EUBRA BIGSEA

EUROPE - BRAZIL COLLABORATION OF BIG DATA SCIENTIFIC RESEARCH THROUGH CLOUD-CENTRIC APPLICATIONS

D8.1 – Assets, Market Analysis & Stakeholder Synergy Plan

Author(s)	Roberto Cascella(TRUST-IT)
Status	Final
Version	1.0
Date	11/01/2017

Dissemination Level

- [x] PU: Public
- [ ] PP: Restricted to other programme participants (including the Commission Services)
- [ ] RE: Restricted to a group specified by the consortium (including the Commission Services)
- [ ] CO: Confidential, only for members of the consortium (including the Commission Services)

Abstract:
This report provides an analysis of assets identified in period one, expected assets resulting from period two; and an analysis of the markets in EU and Brazil, and plan to engage stakeholders identified for each asset.

EUBra-BIGSEA is funded by the European Commission under the Cooperation Programme, Horizon 2020 grant agreement No 690116. Este projeto é resultante da 3a Chamada Coordenada BR-UE em Tecnologias da Informação e Comunicação (TIC), anunciada pelo Ministério de Ciência, Tecnologia e Inovação (MCTI)

Document identifier: EUBra–BIGSEA–WP8-D8.1	
Deliverable lead	TRUST-IT
Related work package	WP8
Author(s)	Roberto Cascella(TRUST-IT)
Contributor(s)	Timea Biro (TRUST-IT), Lorenza Panelli (TRUST-IT), Marco A. S. Netto (IBM), All partners via the asset analysis
Due date	31/12/2016
Actual submission date	11/01/2017
Reviewed by	Marco A. S. Netto (IBM), Enrique Bayonne (UPV)
Approved by	PMB
Start date of Project	01/01/2016
Duration	24 months
Keywords	Market analysis; Asset identification; Stakeholder engagement

Versioning and contribution history

Version	Date	Authors	Notes
0.0	28/11/2016	Roberto Cascella (TRUST-IT)	Table of Contents
0.1	02/12/2016	Marco A. S. Netto (IBM)	Initial market analysis
0.2	13/12/2016	Roberto Cascella, Timea Biro, and Lorenza Panelli (TRUST-IT)	EUBra-BIGSEA assets: integrated information collected from partners
0.3	15/12/2016	Roberto Cascella, Timea Biro, and Lorenza Panelli (TRUST-IT)	Competitors analysis and update of the market analysis
0.4	16/12/2016	Marco A. S. Netto (IBM)	Update of the market analysis
0.5	16/12/2016	Roberto Cascella, Timea Biro, and Lorenza Panelli (TRUST-IT)	Stakeholders engagement plan, Market analysis and Conclusions
0.6	21/12/2016	Roberto Cascella, Timea Biro, and Lorenza Panelli (TRUST-IT)	Addressed comments from the reviewers and added glossary
0.7	03/01/2017	Ignacio Blanquer (UPV)	Revision
1.0	11/01/2017	Sandro Fiore (CMCC), Daniele Lezzi (BSC), Danilo Ardagna (POLIMI), Marco Vieira(UC), NÃądia P. Kozievitch (UTFPR), Timea Biro (TRUST-IT)	Addressed comments, Updates, Final version

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Big Data analytics offers to both private as well as public sectors the promise to provide valuable insight to that can create competitive advantages, foster innovation and scientific discovery and drive efficiency and progress across multiple domains and industries.

EUBra-BIGSEA addresses multiple challenges related to the composition of Big Data applications and their deployment on cloud platforms. The envisioned Big Data services are based on the concept that computing resources are made available to data processing by providing intelligent scheduling for efficient computing, QoS and privacy and security requirements. The value-add of this is demonstrated through real user scenario with high social and business impact, and of interest for both Europe and Brazil, processing of massive data coming from highly connected societies.

The present deliverable is the first one produced under the project WP8 focusing on "Long term Co-operation, Sustainability & Commercial Impact." The document aims to provide an initial individual assessment of 10 project exploitable assets, the market analysis including competitors, stakeholder mapping and engagement plan targeting the customer segments for both the individual assets and new tools and services. The report builds on the findings of the EUBra-BIGSEA D2.2 "User communities engagement and dissemination strategy" and has been developed in parallel with the D8.2 that focuses on the preliminary exploitation plans of the project's assets.

The EUBra-BIGSEA framework integrates 11 main assets focusing on QoS, privacy and security and data analytics. 8 of them are completely developed in the frame of the project and 3 of them result from the evolution and integration of existing components. Lemonade and COMPSs offer a higher abstraction level to application developers extending the capabilities of data analytics programming languages such as Python and Spark, with the data processing capabilities of an enhanced Ophidia. The EUBra-BIGSEA performance guarantee for Big Data applications integrates EC3 with new developed services and commonly used monitoring frameworks to provide performance guarantees for the execution of Big Data applications in a cloud environment. On top of that, EUBra-BIGSEA develops Data Quality and Entity Matching as a service to facilitate the development of Data Analytics applications for Smart Cities. The security and privacy infrastructure offers new capabilities for the development of secure services for Big Data.

The consolidated version, D8.3 "Business & Exploitation plan", due in December 2017, will provide a complete analysis of the assets and defined business and exploitation plans based on potential market and adoption prospectives.

1 INTRODUCTION

1.1 Scope of the document

EUBra-BIGSEA demonstrates the value of using cloud services on applications with high social and business impact, addressing main scenarios of high interest for both Europe and Brazil: processing of massive data coming from highly connected societies. It also demonstrates the value of developing Big Data services for capturing, federating and annotating on the order of PB of data on top of efficient programming models. These Big Data services impose multiple challenges on resource provision, performance, Quality of Service and privacy on a cloud infrastructure and services to support Big Data applications.

The main outcomes of EU-Brazil collaborative work are promoted and disseminated through *WP2 – Community Engagement, Communication & Impact*, with targeted activities for the research community, policy makers, the public and businesses. *Long term Co-operation, Sustainability & Commercial Impact* focuses on analysing the exploitable assets, market positioning and structure, and on concrete exploitation plans.

This deliverable is the first one of WP8 with the purpose of providing an initial analysis of the exploitable assets, and defining a stakeholder engagement plan targeting the customer segments for both the individual assets and new tools and services, including messages and formats best suited for them. With the focus on the general uptake of technologies, this deliverable analyses the market landscape of the relevant sectors for EUBra-BIGSEA to position the preliminary identified assets. The outcome of this deliverable supports D8.2 that focus on the preliminary exploitation plans of the project's assets and produced in parallel.

Future work will focus on providing an updated analysis of the assets, further potential markets, outcomes of the stakeholder engagement plans, final exploitation plans and sustainability potential. This will be the focus of D8.3, due in December 2017.

1.2 Document structure

This document is structured as follows:

Section 2 presents the exploitable assets resulting from the collaboration, in terms of components and applications. It describes the main features and benefits, and provides instructions on usage for researchers or developers outside the project. It also explains briefly the intangible benefits, such as new expertise and knowledge stemming directly from the project. Finally, it outlines the current status of developments and the collaborative work undertaken on specific outputs.

Section 3 focuses on the analysis of the market sectors that will benefit from the EUBra-BIGSEA technology.

Section 4 centres on market positioning of EUBra-BIGSEA technologies, including an analysis of the main competitors.

Section 5 provides the stakeholder engagement plan, mapping the main target segments and the key actions to be taken.

Section 6 covers the conclusions to the present report and plans for the second period.

2 EXPLOITABLE ASSETS RESULTING FROM COLLABORATION

In this section we first analyse each component, and derive the main message for each tangible asset based on key features. We briefly outline the potential intangible assets of the project such as the knowledge creation gathered for the design and development of the EUBra-BIGSEA components. Then we discuss the results of the collaboration in terms of joint and collaborative development of the real user scenario leveraging the components of the project.

2.1 Features and benefits of components

In this section we analyse each component in terms of its overall value proposition (how it helps tackle a specific problem), a snapshot of the main target user groupings, key features and benefits, and essential information for potential adopters.

As shown in Figure 1, the EUBra-BIGSEA framework integrates ten components plus an already identified composite service for elastic cloud management (highlighted as such in a different colour shade), providing 4 different types of services : cloud, programming models, Big Data analytics, security & privacy. Each partner has defined a draft exploitation plan and has also indicated which collaborative partnerships can support their specific exploitation goals.

Each component is then detailed and discussed in each of the following sections.

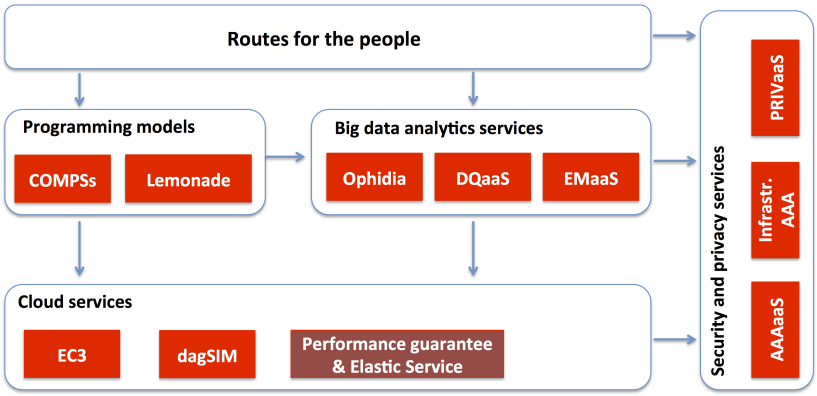

Figure 1: High-level mapping of the EUBra-BIGSEA services

2.1.1 COMPSs

COMPSs is a programming framework that aims to facilitate the parallelisation of existing applications written in Java, C/C++ and Python scripts. For that purpose, it offers a simple programming model based on sequential development in which the user is mainly responsible for identifying the functions to be executed as asynchronous parallel tasks. A runtime system is in charge of exploiting the inherent concurrency of the code, automatically detecting and enforcing the data dependencies between tasks and spawning these tasks to the available resources, which can be nodes in a cluster, clouds or grids. In cloud COMPSs provides scalability and elasticity features allowing the dynamic provision of resources.

The transparent deployment of COMPSs applications in the cloud is delegated to the PMES service that deals with the intricacies of deployment, contextualization operations and the installation of the application packages and required libraries, as well as to the monitoring of the process. A dashboard is also available for the configuration of the user cloud environment.

In a nutshell, the main added value of COMPSs is its focus on different capabilities in the same framework with a low learning curve as developers do not have to deal with application programming interfaces (APIs). What is more, the development and execution of the applications is not restricted to a proprietary infrastructure as interoperability is a key feature.

Target customer segments/user groups.

COMPSs has been adopted and extended in many scientific projects offered as a tool to develop scientific applications and optimize their execution on distributed infrastructures such as VENUS-C (20 user communities), Optimis, EUBrazilOpenBio, EUBrazilCloudConnect and is now leveraged in European Grid Infrastructure (EGI) as enabling technology for the execution of composed workflows on the federated cloud testbed. In EGI Fed Cloud COMPSs has been adopted to implement services for the biodiversity community.

The COMPSs group is collaborating with the BSC Life Sciences department for the development of Virtual Research Environment (MuG) that aims to answer the needs of the emerging worldwide 3D/4D genomics community where the COMPSs framework is used to implement the computational services to provide transparent access to applications and genomic data.

Key features and benefits.

- Automatic parallelization of sequential code without the need to adopt any specific Application Programming Interface (API). Support to Java, C/C++ and Python. The same code can be executed transparently with regards to the underlying infrastructure.
- Automatic scaling and elasticity features so the number of available resources can be adapted to the actual execution needs.
- Interoperability with different cloud providers to run computational loads on multi cloud environments without the need of code adaptation.
- Availability of tools that ease: the COMPSs applications implementation by means of an Integrated Development Environment (IDE); the application deployment in distributed infrastructures by means of the Programming Model Enactment Service (PMES); and the monitoring of executions by means of the Monitoring and Tracing tools.

Essential information for potential users

- COMPSs and PMES are constantly maintained and updated. The software is available as install packages and source code. They are also available on the EGI Application Database (AppDB) for virtual appliances.
- Open source and licensed under Apache 2.
- The installation of the packages automatically resolves the dependencies.
- COMPSs does not include any API, the idea is that the user code is optimized by the runtime. The user is only required to provide information on the tasks composing the application. Tutorials and manuals are available on the BSC website http://www.bsc.es/computer-sciences/grid-computing/comp-superscalar.

2.1.2 Ophidia

Ophidia (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing Big Data challenges in eScience. It exploits advanced parallel computing techniques and a hierarchical storage organization to execute intensive data analysis over multi-terabytes datasets.

Ophidia provides a Big Data analytics framework for parallel I/O and the analysis of multi-dimensional datasets. It leverages the datacube abstraction and comes with an extensive set of OLAP-oriented parallel operators, supporting e.g. datacube sub-setting, datacube aggregation, NetCDF file import and export, datacube intercomparison. Additionally it provides several primitives to operate on n-dimensional arrays that allow, for example, sub-setting, data aggregation, array concatenation, algebraic expressions, predicate evaluation, statistical analysis and regression.

Target customer segments/user groups.

Ophidia is used mainly in scientific sectors like in the climate change domain. It has been extended and used in several research projects like: FP7 EUBRazilCloudConnect, FP7 CLIP-C and H2020 INDIGO-DataCloud.

Key features and benefits.

- Big data analytics framework for scientific multi-dimensional data.
- Hierarchical storage organization to partition and distribute data across multiple nodes.
- Server-side approach with multiple standard interfaces.
- Wide set of parallel (MPI-based) operators and array-based primitives for data analytics.
- Operators for metadata, provenance and search & discovery.
- Workflow support or data analytics experiments

Essential information for potential users

- The latest Ophidia release is v0.10.6 (released in October 2016).
- Open source framework released under the GPLv3 license.
- It can be installed on Linux Debian/RedHat-based operating systems. Most libraries and tools dependencies are automatically solved when installing the binary packages, while MySQL server and Slurm should be manually installed and configured. This is being addressed (as cluster deployment in a cloud environment) in the frame of EUBra-BIGSEA and full automatic installation will be feasible by the end of the project.
- Ophidia can be exploited by users through the Ophidia terminal (shell-like client) or PyOphidia (python bindings). To support the user, the terminal provides auto-completion features and an online manual for all commands and operators available. Web documentation provides a quick start guide (http://ophidia.cmcc.it/documentation/general/quickstart.html) to start using the framework swiftly. Additionally a comprehensive user guide describing all the features, operators and primitives is also available in the documentation (http://ophidia.cmcc.it/documentation/users/index.html). A youtube channel has been also setup to help end users (https://www.youtube.com/user/OphidiaBigData/).

2.1.3 EC3

Elastic Cloud Computing Cluster (EC3) (http://servproject.i3m.upv.es/ec3/) is a tool to create elastic virtual clusters on top of Infrastructure-as-a-Service (IaaS) providers, either public (such as Amazon Web Services, Google Cloud or Microsoft

Azure) or on-premises (such as OpenNebula and OpenStack). EC3 offers recipes to deploy TORQUE (optionally with MAUI), SLURM and MESOS clusters that can be self- managed with CLUES. It starts with a single-node cluster and working nodes are dynamically deployed and provisioned to fit increasing load (number of jobs at the LRMS). Working nodes will be undeployed when they become idle. This introduces a cost-efficient approach for Cluster-based computing. The development has been performed completely by the I3M-UPV.

Target customer segments/user groups.
EC3 is currently used in different scientific and industrial sectors, as it is a horizontal tool capable of deploying general purpose clusters. UPV has experiences of use in Computational Chemistry and Big Data Analytics, among others.

Key features and benefits.
There are several tools in the market for creating self-managed virtual clusters. However, they rely on existing VMIs that the user has to maintain and customize. EC3 can be used on top of generic VMIs (e.g. an UBUNTU 14.0, a CENTOS 6.5, etc.) and reusing the same installation recipes in different cloud platforms with minimal changes. EC3 increases portability, interoperability and reduces vendor lock-in. Moreover, it uses Ansible for the configuration of the software, leveraging from the huge community and existing recipes.

Essential information for potential users

- Current version is in production stage, expecting to have new improvements during the EUBra-BIGSEA lifetime to increase performance (self creation of base images), interoperability (TOSCA support), integration with other components of the project and the support of ahn extended catalogue .
- Open source software released under the Apache 2.0 license.
- EC3 is a lightweight service and it needs to be maintained to give support to the new versions of the external software supported.
- It uses configuration recipes that can be stored in the Ansible Galaxy repository (not necessarily) or locally. It will also support shortly a local repository of components to facilitate the maintenance of the recipes.
- Extending the service with the recipes for the user own software requires a training in Ansible, yaml and similar DevOps concepts and tools.

2.1.4 EUBra-BIGSEA performance guarantee for Big Data applications

This asset is fully developed and integrated in the framework of the EUBra-BIGSEA project and it is based on the combination of three key components: 1) EC3 which automates the deployment and the initial configuration of a Big Data application and provides also mechanisms for runtime re-configuration (developed by Universitat Politecnica de Valencia), 2) a rule-based module for pro-active run-time policies specification and execution (developed by Federal University of Campina Grande), 3) a module implementing optimization based policies able to identify the deployment configuration of minimum costs that provides also performance guarantees (e.g., jobs are executed within a time limit or data streams are executed with no loss at a given rate, developed by Politecnico di Milano and Federal University of Minas Gerais).

Target customer segments/user groups.
Currently, these three components are used within academic environments and are developed and tested within lab deployments.

Key features and benefits.
This asset allows to run Big Data application providing a priori performance guarantees. Since the general trend is to run Big Data applications in cloud environments where resource contention is significant, providing a solution that is able to cope with such issues is of paramount importance. The sets of components and services provided will be able to consider more information about the application in order to enable a smarter initial configuration for periodic jobs and a fast reaction time for the case of disturbances during the execution.

Essential information for potential users

- The three components are developed separately at the four Institutions (UPV, UFCG, Polimi and UFMG). EC3 is more mature and partially in production (basic case). The three components will be integrated during the second year of the project.
- The open source license scheme is still under definition. Most of the components will rely on Apache 2.0, unless the dependencies to other libraries introduced in the implementation are not compatible with Apache 2.0.
- Maintenance is performed on a voluntary basis and it depends on the number of technologies (e.g. COMPSs, Spark, etc.) that need to be supported and integrated within the runtime platform.
- The modules are being developed considering an OpenStack cloud environment and the usage of OpenStack Monasca as a monitoring tool.
- The initial version of the optimization component will be partially based AMPL (a framework for mathematical programming) and IBM CPLEX. The final version will be migrated to the open source frameworks jCMPL and CoinOR and will be released as Open source.

2.1.5 DQaaS

DQaaS (Data Quality-as-a-Service) is a service that aims to provide information about the quality of a requested dataset. Data Quality helps applications and users in understanding the degree with which a dataset is suitable for their goals. In particular, considering a dataset, the service (i) offers the access to different quality metrics periodically evaluated and (ii) allows applications and users to define and assess personalized quality metrics.

Key features and benefits.

DQaaS is designed for dealing with Big Data, thus it addresses volume and velocity requirements. In particular, the algorithms will be developed on architectures able to support parallelization and when applications/users request real time quality analyses, only on a sample of data will be considered. This choices aim to reduce the impact that such service can have on the system performance.

Essential information for potential users

- DQaaS is currently still under development and some preliminary tests have been conducted in the academic environment. A first release is planned by April 2017.
- At the moment DQaaS uses as input data the data sources available for EUBra-BIGSEA use case. Open data can also be used.
- The results that are expressed in terms of data quality dimensions.

2.1.6 EMaaS

Entity Matching-as-a-Service (EMaaS) targets the problem of identifying records that refer to the same entity of the real world. This task is known to be challenging due to its pair-wise comparison nature, especially when the datasets involved in the matching process have a high volume (Big Data). Since the EM task has critical importance for data cleaning and integration, e.g., to find duplicate points of interest in different databases, studies about challenges and possible solutions of how EM can benefit from modern parallel computing programming models, such as Apache Spark (Spark), have become an important demand nowadays. For this reason, the EMaaS service, to be provided by the main API of the EUBra-BIGSEA, consists of a bag of tools and functions that can process the Entity Matching task (e.g., geo/spatial- matching) in parallel by using Apache Spark.

The EMaaS service will attend the requests from applications/systems interested in submitting Entity Matching tasks to the cluster environment. To this end, the service will establish a connection to the Hadoop Eco-system to perform the necessary operations such as submitting artifacts (e.g. datasets) to the HDFS or starting the execution of Spark jobs.

Target customer segments/user groups.

EMaaS targets every scientific/industrial sector interested in a service to process large-scale geo/spatial data matching.

Key features and benefits.
EMaaS is designed for dealing with Big Data and provides a set of Spark approaches to enable the large- scale matching of various geographic entities (e.g., trajectories, streets, city facilities).

Essential information for potential users

- EMaaS is currently still under development. There are four use cases defined and two of them are in progress. The EMaaS release is scheduled to the end of 2017.
- EMaaS requires a cluster of machines configured with an Apache Spark and the desired environment needed to keep the client data and the service's output.

2.1.7 dagSIM

dagSim is a discrete event simulator working on a DAG corresponding to a Map/Reduce Tez and Apache Spark models which is able to estimate Big Data applications performance efficiently. MapReduce is capable of analyzing very efficiently large amounts of unstructured data and it has been adopted in multiple application domains, e.g., machine learning, graph processing, and data mining.

In this context, one of the main challenges is that the execution time of a MapReduce job is generally unknown. Because of this, predicting the execution time of Hadoop jobs is usually done empirically through experimentation, requiring a costly setup. In alternative, it is possible to develop models and software tools for predicting performance. dagSIM addresses this issues.

Target customer segments/user groups.
Currently, dagSim is used within academic environments, though any industrial sectors making use of Big Data could deploy it: Communication, Life Sciences, Manufacturing and Financial Services, and Smart Cities just to mention a few.

Key features and benefits.
dagSim is characterized by the following strategic features: Fast performance, user friendly, no need to train the users and currently no competitors (being its application target highly specific). Existing literature has been focusing on Hadoop 1.0 only, therefore considering Map/Reduce without DAGs and allocating statically the containers to the two stages, while dagSim, performs a dynamic allocation of the resources reflecting a more realistic scenario.

Essential information for potential users

- dagSIM is currently still under development. The internal releases implement different event allocation policies in order to get closer to Big Data and High Performance computing frameworks using models different than DAGs.
- dagSIM uses a model of the DAG, which should be provided manually.
- The tool uses currently LUA language (https://www.lua.org), which however is not an essential component and can be removed.

2.1.8 Lemonade

Lemonade (Live Environment for Mining Of Non-trivial Amount of Data from Everywhere) is an analytics platform that supports intuitive definition of tasks for knowledge discovery, mining, and learning from large amounts of data that come from a wide spectrum of scenarios. The platform interface is a web application in which users may define analytics workflows visually by dragging and dropping operations and data sources, and connecting them. Lemonade is being developed by UFMG as part of the EUBra-BIGSEA project and targets users who do not want to learn a programming language, but need to develop analytics workflows.

Target customer segments/user groups.
Lemonade targets those users from areas such as Mathematics, Statistics, Business Administration, as well as Data Science practitioners from any knowledge area.

Key features and benefits.

Lemonade provides a rich web interface, which is both accessible to learners and powerful to experts. Lemonade scope plan comprises more than 30 different operations of data mining, machine learning and extraction, transformation and loading of data. The platform is also capable of processing massive amounts of data ("Big Data"), since it is being built on top of three scalable processing and storage technologies: Apache Spark, CMCC Ophidia and BSC COMPSs, being the last two technologies developed by partners of EUBra-BIGSEA project.

Users will be able to upload data sets using a service provided by Lemonade. Data are kept in a redundant file system, aimed to provide high-availability and high throughput. Data storage requirements will depend on use cases and installation. Users may process terabytes of data and their volume will directly impact the storage and processing costs.

Lemonade can be scaled to support hundreds of users by increasing cluster capacity. A large number of users can be supported in a modest cluster of commodity computers and a volume of data often found in most of organizations.

Essential information for potential users

- Currently, Lemonade is being developed and a first prototype version was demonstrated to EUBra-BIGSEA partners by UFMG in Sep, 2016. An alpha version is planned to January, 2017.
- Lemonade is an open-source solution. All dependencies (operating system, processing frameworks, infrastructure technologies) are also open source, so there are no licensing costs. The license scheme is under discussion and it will be finalised for the first release.
- To be kept up and running, Lemonade requires a cluster of processing computers and data storages. The size and capacity of the cluster depends on the number of users, data volume and complexity of workflow/tasks.
- Lemonade depends on Apache Mesos (standalone mode) or a distributed processing technology (Apache Spark, BSC COMPSs or CMCC Ophidia), Oracle MySQL database server and a Linux operating system distribution.
- Lemonade requires a reliable infrastructure to run that may be provided by platform-as-a-service (PaaS) companies, such as Google, Amazon or Microsoft or by the organization using Lemonade.
- Three different user roles are supported in Lemonade: a system administrator, a data scientist and a data explorer. System administrator will be responsible for keeping Lemonade running, adding new users, setting permissions and security, and managing data sets. Data scientists must know about Lemonade operations in order to create processing workflows and data being processed, their characteristics and how his/her results can be applied in a real scenario. Data explorers are users who uses existing models/ workflows (built by data scientists) to process their data and create new data sets, visualizations or web services.

2.1.9 AAAaaS

AAAaaS is a software component that provides a set of libraries and tools for application developers in need of Authentication, Authorization and Accounting (AAA) services within the scope of their applications (e.g. to authenticate and authorize the end-users of that (or those) application(s).

These AAA services can be deployed and used directly by the software developer, per application or per application set. Eventually, they may also be provided "as-a-service" by an external entity – although this is not the expected usage model in the scope of EUBra-BIGSEA.

The software provides the general functionalities of traditional AAA and Identity and Access Management (IAM) services, including interfacing with external identity providers, but is deployable and manageable according with cloud principles such as scalability, elasticity and resilience: service instances are automatically deployed in line to the application needs; scale-up/scale-down operations are performed in reaction to workloads; and the lifecycle management is performed using cloud paradigms.

Target customer segments/user groups.

AAAaaS is being developed to be used in the EUBra-BIGSEA use cases. More specifically, it was designed to be used in scenarios where application developers need to use infrastructure-independent AAA and IAM services for their own applications. In the future, AAAaaS (or parts of it) may be made available to the community to be used in other contexts and other sectors.

Key features and benefits.
The main distinguishable feature of AAAaaS will be the already mentioned support for deployment and management according with cloud principles such as scalability, elasticity and resilience. The other features are similar to those provided by current AAA toolkits. In practice:

- AAAaaS provides typical AAA and IAM services using cloud principles for deployment and lifecycle management.
- AAAaaS supports interfacing with external identity provider services.
- AAAaaS supports most relevant protocols for authentication, authorization and identity management.
- AAAaaS supports cloud-based deployment and management.
- AAAaaS is planned to be released as an open source contribution to the community.

Essential information for potential users
- AAAaaS is still under development. The release of the asset is planned in June 2017.
- AAAaaS only needs to be configured by the application developer.
- AAAaaS instances can be deployed per application (one application developer using it) or per application set (one or multiple application developers using it).

2.1.10 Infrastructure AAA

EUBra-BIGSEA Infrastructure AAA is a software component that provides a common Identity and Access Management (IAM) service interface to the EUBra-BIGSEA Infrastructure resources, independently of the underlying cloud framework (e.g. OpenStack, CloudStack, commercial frameworks).

This service corresponds to a high-level abstraction layer, mapping with and extending the native IAM features of each cloud framework to be supported by the EUBra-BIGSEA platform. In practice, it may be considered as an extension component for UPV's IM (Infrastructure Manager) (http://www.grycap.upv.es/im/index.php), providing a high-level abstraction layer for AAA services controlling the access to cloud resources.

Target customer segments/user groups.
Infrastructure AAA is being developed to be used in specific scope of the EUBra-BIGSEA framework. More specifically, it was designed so that application developers/providers have a single, unified authentication and authorization interface for accessing underlying cloud resources when deploying and managing their applications. In the future, it may be made available to the community to be used in (or, most likely, adapted to) other contexts and other sectors.

Key features and benefits.
Infrastructure AAA provides a common, unified access point for authentication and authorization when accessing underlying cloud resources, abstracting the specificities of each of the encompassed cloud frameworks.

Essential information for potential users
- The asset is still under development. The release of the asset is planned in June 2017.
- Infrastructure AAA will integrate with the envisaged underlying cloud frameworks (CloudStack, OpenStack, others). Usage with other frameworks will require adaptation work.
- Infrastructure AAA will operate at infrastructure level, and its end-users are the EUBra-BIGSEA application developers/providers.

2.1.11 PRIVAaaS

PRIVAaaS is a software toolkit that provides a set of libraries and tools that allow to control and reduce the data leakage in the context of Big Data processing and, consequently, to protect sensible information that is part of the EUBra-BIGSEA framework.

The process is divided into two perspectives which model different aspects of the anonymization problem: the first perspective is related to the anonymization of the loaded input data, while the second is related to the anonymization of the data resulting from the data processing algorithms. The result is output data that is anonymized for the intended usage scenario.

The process starts with the definition by users who have privacy knowledge of policies that will guide the anonymization process according to a set of rules which are implemented by an ontology. To maximize data utility while preserving low levels of disclosure risk, these policies govern 2 key anonymization phases: 1) the anonymization of raw data and how each algorithm can use each set of data; and 2) the anonymization of the data provided by the analytics algorithm to the end user, to avoid that knowledge that is extracted by the algorithm is unduly accessed.

Target customer segments/user groups.
PRIVAaaS has been developed to be used in the EUBra-BIGSEA use cases. Concretely, it was designed to be used in scenarios that involve the processing of massive amounts of data that may contain or lead to privacy-sensitive information.

Furthermore, as data anonymization is currently relevant in several sectors, PRIVAaaS can be used in all sectors that take advantage of Big Data analytics techniques: as government, health, private companies, etc.

Key features and benefits.

- Implements different anonymization techniques and models.
- Off-the-shelf component, requiring minimal configuration to be used.
- Based on open source dependencies.
- Data can be uploaded through json/csv files.
- The anonymization policy guides the anonymization process, which can be done automatically, letting privacy specialists free from this task.
- The ontology describes a vocabulary for anonymization and can be adapted to different policies for different scenarios.

Essential information for potential users

- PRIVAaaS is still under development. A release is foreseen for March 2017.
- The anonymization process implemented by PRIVAaaS is based on data definition. So, it is necessary to know the datasets in which it will be applied.
- Users need some knowledge about privacy and anonymization to select and classify correctly the attributes (identifiers, quasi-identifiers, sensible) and the anonymization techniques to be applied to each attribute, as each one has its particularity.
- Future releases of the tool plan to automate the anonymization process, based on predefined rules described by the policy.

2.2 Benefits of the smart city application

The main use case of EUBra-BIGSEA is the management of information related to public transportation on cities. One of the products will be an application which implements concepts related to routes for people. This application leverages descriptive and predictive models to provide city dwellers and urban planners information about location, traffic and spaces in the city. Citizens will use this application through web and mobile interfaces to plan their trips in the city taking into account not only travel time, but also more subjective measures of the space in which they will travel. Such measures may include landscape aesthetics, how pleasant the routes are, connections with bus terminals or traffic. As such, a person may find that a slightly longer route to work trades off a few minutes of extra commuting for a drive through a park, which is often associated with sentiments of calm and happiness by city dwellers in general.

Target customer segments/user groups.
The smart city application primarily targets two groups of end-users: citizens and urban planners associated with the municipality.

From a citizen perspective, mobile and web interfaces will allow the citizen to query for the state of the route options available for a given trip. Once the citizen manifests a travel need, the system will provide multiple route options that maximize different criteria in addition to travel time, such as likely stress, pleasantness, interestingness and liveliness of the routes (according to parameters such as day of the week, hour and location). Such route recommendations will be built processing both long-term historical data to identify trends, and short-term data focusing on recent and relevant events. Furthermore, recommendations will be provided through state-of-the-art and novel predictive models from Data Science.

From the urban planning perspective, the smart city application will provide a descriptive view on the state of the mobility in the city as a whole, identifying its status, trends and the impacts of relevant events. Contrasting with the citizen view, the goal of the application in this perspective is to use data mining descriptive models and to allow urban planners to interactively explore it, as well as to provide notifications and alarms for these stakeholders. This view will inform urban planners in both operating and planning transportation systems in a more effective way.

Essential information for potential users

The application is still not fully developed. The initial parallel prototypes are evolving toward the direction of data integration, which covers the general problem of mechanisms for collecting, cleaning, transforming and integrating all the listed data sources, in order to understand the dynamics of traffic and transportation public services in Brazilian cities. At the moment the initial prototype use PostGIS and GoogleMaps services and integrate the open data from Curitiba municipality for testing purposes. Despite that the EUBra-BIGSEA pilot has been initially planned over the data of the city of Curitiba, where the pilot case is being constructed, the EUBra-BIGSEA framework will support other European and Brazilian cities.

2.3 Benefits of intangible assets

Intangible assets are important for the value creation of an organisation, and refers to patents, copyrights, knowledge that are part of the exploitable resources and play an important role in the decision making process of the organisation. In the context of an innovative project such as EUBra-BIGSEA, the term intangible asset is referred to the knowledge creation, new scientific and technical concepts, thanks to the research results achieved and demonstrated by the advancements in the field.

On the one hand the knowledge created has been transferred and applied for the development of the project assets, both architectural components and the smart city use case. On the other hand, it has been disseminated to the scientific community and to young researchers thought scientific papers and presentations. Thus, we identify a set of intangible assets below listed:

Scientific data management. Partners have acquired relevant expertise in the several domains related to data management: cloud, grid and high performance computing, and Big Data analytics. The project has not generated new data, but on the other hand it is has generated new knowledge in managing different and heterogeneous data sources available in different areas and serving different purposes, such as the smart city application.

Data Quality. Partners have studied deeply the inconsistencies, challenges and data formats from different Geographic data providers, which is of uttermost value for developing accurate applications for route planning. Project Deliverable 7.2 [17] reports on the data challenges addressed, the integration process and scenarios.

Data formats and standards. Partners have increased their understanding and relevance in standards such as TOSCA and RDA groups, which constitutes an important opportunity for contributing to the global adoption of such standards and best practices.

The creation of value from these intangible assets is not an immediate process as it requires understanding the value in a negotiable form, i.e., exchange knowledge for money in the form of product and services or trade with other knowledge. The project's intangible assets do not generate revenue streams directly, but they might be exploited for generating future knowledge and development of new tangible assets leveraging the existing knowledge. Specifically, the knowledge generated in the framework of the EUBra-BIGSEA will serve to boost and sustain future research activities with strong innovation in other frameworks (national, international, and industrial), and it is the bricks to educate students and scientist at large.

2.4 Status of developments

This section briefly outlines the development roadmap of the project's assets.

COMPSs. The current release of COMPSs is 2.0 and it includes the integration with the Mesos framework, the support to run MPI workloads as tasks, and streaming support. In addition COMPSs provide following features: Runtime; Support for Dockers; Support for Chameleon; Object cache for persistent workers; Improved error management; Connector for submitting tasks to Marenostrum supercomputer from external COMPSs applications; Bug-fixes. An additional Enhanced Tracing mechanism has been implemented with reduced overhead using native Java API, support for communications instrumentation and for PAPI hardware counters.

Ophidia. The latest Ophidia release is v0.10.6 (released in October 2016). The next version will be released by the end of 2016. Additional releases will follow in 2017 with the features developed in the context of EUBra-BIGSEA project. These activities will include the extension of the Ophidia framework to elastically scale in a cloud environment based on the workload and in order to meet QoS levels, and integrate security support.

EC3. Current version is in production stage, expecting to have new improvements during the EUBra-BIGSEA lifetime to increase performance (self creation of base images) and interoperability (TOSCA support). Current version has been improved in the frame of the project with the enhanced support to Mesos, the development of configuration recipes for the whole stack of components and the capability of dealing with hybrid deployments (both elastic and static).

EUBra-BIGSEA performance guarantee for Big Data applications. The software integration is planned in 2017 and a release at the end of 2017.

DQaas. There is no current release of Lemonade. The first release is planned in April 2017.

EMaaS. EMaaS is currently still under development. There are four use cases defined and two of them are in progress. The EMaaS release is scheduled to the end of 2017.

dagSIM. There is no current release. The simulator is still under development and testing phase.

Lemonade. There is no current release of Lemonade. An alpha version is planned in January 2017.

AAAaaS. The component is still under development. The release of the asset is planned in June 2017.

Infrastructure AAA. The asset is still under development. The release of the asset is planned in June 2017.

PRIVAaaS. PRIVAaaS is still under development. A release is foreseen for March 2017.

2.5 Main collaboration outputs

EUBra-BIGSEA is a key driver for advancing scientific collaboration across borders, by improving existing components, offering a test ground for new developments, and designing and implementing new components and related services. It has fostered the concepts of open science coupled with a strong focus on open source and open standards.

Table 1 summarises the technical assets developed in the project. The development of the whole EUBra-BIGSEA framework will not be possible without the collaboration between Brazilian and European partners, who contributed with complementary tools and data. One of the main achievements of the collaboration has been the use and development of specific functionalities of existing components to ingrate additional features. 7 plus 1 composite out 11 are new components that have been developed to serve the scope of EUBra-BIGSEA as highlighted in Table 1.

Asset	EUBra-BIGSEA contribution
COMPSs	Component already available and developed by BSC. The component has been extended in EUBra-BIGSEA to include additional functionalities such as the integration of the Mesos framework and the support to run MPI workloads as tasks.
Ophidia	Component already available and developed by CMCC. The component has been extended in EUBra-BIGSEA to include additional functionalities such as elasticity in a cloud environment based on the workload and in order to meet QoS levels, and integrate security support.
EC3	Component already available and developed by UPV. The component has been extended in EUBra-BIGSEA to increase performance (self creation of base images) and interoperability (TOSCA support), as well as other features described above.
EUBra-BIGSEA performance guarantee for Big Data applications	New component fully developed in EUBra-BIGSEA thanks to a coordinated activity between UPV, UFCG, UFMG, and Polimi. It leverages three existing components brought by the project partners.
DQaaS	New component fully developed in EUBra-BIGSEA by Polimi.
EMaaS	New component fully developed in EUBra-BIGSEA by UFCG.
dagSIM	New component fully developed in EUBra-BIGSEA by Polimi.
Lemonade	New component fully developed in EUBra-BIGSEA by UFMG.
AAAaaS	New component fully developed in EUBra-BIGSEA by UC and Unicamp.
Infrastructure AAA	New component fully developed in EUBra-BIGSEA by UC and Unicamp.
PRIVAaaS	New component fully developed in EUBra-BIGSEA by UC and Unicamp.

Table 1: List of EUBra-BIGSEA components developed by project partners. All new components are highlighted as such.

3 MARKET ANALYSIS

This section introduces the market analysis related to the efforts of this project.

3.1 Cloud computing

Cloud computing has become an essential platform for several industries worldwide. Cloud became popular mainly for start-ups that wanted to deploy their products as a service. Over time, as cloud computing became more mature, companies of medium and large sizes also started to explore cloud to run their services. The reasons are mainly to reduce investments in new hardware, but also to reduce costs in labor and electricity bills. However, for many of those medium and large companies, the hybrid model became more promising since it can use the the already purchased on-premise resources, which can also host sensitive applications and data, but also use cloud for meeting peak demands.

According to a 2016 Gartner report, IT spending is moving from traditional sources, such as direct server and software license purchases, into public cloud computing services. The amount of money IT will spend on cloud in 2016 year is $114 billion, and will grow to $216 billion in the year 2020. Therefore the amount of is a total of $1 trillion over the course of the next five years.

IDC, cloud-based solutions will triple by 2020 with public & private clouds increasingly merging into hybrid clouds. IDC baseline scenarios indicate that cloud will help boosting the EU economy with the creation of over 300k new companies and up to 1M jobs starting from 2015. Scenarios show that more and more enterprises move to cloud and cloud open source will become increasingly relevant and a potential differentiating factor for Europe.

The study "Futurescape: Latin America IT Industry 2016 Predictions — Leading Digital Transformation to Scale" conducted by IDC shows that cloud computing is a key priority for executives. In Latin America, it is estimated an increase of 40% in investments in this area and should reach to US$ 3,6 billions. The Brazilian market will have large contribution to these numbers. According to Mario Rachid, Director of Digital Solutions in Embratel in Brazil, Cloud is being used as a main platform to deliver applications to companies of all sizes. The key demand of the sector in Brazil is on infrastructure services with capabilities for processing high volumes of data as well as resilience. There is a growing interest in hybrid cloud too.

According to the same 2016 IDC study in Latin America, companies are just starting their digital transformation journey. The study shows that 26% of large company CIOs say they are beginning their "digital journey" in 2016, moving from inconsistent use of digital experiences and products to repeatable, managed, and ultimately disruptive use of digital technologies and business models. The enabling technologies for this transformation comes from: Social, Mobile, BigData/Analytics, and Cloud.

Cloud-based Big Data and analytics (BDA) solutions are hitting the hype over the next five years with a spending prediction larger than on-premise solutions, pushing the need towards a unified platform architecture for data management and analysis.

In this project we contribute to the advances in the area of cloud computing. As highlighted in market analysis reports, hybrid cloud is a major trend, in addition to having data and computing power spread over different sites. One of the the aspects in cloud computing we focus in this project is resource orchestration, in particular in elasticity of resources, which could be on-premise, public, or in hybrid settings. We believe the findings and technologies developed in this area can contribute to the advances expected in cloud computing.

3.2 Big Data and data analytics

Another strong IT shift is Big Data and data analytics, which relate to solutions that are able to process high volumes of data. According to a 2016 IDC report, "Big Data and Analytics Spending Guide", the Big Data market can get to an amount of US$ 187 billions until 2019, with a growing of 50% in relation to the amount of US$ 122 billions in 2015.

IDC predicts that the "Western European Big Data market will grow by a compound annual growth rate (CAGR) of 24.6% between now and 2018 [12] with European organizations reaching a higher maturity despite small data sets and greater privacy concerns compared to North America. With respect to data sets, the cooperation with Brazil is strategic, where a large diversity of data is available ranging from biodiversity, climate, health, to massive connected societies and smart cities. IDC26 estimates that "50% of BDA investment will focus on citizen service, public health, Smart Cities, and reducing fraud, waste, and abuse." Worldwide, Navigant Research [22] estimates global investment in smart city technologies over the decade is expected to reach $174.4 billion with the global smart city technology market to be worth more than $27.5 billion annually by 2023, compared to $8.8 billion in 2014, with IBM maintaining its leadership in the sector [19].

Big Data is becoming essential in several regions and Brazil is not an exception. Big Data started to take more traction in Brazil in 2013, when companies started to provide more pervasive solutions for fraud detection analytics, real-time analytics with in-memory computing, and on-line and digital analytics, among others in sectors such as finance, telecoms, and retail.

Brazil represents the largest market in Latin America, where Big Data reached almost US$300 million in 2012 and it is forecast to grow around 25% per year until 2017. Frost & Sullivan [24] indicates that the Big Data market is still at its initial phase in Brazil and it is planned to grow very fast in the coming years with a prediction that "one third of Brazilian companies will have invested in Big Data analytics by the end of 2015". The key factor will be the development of more user-friendly data analytics tools and programming models for massive processing of real-time data, and the exploitation of elastic cloud resources.

Another relevant segment of Big Data in Brazil is the oil and gas industry, as it deals with large volumes of data generated by seismic analysis. This comes mainly from discoveries related to pre-salt.

Retail and e-commerce are also making use of Big Data in Brazil. Tools and services in these areas are used by Brazilian retailers to track and establish individual profiles of their customers to offer products with high probability of consumption.

More recently, Digital Agriculture started to adopt Big Data in Brazil, according to a recent report by McKinsey Global Institute. Big Data can bring gains of US$ 8 billions until 2019. As farmers and farming companies get more familiar with the Big Data technologies, more opportunities in agriculture will get traction.

Ernst & Young indicates 7 drivers for a successful adoption of Big Data by companies [26]: very relevant are the security and privacy issues, which on the one end might limit the processing of data, thus their value, and on the other end could be a driver to enforce privacy policies. The advance on SLAs and support for security and data privacy will smash the barriers that block organizations and SMEs in migrating to the Cloud and entrusting public and private data centers for their data. Europe can play a central role in the Big Data market thanks to the excellence from a number of innovation activities aimed at addressing Big Data challenges and to the experience maturated in different initiatives related to federated infrastructure, security, data portability and interoperability, legal aspects and standardization. The Strategic Research Agenda [15] of the Big Data Value Association identifies Security, Data Protection, Privacy and Trust as transversal topics that will have an impact on the Big Data ecosystem along the value chain: data generation and acquisition, data analysis and processing, data storage curation, and data visualisation usage & services.

Big Data is a central area we focus on in this project. Issues such as how we collect data, how we partition data to be processed, and how we define the amount of processing power required to meet the demands of services that leverage the data are relevant for several organizations and the technologies and findings in this project can definitely contribute to the expected advances in this area.

3.3 Privacy and security services

Privacy and security are still major issues for the adoption of public cloud services. The recent cloud security breaches and data leakage have increased the concerns of both governments and companies in trusting public cloud with their data, pushing more towards the adoption of hybrid cloud solutions. According to a Capgemini report, fear of security breaches and issues with data sovereignty is the top ranked impediment preventing the uptake of Cloud in Brazil [16].

In Europe the Digital Single Market strategy has defined a set of challenges that need to be addressed to revitalise the European industry with a strong focus on cloud computing and the data economy. To fully realise its benefits, the investment in cloud technology should be strongly coupled with a common understanding of best practices, such as security and data protection, to build the trust and confidence in cloud services that is currently lacking.

Security and privacy preserving mechanisms are the first step towards a more secure development and experimental environment. The European Commission is moving in this direction with the new definition of the General Data Privacy Regulation, which will entry into force in 2018. On the other side of the Atlantic, the Brazilian government is specifically addressing the need of defining specific laws aimed at the protection of citizens' privacy.

As such, security and privacy remain still of paramount importance with the recent trend of specific cloud access security brokers to monitor and manage private data outsources to the public cloud [13]. A recent survey [21] conducted by 451 Research in late 2015 over 1,100+ senior security executives indicates that there will be "an increased role for more granular access controls" as an intermediate solution for companies that on one hand would like to maintain local control over keys used to encrypt data and on the other hand delegate keys to the cloud service provider.

The same survey points out that compliance to regulations does not alway ensure security as companies reporting to be compliant have indicated past episodes of security incidents. 451 Research estimates that nearly $40 billion is spent annually on

information security products, but data breaches are still frequent. Spending in IT security is forecasted to still increase in the coming years to implement security best practices and be more competitive on the market [21].

EUBra-BIGSEA contributes clearly to the advances in cloud and data security solutions by designing and implementing fine grained access control mechanisms for both the cloud infrastructure and services. Privacy is also a central point of the project, specifically addressing scenarios that involve the processing of massive amounts of data that may contain or lead to privacy-sensitive information.

3.4 Smart cities

Cities need to transform to meet citizens expectations, attract investments, improve quality of life, and improve resilience. These transformations are key to achieve what is known as Smart Cities. There are several driving forces coming from Big Data, Analytics, the Internet of Things, Cloud Computing, and Mobile solutions that are enabling the development of services more efficiently in this area.

Smart cities have a high societal impact where EUBra-BIGSEA has its value in the use case addressing massively connected societies. Schneider Electric [14] predicts that 70% of the world's population will reside in cities, in less than 40 years bringing new concerns about the sustainability of the infrastructures caused by this massive growth: cities consumes 75% of the energy resources and are the cause of 80% of the carbon footprint. Traffic congestion has a large impact on the quality of life and sustainability of cities with a cost in Brazil of roughly $7.2 million in 1998, according to the National Association of Public Transport, and an increase of traffic by three nowadays, thus with far worse consequences [23].

According to a 2014 report from Frost & Sullivan [25], the global market of smart cities will be valued at US$1.565 trillion in 2020. The report also points out that over 26 Global Cities are expected to be Smart Cities in 2025, with more than 50% of these smart cities from Europe and North America. A report from MarketsAndMarkets estimates that IoT in Smart Cities can reach $147.51 Billion by 2020, at an estimated Compound Annual Growth Rate (CAGR) of 23.2% during the next five years. All the data generated by IoT sensors cannot be processed at required speeds and therefore optimized Big Data and cloud computing resources are essential to be able to leverage the data in real time services. One of the areas that is changing over time is the building technology, and other areas such as building automation for facility and emergency response management, parking management, building energy optimization are becoming adopted. There is growing interest also in green buildings, especially across Europe, North America, and Asian-Pacific regions. There is also a trend in reducing the energy consumption and attempts to avoid proliferation of environmental wastes in this area. Transportation is another focus area, especially in major metropolitan cities around the world where citizens may spend ours to go from one place to another.

Large companies are investing resources in Smart Cities. For instance, in 2014 IBM announced the support in China for deploying researchers in a 10-year initiative to transform energy systems. The project "Green Horizon" [20] is working in three areas critical to China's sustainable grows: air quality management, renewable energy forecasting and energy optimization for industry. One of the first partners join this effort was the Beijing Municipal Government. Through a collaboration agreement, the two parties have agreed to work together to develop solutions which can help tackle the city's air pollution challenges. The efforts are leveraging IBM's most advanced technologies such as cognitive computing, optical sensors and the Internet of Things all based on a Big Data and analytics platform, which also leverage IBM's deep experience in weather prediction and climate modelling.

In other to exercise the technologies in Big Data, analytics, and cloud computing in this project, we are focusing on an relevant problem for all efforts in smart cities, i.e., transportation. We are using real data collected from several Brazilian and European cities, as detailed in deliverable D7.2 [17]. Any advance in the area of Big Data and cloud computing that can enable an user construction and deployment of services that can help on transportation in cities has a key value for city managers and citizens.

4 EUBra-BIGSEA market positioning and potential services

4.1 Competitors and trends

The following competitors have been identified in relation to the components. Some of the components are still under development, as such the analysis of the positioning against the competitors will be refined in D8.3, due in December 2017.

Competitor: Pegasus Workflow Management System [9, 18]	
Short description	Workflow manager that automatically maps high-level workflows descriptions, provided as XML files, onto distributed resources as clouds. It has been used for more than twelve years by scientists in a wide variety of domains. It can achieve reliable, scalable workflow execution across a wide variety of computing infrastructures.
Developed by	University of Southern California
Open source or proprietary	Open source
Business models (free, pay-per-use, licence, etc?)	Free
EUBra-BIGSEA technology (plus)	Pegasus has an explicit way to declare workflows, opposite to COMPSs that identifies the tasks of a sequential application and parallelize their execution automatically building a workflow.
EUBra-BIGSEA technology (minus)	Pegasus has a much wider user basis and has also stronger support to data replicas definition, optimization of workflows execution and checkpointing.

Table 2: COMPSs competitor: Pegasus.

Competitor: Apache Spark [3]	
Short description	Data analytics/mining software
Developed by	Apache project – UC Berkeley AMPLab: big data research lab that initially launched Spark
Open source or proprietary	Open Source
Business models (free, pay-per-use, licence, etc?)	Apache v2 license
EUBra-BIGSEA technology (plus)	Ophidia focused more on the eScience side (it provides full support for metadata management). It implements a high-performance data analytics approach, complete OLAP- based functionalities (datacube oriented) and it includes a native workflow support.
EUBra-BIGSEA technology (minus)	Spark provides a wider support for data mining and machine learning functions (MLLib).

Table 3: Ophidia competitor: Apache Spark.

Competitor: Rasdaman [10]	
Short description	Array-based analysis software
Developed by	Jacobs University Bremen.
Open source or proprietary	Open Source
Business models (free, pay-per-use, licence, etc?)	Both open source and commercially supported variants. Apache v2 license
EUBra-BIGSEA technology (plus)	OLAP approach, with a stronger parallel I/O system (more HPC- oriented). Native workflow support.
EUBra-BIGSEA technology (minus)	Rasdaman supports several OGC standard interfaces with a stronger focus on EO data.

Table 4: Ophidia competitor: Rasdaman.

Competitor: Star Cluster [11]	
Short description	Elastic – self managed – computing clusters
Developed by	MIT
Open source or proprietary	Open source
Business models (free, pay-per-use, licence, etc?)	LGPL license
EUBra-BIGSEA technology (plus)	Star Cluster is bound to AWS. EC3 is platform-agnostic, easily extendable, based on de-facto standards (e.g. Ansible), and elasticity is much more transparent.
EUBra-BIGSEA technology (minus)	Optimisation in Amazon AWS.

Table 5: EC3 competitor: Star Cluster.

Competitors: CSP autoscaling mechanisms	
Short description	Some cloud providers (e.g. Amazon or Microsoft Azure) implement autoscaling mechanisms which, however, are proprietary and specific both for the underlying infrastructure and the way autoscaling policies can be specified. Moreover, such solutions cannot provide any performance guarantees for Big Data applications execution. Moreover, the automation is limited to the VM level, and not to the execution framework level.
Developed by	Large cloud vendors
Open source or proprietary	Proprietary/Open source (OpenStack Heat)
Business models (free, pay-per-use, licence, etc?)	Pay-per-use/Free (in case of OpenStack Heat).
EUBra-BIGSEA technology (plus)	Open source, more sophisticated features and potentially technology independent.
EUBra-BIGSEA technology (minus)	Less user friendly and requires some system administration skills.

Table 6: EUBra-BIGSEA performance guarantee for Big Data applications competitors: CSP autoscaling mechanisms.

Competitor: Alces flight [1]	
Short description	Computing clusters for research
Developed by	Alces flight
Open source or proprietary	Partially open source
Business models (free, pay-per-use, licence, etc?)	Two license models per node.
EUBra-BIGSEA technology (plus)	Self-management of elasticity, use of standard software repositories, platform-agnostic.
EUBra-BIGSEA technology (minus)	Wide catalogue of predefined and maintained software packages. In EC3 we need some specialized ICT knowledge to use Ansible Galaxy catalogue.

Table 7: EC3 competitor: Alces flight.

Competitors: Data Quality as a service products (e.g, Data as a Service by Informatica)	
Short description	The application offers a service for data quality assessment and improvement.
Developed by	Different software vendors.
Open source or proprietary	Proprietary.
Business models (free, pay-per-use, licence, etc?)	
EUBra-BIGSEA technology (plus)	Competitors focus on a specific type of data (e.g., customer data) while DQaaS is able to deal with different data types.
EUBra-BIGSEA technology (minus)	Currently, DQaaS cannot compete from a commercial perspective with commercial tools.

Table 8: DQaaS competitors: Data Quality as a service products (e.g, Data as a Service by Informatica)

Competitors: Arena [4] and Java Monitoring Tools [5]	
Short description	These tools are discrete event simulators.
Developed by	Rockwell Automation (Arena) / OpenSource (JMT).
Open source or proprietary	Proprietary (Arena) / Gnu General Public Licence (JMT).
Business models (free, pay-per-use, licence, etc?)	Licence (Arena) / Free (JMT).
EUBra-BIGSEA technology (plus)	dagSim is a highly performant discrete event simulator with respect to other competitors.
EUBra-BIGSEA technology (minus)	Currently, dagSim is at its infancy and it could not compete yet from a commercial perspective with more seasoned commercial tools.

Table 9: dagSIM competitors: Arena and JMT

Competitor: KNIME Analytics Platform [6]	
Short description	KNIME Analytics Platform is the leading open solution for data-driven innovation, helping people discover the potential hidden in data, mine for fresh insights, or predict new futures. It is an enterprise-grade, open source platform, fast to deploy, easy to scale and intuitive to learn.
Developed by	KNIME.COM AG
Open source or proprietary	Open-source version is free, but cannot be used with big data because it has scalability problems (desktop only). Web and scalable version is paid.
Business models (free, pay-per-use, licence, etc?)	Open-source version to learn and build processing workflows using small data and create a community. Offer paid support and scalable environment for those willing to pay.
EUBra-BIGSEA technology (plus)	Scalable version is free. Integrated to other EUBra-BIGSEA technologies, such as COMPSs, Ophidia, authentication/authorization/accounting and infrastructure QoS tools.
EUBra-BIGSEA technology (minus)	KNIME has been developed for more than 10 years and has dozens of operations and features. Large user base and community.

Table 10: Lemonade competitor: KNIME Analytics Platform.

Competitor: Microsoft Azure Machine Learning Studio [7]	
Short description	Microsoft Azure Machine Learning is a suite of offerings designed to enable customers to easily build, deploy and share advanced analytics solutions in the cloud.
Developed by	Microsoft Corporation
Open source or proprietary	Proprietary
Business models (free, pay-per-use, licence, etc?)	Machine Learning Studio is offered in two tiers: Free and Standard. Free is available with limitations (processing capacity, number of users, data volume). Standard tier offers a pay-as-you-go (by usage). Real applications could cost hundreds of dollars.
EUBra-BIGSEA technology (plus)	Scalable version is free. Integrated to other EUBra-BIGSEA technologies, such as COMPSs, Ophidia, authentication/authorization/accounting and infrastructure QoS tools.
EUBra-BIGSEA technology (minus)	Large number of operations, paid support, very good documentation, scalability (paid), many modern implemented machine learning and text processing algorithms.

Table 11: Lemonade competitor: Microsoft Azure Machine Learning Studio.

Competitor: OpenAM [8]	
Short description	OpenAM is an opensource AAA and IAM product, developed by ForgeRock. It supports most authentication and authentication protocols of relevance for AAA and IAM services.
Developed by	ForgeRock.
Open source or proprietary	Open source, licensed under CCDL.
Business models (free, pay-per-use, licence, etc?)	Free (a commercial version also exists, with more extensive support from ForgeRock).
EUBra-BIGSEA technology (plus)	EUBra-BIGSEA AAAaaS is complementing, rather than competing with OpenAM. In fact, AAAaaS will be based on OpenAM components, adjusted for enhanced deployment and managed in cloud environments.
EUBra-BIGSEA technology (minus)	Due to the extensive size of OpenAM features and components, it is possible that AAAaaS will not support all the range of OpenAM functionalities, focusing instead on those relevant for the scope of EUBra-BIGSEA.

Table 12: AAAaaS competitor: OpenAM.

Competitor: Platform-dependent IAM services such as Keystone	
Short description	IAM service for accessing cloud resources such as virtual machines, network and storage resources.
Developed by	Several entities.
Open source or proprietary	Depends on the related cloud framework. Open source in the case of Keystone and CloudStack IAM plugin.
Business models (free, pay-per-use, licence, etc?)	Free in the case of open source frameworks. Integrated in the overall commercial service in the case of commercial services such as Amazon.
EUBra-BIGSEA technology (plus)	Unified interface/access point, making it simpler to use the EUBra-BIGSEA framework.
EUBra-BIGSEA technology (minus)	Provided functionalities still depend on the underlying infrastructure. A unified framework may result in the loss of more specific provided by underlying services.

Table 13: Infrastructure IAAA competitor: Platform-dependent IAM services such as Keystone.

Competitor: Anonymization tool ARX [2]	
Short description	ARX is an open source software for anonymizing sensitive personal data.
Developed by	Technical University of Munich (TUM), Germany.
Open source or proprietary	Depends on the related cloud framework. Open source in the case of Keystone and CloudStack IAM plugin.
Business models (free, pay-per-use, licence, etc?)	Free.
EUBra-BIGSEA technology (plus)	This alternative requires users with advanced privacy knowledge and does not provide ways for the data owners to express their preferences I terms of anonymization. Implementing anonymization policies will allow anonymization be processed in an automatic way, providing a comprehensive tool that will reduce the need for an expert and avoid mistakes that can be made when the process is manual.
EUBra-BIGSEA technology (minus)	

Table 14: PRIVaaS competitor: Anonymization tool ARX.

5 ENGAGEMENT ACTION PLAN

5.1 EUBra-BIGSEA Stakeholders and User Communities

The present section builds on the findings of the EUBra-BIGSEA D2.2: "User communities engagement and dissemination strategy" report delivered in Month 8, August 2016, and on the ongoing user communities engagement work performed under the EUBra-BIGSEA WP2 "Community Engagement, Communication & Impact". The main objective of WP2 is to support the project goals through an effective communication and engagement approach, which ensures the wide promotion and high visibility around the innovation, assets and benefits provided by the EUBra-BIGSEA to its stakeholders and particularly its user communities.

The D2.2 report provided an initial overview of the user communities targeted by EUBra-BIGSEA as relevant stakeholders to facilitate the sustainability of the project on the one hand, and on the other hand to lay the ground for the technology transfer of the EUBra-BIGSEA assets, objectives which are both the focus of the EUBra-BIGSEA WP8 "Long term Co-operation, Sustainability & Commercial Impact."

For the mapping exercise we specifically considered a very broad definition of project stakeholders: any person, organisation or group with an interest in the project and that might, at some point in time, be affected by the progress or outcomes. Users are the most obvious group of stakeholders and are ultimately the people who make use of or are intended to make use of a particular product or service developed by the project. Based on their specific requirements, use, context etc. users are classified in different groups often referred to as user communities.

Based on their characteristics and relationship to the EUBra-BIGSEA project outputs and assets, the following user communities have been identified and defined:

- Architecture & components user communities.
  - Data scientists.
  - Administrators & Infrastructure providers.
  - Open source communities and application developers.
  - Private sector: SMEs, Start-Ups, large enterprises.
- Use case / Application field user communities.
  - Public sector: Public authorities; Urban planners.
  - General public – citizens.

Policy Makers as well as standardisation bodies have also been identified as key stakeholders for the project and they are depicted in Figure 2.

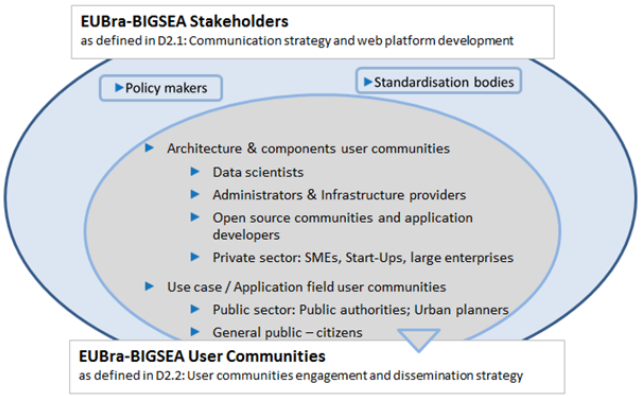

Figure 2: EUBra-BIGSEA Stakeholders and User communities

5.1.1 Stakeholder and asset mapping

Table 15 below provides a mapping for stakeholders, needs and EUBra-BIGSEA assets.

Category	Overview	Needs	Mapped Assets
Data scientists	Data managers and data providers belonging to different scientific domains.	Process large amounts of data or to develop a repeatable process for their data analyses.	COMPSs, Ophidia, EMaaS, DQaaS, Lemonade
Administrators & Infrastructure providers	Providers or administrator of an cloud infrastructure and related services	Addressing challenges related to the composition of Big Data applications and their deployment on the cloud platform with a specific highlight on QoS.	EC3, EUBra-BIGSEA performance guarantee for Big Data applications, DQaaS, dagSim, AAAaaS, Infrastructure AAA, PRIVAaaS
Open source communities and application developers	Communities and individual application developers working on open source software (developed in a collaborative manner; open for change and distribution).	Leverage on the open source software developed in the framework of the project and released under open source software	COMPSs, Ophidia, EMaaS, Lemonade, EC3, EUBra-BIGSEA performance guarantee for Big Data applications, DQaaS, dagSim, AAAaaS, Infrastructure AAA, PRIVAaaS
Private sector: SMEs, Start-Ups, large enterprises	Private organisations (SMEs, large enterprises, etc) belonging to ICT intensive industry sectors.	Knowledge and technology transfer for the development on added-value services.	COMPSs, EMaaS, Lemonade, EC3, EUBra-BIGSEA performance guarantee for Big Data applications, DQaaS, dagSim, AAAaaS, Infrastructure AAA, PRIVAaaS
Public sector: Public authorities; Urban planners	Public authorities and urban planners.	Providing better and smarter services for citizens and improving urban environments and mobility in particular.	The smart city application will provide a descriptive view on the state of the mobility in the city as a whole, identifying its status, trends and the impacts of relevant events.
General public – citizens	General public – citizens	Rute options for any given trip.	Smart city application will provide multiple route options based different criteria as well as travel time, sightseeing etc

Table 15: EUBra-BIGSEA Stakeholder and asset mapping.

5.1.2 Engagement action plan and key messages

To serve the purposes of the project WP2, the engagement plan targeting key stakeholders outlined in D2.2: "User communities engagement and dissemination strategy" offers a mapping based on the category of stakeholder and needs. Building on this assessment, while aiming to maximize the project exploitation potential and commercial impact, the engagement activities under WP8 will be focused on profiling the EUBra-BIGSEA stakeholders and community by clustering them around the main exploitable assets of the project.

Following this approach, WP8 activities will place a high emphasis on strengthening the EUBra-BIGSEA offer. For each customer segment and asset, marketing and information packages will be delivered and effort will be dedicated to their continuous update as new releases will be made available. Promoted across multiple formats (web, social media, promotional materials etc) the marketing and information packages will cover:

- Asset value statement.
- Benefits for the different stakeholders.
- "How to get started instructions" and essential information for potential users.
- Testimonials and adoption stories.

The key messaging tailored to both asset added value and user community will consider the specific information needs as well as the level of knowledge of the target audience.

6 CONCLUSIONS AND PLANS FOR SECOND PERIOD

This deliverable is the first document of WP8 – Long term Co-operation, Sustainability & Commercial Impact, which focuses on analysing the exploitable assets, the market and positioning of the EUBra-BIGSEA assets. It provides an initial analysis of the exploitable assets, and defines a stakeholder engagement plan to promote the assets, tools and services to the identified target groups.

EUBra-BIGSEA has several tangible and intangible assets. The tangible assets are the 10 identified components both for cloud resource management, security, and data analytic services, in addition to the smart city use case. In this document, we highlight the maturity level of each one of the assets. In particular, 7 out of 10 assets are new components developed to address specific gaps in technology. We have described both the main features and benefits of these assets and the instructions on usage for researchers and developers outside the project. The intangible assets refer to the knowledge and expertise about scientific data management stemming directly from the project.

The other topic covered is the market, including an analysis of the 4 main aspects addressed in EUBra-BIGSEA project: cloud computing, big data, security & privacy, and smart cities. In addition to insights from the market analysis to position the project as whole, we have also provided a first market positioning in terms of competitors by highlighting the pluses of the EUBra-BIGSEA technology. It is worth noticing that some of the assets are still not fully developed, as such the analysis will evolve in the second period of the project once the technology will be mature and tested.

Finally, the document concludes with the stakeholder plan, mapping the main target segments and assets relevant for the group.

During the second period of the project, EUBra-BIGSEA will monitor the activities of the EUBrasilCloudForum project and specifically its marketplace under development. EUBra-BIGSEA plans to contribute with the description of the assets, first with the mature ones that have been tested and validated, and then with the ones currently still under development. The marketplace will help disseminating the assets and contribute to the engagement strategy of the project stakeholders.

This deliverable already includes a preliminary analysis of the assets that will be fully deployed and available only in Period 2. The extent of the analysis has been limited to the current status of development and features planned. The final assessment of the EUBra-BIGSEA assets, including the market positioning and competitor analysis will be reported in the D8.3 deliverable planned in December 2017.

REFERENCES

[1] Alces flight. http://alces-flight.com.

[2] Anonymization tool | arx - powerful data anonymization. http://arx.deidentifier.org/anonymization-tool/.

[3] Apache spark: Lightning-fast cluster computing. http://spark.apache.org.

[4] Arena simulation: Discrete event simulation software. https://www.arenasimulation.com.

[5] Java modelling tools. http://jmt.sourceforge.net.

[6] Knime analytics platform. https://www.knime.org/.

[7] Microsoft azure machine learning studio. https://studio.azureml.net/.

[8] Openam - access management & sso. https://forgerock.org/openam/.

[9] Pegasus: Workflow management system. http://pegasus.isi.edu/.

[10] Rasdaman: raster data manager. http://www.rasdaman.com.

[11] Star cluster. http://star.mit.edu/cluster/index.html.

[12] Western european big data technology and services market to grow by 24.6 % cagr by 2018. http://ap-verlag.de/western-european-big-data-technology-and-services-market-to-grow-by-24-6-cagr-by-2018/2683/. Accessed on December 16, 2016.

[13] US International Trade Administration. 2016 top markets report cloud computing: A market assessment tool for u.s. exporters. http://trade.gov/topmarkets/pdf/Cloud_Computing_Top_Markets_Report.pdf, April 2016.

[14] Charbel Aoun. The smart city cornerstone: Urban efficiency. http://www.digital21.gov.hk/sc/relatedDoc/download/2013/079%20SchneiderElectric%20(Annex).pdf. Accessed on December 16, 2016.

[15] Big Data Value Association. European big data value strategic research & innovation agenda. http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf, January 2015.

[16] Capgemini. Business cloud in brazil: Research report 2014. https://www.br.capgemini.com/resource-file-access/resource/pdf/business_cloud_in_brazil_20140827_v15.pdf.

[17] EUBra-BIGSEA Consortium. Eubra-bigsea deliverable 7.2:ges3dataintegration.

[18] Ewa Deelman, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J. Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger. Pegasus, a workflow management system for science automation. Future Gener. Comput. Syst., 46(C):17–35, May 2015.

[19] IBM press. Frost & sullivan leadership award names ibm best-in-class global smart city integrator. https://www-03.ibm.com/press/us/en/pressrelease/43562.wss. Accessed on December 16, 2016.

[20] IBM press. Ibm research launches project "green horizon" to help china deliver on ambitious energy and environmental goals. http://www-03.ibm.com/press/us/en/pressrelease/44202.wss. Accessed on December 16, 2016.

[21] 451 Research. 2016 vormetric data threat report. http://www.ciosummits.com/Vormetric_2016_Data_Threat_Report_Global_WEB.PDF.

[22] Navigant Research. Navigant research leaderboard report: Smart city suppliers. 4q 2014. http://az370354.vo.msecnd.net/whitepapers/Navigant%20Research%20Leaderboard%20Report-Smart%20City%20Suppliers-%202014%20Executive%20Summary.pdf.

[23] SmartCitiesCouncil. Predicting traffic snarls an hour in advance? microsoft is on it. http://smartcitiescouncil.com/article/predicting-traffic-snarls-hour-advance-microsoft-it. Accessed on December 16, 2016.

[24] Frost & Sullivan. More than one third of brazilian companies will have invested in big data analytics by the end of 2015, finds frost & sullivan. http://ww2.frost.com/news/press-releases/more-one-third-brazilian-companies-will-have-invested-big-data-analytics-end-2015-finds-frost-sullivan/. Accessed on December 16, 2016.

[25] Frost & Sullivan. Smart city as a service: Using analytics to equip communities for data-driven decisions. http://corpcom.frost.com/forms/IBMSmartCityWhitepaperDownloadLeadgen. Accessed on December 16, 2016.

[26] Ernst & Young. Big data: changing the way businesses compete and operate. http://www.ey.com/gl/en/services/advisory/ey-big-data-seven-big-data-drivers. Accessed on December 16, 2016.

Glossary

Term	Explanation
AAA	Authentication, Authorization and Accountability
AAAaaS	Authentication, Authorization and Accountability as a Service
API	Application Programming Interface
COMPSs	COMP Superscalar (COMPSs) - programming model which aims to ease the development of applications for distributed infrastructures, such as Clusters, Grids and Clouds
CLUES	CLUster Energy SavingS
DAG	Directed Acyclic Graph
dagSIM	Event simulator working on a DAG able to estimate Big Data applications performance
DQaaS	Data quality as a service
EC3	Elastic Compute Clusters in the Cloud
EM	Entity matching
EMaaS	Entity Matching-as-a-Service
I/O	Input/Output
IAM	Identity and Access Management
IM	Infrastructure Manager
IDE	Integrated Developement Environment
MapReduce	Programming model for processing large data sets
MESOS	A Resource Management platform that abstracts CPU, memory, storage, and other compute resources away from machines
OLAP	Online Analytical Processing
Ophidia	A CMCC Foundation research project addressing Big Data challenges for eScience
OpenStack	OpenStack cloud management platform
PRIVAaaS	Software toolkit providing a set of libraries and tools that allow to control and reduce the data leakage in the context of Big Data processing
PMES	Programming Model Enactment Service
SLA	Service Level Agreements
Spark	Apache Spark - engine for large-scale data processing
TOSCA	Topology and Orchestration Specification for Cloud Applications
QoS	Quality of Service
VM	Virtual machine