



**Big Data to Enable Global Disruption of the Grapevine-powered Industries**

## **D1.3 - Annual Public Report**

<b>DELIVERABLE NUMBER</b>	D1.3
<b>DELIVERABLE TITLE</b>	Annual Public Report
<b>RESPONSIBLE AUTHOR</b>	Dimitris Fotakidis (Agroknow)



Co-funded by the Horizon 2020  
Framework Programme of the European Union

<b>GRANT AGREEMENT N.</b>	780751
<b>PROJECT ACRONYM</b>	BigDataGrapes
<b>PROJECT FULL NAME</b>	Big Data to Enable Global Disruption of the Grapevine-powered industries
<b>STARTING DATE (DUR.)</b>	01/01/2018 (36 months)
<b>ENDING DATE</b>	31/12/2020
<b>PROJECT WEBSITE</b>	<a href="http://www.bigdatagrappes.eu/">http://www.bigdatagrappes.eu/</a>
<b>COORDINATOR</b>	Nikos Manouselis
<b>ADDRESS</b>	110 Pentelis Str., Marousi, GR15126, Greece
<b>REPLY TO</b>	<a href="mailto:nikosm@agroknow.com">nikosm@agroknow.com</a>
<b>PHONE</b>	+30 210 6897 905
<b>EU PROJECT OFFICER</b>	Ms. Annamária Nagy
<b>WORKPACKAGE N.   TITLE</b>	WP1   Project Management
<b>WORKPACKAGE LEADER</b>	Agroknow
<b>DELIVERABLE N.   TITLE</b>	D1.3   Annual Public Report
<b>RESPONSIBLE AUTHOR</b>	Dimitris Fotakidis (Agroknow)
<b>REPLY TO</b>	<a href="mailto:Dimitris.fotakidis@agroknow.com">Dimitris.fotakidis@agroknow.com</a>
<b>DOCUMENT URL</b>	<a href="http://www.bigdatagrappes.eu/">http://www.bigdatagrappes.eu/</a>
<b>DATE OF DELIVERY (CONTRACTUAL)</b>	31 December 2018 (M12), 31 December 2019 (M24, Updated Version), 31 December 2020 (M36, Final Version)
<b>DATE OF DELIVERY (SUBMITTED)</b>	20 December 2018 (M12), 1 January (M25, Updated Version), 31 December 2020 (M36, Final Version)
<b>VERSION   STATUS</b>	3.0   Final
<b>NATURE</b>	Report (R)
<b>DISSEMINATION LEVEL</b>	Public (PU)
<b>AUTHORS (PARTNER)</b>	Eliana Giannelou (Agroknow), Nikoletta Nikolopoulou (Agroknow), Dimitris Fotakidis (Agroknow)
<b>CONTRIBUTORS</b>	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (GEOCLEDIAN), Nikola Rusinov (ONTOTEXT), Simone Parisi (Abaco), Raffaele Perego (CNR), Salvatore Trani (CNR), Franco-Maria Nardini (CNR), Nyi-Nyi Htun (KULeuven)
<b>REVIEWER</b>	All partners

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Table of Contents	03/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow)
0.4	Section 1, 2, 3, 5	10/12/2018	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Coraline Damasio (INRA), Florian Schlenz (GEOCLEDIAN), Simone Parisi (Abaco), Eleni Foufa (APIGEA), Pythagoras Karampiperis (Agroknow)
0.6	Section 4	14/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow), Milena Yankova (ONTOTEXT), Nikola Rusinov (ONTOTEXT) Raffaele Perego (CNR), Nicola Tonello (CNR), Franco-Maria Nardini (CNR), Nyi-Nyi Htun (KULeuven)
1.0	Final version	20/12/2018	Pythagoras Karampiperis (Agroknow), Panagiotis Zervas (Agroknow)
1.8	Partners review, final comments and edits	21/12/2019	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Coraline Damasio (INRA), Florian Schlenz (GEOCLEDIAN), Simone Parisi (Abaco), Eleni Foufa (APIGEA), Panagis Katsivelis (Agroknow), Eva Bozou (Agroknow)
2.0	Updated version	1/1/2020	Eva Bozou (Agroknow)
2.1	Update TOC	15/11/2020	Eliana Giannelou (Agroknow), Nikoletta Nikolopoulou (Agroknow), Dimitris Fotakidis (Agroknow)
2.5	Update Section 2 and Section 3 (Pilots)	30/11/2020	Eliana Giannelou (Agroknow), Nikoletta Nikolopoulou (Agroknow), Dimitris Fotakidis (Agroknow)
2.5	Update Section 4 and Section 5	30/12/2020	Eliana Giannelou (Agroknow), Dimitris Fotakidis (Agroknow)
2.9	Contributions and Internal Review	31/12/2020	Aikaterini Kasimati (AUA), Maritina Stavrakaki (AUA), Florian Schlenz (GEOCLEDIAN), Nikola Rusinov (ONTOTEXT), Simone Parisi (Abaco), Raffaele Perego (CNR), Salvatore Trani (CNR), Franco-Maria Nardini (CNR), Nyi-Nyi Htun (KULeuven)
3.0	Final Version	31/12/2020	Eliana Giannelou (Agroknow), Dimitris Fotakidis (Agroknow)

PARTICIPANTS		CONTACT
<p>Agroknow IKE (Agroknow, Greece)</p>		<p>Nikos Manouselis Email: <a href="mailto:nikosm@agroknow.com">mailto:nikosm@agroknow.com</a></p>
<p>SRIMA AI (SAI, Bulgaria)</p>		<p>Todor Primov Email: <a href="mailto:todor.primov@ontotext.com">todor.primov@ontotext.com</a></p>
<p>Consiglio Nazionale DelleRicerche (CNR, Italy)</p>		<p>Raffaele Perego Email: <a href="mailto:raffaele.perego@isti.cnr.it">raffaele.perego@isti.cnr.it</a></p>
<p>Katholieke Universiteit Leuven (KULeuven, Belgium)</p>		<p>Katrien Verbert Email: <a href="mailto:katrien.verbert@cs.kuleuven.be">katrien.verbert@cs.kuleuven.be</a></p>
<p>Geocledian GmbH (GEOCLEDIAN Germany)</p>		<p>Stefan Scherer Email: <a href="mailto:stefan.scherer@geocledian.com">stefan.scherer@geocledian.com</a></p>
<p>Institut National de la Recherché Agronomique (INRA, France)</p>		<p>Pascal Neveu Email: <a href="mailto:pascal.neveu@inra.fr">pascal.neveu@inra.fr</a></p>
<p>Agricultural University of Athens (AUA, Greece)</p>		<p>Katerina Biniari Email: <a href="mailto:kbiniari@aua.gr">kbiniari@aua.gr</a></p>
<p>Abaco SpA (ABACO, Italy)</p>		<p>Simone Parisi Email: <a href="mailto:s.parisi@abacogroup.eu">s.parisi@abacogroup.eu</a></p>
<p>SYMBEEOSIS EY ZHN S.A. (Symbeeosis, Greece)</p>	 <p>Symbeeosis</p>	<p>Konstantinos Rodopoulos Email: <a href="mailto:rodopoulos-k@symbeeosis.com">rodopoulos-k@symbeeosis.com</a></p>

## ACRONYMS LIST

BDG	BigDataGrapes
APIs	Application Programming Interfaces
AUA	Agricultural University of Athens
BDG	BigDataGrapes
UEPR	Unit of Pech Rouge
LDBC	Linked Data Benchmark Council
DSS	decision support systems
PA	Precision Agriculture
SUS	System Usability Scale

## EXECUTIVE SUMMARY

This report presents the BigDataGrapes project vision and ambition and summarises the project's achievements. Its target audience comprises representatives of external interested communities as well as the general public.

The document summarizes the technical and implementation details of the BigDataGrapes infrastructure, describes the technical choices and the rationale behind them, and discusses the research and scientific particularities faced by each of the BigDataGrapes pilots.

More specifically, the document establishes the main objectives of the BigDataGrapes project and summarizes the core outcomes of the five pilot communities represented in the project. Based on these two axes, the technical advancements achieved during the reporting period are contextualised and discussed at a high level. Finally, the report concludes with the presentation of the **BigDataGrapes platform** and the **BigDataGrapes Data Market Place**.

The document is structured as follows. **Chapter 1** provides an introduction to the main issues tackled by the BigDataGrapes project, whereas **Chapter 2** provides an overview of the use cases with details on the methodology followed in order to define them and associate them with BigDataGrapes pilots. **Chapter 3** identifies the progress and results of each of the five selected pilots, while **chapter 4** describes the technical advancements of the project and the BigDataGrapes Platform. **Chapter 5** presents the BigDataGrapes Data Market Place providing the necessary proof in action that grapevine-powered data assets are shared and exchanged in interoperable formats and versions, by companies and organisations responsible for them. The **last Chapter**, discusses the conclusions of the deliverable. **Finally, in the annexes** we report the best practices for using the BigDataGrapes platform and the GaCoVi usage instruction. All the necessary components to set up an instance of the Big Data Platform are available in the projects Docker hub and can be deployed easily at any infrastructure. The use of dockers is straightforward so in this section, we are describing how one can set up and deploy an instance of the Big Data Platform for a specific use case: collecting and processing food safety incidents that are announced by National Authorities all around the world.

## TABLE OF CONTENTS

<b>1</b>	<b>Introduction.....</b>	<b>10</b>
<b>2</b>	<b>Pilots Overview .....</b>	<b>11</b>
<b>3</b>	<b>BigDataGrapes Pilots Advancement and Results .....</b>	<b>12</b>
3.1	Table and wine grapes pilot.....	12
3.1.1	Pilot description.....	12
3.1.2	Specific Goals .....	13
3.1.3	Site Description.....	13
3.1.4	Envisaged Outcomes .....	14
3.1.5	Advancement during the two first years .....	15
3.1.6	Data flow experiments end to end report .....	15
3.1.7	Quantitative Evaluation Against KPIs.....	15
3.1.8	Final results of the pilot.....	17
3.2	Wine making pilot.....	19
3.2.1	Pilot description.....	19
3.2.2	Specific goals .....	19
3.2.3	Site Description.....	19
3.2.4	Envisaged Outcomes .....	20
3.2.5	Advancement during the two first years .....	20
3.2.6	Data flow experiments end to end report .....	20
3.2.7	Quantitative Evaluation Against KPIs.....	21
3.2.8	Final results of the pilot.....	24
3.3	Farm management pilot.....	26
3.3.1	Pilot description.....	26
3.3.2	Specific goals .....	27
3.3.3	Site Description.....	29
3.3.4	Envisaged Outcomes .....	30
3.3.5	Advancement during the two first years .....	30
3.3.6	Data flow experiments end to end report .....	31
3.3.7	Quantitative Evaluation Against KPIs.....	31
3.3.8	Final results of the pilot.....	33
3.4	Natural cosmetics pilot .....	35
3.4.1	Pilot description.....	35
3.4.2	Specific goal .....	35
3.4.3	Site Description.....	35
3.4.4	Envisaged Outcomes .....	36
3.4.5	Advancement during the two first years .....	36
3.4.6	Data flow experiments end to end report .....	37
3.4.7	Quantitative Evaluation Against KPIs.....	37
3.4.8	Final results of the pilot.....	39
3.5	Food protection pilot .....	41
3.5.1	Pilot description.....	41
3.5.2	Specific goal .....	42
3.5.3	Site Description.....	42
3.5.4	Envisaged Outcomes .....	42
3.5.5	Advancement during the two first years .....	42
3.5.6	Data flow experiments end to end report .....	43



- 3.5.7 Quantitative Evaluation Against KPIs .....43
- 3.5.8 Final results of the pilot..... 46
- 4 BigDataGrapes technical solution..... 48**
  - 4.1 Overall approach ..... 48
  - 4.2 Advancement on Data & Semantics ..... 48
  - 4.3 Advancement on Data Analytics & Processing ..... 49
  - 4.4 Advancement on Visualisation & Decision Support .....50
  - 4.5 The BigDataGrapes Software Stack..... 51
    - 4.5.1 Persistence Layer..... 51
    - 4.5.2 Data Ingestion Layer..... 52
    - 4.5.3 Data Retrieval Layer ..... 52
    - 4.5.4 Processing Layer..... 52
    - 4.5.5 Management Layer..... 53
    - 4.5.6 Support Layer ..... 53
    - 4.5.7 Presentation Layer ..... 53
    - 4.5.8 Integration Components..... 53
  - 4.6 Docker platform.....54
    - 4.6.1 Docker image .....54
    - 4.6.2 Docker compose.....54
- 5 BIGDATAGRAPES DATA MARKETPLACE ..... 55**
  - 5.1 BigDataGrapes Data Marketplace Goal.....55
  - 5.2 Introduction to data marketplaces..... 55
  - 5.3 BigDataGrapes marketplace design .....58
  - 5.4 BigDataGrapes marketplace approach .....62
  - 5.5 Data Marketplace Services.....63
  - 5.6 Analysis of data marketplace business models ..... 65
- 5 CONCLUSIONS ..... 68**
- ANNEX A BigDataGrapes Platform Best Practices .....70**
  - A.1 Setting up an Instance of the Big Data Platform .....70
  - A.2 ID assignment in Big Data Projects..... 71
  - A.3 Orchestrating ETL Pipelines.....74
  - A.4 Automating Data Engineering Tasks in a Big Data Platform .....78
  - A.5 Container orchestrating and deployment best practices for the BigDataGrapes GraphDB instance .....79
- ANNEX B GaCoVi USAGE INSTRUCTIONS ..... 80**

**LIST OF TABLES**

- Table 1: Use Cases and Scenarios..... 11
- Table 2: Table and Wine Grapes Pilot Domain Specific KPIs Catalogue..... 16
- Table 3: Table and Wine Grapes Pilot Technological KPIs Catalogue ..... 16
- Table 4: Wine Making Pilot Domain Specific KPIs Catalogue ..... 21



Table 5: Wine Making Pilot Technological KPIs Catalogue..... 22

Table 6: Farm Management Pilot Domain Specific KPIs..... 32

Table 7: Farm Management Pilot Technological KPIs Catalogue..... 32

Table 8: List of the vineyards chosen for sample collection and their location..... 36

Table 9: Natural Cosmetics Pilot Domain Specific KPIs Catalogue ..... 37

Table 10: Natural Cosmetics Pilot Technological KPIs Catalogue ..... 38

Table 11: Food Protection Pilot Domain Specific KPIs Catalogue..... 43

Table 12: Food Protection Pilot Technological KPIs Catalogue – Lab Data ..... 44

Table 13: Food Protection Pilot Technological KPIs Catalogue –Food recalls and border rejections..... 44

Table 14: Food Protection Pilot Technological KPIs Catalogue – Price data ..... 45

Table 15: Indicative data model for the data marketplace ..... 66

Table 16: Example of a Business Model..... 67

Table 17: Example of a Business Model ..... 67

**LIST OF FIGURES**

Figure 1: BigDataGrapes device installations from AUA..... 12

Figure 2: Palivou Estate test site (Google Earth Pro)..... 13

Figure 3: Kontogiannis Estate test site (Google Earth Pro)..... 14

Figure 4: Fasoulis Estate test site (Google Earth Pro) ..... 14

Figure 5: SUS score of the correlation interface developed for the table and wine grape pilot ..... 18

Figure 6: Landfield of Pech Rouge (INRA, France) (Google Maps) ..... 19

Figure 7: INRA Pech Rouge Experimental Unit ..... 20

Figure 8: SUS score of the leaf counting interface developed for the wine making pilot ..... 25

Figure 9: SUS score of the correlation interface developed for the wine making pilot ..... 26

Figure 10: SUS score of the vine to wine exploration interface developed for the wine making pilot ..... 26

Figure 11: Drones and sensors operating in BigDataGrapes pilot sites in Tuscany ..... 27

Figure 12: A satellite NDVI image (from Geocledian APIs) for a parcel in SITI4Farmer..... 28

Figure 13: A prescription map created on the basis of NDVI and Soil maps applying a cluster analysis in order to identify different management zones..... 28

Figure 14: The selection mask for different kinds of best practices. .... 29

Figure 15: 12 HA of wineyards of Brunello of Montalcino ..... 29

Figure 16: 35 HA of Wineyards of CHIANTI D.O.C..... 30

Figure 17: SUS score of the irrigation interface developed for the farm management pilot..... 34

Figure 18: Dispersion of samples across the Greek territory..... 35

Figure 19: SUS score of the bio-efficacy correlation interface developed for the natural cosmetics pilot ..... 41

Figure 20: Food Protection dashboard of FOODAKAI system ..... 42

Figure 21: SUS score of the risk assessment interface developed for the food protection pilot..... 47

Figure 22: The BigDataGrapes top-level architecture ..... 48

Figure 23: GaCoVi interface ..... 50

Figure 24: BigDataGrapes software stack layers..... 51

Figure 25: Monetization in a data marketplace ..... 56

Figure 26: Types of data marketplaces..... 56

Figure 27: Main menu of the data marketplace ..... 59

Figure 28: How it works wireframe ..... 60

Figure 29: Data discovery wireframe..... 60

Figure 30: Dataset page wireframe ..... 61

Figure 31: Dataset details wireframe ..... 61

Figure 32: Share your data wireframe ..... 61



Figure 33: BigDataGrapes technical architecture .....62  
Figure 34: Operation model of data marketplace .....63  
Figure 35: Analytics for the data that is available through the data marketplace ..... 64  
Figure 36: A (canadian) example of the YML configuration expected by our crawlers .....70  
Figure 37: State of the data platform entities as taken from our internal Kibana dashboard instance ..... 71  
Figure 38: Heatmap for our daily cronjobs generated using Cron Heatmap .....74

## 1 INTRODUCTION

BigDataGrapes is a 36-month Research and Innovation action, supported by the European Commission through the H2020 Research and Innovation programme, under grant agreement no. 780751.

BigDataGrapes aspires to help European companies in the wine and natural cosmetics industries become more competitive in the international markets. Specifically, it tries to help companies across the grapevine-powered value chain ride the big data wave, supporting business decisions with real time and cross-stream analysis of very large, diverse and multimodal data sources.

In particular, BigDataGrapes aims to improve the competitive positioning of companies in the European IT sector that are serving companies and organizations with software applications:

- Software companies developing farm management and precision agriculture systems for companies in the agriculture sector.
- Software companies developing food risk assessment monitoring and prediction systems for companies in the food sector.
- Software companies developing quality control and compliance software for companies in the beauty and cosmetics sector.

To this end, the project develops, extends and provides the necessary specifications, mechanisms, fault-tolerant tools and components for allowing the rapid and intuitive development of variegating data analysis workflows, where the functionalities for data collection and storage, dataset creation, results visualization and deployment are provided by specialized services utilizing European large-scale, cloud-based infrastructures.

Thus, the vision of BigDataGrapes project is to manage technology challenges of the grapevine-powered data economy as its business problems and decisions requires processing, analysis and visualisation of data with rapidly increasing volume, velocity and variety: satellite and weather data, environmental and geological data, phenotypic and genetic plant data, food supply chain data, economic and financial data and more. It therefore makes a perfectly suitable cross-sector and cross-country combination of industries that are of high European significance and value.

The main objectives of BigDataGrapes are to build upon the rich historical, cultural and artisan heritage of Europe, aiming to support all European companies active in two key industries powered by grapevines: the grape and wine industry and the natural cosmetics one. It will help them respond to the significant opportunity that big data is creating in their relevant markets, by pursuing two ambitious goals:

- To develop and demonstrate powerful data processing technologies that will increase the efficiency of companies that need to take important business decisions dependent on access to vast and complex amounts of data.
- To catalyse the creation of a data ecosystem and economy that will increase the competitive advantage of companies that serve with IT solutions these sectors.

All the above drive us to the development of the **BigDataGrapes platform** and the **BigDataGrapes Data Market Place**.

## 2 PILOTS OVERVIEW

BigDataGrapes collects and monitors sensor data derived from all test sites owned or accessible by consortium members, bringing an expansive and diverse collection of datasets. These streams of data and datasets serve as the basis for carrying out research and technical work and are used as the testbed for enabling the implemented technical components to efficiently handle the volume and intricacies of these data, clearly acquired from realistic in- field conditions. The data analysis phase is part of the definition of the BigDataGrapes use cases and the BigDataGrapes pilots. Thus, five (5) overarching use cases have been identified and were then further divided in different scenarios.

Table 1: Use Cases and Scenarios

Use Cases (Generic)	Use Case Scenarios
A. Data Anomaly Detection & Classification	A. Earth Observation Data Anomaly Detection & Classification
B. Prediction	B1. Yield Prediction B2. Predicting Biological Efficacy B3. Crop Quality Prediction <ul style="list-style-type: none"> <li>● for Optimizing Post Harvest Treatments of Table Grapes (B3-1)</li> <li>● for Optimizing Winemaking (B3-2)</li> </ul>
C. Farm Management	C1. Optimization of Farm Practices in the Vineyard C2. Management Zones Delineation for Vineyards
D. Food protection	D1. Supply Chain Risk Prediction Dashboard D2. Price Prediction Dashboard D3. Price & Fraud Correlation Dashboard D4. Marketing Automation Dashboard

Moving from testing in laboratory conditions to testing in real-world settings, BigDataGrapes has designed and is executing human-centred assessment activities, the application pilots, pertaining to the defined use cases. The pilots defined, namely the Table and Wine Grapes pilot, the Wine Making pilot, the Farm Management pilot, the Natural Cosmetics pilot, and the Food Protection pilot constitute instantiations of these use cases. They are fully defined grapevine-powered industry use cases’ demonstrators, developed in order to allow the evaluation of the BigDataGrapes components within real-world settings, fulfilling industry-centred and specific end-user requirements.

### 3 BIGDATAGRAPES PILOTS ADVANCEMENT AND RESULTS

#### 3.1 TABLE AND WINE GRAPES PILOT

##### 3.1.1 Pilot description

Table and Wine Grapes Pilot aims to denote correlations between precision agriculture information and phenological data and grape and wine chemical analysis. Another goal is to associate the aforementioned data with earth observation data in order to examine the effectiveness of applying machine learning techniques and eventually train the relevant machine learning components.



Figure 1: BigDataGrapes device installations from AUA

The responsible partner of this pilot, Agricultural University of Athens (AUA), collects and monitors sensor, farming and phenological data derived from all test sites located in Greece. Soil properties, climate conditions and cultivation techniques constitute significant variables, which affect the quality of the final product. In particular, soil data affect both crop quality data and crop quantity data. Deriving meaningful knowledge from many relevant, yet heterogeneous data sources is important, acting as the basis for future decision-making processes.

### 3.1.2 Specific Goals

Some of the goals to be achieved through sensor and farming data collection is to denote correlations between precision agriculture information and phenological data and grape and wine chemical analysis. Finally, the ultimate goal is to correlate the aforementioned data with earth observation data in order to examine the effectiveness of applying machine learning techniques and eventually train the relevant machine learning components.

### 3.1.3 Site Description

Three test sites have been chosen for data collection for BigDataGrapes in Greece. These are situated in the regional unit of Corinthia, in the north-eastern part of Peloponnese. The following have been selected: for winemaking Palivou Estate and Kontogiannis Estate and for table grapes Fasoulis Estate.

Palivou Estate: is located in Nemea, planted with *Vitis vinifera* L. cv. ‘Agiorgitiko’ and ‘Merlot’ for winemaking. The row orientation is northeast-southwest, and the training/trellis system is VSP (vertical shoot positioned)-cane pruning, double Guyot.



Figure 2: Palivou Estate test site (Google Earth Pro)

Kontogiannis Estate: in Ancient Corinth having the same VSP -double Guyot or double Royat- training/trellis system planted with ‘Roditis’, ‘Savatiano’, ‘Mavroudi’ and ‘Soultanina’ for winemaking. Its row orientation is north to south.



**Figure 3: Kontogiannis Estate test site (Google Earth Pro)**

Fasoulis Estate: situated in Nemea, cultivated with 22 different table grape varieties, where each line has a different variety. The orientation is southeast to northwest.



**Figure 4: Fasoulis Estate test site (Google Earth Pro)**

### 3.1.4 Envisaged Outcomes

The expansive and diverse collection of datasets for BigDataGrapes will serve as the basis for carrying out research and technical work. These data assets will contribute to a data marketplace demonstrator that will serve as the project’s experimentation environment. The streams will be used as the testbed for enabling the

implemented technical components to efficiently handle the volume and intricacies of these data (correct sensor measurements, fill in missing values, corrupted or inconsistent data, adjust outliers, etc.), clearly acquired from realistic in-field conditions.

### 3.1.5 Advancement during the two first years

During the first year of the project's lifetime, AUA designed a detailed plan for the development and the execution of the Table and Wine Grapes pilot and defined the experimental protocols and processes to be employed in accordance to the piloting plan. AUA also defined the data and datasets to be collected in this piloting session and both the first and the second year engaged to the collection of these data from all three test sites chosen in Greece, namely Palivou Estate and Kontogiannis Estate for winemaking production and Fasoulis Estate for table grapes production. In particular, the collected data is comprised of the following: spatial data including topographical and elevation, geo-reference soil electrical conductivity data, weather data, data related to the quantity and quality of the grapes as well as canopy characteristics. For the realization of the data collection, special equipment was used. For example, the first year a HiPer V RTK GPS was used to record positioning data, such as field boundary points, and elevation data, and a Geonics EM38-MK2 Ground Conductivity Meter was used to measure the soil electrical conductivity along the fields. In order to record the canopy characteristics different pieces of equipment were used. In both years Crop Circle, Rapid Scan and SpectroSense2+GPS were used to retrieve classic spectral vegetative index data including NDVI, NDRE and LAI. Additionally, two drones with Multispectral and Thermal Sensors scanned the field on six occasions over the course of the second year, namely in the summer in 2019. Satellite imagery was also retrieved from Geocledian for the specific fields belonging to the Table and Wine Grapes pilot during the same dates. Moreover, two weather stations were installed at Palivou and Kontogiannis Estates respectively, in order to measure the wind speed and direction, air temperature, air humidity and atmospheric pressure. Last but not least, after harvesting the grapes at the end of each season and measuring the total grape yield, the collected samples were transferred to the Laboratory of Viticulture for further qualitative analysis (pH, Sugar Content, Titratable Acidity etc.).

### 3.1.6 Data flow experiments end to end report

In this section we conducted a thorough experimentation on the steps the data provided by the Table and Wine Grapes pilot follow inside the BDG stack. The initial step of the dataset upload shows excellent performance and scalability regardless of the increase in concurrency. Since the CPU and network usage show minor increases as we increase the concurrency, we consider this step as a highly performant one. The data pipeline step also shows very good performance as far as the real-life scenario is concerned. Following the conclusions, we also came up for the Farm Management pilot higher concurrency should be employed for this step to overcome high execution time as the volume of data increases. In terms of rdfizing the data, as we observed for the previous cases as well, the steps show good performance using the respective command line tool. The extraction of the semantic enriched data is the one showing the poorest performance for this specific pilot. As we have also described in the previous cases, we consider the increase in terms of concurrency to help greatly in improving the performance of this step so that no bottleneck is observed. Moreover, the extraction and storage of the satellite image processing dataset demonstrates great performance, keeping under consideration the low number of fields required to cover the pilot's needs. Finally, the analytic step shows good performance in terms of average latency of the correlation APIs, disregarding the quantity of data taken into account for the correlation.

### 3.1.7 Quantitative Evaluation Against KPIs

#### Domain Specific KPIs

AUA has created the list of domain specific KPIs for the Table and Wine Grapes Pilot and has defined their baseline values, which are presented in the following table.



Table 2: Table and Wine Grapes Pilot Domain Specific KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Total soluble solids</b>	The minimum sugar content of the must at harvest for the production of red dry wine	Brix	20	20	20
<b>Total titratable acidity</b>	The minimum total titratable concentration of the must at harvest for the production of red dry wine	g tartaric acid/L of must	3.5	3.5	3.5
<b>Total anthocyanin content</b>	Minimum total anthocyanin content for the production of red dry wine	mg malvidin/g of fresh skin	3.00	3.00	3.00
<b>Selective harvesting</b>	The purpose is to achieve different harvest dates depending on the grape quality characters per plot/cell instead of harvesting the entire vineyard on the same date	Number of harvesting dates per plots per vineyard	1	1	1

**Technological KPIs**

Additionally, in order to perform a complete quantitative evaluation for the Table and Wine Grapes Pilot, a Technological KPIs list along with baseline values have been defined by AUA.

Table 3: Table and Wine Grapes Pilot Technological KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Focusing Big Data</b>					
<b>Volume</b>	Variation in raw data volume – Proximal sensor data	MB	26	19.5	13.5
<b>Volume</b>	Variation in raw data volume – Weather data	MB	5	8.5	8.4
<b>Volume</b>	Variation in raw data volume – Earth observation data	GB	300	350.75	346.0**
<b>Volume</b>	Variation in raw data volume – Drone Imagery	GB	*	37.8	35

<b>Volume</b>	Variation in raw data volume – Yield and Quality	MB	2	2	2
<b>Variety in Data Source Types</b>	Number of different data source types	Data sources	15	17	17
<b>Variety in Data</b>	Number of different types of data (in different resolutions)	Datasets	10	12	12
<b>Velocity</b>	Speed of data generated – Proximal sensor data	MB/crop season	26	19.5	13.5
<b>Velocity</b>	Speed of data generated – Weather data	MB/year	5	8.5	8.4
<b>Velocity</b>	Speed of data generated – Earth observation data	GB/crop season	125	146.15	157.25
<b>Velocity</b>	Speed of data generated – Drone Imagery	GB/crop season	-	37.8	35
<b>Velocity</b>	Speed of data generated – Yield and Quality	MB/crop season	2	2	2

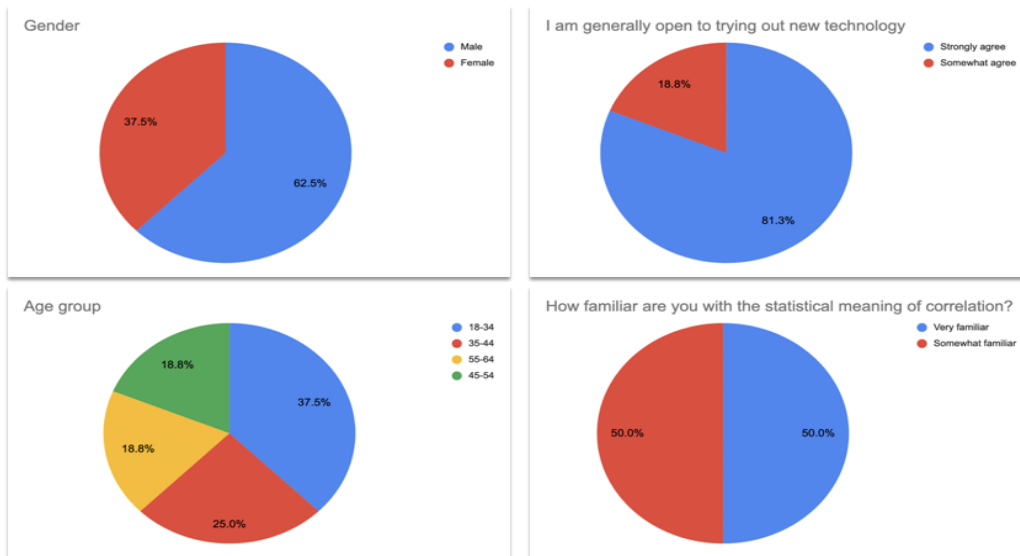
\*Drone imagery is expected to add up to another 200GB from 2019, when the data collection starts.

### 3.1.8 Final results of the pilot

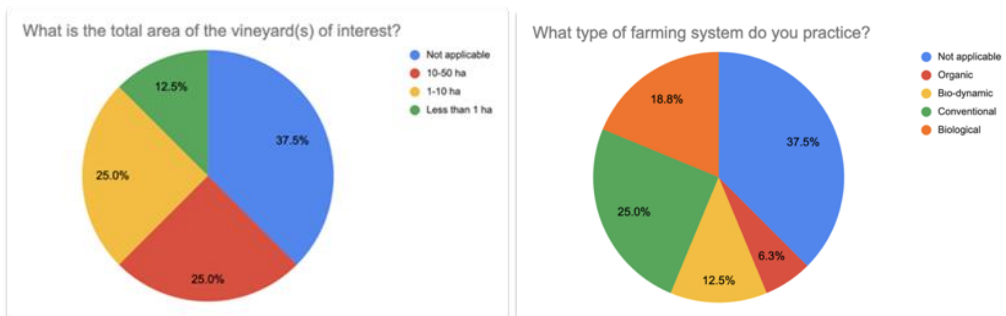
#### Table and Wine Grape Pilot Vineyard Information and Participant Demographics

##### Participant Demographics

Sixteen people participated in the Evaluation of the Table and Wine Grape Pilot, representing all aspects of the grapevine industry: vine growers, wine makers, agronomists, oenologists, as well as representatives from the research sector and the industry. The majority of the participants (62.5%) were male, of postgraduate education ageing from 18-44 years old. All participants proved to be open to try out new technology and quite familiar with the statistical meaning of correlation. Among the participants who owned a vineyard, they all had less than 50 ha of land, while all types of farming systems (conventional, organic, biodynamic) were among the answers provided.

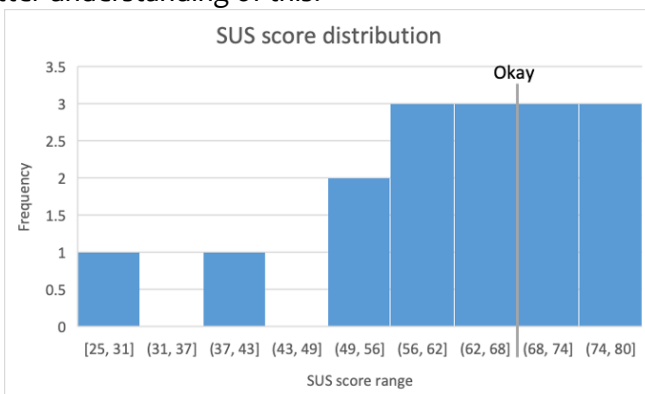


**Vineyard Information**



**Table and Wine Grape Pilot - Correlation Task**

The figure below shows results of the SUS questionnaire. A median of 66.3 puts the interface in a slightly lower percentile from the “Good” region of SUS scale, which is between 68 and 80.3. However, with the standard deviation of 14.7, we conclude that the SUS score does not significantly deviate from the acceptable region. Regardless, further analyses using the participants’ expertise and background as factors will help us gain a better understanding of this.



Median 66.3

SD 14.7

Reference	
SUS Score	Adjective Rating
> 80.3	Excellent
68 – 80.3	Good
68	Okay
51 – 68	Poor
< 51	Awful

**Figure 5: SUS score of the correlation interface developed for the table and wine grape pilot**

## 3.2 WINE MAKING PILOT

### 3.2.1 Pilot description

The Wine Making Pilot is dedicated to research in the fields of viticulture and oenology with an integrated point of view that allows a transversal approach from the vineyard to the packaged final product.

### 3.2.2 Specific goals

The main goals of this pilot are: (i) a better knowledge and better control of grape quality; (ii) quality potential existing in the grapes and wines and the on-line monitoring and control of the alcoholic fermentation; (iii) propose and study innovative technologies applicable to various steps of winemaking; (iv) valuation of coproducts, extraction of molecules and environmental impacts.

### 3.2.3 Site Description

The INRA Pech Rouge Experimental Unit is located N43°08'47", E03°07'19' WGS84, in the Languedoc-Roussillon region (Aude department) of France. The landfield of Pech Rouge includes a total area of 170 ha of land planted with 38 hectares of vines, distributed in three areas. The INRA Pech Rouge Experimental Unit also contains analytical laboratories, technological tools and finally a Sensory Analysis Laboratory which enables the tasting of different wines.



*Figure 6: Landfield of Pech Rouge (INRA, France) (Google Maps)*

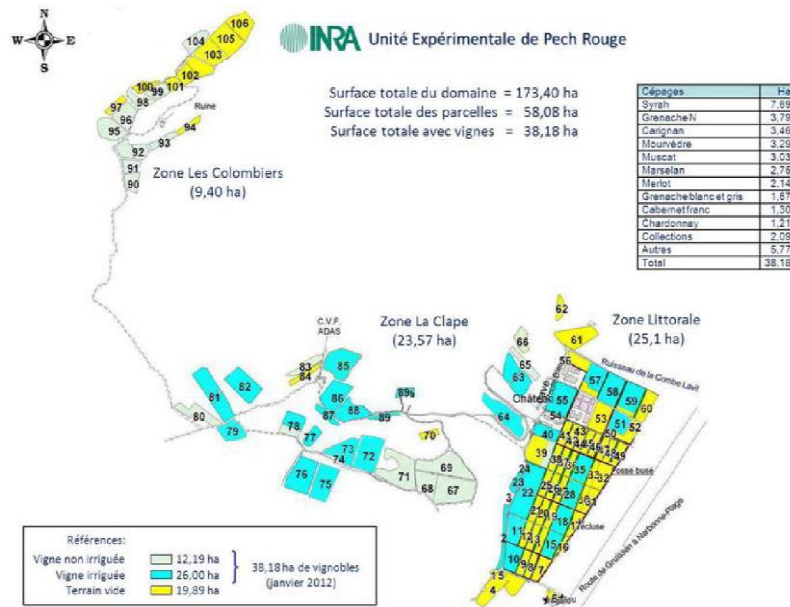


Figure 7: INRA Pech Rouge Experimental Unit

### 3.2.4 Envisaged Outcomes

Previous and on-going experimentation on Pech Rouge experimental Unit provided a large-scale datasets about winemaking and the vine-grape-wine continuum. Those data and datasets were benefit for:

- The application and test of the BigDataGrapes solution
- The validation of the BigDataGrapes components in real-life conditions and with complex dataset.

Our goal is also:

- To have a device to improve data quality (correction) and make FAIR data
- To have a better understanding of ‘How data from the field can affect the wine quality?’ and ‘How vine water status can affect the wine quality?’
- To discovery knowledge in order to design new viticulture / vinification systems.

### 3.2.5 Advancement during the two first years

INRA, the responsible partner of the Wine Making Pilot, was engaged to create a dataset with data from the vine to the wine. INRA’s experimental unit of Pech Rouge has conducted a lot of research experiments for private wine companies and INRA in the frame of its engagement, made a special effort to link all the datasets gathered by diverse teams in order its large-scale dataset to be enriched.

### 3.2.6 Data flow experiments end to end report

In this section we experimented on the datasets provided by the Wine Making pilot and the datasets it has provided. In terms of the data flows specific for this pilot we have identified that the initial step the dataset upload step, presents great performance in respect to the completion time as well as the CPU, network and memory usage. It is a step that can be easily made with a high degree of concurrency without seriously affecting the rest of the stack. Following the experimentation, we performed on the rdfization step, we consider this step to have a good performance regardless of the volume of the data, using the command line tool developed for this step. Interestingly for the step, that of the semantically enriched data extraction and storage into Elasticsearch, we observed a slightly different behaviour than the other pilots. This difference can be explained considering the differences in the data model for each pilot and the better performance of this step is directly

affected by the work done in D3.1. Finally, also the prediction step shows very good performance, taking into consideration the difficult nature of the task and the complexity of the prediction model.

### 3.2.7 Quantitative Evaluation Against KPIs

#### Domain Specific KPIs

INRAE has generated the list of domain specific KPIs for the Wine Making Pilot and has defined their baseline values, which are presented in the following table.

Table 4: Wine Making Pilot Domain Specific KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Product yield per plot</b>	Kg per plot of grapes harvested	Kg/plot	1425	1347	1198
<b>Product yield</b>	Kg per ha of grapes harvested	Kg/ha	4107	4721	4115
<b>Wine volume per kilogram harvested Red wines</b>	Red wines It corresponds to the volume of wine produced per kilogram of harvested grape	L/kg	0.65	0.67	0.66
<b>Wine volume per kilogram harvested White and rosé wines</b>	White and rosé wines. It corresponds to the volume of wine produced per kilogram of harvested grape	L/kg	0.29	0.39	0.26
<b>Residual sugar content in wine – red wines</b>	Sugar content in wine. We have to check that the value of residual sugar is below 2 g/L. Red wines Calculation: $100 * \text{nber of conformed wine} / \text{total wine number}$	%	100	100	100
<b>Residual sugar content in wine – white and rosé wines</b>	Sugar content in wine. We have to check that the value of residual sugar is below 2 g/L. White and rosé wines Calculation: $100 * \text{nber of conformed wine} / \text{total wine number}$	%	100	100	100
<b>Volatile acidity after alcoholic</b>	It must be $0,10 < x < 0,98$ . Red wines	%	100	100	100

<b>fermentation for red wines</b>	Calculation: $100 * \text{nber of conformed wine} / \text{total wine number}$				
<b>Volatile acidity after alcoholic fermentation for white and rosé wines</b>	It must be $0,10 < x < 0,88$ White and rosé wines Calculation: $100 * \text{nber of conformed wine} / \text{total wine number}$	%	100	100	100
<b>Color intensity (darkness) for red wines – visual analysis Before bottling</b>	The purpose is to have a dark color for red wines. The ratio calculated corresponds to the number of judges who found the wine dark / total number of judges who are able to detect the characteristic	$0 < \text{Ratio} < 1$	0.66	0.57	0.33
<b>Color intensity (clearness) for white and rosé wines – visual analysis Before bottling</b>	The purpose is to have a clear wine for white and rosé wines. The ratio corresponds to the number of judges who found the wine clear / total number of judges who are able to detect the characteristic	$0 < \text{Ratio} < 1$	0.52	0.60	0.56
<b>Fruity flavor Before bottling</b>	The fruity flavor is well desired for all wine types. The ratio corresponds to the number of judges who detected this aroma / total number of judges	Ratio	0.48	0.55	0.49

### Technological KPIs

Additionally, in order to perform a complete quantitative evaluation for the Wine Making Pilot, a Technological KPIs list along with baseline values have been defined by INRAE.

Table 5: Wine Making Pilot Technological KPIs Catalogue

Variable	Definition	Units	2018 Baseline First year of the project	2019	2020
<b>Focusing Big Data</b>					
<b>Volume</b>	Variation in raw data volume – Plot Management	MB	2.4	2.4	2.4

<b>Volume</b>	Variation in raw data volume – Climatic data	KB	265	265	265
<b>Volume</b>	Variation in raw data volume – Grape and berry mechanical and chemical properties	KB	139	144	61
<b>Volume</b>	Variation in raw data volume – Qualitative and quantitative characteristics of must and wine	KB	408	336	492
<b>Volume</b>	Variation in raw data volume – Winemaking activities	MB	8.0	7.6	5.6
<b>Volume</b>	Variation in raw data volume – Sensory Analysis	KB	544	388	1550
<b>Volume</b>	Variation in raw data volume – Satellite Data	GB	47	55	44.5**
<b>Velocity</b>	Speed of data generated during harvesting period	GB/ harvesting period, 4 months	14.9	18.3	16.2**
<b>Velocity</b>	Speed of data generated – Satellite data	GB/month	2.6 S2 1.3 L8	3.3 S2 1.3 L8	2.9 S2** 1.2 L8**
<b>Variety in Data Source Types</b>	Number of different data source types	Data sources	18	19	20
<b>Variety in Data</b>	Number of different types of data (in different resolutions)	Datasets	9	10	11
<b>Data transformation</b>	Number of rdf triplets, from raw data	Number	0	62157	207190
<b>Data linked</b>	% of data linked, data connection – dataset linked divided by the total number of datasets	%	11%	67	100
<b>Level of FAIR-ness</b>	Fair data assessment tool	RDA SHARK evaluation	12/18 Never	3/18 Never	1/18 Never



	especially for winemaking activities	(David et al., 2019)	6/18 If Mandatory 0/18 Sometimes 0/18 Always	7/18 If Mandatory 7/18 Sometimes 1/18 Always	0/18 If Mandatory 10/18 Sometimes 7/18 Always
Big Data Process Metrics					

	Steps number needed for data to be available for analysis and processing Winemaking activities	Number	7	5	3
<b>Data Normalization (Homogenization)</b>					

It is important to underline that these variables make sense if they are well described in ontologies using semantic web to be able to do machine learning on them.

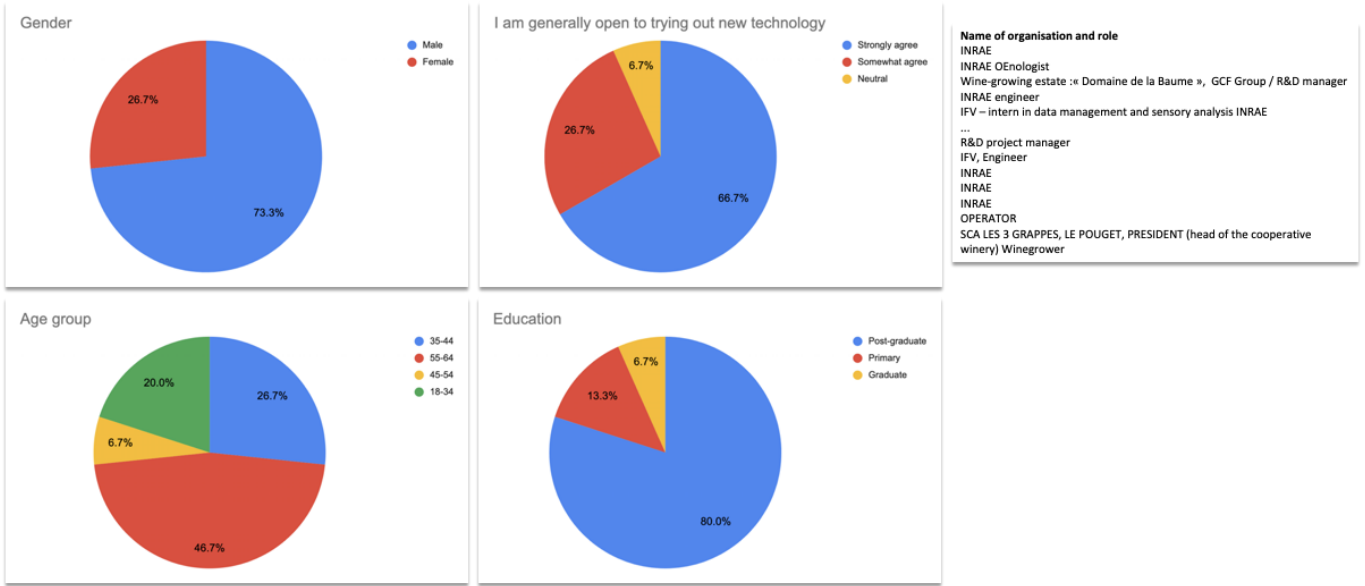
\*\* measured until November 30, 2020

### 3.2.8 Final results of the pilot

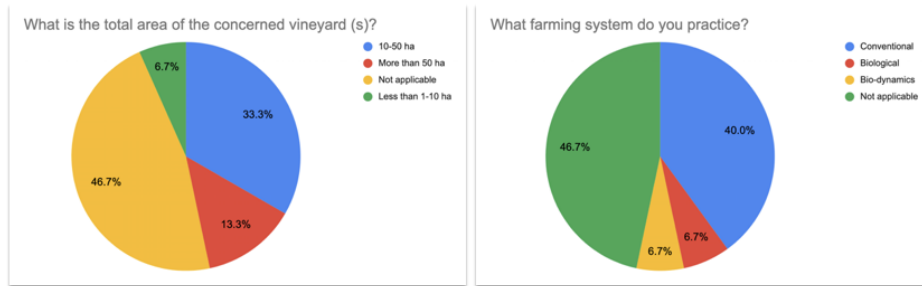
#### Wine Making Pilot Vineyard Information and Participant Demographics

##### Participant Demographics

In terms of participants, a large panel of jobs from the wine industry were represented: vine growers, a head of a cooperative winery, engineers from R&D departments or consulting, etc. Moreover, researchers working on viticulture and winemaking in diverse research units were also doing the evaluation. First of all, most end users were male, postgraduate and almost half of them were between 55 and 64 years old. They were open to try new technology. This result is not surprising because they accepted our invite to this trial which was focusing on IT tools for professionals. Secondly, participants who owned their vineyard had a surface between 10 and 50ha conducted in a conventional farming system. For almost 50%, the question related to vineyard information is not applicable as people working in companies, in consulting or in the field of research do not have their own vineyard.

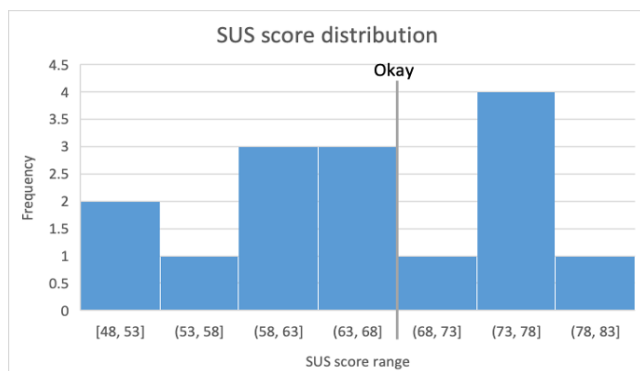


**Vineyard Information**



**Wine Making Pilot - Leaf Counting Task**

The median of 67.5 puts the interface in an acceptable region of the SUS scale. Looking at the SUS score distribution, we can observe that the scores between 73 and 78 were given by a relatively large portion (4/15) of the participants but the majority (9/15 participants) yielded the scores below the 68th percentile.



Median 67.5

SD 10.6

Reference	
SUS Score	Adjective Rating
> 80,3	Excellent
68 – 80,3	Good
68	Okay
51 – 68	Poor
< 51	Awful

**Figure 8: SUS score of the leaf counting interface developed for the wine making pilot**

**Wine Making Pilot - Correlation Task**

Figure 9 shows results of the SUS questionnaire. This interface was exactly the same as the one developed for the table and wine grape pilot. Thus, it had a similar SUS score with the median of 65 and standard deviation of

15.1 across 11 participants. We therefore conclude that although the score does not significantly deviate from the acceptable region, further analyses will help us understand it better.

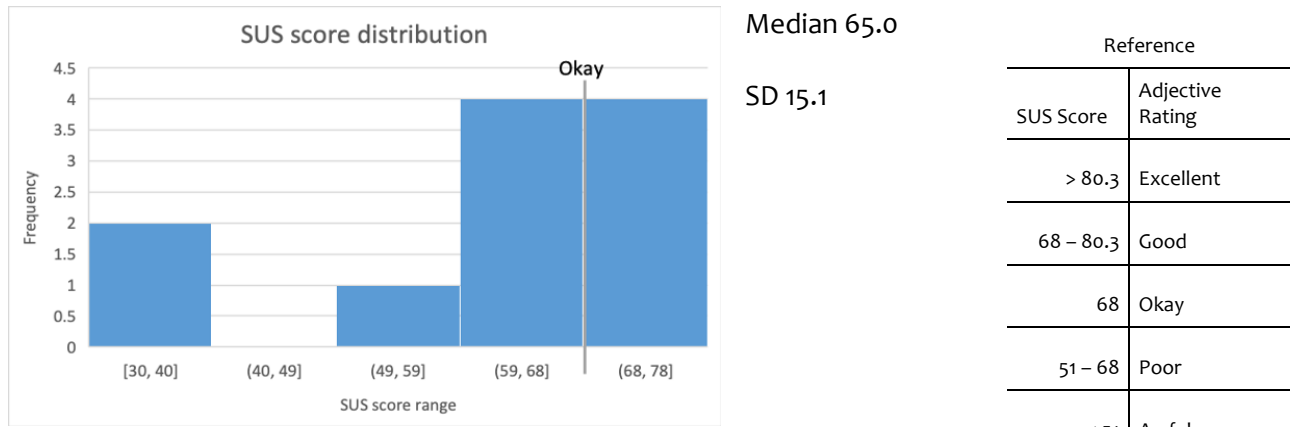


Figure 9: SUS score of the correlation interface developed for the wine making pilot

### Wine Making Pilot - Vine to Wine Exploration Task

Figure 10 shows results of the SUS questionnaire. The median of 67.5 puts the interface in an acceptable region of the SUS scale. The SUS score distribution tells us that 8 out of the 14 responses put the interface below the 68th percentile.

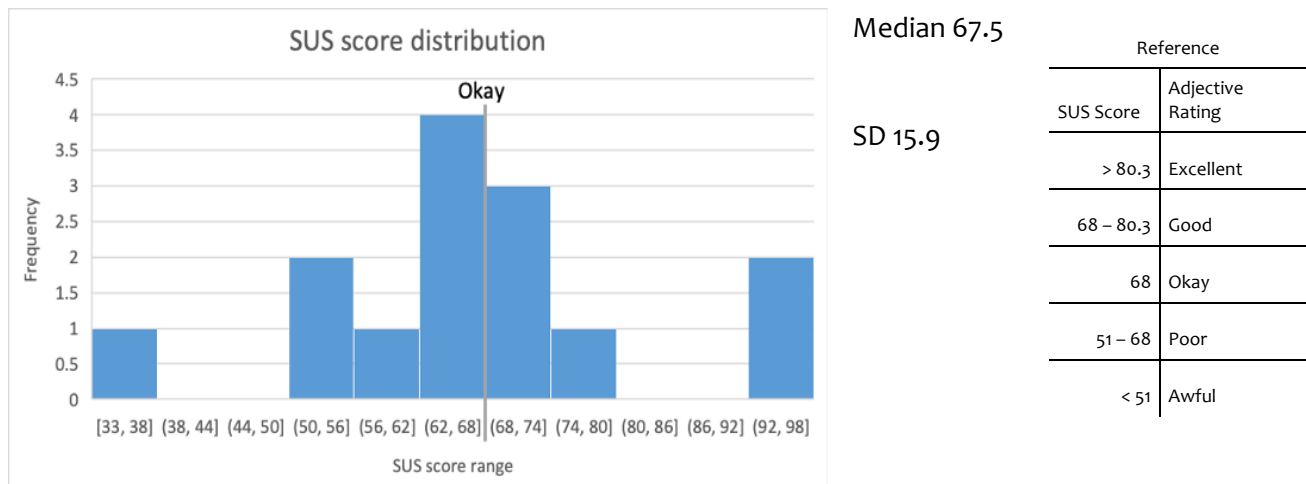


Figure 10: SUS score of the vine to wine exploration interface developed for the wine making pilot

## 3.3 FARM MANAGEMENT PILOT

### 3.3.1 Pilot description

The Farm Management Pilot aims to develop a unique system that satisfies the following needs:

- A Farm Management system with all the functionalities to support the farmer in his day by day activities and in gathering data from the field

- Hosting data from different sources with proper tools and functionalities for comparisons and easy data management
- Data exchange. A “day by day” data producer, to feed the generated data into the other BDG components, and make use of the incoming information from the other BDG components.
- Data visualization. The data related to the farmer should be displayed in a way that provides an added value and new insights to the farmer for his activities.

### 3.3.2 Specific goals

The specific goal is the development of a unique system that satisfies the needs of: a farm management with all the functionalities to support farmers, data hosting, data exchange and data visualization.

Two wine makers were identified as actors in this pilot. They will be involved in the pilot in two ways:

- Their work will be supported by making the developed products and systems available to them. In addition to the farm management system itself, this includes sensors and measurements that will provide data as basis for decision support.
- On the other hand, these actors can help in designing the new system by providing input and knowhow about their needs and activities. They can also give insights on how to disseminate results, approach and ideas of the BigDataGrapes Project.



**Figure 11: Drones and sensors operating in BigDataGrapes pilot sites in Tuscany**

In the following figures, the SITI4Farmer platform, the platform that is going to be used by the two Tuscany pilots winegrowers is demonstrated. It can be used to support various activities: e.g. to load best practice data, to manage variable rate fertilizer maps, to manage different information layers on soil, meteorology, satellite data and so on.

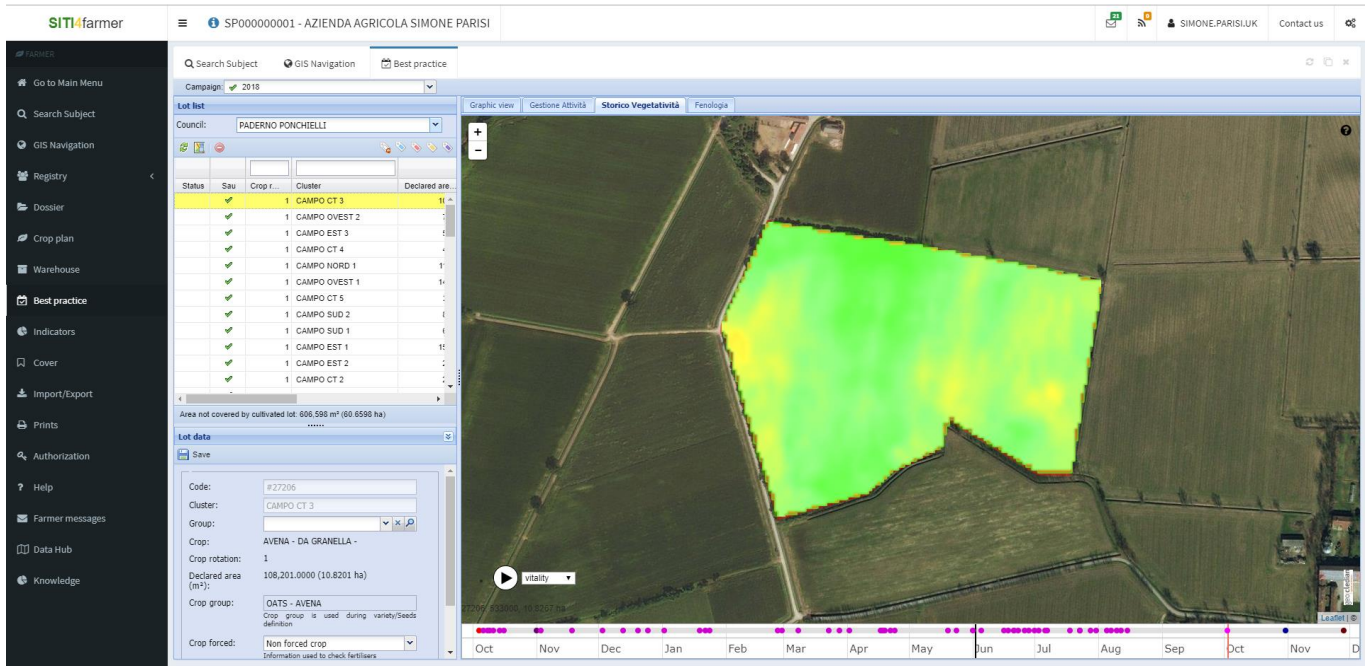


Figure 12: A satellite NDVI image (from Geocledian APIs) for a parcel in SITI4Farmer.

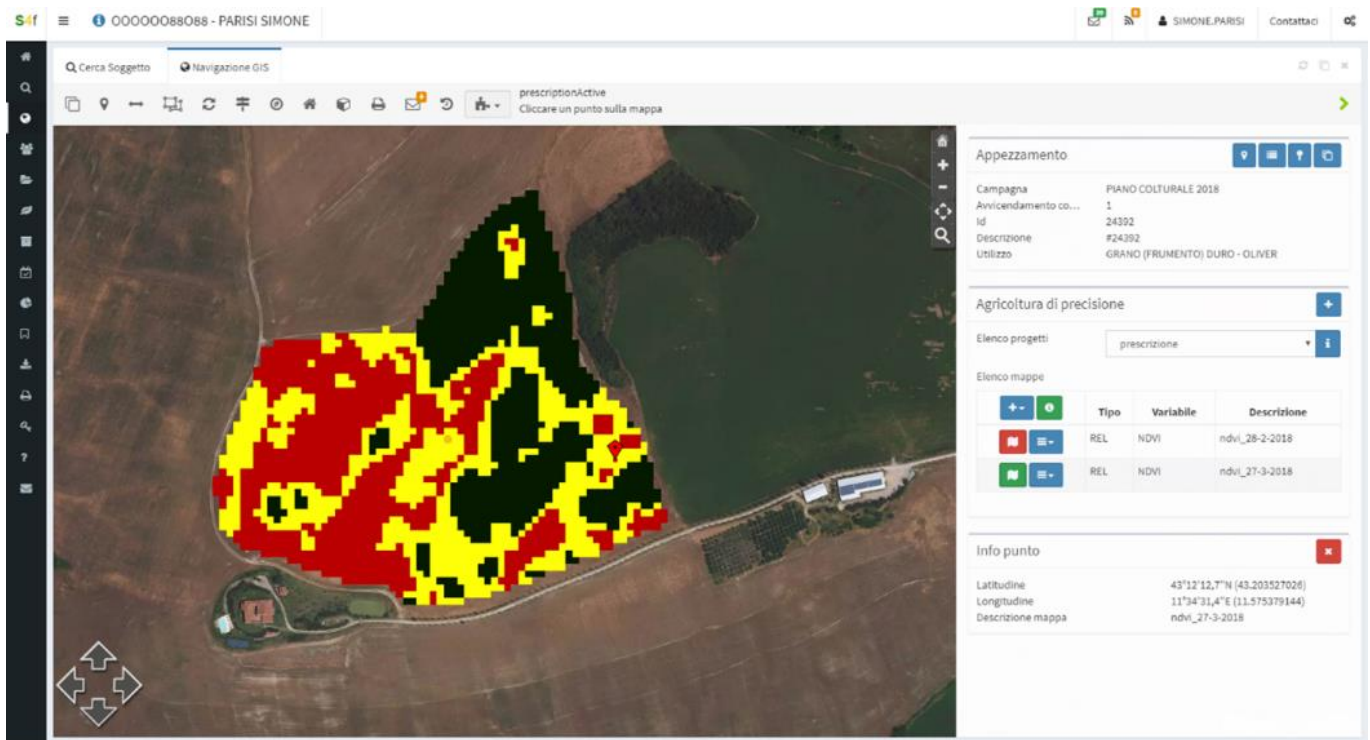


Figure 13: A prescription map created on the basis of NDVI and Soil maps applying a cluster analysis in order to identify different management zones.

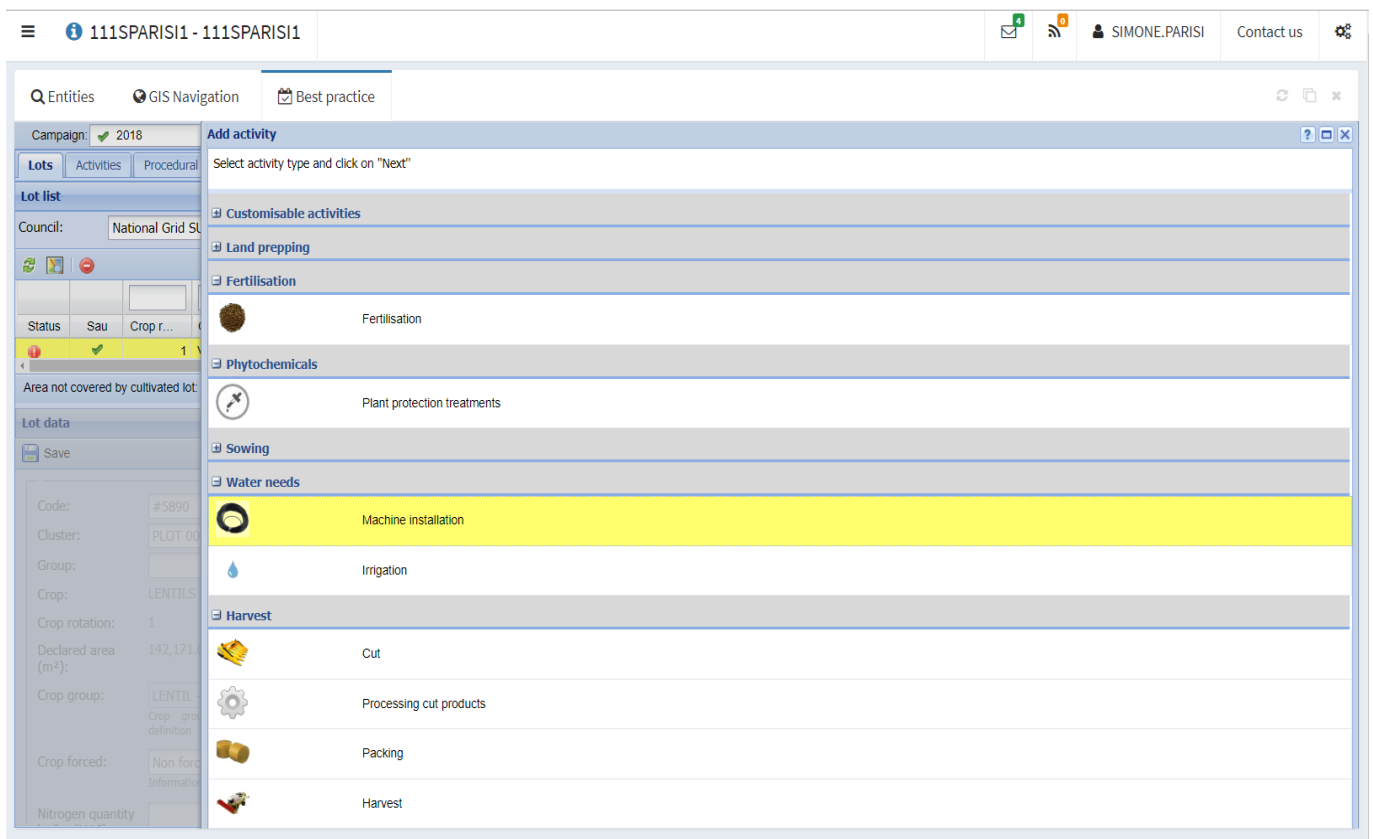


Figure 14: The selection mask for different kinds of best practices.

### 3.3.3 Site Description

The approach expects to involve 2 wineries, making them an active part of the project, collecting data from the field, in automatic and manual manners, and therefore contribute to the results.

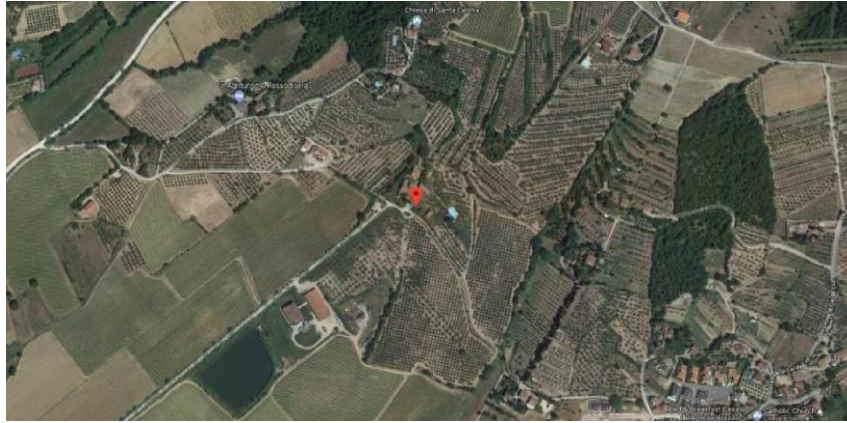
Company Name: CASATO PRIME DONNE CIRCA  
 Address: Località Casato – Montalcino, Tuscany, IT  
 GPS Coordinates : 43.088196° N 11.464319° E  
 Internet Site: [www.cinellicolombini.it](http://www.cinellicolombini.it)  
 12 HA of Vineyards of Brunello of Montalcino



Figure 15: 12 HA of vineyards of Brunello of Montalcino

Company Name: CANTINA IL PALAZZO

Address: Loc. Antria, Arezzo, Tuscany, IT  
 GPS Coordinates: 43.502773, 11.904402  
 Internet Site: [www.tenutailpalazzo.it](http://www.tenutailpalazzo.it)  
 35 HA of Vineyards of CHIANTI D.O.C.



**Figure 16: 35 HA of Vineyards of CHIANTI D.O.C.**

### 3.3.4 Envisaged Outcomes

In the frame of the pilot, Geocledian further developed the current data processing platform into a Big Data Processing Platform that allow the scalable production, provision & analysis of large scale data sets. In particular, this allows the new vineyard-specific products of all test sites of the project to be integrated into farm management systems like Abaco’s SITI4Farmer.

The combination of remote sensing with in situ field & weather data enabled the following developments:

- Management Zones Maps;
- Combined analysis methods of combined field & weather data provided by Abaco with remote sensing data;
- New, grape-specific higher level information products - Integration of additional data sources;
- Data anomaly detection procedures to detect features in the satellite data that allow issuing warnings to farmers when potentially interesting farm management related issues are detected;
- User-specific Visualization of big data analytics that are relevant for the farmer

Abaco used Geocledian’s satellite data products, from sensors, and from the users of the system, to create knowledge maps and data systems to put in relation the culture quality with all the other variables.

### 3.3.5 Advancement during the two first years

During the first two years the two wine makers pilot partners were selected, onboarded and trained after the hard- and software was deployed. For the collection of weather and soil data a weather station and soil sensors have been installed at the vineries. The acquisition, processing and delivery of satellite data for all pilot sites was started. ABACO implemented connectors to improve communication with weather station and in general sensors from external providers. Since the project start, SITI4farmer has been updated with a series of developments related to the farm management and related data. These developments form the basis to support the above described use cases. ABACO is continuously upgrading interface and functionalities related to Best Practices, Precision Farming issues and remote sensing image import and visualization coming from GEOCLEDIAN. GEOCLEDIAN has delivered all available USGS Landsat 8 and Copernicus Sentinel 2 satellite images for 2013 – 2019 for all fields in the system. Visible images and Vegetation Index Maps have been produced, and the data provided to SITI4Farmer. After a detailed system analysis a series of developments have

been implemented and deployed successfully to improve data download, processing, quality, performance monitoring, scalability and data visualization and to enable the delivery of the new data products and 7 new vegetation indexes that were developed for ABACO (NDVI, NDRE1, NDRE2, NDWI, SAVI, EVI2, CI-RE).

### INVOLVEMENT OF END USER EVALUATION PANEL

The new Dashboard for the Water Stress and irrigation decision support system was evaluated by a panel of end users comprising the 2 already engaged winemakers, plus a winemakers consortium of the Oltrepo Pavese DOC Area (South west Lombardy Region), and 2 other consortia of Corn, Wheat and Tomato producers of the Po plain valley and Tuscany province of Siena.

The involvement of farmers of different crops enlarged the representativeness of the evaluators panel.

End-User Name	Type of Organisation	Best way to interact
TORREVILLA	Winemakers	Face to Face
CONSORZIO DI SIENA	Farmers	Face to Face
CONSORZIO DI CREMONA	Farmers	Face to Face
IL PALAZZO	Winemakers	Face to Face
CASATO PRIME DONNE	Winemakers	Face to Face

#### 3.3.6 Data flow experiments end to end report

In conclusion to our experimentation for the data flows, for this specific pilot we can see that overall the performance of the stack shows the necessary scalability. In particular the data pipeline step, involving Apache Kafka, MongoDB and Elasticsearch shows very high scalability when performed with high parallelization and low performance in terms of completion time if done otherwise. For the satellite image processing dataset ingestion, we consider that the best performance is also achieved when increasing its concurrency. However, high network usage is observed in this case, which leads us to believe that further experimentation is needed to ensure that this observation does not create a problem when the stack is used by many concurrent users and pilots. The rdfization process shows a nice performance when performed using the command line tool. Moreover, for the extraction of semantically enriched data, an excellent performance is observed in the real-life scenario. However, as the data volumes increase, high completion times are observed which leads us to believe that further experimentation is needed, along with possible changes in terms of employed technologies/frameworks for this specific step. Finally, the prediction step shows interesting performance in terms of latency and resource usage, disregarding the number of observations to use for prediction.

#### 3.3.7 Quantitative Evaluation Against KPIs

##### Domain Specific KPIs

ABACO has acquired the list of domain specific KPIs and their baseline values for the Farm Management Pilot from the IL Palazzo test site in Italy, which are presented in the following table.



Table 6: Farm Management Pilot Domain Specific KPIs

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Harvested Area</b>	ha of harvested area	ha	29	34	34
<b>Product Yield</b>	Kg per ha of grapes harvested, wine produced, raisins produced	Kg/ha	100	90	80
<b>Grape Product Quality</b>			High	High	High
<b>Production Costs</b>	Costs in Euros per year	Euros/year	5500	5500	5600
<b>Organic Fertilizer Use</b>	Kg fertilizer used per kg grapes harvested per year	Kg/kg y	500	500	500
<b>Organic Pesticides Use</b>	Kg pesticides used per kg grapes harvested per year	Kg/kg y	18	23	30
<b>Irrigation Cost</b>	Euros per ha	Euros/ha	0	0	50
<b>Fertilization Cost</b>	Euros per kg	Euros/kg	400	500	650
<b>Pesticides Cost</b>	Euros per kg	Euros/kg	600	600	700
<b>Labour Cost</b>	Euros per hour	Euros/hr	2500	2600	2600

### Technological KPIs

Additionally, in order to perform a complete quantitative evaluation for the Farm Management Pilot, a Technological KPIs list along with baseline values have been defined by ABACO and Geocledian.

Table 7: Farm Management Pilot Technological KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019 (Up to M18)	2020
<b>Focusing Big Data</b>					
<b>Volume</b>	Variation in raw data volume – Sentinel2	GB	130.5	139.5	127.5**
<b>Volume</b>	Variation in raw data volume – Landsat8	GB	54	57	60**
<b>Volume</b>	Variation in raw data volume – Pessl Instrumens	MB	2.5	3.3	-

<b>Variety in data</b>	Sentinel 2	Number of scenes	174	186	170**
<b>Variety in data</b>	Landsat 8	Number of scenes	54	57	60**
<b>Variety in data</b>	Pessl Instrumens	Hours	24h * 90 days	24h * 365 days	
<b>Variety in Data Source Types</b>	Data sources (Sentinel 2, Landsat 8, Pessl Instruments)	-	3	3	-
<b>Variety inter Data</b>	All variables measured (Satellite vegetation Indices, Soil data, Weather data, Canopy data)	Datasets	47	47+7 (new satellite indices)	-
<b>Velocity</b>	Speed of data generated – Sentinel 2	GB / month	10.88	11.63	11.59**
<b>Velocity</b>	Speed of data generated – Landsat 8	GB / month	4.50	4.75	5.45**
<b>Velocity</b>	Speed of data generated – Pessl Instrument	GB / month	0.25	0.3	-

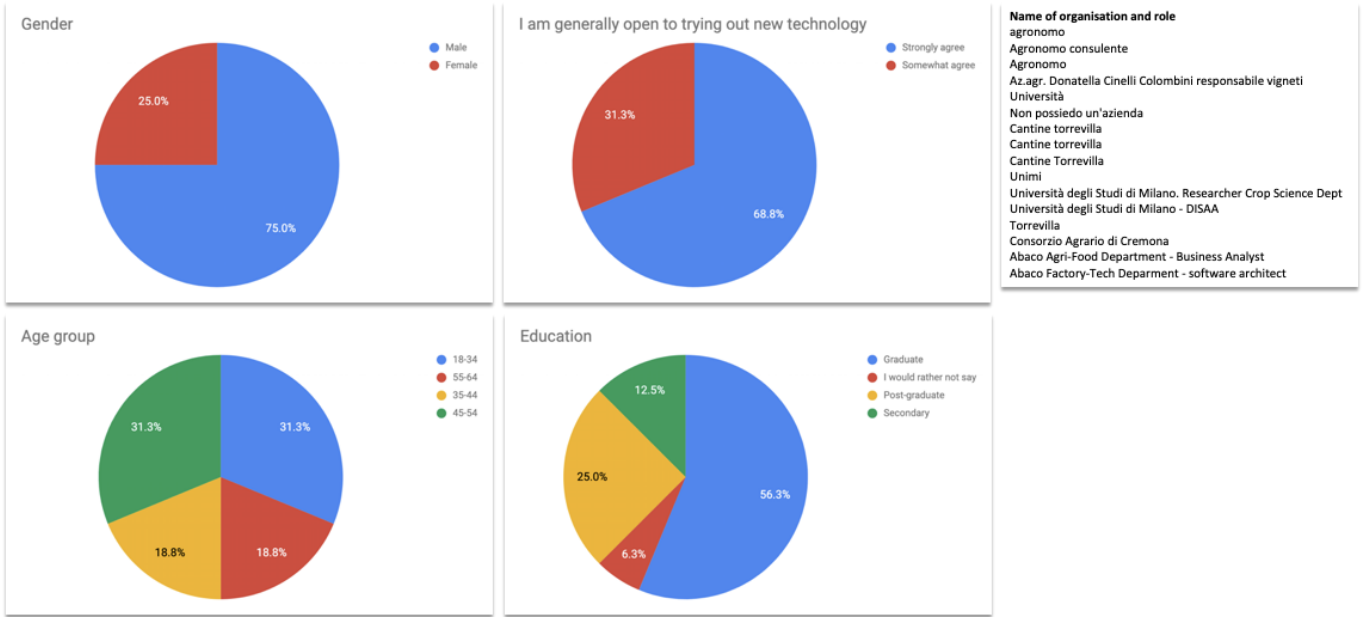
\*\* measured until November 30, 2020

### 3.3.8 Final results of the pilot

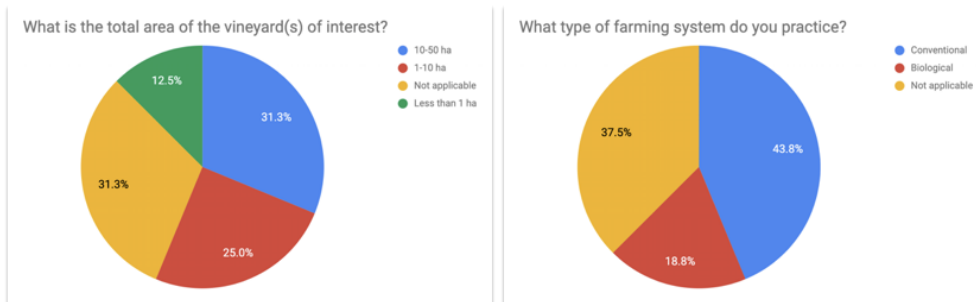
#### Farm Management Pilot Vineyard Information and Participant Demographics

##### Participant Demographics

The participant demographics has been represented by wine makers agronomist and farmers, crop science university researchers and also software engineers. In particular the age of the participants in the majority of cases were male, below 45 years showing a particular predisposition to the new technologies. Participants were in the majority of cases graduated in crop science faculties and for over the 30 % not owners of vineyards. Among those owners, the vineyards surfaces were under 30 hectares, with conventional management. There were also agronomist working not only on vineyards but also with corn, tomato and wheat farms.

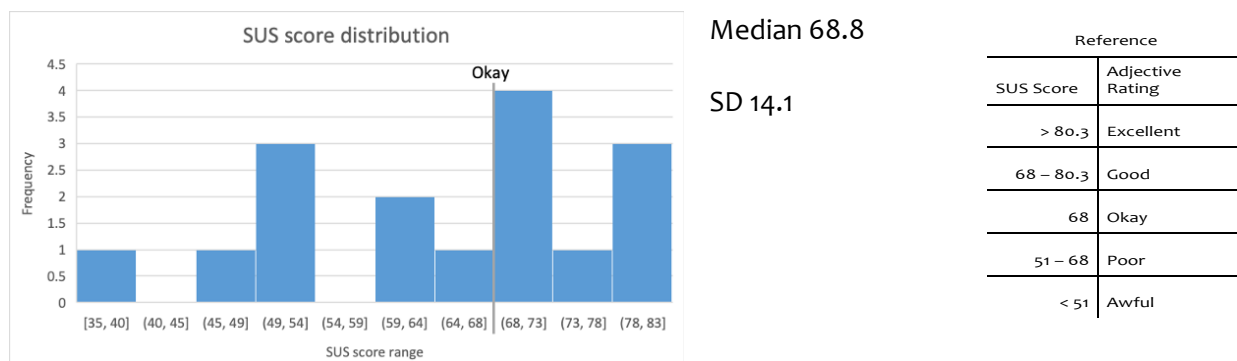


**Vineyard Information**



**Farm Management Pilot - Irrigation Task**

Figure 17 shows results of the SUS questionnaire. The median of 68.8 puts the interface slightly above the acceptable region on the SUS scale. The SUS score distribution tells us that half (8/16) of the participants yielded the scores below 68.



**Figure 17: SUS score of the irrigation interface developed for the farm management pilot**

### 3.4 NATURAL COSMETICS PILOT

#### 3.4.1 Pilot description

The Natural Cosmetics pilot intends to gather samples of vineyard by-products across the Greek territory. There is a need to extract the most out of pharmaceutical plants for both economic and environmental reasons. A real challenge is to add high value to by-products. Wine making produces a lot of by-products that may have a significant biological value if there are adequate data concerning farm management. These data can lead to decisions concerning the processing of by-products in order to produce high added value active ingredients for cosmetics and food supplements. Bioactive compounds from winery by-products have disclosed interesting health promoting activities both in vitro and in vivo. If properly recovered, they show a wide range of potential and remunerative applications in many industrial sectors, including cosmetics, pharmaceuticals, biomaterials and food. In fact, winemaking by-products are outstanding sources of oil, phenolic compounds and dietary fibre and possess numerous health benefits and multifunctional characteristics, such as antioxidant, colouring, antimicrobial and texturizing properties.

#### 3.4.2 Specific goal

The scenario presumes that precision farming and control of parameters linked to the quality of wine (soil characteristics, GIS data etc) may provide by-products of superior quality. In particular, the pilot intends to gather samples of vineyard by-products across the Greek territory and more specifically vine leaves of two different grape varieties (Agiorgitiko and Mandilaria) and test their phytochemical profile and biological value after extraction.

#### 3.4.3 Site Description

For the first and second year of the project, sixteen vineyards from 5 Greek geographic regions have been chosen for sample collection, i.e. dried vine leaves of two different grape varieties (Agiorgitiko and Mandilaria). Also, samples of both grape varieties from the vineyard of Hellenic Agricultural Organization “DIMITRA” located in Attica will be tested. The dispersion and origin of the samples is shown in the following map, where the samples of Agiorgitiko are pictured in green and the samples of Mandilaria in red.

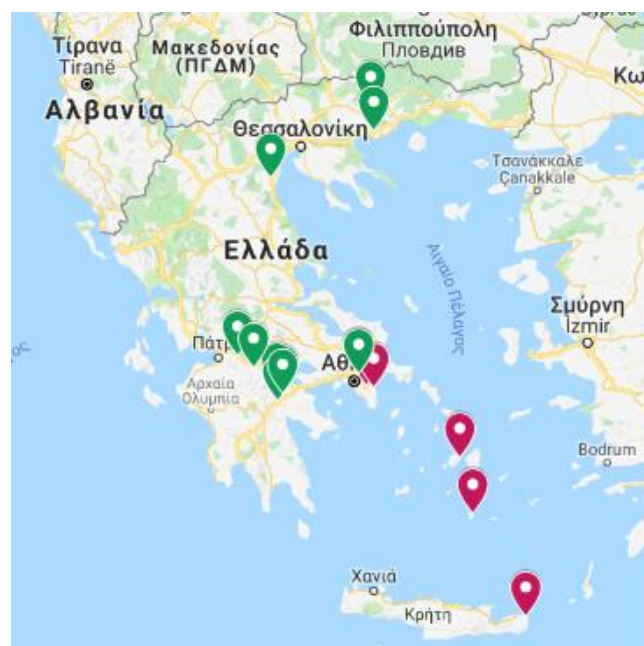


Figure 18: Dispersion of samples across the Greek territory

In the following table there is a list of the vineyards chosen for sample collection and their location.

**Table 8: List of the vineyards chosen for sample collection and their location**

Vineyard	Grape Variety	Region	City
Semeli Wines	Agiorgitiko	Peloponnese	Nemea
Pavlidis Estate	Agiorgitiko	Northern Greece	Drama
RIRA Vineyards	Agiorgitiko	Peloponnese	Aigio
Vassaltis Vineyards	Mandilaria	Aegean	Santorini
Strofilia Estate Winery	Agiorgitiko	Peloponnese	Stimfalia
Papagiannoulis Winery	Agiorgitiko	Northern Greece	Katerini
Tetramythos Wines	Agiorgitiko	Peloponnese	Ano Diakopto
Skouras Domaine	Agiorgitiko	Peloponnese	Argos
Moraitis Winery	Mandilaria	Aegean	Paros
Toplou Winery	Mandilaria	Crete	Sitia
Aoton Winery	Mandilaria	Attica	Peania
Biblia Chora Estate	Agiorgitiko	Northern Greece	Kavala
Papagiannakos Domaine	Mandilaria	Attica	Markopoulo
Hellenic Agricultural Organization "DIMITRA"	Mandilaria	Attica	Lykovrisi
Hellenic Agricultural Organization "DIMITRA"	Agiorgitiko	Attica	Lykovrisi
Agricultural University of Athens	Agiorgitiko	Peloponnese	Nemea

### 3.4.4 Envisaged Outcomes

Bioactive compounds found in wine-making by-products such as vine leaves possess multifunctional characteristics and show a wide range of potential and remunerative applications, concerning health promoting activities. Nevertheless, the quality of these by-products and more specifically their biological efficacy can vary depending on multiple parameters, such as the origin of the sample, the recovery process and more.

The collected data from the natural cosmetics pilot provided the necessary information for the evaluation of the quality of each sample, linked with the special characteristics of the vineyard of origin. The goal is to face the challenge: "how data from the field can be linked to the biological efficacy of final products - an application on wine making by-products".

### 3.4.5 Advancement during the two first years

From the very first months of the project, Symbeeosis (ex-APIGEA), as pilot’s responsible partner trained producers in the correct collection of the leaves and stems and especially the drying process that is very important in order to avoid contamination and be able to process further the products for the pilot’s deliverable. Two important indigenous grape varieties (Agiorgitiko and Mandilaria) were chosen for the needs of the Natural Cosmetics pilot and sixteen over twenty-four vineyards that Symbeeosis approached have agreed to gather samples of dried leaves of the two different varieties from their vineyards for the three years of the project. The first data on the results from the gathered samples on the biological efficacy (pH, refractive index, Total microbial count, Yeasts & Moulds, TPC, TFC, Antioxidant activity) have been analysed by Symbeeosis, while second year’s samples have been already delivered to the laboratory for their analyses. Data of the first year were already correlated with data of vegetation indices from satellites Sentinel2 and Landsat 8 collected and processed by GEOCLEDIAN. In order to test the hypothesis whether the location and field management are correlated with the BA parameters measured in the laboratory and to deliver a mathematical process that could

serve for their prediction CNR studied the datasets and presented significant correlations between SVIs and BA parameters. This approach was repeated as soon as second year samples' analyses were completed, with both years' data for BA, SVIs and additionally weather data. The outcome of mathematical processing was then used for the creation of the DSS dashboard, while third year's data were used for validation of the system performance.

### 3.4.6 Data flow experiments end to end report

In conclusion to our experimentation for the Natural Cosmetics pilot and the data flows its dataset have inside the BDG stack, we consider the overall performance of the distinct steps as very successful. In the context of the data uploading step we observed that this step is an easily scalable one, since even with 10K concurrent requests the respective components did not show any downtime. However, due to the high network usage as the concurrency increases, we note as an upper bound for the performance of the stack that of 5,000-7,000 concurrent requests at most. The rdfization step also shows very good performance when executed using the command line tool, a tool triggered by cron jobs installed in the platform. The extraction of the semantically enriched data is the one presenting a bottleneck for this pilot (as was the case for the previous). This leads us to believe that to achieve the best performance for this step, the exported data should either be split into smaller batches or further experimentation employing different components of the stack should be investigated. In the context of the satellite image processing data, we consider this step as a highly performant one, since its completion time and the monitored metrics show very good values for the provided fields of this pilot. Finally, the correlation of data, also shows very good performance, taking into consideration the nature of this specific step.

### 3.4.7 Quantitative Evaluation Against KPIs

#### Domain Specific KPIs

SYMBEEOSIS has generated the list of domain specific KPIs for the Natural Cosmetics Pilot and has defined their baseline values, which are presented in the following table.

Table 9: Natural Cosmetics Pilot Domain Specific KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019	2020
Agiorgitiko Samples/ parcel	Number of samples per vineyard (parcel)	Number	1	1	1
Mandilaria Samples/ parcel	Number of samples per vineyard (parcel)	Number	1	1	1
Agiorgitiko Samples	Samples of vine leaves to be analysed	Number	16	16	16
Mandilaria Samples	Samples of vine leaves to be analysed	Number	16	16	16
UAE and MAC efficiency	Percentage of extract from incoming raw material	%	>60	>60	>60
Extract pH	Ranges for acceptable pH	pH	> 3.5	> 3.5	> 3.5
Extract RI	Ranges for acceptable % for RI	%	22±4	22±4	22±4

Extract TMC	Ranges for acceptable Total Microbial Count	CFU	< 10	< 10	< 10
Extract Y&M	Ranges for acceptable Yeasts and Moulds counts	CFU	< 10	< 10	< 10
Processing Time	Overall time for extraction, required analysis, and assessment of new product	Months	3	3	3

### Technological KPIs

Additionally, in order to perform a complete quantitative evaluation for the Natural Cosmetics Pilot, a Technological KPIs list along with baseline values have been defined by SYMBEEOISIS.

Table 10: Natural Cosmetics Pilot Technological KPIs Catalogue

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Focusing Big Data</b>					
<b>SVIs Volume Data</b>	Sentinel-2 A/B MSI visible & NIR bands, NDVI time series	GB	301.5	294.0	316.5**
<b>SVIs Volume Data</b>	Landsat 8 A/B MSI visible & NIR bands, NDVI time series	GB	160.0	163.0	143.0**
<b>Agiorgitiko and Mandilaria Samples UAE BA parameters Volume</b>	Data on biological efficacy of samples of Agiorgitiko and Mandilaria dried vine leaves, developed with Ultrasound Assisted Extraction	KB	58	81	120
<b>Agiorgitiko and Mandilaria Samples MAC parameters Volume</b>	Data on biological efficacy of samples of Agiorgitiko and Mandilaria dried vine leaves, developed with Maceration	KB	58	81	120
<b>Variety in Data Source Types</b>	BA parameters, SVIs	Data sources	8	8	8
<b>Variety in Data</b>	Number of different types of data (in different resolutions)	Datasets	6	6	6

<b>SVIs Velocity Data</b>	Sentinel-2 A/B MSI visible & NIR bands, NDVI time series	GB/month	25.13	24.50	28.77**
<b>SVIs Velocity Data</b>	Landsat 8 A/B MSI visible & NIR bands, NDVI time series	GB/month	13.33	13.48	13.00**
<b>BA Parameters Velocity</b>	Speed of data generated – BA Parameters	KB/season	58	23	39
<b>Weather Data Velocity</b>	Speed of data generated - WD	MB/seasons	17.6	17.5	17.4
<b>Big Data Process Metrics</b>					
<b>Data Normalization (Homogenization)</b>	Time needed for data to be available for analysis and processing	Months	3	3	3

\*\* measured until November 30, 2020

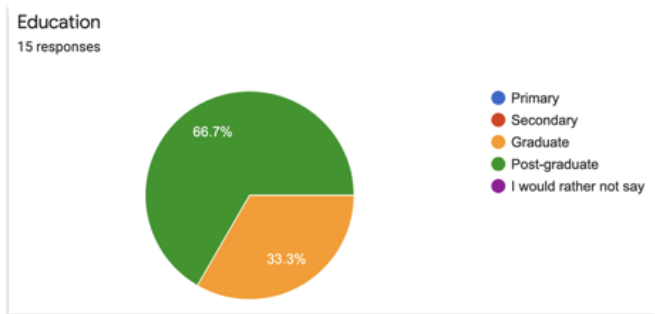
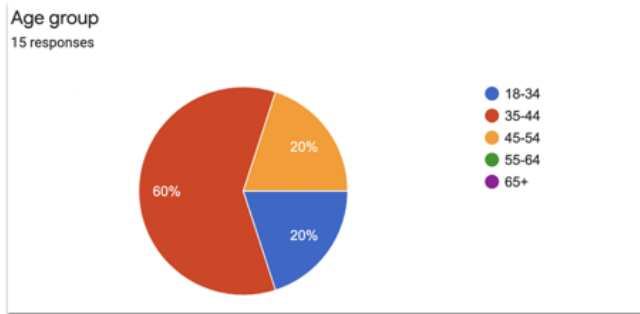
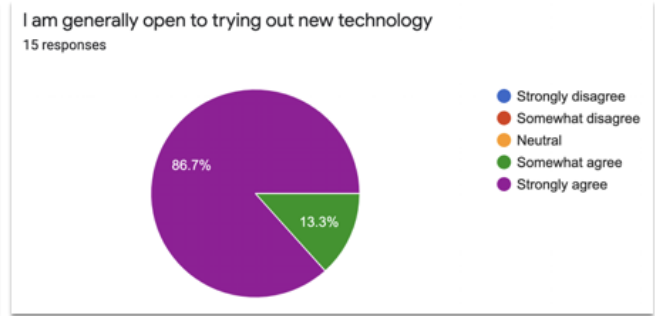
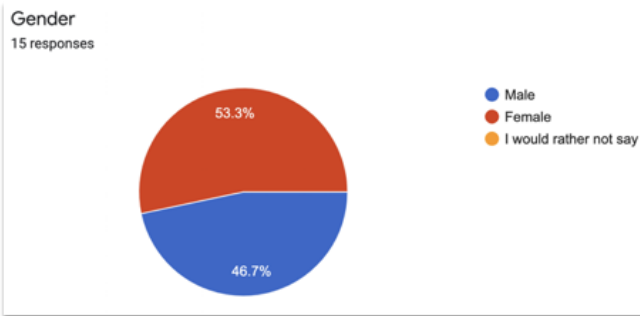
### 3.4.8 Final results of the pilot

#### Natural Cosmetics Pilot Vineyard Information and Participant Demographics

##### Participant Demographics

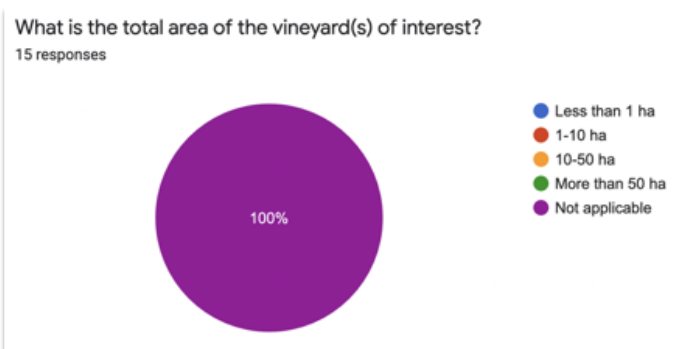
The participant individuals to the Evaluation Process run were composed by potential end-users of the Demonstrator of Natural Cosmetic Pilot, “Grapevine By-Products Biological Efficacy Predictor”. All end users were selected based on the fact that during their everyday practices are facing the competence questions on which the Pilot was developed, and thus can assess the advantages of such a DSS, dashboard’s visualisation and handling. A total of 15 end-users, with 10 of them being natural cosmetic industry’s end-users, 2 were researchers engaged with grapevine related disciplines, and 1 was a grapevine grower of a winery, have successfully completed the evaluation of the Demonstrator. Regarding the demographic of the sample, one third were graduated, while two thirds had postgraduate degrees, gender was equally distributed between male and female, all of them were younger than 54 years old with the majority between 35 to 44 years old, and all participants were open to try new technology. For the demonstrator assessment 93.3% stated that they would use the system frequently, 86.7% did not find the system complex, and 100% found it easy to handle. Although 26.7% of the sample would need the help of a technician and would need to learn a lot of things to use the platform, finally a 93.4% would learn to handle it very quickly and 86.6% felt very confident using it. All participants found the system easy to use and learn to operate it, with a clear and understandable interaction with it, and without feeling the need of additional knowledge or resources to successfully operate it. Encouraging were also the findings that 80% found the system useful for their work, 74.4% could accomplish their tasks more easily, and 66.6% could increase their productivity, with it.





**Name of organisation and role**  
 APIVITA - RD DIRECTOR  
 SYMBEEOISIS - COO  
 QS Professional (Cosmetics) - Quality control  
 R&D MANAGER FREZYDERM S.A.  
 Frezyderm S.A Plant Manager  
 GR. Sarantis S.A. - Brand Manager  
 Hellenic Agricultural Organization/Researcher  
 The NuClab  
 The NuClab, rnd cosmetologist  
 APIVITA, R&D & Extracts Development Scientist  
 APIVITA S.A. R&D Lab Scientist  
 Aoton Winery, Owner  
 Agricultural University of Athens; Lab of Food Microbiology & Biotechnology, Associate Researcher  
 The name of organisation is SYMBEEOISIS Eu Zην SA and my role is "Raw Materials Procurement and Sustainable Sourcing Manager"  
 Sales associate/Cosmetic scientist at Beautylab the Store

**Vineyard Information**



**Natural Cosmetics Pilot - Bio-efficacy Correlation Task**

Figure 19 shows results of the SUS questionnaire. This interface, among the rest, received the highest SUS score with the median of 75 and the standard deviation of 11.3. Only 5 out of the 15 responses yielded the scores below 68, leaving the rest to put the interface in the “Good” and “Excellent” regions.

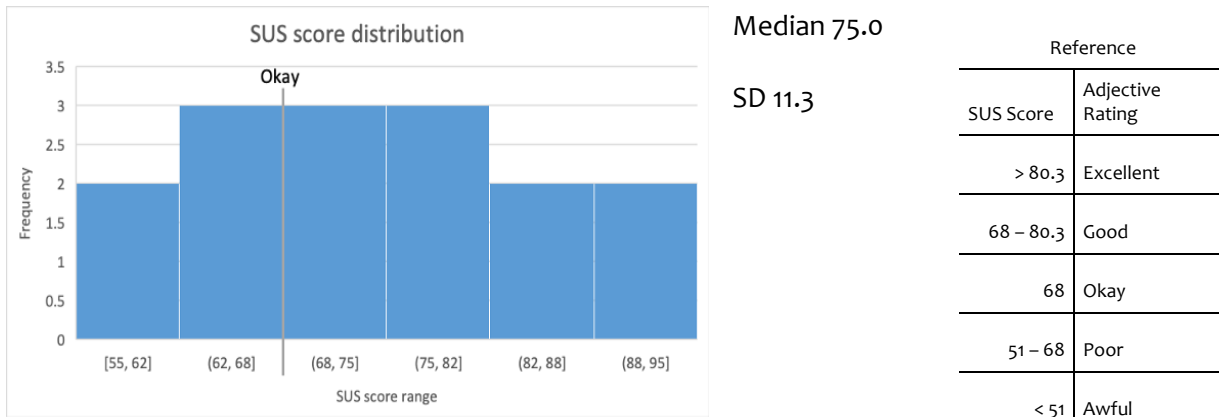


Figure 19: SUS score of the bio-efficacy correlation interface developed for the natural cosmetics pilot

### 3.5 FOOD PROTECTION PILOT

#### 3.5.1 Pilot description

Food protection, including safety and fraud, is one of the most critical parameters in food production highly affecting the food companies from the financial and brand point of view. Agroknow provided a digital solution for the food industry that delivers trends and risk estimation for raw materials, ingredients and finished products. The solution is helping the Quality Assurance (QA) and Food Safety (FS) experts working in the food industry to identify risk in their supply chain. The current solution is limited to alarms, statistics, simple trends and search mechanisms.

During the first two years of the project, Agroknow has performed a series of focused group and consultation meetings with several companies of the food industry, such as Gallo Winery, Conagra, Campbell, Pepsico, Hershey and Lamb Weston. The meetings were held during large food safety events like the GMA Science Forum. During these meetings Agroknow team validated the need for new FOODAKAI extensions that will enable risk predictions in the supply chain.

Thus, the main objective of this pilot was to enhance the current digital solution with new modules that would address further needs of the grape and wine supply chain. The enhancement focused on the further development of Agroknow’s Big Data platform with new software modules that would enable advanced data analysis and risk prediction using machine learning and deep learning methods.

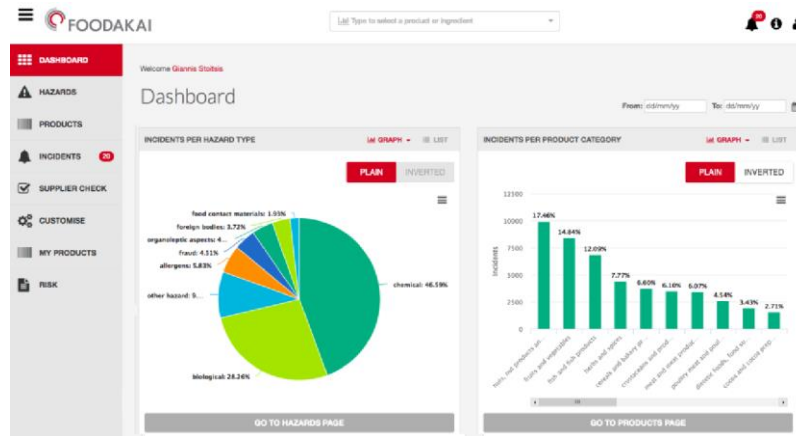


Figure 20: Food Protection dashboard of FOODAKAI system

### 3.5.2 Specific goal

The specific goals of the food protection pilot were:

- To develop a software module able to predict emerging and increasing risks for chemical hazards in the grapes and wines supply chain;
- To develop a price prediction dashboard that includes algorithms able to predict the prices of agricultural products, including grapes;
- To develop a food fraud dashboard that helps experts working in the food industry to perform an effective vulnerability assessment for products;
- To develop a marketing automation module that facilitate the exploitation of the food safety and fraud solutions that would be developed by Agroknow in the context of the project.

### 3.5.3 Site Description

This pilot did not have a physical location, all used data is retrieved from the internet.

### 3.5.4 Envisaged Outcomes

The QA and FS experts that are working in the food industry, and specifically in the grape supply chain, were able to identify early enough potential risks for their supply chain so they could take the required corrective measures and finally to prevent a food recall or a border rejection for their products. The risk prediction covered food safety and fraud risks. An existing digital solution that was already provided by Agroknow to the food industry, was enhanced and exploited in several food sectors.

### 3.5.5 Advancement during the two first years

During the first two years of the project, based on what has been described in the piloting plan, have been achieved the following aspects:

- Pesticides monitoring data (Laboratory analysis results) from 34 countries have been collected and processed by the Agroknow Data Platform;
- Pricing data from countries and EU have been collected and processed by the Agroknow Data Platform. In particular, data sources are the Food and Agriculture Organization of the United Nations, the European Commission and the Greek market. These datasets have been cleaned, prepared and stored in Agroknow's Big Data Platform;

- A lab data analysis dashboard has been developed and tested by a control group of end-users. The Lab data analysis module was presented to food companies like Gallo Winery and AB Vasilopoulos;
- We developed a Bayesian network and applied it to food recalls, border rejections, pricing, country risk and corruption data to predict risk for specific products.
- Deep and Machine learning algorithms have been applied to pricing data in order to predict the future price for several agricultural products. In particular, have been tested 6 algorithms in order to find the best one. These algorithms are Moving Average, K-Nearest Neighbors, Linear Regression (Machine learning) and Arima, Prophet, Long Short Term Memory (Deep Learning). A price prediction dashboard has been developed in collaboration with KUL;
- A process for personalized marketing messages based on the profile of the target company and on the data powered reports have been developed and deployed.

### 3.5.6 Data flow experiments end to end report

In this section we experimented on the datasets provided by the food protection. In terms of the data flows specific for this pilot we have identified that the initial step the dataset upload step, presents good performance in respect to the completion time as well as the CPU, network and memory usage. It is a step that can be easily made with a high degree of concurrency without seriously affecting the rest of the stack. Following the experimentation part, recall prediction API presents very good performance in respect to the completion time as well as the CPU, network and memory usage. Performance increased by lowering the volumes of data handled. Increasing data needs more time, more CPU and memory to be trained. The price prediction API as well shows good performance, with a constant latency per product and an overall latency that is dependent from the number of products in the dataset. The CPU and memory usages are constant as well, indicating the resources are not limiting the performance of the module.

### 3.5.7 Quantitative Evaluation Against KPIs

#### Domain Specific KPIs

For the cost reduction we developed a Return-on-Investment calculator and we used it to estimate the cost reduction in collaboration with the food companies that are using the FOODAKAI risk estimation and prediction. Agroknow has generated the list of domain specific KPIs for the Food Protection Pilot and has defined their baseline values, which are presented in the following table.

Table 11: Food Protection Pilot Domain Specific KPIs Catalogue

Variable	Definition	Units	2018	2019 Baseline	2020
Cost reduction	Reduce cost of running risk estimation, including travelling costs	Euros/year	N/A	0.5M	0.25M
Productivity increase	Reduce the time that is needed to perform risk estimation and prediction	Hours	N/A	2 months every year	1 month per year (50% reduction)

#### Technological KPIs

Additionally, in order to perform a complete quantitative evaluation for the Food Protection Pilot, a Technological KPIs list along with baseline values have been defined by Agroknow.

Table 12: Food Protection Pilot Technological KPIs Catalogue – Lab Data

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Focusing Big Data</b>					
<b>Volume Data</b>	The size of all the data that are stored in the Big Data Platform	GB	3	100	117.6
<b>Variety in Data Source Types</b>	The number of the data sources from which we are collecting information about food safety incidents	Data sources	10*	29	32
<b>Variety in Data</b>	The different types of data that we are processing in Big Data Platform	Datasets	1	1	1
<b>Velocity Data</b>	The growth of all the data types in the Big Data Platform	MB/month	9.9	8,000	1,460
<b>Velocity Data</b>	The growth of all the data types in the Big Data Platform	GB/month	0.099	8.08	1.46
<b>Big Data Process Metrics</b>					
<b>Data Normalization (Homogenization)</b>	Time needed for data to be available for analysis and processing	Months	0.1	2.5	0.4

Table 13: Food Protection Pilot Technological KPIs Catalogue –Food recalls and border rejections

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Focusing Big Data</b>					
<b>Volume Data</b>	The size of all the data that are stored in the Big Data Platform	GB	2.32	2.55	2.8
<b>Variety in Data Source Types</b>	The number of the data sources from which we are collecting information about	Data sources	45*	45*	45*

	food safety incidents				
<b>Variety in Data</b>	The different types of data that we are processing in Big Data Platform	Data Types	1	1	1
<b>Velocity Data</b>	The growth of all the data types in the Big Data Platform	MB/month	19.1	20.1	14.4

**Big Data Process Metrics**

<b>Data Normalization (Homogenization)</b>	Time needed for data to be available for analysis and processing	Months	0.5	0.7	0.85
--	--	--------	-----	-----	------

Table 14: Food Protection Pilot Technological KPIs Catalogue – Price data

Variable	Definition	Units	2018 Baseline	2019	2020
<b>Focusing Big Data</b>					
<b>Volume Data</b>	The size of all the data that are stored in the Big Data Platform	GB	2.27	2.5	2.68
<b>Variety in Data Source Types</b>	The number of the data sources from which we are collecting information about food safety incidents	Data sources	3	3	5
<b>Variety in Data</b>	The different types of data that we are processing in Big Data Platform	Data Types	1	1	1
<b>Velocity Data</b>	The growth of all the data types in the Big Data Platform	MB/month	19.3	14.75	15

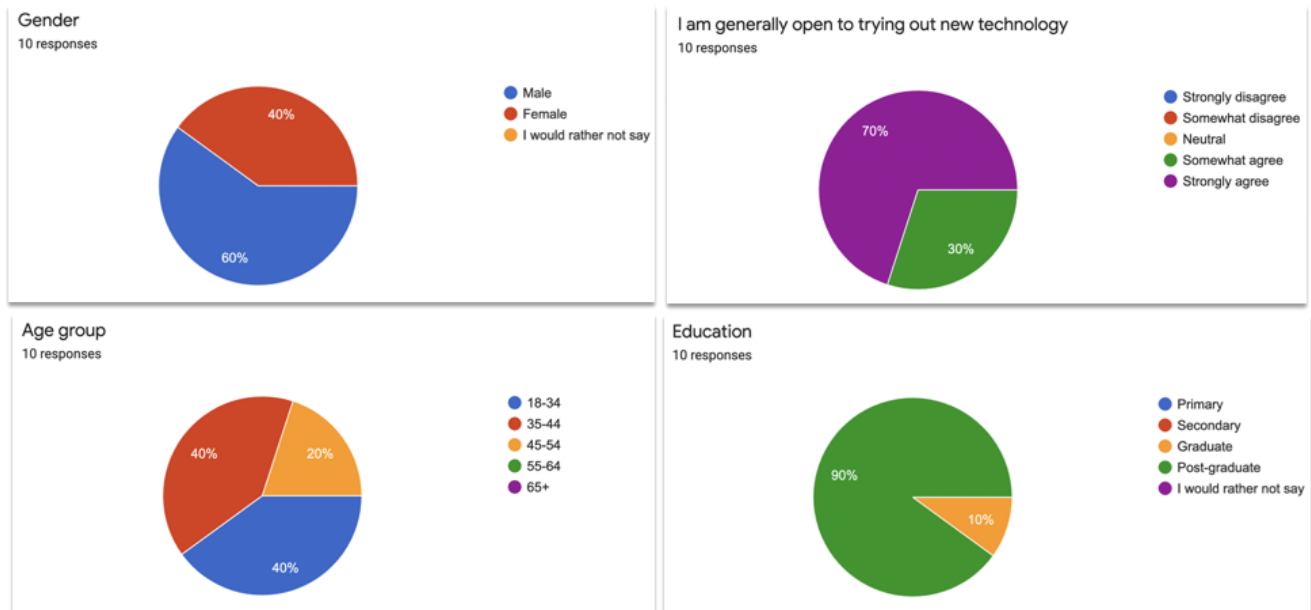
Data Normalization (Homogenization)	Time needed for data to be available for analysis and processing	Months	0.45	0.65	0.75
-------------------------------------	--	--------	------	------	------

### 3.5.8 Final results of the pilot

#### Food Protection Pilot Vineyard Information and Participant Demographics

##### Participant Demographics

The participants in the Food protection pilot, ten in total, were recruited through invitations to a variety of organizations so as to ensure the perspective of different types of experts, including food science and quality assurance experts (3), food scientists and researchers (5), as well as business and marketing professionals familiar to prediction processes as part of their work (2). Six men and four women participated in the evaluation, of diverse age groups, from 18 - 64 and all had reached a level of post-graduate education, except one. All participants reported interest and openness to try new technologies.

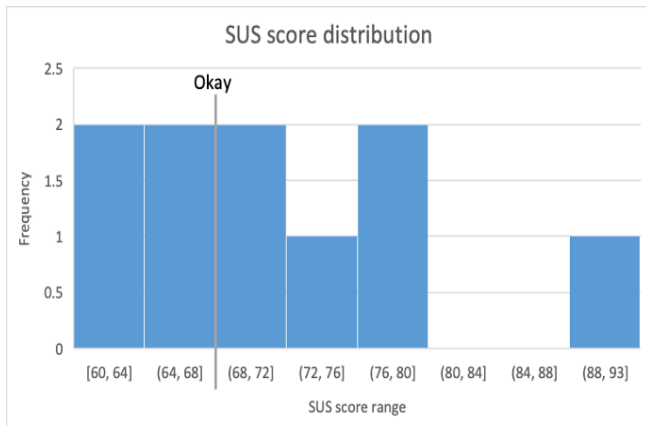


- Name of organisation and role**
- Certification Body
  - Agricultural University of Athens, Assistant Professor
  - AUA - Research Associate
  - TUV AUSTRIA HELLAS - Food Safety Auditor
  - Cool Bananas Corporation, Marketing Manager
  - INRAE- pilot leader for BDG project
  - AUA
  - National Institute of Research and chemico-physicals analysis Tunisia, Ph.D student
  - Hellenic Agricultural Organisation - Institute of Technology of Agricultural Products
  - Teaching and Research Staff at Laboratory of Viticulture (AUA)

#### Food Protection Pilot - Risk Assessment Task

Figure 21 shows results of the SUS questionnaire. This interface received the second highest SUS score, following the highest scoring interface of the natural cosmetic pilot, with the median of 70 and the standard deviation of 9.3. The score distribution shows that 4 out of the 10 responses yielded the scores below 68,

meanwhile the majority (5/10 participants) put the interface in the “Good” region and one put it in the “Excellent” region.



Median 70.0

SD 9.3

Reference	
SUS Score	Adjective Rating
> 80.3	Excellent
68 – 80.3	Good
68	Okay
51 – 68	Poor
< 51	Awful

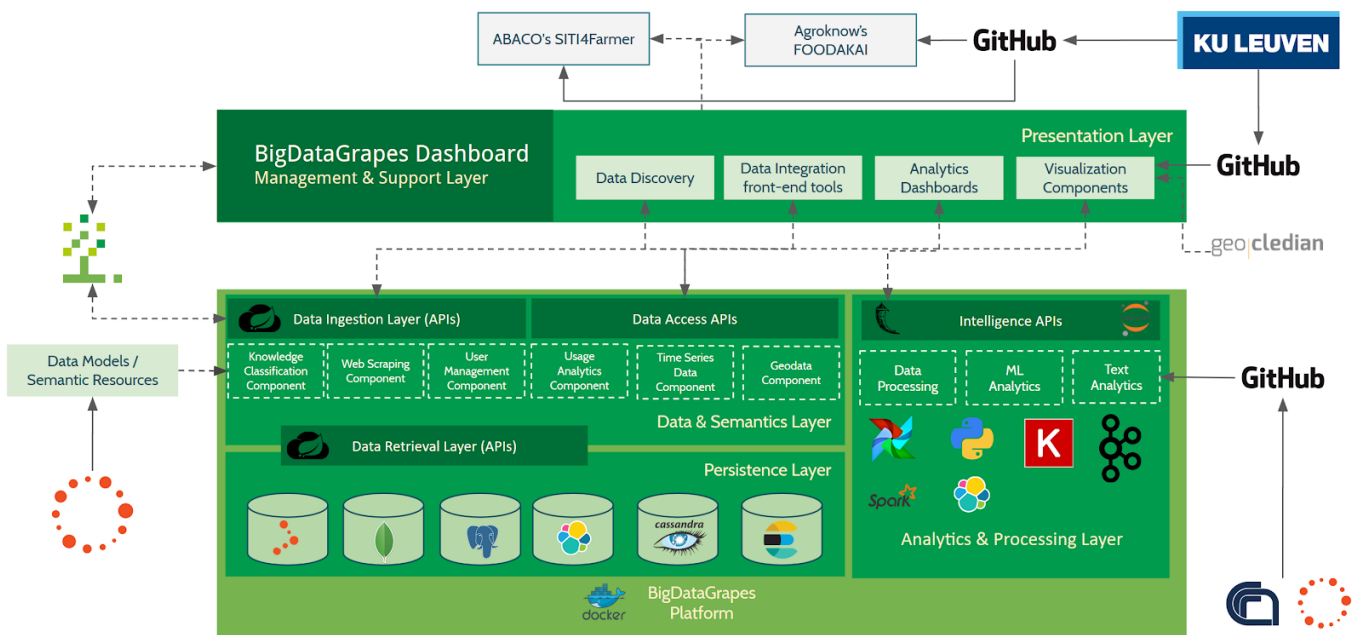
**Figure 21: SUS score of the risk assessment interface developed for the food protection pilot**



## 4 BIGDATAGRAPES TECHNICAL SOLUTION

### 4.1 OVERALL APPROACH

The data problems and their respective application scenarios demand the provision of a complete computational solution that serves all aspects of the Big Data management value chain. To this end, BigDataGrapes build and deploy a set of components that carry out the various processes in order to (a) solve the data problems of the grapevine-powered industry and (b) ensure transferability and extensibility to other data-driven businesses by adopting a coherent architectural approach.



**Figure 22: The BigDataGrapes top-level architecture**

A critical goal for the overall system is to ensure that the relevant data sources are semantically annotated and integrated as parts of a common data pool comprising disparate yet interconnected data assets. To this end, BigDataGrapes designs and develops methods and components for the semantic linking and enrichment of the available data and furthermore, makes available to the domain experts' tools for annotating and describing their data in order to be incorporated in the BigDataGrapes pool.

At the processing stage, BigDataGrapes employs the necessary components for carrying out typical analytics processes, making sure that the execution environment and methodology retain scalability and efficient use of computational resources. Additionally, the project designs and implements inference and machine learning methods to produce advanced predictive analytics over the entirety of the available data pool, tackling open issues like the movement to distributed architectures for these tasks.

Finally, BigDataGrapes leverages the value elicited from these data, by translating them to intuitive and actionable knowledge and using them as the foundation for decision support in complex environments.

### 4.2 ADVANCEMENT ON DATA & SEMANTICS

During the first year of the project, the initial and most important steps for identifying and making available the tools and components pertaining to BigDataGrapes semantic layer were carried out. More specifically, the Data & Semantics Layer presented:

1. Extensive overview of existing ontologies;

2. Dataset analysis of all of the data provided by the partners;
3. Definition of a modelling methodology;
4. Establishment of a harmonized data model;
5. Establishing of a shared vocabulary;
6. Mapping of source data;
7. The mapping of pilot needs to a list of competency questions, so that the first pilot scenarios would emerge.

Furthermore, regarding the architectural components that were identified and put in place, the Data and Semantics layer has been extended to support the use of GraphDB, MongoDB, PostgreSQL, Apache Cassandra and the ELK stack for the necessary data ingestion and integration routines. Additionally, an end-to-end ingestion workflow has been developed for sensor data, along with adjustments on the Wrapper API to support querying and storage of the underlying data. To further encourage the adoption of GraphDB, an RDFization API endpoint has been developed, that transforms tabular data to RDF following the generated BDG data models.

To address the big data indexing needs of the project, novel time and space efficient data structures for indexing structured and unstructured data such as labelled trees, graphs, and text documents were implemented. The data structures focused on text documents and RDF data and the RDF index based on said structures outperformed existing solutions.

Lastly, ONTOTEXT has presented a Semantic Enrichment approach by formulating a use case together with SYMBEEOSIS, based on scientific and medical unstructured content. The relevant activities were: requirements analysis for the use case, development of crawling and ETL workflows, analysis of resulting linked datasets, ingestion of ontologies and vocabularies, application of text analysis pipelines and definition of the resulting annotation model and its correspondent in RDF.

### 4.3 Advancement on Data Analytics & Processing

The BigDataGrapes platform has been enriched with methods and components towards the end-goal of enabling real-time answers for critical business decisions over heterogeneous data sources. Said components have been designed and implemented, deployed and used for efficient processing of large datasets.

The first step was the design of the architecture of components in the wider context of the BigDataGrapes platform, but also the selection of technologies used for efficient data processing over extremely large datasets. With the requirements that were elicited during the process, two demos were identified, performing scalable operations on geospatial raster data using the Spark-based GeoTrellis geographic data processing engine of the platform.

To address inference scenarios over big data, Ontotext's GraphDB was selected as the most appropriate candidate technology. The developed distributed inference engine of BigDataGrapes was defined as a set of external to GraphDB instances which are configured to access data in real time and synchronize inference indexes, power inference algorithms and provide provenance of newly inferred facts. To that end, a number of extensions on GraphDB technology have been proposed and detailed steps for evaluation of possible use cases have been planned.

All analytics tasks were carried out in the scope of a scalable platform, based on the software stack of BigDataEurope. On this basis, four demonstrators were released in the form of Jupyter Notebooks, showcasing four different machine learning tasks based on the proposed platform, in the scope of a generic market penetration scenario that resulted in three different scientific publications.

To define the optimization needs of the platform, the concept of OnLine Data Intensive systems was explored, since the BigDataGrapes platform follows an identical nature. To that end, an OLDI Simulator was developed to

simulate the performance of the platform and its underlying physical components, the results of which were communicated in two separate scientific publications.

### 4.4 Advancement on Visualisation & Decision Support

Decision Support System (DSSs) are designed to assist users with decision making activities while dealing with massive amounts of data. In the field of agriculture, different stakeholders such as farmers, advisers and policymakers use DSSs to often facilitate farm management and planning tasks. Depending on the type of decision support required, data is first gathered from multiple sources including sensors, satellites and in-field observations, and analysed using a series of statistical models. The output is then presented to users in a number of ways such as tables and/or graphs.

To develop a trust aware DSS, the system must be transparent, meaning it must be able to clearly communicate the prediction model with users and show differing effects of input variables on the model’s output. Thus, we delivered a DSS, named AHMoSE (Augmented by Human Model SElection), which allows viticulture experts with little to no ML (machine learning) experience to answer the following questions in viticulture:

- 1) “Which machine ML model should I use with the data that is specific to this vineyard/grape variety?”,
- 2) “How do various grape parameters affect the quality predictions of different ML models?” and
- 3) “Which of the different ML models produces an output that is in-line with my knowledge?”

AHMoSE compares and explains the predicted outcomes of various machine learning models and helps domain experts to select the models that fit their knowledge the most. Through user evaluations, it has proven to be a potentially useful tool for viticulture experts. However, AHMoSE had a limitation. There are other factors that influence the interpretability of a system, such as end-users' understanding of individual features and the total number of features. Simply put, if the number of variables rises or if the end-user is unfamiliar with the selected features, AHMoSE can become difficult to use. Thus, involving end-users in the feature selection process may be key to achieving interpretability. Besides, previous work has suggested that to obtain satisfactory interpretability and predictive performance, the feature selection process should look for a subset of features that are highly correlated with the response variable yet uncorrelated to each other. **Taking this into account, we developed a system for correlation visualisation, named GaCoVi (Gapped Correlation Visualisation). It is designed to put viticulture experts in the loop of the feature selection process which is a preliminary to the decision support offered by AHMoSE.**

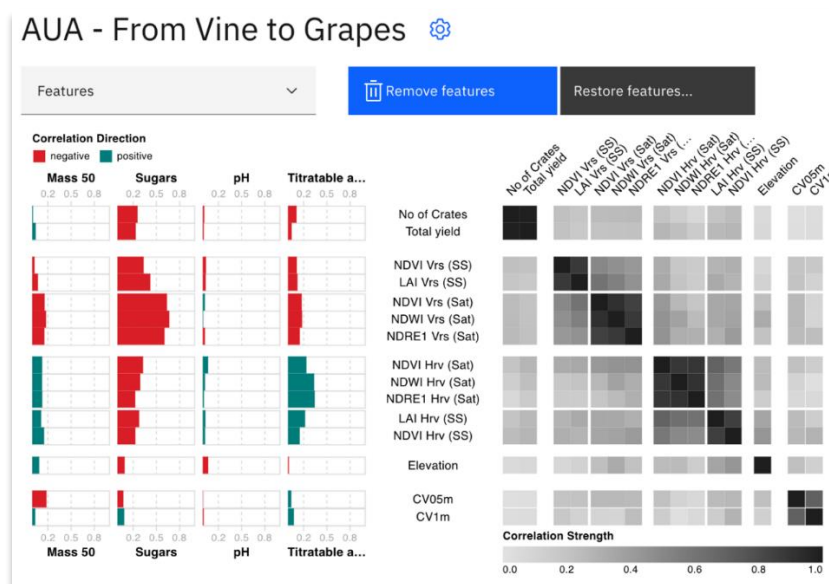


Figure 23: GaCoVi interface

## 4.5 The BigDataGrapes Software Stack

The individual technical components produced in the context of BigDataGrapes, are integrated in a configurable and deployable software stack. The BigDataGrapes software stack is logically organised into three main layers as presented in Figure 19.

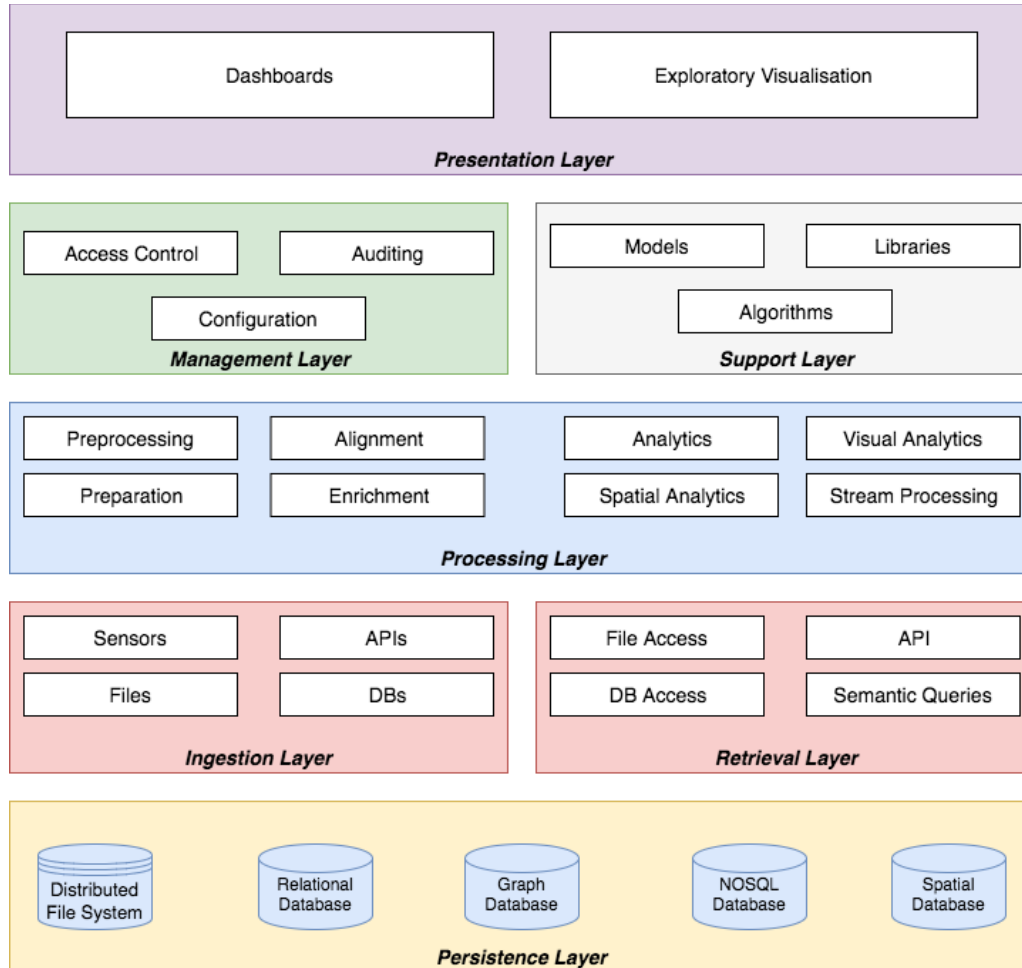


Figure 24: BigDataGrapes software stack layers

The purpose and scope of each of the three layers are summarised in the following subsections.

### 4.5.1 Persistence Layer

The layer deals with the long-term storage and management of data handled by the software stack. Its purpose is to consistently and reliably make the data available to the processing layer. The layer incorporates schema-less persistence technologies, that do not pose processing overheads either when storing the data or retrieving them. Therefore, the storing and retrieving complexity is minimized. The components used are:

- MongoDB
- HADOOP distributed file system
- HBASE
- ELASTICSEARCH
- MySQL
- GRAPHDB
- Virtuoso

- NEO4J
- APACHE Cassandra

#### 4.5.2 Data Ingestion Layer

Data ingestion is the first step for building data pipelines and also, one of the toughest tasks in Big Data processing. Big Data Ingestion involves connecting to several data sources, extracting the data, and detecting the changed data. It is about moving data from where it is originated, into a system where it can be stored, processed and analysed. Furthermore, these several sources exist in different formats such as: Images, OLTP data from RDBMS, CSV and JSON files, etc. Therefore, a common challenge faced at this first phase is to ingest data at a reasonable speed and further process it efficiently so that data can be properly analysed to improve business decisions.

In the data ingestion layer Apache Flume is used. Flume is a tool which has been designed specifically for ingesting stream data. Flume is distributed in nature, and its flexible architecture makes it a robust solution. Also, it provides a tunable fault-tolerant mechanism that can be customized to satisfy the different requirements of different sources. Its distributed nature encapsulates a variety of failover and recovery mechanisms.

#### 4.5.3 Data Retrieval Layer

In similar fashion, components of the Data Retrieval layer are responsible for exposing the stored data to the Processing and Presentation layers. Depending on the needs of each processing component, data can be retrieved via:

- Access to files lying on a distributed file system;
- Direct access to relational or NOSQL databases, via the execution of custom or pre-defined queries (depending on the use case and the degree of control that the end-users should have);
- Direct access to Graph databases and triple stores, via the execution of custom or pre-defined semantic queries (depending on the use case and the degree of control that the end-users should have);
- Calls on Application Programming Interfaces (APIs) that expose the underlying stored data in a controlled fashion.

#### 4.5.4 Processing Layer

The Processing Layer implements the core processes for data management and analysis towards serving the analytics and decision support requirements of the BigDataGrapes use cases. These operations are classified under the following main categories:

- **Pre-processing:** processes designed to validate and pre-process incoming datasets. Different data sources evidently required different pre-processing mechanisms. Exemplary operations carried out by pre-processing components are generation of provenance metadata, validation and anomaly detection, feature extraction, etc.
- **Alignment:** the alignment components are responsible for discovering and proposing links between semantic resources imported into the platform, either at the schema (ontologies, taxonomies, vocabularies) or at the instance level (identity or similarity between datasets or data items).
- **Enrichment:** the enrichment components carry out the automatic annotation of the available data with semantic information, using the semantic resources available to the platform.
- **Preparation:** The structure in which data are stored after the integration may not be suitable to perform the target analysis. The preparation modules adapt the data to match the format expected by the analytics components. Since each analytical model may expect data in a different format, data preparation is specific to each analytic model and the preparation components were enriched and extended as necessary.

- **Stream Processing:** The stream processing components are responsible for carrying out analysis over live data streams, as opposed to persistent data collections.
- **Data Analytics:** The components receive as input the prepared data from the preparation components and apply statistical and machine learning methods to extract knowledge and make predictions. At this level, descriptive analysis is done providing some statistical insights on the characteristics and behaviour of the variables under study. In turn, the predictive techniques are based on machine learning models used to explain, classify and predict the targeted variables.
- **Visual Analytics:** In similar fashion, the visual analytics components apply analytical algorithms and methods over the appropriately prepared data to generate the augmented visualisations to be presented to the end-user/ decision maker via the relevant components in the presentation layer.
- **Spatial Analytics:** The spatial analytics components entail the functionality for geospatial analysis, making use of the relevant geospatial information, as well as, invoking and using directly the other analytics components of the platform.

The components used in the processing layer are:

- Hadoop
- Sparkl
- Flink
- Sparkling Water
- Flask
- Django

#### 4.5.5 Management Layer

The layer incorporates the tools for managing and configuring the operation of the BigDataGrapes platform itself. It targets administrators and managers of a deployment and provides the functionalities for user and role definition and access credentials, log monitoring and auditing, component configuration etc.

#### 4.5.6 Support Layer

The support layer incorporates the modules and functions to be used by the processing components of the platform. These tentatively include implementations of machine learning algorithms, analytical models, geospatial operators, transformation libraries, etc.

#### 4.5.7 Presentation Layer

The presentation layer entails all the user-facing components and environments of the platform. These include platform interfaces in the different modalities supported by BigDataGrapes (web, mobile, on-site equipment), content browsing and management dashboards, administration platforms, etc. The layer also includes the components for presenting analytics results as derived from the operations of the processing layer, and the appropriate environments for directly executing data retrieval queries.

#### 4.5.8 Integration Components

Apache Kafka<sup>1</sup> is a messaging framework, that is distributed in nature and runs as a cluster in multiple servers across multiple datacentres. Moreover, Kafka allows the real-time subscription and data publishing of large numbers of systems or applications. This allows streamlined development and continuous integration facilitating the development of applications that handle either batch or stream data. An important factor in data ingestion technology, especially when handling data streams, is the fault tolerance capability of the chosen technology. Kafka ensures the minimization of data loss through the implementation of the Leader/Follower

---

<sup>1</sup> <https://kafka.apache.org/>

concurrency architectural pattern. This approach allows a Kafka cluster to provide advanced fault tolerant capability, which is a mandatory requirement for streaming data applications.

## 4.6 Docker platform

The Docker platform<sup>2</sup> is a suite of tools that offers containerization functionalities that ensure smooth building, delivery and deployment of complex software systems. The basic functionality of the Docker platform offers a template-way of packaging software components, to be easily deployed, delivered and extended. Moreover, the docker platform includes the Docker Composer, which is an orchestration engine, that is used to simplify the deployment of large and complex systems, independent of the underlying infrastructure.

### 4.6.1 Docker image

An image is package that is executable and includes everything needed, from source code to environment variables to run an application. A container is an image with state, thus whenever an image is executed it becomes a container.

### 4.6.2 Docker compose

Compose is a tool used to define and run application that use many different containers. It allows the deployment of every component and service with a single command. Compose takes as input a simple yml configuration file, that describes all the different docker images. Moreover, compose provides commands that allow the managing of the whole lifecycle of an application, i.e.

- Start, stop and rebuild services
- Monitor the current status of the services
- Monitor the log output of the running services

The documentation of deploying the BigDataGrapes software stack through Docker Compose and Docker Image can be found in <https://github.com/BigDataGrapes-EU/deliverable-D6.1>

The dockerized version of the components can be accessed through <https://hub.docker.com/u/bigdatagrapes>

---

<sup>2</sup> <https://www.docker.com/>

## 5 BIGDATAGRAPES DATA MARKETPLACE

### 5.1 BigDataGrapes Data Marketplace Goal

The goal of setting up the data marketplace for BigDataGrapes project, is to create a large-scale, multifaceted marketplace for grapevine-related data assets, increasing the competitive advantage of companies that serve with IT solutions these sectors and helping companies and organisations evolve methods, standards and processes to help them achieve free, interoperable and secure flow of their data.

The development of the data marketplace provides the necessary proof in action that grapevine-powered data assets are shared and exchanged in interoperable formats and versions, by companies and organisations responsible for them. It triggers the facilitation of free, interoperable and secure data flows, as well as the adoption, implementation and revision of data standards.

Corporate and public organisations producing and collecting those data assets can contribute them to the data marketplace demonstrator (<http://marketplace.bigdatagrapes.eu>) that serves as the project's experimentation environment.

BigDataGrapes has decided to evolve this concept into a data marketplace where data will be eventually commercially shared, exchanged and (whenever applicable) traded in a secure and confidential manner. The data marketplace focuses on the food industry and specifically on the food safety data where the European companies and organisations managing large data assets may share, exchange and trade their data.

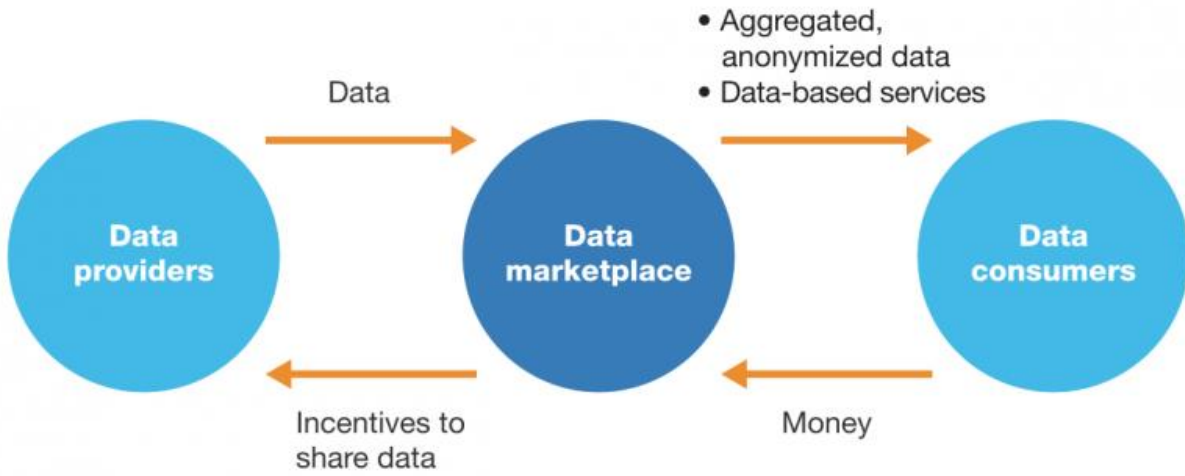
### 5.2 Introduction to data marketplaces

A data marketplace is a platform where users buy or sell different types of data sets and data streams from several sources. Data marketplaces are mostly cloud services where individuals or businesses upload data to the cloud. Those platforms enable self-service data access while ensuring security, consistency and high quality of data for both parties.

The expansion and publicity of data marketplaces relates to the growth of Big Data. Companies and organizations have started to handle and manage data as a new type of asset. Businesses constantly generate more data either internally or by collecting external data. Some of this data is valuable for other companies, too. Data marketplaces are the medium through which organizations have the ability to monetize the data. They monetize the data by offering it to other companies or individuals. With data marketplaces, monetization can be in the form of:

- Selling the data or products derived from the data
  - o Based on subscription
  - o Based on selling datasets
  - o Purchasing data from a marketplace to train and sell an AI-based product
- Using external data internally to generate value, by adding another dataset to your own business data to create better insights or new work stream





Source: McKinsey

Figure 25: Monetization in a data marketplace

Data marketplaces offer incentives such as cash or gifts to promote data sharing. With the advancements in Blockchain, data marketplaces are evolving to be more secure. Several data marketplaces have integrated blockchain technology into their solution, by using blockchain to encrypt and anonymize access to the submitted data streams from data providers. In that sense, buyers may purchase data streams through an automated smart contract. Once the transaction has been completed the tokens are distributed among the parties according to agreed prices.

**Analysis of existing data marketplaces**

Before starting the design and the development of the data marketplace, we conducted a desktop and market research on existing data marketplaces. There are various types of data marketplaces, supporting different features and targeting different segments, as depicted in the following table:

Types of data marketplaces			
Key features	Personal	Business	Sensor
Value proposition	Allows consumers to monetize their data	Allows organisations to exchange data	Allows sensor owners to monetize their devices
Transaction type	Business-to-consumer	Business-to-business	Machine-to-machine
Data type	Personal & sensitive	Public & fact-level	IoT sensor stream
Interface	App (sellers) & API (buyers)	API	API
TX confirmation	Wait for seller	Immediate	Immediate
Quality assurance	Trusted sellers	Crowdsourced reputation	Trusted marketplace operator
Pricing	Pay-per-user	Pay-per-datapoint	Pay-per-hour

Source: Towards Data Science

Figure 26: Types of data marketplaces

### Personal data marketplace

In this type of data marketplaces, individuals monetize their data by selling it to platforms. The data shared may be related to anything such as location, food preferences or website designs. Individuals either set the price for their data waiting for a buyer or accepting incentives such as sign-up cash or gift cards provided by marketplaces. Personal data marketplaces are fully GDPR-compliant since individuals are sharing their data purposely.

### B2B data marketplace

B2B data marketplaces collect and store company data from various data providers in one platform. They enable data consumers (other companies or organizations) to access an aggregate of pre-processed information from multiple sources that can be used for marketing, sales, research and BI purposes. Compared to personal data marketplaces, larger amounts of datasets are shared.

### Sensor/IoT data marketplace

An effective way to cash-in IoT data is by selling information to third parties. With a sensor or IoT data marketplace, organizations can buy or sell real-time data that is collected from an IoT device. Data collected from sensors help organizations understand consumer behavior, improve sales, and build better marketing strategies.

Significant presence in the data marketplace segment have Quandl (<https://www.quandl.com/>) and Advaneo (<https://www.advaneo.de/en/>) and Icarus 2020 (<https://www.icarus2020.aero/>). Those data marketplaces have been used as a reference for the development of the BigDataGrapes marketplace.

**Quandl** is a data marketplace that provides financial, economic, and alternative datasets, aiming to serve investment professionals. It provides a series of data for institutional clients and a series of data to individuals. The first includes data assets that have been sourced, evaluated and productized to be transformed into quantified, actionable intelligence for institutional clients. The latter provides a large set of data assets available to individuals who are interested in financial and investment activities.

**Advaneo** provides comprehensive, freely scalable solutions for data-driven business models and AI-based applications, aiming to support companies in their digital transformation. To this end, they have developed solutions that ensure data sovereignty. Its marketplace provides features such:

- Data catalog, used to manage all data entries and make them usable. All data entries of the marketplace are described and indexed in detail using a metadata standard.
- Open data for AI and ML applications, such as machine-readable data which can be used freely available by anyone under an open license.
- Closed User Groups, which are a function to make certain data sets available to certain users. A CUG consists of an exclusive group of users from one or more companies who share one or more specific data sets. The data is usually confidential.
- Selling data, which makes it possible to offer data either freely or commercially.
- Data recommendation, supporting visualization and AI development, making data recommendations and provision of the best methods of recommender systems with the rich database.
- Data science workbench, providing a set of tools for data processing.
- IDS connector, which is a software application that can be connected to a wide range of different data source.

**Icarus 2020** is an aviation data marketplace. Big Data from airlines, airports, aircrafts, and extra-aviation service providers combined with open linked data (e.g. for weather, environment, population, etc.) have the

credentials to reassess the mentality of the aviation ecosystem by early predicting critical failures and maintenance needs, optimizing flight paths, rescheduling routes at real-time, improving operational efficiency, serving a seamless ground/air passenger experience, safeguarding the environment, and monitoring safety and risk threats (like epidemics and terrorist attacks). The marketplace aims to build a novel data value chain in the aviation industry towards data-driven innovation and collaboration across currently diversified and fragmented industry players, acting as multiplier of the “combined” data value that can be accrued, shared and traded, and rejuvenating the existing, increasingly non-linear models / processes in aviation.

It provides the following features:

1. End-to-End data security allowing the data providers to process and encrypt their data on-premise and transfer them to the ICARUS Core Platform in an already encrypted form.
2. Trusted data sharing for creating, signing and validating smart data contracts in a way that dictates the terms of data acquisition between a data provider and a data consumer.
3. Advanced access control to regulate access to the privately owned data assets.
4. Secure and private analytics spaces for designing and executing analytics in private, sandboxed environments spawn on demand.
5. Intuitive data exploration in order to find, understand and explore aviation-related data,
6. Effortless Data Linking that aims at mapping and linking the privately owned data assets with external data based on a common data model.

### 5.3 BigDataGrapes marketplace design

#### *Personas*

The Personas that had been used for the definition of user stories are the following:

- **Data scientist** that is working in the R&D department of a company that conducts **research** in the food (safety) sector.
- **Researcher** that is working in a Research Center and conducts research in the food safety and quality center.
- **Data scientist** that is working in a startup that wants to **develop models for risk prediction**.

#### *Functional specifications - User Stories*

As a **Data Scientist**, I want to browse and see the available categories of data, so I can quickly browse through available datasets.

As a **Data Scientist**, I want to see the details of a dataset, so I can find a dataset that fulfils my requirements (data, documentation, usage – API, Python, R or Excel).

As a **Data Scientist**, I want to download a sample of a dataset that I found in the preferred format (API, Python, R or Excel), so I can verify if I can use it in my model.

As a **Researcher** working on Food Safety Modeling, I want to find a dataset for recalls and border rejections for a specific hazard or ingredient, so I can use the dataset in the risk assessment model that I am building.

As a **Data Scientist**, I want to search for a dataset using a free text, so I can find quickly the dataset that I am looking for.

As a **Data Scientist**, I want to use filters for the type of data, the publisher, the region, so I can find the dataset that I am looking for.

As a **Data Scientist**, I want to see details about the publishers so I can learn more information about the organization that created and/or uploaded the dataset.

As a **Data Scientist**, I want to see featured datasets so I can quickly navigate and see details about the dataset.

As a **Data Scientist**, I want to see the details about the data marketplace so I can understand who is providing the service and how it works.

As a **Data Scientist**, I want to see a documentation and test the data API for a specific period of time (e.g. 6 weeks trial) so I can verify that the data has all the specific fields that I need.

As a **Researcher** working on Food safety modeling, I want to share a dataset that I have created so also other researchers can download it.

As a **Data Provider** of food safety data, I want to share a dataset that my company is producing so I can monetize my data.

As any type of user, I want to view the available plans or license options of the data package provided on the website so I can select the plan or license that best fits my needs

As any type of user, I want to subscribe to a plan or license by adding it to my shopping cart, so I can proceed with buying the subscription

As any type of user, I want to proceed to the checkout page to enter my information and payment details, so I can complete my payment

As any type of user, I want to be added to the e-shop as a customer, so I can get customer support in the future

As any type of user, I want to access the Food alerts, Food recalls or Outbreak reports, so I can do what is required by my job description

As any type of user, I want to access Journal articles, Conference papers or Research datasets, so I can do what is required by my job description

As any type of user, I want create my custom dataset, so I can do what is required by my job description

### Wireframes

The main menu of the marketplace is depicted in the following figure:

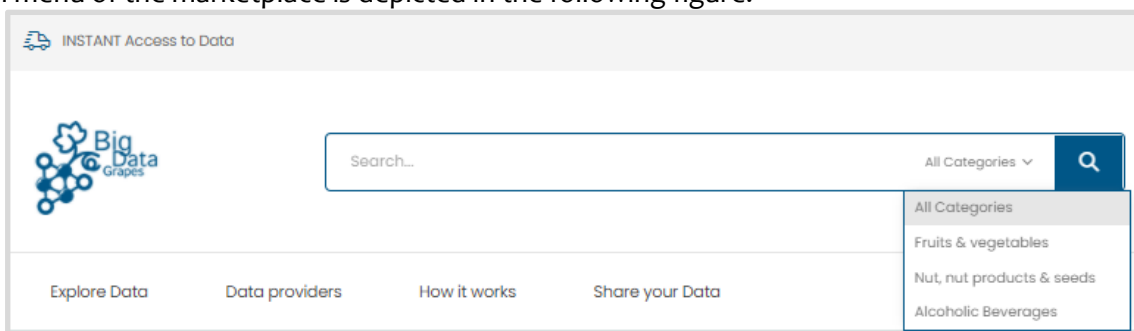


Figure 27: Main menu of the data marketplace

The user is able to perform a free text search of the dataset of interest and select one of the existing dataset categories.

He is also able to make one of the following selections:

- Explore Data, to start searching for the data of interest
- Data providers, to see who are the providers that contribute to the marketplace
- How it works, for further information regarding the marketplace’s functionality
- Share your data, for those who would like to express their interest of becoming data providers.

The information regarding how the marketplace works, is the following:

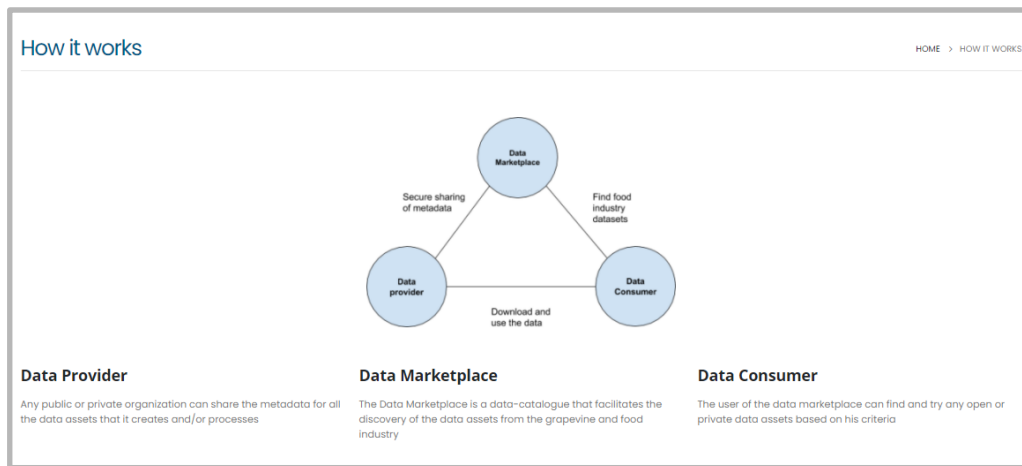
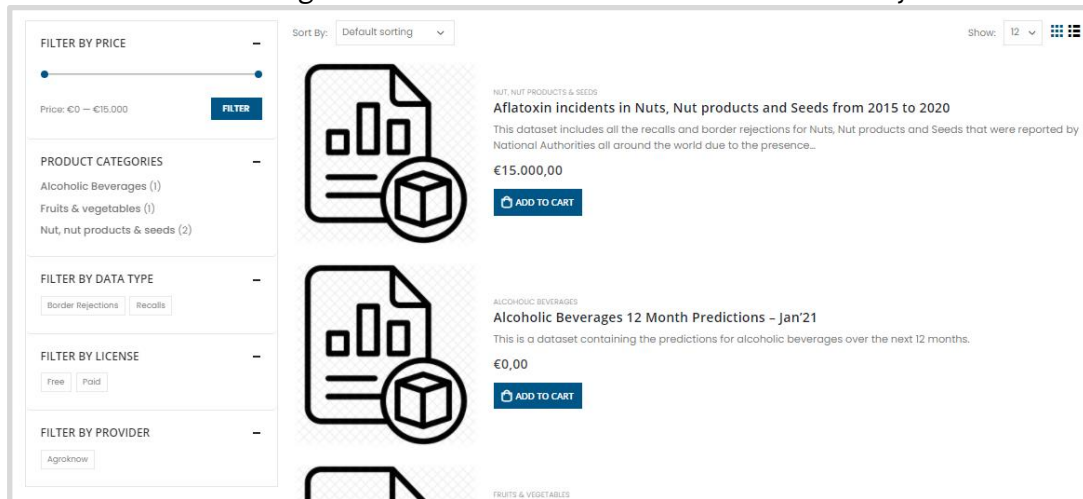


Figure 28: How it works wireframe

The user is able to use the following filters to narrow down his selection and identify the dataset of interest:



**Filter by Price**: Price: €0 – €15,000 [FILTER]

**Product Categories**: Alcoholic Beverages (1), Fruits & vegetables (1), Nut, nut products & seeds (2)

**Filter by Data Type**: Border Rejections, recalls

**Filter by License**: Free, Paid

**Filter by Provider**: Agroknow

**Dataset 1**: **Aflatoxin incidents in Nuts, Nut products and Seeds from 2015 to 2020**  
 This dataset includes all the recalls and border rejections for Nuts, Nut products and Seeds that were reported by National Authorities all around the world due to the presence...  
 €15,000,00 [ADD TO CART]

**Dataset 2**: **Alcoholic Beverages 12 Month Predictions - Jan'21**  
 This is a dataset containing the predictions for alcoholic beverages over the next 12 months.  
 €0,00 [ADD TO CART]

Figure 29: Data discovery wireframe

The information presented for a dataset of interest is:

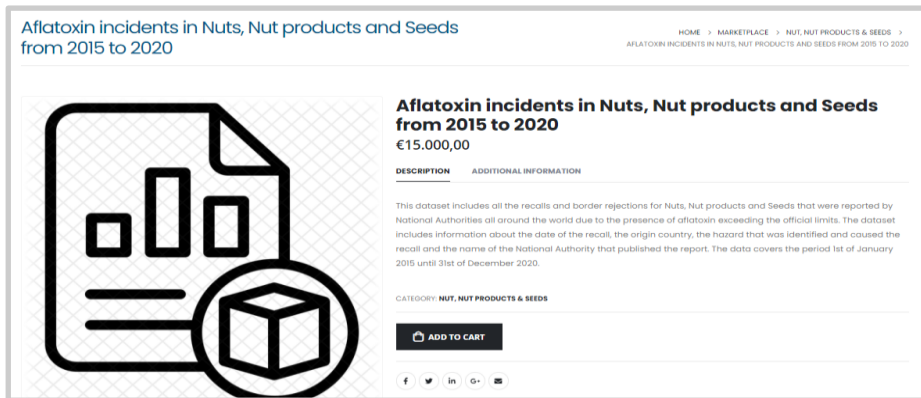


Figure 30: Dataset page wireframe

In case that the user wishes to view additional information for the dataset of interest, the following information is available, as well:

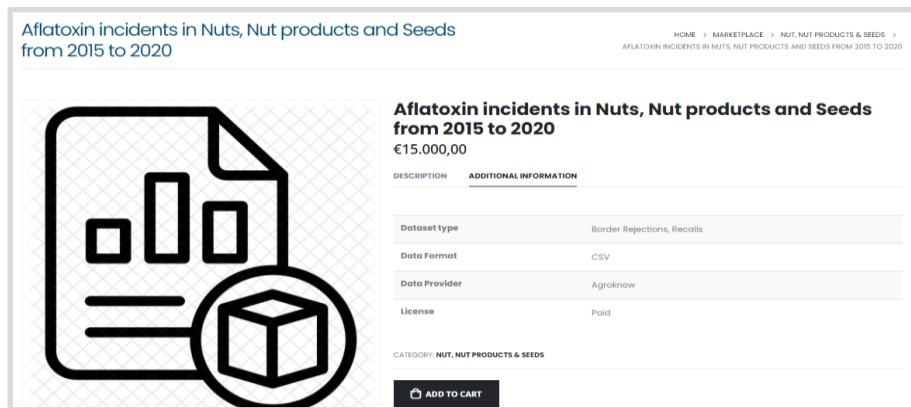


Figure 31: Dataset details wireframe

For users who wish to contribute to the marketplace as data providers, the following webpage is available:

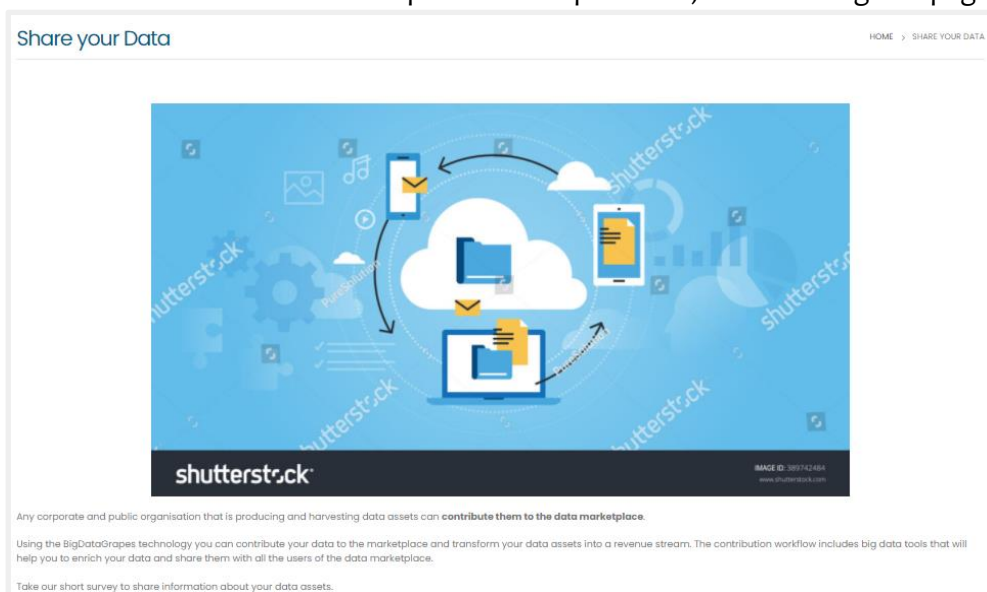


Figure 32: Share your data wireframe

### BigDataGrapes marketplace architecture

The architecture of the marketplace followed the overall BigDataGrapes architectural approach. As presented in the figure below, the data marketplace uses a set of APIs to get the information about the datasets and data providers that the big data platform is managing. A curation tool has been developed to enable the description of data providers and datasets with metadata that will facilitate their discovery in the marketplace. This will allow user to perform a selection based on various filtering criteria and the to select the downloading of the dataset of interest.

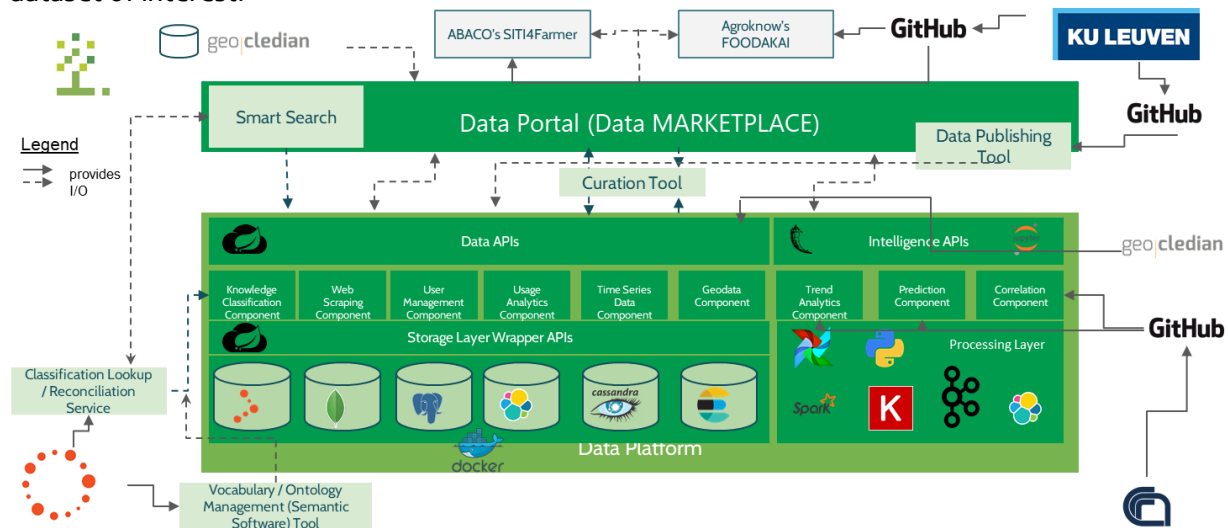


Figure 33: BigDataGrapes technical architecture

A web service API has been developed on top of the data wrapper that has been developed on Big Data Platform's data sets. As soon as the marketplace user has reached the dataset of interest and requests the dataset downloading, a request is sent to the web service API. The result of the call on the particular URL is the creation of a dataset that is then downloadable by the user. For access control and security reasons, and for the marketplace to have access to the API calls, a bearer token in the authentication header of the request is used.

The API was developed using Node.js and Elasticsearch. The development of the marketplace front end was done using an open source content management system and custom operations were developed using php and Javascript.

## 5.4 BigDataGrapes marketplace approach

### How it works

The following figure depicts how the data marketplace is working by connecting the different actors that can benefit from data discovery and data sharing services.

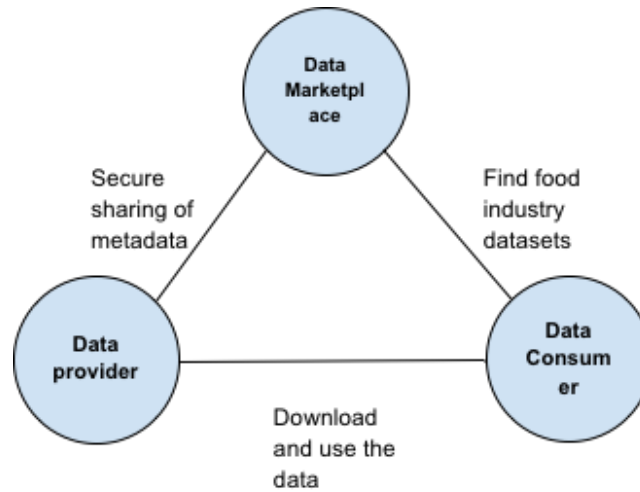


Figure 34: Operation model of data marketplace

- **Data Provider:** Any public or private organization can share the metadata for all the data assets that it creates and/or processes
- **Data Marketplace:** The Data Marketplace is a data-catalogue that facilitates the discovery of the data assets from the grapevine and food industry
- **Data Consumer:** The user of the data marketplace can find and try any open or private data assets based on his criteria.

## 5.5 Data Marketplace Services

The main services of the BigDataGrapes data marketplace are:

- Data catalogue services based on metadata
- Data sharing services
- Data enrichment services
- Access to Open Data and Commercial Data
- Selling data
- Personal data space with favorite datasets and dataset card

### *Data discovery*

The data marketplace provides access to millions of records of different data types. The data assets are updated daily and this means that the volume and the coverage is continuously growing.





Figure 35: Analytics for the data that is available through the data marketplace

The data marketplace aims to promote the usage of the data sets that were generated within the context of BigDataGrapes project. The approach is similar to that of an e-commerce platform, the products of which are the data sets provided, for the time-being, by Agroknow and other data providers of the project.

The marketplace user is able to view information regarding the datasets that are available and the respective data providers. User is also able to define the criteria based on which he wishes to filter the datasets of interest and to view further details regarding a particular dataset. The criteria used are:

- Price of dataset
- Product category
  - o Alcoholic beverages
  - o Fruits & vegetables
  - o Nut, nut products & seeds
  - o ...
- Type of dataset
  - o Border rejections
  - o Recalls
  - o Lab tests
  - o ...
- License
  - o Free
  - o Paid
- Provider
  - o Agroknow
  - o FDA
  - o RASFF
  - o USDA FSIS
  - o ...

For each dataset the following properties are provided:

- Title
- Description
- Dataset type
- Data format
- Data provider
- License
- Sample datasets
- Price

### *Sharing data*

Any corporate and public organisation that is producing and harvesting data assets can **contribute them to the BigDataGrapes data marketplace.**

Using the BigDataGrapes technology any organization can share information about its data and transform the data assets into a revenue stream. The contribution workflow includes big data tools that help the data provider to enrich data and share them with all the users of the data marketplace.

## **5.6 Analysis of data marketplace business models**

With respect to the marketplace's governance, two business models are currently being adopted:

1. Independent platform, where data sets are bought and sold, while fully owned data-as-a-service providers sell primary data in specific segments or with services and solution wraps.
2. Limited ownership hybrids where the marketplace collects and aggregates data from multiple publishers or data owners and then sells the data.

Due to the fact that the business model has not been clearly defined yet, it is considered useful to provide all related parameters.

Marketplace users / buyers may be grouped into three categories: non-commercial users, commercial users and data providers / data sharing users / sellers.

The provided assets / products from the marketplace could be the following:

- Data sets to be downloaded.
- Access to APIs to perform custom queries on available datasets.
- Analytics and statistics.

The pricing options are:

- Free
  - Based on dataset type (i.e. predefined samples)
  - Based on number of records per dataset (i.e. less than 100 records to download)
  - Based on number or download request (i.e. maximum number of dataset downloads: 10)
- One-off offline payment per transaction (after order request submission and offline pricing of order request))
- One-off online payment per transaction (completed transaction using credit card or other online payment method using predefined price catalogue of datasets)

- Subscription-based access. Different subscription types based on:
  - o Number of datasets accessible to download.
  - o Number of total records to download.
  - o Combination of access to datasets and services (analytics and statistics).
  - o Frequency of dataset updates (i.e. 6 months, 3 months, monthly, custom).

Additional features, such as data deduplication, special data format, type of metadata schema, etc.

The final business model for the data marketplace, could be defined using the above parameters. An indicative business model could be similar to the one presented in table 1. This business model was defined based on the feedback that we got from startups, food companies and agrochemical businesses that have tested the Data API that was developed in the context of the BigDataGrapes project.

LEVEL	Basic - Free	Plus	Standard	Advanced	Professional
Description	Pre-defined Open Access package	Pre-defined Open Access package	Pre-defined data package (or databases)	Pre-defined data package (or databases)	Custom data packages / subsets
<b>DATA VERSION</b>	OPEN	OPEN	PREMIUM	PREMIUM	PREMIUM
Frequency of updates	6 months	6 months	3 months	Monthly	Custom
De-duplication	No	No	Yes	Yes	Yes
Enrichment with Hazards and Product terms	If originally	If originally	Yes	Yes	Yes
<b>SERVICE FEES (DOWNLOAD / API)</b>	<b>DOWNLOAD</b>	<b>API</b>	<b>DOWNLOAD (XML, XLS, ...)</b>	<b>API</b>	<b>DOWNLOAD / API)</b>
< 100 records	FREE	500€ per month	250	1.000€ per month	Contact us
< 1.000 records	FREE		500€		
< 10.000 records	FREE		5.000€		
< 100.000 records	FREE		10.000€		
< 1.000.000 records	FREE		20.000€		
Support Service Level	No	Monday-Friday 9am-6pm CET, Business day response			

Table 15: Indicative data model for the data marketplace

Best practices from existing business models presented are those of Advaneo and Quandl.

Advaneo is providing the following membership models:

LEVEL	Free	Premium	Small Business	Enterprise
Service level	Community level service	Basic level service	Standard level service	Business level service
Visualization tools	Basic	Some	Some	All
Connectors	IDS-Connector (trial)	Basic IDS-Connector	Trusted IDS-Connector	Trusted+ IDS-Connector
Workbench	15MB	100MB	10GB	Customizable space
Integration	-	Restricted 3 <sup>rd</sup> party	Basic 3 <sup>rd</sup> party	Custom 3 <sup>rd</sup> party
Dashboard	-	-	Yes, fixed	Yes, customizable

Table 16: Example of a Business Model

Quandl provides the following subscription models:

LEVEL	Personal	Academic	Business
Description	Data for personal use only	Data to be used in an academic environment	For a business to access data for a specific, defined use

Table 17: Example of a Business Model

For the first phase of the marketplace development, all users are able to access the available sample datasets for free. The aim is for the marketplace to become gradually aware of the most popular types of services and data required by users / visitors and for users / visitors to become familiar with the assets (current and future) available in the marketplace, as well as with the ability to share / publish / sell their own datasets.

On a later phase, apart from free sample datasets, paid datasets will be available, too. For paid datasets, the user will be able to submit an order request for a particular dataset, after applying the appropriate filters and reading the dataset-related information.

On the final phase of the development, users will be able to perform online transactions by means of credit card and have access directly to downloadable datasets in various formats (XML, CSV, JSON) or even to API, through which they will be able to build their own query.

## 5 CONCLUSIONS

This deliverable aims to provide a public report on the results of the application piloting sessions, in line with the defined experimental protocols and in accordance with the evaluation methodology, providing an overview regarding each of the five pilots' results. It states and explains the development of the pilots, while the implementation and achieved performance of the BDG pilots are assessed and demonstrated.

Evaluation was to be both formative and summative. The former is essentially self-assessment and was carried out by all partners through filling the "Qualitative and Quantitative Evaluation", which consists of a total of five reports that displays the current status of the piloting activities and thus, it is providing tangible results. The summative evaluation involved external as well as internal evaluation in the form of "BigDataGrapes Pilots' Survey".

Throughout the piloting sessions, all five pilots successfully gathered data from their respective experimental sites. The individual reports have been analysed and the results have shown that the gross data volume resulted in a total of over ~ 3.5 TB throughout the projects lifetime. More specifically, all pilot partners used more than 90 different data sources to generate almost 70 unique datasets. One of the most popular characterization methods of Big Data is the "3V", representing Volume, Variety and Velocity of data generated respectively. From the pilots' data, it becomes obvious that out of the 3 "V"s, the Variety aspect has met the sufficient requirements, along with extended efforts to also cover sufficiently Volume and Velocity. The KPIs list was continuously updated during the project's lifetime.

The BDG consortium selected and defined eight (8) main Software Demonstration scenarios, reflecting the work that has been done in the five pilots, to focus on, fine-tune and showcase. These scenarios were selected to show how different software tools and components produced by the BDG project, together with the critical business decisions to be supported, relevant data and data sources, intelligence and data competence questions to be answered, and algorithm implementations, may support industrial end-users and other key stakeholders in the grapevine-powered industry in new innovative ways.

Following the concept of gradual extension of functionality, intended audience and assessment of this scheme, the pilots interacted with the community and the pilot evaluators accordingly. Thus, this survey was distributed to all relevant stakeholders involved in the BigDataGrapes piloting activities. Feedback was asked from the end-user, with a focus on the "Industry End-User". Representatives from 46 Industries/organisations participated with a total number of 83 Survey completed!

All the above drive us to the development of the **BigDataGrapes platform** and the **BigDataGrapes Data Market Place**.

The **BigDataGrapes Platform** employs the necessary components for carrying out typical analytics processes, making sure that the execution environment and methodology retain scalability and efficient use of computational resources. Additionally, the platform consists of inference and machine learning methods to produce advanced predictive analytics over the entirety of the available data pool. Finally, BigDataGrapes platform leverages the value elicited from these data, by translating them to intuitive and actionable knowledge and using them as the foundation for decision support in complex environments.

A critical goal for the overall system is to ensure that the relevant data sources are semantically annotated and integrated as parts of a common data pool comprising disparate yet interconnected data assets. To this end,



BigDataGrapes designs and develops methods and components for the semantic linking and enrichment of the available data and furthermore, makes available to the domain experts' tools for annotating and describing their data in order to be incorporated in the BigDataGrapes pool.

The **BigDataGrapes Data Market Place** is a large-scale, multifaceted marketplace for grapevine-related data assets, increasing the competitive advantage of companies that serve with IT solutions these sectors and helping companies and organisations evolve methods, standards and processes to help them achieve free, interoperable and secure flow of their data. The development of the data marketplace provides the necessary proof in action that grapevine-powered data assets are shared and exchanged in interoperable formats and versions, by companies and organisations responsible for them. It triggers the facilitation of free, interoperable and secure data flows, as well as the adoption, implementation and revision of data standards.

To conclude, this report clearly proves that BDG project has a disruptive innovative potential to bring new and market driven ICT technologies into the grapevine-powered industries.

## ANNEX A BIGDATAGRAPES PLATFORM BEST PRACTICES

### A.1 SETTING UP AN INSTANCE OF THE BIG DATA PLATFORM

All the necessary components to set up an instance of the Big Data Platform are available in the projects Docker hub and can be deployed easily at any infrastructure. The use of dockers is straightforward so in this section, we are describing how one can set up and deploy an instance of the Big Data Platform for a specific use case: collecting and processing food safety incidents that are announced by National Authorities all around the world.

In this case we need to **crawl** and **scrape** various data sources announcing **food recalls** and **border rejections** worldwide. This data may come in various formats; just to give a quick overview: multilingual PDF files, custom formatted Excel files, RSS feeds and of course HTML pages.

How to crawl these data? You can use a tool like the [Crawler4J](#), but we need to customize it a bit in order to cover our needs. Thankfully it's an open source project so that was easy. Currently all it needs is a **yml configuration file** and it takes care of the rest.

```

seeds:
  - https://www.inspection.gc.ca/food-recall-warnings-and-allergy-alerts/eng/1351519587174/1351519588221
  #these are connected through OR
patterns:
  - inspection.gc.ca/about-the-cfia/newsroom/food-recall-warnings/complete-listing
  - inspection.gc.ca/food-recall-warnings-and-allergy-alerts/
neg_patterns:
  - complete-listing/eng/
  - ay=
  - print=
#these are connected through AND
stored:
  - inspection.gc.ca/food-recall-warnings-and-allergy-alerts/
output:
  action: mongo
  belongs: cfia

```

Figure 36: A (canadian) example of the YML configuration expected by our crawlers

Ok, we got the data, but we need to store it. We have a wide variety of data, each with its own properties. Data concerning food recalls that have a velocity of at most under a hundred per day and data coming from sensors deployed on a field level which presented a velocity of thousands per day; no one framework would be able to meet our needs. So we decided to split the data based on its velocity. Those presenting lower velocity (eg. raw html, xls) would be stored into a [MongoDB](#) instance and those with a higher one into an [Apache Cassandra](#) cluster.

But data is useless if you cannot process it, so the next part of our stack is our **Transformer** into our internal schema; of course based on the entity type we are processing. To that end, we employed into our stack python

and PHP scripts as well as a custom Java project, all of which take care of the harmonisation of the collected data.

The next step is how to identify important terms in the collected data like the ingredient that was recalled, the reason behind a recall or the company involved in it. This is where data mining, NLP, NER, ML and DL techniques are employed. A number of projects and respective API endpoints are deployed taking care of this tasks. As far as technologies and frameworks are concerned, we have **Spring {Boot, Data}** projects, **Flask** endpoints taking advantage of **scikit** and **Keras** classifiers all communicating with **Elasticsearch** instances and internally trained models. Each producing an accuracy score, that if above a threshold is accepted as valid response.



Figure 37: State of the data platform entities as taken from our internal Kibana dashboard instance

The collected data is now harmonised and enriched. If we want to enable human curation we can store the data into an internal CMS (**Drupal**, CKAN, DKAN) in order for it to be easily accessible by our internal food expert team to review, correct and approve for publishing.

And the collected, automatically enriched and human curated data are ready to be published over to our production instances of **Elasticsearch**, ready to be queried by our custom developed **Smart Search API**, visualized over to our application layer.

Such an instance of the Big Data Platform, includes:

- **131 different API endpoints** wrapped on top of each component of the stack;
- over the past year, **14.809.924 requests** have been served by our endpoints;
- with an average response time of **200ms**;
- **9 Elasticsearch** instances;
- **2 Apache Cassandra** nodes;
- **1 MongoDB**;
- **3 Graph Databases** (2 [Neo4j](#) instances and 1 [GraphDB](#)).

All of the above stats were generated by the [Elastic Stack](#), dedicated to monitoring our whole infrastructure.

## A.2 ID ASSIGNMENT IN BIG DATA PROJECTS

Using the Big Data Software stack developed in the context of BigDataGrapes you may build (Big) Data Platform instances. One very important aspect that you need to take into consideration is **“How to assign IDs at records in the Data Platform?”**



Now let's take a step back; this may seem like a trivial task for those out there dealing with relational DBs or rather traditional architectures and platforms. A simple **object = new Object()** or **findById** may suffice.

- But what about complex platforms?
- Data platforms receiving data from various sources?
- Data stored in any kind of storage engines or scraped one-off from another platform?

How can one be certain that everything will be correctly matched throughout a pipeline? The answer is that **special care should be taken as far as id assignment is concerned.**

Many practises may be applied to tackle this problem:

- applying a **hash function** over crawled/scraped urls,
- some kind of **internal identification** process,
- attempting to **identify/extract each source's unique identification** method (everyone has one!).

Let's review each of the above.

### **Hash function over crawled urls**

This is a somewhat safe approach; urls are unique throughout the web so chances are a hash function on top can prove to be successful. It however does not come without any drawbacks.

*What if there are updates to the content crawled?*

It is not uncommon for urls of websites to be generated based on the title of the source. It is the piece of text containing the most important information on the generated content; and the most SEO friendly one.

**So what about updates to the titles?** This can lead to updates to the url as well. So even though that is a rather straight-forward choice, special care should be taken to such updates in order to avoid duplicates.

### **Internal Identification Process**

Time for the another approach; an internal identification process. This can be implemented either deploying an API endpoint responsible for assigning an ID to each resource collected (if your architecture follows the microservice one), or a simple method/function/bash script if you follow a monolithic approach.

The above suggested method has some very important pros; most important of them being its **blackbox way of working**. Once it has been perfected, you no longer have to worry about duplicates in your platform or assigning the same ID to 2 different resources.

*But what about cons?*

First and foremost, **time should be spent perfecting** such a mechanism. We cannot stress enough the important of ID assignment in (Big) Data Projects/Platforms, so you should definitely allow many hours (or story points) to such a project/task since it will be the backbone of pretty much everything you build.

Another drawback we should point out is the **rationale behind the identification** process. Basing it uniquely on the collected content can lead to duplicates as described in the previous case. Having some kind of complex process involving various factors (possibly differentiating based on the collected source) may prove more suitable.

### **Remote Source Identification**

Let's switch our attention to the most challenging choice available. Time to attempt to identify the collected source's ID assignment method.

*Why is it a challenging one?*

It is because it requires knowledge of the remote source's tech stack. Although one may think of this trivial if the data collected is in an xls or csv format where identification is rather straight-forward what if a CMS is employed?

**Knowledge of it should be present** if one wants to successfully assign a unique ID able to avoid duplicates. For instance Drupal assigns a unique id to each piece of content (*nid*) always present in meta tags and by-default in CSS classes of article tags.

However not everything is a drawback for this method! If employed correctly one should never worry for her ID assignment; or almost never. Care should be taken only when **some major migration takes place** on the remote source's side, a rather infrequent case.

This concludes our analysis over various methods that can be employed as far as ID assignment in (Big) Data Platforms is concerned.

All of the above have pros and cons as is the case with everything out there. Similarly to every choice one has to make, you should weight these pros and cons.

**BigDataGrapes suggested approach?**

Apply some kind of hybrid approach, taking advantage of pros from various methods. It is what we have deployed so far in our platform and seems to work well.

The most important is to know your data and your sources. Keeping such knowledge in mind can prove crucial when assigning a unique identification method.

**A.3 ORCHESTRATING ETL PIPELINES**

As already described in 4.1 for the case of the Food Protection Pilot, we collect, translate and enrich global food safety data. This data covers:

- **food recalls** and **border rejections**,
- **price data** on agricultural commodities and animal products,
- **news items** related to food safety,
- **fraud cases**,
- **laboratory testing** performed by Food Safety Authorities worldwide,
- **inspections** and **warning letters** on food companies’ plants and premises,
- **country level indicators** concerning food safety.

A heatmap for our daily cronjobs generated using [Cron Heatmap](#)

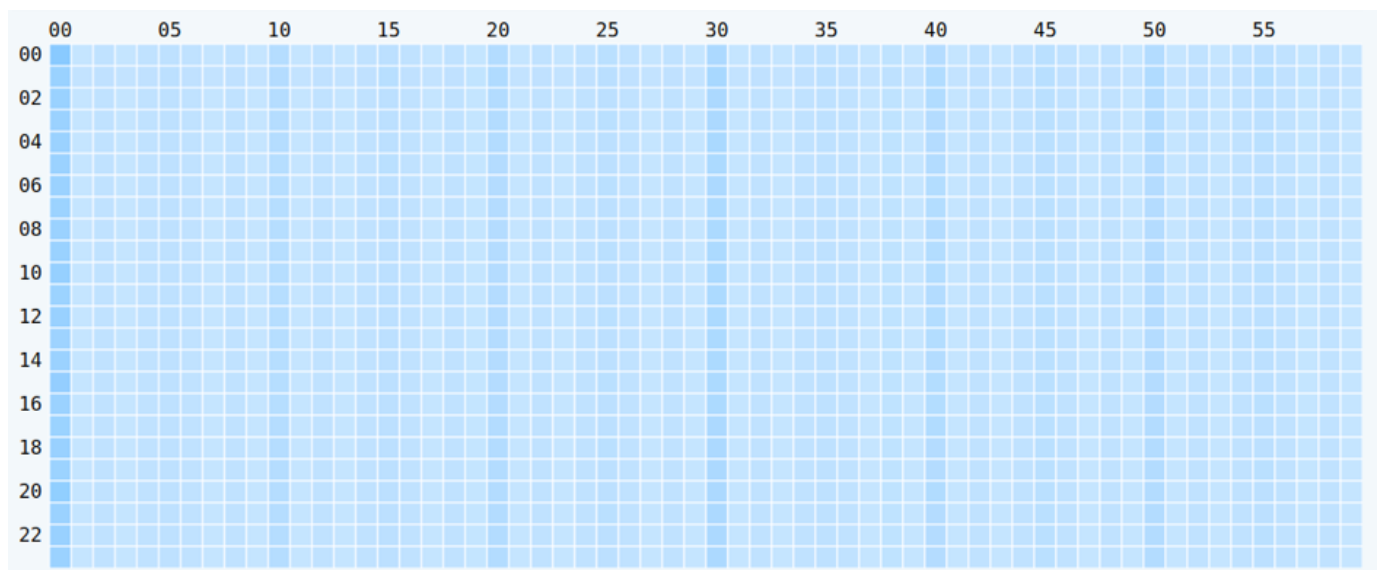


Figure 38: Heatmap for our daily cronjobs generated using Cron Heatmap



As you can imagine though, a number of workflows are involved in the process. Tasks triggering one another, signifying the collection, processing, enrichment of each of the close to 200M (taking into account the hierarchical model employed) data points that are present in our infrastructure.

There is a very important challenge that we need to take into account in such a Big Data Platform instance; **that of the overall orchestration.**

How did we tackle this challenge and what tools did we enlist for help?

How can we synchronize all these flows?

Back in 2018, when we first started implementing and deploying the Big Data Platform stack, cronjobs were an initial choice. Unfortunately popular choices like [Apache Airflow](#) and [Spotify's Luigi](#) had not gained that attention at the time!

What did we do? Bash scripts and crontab -e commands. Every source we track has its dedicated directory in our backend/processing servers and within each of these directories lies a run.sh script. This is the script that manages all the action. Every single task in each workflow triggered is managed by such a script, calling other scripts created with the responsibility to handle each task.

And this **run.sh** is triggered by crontab. For many of you accustomed with cronjobs, execution space of each cronjob may come as natural; we had to learn it the hard way. A base sample of the run.sh scripts that can be used.

```
#!/bin/bash

curr_dir="$( cd "$( dirname "${BASH_SOURCE[0]}" )" && pwd )"

cd ${curr_dir}

contents=$(cat ${curr_dir}/sync.lock)

if [ "$contents" != "o" ];then

    exit

fi

echo "1" > ${curr_dir}/sync.lock

# main code goes here

echo "o" > ${curr_dir}/sync.lock
```

The above depicted lines of code accompany every single script existent in our (dedicated) servers.

- The first thing we need to do is switch over to the directory of our to-be-crawled source. (lines: 2–3)
- Then we need to check if the previously triggered script has completed its work, this is done by checking the respective lockfile. (lines: 5–10)
- Once we are done with the main work we should update the lockfile for the next trigger of our script. (line: 14)

Depending on the source, **translation** endpoint triggering scripts may be present. **Text mining** or **text classification** workflows may take place with their respective scripts. All initiating calls to the respective projects and endpoints.

Say we are done with the collection and processing of each data record, we now have to let our **internal CMS** (Drupal) know of the new data. Time to sftp over there and upload the respective transformed and enriched records. That system will take care of the rest.

```
{  
  "_id": ObjectId("5df60da920252e02c73c94bf"),  
  "url": "https://www.inspection.gc.ca/about-the-cfia/newsroom/food-recall-warnings/complete-listing/2019-12-08/eng/1575846136193/1575846136974",  
  "belongs": "cfia",  
  "accessed": NumberLong(1576406528),  
  "html_size": 23584,  
  "text_size": 6093  
}
```

Is this enough? You have most probably already guessed it. No. **What about data we have already processed?**

We should not stress our (already working at maximum capacity) servers any more than they have to; only new data needs to be taken into account. This is where our **MongoDB** kicks in; the place where all the raw data is stored, along with a collected on timestamp and a flag signifying whether or not a record has already been processed.

Sample of the metadata stored for each resource in our MongoDB instance

Is this enough? Again the (obvious) answer is no. **What about firewall limitations, fail2ban or overall crawler traffic restrictions?**

Although it would make our life way easier, firing up crawlers every minute towards each of the sources tracked regardless of the publishing rate may prove fatal in our endeavor. We need to configure our **ETL workflows** to be triggered only when chances are new data are present.

This although easily configurable through cron expressions requires some manual labor.

We need to dive into our data and identify the rate at which new records are published. Only then can we define an acceptable rate at which we can dispatch our workflows.

#### **What about hardware limitations?**

This is the most tricky question of all. Implementing and deploying a workflow capable of executing regardless the stress levels of a server is really challenging. Our choice at this point was splitting our workflow into **atomic operations** this ensures that even though a task or a workflow may not complete, no data loss will be observed since each new workflow triggered will always check for previous workflows' leftovers.

We should also consider some additional aspects:

- what about CPU/RAM intensive tasks?
- error logging?
- tools out there (Apache Airflow for instance!) that can make our life easier?

Since we are currently switching to **Apache Airflow** for our ETL pipelines, however let us stress at this point that all of the above are crucial. One cannot have a robust ETL pipeline if the above remarks are not addressed. And always researching and exploring new technologies and frameworks out there should be present in day to day tasks!

Just to give a quick overview in terms of numbers, in the current infrastructure of the Big Data Platform:

- **113 ETL workflow cronjobs** are present;
- on average workflows are triggered **once every 10 minutes**;
- **9 dedicated servers** are involved in this part of the infrastructure;
- **11 workflow jobs** have been switched to **Apache Airflow DAGs**;
- **1 Elastic Stack** instance (involving [Elasticsearch](#) & [Kibana](#) & [Metricbeat](#)) is employed to keep track of the health of our infrastructure.

## A.4 AUTOMATING DATA ENGINEERING TASKS IN A BIG DATA PLATFORM

For every data engineer that is responsible for the operation and maintenance of a Big Data Platform instance there are 2 categories tasks:

- tasks that are challenging, meaning interesting, and
- tasks that are somewhat trivial.

To start with, **every task starts in the first category** and ends up in the second. Every single task each individual out there handles for the first time is a challenging one. One needs to perform it over and over again for it to fall under the category of the trivial tasks.

A lot of time is spend to conduct many times that same tasks such as:

- create a new ETL workflow
- performing the same operations
- calling the same endpoints
- saving results to a file system directory or to a storage engine
- moving data from one server to another

And there are many more questions where that came from and many ways to tackle each task. One can always click on Sublime, IntelliJ, PyCharm or whatever editor you favour and click on create new project.

Nothing compares with the excitement of a blank page! This is true only for **tasks never handled before though**. For all the rest you mainly copy-paste stuff from previous projects or Stack Overflow.

And although Stack Overflow will (hopefully) always be there for you to copy-paste, if you turn to previously implemented projects too often it may mean that there is some room for automation to take place.

- **Why copy-paste** the same code over and over again when you can *create a new API endpoint* to handle your requests?
- **Why manually** triggering a workflow when [ETL pipelines](#) are out there?
- **Why (s)ftp uploading** stuff with Filezilla when you can *write a script* to do that?
- **Why having reminders** to your calendar for backup ops when one or more *scripts can take care* of that?
- **Why manually** adding cronjobs when you can automatically *generate a DAG file* and put it in your [Apache Airflow](#) instance?

And although the above depicted list can most probably go on forever, there is a key take away message here: This can be fully automated.

- **A new project is always fun** to build and can help packing the same ops under one roof.
- **Cronjobs and crontabs** are there for you; all you need to do is create the script and pick the desired frequency.
- **Upload files & directories scripts** are easily writable and do not include the possibility of something going wrong.
- You may have more than 10 different storage engines; taking the time to write a **backup script** for each new instance means you never have to worry about it again, no small thing!
- **ETL workflow frameworks** can make your life way (way) easier, why not employ them?

If one attempts to always keep in mind the reusability and automation of processes then the result can be amazing. *Both in terms of code quality and component reusability, as well as in terms of robustness and bugless components.*

## **A.5 CONTAINER ORCHESTRATING AND DEPLOYMENT BEST PRACTICES FOR THE BIGDATAGRAPES GRAPHDB INSTANCE**

In cases where a microservice architecture is used with multiple services having different functionalities and system requirements, the industry standard best practices suggest the use of a container orchestration system. In the case of BigDataGrapes, we have selected Kubernetes for its wide use in the field, comprehensive documentation, multiple plugins and all of the required functionalities. Following this approach, services are packed as docker images and can be automatically deployed from a docker repository. In order for Kubernetes to be able to monitor all deployed software and take action to restore services back to their operational state if needed, health check endpoints have been implemented.

Best practices suggest maintaining separate environments for development and production. Before services are released and deployed, they should pass through automated deployment and integration tests. A continuous integration system is used for this purpose, which in the case of BigDataGrapes is Jenkins. Once a newly developed software component is committed, the automated tests would run and in case of a success, the CI system would deploy the new artefacts.

For deployment, the de facto standard script for Kubernetes is Helm. This allows for easy deployment while allowing the deployment process to remain predictable and reliable.



## ANNEX B GACoVI USAGE INSTRUCTIONS

A demo of GaCoVi is available at: <http://picasso.experiments.cs.kuleuven.be:3604/>. The source code of the system has been uploaded to the GitHub repository of BigDataGrapes group: <https://github.com/BigDataGrapes-EU/d5.3-gacovi>. Please follow the following steps to download the source code and run locally.

Step 1. Install and run **Docker** (download Docker at <https://docs.docker.com/get-docker/>)

Step 2. Download or clone the project:

```
$ git clone https://github.com/BigDataGrapes-EU/d5.3-gacovi.git
```

Step 3. Navigate to the cloned/downloaded folder:

```
$ cd d5.3-gacovi
```

Step 4. Build the Docker image:

```
$ docker build -t d5.3-gacovi .
```

Step 5. Run the image as a Docker container:

```
$ docker run -dit --name d5.3-gacovi -p 8000:80 d5.3-gacovi
```

Step 6. Open <http://localhost:8000> in browser