Project Title	High-performance data-centric stack for big data applications and operations
Project Acronym	BigDataStack
Grant Agreement No	779747
Instrument	Research and Innovation action
Call	Information and Communication Technologies Call (H2020-ICT-2016-2017)
Start Date of Project	01/01/2018
Duration of Project	36 months
Project Website	http://bigdatastack.eu/

D1.4 – Data Management Plan

Work Package	WP 1 – Project Coordination
Lead Author (Org)	Yosef Moatti (IBM)
Contributing Author(s) (Org)	
Due Date	30.06.2018
Date	30.06.2018 (Re-submission: 03.10.2019)
Version	2.1

Dissemination Level

X	PU: Public (*on-line platform)
	PP: Restricted to other program participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

The work described in this document has been conducted within the project BigDataStack. This project has received funding from the European Union's Horizon 2020 (H2020) research and innovation programme under the Grant Agreement no 779747. This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of such content.

Versioning and contribution history

Version	Date	Author	Notes
1.0	25.05.2018	Eliot Salant (IBM)	Initial version
1.1	18.07.2018	Eliot Salant (IBM)	Explanation to section 4.2 about opting out of the Pilot for Open Research Data
1.2	27.09.2019	Yosef Moatti (IBM)	Draft of major revision following rejection as of Sept. 4th
1.2	28.09.2019	Bernat Quesada Navidad (ATOS-WL)	Review
1.3	28.09.2019	Maurizio Megliola (GFT)	Review
2.0	30.09.2019	Dimosthenis Kyriazis (UPRC)	Candidate final
2.1	03.10.2019	Yosef Moatti (IBM)	Artifacts details added

CC BY NC ND

Table of Contents

CC BY NC ND

1 Executive Summary

This new deliverable version aims to present a plan for the data management, collection, generation, storage and preservation related to BigDataStack activities. In this action, we envision five different types of data:

- data provided by the use cases
- publications (e.g., conference papers)
- public deliverables
- open source software
- artifacts of research value produced by application executions (e.g., logs, playbooks)

The document presents, following the EC template [1], how these different types of data will be collected, who the main beneficiaries are, and how BigDataStack will store them, manage them, and if the project will make them accessible, findable and re-usable. The text continues with the foreseen resources needed for the openness and data to finalize with security and ethical aspects that will be taken into consideration in the context of BigDataStack.
This plan may be updated in the M36 version of Technical Reports, having as input the work carried out in the use cases (WP6), the technical work packages (WP2 – WP5) and the dissemination activities (WP7).

2 Introduction

2.1 About this Deliverable

This deliverable focuses on the management of the data in BigDataStack.

The following kind of data will have to be handled:

First one are the **data sets provided by the use cases** (either directly such as in the case of Danaos, or through their customers such as in the cases of Atos Worldline and GFT) and used to validate the project. Each use case data has its specific requirements. Although anonymized, these data sets are consortium confidential. Moreover, they may be complemented by open data sets. This already has happened in the first half of the project where NOAA [2] weather data was used to complement the vessels data provided by Danaos. In the second half of the project, GFT intends to augment insurance data with public data sets for the Insurance use-case.

The second kind of data are the **publications** that have and will be published. Here the main concern is to make sure that FAIR principles will be adhered to.

The third kind of data are the **deliverables**. Except for the very few that are consortium private, all these deliverables have and will be made publicly and freely accessible from the Project Web site.

The fourth kind of data are the **open source software artifacts**. As of September 2019, substantial code output of the project has already been up streamed to big Open Source

CC BY NC ND

projects such as OpenStack. During the second part of the project, we envision further contributions to OS projects.

Fifth and last kind of data: **artifacts of research value** obtained from the BigDataStack infrastructure. For instance, logs or playbook generated by applications being executed over BigDataStack may be of research interest. In this case, FAIR principles will be applied to them.

2.2 Document structure

This version of the document follows the established H2020 template for a Data Management Plan (DMP) [1]. Section 2 presents a summary of what the purpose of the data collection and generation is in the case of BigDataStack. Section 3 explains how the data and metadata will be made FAIR, and thus Findable, Accessible, Interoperable and Reusable. Section 4 briefly explains how the financial resources for this openness are envisioned at this stage to be allocated. Section 5 and 6 focus on the security and ethical aspects respectively. Section 7 presents the conclusions and future work.

CC BY NC ND

3 Data Summary

In this chapter, we provide a data summary of the 5 kind of data relevant to the project (as described previously in Section 2.1):

- **Data provided by the use cases:** The purpose is obviously to validate the project with real use case. Version 2.0 of deliverable D6.1 (Use case description and implementation) details in length how the use case datasets (Connected Customer of ATOS WorldLine and Insurance Use Case of GFT) are anonymized to avoid any GDPR concern.

  The ship management data is Danaos proprietary. The vessel name has been replaced by an id number. The Danaos dataset has to be kept consortium confidential. It was uploaded to IBM Cloud Object Storage after agreement of Danaos and taking into account the high security provided by this cloud service. This data will typically be augmented with publicly available data, such as weather data [2], which will then be correlated based on geographical coordinates with the original data.

  The Connected Customer data provided by ATOS originates from one of its clients: EROSKI, which is a leading vendor in Spain. The data is anonymized, the customer identity being represented by a customer ID, which is EROSKI private information. In addition, the other data fields which pertain only to the sale details could not permit to reconstruct the identity of the customer. EROSKI has requested and obtained special caution for this data set: its storage requires compliance with security standard ISO 27001 which is the case for the testbed provided by UBITECH partner. This requirement has been added in the Consortium Agreement amendment #2, which was completed and signed by all partners in February 2019.

  The Insurance data to be provided by GFT originates from HDI Assicurazioni, which is part of Talanx Group of Hannover: a large German insurance group. This data has high GDPR sensitivity, however an ad-hoc process (for details, please refer to version 2.0 of D6.1 deliverable) has been set-up. The dataset goes through an anonymization process, according to MD5 encryption, which removes all possibility to reconstruct the identity of the customer. Prior to this customization process, a GFT person is allowed to augment the dataset with open data, so that even, if, for instance, the address of the customer is anonymized, it can be used prior to the anonymization to augment the data record with open information related to the address.

  All these three datasets come as formatted text data files (e.g., CSV).

- **Publications** (e.g., conference and journal peer review papers): In the H2020 framework programme, publications are required to be openly accessible. All the partners have been made aware of this and are expected, if relevant, to pay the fees to publishers needed so that their publications are openly accessible.

- **Public deliverables:** All the deliverables which are marked with Public access have

CC BY NC ND

and will be put on line in the project site

- **Open source software artifacts:** Software that are produced in relation to the project will be of one of the following types:
  - Closed software. An example is the LeanXcale code, which is private to the company and will not be published as open sources.
  - Open source code up streamed to big open source projects. An example is the OpenStack project to which RedHat has already contributed.
  - Open source code that will be made available through open repositories (e.g. Github) allowing its utilization by other entities (e.g. stakeholders, researchers, etc) following partners' exploitation paths and plans.

- **Artifacts of research value produced by application executions** (e.g., logs, playbooks, etc): Various components of the BigDataStack infrastructure when running big data applications produce artifacts may be of research interest. Application, input data and deployment related artifacts are of interest. The following gives more details on their possible structure:

Application
- ID
- Attributes (e.g. typology, owner, etc.)
- Application components
  - Pods / Kubernetes services
  - Containers

Input Data
- Input rate (expected)
- Size (expected)
- …

Deployment
- ID
- Date
- Attributes (e.g. typology such as "production", "test" or "experiment", owner, etc.)
- Resources requested
  - HDD/Disk, RAM/Mem, CPU per container
  - Number of replicas per pod
- Logging data
  - Raw logging of the containers' resources use
    - CPU
    - Mem
    - Network
  - Resource cluster metrics (i.e. Openshift, to monitor how busy are the nodes of the Kubernetes cluster).
  - Data services metrics (e.g. LeanXcale, Spark, CEP, etc.).
  - Application performance metrics (i.e. indicators such as response time, throughput, error rate, etc.)

CC BY NC ND

- Logs of QoS alerts (notifying violations of the SLOs)
- Logs of decisions and actions by the data-driven infrastructure self-adapting process (e.g. decisions made by the Dynamic Orchestrator and LeanXcale, decisions taken by the ADS-Ranking & Deploy

The artifacts will be packaged (e.g., zipped) and uploaded to the BigDataStack OpenAIRE page. This will permit open access and reuse of such artifacts.

The collected data content may be of interest to both the commercial sectors from which they were collected, as well as to a wider community of data scientists, or students of data science, to carry out machine learning research. Big Data infrastructure developers may be interested in the data as a means of performance testing against large volumes of data.

CC BY NC ND

4 FAIR (Findable, Accessible, Interoperable, Reusable) data

4.1 Making data findable, including provisions for metadata

The datasets for the three use cases are consortium-confidential due to their business value, and may not be circulated outside of the consortium.

Concerning the possible creation and publication of research value artifacts, we think it is premature to plan naming conventions, keywords, etc. since it is not yet clear exactly which datasets will be made available. The technical leaders have been made aware of these requirements and following the second release of all project components, the corresponding datasets will be identified and communicated to the public as described in Section 3 of this document.

4.2 Making data openly accessible

As of September 2019, all the project datasets should still be kept consortium confidential. This obviously does not apply to the open data sets brought in to supplement the use case partner supplied data sets.

This stems from the fact that the datasets received up until now, as well as future datasets expected to be supplied in the future, are all business sensitive data. They are made available to project as consortium-confidential and should not be made accessible to the general public. This was not anticipated at the time of proposal writing.

As in first version of this deliverable, the project will still request from the Commission to opt out of the Pilot on Open Research Data in H2020.

The datasets will be accessible from either the LeanXcale database of from Object Stores in the Ubitech testbed. Note that, as was done in 2019 for the Danaos dataset, certain datasets may also be stored in IBM Object Cloud, since it offers enough security (even for the EROSKI data set). Same may occur in second half of the project for other datasets, possibly also for other Cloud storages.

Access to the datasets is made possible in one of the three possible manners:

1. through the regular LeanXcale database
2. through Object Store interfaces
3. through the Seamless interface which itself accesses the data either in LeanXcale or in Object Store. The Object Store may either be local (in Ubitech) or remote.

Furthermore, as mentioned earlier, artifacts of research value may be generated. It is not yet clear which exact datasets (e.g. monitoring information of the infrastructure, deployment patterns generation results, etc) will be made available and have research potential (as of September 2019). If that will happen, these artifacts will be uploaded to the BigDataStack OpenAIRE page. The consortium plans to re-examine the aforementioned artifacts in the second release of the components and make them public along with the associated metadata and the licensing information.

4.3 Making data interoperable

Prior to packaging produced artifacts we will research if there are established formats and will stick to them.

CC BY NC ND

4.4 Increase data re-use (through clarifying licenses)

Similarly, we will grant a clear licensing for the artifact that will not hinder FAIR access and usage of them.

CC BY NC ND

5 Allocation of resources

As of September 2019, the sole envisioned costs for making data FAIR in BigDataStack are possible fees to be paid to publishers to make publications open.

In this case, these fees, which are eligible expenses, will be reported as regular project expenses.

Ubitech partner is responsible for the data management of BigDataStack, providing an ISO certified testbed to host the datasets (as required by the use case partners of the project).

As detailed at end of the Data Security section, the sole case of long term preservation will be solved thanks to the BigDataStack OpenAIRE page will not generate costs. Since OpenAIRE is an EU site, the length of the preservation will automatically be in line with EU directives.

CC BY NC ND

6 Data security

Best practices for securely storing data will be implemented by Ubitech. In particular, Ubitech is ISO 27001 certified, which is required by EROSKI for storing its dataset (even after PI removal).

The preservation and curation of the use case datasets beyond the length of the project is not required. In the case artifacts of research value will be produced by the project, they will be uploaded to the BigDataStack OpenAIRE page as explained in the Data Summary section of this document.

CC BY NC ND

7 Ethical aspects

As detailed in version 2.0 of deliverable D6.1, all the datasets that will reach the BigDataStack testbed, be it at Ubitech or elsewhere, will have been anonymized in a manner which not only removes personal identity in case it appeared in the original data but also other fields (e.g., address) which could have permitted to reconstruct the identity of person. Therefore, for all practical purpose, the data can be considered as not containing any personal information. However, the datasets of the three use cases will be business confidential, and hence, not made available outside of the consortium.

CC BY NC ND

8 Other issues

BigDataStack has no usage of other national/funder/sectorial/departmental procedures for data management.

CC BY NC ND

9 Conclusions

The datasets that are and will be used in the project do not and will not contain any PII (Personally Identifiable Information), however they will be Business Confidential, and hence not circulated outside of the consortium. The data will be securely stored in a store of the project testbed (ISO certified). Artifacts of research value that will be generated in the second half of the project will be uploaded to the BigDataStack OpenAIRE page.

CC BY NC ND

References

[1] Template for the Data Management Plan

[2] National Climatic Data Center

CC BY NC ND