intellectual Property and privacy protection.				
	KPI.D.7 (II-15)	**Availability of metrics for measuring the quality, diversity and value of data assets.** It is not only the amount of data made available to perform data analysis; of utmost importance are the quality, diversity and value of the data. The ultimate goal is to create value out of Big Data, to derive analytical findings from a minimal, yet most significant dataset, allowing faster data processing and the management of data for data analytics. During the PPP relevant metrics will be derived.				

Technical	KPI.D.8 (II-16)	**The speed of data throughput is increased by 100 times compared to 2014.** One of the main problems regarding today's data storage and processing techniques is the time required for accessing large datasets in order to analyse them. Techniques to be implemented in the scope of the Data Management priority will make data access for analysis much more efficient.				
	KPI.D.9 (II-6)	**The energy required to process the same amount of data is reduced by 10% per year.** One of the main problems today is the amount of energy consumed processing data due to the huge amount of data coupled with a lack of algorithms. New hardware for devices will reduce the energy required to process data. Beyond hardware optimisation, new tools and algorithms will require fewer resources and time to provide the same quality of analytics.				
	KPI.D.10 (II-4)	**Enabling advanced privacy and security respecting mechanisms (including anonymisation) for data access, process and analysis.** The availability of suitable privacy and security respecting mechanisms will encourage data users to provide closed data for experiments and analyses in i-Spaces and Lighthouse projects.				

Additional KPIs listed in the Contractual Arrangement

Besides the KPIs listed above, Article 7 of the Contractual Arrangement of the PPP specifies further KPIs. For the sake of completeness the following table contains the full list of KPIs detailed in Article 7, and where applicable, we refer to the two KPI tables above.

KPIs in Article 7 of the Contractual Arrangement		Societal	Competitiveness	Innovation	Operational
CA-KPI	**Description**				
KPI.CA.1	See corresponding strategic KPI		■		
KPI.CA.2 (I-4)	**PPP investments leveraged through sector investments by 4 times the PPP's total budget**		■		
KPI.CA.3 (I-5)	**SMEs participating in the PPP projects under this initiative represent at least 20% of participant organisations**		■		
KPI.CA.4 (I-3)	**Increased competitive European provision of big data value creation systems and technologies**		■		
KPI.CA.5	See corresponding strategic KPI		■	■	
KPI.CA.6	See corresponding strategic KPI		■		
KPI.CA.7	see KPI.D.10		■		
KPI.CA.8	See corresponding strategic KPI	■	■	■	
KPI.CA.9	see KPI.D.9	■	■	■	
KPI.CA.10 (II-7)	**New economically viable services of high societal value developed by PPP projects**	■	■	■	
KPI.CA.11	See KPI.D.3 and KPI.D.4	■	■		
KPI.CA.12 (II-9)	**Ensure efficiency, transparency and openness of the PPP's consultation process**				■
KPI.CA.13 (II-10)	**Ensure that the technology is in line with the established multi-annual roadmap**				■

BDV BIG DATA VALUE ASSOCIATION

6 ANNEXES

6.1 Acronyms and Terminology

Acronym/Term	Name/Description
General	
AI	Artificial Intelligence
AIOTI	Alliance for Internet of Things Innovation
API	Application Programming Interface
BDV	Big Data Value
BDVA	Big Data Value Association
BDVe	Big Data Value ecosystem
BoD	Board of Directors
BPM	Business Process Management
CA	Contractual Agreement
CAGR	Compound annual growth rate
CASD	Secure Remote Data Access Centre
CSA	Coordination and Support Action
CEP	Complex Event Processing
CPS	Cyber Physical Systems

DaaS	Data-as-a-Service
DEI	Digitising European Industry
DIH	Digital Innovation Hub
DSM	Digital Single Market
DSMS	Data Stream Management System
ECR	Efficient Consumer Response
ECSO	European Cyber Security Organisation
EDSA	European Data Science Academy
EFFRA	European Factories of the Future Research Association
EHR	Electronic Health Record
EIP	European Innovation Partnership
EOSC	European Open Science Cloud
EU	European Union
ETP4HPC	European Technology Platform for High Performance Computing
FI	Future Internet
FIRE	Future Internet Research & Experimentation
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
HPC	High Performance Computing
HPDA	High Performance Data Analytics
IA	Innovation Action
ICT	Information and Communication Technologies
IDP	Industrial Data Platform
IoT	Internet of Things
IPRs	Intellectual Property Rights
i-Space	(European) Innovation Space
KPI	Key Performance Indicators
MOOC	Massive Open Online Course
MOU	Memorandum of Understanding

MPP	Massively Parallel architectures
NLP	Natural Language Processing
NoSQL	Not only SQL (Structured Query Language) (referring to databases)
OECD	Organisation for Economic Co-operation and Development
OLAP	On-line analytical processing
PDP	Personal Data Platform
PII	Personally identifiable information
PIMS	Personal Information Management Systems
PPP	Public Private Partnership
QoS	Quality of Service
RDP	Research Data Platforms
RIA	Research and Innovation Action
SDO	Standards Development Organisation
SMEs	Small and Medium-sized Enterprises
SMI	Smart Manufacturing Industry
SRIA	Strategic Research & Innovation Agenda
SWOT	Strengths, Weaknesses, Opportunities and Threats
TRL	Technology readiness level
UDP	Urban/City Data Platform
VC	Venture Capital
WP	Work Programme
5G	5th generation mobile networks
Data Orientated	
Open Data	Data available to everyone to use and republish
Private Data	Data which are generated by organisations, typically by companies and in particular related to users, which have not been made 'open' and often are kept internally or under restricted conditions (e.g. Non-Disclosure Agreements (NDAs))
Closed Data	Data that have restrictions on their access or reuse (i.e. charges, technology, memberships, etc.). Typically Closed Data include Private Data
Free Data	Data that can be accessed or reused without a charge
Non-Free Data	Data which have a charge associated with use or reuse

6.2 Contributors

SRIA Version 4 in October 2017

The following individuals and organisations are thanked for their involvement in creating this updated version of the SRIA document or other documents to which it closely relates.

SRIA Core Team		
Sonja Zillner	Chair of Editorial Committee	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Andreas Metzger	Co-Editor	Paluno, Univ. Duisburg-Essen
Robert Seidl	Co-Editor	Nokia
Contributors		
Ana Garcia Robles	Contributor	BDVA Secretary General
Jim Keneally	Contributor	Intel
Maria Perez	Contributor	UPM
Souleiman Hasan	Contributor	Insight @ NUI Galway
Paul Czech	Contributor	Know-Center
Davide Dalle Carbonare	Contributor	Engineering
Simon Scerri	Contributor	Fraunhofer
Thomas Hahn	Contributor	Siemens AG
Caj Södergård	Contributor	VTT
Roberta Piscitelli	Contributor	EGI Foundation
Yannick Legré	Contributor	EGI Foundation
Geraud Canet	Contributor	CEA List/DIASI
Claire Tonna	Contributor	Catapult
Ray Walshe	Contributor	Insight
Arne Berre	Contributor	SINTEF
Marija Despenic	Contributor	Philips
Milan Petkovic	Contributor	Philips
Ernestina Menasalvas	Contributor	UPM
Other participants involved in this update of the SRIA		
Numerous BDVA members participating in the various BDVA task forces working on state-of-the art research related to the SRIA priorities as well as the development of the European Data Value Ecosystem.		

SRIA Version 3 in January 2017

The following individuals and organisations are thanked for their involvement in creating this updated version of the SRIA document or other documents to which it closely relates.

SRIA Core Team		
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Andreas Metzger	Co-Editor	Paluno, Univ. Duisburg-Essen
Robert Seidl	Co-Editor	Nokia
Contributors		
Dirk Mayer	Contributor	Software AG
Nuria de Lama	Contributor	ATOS
Jim Keneally	Contributor	Intel
Souleiman Hasan	Contributor	Insight @ NUI Galway
Dumitru Roman	Contributor	SINTEF
Carlos A. Iglesias	Contributor	UPM
Meilof Veeningen	Contributor	Philips
Bjarne Kjær Ersbøll	Contributor	TNO
Ernestina Menasalves	Contributor	UPM
Other participants involved in this update of the SRIA		
Numerous BDVA members participating in the various BDVA task forces working on state-of-the art research related to the SRIA priorities.		
All BDVA members participating in the BDVA Member Expression of Interest as well as the participants of the BDVA Community Survey that was open from June to October 2016.		
More than 200 participants of the BDVA Mini Summit in March 2016 in Den Hague and more than 350 participants of the BDVA summit in November 2016 in Valenica who have actively contributed to a large number of workshops dedicated to the various technical and non-technical topics. Workshop outputs that were appropriate were used as input for the SRIA update.		

SRIA Version 2 in 2016

The following individuals and organisations are thanked for their involvement in creating this updated version of the SRIA document or other documents to which it closely relates.

SRIA Core Team		
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Arne Berre	Co-Editor	SINTEF
Andreas Metzger	Co-Editor	Paluno, Univ. Duisburg-Essen
Colin Upstill	Co-Editor	IT Innovation
Contributors		
Wolfgang Gerteis	Contributor	SAP
Ernestina Menasalves	Contributor	UPM
Nuria de Lama	Contributor	ATOS
Pierre Pleven	Contributor	INSTITUT MINES TELECOM Paris
Corinna Schulze	Contributor	SAP
Freek Bomhof	Contributor	TNO
Carlos A. Iglesias	Contributor	UPM
Souleiman Hasan	Contributor	Insight @ NUI Galway
Dumitru Roman	Contributor	SINTEF
Aris Gkoulalas-Divanis	Contributor	IBM Research
Bjarne Kjær Ersbøll	Contributor	DTU
Other participants involved in this update of the SRIA		
Numerous BDVA members participating in the various BDVA task forces working on state-of-the art research related to the SRIA priorities.		
All BDVA members participating in the BDVA Member Expression of Interest as well as the participants of the BDVA Community Survey that was open from June to October 2015.		
More than 300 participants of the BDVA Summit in June 2015 in Madrid who have actively contributed to over 60 workshops. Workshop outputs that were appropriate were used as input for the SRIA update.		
Other European Technology Platforms, for example ETP4HPC, that contributed through joint workshops and discussion, in particular focusing on the alignment of related technical priorities and requirements.		

BDV BIG DATA VALUE ASSOCIATION

SRIA Version 1 in 2015

The following individuals and organisations are thanked for their direct involvement in creating the first version of the SRIA document or other documents to which it closely relates.

SRIA Core Team		
Nuria de Lama	Co-Editor	ATOS
Julie Marguerite	Co-Editor	Thales
Klaus-Dieter Platte	Co-Editor	SAP
Josef Urban	Co-Editor	Nokia
Sonja Zillner	Co-Editor	Siemens AG
Edward Curry	Co-Editor	Insight @ NUI Galway
Primary Editing Team		
Antonio Alfaro	Contributor	Answare
Ernestina Menasalves	Contributor	UPM
Andreas Metzger	Contributor	Paluno, Univ. Duisburg-Essen
Robert Seidl	Contributor	Nokia
Colin Upstill	Contributor	IT Innovation
Walter Waterfeld	Contributor	Software AG
Stefan Wrobel	Contributor	Fraunhofer IAIS
Contributors		
Paolo Bellavista	Contributor	CINI
Stuart Campbell	Contributor	TIE Kinetix/BDVA SG
Thomas Delavallade, Yves Mabiala	Contributor	Thales
Nuria Gomez, Paolo Gonzales, Jesus Angel	Contributor	INDRA
Thierry Nagellen	Contributor	Orange
Dalit Naor, Elisa Molino	Contributor	IBM Research
Stefano de Panfilis, Stefano Scamuzzo	Contributor	Engineering
Nikos Sarris	Contributor	ATC
Bjørn Skjellaug, Arne Berre, Titi Roman	Contributor	SINTEF
Tonny Velin	Contributor	Answare
Alexandra Rosén, Francois Troussier	Contributor	NESSI Office

6.3 SRIA Preparation Process and Update Process

SRIA Preparation Process

Within the SRIA preparation process, the proposers were closely engaged with the wider community. Multiple workshops and consultations took place to ensure the widest representation of views and positions, including those from the full range of public and private sector entities. These activities have been carried out in order to identify the main priorities in this area, with approximately 200 organisations and other relevant stakeholders physically participating and contributing. Extensive analysis reports were then produced which helped both formulate and construct this SRIA.

The series of workshops gathered views from different stakeholders in the existing value chains of different industrial sectors, including: energy, manufacturing, environment, health, public sector, content and media. Additional workshops were organised to gather feedback on cross-sectorial aspects, for example, the view of SMEs. The selection of sectors was based on the criteria of their weight in the EU economy and the potential impact of their data assets (source: demosEUROPA). The community involved in the workshops included actors such as: AGT International (DE), Hospital de la Hierro (ES), Press Association (UK), Reed Elsevier (NL); BIODONOSTIA (ES), Merck 8 (ES), Kongsberg Group (NO), and many more.

In addition, NESSI, together with partners from the FP7 project BIG[49], ran an on-line public consultation on the BDV Strategic Research and Innovation Agenda between 9 April and 15 May 2014. The aim was to validate the main ideas put forward in the SRIA on how to advance Big Data Value in Europe in the next 5–10 years. A total of 195 organisations from all over Europe participated in the consultation, including companies such as Hitachi Data Systems, OKFN Belgium, TNO Innovation for Life, Euroalert, Tecnalia Research and Innovation, ESTeam AB and CGI Nederland B.V. Furthermore, around another twenty organisations and companies such as Wolters Kluwer Germany, Reed Elsevier and LT-Innovate shared in more detail their views on the content of the SRIA.

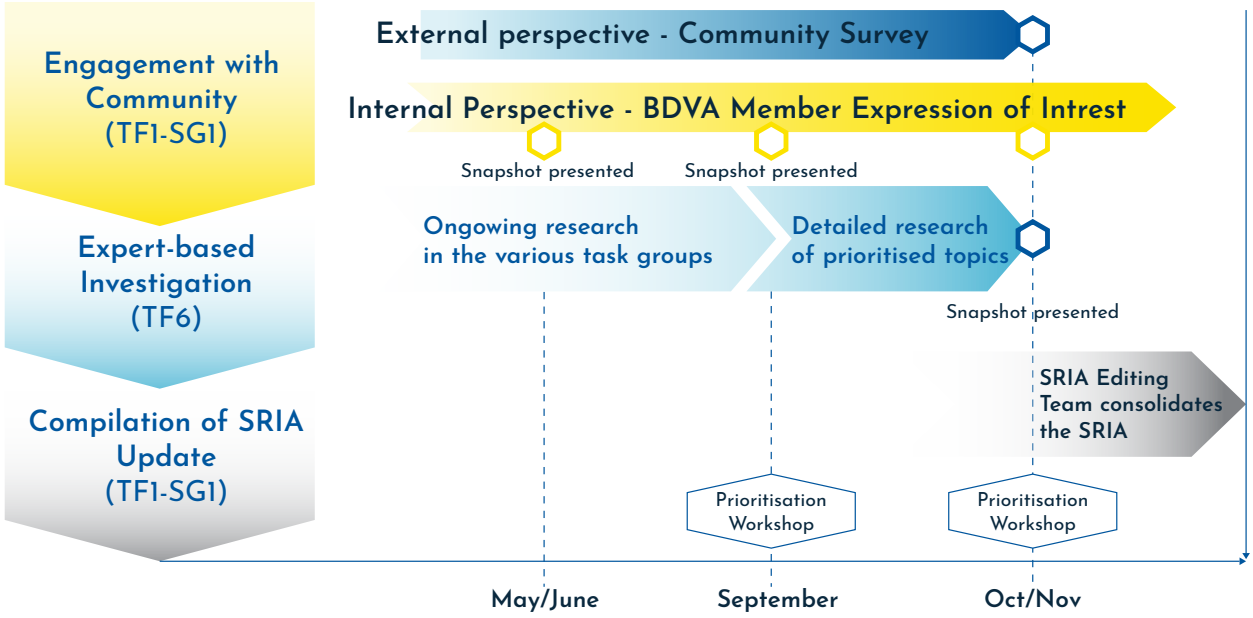
Although the primary target of the SRIA is to create impact at a European-level, cooperation with stakeholders outside Europe will allow the transfer of knowledge and experiences around the globe. For future collaborations, NESSI has already established links to the following regions through NESSI partners: the Mediterranean countries[50], the LatAm countries[51], the South East Asian countries[52] and the Russian-speaking countries.

[49] M. Cavanillas, E. Curry and W. Wahlster, New horizons for a data-driven economy: a roadmap for big data in Europe, Berlin, Springer International Publishing, 2016.
[50] MOSAIC (http://www.connect2sea.eu/) and MED-Dialogue (http://www.med-dialogue.eu/).
[51] CONECTA 2020 (https://www.conecta2020.net/).
[52] CONNECT2SEA (http://www.connect2sea.eu/)

SRIA Update Process

The Big Data Value Association (BDVA) is responsible for providing regular (yearly) updates of the SRIA, defining and monitoring the priorities as well as the metrics of the PPP.

Figure 8: Overview of the annual SRIA update process



The purpose of this process is to understand whether the SRIA (e.g. the technical the non-technical priorities) needs to be updated in terms of the following questions:

- How important are the priorities already covered in the SRIA?
- Are there other priorities relevant for BDV SRIA? How important are those priorities?

Engagement with the community: Within the updating process, the BDVA engages with (a) the BDVA members as well as (b) the wider community to ensure a comprehensive perspective concerning the technical and business impacts of the SRIA technical and non-technical priorities, as well as to identify emerging priorities with high impact.
For engaging with the internal and wider community, two parallel interaction streams have been set up:

- **BDVA Member Expression of Interest**: Each BDVA member is requested to express interest concerning research and innovation activities once a year (only one consolidated vote per member organisation will be reflected). This interaction stream does not follow a fixed deadline. The process is motivated by the fact that the BDVA is a fast-growing community and new members' opinions should be included

[53] In order to establish a contractual counterpart to the European Commission for the implementation of the PPP, the Big Data Value Association, a fully self-financed not-for-profit organisation under Belgian law, was founded by 24 organisations including large businesses, SMEs and research organisations.

as soon as possible. In order to learn about the BDVA members' expressed interest, snapshots of the survey results are taken on a regular basis.

- **Community Survey:** The community survey is open to everybody for a fixed period of time. The consolidated votes of the wider community will establish an important outside perspective.

The results of both community engagement streams provide important insights into the relevance of covered SRIA priorities and highlight emerging topics that require detailed analysis.

Expert-based Investigations: The BDVA has established task groups for all technical and non-technical priorities. The task groups are working continuously in order to produce related state-of-the-art analysis or working papers.

In accordance with the outcome of the community engagement process, particular task groups are consolidated in order to discuss the scope and level of detail the SRIA required for the various updates. A dedicated prioritisation workshop was also organised. As an outcome of the workshop, the task groups conducted detailed state-of-the-art research on the agreed priorities, which was presented in a dedicated SRIA update consolidation workshop.

6.4 History of Document Changes

HISTORY OF CHANGES

Version	Publication	Changes
1.0	January 2015	• Initial version
2.0	January 2016	The main changes compared to version 1 of BDV SRIA document are as follows: • Section 1: structure expanded by integrating more subheadings • Section 1.1: the most relevant Big Data market numbers were consolidated and updated • Section 1.3: adjustment of argumentation to reflect the situation that the PPP has already been running for one year • Section 1.4: a more condensed version of the original section • Section 1.5: a new section covering the objectives of the Contractual Arrangement was added • Section 1.6: a new section documenting the SRIA document history was added

		• Section 2.1: the text relating to i-Spaces and Lighthouse projects was updated by incorporating respective material from BDVA task forces • Section 2.2: a new section describing the BDV methodology was added • Section 2.3: text from the original cPPP document was reused to improve the stakeholder platform description; a sub-section describing the ongoing cooperation with ETP4HPC was added • Sections 3.2–3.6: several updates motivated by the SRIA survey results and proposed by the BDVA task forces were incorporated; overlaps across technical priorities were removed; and titles of sections were aligned to achieve consistency • Section 3.7: an innovation roadmap derived from SRIA survey results was added • Sections 4.1, 4.3 and 4.4 were improved by including information about already existing approaches, more precise definitions and relevant background information • The indicators represented in Section 5.2 were updated to achieve a strong alignment with the specific KPIs in the Contractual Arrangement; KPI.D.6 was adjusted to '10x Exabyte range' in order to establish a sound basis for evaluation • To ensure completeness, the KPIs in Article 7 of the Contractual Arrangement were listed • The Annex was extended by several sections • Annex 6.3 provides a detailed description of the BDV SRIA Update process • Annex 6.4 encompasses the history of document changes • Other minor drafting changes and corrections of clerical mistakes have been carried out across the document
3.0	January 2016	The main changes compared to version 1 of BDV SRIA document are as follows: • Section 1: recent developments in the European data market have been reflected throughout the whole section • Section 1.5: the objectives were described in a more specific way • Section 2.1: the definition of i-Spaces was updated in accordance with the discussion in the i-Spaces task force • Section 2.2: information about the nature of future Lighthouse projects was added • Section 2.3: a listing of projects that were funded in the 2016 calls was added • Section 2.4: encompasses an update concerning the collaboration between ETP4HPC and BDVA

		• Sections 3.2–3.6: the outcome descriptions were consolidated in order to prioritise the technical aspects • Section 3.7: the innovation roadmap was updated based on a data analysis in addition to the survey results of 2015 and 2016. • Section 4.1: activities related to skills development covered in the CSA BDVe were incorporated • Section 4.2: minor updates fostering alignment with ongoing activities in the Business task force were made • Section 5: minor updates were made to incorporate recent numbers from the 2016 IDC report • Other minor drafting changes and corrections of clerical mistakes were made across the document

BDV BIG DATA VALUE ASSOCIATION

www.bdva.eu