

D2.3: Preliminary Action Plan Report

Author(s)	Lorenza Panelli (Trust-IT)
Status	Review
Version	V1.0
Date	25/01/2017

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)

Abstract:

This preliminary report will highlight the initiatives across Europe and Brazil among five reference domains in which EUBra-BIGSEA operates: (i) Quality of Service for Cloud Computing Infrastructure, (ii) Big Data, (iii) Security & Privacy (with a focus on free flow of data), (iv) Smart Cities and Urban Mobility planning, (v) Standardisation, Interoperability & Portability. This preliminary report discusses the main challenges in these domains and common to Europe and Brazil, covering the current gaps in innovation and research that need to be addressed in future research activities.

Document identifier: EUBRA BIGSEA -WP0-D0.0	
Deliverable lead	TRUST-IT
Related work package	WP2
Author(s)	Lorenza Panelli (Trust-IT)
Contributor(s)	Sara Pittonet (Trust-IT), Silvana Muscella (Trust-IT), Timea Biro (Trust-IT), Danilo Ardagna (Polimi), Daniele Lezzi (BSC), Sandro Fiore (CMCC), Ignacio Blanquer (UPV), Marco Vieira (UC), Wagner Meira Jr (UFMG), Nadia P. Kozievitch (UTFPR)
Due date	31/12/2016
Actual submission date	26/01/2017
Reviewed by	UFMG, UPV
Approved by	PMB
Start date of Project	01/01/2016
Duration	24 months
Keywords	Quality of Service, Big Data, IoT, IaaS, SaaS, Cloud Computing, Interoperability, Standards, Security, Privacy, Authentication Authorization Accounting

Versioning and contribution history

Version	Date	Authors	Notes
0.1	18/11/2016	Trust-It	ToC; Initial version
0.2	16/12/2016	Trust-It, POLIMI, BSC, UPV, UFMG, CMCC, UC, UTFPR	First draft of contributions
0.3	06/01/2017	Trust-It, POLIMI, UPV, UFMG, CMCC, UC	Second round of contributions
0.4	13/01/2017	Trust-It	Updates and minor additions
0.5	16/01/2017	Trust-It	Glossary; Minor revisions
0.6	20/01/2017	UPV, UFMG, Trust-It	Review, Updates
0.7	23/01/2017	Trust-It	Addressed review comments and suggestions
0.1	25/01/2017	All	Final internal review

Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 license. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0>.

Disclaimer: The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EUBra-BIGSEA Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EUBra-BIGSEA Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EUBra-BIGSEA Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENT

EXECUTIVE SUMMARY	6
1. INTRODUCTION	7
1.1. EU-Brazil co-operation: an overview	7
1.2. Technology & Innovation relevant to EUBra-BIGSEA	9
1.3. Specific challenges of EUBra-BIGSEA.....	10
2. QUALITY OF SERVICE FOR CLOUD COMPUTING INFRASTRUCTURES	11
2.1. Introduction.....	11
2.2. Challenges.....	11
2.3. Analysis of Priorities	13
3. BIG DATA	14
3.1. Introduction.....	14
3.2. Challenges.....	14
3.3. Relevant Initiatives	15
3.3.1 Research Data Alliance	15
3.3.2 NIST Big Data Public Working Group	15
3.3.3 OGC Big Data Domain Working Group	16
3.3.4 Big Data Value Association	16
3.3.5 US Big Data Research and Development Initiative.....	16
3.3.6 European Open Science Cloud & European Data Infrastructure	16
3.3.7 EUDAT Collaborative Data Infrastructure (CDI).....	16
3.4. Analysis of Priorities	17
4. SECURITY & PRIVACY ENSURING FREE FLOW OF DATA	17
4.1. Introduction.....	17
4.2. Challenges.....	18
4.2.1 Legal challenges – common legal framework for managing security and privacy aspects.....	18
4.2.2 Public awareness challenges – awareness about the privacy of the data, and which data can or should used and for which purposes.....	19
4.2.3 Standardisation challenges – standards to allow data owners to express privacy preferences....	19
4.2.4 Technological challenges – technological solutions to deal with the massive sizes and heterogeneity of the big data datasets	19
4.2.5 Analytics challenges – solutions to deal with the analytics algorithms that extract privacy sensitive information from anonymised data	20
4.3. Relevant Initiatives	20
4.4. Analysis of Priorities	22
5. SMART CITIES AND URBAN MOBILITY PLANNING	23

5.1. Introduction.....	23
5.2. Challenges.....	24
5.3. Relevant Initiatives	24
5.4. Analysis of Priorities	25
6. STANDARDS, INTEROPERABILITY & PORTABILITY	26
6.1. Introduction.....	26
6.2. Challenges & Priorities.....	26
6.3. Related initiatives	27
7. PRELIMINARY ACTION PLAN.....	27

LIST OF TABLES

Table 1: EUBra-BIGSEA Roadmap for Preliminary Action Plan.....	29
--	----

EXECUTIVE SUMMARY

EUBra-BIGSEA is committed to generating significant impact on the cooperation between Europe and Brazil in the area of advanced cloud services for Big Data applications. EUBra-BIGSEA facilitates the integration of European and Brazilian technologies and experiences to bring forward scientific innovation through a use case scenario approach that is important for both Europe and Brazil.

The Preliminary Action Plan report (D2.3) provides an initial overview of current challenges, research and innovation opportunities, including an analysis of the priorities identified, highlighting as well the relevant EU-Brazil joint effort initiatives in the areas addressed.

The Plan is the first of 2 reports produced under EUBra-BIGSEA Task 2.3 “Joint EU Brazil Cloud Computing Research & Innovation Action Plan” focusing on the joint EU Brazil cloud computing research and innovation action plans in the area of Quality of Service for cloud computing infrastructures; Big Data analytics; security and privacy; standards, portability and interoperability, smart cities and urban mobility planning. As a Preliminary Joint Action Plan, this document covers thus the main challenges that EUBra-BIGSEA is addressing, providing an overview of recent developments and opportunities, while offering an initial plan aimed at boosting international co-operation and sharing of new expertise with other relevant initiatives in Europe and Brazil.

The Final Research & Innovation Action Plan (D2.5), due December 2017, will further develop this analysis and discuss the potential socio-economic impact of the research roadmap proposed in EUBra-BIGSEA.

1. INTRODUCTION

1.1. EU-Brazil co-operation: an overview

Cloud computing, Big Data technologies, the Internet of Things (IoT), 5G communications and cyber security are the building blocks of the digital economy. The uptake of cloud computing and virtualised infrastructures plays an essential role in enabling the transition towards a distributed global community, enhancing collaborative work and tackling the challenges of big data.

The cooperation between Europe and Brazil seeks to sustain and enhance the social and economic conditions, increase competitiveness, creating jobs, and addressing common global challenges in areas like energy, international cyber policy, sustainable development, climate change, and the environment.

Policy collaboration to date has included work on identifying barriers that may preclude the adoption of cloud-based services in Europe and in Brazil and on identifying concrete joint initiatives to minimise such barriers.

These barriers include:

- Interoperability - in reference to the different architecture layers, big data services or metadata standards.
- Complexity of the legal framework (users' rights, data location, data protection and privacy, including global aspects of these issues).
- Security - lack of a common, standardised approach to security levels.
- Standards - lack of common standards for (interoperable) cloud computing services.

EUBRA-BIGSEA¹ (2016-2018) and SecureCloud² (2016-2018) are the two initiatives on cloud computing, including security aspects under the first Horizon 2020 Work Programme 2014/2015 as part of the third Coordinated Call jointly funded by the EU and Brazil: Advanced Cyber Infrastructures. Another initiative, EUBraSiCloudFORUM³ (2016-2018) is a Coordination and Support Action that will define the research priorities for future collaboration under work programme 2018-2020. The aim is to ensure that projects will develop innovative technologies for cloud-based service provision, Big Data analytics while also taking into account security concerns.

Current EU-Brazil collaboration is expected to advance cloud-centric applications for big data, and move forward towards facilitating policy coordination between the EU and Brazil.

With regard to international cooperation on research and innovation, the European Commission published its second progress report in October 2016, highlighting that:

"No single country or region can face global challenges alone. That's why our research and innovation needs to be Open to the World. This report clearly shows that we have come a long way in engaging with our global partners, which enables us to maintain our excellence in science and technology, create new business opportunities and have a leading role in global developments", Carlos Moedas, European Commissioner for Research, Science and Innovation⁴.

[1]

¹ <http://www.eubra-bigsea.eu/>.

² <https://www.securecloudproject.eu/>.

³ <https://eubrasilcloudforum.eu/>.

⁴ EU international research cooperation helps reach solutions to global challenges; October 2016;

<http://ec.europa.eu/research/index.cfm?pg=newsalert&year=2016&na=na-131016>

www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

The report reaffirms the importance of EU-Brazil cooperation on research and innovation that addresses shared economic, environmental and social challenges. Accompanying the progress report, the EC has also published a roadmap⁵ identifying 10 priority areas for joint coordinated co-operation strategies, highlighting ICT as the prominent area for collaborative work with Brazil.

Future challenges in cloud computing and Big Data joint research activities between Brazil and Europe include:

Data integration and **data harmonisation** with **data analytics** playing a key role in the development of future technologies across a variety of domains. With regard to finance and insurance services, Brazil counts some of the most advanced technologies, some of which are evolving in the cloud, while Europe boasts a thriving fintech industry.⁶ Cooperation could therefore help accelerate research and innovation with a particular focus on security, privacy, interoperability and Quality of Service (QoS).

The management and analysis of large amounts of data are highly connected with quality of data and the “trust factor”: where secure-by-design approaches can help create **trustworthiness** and **dependability**, where research and innovation needs to address challenges related to the protection of data, **transatlantic flows of data**, **portability** and **interoperability** of data, also in compliance with new EU regulations, most notably the General Data Protection Regulation⁷ (GDPR), but also the Directive on Network and Information Systems⁸ (NISD). While Europe is playing a leading role in data protection and IT security, Brazil is currently defining its regulatory framework. EU-Brazil co-operation can facilitate understanding of the EU legal framework and how a similar approach can benefit Brazil. The EU regulation and directive can both provide relevant inputs for the Brazilian framework on data protection and cyber security by requiring organisations to raise the bar on addressing vulnerabilities. Such a collaborative approach was highlighted during the last EU-Brazil Policy Dialogue (Brussels, November 2016), to which EUBra-BIGSEA contributed.

The new EU regulations are expected to have a significant impact on the provision of public cloud computing services. One of the major barriers concerning the wide adoption of cloud services is indeed represented by the lack of trust in cloud services, more in terms of data privacy in addition to reliability of the cloud infrastructure in coping with the application requirements within a given budget. This has pushed towards hybrid cloud solutions where sensitive data are restricted within the boundaries of the organisation and kept in the trusted domain: **trustworthiness** of services and solutions provided, where trust is also mainly connected with public versus private cloud⁹ adopted by different sectors.

As a result, **data protection, privacy and security** are among the crucial aspects for the adoption of cloud computing in different sectors, across Europe and globally. Considerable investments have been made in Europe on cloud computing initiatives funded through DG CONNECT (Software, Services, Cloud Computing), and with cross-project work taking place in the EC Cluster on Data Protection, Security and Privacy¹⁰ (DPSP Cluster).

Leveraging this work, a key challenge for both Europe and Brazil is to create a solid framework where users and organisations can use applications that are secure and resilient to data breaches. This also means covering data traceability, amplified further by the use of multi-cloud infrastructures, at the same time

⁵ Priorities for international cooperation in research and innovation; October 2016; http://ec.europa.eu/research/iscp/pdf/policy/annex_roadmaps_oct-2016.pdf#view=fit&pagemode=none

⁶ In Q1 2014 financial technology companies raised a total of €166 million (+201% compared to the previous quarter), a peak not seen in the industry since 2000 when €263 million flowed into these type of companies, <https://startupxplore.com/en/blog/fintech-startups-europe/>.

⁷ http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

⁸ <https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive>.

⁹ G. M. Gonzalez, D. M., L. Ferraz, O. Duarte. Sistema Automatizado de Gerencia de Recursos para Ambientes Virtualizados. In Proceedings of the XXXII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SRBC'14).

¹⁰ <https://eucloudclusters.wordpress.com/data-protection-security-and-privacy-in-the-cloud/>.

providing, from the service point of view, transparent and secure systems - especially if regarding public services.

Other critical aspects where Europe and Brazil see the need for widely accepted solutions are the **standardisation of contracts**, the need for harmonised legal and technical guarantees, including security and data protection levels delivered in the cloud service, adequately mapped and described in Service Level Agreements. These are crucial factors for boosting the economies of both Brazil and Europe.

In this perspective, the EUBra-BIGSEA project demonstrates the value of cloud services and big data services for applications with a high social and economic impact for both EU and Brazil such as the processing of massive data coming from highly connected societies.

As a Preliminary Joint Action Plan, this document covers the main challenges that EUBra-BIGSEA is addressing, providing an overview of recent developments and opportunities, while offering an initial plan aimed at boosting international co-operation and sharing of new expertise with other relevant initiatives in Europe and Brazil.

1.2. Technology & Innovation relevant to EUBra-BIGSEA

EUBra-BIGSEA aims at providing a cloud service platform that could provide QoS guarantees to Big Data Analytics applications, covering both performance and security. EUBra-BIGSEA is focused on integrating state-of-the-art framework and tools with innovative components that could extend their functionality.

EUBra-BIGSEA bases its developments on the following:

1. An advanced resource management framework, which provides the necessary resources for the execution of the applications. The resource management includes both the physical and the virtual resources that are provisioned for the services, and based on Apache Mesos, Marathon, Chronos, OpenNebula, OpenStack, and KVM hypervisor.
2. An efficient monitoring system for providing the metrics in relation to the applications and resources, enabling decision-making about elasticity, by switching on/off nodes to the resource management framework and allocating more or fewer resources to the applications, based on OpenStack Monasca.
3. An effective environment to code data analytic applications on top of the cloud services, aiming both at general-purpose languages and data analytics specific frameworks. Based on COMPSs, Ophidia and Apache Spark.
4. Security and privacy constraints and assessment on the applications that consume the data, focusing on the challenges of Big Data storage.
5. An ecosystem of components, software libraries and programmes that address data entity matching, data clustering, data quality analysis and other services needed to build data analytic applications ready to be used in the system.
6. A family of applications for traffic recommendation consuming the previous services for citizens and municipalities, focusing on mobile and web-based applications.
7. QoS proactive policies and optimisation-based policies which provide a priori QoS guarantees to application execution.

EUBra-BIGSEA develops the following innovations to each of the above:

1. A two-layered resource management framework comprising a Mesos Framework on top of an IaaS that provides horizontal and vertical self-elasticity. The innovative aspects here are: automatic configuration based on standard templates, platform agnosticism, consistent, self-elasticity for memory (Vertically) and CPU (horizontally).
2. Extension of probes and agents for OpenStack Monasca focusing not only on IaaS specific metrics (relevant for the system administrators) but also at the application level (relevant for QoS).
3. A from-scratch development of a Data Analytics IDE called LEMONADE that enables registering and integrating data analytics components in a graphical format, producing Spark, Ophidia, and COMPSs code. EUBra-BIGSEA also extends COMPSs with the capability of executing containerised

jobs on a Mesos Framework. COMPSs extracts parallelism from the dependencies and executes on a distributed platform without explicit parallelisation of the code. Additionally, EUBra-BIGSEA extends Ophidia with QoS-based extensions to address dynamic and elastic analytics scenarios in the cloud.

4. Privacy annotation for the specificities of Big Data applications with the management of Data privacy constraints for datasets and applications, in a way that privacy warnings can arise when combining non-privacy restricted datasets.
5. A catalogue of components that are bound to specific unsolved problems such as the geographic layout discrepancies among providers, the multi-level route selection, etc.
6. HTML5 and mobile-based applications tailored to the platform addressing functionalities such as the most pleasant, secure, convenient or touristic route.
7. Innovation combining predictive techniques and analytical fast performance analysis tools for large scale applications and clusters.

1.3. Specific challenges of EUBra-BIGSEA

This first Preliminary Joint Action Plan aims to demonstrate the commitment of EUBra-BIGSEA in driving forward innovations in cloud computing and big data through collaborative research. It also aims to show the project's commitment to making a contribution in identifying key challenges, gaps, and opportunities that need to be addressed to foster the acceleration of the cooperation.

The Plan describes three major challenges addressed by the project and identifies areas that can boost international co-operation.

The first challenge addressed in this document regards the **Quality of Service of Cloud Computing infrastructures** (Section 2). As big data systems become critical in businesses and society, quality of service, reliability and performance guarantees will play a key role, due to the increasing complexity, quality and quantity of data that needs to be handled and the heterogeneity of cloud resources.

Another relevant challenge affecting several industrial sectors is **Big Data Analytics** (addressed in Section 3) due to the growing demand for the design and development of big data processing applications in order to provide solid and efficient tools and services.

Challenges and priorities related to **Security and Privacy** in both Europe and Brazil that research and innovation activities need to cover are covered (Section 4), looking at key related aspects such as the free flow of data.

Thereafter we describe the framework in which EUBra-BIGSEA operates in relation to cloud computing and big data analytics challenges and to **smart cities and mobility**, from the public transportation point of view (Section 5).

Standards, interoperability and data portability aspects are then addressed (Section 6), taking into account the impelling necessity for standardisation of contracts among European States and between Europe and Brazil.

Preliminary Joint Action Plan Roadmap

The Roadmap sets out concrete actions that EUBra-BIGSEA will undertake with the most relevant initiatives identified during the second year of the project based on synergies established.

2. QUALITY OF SERVICE FOR CLOUD COMPUTING INFRASTRUCTURES

2.1. Introduction

Cloud computing infrastructures have added tremendous flexibility to application developers, as they can customise virtual infrastructures to their application specific requirements and match allocated resources to a specific workload. The use of cloud computing infrastructures though has also brought new challenges:

- Increased DevOps at user or application-developer side. Configuration and maintenance of servers decreases (or changes) at the infrastructure-side, as the software to be installed is concentrated on the resource management. However, at the user level there is a need for higher configuration skills to operate an IaaS. There is also a need to automate and facilitate the deployment of data analytic infrastructures. Providers start to distribute Virtual Machine Sandboxes, Docker composes the specification to facilitate the deployment of reduced and static infrastructures. Public cloud providers offer Data analytics as a Service (AWS EMR, Azure Insight, etc.). Recent approaches move towards “Zero-DevOps” and serverless orientations (e.g. Lambda, google containers) to totally remove IaaS deployment at the user level. By facilitating infrastructure deployment on a single CLI, we aim to address such an issue and provide a “ready-to-use” complete infrastructure without hacking any configuration line.
- Resource Elasticity. Existing metrics make it easy for users to define in terms of memory starvation, CPU or network traffic overload. There are simpler measures to be applied at infrastructure level that do not require user definition. This way, the data analytics infrastructure works “as a Service” hiding some particularities from the user. EUBra-BIGSEA does this at the level of the resources. When a framework request arrives, the platform checks for available resources and spawns new resources if needed. The allocation of physical memory is transparently provisioned to the VM, as we support memory ballooning (only if deployed in a specifically modified KVM hypervisor).
- Quality of Service. Elasticity is good, but choosing a priori the right configuration is even better. The QoS system predicts the expected workload based on the experience and defines the proper allocation of resources. If the proactive allocation fails, the elasticity will reallocate resources to meet the QoS. This way we can reduce the reallocation overhead.

2.2. Challenges

Big data frameworks are evolving very fast. They include very complex software systems involving many stacks where application execution is spread across a very large number of physical nodes. Clouds are cost-effective platforms to support such systems as resources (e.g., nodes) can be allocated and deallocated, on demand, by cloud providers, in response to the applications requirements and QoS needs (elasticity). As a matter of fact, **IDC predicts that by 2020 nearly 40% of Big Data analytics will be supported by public clouds¹¹.**

In such a context, defining mechanisms to reduce costs for application execution while supporting the “on demand” approach implies the need to run highly intensive analytics tasks for business critical applications. As big data systems become a central force in society, the field should shift from simply building systems to developing intelligent systems that provide quality of service.

Yet, predicting the performance of big data applications in scenarios of technical interest, notably a mix of applications running concurrently in a cloud system, which is very challenging. Big data applications are characterised by changing behaviour during execution: e.g. they require initially a lot of CPUs, then a lot of network capacity to later switch between the two with sometimes complex patterns. Moreover, to cope with the large amount of data, such applications often run in parallel stages. The performance (and thus QoS) of parallel applications is often harder to predict due to synchronisation overheads.

In the remainder of the section, we detail the challenges associated with providing QoS for cloud infrastructures, contention and variable communication costs.

¹¹ The digital universe in 2020. - <http://idcdocserv.com/1414>

Execution of QoS-aware applications within a budget. Big data applications are supported by cloud infrastructures which, for resource contention, can be affected by performance decline. For this reason, one of the major challenges for big data applications is to define mechanisms and policies that implement resource partitioning and resource management in a way that cloud data centers' resources are used efficiently, providing differentiated service levels to customers according to the price of the resources.

Cost predictability of big data applications in the cloud. Designing new models to estimate the costs in terms of cloud resources to run big data applications is key. An accurate estimate will enable more efficient scheduling of resources including cost-effective utilisation of data centers.

Models to predict big data application performance. Similarly, designing new models to predict the performance of applications, in terms of execution time, given certain resources, is also key to providing QoS to application customers. Such models should be accurate and efficient (i.e., provide an estimate quickly). The use of accurate models is beneficial for both cloud providers and end users: for cloud providers models can trigger runtime adaptations to provide QoS guarantees, for end-users they can support what-if analysis and take more informed decisions on the resources to be used. Similarly, it is important that the models run reasonably fast (e.g., provide response quick enough to drive runtime adaptations). However, model accuracy and efficiency are two often conflicting objectives. More sophisticated models, capturing in more details different aspects of the application execution, are often more accurate, but also very costly to run. Thus, finding the best tradeoff between these two goals is a major challenge.

Cross-layer orchestration of big data applications. Another key challenge relates to the definition of new mechanisms to manage efficiently cloud resources and leverage cloud monitoring systems to adaptively orchestrate cloud services and applications so as to cope well with the lack of support for elasticity mechanisms for big data processing applications.

Relevant Initiatives

Current solutions are typically reactive. Horizontal elasticity is supported in public and on-premise clouds (e.g. AWS, VMWARE VDC, Heat autoscaling groups, etc.), by defining the metrics and the triggers that fire them up. Vertical elasticity is also provided through the restart of containers and VMs. EUBra-BIGSEA addresses both proactive and reactive scalability and horizontal and vertical elasticity in the same solution.

With respect to the configuration and deployment of TOSCA-based application descriptions, there are other solutions such as Cloudify and OpenTOSCA. EUBra-BIGSEA solution provides a wider support of IaaS, and the capability of dealing with both VMIs and Container images.

For what concerns big data applications performance evaluation, there is a large market of discrete event simulation solutions, e.g., Arena (Rockwell Automation), OMNeT++ (licence required for commercial use) and JMT (openSource).

The research carried out within EUBra-BIGSEA consists of an open-source, high performance, light-weight, ad-hoc discrete event simulator for DAG models corresponding to Map/Reduce, Tez, and Spark jobs.

Being still in its infancy, the solution developed has not reached yet its full capacities that would go beyond other simulation tools. That said, its added value lies in i) ease to use, ii) no training required, iii) fast simulation iv), and may not have any direct competitors (the simulator has been designed and optimised for a specific kind of model), and v) accuracy.

With regard to competitors, the simulator has proven to be sensibly faster than JMT against equivalent models.

New software releases implement different event allocation policies. The technical Roadmap will include the implementation of fluid models to achieve high accuracy and efficiency to fulfil runtime control requirements.

Monitoring

One especially important aspect for implementing quality of service mechanisms is the quality of monitoring. There is a large number of systems that collect and aggregate data for display or triggering of alarms. For example, Nagios (<https://www.nagios.org/>) and Zabbix (<http://www.zabbix.com>) are very common in data-centers in general, while Ceilometer (<https://wiki.openstack.org/wiki/Telemetry>) and Monasca (<https://wiki.openstack.org/wiki/Monasca>) are typically used to monitor a cloud infrastructure managed by OpenStack.

Nevertheless, in EUBra-BIGSEA we target the provisioning of applications with non-trivial quality of service requirements. These requirements impose needs such as modelling the performance behavior of some applications so that when a (periodic) instance of such application is initiated, a cluster with adequate size can be provisioned. EUBra-BIGSEA applications also require scalable databases for storing long-term metrics and mechanisms that can dynamically adjust cluster performance to ensure the target QoS.

Therefore, the monitoring system should be flexible and well-integrated with the cloud infrastructure. Flexibility helps to define metrics, monitoring frequencies that are suitable both for the optimisation and for the dynamic actuation to compensate for deviations on expected performance or input data.

QoS

In the international scenario, the most recent work closely related to ours are the ones from Khazaei et al.¹² and Nanda et al.¹³ The work from Khazaei et al. proposes a performance analytical model supported by experiments to study the provisioning performance of microservice platforms. The microservices are deployed using containers technologies inside virtual machines. The performance analysis consists in understanding the influence of different aspects (for example, the number of containers per VM, the number of VMs allocated, the rate of user requests, etc.) in the performance of the executed application. The main contribution of the work consists in a tractable analytical performance model that showed a high fidelity to experiments and enables the study the provisioning performance of microservice platforms at large scale.

Then, the work from Nanda et al. proposes a predictive model for dynamic and vertical scaling of multi-tier web applications on cloud environments. The solution seeks to both minimize the application's resource allocation and its related costs, and maximise the resource utilisation of the cloud's infrastructure. Thus, the QoS of the application services is ensured by a mapping between the resource utilisation levels from the virtual infrastructure and the end-user performance of the application running with different layers. The main contribution is the use of an ARIMA prediction model combined with a bandpass filter to estimate accurately short-term future application demands and perform a proactive vertical auto-scaling. The approach has an average response time well within the considered SLA range, ranging from 14ms to 17ms for a 95% confidence interval, but has a significant SLA violation rate, around 7.9%, for some scenarios.

2.3. Analysis of Priorities

In relation to monitoring, there is a need to provide precise metrics for containerised applications, which are normally mixed with the machine-level metrics. This will provide more realistic information about the exact state and health of an application during its execution.

¹² H. Khazaei, C. Barna, N. Beigi-Mohammadi, M. Litoiu. Efficiency Analysis of Provisioning Microservices. In Proceedings of the 8th IEEE International Conference on Cloud Computing Technology and Science (CloudCom'16)

¹³ S. Nanda, T. J Hacker, Y.-H. Lu. Predictive Model for Dynamically Provisioning Resources in Multi-Tier Web Applications. In Proceedings of the 8th IEEE International Conference on Cloud Computing Technology and Science (CloudCom'16)

With respect to elasticity, the challenge is the compromise between flexibility and performance. Application topology descriptions enable quick migration of applications but impose configuration overhead. Elasticity on such applications imply configuring and reconfiguring VMs, which impacts on performance and sometimes implies VMs reboot (which might introduce application disruption, performance degradation and does not fit with run-time management requirements). The combination of preexisting VMs, on-the fly VMs, adaptive recipes and the embedding of dependencies in containers can reach the perfect equilibrium.

With respect to QoS prediction, a key challenge is to handle the tradeoff between model accuracy and model efficiency. It is important to properly assess to what extent more sophisticated models that capture, for example, details of the parallel execution are beneficial, particularly when coupled with the optimisation models. A promising direction, investigated during the project, is to mix traditional white box analytical models (like queuing networks and stochastic Petri Nets), with black box models based on machine learning, trying to obtain the advantages of the two worlds. Experiments with different scenarios - different applications and setups -- must be run to support such analysis.

A key point with respect to adaptation activities to support QoS elastic and dynamic scenarios in the cloud is to extend the data analytics applications in a way they are able to provide metrics reflecting the current status and health of the runtime system about the target applications. Based on that we can build reactive/proactive scenarios where the analytics applications/frameworks can make use of additional resources from the infrastructure. Worth mentioning are also changes in the applications themselves that can enable elastic and dynamic scenarios in the cloud. Such changes are often mandatory to move from static toward more challenging dynamic scenarios.

3. BIG DATA

3.1. Introduction

Big data is often defined as a collection of data that due to its size or complexity cannot be adequately processed with traditional data processing applications. Such data often result from the *datafication* phenomenon that refers to the ability to capture and turn different aspects of daily life into digital data and to exploit it to make decisions and thus to gain maximum benefit.

The EUBra-BIGSEA project aims to ease the development of big data processing applications by providing a set of services designed for supporting Big Data Analytics. Specifically, such services provide functionalities for managing big data from data access to advanced data analysis using data analytics and mining tools and by also guaranteeing the satisfaction of non-functional requirements (i.e., quality). The service has to be clearly designed in order to address the main big data challenges due to the volume of data, the variety of data sources, the velocity requirements and the data uncertainty.

3.2. Challenges

There are several challenges that need to be faced in the big data landscape. In this section the most prominent ones are presented.

First of all, the need for a **fast environment** where results can be delivered in real-time despite the large amount of data that needs to be processed. In this regard, several issues need to be solved both at the hardware and software side. From a software point of view, in-memory analytics jointly with high-performance approaches leveraging parallel paradigms can help a lot. From a hardware standpoint, infrastructures leveraging new generation of memories and SSD (solid-state drive) devices can better help. EUBra-BIGSEA is trying to address from a software point of view such need by integrating WP4 technologies that incorporate such paradigms (e.g. Spark, Ophidia).

Data quality is really key to delivering high-quality answers to real problems. Big data analytics and mining services should consider the relevant data to provide the output for good and valuable decisions. The

www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

problem is that not all the data are relevant, in fact gathered data can be unreliable since incorrect, outdated, or incomplete. Data quality services are strongly needed. Data quality techniques are consolidated only for structured data. New methods are needed, but velocity requirements impose the use of approximate approaches for the data quality assessment.

Data pre-processing is a challenge itself as before starting any kind of data analysis the data needs to be gathered, cleaned, and integrated. Only in a few cases data can be directly ingested/analysed by data analytics tools. So, pre-processing is somehow at the border, but needs to be faced as well as it can require a lot of work, resources and time to be addressed. The categorisation of the data sources can be an important starting point to find solutions for classes of data.

QoS is a key point as mentioned in the previous section. Making a data analytics application QoS-enabled or QoS-aware can imply some reworking, redesign, adaptation activities. Additionally, QoS-aware cloud environments are today almost still under development (e.g. cloud monitoring and cloud orchestrators components). QoS is strongly related to cost predictability of big data applications in the cloud, another critical point representing a priority for the present and near future. Predicting the performance of a mix of applications (running concurrently in a cloud system) is a very complex task due to the data-intensive nature of big data applications with regard to more classical cpu-intensive scenarios.

One size fits all solutions vs multiservice analytics platforms is another question. While single solutions are very effective to face specific questions, data analytics platforms can better deal with real-world questions due to the complexity of big data challenges, the increase of expectations from the end-users standpoint, the increase of complexity in terms of number of data sources (and their heterogeneity) to be managed for a specific use case/objective.

3.3. Relevant Initiatives

Key initiatives related to the big data landscape are the Research Data Alliance, the NIST Big Data public working group, the OGC Big Data domain working group, the Big Data Value Association and the US Big Data Research and Development Initiative. In the following sub-sections they are presented in more detail.

3.3.1 Research Data Alliance

The Research Data Alliance (RDA) was launched as a community-driven organisation in 2013 by the European Commission, the United States National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation with the goal of building the social and technical infrastructure to enable open sharing of data. The Research Data Alliance focuses on enabling data sharing across barriers through focused Working Groups and Interest Groups, formed of experts from around the world – from academia, industry and government¹⁴. RDA enables data to be shared across barriers through focused Working Groups and Interest Groups.

EUBra-BIGSEA partners are involved into RDA Working Groups and Interest Groups. Two key examples are the Big Data IG and the Array Database Assessment WG. In particular, the Ophidia analytics framework is one of the tools being evaluated in the Array Database Assessment WG. The main outcomes of the WG are expected in 2017.

3.3.2 NIST Big Data Public Working Group

The National Institute of Standards and Technology (NIST) has released the NIST Big Data interoperability framework, a huge set of documents aimed at creating standards around everything in big data from definitions to architectures¹⁵.

¹⁴ Research Data Alliance website - <https://www.rd-alliance.org/>.

¹⁵ NIST Big Data Public Working Group - <https://www.nist.gov/el/cyber-physical-systems/big-data-pwg>.
www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

Such documents represent a strong reference for scientists in general and in the EUBra-BIGSEA context as well.

3.3.3 OGC Big Data Domain Working Group

The purpose of the OGC Big Data Domain Working Group is to provide an open forum for work on Big Data interoperability, access, and analytics. To this end, the open forum pursues collaborative information collection and liaisons with other Big Data working groups. Such an initiative is mainly related to the geospatial context and to spatial data infrastructures.

EUBra-BIGSEA partners are following the activity in the WG and in particular the adoption of OGC standards like OGC-Web Processing Service for geospatial platforms.

3.3.4 Big Data Value Association

The Big Data Value Association (BDVA) is the private counterpart to the EU Commission to implement the BDV PPP programme (Big Data Value PPP). The objectives of the Big Data Value Association are to boost European Big Data Value research, development and innovation and to foster a positive perception of Big Data Value.

The interest from EUBra-BIGSEA partners on BDVA is mainly related to the link with the European Technology Platform for High-Performance Computing, with specific regard to *high performance data analytics* aspects.

3.3.5 US Big Data Research and Development Initiative

In 2012 the Obama administration unveiled the “BIG DATA” initiative, a plan of investment of about \$200 million in R&D for big data, with the aim of improving the ability to extract knowledge and insights from large and complex collections of digital data, to help solve some the Nation’s most pressing challenges¹⁶.

3.3.6 European Open Science Cloud & European Data Infrastructure

Europe is considered to be the largest producer of scientific data worldwide but due to the fragmented landscape of scientific communities and infrastructures, the potential of the Big Data generated is not fully exploited. The European Open Science Cloud initiative is currently piloted¹⁷ with the ultimate goal of offering researchers and technologists a virtual environment to store, share and re-use their data across disciplines and borders. The European Data Infrastructure is underpinning this by employing the high-bandwidth networks, large-scale storage facilities and super-computer capacity necessary to effectively access and process large datasets stored in the cloud.¹⁸ EUBra-BIGSEA is following closely the activities and partners are involved in synergy activities to be built.

3.3.7 EUDAT Collaborative Data Infrastructure (CDI)

The EUDAT CDI is essentially a European e-infrastructure of integrated data services and resources to support research. This infrastructure and its services have been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines and involved at all stage of

¹⁶ https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

¹⁷ EOSCpilot project http://cordis.europa.eu/project/rcn/207500_en.html Start date 01/01/2017.

¹⁸ European Cloud Initiative to give Europe a global lead in the data-driven economy – European Commission Press Release - http://europa.eu/rapid/press-release_IP-16-1408_en.htm.

the design process. The EUDAT CDI was established in light of the European Open Science Cloud Initiative and has launched an open call for collaboration targeting both researchers and other organisations.¹⁹

3.4. Analysis of Priorities

Based on the current landscape, a set of priorities have been defined to help big data applications taking full advantage of a QoS-based analytics environment in the cloud. Such priorities reflect the challenges identified in the previous sections. As big data systems become a key economic and social driver, there is a strong need for a paradigm shift from simply building data systems to building QoS-enabled, fast, and secure data systems.

In particular, the priorities for the future research agenda are:

- Software and hardware solutions targeting high performance in-memory analytics to address near and real-time data analysis. That would help moving toward smart fast data analytics environments.
- Mechanisms and policies that implement resource partitioning and resource management on clusters in a way that cloud data centers resources are used efficiently providing differentiated service levels to customers according to the price of the resources.
- Community-oriented components for developing complex data analytics applications (e.g. interacting with multiple, heterogeneous data sources). At the same time, community-based toolbox repositories to foster machine learning libraries re-use among multiple users and also across disciplines.
- Better support for quality of service enabled big data applications, as the proper basis for elastic (vertical and horizontal) and dynamic management of resources in a cloud environment. Additionally, mechanisms for flexible management of data security controls would complement the security aspects still not properly addressed. Models to predict big data applications performance. The use of accurate models would be beneficial both for cloud providers and end users:
 - Modelling enables cloud providers to estimate applications performance based on provisioned resources and to manage effectively data center infrastructures at runtime in a way that applications can be executed with quality of service guarantees (e.g., a job will be completed within a deadline, or a data stream can be processed at a given rate without any drop). Models can be used for example in predicting the expected end of a job and to trigger the allocation of additional resources if the deadlines cannot be met.
 - Moreover, based on the performance modeling feature, cloud providers can offer what-if analysis for customers to perform trade-off between cost and application execution time. For example, customers might be willing to pay additional 10% to speedup their application by two if there is such a choice but choose not to do so if they have to pay twice more for 10% performance improvement [DA1].

4. SECURITY & PRIVACY ENSURING FREE FLOW OF DATA

4.1. Introduction

While there is increasing awareness about data privacy among consumers, it is important to bear in mind different protection levels required for different types of data. It is also important to address vulnerabilities in an evolving cyberspace and in relation to forthcoming EU regulations. EUBra-BIGSEA will make it easier to develop distributed big data applications, which will enable more organisations to use data analytics over massive amounts of data. However, added capabilities also bring privacy and security concerns. Big data analytics raises privacy concerns as the trade-off between disclosure risk and data utility, the need of

¹⁹About the EUDAT Collaborative Data Infrastructure <https://www.eudat.eu/eudat-cdi/about>.
www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

means for data owners to express their privacy preferences and the algorithms that can extract sensitive information by combining information that was already anonymised.

The big data volumes, the need for continuous movement of data, the demand for portability across systems, and the development of new systems based on the integration of existing ones, all factor in the challenges of assuring and providing evidence about privacy and security of data and systems. Allowing large volumes of data to flow across systems, countries, regions, and continents, to be used by different parties or processed close to the location where they are needed, creates the need for moving large volumes of data across different cloud infrastructures, services, providers, and consumers, distributed worldwide, in a way that allows near real-time access to data and resources. Another key aspect related to effective data flow is the need for assuring data quality (correctness, completeness, integrity, etc.), thus respecting regulations and quality of service agreements that must be specified in advance. Data flow also raise challenges related to portability, standardization, and national/regional regulations and agreements, among others.

4.2. Challenges

The protection of security and privacy in big data environments with free flow of data raises several kinds of challenges:

- Legal challenges – a common legal framework for managing security and privacy aspects.
- Public awareness challenges – awareness about the privacy of the data, and which data should be allowed to be used or not, and for which purposes.
- Standardisation challenges – standards to allow data owners to express privacy preferences.
- Technological challenges – technological solutions to deal with the massive sizes and heterogeneity of the big data datasets.
- Analytics challenges – solutions to deal with the analytics algorithms that extract privacy sensitive information from anonymised data.

4.2.1 Legal challenges – common legal framework for managing security and privacy aspects

There are many differences between Europe and Brazil regarding the legal framework for privacy management. Also, priorities between EU and Brazil are not always aligned in terms of security/privacy/infrastructure needs. Privacy-related regulations in Europe are more advanced in terms of definition and implementation. This is an area that needs to be addressed in co-operation between the two regions, in an integrated fashion supported by all the relevant stakeholders including research, industry, and policy makers. In this context, it is necessary to develop a common legal framework to ensure data privacy aspects in cloud computing, specially in the context of big data processing.

Public administrations in Brazil have been moving towards adopting cloud technology, with the Ministry of Planning investing huge resources to address barriers including security, data protection, data storage, and governance. This is demonstrated by the new legal package on the protection of personal data, underpinning a free and secure Internet, and the new cyber security law soon to be approved by the Brazilian congress. Despite these efforts, the lack of knowledge about cloud computing laws, data protection issues and taxation is causing difficulties for companies in Brazil. With the advent of cloud services in Brazil, some regulations have caused problems: companies using applications based on IaaS and SaaS must pay an import tax or alternatively hire services from companies not operating in Brazil, which increases the costs for start-ups wanting to start a business born in the cloud. The regulation exists now, but there is a need to make it operational.

4.2.2 Public awareness challenges – awareness about the privacy of the data, and which data can or should be used and for which purposes

Privacy has mostly to do with individuals, although vulnerabilities can also affect privacy, while security is a major issue for companies. Privacy, especially in public clouds, is a major challenge as it is hard to say there is a guarantee of QoS and user monitoring to observe what is actually being provided. Only after having an agreed vision of what is private data, what type of data is being moved to the cloud and what needs to be protected, we can move towards solving the technical issues. An open issue is how to enforce and control the privacy agreements, like Service Level Agreements, to assure users that the policy is conforming existing contracts.

A serious issue concerning privacy is user awareness. Privacy is a social aspect and a growing issue in the data mining/analytics communities. Solving the awareness issue is key to moving forward on technical aspects. It is important to understand if the infrastructure can support applications that are security and privacy aware, how to promote traceability of data and revoke data if lost, and what are the weakest security and privacy points in going multi-cloud. To have the results moving between platforms and sharing data with business partners, it is necessary to know how much these systems can be trusted. The Cloud must just not be a repository of data but should also provide transparency of algorithms and clear and adequate guarantees. Empowering users and raising awareness of their privacy is a key challenge for the future.

4.2.3 Standardisation challenges – standards to allow data owners to express privacy preferences

There is a lack of standards related to big data, the expression of privacy preferences and the representation of the privacy concerns associated with data analytics algorithms. Particularly, there are no standardised ways to represent the data and meta-data between the two regions, which substantially limits the applicability of new solutions, especially for innovative SMEs. Standards are essential to ensure the exchange of data between machines, systems and software within a networked value chain. If data and communication protocols are proprietary or related only for certain regions, this may impact competition and flexibility. In cloud computing, standardisation is part of the solution for the coordination problem, in which all parties can realise mutual gains, but only by making consistent decisions.

Standardisation is key to maximising compatibility, interoperability, repeatability, or quality. It can facilitate custom processes, procedures, products, and services. In practice, the demand for standardisation will be exacerbated by the need for sharing data across systems, countries, regions, and continents, both for the cases of open data and proprietary data. Standards will be essential for managing security and privacy aspects of big data in the cloud, including enforcing application developers to handle personal data adequately, while allowing people to define fine-grained access control policies for their data.

4.2.4 Technological challenges – technological solutions to deal with the massive sizes and heterogeneity of the big data datasets

Protecting large volumes of data during storage and processing against unauthorised and malicious access (attackers) requires disruptive approaches for data encryption, obfuscation and masking, including the need for mechanisms that are able to decide what, when and how data and information can be released. Moreover, trusting data provenance in cloud environments, in which the data can be accessed at any time, from anywhere and any terminal, is a challenging issue. Obviously, these can only be achieved if security aspects are assured at the storage, communication and processing levels, which will require the real application of defence-in-depth approaches that cover all aspects and layers of cloud infrastructures and services.

Policies that enforce application developers to handle personal data adequately should be put in place. Application developers and their users should be aware of the risks when data is not handled in a secure and privacy-friendly way. This includes the development of a framework for data collection and protection regulations between EU and BR, as the way forward to cross-regional innovation. Adding to the expected big data volumes, privacy and security challenges will be intensified by the need for continuous movement of data, by the demand for portability across systems, and by the development of new systems based on the integration of existing ones, among others. Assuring and providing (communicating) evidences about privacy and security will also be major challenges in the future.

Allowing large volumes of data to flow across systems, countries, regions, and continents, to be used by different parties or processed close to the location where they are needed, will raise several challenges. Of importance will be the need for moving large volumes of data across different cloud infrastructures, services, providers, and consumers, distributed worldwide, in a way that allows nearly real-time access to data and resources. Data flow raises challenges related to privacy (where the data should maintain their privacy properties), portability (data should be readable by different systems and services), standardisation (to define clear data formats and flow rules, thus supporting privacy and portability), and national/regional regulations and agreements (to support privacy, portability, and standardisation, while regulating what, when and how data can be shared), among others.

4.2.5 Analytics challenges – solutions to deal with the analytics algorithms that extract privacy sensitive information from anonymised data

Big data analytics processes are interested in statistical data to identify trends, models and patterns among larger groups of information (which includes sensitive and personally identifiable information). However, in some cases, it is possible to obtain information that identifies specific individuals, violating their privacy. Thus, information must be obtained from databases only for the purpose of analysis, description and prediction, without revealing specific data of an individual. Mechanisms for anonymisation must be implemented to protect the user identification. However, excessive anonymisation can make the disclosed data less useful to the recipients because some analysis becomes impossible or the analysis produces biased and incorrect results. So, it is necessary to understand the utility of anonymised data, as the information extracted from an anonymised database must remain useful and relevant. Another important measure is the disclosure risk. Even when anonymisation methods are applied, there is a risk of data disclosure, i.e., a probability of data re-identification.

Besides researching techniques to measure the data utility and disclosure risk, it is necessary to devise techniques that can prevent that privacy sensitive information is disclosed as result of the algorithms. There are three main types of algorithms: 1) algorithms that produce data with less privacy concerns; 2) algorithms that maintain the privacy concerns of the data; and 3) algorithms that extract data with higher privacy concerns. The last case is the challenge that needs to be addressed, both by providing the developers with means to characterize their algorithms and by researching means to classify the algorithms in an automated fashion.

4.3. Relevant Initiatives

There are several on-going projects that are currently addressing parts of the challenges mentioned above. We highlight the following relevant initiatives:

- **EC Cluster on Data Protection, Security and Privacy (DPSP Cluster)²⁰**

The EC Cluster on data protection, security and privacy (DPSP) brings together over 20 EU projects on cloud computing that are addressing research and innovation on diverse solutions ensuring data protection, security and privacy.

²⁰ <https://eucloudclusters.wordpress.com/data-protection-security-and-privacy-in-the-cloud/>.
www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

The main objectives of the cluster are to: maximise the impact of EU-funded research and innovation project results; ensure market orientation and adoption; help define the research and innovation needs in H2020. The Cluster has produced a white paper on Challenges for trustworthy (multi-)cloud-based services in the Digital Single Market (January 2016) and is currently working on a paper on the Free Flow of Data, again in the context of the DSM.

The main findings of the Cluster are relevant to EUBra-BIGSEA in relation to the topics covered, providing insights into emerging solutions, as well as on future challenges that have yet to be addressed.

- **H2020 EUBrasilCloudFORUM²¹**

EUBrasilCloudFORUM: Fostering an International dialogue between Europe and Brazil aims to facilitate the policy and technical dialogues between the European Union (EU) and Brazil in focus areas related to cloud computing, including security aspects. The project will establish an organisational co-operation forum that enables the European Union and Brazil to formulate and develop a common strategy and approach for research and innovation in cloud computing, including security aspects.

EUBrasilCloudFORUM is clearly of interest to EUBra-BIGSEA, by offering a platform for communicating project results on the one hand, and by sharing policy priorities on EU-BR research and innovation on the other.

- **H2020 CLARUS²²**

CLARUS aims at improving trust in cloud computing and securely unlocking sensitive data to enable new and better cloud services. CLARUS is developing a secure framework for storing and processing data outsourced to the cloud so end-users can monitor, audit and control their stored data while gaining the cost-saving benefits and capacity that cloud services bring. Trust can be established with solid mechanisms that ensure greater control over the security and privacy of the user's data. Confidentiality and privacy are still major concerns when it comes to moving to the cloud. Also, the project will provide means to protect business assets from vulnerabilities or attacks and ensure that applications continue to operate and provide a good level of service even during an attack.

CLARUS results are interesting for EUBra-BIGSEA in terms of the privacy-preserving mechanisms and are expected during the project's lifetime. CLARUS is also a contributing project in the EC DPSP Cluster.

- **H2020 CloudWATCH2²³**

CloudWATCH2 will support the European Research and Innovation (R&I) project with strategical thinking and guidelines in order to improve the projects ability to provide results with an impact on the market, so they can be sustainable. It will help these initiatives with strategic guidance on the value proposition and business case so that they can take their outputs to market through pricing transparency, improved risk assessment, security and legal guides, an evolved portfolio of standards for interoperability and security, and a mapping of technologies, development status and practical support activities.

CloudWATCH2 results are interesting for EUBra-BIGSEA because it can help the project with the strategy to make the framework to be developed sustainable.

- **H2020 SecureCloud²⁴**

The **SecureCloud** project aims to remove technical impediments to dependable cloud computing, i.e., SecureCloud will ensure the confidentiality, integrity, availability and security of applications and their data. Thereby, SecureCloud will encourage and enable a greater uptake of cost-effective, environment-friendly, and innovative cloud solutions, in particular, for critical infrastructure applications throughout Europe and Brazil. The primary goal of SecureCloud is to ensure the dependability of critical applications that are executed in distributed, potentially untrusted cloud infrastructures.

²¹ <https://eubrasilcloudforum.eu/>.

²² <http://www.clarussecure.eu/>.

²³ <http://www.cloudwatchhub.eu/>.

²⁴ <https://www.securecloudproject.eu/>.

SecureCloud is addressing the challenge of data confidentiality (among other challenges) in untrusted cloud infrastructures. The solutions proposed can be used to improve the solutions of EUBra-BIGSEA, extending them to take advantage of a broader spectrum of infrastructures.

- **H2020 SafeCloud²⁵**

There are major privacy and security concerns about data located in the cloud, especially when data is physically located, processed, or must transit outside the legal jurisdiction of its rightful owner. **SafeCloud** will re-architect cloud infrastructures to ensure that data transmission, storage, and processing can be 1) partitioned in multiple administrative domains that are unlikely to collude, so that sensitive data can be protected by design; 2) entangled with inter-dependencies that make it impossible for any of the domains to tamper with its integrity.

SafeCloud will ensure the users that their data is stored and processed in a partitioned fashion, which will endow the users with more control over the confidentiality and integrity of their data. These techniques can improve the solutions of EUBra-BIGSEA to allow it to be used in domains with higher criticality.

- **Assured Cloud Computing²⁶**

The Assured Cloud Computing-University Center of Excellence (ACC-UCoE) is a joint effort of the Air Force Office of Scientific Research (AFOSR), Air Force Research Laboratory Technology Directorate (AFRL), the Information Trust Institute (ITI) and the University of Illinois at Urbana-Champaign (Illinois) performing state of the art research by providing technical exchange and educating students in vital secure cloud computing sciences and technologies needed to fly, fight, and win in air, space, as well as cyberspace. To meet these needs, the ACC-UCoE offers the expertise of over 85+ research faculty at the Information Trust Institute (ITI). Since 2004, ITI has supported almost \$60M in sponsored research into trustworthy systems.

ACC will research algorithms that detect security policy or reliability requirement violations in cloud environments, which are also useful for the privacy challenges being addressed in EUBra-BIGSEA.

4.4. Analysis of Priorities

From the **point of view of the EUBra-BIGSEA project, and its timeframe**, several research goals should be priority, such as:

- **Definition of a Standard Privacy Policy Format**

Although some standard privacy policies exist (e.g., P3P, EPAL), they are focused on the end user, allowing them to express their preferences. It is necessary that privacy policies are defined by the data source owners and we need to define a machine-readable format for this policy, which allows specifying the information that must be protected. These policies shall be interpreted and enforced by the framework, through appropriate anonymisation and access control techniques. Privacy-related policies can be organised in a hierarchy: high-level policies are described in natural language; low-level policies are specified in machine-readable format, and used by the application itself. Reproducing high-level statements in machine-readable statements is a big challenge due to the semantics involved. The lower the level, the lower is the impact of semantics.

- **Avoidance of Statistical Disclosure**

Information obtained from the datasets must not reveal specific individual data. Statistical disclosure control methods must be used to protect the user identification. The database anonymisation will be implemented according with privacy policies defined by data source owner. The SDC must also help preventing privacy attacks (e.g., attack of attribute connection, attack of register connection and attack of table connection). It will be necessary to follow the data owner anonymization police for the Big Data.

- **Efficient Data Anonymisation/Obfuscation Techniques**

²⁵ <http://www.safecloud-project.eu/>.

²⁶ <http://assured-cloud-computing.illinois.edu/>.

Efficient techniques for data anonymisation and obfuscation are necessary (e.g., generalisation, suppression, encryption, perturbation, masking, etc.). These techniques can be used or combined with each other to reach the best results. Due to the context of big data, it is imperative that the techniques used have reduced performance impact, while keeping acceptable levels of privacy protection. The action to be taken in this context is to select solutions that, in the context of the project, have this impact reduced and keep acceptable levels of privacy protection.

- **Query Anonymisation**

The data analytics team may prefer to access the raw data of the database or, in other words, prefer to access the non-anonymised tables. In these cases, it is important to anonymize the results of the queries performed by the data analytics team. There are some techniques that enable query anonymisation, but these techniques have not been applied in the Big Data and Data Analytics context, especially regarding the use of data policies as basis for the anonymisation. It is thus necessary to provide a technique to execute the query anonymisation in the context of this project.

Furthermore, there are a set of research goals that are of **lower priority in the context and time frame of EUBra-BIGSEA** project, but that are also essential to overcome these goals, as follows.

- **Measurement of Data Utility and Disclosure Risk after anonymisation**

There are some state-of-the-art techniques that enable measuring Data Utility and Disclosure Risk, but these techniques were not applied in the Big Data context. It is necessary to evaluate (or improve) the use of these techniques in the context of this project, provide improvements to the measurement techniques, and investigate the feasibility of using algorithms input/output privacy, to be expressed by developers, in the Disclosure Risk Measurement.

- **Privacy Violation Detection**

There are some tools that verify if the user's privacy is violated, but their focus is on privacy policies compatibility and negotiation protocols, with the goal of not allowing information traffic between clients and services with inconsistent privacy policies. For data leakage detection, some models were proposed but their focus are in detecting the agent who leaked the information. We need automatic tools that, similarly to intrusion detection systems, and based on privacy policies defined by data source owners, detect and avoid data leakage and privacy attacks.

5. SMART CITIES AND URBAN MOBILITY PLANNING

5.1. Introduction

The concept of smart cities is usually linked to efficiency in the use of natural resources [Souza et al. 2015]. Public transportation is a critical factor for the functioning of a city. It provides mobility to the masses and helps to mitigate traffic and pollution. With the advent of smart technologies, urban transportation systems are able to capture a lot of useful data. Such data can be used to shed light on a number of factors, including user trends and traffic patterns. These items are essential for urban planning, optimising the transportation system (e.g., fuel, time), reducing environmental impacts of mass transport (e.g., noise pollution, air pollution).

Public transportation is one of the most critical areas of smart cities. In Brazil, the vehicle fleet in major cities grew more than the road structure²⁷. Mobility challenges have already gained attention of computer science community in Brazil²⁸. In particular, these challenges can be grouped in the following areas: (i) discovery of patterns, (ii) data statistics, (iii) data integration, (iv) location and tracking, (v) open and

²⁷<http://www1.folha.uol.com.br/cotidiano/2014/08/1503030-frota-de-veiculos-cresce-mais-rapido-que-a-estrutura-viaria-no-pais.shtml>.

²⁸ <http://www.sbc.org.br/documentos-da-sbc/send/141-grandes-desafios/802-grandesdesafiosdacomputaonobrasil>. – www.eubra-bigsea.eu | contact@eubra-bigsea.eu | [@bigsea_eubr](https://twitter.com/bigsea_eubr)

connected data, (vi) contextual information, (vii) security and privacy, (viii) energy and management, (ix) use of cloud resources, and (x) trajectories with semantic information, among others.

EUBra-BIGSEA addresses the mobility challenge (by the public transportation view), from the top 5 perspectives previously mentioned, for a case study in Curitiba. Curitiba, a city located in the south of Brazil, with 1.8 million people in a total area of 430.9 km², according to the Brazilian Institute of Geography and Statistics (IBGE)²⁹. This area encompasses 75 districts, and is surrounded by other 29 cities, known as the metropolitan region of Curitiba (with the Portuguese acronym RMC).

Curitiba has developed and implemented mass transport corridors, densification of land-use along these corridors, and mobility solutions using Bus Rapid Transit (BRT) systems in the 1970s, where one main feature of the success of the system is its complex network of feeder lines [Duarte et al. 2016]. The city has also been participating in the open data initiative, through several government stakeholders (such as Instituto de Planejamento de Curitiba (IPPUC)³⁰ and the Municipality of Curitiba³¹) and competitions (such as Hackathons³² using open data).

5.2. Challenges

The Acquisition and Engineering of Georeferenced Environmental, Stationary, Streaming and Social (GES³) data are related to a Data Acquisition use case (see EUBra-BIGSEA Deliverable *D7.1: End User Requirements Elicitation*). In particular, these data sources are related to urban traffic and cover four main data types: stationary data, dynamic spatial data, environmental data and social network data. Despite that they are related to the city of Curitiba, where the pilot case is being constructed, the EUBra-BIGSEA framework should be extensible to other scenarios.

Therefore, the data integration 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. It should be noted that some of them are official (original data from municipalities) and some of them are informal (Google/OpenstreetMaps). Issues surrounding integration include missing entities (new streets and parks at a new district), difference among sources (Google has an average of 60% of the available bus stops in Curitiba, GTFS has not the same bus lines), mixing non-structured with tabular and spatial data, as long as different precision, accuracy and consistency. In summary, issues related to data quality, entity matching, and data mining. Several preprocessing steps will be applied within this phase, so the final user can interact (query) an integrated source.

Additional challenges include: validation of the suitability of the EUBra-BIGSEA platform (scalability, interoperability tests and comparisons with different technologies), and how to use the application to provide a feedback to the municipality stakeholders, citizens and non official applications (among others).

In summary, these challenges are related to the mobility challenges already cited by Brazilian Computer Science community: (i) discovery of patterns, (ii) data statistics, (iii) data integration, (iv) location and tracking, (v) open and connected data, (vi) contextual information, (vii) security and privacy, (viii) energy and management, (ix) use of cloud resources, (x) trajectories with semantic information, among others.

5.3. Relevant Initiatives

Of particular interest for EUBra-BIGSEA is the Open & Agile Smart Cities Task Force (OASC), an open innovation initiative for improving smart city services, first presented at CeBIT 2015 by 25 cities from 6 EU countries –Belgium, Denmark, Finland, Italy, Portugal and Spain– and by 6 cities from Brazil.

²⁹ <http://www.ibge.gov.br>.

³⁰ <http://ippuc.org.br/>.

³¹ <http://www.curitiba.pr.gov.br/DADOSABERTOS/>.

³² www.hackathonparana.pr.gov.br/.

The aim of the initiative is to make it easier for city councils and start-ups to improve services as transport, energy efficiency or e-health. This project will be achieved thanks to FIWARE, an EU-funded open source platform and cloud-based building blocks that can be used to develop a lot of different applications in the huge range of topics that Smart Cities tackle.

The EU and Brazilian signing cities will be able to share open data which will also allow start-ups to develop apps that could have a benefit for all citizens. Furthermore, the new systems that emerge from this platform will be shared between cities.

The EU signing cities are: Brussels, Ghent, Antwerp (Belgique), Copenhagen, Aarhus, Aalborg (Denmark), Helsinki, Tampere, Espoo, Vantaa, Oulu, Turku (Finland), Milan, Palermo, Lecce (Italy), Lisbon, Porto, Penala, Fundão, Palmela, Águeda (Portugal), Valencia, Santander, Málaga, Sevilla (Spain).

The Brazilian signing cities are: Olinda (Recife), Anapólis (Goiás), Porto Alegre (Rio Grande do Sul), Vitória (Espírito Santo), Colinas de Tocantins (Tocantins) and Taquaritinga (São Paulo).

Relevance for EUBra-BIGSEA

The initiative is relevant to EUBra-BIGSEA in terms of smart city application deployments and usage of open data. Both topics can be proposed to EUBrasilCloudFORUM as a potential topic for Cloudscape 2017 to discuss how developments within the OASC and EUBra-BIGSEA can benefit EU-Brazil co-operation and how they are impacting the use of cloud-based approaches, as well as how open innovation initiatives could be taken forward in future joint co-operation.

From the data integration perspective, data integration portals not only provide information, but also enable users to suggest datasets, give feedback and share the use that they have made with the data from the portal.

In most cases, the data integration portals include urban data of education, security, economy, finance and so on from countries such as Europe, through European Data Portal³³ and Brazil, through Brazilian Open Data Portal³⁴. However, these portals may encompass the most heterogeneous types of data, such as the CDP – Carbon Disclosure Program portal³⁵ that provides global data about the environment (e.g. forest risks assessment, global emissions, water).

There are also interest groups that created open data portals to analyze government data, such as Working Group on Open Government Data³⁶ and the European Union that created the European Union Open Data Portal³⁷.

5.4. Analysis of Priorities

An important priority is dealing with the interoperability of the GIS information in different providers. Interoperability is not only a matter of data formats, but data consistency. Urban maps of Google Maps, MapBox and other providers differ and therefore reduce the capability of integrating and validating different layers. This is one of the objectives of the Entity Matching-as-a-Service (EMaaS) and Data quality as a service (DQaaS) in EUBra-BIGSEA.

Another priority identified is the provision of routes according to other criteria rather than the quickest path. There are many examples on daily news about the risks of blindly using routing information from car navigators. The quickest path may not be the most pleasant, less crowded, more touristic or safer route.

³³ <http://www.europeandataportal.eu>.

³⁴ <http://dados.gov.br>.

³⁵ <https://data.cdp.net>.

³⁶ (<http://opengovernmentdata.org>.

³⁷ <https://data.europa.eu/euodp/en/data>.

We would prefer to skip conflictive streets, crowded buses, industrial areas, old buses, etc. and we would prefer to follow paths close to landmarks, new suburbs or more popular routes.

6. STANDARDS, INTEROPERABILITY & PORTABILITY

6.1. Introduction

Standards and interoperability are very relevant and strictly related issues. Standards are crucial for enabling interoperable, secure and trusted clouds, and thereby for increasing confidence and uptake. Moreover, standards are one of the most important means to bring new technologies to market. By transferring research findings into guidance documents, standards provide a bridge connecting research to industry. This connectivity is critical to successful commercialisation and meeting growing consumer demands for more standards in cloud computing to address fears over issues such as lack of control, security and vendor lock-in. Consumers are increasingly concerned about the lack of control, interoperability and portability, which are central to avoiding vendor lock-in. Public open standards offer protection from vendor lock-in and licensing issues, therefore avoiding significant migration costs if not provided.

The design of the EUBra-BIGSEA platform takes into account standards and interoperability. In the development of the fast and Big Data ecosystem several aspects are taken into account like the adoption/implementation of (i) interoperable data service interfaces and metadata standards (e.g. OGC-WPS/WCS and ISO19115/19139, respectively, for geospatial data) and (ii) standard data formats/conventions (e.g. NetCDF-CF for environmental data). EUBra-BIGSEA also addresses the interoperability between different layers of the architecture and focuses on the integration and interoperability at the level of the services. These developments rely on the adoption of existing open standards and will be shared with the community thanks for example to partner's participation to bodies as the TOSCA consortium. The extension to the TOSCA specification will specify deployment options to automate application, data packaging and delivery strategy for the considered class of Big Data cloud applications. Through its partners, EUBra-BIGSEA is investigating the ways in which the project technologies may influence either the TOSCA Interop sub-committee, by providing big data interoperability packages, or the TOSCA notation in general, by providing big data-specific TOSCA-based node types and blueprints.

6.2. Challenges & Priorities

Challenges related to programming frameworks

- Smart computation. There is a need to extend the interoperability of the programming frameworks in order to include new devices that build smart infrastructures (IoT, smart cities related, etc) and are not as easy to manage as cloud resources.
- Smart runtime. The runtime needs to be able to handle distribution, parallelism and heterogeneity in the resources transparently to the application programmer, and on the other hand has to be able to handle data regardless of location by supporting a single and unified data model. The availability of these hybrid deployments further implies the need to reinforce the enactment of interoperability and portability.
- Standards to support cloud interoperability and portability exist, but gaps remain in standardisation, specifically in the PaaS area. Moreover, some of the current standards need to mature in order to describe how services interoperate and how data can be readily ported between cloud offerings.

Priorities

- Evolve the programming environments to emerging paradigms as fog computing.
- Further improve the programming framework to be more data centric. BSC for example has as one of its main activities the integration between programming models and persistent storage. The aim is to enable data to include its processing code as an indivisible part to ease the deployment of

code into devices as well as the need to offload the data to be processed in larger nodes seamlessly.

6.3. Related initiatives

NIST³⁸

Reference studies and frameworks for cloud computing, cyber security and cyber physical systems, as well as participation in ISO standardisation in relation to SLAs and metrics.

[NIST 800-53 Rev.4 Security Controls](#) – NIST: holistic approach to information security and risk management with security controls needed to strengthen their information systems and environments in which they operate, contributing to systems that are more resilient in the face of cyber and other threats.

[NIST Security Reference Architecture](#): risk-based approach of establishing responsibilities for implementing security controls throughout the cloud lifecycle.

CloudWATCH2³⁹

Coordination and Support Action for Cloud Computing under EC Unit E2. Focus areas include risk management, legal aspects of cloud contracts, exploitation of results from EU cloud R&I, and cloud standards profiling and mapping.

This is the EU reference project against which to measure progress on the state of the art in cloud computing, as well as for monitoring progress on relevant standards and their status. With regard to standardisation, CloudWATCH2 provides:

- Cloud Standards Catalogue (February 2017): mapping and analysing standards implementation in the context of European research and innovation projects. The aim is to gain clarification on the value created from standards implementation for interoperability and security (<http://www.cloudwatchhub.eu/cloud-standards-mapping>).
- Standards Plugfests: interoperability testing in virtual settings (<http://www.cloudwatchhub.eu/standards-plugfests>).

DICE project⁴⁰

The DICE project is developing a framework and an UML profile for the design and quality analyses of data intensive applications. DICE UML models, relying on model to model and model to text transformations, can be automatically translated into TOSCA blueprints and deployed through an extended Cloudify orchestration engine. Thanks to the participation of Polimi to the DICE consortium, EUBRA-BIGSEA will follow DICE activities.

7. PRELIMINARY ACTION PLAN

The preliminary roadmap that defines the proposed action plan and is presented below. The Roadmap covers actions in relation to big data, security and privacy, smart cities and mobility, and standards, interoperability and portability.

Theme: Big Data		
Initiative	EUBra-BIGSEA relevance	Action plan
RDA – Big Data IG and Array	EUBra-BIGSEA partners are	EUBra-BIGSEA will continue to

³⁸ <https://www.nist.gov/>.

³⁹ <http://www.cloudwatchhub.eu/>.

⁴⁰ <http://www.dice-h2020.eu/>

Database Assessment WG	following the activity in the WG and in particular the adoption of OGC standards like OGC-Web Processing Service for geospatial platforms.	contribute to the RDA groups and promote the outcomes to the project's community members.
Big Data Value Association (BDVA)	The link with the European Technology Platform for High-Performance Computing, with specific regard to high performance data analytics aspects.	Monitoring of activities, identifying opportunities to play
Theme: Security and privacy – Free flow of data		
Initiative	EUBra-BIGSEA relevance	Action plan
EC DPSP Cluster	Focuses on cloud computing and related issues of security, data protection and privacy.	Monitor on-going analysis of challenges related to the free flow of data and the role played by relevant projects within the cluster. Promote outcomes via the project's stakeholder community.
EUBrasilCloudFORUM	CSA for EU-BR funded initiatives (Call 3).	Use as a forum for raising awareness about EUBra-BIGSEA. Jointly share insights on cloud computing R&I in EU and Brazil. Contribute to the Policy Dialogue. Promote outputs on forthcoming marketplace.
CLARUS	Privacy-preserving mechanisms, use case on geo-reference data, standardisation guide.	Promote relevant outcomes of CLARUS to the project's stakeholder community.
H2020 SecureCloud	Data confidentiality in untrusted cloud infrastructures.	Evaluate potential for exploitation of results. Jointly share insights on R&I outputs and new expertise.
H2020 SafeCloud	Data-intensive cloud applications	Evaluate potential for exploitation of results. Jointly share insights on R&I outputs and new expertise.
EOSC Pilot Project	EUBra-BIGSEA partners are involved in the project.	Follow the developments and identify synergy potential.
EUDAT Collaborative Data	EUBra-BIGSEA partners are	Follow developments and

Initiative	involved in the project.	identify synergy potential. The data focused discipline communities are a key stakeholder for EUBra-BIGSEA.
Theme: Smart cities & urban mobility planning		
Initiative	EUBra-BIGSEA relevance	Action plan
Open & Agile Smart Cities Task Force (OASC)	Complementary smart city developments, based on open data and cloud approaches. Potential for future joint actions under EU-Brazil co-operation funding.	Discuss a potential joint synergy through EUBrasilCloudFORUM, also to gauge interest in Cloudscape Brasil 2017 participation fostering a two-way dialogue on common topics. Liaise also with FIWARE-Mundus, which covered some aspects of Cloudscape Brasil 2016.
Initiative	EUBra-BIGSEA relevance	Action plan
Data Integration Portals such as the European Data Portal or the Brazilian Open Data Portal	Addressing the challenge of GIS information interoperability	Follow developments – particularly relevant for the challenges related to collecting, harmonising and integrating diverse data related to urban traffic and management: stationary data, dynamic spatial data, environmental data and social network data.
Theme: Standards, Interoperability and Portability		
Initiative	EUBra-BIGSEA relevance	Action plan
CloudWATCH2	Standards Catalogue and Interoperability testing	EUBra-BIGSEA will use the standards catalogue as a guide for implementation of relevant standards to assess best practices in current R&I initiatives. Draw on the results of last plugfest with its focus on security standards as the basis for participation in future plugfests. Use the hub as a forum for sharing insights and results.

Table 1: EUBra-BIGSEA Roadmap for Preliminary Action Plan

GLOSSARY

Term	Explanation
API	Application Programming Interface
COMPSSs	COMP Superscalar (COMPSSs) - programming model which aims to ease the development of applications for distributed infrastructures, such as Clusters, Grids and Clouds
DAG	Directed acyclic graph
DPSP	EC Cluster on data protection, security and privacy
DQaaS	Data quality as a service
EMaaS	Entity Matching-as-a-Service
GES ³	Georeferenced Environmental, Stationary, Streaming and Social
GDPR	EU General Data Protection Regulation
IaaS	Infrastructure-as-a-Services
IDE	Integrated Development Environment
IG	Interest group
IoT	Internet of Things - connection via the Internet of various computing devices embedded in objects enabling them to transfer data
LEMONADE	Live Environment for Mining Of Non-trivial Amount of Data from Everywhere
MESOS	A Resource Management platform that abstracts CPU, memory, storage, and other compute resources away from machines
NISD	EU Network and Information Security Directive
Ophidia	A CMCC Foundation research project addressing big data challenges for eScience
OpenStack	OpenStack cloud management platform
SaaS	Software as a 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
WG	Working Group