BIG DATA EUROPE

Coordination and Support Action

Big Data Europe – Empowering Communities with Data Technologies

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

Deliverable 6.2: Pilot Deployment I

Dissemination Level	Public
Due Date of Deliverable	M18, 30/06/2016
Actual Submission Date	M20, 17/08/2016
Work Package	WP6, Real-Life Deployment & User Evaluation
Task	T 6.2
Type	Other
Approval Status	Approved
Version	1.0
Number of Pages	12
Filename	D6.2_Pilot_Deployment_I.pdf

Abstract: This deliverable presents the implementation and deployment efforts undertaken to realize the various pilots for the seven Societal Challenges identified in H2020. Main core of the document form the source code, scripts, recorded demonstrations and manuals for deploying the pilots. Thus, the deliverable serves as an index for the various online sources available for the first cycle of pilots.

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

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

History

Version	Date	Reason	Revised by
0.0	04/07/2015	Initial version	Ronald Siebes
0.1	18/07/2015	Outline	Ronald Siebes
0.2	12/08/2016	Final Draft	Ronald Siebes
0.5	15/08/2016	Reviewed version	Aad Versteden
0.6	16/08/2016	Final draft corrected according to review comments	Ronald Siebes
1	17/08/2016	Final version	Ronald Siebes

Author List

Organisation	Name	Contact Information
OpenPhacts	Bryn Williams-Jones	bryn@openphactsfoundation.org
VUA	Victor de Boer	v.de.boer@vu.nl
VUA	Ronald Siebes	rm.siebes@few.vu.nl
NCSR-D	S. Konstantopoulos, A. Charalambidis, I. Mouchakis, G. Stavrinos	konstant@iit.demokritos.gr acharal@iit.demokritos.gr gmouchakis@iit.demokritos.gr
TenForce	Aad Versteden, Erika Pauwels	aad.versteden@tenforce.com erika.pauwels@tenforce.com

Executive Summary

This document serves as an index for references to the implementation and deployment efforts that realized the first cycle of pilots for the seven Societal Challenges. The source code, scripts, recorded demonstrations and manuals for deploying the pilots are the core of this deliverable. The good collaboration between the technical and domain partners combined with deliverance of a working infrastructure, generic Big-Data components, user interfaces and documentation according to internal deadlines resulted in a versatile but coherent set of demonstrators. Various webinars, hangouts and presentations during the last year are recorded and serve as excellent reference material in this document together with the public code repositories where the more technically skilled reader can get more insight into the details.

The pilots demonstrate how relevant large-scale datasets or data-streams for the respective seven SC communities can be processed by the BigDataEurope infrastructure and provide novel insights that are promised by the Big Data community.

This document refers to online demonstrations, presentations, instructions and code-bases that form the content of D6.2.

Abbreviations and Acronyms

BDE	Big Data Europe
LOD	Linked Open Data
SC	Societal Challenge

Table of Contents

List of Tables

List of Figures

1. Introduction

Implementation and instantiation are the two main aspects of the pilot deployment. These consist of having a code-base, a CPU/GPU cluster to test and deploy the code, creating modules by the various partners either from scratch or, in most cases modifying existing code into docker components. Close collaboration with the technical partners (TenForce, SWC, NCSR-D and FhG) that are responsible for creating the BDE infrastructure and the generic components is essential. It is the role of WP6 to facilitate the communication between the technical partners and the domain partners (those who have the specific knowledge for each Societal Challenge) and guide the pilot implementation process. This report serves as an index referencing the current implementation and deployment work for the first cycle of pilots for the seven Societal Challenges which are available online and which are part of this deliverable.

The best way to get a concise technical overview of the development and deployment status of all the pilots is to watch this online webinar: https://www.youtube.com/watch?v=Jy-6ImpL1rE

The best way to get a written overview of each pilot and the relation to the generic BDE components and the infrastructure is reading deliverable D5.2.

The best way to dive into the source code and deployment scripts is to browse through the sub-repositories starting with 'pilot-SCxxx' on the main BDE GitHub repository: https://github.com/big-data-europe

2. Planning

Figure 1 shows the initial planning for the implementation and deployment of the pilot cycles as described in D6.1.

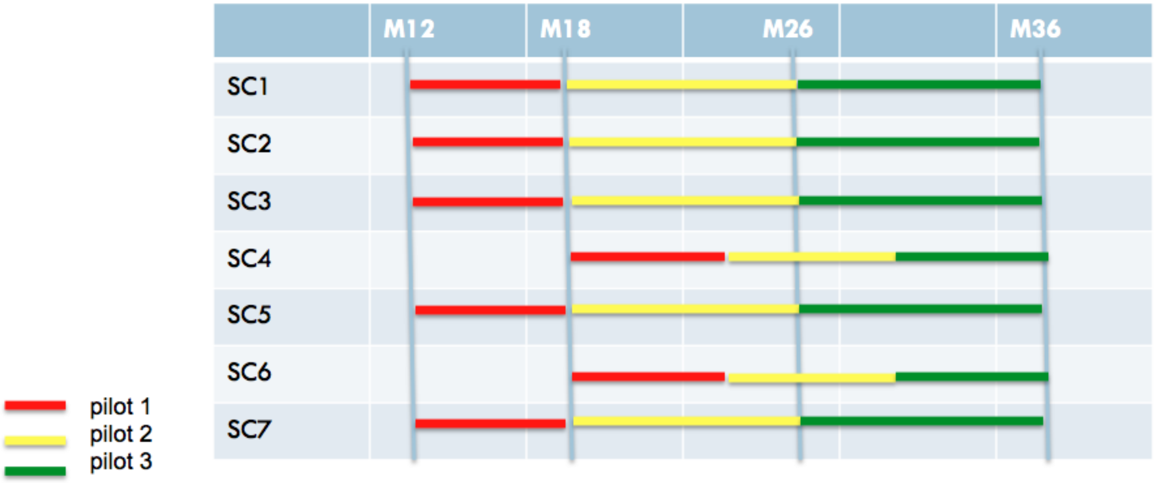

Figure 1: Pilot Planning

Currently, the development and deployment is progressing according to schedule and we are entering the evaluation phase (deliverable D6.3).

3. SC1 – Health Pilot

The pilot is carried out by OpenPhacts and VUA in the frame of SC1 *Health, Demographic Change and Wellbeing.* NCSR-D has joined to work on the distributed RDF storage. The pilot is described in section 2 from deliverable D5.2. It implements the workflow of reproducing the functionality of an existing data integration and processing system (the OpenPhacts Discovery Platform) on BDI.

The most important challenge for this pilot is to replace the commercial cluster RDF[1] store, Virtuoso, with an open source variant. There is currently only one promising candidate, 4Store[2], which has received virtually no updates in the past year. Therefore the VUA, OpenPhacts and NCSR-D have decided to adopt the code and improve it in such a way that it can serve as a generic component on the BDE infrastructure. This component will not only serve SC1 but also SC2 (Agriculture) and SC6 (Social Sciences) because the *veracity* and *variety* aspects also apply to them. Linked-Data is the likely candidate to deal with this for which a distributed RDF datastore is required when dealing with Big Data. The ongoing implementation work on the 4Store BDE docker component can be followed here: https://github.com/big-data-europe/docker-4store

The first test results are being demonstrated at the FLINK workshop part of the Int. Semantic Web Conference this October in Kobe, Japan.[3]

Presentations	○ https://www.big-data-europe.eu/the-open-phacts-pilot-second-hangout-for-the-health-societal-challenge/ ○ https://www.big-data-europe.eu/event/webinar-pilot/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc1-cycle1

Table 1: SC1 - Health Pilot - Cycle 1 references

4. SC2 – Agriculture Pilot

The pilot is carried out by AgroKnow, FAO, and SWC in the frame of SC2 *Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy*. It is described in section 3 of deliverable D5.2.

One of the key challenges for this pilot is the automatic data extraction from heterogeneous scientific publications and tables. SWC took the lead here and worked hard on PoolParty[4] code to extract data from viticulture publications, to develop queries based on real-life research questions and to deploy efficiently on one of the BDE RDF storage components which are accessed via one of the graphical web interfaces which are still under development.

[1] http://virtuoso.openlinksw.com/features-comparison-matrix/
[2] https://github.com/garlik/4store
[3] http://project-hobbit.eu/events/blink-2016/
[4] https://www.poolparty.biz/

This data preparation step took more time than expected and resulted in the decision to reduce the number of features that the demonstrator will have in the first pilot cycle.

Presentations	○ https://www.big-data-europe.eu/the-open-phacts-pilot-second-hangout-for-the-health-societal-challenge/ ○ https://www.big-data-europe.eu/event/webinar-pilot/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc2-cycle1

Table 2: SC2 - Agriculture Pilot - Cycle 1 references

5. SC3 – Energy Pilot

The pilot is carried out by CRES in the frame of SC3 *Secure, Clean and Efficient Energy.* It is described in section 4 of deliverable D5.2.

One of the biggest challenges was to break up the daily 0.6TB stream of sensor data from the wind-turbine farms into coherent chunks of the distributed file storage. This coherence is needed because the algorithms that do the analysis need to have data in an exact format and size. Also the third-party data is transformed into that same format which is used for correlation comparisons.

Presentations	○ https://www.big-data-europe.eu/bde-pilot-case-in-energy-system-monitoring-in-wind-energy-production-unit/ ○ https://www.big-data-europe.eu/big-data-europe-on-line-hangout-in-energy-domain/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc3-cycle1

Table 3: SC3 - Energy Pilot - Cycle 1 references

6. SC4 – Transport Pilot

The pilot is carried out by FhG and CERTH in the frame of SC4 *Smart, Green and Integrated Transport*. It is described in section 5 of deliverable D5.2.

Thessaloniki is currently collecting data via a static sensor network, dynamic sensors. An important implementation part by CERTH in this pilot is writing 'R' scripts[5] that consume JSON data from BLUEtooth data and other data on positions of taxi's in order to map the position of the vehicles on a map. The data comes from a JSON service, is consumed by Kafka, pre-processed via FLINK, and send to 'R-server' which performs the calculations (e.g. how busy it is on a certain street). For the pilot MongoDB with Elasticsearch is chosen as the datastore

[5] https://www.r-project.org/

allowing efficient visualisation via geo-spatial queries combined with static OpenStreetMap data.

Presentations	○ https://www.big-data-europe.eu/behind-the-scenes-of-the-bigdataeurope-transport-pilot-recap-of-our-hangout/ ○ https://www.big-data-europe.eu/more-big-data-less-traffic-congestion/ ○ https://www.big-data-europe.eu/big-data-for-transport-webinar-wrap-up-the-tech-the-business-and-the-policy/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc4-flink-kafka-consumer ○ https://github.com/big-data-europe/pilot-sc4-kafka-producer ○ https://github.com/big-data-europe/pilot-sc4-pipeline ○ https://github.com/big-data-europe/pilot-sc4-mapmatcher ○ https://github.com/big-data-europe/pilot-sc4-docker-r

Table 4: SC4 - Transport Pilot - Cycle 1 references

7. SC5 – Climate Pilot

The pilot is carried out by NCSR-D in the frame of SC5 *Climate Action, Environment, Resource Efficiency and Raw Materials.* It is described in section 6 of deliverable D5.2.

Dynamical downscaling of climatic and/or meteorological data is the process where output from a “large-scale” model (such as a GCM) is used to drive a regional or local model in higher spatial and temporal resolution, which is able to simulate local conditions in greater detail. It is an essential first step for any further analysis, assessment or processing in climate and related domains, such as environmental impact assessment, air pollution etc. The downscaling procedure can consist of several successive steps with nested (geographical) computational domains. The current features of the 1st Climate pilot were shown to be the following:

- Data Ingestion from NetCDF file
- Data export to NetCDF files
- It can start and monitor WRF-based downscaling on institutional (i.e., non-BDE platform) resources, depending on whether the requested results already exist in BDE database
- It maintains provenance record of results on the BDE platform, for monitoring and further analysis
- It supports basic analytics on the BDE platform (Hive queries)
- It has a console-based user interface, targeted on climate/atmospheric scientists(Python/Jupyter interface for demonstration)

Presentations	○ https://www.big-data-europe.eu/description-and-evaluation-of-1st-climate-pilot-use-case-2nd-online-hangout-wrap-up/ ○ https://www.big-data-europe.eu/1st-pilot-to-be-developed-in-the-frame-of-bigdataeurope-under-societal-challenge-5-climate-action-environment-resource-efficiency-and-raw-materials/ ○ https://www.big-data-europe.eu/sc5-bde-presentation-egu-general-assembly-vienna-17-22-april-2016/ ○ https://www.big-data-europe.eu/big-data-in-the-climate-domain-online-hangout-wrap-up/
Instructions, deployment scripts and source code	○ https://github.com/iaklampanos/bde-climate-1

Table 5: SC5 - Climate Pilot - Cycle 1 references

8. SC6 – Social Sciences Pilot

The pilot is carried out by NCSR-D, and SWC in the frame of SC6 *Europe in a Changing World - Inclusive, Innovative and Reflective Societies.* It is described in section 7 of deliverable D5.2.

Like SC2 the biggest challenge for this pilot is to extract the data for the various heterogeneous sources, transform it to RDF and write efficient queries that implement the required returned aggregations that are presented via a graphical web interface defined by the use cases. Here too, SWC is the responsible technical partner and uses for example their PoolParty toolkit to achieve an important part of this goal.

Presentations	○ https://www.big-data-europe.eu/recording-of-the-sc6-bde-hangout-webinar/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc6-cycle1

Table 6: SC6 – Social Sciences Pilot - Cycle 1 references

9. SC7 – Security Pilot

The pilot is carried out by SatCen, UoA, and NCSR-D in the frame of SC7 Secure Societies – Protecting Freedom and Security of Europe and its Citizens. It is described in section 8 of deliverable D5.2.

The biggest implementation challenge is todistribute the image analysis load over multiple nodes in a CPU cluster and in particular dealing with splitting images in tiles where the results will need to be integrated again at the end of the processing pipelines/workflows.

These flows are as follows:

- The *Change Detection* workflow ingests satellite images to detect areas with changes on land cover or land use by using change detection techniques; the identified Area of Interest (AoI) is then associated with social media and news agencies items and presented to the user for cross-validation;
- The reverse procedure is applied to the *Event Detection* workflow. Event detection is triggered by social media and news agencies information, where trending topics with geospatial connotation constitute a time- and space- localized event. Provided such an event, the corresponding satellite images are acquired and processed in order to check for changes in land cover or land use.

These workflows are successfully deployed on the NCSR-D cluster reserved for this project.

Presentations	○ https://www.big-data-europe.eu/second-hang-out-big-data-in-secure-societies-outcome/ ○ https://www.big-data-europe.eu/technical-update-on-societal-challenge-7-pilot/ ○ https://www.big-data-europe.eu/introducing-the-bde-societal-challenges-secure-societies/ ○ https://www.big-data-europe.eu/online-hang-out-big-data-in-secure-societies-outcome/
Instructions, deployment scripts and source code	○ https://github.com/big-data-europe/pilot-sc7-image-aggregator ○ https://github.com/big-data-europe/pilot-sc7-geotriples ○ https://github.com/big-data-europe/pilot-sc7-change-detector ○ https://github.com/big-data-europe/pilot-sc7-lookup-service

Table 7: SC7 – Security Pilot - Cycle 1 references

10. Conclusion

This report provides references to the online material that describes the implementation and deployment efforts making up this deliverable. It also provides a very brief overview of the current technical approaches used in each Societal Challenge. Managing to have an interesting pilot for each Societal Challenge being deployed on the BDE infrastructure within the provided internal deadlines was an ambitious goal which due to the good synergy between the partners was achieved satisfactory.