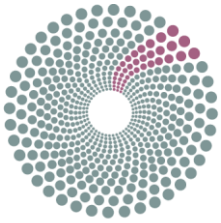

BIG DATA EUROPE

Support Action

Big Data Europe – Empowering Communities with Data Technologies

Project Number: 644564 Start Date of Project: 01/01/2015 Duration: 36 months

Deliverable 1.2

Annual Public Report I

Dissemination Level	Public
Due Date of Deliverable	M12, 31/12/2015
Actual Submission Date	M14, 29/02/2016
Work Package	WP1
Task	T1.3
Type	Report
Approval Status	Approved
Version	1.0
Number of Pages	15
Filename	D1.2_Annual_Public_Report_I.pdf

Abstract: This deliverable reports on the progress of the BigDataEurope project accomplished during its first year. Mainly targeting a general audience, it describes the project's work conducted in 2015 and its first achievements across Europe's seven Societal Challenges (SC) in brief and succinct terms and an easy-to-understand language. The central aim is to inform the public about the contribution and importance of Big Data Europe towards facing these challenges.

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

History

Version	Date	Reason	Revised by
0.0	02/02/16	Draft	Nadine Jänicke (Fraunhofer)
0.1	23/02/16	Report completed	Nadine Jänicke (Fraunhofer)
0.2	25/02/16	Report reviewed	Simon Scerri (Fraunhofer)
1.0	29/02/16	Report revised and finalized	Nadine Jänicke (Fraunhofer)

Author List

Organisation	Name	Contact Information
FHG	Nadine Jänicke	Nadine.Jaenicke@iais.fraunhofer.de
FHG	Simon Scerri	simon.scerri@iais.fraunhofer.de
Open Phacts	Bryn Williams-Jones	bryn@openphactsfoundation.org
FAO	Valeria Pesce	valeria.pesce@fao.org
CRES	Fragiskos Mouzakis	mouzakis@cres.gr
Ertico	Andrea Toth	a.toth@mail.ertico.com
Ertico	Maxime Flament	m.flament@mail.ertico.com
CESSDA	Ivana Versic	ivana.versic@cessda.net
SWC	Martin Kaltenböck	m.kaltenboeck@semantic-web.at
Tenforce	Erika Pauwels	Erika.Pauwels@tenforce.com
NCSR	Mandy Vlachogianni	mandy@ipta.demokritos.gr
SatCen	Michele Lazzarini	Michele.Lazzarini@satcen.europa.eu

Executive Summary

The following deliverable provides the general public with an overview of the main project results accomplished in 2015 by the BigDataEurope (BDE) consortium. Collaborating within the EU Horizon 2020 project entitled “Integrating Big Data, Software and Communities Addressing Europe’s Societal Challenges”, our major goal is to develop a platform that facilitates big data usage across the seven targeted societal sectors. That is, an adaptable and ready-to-use technical solution is sought for stakeholders and user groups interested in Big Data management in the fields of Health, Food & Agriculture, Energy, Transport, Climate, Social Sciences and Security.

In this respect, the project tackles two key aspects. First, (Coordination) BDE aims to build an extensive stakeholder network spread within relevant communities from across the different SC domains; cover the whole process of data usage within each, from data collection, processing, storage and visualization to the development of data services. A series of workshops have brought together these communities with the objective of eliciting domain-specific requirements. Interest groups modelled after the W3C scheme have also been launched to centralise discussions on the needs of each sector, to dissimenate project results and maximise impact. The second aspect of the project (Support) sees that the requirements collected in the workshops and other stakeholder engagement activities guide the technical development and implementation of the open BDE Platform. This infrastructure is to push and extend the use of data technologies within the above key sectors by providing an integrated stack of tools that can be installed and used freely in a customized data processing chain with minimal knowledge of the technologies involved.

During the first year of the project work revolved, on the one hand, strongly around dissemination and community-building activities and, on the other hand, around the foundational technical tasks. Through various networking efforts the domain-leading partners were able to engage with stakeholders and relevant communities from across Europe’s SC domains. This engagement has led to the collection and evaluation of the needs and requirements that stakeholders and user groups identified for the adoption or extension of Big Data technologies in their domains. Moreover, through the workshops, hangouts, webinars or interviews partners organized and carried out the consortium has gained vital feedback for the development of the Big Data Integrator Platform. This information has also already driven the outline of the diverse domain-specific use cases that should essentially guide the adaptation and implementation of tools and technologies for Big Data management within the project. Likewise, first steps were taken towards translating the needs and requirements collected from the different domains into specifications for and an architectural blueprint of the Big Data integrator platform. This effort has led to a first internal release of the BDE platform. Its current set of components presents a flexible basis that allows its use for specific Big Data problems as well as its extension by new technologies within the different domain fields. In the coming year, project activities will focus on instantiating this generic BDE platform for the seven identified SC-oriented pilot use-cases, and the dissemination of intermediate results across the established stakeholder communities.

This deliverable consists of two sections compiling information from project beneficiaries leading the main tasks performed in 2015. The first section summarizes the various networking activities for each of the seven societal challenge domains. Focus will be on the overall goals, work performed in 2015 and next steps to follow by the domain-leading partners. A similar focus is taken in the second part to describe advances made by the technical team during the first 12 months of the project; and also to provide a brief outlook for 2016. 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

BDE	Big Data Europe
SC	Societal Challenge
H2020	Horizon 2020

Table of Contents

List of Figures

1. Summaries of the Seven Societal Challenges for the First Year

1.1 Societal Challenge 1 - Health, Demographic Change and Wellbeing

In common with the other Societal Challenge (SC) areas, the tasks in the first period mainly concern the development of focussed community engagement and outreach activities, delivery of the first workshop and webinars, as well as developing plans and preparing the pilot activities in this domain.

Pilot BDE Use Case:

In a large and diverse bioscience community with an already crowded and vibrant selection of data resources and infrastructures, many of the activities relate to the impact of data variety in the healthcare and demographic change space. Building on the lead of the Open PHACTS Discovery platform in the application of semantic data integration to tackle data diversity in the space, the pilot activities in this domain centre on the delivery of equivalent data services functionality on the generic BDE tech stack. Determined as a key output of the first pilot activities in BDE, this focus has been prominently pursued in the engagement with key stakeholders involved in the Open PHACTS Foundation

Stakeholder Workshop:

The first community workshop for SC1 was held in Brussels on May 21, 2015. This workshop covered perspectives from both industry and academic experts in handling biomedical data. The workshop also featured a working session with attendees to better understand the big data challenges in this domain, and how big data approaches may in future be able to impact the issues faced. The key challenges identified in applying Big Data approaches to healthcare were identified, although there was wide recognition that many of them - particularly those relating to security and privacy of patient data - were beyond the scope of the BDE project. These do, however, remain key issues for this domain as a whole.

Engagement and Outreach:

In this domain, engagement of stakeholders in the absence of a working platform is always a tough challenge given the widespread commercial and other options already available. The emphasis has been on raising awareness with the community in general in advance of the pilot being available, as well as highlighting the role that the Open PHACTS Foundation has in the BDE project with all stakeholders at various forums, conferences and discussions with the biomedical science data community.

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

Most of the activities in SC2 in the first year revolved around stakeholder engagement and elicitation of big data requirements from our community.

In order to ensure the involvement of relevant stakeholders, the Food and Agriculture Organization of the United Nations (FAO), together with AgroKnow and in collaboration with the Global Forum on Agricultural Research, set up a list of important stakeholders with contact information and updated this list throughout the year, using it for dissemination of announcements and for invitations to workshops and webinars. The list included around 140 people at the end of the first year.

A first round of interviews and surveys was conducted in the first year with selected representatives of the domain. A total of 15 interviews were conducted with data managers from key stakeholders like the French National Institute for Agricultural Research (INRA), the World Bank, FAO, Wageningen UR, the International Food Policy Research Institute (IFPRI) of the CGIAR and others.

The main requirements elicitation event in the first year was the workshop on big data requirements in Social Challenge 2[1] held on September 22, 2015 in Paris and organized by Agro-Know, FAO and the Semantic Web Company. The workshop took place in conjunction with the Interest Group on Agricultural Data (IGAD) Pre-Meeting of the sixth Research Data Alliance (RDA) Plenary and was hosted by the *Institut National de la Recherche Agronomique* (INRA). It included six presentations by selected professionals from high-class research institutions and an interactive part to get input from more than 40 experts from around the world. The main finding of the event were that the biggest big data challenge in the area of food and agriculture is the high data variety (multiple, heterogeneous data types and formats from various sources have to be combined) and that researchers still use basic general-purpose software and there is a need for more specialized tools.

Furthermore, in order to raise awareness about big data among data managers in the SC2 community, since November 2015 Agro-Know and FAO have started a BDE webinar series in conjunction with the Webinars@AIMS[2], that are regularly conducted by the AIMS community since 2012. The BDE/AIMS webinar series started in December 2015 and is continuing in 2016. The following topics were presented in 2015: 1. INRA's Big Data Perspectives and Implementation Challenges[3] (on 03/12/2015 by Pascal Neveu) and 2. Big Data challenges and solutions in agricultural and environmental research[4] (on 17/12/2015 by Rob Lokers). 48 people registered for these online seminars that have been recorded and published on the AIMS platform of FAO (http://aims.fao.org).

From the technical side, one of the objectives of the requirements elicitation process was to identify a first interesting pilot for SC2 to be set up in the initial BDE technology stack. After the described round of requirements elicitation activities and based on some of the feedback received, a first concept for the pilot description for SC2 was drafted at the end of the first year. The name of the pilot is “Vitis” and it will focus on aggregating and visualizing experimental data on grape variety identification.

[1] http://www.big-data-europe.eu/event/sc2-paris-2015/
[2] http://aims.fao.org/capacity-development/webinars
[3] http://aims.fao.org/capacity-development/webinars/webinaraims-inras-big-data-perspectives-and-implementation-challenges
[4] http://aims.fao.org/capacity-development/webinars/webinaraims-big-data-challenges-and-solutions-agricultural-and

1.3 Societal Challenge 3 - Secure, Clean and Efficient Energy

Tasks in 2015 for the SC3 in the BDE project mainly focused on community building and requirements elicitation, pilot show case preparation and supportive dissemination actions.

Community Building

The first step in our community building efforts was the identification and contact of data management-related stakeholders in the given domain. The current stakeholder list (>170 persons) comprises representatives from industry, international organizations and associations, academia, related research projects, EU officials and private companies. These groups cover the fields of electricity production, transmission and distribution, renewable energy production, distributed production and smart grids, energy saving and energy policy planning.

Apart from the direct contact to already established networks, other supportive actions were performed for the purpose of community building, namely participation in domain-thematic conferences for networking (e.g. European Wind Energy Association Annual event), presentations in workshop (e.g. EMENDER 2015 - International Workshop on Energy Management, Prediction and Big Data Elaboration) and establish contacts with other user groups.

Data Requirements

The task of evaluation and identification of Big Data management needs in the energy domain was supported by several stakeholder interviews, the first topic-specific workshop (held in Brussels on June 15, 2015) and the first on-line hangout (held December 21, 2015). The domain-thematic workshop attracted prominent speakers from the energy industry who, along with the participants, supported the identification of data management challenges in the domain. The topics of energy production, renewables and smart grids were discussed. During the first online hangout further information was delivered including the technical background for the BDE platform along with the first pilot use case description.

Pilot Preparation

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 will include exemplary analytics modules.

Dissemination Activities

The dissemination activities were supported through the BDE website (thematic posts) as well as through the participation to the afore-mentioned events (EWEA 2015 and EMENDER 2015).

Next Steps

As the domain is quite wide the data requirements-related actions are continuing and the effort for the strengthening of the involvement of the stakeholders will increase. These actions will be strongly supported by the tangible results of the progress for the first pilot implementation.

For the organization of the next workshop, 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.

1.4 Societal Challenge 4 - Smart, Green and Integrated Transport

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.

Elicitation of Requirements

Prior to the first workshop held in October, ERTICO conducted a series of 12 interviews with business, strategically, technical and domain experts with the aim of gathering a deeper understanding of the requirements.

In October 2015, ERTICO organised a workshop focusing on the elicitation of requirements in the transport area. This workshop took place during the 22nd ITS World Congress in Bordeaux. The overarching aim of the workshop was to define the requirements necessary for big data management in the intelligent transport domain. The workshop was glad to host speakers from various backgrounds, be it policy, industry, research institutions or universities, who presented on a wide range of topics ranging from the role of social media in transport, to open logistics and traffic management as well as to data analyses techniques.

The workshop revealed that big data with regard to transport holds considerable benefits for citizens with new and better services in transport, for society at large with, for example, a more environmental traffic flow, for the public sector with an optimisation of traffic management, and for service providers for whom big data provides an important business opportunity. Big and open data also plays an important role in how smart cities deploy and use ICT to enhance their transportation networks.

Community Building and Dissemination

As a first step to community building ERTICO contributed to the identification of key stakeholders within the transport domain. As of today over 230 stakeholders were identified

representing industry, international organizations and associations, academia, related research projects, EU officials and industry.

A webinar was organized prior to the workshop as an introduction to the BigDataEurope project and in particular to the opportunities and challenges within the transport domain. The webinar was attended by 83 participants and featured the following agenda: Simon Scerri, University of Bonn gave an overview and introduction to the BigDataEurope project; Philippe Crist from OECD summarized the findings of the International Transport Forum report on big data; and finally Maxime Flament from ERTICO gave an introduction to the workshop coming up in October.

A second webinar took place in mid-December to build upon the learnings of the workshop and to discuss the needs and requirements of the platform from the viewpoints of Policy, Business, and Technology. Speakers included Dave Marples, Technolution (Business), Sean Gaines, Vicomtech (Technology) and Maxime Flament, ERTICO (Policy).

The team has contributed towards raising awareness on big data opportunities with a series of blog posts on the BDE website as well as on the ERTICO Network.

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 selected taking into account the lessons learnt from the first workshop as well as the webinars where domain-specific requirements and needs were pinpointed.

Next Steps

A second workshop is planned in the fall of 2016 and will focus on the alignment and improvement of the specifications, and the third on the evaluation of project technologies on selected use cases. Furthermore, two webinars related to the identified pilot use case will be organized, foreseeably in June and November. 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

Introduction

The BDE project aims, amongst others, to provide an integrated stack of tools to manipulate, publish and use large-scale data resources. The tools will be installed and used freely in a customised data processing chain, integrating key open-source Big Data technologies and European research prototypes into a Big Data Integrator Platform. During the first year of the BDE project, specific actions were taken to achieve the contractual goals determined within the SC5 domain, which mainly focused on analysing the technological demands and requirements pertaining to managing and exploiting Big Data of European stakeholders. In relation to community building & support, the establishment of a SC5 Big Data interest group has been achieved by using various channels (workshop, interviews, online survey, etc.) to include as many stakeholders and / or projects of the H2020 work programme. In addition, the first BDE SC5 pilot was planned. More information follows below.

Stakeholders

A number of relevant stakeholders, representing the H2020 SC5 'Climate Action", has been identified to collect requirements for the ICT infrastructure, needed by data-intensive science practitioners. The range of stakeholder groups includes public and private organisations, universities and research centres as well as networks and initiatives in this domain. Overall, a total number of 146 stakeholder representatives has been registered in the list as observers, followers, endorsers and contributors. Several of them have participated in the events organised by the SC5 BDE team (workshop, webinar) and have been following the project developments via the website, newsletters and the w3c group posts.

Questionnaires / Interviews

Quality face-to-face interviews with at least five representatives from each of the two stakeholder groups (climate domain experts and climate technical experts) were successfully performed. The interviews focused on the requirements elicitation of Big Data needs and problems and the specification of comprehensive use cases and user stories. The identified requirements and information collected from the interviews were documented in the form of questionnaires which served the compilation of the final requirements specification document.

Workshop

The first BDE SC5 workshop was held with the intention to identify the current as well the future Big Data challenges in the Climate domain (in Brussels on June, 15, 2015). The challenges and complexities faced during Big Data implementations in the area were identified and discussed, while typical applications of Big Data technologies were presented. Stakeholder requirements were collected to support the design and realization of the required computational infrastructure that will make use of the Lambda-Architecture. The key challenges identified in applying Big Data approaches to the Climate Action were the following:

- Data management tools and practices currently common in the domain and /or in the participating organizations; problems in getting, processing, analysing and storing the data; Identification of tasks using large (internal and/or external) datasets
- Discussion of use cases of interest; functional and non-functional requirements for the big data platform Identification of tools/systems/databases/resources -gaps applied; practical or technical restrictions that might impede access and processing of the data
- Implementation of the Open Access to Research Data policy promoted by the EC to all projects funded under H2020; legal concerns that could be encountered in the usage of (Big) Data management solutions; identification of key roadblocks influencing policy requirements.

Pilot

The organisation responsible for the compilation of the first BDE SC5 pilot is NCSR "Demokritos". The pilot has been designed by taking into account the user requirements collected and analysed through the project actions (workshop, questionnaires, etc.). The pilot aims to facilitate the process of dynamical downscaling from global climate data to regional / local scales with the support of tools aggregated on the BDE platform. The overall first pilot objectives are focused on improving the productivity of climate researchers (through e.g. easier

management, ingestion and transformation of external data, making use of existing infrastructure and procedures) and creating opportunities for pilots across communities within the BDE platform (by e.g. climate change impact assessment studies on sectors such as energy, food and agriculture). The first version of the technical document describing the pilot has been produced.

Online Hangout Preparation

Towards the end of the first year, specific actions were taken to prepare a BDE SC5 online hangout (webinar) for the climate community by sending private invitations and informing the stakeholders via emails and website announcements. The aim of the webinar was to introduce the activities and recent developments on the use of Big Data in Climate action within BDE and present the first pilot use case that will be developed and implemented in the BDE integrator platform.

1.6 Societal Challenge 6 - Inclusive, Innovative and Reflective Societies

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.

During 2015 most work was done within the area of community building and requirements. Stakeholders for SC6 were identified, and requirements were elicited, defined and prioritized for this particular SC. During the webinar held on October 13, 2015 a general presentation was given on the project and recent developments. In order to further and explore in more detail the requirements already identified through the questionnaire, 13 face-to-face interviews with domain experts in SSH were accomplished and a stakeholder workshop was organized as back-to-back event with the European Data Forum (http://2015.data-forum.eu) on November 18, 2015 in cooperation with Eurostat (see more details: http://bit.ly/1fRc3eJ). This workshop offered more insight in specific risks and challenges related to SC6 and social sciences and humanities in particular when it comes to big data management and storage.

By the end of the year work began in planning and outlining interesting and useful application scenarios in the given domain, and the identification of a promising showcase involving key large-scale datasets being relevant for the SC6 (social sciences and humanities) started.

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 (CESSDA and BDE website), but also via social media (twitter) and a W3C community group.

As a technical SC6 domain partner, Semantic Web Company (SWC) worked together with CESSDA on the above listed activities. The SWC team gathered the requirements of the stakeholders along the 7 societal challenges (by a stakeholder survey, 100+ domain experts’

interviews, 7 workshops et al). In all the above mentioned activities SWC covered more the technical part regarding Big Data technologies and related data management requirements and supported CESSDA in respect to raising technical questions and issues. For concrete outcomes of requirements elicitation in the SC6, please see also the respective slides of the talk of SWC at the first SC6 workshop taking place in November 2015 at Eurostat: http://www.big-data-europe.eu/a-brief-summary-of-the-first-big-data-europe-sc6-workshop-the-challenges-of-big-data-for-societies-in-a-changing-world/.

Finally, SWC supported SC6 activities regarding the identification and specification of the SC6 Pilot / Use Case (that will be around the topic of budget / spending data – and thereby in the field of economic data) that will be finally specified and also implemented in 2016.

1.7 Societal Challenge 7 - Secure Societies

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).

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. In 2015 several activities were conducted by SatCen 1) to build and consolidate a User Community in the Security domain with an interest in the Big Data issue and 2) to evaluate new solutions needed for the management and exploitation of big amount of heterogeneous data for Secure Societies.

The community was initially built through the submission of customized questionnaires and discussions on Big Data management practices with a number of entities active in the Security domain such as SatCen and Security Stakeholders (in particular EU entities, EU Member States, EC representatives, International Organizations and Industry representatives). Another milestone was the first “Big Data in Secure Societies” workshop held in Brussels on September 30, 2015; the workshop attracted people coming from different EC DGs and a number of other entities and private companies covering the domains of Space and Security, Cybersecurity, Fight against Crime and Data/Infrastructure management.

According to the user requirements collected during these activities in the first phase of the project, a pilot was developed with the Secure Societies technical domain leader (the University of Athens) and the support of NCSR “Demokritos”. The pilot considers the fusion and analysis of information coming from remote sensing (mainly satellite data) and social sensing (news from Reuters and Twitter); more in detail, the analysis of satellite images to detect areas with changes on land cover or land use will be enhanced using information extracted from social media and news items.

2016 will be an important year for the project. The 1st testing version of the pilot will be available in spring and it will be presented at the 2nd Big Data from Space Conference which will be held in Tenerife. The community building and dissemination activities will continue with the organization of the 2nd workshop on “Big Data in Secure Societies” (autumn 2016), periodic webinars and internal SatCen events with the aim of maintaining the involvement of the stakeholders in the project to collect their feedback and to refine the user requirements.

2. Summary of Technical Advances in the First Year

The requirements elicitation of WP2 reveals that there is not one V that overshadows the others. Not even within a specific societal challenge. The BDE platform therefore needs to be flexible in supporting Big Data pipelines characterized by any of these Vs. The base platform (see Figure 1) designed in WP3 including Mesos and Docker offers a solid foundation to this end. Mesos 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 on its turn provides a container abstraction allowing to incorporate virtually any technology. One container could run Spark, another could run Storm, 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, a preview of the base platform can be installed using Vagrant. An automated installation of the base platform is available in the form of a Chef cookbook.

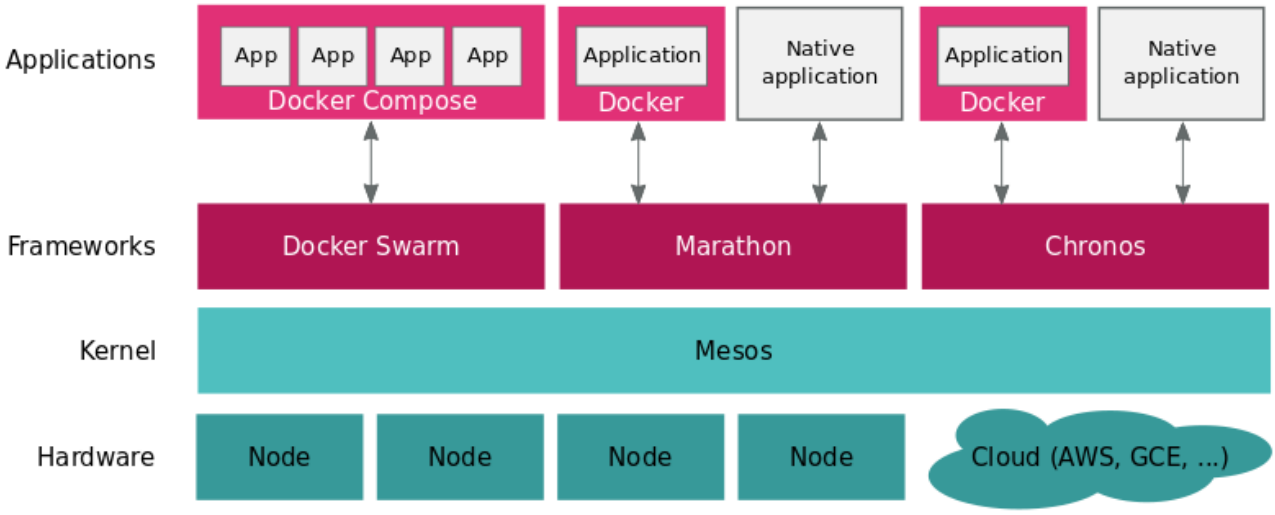

Figure 1: BDE platform architecture

An application running on the BDE platform can be seen as a pipeline consisting of multiple components, which are wired together in order to solve a specific Big Data problem. The components will be packaged in Docker containers. In order to encourage technology reuse and to facilitate the development of a component, the BDE platform will provide 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. Currently, a Spark base image is available together with a Spark demo application that extends this base image. More base images will be developed in the second year. The selection of technologies that will be provided as base images 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.

In the second year, we will gather experience with the main technologies and develop support in the form of base Docker images. We will also follow-up the evolution of Docker and its ecosystem to validate whether tools as Docker Compose and Docker Swarm could be integrated in the platform to facilitate the setup and deployment of Big Data pipelines. Finally, some topics that are not yet elaborated, like monitoring, logging and GUIs, will also be investigated in more depth.