Deliverable 1.5
Annual Public Report II

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date of Deliverable</td>
<td>M24, 31/12/2016</td>
</tr>
<tr>
<td>Actual Submission Date</td>
<td>M25, 31/01/2017</td>
</tr>
<tr>
<td>Work Package</td>
<td>WP1</td>
</tr>
<tr>
<td>Task</td>
<td>T1.3</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Approval Status</td>
<td>Approved</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of Pages</td>
<td>30</td>
</tr>
<tr>
<td>Filename</td>
<td>D1.5_Annual_Public_Report_II.pdf</td>
</tr>
</tbody>
</table>

**Abstract:** This deliverable reports on the project’s progress, targeting the general public. It will focus on the impact of the conducted work across Europe’s seven Societal Challenges, exemplifying the contribution and importance of Big Data Europe towards facing these challenges.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.

Project funded by the European Union’s Horizon 2020 Research and Innovation Programme (2014 – 2020)
History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason</th>
<th>Revised by</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>19/12/16</td>
<td>Draft</td>
<td>Nadine Jochimsen (FhG)</td>
</tr>
<tr>
<td>0.1</td>
<td>20/12/16</td>
<td>Updating SC4</td>
<td>Andrea Toth</td>
</tr>
<tr>
<td>0.2</td>
<td>20/12/16</td>
<td>Updating section 2</td>
<td>Ivan Ermilov</td>
</tr>
<tr>
<td>0.3</td>
<td>21/12/16</td>
<td>Updating SC4</td>
<td>Luigi Selmi</td>
</tr>
<tr>
<td>0.4</td>
<td>23/12/16</td>
<td>Updating SC7 (final)</td>
<td>Michele Lazzarini</td>
</tr>
<tr>
<td>0.5</td>
<td>02/01/17</td>
<td>Updating SC3</td>
<td>Fragiskos Mouzakis</td>
</tr>
<tr>
<td>0.6</td>
<td>03/01/17</td>
<td>Updating SC2 (final)</td>
<td>Valeria Pesce (FAO), Panagiotis Zervas (AgroKnow)</td>
</tr>
<tr>
<td>0.7</td>
<td>12/01/17</td>
<td>Updated section on SC5</td>
<td>Iraklis Klampanos (NCSR), Andreas Ikonomopoulos (NCSR), Spyros Andronopoulos (NCSR)</td>
</tr>
<tr>
<td>0.8</td>
<td>16/01/17</td>
<td>Finalized section 2</td>
<td>Ivan Ermilov, Aad Versteden</td>
</tr>
<tr>
<td>0.9</td>
<td>23/1/17</td>
<td>Added SC1 section</td>
<td>Kiera McNeice</td>
</tr>
<tr>
<td>1.0</td>
<td>08/02/17</td>
<td>Updated dissemination activities</td>
<td>Phil Archer</td>
</tr>
</tbody>
</table>

Author List

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Name</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>FhG</td>
<td>Nadine Jochimsen</td>
<td><a href="mailto:Nadine.Jochimsen@iais.fraunhofer.de">Nadine.Jochimsen@iais.fraunhofer.de</a></td>
</tr>
<tr>
<td>FhG</td>
<td>Simon Scerri</td>
<td><a href="mailto:scerri@iai.uni-bonn.de">scerri@iai.uni-bonn.de</a></td>
</tr>
<tr>
<td>FAO</td>
<td>Valeria Pesce</td>
<td><a href="mailto:valeria.pesce@fao.org">valeria.pesce@fao.org</a></td>
</tr>
<tr>
<td>CRES</td>
<td>Fragiskos Mouzakis</td>
<td><a href="mailto:mouzakis@cres.gr">mouzakis@cres.gr</a></td>
</tr>
<tr>
<td>Ertico</td>
<td>Andrea Toth</td>
<td><a href="mailto:a.toth@mail.ertico.com">a.toth@mail.ertico.com</a></td>
</tr>
<tr>
<td>Organization</td>
<td>Name</td>
<td>Email</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>InfAI</td>
<td>Ivan Ermilov</td>
<td><a href="mailto:iermilov@informatik.uni-leipzig.de">iermilov@informatik.uni-leipzig.de</a></td>
</tr>
<tr>
<td>SatCen</td>
<td>Michele Lazzarini</td>
<td><a href="mailto:Michele.Lazzarini@satcen.europa.eu">Michele.Lazzarini@satcen.europa.eu</a></td>
</tr>
<tr>
<td>TF</td>
<td>Aad Versteden</td>
<td><a href="mailto:aad.versteden@tenforce.com">aad.versteden@tenforce.com</a></td>
</tr>
<tr>
<td>FhG</td>
<td>Luigi Selmi</td>
<td><a href="mailto:luigi.selmi@iais.fraunhofer.de">luigi.selmi@iais.fraunhofer.de</a></td>
</tr>
<tr>
<td>NCSRD</td>
<td>Iraklis Klampanos</td>
<td><a href="mailto:iaklampanos@iit.demokritos.gr">iaklampanos@iit.demokritos.gr</a></td>
</tr>
<tr>
<td>NCSRD</td>
<td>Andreas Ikonomopoulos</td>
<td><a href="mailto:anikon@ipta.demokritos.gr">anikon@ipta.demokritos.gr</a></td>
</tr>
<tr>
<td>NCSRD</td>
<td>Spyros Andronopoulos</td>
<td><a href="mailto:sandron@ipta.demokritos.gr">sandron@ipta.demokritos.gr</a></td>
</tr>
<tr>
<td>Open PHACTS</td>
<td>Kiera McNeice</td>
<td><a href="mailto:kiera@openphactsfoundation.org">kiera@openphactsfoundation.org</a></td>
</tr>
<tr>
<td>Foundation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W3C/ERCIM</td>
<td>Phil Archer</td>
<td><a href="mailto:phila@w3.org">phila@w3.org</a></td>
</tr>
</tbody>
</table>
Executive Summary

The second deliverable in this series provides an overview of the main project results accomplished in 2016 by the BigDataEurope (BDE) consortium. Collaborating within the EU Horizon 2020 project entitled “Integrating Big Data, Software and Communities Addressing Europe’s Societal Challenges”, the CSA’s primary objective was to build and maintain seven communities revolving around the societal challenges and to support them by developing an adaptable and ready-to-use platform that facilitates big data usage across the seven targeted societal sectors: Health, Food & Agriculture, Energy, Transport, Climate, Social Sciences and Security.

In the second year of the project, the coordination of the seven communities continued; extending the stakeholder network to cover the whole process of data usage within each, from data collection, processing, storage and visualisation to the development of data services. A second series of workshops and various webinars and hangouts throughout 2016 have brought together these communities to serve as a base for discussion and further understand their big data management needs, and to disseminate project results so far.

The support provided to the communities started to become concrete in 2016 with the release of the first public release of the Big Data Integrator Platform, and the demonstration of its value through the initial societal pilots. The platform consists of the intended integrated stack of tools that can be installed and used freely in a customized data processing chain with minimal knowledge of the technologies involved, and is based on the stakeholder requirements that were collected in the first year of the project.

The main technical results achieved in the project’s second year are the realisation of the architectural Big Data integrator platform blueprint into seven generic instances applicable to each societal domain, the release of a generic domain-independent platform, and the development of the first phase of the seven pilots. The generic instances were based on the domain requirements elicited in the project’s first year. The pilots showcase the use of these instances to address seven real-world use-cases.

Project efforts in the third and last year of the project will concentrate on maximising impact through i) intensified dissemination activities to promote the integrator platform and components and promote its take-up by stakeholders, and ii) further demonstrate its value through the showcasing of the second and third phase of the seven pilots. The dissemination will be carried out in the yearly workshops and monthly webinars and hangouts, through the participation of consortium members in other relevant venues, and through direct correspondence with stakeholders who have expressed interest in using the BDE technology for their needs.
The reported activities in this deliverable are structured along two sections. The first summarizes the various networking activities that were carried out within each of the seven societal challenge communities, explaining the efforts performed in 2016 and the next steps planned. The second section describes advances made by the technical team during the second 12-month period and outline the plans for the last year of the project. This report is public and shall be made available for unrestricted download on the BDE webpage http://www.big-data-europe.eu/results/. 
## Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDE</td>
<td>Big Data Europe</td>
</tr>
<tr>
<td>SC</td>
<td>Societal Challenge</td>
</tr>
<tr>
<td>H2020</td>
<td>Horizon 2020</td>
</tr>
</tbody>
</table>
Table of Contents

1. Domain Summaries for the Second Year (2016) ................................................................. 9
   1.1. Societal Challenge 1 - Health, Demographic Change and Wellbeing ......................... 9
       1.1.1. General Objectives .................................................................................................. 9
       1.1.2. Pilot Use Case ....................................................................................................... 9
       1.1.3. Stakeholder Workshop .......................................................................................... 10
       1.1.4. Engagement, Outreach, Dissemination .................................................................. 11
       1.1.5. Next Steps ........................................................................................................... 11
   1.2. Societal Challenge 2 - Food, Agriculture, Forestry, Water and Bioeconomy ............ 11
       1.2.1. General Objectives ............................................................................................... 11
       1.2.2. Pilot Use Case ....................................................................................................... 12
       1.2.3. Stakeholder Workshop .......................................................................................... 13
       1.2.4. Engagement, Outreach, Dissemination .................................................................. 13
       1.2.5. Next steps ............................................................................................................ 14
   1.3. Societal Challenge 3 - Secure, Clean and Efficient Energy ........................................... 14
       1.3.1. General Objectives ............................................................................................... 14
       1.3.2. Pilot Use Case ....................................................................................................... 14
       1.3.3. Stakeholder Workshop .......................................................................................... 15
       1.3.4. Engagement, Outreach, Dissemination .................................................................. 15
       1.3.5. Next Steps ........................................................................................................... 15
   1.4. Societal Challenge 4 - Smart, Green and Integrated Transport .................................... 16
       1.4.1. General Objectives ............................................................................................... 16
       1.4.2. Pilot Use Case ....................................................................................................... 16
       1.4.3. Stakeholder Workshop .......................................................................................... 18
       1.4.4. Engagement, Outreach, Dissemination .................................................................. 19
       1.4.5. Next Steps ........................................................................................................... 19
   1.5. Societal Challenge 5 - Climate, Environment, Resource Efficiency and Raw Materials . 20
       1.5.1. General Objectives ............................................................................................... 20
       1.5.2. Pilot Use Case ....................................................................................................... 20
1.5.3. Stakeholder Workshop ........................................................................................................... 21
1.5.4. Next Steps .......................................................................................................................... 22
1.6. Societal Challenge 6 - Inclusive, Innovative and Reflective Societies ............................... 22
  1.6.1. General Objectives .......................................................................................................... 22
  1.6.2. Pilot Use Case ................................................................................................................ 22
  1.6.3. Stakeholder Workshop ..................................................................................................... 23
  1.6.4. Engagement, Outreach, Dissemination ............................................................................. 23
  1.6.5. Next Steps ..................................................................................................................... 24
1.7. Societal Challenge 7 - Secure Societies ........................................................................... 24
  1.7.1. General Objectives .......................................................................................................... 25
  1.7.2. Pilot Use Case ................................................................................................................ 25
  1.7.3. Stakeholder Workshop ..................................................................................................... 25
  1.7.4. Engagement, Outreach, Dissemination ............................................................................. 25
  1.7.5. Next Steps ..................................................................................................................... 26

2. Summary of Platform Development and Pilot Deployment in the Second Year (2016) .... 26
3. Dissemination ......................................................................................................................... 28
1. Domain Summaries for the Second Year (2016)

1.1. Societal Challenge 1 - Health, Demographic Change and Wellbeing

1.1.1. General Objectives
In the second year of the project, SC1 focused on implementing and evaluating the first cycle of the Pilot, and continued engaging with stakeholders, in particular to consider how the pilot might be expanded.

1.1.2. Pilot Use Case
The first cycle of the Pilot Use Case has been completed, with VU Amsterdam joining the project to contribute to implementation of the Pilot. The goal of SC1 Pilot Cycle 1 was to replicate the functionality of the Open PHACTS Discovery Platform on BDE infrastructure, replacing the commercial components of the Open PHACTS Platform with open source alternatives. This would make it possible for any user to deploy an instance of the Open PHACTS Platform in-house, an important consideration for security-conscious pharmaceutical researchers. And as the Pilot's technology is domain-agnostic, it would open up the possibility of connecting to data from other domains, for example SC2.

All requirements for the first Cycle were met satisfactorily. The Open PHACTS Discovery Platform can now be installed using Docker modules on a high-end consumer PC in around 4 hours, with functional RDF storage and API access. This has been achieved using the open-source version of Virtuoso software, replacing the commercial Virtuoso software used in the existing Open PHACTS Platform. The Open PHACTS Docker modules are compatible with BDE allowing for custom integrated solutions; the IMS has been set up as an independent Docker module, allowing for customisation of various parameters that deal with communication between components; and running an entire Open PHACTS system on the same local area network reduces delays compared to the existing Open PHACTS Platform.
Using the open source Virtuoso RDF Docker component, the Pilot can answer a total of 18 of the original 21 use cases\(^1\) for the Open PHACTS Discovery Platform. The remaining three cannot be answered without access to non-open datasets.

The plan for Cycle 2 of the Pilot Use Case is to look into broadening the applicability of the Pilot beyond drug discovery. For example, in SC2 (Food Security and Sustainable Agriculture), the effects of chemistry on biology are an important consideration, and could exploit the linked data services currently within the Open PHACTS Platform. This will depend on the specifics of the SC2 Pilot Use Case and its data. Cycle 2 will also involve a complete refresh of the datasets integrated into the Pilot, homogenising and integrating new data available for these datasets, and simplifying existing workflows for querying the API for example via software tools such as KNIME. In Cycle 1 the Virtuoso component has also been successfully exchanged for 4Store; Cycle 2 will include experimenting with this and other RDF stores for the Pilot, and testing the feasibility of using SANSA stack as an alternative for SPARQL query processing.

1.1.3. Stakeholder Workshop

The second SC1 Stakeholder Workshop was held in December 2016 at KoWi in Brussels. The aim of the workshop was to demonstrate the progress made in Cycle 1 of the Pilot Use Case; to gather input about possible ways to expand on the Pilot; and to solicit the input of representatives of the European Commission in the domain of SC1 (Health, Demographic Change and Wellbeing), about how the SC1 Pilot might engage with other Horizon 2020 instruments and programmes to increase the pace of knowledge sharing within this domain. Attendees came from a variety of backgrounds related to big data in health.

Full details of the workshop will be published in the workshop report. An early point to emerge from the discussions was that data variety remains a major challenge in this domain; this is being addressed in the Open PHACTS platform by focussing on semantic data integration, driven by real-life research use cases. Further discussion was guided by three questions the attendees were asked to discuss:

- Which projects should we be engaging with that could benefit from a big data platform?
- What pilot use cases can you foresee in your area?
- Which areas are a future focus for new collaborations?

Several projects were suggested for SC1 to reach out to and discuss possible collaborations. This was followed by a long discussion about the potential benefits and challenges of integrating new kinds of data into the Pilot, from patient clinical data, to clinical trial data, to data gathered via social media.

Finally several options for collaboration across other Societal Challenges were suggested. All in all the

\(^1\)http://www.openphacts.org/documents/registered/deliverables/D%206.1%20Prioritised%20Research%20Questions_final%20version.pdf
second Stakeholder Workshop provided valuable insights on potential directions for both technical development and outreach in the next year of the project.

1.1.4. Engagement, Outreach, Dissemination
As well as hosting the second Stakeholder Workshop, SC1 has been engaged in several other outreach activities this year.

SC1 was represented at the joint webinar on Technical Insights into the BDE Project in March 2016, and hosted a dedicated webinar about the Open PHACTS Pilot Use Case in July 2016. The latter was presented by representatives from VU Amsterdam, the Open PHACTS Foundation, TenForce, and BioExcel; it gave an overview of the purpose and infrastructure of the Pilot Use Case, as well as a detailed look at the generic components of the Pilot, and how to set up and use them. The webinar was aimed at people with an interest in big data and drug discovery who are acquainted with basic development tools, and allowed time for questions and discussion to engage with attendees. The full webinar is available on YouTube.²

1.1.5. Next Steps
In terms of specific outreach to other projects, SC1 has had initial discussions with OpenTrialsFDA,³ connected to the OpenTrials initiative which aims to locate, match, and share data on trials conducted on medicines and other treatments. Discussions centred around what the BDE infrastructure could do in this area, and how it might ultimately be linked to OpenTrialsFDA data, once that data is available.

1.2. Societal Challenge 2 - Food, Agriculture, Forestry, Water and Bioeconomy

1.2.1. General Objectives
Activities in SC2 in the second year of the project revolved around two main areas: stakeholders’ engagement and design of the big data pilot for SC2.

² https://www.youtube.com/watch?v=L6sF_0rgocA
³ https://fda.opentrials.net/about
1.2.2. Pilot Use Case

The SC2 Pilot Cycle 1 was completed. The pilot was carried out by SWC, AK and FAO. The goal of the SC2 Pilot Cycle 1 was to demonstrate / evaluate the ability of BDE proposed technologies to complement existing community-driven systems (e.g. VITIS for the Viticulture Research Community) with efficient large-scale back-end text-mining workflows, which can utilise any spark-based Natural Language Processing (NLP) module.

The goal of the text-mining workflow demonstrator was to automatically annotate scientific publications relevant to Viticulture, available at FAO/AGRIS\(^4\) and NCBI/PubMed\(^5\) in PDF format (about 26K and 7K publications respectively) by extracting (a) named entities (locations, domain terms), (b) images / figures and tables as digital objects, and (c) the captions of images / figures and tables.

The extracted information (metadata and digital objects) extends the Knowledge Base of the VITIS application: Metadata are stored as triples in GraphDB, and digital objects (files) are stored in HDFS.

SC2 Pilot Cycle 1 demonstrated the ability of BDE proposed technologies to setup efficient large-scale back-end processing workflows. The ability to effectively handle pipeline failures showcases the choice of Apache Kafka, especially when combined with Apache Flume for seamless data input. Recommendations for the next piloting cycle include the extension of functionality, by extending the Flume/Kafka pipeline to handle other than bibliographic data (e.g sensor / weather data) and inclusion of use case scenarios which combine / link more heterogeneous data sources.

Finally, the plan for SC2 Pilot Cycle 2 was prepared. More specifically, the goal of this pilot cycle is to showcase the ability of scalable processing workflows to handle a variety of data types (beyond bibliographic data) relevant to Viticulture.

The content pool will be extended to include:

- Weather Data, available via publicly available APIs (e.g. OpenWeatherMap, Weather Underground, AccuWeather etc.)
- User-generated data, e.g. geotagged photos from leaves, young shoots and grape clusters, ampelographic data, SSR-marker data etc.

\(^4\) http://agris.fao.org/agris-search/home
\(^5\) http://www.ncbi.nlm.nih.gov/pubmed
The goal of the inclusion of these data is to complement the existing SC2 Pilot Demonstrator Knowledge Base, so as to support complex real-life research questions, based on the correlation of environmental conditions with real observations on crop production and quality.

1.2.3. Stakeholder Workshop

The main event around stakeholders’ engagement in the second year was the second societal challenge workshop. Based on lessons learnt from the first workshop, especially the weak involvement of EC Units, as well as on the recommendations of the Project Officer, the objective of this second workshop was to involve a selected group of EC representatives that are working on agendas or initiatives of relevance (from DG AGRI, DG CONNECT, DG RTD as well as the EU Publications Office) to meet a number of experts from key non-EC organizations and companies working on data for food and agriculture (FAO, CGIAR, GODAN, INRA, Alterra Wageningen, TNO, Syngenta, AgroKnow) to share recent developments.

The one-day workshop took place at the CLORA premises in Brussels on 30/09/2016 and followed the DG AGRI “Digitising the Agri-food Sector” meeting that had taken place the previous two days also in Brussels. Organizing the workshop back-to-back with the DG AGRI meeting gave us the opportunity to have some members of the BDE project and some key partners invited to the DG AGRI meeting and to get a full report of the meeting during the BDE workshop.

1.2.4. Engagement, Outreach, Dissemination

Besides the main SC2 BDE workshop, BDE was also presented at the “Open Harvest” workshop held in Chania, Crete, in May 2016, where representatives from different types of institutions in the area of food and agriculture (research, private sector, international) met to discuss elements of a roadmap for a data infrastructure for food and agriculture.

Furthermore, AgroKnow and FAO have continued with the BDE webinar series held on the Webinars@AIMS\(^6\) platform. In February 2016, a webinar on “Perspectives on Big Data in the CGIAR” was conducted by Elizabeth Arnaud of Bioversity International. A second webinar focused on the pilot was planned for November 2016 but has been postponed to early 2017 so as to align with the developments on the pilot.

News about BDE (webinar, workshop, reports) have been disseminated to the community through various channels (the FAO AIMS website, the GFAR website, the AgroKnow’s blog and related newsletters). Eight thematic articles/blogs on data for food and agriculture have been posted to the BDE website.

\(^6\) http://aims.fao.org/capacity-development/webinars
Lastly, the list of stakeholders for outreach by the project (experts made aware of the project and involved in consultations) has increased from 140 to 193, with the addition of participants in the second workshop held in Brussels in September 2016 and participants in the “Open harvest” workshop held in May.

1.2.5. Next steps
In the third year of the project, the main priority will be the development of the next pilot cycles, while continuing the engagement of stakeholders through two webinars and the third workshop. The webinars will be focused mainly on the pilot, which uses instances of the BDE platform for our societal challenge, whereas the workshop will focus on future developments for the definition of a roadmap for a big data infrastructure for food and agriculture.

Other events and initiatives where the SC2 pilot could be presented in 2017 are: the second Open Harvest workshop in May 2017, the workshops that will be organized by the newly started H2020 e-ROSA project, as well as the GODAN Data Ecosystem Working Group.

1.3. Societal Challenge 3 - Secure, Clean and Efficient Energy

1.3.1. General Objectives
Tasks in 2016 for the SC3 in the BDE project mainly focused on pilot show case preparation, supportive dissemination actions and on the continuation of community building and requirements elicitation.

1.3.2. Pilot Use Case
As the first showcase in the energy domain the System Monitoring example was selected (this example was also identified in the proposal stage). The use case is typical in the energy domain with its strong industrial footprint and relevant to other domains with systems and processes monitoring challenges (i.e. transport or manufacturing).

The pilot aims to present to the community the capabilities of the available Big Data technology tools (in the form of the BDE platform) for the task of monitoring a wide network of sensors producing a high volume of data; the paradigm being the operation and condition monitoring of wind energy converters. The pilot includes exemplary analytics modules.
The next cycle of the pilot case was investigated for various energy fields ranging from energy efficiency to resource assessment, and from condition monitoring to renewables forecasting. The evaluation was based on pilot development readiness, the innovation from the SC perspective, the data availability and volume, the identification of the main stakeholder as well as on evaluation of its contribution towards the domain widening, the improvement of technology and the broaden of the community. The extension of the first pilot, towards the implementation a real time component and the addition of a research workflow, has been promoted as the best option for the second cycle. The case of resource assessment and forecasting is the current candidate for the last cycle of the pilot cases.

1.3.3. Stakeholder Workshop

The community building efforts are on-going, and the current stakeholder list (>190 persons) comprises representatives from industry, international organizations and associations, academia, related research projects, EU officials and private companies. The fields of electricity production, transmission and distribution, renewable energy production, distributed production and smart grids, energy saving and energy policy planning as well as IT (Information technology) and OT (operational technology) service are represented.

The actions were supported by the scheduled thematic workshop and hangout. The second thematic workshop, held at Brussels on the 4th of October 2016) focused on energy production and energy efficiency. The workshop attracted prominent speakers from the energy industry who, along with the participants, supported the identification of data management challenges in the domain. The second online hangout was focused on the system monitoring data acquisition technology. In the events the BDE platform technical details along with the pilot case (including the details on the developed data acquisition infrastructure and the BDE platform component use) were highlighted.

1.3.4. Engagement, Outreach, Dissemination

The dissemination activities were supported through the BDE website (thematic posts) as well as through the organization of the workshop and hangout.

1.3.5. Next Steps

During the last year, the primary focus will be the development of the next pilot cycles.

The third workshop will focus on the pilot applications and the BDE application options for the Energy community. Except from the presentation of the first pilot, it will also be strongly considered and aimed to attract stakeholders from the electricity grid operation and monitoring field (i.e. smart grids). For the next online hangouts the focus will be on the pilot presentation and the related BDE technical solutions.

The foreseen webinars in 2017 regard the presentation and discussion per pilot cycle, the BDE Platform options for Energy applications and optionally a review of Big Data application in the Energy domain.

1.4. Societal Challenge 4 - Smart, Green and Integrated Transport

1.4.1. General Objectives

The activities in the transport domain (Societal Challenge SC4) focus in particular on streaming sensor network and geo-spatial data integration. ERTICO - ITS Europe has the task to organise three workshops as well as provide support in dissemination and community building activities with regular blog posts and the organisation of multiple webinars.

1.4.2. Pilot Use Case

ERTICO has contributed to the identification of a series of SC4 pilot candidates to implement the BDE architecture for a real use case within the transport domain. The pilot was demonstrated in multiple webinars as well as being the focal point of the second workshop.

The 4th societal challenge, Smart Green and Integrated Transport, is a broad topic ranging from urban mobility, to safety, logistics and transport system integration. Transport systems consume huge flows of data to provide services, monitor infrastructures and their usage and forecast what will be the usage
patterns in the near or distant future. All these systems need to consume streams of unbounded data sets in different formats and integrate them. As a consequence for the 4th Societal Challenge we have decided to build a pilot that can ingest streams of data that have spatial and temporal dimensions, process them and store the result in order to enable monitoring and forecasting. We have focused our pilot on urban mobility and particularly in traffic flow monitoring for the 1st cycle of the pilot. Our partner CERTH is managing a system that monitors traffic flow using floating car data from the Thessaloniki Taxi Association and from a Bluetooth sensors network deployed in the city. The legacy system currently used by CERTH is based on proprietary software and its limit in scalability have been already reached. The objective of the pilot is to investigate the use of Big Data frameworks to provide a scalable, fault tolerant platform that could replace the system currently in use.

The size of current data sets from CERTH amount to 44 GB for the FCD historical data, up to 200 MB/day for the FCD near real-time data and 50 MB/day from the BT sensors. The architecture has been designed in order to handle many data sources, process the data in time windows and store the result so that critical features such as geographic information and time can be indexed enabling low latency for search, queries and analysis. The architecture of the pilot is based on microservices. The best available open source frameworks for the SC4 tasks have been selected from those available in the BDE platform. Apache Kafka can ingest a high data throughput and distribute the data to different processing components or microservices. The microservices used by an application can communicate, while being decoupled, sending the messages to a Kafka topic. A producer send the data to a Kafka topic from which a consumer can read the data, process it and send the result to another topic, in order to be consumed by another component, or to a sink (database). More than one consumer can have access to the same topic and any of them can read the data from the beginning or from a specified message. Apache Kafka can be configured to store the messages as long as it is needed before being processed. It is used in many large companies as a backbone for their data transformation pipelines. The processing tasks are handled by Apache Flink that provides support for stream and batch processing, and libraries to work with graphs and machine learning algorithms. The road network data from OpenStreetMap has been imported in PostGis with some SQL scripts that implement the functions called by the R script for the map-matching. All the components are released with a Dockerfile and the corresponding image is available on Docker Hub. Docker provides a set of tools to deploy and run applications in an isolated environment. The environment currently supported by Docker is Linux. The concrete form of this isolated environment is a Docker image. A developer can add to such image the binary files or the application source code with the tools, libraries and configuration files it needs to be executed. A Docker image runs within its own container and is isolated from other containers and from the host. The isolation provided by the container ensures that an application that runs in a container can access another container only through the networking service provided by Docker. An application can be based on different containers that communicate
through a network. As an example, a web application can be added to an image that contains a web server that sends requests to a database running in another container. Docker provides an API that enables a user to build an image from its definition, or Dockerfile, pull or push images to a repository, run or stop a container, list the images available in the host or the containers currently running on it. All these functionalities are implemented in a tool, Docker Engine, that must be installed in the host. Docker Engine can be run in Swarm mode to support multi-host environments. Docker containers deployed in different nodes that are members of a swarm can communicate and form a cluster of distributed containers.

The pilot that has been developed in the 1st cycle consumes the near real-time and the historical floating car data from CERTH applying the map matching algorithm so that we are able to make aggregations, such as the average speed and the flow (i.e. number of vehicles) in each road segment within a time window, and visualize the result. The map matching algorithm integrates the traffic data and the road network data from OpenStreetMap matching the coordinates pairs from the FCD data to the road segments. This algorithm is particularly relevant as it the starting point for more interesting use cases such as finding the fastest path that connect two points. We use the dockerized versions of the frameworks provided by the BDE project as General Infrastructure Components and other pilot specific components as it makes it easy to deploy and scale the pilot in different environments, from a set of VMs in a laptop for development and testing, to physical nodes and cloud environments. The map matching algorithm and the prediction algorithm have been developed in R since it provides many software packages used in the algorithms. In order to integrate the algorithm in the pipeline we had to map the data structure used by R to the data structures used by Apache Flink and send the data from Flink to R for the map matching and prediction. A limitation of R is that it supports by default only one thread but since it is used by many data analysts there have been efforts in the industry to provide support for parallelization and access to big data sets. Software packages such as SparkR, RHadoop, MRAN, H2O R package, allow R users to run their script using all the cores of their machines in parallel and to process data sets from Hadoop or Spark.

1.4.3. Stakeholder Workshop

ERTICO continued its effort in identifying and involving key stakeholders in the transport domain. As of today over 380 stakeholders were identified representing industry, international organizations and associations, academia, related research projects, EU officials and industry.

The primary objective of the second workshop organised by ERTICO, held 22nd September 2016, was to showcase the platform applied to the Transport pilot site in Thessaloniki. The pilot demonstrates the features and flexibility of the big data integrator platform, allowing participants to understand the
potential of the Big Data technologies made available by the project. Further Big Data needs from workshop participants were taken into account to improve future prototypes.

1.4.4. Engagement, Outreach, Dissemination

The 45 participants of the workshop came from a diverse background, including representatives of associations, research institutes and universities, industry, the European Commission, public authorities and consultancies. Having such a diverse attendance revealed its benefits especially during the breakout sessions where the participants could voice, exchange, discuss and learn from each other on various topics in the transport and big data domains. It was a valuable opportunity to meet people that deal with transport data, using or providing services based most of the time on legacy systems that might benefit from the outcomes of the pilot.

Two webinars were organised over the course of 2016, one in June, prior to the workshop and following the release of the big data platform. The webinar was attended by 46 participants and featured the following agenda: Simon Scerri gave an introduction and overview of the project, Hajira Jabeen provided a general technical background to the BDE platform, and Josep Maria Salanova and Luigi Selmi gave a detailed presentation on the Transport-specific pilot.

A second webinar took place in the beginning of December to build upon the learnings of the workshop and to discuss the potential of big data technologies and the future of the BDE platform in its second cycle phase coming up next year. An introduction to the transport domain was made by Maxime Flament, followed by a presentation on the potential of big data technologies by Paul Kompfner. Finally, Josep Maria Salanova and Luigi Selmi gave a short background and demonstration of the platform applied in Thessaloniki.

The team has contributed towards raising awareness on big data opportunities with a series of blog posts and other publications on the BDE website, the ERTICO Network and several other printed journals.

1.4.5. Next Steps

The next cycle will be an evolution of the first pilot and the main use case will be the forecasting of the traffic level (speed and flow) by implementing the prediction algorithm, based on a neural network, and the training of a model for the road segments. One other target is to improve the map-matching algorithm using the topological information available from the OSM data, currently the
algorithm uses only geometric information (i.e. distance between the vehicle and the road segment and their orientations).

A third workshop is planned in the fall of 2017 and will focus on the alignment and improvement of the specifications and evaluation of project technologies on selected use cases. Furthermore, two webinars related to the next cycle of the pilot use case will be organized. ERTICO will continue activities to further engage the community and disseminate progress and results via blog posts on the website as well as through the ERTICO Network.

1.5. Societal Challenge 5 - Climate, Environment, Resource Efficiency and Raw Materials

1.5.1. General Objectives
The SC5 BDE action “Climate Action, Environment, Resource Efficiency and Raw Materials” aims to support a large and diverse community with big data technologies via communication and networking actions and through the development of relevant use-case pilots and software tools.

During the last reporting period, specific actions were taken to achieve the goals set for the SC5 domain. These actions have been focused on developing the first SC5 reference pilot, by further eliciting technological requirements via workshops and online tutorial and hangout sessions and by giving presentations at conferences of interest. The BDE party responsible for the SC5 actions is NCSR-“Demokritos”.

1.5.2. Pilot Use Case
The first SC5 pilot was designed and implemented according to previously acquired requirements, via online channels (hangouts and correspondence) and the first SC5 workshop. This pilot demonstrates ways in which weather and climate data ingestion, transformation, analytics, downscaling and lineage can be made more productive and traceable via the use of big data technologies. These pilot operations are designed to work on WRF-compatible NetCDF files. The four main functionalities implemented are (1) climate NetCDF data ingestion and exporting, (2) basic analytics, (3) facilitation of climate downscaling and (4) data lineage. The pilot was evaluated during a hands-on online tutorial session on

7 The Weather Research and Forecasting Model: http://www.wrf-model.org
12 July 2016. This evaluation pointed to the need for the expansion of analytics functions, something which is facilitated by the pilot design and the BDE platform. It also confirmed the usefulness of metadata and data lineage for large experiments involving multiple streams of complex data. The first pilot was presented at the European Geosciences Union (EGU) General Assembly 2016.

Aiming to extend our work to additional SC5 domains, the follow-up SC5 pilot will focus on the theme of emergency response as a result of a hazardous substance release into the atmosphere, and will demonstrate how big data technologies can aid in estimating the origin of radioactive or toxic releases based on contamination detections. The second piloting cycle will make use of some of the components implemented by the first, regarding the management of NetCDF data. For discussing the design and approach, and for eliciting further requirements for the 2nd pilot, an online hangout was carried out on 20 December 2016. This pilot has been technically specified and cross-checked with the platform team for technical soundness, and is currently being implemented.

1.5.3. Stakeholder Workshop

A number of relevant stakeholders has been identified to aid in requirements elicitation, thus shaping the practical work of BDE by collecting requirements for the ICT infrastructure needed by data-intensive science practitioners. Stakeholders include public and private organisations, universities and research centres, relevant networks and initiatives and decision makers. At the time of writing, a total of 285 stakeholder representatives have been registered as observers, followers, endorsers and contributors (208 relating to the climate action, 58 relating to the emergency response actions, 19 EC officials). Several of them have participated in the events organised by the SC5 BDE team (workshops, webinars) and have been following the project developments via the website/blog, newsletters and targeted correspondence.

The second BDE SC5 workshop was held to present the 1st pilot developed, the 2nd pilot under specification, as well as to identify current and future Big Data challenges in the domains covered by SC5 (Brussels on 11 October 2016). The main challenges and complexities faced during Big Data implementations in the area were identified and discussed, while typical applications and potential directions regarding Big Data technologies were presented by a series of internal and invited speakers.

The presentations were followed by a discussion on the utilisation of the BDE platform and issues in relation to the planned 2nd SC5 Pilot case. In particular discussions were held regarding the BDE platform and how easy (or not) it would be for a third party to acquire, deploy and use it. Another matter that was discussed was related to the NetCDF format, its integration to the BDE platform and

---

ways to enable analytics to be drawn from relevant data. With respect to the 2nd SC5 pilot, the discussion revolved around current practices for identifying the location of a source using inverse modelling and the need for complementary and more efficient solutions.

**Summary of events organised:**
- 2nd SC5 online hangout, 12 July 2016
- 3rd SC5 online hangout, 20 December 2016
- 2nd SC5 workshop, 11 October 2016

1.5.4. Next Steps
Currently focus is given on implementing the specifics of the 2nd SC5 pilot. This involves applying data-scientific methods to voluminous weather data sets. We are also planning the next dissemination and training actions. This will include a presentation of the pilot at EGU’17, and the organisation of follow-up online sessions and workshops.

1.6. Societal Challenge 6 - Inclusive, Innovative and Reflective Societies

1.6.1. General Objectives
Role of CESSDA AS in the BigData Europe project is to coordinate the SC6 interest group for “Europe in a changing world - inclusive, innovative and reflective societies”, and potential users of big data in the fields of social sciences and humanities (SSH). Furthermore, it should build this interest group, collect its requirements, assist the building of an ICT big data infrastructure access point for SSH, explore and evaluate the input data, and discover the implications for the future of big data in SSH. In 2016, CESSDA coordinated a number of activities in SC6 area:

1.6.2. Pilot Use Case
As a technical SC6 domain partner, Semantic Web Company (SWC) worked together with CESSDA on all of the above listed activities. Building up on work done in 2015, the SWC team supported SC6
activities regarding the identification and specification of the SC6 Pilot / Use Case in the field of economic data that was specified and implemented in 2016.

The main question that has driven this pilot so far is whether municipal/public budgets can be made more useful for citizens, researchers and decision makers? Different components within the creation of an online dashboard on Economic Data were highlighted: harvesting, normalising, linking & mapping the data, analysing the data, visualising and finally providing raw data. Vast range of potential users and usages was covered, from citizens and journalists to decision makers and researchers. The Linked Economy model was introduced as a common semantic model that captures an integrated view of the economic activity, being economic-theory driven, based on existing vocabularies, structured and open development process, easy to be consumed by any system and scalable as well as reusable. Data used and produced in the pilot (including budget data and budget execution data) are coming from the municipalities of Athens, Thessaloniki and Kalamaria, are in detail explained, along with the process of converting mass data input into exploitable output.

SC6 pilot architecture and technical components were elaborated and 1st MockUp/Wireframe was developed as well as the planned Evolution of the Pilot.

1.6.3. Stakeholder Workshop

Second SC6 workshop - “The Challenges of Big Data for societies in a changing world” was held on 5th December 2016 in Cologne, Germany, and was co-located with the EDDI2016 conference. The central theme of this workshop was the SC6 Pilot: Citizens Budget on Municipal Level built on the BDE Platform. SC6 is dealing with data from three municipalities, Athens, Kalamaria and Thessaloniki budget execution data. Further data from additional data sources (also from outside of Greece) should be integrated e.g. in a 2nd step of the Pilot realisation.

1.6.4. Engagement, Outreach, Dissemination

Community building and Dissemination

One of the ongoing activities of the SC6 is regular reporting and dissemination of activities throughout the project, as well as distribution of all relevant information and materials related to the SC6 domain through the already established communications channels within the CESSDA and BDE consortiums (through CESSDA and BDE website), but also via social media (twitter) and a W3C community group.
In 2016, CESSDA posted 8 blogs on BDE community website, and organised and held 2 hang-outs and 1 workshop aiming at SC6 community members (decision makers/EC, ministries, Research Councils, domain experts/social sciences and humanities researchers, technical staff).

**Second SC6 hang-out** was held on 25th May 2016 on New General Data Protection Regulation (GDPR) adopted by the EC. The Regulation shall apply 2 years after its formal adoption by the European Parliament and Council which now is the case. The regulations for processing personal data affect large areas of European research and are of great importance to the scientific communities and research infrastructures of Europe.

**Third SC6 hang-out** was held on 28th September 2016 and its main purpose was to introduce the ongoing work on the SC6 pilot - “Citizens’ budget at municipal level” to the domain community. Local government budgets are filled with so many numbers and technical jargon that the ordinary readers cannot easily understand what they mean. It was needed to have a clear and simple summary guide (visualisations) for citizens (the so-called citizens’ budget), researchers and decision makers. Expected impact included the opportunity for citizens to understand, comment, and propose actions related to the budget of their municipality, and for researchers and/or general public to download and analyse available data. Pilot would enrich the opportunities for accountability, transparency and research innovation.

1.6.5. Next Steps

To ensure scalability and performance with a growing amount of data, as well as smooth data integration from different sources in several formats the Big Data Europe Aggregator Platform is being used to realise this BDE SC6 Pilot on top of Big Data technology and principles.

1.7. Societal Challenge 7 - Secure Societies
1.7.1. General Objectives
In the framework of the BigDataEurope project, SatCen is the domain leader addressing the “Secure Societies” Challenge focusing on Big Data opportunities and requirements as well as exploring and evaluating relevant Big Data end-to-end management approaches and techniques.

1.7.2. Pilot Use Case
The “Secure Societies” H2020 Societal Challenge is related to the protection of freedom and security of Europe and its citizens. A major activity in supporting the primary aims of this Societal Challenge (in particular to enhance the resilience of our society against natural and man-made disasters, to develop novel solutions for the protection of critical infrastructure, to improve border security and to support the Union’s external security policies) is the provision of geospatial products and services, mainly resulting from satellite data. In fact datasets used in the Space and Security domain comply with the definition of Big Data in terms of variety (data are coming from different sensors in orbit on several governmental and commercial satellites), volume (data received each day from satellites are on the order of terabytes), velocity (data have to be delivered and processed in a short time frame to provide users that require fast responses with 24/7 information), veracity (decision making and operations require reliable sources) and value (information provided have to be useful and clear).

1.7.3. Stakeholder Workshop
BigDataEurope project and the related SatCen activities were presented during the 2016 conference Big Data from Space (Tenerife, 15-17 March), the ESA Living Planet Symposium 2016 (Prague, 9-13 May), the SatCen Expert User Forum (Madrid, 17 June) and the Earth Observation Open Science 2016 conference (Frascati, 12-14 September). The community was consolidated with the 2nd “Big Data in Secure Societies” Workshop, held in Brussels on 18th of October: almost 60 participants (in particular, stakeholders from EU entities, International Organizations, Industry and Academia) attended this event, representing a consistent increment with respect to the previous edition, held in 2015. Moreover, two Hangout sessions (held in April and December respectively) allowed to present the BigDataEurope activities through dedicated webinars.

1.7.4. Engagement, Outreach, Dissemination
In 2016 SatCen activities were mainly focused on raising awareness on “Big Data in Secure Societies” amongst the stakeholders of the Security domain. In the framework of these Community Building activities, the Secure Societies pilot was presented during key events of the security domain.
According to the user requirements collected during the first phase of the project and their refinement during the second year of the project, the pilot developed with the Secure Societies technical domain leader (the University of Athens) and the support of NCSR “Demokritos” reached a first milestone and it was presented in the framework of the above listed events. The pilot considers the fusion and analysis of information coming from remote sensing (satellite data) and social sensing (news from Reuters and Twitter); more in detail, the user through a user-friendly interface can analysis satellite images to detect areas with changes on land cover or land use and integrate the output extracting information from social media and news items. The second phase of the pilot represents an improvement of the first one, where new functionalities (e.g. authentication/authorization mechanisms) are added to widen the Secure Societies domain to different subdomains (e.g. cybersecurity).

1.7.5. Next Steps

In 2017 several activities will continue the work of the first two years, where new key events to present the BigDataEurope project have been already identified (e.g. Big Data from Space 2017) and the Community Building and Dissemination activities will continue with the organization of the 3rd workshop on “Big Data in Secure Societies” (autumn 2017). In addition, periodic webinars and internal SatCen events will also be organized to maintain the involvement of the stakeholders in the project and to collect further feedbacks and to refine the user requirements.


The requirements elicitation of WP2 reveals that there is not one V (out of four Vs, volume, velocity, variety and veracity) that overshadows the others, not even within a specific societal challenge. The BDE platform, therefore needs to be flexible in supporting Big Data stacks characterized by any of these Vs. The base platform (see Figure 1) designed in WP3 based on Docker Swarm offers a solid foundation to this end. Docker Swarm can be thought of as a kernel for a distributed operating system. It abstracts from the underlying hardware making the cluster behave as one big machine to the user. Docker provides a container abstraction allowing to incorporate virtually any technology. One container could run Spark, another could run Elasticsearch, and yet another container can use a completely custom implementation. In order to get a basic idea of the platform, and how it can be used, we provide a video instruction for installation and a preview of the base platform can be installed.
following this instruction. An automated installation of the base platform is available in the form of Ansible playbooks.

![BDE platform architecture](image)

An application running on the BDE platform can be seen as a stack consisting of multiple components, which are wired together in order to solve a specific Big Data problem. The components are packaged in Docker images. In order to encourage technology reuse and to facilitate the development of a component, the BDE platform provides base Docker images. A base image offers a template implementation for a specific technology. This template can be easily extended by a programmer with his/her own custom implementation to solve a particular problem. For example, a Spark base image is available together with a Spark demo application that extends this base image. Overall, the BDE platform provides 20 various docker images including Hadoop, Hive, HBase, Flink and others. The selection of technologies mainly depends on the assessment made during the first year and the pilot cases that need to be implemented. Nevertheless, the usage of Docker keeps options open for the future. The BDE platform is able to embrace new technologies without modifications to the base platform. We also follow-up the evolution of Docker and its ecosystem to validate whether tools as Docker Compose and Docker Swarm could be integrated into the platform to facilitate the setup and deployment of Big Data stacks. Finally, monitoring, logging and helpful GUIs, are being investigated in more depth.

In the third year, we will continue to track cutting edge community developments, ensuring the Big Data Europe platform stays stable and easy to use. This will consist of upgrading Docker and Docker Swarm when necessary, and adopting our platform components to the latest and greatest. We will integrate new external components if they are necessary for the pilot cases, or if they have sufficient community value. We will integrate the current user interfaces more tightly and extend them. This
will make the platform easier to use and will further lower the barrier to entry. Lastly, we will continue our research towards the Semantification of Big Data, turning our initial research into practical cases.

3. Dissemination

The efforts to promote the Big Data Europe project and to disseminate its results continue to pay dividends. Thousands of people a month are seeing the growing amount of high quality content on the project’s website and the number of Twitter followers is just shy of 1,000. Over 150 presentations have been uploaded to SlideShare and 5 videos have been uploaded to YouTube. The release of the beta version of the Big Data Integrator (BDI) platform raised significant interest and the project has been promoted directly at high profile events, including the Apache Software Foundation’s premier event in Europe. This was one of the three high profile events that the project sponsored; in all around 10,000 people have heard about the project at different events during 2016. This is in addition to the webinars and workshops organised by the societal challenge partners.

The project’s flyer has been updated and two iterations of a separate handout giving an overview of the technology used in the BDI have been created. The project’s newsletter has over 900 subscribers and the partners have collected over 1500 e-mail addresses of stakeholders with interests in each of the societal challenges. It is noted that more effort needs to be made to make the newsletter more attractive to boost engagement.

In the final year of the project, dissemination activities will be combined with the community building activities of work package 2. Work has already begun to contact the coordinators of new projects beginning in 2017 with a view to offering the BDI platform as a means to handle their data and the partners are actively seeking ways to expand the audience so far reached.