DataBench

Evidence Based Big Data Benchmarking to Improve Business Performance

D6.1 Dissemination and Liaison Plan

Deliverable D6.1	Dissemination and Liaison Plan
Work package	WP6
Task	6.1
Due date	15/09/2018
Submission date	15 /09/2018
Deliverable lead	ATOS
Version	2.0
Authors	ATOS (Nuria de Lama, Ana Morales) IDC (Cristina Pepato, Stefania Aguzzi)
Reviewers	Frankfurt Big Data Lab (Todor Ivanov) IDC (Richard Stevens)

Keywords

Benchmarking, big data, big data technologies, business performance, economic indicator, European significance, industrial relevance, performance metrics, tool, use cases, dissemination, marketing, Big Data Value PPP, image.

Disclaimer

Copyright Notice

Abstract

DataBench has the potential to become a key element of the Big Data Value Public-Private Partnership, since it will contribute to measure the real impact of the investments made by industry and the European Commission in this partnership. It will support the benchmarking of Big Data Technologies, therefore helping to understand the progress with respect to the state of the art but will also relate technical performance indicators with business indicators, establishing a bridge that has never been built before.

The challenges are however important. Among them, DataBench will require tight collaboration with a number of players and communities. Working hand-in-hand with benchmarking communities as well as with pilots and use cases (ideally coming from PPP projects) is key. They will be contributors but also validators of the so called DataBench ToolBox, main product to be produced by the project.

Along the three years duration of DataBench we will have to gain the credibility and recognition of those communities, since -in most cases- they will be the first adopters/users of the DataBench outcomes.

WP6 will support DataBench in creating the connections and engaging with those communities, will define and will implement a dissemination and communication strategy and will pave the path towards the self-sustainability of the results beyond the project duration. This deliverable describes the initial steps in that endeavour by sharing the Dissemination and Liaison Plan. The document identifies specifically: (1) target communities/audience, (2) Phases of the dissemination strategy, (3) tools, channels and mechanisms that will be used and (4) Key Performance Indicators.

Table of Contents

Table of Figures

Table of Tables

Acronyms and Abbreviations

Acronym	Title
AI	Artificial Intelligence
BDBC	Big Data Benchmarking Community
BDT	Big Data Technologies
BDV	Big Data Value
BDVA	Big Data Value Association
BSS	Business Support System
CMS	Content Management System
CoE	Center of Excellence
CRM	Customer Relationship Management
DG	Directorate General
DIAS	Data Information Access Services
DIH	Digital Innovation Hub
EBDVF	European Big Data Value Forum
EC	European Commission
ECSO	European Cyber Security Organisation
EFFRA	European Factories of the Future Research Association
ENoLL	European Network of Living Labs
EU	European Union
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
HIPEAC	High Performance and Embedded Architecture and Compilation
HPC	High Performance Computing
HTML	HyperText Markup Language
ICT	Information and Communication Technologies
IoT	Internet of Things
JRC	Joint Research Center
KPI	Key Performance Indicator
LDBC	Linked Data Benchmarking Council
NFV	Network Function Virtualization
OSS	Operational Support System
PPP	Public Private Partnership
RoI	Return on Investments
SRIA	Strategic Research and Innovation Agenda
SC	Steering Committee
SDIL	Smart Data Innovation Lab
SDN	Software Defined Networks
SME	Small and Medium Enterprise
TC	Technical Committee
TF	Task Force
TFP	Total Factor of Productivity
TRL	Technology Readiness Level
UK	United Kingdom
USA	United States of America
WP	Work Package

Executive Summary

WP6 will support DataBench in creating connections and engaging with the relevant communities and stakeholders, will define and will implement a dissemination and communication strategy and will pave the path towards the self-sustainability of the results beyond the project duration. This deliverable describes the initial steps in that endeavour by sharing the Dissemination and Liaison Plan. The document identifies specifically: (1) target communities/audience, (2) Phases of the dissemination strategy, (3) tools, channels and mechanisms that will be used and (4) KPIs.

Among the target communities it is worth mentioning the PPP project portfolio, benchmarking communities, I-Spaces (data-driven experimentation environments) and networks of research or innovation structures that could well become potential users of the DataBench ToolBox (e.g. Big Data CoE, DIH). All of them are briefly described in the document and the value proposition is analysed.

1. Introduction and Objectives

It is not by chance that we use the criterion of impact when we evaluate projects and investment opportunities. And the reason is that we all know that scientific and technical excellence is not a guarantee to be successful. Your system may work perfectly well and react as expected, but you may be doing something where the value is not evident for users, or where the value proposition is not relevant enough as to invest in it. So, being technical performance a pre-requisite but not a guarantee for adoption, we have to create an environment that maximizes the probability of convincing users. This is even more important in the case of DataBench, since the major outcome of the project will be a tool that relates technical performance of Big Data Technologies (sometimes referred to as BDT) with business indicators. The tool, known as DataBench Toolbox, is fully dependent on the collaboration with different stakeholders along the product cycle, from its inception till its usage:

- Before it is launched (design and implementation phase) it will be fed by use cases that help us understand relationships between indicators and that allow the system to learn and make the right recommendations: this implies working with many companies in diverse sectors and with heterogeneous technical environments;
- Before it is launched (design and implementation phase) it will require integration of existing benchmarks: the DataBench Toolbox will not replace benchmarks that work well but will provide an environment that facilitates the work of those people that run the benchmarking processes (for example, by including several benchmarks in the same system). It is therefore important to work closely with the benchmarking community;
- When it is launched, the DataBench Toolbox will make sense only if there is a community of users ready to adopt the methodology and tool developed by DataBench.

Therefore, developing a solid network of collaborators, contributors and adopters/users is essential to the success of DataBench.

We will have to work on gaining the credibility of those stakeholders, starting with some ambassadors and later on using networks to multiply the DataBench effect. On the other hand, we should not neglect the impact of running a well-designed communication campaign. Communication and dissemination actions built on top of a good product will for sure contribute to achieve the expected impact.

WP6 is the WP in charge of Consensus Building, Dissemination and Exploitation. As such, it will have a major responsibility in achieving the biggest possible impact. The work in this WP is organized around three major lines:

- Community engagement or active involvement of stakeholders
- Awareness, dissemination and marketing & communication campaign
- Exploitation, DataBench adoption and sustainability

D6.1 provides the initial version of the Dissemination and Liaison Plan, addressing elements of the two first action lines. The plan is not a closed and hermetic document. On the contrary, it will be revised on a periodic basis. Updated versions and potential changes of direction will be reported in the follow-up deliverables (D6.3 and D6.4), which will also contain the report of activities carried out by DataBench according to

this strategy as well as a critical assessment of the achievements against the defined Key Performance Indicators (KPIs). This framework will undoubtedly feed D6.5 Exploitation and Sustainability plans. D6.5 will address the third action line of the list and will pave the way to the self-sustainability of DataBench.

Connections with the other WPs of the project are important, since they provide the content, as well as the methodologies to work with the different stakeholders and communities once they have been reached out by WP6.

We could not finish this introduction without referring to the special environment DataBench belongs to. DataBench is part of the Big Data Value PPP (also referred to as BDV PPP or PPP in this document). The private part of the PPP is the Big Data Value Association (or BDVA), which represents the views of industry; the public part of the contract is represented by the European Commission (EC). While planning, monitoring and some content-related activities are part of the activities performed by BDVA, the implementation of the so-called Strategic Research and Innovation Agenda (SRIA) released by BDVA happens mainly through the projects. DataBench is one of those projects and as part of the program, collaboration and cooperation with its peers is expected for a coherent development of the PPP. Furthermore DataBench has the potential to become an essential element of the program since a) it could help to assess the impact of BDT in the different pilots/use cases/environments of the PPP and b) thanks to the learning experience of the system, it could help companies to understand and take the right decisions when it comes to decide which technologies, frameworks or architectures are needed to fulfil their real needs and business requirements.

2. DataBench Concept, Objectives and Assets

The goal of DataBench is to design a benchmarking process helping European organizations developing Big Data Technologies to reach for excellence and constantly improve their performance. This will be done by measuring their technology development activity against parameters of high business relevance. In order to achieve this, DataBench has defined the following concrete objectives, as stated in our initial dissemination material:

1) Provide the BDT stakeholder communities with a comprehensive framework to integrate business and technical benchmarking approaches for BDT.
2) Perform economic and market analysis to assess the "European economic significance" of benchmarking tools and performance parameters.
3) Evaluate the business impacts of BDT benchmarks of performance parameters of industrial significance.
4) Develop a tool applying methodologies to determine optimal BDT benchmarking approaches.
5) Evaluation of the DataBench Framework and Toolbox in representative industries, data experimentation/integration initiatives (ICT-14) and Large-Scale Pilot (ICT-15).
6) Liaise closely with the BDVA, ICT-14 and 15 projects to build consensus and to reach out to key industrial communities, to ensure that benchmarking responds to real needs and problems.

Those statements represent the "how", i.e., the way the project will realize its major goal, but for the purpose of communicating the project, we will focus much more on the "what", i.e. the assets and outcomes that will be generated by the project and constitute its main essence. They will be the tangible tools that will be provided to the community.

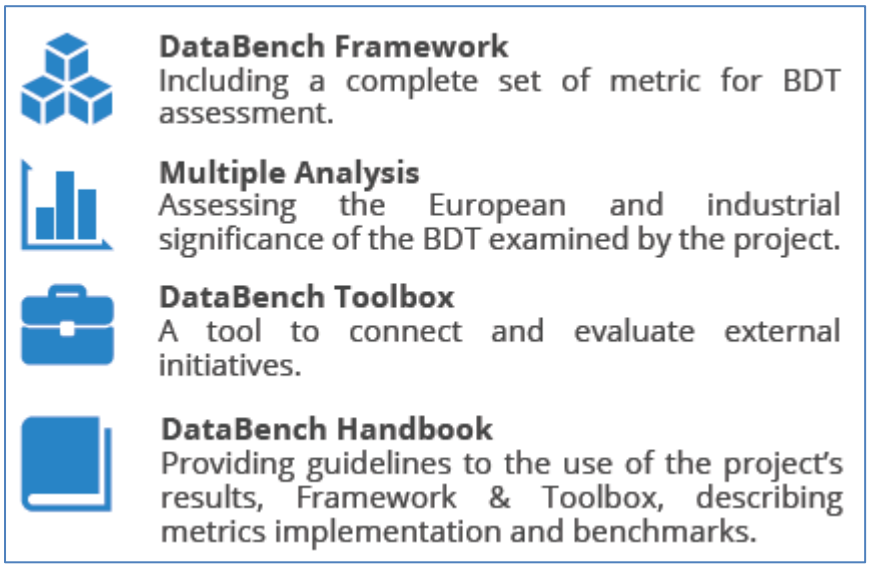

Figure 1 DataBench expected results

Each of the assets listed in the picture above relates to the others and adds value to the potential usage of DataBench. However, and as we will see later on, the main product is the so called **DataBench Toolbox**.

As described in [1], the DataBench Toolbox will allow users to select from the plethora of existing benchmarks the one(s) that suits their needs, deploy and use it, as well as the possibility to upload the results of the benchmark execution to get a homogenized set of metrics, including business insights.

For the purpose of an overall understanding of the DataBench ecosystem we include here the functional view with its elements, as depicted in [1]. More detailed technical descriptions can be found in that document.

- **The DataBench Toolbox**: The DataBench Toolbox is the core technical component of the DataBench Framework. It will be the entry point for users that would like to perform Big Data benchmarking and will ultimately deliver recommendations and business insights. It will include features to reuse existing big data benchmarks, and will help users to search, select, download, execute and get a set of homogenized results. The Toolbox is based on the analysis and work performed in other areas of the project.
- **Benchmarks integrated into the Toolbox**: external Big Data benchmarks are the input to the whole DataBench framework, and in particular to the DataBench Toolbox. These benchmarks will be made available to the users through an easy-to-use Toolbox user interface. The degree of integration of the Toolbox with the existing benchmarks will vary depending on the cases.
- **Business KPIs**: Deriving business KPIs from the results of running Big Data benchmarks is not a straightforward process. It will depend on the business context and therefore it is not completely clear to what extend it could be automated. However, different approaches to derive business insights are currently under study in the context of DataBench and they will provide the basis for one of the most important added values of this initiative, with an extremely high potential impact for the Big Data Value PPP and the data ecosystem in Europe.
- **AI framework**: The framework will provide recommendations to the benchmarking community based on past experiences. This component will also be integrated with the Toolbox.

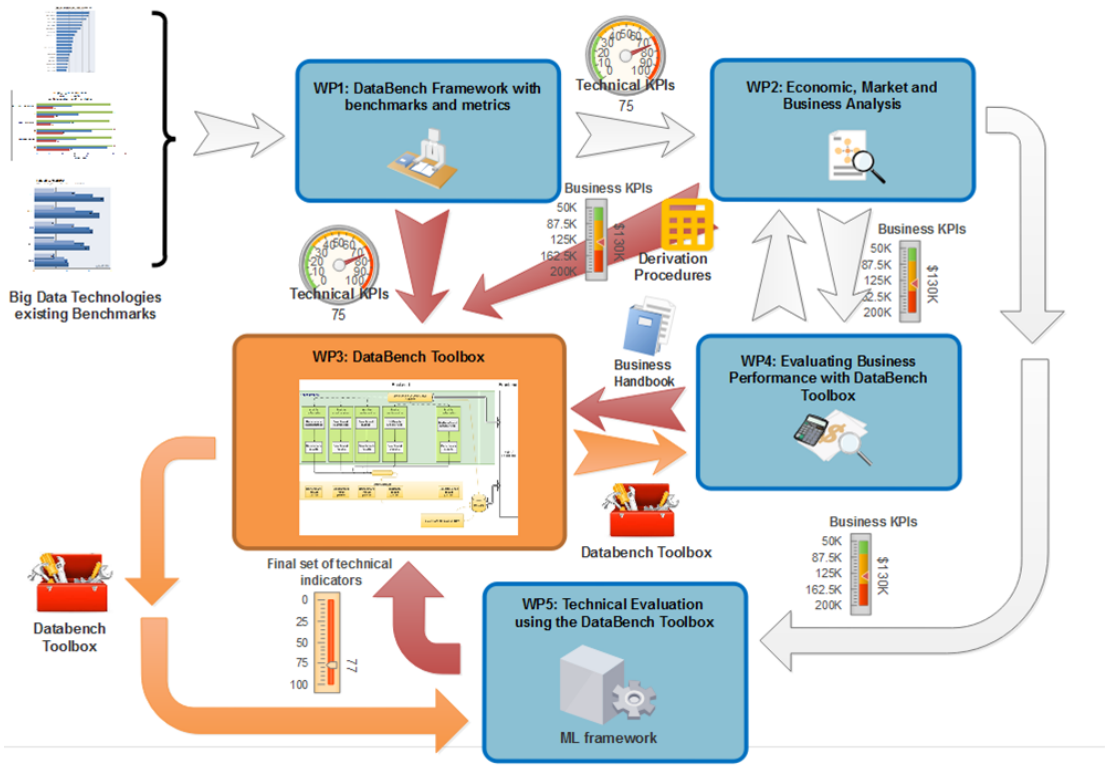

Figure 2 Functional view of the DataBench ecosystem [1]

Looking deeper into the Unique Selling Propositions of the DataBench Toolbox we can derive the overall benefits proposed by DataBench and the way they fit with the needs of the market. Some simple messages that will be the basis of our initial communication campaign are depicted here:

- Organizations **need** methods to select the best Big Data Technologies (BDT) for their business needs, particularly in emerging application areas such as the IoT smart environments, or the AI and robotics implementations;
- There is still a **wide gap** between the capability to monitor technical performance of BDT and the assessment of their business impacts, and this is one of the barriers slowing down adoption;
- There is a **lack** of objective, evidence-based methods measuring the correlation between BDT technology benchmarks and business benchmarks as well as BDT capability to impact the competitiveness and market success of organizations
- No one has yet bridged the **gap** between technical benchmarks and economic performance

In summary, there is a plethora of Big Data tools and frameworks, but how can companies entering the Big Data "era" know which one(s) can fulfil their needs and their business/technical environment? Sometimes a technical benchmark is directly related to an understandable business benefit. For example, we all agree that if you have to provide a result of an operation in a financial environment (investment) and your velocity (number of operations processed per second) increases, you will be able to react faster. In such environment reacting faster is a clear benefit, but better technical performance of some indicators sometimes does not translate so easily into

a better RoI. DataBench will help to relate technical indicators to business ones, and hence will provide the tools to better understand the economic impact of some technical decisions associated to different choices of BDT.

For DataBench there are obvious challenges; **ensuring that the benchmarks identified respond to actual business needs** is one of them and getting the **acceptance and recognition by the industrial community** is another one. If the first one is successfully implemented, then the path towards the second one will be eased.

The main objective of the work proposed in this Dissemination Plan is precisely to support DataBench in those challenges; to create the relationships, mechanisms and tools that will help DataBench to work hand-in-hand with the right communities, the ones that a) will help DataBench to create the right product and/or b) will become users or customers of the DataBench Toolbox.

The extent to which the DataBench Toolbox fulfils its function (i.e. provides the right recommendations and relates technical benchmarks with business indicators in an accurate way) will be the major indicator of (technical) success of the project.

The extent to which DataBench is adopted will be the main indicator of (business) success of the project.

Exactly as mentioned few paragraphs above, the technical success of DataBench will be a pre-condition for the business success. However, the second will also be dependent on additional forces and elements. Here is where the communication campaign designed by the project will be particularly relevant.

The Dissemination strategy includes three major elements, besides the definition of the "product":

- **Identification of target communities** (who is our customer/recipient), including the value proposition for them (why should they be interested about DataBench)
- **Dissemination phases** (when); understanding when different actions have to be implemented is important. For example, too much marketing when the product is not available can lead to a situation where potential users get tired. The ones who are not retained will probably not come back. Messages will obviously evolve over the time too, as well as the tools to be used and the way they will be used.
- **Tools, mechanisms and channels** (how or where): the range of channels that will be used by DataBench include web tools (website, social networks), dissemination material, events or publications, to name some of them. They will also evolve over the time from different points of view (message, frequency of use).

3. Target Audience

3.1 The Big Data Value PPP framework and projects

In October of 2014 the European Commission and the Big Data Value Association (from now on referred to as BDVA [2]) signed a contractual agreement for a Public Private Partnership focused on Big Data Technologies, or, as the formal name says, on Big Data Value (i.e. the ability to transform data into value thanks to BDT). The contractual agreement includes a wide umbrella of objectives, but some of them -that we list below- are specifically relevant in the context of the work DataBench proposes [3]:

Objectives for improved competitiveness:

- Develop solutions leading towards the use of Big Data Technology for increased productivity, optimized production, more efficient logistics, and effective service provision from public and private organizations,
- Develop and diffuse a better understanding of business opportunities of the Big Data sector,
- Drive the take-up and integration of Big data value services in private and public decision-making systems,

Innovation objectives:

- Validate technologies from a technical and business perspective through early trials in cross-organizational, cross-sector and cross-lingual innovation environments.

As we can see, the PPP plays a very important role in driving the use of Big Data Technologies, or said in other words, in driving the demand side. One of the reasons that is usually argued to explain the reluctance of some companies to invest in Big Data is precisely the lack of success stories or clear RoI cases. That is why investing in demonstrators (through large scale pilots or lighthouse projects) and use cases that bring the general concept of Big Data to real processes, products and operational environments is crucial.

From the Monitoring Report of the PPP [4] we can get a good overview of the existing projects in the portfolio as well as their focus. Furthermore, interviews and replies from the projects to this assessment process show that the program accounts for:

- **26 Large Scale experiments** (19 involving closed data) and over 150 use cases and experiments.
- **4 major sectors** (Bio Economy; Transport, Mobility and Logistics; Healthcare; Smart Manufacturing) covered with close to the market large-scale implementations, and **over 10 different sectors** covered in total.

Projects in the PPP reported **151 use cases or/and experiments** conducted during 2017 with a contribution from 8 different projects. BDVA members, including those running BDVA labelled I-Spaces, reported independently and additionally (to the BDV PPP projects) **130 data experiments** during 2017 most of which have been developed in BDVA I-Spaces. In relation to the amount of data made available for experimentation, projects have reported a total of 0,0854 Exabytes (**85,4 Petabytes**) for 2017.

That is the picture of what was developed the previous year. However, a second wave of projects resulted from the last call, giving birth to new initiatives, many of them falling under the topic of data integration and thus, of high interest to DataBench. One of the conclusions we can extract is that, even though projects will be a first source of data for the purpose of testing the DataBench Toolbox, we should not neglect other interesting environments, also within the PPP framework, which could give us access to a wide umbrella of data sets, sectors and types of experiments. This is especially the case, as mentioned above, of the **Innovation Spaces**, or in short, I-Spaces.

For a better overview of the current state of play of sectors in the Big Data PPP we add a comprehensive analysis made by BDVA and the BDVe project in Annex 1 in the current document. Notice that this information is available in the monitoring report 2017 [4], but no public version of such document is available at the time of submitting this deliverable. Nevertheless we consider this information of interest for the overall understanding of the portfolio and therefore the environment DataBench needs to deal with.

SRIA Mechanism	WP Topics	Projects calls 2016	Projects calls 2017
i-Spaces	ICT-14	**EW-Shopp:** Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey **AEGIS:** Advanced Big Data Value Chain for Public Safety and Personal Security **QROWD:** Because Big Data Integration is Humanly Possible **FashionBrain**: Understanding Europe's Fashion Data Universe **euBusinessGraph**: Enabling the European Business Graph for Innovative Data Products and Services **SLIPO**: Enabling the European Business Graph for Innovative Data Products and Services **BigDataOcean**: Exploiting Ocean's of Data for Maritime Applications	**BODYPASS:** API-ecosystem for cross-sectorial exchange of 3D personal data **Fandango:** FAke News discovery and propagation from big Data ANalysis and artificial intelliGence Operations **ICARUS:** Aviation-driven Data Value Chain for Diversified Global and Local Operations **TheyBuyForYou**: Enabling procurement data value chains for economic development, demand management, competitive markets and vendor intelligence **Lynx**: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe **Cross-CPP**: Ecosystem for Services based on integrated Cross-sectorial Data Streams from multiple Cyber Physical Products and Open Data Sources
		Data Pitch: Accelerating data to market **(INCUBATOR)**	**EDI**: European Data Incubator **(INCUBATOR)**
Lighthouse Projects	ICT-15	**DataBio**: Datan-Driven Bioeconomy **TT**: Transforming Transport	**BOOST 4.0:** Big Data Value Spaces for COmpetitiveness of European COnnected Smart FacTories 4.0 **BigMedilytics:** Big Data for Medical Analytics
Cooperation and Coordination Projects	ICT-17.a	**BDVe:** Big Data Value ecosystem	N/A
	ICT-18.b	**e-Sides**: Ethical and Societal Implications of Data Sciences	N/A
Technical Projects	ICT-16	N/A	**BigDataGrapes**: Big Data to Enable Global Disruption of the Grapevine-powered Industries **BigDataStack**: High-performance data-centric stack for big data applications and operations **CLASS**: Edge and CLoud Computation: A Highly Distributed Software Architecture for Big Data AnalyticS **E2Data**: European Extreme Performing Big Data Stacks **I-BiDaaS:** Industrial-Driven Big Data as a Self-Service Solution **Track and Know:** Big Data for Mobility Tracking Knowledge Extraction in Urban Areas **TYPHON:** Polyglot and Hybrid Persistence Architectures for Big Data Analytics
	ICT-18	**SODA**: Scalable Oblivious Data Analytics **MH-MD**: My Health - My Data **SPECIAL:** Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compliance	
	ICT-17.b	N/A	**DataBench**: Evidence Based Big Data Benchmarking to Improve Business Performance

Figure 3 Portfolio analysis: mapping with the BDV SRIA implementation mechanisms (source: [4])

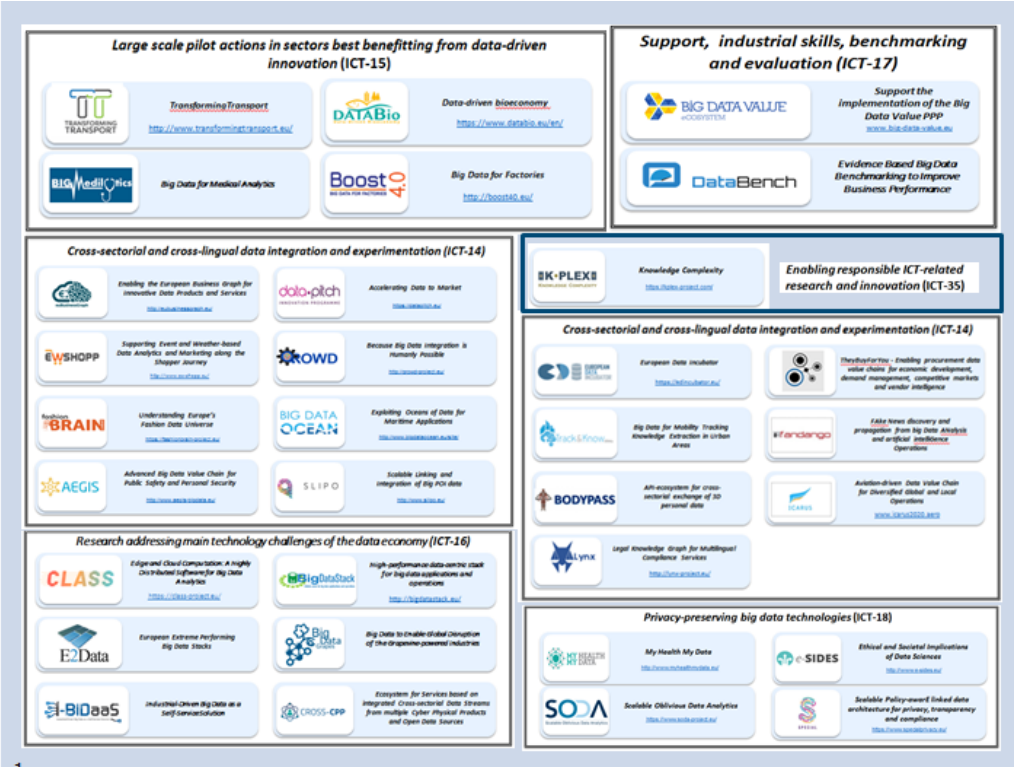

Figure 4 Visual representation of the PPP project portfolio

Benefits (why):

- For DataBench: environment for validation and testing; in addition, potential creation of customer/user base
- For the projects: a way to identify additional areas where benchmarks are needed and redirect efforts towards them, contribute to the model (research-oriented), get advantage to benchmark the solutions they are producing at technical level (when BDT are generated) and benchmark vertical pilots in their performance with respect to business indicators that are affected by BDT
- For the PPP as a whole: DataBench could become an extremely interesting tool to evaluate and benchmark the solutions produced by the PPP and understand the progress with respect to state-of-the-art technologies. As such, the DataBench Toolbox has the potential to become a very relevant tool for the program (including decision makers such as the European Commission) to understand the impact of the investments made in Big Data.

How (the way the collaboration will be organized):

DataBench will address projects in the PPP portfolio through the well-established **governance structure of the program** (included either in the Grant Agreement or Contractual Agreement of each project); the governance structure governs the relationships and collaborations within the portfolio. Such structure is composed of two formal committees, the so call **Steering and Technical committees**. In the first one all project coordinators share knowledge, experience and outcomes of their projects and define together actions in areas like PPP monitoring (exercise known as the "monitoring report"), marketing and communications (including events) and in general topics where we find commonalities among projects (e.g. impact of GDPR). In the case of the Technical Committee, it is composed by the technical coordinators of

all projects; so far the focus has been on mapping their outcomes to the **BDVA reference model** as an attempt to understand the progress of the PPP against the objectives defined in the SRIA[1] [5]. This committee has also been used as a vehicle to gather input for standardization bodies at program level. The committees include experts from different task forces of BDVA when it is needed (a clear example of this is precisely the group on standards but also the one on benchmarks). The active involvement of some of the DataBench partners in the management structure (Atos as chair of the SC), but also the working structure of BDVA (it is of special relevance the leadership of TF6 on technical aspects by SINTEF, where activities on benchmarks take place) give DataBench a privileged position to relate to the projects. All these channels will be used as follows:

- Steering Committee: to reach out to all project coordinators; provision of general information on events, achievements and availability of results; overall coordination and knowledge exchange
- Technical committee and BDVA TF on benchmarks: they will be used in a coordinated way. While the first one will be a channel to address all projects with technical info (in many cases in conjunction with SC), the second one will be used as working environment to develop contents and actions in a deeper way.

DataBench will develop pilots with some of these projects. This will require a closer relationship with individual projects and their actors. As such, all projects to be addressed by DataBench will be distributed among DataBench partners to make sure that they all have a single point of contact in the project. This will make communications more agile and should also lead to a trusted (personal) relationship.

3.2 Innovation Spaces or I-Spaces

As it was mentioned before, a good range of data-driven experiments is happening in the context of the I-Spaces. I-Spaces are secured infrastructures for experimenting with data (open and closed data) and different analytics platforms. They do not only provide technical facilities but also additional services such as legal advice for management of contractual aspects between data owners and users.

Value proposition (why):

DataBench could help them to assess different platforms and tools and also to understand the impact between specific technical architectures/approaches and business benefits derived from them.

For DataBench they are an opportunity that is worth exploring, since they could become potential beneficiaries/customers of the DataBench framework. We provide here a sample of some of the most relevant I-Spaces. They will be further analyzed in order to select cases for DataBench validation and with the aim of defining additional exploitation roots. The complete list, labelling process and full descriptions can be found in [2].

1 SRIA stands for Strategic Research and Innovation Agenda.

I-SPACE	LOCATION	DESCRIPTION	PLATFORMS	SERVICES
Smart Data Innovation Lab (SDIL) [6]	Karlsruhe (Germany)	SDIL offers big data researchers access to a large variety of big data and in-memory technologies. Industry and science collaborate closely to find hidden value in big data and generate smart data. Projects focus on the strategic research areas of Industry 4.0, Energy, Smart Cities and personalized Medicine	SAP Hana, Software AG Terracotta, IBM Watson Foundation Power 8, Huawei Fusion Insight, System: HT Condor	• Infrastructure providing: The infrastructure, including technical support, is provided free-of-charge by the SDIL operation partners to any SDIL project. • Communities: SDIL provides access to experts and domain-specific skills within Data Innovation Communities fostering the exchange of project results. They further provide the possibility for open innovation and bilateral matchmaking between industrial partners and academic institutions. • Data curation: The SDIL guarantees a sustainable invest to all partners by curating industrial data sources, best practices, and code artefacts, that are contributed on a fair share basis. • Data Anonymization: The SDIL offers various anonymization tools to its projects which are applicable to data from research and industrial sources.
Teralab [7]	Paris (France)	Teralab is a Big Data platform to enhance the value of industrial data in partnership with laboratories or collaborative projects. It has been operational for more than 3 years and offers state-of-the-art infrastructure and tools. TeraLab's infrastructure is secure, sovereign and neutral. It provides the necessary security guarantees for industrial partners so that they can make available, within a defined framework, their high-value data for research or innovation projects.	Apache Suite Open Source on top of in memory analytics based on Atos Bull technology.	• ability to perform experiments on more data-sets from different sectors across EU; • access for industry from any member state to SotA experiment platforms & tools; • market-realistic use-conditions to test & validate new tool concepts from Academia; • enabling cross-regional access to SotA Academic knowhow; • sharing best governance and incubation support methods between existing i-Spaces; • wider access to industry data and challenges for education and training of students.
Know-Center [8]	Graz (Austria)	Know-Center is Austria's leading research center for data-driven business and big data analytics. It conducts applied and interdisciplinary research in the field of computer science in the areas of data-driven business, big data and cognitive computing. Main research areas fall under: Knowledge Discovery, Knowledge Visualization, Social Computing, Ubiquitous Personal Computing & Business Models, Data Management and Data Security.	HDFS, MapReduce2, YARN, Tez, Hive, Pig, ZooKeeper, Kafka, Kerberos, Slider, Zeppelin Notebook,Jupyter Notebook, Spark, Spark2, Apache Solr.	The offer includes consultations, data analysis and trainings. The Big Data Lab offers a simple and direct access to both expertise and infrastructure. In addition to Apache Hadoop, the Big Data Lab is equipped with other big data technologies such as Apache Spark and Apache Storm on its computer clusters. Integrated within an international network around the topic of Big Data and Data Science, the Know-Center provides its partners with access to the latest trends and findings in this area.

		From the business point of view, main areas covered by Know-Center are: Industrial Data Analytics, Data-Driven Markets, Strategic Intelligence, Data-driven Process and Decision Support, Learning 4.0 and Digital Life Science.		
RISE SICS North ICE [9]	Luleå (Sweden)	ICE, the Infrastructure and Cloud research & test Environment, is a research data center inaugurated in January 2016. The facility is open to use primary for European projects, universities and companies. However, customers and partners from all over the world are welcome to use ICE for their testing and experiments.	• HOPS - Hadoop as-a-service; •Tensorflow-as-a-service; • Streaming analytics-as-a-service; • Apache Spark; • Apache Flink; • Customized common development environment	The range of the offering extends from choosing to just use our Hadoop application HOPS, all the way to full service with tool experts, analysts and even the possibility to use the data owned and stored by the center. ICE offer covers all parts of the stack: • Big data and machine learning - Computing capacity, platforms and tools for handling big data and machine learning; • IT and cloud - testing and experiment environments for software development, scaling and infrastructure optimization; • Facility and IT HW - possibilities for testing disruptive innovations concerning the facility and hardware of a datacenter; • Utility - measurements and research securing a sustainable society with efficient datacenters as a part of the energy system.
EURECAT: Big Data Centre of Excellence [10]	Barcelona (Spain)	The Big Data Centre of Excellence in Barcelona is an initiative led by Eurecat which has been launched in February 2015 with the support of the Government of Catalonia, the Barcelona City Council and Oracle. The Big Data CoE constructs, evolves, integrates and makes available to companies differential Big Data-related specialised knowledge, tools, data sets and infrastructures that will allow them to define, experiment with and validate Big Data models and their impact on business, as well as define innovative solutions within a collaborative framework with key agents from the sector.	DATURA is an OpenStack based platform developed by Eurecat to provision, configure and deploy a whole Big Data stack through an UI assistant matching different project requirements. ATURA currently allows to deploy Hadoop environment clusters (including HDFS, YARN, Hive, Pig and Sqoop) as well as Spark, Elasticsearch and Kafka clusters.	The portfolio of services provided by EURECAT ranges from Big Data Strategy and Planning to Big Data Value and Discovery or Big Data deployment support.

Table 1: Description of i-Spaces as potential users of DataBench

3.3 Other potential users of DataBench

I-Spaces represent a very obvious case of potential user of the DataBench outcomes, but there are other instruments with similarities that could well become customers of the DataBench solution and in a first step, validators/contributors to it. Since our resources are limited they will be kept as part of the dissemination strategy but if some interest is raised by our marketing material we will address them in a more focused way.

In that group we have identified specifically the network of Big Data Centers of Excellence (CoE) as well as Digital Innovation Hubs (DIH) with competences on Big Data, data mining and database management. The network of CoE currently accounts for 55 centers. The case of DIH is more difficult to quantify, since the tool proposed by JRC -where DIH can be checked- is based on entries filled in by the DIH as such.

Figure 5 Distribution of Big Data Centers of Excellence [11]

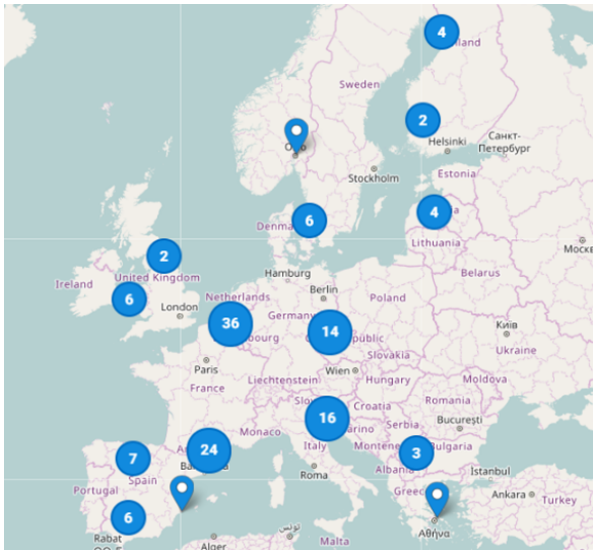

Figure 6 DIH with competences in Big Data, data mining and database management [12]

Finally, as an extension of the networks listed here we would like to mention the investment made by some countries in Big Data, resulting in a myriad of initiatives where data ecosystems have emerged at national level. These initiatives operate as BDVA, but in their respective countries or regions and in some cases they are extremely powerful. It is not by chance that some overlaps/synergies take place between some of the instruments described in this section (CoE/DIH/national initiatives). They can also be very interesting channels to disseminate the results of DataBench and depending on the nature of the initiative they could also become users (that could be the case of an I-Space-type of initiative) or multiply the effects of the dissemination strategy of the project.

An overview of such initiatives is represented in the following diagram. They will be analysed with the support of the BDVe project [13] to understand their potential value for DataBench.

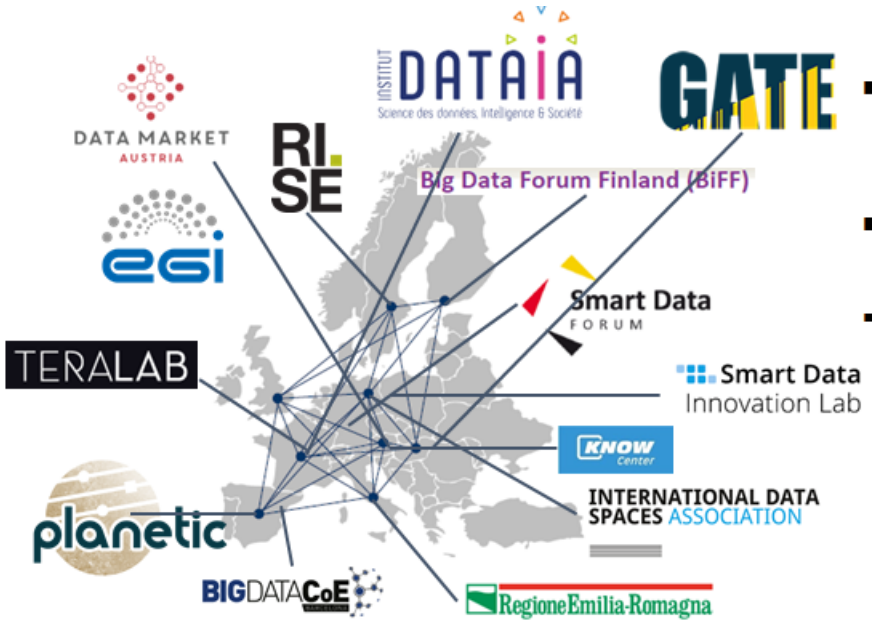

Figure 7 Sample of National Initiatives on Big Data [13] [14]

3.4 Benchmarking communities

One of the target communities of DataBench is the one composed by different benchmarking efforts in the area of Big Data Technologies, where many benchmarks for particular indicators of technical nature exist. At proposal stage DataBench already identified a list of potential benchmarks we would like to relate to that we reproduce here. Some of their benchmarks will be integrated into the DataBench Toolbox. They will therefore be direct contributors to DataBench. In exchange they will benefit from the opportunity to increase their user base. The engagement with them will be realized through the setting up of the **Big Data Benchmarking Community (BDBC)** established by WP1, which will help to build synergies between the main international benchmark communities (such as TPC, SPEC, BigDataBench, Linked Data Benchmarking Council (LDBC), Hobbit and others) to create a self-sustainable interaction model likely to continue after the end of the project.

Initiative	Description
YCSB	A benchmark designed to compare emerging cloud serving systems like Cassandra, HBase, MongoDB, Riak and many more, which do not support ACID. It provides a core package of 6 pre-defined workloads A-F, which simulate a cloud OLTP application.
BigBench	It is an end-to-end Big Data benchmark that represents a data model simulating the volume, velocity and variety characteristics of a Big Data system, together with a synthetic data generator for structured, semi-structured and unstructured data, consisting of 30 queries. Currently, the most popular implementation of BigBench is for the Hadoop platforms using also Hive, Mahout, Spark and the Natural Language Processing Toolkit (NLTK). BigBench was adopted by TPC as TPCx-BB.
BigData Bench	It is an open source Big Data benchmark suite consisting of 15 data sets (of different types) and more than 33 workloads. It is a large effort organised in China available with a toolkit that might be used for different adaptions.
LDBC-1: Semantic Publishing Benchmark (SPB)	It is a European project-based approach resulting in the establishment of the Linked Data Benchmark Council (LDBC) benchmark for RDF database engines inspired by the Media/Publishing industry, particularly by the BBC's Dynamic Semantic Publishing approach.
LDBC-2: Social Network Benchmark	It consists of a data generator that generates a synthetic social network, used in three workloads: Interactive, Business Intelligence and Graph Analytics.
LDBC-3: Graphalytics	It is an LDBC benchmark for graph analysis platforms such as Giraph. It consists of six core algorithms, standard datasets, synthetic dataset generators, and reference outputs, enabling the objective comparison of graph analysis platforms.
TPCx-HS	It stresses both the hardware and software components including the Hadoop run-time stack, Hadoop File System and MapReduce layers. The benchmark is based on the TeraSort workload, which is part of the Apache Hadoop distribution.
SparkBench	SparkBench, developed by IBM, is a comprehensive Spark specific benchmark suite that comprises of four main workload categories: machine learning, graph processing, streaming and SQL queries.
TPCx-BB	TPCx-BB is a measure the performance of Hadoop based Big Data systems. Based on BigBench, it measures the performance of both hardware and software components by executing 30 frequently performed analytical queries in the context of retailers with physical and online store presence.
Yahoo Streaming Benchmark (YSB)	It is an end-to-end pipeline that simulates a real-world advertisement analytics pipeline. Currently implemented in Kafka, Storm, Spark, Flink and Redis.
DeepBench	Benchmarking tool for evaluating the performance of deep learning operations across hardware platforms. DeepBench initially provided benchmark results for four platforms: NVIDIA TitanX, NVIDIA M40, NVIDIA TitanX Pascal and Intel Xeon Phi processors.
DeepMark	Deep Learning benchmark on use of multi-GPU platform for a set of Big data types – in particular for Image/Video/Audio and Text.
Stream Bench	It covers 7 micro-benchmark programs that intend to address typical stream computing scenarios, implemented in Spark Streaming and Storm.
RIoTBench	A Real-time IoT Benchmark suite, consisting of 27 IoT micro-benchmarks and 4 real-application benchmarks reusing the micro-benchmark components, along with performance metrics. The goal of the benchmark suite is to evaluate the efficacy and performance of Distributed Stream Processing Systems (DSPS) in cloud environments.
Hobbit Benchmark	A European project-based Holistic and vertical benchmark approach derived from business data requirements and focusing on 4 data life cycle areas with a basis in graph and linked data representations. Hobbit I Generation & Acquisition, Hobbit II Analysis & Processing, Hobbit III – Storage & Curation, Hobbit IV – Visualisation & Services. Open source available benchmark framework.
BigBench V2	University of Frankfurt is involved in the development of BigBench V2 (in progress in 2017), BigBench V2 separates from TPC-DS with a simple data model. The new data model still has the variety of structured, semi-structured, and unstructured data as the original BigBench data model. The difference is that the structured part has only six tables that capture necessary information about users (customers), products, web pages, stores, online sales and store sales. A scale factor-based data generator for the new data model has been developed. The web-logs are produced as key-value pairs with two sets of keys. The first set is a small set of keys that represent fields from the structured tables like IDs of users, products, and web pages. The other set of keys is larger and is produced randomly. This set is used to simulate the real life cases of large keys in web-logs that may not be used in actual queries. Product reviews are produced and linked to users and products as in BigBench but the review text is produced synthetically contrary to the Markov chain model used in BigBench.

Figure 8 List of benchmarking initiatives analysed by DataBench

4. DataBench timing: Dissemination Phases

DataBench has defined four major phases of dissemination and communication activities that will be revisited during the project if needed. Initial input was already sketched at proposal stage and has been used here with minor modifications and adaptations as main basis for the upcoming work.

4.1 Phase 1 - Create awareness (M1-M12)

The initial phase will focus on creating awareness about the project, identifying the key stakeholder communities and building working relationships with them (some of these elements have been depicted in previous sections). Key to this objective will be on the one hand the Big Data Benchmarking Community (BDBC) that will allow DataBench to work hand-in-hand with the most relevant benchmarking initiatives, but also the relationship with use cases and pilots, especially through the connection with the PPP project portfolio. From a communication point of view the strategy will focus on **ensuring that as many people and organisations as possible that are stakeholders in the Big Data benchmarking and business analysis arena know about the project, its objectives as well as the expected benchmarks and metrics that will be delivered**. This will include presentations and visibility at some Big Data conferences and workshops as well as the elaboration of the first dissemination kit comprising at least a project fiche, handout/flyer, roll-up and a poster, besides a DataBench presentation template. This material will be used internally to ensure that the project team communicates the same coherent message, but also externally to assess the potential acceptance of the concept. This will be assessed and input will be later on used to refine messages. The web strategy will also be launched in this initial phase, opening our main "window" to the world through the website and first steps in social networks.

4.2 Phase 2 - Increase the potential impact (M13-M24)

During the second year, the project will deliver the preliminary benchmarks, the first metrics and the Alpha version of the Toolbox. The dissemination strategy will then focus on **creating an appealing and "marketable" version of the results, and to communicate and share them with the stakeholder communities** targeted in Phase 1, increasing the willingness to participate in the test and validation of these methods. The project will organise webinars and participate in community events to support the collaborative working relationships established with the projects and other instruments that may be useful to this purpose (as identified in section 3 of this document). Usage of channels like newsletters or blogs will be maximized. This phase is expected to leverage the project website and social media to ensure also the wider dissemination of results and to attract the attention of potential users of the Toolbox and the proposed benchmarks.

4.3 Phase 3 - Maximise Results (M25-M34)

During the next ten months, as the project develops the final version of the benchmarks and toolbox, the main objective will be to **maximise the potential impacts by continuing the communication and networking activities**, as depicted in Phase 2, but with a focus on **building consensus and recognition**. Beyond use of webinars, participation in workshops and conferences, we will also begin to publish technical papers detailing the Benchmarks and the Toolbox, which will be presented in at least three international Big Data

conferences. In the final year, we will organize specific community events to present complete DataBench findings and to run demos to exhibit the effectiveness of the Toolbox. The primary goal is to increase the impact through the external benchmarking initiatives and to attract potential users or clients including Big Data industry and policy makers. This will constitute the "**DataBench product roadshow**".

4.4 Phase 4 - Valorisation (M34-M36+M>36)

From the near end (final two months) of the project and after the formal end of the project duration, **demonstrations for specific audiences** (research community, industry, policy makers) will be carried out. In the same period, the contents and findings related to Horizontal Benchmarks, Benchmarks of European and industrial significance, and DataBench Handbook associating the DataBench Framework and the DataBench Toolbox describing metrics implementation and benchmarks will be widely **published in national/international journals and in on-line media**. The primary goal will be to demonstrate to internal and external potential adopters, existing Big Data research communities and policy makers the relevance of benchmarking approaches for Big Data Technologies.

5. Implementation of the Strategy: tools and channels

This section describes the "how-to" part of the strategy, essentially channels, tools and mechanisms that will be used by DataBench to create awareness and disseminate the different messages to the communities of interest that have been previously described.

5.1 DataBench branding

The first step consists in the definition and agreement on a branding that reflects the **identity of DataBench**. This was done at the beginning of the project with the support of professional designers resulting in several creative elements, including the logo, whose final version is:

Figure 9 DataBench logo

Details about the meaning of font, colours and other elements, as well as the process behind the artistic creation can be found in D6.2.

5.2 Marketing Material

Once the identity is created, all the other materials should follow the same graphical image. DataBench has prioritized the elaboration of the following material for the first phase of the project: project fiche, handout and roll-up. This is complemented by a series of templates that go from deliverables to presentations and press releases. All of them can be checked in D6.2. This initial marketing toolkit should be enough at this stage of the project where main content is being created. For the second part of the year, and looking at some important and big events of the research community where DataBench will be present (EBDVF, ICT event), we may create an additional poster focused on specific technical contents and a video. We are also in the process of selecting a catalogue of *goodies* that should serve the purpose of attracting visitors to our booth/exhibition area in events where a lot of competitors *fight* to get the attention of people.

5.3 Strategy on the use of web channels

General overview

DataBench will adopt a multi-channel online strategy in order to implement dissemination initiatives and engage with the stakeholders, maximising its impact and visibility. The web channels include the project website (including the community space and ToolBox testing area), the social media channels and the newsletter. A brief description of all of them is provided here, although we redirect interested readers to D6.2, where they can find more information.

Project website

The project website, available at the URL www.databench.eu, is based on the open source content management system (CMS) by WordPress that allows users with an account to create web content via a text-based interface through a web browser without any HTML editing involved. The website presents the vision and the most important information about the project; as it was said before, it is considered our main window to the world, and as such, it also provides access to all public materials. Find below a glimpse on main sections and information that can be found in the current version:

- Home page: designed with an attractive and eye-catching layout to engage the visitors, and in line with the visual identity of the project, it gives an overview of the focus of the project, as well as quick access to the latest content uploaded on the website and posts published on the social media channels.
- The project: this section provides a short description of the project scope, its main objectives and expected outcomes. It also presents project partners.
- Resources: this section is a repository of the public materials produced by the project, available for visitors' download. The resources available include public deliverables, scientific publications, presentations, marketing materials and webinar recordings.
- News and events: this section features the updates about the development of the project and dissemination activities, in the format of journalistic news and blog posts. It also gives information about the participation and contributions to external events and/or events/workshop/sessions organized by the project. A calendar of relevant events, including webinars, is also available in this section.
- Community space and testing area: one of the key outputs of the project will be the DataBench Toolbox, as mentioned several times along this document. It is therefore important to create a place or area on the website fully devoted to it and, in particular, to the potential testing of the tool. The website will feature a restricted area, available upon registration and approval of the project team, which will give access to a collaborative space and to the DataBench Toolbox for testing purposes. The DataBench Handbook providing guidelines on when and how to use the DataBench Toolbox will be made available in the restricted area.

The project team will periodically update the content to make sure that the website provides up-to-date information.

Social media

DataBench will establish a strong social media presence through the following platforms to boost the dissemination and communication activities. Social media channels will be used to connect with the community of stakeholders and to drive traffic to the project website through links to the content uploaded.

- Twitter: this platform has been identified as one of the most powerful means to connect with the research community as well as with other audience groups. It will also be used to attract attention to the activities

of the project, to drive traffic to the website and to cross-publish any relevant third-party content. The account handle is @DataBench_eu.

- Facebook: the project team is aware that Facebook is less used by our target audience, however the DataBench page will be leveraged to reach out to both the general public as well as to the targeted stakeholder groups that are also on Facebook. The account handle is @DataBenchEU.
- LinkedIn: an open group (DataBench Project) has been created to facilitate the discussion with the data community in a forum-format approach.
- YouTube: a dedicated channel (DataBench Project) has been created to share any video produced by the project, including the recorded videos of the webinars.
- SlideShare: the project account (DataBench) will be used to periodically share DataBench public presentations.

All social media channels reflect the visual identity of the project. Each post will be accompanied by visual content and will encourage engagement.

With the aim of generating traffic, the link to the DataBench website is displayed on all social media profiles.

Newsletter

At the time of submitting this document the project acknowledges that newsletters are important tools to inform people interested in the topic on a periodical basis. They enable the inclusion of more detailed information about the project, thus engaging the community in a more solid way. Main objectives pursued by DataBench would be:

- Raise awareness about DataBench and its activities, in particular in the first phase of the project;
- Give regular updates on the project activities;
- Drive traffic to the project website where the complete text of the newsletter will be available;
- Increase community engagement.

Nevertheless, DataBench is still analysing the suitability of integrating the content into the BDVA one or having its own newsletter. The first option seems more promising, since we would target immediately more than 1000 recipients (close to 2K before GDPR). At the time of submitting this document, DataBench has already contributed to the BDVA/PPP newsletter with specific content.

5.4 BDVA tools and engagement with the PPP community

Even though we have already tackled the collaboration with projects of the PPP portfolio in a previous section of this document, it is worth going deeper into the possibilities offered by becoming part of a program. In this particular case, DataBench will capitalize the tight relationships with the BDVA working structures and the opportunities managed by the BDVe project through the communications committee (not a formal committee though), which works in alignment with BDVA. Main channels and tools of interest identified at the beginning of the project and that will be worked out are:

- **BDVA website, PPP portal, newsletters and social networks**: BDVe is responsible for the communication activities at program level, and as part of that responsibility

it runs the Big Data Value Portal, a place where you can find descriptions of all projects, among other things. DataBench information has been included as part of this ecosystem and the project contributes to the different communication campaigns launched in that framework. One of the most successful ones is called "project of the week", where a project is selected, and more detailed information is provided to the PPP community so that people get to know it in depth. This portal, which also provides access to different tools (as you will see below) is complementary to the BDVA website, which is used as additional channel for promotion. BDVe manages some of the web tools of BDVA/PPP and in particular the newsletter and social network channels. A lot of attention is paid to the projects of the program, and they are heavily used to highlight their evolution, news, outcomes, etc. This is a clear advantage to any project, since these tools open up the door to thousands of people that otherwise would be difficult to reach out by a single project. The **use of program-level tools is therefore encouraged**. As such, DataBench has already sent out information and used these tools in the last months.

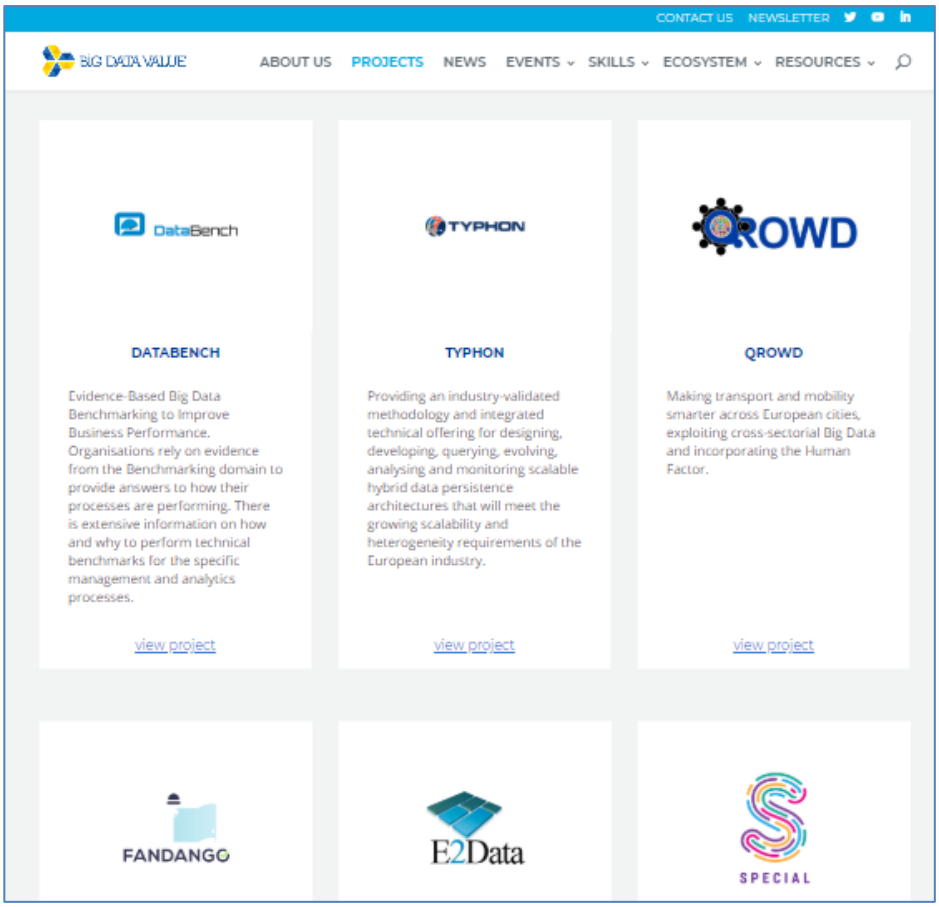

Figure 10 DataBench @PPP portal [13]

(Detail of the section devoted to PPP projects, which runs an algorithm to ensure that projects are shown in aleatory positions)

- **The Big Data landscape**: this tool will be accessible from the PPP portal and represents major actors of the data economy in Europe on a map. It builds on previous efforts, such as the datalandscape.eu developed by IDC and Open Evidence.

Some of the advantages of the landscape are that it allows users to make intelligent search by using multiple filters and it also features different layers of information, including stakeholders that are not single organizations but initiatives that could enable other actors to get engaged in the Data economy. Examples of enablers represented on the map are the network of Big Data Centers of Excellence, a multiplicity of national initiatives and programs focused on Big Data and the existing network of data-driven experimentation environments known as i-Spaces. The map will also show the distribution of pilots implemented in the program. As such, it will be a useful tool for DataBench even though more suited for connections than for pure dissemination purposes.

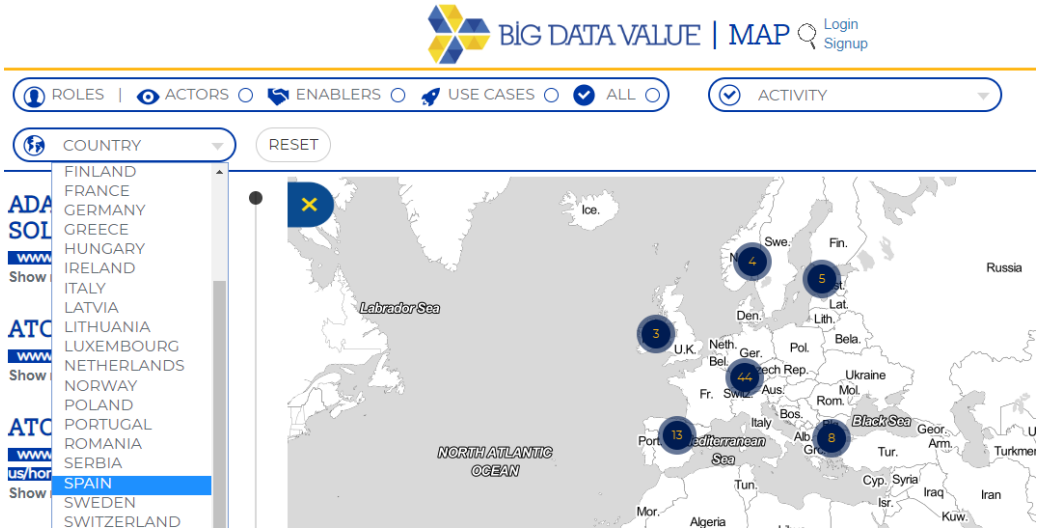

Figure 11 The Big Data Landscape

- **The Innovation Marketplace**. This tool provides visitors to the PPP portal with a catalogue of Big Data solutions that are searchable by using different criteria, such as TRL, domain/sector, typology of solution, etc. It will include all the outcomes produced by PPP projects. The DataBench Toolbox could obviously be one of those even though the way it will be exposed and marketed will be discussed later on in the project, as it could have a high impact on exploitation.

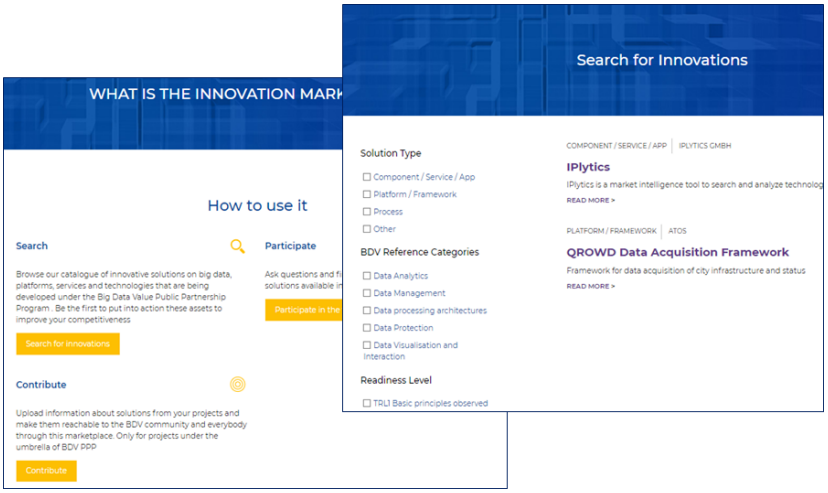

Figure 12 The PPP Innovation Marketplace

- **The PPP Webinar Series**: from November 2018, the BDVe project will launch a very promising tool to spread the knowledge generated in the PPP. Webinars will give visibility to the project results, but also to topics of general interest to the PPP. Speakers will be from PPP projects, BDVA but also external experts. The Webinar Series will have two tracks: a technical one and an industrial track. Some of the pre-selected topics include "Tips to apply GDPR" (i.e. how does GDPR affect a project or a sector, practical examples, implications, etc.), "Data sharing architectures and solutions", "Data value and monetization" or wide information about existing data sources in Europe, as it is the case of Copernicus/DIAS platforms/European Data Portal. It is our pleasure to announce that **the Technical Track will be kicked off with DataBench.**

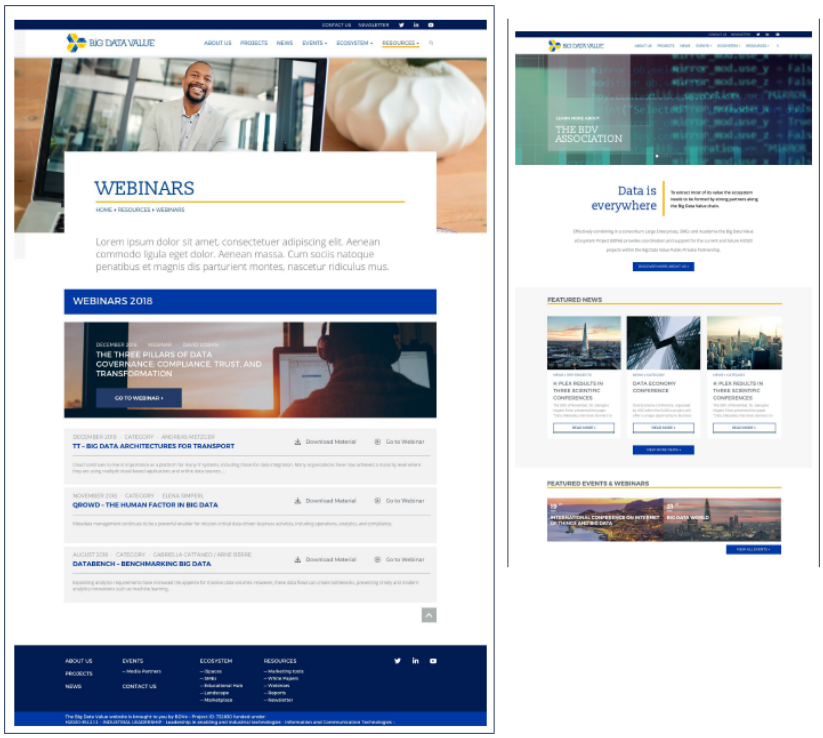

Figure 13 The PPP Webinars Series to be launched in Autumn 2018 with DataBench

- **Events:** as part of the Communications Committee, DataBench will take place in the meetings and actions leading to a global presence of the PPP in events. This will be particularly important to manage the presence of the projects in big events where it is difficult to get a visible place either in the program or in the exhibition area if you go as a single project. Critical mass will be used to get better exposure in PPP flagship events, such as the EBDVF or PPP Meet-ups, EC-related events such as the ICT event or industrial-like events, as it could be the case of CEBIT, where economic reasons prevent a single project from being represented there. Some of these events are mentioned later on in section 5.5.

5.5 Events

DataBench has already identified a list of events of potential interest to the project. They will be further analysed during the course of the project and decisions will be made based on the availability of results (dependent on the implementation phase of the project), budget constraints and opportunities that could be materialized.

Due to the fact that this plan is submitted later than initially expected, we can take advantage of the availability of additional information about events in the first period. A number of opportunities have already been exploited by DataBench. They will be reported in detail in D6.3 Dissemination and Liaison Report (M18). However, we anticipate here some of the actions already carried out by the project.

DataBench in H1 2018

International Conference on Performance Engineering (ICPE) [15]

- **When:** 9-13 April
- **Where:** Berlin (Germany)
- **Short description:** ICPE integrates theory and practice in the field of performance engineering by providing a forum for sharing ideas and experiences between industry and academia.
- **Role of DataBench:** Within the conference program we can find multiple workshops, being one of them the 4th International Workshop on Performance Analysis of Big data Systems (PABS) [16]. It is precisely in this one where DataBench was represented.

Big Data PPP Meet-up Sofia [17]

- **When:** 14-16 May 2018
- **Where:** Sofia (Bulgaria)
- **Short description:** First edition of a PPP event that comes to stay. It is organized by the BDVe project and BDVA with the aim of bringing together the main communities implementing the Big Data PPP: projects and BDVA membership/Task Forces (content-driven working groups of the association). The event was composed by two working days (with a good number of workshops running in parallel) and an Open Day intended to give visibility to the data ecosystem in Bulgaria that was supported by the Commissionaire Gabriel.
- **Role of DataBench:** DataBench was the organizer of a workshop together with the benchmarking TF of BDVA**.** It was the first opportunity for the project to relate to the other projects of the program in an interactive mode.

Figure 14 Big Data PPP Meet-up Sofia, the place where DataBench kicked off collaboration with the PPP projects

Big Data Innovation Conference [18]

- **When**: 7-8 June
- **Where:** Frankfurt (Germany)
- **Short description:** The conference is intended to help companies understand & utilize data-driven strategies and discover what disciplines will change because of the advent of data**.**
- **Role of DataBench:** Presentation of the project at the conference.

KDD 2018 (24TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING) [19]

- **When:** 19-23 August
- **Where:** London (UK)
- **Short description:** With several thousands of visitors, KDD is one of the most well-known international conferences of the data community. The 2018 edition has counted on keynote speakers, invited talks, 27 workshops, 8 hands-on tutorials and 29 conventional tutorials.
- **Role of DataBench:** DataBench was one of the projects selected for the Project Showcase Track, which offered a full day agenda focused exclusively on innovative projects on Machine Learning and Data Analytics to show the state-of-the-art in research and applications in this field. The selected paper *"DataBench: Evidence Based Big Data Benchmarking to Improve Business Performance"* together with additional information will be made available for download on the DataBench website.

DataBench Priorities in H2 2018-H1 2019 (short-term plan)

ItAIS 2018: 15th conference of the Italian Chapter of AIS (Association for Information Systems) [20]

- **When:** 12-13 October
- **Where:** Pavia (Italy)
- **Short description:** the 2018 edition of ItAIS will run under the slogan "Living in the digital ecosystem: technologies, organizations and human agency".
- **Role of DataBench:** Presentation of accepted paper.

  Barbara Pernici, Chiara Francalanci, Angela Geronazzo, Lucia Polidori, Stefano Ray, Leonardo Riva, Arne Jørgen Barre and Todor Ivanov, "Big Data key performance indicators", accepted for presentation at the XV edition of the itAIS conference, Pavia on Oct 12th -13th 2018.

  The paper will be made available for download on the DataBench website.

European Big Data Value Forum (EBDVF 2018) [21]

- **When:** 12-14 November
- **Where:** Vienna (Austria)
- **Short description:** EBDVF is a key European event for industry professionals, business developers, researchers, and policy makers to discuss the challenges and

opportunities of the European data economy and data-driven innovation in Europe. This year, the event will be part of the EU Presidency agenda and will count on the presence of Commissionaire Gabriel, among other relevant speakers. The program will devote two days to keynotes and parallel sessions where many projects of the PPP will be represented. A third day will enable projects to interact in working workshops resulted from an Open Call for contributions.

- **Role of DataBench:** DataBench is one of the organizers of the session 1.8 on Big Data Benchmarking[2]. Two speakers from the project are already confirmed there. An additional workshop to collaborate with PPP projects was also proposed, but due to the acceptance of the Benchmarking session on the main program, the workshop was discarded. DataBench is waiting for instructions regarding the exhibition area in case we can get visibility through the PPP booth.

ICT Event 2018 (Imagine Digital - Connect Europe) [22]

- **When:** 4-6 December
- **Where:** Vienna
- **Short description:** ICT 2018 will focus on the European Union's priorities in the digital transformation of society and industry. It will present an opportunity for the people involved in this transformation to share their experience and vision of Europe in the digital age. As a general ICT event, it will feature many sessions and workshops focused on many different topics, some of which will for sure address data-related topics.
- **Role of DataBench:** DataBench submitted a proposal for a networking session that has not been successful. Nevertheless we will take advantage of the sessions proposed by the Big Data PPP together with other initiatives and platforms with the aim of exposing DataBench. Negotiations are ongoing with BDVe/BDVA to promote a speaker in networking session (1) of the following list of accepted proposals:

    1. "Impact of Data-Driven AI in business sectors" in collaboration with euRobotics/SPARC PPP, which will convene the 4 BDV PPP lighthouse projects;

    2. "Data democratisation: empowering the citizens in the digital transformation" in collaboration with MyData community, Open and Agile Smart Cities and the Living Labs (ENoLL);

    3. "Personal and machine-generated data: impact on privacy and security" in collaboration with ECSO;

    4. "Data Science Skills for Society: From Big Data to AI, robotics and beyond" done in collaboration with euRobotics/SPARC PPP;

    5. "Common priorities of HPC, Big Data and HiPEAC for post-H2020 era", result of the collaboration between ETP4HPC, BDVA and HiPEAC;

2 See https://www.european-big-data-value-forum.eu/program/

6. "Would you buy a second-hand car from this algorithm?" (Evolution of the Fate of AI, ellaborated by BDVA in collaboration with Robotics PPP).

A booth for the PPP has also been accepted. It will be the meeting place for PPP projects at the exhibition area. That is why it has been called "The BIG DATA VALUE Public Private Partnership village". DataBench will be part of the *village*.

Overall list of events of interest to DataBench objectives

The following table filters some of the events that have been selected as potential opportunities for DataBench. They will be revised on a continuous basis in order to confirm which ones will be attended by the project, actions needed to carry out the work and to update the list with emerging opportunities that we may have not considered so far. The list should therefore not be taken as a complete or stable one. Events pointed out above and already confirmed have not been included.

Event	When and where	Further information
CEN-CENELEC Stakeholder Workshop - Trustworthy Artificial Intelligence	Brussels (Belgium); 18 September	Standardization for take-up of AI technologies;
Data for AI	Brussels (Belgium); 19 September	The focus of this workshop, jointly organized between the EC, BDVA and the euRobotics platform is on identifying the main challenges to develop the European Data Space as essential component for AI;
12th European Conference on Software Architecture (ECSA)	Madrid (Spain); 10 October	Conference that brings together major experts on software architecture research and practice. One of the workshops, which will count on participation of BDVA is of special interest to us: "Workshop on Software Architecture Challenges in Big Data - SACBD@ECSA2018";
High Tech Summit	Copenhagen (Denmark); 10 October	A DTU[3] -powered event that is the largest research-based meeting place in Denmark within the fields of digitalization, robotics and artificial intelligence;
6th Global Summit on Artificial Intelligence and Neural Network	Helsinki (Finland); 15 October	Scientific conference addressing knowledge and sharing of new ideas amongst the professionals, industrialists, researchers, and students from research area of Artificial Intelligence;
IoT Solutions World Congress	Barcelona (Spain); 16 October	Major meeting point for people interested in IoT for Digital Transformation or Industrial IoT;

3 Technical University of Denmark

ODBASE 2018 - The 17th International Conference on Ontologies, DataBases, and Applications of Semantics	La Valetta (Malta); 16 October	Forum focused on the use of ontologies, rules and data semantics in novel applications;
BigSurv 2018: Exploring new statistical frontiers at the intersection of survey science and big data	Barcelona (Spain); 25-27 October	This event hosted by the European Survey Research Association (ESRA) addresses how researchers produce, analyze, and use statistics;

Table 2: List of potential events for DataBench

6. KPIs and Monitoring Framework

At proposal stage DataBench described some of the targets for marketing and communications. They will be used as starting point to monitor the performance of the project. They will be revisited during the project and will consider the different phases explained in Chapter 4, since expectations in numbers will not necessarily grow in a linear way. D6.3, as the deliverable that will report the activities developed by the project, will address this aspect in a more detailed way.

Activity	Indicator	Target KPI
Technical papers	Number of papers and number of conferences where papers are presented	At least 5 papers presented in at least 3 international Big Data conferences
Project website	N. of unique visitors to the website (average per year)	Min. 2000
Social media - Twitter	N. of followers New followers per year	Y1 300 +100
Social media - Facebook	N. of followers New followers per year	Y1 50 +100
Social media - LinkedIn	N. of followers New followers per year	Y1 100 +100
Social media - YouTube	N. of videos published N. of views	Min. 4 100 views per video
Social media - SlideShare	N. of overall views	200
Newsletter	N. of subscribers per year Number of newsletters	100 3 per year
Webinars	N. of webinars N. of participants per webinar	Min. 4 10

Table 3: Dissemination and communication KPIs

7. Conclusions

DataBench has the potential to become a key element of the Big Data Value PPP, since it will contribute to measure the real impact of the investments made by industry and the EC in this partnership. It will support the benchmarking of BDT, therefore helping to understand the progress with respect to the state of the art, but will also relate technical performance indicators with business indicators, establishing a bridge that has never been built before.

The challenges are however important. Among them, DataBench will require tight collaboration with a number of players and communities. Working hand-in-hand with benchmarking communities as well as with pilots and use cases (ideally coming from PPP projects) is key. They will be contributors but also validators of the so called DataBench ToolBox, main product to be produced by the project.

Along the three years duration of DataBench we will have to gain the credibility and recognition of those communities, since -in most cases- they will be the first adopters/users of the DataBench outcomes.

WP6 will support DataBench in creating the connections and engaging with those communities, will define and will implement a dissemination and communication strategy and will pave the path towards the self-sustainability of the results beyond the project duration. This deliverable describes the initial steps in that endeavour by sharing the Dissemination and Liaison Plan. The document identifies specifically: (1) target communities/audience, (2) Phases of the dissemination strategy, (3) tools, channels and mechanisms that will be used and (4) KPIs.

Among the target communities it is worth mentioning the PPP project portfolio, benchmarking communities, I-Spaces (data-driven experimentation environments) and networks of research or innovation structures that could well become potential users of the DataBench ToolBox (e.g. Big Data CoE, DIH).

Due to the submission of this document later than initially planned, we have taken advantage not only to identify actions but also to report some of the activities already carried out by the project in different areas: web strategy, publications, or visibility at events.

8. References

[1] D3.1 DataBench Architecture

[2] http://www.bdva.eu/

[3] Contractual arrangement: Setting-up a Public-Private Partnership in the area of Data between the Big Data Value Association and the European Commission (Brussels, 13 October 2014)

[4] Monitoring report Big Data PPP 2017

[5] Strategic Research and Innovation Agenda BDVA v4; http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

[6] https://www.sdil.de/en/

[7] https://www.teralab-datascience.fr/en/home/

[8] http://www.know-center.tugraz.at/

[9] https://www.sics.se/

[10] https://eurecat.org/es/eurecat/centros-de-excelencia/big-data-coe-barcelona/

[11] Big Data Labs Europe: http://magazin.know-center.tugraz.at/i-know-2016/program/big-data-labs-europe/

[12] http://s3platform.jrc.ec.europa.eu/digital-innovation-hubs-tool

[13] http://www.big-data-value.eu/

[14] BDVe project. D3.11 Report on Big data National and Regional Outreach

[15] https://icpe2018.spec.org/conference-program.html

[16] https://web.rniapps.net/pabs

[17] http://www.big-data-value.eu/big-data-value-meet-up-sofia/

[18] https://pgsolx.com/IT-Tech/BDIC/

[19] http://www.kdd.org/kdd2018/

[20] http://www.itais.org/conference/2018/

[21] http://www.european-big-data-value-forum.eu/

[22] https://ec.europa.eu/digital-single-market/en/events/ict-2018-imagine-digital-connect-europe

[23] BDVe project. D3.3. User Ecosystem Characterization

9. Annex I. State of play of sectors in the Big Data PPP [4] [23]

As part of the work developed for the monitoring report, BDVA, supported by BDVe, has gathered data to understand the current situation of activities associated to the deployment and evaluation of Big Data in concrete industrial sectors. Input comes essentially from two sources: the large scale pilots (also known as lighthouse projects) of the program and the sectorial working groups that are operating in BDVA (under what it is known Task Force 7 on Applications). The information provided here reflects the **current state of play and has been extracted directly from [4] with just minor changes. You can find this information directly in [23]**, which is offered here for context awareness.

Transport, mobility and logistics

Relevance of the Sector

Big Data will have profound economic and societal impact on mobility and logistics. **Mobility and logistics is one of the most-used industries in the world** - with a market size of about 1305 B€ contributing to approximately 15% of GDP and to employment of around 11.2 million persons in EU-28, i.e., some 5.0 % of the total workforce. The improvements in operational efficiency empowered by Big Data are expected to lead to 500 billion USD in value worldwide in the form of time and fuel savings, as well as savings of 380 megatons CO_2 in mobility and logistics. With freight transport activities projected to increase, with respect to 2005, by 40% in 2030 and by 80% in 2050, transforming the current mobility and logistics processes to become significantly more efficient, will have a profound impact. The logistics sector is ideally placed to benefit from Big Data technologies, as it already manages massive flow of goods and at the same time creates vast data sets - a **10% efficiency improvement will lead to EU cost savings of 100 B€**.

There are huge untapped opportunities for improving operational efficiency, delivering improved customer experience, and creating new business processes and business models, yet **only 19% of EU mobility and logistics companies currently employ Big Data solutions** as part of value creation and business processes. The mobility and logistics sector is ready to exploit Big Data solutions to significantly increase the EU market in the mobility and logistics sector.

As cited above, **mobility and logistics is one of the most-used industries in the world**, contributing significantly to GDP and employment. However, with total goods transport activities in EU-28 estimated to have amounted to **3,768 billion tonne-kilometres**[4] and **6,391 billion person-kilometres** (on average around 12,652 km per person), mobility and logistics is also a key contributor to CO_2 emissions with total greenhouse gas emission of 4,824 megatonnes CO_2. Therefore, any improvement in efficiency will have **profound impact on sustainability**. This means the mobility and logistics sector will hugely benefit from the introduction, adoption and large-scale take-up of Big Data solutions. As mentioned above, Big Data solutions are only used by a small fraction of EU mobility and logistics companies, thus presenting a huge opportunity for **increasing the market share** of Big Data solutions. In addition, the nature of the majority of mobility and logistics data does not imply major legal barriers to overcome before adaption for data-driven innovation and **immediate take up** of Big Data Value PPP outcomes. As a consequence, the mobility and

4 A tonne-kilometre is the product of transported mass in tonnes and the distance in km.

logistics sector - compared with other sectors (such as healthcare, finance, media, or telecommunications) - is optimally positioned and ready to deliver value from Big Data solutions.

Nature of Data Assets in the Sector

The data characteristics expected for the transport, mobility and logistics sector have been presented in a recent study, in which 69% of corporate executives named greater data variety as the most important factor, followed by volume (25%), with velocity (6%) trailing - indicating that the big opportunity lies in integrating more sources of data, not bigger amounts.

Compared with other sectors (such as health, finance, or telecommunications), the mobility and logistics sector **faces less legal barriers** for what concerns the use of data to deliver value; e.g., due to the fact that the many of the available data sources do not include personal-identifiable data. This means the mobility and logistics sector - compared with other sectors - is very well positioned and ready to take up Big Data solutions. Nevertheless, the European Data Protection Supervisor has identified general concerns related to personal data in the mobility and logistics domain, which should be addressed. Reasonable safeguards should be applied to protect the personal information from unauthorised access, loss, misuse, modification and disclosure.

As an example of the indicative volume, velocity and variety of data assets in the application domain we are referring to **average numbers computed from the 13 pilots of the ICT-15 Big Data PPP Lighthouse project Transforming Transport: 60,000 GB (volume), 24.5 GB/day (velocity), 22 different data sources.**

Impact

Mobility and logistics stakeholders are **already adopting technology and massively storing data**. Logistics providers already now manage a massive flow of goods and at the same time create vast data sets. The volume, velocity and variety of data generated today is unprecedented - it will fundamentally transform the sector. Today these additional data sources augment traditional sources of transport data collection, while in the future they will likely replace them. From the transport point of view, **vehicles are massively incorporating telematics systems** either as OEM (Open On-start, BMW connected drive, Mercedes Embrace) or as existing **plug-and-play and retrofitting solutions** available in the market (Openmatics, Geotab, Mobile Devices, TomTom, etc.).

Improvements in transport processes will have **significant societal impact**, in particular in terms of CO_2 emissions. It is estimated that a **10% efficiency improvement will lead to EU cost savings of 100 B€** and time and **fuel savings of 90 megatons CO_2** within the EU. In addition Big Data applied to transport will produce a safer mobility in the EU, contributing to reduce deaths in the transport sector (26.000 people die every year in the EU in traffic accidents).

The mobility and logistics sector **faces less legal barriers** for what concerns the use of data to deliver value; e.g., due to the fact that the many of the available data sources do not include personal-identifiable data. In transport and logistic sector, both technologies and datasets are much more harmonised and standardised contributing to much faster pan-EU implementation and adoption of big data solutions than other sectors in which datasets curation and country-by-country adaptation is needed. This project demonstrates through

pilot replications that solutions developed in this sector are immediately implemented and adopted by stakeholders in different countries and with different contexts.

The EU is pursuing the political aims (1) to reduce CO_2 emissions by 40% by 2030, (2) to reduce the number of deaths by car accidents by half by 2020 and (3) to support methods for more effective use of cars and automobile fleets. These goals stem from the circumstance that every year 26,000 people in the EU die in car accidents. 600 million tons of CO_2 are produced, and of that amount a major proportion is caused by poorly maintained vehicles and substandard traffic routing. Further, logistics is considered as one driver for the **Digital Single Market** (e.g., currently 62% of companies trying to sell online say that too-high parcel delivery costs are a barrier").

Bioeconomy

Relevance of the Sector

Launched and adopted on 13 February 2012, **Europe's Bioeconomy Strategy** addresses the production of renewable biological resources and their conversion into vital products and bio-energy. Experiences from US show that bioeconomy and specifically agriculture can get a significant boost from Big Data. In Europe, this sector has until now attracted few large ICT vendors. Nevertheless, the share of bioeconomy is remarkably large in the national economy in EU countries. The European bioeconomy is already worth more than **€2 trillion annually** and employs over **22 million people**, often in rural or coastal areas and in Small and Medium Sized Enterprises (SMEs).

Farm machines, fishing vessels and forestry machinery in use today collect large quantities of data in new and previously unimaginable ways. Remote and local sensors and imagery, and many other technologies, are all working together to give details about **soil content, marine environment, weeds and pests, sunlight and shade, and many other factors**. Analyzing these data can help the farmers, foresters and fishers to adjust their activities. The challenge is that **applying this data in practice requires a great deal of analysis and synthesis**, both data derived from the farm itself and from other sources. Furthermore, large data sets - such as those coming from the **Copernicus earth monitoring infrastructure** - are becoming more and more available on different levels of granularity, but they are heterogonous (in some cases also unstructured, hard to analyze and distributed across various sectors and different providers).

Nature of Data Assets in the Sector

The agriculture sector data assets include Sentinel-1/2 data, machinery monitoring and sensor measurements. As an example, *DataBio*, the Big Data PPP Lighthouse project focused on this sector, is planning agriculture pilot implementations expected to utilize **53TB volume and 197TB/year velocity of data**. The European Copernicus space program has currently launched its third Sentinel satellite whose data will be added to the above.

The forestry sector data assets include Sentinel-1/2 forest image data, national forest resource databases (like the Finish Metsään.fi database), forest canopy height models, drainage basin data and aerial image data. Planned forestry pilot implementations in *DataBio* are expected to utilize 12TB volume and 12TB/year velocity of data. In addition, aerial/UAV data offer hyperspectral remote sensing imaging with sample velocities of 100GB/h.

The fisheries sector data assets include ship operational and motion data, data from FAD and buoys, ship engine sensors, VMS, AIS, first sale notes from fisheries' auctions, big vessel activities as recorded by authories, earth observation and earth models. Planned fisheries pilot implementations in *DataBio* are expected to utilize 9TB volume and 7TB/year velocity of data.

In the **time scale**, Big Data experts provide common analytic technology support for the main common and typical Bioeconomy applications/analytics that are now emerging:

- Past: Managing and analysing data from the past - including many different kind of data sources, i.e. Descriptive analytics and classical query/reporting (in need of Variety management - and handling and analysis of all of the data from the past, including performance data, transactional data, attitudinal data, descriptive data, behavioural data, location-related data, interactional data, from many different sources).
- Present: Monitoring and real-time analytics - pilot services (in need of Velocity processing - and handling of real-time data from the present) - trigging alarms, actuators etc.
- Future: Forecasting, Prediction and Recommendation analytics - services (in need of Volume processing - and processing of large amounts of data combining knowledge from the past and present, and from models, to provide insight for the future).

Impact

Big Data technologies can especially contribute to improving the processes in production of best possible raw materials from agriculture, forestry and fishery for the bioeconomy industry to produce food, energy and biomaterials. Farm machines, fishing vessels, forestry machinery and remote and proximal sensors collect large quantities of data. Large scale data collection and collation enhances knowledge to increase performance and productivity in a sustainable way.

In his **Agenda for Jobs, Growth, Fairness and Democratic Change**, President Juncker identified 10 key priorities for the European Commission. The bioeconomy and ICT are **central to at least three of them**:

- New Boost for Jobs, Growth and Investment**.** The innovative bioeconomy is an important source of new jobs - especially at local and regional level, and in rural and coastal areas - and there are big opportunities for the growth of new markets, for example in bio-fuels, food and bio-based products
- Resilient Energy Union with a Forward-Looking Climate Change Policy**.** Europe needs to diversify its sources of energy and can support breakthroughs in low-carbon technologies with coordinated research. Replacing fossil raw materials with biological resources is an indispensable component of a forward-looking climate change policy.
- Deeper and Fairer Internal Market with a Strengthened Industrial Base**.** Innovative bio-based and food industries will contribute in raising the share of industry in GDP from 16% to 20% and to creating a circular, resource-efficient economy. The food and drink industry is already the largest manufacturing sector in the EU.

In addition, marine issues and food security are two aspects of the bioeconomy where Europe can and should lead the global agenda as part of President Juncker's strategy to make the EU a stronger global actor.

Agricultural productivity grows through bringing new resources into production (new land, extension of irrigation, or input intensification per hectare) or through raising the productivity of existing resources. The appropriate measure of productivity growth in this context is *Total Factor Productivity (TFP) growth*, which is defined as the aggregate quantity of outputs produced by the agricultural sector divided by the aggregate quantity of inputs used to produce those outputs. TFP growth in the EU-27 over the past decade was according to EU DG AGRI a disappointing 0.6% per annum (ibid). The USDA statistics showed higher numbers - an average growth of **2.5 %** varying between 4.4 % and -0.1 % (ibid). Big Data technologies are estimated to contribute over 30% productivity increase, which in 5 year perspective means more than a double annual growth rate (USDA numbers).

In **forestry**, the ratio of value added generated within the forestry and logging sector compared with the forest area available for wood supply is one indicator of the productivity. The indicator shows that in 2011 the highest shares of value generated per forest area in the EU were in Portugal (419 €/hectare) and the lowest in Greece (18 €/hectare). When weighted with the forest area, the annual growth for the EU countries with available data has during 2005 - 2012 been about 3.5 %. Productivity in certain dimensions (mainly cost) can be at least doubled from the current annual one. Some benefits will materialise over a longer time with annual increases under 1 %.

Regarding **fishery productivity**, it is a general trend that the production of the world fisheries has stagnated the latest decades. This is caused by the fact that most fish stocks are either overexploited or at its maximum yield, and for most species this is governed by the set quotas. To increase productivity, Big Data can contribute in the following areas:

- Better stock management: By improving the monitoring of the fish stocks, quotas can be better adapted to actual situations. This will decrease the risk of collapse, and it will help keep the stocks at the size where it is most productive, leading to estimations of 10% increase in productivity.
- Improved energy efficiency: The oil consumption in the fisheries depends partly on immediate operational choices, partly on vessel design and partly on planning of the fisheries. Better immediate decision making and planning could lead to 10% decrease of oil consumption.
- Improved market adaptation: The value of the fish is very dependent on both quality and timing. By helping the fishing vessel to catch the correct species at the correct time, a production increase of 5% is expected. By also improving quality through decreased vessel movements, another 2% increase in delivered value is expected.

In addition, **Big Data technology providers** have the opportunity to expand to **new markets with new products/services**, where, in some sectors, there is very small adaptation of Big Data services and products, offering thus opportunities for very large growths. For example, currently there are no Big Data based commercial services for agricultural professionals in several European countries and the number of Big Data users in fishery is negligible. In forestry, new Big Data products include forest health monitoring in near real time, species classification based on temporal behaviour of each species, or illegal logging detection.

Even though the DataBio project is the most prominent example of activities in this sector in the PPP, other projects tackle sub-sectors or specific use cases in bio-domains, as it is the case of BigDataGrapes, and to some extent BigDataOcean.

Smart Manufacturing Industry

Relevance of the Sector

The Manufacturing sector contributes per se for more than 15% to the overall GDP of EU28, additionally mobilizing huge resources from other sectors. Industry 4.0 was born in Germany in 2011 as the fourth industrial revolution, driven by the adoption in production of the so-called Cyber Physical Systems, but very soon became the flag for any ICT-driven modernization of manufacturing industry. The EC Communication "Digitising EU Industry" of April 19th 2016, identified **three main groups of technologies implementing Industry 4.0.namely Internet of Things, Big Data and Artificial Intelligence** - Deep Learning, all of them concurring in synergy to the development of new smart products (digital inside), new digitized processes and new ICT-enabled business models. In almost all EU Countries and Regions (e.g. the Smart Specialisation Strategy on Advanced Manufacturing and the Vanguard Initiative on Efficient and Sustainable Manufacturing), initiatives have been developed to allow SMEs to access technology and knowledge and to stimulate the adoption of Industry 4.0 principles.

Nature of Data Assets

Scenarios in the manufacturing industry can be grouped in three main categories. **Smart Factory** scenarios are related to the Real World of production systems and plants along with their interaction with the Digital World (Digital Twin). Data are mostly generated from the field by production systems (e.g. robots, machine tools, production lines, conveyors, workplaces, sensors) and generally need to be processed at high speed, in order to take the proper factory management decisions. Typical applications concern the optimization of the production, the management of waste and energy, the zero defect quality of products, the diagnosis and predictive maintenance as well as the safety and wellbeing at the workplace for Blue Collar Workers. **Smart Lifecycle** scenarios are related to the product lifecycle and its phases of ideation, design, manufacturing, operations, maintenance and end-of-life. Structured Data are often enriched by sentiments and opinions sometimes also gathered from social networks. Usually, real time processing performance is not required, but the amounts of data are often very large and in specific cases require specialized computing architectures and resources (HPC). Typical applications concern the modelling and simulation of product properties, the design of new product-services, the implementation of post-sales services such as training and maintenance and the support to new business models dictated by Sharing and Circular Economy paradigms. **Smart Supply Chain** scenarios are related to the digitalization of the value chain and involve different stakeholders such as suppliers, distributors, retailers, all linked to the manufacturers. Data to be interoperated are not just huge, but also very heterogeneous and could concern cross-border legal and policy issues. Semantic technologies are often in use to achieve interoperability. Ownership, confidentiality and privacy concerns can also be relevant in these scenarios. Typical applications concern the optimization of the various tiers of supply chain, the creation of non-hierarchical business ecosystems, the management of distribution and logistics processes, the synchronization of online and physical retail, the development

of proper Sales & Operations replenishment policies and the improvement of consumers' experience at the point of sale.

Impact

In its recent "The European Data Monitoring Tool", IDC reports a considerable expansion of the EU Data Market (60 bn EUR) also in relations with the total ICT market (620 bn EUR). Three future scenarios have been developed (High Growth, Baseline, Challenge) bringing the current 60 bn EUR to 107-80-70 bn EUR in 2020 respectively. **Manufacturing is always considered in top position regarding the current situation and future projections of the Data Market**. In the current 60 bn EUR scenario, Manufacturing is leading the group of applications by 12.8 bn EUR, with Financial and Professional services just behind. In the future Baseline scenario 2020 (80 bn EUR) Manufacturing is registering an important jump to 17.3 bn EUR by consolidating its leading position with respect to Finance and Professional Services. There is therefore no doubt that Industry 4.0 and industrial modernization are opening interesting new business opportunities in the Data Market, also in the most prudent projections.

Activities on Smart Manufacturing Industry in the PPP are very dynamic, with a group that holds periodic meetings, workshops, keeps a direct relationship with EFFRA (with whom BDVA has signed a collaboration agreement) and has produced a white paper). A "Connected Smart Factory" input has also been delivered to the Digitising European Industry WG2 on Digital Platforms, as example of dissemination and awareness to communities beyond the BDVA. Activities in this community count on the direct participation of the lighthouse project BOOST4.0.

A glimpse on the SMI group history

The roadmap of the Smart Manufacturing Industry group in 2016 followed 4 major steps: (1) engage an internal critical mass of members participating to the topic (15-20 attendees at AG breakout sessions and more than 60 emails in the list) (2) get the approval from the Board of Directors as official subgroup of TF7 applications, (3) agree among the strategy and rationale for a position paper and (4) run an effective interactive workshop at the Summit in order to collect inputs and get consensus from the community on the R&I topics with the final aim to elaborate a position paper.

The activities of the group started in summer 2016 with a series of breakout sessions at the Activity Group Meetings in September and October. The successful outcomes of these sessions both in terms of attendance and of expressions of interest resulted to the BDVA BoD approval for a TF7 subgroup.

The newly established subgroup organized a session at the Valencia BDVA Summit on December 1st 2016 with two subsequent slots of 90 minutes each. The first part, attended by more than 70 people, was dedicated to set the context among all the participants, providing presentations on concrete experiences reported by BDVA members, industrial visions and investments and relevant references coming from other association (e.g. EFFRA). The second session, attended by nearly 50 people and leveraging on the interaction and collaboration among attendees, has identified the most urgent R&I challenges for fully adopting Big Data technologies in Manufacturing Industry. Attendants were split in three groups, addressing in turn the identified Manufacturing Industry Grand Challenges: (1) Smart Factory (factory automation and workplace interaction scenarios), (2) Smart Product Lifecycle (Beginning, Middle and End of Life scenarios) and (3) Smart Business Ecosystems (Supply and Distribution chain scenarios). Each one of these three scenarios was analyzed in relation to the technical priorities defined in the BDVA SRIA. As a conclusion of the group work, three selected rapporteurs summarized and exposed the main findings of the whole attendance. The complete group work done on posters with post-its was captured in pictures, reported in a document and stored in the BDVA internal document repository as common resource and predominantly as the driving force for the position paper.

Healthcare

Relevance of the Sector

The healthcare sector currently accounts for 8% of the total European workforce and for 10% of the EU's GDP. However, public expenditure on healthcare and long-term care is expected to increase by one third by 2060. This is primarily due to a rapidly aging population, rising prevalence of chronic diseases and costly developments in medical technology.

Big Data technologies have already made impact in fields related to healthcare: medical diagnosis from imaging data in medicine, quantifying lifestyle data in the fitness industry, etc. Nevertheless, the healthcare has been lagging in taking up the Big Data technologies due to several challenges.

Nature of Data Assets

There are mainly **four key verticals generating vast amounts of data within the healthcare sector**: *Healthcare providers* (e.g. hospitals, GP, laboratories), *Insurance*, *Pharmaceuticals* and the *HealthTech* sectors. Over the past decade, each vertical has started analysing its own data sets in order to derive insights which are subsequently used to improve the quality of their product offerings. Such siloed approaches to deriving insights only from data generated by the vertical itself, places a limit on the ability to introduce innovations that can make a genuine difference. This has a serious impact on the ability to meet the ever increasing demands of the healthcare sector. Therefore, collecting, integrating and analysing health data is a complex and challenging task due to lack of interoperability and harmonization of data formats, processing techniques, data storages and transfers and different legal frameworks. As a result, deriving insights and value from the aggregation of these datasets is not possible at the moment.

In healthcare, different types of information are available from different sources such as electronic health care records, patient summaries, genomic and pharmaceutical data, clinical test results, imaging (e.g. x-ray, MRI, etc.), insurance claims, vital signs from e.g., telemedicine, mobile apps, home monitoring, on-going clinical trials, real-time sensor data, and information on wellbeing, behaviour and socioeconomic indicators. This data can be both structured and unstructured. Analysing health data coming from the different sources is challenging. On the other hand, health data presents valuable opportunities. Better clinical outcomes, more tailored therapeutic responses, and disease management with improved quality of life are all appealing aspects of data usage in health.

However, because of the personal and sensitive nature of health data, special attention needs to be paid to legal and ethical aspects concerning privacy. To unlock its potential, health (and genomic) data sharing, with all the challenges it presents, is often necessary to ensure such endeavours are undertaken responsibly.

Achieving such a vision which involves the integration of such disparate healthcare datasets (in terms of data granularity, quality, type (e.g. ranging from free text, images, (streaming) sensor data to structured datasets) poses major legal, business and technical challenges from a data perspective, in terms of the volume, variety, veracity and velocity of the data sets. All these aspects lead to the need for new algorithms, techniques and approaches to handle these challenges. Besides, new or consolidated standards are a necessity for solving the problems related to data exchange and interoperability.

As a result, Big Data must not only be addressed from a technology perspective, but also focus on the legal and ethical issues that come into play when sensitive healthcare data needs to be shared beyond the traditional boundaries of the healthcare provider or across international borders. Furthermore the whole end-to-end value chain in the health continuum needs to be covered starting from Prevention, to Diagnosis, Treatment and eventually Home care where the data generated not only within the boundaries of the hospital but also in primary care settings and by patients themselves will be utilized.

Impact

The key challenges that are relevant to be tackled in order to achieve a significant impact in the Healthcare sector are:

- Bringing together the key players within the Healthcare sector: Healthcare providers and HealthTech, Pharmaceutical and Insurance industries.
- Integrating data value chains across Healthcare providers, HealthTech, Pharmaceutical and Insurance industries to derive new, actionable insights to achieve breakthrough quality, access and yet reduce costs.
- Integrating heterogeneous data sources with very different characteristics in terms of Volume, Variety, Velocity, Veracity and Value, e.g. data from Primary and secondary care, EMR, Laboratory data, Prescription, Imaging, Patient feedback, Real-time location tracking, Patient monitors (both hospital and home-based), etc.
- Bridging the language gap inherent in EMR systems across Europe in order to have a unified approach to perform Big Data analytics on healthcare datasets spread across multiple European countries.
- Coping with different legal and ethical frameworks across Europe so that the Big Data concepts developed in a lighthouse project can be rolled out across Europe.

The fusion of healthcare data from multiple sources could take advantage of existing synergies between data to improve clinical decisions and to reveal entirely new approaches to treating diseases and keep citizens healthy. Insights derived from such data generated by the linking among EMR data, vital data, laboratory data, medication information, symptoms (to mention some of these) and their aggregation, even more with doctor notes, patient discharge letters, patient diaries, medical publications, namely linking structured with unstructured data, can be crucial to design coaching programmes that would help improve peoples' lifestyles and eventually reduce incidences of chronic disease, medication and hospitalization.

Evidence suggests that by improving the productivity of the health care system, public spending savings would be large, approaching 2% of GDP on average in the OECD which would be equivalent to €330 billion in Europe based on GDP figures for 2014. The Big Data technologies have the potential to unlock vast productivity bottlenecks and radically improve the quality and accessibility of the healthcare system by disrupting the **Iron Triangle of Healthcare** and allowing simultaneous optimization of all its components - quality, access and cost, which is impossible at the moment.

The activities regarding this sector are driven in the PPP by BDVA TF.7 Healthcare subgroup and have resulted in a whitepaper entitled "Big Data Technologies in Healthcare". The process has matchmaking events and workshops in different events. BDVe has had a supportive role to the current structures so far without an especially active approach in this domain. However, the fact that a lighthouse project is running since the beginning of 2018

(BigMedilytics) will help to establish additional synergies with other projects that hold use cases in this sector and communities beyond the PPP.

Telecommunications

Relevance of the Sector

Big Data is commonly recognized as a fundamental game changing set of technologies across many industrial sectors and societal solutions. The biggest challenge for Big Data is to overcome the fragmentation of the market and to achieve cross-fertilization between many established silos. Telco operators could serve as platform bringing different other sectors together to exchange data and build new business models. This supports the monetization of data, the digital transformation of vertical industries and the fostering of entrepreneurship of data. Upcoming technologies like Edge Computing support this transformation, by providing real time information to verticals at the edge (vertical edge intelligence), supporting use cases like autonomous driving, industrial internet, transport/logistics and health care.

In the telecommunication sector technologies are evolving much faster than previously based on softwarization (e.g. SDN, NFV, 5G, IoT, Industrial Internet, Edge Computing, Cloud). This has as well impact on the business/operational side and plays a major role to support the transformation of telco operators towards new business opportunities.

Nature of Data Assets in the Sector

Telecom operators hold large amounts of data in a variety of locations - central and distributed databases, as well as OSS, BSS, and CRM systems complemented by data collected thru verticals (e.g. IoT, Industry 4.0). The data has great value in terms of cross-fertilization and analytics, which by far exceeds the value currently assigned to the data by itself. Given the fact that data is a key asset, the telecommunication sector could leverage on its unique position with respect to the ever increasing amount of data sources. Taking the legal frame work into account (accountability, transparency, privacy,…) telecom operators can provide the right technological platforms and techniques to ensure its enforcement and support the concept of Personal Data Spaces.

Impact

Technical challenges with respect to data analytics in the telecommunication sector are unsupervised learning, deep learning for networks (to solve difficult tasks in the network), reinforcement learning, predictive analytics, cognitive knowledge (generation of value, i.e. knowledge, from data), real-time processing of large amounts of data (generating new knowledge, not previously available), and confidential data processing in e.g. cloud environments. This underlines the impact of Big Data and Data Analytics from a technological angle, while it impacts as well business models today and in the future.

PPP activities in this sector are basically led by the subgroup on telecommunications (under TF7), which has organized several workshops in the context of PPP event and is in the process of elaborating a position paper. Some interactions have been produced with the 5G PPP; however, with little or no impact and the activity has not been prioritized so far.

Media

Relevance of the Sector

The Big Data trend has impacted all industries, including the media industry, as new technologies are being developed to automate and simplify the process of data analysis, So, Big data is no longer a buzzword in the media context, but it must be considered as a mainstream technology especially for the content and news industry. In addition, **media sector provides enormous data resources** and the European contribution is of leading role in this amount of data.

The notion of media industry in this context covers a full range of different industries active in the Media Market: from TV channels, to newspapers, from Large Broadcasters to SMEs. The synergy between Data Science technologies and media applications is one of the most promising dimension of ICT applications, this is particularly true for SMEs. This synergy can bring together competencies from industries and research domains, with a strong cross disciplinary or intercultural dimension and a pervasive impact on the full value chain of media companies, namely: production, distribution, personalisation services and advertising. In addition, a better and more integrated technology usage inside news, media and content producer is becoming urgent to protect our democracy from fast growing of populists and fake news distribution. Despite the large gap in the media industries in term of Big Data competencies, according to Financial Times, half of leaders in the sector also see alliances - particularly with technology partners - as a strategic priority. Some of the specific competencies that will be impacted are (1) video and audio analytics, (2) recommendation systems, (3) targeting advertising and (4) audience analytics and screening.

Big Data and Media means also the birth of new professional specialty, such as **Data Journalist.** Data journalism is a journalism specialty reflecting the increased role of numerical data in the production of information in the digital era. It reflects the increased interaction between content producers (journalist) and several other fields such as design, computer science and statistics.

Nature of Data Assets in the Sector

The main data sources of the media industry cover media content assets (videos, photos, text and graphics), user behaviour data, networked data, customer data and many other data sources. Media content and data represent about the 70% of the Internet's stored and shared data, which is growing exponentially. However, the interlinking of these data sources is currently only emerging and the full potential for exploitation of all these data sources is not yet in place. **The diversity of European content sources has a huge potential.** However, the fragmentation of European media organisations prevents the full exploitation of the European data sources versus the monopolistic approach of Large USA Web companies, such as Google or Facebook. The media related data sources are stored and managed in many different data silos across Europe.

Impact

The integration of a large set of diverse and new information can provide, indeed, relevant insights to aid organisations in tackling world's hardest social problems. However, until now the societal impact of big data has been affected by incomplete, low quality, and even

incorrect data[5]. By allowing the fusing the data from disparate sources of information, Big Data and Media can empower journalists and citizens to comprehend an issue in very complete ways and protect them from the fake news and "post-true" effects.

The use of Big Data in media can impact the social good given the peculiar role media industry plays at a societal level at the crossroads of multiple sets of data. Of course, data has to be collected in ways that match our value systems and respect ethics, privacy, and informed consent. Big Data applied to Media has an enormous potential to improve society - much like other fundamental discoveries and technology advancement, it can give us a much more accurate and timelier understanding of how our society works, so that decisions are based on actual facts rather than doubtful hunches.

It is important that public media uses and shows results from working with "Big Data" - to raise expectations of the general public across the whole of Europe of how data can be used. While other areas such as education, health and traffic are sure to change if data were used, media is the area where things are seen by many, not just few. The impact of the "Panama papers" published in early April shows to some extent how the application of Big Data analytics and data capabilities from raw data to understandable narration is utterly important.

Activities in the Media sector in the PPP are driven by the corresponding subgroup of BDVA. Activity has been limited so far even though some community building workshops have been held and discussions have involved Media players. Connections with the NEM community are in place and we forecast increasing activity in the coming period as a result of the development of related projects in the program, such as Fandango (unfortunately some of the activities of the second wave of PPP projects cannot be reported because they have all started in the first half of 2018).

5 World Bank has identified the ability to produce, capture, communicate and analyze big data as the fundamental action to drive new forms of growth and social development, unlocking its data through its Open Data Initiative.