HORIZON 2020 - ICT-14-2016-1

AEGIS

Advanced Big Data Value Chains for Public Safety and Personal Security

WP5 – AEGIS Data Value Chain
Early Community Demonstrators

D5.6 – Final Evaluation, Impact Assessment and Adoption Guidelines

Version 1.0

Due date: 30.06.2019 **Delivery Date**: 29.06.2019

Author(s): Dimitrios Miltiadou, Konstantinos Perakis, Stamatis Pitsios (UBITECH), Marios Phinikettos (SUITE5), Cesar Lisi, Alessandro Testa (HDI), Alexander Stocker, Christian Kaiser (VIF), Elisa Rossi (GFT), Anna Maria Hounti. Evgenia Alexopoulou, Fotios Manesis (KONKAT), Alexandru Ormenisan, Mahmoud Ismail (KTH), Fabian Kirstein (Fraunhofer)

Editor: Dimitrios Miltiadou (UBITECH)

Lead Beneficiary of Deliverable: UBITECH

Dissemination level: Public **Nature of the Deliverable:** Report

Internal Reviewers: Gianluigi Viscusi (EPFL), Alessandro Testa (HDI)

EXPLANATIONS FOR FRONTPAGE

Author(s): Name(s) of the person(s) having generated the Foreground respectively having written the content of the report/document. In case the report is a summary of Foreground generated by other individuals, the latter have to be indicated by name and partner whose employees he/she is. List them alphabetically.

Editor: Only one. As formal editorial name only one main author as responsible quality manager in case of written reports: Name the person and the name of the partner whose employee the Editor is. For the avoidance of doubt, editing only does not qualify for generating Foreground; however, an individual may be an Author – if he has generated the Foreground - as well as an Editor – if he also edits the report on its own Foreground.

Lead Beneficiary of Deliverable: Only one. Identifies name of the partner that is responsible for the Deliverable according to the AEGIS DOW. The lead beneficiary partner should be listed on the frontpage as Authors and Partner. If not, that would require an explanation.

Internal Reviewers: These should be a minimum of two persons. They should not belong to the authors. They should be any employees of the remaining partners of the consortium, not directly involved in that deliverable, but should be competent in reviewing the content of the deliverable. Typically this review includes: Identifying typos, Identifying syntax & other grammatical errors, Altering content, Adding or deleting content.

AEGIS KEY FACTS

Topic:	ICT-14-2016 - Big Data PPP: cross-sectorial and cross-lingual data integration and experimentation
Type of Action:	Innovation Action
Project start:	1 January 2017
Duration:	30 months from **01.01.2017** to **30.06.2019** (Article 3 GA)
Project Coordinator:	Fraunhofer
Consortium:	10 organizations from 8 EU member states

AEGIS PARTNERS

Fraunhofer	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.
GFT	GFT Italia SRL
KTH	Kungliga Tekniska högskolan
UBITECH	UBITECH Limited
VIF	Kompetenzzentrum - Das virtuelle Fahrzeug , Forschungsgesellschaft-GmbH
NTUA	National Technical University of Athens – NTUA
EPFL	École polytechnique fédérale de Lausanne
SUITE5	SUITE5 Limited
KONKAT	ANONYMOS ETAIREIA KATASKEVON-TECHNIKON ERGON, EMPORIKON, VIOMICHANIKONKAI NAUTILIAKON EPICHEIRISEON KON'KAT
HDIA	HDI Assicurazioni S.P.A

Disclaimer: AEGIS is a project co-funded by the European Commission under the Horizon 2020 Programme (H2020-ICT-2016) under Grant Agreement No. 732189 and is contributing to the BDV-PPP of the European Commission.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Executive Summary

The scope of deliverable D5.6, which is the final deliverable of WP5, is to document the efforts undertaken within the context of the tasks 5.3, 5.4, 5.5, and 5.6. Towards this end, the deliverable builds on top of the works and outcomes of all deliverables of WP5, namely D5.1, D5.2, D5.3, D5.4 and D5.5, in order to report the final evaluation of both the AEGIS demonstrators and the AEGIS platform taking into consideration of all the activities that were performed from M7 till M30 of the project. The final evaluation presents the summary of the key outcomes of the holistic quantitative and qualitative evaluation that performed during the project, as dictated by the AEGIS evaluation framework.

Within the context of the deliverable D5.6, an overview of all the demonstrators activities is documented, presenting all the detailed information with regards to the execution scenarios, the demonstrators' implementation phases, the demonstrators' execution, and the demonstrators' evaluation. For each of the three demonstrators, a detailed overview is presented, highlighting all the demonstrators' activities during the project, the digital artefacts that were implemented within the context of each demonstrator, as well as the achievements and the lessons learnt from each demonstrator.

Following the final evaluation of the AEGIS demonstrators, the deliverable also presents the final evaluation of the AEGIS platform. The deliverable presents the aggregated results of all three quantitative and qualitative evaluations that were performed during the project period, as well as the description of the insights that were extracted from this holistic evaluation. In addition to quantitative and qualitative results, the deliverable discusses the security assessment of the final version of the platform that was performed in order to evaluate the security mechanisms of the platform.

Finally, the deliverable presents the AEGIS platform documentation and adoption guidelines. In detail, a comprehensive documentation of the AEGIS platform usage is documented, providing useful insights in all the functionalities of the platform and the necessary guidelines for the development activities that can be performed in the platform. Following the practical guidelines, the deliverable presents the AEGIS platform's adoption guidelines that are based on the knowledge obtained from the partners involved during the implementation phase of the demonstrators. The guidelines are presented in order to be exploited from the potential users of the platform beyond the project consortium.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AAL	Active and Assisted Living
API	Application Programming Interface
CO	Confidential, only for members of the Consortium (including the Commission Services)
CPU	Central Processing Unit
CSP	Care Service Provider
CSV	Comma Separated Values
D	Deliverable
DoW	Description of Work
H2020	Horizon 2020 Programme
GUI	Graphical User Interface
HVAC	Heating Ventilation and Air Conditioning
IT	Information Technology
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
PSPS	Public Safety and Personal Security
R	Report
RTD	Research and Development
SHAL	Smart Home and Assisted Living
UI	User Interface
VOC	Volatile Organic Compounds
XML	Extensible Markup Language
WP	Work Package
Y2	Year 2

1. Introduction

The scope of this section is to introduce the deliverable and familiarise the reader with its contents. To this end, the current section summarises the objective of the deliverable, its relationship with the other work packages and corresponding deliverables and analyses its structure.

1.1. Objective of the deliverable

The scope of deliverable D5.6, which is the final deliverable of WP5, is to document the efforts undertaken within the context of the tasks 5.3, 5.4, 5.5, and 5.6. Within the context of this deliverable, both the AEGIS demonstrators' final evaluation and the AEGIS platform's final evaluation are presented, summarizing the most relevant aspects of the holistic quantitative and qualitative evaluation that was conducted in the course of the project as an overall assessment of the AEGIS platform. Additionally, it presents the results of a security assessment conducted with the final version of the AEGIS platform.

Furthermore, the deliverable includes for each of the three demonstrators (Automotive, Smart Home & Assisted Living, and Insurance demonstrators), an overview of all demonstrators' activities that were performed during the project. In this overview, the digital artefacts developed within the context of the demonstrator implementation are presented, a summary of the most relevant information regarding each demonstrator execution scenarios and implementation phases is documented, as well as the lessons learned during the comprehensive demonstrator implementation phase on the AEGIS platform.

Furthermore, the deliverable presents the AEGIS platform documentation and adoption guidelines as practical guidelines on how to exploit the value of the AEGIS platform in the best possible way. It includes a comprehensive documentation of the AEGIS platform usage, as well as a guideline on how to develop your own data driven applications with the AEGIS platform. This information was compiled considering the vast knowledge of the data scientists within the AEGIS project implementing the three demonstrators on the AEGIS platform, which is to be transferred to potential users beyond the project consortium.

1.2. Insights from other tasks and deliverables

The deliverable builds on top of the work reported in WP5. In particular, the previous outcomes of the work performed in WP5, as reported in D5.1, D5.2, D5.3, D5.4, and D5.5 provided the AEGIS evaluation framework, as well as the methodology on how to implement the framework during the implementation phase of the three demonstrators of the AEGIS project.

The outcomes of the deliverable D5.2 served as guidance on how the evaluation of both the AEGIS platform and the AEGIS demonstrators will be performed. The AEGIS platform evaluation plan, as well as the scenarios defined for each demonstrator and the documented evaluation plan for each demonstrator have driven the assessment performed during all versions of the demonstrators.

The outcomes of the deliverables D5.3, D5.4, and D5.5, documenting the evaluation of the first (early), second (medium), and third (final) versions of the demonstrators, provided the baseline for D5.6.

1.3. Structure

Deliverable D5.6 is organised in five main sections as indicated in the table of contents:

- The first section introduces the deliverable. It documents the scope of the deliverable and briefly describes how the document is structured. It also documents the relation of the current deliverable with the other deliverables, and how the outcomes of other deliverables are received as input to the current deliverable.
- Following the introductory section, the second section provides a summarized result of the AEGIS demonstrators' final evaluation, covering both the quantitative and the qualitative evaluation. It provides an overview of each implemented demonstrator along with the activities performed during the project and the artefacts developed.
- The third section summarizes the results of the evaluation of the AEGIS platform that was conducted during the project covering both the quantitative and the qualitative evaluation. The section presents the aggregated results of all three evaluations that were performed during the AEGIS project execution phase as an overall assessment of the AEGIS platform as well as the insights that were extracted. Additionally, it presents the results of a security assessment which was conducted for the final version of the AEGIS platform.
- Section four is dedicated to present the AEGIS platform documentation and adoption guidelines and presents the general usage of the AEGIS platform, as well as the usage of the different AEGIS Jupyter tools Query Builder, Visualiser and Algorithm Execution Container. Furthermore, it introduces into the use of jobs, model serving (TensorFlow), the Cleaning tool, and the Anonymisation tool. Finally, this section provides the AEGIS platform adoption guidelines as a practical how-to for creating own data science applications on the AEGIS platform.
- Section 5 concludes the deliverable. It outlines the main findings of the deliverable, which will guide the future research and technological efforts of the consortium.

2. AEGIS Demonstrators Final Evaluation

2.1. Demonstrators overview

In this section, an overview of the demonstrator activities that were performed during the project is provided. It includes information with regards to the execution scenarios, the demonstrators' implementation phases, the demonstrators' execution, and the demonstrators' evaluation (interviews, focus groups, workshops).

In the following table, the demonstrator-specific execution scenarios and required functionalities that were implemented within each demonstrator are summarized. In total, **12 scenarios** were **implemented**.

Scenario ID	Scenario	Functionalities	Demonstrator Version
Automotive Demonstrator			
1	Broken Road Indicator	The broken road indicator scenario provides the baseline for the two further scenarios. Data is collected by a data logger developed at VIF. 1. Create project 2. Upload driving data (in bulk) and do minimal pre-processing Raw data files of individual sensors are merged, and all trips contained in the data are extracted. 3. Transform driving data and save results Extracted trips are resampled to a fixed, regular-spaced time grid of 10Hz. Coordinate system of the sensors is aligned with coordinate system of the vehicle for each of these trips. 4. Identify road damage and save results (An artificial event-signal is computed. This event-signal has large values in presence of a speedbump/pothole and low values in other situations.) 5. Provide visualisation of road damage	Early
2	Safe Driving Indicator	Additional functionalities required are: 1. Identify safe driving events and save results 2. Assess driving risk and save results Trip-specific and driver-specific risk scores (safe driving scores are computed) Provide visualisation of save driving events	Medium
3	Regional Driving Safety Risk Estimator	Additional functionalities required are: 1. Assess regional driving safety risk with a geographic risk estimator and visualise it in a heat map	Advanced

Smart Home and Assisted Living Demonstrator			
1	(At-risk) Individuals data fetching, processing and classification	Registration of personal devices (for data collection and aggregation of activity and health data), registration of external data streams (events, weather, etc.), data processing, creation of personas based on data classification, definition of generic rules per persona, definition of outlier detection algorithms, generation of simple alerts to (at-risk) individuals.	Early
2	Smart home data monitoring and processing	Registration of smart home data streams, Smart Home Data Monitoring, Pre-processing and normalization.	Early
3	Notifications and alerts for (at-risk) individuals	More information-rich Personas, definition of medical rules for the notification and alert engine, advanced and non-personalized alerts to (at-risk) individuals, personalised tracking of (at-risk) individuals following their consent and simple notifications to carers	Medium
4	Smart home comfort profiling and notifications	Thermal and Visual Comfort Profiling, notification and alert services for adverse indoor environmental conditions	Medium
5	Personalised notifications and recommendations for (at-risk) individuals and their carers	Definition of personalized medical rules, personalised alerts and recommendations to (at-risk) individuals and notifications to informal carers	Advanced
6	Smart home automation services	Optimization and automated control of HVAC and lighting devices	Advanced
Insurance Demonstrator			
1	Financial impact, customer support and services	Definition of the financial impact that an event (weather or socio-political) should have on the company. Identification of the customers that should be involved and customer support service initialisation.	Early
2	Personalised early warning systems for asset protection	Detection of a foreseen event that could impact on the HDI customers, identification of the customers, notification to their Mobile App and to their agent. Personalised offers based on the guarantee hold and on the event.	Medium
3	Business plan and marketing strategy	Business analysis request for analysis on the customers' portfolio in different years in order to define an accurate marketing strategy as well as a successful business plan.	Advanced

Table 2-1: AEGIS Demonstrators execution scenarios

In the following table, all demonstrator specific test cases and their execution states are summarized. All test cases have been successfully executed. In total, **58 test cases** were **executed**. The test cases were presented in detail in the deliverables D5.3, D5.4 and D5.5.

Test case ID	Test case Name	Execution Status	Demonstrator version
Automotive Demonstrator			
Scenario1 - 1.1	Create project	Completed	Early
Scenario1 - 1.2	Upload driving data (in bulk)	Completed	Early
Scenario1 - 1.3	Transform driving data and save results	Completed	Early
Scenario1 - 1.4	Identify road damage and save results	Completed	Early
Scenario1 - 1.5	Provide visualisation of road damage	Completed	Early
Scenario2 - 2.1	Identify safe driving events and save results	Completed	Medium
Scenario2 - 2.2	Assess driving risk and save results	Completed	Medium
Scenario2 - 2.3	Provide visualisation of save driving events	Completed	Medium
Scenario3 - 3.1	Assess regional driving safety risk with a geographic risk estimator and visualise it in a heat map	Completed	Advanced
Smart Home and Assisted Living Demonstrator			
Scenario1 - 1.1	(At-risk) Individuals Profile Building and Devices Registration	Completed	Early
Scenario1 - 1.2	Carers Profile Building	Completed	Early
Scenario1 - 1.3	Linking (at-risk) Individuals with Carers	Completed	Early
Scenario1 - 1.4	CSP services-relevant External Data Sources registration	Completed	Early
Scenario1 - 1.5	“Personas” Building and Clustering	Completed	Early
Scenario1 - 1.6	Identification of Critical External Conditions/Events	Completed	Early
Scenario1 - 1.7	Simple Notification Issuing	Completed	Early
Scenario2 - 2.1	Establishment of smart home monitoring infrastructure and data collection	Completed	Early
Scenario2 - 2.2	Smart home data processing and normalization	Completed	Early
Scenario2 - 2.3	Real-time smart home data monitoring	Completed	Early
Scenario3 - 3.1	Design of persona outlier identification model	Completed	Medium

Scenario3 - 3.2	Design of outlier notification model	Completed	Medium
Scenario3 - 3.3	Execution of Outlier Notification Model	Completed	Medium
Scenario3 - 3.4	Mapping Actions to Notifications for personas of (at-risk) individuals	Completed	Medium
Scenario3 - 3.5	Mapping Actions to Notifications for carers	Completed	Medium
Scenario4 - 4.1	Software development for comfort profiling and notifications	Completed	Medium
Scenario4 - 4.2	Receive alerts regarding uncomfortable or health-endangering conditions	Completed	Medium
Scenario5 - 5.1	Enhanced (at-risk) individuals profile and provision of access to CSPs for personalised notifications	Completed	Advanced
Scenario5 - 5.2	Registration of medical rules	Completed	Advanced
Scenario5 - 5.3	Re-Classification of Individuals	Completed	Advanced
Scenario5 - 5.4	Dynamic Dashboard following Algorithm Execution	Completed	Advanced
Scenario5 - 5.5	Persona and Personalised notifications	Completed	Advanced
Scenario6 - 6.1	Smart home automation recommendation system	Completed	Advanced
Scenario6 - 6.2	Smart home automation implementation	Completed	Advanced
Insurance Demonstrator			
Scenario1 - 1.1	Event Detection tool configuration and training	Completed	Early
Scenario1 - 1.2	Event Detection Tool Streaming	Completed	Early
Scenario1 - 1.3	Create Account and Project	Completed	Early
Scenario1 - 1.4	Anonymised in-house dataset upload	Completed	Early
Scenario1 - 1.5	Identification and visualisation of the possibly involved customers	Completed	Early
Scenario1 - 1.6	Priority list creation	Completed	Early
Scenario1 - 1.7	Priority list (report) de-anonymization and assignment, risk exposure evaluation	Completed	Early
Scenario1 - 1.8	Priority list (report) de-anonymization and assignment	Completed	Early
Scenario2 - 2.1	Event Detection tool training (version 2)	Completed	Medium
Scenario2 - 2.2	Event Detection notification configuration	Completed	Medium
Scenario2 - 2.3	Event Detection Tool Streaming (version 2)	Completed	Medium

Scenario2 - 2.4	Create Project	Completed	Medium
Scenario2 - 2.5	Uploading datasets	Completed	Medium
Scenario2 - 2.6	Mobile App data enrichment	Completed	Medium
Scenario2 - 2.7	Identification and visualisation of the possibly involved customers	Completed	Medium
Scenario2 - 2.8	Priority list creation	Completed	Medium
Scenario2 - 2.9	Priority list (report) de-anonymization, sharing and personalised offer	Completed	Medium
Scenario3- 3.1	Business analysis request	Completed	Advanced
Scenario3- 3.2	Create project	Completed	Advanced
Scenario3- 3.3	Uploading anonymised datasets	Completed	Advanced
Scenario3- 3.4	Business Analysis – open datasets search	Completed	Advanced
Scenario3- 3.5	Data preparation with Query Builder	Completed	Advanced
Scenario3- 3.6	Data analysis with Algorithm Execution Container	Completed	Advanced
Scenario3- 3.7	Report visualisation	Completed	Advanced
Scenario3- 3.8	Report sharing	Completed	Advanced

Table 2-2: AEGIS Demonstrators test cases

A series of evaluations have been made during the progress of the demonstrator development. Evaluations concerning the demonstrators (and not the AEGIS platform) are summarized in this section, while evaluations concerning the platform are provided in section 3.

With respect to the automotive demonstrator an evaluation involving a focus group of drivers was conducted after the development of the medium demonstrator, safe-driving indicator, targeting at drivers as beneficiaries. In general, the drivers (who were also providing their driving data for the demonstrator) understood the concept and liked the services provided, visualising trips and driving styles and providing a risk score as a safety benchmark on trip and driver level. They would use such a service to further improve their driving style in order to learn how to driver safer, if it is commercially available. Furthermore, they provided a list of ideas and suggestions on how to further improve the demonstrator including e.g. self-exploration of the data using a mobile application for drivers, increase the number of interactive elements of data exploration and visualisation, provide further contextual information to drivers, benchmark the own driving style against the driving styles of other drivers and detect also speeding events. Many of their ideas and suggestions (e.g. benchmarking of driving styles, mobile driver app) have been implemented in the later phases of the project.

Another evaluation of the automotive demonstrator was conducted involving a focus group of people related to traffic planning after the development of the advanced demonstrator, regional driving safety risk estimator, targeting at such stakeholders as beneficiaries. This evaluation was conducted together with a focus group of persons knowledgeable in traffic planning to generate additional feedback on how the experts experienced the service. Results have shown that in general the information provided to the traffic planning experts in the dashboard and the approach pursued are perceived to be useful to better understand driving risks within urban areas and to further improve interventions to mitigate them. The intense discussion between the experts triggered a plethora of useful ideas, many of them going even beyond the scope of the demonstrator. In general, the experts were interested to get more contextual information about areas of Graz covered by the heatmap and about the underlying trip data. Displaying road use as an (additional) heatmap or route-based visualisation would be a great add-on. Having a context menu leading to further information while marking certain areas of the heatmap would be interesting to better judge the severity of the indicated risks. Traffic flow prediction and visualisation were topics of great interest for the participating experts. Adding a speeding event (besides braking, acceleration and curving/cornering) would be interesting as speeding leads to many risky situations and accidents. Too many events are used in the calculation of the heatmap (based on too sensible set thresholds for event detection, which was required for better testing the AEGIS Platform).

With respect to the smart home and assisted living demonstrator, an internal focus group was organized among data scientists and developers after the development of the medium demonstrator, notifications and alerts for (at-risk) individuals & smart home comfort profiling and notifications, discussing advantages of the service offered, perception of the development of the service application, and the level of refinement of the offered services. One major result was the perceived usefulness of the data scientists that the utilization of the big data analytics platform really provides benefits such as high scalability of services offered or the utilization of components to help overcome difficulties in exploring and visualizing data. The demonstrator development process has so far been relatively free from unexpected issues. The service application workflow was stable and streamlined and the UI experience of the demonstrator was perceived to be good supporting both mobile and web interfaces.

Another evaluation of the smart home and assisted living demonstrator with a focus group of healthcare experts was conducted after the development of the third demonstrator version. These persons were experienced users of (similar) healthcare services and so collecting their unbiased feedback was very valuable. The participants showed a high interest in the pursued approach using the features of the AEGIS platform (especially the notebook feature), as well as on the designed classification model and its interactions with the underlaying rule engine. They acknowledge the persona approach, the classification of the (at-risk) individuals on these personas and the ability of a care service provider to define a set of medical rules on each persona. They also liked the utilisation of wearable and smart home device for a near to real-time monitoring of an individual' well-being and home conditions. The evaluation also led to many new ideas on how to enhance the demonstrator with new features, which will be considered in the exploitation phase. They also foresee a number of challenges to solve such as challenges related to the continuous data retrieval, processing and analysis, extensibility of the data pipeline with new devices, import of the anonymized SHAL database in AEGIS with growing database size, data protection and anonymization, and the inclusion of a 'right to be forgotten'.

With respect to the insurance demonstrator an internal focus group involving data scientists and demonstrator developers has been conducted after the development of the medium demonstrator. One result was the effective run of the query builder and visualiser components in order to be provided with a better overview on customers affected by particular (detected) events using geographic maps. A developed web application allows a seamless information exchange between the actors involved in the process. Data scientists have gained more experience of using the platform and problems encountered during the first version of the demonstrators have been fixed. The status of the insurance demonstrator is on a good direction with clearly defined steps. The event detection tool has been trained in Italian language and the respective machine learning algorithm has been significantly improved. The anonymiser was exploited for managing sensitive in-house data following the rules of the ethical advisory board. Furthermore, the developed mobile app was tested and found to work satisfactorily. Finally, suggestions to new functionalities such as enhanced technical support in the app, or information about traffic and statistics about vehicle usage were made. Security and privacy regulations seem to be the main issues, such as customers as app users have to give consent for their geo-location to be used. Finally, a close cooperation between HDI and the technical team has found to be fundamental and fruitful during demonstrator development.

Another evaluation of the insurance demonstrator, was conducted based on a workshop held at HDI with insurance experts and data scientists of different business domains, where the AEGIS project and the insurance demonstrator were presented, followed by a demo session of the insurance demonstrator and a feedback phase from the experts. On very interesting wish of the participants was to explore the AEGIS platform on its own, showing the huge interest in the project. All of them recognized that the implemented scenarios bring value to the company and the insurance business. Work practices, which currently need many manual steps based on personal relationships could be automated using the features developed. In general, the experts were satisfied with the workflow implementation and the insurance demonstrator (as well as the use of the AEGIS platform) were positively evaluated from a business perspective.

2.2. Automotive Demonstrator overview, achievements and lessons learnt

The automotive demonstrator was developed according to three different scenarios, (1) broken road indicator, (2) safe driving indicator, and (3) regional driving safety risk estimator. All three scenarios have been successfully implemented and evaluated.

The figure below shows the **automotive demonstrator project page** as well as all datasets created on the AEGIS platform, which are necessary to run the three demonstrators. In total 2.163 trips were processed and analysed on the AEGIS platform resulting in ~47 GB Vehicle Raw Data, ~17 GB TripData_raw and ~18 GB TripData_prepared and ~22 GB TripDensity in the corresponding folders as shown in the figure below.

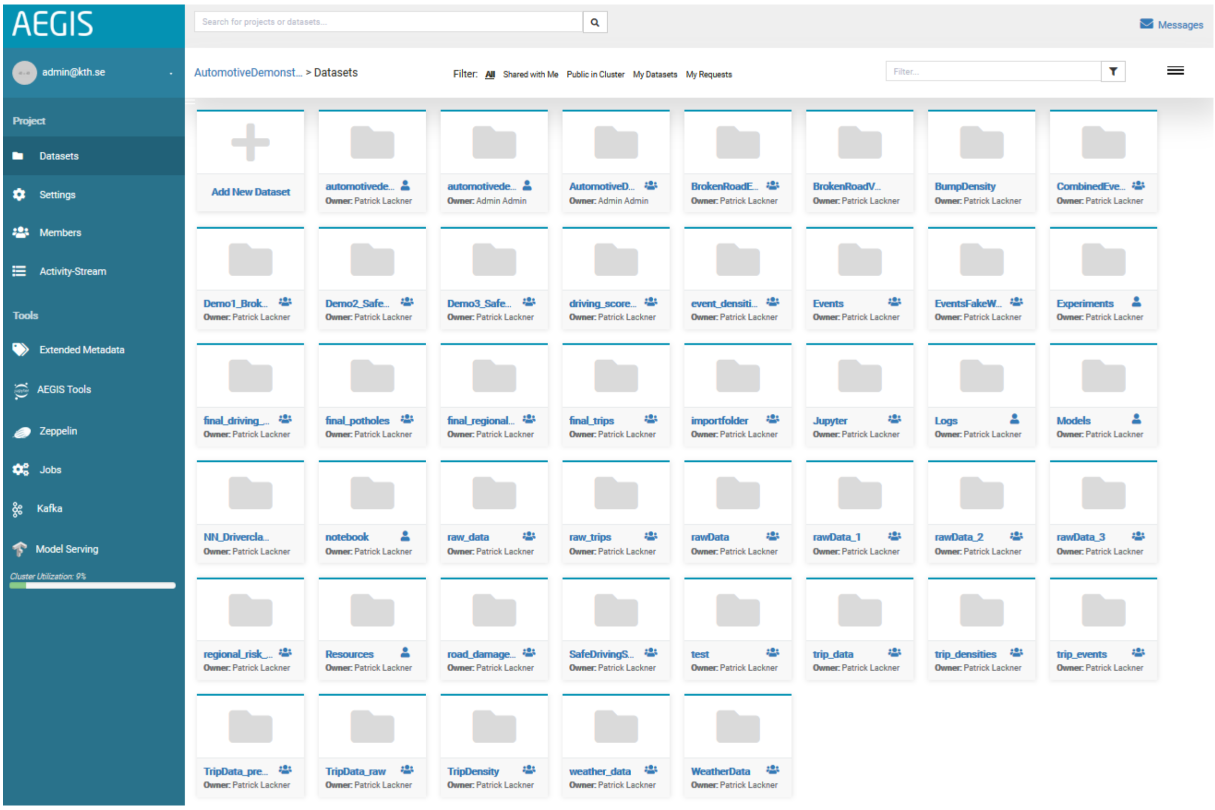

AEGIS
Search for projects or datasets...
Messages
admin@kth.se
AutomotiveDemonst... > Datasets
Filter: All Shared with Me Public in Cluster My Datasets My Requests
Filter...
Project
Datasets
Settings
Members
Activity-Stream
Tools
Extended Metadata
AEGIS Tools
Zeppelin
Jobs
Kafka
Model Serving
Cluster Utilization: 9%
Add New Dataset
automotivede... Owner: Patrick Lackner
automotivede... Owner: Admin Admin
AutomotiveD... Owner: Admin Admin
BrokenRoadE... Owner: Patrick Lackner
BrokenRoadV... Owner: Patrick Lackner
BumpDensity Owner: Patrick Lackner
CombinedEve... Owner: Patrick Lackner
Demo1_Brok... Owner: Patrick Lackner
Demo2_Safe... Owner: Patrick Lackner
Demo3_Safe... Owner: Patrick Lackner
driving_score... Owner: Patrick Lackner
event_densiti... Owner: Patrick Lackner
Events Owner: Patrick Lackner
EventsFakeW... Owner: Patrick Lackner
Experiments Owner: Patrick Lackner
final_driving_... Owner: Patrick Lackner
final_potholes Owner: Patrick Lackner
final_regional... Owner: Patrick Lackner
final_trips Owner: Patrick Lackner
importfolder Owner: Patrick Lackner
Jupyter Owner: Patrick Lackner
Logs Owner: Patrick Lackner
Models Owner: Patrick Lackner
NN_Drivercla... Owner: Patrick Lackner
notebook Owner: Patrick Lackner
raw_data Owner: Patrick Lackner
raw_trips Owner: Patrick Lackner
rawData Owner: Patrick Lackner
rawData_1 Owner: Patrick Lackner
rawData_2 Owner: Patrick Lackner
rawData_3 Owner: Patrick Lackner
regional_risk_... Owner: Patrick Lackner
Resources Owner: Patrick Lackner
road_damage... Owner: Patrick Lackner
SafeDrivingS... Owner: Patrick Lackner
test Owner: Patrick Lackner
trip_data Owner: Patrick Lackner
trip_densities Owner: Patrick Lackner
trip_events Owner: Patrick Lackner
TripData_pre... Owner: Patrick Lackner
TripData_raw Owner: Patrick Lackner
TripDensity Owner: Patrick Lackner
weather_data Owner: Patrick Lackner
WeatherData Owner: Patrick Lackner

Figure 2-1: Automotive demonstrator dataset overview

The **automotive demonstrator (V1, V2, V3)** features a comprehensive **data processing pipeline** implemented on the AEGIS platform. Thereby the four steps (01) data extraction & preparation, (02) event calculation & aggregation, (03) analytics & result generation and (04) output preparation are executed before the result is visualized, using the AEGIS Visualiser component. The figure below explains all created datasets (*raw_data, weather_data, raw_trips, trip_data, trip_events, trip_densities, event_densities, road_damage_quality, regional_risk_density, driving_score_data, final_potholes, final_regional_risk, final_trips, and final_driving_scores*) including their substructures as well as all python scripts executed to transform data from one stage to another. The figure also includes all scripts required to compute the required data for the Visualiser component (*01a_extract_trips, 01b_prepare_trips, 02a_compute_events, 02_compute_trip_density, 02b_compute_event_density, 03_compute_damage_density, 03_compute_risk_density, 03_compute_driving_scores, 04_compute_pothole_vis, 04_compute_regional_risk_vis, 04_compute_trip_vis, and 04_compute_driving_score_vis*). Finally, the script *visualizer* is run to show the results.

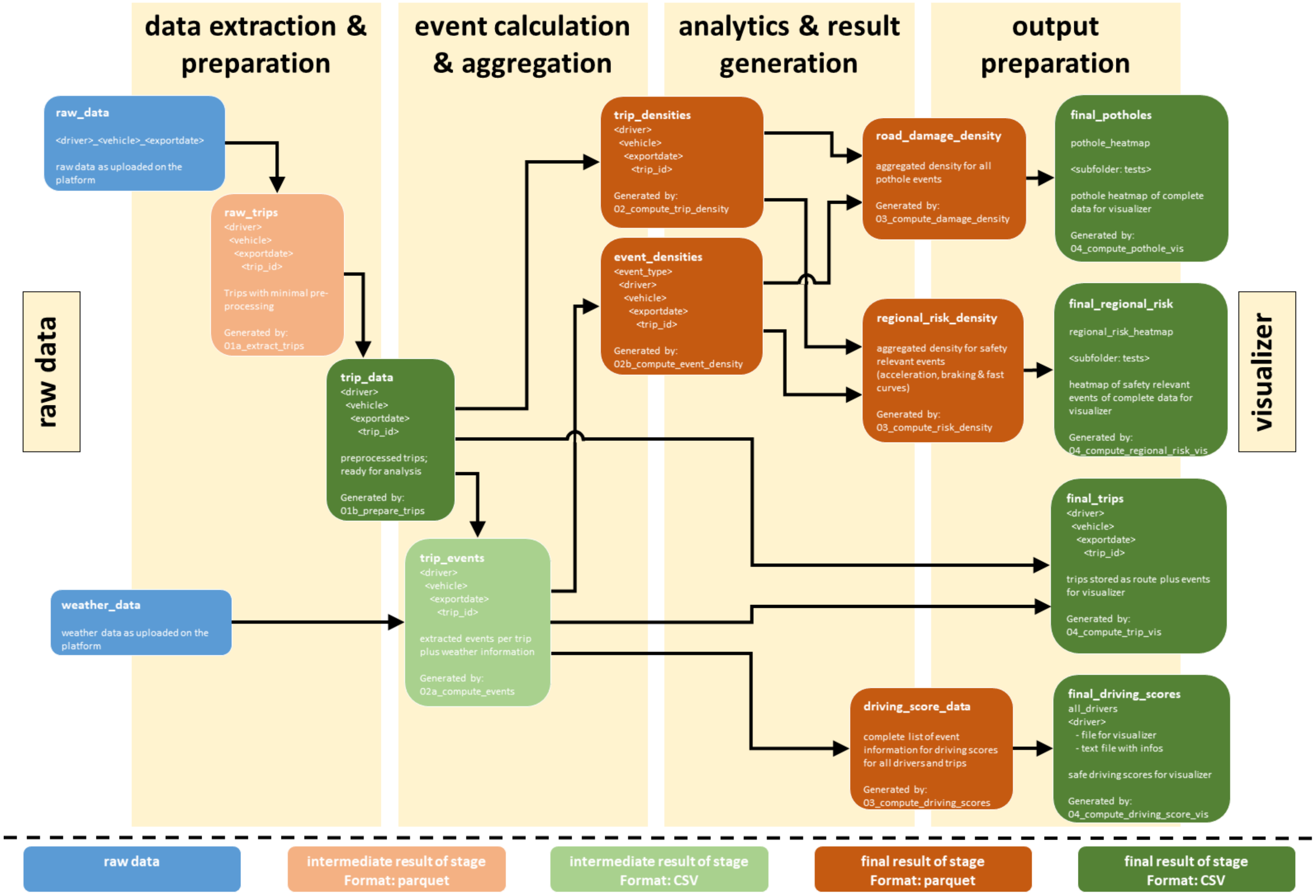
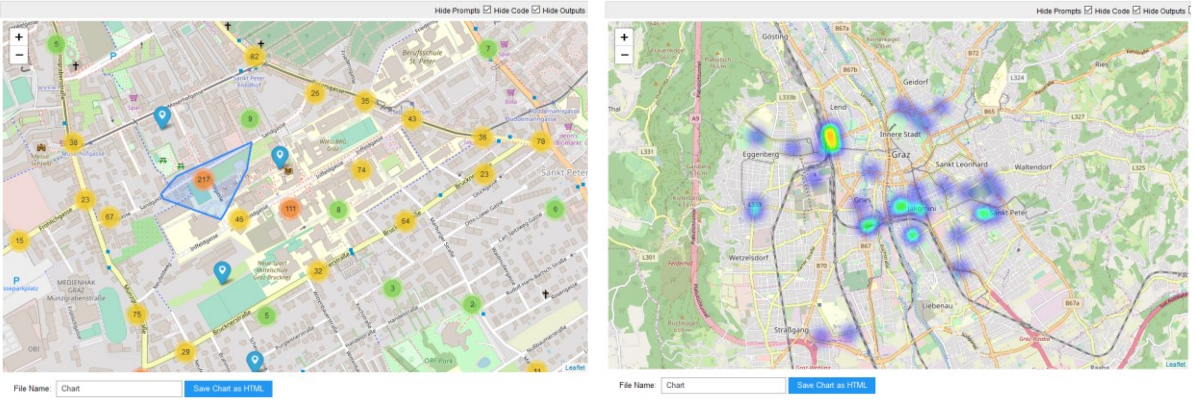

data extraction & preparation
event calculation & aggregation
analytics & result generation
output preparation
raw data
raw_data
<driver>_<vehicle>_<exportdate>
raw data as uploaded on the platform
raw_trips
<driver>
<vehicle>
<exportdate>
<trip_id>
Trips with minimal pre-processing
Generated by:
01a_extract_trips
trip_data
<driver>
<vehicle>
<exportdate>
<trip_id>
preprocessed trips; ready for analysis
Generated by:
01b_prepare_trips
weather_data
weather data as uploaded on the platform
trip_events
<driver>
<vehicle>
<exportdate>
<trip_id>
extracted events per trip plus weather information
Generated by:
02a_compute_events
trip_densities
<driver>
<vehicle>
<exportdate>
<trip_id>
Generated by:
02_compute_trip_density
event_densities
<event_type>
<driver>
<vehicle>
<exportdate>
<trip_id>
Generated by:
02b_compute_event_density
road_damage_density
aggregated density for all pothole events
Generated by:
03_compute_damage_density
regional_risk_density
aggregated density for safety relevant events (acceleration, braking & fast curves)
Generated by:
03_compute_risk_density
driving_score_data
complete list of event information for driving scores for all drivers and trips
Generated by:
03_compute_driving_scores
final_potholes
pothole_heatmap
<subfolder: tests>
pothole heatmap of complete data for visualizer
Generated by:
04_compute_pothole_vis
final_regional_risk
regional_risk_heatmap
<subfolder: tests>
heatmap of safety relevant events of complete data for visualizer
Generated by:
04_compute_regional_risk_vis
final_trips
<driver>
<vehicle>
<exportdate>
<trip_id>
trips stored as route plus events for visualizer
Generated by:
04_compute_trip_vis
final_driving_scores
all_drivers
<driver>
- file for visualizer
- text file with infos
safe driving scores for visualizer
Generated by:
04_compute_driving_score_vis
visualizer
raw data
intermediate result of stage Format: parquet
intermediate result of stage Format: CSV
final result of stage Format: parquet
final result of stage Format: CSV

Figure 2-2: Data workflow of the Automotive Demonstrator (V1, V2, V3)

Though each demonstrator version can be used independently, they require similar data extraction & preparation as well as event calculation and aggregation steps using the data processing pipeline presented above.

The first demonstrator, **broken road indicator**, creates either a marker visualisation of road damage (potholes), or a heatmap visualisation of road damage (potholes) for the Visualiser, as shown in the figure below, to further discuss it with road maintenance experts.

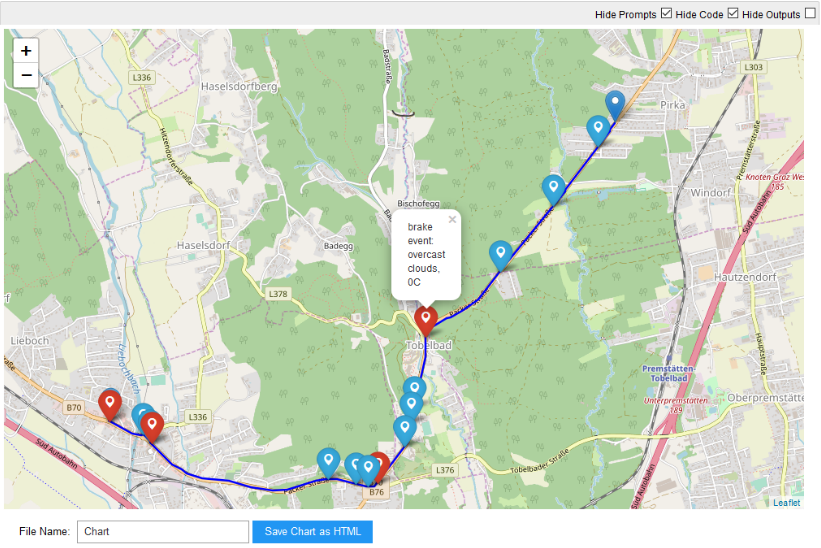

Hide Prompts ☑ Hide Code ☑ Hide Outputs
File Name: Chart
Save Chart as HTML
Hide Prompts ☑ Hide Code ☑ Hide Outputs
File Name: Chart
Save Chart as HTML

Marker Visualisation of Road Damage (all drivers)

Heatmap Visualisation of Road Damage (driver-id3)

Figure 2-3: Broken Road Indicator (Automotive Demonstrator V1)

The second demonstrator, **safe driving indicator**, creates a (combined) trip and marker visualisation of all trip-specific safe driving events detected within one trip to further discuss it with drivers.



brake event: overcast clouds, 0C

Marker and Route Visualisation of Safe Driving Events (one trip)

Figure 2-4: Safe Driving Indicator I (Automotive Demonstrator V2)

Furthermore, the second demonstrator, **safe driving indicator**, calculates individual risk scores for all trips and all drivers, as well as the aggregated risk scores, for all drivers for driver/trip benchmarking, which can be visualised with time series visualisation (all risk score of all trips) and scatter plot (all aggregated risk scores of all drivers) to further discuss it with drivers.

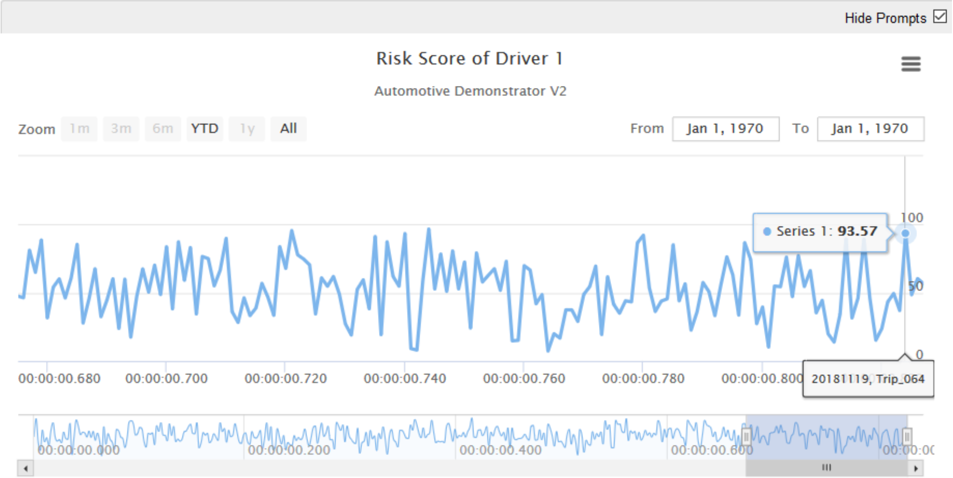


Figure 2-5: Visualisation of all individual risk scores of a single driver (Automotive Demonstrator V2)

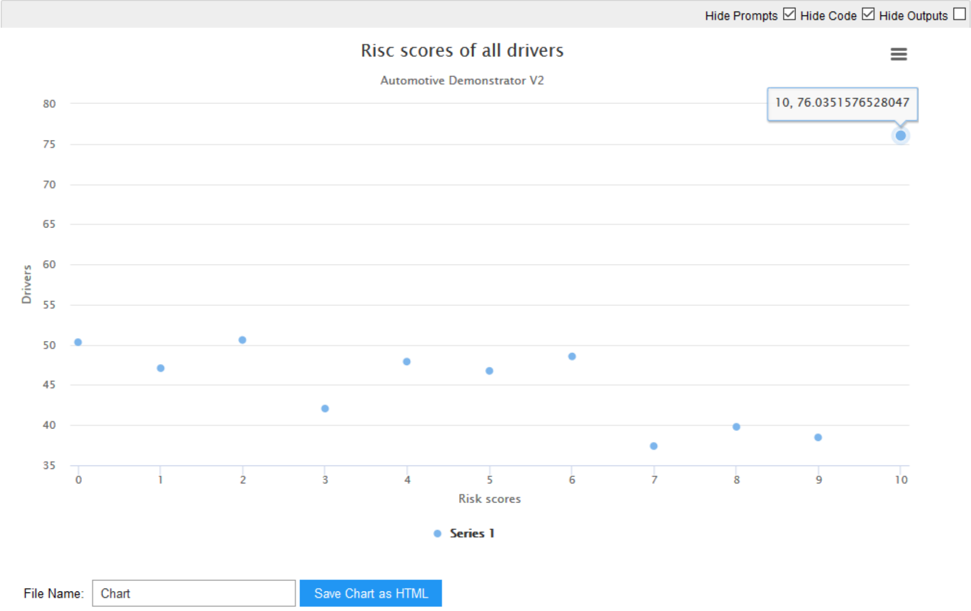


Figure 2-6: Visualisation of aggregated risk scores of all drivers(Automotive Demonstrator V2)

The third demonstrator, **regional driving risk estimator**, creates a heatmap of all safety relevant events identified in the complete data for the visualizer to further discuss it with traffic planning experts.

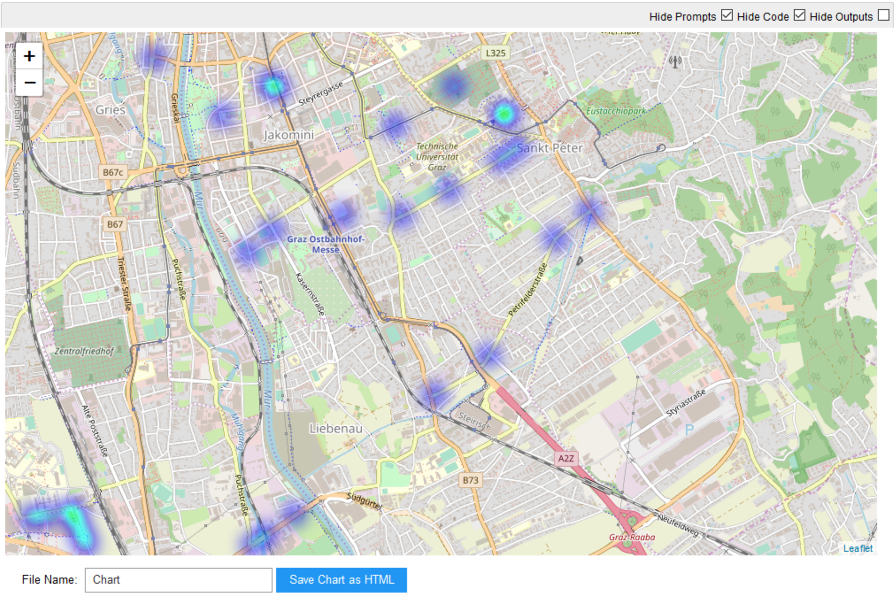


Figure 2-7: Regional Driving Risk Estimator Indicator (Automotive Demonstrator V3)

During the implementation of the **automotive demonstrator**, many **challenges** were raised, which are described in the deliverables D5.3-D5.5. One particular challenge to mention was the alignment of the demonstrator development with the platform development. However, during the progress of the project, the cooperation quality between demonstrator developers and the platform developers highly increased. Furthermore, after exploring the platform's capabilities by developing the first demonstrator, broken road indicator, a redesign of data structures, workflows, and scripts has been initiated, as experiences with the AEGIS methodology have increased, too. This redesign process lasted until the end of the AEGIS project and the resulting final data workflow of the automotive demonstrator has been already presented in this deliverable.

The evaluation with users (of the services) generated many relevant **insights**; however, not all of them can be implemented within the project's runtime. Many of them are out of scope and therefore relevant for the latter exploitation phase of the project with respect to automotive data science. They will be carefully considered in further exploitation actions. Furthermore, the three demonstrators enabled VIF to meet virtually and face to face with external stakeholders of the automotive ecosystem (car manufacturers, engineers, and technology providers) and engage them in commercially relevant discussions for future cooperation in automotive data science projects. This is something which would not have been possible without the results of the AEGIS project, as these clearly indicates the data science competencies of VIF leading to three relevant demonstrator's versions.

During the cause of the demonstrator implementation, even a **mini use case** on the topic of driver/vehicle classification was implemented to explore the machine learning/ AI capabilities of the AEGIS platform in the course of a computer science seminar project using the data of the automotive demonstrator too. Thereby the following research questions have been formulated to set the direction: Can a deep neural network identify the driver given a trip segment of raw data? Is the network able to construct its own relevant features? Is the network able to analyse only the comparable parts when given a larger segment? What accuracies can be achieved? How do recurrent networks compare to non-recurrent ones? The figure below shows already quite promising results for driver classification, using the data of vehicle standstill events calculated on the AEGIS platform. The project will continue till the end of July 2019.

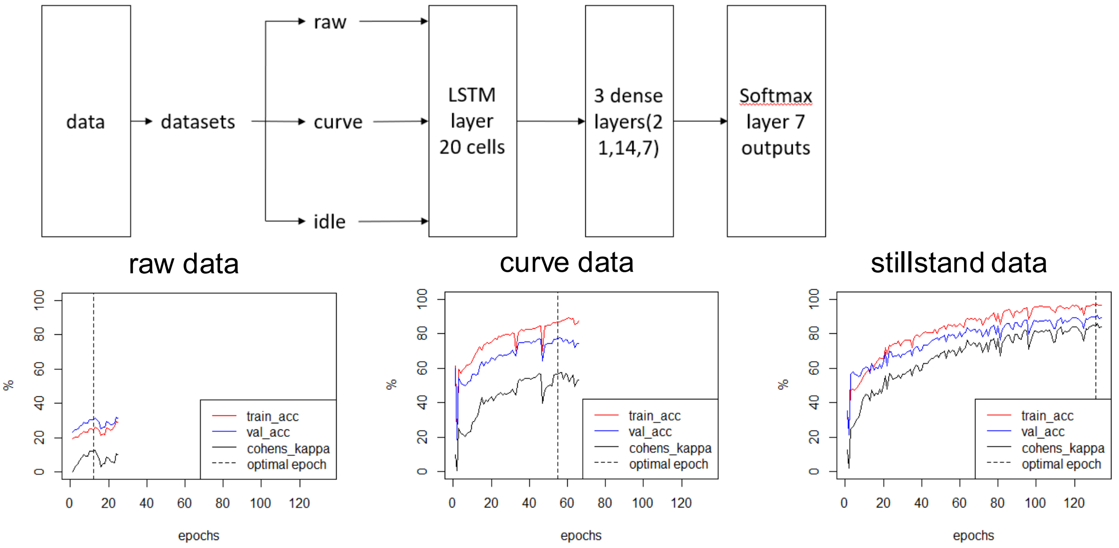


Figure 2-8: Automotive Demonstrator: Mini Use Case Driver Classification

Finally, a lightweight **mobile application for drivers** has been developed as a further mini demonstrator to visualise their driving data (computed on the AEGIS platform) without having to access the AEGIS platform. The figure below shows screenshots of the application running on an Android phone (Samsung Galaxy S9 Plus). This was also a requirement voiced by the voluntary drivers providing data to the AEGIS platform so that they can investigate their driving data without the assistance of a VIF data scientist operating the AEGIS platform.

The driver can visualise an overall risk score, as well as the individual risk scores of the last 20 trips. Finally, the driver receives a summary of the total number of trips examined, the total distance travelled, the total time driven, the total number of events detected in the driving data and the total number of events detected per category (brake, acceleration, standstill), as shown in the three screenshots of the driver app below.

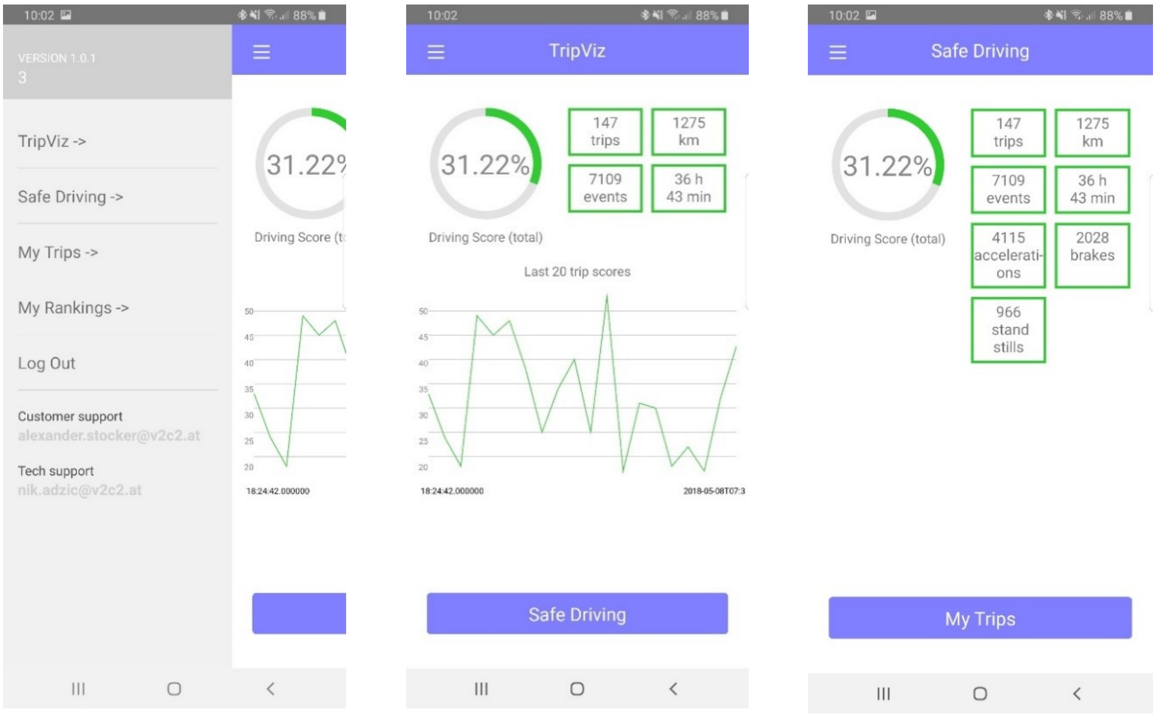
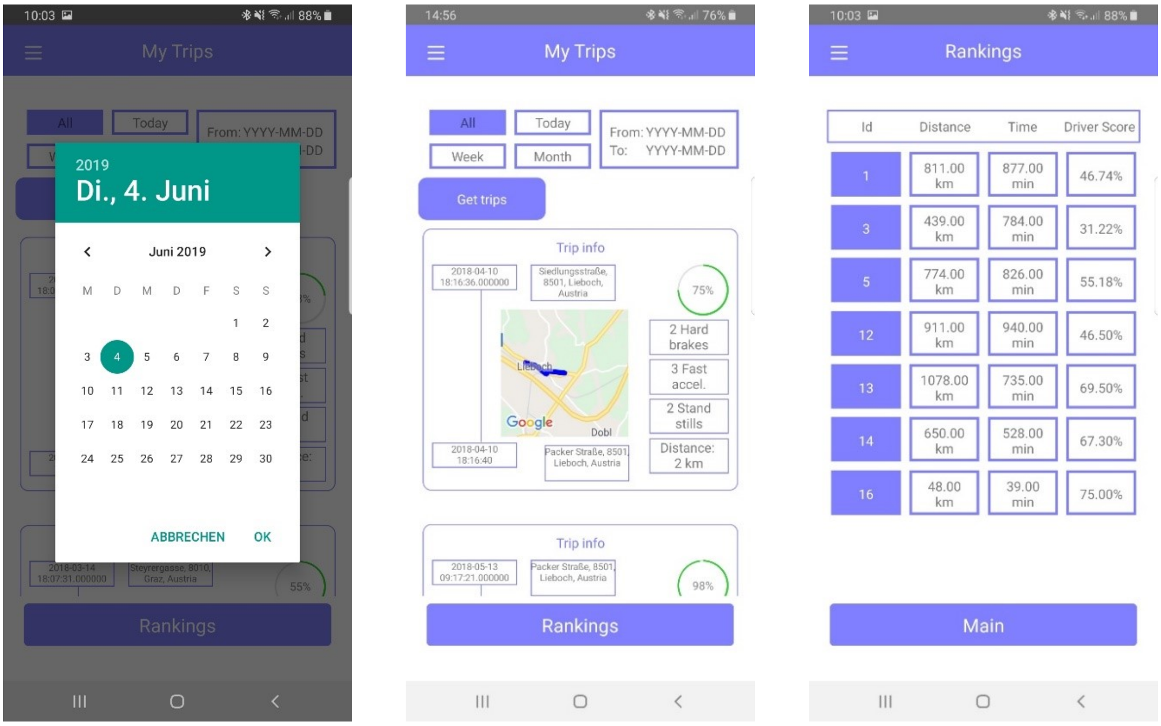


Figure 2-9: Automotive Demonstrator: Driver App I

The driver can select a trip from all trips of today, the current week or the current month and visualise it with the mobile app. All trip-specific events and a trip-specific risk score are displayed. Finally, the driver can compare his individual driving style with that of his colleagues, as the three screenshots of the driver application below show.



Id	Distance	Time	Driver Score
1	811.00 km	877.00 min	46.74%
3	439.00 km	784.00 min	31.22%
5	774.00 km	826.00 min	55.18%
12	911.00 km	940.00 min	46.50%
13	1078.00 km	735.00 min	69.50%
14	650.00 km	528.00 min	67.30%
16	48.00 km	39.00 min	75.00%

Figure 2-10: Automotive Demonstrator: Driver App II

2.3. Smart Home and Assisted Living Demonstrator overview, achievements and lessons learnt

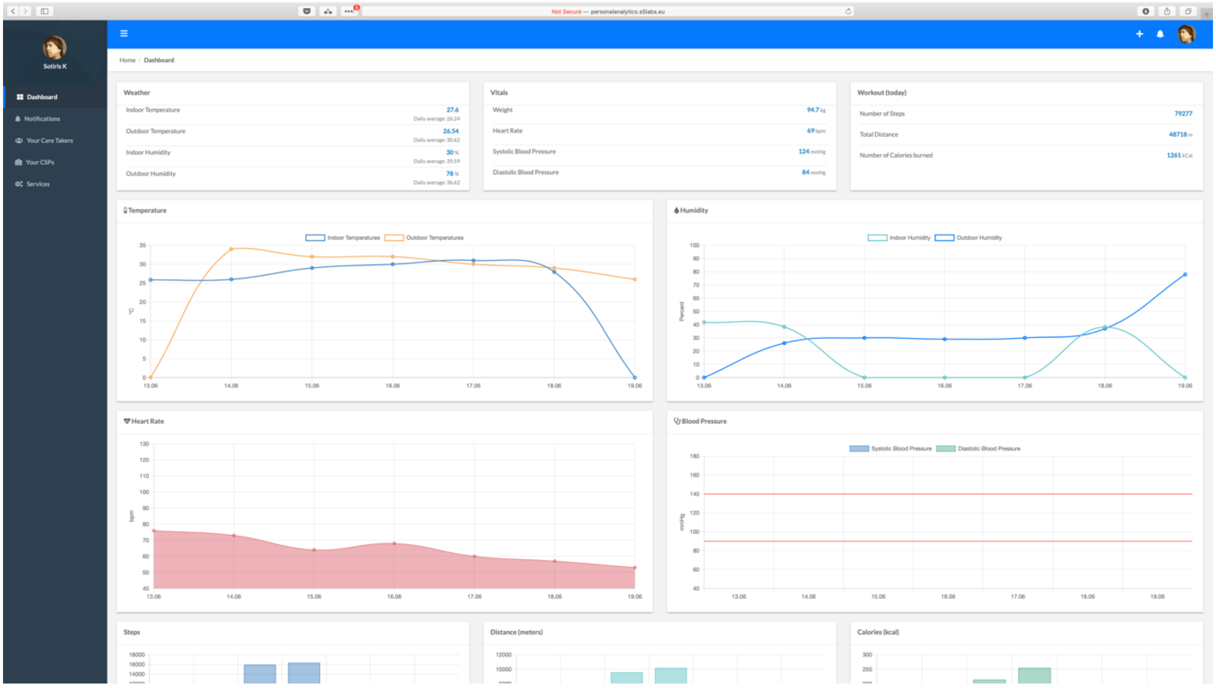
As part of the Smart Home and Assisted Living (SHAL) demonstrator, the responsible partners developed an infrastructure that collects health, well-being, activity and smart-home data of (at risk) individuals and makes use of the AEGIS platform to perform certain analyses and retrieve data towards producing services that are able to improve the safety of such individuals. Although the end-beneficiary of the service are (at-risk) individuals and their informal carers, the services are offered to Care Service Providers (CSPs) that are in need of a system that is able to collect data from individuals they treat, and to use big data infrastructures such as the AEGIS platform to minimise the cost of running in-house analytics.

In more detail, the system that has been developed makes use of the data coming from individuals and that are (after consent) offered by those individuals to CSPs, in order to help data analysts working in such organisations to combine such heterogenous data coming from diverse proprietary and public sources, and conduct analyses that may result to better motoring and more evidence-based recommendations to such user groups, targeting also personalisation as well.

During the course of the demonstrator, the responsible partners successfully developed the demonstrator envisaged in the project's DoA, and in total a number of 24 test cases were executed in the three stages of the demonstrator, which has been constructed in such a procedural manner to allow for better testing the assumptions that were made during its initial conceptualisation.

Within the context of the SHAL demonstrator the key main artefacts produced were: (a) the SHAL Web App that is offering the extended list of functionalities to all stakeholders of the demonstrator (CSPs, informal carers and individuals), (b) the mobile applications that were developed for both Android and iOS devices that is offering the main functionalities for the (at-risk) individuals and (c) a sophisticated lightweight but still powerful backbone was developed in order to orchestrate the rest of components, but also to support the required backend functionalities such as the storage, homogenisation and anonymisation of the collected data and the connectivity with the AEGIS platform in order to leverage from the platform's big data capabilities.

The Web App is offering an interface for both individuals, as well as their CSPs (and informal carers) to visually explore some of the data that are collected by sensors that individuals register to the system (wearables, smart home sensors, etc). From the point of individuals, they are served with a dashboard that contains such information and allows them to share it with certain CSPs, either anonymously (in that case personal data is stripped from personal identifiable information and is aggregated under a pre-selected "persona" under which the user fits best), or even in full, by granting access to their dashboard to CSP. Moreover, they are also able to link to carers, which are receiving only notifications that have to do with the wellbeing of each individual (the exact same notifications that individuals receive), but are not able to access any other data that relates to those people.



Weather

Indoor Temperature	27.6 Daily average: 26.24
Outdoor Temperature	26.54 Daily average: 30.62
Indoor Humidity	30 % Daily average: 39.59
Outdoor Humidity	78 % Daily average: 36.62

Vitals

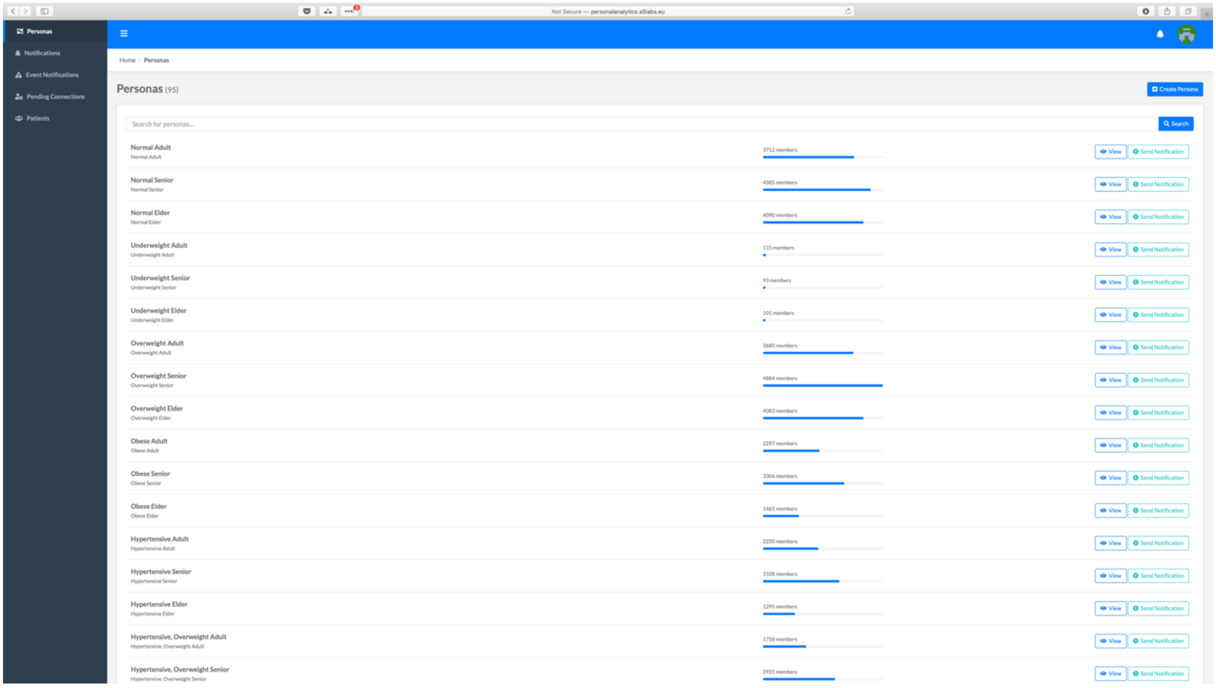
Weight	94.7 kg
Heart Rate	69 bpm
Systolic Blood Pressure	124 mmHg
Diastolic Blood Pressure	84 mmHg

Workout (today)

Number of Steps	79277
Total Distance	48718 m
Number of Calories burned	1261 kCal

Figure 2-11: SHAL Web App individuals' dashboard

For CSPs, the web app provides a similar interface, where they are able to witness stats that relate to "personas", as showing in the next figures, where individuals are anonymously categorised and aggregated stats of the persona are displayed. Options exists to extract the data of the persona for further analysis on the AEGIS platform, or locally, as well as to manually send notifications to groups of a persona.



Personas (95)

Normal Adult Normal Adult	3712 members
Normal Senior Normal Senior	4385 members
Normal Elder Normal Elder	4090 members
Underweight Adult Underweight Adult	115 members
Underweight Senior Underweight Senior	93 members
Underweight Elder Underweight Elder	101 members
Overweight Adult Overweight Adult	3680 members
Overweight Senior Overweight Senior	4884 members
Overweight Elder Overweight Elder	4083 members
Obese Adult Obese Adult	2297 members
Obese Senior Obese Senior	3306 members
Obese Elder Obese Elder	1465 members
Hypertensive Adult Hypertensive Adult	2250 members
Hypertensive Senior Hypertensive Senior	3108 members
Hypertensive Elder Hypertensive Elder	1295 members
Hypertensive, Overweight Adult Hypertensive, Overweight Adult	1758 members
Hypertensive, Overweight Senior Hypertensive, Overweight Senior	2931 members

Figure 2-12: SHAL Web App list of Personas with action buttons

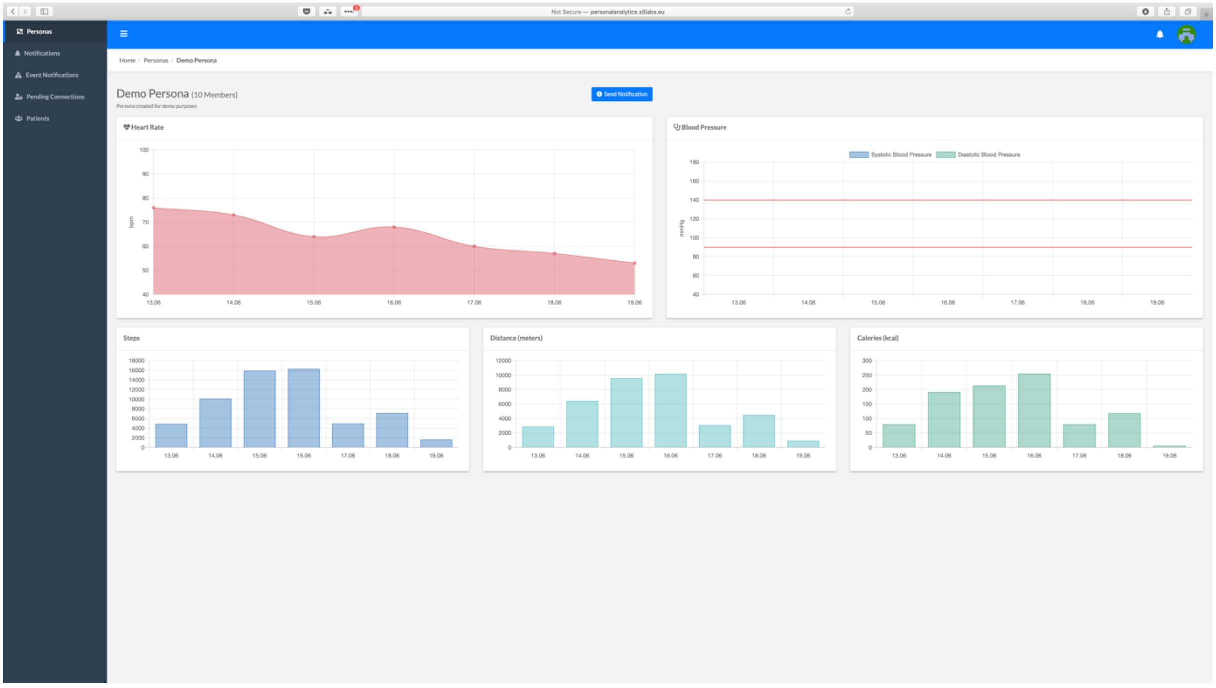


Figure 2-13: SHAL Web App aggregated stats of a Persona

The mobile applications are offering an interface where at risk individuals can check some of the more important values that have to do with their daily activities and biometrics, such as heartrate, blood pressure, steps walked, etc., while also information from smart home devices that they have linked to their profile is displayed.

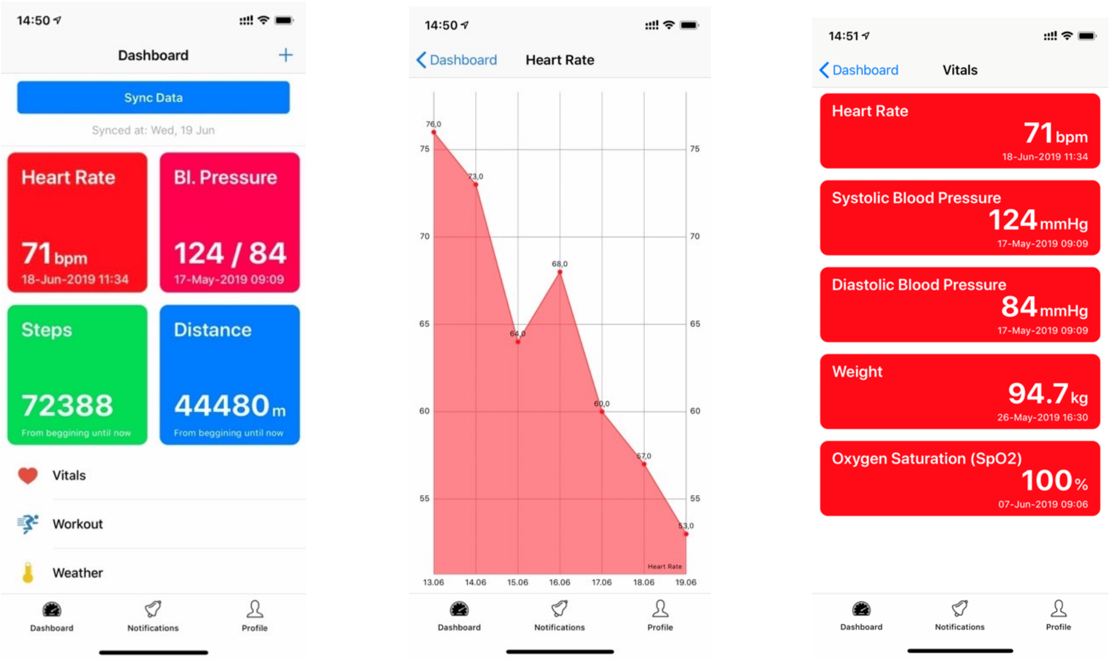
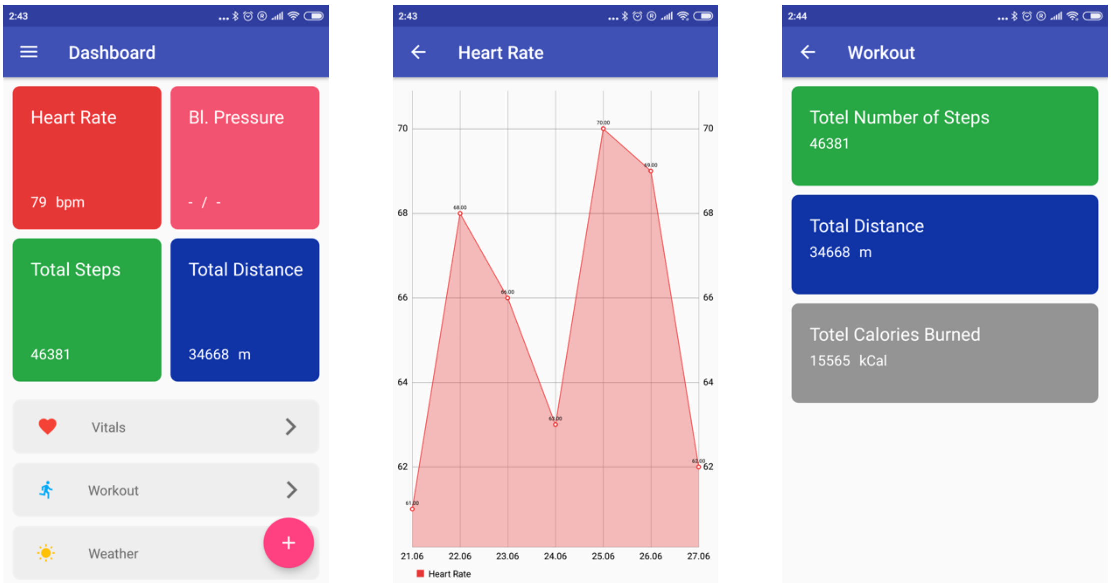


Figure 2-14: SHAL iOS Dashboard and Graphs



2:43

Dashboard

Heart Rate
79 bpm

Bl. Pressure
- / -

Total Steps
46381

Total Distance
34668 m

Vitals

Workout

Weather

2:43

Heart Rate

2:44

Workout

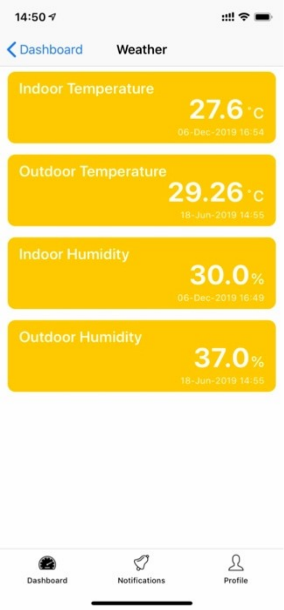
Totel Number of Steps
46381

Total Distance
34668 m

Totel Calories Burned
15565 kCal

Figure 2-15: SHAL Android Dashboard and Graphs

The App works as a data hub, which collects information from various sources that are connected to a smartphone and pushes these data to the backend SHAL server. Moreover, it retrieves other data available in SHAL to showcase to the user, such as for example indoor and outdoor conditions.



14:50

Dashboard Weather

Indoor Temperature
27.6 °c
06-Dec-2019 16:54

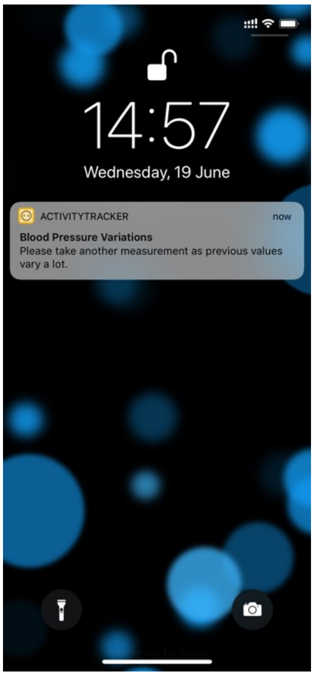
Outdoor Temperature
29.26 °c
18-Jun-2019 14:55

Indoor Humidity
30.0 %
06-Dec-2019 16:49

Outdoor Humidity
37.0 %
18-Jun-2019 14:55

Figure 2-16: SHAL Display of third-party sensor measurements (smart home)

Moreover, the app allows the receipt of notifications, which are pushed to such individuals either automatically based on the trained models that relate to health issues or smart home comfort zones, or manually, once this is instructed by a CSP to inform the people monitored.



ACTIVITYTRACKER now

Blood Pressure Variations

Please take another measurement as previous values vary a lot.

Figure 2-17: SHAL Notification as shown in an iPhone locked screen

Apart from the successful technical results of the demonstrator, that were tested during this period, of special importance for the partners implementing the demonstrator was the feedback received by potential stakeholders of the domain that the demonstrator addresses, which are CSPs that have been engaged in face-to-face meetings and in an online focus group (webinar), as well as individuals that have tested the infrastructure as volunteers. Both those groups were very positive both on the concept and the solution developed and were eager to use the solution to understand better its operation and its potential impact.

From a technological perspective, the consortium has overcome some challenges that had to do with data homogenisation in order to facilitate analyses, while the privacy and anonymisation features have been successfully tackled by the chosen approach. The existence of the AEGIS platform and its analytics processing features, as well as the presence of some open datasets allowed the partners to develop a solution that is lightweight and requires little processing resources, which is ideal for CSPs of medium to small sizes, that do not want to invest in computing resource, but outsource data analysis and data linkage tasks. One major technical challenges faced, which has not been fully tackled due to external conditions have to do with the real-time flow of information (as most sensors and mobile devises do not support the continuous streaming of measurements mostly due to energy constrains).

CSPs commented that the overall concept as well as the insights that are delivered by the platform see to be of great importance, as they allow CSP personnel after performing certain analyses to better categorise subjects, not only on pure medical and health data, but also on other activity which are in the majority of cases unknown to CSPs and they are not able to track them down. Although such data are by themselves not of great significance to CSPs, when combined with the medical data and the health profiles of individuals, they become an extra asset to understand the behaviour of individuals and to offer more informed and personalised recommendations.

Individuals are also providing very positive comments for the SHAL idea, both as a concept, but also liked the current deployment which allows them to understand better their day-to-day routines, and also receive recommendation for tasks they would otherwise either ignore. In principle all of the individuals which voluntarily participated in the demonstrator were concerned about the use of their personal data, nevertheless the anonymisation method used in SHAL, as well as the option for them to choose with who to share their data and to which extent (e.g. anonymised or actual data), played down their concerns, whereas they all agreed to share more data for getting more personalised notifications.

Finally, it has to be noted that the overall operation of the demonstrator has showcased that there is a huge need for data availability and homogenisation, as the more and better data available, the better the insights can be. As an example, the availability of (open) APIs from third party sources that provide information about weather and environmental conditions are at the moment quite high, nevertheless the data they provide are quite poor; they are restricted to certain regions, updating times are slow and data frequencies are quite low, while the accuracy of those is poor as well. In principle, the partners believe that the availability of such data would further improve the impact of infrastructures that have similar concepts with SHAL, and that the sharing of those data could be facilitated by a platform as AEGIS, where data providers would be able to distinguish a real need for their data, and work towards improving them in return for some value.

2.4. Insurance Demonstrator achievements and lessons learnt

The overall goal of the AEGIS Insurance demonstrator is to exploit the AEGIS platform big data technologies in order to access and analyse information coming from diverse and heterogeneous data sources including the in-house data (e.g. customers location, customers portfolio). Exploring with the AEGIS tools weather, news and crime open data, the HDI data scientists would be able to manage in an efficient way events (that are going to happen or just happened), while the use of the AEGIS analytic tools would allow the company to set a strategy to minimise the impact of the event on the company itself, while offering a support to the customers.

Summarizing, the expected benefits of the AEGIS adoption by HDI Assicurazioni include the possibility to rapidly analyse and process huge volumes of cross domain data, an improvement of the customer’s satisfaction through personalised offering and information about natural or social events of interest. Moreover, AEGIS services would constitute a key for the development of an accurate and successful business plan through the pricing analysis and the analysis of the trend of the company.

The Insurance demonstrator has been presented with three scenarios, as summarized in Figure 2-18:

- ***Scenario 1:*** *Financial impact, customer support and services*
- ***Scenario 2:*** *Personalised early warning services for asset protection*
- ***Scenario 3:*** *Marketing strategy and pricing support services*

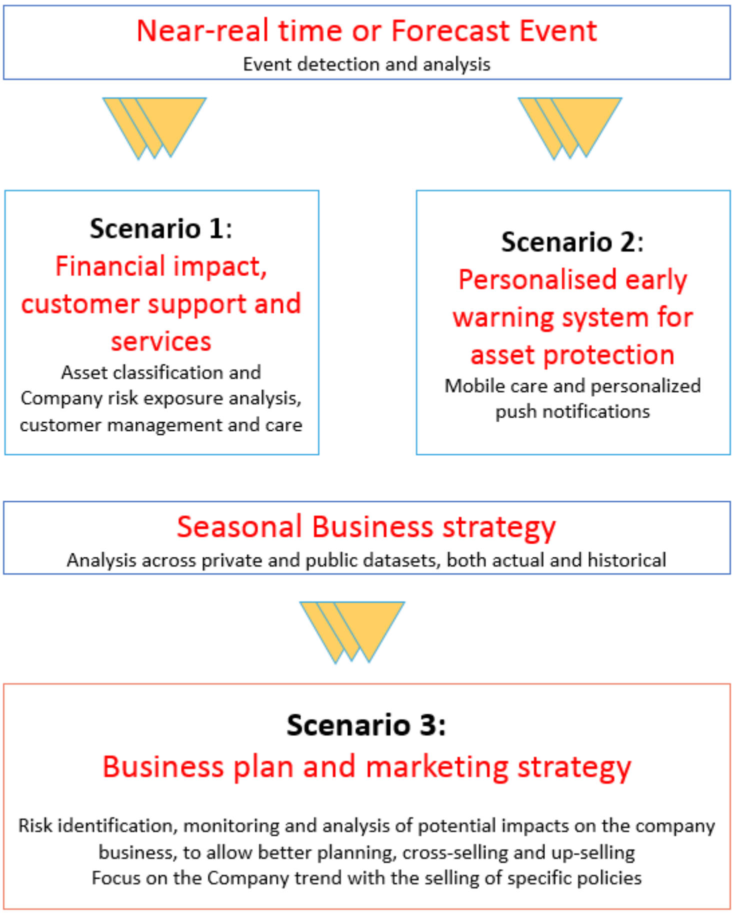

Near-real time or Forecast Event
Event detection and analysis
Scenario 1:
Financial impact, customer support and services
Asset classification and Company risk exposure analysis, customer management and care
Scenario 2:
Personalised early warning system for asset protection
Mobile care and personalized push notifications
Seasonal Business strategy
Analysis across private and public datasets, both actual and historical
Scenario 3:
Business plan and marketing strategy
Risk identification, monitoring and analysis of potential impacts on the company business, to allow better planning, cross-selling and up-selling
Focus on the Company trend with the selling of specific policies

Figure 2-18: Insurance demonstrator - scenarios overview

Each scenario aims to profit of the AEGIS platform functionalities towards:

- Improving the customer centricity (Scenario 1 and 2),
- Preventing frauds (Scenario 1),
- Exploiting the internet of things and real-time notifications (Scenario 1 and 2),
- Setting the pricing and selling strategies (Scenario 2 and 3),
- Performing risk analysis (Scenario 3).

The main technical components of the Insurance demonstrator are depicted in Figure 2-19:

- The Anonymiser that is an offline tool locally installed. With the use of the Anonymiser, the Data Scientist anonymizes the in-house datasets before their upload on the AEGIS platform. This data pre-processing is needed for compliance with the privacy regulation. The Anonymiser is provided as an offering of the AEGIS project.
- The HDI Mobile App that provides the customers geolocation to the Web App (Scenario 2) and receives notifications from the Web App (Scenario 1 and 2). A Google service (Firebase) is called by the Mobile App Back-End in order to enable the push notifications. The HDI Mobile App was enhanced for the scenario execution with the two functionalities described.
- The HDI Web App that is an HDI environment (accessed by all the HDI actors defined) that allows the communication between different roles. It has been developed within the

project by the demonstrator in order to implement the correct workflows between the actors in the three scenarios. From the HDI Web App it is possible to gather the geolocation data from the HDI Mobile App (Scenario 2). Finally, the HDI Expert is informed that an event has been detected (Scenario 1 and 2) through a notification that is sent by the AEGIS platform.

- The Event Detection Tool/ Data Harvester (Scenario 1 and 2): through the creation of a Job on the AEGIS platform, the Data Harvester uploads the event data (detected by the Event Detection tool) on the AEGIS platform. The Kafka service of AEGIS detects this and sends notifications to the Web App.
- The AEGIS platform notebooks (Query Builder, Algorithm Execution Container and Visualizer) are the actual tools used by the Data Scientists to perform his/her analysis (Scenario 1, 2 and 3).

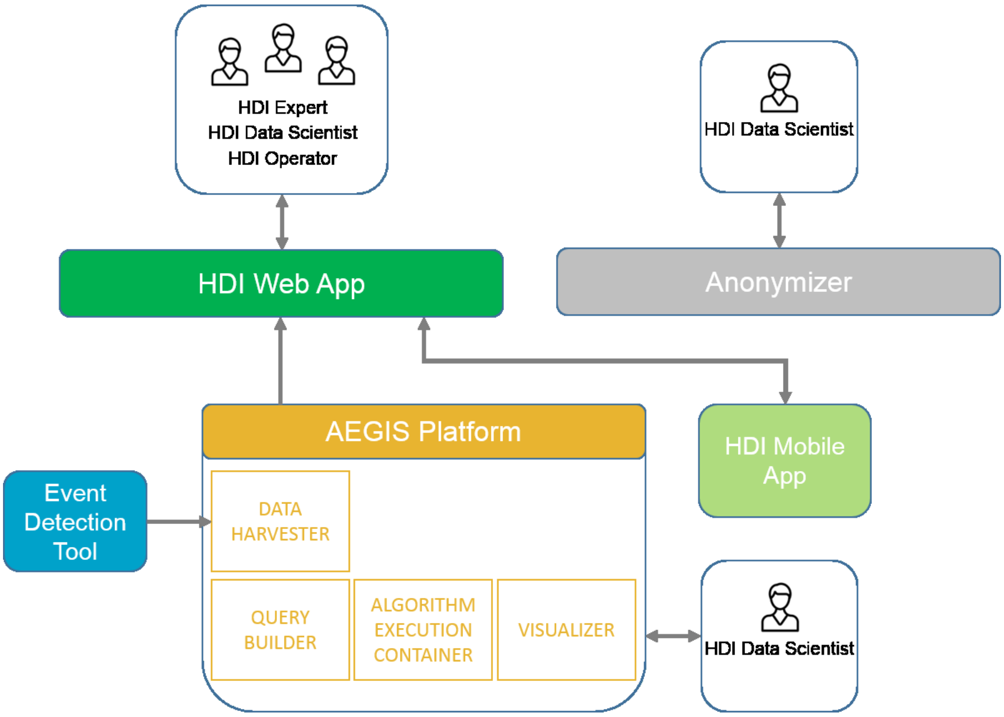


Figure 2-19: Insurance demonstrator - technical components and user(s) for each component

Summarising, from this list of components, the following have been developed within the context of the demonstrator for the execution of the three scenarios of the Insurance demonstrator:

- the HDI Web App, for the sharing of information within the HDI actors. The application has the aim to ease the communication between different departments of the company.
- the existing HDI Mobile App has been enriched with some functionalities, in particular the geolocation feature and push notifications.
- the Event Detection Tool has been trained for four events (flood, whirlwind, hailstorm and socio-political) in the Italian language.

- the Query Builder and the Visualizer have been customized to fulfil the demonstrator requirements: customers priority feature, the earthquake risk Heatmap and the customers pin related with the kind of policy held.

It is important to point out that the scenarios have been developed and executed in different times according to the demonstrator implementation and execution plan: on M18 for the early demonstrator, Scenario 1 was executed, on M24 for the medium demonstrator Scenario 2 was executed, and on M30 for the advanced demonstrator, Scenario 3 was executed. They have been deeply described respectively in D5.3, D5.4 and D5.5 and their execution results reflected the version of the platform at that time.

While the platform maturated, some corrections have been done to the workflows developed. The major achievement that has not yet been described is the real-time notification service that enhances the workflow of the Scenario 1 and 2. Through a Job on the AEGIS platform in fact, from the HDI Web App it is possible to receive notifications by the Kafka service of the platform when the Event Detection tool detects a new event. The increasing integration of the platform and its components, lead to store the events data detected by the Event Detection Tool as files through the Data Harvester on AEGIS. If these files are updated, the Kafka service subscribed with a Java class by the HDI Web App retrieves this change and pushes a notification.

The scenarios execution has been performed by GFT and HDI, however during the development phases the consortium partners that developed the three notebooks, the Anonymizer and the Even Detection Tool, namely UBITECH, NTUA and SUITE5, collaborated with the demonstrator development team to facilitate the process.

For the final evaluation of the Insurance demonstrator, as described also in D5.5, a workshop with six HDI people that were not involved in the team of the AEGIS project has been organised. The participants were three Data Scientists/Analysts and three Experts (belonging to different HDI departments: claims, commercial, portfolio). They had the opportunity to have an overview of the project with a focus on the Insurance demonstrator, a demo and some time to use the AEGIS platform and the HDI Web App.

Due to the privacy regulation, as stated also in the previous deliverables of WP5, it was not possible to involve the HDI customers in the scenarios execution since a large amount of them could have signed a specific consent to make the results significant (for further details see section 6.3.1 of D5.5 - Demonstrators Evaluation and Feedback – v3).

During the three scenarios development phase, it was acknowledged that apart from the ‘accessory tools’, the platform functionalities have been exploited almost as they are offered by the AEGIS platform and only minor changes were needed to the Jupyter notebooks. This easy adaptation to different scenarios from the available default options of the three notebooks has been pointed out also during the workshop as the extended usability level of the platform. Furthermore, during the workshop the events identified for the EDT training have been evaluated as suitable which highlight the usefulness of the Event Detection Tool. All the workshop’s participants showed interest on the project and on the platform, and evaluated successfully the usability both of the AEGIS platform and of the HDI Web App. They expressed their interest to use more the platform; in fact the two Data Scientists that participated to the workshop have already started using the platform.

3. AEGIS PLATFORM FINAL EVALUATION

3.1. Quantitative Evaluation

This section presents an overview of the quantitative evaluation of platform that was conducted during the project. As instructed by the AEGIS evaluation framework, which was documented in deliverable D5.1 and updated in deliverable D5.2, the holistic evaluation of the platform was performed in three different iterations, one for each of the phases of the demonstrators, giving valuable feedback to the AEGIS platform developers during the different implementation phases of the AEGIS platform. Especially for the third evaluation of the platform, as it was performed in the final month of the project (M30), in which the final release of the AEGIS platform was also delivered, the valuable feedback collected will drive the enhancements and adjustments that will be introduced in the platform during the project's exploitation phase.

The quantitative evaluation of the platform was based on a set of technical Key Performance Indicators (KPIs) that were defined based on the eight characteristics (and its' sub-characteristics) of the international standard ISO/IEC 25010:2011: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE). The scope of the quantitative evaluation was to provide the required quality assurance and control during the implementation phase of the platform.

Table 3-1 presents the aggregated results of the three evaluations that were performed during the evaluation phase. From these results, the following main insights were extracted:

- As the implementation of the platform progressed, some key characteristics were significantly improved towards the final version of the platform, such as the performance efficiency, the usability and the reliability characteristics. This indicates the incremental maturity of the platform, as well as the focus of the development team to offer additional value to the platform's stakeholders.
- For several key characteristics, the platform has showcased the added value that it offers from the early releases, such as the functional suitability, the portability and maintainability, and it is obvious that the development team focused on providing a stable platform with advanced characteristics from the very beginning. Furthermore, in some cases these characteristics were further improved indicating that the development team provided additional features towards the better addressing of the stakeholders' requirements.
- Critical characteristics such as the security and compatibility of the platform remained in high standards from the early releases. Nowadays, security has been one of most critical aspects of any software solution and the results indicate the robustness and effectiveness of the applied security mechanisms. Additionally, as the technologies change at such rapid rate, the high-level of compatibility of a platform is also a critical aspect for its success and sustainability in the future.

Sub-characteristics	KPIs	Calculation Type	Mandatory / Optional	1st evaluation	2nd evaluation	3rd evaluation	Comments
Functional suitability							
Functional completeness	Portion of completed User Stories	[Completed User Stories] / [Iteration Cycle of User Stories] * 100%	M	100%	100%	100%	All use cases planned for each version were successfully executed.
Functional correctness	Portion of User Stories without reported bugs	[Completed User Stories without bugs] / [Iteration Cycle of User Stories] * 100%	M	90%	93%	95%	During each implementation phase a small series of bugs were identified and they were successfully addressed.
Functional appropriateness	Straightforward task accomplishment	Are tasks completed without the use of unnecessary steps? [Yes/No]	O	No	No	No	Due to the nature of the accomplished tasks, assistance from the respective persons is required in some cases.
Performance efficiency							
Time behaviour	Average latency	[Total response time] / [Number of requests]	M	~1.2 sec	~1.1 sec	~1.0 sec	Average latency was measured with tools such as Chrome Dev Tools.
	Throughput	[Total Number of Kilobytes] / [Total Time of Operation]	M	~ 270 KB/sec.	~ 300 KB/sec.	~ 300 KB/sec.	Value documented while previewing files and downloading files.

Resource utilisation	Mean CPU Utilisation	[Σ[%CPU utilisation probes]] / [Number of probes]	M	<45%	<40%	<38%	Based on the resource monitoring tool of the platform
	Mean memory usage	[Σ[RAM Megabytes used in each probe]] / [Number of probes]	M	<20%	<18%	<20%	Based on the resource monitoring tool of the platform
	Maximum memory usage	Maximum % RAM Memory utilisation recorded	M	37%	40%	42%	Based on the resource monitoring tool of the platform
	Maximum processing power used	Maximum % CPU utilisation recorded	M	90%	90%	90%	As the resource management is performed by YARN (see deliverable D3.5), the appropriate resource allocation is always performed according to the provided configuration.
Capacity	Maximum file size upload	Total number of Kilobytes of files	M	250MB	450MB	5.1 GB	Note: This is size of the current biggest individual file available.
	Maximum file system size[1]	Total number of Kilobytes of files	M	76GB	89GB	163.5 GB	Note: This is the current size of HopsFS that can

[1] AEGIS platform utilises the distributed file system HopsFS. Thus, the database size metric was modified.

							scale according to the needs of the project.
Compatibility							
Co-existence	Ability to Co-Exist (host in a single environment)	Can the AEGIS platform operate in shared environment? [Yes/No]	O	Yes	Yes	Yes	
Interoperability	% of APIs coverage	[Number of integrated systems exposing or consuming data through API] / [Total number of integrated systems] * 100%	M	100%	100%	100%	All integrated components / services are integrated through APIs
	Ability to handle different datasets	Can the AEGIS platform consume datasets from different formats (e.g. CSV, JSON, XML files)? [Yes/No]	M	Yes	Yes	Yes	No limitations on the file formats HopsFS can store. Files can be processed using the appropriate libraries by the data scientist.
		Can the AEGIS platform provide datasets in various formats (e.g.	M	Yes	Yes	Yes	No limitations on the file formats HopsFS can store and provide.

		CSV, JSON, XML files)? [Yes/No]					
Usability							
Appropriateness recognisability	% Positive feedback on appropriateness based on the available documentation	[Number of positive response] / [Total number of responses] * 100%	O	Not applicable.	80%	90%	The final version of the documentation of the platform is delivered as part of D4.4
Technical Learnability	% Coverage of features with learning documents	[Unique number of help documents mentioning a feature] / [Total number of features available] * 100%	M	Not applicable.	90%	100%	In the final version of the documentation all the features of the platform are properly documented.
Ease of Use	Dashboard availability	Is there an available dashboard or wizard with easy navigation? [Yes/No/Partially]	O	Partially	Partially	Yes	The platform offers an intuitive and user friendly UI
User error protection	% Coverage of input fields with error protection methods	[Number of error protected fields] / [Total number of critical input fields] * 100%	M	100%	100%	100%	All input fields in the UI are protected.

User interface aesthetics	% Positive feedback on user interface aesthetics poll	[Number of supported screens] / [Total number of different screens] * 100%	O	80%	85%	90%	The platform offers an intuitive and user friendly UI, with enhanced aesthetics and user experience.
	Responsiveness	[Number of supported screens]/[Total number of different screens] * 100%	M	100%	100%	100%	No inaccessible or malformed screens were identified.
Technical Accessibility	WCAG 2.0 Conformance Level[2]	[None/ A/ AA/ AAA]	M	A	A	A	
Reliability							
Maturity	Maximum Concurrent users	Maximum number of concurrent users recorded	M	27	31	40	
	Simultaneous requests	Maximum number of simultaneous requests	M	Max Containers allocated = 12, Max applications = 5 Concurrent application, requests to filesystem =	Max Containers allocated = 17, Max applications = 9 Concurrent application, requests to filesystem =	Max Containers allocated = 22, Max applications = 11 Concurrent application, requests to filesystem =	Based on the resource monitoring tool of the platform

[2] WCAG 2.0: https://www.w3.org/WAI/WCAG20/quickref/

				11 concurrent requests.	15 concurrent requests.	20 concurrent requests.	
Availability	% Monthly availability	[1-[Downtime in minutes] / [Total month minutes]] * 100%	M	>95%	~97%	~98%	All downtime recorded was due to infrastructure upgrades
	Success rate	[Number of correctly completed requests] / [Total number of requests]	M	~90%	~95%	~96%	All problematic requests were successfully addressed with bug fixing.
Fault tolerance	% of identified Software problems affecting the platform	[Critical Software Issues] / [Total number of Software faults detected] * 100%	M	~25%	~22%	~19%	All problems have been identified and fixed.
	% of identified Hardware problems affecting the platform	[Critical Hardware Issues] / [Total number of Hardware faults detected] * 100%	M	100%	100%	100%	Critical hardware issues were identified and fixed in short time.
Recoverability	Mean recovery time from Software problems	[Total recovering time from Software issues] / [Total number of Software issues in need of recovery]	M	~ 1 hour	~ 1 hour	~ 1 hour	

	Mean recovery time from Hardware problems	[Total recovering time from Hardware issues] / [Total number of Hardware issues in need of recovery]	M	~ 1 hour	~ 1 hour	~ 1 hour	
Security							
Confidentiality	Incidents of ownership changes and accessing prohibited data	Number of recorded incidents	M	None	None	None	
Integrity	Incidents of authentication mechanisms breaches	Number of recorded incidents	M	None	None	None	
Non-repudiation	% Activities reporting	[Number of log categories] / [Total number of system operations]	M	90%	95%	97%	The platform provides advanced logging mechanism
Accountability	User actions traceability	Are usernames included in each activity log entry uniquely? [Yes/No]	M	Yes	Yes	Yes	Logging mechanisms provide all the appropriate reporting information
Authenticity	Level of User authenticity	Can you identify that a subject is the one it claims to be? [Yes/ No/ Partially]	M	Yes	Yes	Yes	

Maintainability							
Modularity	% of modularity	[Number of components that can operated individually] / [Total number of components] * 100%	M	100%	100%	100%	
Reusability	% of reusable assets	[Number of assets that can or are reused] / [Total number of assets] * 100%	M	100%	100%	100%	
Analysability	Level of analysability	Can the changes in the performance of the AEGIS platform be efficiently evaluated after each upgrade? [Yes/No]	O	Yes	Yes	Yes	The system offers monitoring tools with performance indications.
Modifiability	% of update effectiveness	[Number of updates performed without operational issues] / [Total number of updates] * 100%	M	95%	95%	97%	The updates were performed successfully with minor issues
Testability	Level of testing	Are tests able to probe the behaviour of the	M	Yes	Yes	Yes	

		AEGIS platform? [Yes/No]					
Portability							
Adaptability	Mean number of errors per hardware change/ upgrade	[Total number of errors recorded] / [Total number of hardware changes]	M	None	None	None	
	Mean number of errors per software change/ update	[Total number of errors recorded] / [Total number of software changes]	M	<2%	<2%	<2%	Minor errors were identified and fixed.
Replaceability	% of software products replaceability within AEGIS platform	[Number of replaceable software components] / [Total number of used software components] * 100%	M	100%	100%	100%	

Table 3-1: Summary of the AEGIS platform quantitative evaluation results

3.2. Qualitative Evaluation

This section provides an overview on the results of the qualitative evaluation of the platform performed after each demonstrator development phase. Thereby interviews and/or small focus groups with data scientists in charge of demonstrator implementation as well as with demonstrator stakeholders (e.g. end users) have been conducted. In general, all data scientists in charge of demonstrator development have already gained experiences with different tools and languages for data analysis as well as experiences in analysing data within their own industrial domains. They are well qualified to execute the implementation of the demonstrators and to provide feedback on the AEGIS platform based on their experiences from using it.

The following summarized evaluation results have been gained after the implementation of the early demonstrator, by using a very early version of the AEGIS platform, providing only basic features to support data scientists in their demonstrator development. It can be said that the data scientists liked the ‘platform as a service concept’ (Hadoop file system / Spark / Zeppelin / Jupyter). It clearly makes sense and saves data scientists a lot of work (since setting such an environment up and maintaining it requires a lot of experience and infrastructure and should not be the work of data scientists). They see in AEGIS a great platform that combines a lot of things and capabilities that people working in the data science sector need and tools that are hot in the current data industry. However, for less experienced users (e.g. business users with no background in programming) the platform might not seem as the most appropriate tool to solve their challenges. Furthermore, the extendibility of the platform was questioned, as extending the platform would in its current state just be possible with notebooks, or tools that are outside the platform. Data scientists working with geo-referenced data (as for example in automotive) were interested in been provided with more specialised tools for geographic data visualisation by the platform. Some data scientists criticised user guidance and usability of the platform as well as the platform performance, which should be both further improved. For the first version of the demonstrator, the required code was mainly developed offline (in R or Python) first and then ported to python and/or simply copied into a notebook on the platform. So far, the data scientists were already satisfied with the platform, its capabilities and the current development process (besides some technical issues where the platform development team had to be contacted).

Challenges and recommendations made after the first evaluation mainly concerned an improvement of the user interface of the platform, an improvement of the standard settings of the platform with respect to performance, a user guide for the platform to reduce initial efforts of trying out things on the platform, improved capabilities for geo-referenced data visualisation, and a redesign of the platform’s landing page.

The following summarized evaluation results have been gained after the implementation of the medium demonstrator, where a more mature version of the AEGIS platform has been used providing improved features to support data scientists.

Data scientists acknowledged the improvement of the platform in terms of comfortability and usability of the platform, as well as usefulness of the platform for performing the tasks required for the implementation of the demonstrator. Offerings of the platform were significantly increased with the list of predefined tools such as the Query Builder, the Visualiser and the Algorithm Execution Container. They also acknowledged the improved look-and-feel of the platform, which was a major issue in the first evaluation. However, they still see further potential for graphical user interface improvement and enhancement. Furthermore, the improved Visualiser offers more capabilities for visualising geo-referenced data (e.g. heatmaps or enhancements to marker visualisations), which was requested by some data scientists (e.g. from automotive demonstrator). Another perceived major advantage of the platform update was the improved configuration of Jupyter, where certain options like adding execution memory to a project can be more easily accessed. The shift from Zeppelin to Jupyter was perceived to be a good decision as Jupyter seems to run more smoothly. One issue which has been voiced by the data scientists was the cooperation between the platform team and the demonstrator developers, which could be further improved. Demonstrators must meet certain deadlines for implementation and evaluation (and so does the platform development). If critical platform updates conflict with demonstrator deadlines, demonstrator developers face a challenge. Furthermore, the update process of the platform could be improved, because in case of an update, demonstrator owners must (re)create their user accounts on the platform, then re-create the project, and the datasets (folders), before data (files) can be restored into the correct datasets (folders) by the platform team. Finally, cooperation between the platform team and the demonstrator developers was perceived to be good, although some deadlines were not met causing some issues in the demonstrator's processes. A closer coordination between the teams should have been considered.

Challenges and recommendations made after the second evaluation mainly concerned the alignment of the update process of the platform/testbed with the milestones of demonstrator development, a general improvement of the update process of the platform, the fine-tuning of the platform's user interface so that it becomes even more appealing to the data scientists, automated enabling of Query Builder, Visualiser and Algorithm Execution Container tools within a project for users, improved or better documented accessing and processing of available datasets from the Jupyter notebook, improved navigation through the various functionalities and services offered by the platform, automated process of restoring user accounts/datasets following a platform upgrade, and automated import of notebooks when creating a new project.

The following summarized evaluation results have been gained after the implementation of the advanced demonstrator, where the final version of the AEGIS platform has been used. Having used the final version of the AEGIS platform for demonstrator implementation, the data scientists' perception of the platform in terms of usefulness, usability and ease-of-use has again improved significantly. The redesign of the user interface of the platform was well received, and the data scientists acknowledged that the new interface is far more appealing, while also offering an improved user experience. It even facilitates the usage of the platform for non-experienced users focused on the data science sector (e.g. who show interactive visualisations provided by the visualiser). Furthermore, the documentation of the platform has improved significantly and reduces the training period for new platform users. The services of the platform have become more customisable and can be better tailored to the developers' or data scientists' needs. Furthermore, the collaboration between the demonstrator developers and the platform developers was further improved. Technical challenges during demonstrator implementation, such as restarting the core or increasing the computing power or RAM required for data processing workflows, have been quickly solved by the platform team increasing both user satisfaction and trust. Because the AEGIS platform received several refinements and updates during the demonstrator development period, minor issues were faced, especially in the user interface environment that was completely redesigned, and developers had to wait for the stable intermediate release of the platform before they could continue their development activities, which is understandable as both, the demonstrators, and the platform, are developed in parallel. Finally, the upgrade process has been further improved and most of the manual processes that were identified in the previous version were resolved, which was a major concern during the first and second evaluation. Thanks to the weekly technical telephone conferences demonstrator deadlines have been better aligned with platform development deadlines and updates.

Challenges and recommendations made after the third evaluation mainly concerned to remove any manual intervention in the update process of the platform for users, provide a guideline on how to best use the platform for own data science projects and expand the documentation, implement minor improvements in the user interface based on further experiences made while using the platform, further highlight the capabilities of the platform such as pre-defined or scheduled job execution.

3.3. Platform Security Assessment

As part of the final evaluation of the AEGIS platform, and as per the comments received during the first project review, the consortium decided to perform a thorough security assessment of the platform in order to ensure that the security mechanisms that are incorporated in the platform are operating efficiently and effectively. Furthermore, the scope of the security assessment conducted was to identify security weaknesses and deficiencies in order to be resolved and to provide possible recommendations on how to correct these identified weaknesses or deficiencies with the aim to reduce or eliminate any vulnerabilities.

In detail, the security assessment was performed by the specialised security team of UBITECH that analysed the security standards and tools utilised in the platform in order to: (a) identify any possible vulnerabilities, risks or threats that may result in a possible exploitation that may result in an intentional or unintentional compromise of the platform and (b) to identify the ways that these vulnerabilities, risks or threats can be exploited in order to defeat the deployed security mechanisms of the platform. A thorough penetration test was performed, which included amongst others the following methods (attacks): SQL injection, File traversal, Server-Side Request Forgery (SSRF), Insecure Direct Object References (IDOR), Carriage Return Line Feed (CRLF) injection, Cross-Site Scripting (XSS), Remote Code Execution (RCE), Cross-Site Request Forgery (CSRF) and Open Redirect.

During this thorough test, the platform performed well in the attempted attacks, as all of them failed, with the exception of the SSRF method. More precisely, during testing, a set of endpoints was discovered with a SSRF vulnerability, which could allow the attacker to force the server to perform requests on their behalf, for example invoking internal services. On the following request, the development team tried to mitigate this by searching if the word /static/ is included in the URL, but that can easily be bypassed as shown below:

```
REQUEST

GET /hopsworks-api/yarnui/http://IP:8080/test;/static/ HTTP/1.1

RESPONSE

  HTTP/1.1 200 OK
 <!DOCTYPE html PUBLIC "-//W3C//DTD HTML 3.2 Final//EN"><html>
 <title>Directory listing for /</title>
 <body>
```

```
<h2>Directory listing for /</h2>
<hr>
<ul>
<li><a href="/hopsworks-api/yarnui/http://IP:8080/PostgreSQL.564362641">PostgreSQL.564362
</ul>
<hr>
</body>
</html>
```

As it can be seen from the example, the request was performed on a test server that the security team of UBITECH controlled, however if the IP is changed to “127.0.0.1” requests can be performed on services listening on localhost. The spotted vulnerability was reported immediately to the development team of the AEGIS platform and at the time of writing of this deliverable, the development team has already started working on the corrective actions towards the elimination of this vulnerability.

4. AEGIS Platform Documentation And Adoption Guidelines

4.1. AEGIS Platform Usage Documentation

The AEGIS platform provides data management, analysis, and processing as well as user management through the use of Hopsworks integrated services. On top of that, the AEGIS platform provides tools for data anonymization and cleansing as well as data harvesting. AEGIS uses the notions of Project, Dataset, and User to enable multi-tenancy within the context of data management. The AEGIS platform is equipped with a list of services for data processing including data parallel processing frameworks such as MapReduce, Spark, and Flink, as well as interactive analytics using interactive notebooks such as Jupyter. In addition, it offers real time analytics using the Kafka service, full-text search using the Elasticsearch service, as well as development and serving of deep learning models using the Tensorflow service. Moreover, the AEGIS platform provides tools for data validation, visualisation, exploration, and processing through the use of the AEGIS integrated notebooks.

4.1.1. AEGIS Usage

The AEGIS platform can be deployed on-premise or in a cloud environment. Once deployed, the AEGIS platform is accessible through a web interface. A user can create an account that is later verified by an administrator. Once a user is verified, he/she can create projects on the platform by clicking on the New Project button as shown in Figure 4-1.

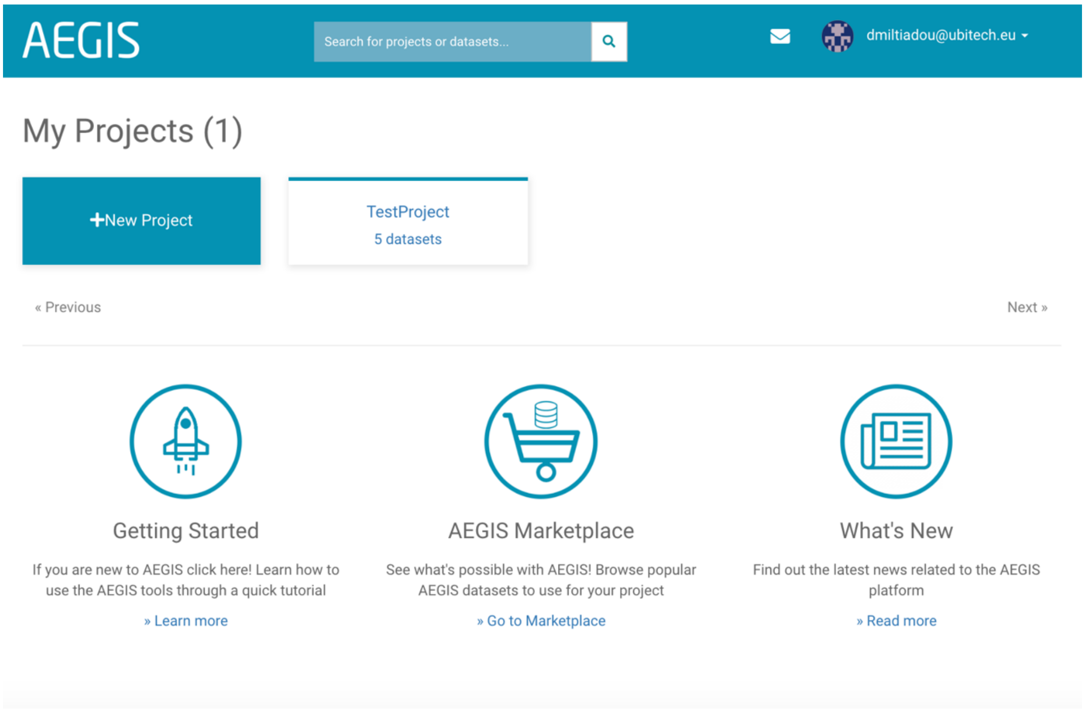

AEGIS
Search for projects or datasets...
dmiltiadou@ubitech.eu
My Projects (1)
+New Project
TestProject
5 datasets
« Previous
Next »
Getting Started
If you are new to AEGIS click here! Learn how to use the AEGIS tools through a quick tutorial
» Learn more
AEGIS Marketplace
See what's possible with AEGIS! Browse popular AEGIS datasets to use for your project
» Go to Marketplace
What's New
Find out the latest news related to the AEGIS platform
» Read more

Figure 4-1: AEGIS User Main Page

Once the project is created, the user can navigate to the project main page as shown in Figure 4-2, browse its default datasets, as well as create new datasets and upload data. The project is pre-populated with a list of default datasets such as Logs, Resources, and Jupyter. The left side pane shows the different project associated settings and tools. Firstly, the user can manage the project general settings such as the description of the project by navigating to the Settings page.

Additionally, the user can view, add, and delete members of the project as well as changing members' roles using the Members page. The activity stream page provides an overview over the changes happening during the project lifetime since its creation such as creating, sharing, and deleting a dataset. The Extended Metadata tool is part of the AEGIS metadata service and provides a user interface for users to associate metadata to the project. The AEGIS tools page provides the three default integrated notebooks offered by the AEGIS platform. The Jobs page offers the ability to create and manage data processing jobs using Spark and Flink. The Kafka page offers the ability to create and manage Kafka topics and their associated schemas. The Model Serving page provides the ability to serve your deep-learning models using the Tensorflow serving.

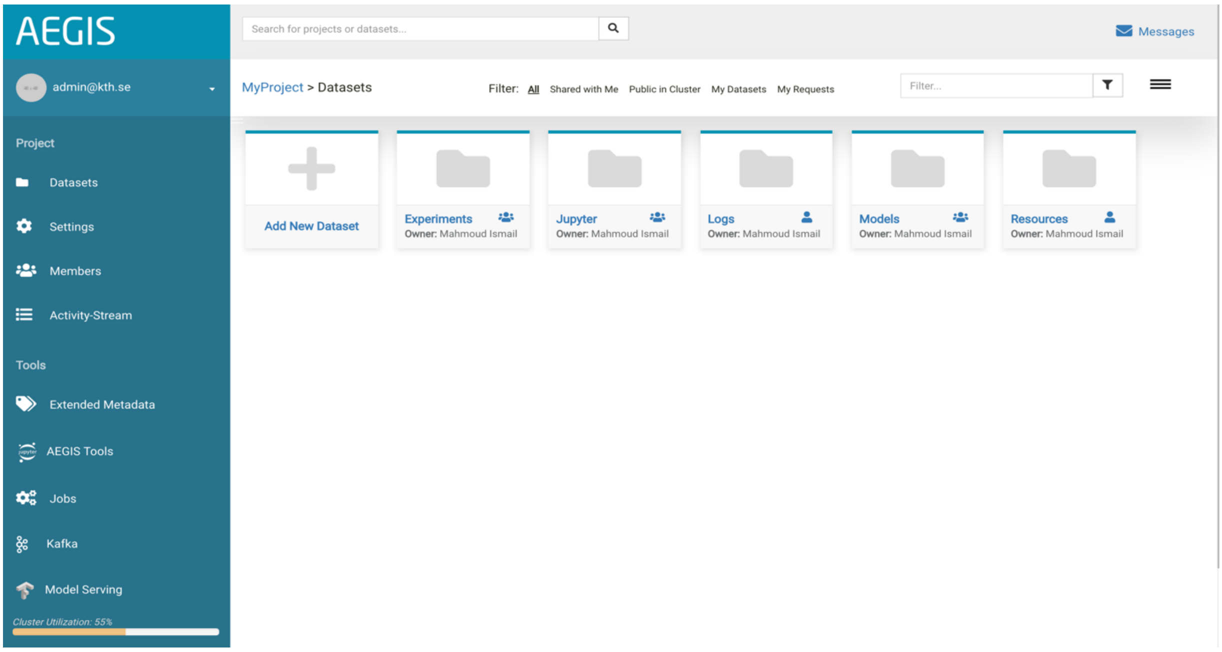
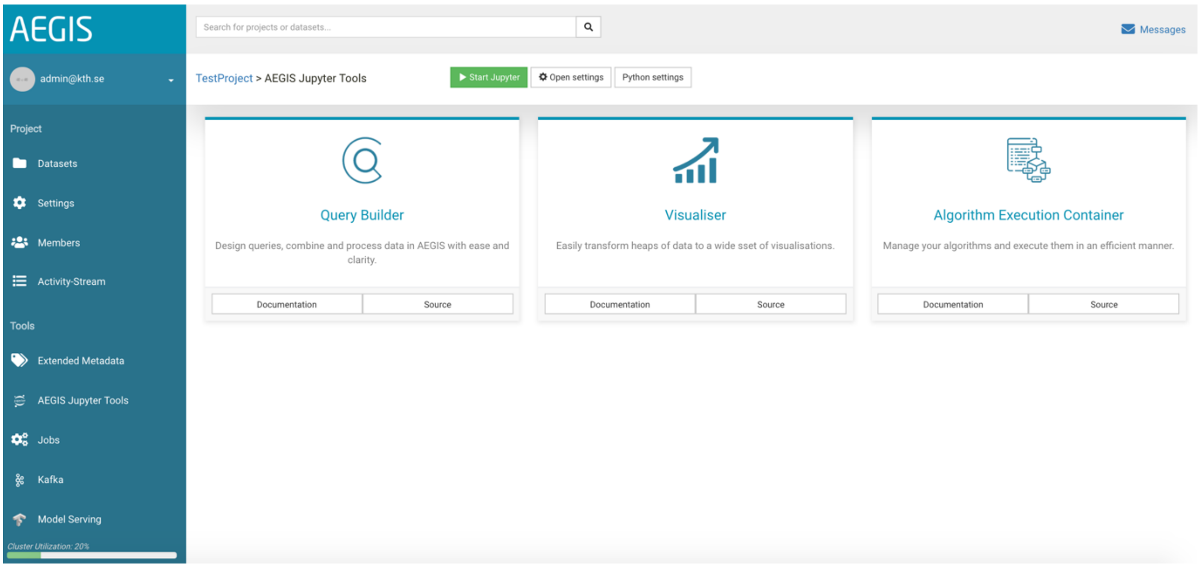

AEGIS
Search for projects or datasets...
Messages
admin@kth.se
MyProject > Datasets
Filter: All Shared with Me Public in Cluster My Datasets My Requests
Filter...
Project
Datasets
Settings
Members
Activity-Stream
Tools
Extended Metadata
AEGIS Tools
Jobs
Kafka
Model Serving
Cluster Utilization: 55%
Add New Dataset
Experiments Owner: Mahmoud Ismail
Jupyter Owner: Mahmoud Ismail
Logs Owner: Mahmoud Ismail
Models Owner: Mahmoud Ismail
Resources Owner: Mahmoud Ismail

Figure 4-2: Project Main Page

4.1.2. AEGIS Jupyter Tools

The AEGIS Jupyter tools contain the three default tools provided by AEGIS to help users to easily preform data exploration, validation, processing, and visualisation, namely the Query Builder, the Visualiser and the Algorithm Execution Container, that are presented in the paragraphs below.



Query Builder: Design queries, combine and process data in AEGIS with ease and clarity.

Visualiser: Easily transform heaps of data to a wide sset of visualisations.

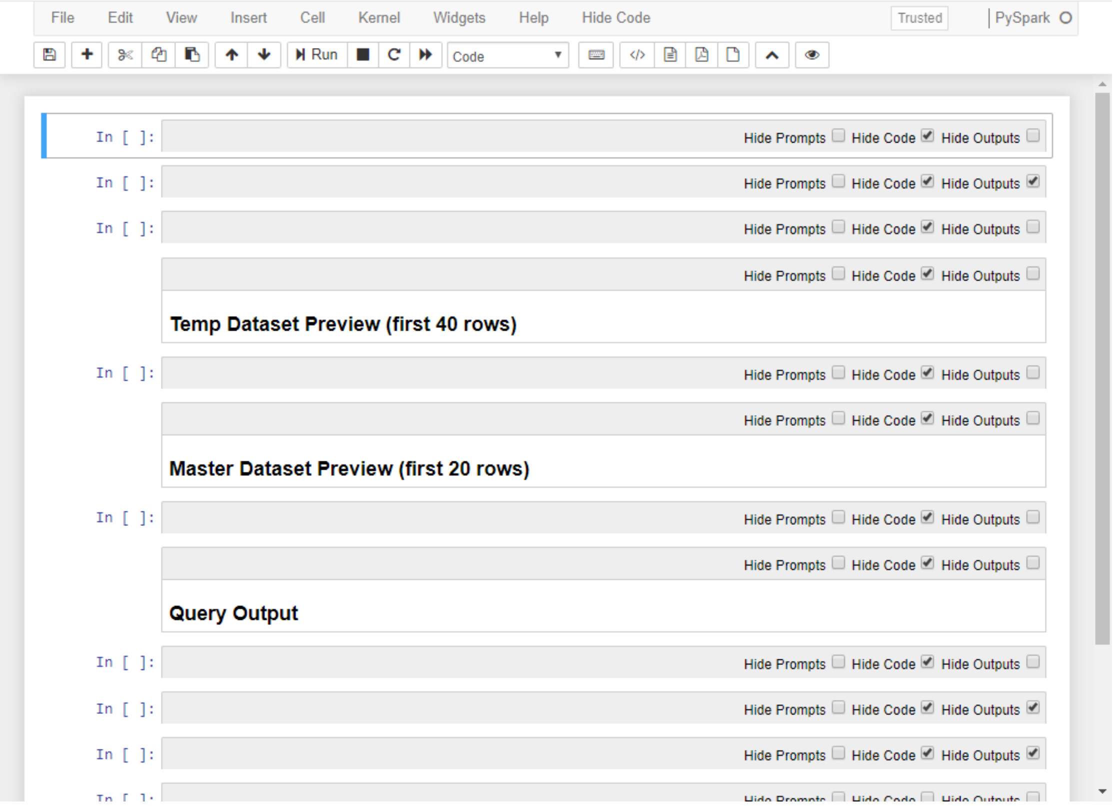
Algorithm Execution Container: Manage your algorithms and execute them in an efficient manner.

Figure 4-3: AEGIS Tools

Query Builder

The main purpose of the Query Builder to enable the capability of interactively defining and executing queries on top of the data available in the AEGIS platform.

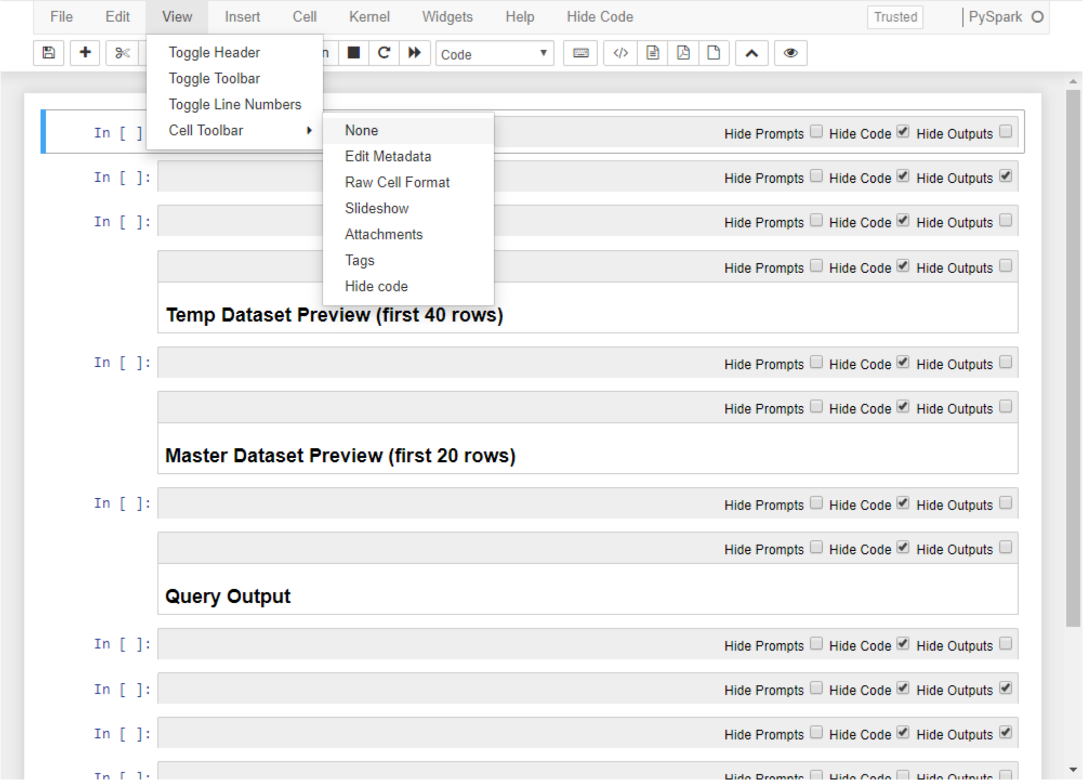
Once the user clicks on the Query Builder tool, a preconfigured Jupyter notebook is opened as shown in the following figure.



```
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs □
          Hide Prompts □ Hide Code ☑ Hide Outputs □
          Temp Dataset Preview (first 40 rows)
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs □
          Hide Prompts □ Hide Code ☑ Hide Outputs □
          Master Dataset Preview (first 20 rows)
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs □
          Hide Prompts □ Hide Code ☑ Hide Outputs □
          Query Output
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:   Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:   Hide Prompts □ Hide Code □ Hide Outputs □
```

Figure 4-4: Query Builder - Initial state

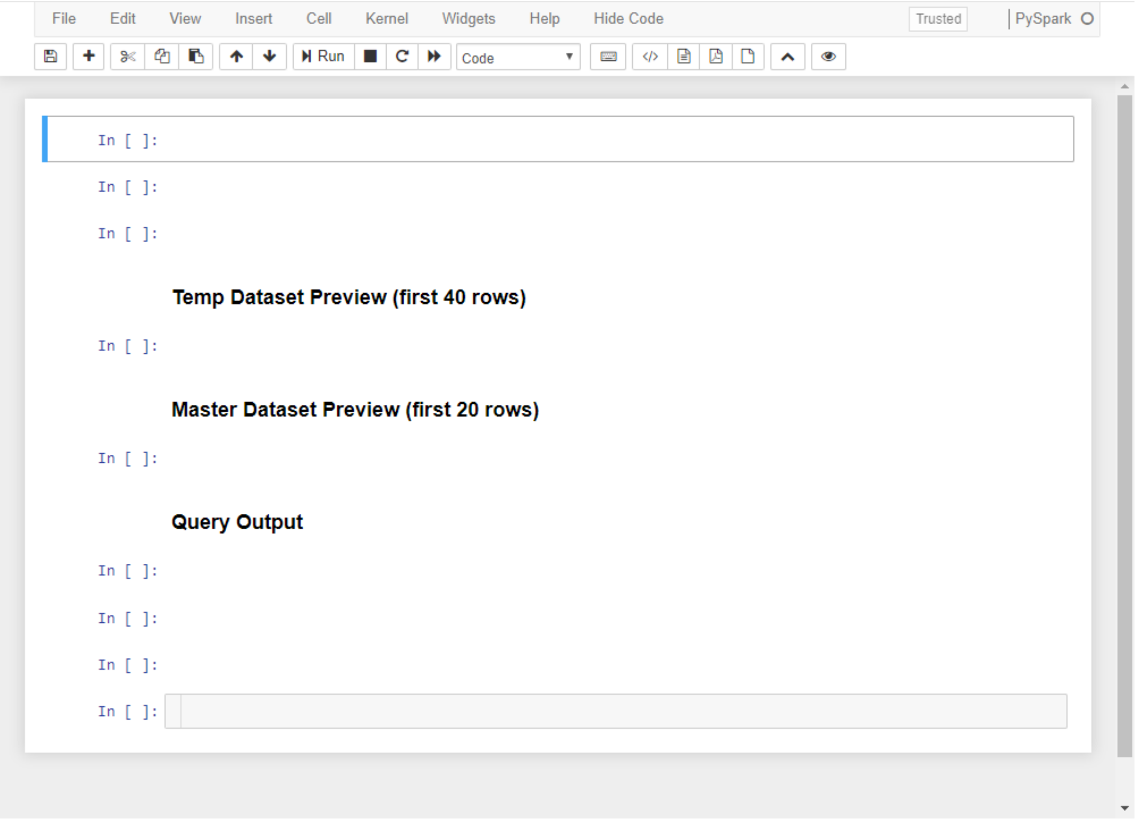
The Query Builder is now in its initial state. All code cells are hidden using the hide code plugin and all output has been cleared. If the user is not interested in reviewing or altering the code (it is not required in the standard workflow) he/she may proceed with hiding the cell toolbars. In order to do that, the user can select "View" from the menu, then "Cell Toolbar" and click on the "None" option, as shown below.



```
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs □
         Hide Prompts □ Hide Code ☑ Hide Outputs □
         Temp Dataset Preview (first 40 rows)
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs □
         Hide Prompts □ Hide Code ☑ Hide Outputs □
         Master Dataset Preview (first 20 rows)
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs □
         Hide Prompts □ Hide Code ☑ Hide Outputs □
         Query Output
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:  Hide Prompts □ Hide Code ☑ Hide Outputs ☑
In [ ]:  Hide Prompts □ Hide Code □ Hide Outputs □
```

Figure 4-5: Query Builder - Hide code

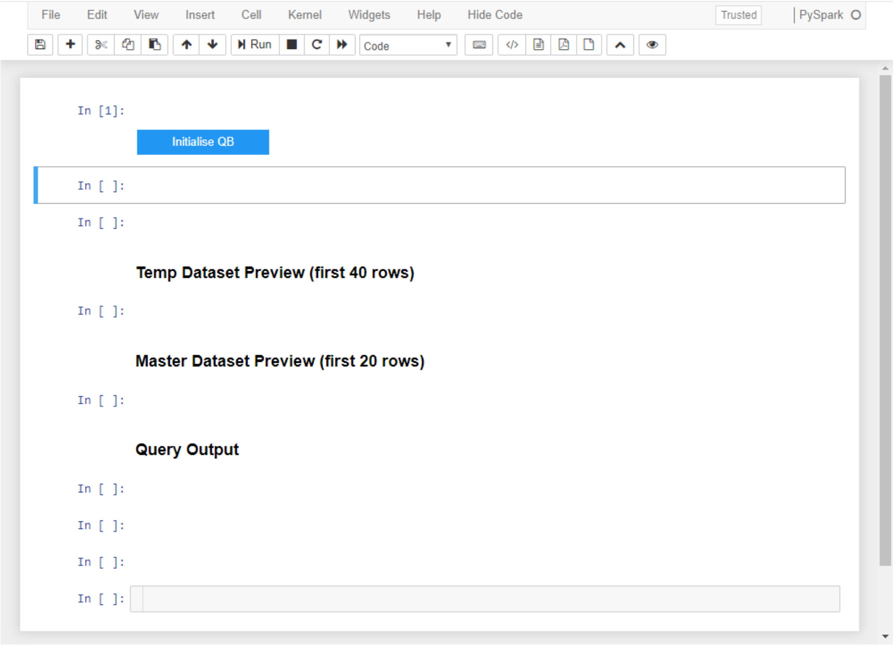
By clicking on this option, the user is able to see the standard view of the Query Builder as depicted in the following screen.



```
In [ ]:

In [ ]:

In [ ]:

Temp Dataset Preview (first 40 rows)

In [ ]:

Master Dataset Preview (first 20 rows)

In [ ]:

Query Output

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```

Figure 4-6: Query Builder - Standard view

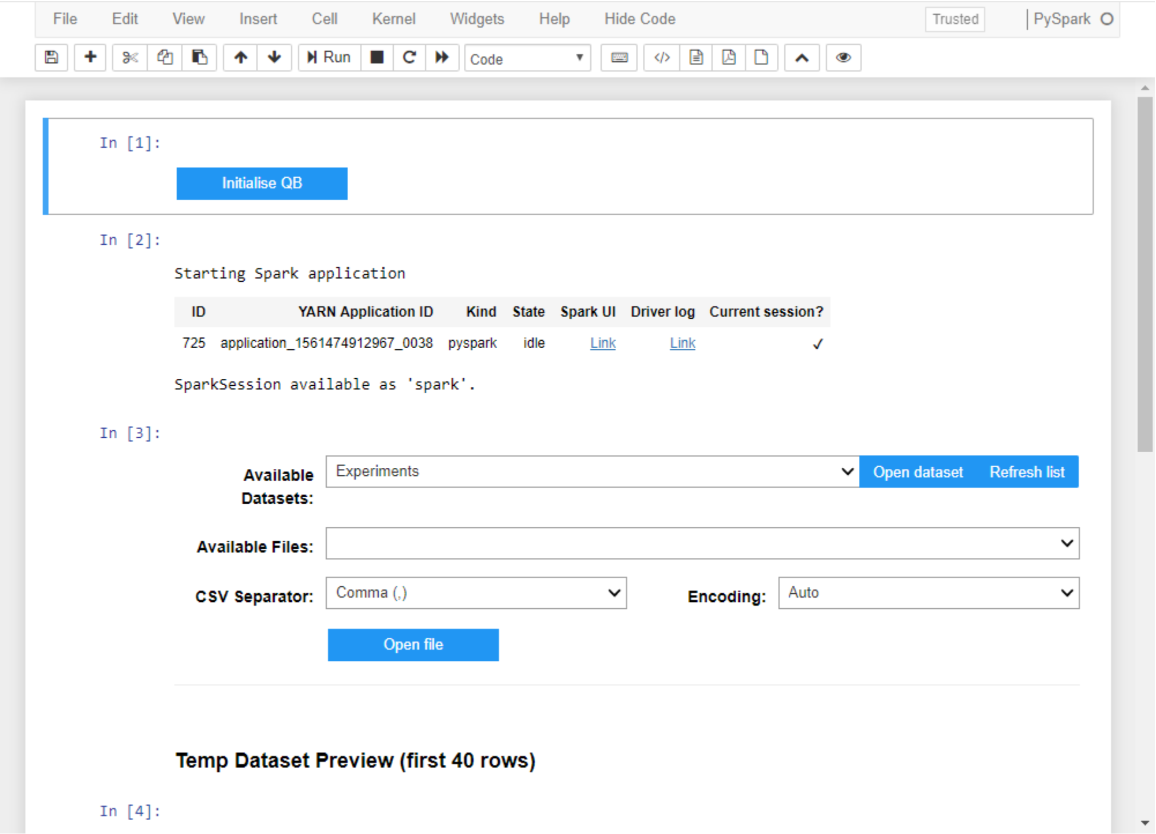
In order to proceed with the usage of the Query Builder, the user must initialize the notebook. To perform this the user must run the first cell. To achieve this, the user must make sure that the first cell is selected and either click on the "Run" button from the top Jupyter menu, or press Ctrl + Enter. At this point, the user should be able to see a button appearing as shown below.



```
In [1]:

In [ ]:

In [ ]:

Temp Dataset Preview (first 40 rows)

In [ ]:

Master Dataset Preview (first 20 rows)

In [ ]:

Query Output

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```

Figure 4-7: Initialise Query Builder

The user should click on the "Initialise QB" button and wait for the execution to stop. When this is done, the user should see a screen like the one below, where all cell prompts have been replaced by numbers and the file selection UI has appeared.



```
In [1]:

In [2]:
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
725	application_1561474912967_0038	pyspark	idle	Link	Link	✓

```
SparkSession available as 'spark'.

In [3]:
```

Available Datasets: Experiments

Available Files:

CSV Separator: Comma (,)   Encoding: Auto

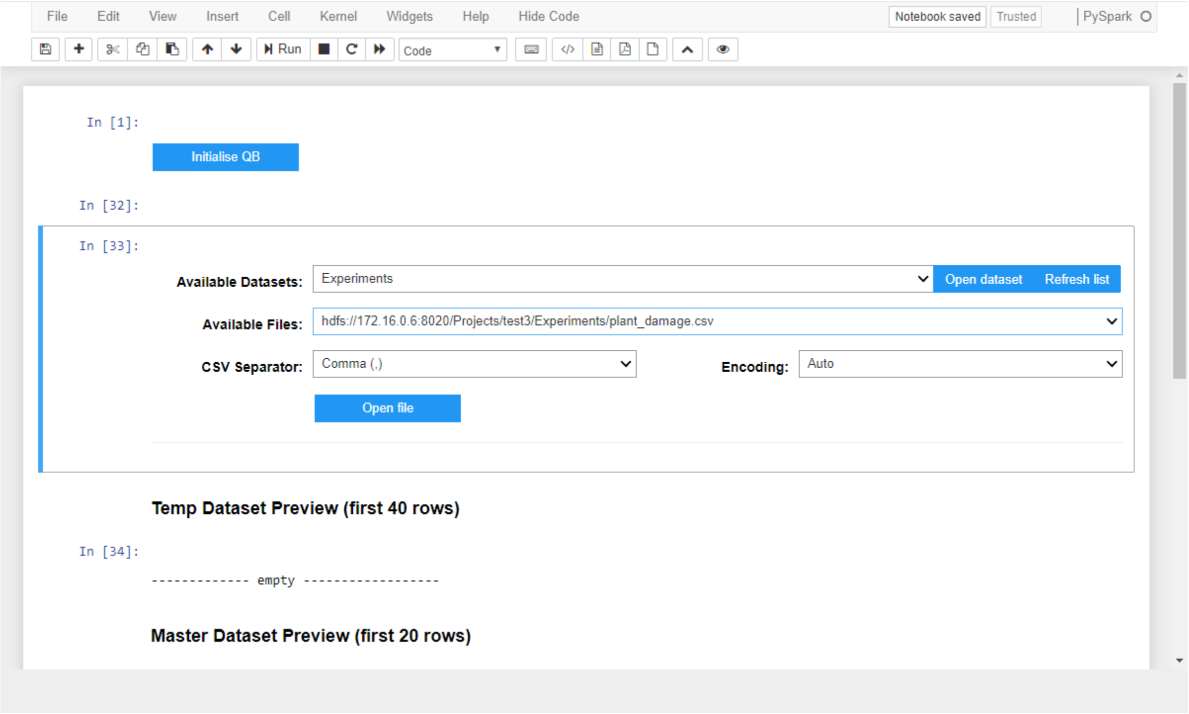
Temp Dataset Preview (first 40 rows)

```
In [4]:
```

Figure 4-8: Query Builder options

Note that the output regarding the Spark application will only appear when the interpreter is started. This means that if the "Initialise QB" button is pressed again, the message will disappear, but the tool will be still functional. Now the query builder in initialized and the user can start working with different files.

User can see a list of all "Available Datasets" in the current project. The list can also be refreshed. Select the desired dataset and click "Open dataset", this will populate the "Available Files" list.



Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Temp Dataset Preview (first 40 rows)

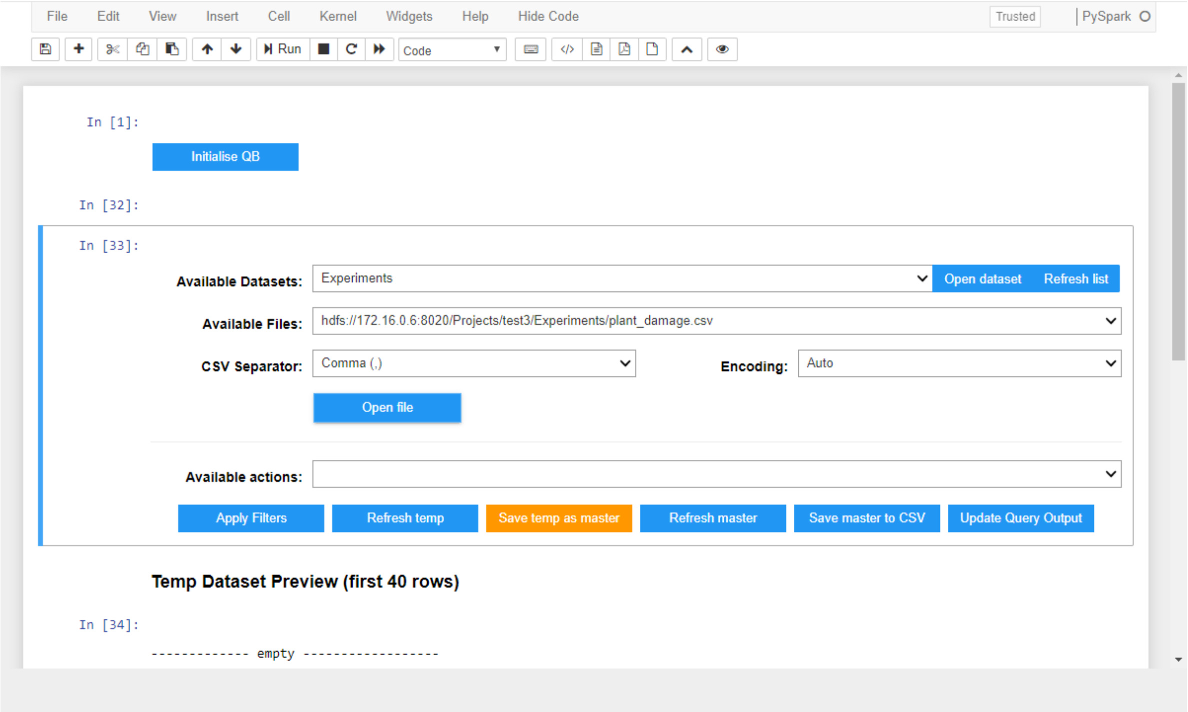
```
------------ empty -----------------
```

Master Dataset Preview (first 20 rows)

Figure 4-9: Query Builder Local Browser

Once the user selected a file and presses the "Open file"button, the file is loaded (as a PySpark Dataframe) and become available for further processing/querying.

Dataframes are essentially data tables. At each moment, Query Builder has two Dataframes: the temporary dataframe (TempDF) which holds the result of the last performed action, i.e. the contents of a file or the result of applying a filter/query on the previous contents of TempDF and the master dataframe (MasterDF) which is explicitly updated by pressing a button that moves the TempDF contents to the MasterDF.



In [1]:

In [32]:

In [33]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Available actions:

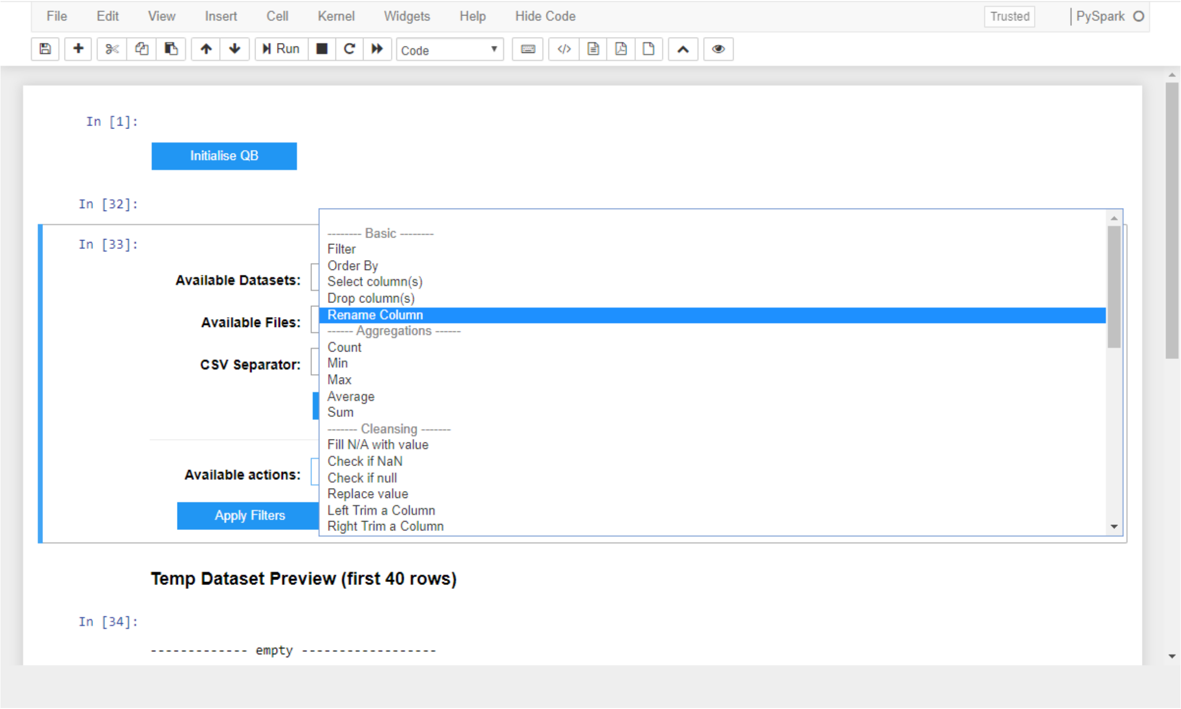
Temp Dataset Preview (first 40 rows)

In [34]:

```
------------- empty ------------------
```

Figure 4-10: Query Builder - Data filter and processing options (1)

When a file is loaded, a form appears with several options for simple data filtering and processing.



In [1]:

In [32]:

In [33]:

Available Datasets:

Available Files:

CSV Separator:

Available actions: Rename Column

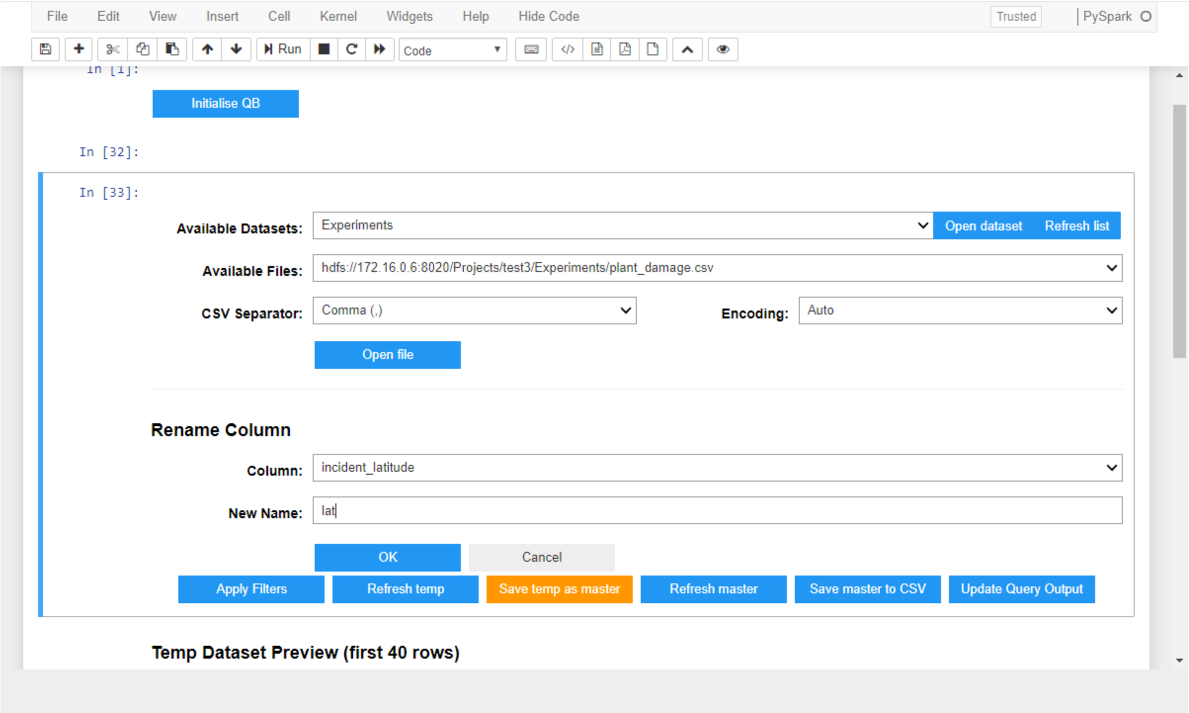
Temp Dataset Preview (first 40 rows)

In [34]:

```
------------- empty ------------------
```

Figure 4-11: Query Builder - Data filter and processing options (2)

Each selected action will open a new form with the necessary parameters. Once the user fills the required information and presses OK a new filter will be added in the queue to be applied when the user wants. As an example, if the user chose the "Rename Column" filter he/she will then need to select which column to rename and what should be the new name.



In [32]:

In [33]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Rename Column

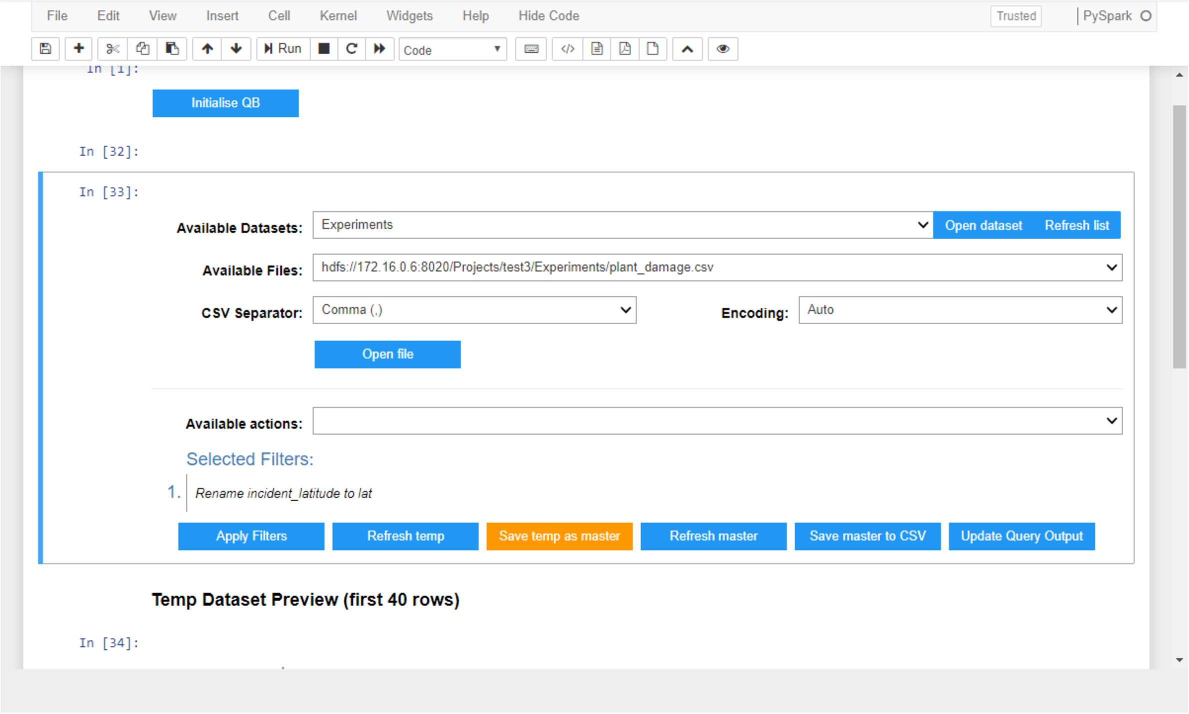
Column: incident_latitude

New Name: lat

Temp Dataset Preview (first 40 rows)

Figure 4-12: Query Builder - Rename column

Once the user presses “OK”, the filter will be added in the queue, under the "Selected Filters" label.



In [32]:

In [33]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Available actions:

Selected Filters:

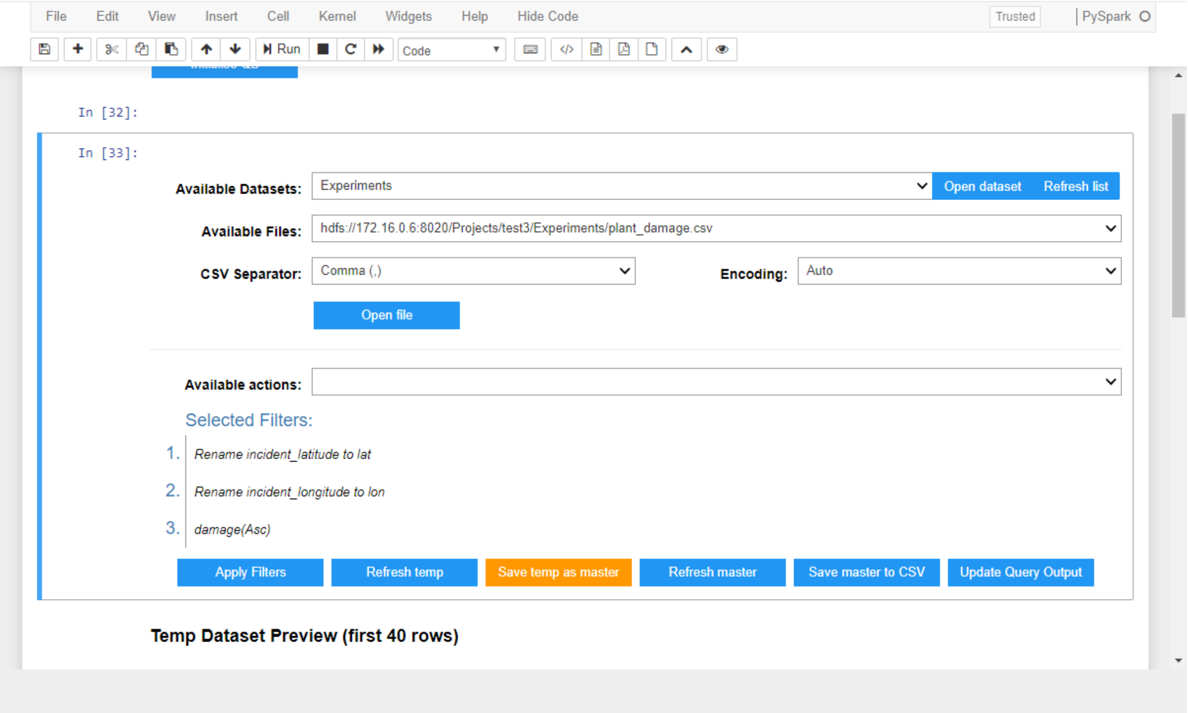
1. *Rename incident_latitude to lat*

Temp Dataset Preview (first 40 rows)

In [34]:

Figure 4-13: Query Builder - Selected filters

The user may add more than one filter before executing them. They will be added in the same list:



In [32]:

In [33]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Available actions:

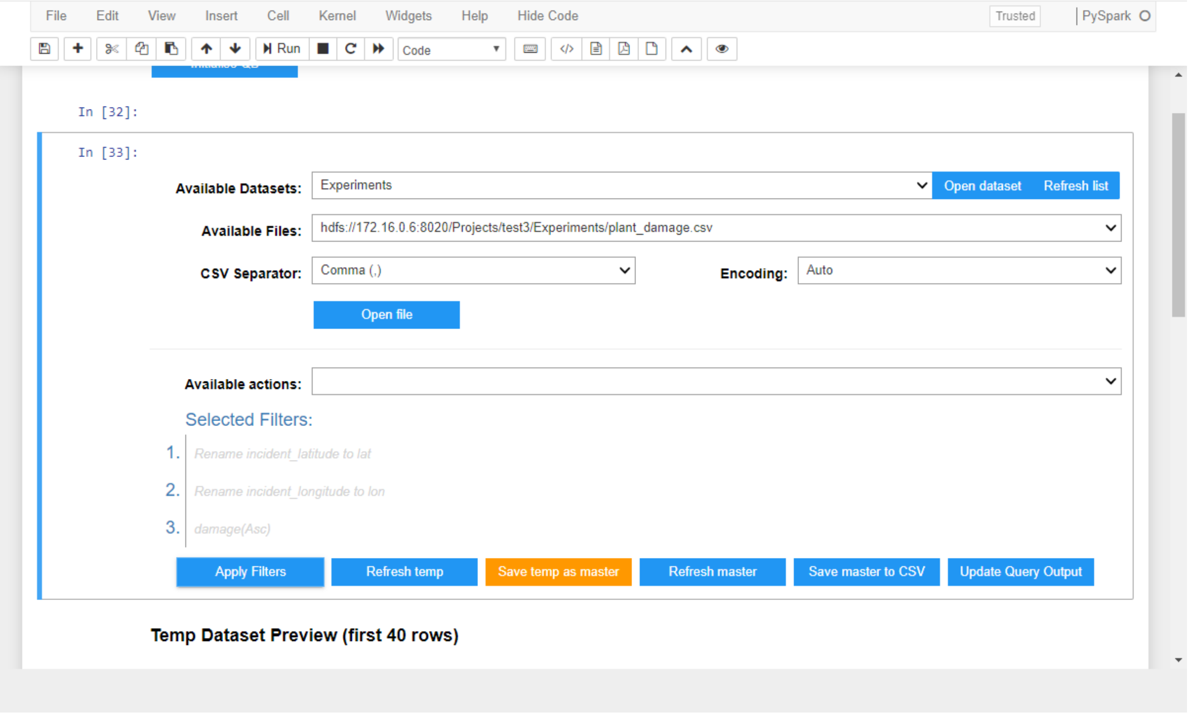
Selected Filters:

1. *Rename incident_latitude to lat*
2. *Rename incident_longitude to lon*
3. *damage(Asc)*

Temp Dataset Preview (first 40 rows)

Figure 4-14: Query Builder - Multiple filters

When the user decides to apply them, he/she should click on the "Apply Filters" button. When the execution is finished, all successfully applied filters will change colour:



In [32]:

In [33]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Available actions:

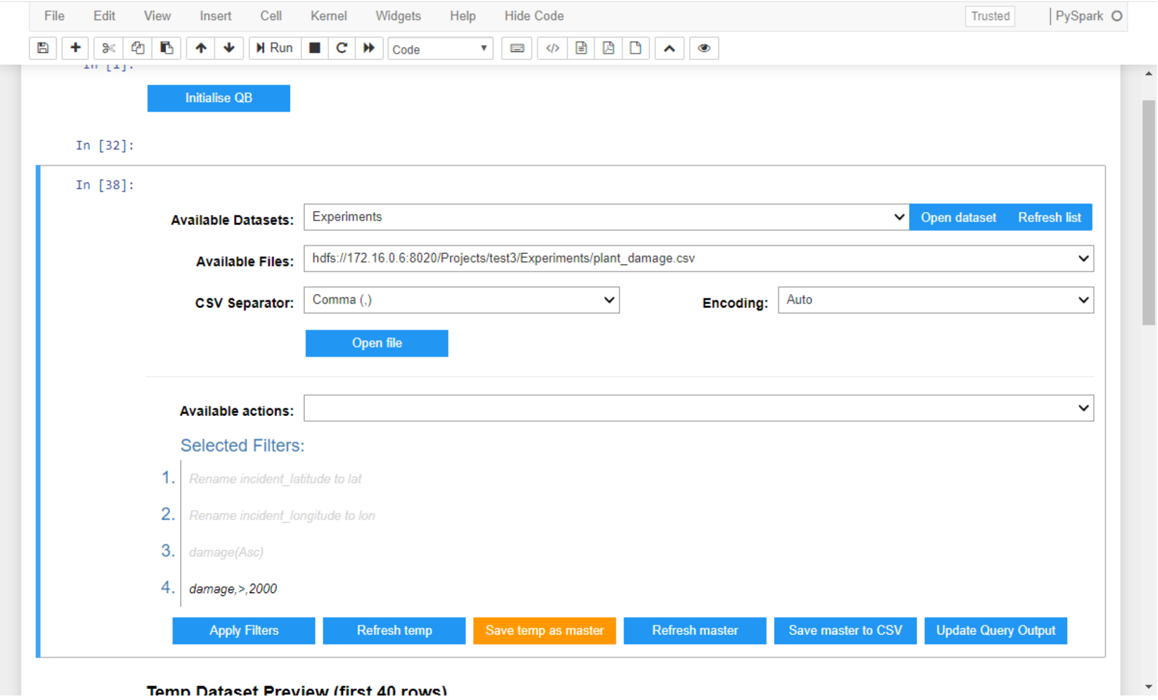
Selected Filters:

1. *Rename incident_latitude to lat*
2. *Rename incident_longitude to lon*
3. *damage(Asc)*

Temp Dataset Preview (first 40 rows)

Figure 4-15: Query Builder - Apply filters

The user may continue adding new filters after that point. The list will show which filters have been already applied and which are pending:



In [32]:

In [38]:

Available Datasets: Experiments

Available Files: hdfs://172.16.0.6:8020/Projects/test3/Experiments/plant_damage.csv

CSV Separator: Comma (,)

Encoding: Auto

Available actions:

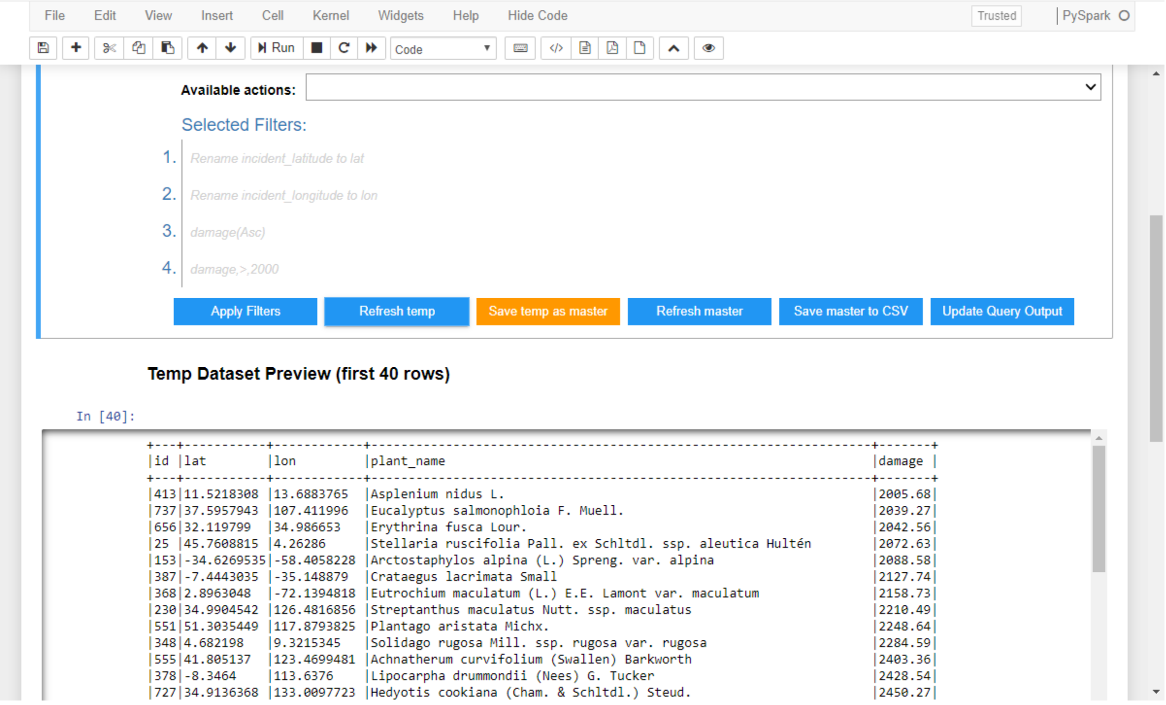
Selected Filters:

1. Rename incident_latitude to lat
2. Rename incident_longitude to lon
3. damage(Asc)
4. damage,>,2000

Temp Dataset Preview (first 40 rows)

Figure 4-16: Query Builder - Add more filters

By pressing the "Refresh temp" button, the user will get the first 40 lines of the TempDF. This is useful both when the user first loads a file but also to review the result of applying some filters.



Available actions:

Selected Filters:

1. Rename incident_latitude to lat
2. Rename incident_longitude to lon
3. damage(Asc)
4. damage,>,2000

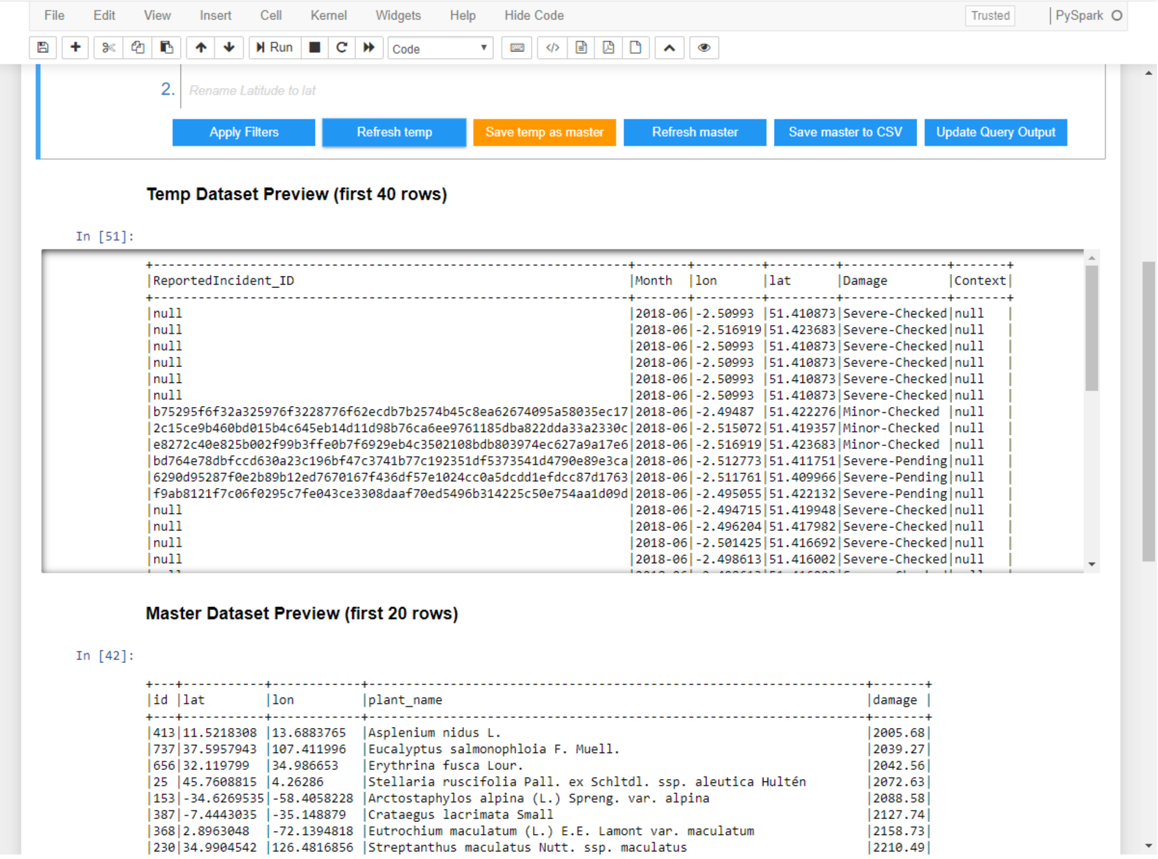
Temp Dataset Preview (first 40 rows)

In [40]:

```
+---+-----------+-----------+-------------------------------------------------------------------+-------+
|id |lat        |lon        |plant_name                                                         |damage |
+---+-----------+-----------+-------------------------------------------------------------------+-------+
|413|11.5218308 |13.6883765 |Asplenium nidus L.                                                 |2005.68|
|737|37.5957943 |107.411996 |Eucalyptus salmonophloia F. Muell.                                 |2039.27|
|656|32.119799  |34.986653  |Erythrina fusca Lour.                                              |2042.56|
|25 |45.7608815 |4.26286    |Stellaria ruscifolia Pall. ex Schltdl. ssp. aleutica Hultén        |2072.63|
|153|-34.6269535|-58.4058228|Arctostaphylos alpina (L.) Spreng. var. alpina                     |2088.58|
|387|-7.4443035 |-35.148879 |Crataegus lacrimata Small                                          |2127.74|
|368|2.8963048  |-72.1394818|Eutrochium maculatum (L.) E.E. Lamont var. maculatum               |2158.73|
|230|34.9904542 |126.4816856|Streptanthus maculatus Nutt. ssp. maculatus                        |2210.49|
|551|51.3035449 |117.8793825|Plantago aristata Michx.                                           |2248.64|
|348|4.682198   |9.3215345  |Solidago rugosa Mill. ssp. rugosa var. rugosa                      |2284.59|
|555|41.805137  |123.4699481|Achnatherum curvifolium (Swallen) Barkworth                        |2403.36|
|378|-8.3464    |113.6376   |Lipocarpha drummondii (Nees) G. Tucker                             |2428.54|
|727|34.9136368 |133.0097723|Hedyotis cookiana (Cham. & Schltdl.) Steud.                        |2450.27|
```

Figure 4-17: Query Builder - Review temp dataset

A "Refresh Master" button is also available to preview the contents of the master dataset. Keep in mind that the two dataframes may hold completely different data. The user may also join the two dataframes through the corresponding option in the filters form. The result will be kept in the TempDF.



Temp Dataset Preview (first 40 rows)

In [51]:

```
+----------------------------------------------------------------+-------+---------+---------+--------------+-------+
|ReportedIncident_ID                                             |Month  |lon      |lat      |Damage        |Context|
+----------------------------------------------------------------+-------+---------+---------+--------------+-------+
|null                                                            |2018-06|-2.50993 |51.410873|Severe-Checked|null   |
|null                                                            |2018-06|-2.516919|51.423683|Severe-Checked|null   |
|null                                                            |2018-06|-2.50993 |51.410873|Severe-Checked|null   |
|null                                                            |2018-06|-2.50993 |51.410873|Severe-Checked|null   |
|null                                                            |2018-06|-2.50993 |51.410873|Severe-Checked|null   |
|null                                                            |2018-06|-2.50993 |51.410873|Severe-Checked|null   |
|b75295f6f32a325976f3228776f62ecdb7b2574b45c8ea62674095a58035ec17|2018-06|-2.49487 |51.422276|Minor-Checked |null   |
|2c15ce9b460bd015b4c645eb14d11d98b76ca6ee9761185dba822dda33a2330c|2018-06|-2.515072|51.419357|Minor-Checked |null   |
|e8272c40e825b002f99b3ffe0b7f6929eb4c3502108bdb803974ec627a9a17e6|2018-06|-2.516919|51.423683|Minor-Checked |null   |
|bd764e78dbfccd630a23c196bf47c3741b77c192351df5373541d4790e89e3ca|2018-06|-2.512773|51.411751|Severe-Pending|null   |
|6290d95287f0e2b89b12ed7670167f436df57e1024cc0a5dcdd1efdcc87d1763|2018-06|-2.511761|51.409966|Severe-Pending|null   |
|f9ab8121f7c06f0295c7fe043ce3308daaf70ed5496b314225c50e754aa1d09d|2018-06|-2.495055|51.422132|Severe-Pending|null   |
|null                                                            |2018-06|-2.494715|51.419948|Severe-Checked|null   |
|null                                                            |2018-06|-2.496204|51.417982|Severe-Checked|null   |
|null                                                            |2018-06|-2.501425|51.416692|Severe-Checked|null   |
|null                                                            |2018-06|-2.498613|51.416002|Severe-Checked|null   |
```

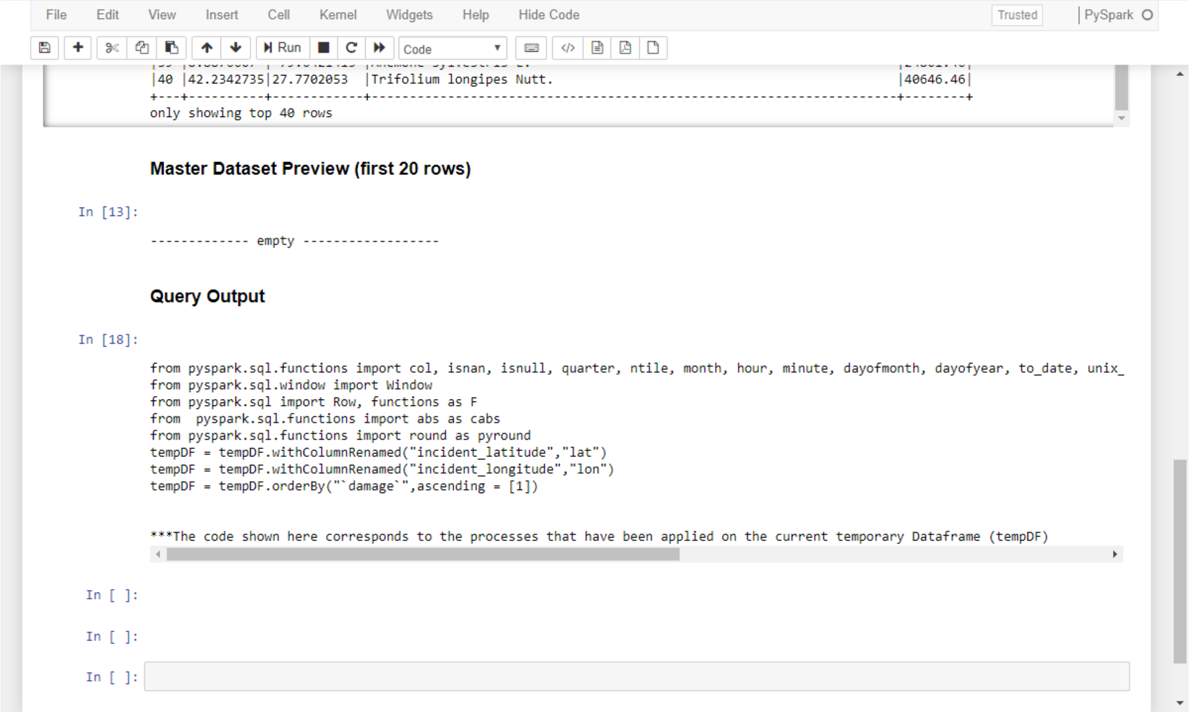
Master Dataset Preview (first 20 rows)

In [42]:

```
+---+-----------+-----------+----------------------------------------------------------------------+-------+
|id |lat        |lon        |plant_name                                                            |damage |
+---+-----------+-----------+----------------------------------------------------------------------+-------+
|413|11.5218308 |13.6883765 |Asplenium nidus L.                                                    |2005.68|
|737|37.5957943 |107.411996 |Eucalyptus salmonophloia F. Muell.                                    |2039.27|
|656|32.119799  |34.986653  |Erythrina fusca Lour.                                                 |2042.56|
|25 |45.7608815 |4.26286    |Stellaria ruscifolia Pall. ex Schltdl. ssp. aleutica Hultén          |2072.63|
|153|-34.6269535|-58.4058228|Arctostaphylos alpina (L.) Spreng. var. alpina                        |2088.58|
|387|-7.4443035 |-35.148879 |Crataegus lacrimata Small                                             |2127.74|
|368|2.8963048  |-72.1394818|Eutrochium maculatum (L.) E.E. Lamont var. maculatum                  |2158.73|
|230|34.9904542 |126.4816856|Streptanthus maculatus Nutt. ssp. maculatus                           |2210.49|
```

Figure 4-18: Query Builder - Preview temp and master Dataset

The user can use the "Update Query Output" button which will reveal a ready to copy and use python snippet which corresponds to the processes that have been applied on the TempDF.



```
|40 |42.2342735|27.7702053  |Trifolium longipes Nutt.                                                    |40646.46|
+---+----------+------------+----------------------------------------------------------------------------+--------+
only showing top 40 rows
```

Master Dataset Preview (first 20 rows)

In [13]:

```
------------ empty -----------------
```

Query Output

In [18]:

```
from pyspark.sql.functions import col, isnan, isnull, quarter, ntile, month, hour, minute, dayofmonth, dayofyear, to_date, unix_
from pyspark.sql.window import Window
from pyspark.sql import Row, functions as F
from  pyspark.sql.functions import abs as cabs
from pyspark.sql.functions import round as pyround
tempDF = tempDF.withColumnRenamed("incident_latitude","lat")
tempDF = tempDF.withColumnRenamed("incident_longitude","lon")
tempDF = tempDF.orderBy("`damage`",ascending = [1])


***The code shown here corresponds to the processes that have been applied on the current temporary Dataframe (tempDF)
```

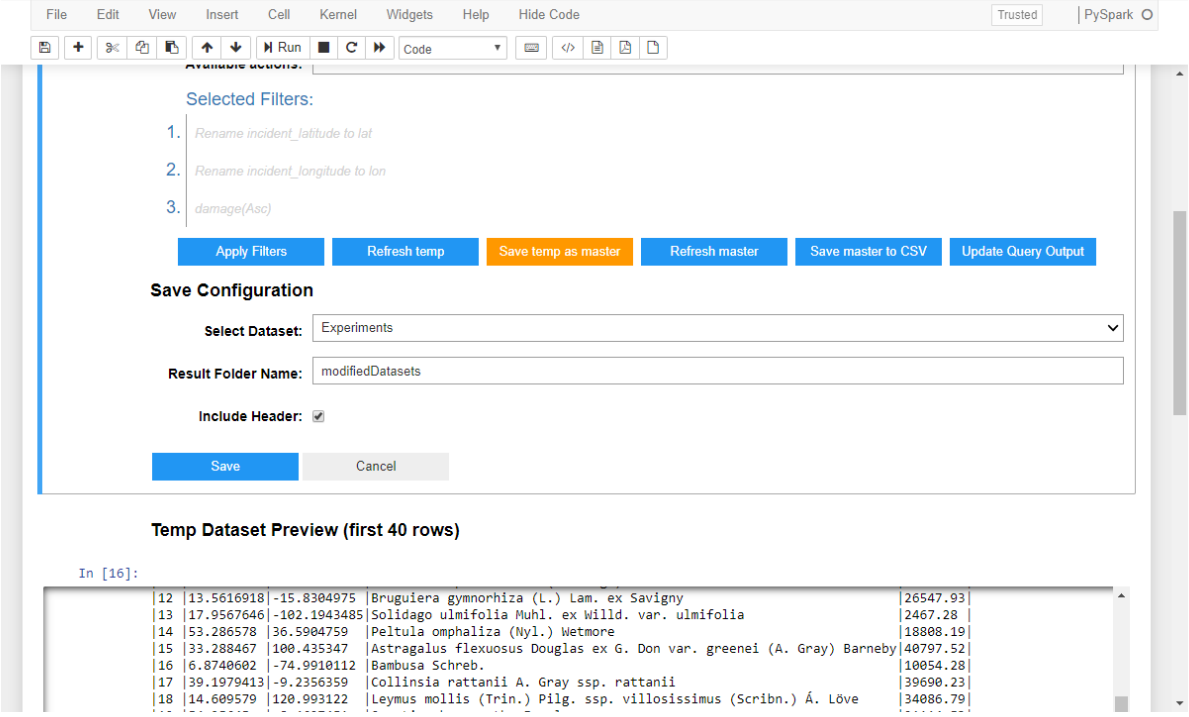
In [ ]:

In [ ]:

In [ ]:

Figure 4-19: Query Builder - Preview Query Output

By clicking the "Save master to CSV" button, the Save configuration form will appear where the user can select where to store the MasterDF's data.



Selected Filters:

1. Rename incident_latitude to lat
2. Rename incident_longitude to lon
3. damage(Asc)

Save Configuration

Select Dataset: Experiments

Result Folder Name: modifiedDatasets

Include Header: ☑

Temp Dataset Preview (first 40 rows)

In [16]:

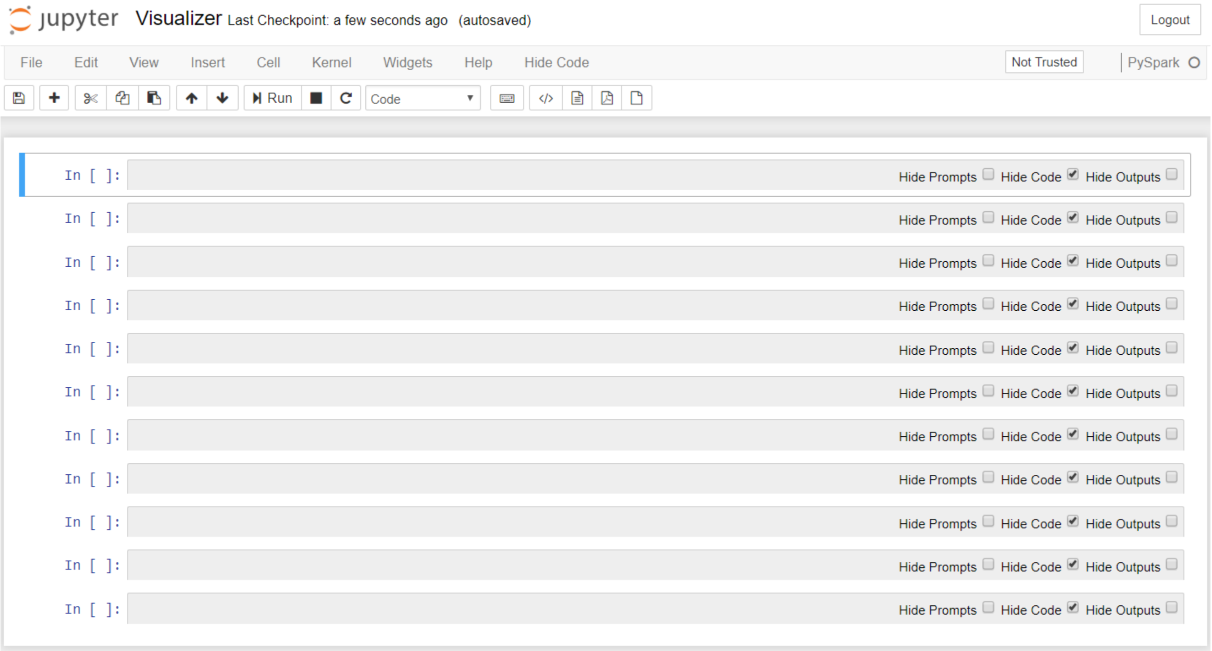
```
|12 |13.5616918|-15.8304975 |Bruguiera gymnorhiza (L.) Lam. ex Savigny                          |26547.93|
|13 |17.9567646|-102.1943485|Solidago ulmifolia Muhl. ex Willd. var. ulmifolia                  |2467.28 |
|14 |53.286578 |36.5904759  |Peltula omphaliza (Nyl.) Wetmore                                   |18808.19|
|15 |33.288467 |100.435347  |Astragalus flexuosus Douglas ex G. Don var. greenei (A. Gray) Barneby|40797.52|
|16 |6.8740602 |-74.9910112 |Bambusa Schreb.                                                    |10054.28|
|17 |39.1979413|-9.2356359  |Collinsia rattanii A. Gray ssp. rattanii                           |39690.23|
|18 |14.609579 |120.993122  |Leymus mollis (Trin.) Pilg. ssp. villosissimus (Scribn.) Á. Löve   |34086.79|
```

Figure 4-20: Query Builder – Save dataframe to CSV

The user may load, process and combine data following the previous steps in any order, until he/she is satisfied with the end-result. At any time, the user may press the "Initialise QB" button to clear the screen and start over. When the user has completed his/her work with the Query Builder, he/she may either directly exit the Notebook or clear all the output first so that he/she can have a fresh start without any leftovers the next time. This can be achieved from Jupyter's top menu. Choose Cell -> All Output -> Clear.

Visualiser

Once the user clicks on the Visualiser tool, a preconfigured Jupyter notebook is opened as shown in the following figure.

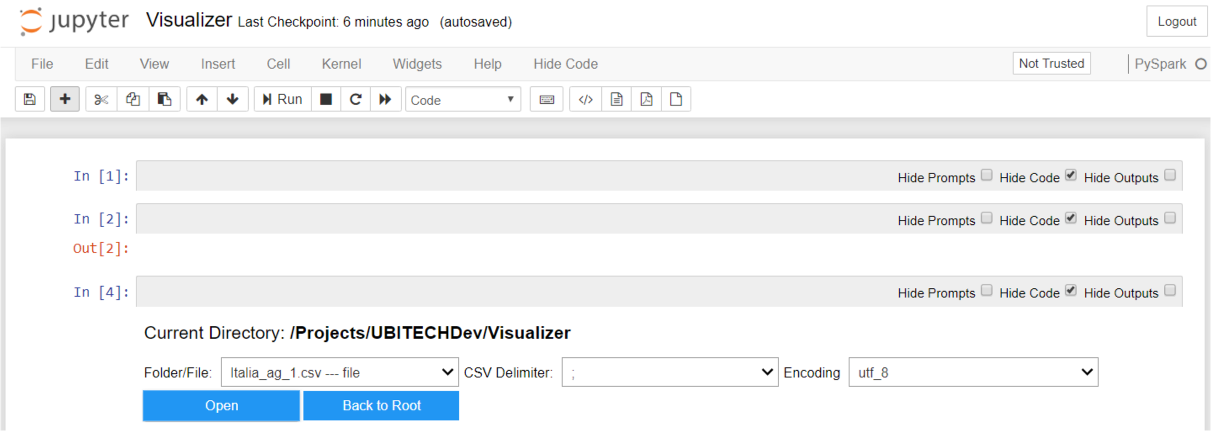


```
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
In [ ]:    Hide Prompts □ Hide Code ☑ Hide Outputs □
```

Figure 4-21: Visualiser initial state

The Visualiser is now in the initial state. All code cells are hidden by using the hide code plugin and also all output has been cleared. To start the tool, the user should simply execute the first paragraph by either selecting in and press ctrl+enter or by clicking the “Run” button on the top menu.

At that stage, the Visualizer is set and can be used, the first widget that appears looks like a file picker, where the user can navigate between the files stored in the current project and open a desired one. Note that only CSV files are supported.

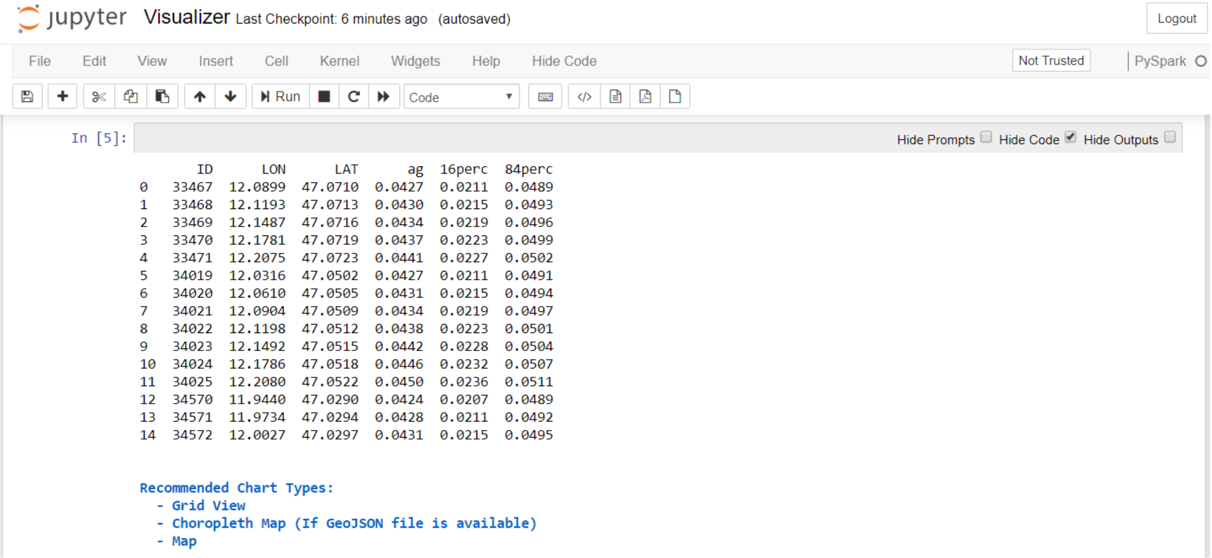

jupyter
Visualizer Last Checkpoint: 6 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Hide Code
Not Trusted
PySpark
Run
Code
In [1]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
In [2]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
Out[2]:
In [4]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
Open
Back to Root

Current Directory: **/Projects/UBITECHDev/Visualizer**

Folder/File: Italia_ag_1.csv --- file CSV Delimiter: ; Encoding utf_8

Figure 4-22: Visualiser - File Picker

Once a file is opened, a small preview in tabular format will be shown.

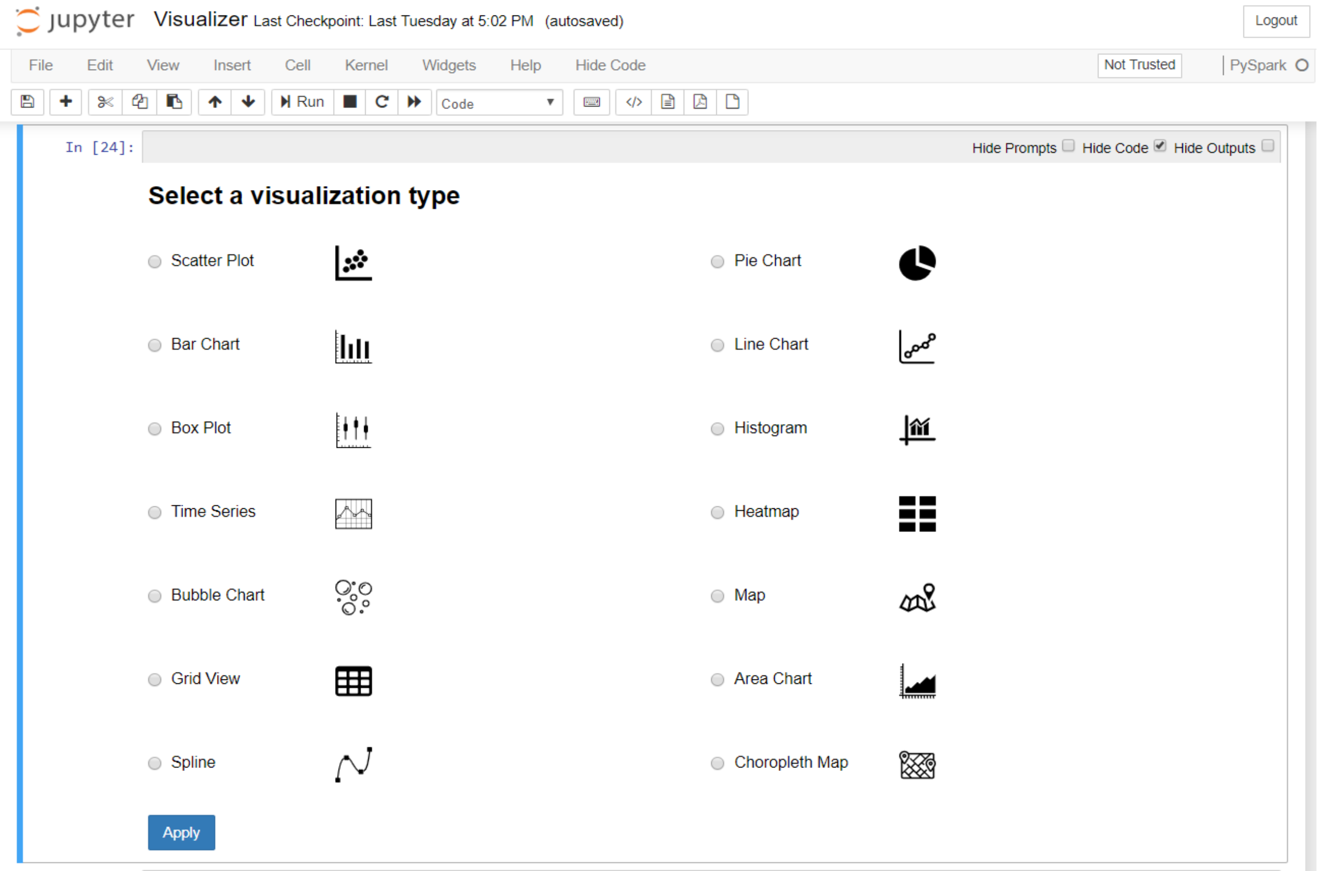

jupyter
Visualizer Last Checkpoint: 6 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Hide Code
Not Trusted
PySpark
Run
Code
In [5]:
Hide Prompts □ Hide Code ☑ Hide Outputs □

```
       ID      LON      LAT      ag  16perc  84perc
0   33467  12.0899  47.0710  0.0427  0.0211  0.0489
1   33468  12.1193  47.0713  0.0430  0.0215  0.0493
2   33469  12.1487  47.0716  0.0434  0.0219  0.0496
3   33470  12.1781  47.0719  0.0437  0.0223  0.0499
4   33471  12.2075  47.0723  0.0441  0.0227  0.0502
5   34019  12.0316  47.0502  0.0427  0.0211  0.0491
6   34020  12.0610  47.0505  0.0431  0.0215  0.0494
7   34021  12.0904  47.0509  0.0434  0.0219  0.0497
8   34022  12.1198  47.0512  0.0438  0.0223  0.0501
9   34023  12.1492  47.0515  0.0442  0.0228  0.0504
10  34024  12.1786  47.0518  0.0446  0.0232  0.0507
11  34025  12.2080  47.0522  0.0450  0.0236  0.0511
12  34570  11.9440  47.0290  0.0424  0.0207  0.0489
13  34571  11.9734  47.0294  0.0428  0.0211  0.0492
14  34572  12.0027  47.0297  0.0431  0.0215  0.0495


Recommended Chart Types:
  - Grid View
  - Choropleth Map (If GeoJSON file is available)
  - Map
```

Figure 4-23: Visualiser - File Preview

Depending on whether or not the selected file has been associated with extended metadata, a small section with a few recommended chart types will be displayed.

Afterwards, a list with the available visualisation types will be displayed. The user can pick a desired type and click on the Apply button.

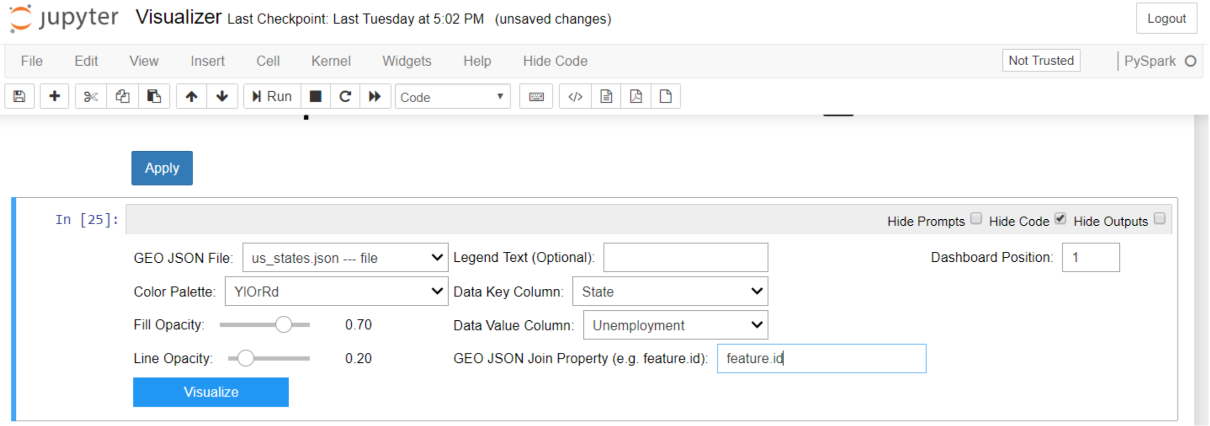

jupyter
Visualizer Last Checkpoint: Last Tuesday at 5:02 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Hide Code
Not Trusted
PySpark
Run
Code
In [24]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
Apply

Select a visualization type

○ Scatter Plot	○ Pie Chart
○ Bar Chart	○ Line Chart
○ Box Plot	○ Histogram
○ Time Series	○ Heatmap
○ Bubble Chart	○ Map
○ Grid View	○ Area Chart
○ Spline	○ Choropleth Map

Figure 4-24: Visualiser - Select visualisation

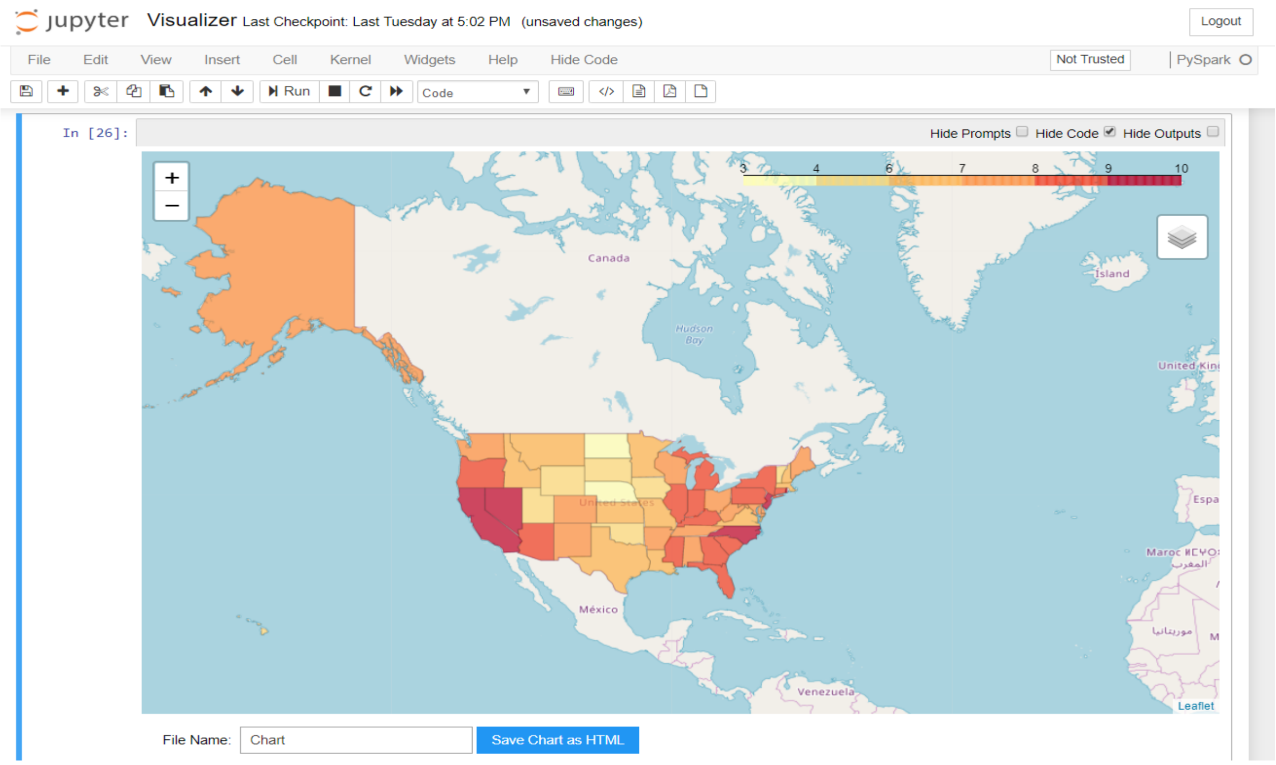
For every visualization type, a widget with various options depending on the type will be opened. The user should fill the mandatory fields with the desired values and then click on the Visualize button.


jupyter
Visualizer Last Checkpoint: Last Tuesday at 5:02 PM (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Hide Code
Not Trusted
PySpark
Run
Code
Apply
In [25]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
Visualize

GEO JSON File: us_states.json --- file
Legend Text (Optional):
Dashboard Position: 1
Color Palette: YlOrRd
Data Key Column: State
Fill Opacity: 0.70
Data Value Column: Unemployment
Line Opacity: 0.20
GEO JSON Join Property (e.g. feature.id): feature.id

Figure 4-25: Visualiser - Select visualisation options

The user should pay attention to the parameter called Position. This actually refers to the position on the dashboard where the chart will be placed. In the previous figure, there are five empty cells in the end of the notebook that act as placeholders for the dashboard. Currently only up to five different charts at the same time are supported, but this can be easily extended. The dashboard can contain several charts from the same dataset or even from different ones. When a new dataset is opened, previous charts are not deleted. A sample chart that appears after filling the necessary options can be seen below.

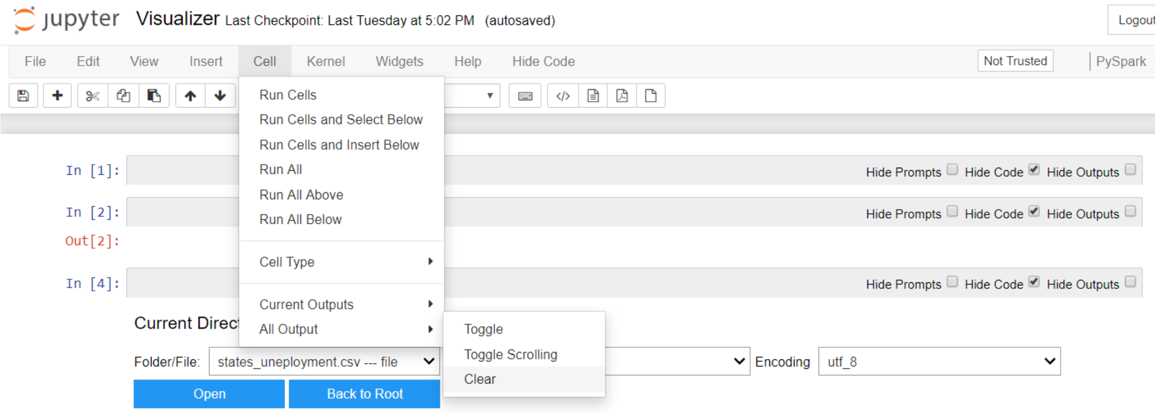

jupyter
Visualizer Last Checkpoint: Last Tuesday at 5:02 PM (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help Hide Code
Not Trusted
PySpark
Run
Code
In [26]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
3 4 6 7 8 9 10
Canada
Hudson Bay
Ísland
United Kingdom
España
Maroc
United States
México
Venezuela
Leaflet
Save Chart as HTML

File Name: Chart

Figure 4-26: Visualiser - Sample chart

At the bottom of every chart, the user is presented with a choice to save the displayed chart. The only thing that needs to be done is to provide a desired name and the press on the "Save Chart as HTML" button. This will store the chart as a dynamic HTML file inside HopsFS.

When the user has completed his/her work with the Visualiser, it is highly recommended to clear all the output, so he/she can do a fresh start without any leftovers the next time. This can be achieved from Jupyter's top menu. Choose Cell -> All Output -> Clear.


jupyter
Visualizer Last Checkpoint: Last Tuesday at 5:02 PM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help Hide Code
Not Trusted
PySpark
Run Cells
Run Cells and Select Below
Run Cells and Insert Below
Run All
Run All Above
Run All Below
Cell Type
Current Outputs
All Output
Toggle
Toggle Scrolling
Clear
In [1]:
In [2]:
Out[2]:
In [4]:
Hide Prompts □ Hide Code ☑ Hide Outputs □
Current Direc
Open
Back to Root

Folder/File: states_uneployment.csv --- file

Encoding utf_8

Figure 4-27: Visualiser - Clear output

Note: There are two alternatives ways for opening the Visualizer besides the one described earlier. The first one is accessible by visiting the Datasets Page and then open "Jupyter" dataset which is available by default in all projects. The user should browse for the Visualizer notebook, right click on it and then select the "Open Jupyter Notebook" context menu option.

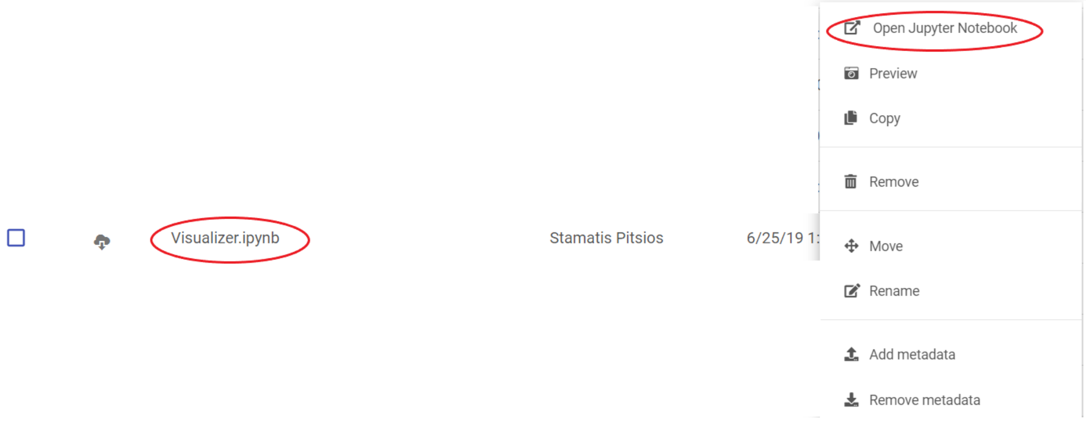


Figure 4-28: Visualiser - Open with Jupyter

The second way, makes again use of the context menu, but in a completely different way. Again, the user should navigate to the Datasets Page and then open a dataset of her choice. Then, by right-clicking on a specific CSV file, she is given the ability to open that file directly into the Visualizer, as it can be seen in the screenshot above.

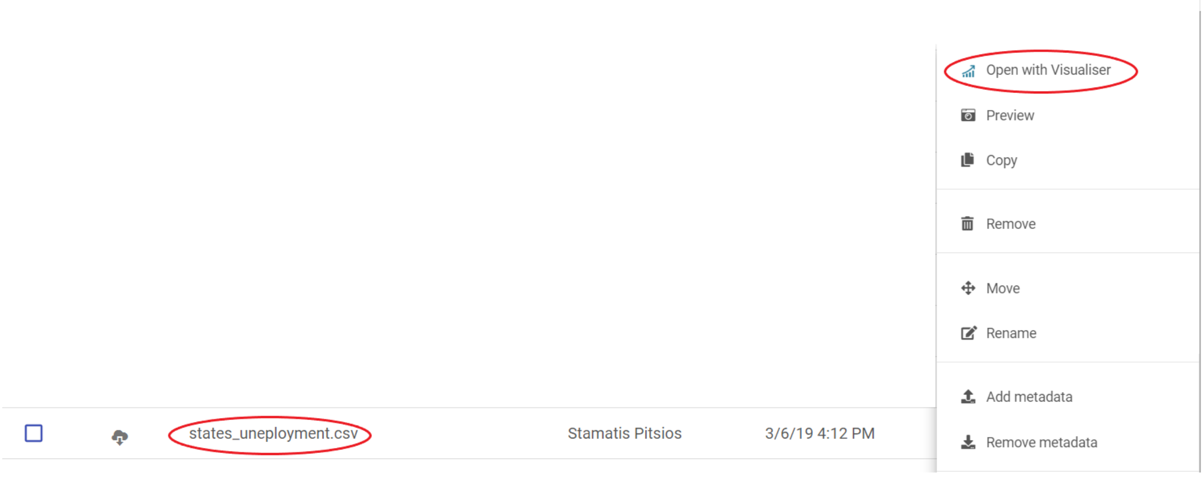


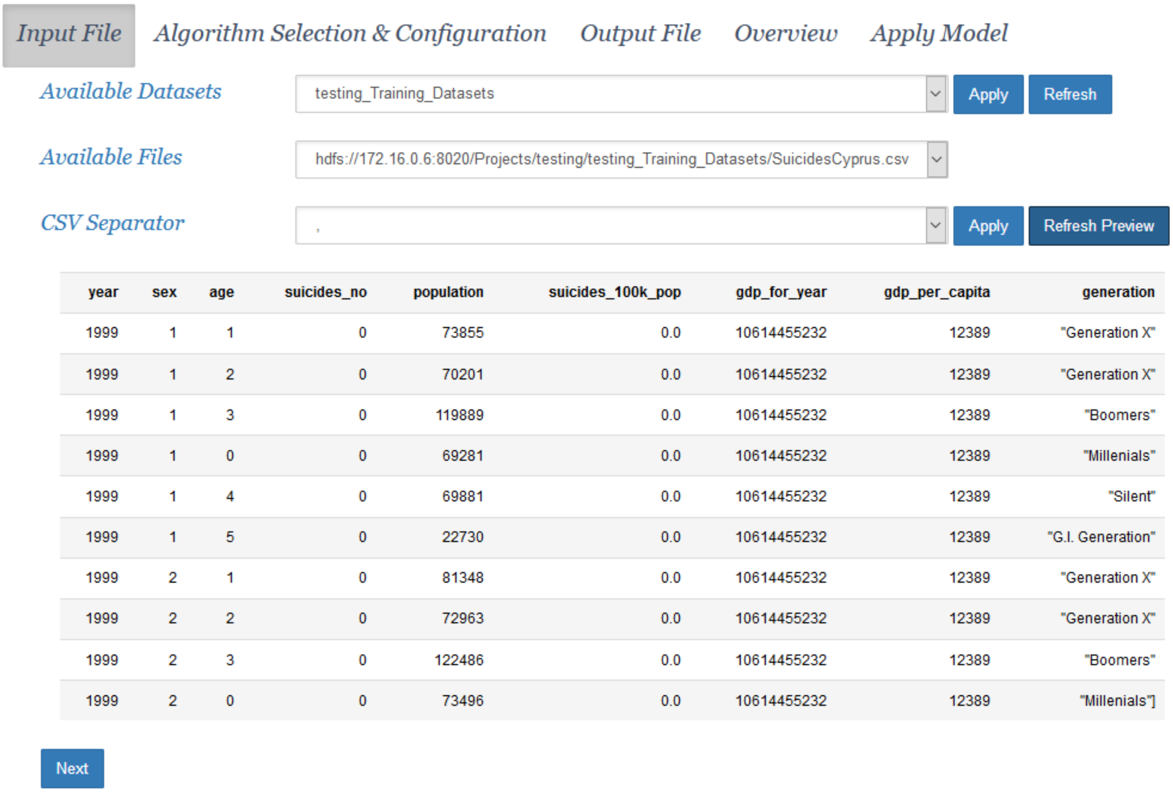
Figure 4-29: Visualiser - Open with context menu

Algorithm Execution Container

The Algorithm Execution Container is implemented with Python 3.6 as a Jupyter notebook. Upon launch of the corresponding Algorithm Execution Container (hereinafter container), the user has to initialise it through a dedicated button. The initialisation ensures that the Spark interpreter of the AEGIS platform is started and also creates the basic UI of the container, which is a simple five-tab window:

- An input file selection tab
- An algorithm selection and configuration tab
- An output folder selection tab
- An overview tab showing the current user selections and, when ready, the execution results
- A model application tab, to perform regression or classification with existing trained models on other datasets.

Apart from the overview tab, the other three correspond to the basic steps towards applying an algorithm through the container. The first step is to select the file that contains the data to be used by the algorithm, through the interface shown in the next figure, which brings up also a preview of the selected file.



Available Datasets: testing_Training_Datasets

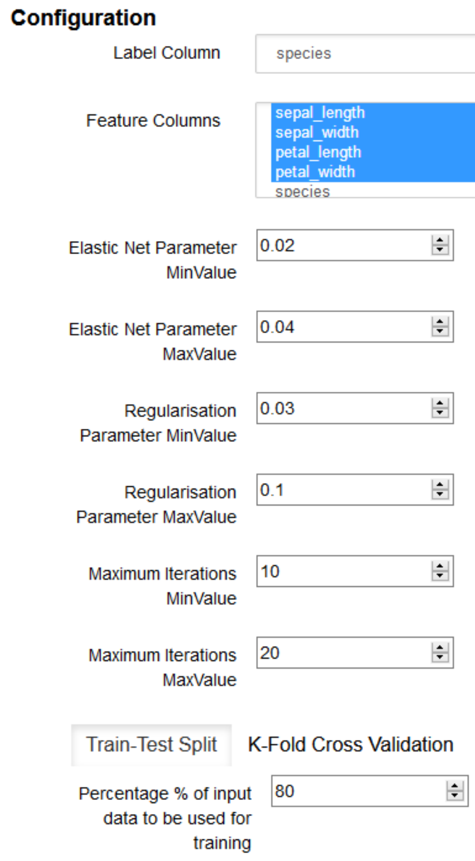
Available Files: hdfs://172.16.0.6:8020/Projects/testing/testing_Training_Datasets/SuicidesCyprus.csv

CSV Separator: ,

year	sex	age	suicides_no	population	suicides_100k_pop	gdp_for_year	gdp_per_capita	generation
1999	1	1	0	73855	0.0	10614455232	12389	"Generation X"
1999	1	2	0	70201	0.0	10614455232	12389	"Generation X"
1999	1	3	0	119889	0.0	10614455232	12389	"Boomers"
1999	1	0	0	69281	0.0	10614455232	12389	"Millenials"
1999	1	4	0	69881	0.0	10614455232	12389	"Silent"
1999	1	5	0	22730	0.0	10614455232	12389	"G.I. Generation"
1999	2	1	0	81348	0.0	10614455232	12389	"Generation X"
1999	2	2	0	72963	0.0	10614455232	12389	"Generation X"
1999	2	3	0	122486	0.0	10614455232	12389	"Boomers"
1999	2	0	0	73496	0.0	10614455232	12389	"Millenials"]

Figure 4-30: Algorithm Execution Container - Select input

The next step is to select the algorithm to be applied. The provided algorithms are grouped under five categories (algorithm families). Once an algorithm is selected, a form from which to configure its parameters is shown to the user. A basic form validation is done for the case that a selected value for a parameter is not within the specified boundaries. In this point, the final version of the Algorithm Execution Container includes the option to set a grid of parameters, so that the algorithm may run within those spans, and it automatically selects the best model based on the results produced. A preview of the configuration of the grid parameters is shown in the next figure.



Configuration

Label Column: species

Feature Columns: sepal_length, sepal_width, petal_length, petal_width, species

Elastic Net Parameter MinValue: 0.02

Elastic Net Parameter MaxValue: 0.04

Regularisation Parameter MinValue: 0.03

Regularisation Parameter MaxValue: 0.1

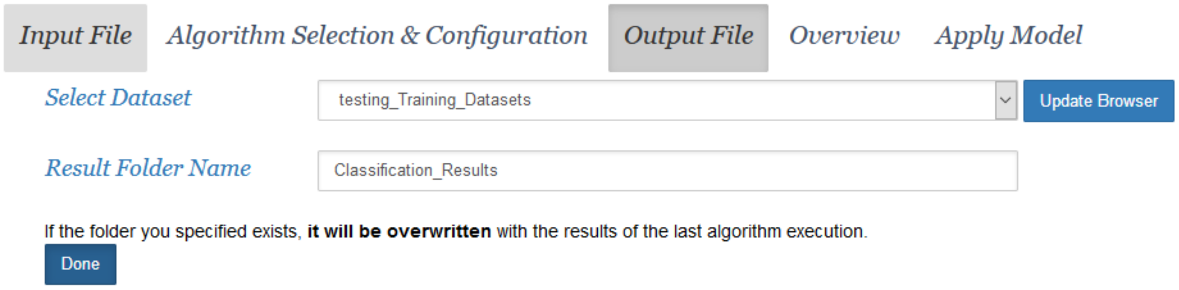
Maximum Iterations MinValue: 10

Maximum Iterations MaxValue: 20

Percentage % of input data to be used for training: 80

Figure 4-31: Algorithm Execution Container – Algorithm configuration

The model that will be created when an algorithm is applied is saved in the user's datasets, in a folder specified in the last container tab, as shown below:



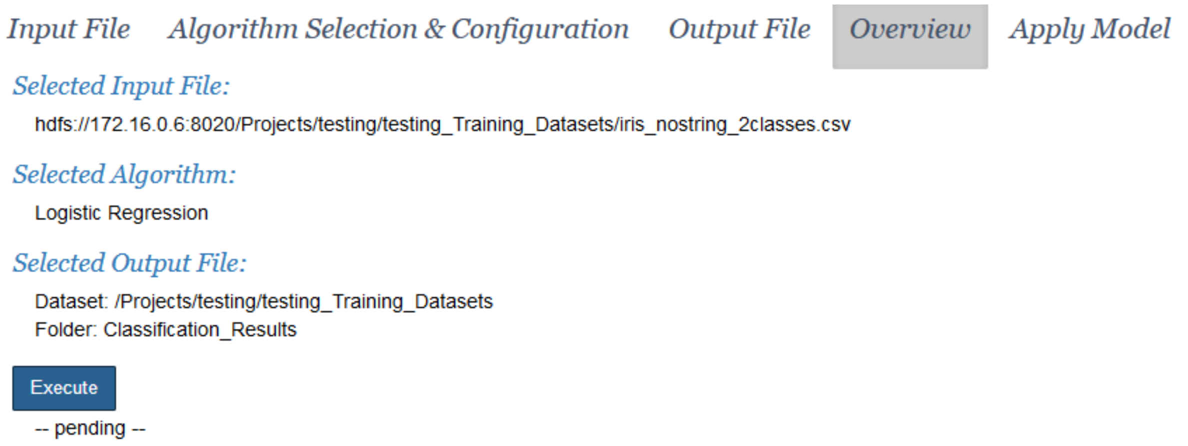
Select Dataset: testing_Training_Datasets

Result Folder Name: Classification_Results

If the folder you specified exists, **it will be overwritten** with the results of the last algorithm execution.

Figure 4-32: Algorithm Execution Container - Output folder

Once this last step is concluded, by pressing the "Apply" button, the user is taken to the Overview Tab:



Selected Input File:

hdfs://172.16.0.6:8020/Projects/testing/testing_Training_Datasets/iris_nostring_2classes.csv

Selected Algorithm:

Logistic Regression

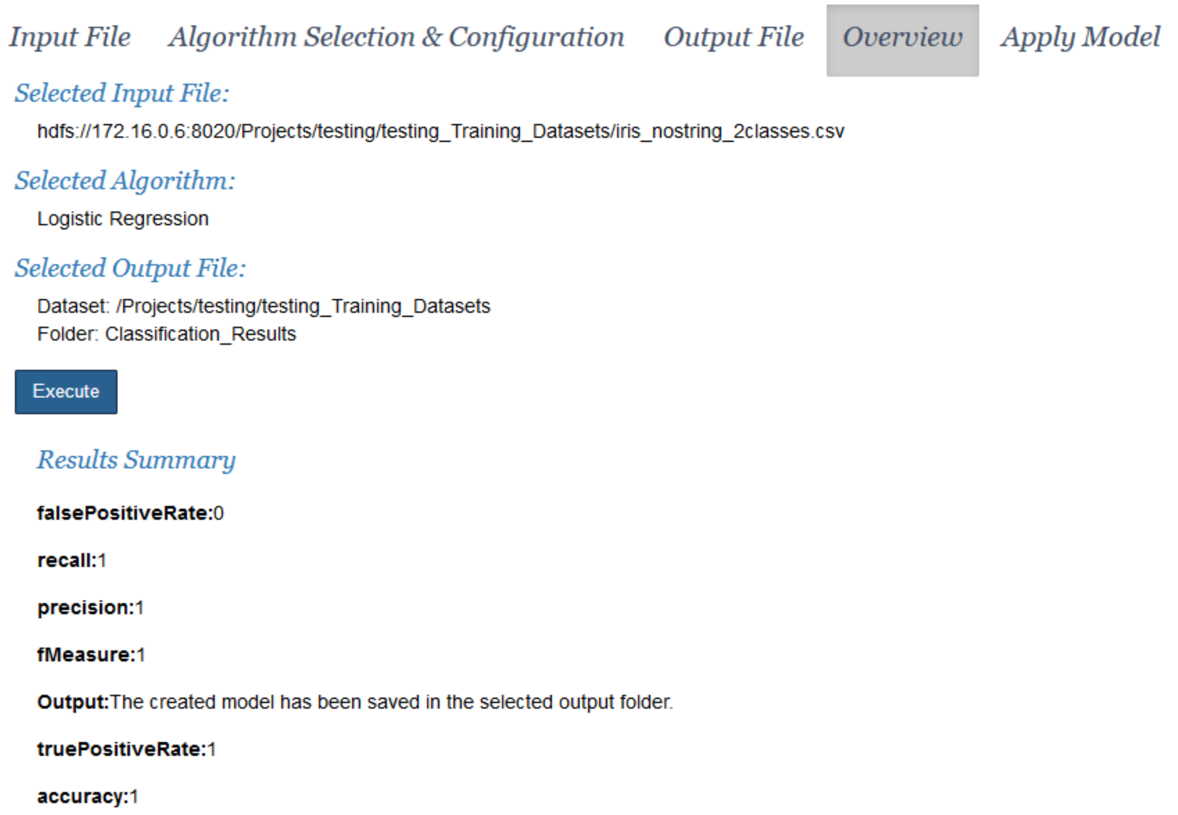
Selected Output File:

Dataset: /Projects/testing/testing_Training_Datasets
Folder: Classification_Results

-- pending --

Figure 4-33: Algorithm Execution Container - Overview

There, by pressing Done, the algorithm is applied and when the execution is completed, some algorithm-dependent results are provided to the user in the same tab (next figure). The final output of the analysis is stored back in the AEGIS Data Store, while the model that was used for the analysis, is also stored alongside with the analysis results.



Selected Input File:

hdfs://172.16.0.6:8020/Projects/testing/testing_Training_Datasets/iris_nostring_2classes.csv

Selected Algorithm:

Logistic Regression

Selected Output File:

Dataset: /Projects/testing/testing_Training_Datasets
Folder: Classification_Results

Results Summary

falsePositiveRate:0

recall:1

precision:1

fMeasure:1

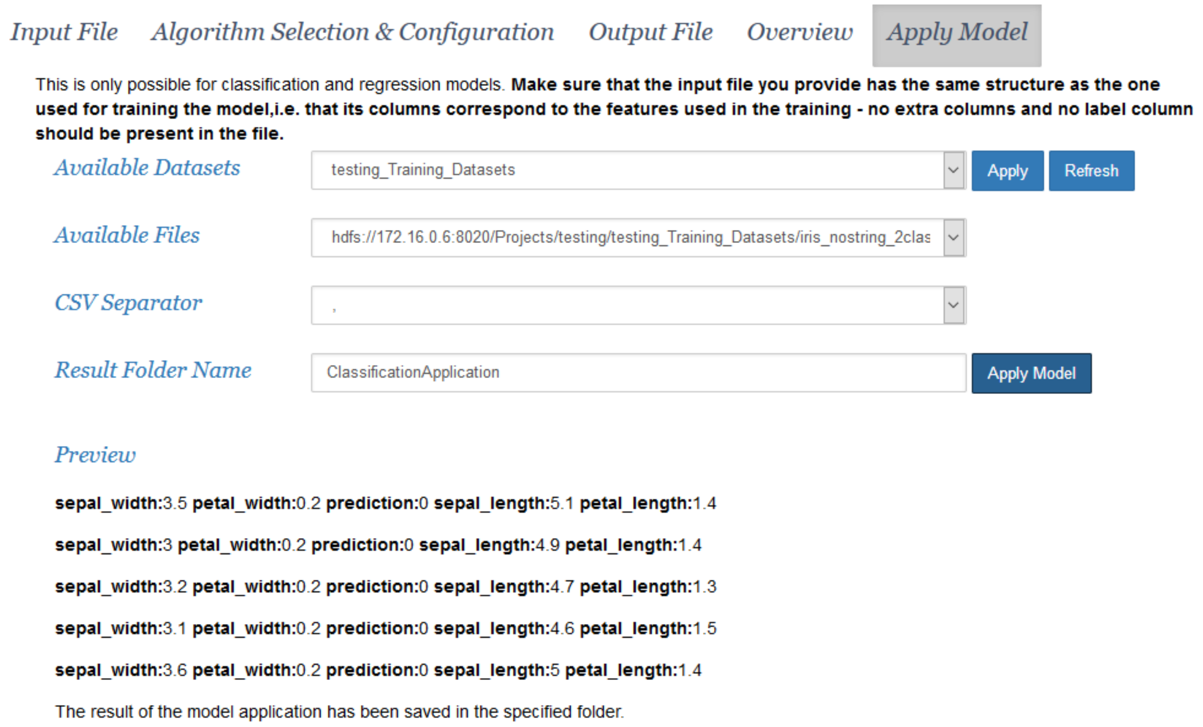
Output: The created model has been saved in the selected output folder.

truePositiveRate:1

accuracy:1

Figure 4-34: : Algorithm Execution Container - Results

In case the analysis that has been selected is of a classification or a regression, the model that has been created in the previous steps can be re-applied on new datasets, using the "Apply Model" tab, provided that the format of the input file is same as that of the one used to generate the model. This feature is shown in the next figure, where the user is able to select from existing files and re-apply the model.



This is only possible for classification and regression models. **Make sure that the input file you provide has the same structure as the one used for training the model,i.e. that its columns correspond to the features used in the training - no extra columns and no label column should be present in the file.**

Available Datasets: testing_Training_Datasets

Available Files: hdfs://172.16.0.6:8020/Projects/testing/testing_Training_Datasets/iris_nostring_2clas

CSV Separator: ,

Result Folder Name: ClassificationApplication

Preview

sepal_width:3.5 **petal_width:**0.2 **prediction:**0 **sepal_length:**5.1 **petal_length:**1.4

sepal_width:3 **petal_width:**0.2 **prediction:**0 **sepal_length:**4.9 **petal_length:**1.4

sepal_width:3.2 **petal_width:**0.2 **prediction:**0 **sepal_length:**4.7 **petal_length:**1.3

sepal_width:3.1 **petal_width:**0.2 **prediction:**0 **sepal_length:**4.6 **petal_length:**1.5

sepal_width:3.6 **petal_width:**0.2 **prediction:**0 **sepal_length:**5 **petal_length:**1.4

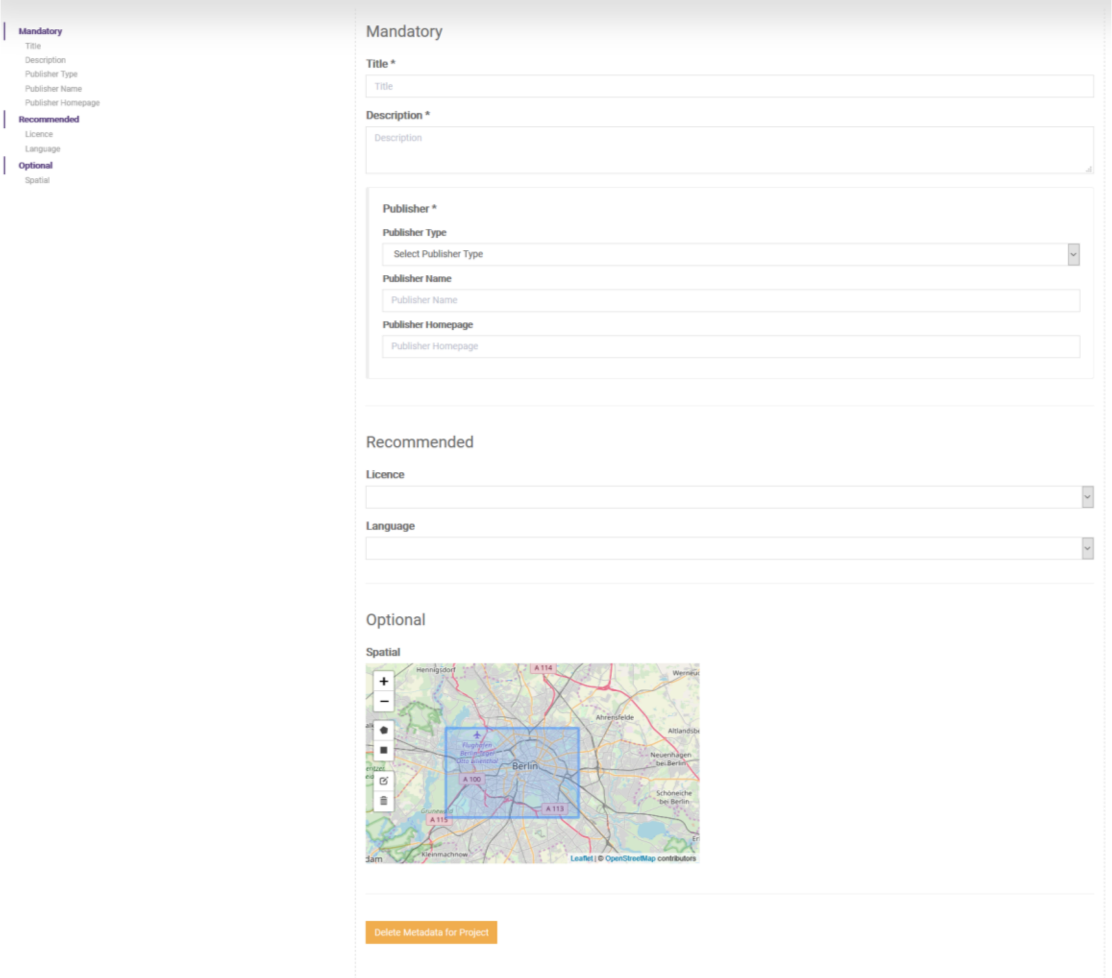
The result of the model application has been saved in the specified folder.

Figure 4-35: Algorithm Execution Container - Select existing model

4.1.3. Extended Metadata

The AEGIS platform supports the provision and usage of extended metadata based on DCAT-AP and the AEGIS ontology. It is serialized as Linked Data and can retrieved via a SPARQL endpoint or via a RESTful-API. Each entity type (project, dataset and file) have their respective metadata model and forms.

The project metadata form can be accessed in Tools menu in right pane: “Extended Metadata”:



Mandatory

Title *

Title

Description *

Description

Publisher *

Publisher Type

Select Publisher Type

Publisher Name

Publisher Name

Publisher Homepage

Publisher Homepage

Recommended

Licence

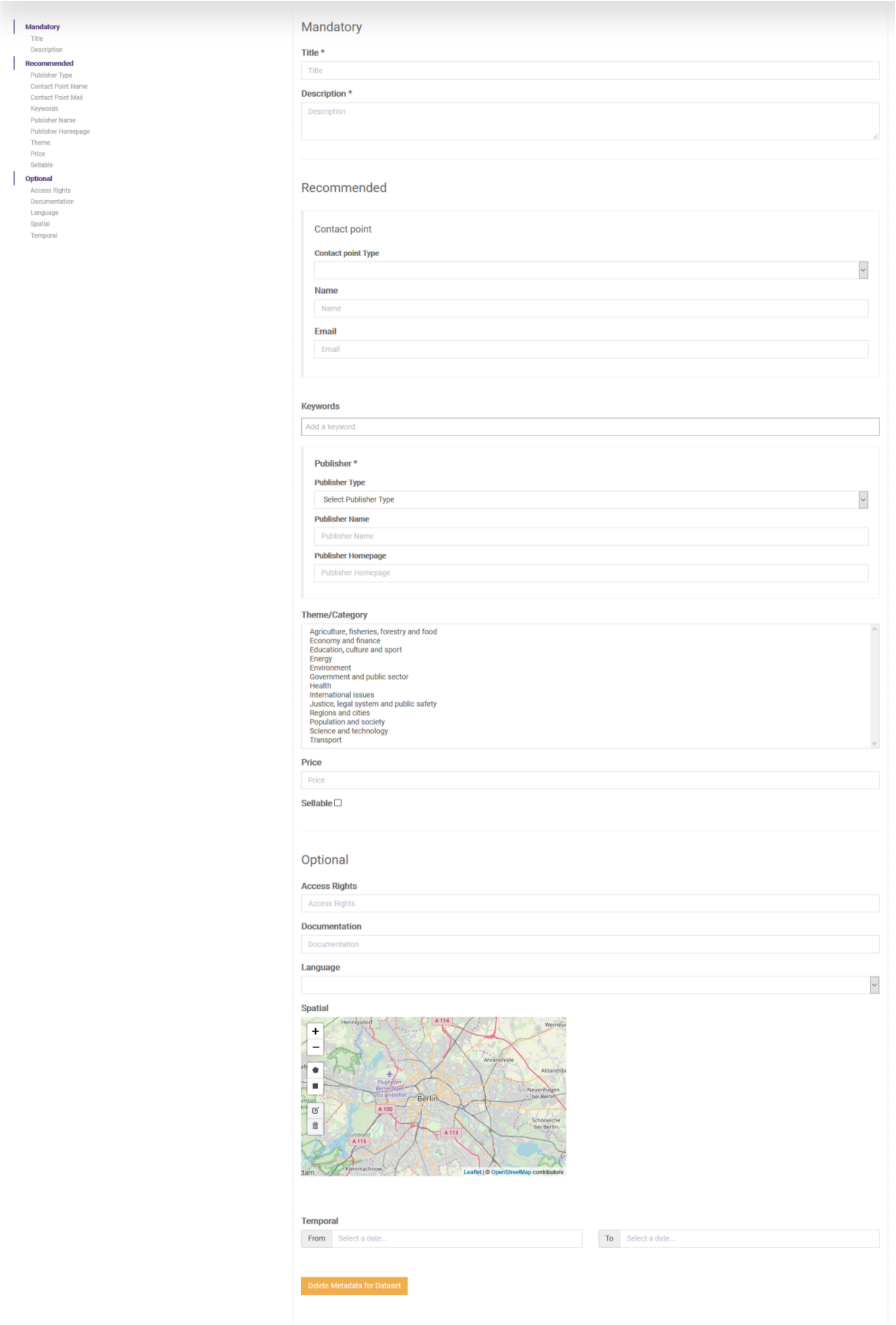
Language

Optional

Spatial

Figure 4-36: Metadata Form for Projects

The metadata form for datasets is available with a right-click on any dataset: “Add or edit extended metadata”:



Mandatory
Title
Description
Recommended
Publisher Type
Contact Point Name
Contact Point Mail
Keywords
Publisher Name
Publisher Homepage
Theme
Price
Sellable
Optional
Access Rights
Documentation
Language
Spatial
Temporal

Mandatory

Title *
Title

Description *
Description

Recommended

Contact point

Contact point Type

Name
Name

Email
Email

Keywords
Add a keyword

Publisher *

Publisher Type
Select Publisher Type

Publisher Name
Publisher Name

Publisher Homepage
Publisher Homepage

Theme/Category
Agriculture, fisheries, forestry and food
Economy and finance
Education, culture and sport
Energy
Environment
Government and public sector
Health
International issues
Justice, legal system and public safety
Regions and cities
Population and society
Science and technology
Transport

Price
Price

Sellable □

Optional

Access Rights
Access Rights

Documentation
Documentation

Language

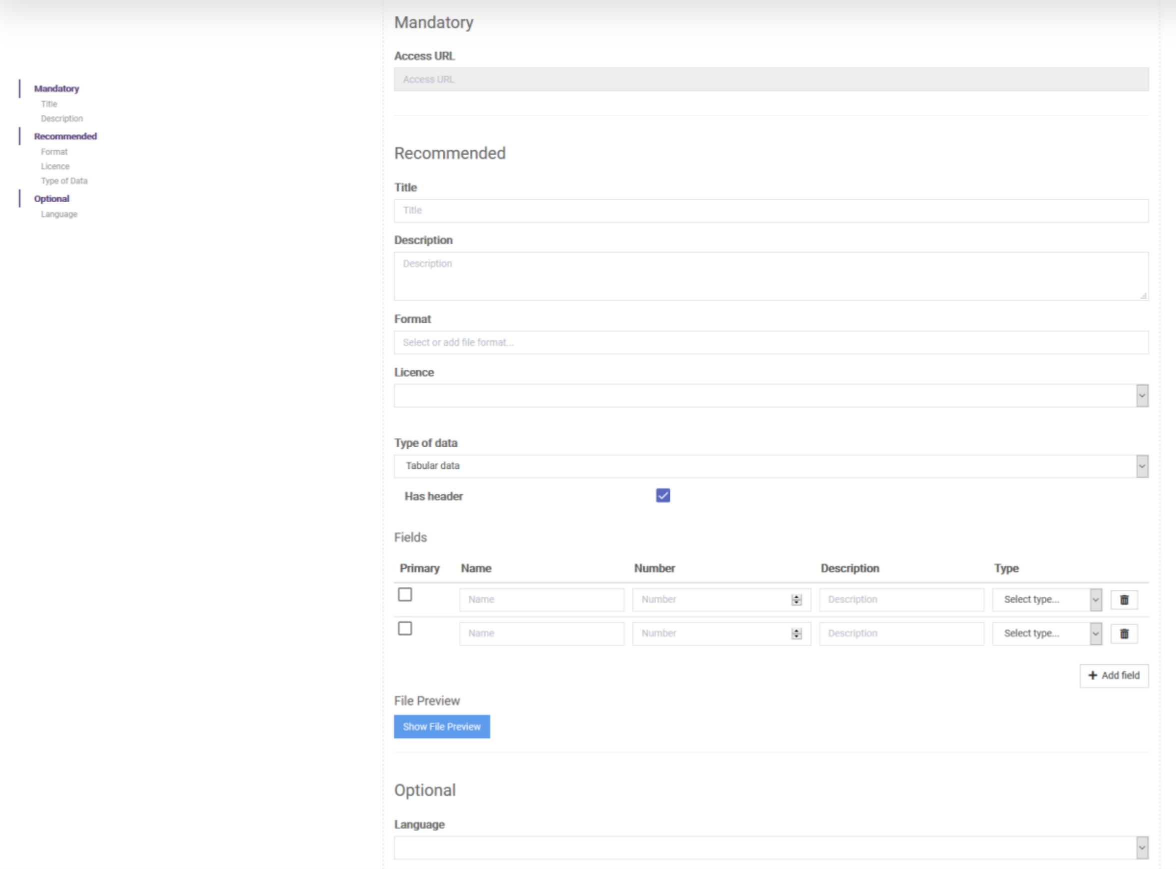
Spatial

Temporal
From Select a date... To Select a date...

Delete Metadata for Dataset

Figure 4-37: Metadata Form for Datasets

The metadata form for files is available with a right-click on any file: “Add or edit extended metadata”:



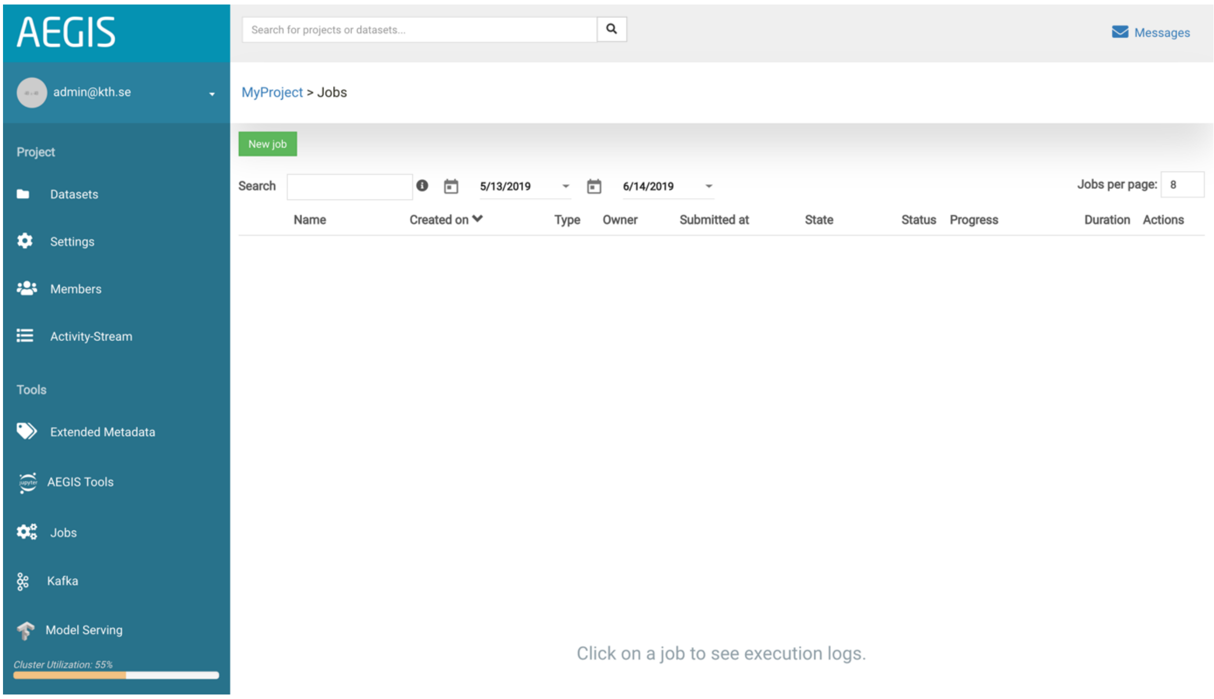
Type of data: Tabular data

Has header ☑

Figure 4-38: Metadata Form for Files

4.1.4. Jobs

The AEGIS platform provides support for data parallel processing frameworks such as MapReduce, Spark, and Flink through the use of the Jobs service. To run a job, the user needs to first upload his/her job files (.jar, .py, .ipynb) to a dataset. Then, the user can add a new job by clicking on the New job button as shown in the following figure.



MyProject > Jobs

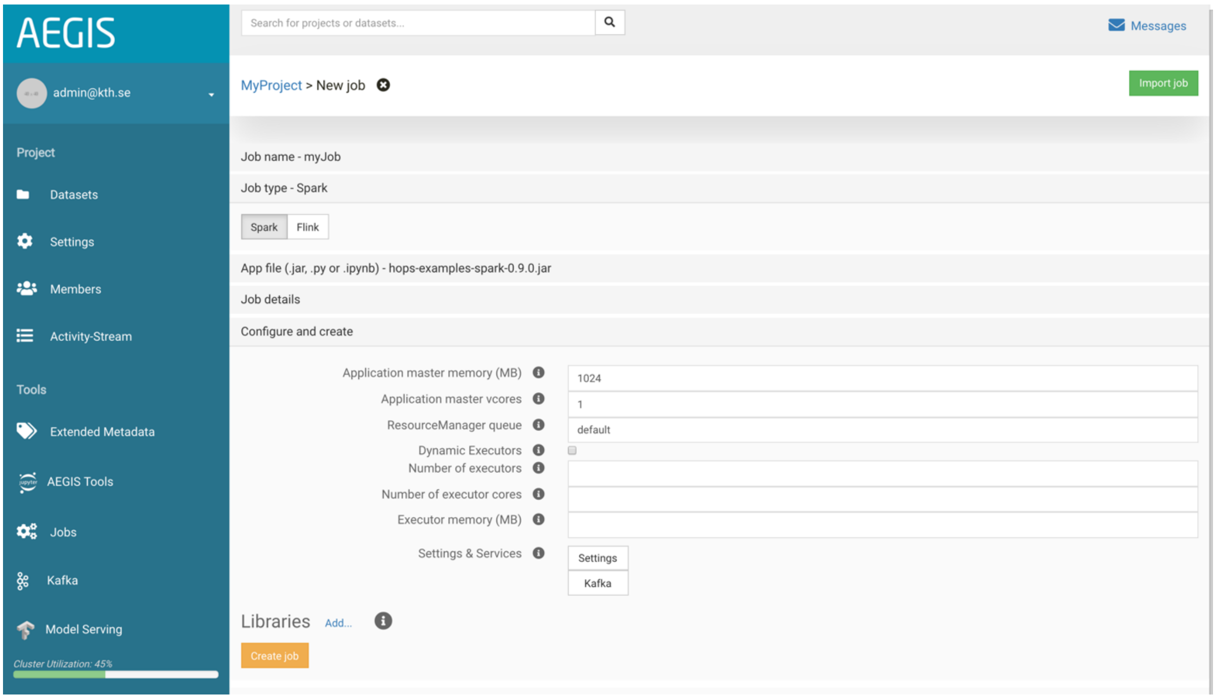
Search 5/13/2019 6/14/2019 Jobs per page: 8

Name	Created on	Type	Owner	Submitted at	State	Status	Progress	Duration	Actions

Click on a job to see execution logs.

Figure 4-39: The Jobs tool

To create a new job, the user provides the job name, type, configuration, and other external libraries as shown in the figure below.



MyProject > New job

Job name - myJob

Job type - Spark

App file (.jar, .py or .ipynb) - hops-examples-spark-0.9.0.jar

Job details

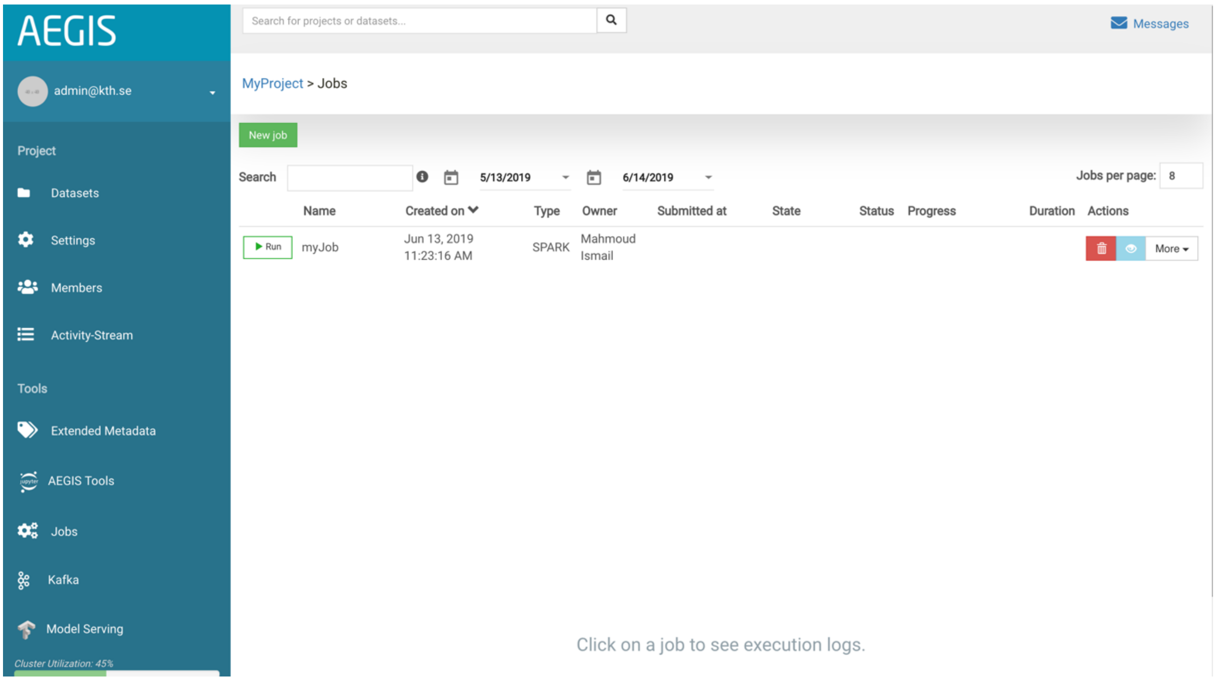
Configure and create

Application master memory (MB)	1024
Application master vcores	1
ResourceManager queue	default
Dynamic Executors	□
Number of executors	
Number of executor cores	
Executor memory (MB)	
Settings & Services	Settings Kafka

Libraries

Figure 4-40: Adding a new Job

Once the job is created, the user can run the job, and then look at the execution logs listed at the end of the page. Also, the user can rerun, edit, delete, and export the existing job or add new jobs.



Search 5/13/2019 6/14/2019 Jobs per page: 8

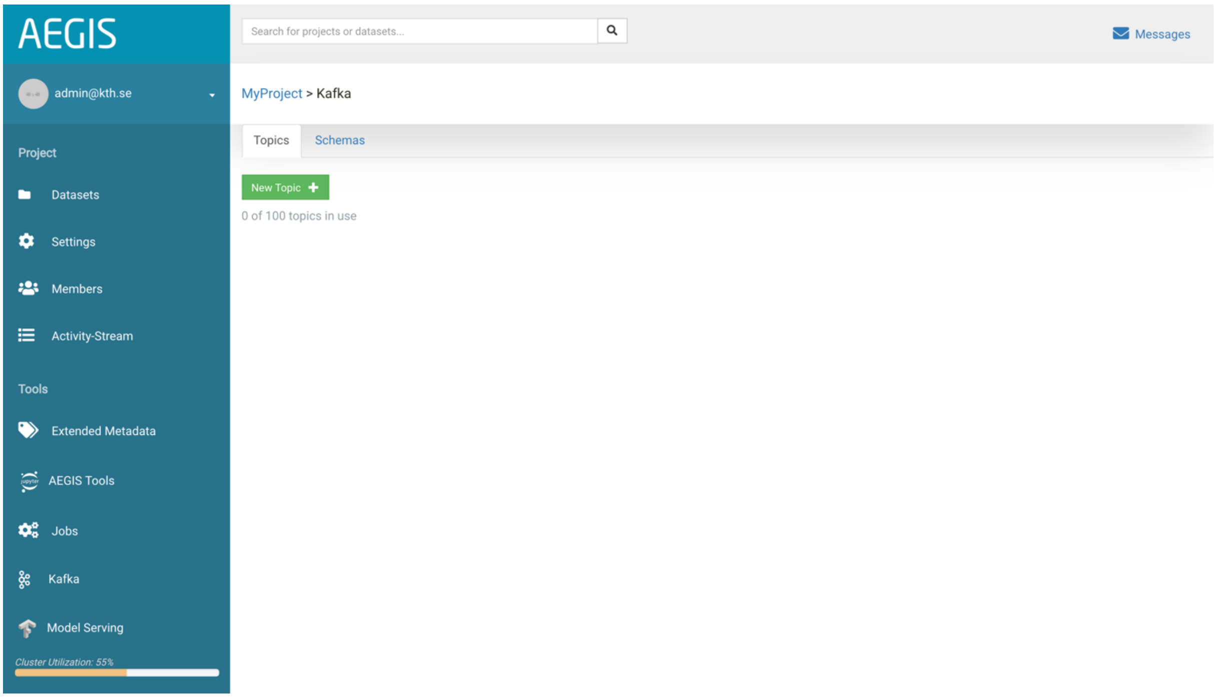
	Name	Created on	Type	Owner	Submitted at	State	Status	Progress	Duration	Actions
▶ Run	myJob	Jun 13, 2019 11:23:16 AM	SPARK	Mahmoud Ismail						More ▾

Click on a job to see execution logs.

Figure 4-41: new Job is added

4.1.5. Kafka

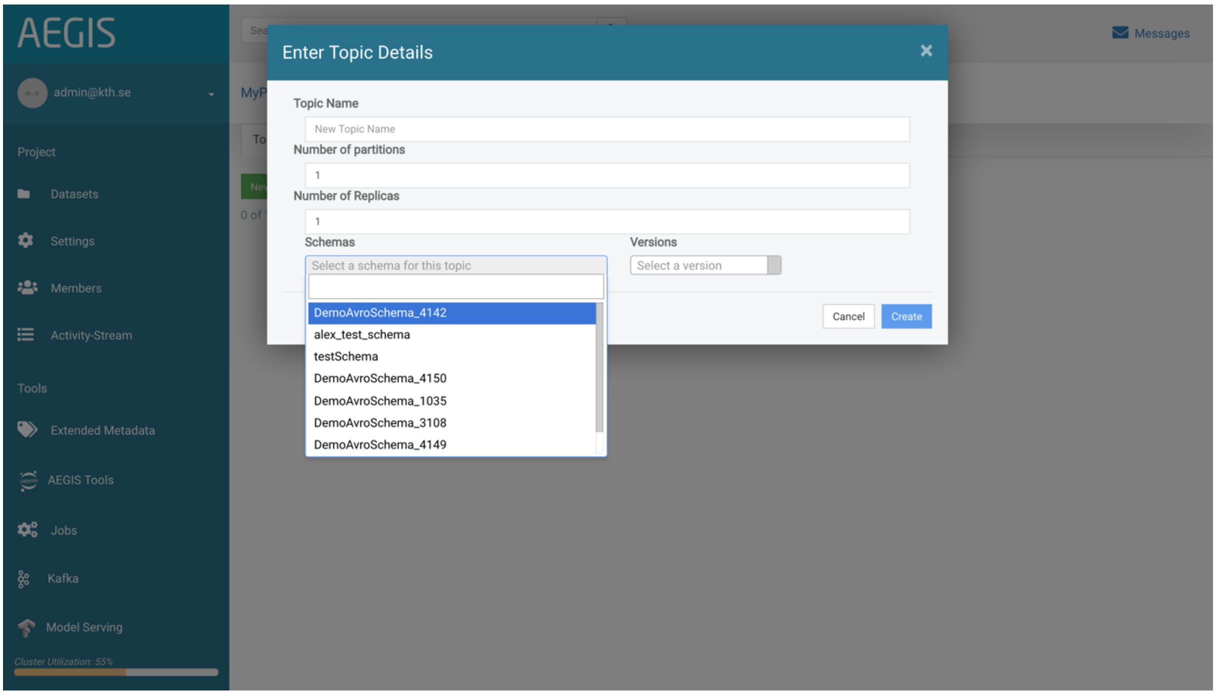
The AEGIS platform provides real time analytics using the Kafka service. The user can create topics and their respective schemas using the user interface provided by AEGIS, as shown in the following figure.



0 of 100 topics in use

Figure 4-42: Kafka

After clicking on the new topic button, the user is required to fill the topic details such as topic name, number of partitions, number of replicas, and the topic schema, as shown in Figure 4-43. The user can select a schema from the existing schema in the platform or create his own schema first by navigating to the schemas tab, as shown in Figure 4-44. The user can create his/her own schema by clicking on the new avro schema button and then provide the schema's details, as shown in Figure 4-45.



Enter Topic Details

Topic Name: New Topic Name

Number of partitions: 1

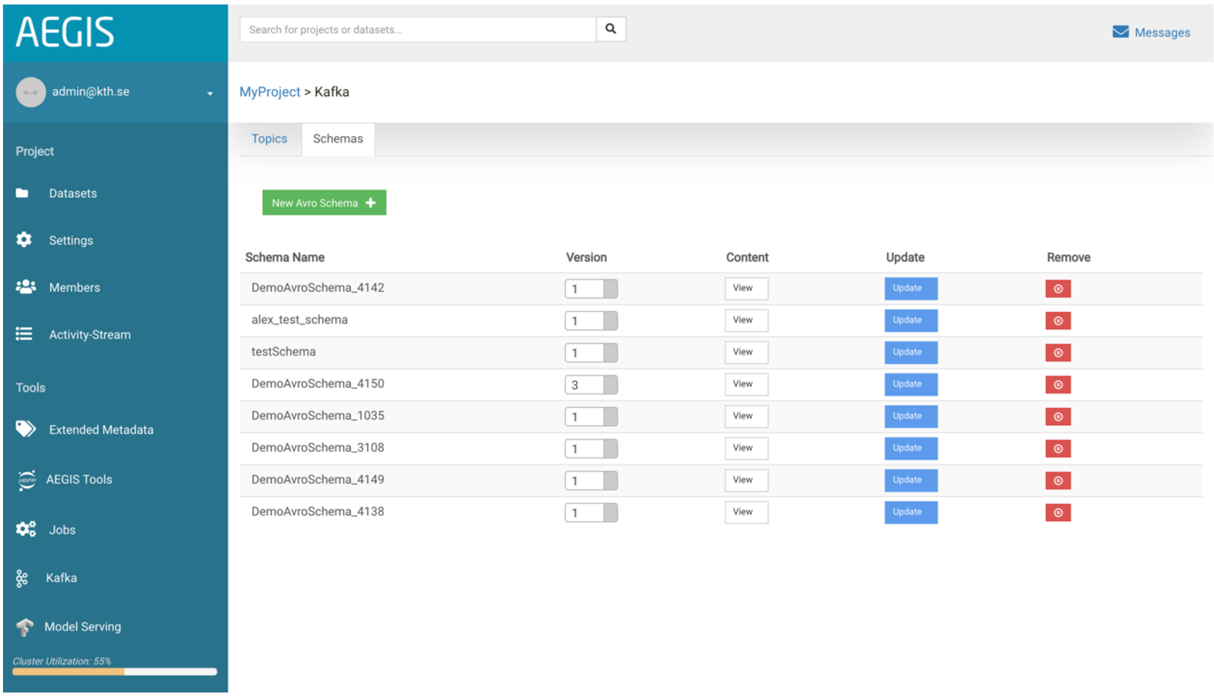
Number of Replicas: 1

Schemas: Select a schema for this topic

- DemoAvroSchema_4142
- alex_test_schema
- testSchema
- DemoAvroSchema_4150
- DemoAvroSchema_1035
- DemoAvroSchema_3108
- DemoAvroSchema_4149

Versions: Select a version

Figure 4-43: Create new Kafka topic



Schema Name	Version	Content	Update	Remove
DemoAvroSchema_4142	1	View	Update	⊗
alex_test_schema	1	View	Update	⊗
testSchema	1	View	Update	⊗
DemoAvroSchema_4150	3	View	Update	⊗
DemoAvroSchema_1035	1	View	Update	⊗
DemoAvroSchema_3108	1	View	Update	⊗
DemoAvroSchema_4149	1	View	Update	⊗
DemoAvroSchema_4138	1	View	Update	⊗

Figure 4-44: Kafka schemas tab

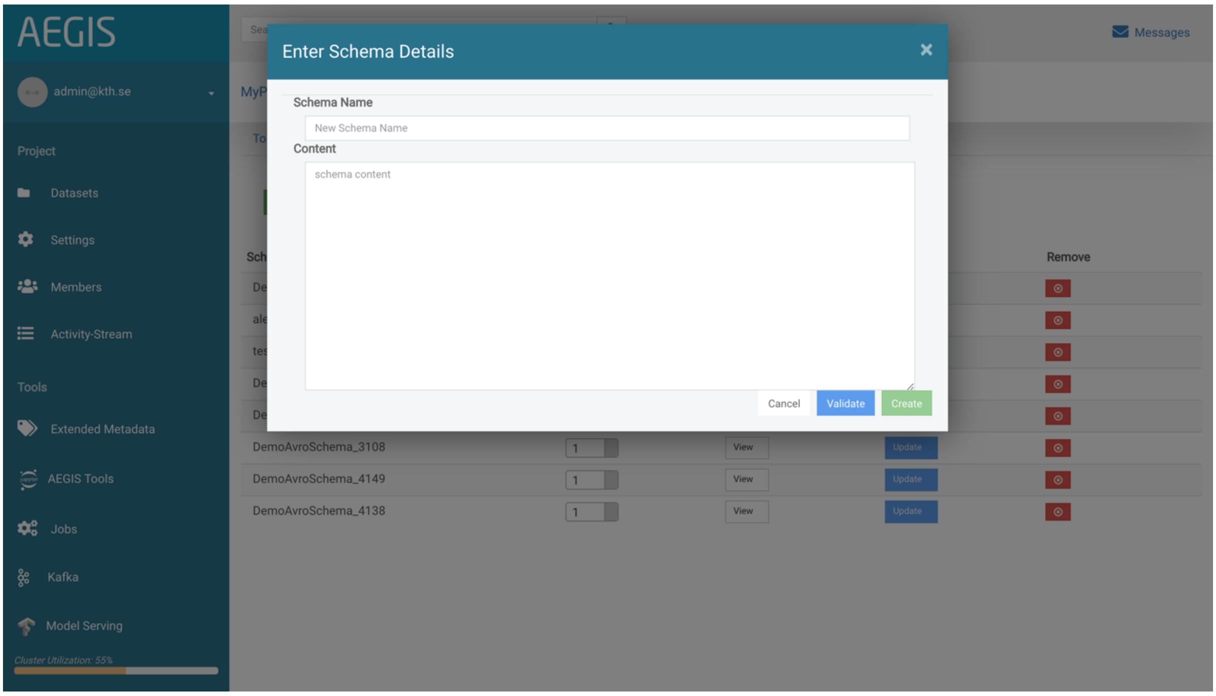


Figure 4-45: Create new Avro schema

4.1.6. Model Serving

The AEGIS platform uses Tensorflow for developing deep learning models. It provides a model serving service to deploy pre-trained model using the Tensorflow serving, as shown in Figure 4-46. The user can create a new serving by providing a model name, version, and a path, as shown in Figure 4-47.

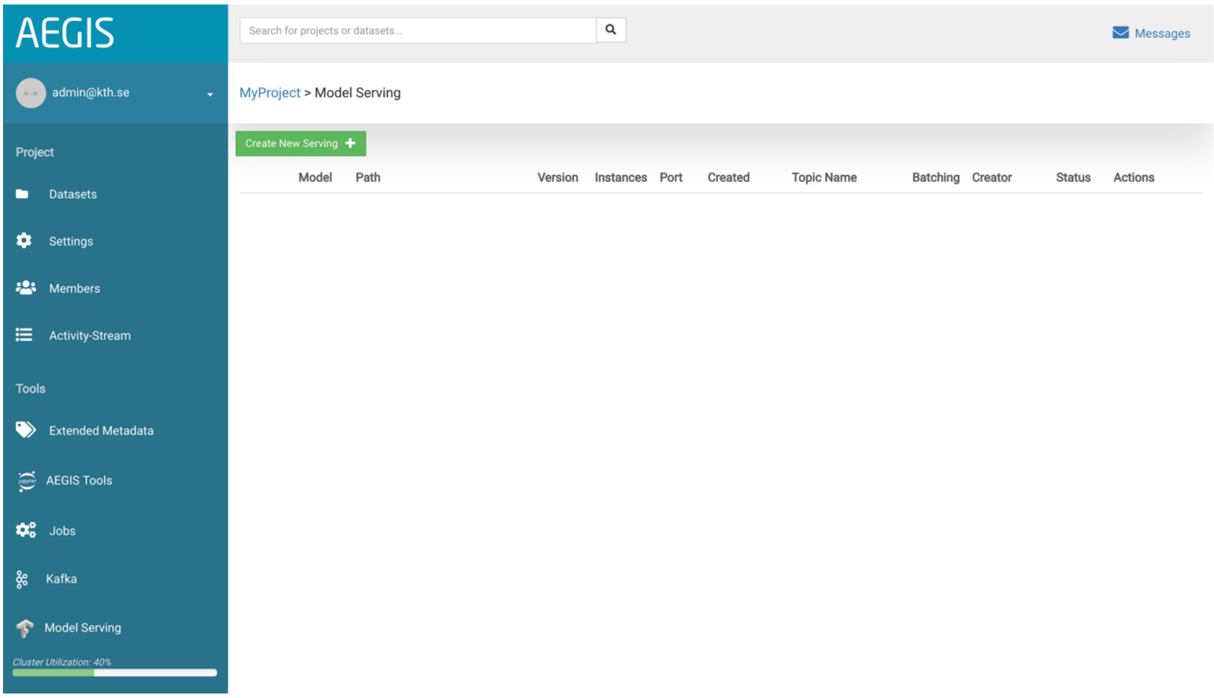


Figure 4-46: Model Serving

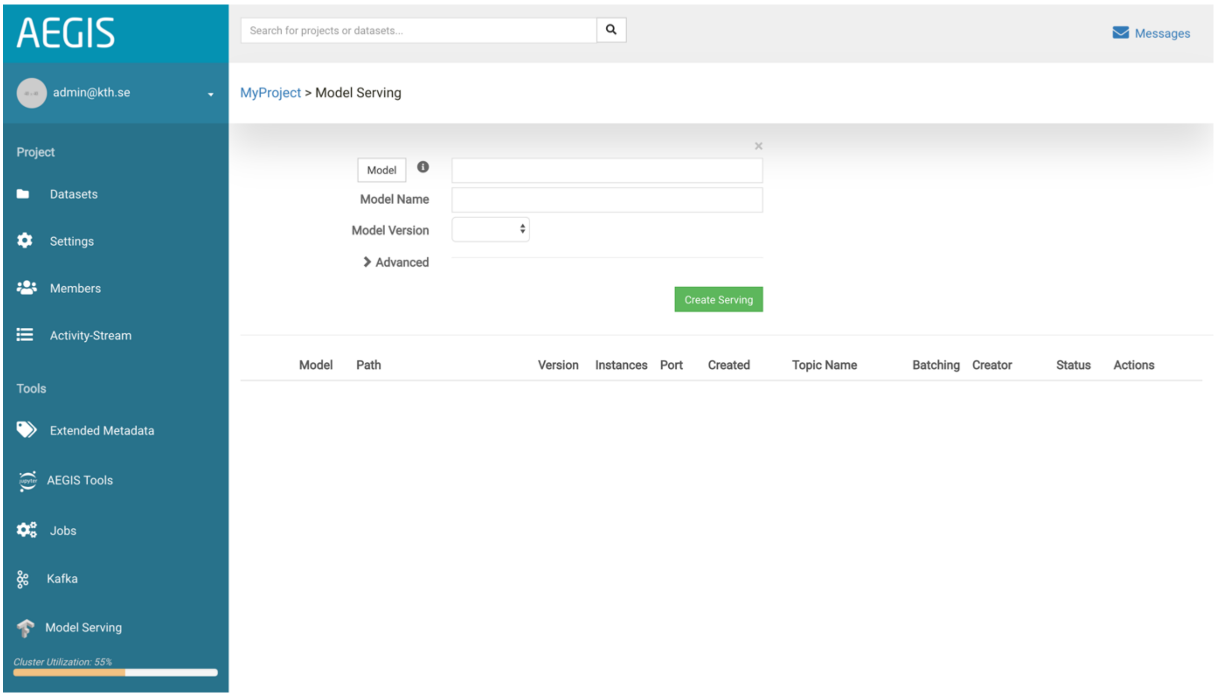


Figure 4-47: Create a new Serving

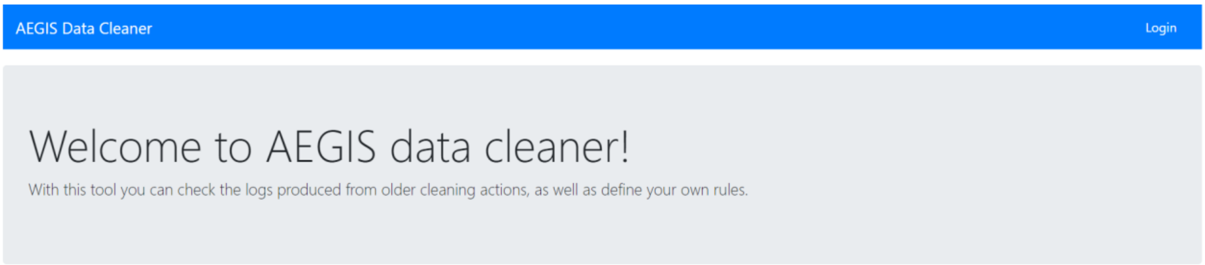
4.1.7. Cleansing Tool

Data Cleanser is an offline tool that can help data owners to apply a set of validations, cleansing and data completion actions before they upload their data in the main platform. It is offered as a Docker image to allow the easy installation in every system.

In order for a user to get started with it, she must first clone the repository that is publicly available in GitHub. Inside the project's folder there is a .env file which is used to define a series of environmental variables that are required for the application to run. The environmental variables can be set following the instructions below:

- MONGO_INITDB_ROOT_USERNAME: The username to access the MongoDB instance (MongoDB is used to store the rules configuration). The user should set it to value of his/her choice, or leave the default one.
- MONGO_INITDB_ROOT_PASSWORD: The password of the MongoDB instance.
- MONGO_PORT: The port at which MongoDB will be running on the host machine.
- SECRET_KEY: A secret to be used for securing Flask sessions. The user should set it to a long alphanumeric value.
- JWT_SECRET_KEY: A secret which is used to sign the JWT created by the application. The user should set it to a long alphanumeric value.
- FLASK_ENV: The environment under which Flask application will be running. Possible values are **dev**, **test** and **prod**. The latter is recommended.
- APP_PORT: The port under which the application will be running on the host machine.
- MONGO_URI: The full MongoDB URI which the application will use to connect to the MongoDB instance.

After environmental variable setup is complete, the user should execute the command `docker-compose -f docker-compose-prod.yml up --build -d`. This will take a couple of minutes. After that, the user should visit **http://{IP}:{PORT}** and the landing page will be displayed:



Welcome to AEGIS data cleaner!

With this tool you can check the logs produced from older cleaning actions, as well as define your own rules.

Figure 4-48: Cleansing tool - Landing page

In order to use the features of the cleansing tool, one must login first. The default username/password is admin/adminadmin123.

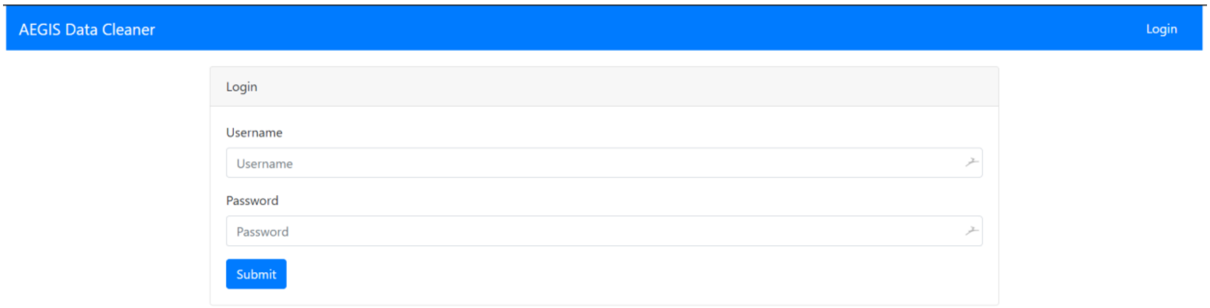
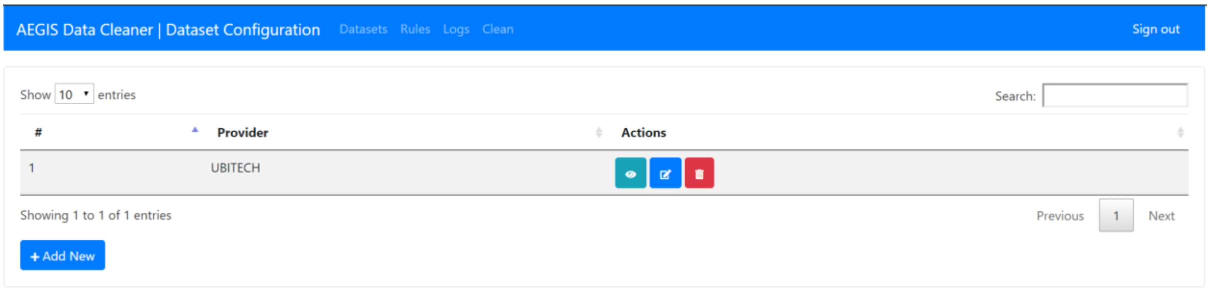


Figure 4-49: Cleansing tool - Login page

Before starting setting up cleaning rules, the datasets of interest and their variables should be registered first. The tool assumes the following hierarchy. First, the user defines the providers. A provider or data owner is the name of the company/organization/individual who possesses the data. Each provider has a set of datasets and each dataset a set of variables. Note that it is not necessary to register all of the datasets/variables, but only the ones that you are interested to clean.

The tool offers an easy to use UI for creating the aforementioned structure.

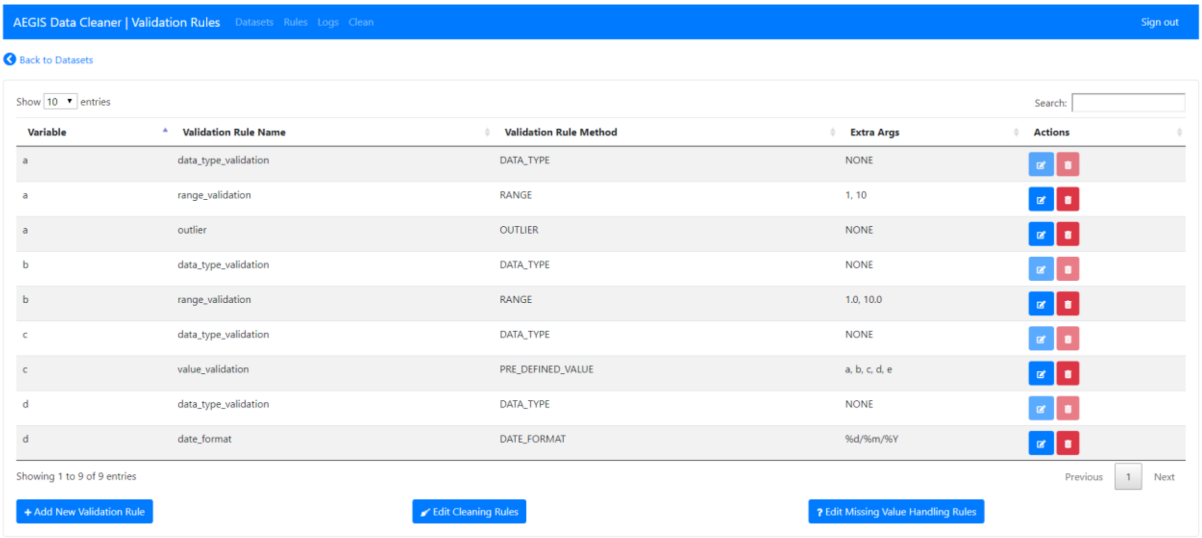


#	Provider	Actions
1	UBITECH	

Showing 1 to 1 of 1 entries

Figure 4-50: Cleansing tool - Set provider

Next, the user creates the cleaning rules of interest.



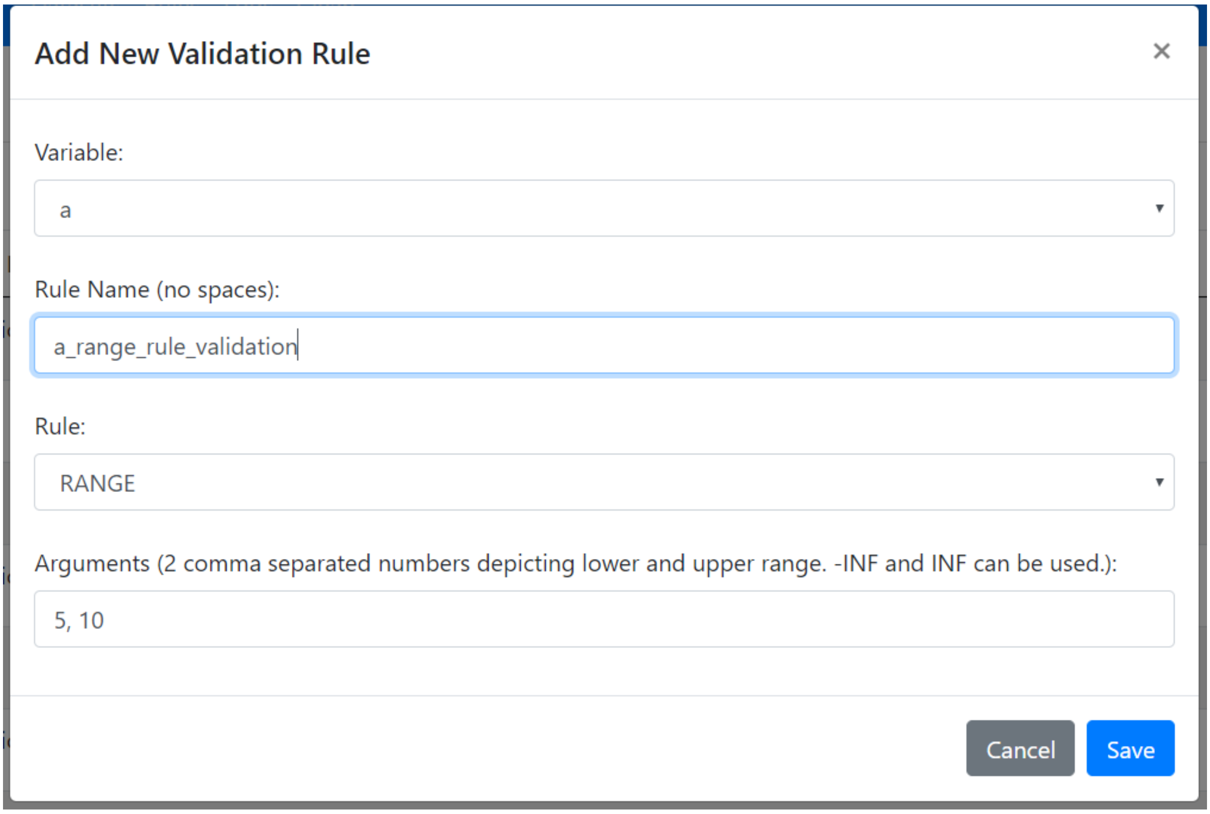
Variable	Validation Rule Name	Validation Rule Method	Extra Args	Actions
a	data_type_validation	DATA_TYPE	NONE	
a	range_validation	RANGE	1, 10	
a	outlier	OUTLIER	NONE	
b	data_type_validation	DATA_TYPE	NONE	
b	range_validation	RANGE	1.0, 10.0	
c	data_type_validation	DATA_TYPE	NONE	
c	value_validation	PRE_DEFINED_VALUE	a, b, c, d, e	
d	data_type_validation	DATA_TYPE	NONE	
d	date_format	DATE_FORMAT	%d/%m/%Y	

Showing 1 to 9 of 9 entries

Figure 4-51: Cleansing tool - Rules creation

Rules fall under three different categories:

- Validation Rules: They define constraints that should be checked (e.g. if values of a column are in a desired range)



Variable: a

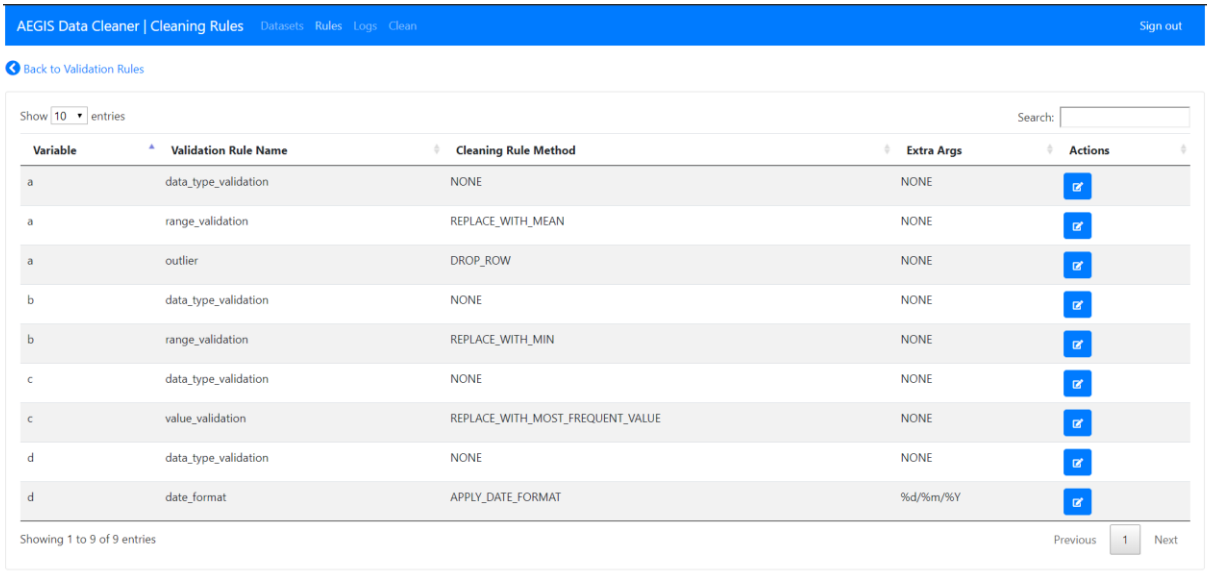
Rule Name (no spaces): a_range_rule_validation

Rule: RANGE

Arguments (2 comma separated numbers depicting lower and upper range. -INF and INF can be used.): 5, 10

Figure 4-52: Cleansing tool - Validation rule

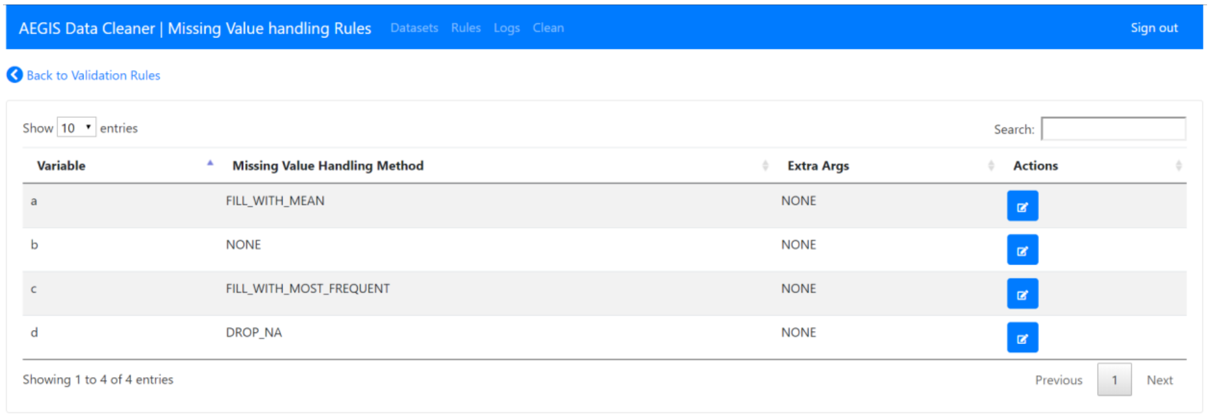
- Cleaning Rules: They define actions that should be taken, in case of a specific validation rules is being violated (e.g. if values of a column are outside a desired range, replace those values with a predefined value).



Variable	Validation Rule Name	Cleaning Rule Method	Extra Args	Actions
a	data_type_validation	NONE	NONE	
a	range_validation	REPLACE_WITH_MEAN	NONE	
a	outlier	DROP_ROW	NONE	
b	data_type_validation	NONE	NONE	
b	range_validation	REPLACE_WITH_MIN	NONE	
c	data_type_validation	NONE	NONE	
c	value_validation	REPLACE_WITH_MOST_FREQUENT_VALUE	NONE	
d	data_type_validation	NONE	NONE	
d	date_format	APPLY_DATE_FORMAT	%d/%m/%Y	

Figure 4-53: Cleansing tool - Cleansing rules

- Missing Values Rules: Define actions that should be taken, in case where there are empty values in a column.


AEGIS Data Cleaner | Missing Value handling Rules

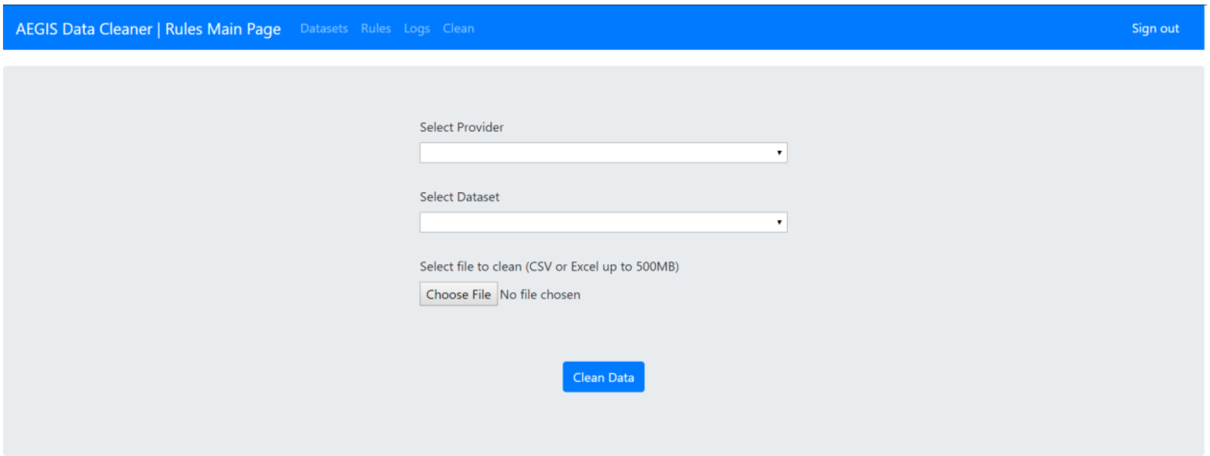
Show 10 entries Search:

Variable	Missing Value Handling Method	Extra Args	Actions
a	FILL_WITH_MEAN	NONE	
b	NONE	NONE	
c	FILL_WITH_MOST_FREQUENT	NONE	
d	DROP_NA	NONE	

Showing 1 to 4 of 4 entries

Figure 4-54: Missing values rules

Having registered the necessary rules, the user can proceed to the cleaning process. The tool offers a simple UI for choosing the necessary provider and dataset and then upload a file to clean. The only supported file types are CSV and XLSX up to 500MB in size. For larger datasets, it is advised to use the tool's API instead.


AEGIS Data Cleaner | Rules Main Page

Select Provider

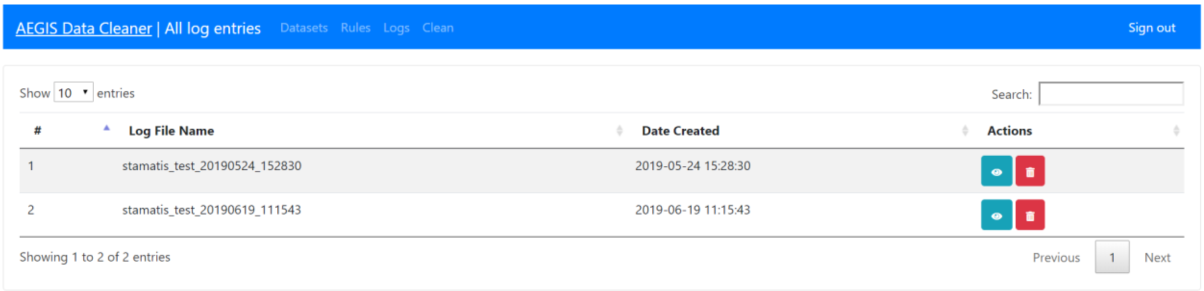
Select Dataset

Select file to clean (CSV or Excel up to 500MB)

Choose File No file chosen

Figure 4-55: Cleansing tool - Upload dataset

Once the process completes, a new cleaned file will be returned. Additionally, the cleansing tool offers a feature where the user can check all the actions in detail that took place during the cleaning process. Actions are stored as log files.


AEGIS Data Cleaner | All log entries

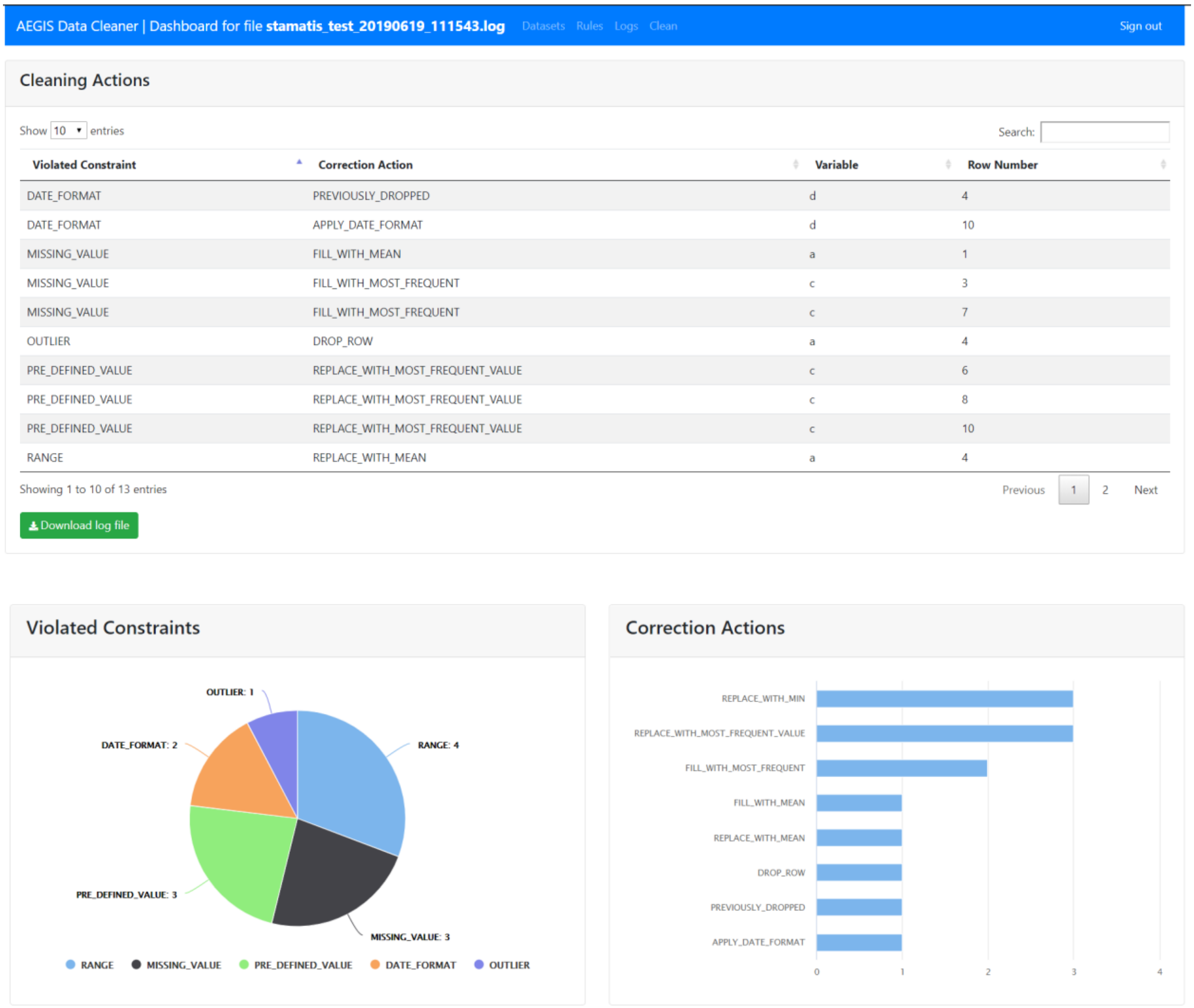
Show 10 entries Search:

#	Log File Name	Date Created	Actions
1	stamatis_test_20190524_152830	2019-05-24 15:28:30	
2	stamatis_test_20190619_111543	2019-06-19 11:15:43	

Showing 1 to 2 of 2 entries

Figure 4-56: Cleansing tool - Log files

Upon opening a specific log file, the user can get a detailed explanation of the actions that took place, as well as a nice dashboard which displays some interesting statistics.



AEGIS Data Cleaner | Dashboard for file **stamatis_test_20190619_111543.log**
Datasets
Rules
Logs
Clean
Sign out

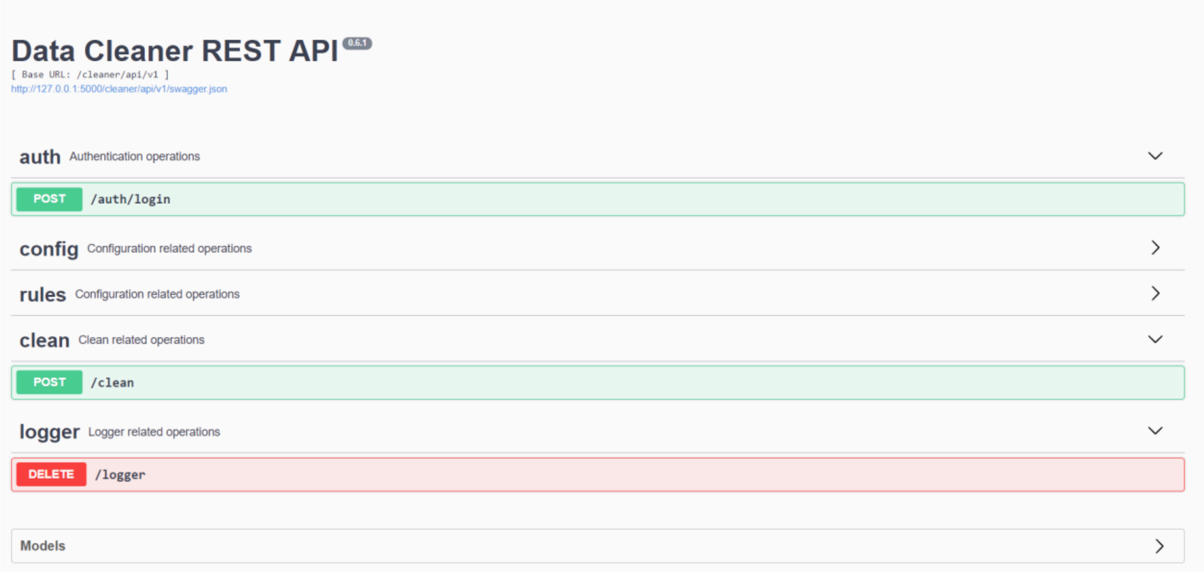
Cleaning Actions

Violated Constraint	Correction Action	Variable	Row Number
DATE_FORMAT	PREVIOUSLY_DROPPED	d	4
DATE_FORMAT	APPLY_DATE_FORMAT	d	10
MISSING_VALUE	FILL_WITH_MEAN	a	1
MISSING_VALUE	FILL_WITH_MOST_FREQUENT	c	3
MISSING_VALUE	FILL_WITH_MOST_FREQUENT	c	7
OUTLIER	DROP_ROW	a	4
PRE_DEFINED_VALUE	REPLACE_WITH_MOST_FREQUENT_VALUE	c	6
PRE_DEFINED_VALUE	REPLACE_WITH_MOST_FREQUENT_VALUE	c	8
PRE_DEFINED_VALUE	REPLACE_WITH_MOST_FREQUENT_VALUE	c	10
RANGE	REPLACE_WITH_MEAN	a	4

Showing 1 to 10 of 13 entries

Figure 4-57: Cleansing tool – Log file view

Finally, several API endpoints are being exposed from the tool. Those endpoints are documented using Swagger which can be accessed under: http://{IP}:{PORT}/cleaner/api/v1/docs



Data Cleaner REST API 0.6.1

[Base URL: /cleaner/api/v1]

http://127.0.0.1:5000/cleaner/api/v1/swagger.json

auth Authentication operations

POST /auth/login

config Configuration related operations

rules Configuration related operations

clean Clean related operations

POST /clean

logger Logger related operations

DELETE /logger

Models

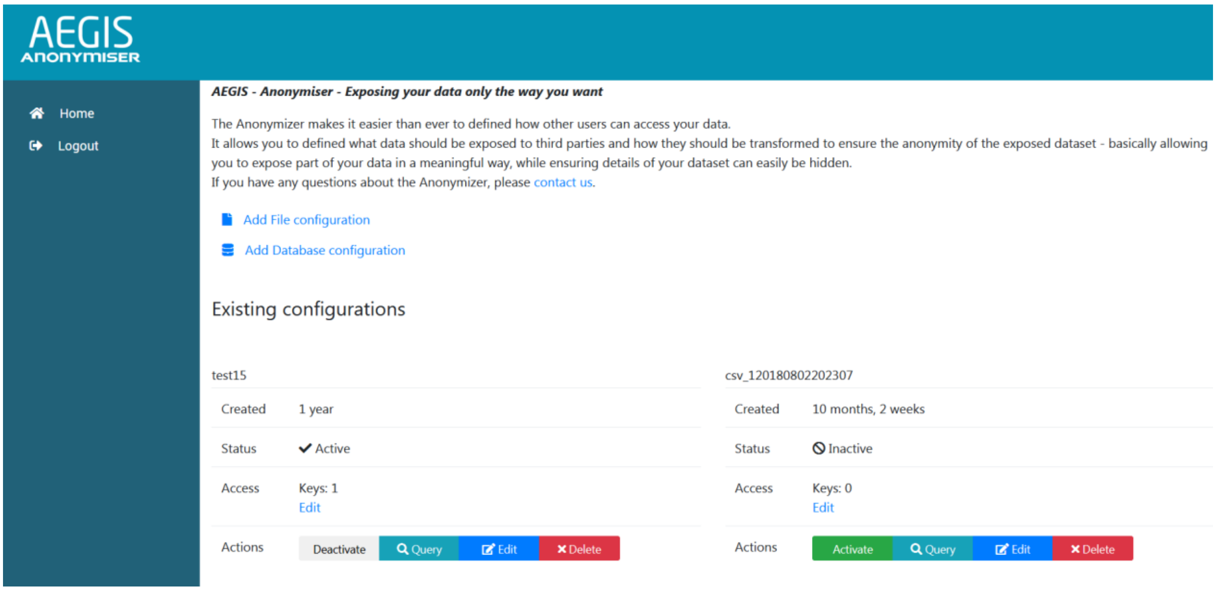
Figure 4-58: Cleansing tool - Rest API

4.1.8. Anonymisation Tool

The Anonymisation Tool is an extensible, schema-agnostic plugin that allows real-time efficient data anonymisation. The tool can used as an offline tool on the premises of the user and can configured to provide the anonymised output through a secure web API. The input of the Anonymisation Tool can be provided through a sync with a private dataset or through a file that can be imported to the tool.

In the first screen of the tool, the user is able setup or edit a saved configuration. The configuration is composed by a set of various database back-ends or text files and the corresponding anonymisation functions that will be applied on top of them.

When creating a new configuration, the tool will prompt the user to provide the connection details to the desired private database backend or select the local file to open. Additionally, the user is prompt to select the entities / tables to anonymise.



AEGIS - Anonymiser - Exposing your data only the way you want

The Anonymizer makes it easier than ever to defined how other users can access your data.
It allows you to defined what data should be exposed to third parties and how they should be transformed to ensure the anonymity of the exposed dataset - basically allowing you to expose part of your data in a meaningful way, while ensuring details of your dataset can easily be hidden.
If you have any questions about the Anonymizer, please contact us.

Existing configurations

test15

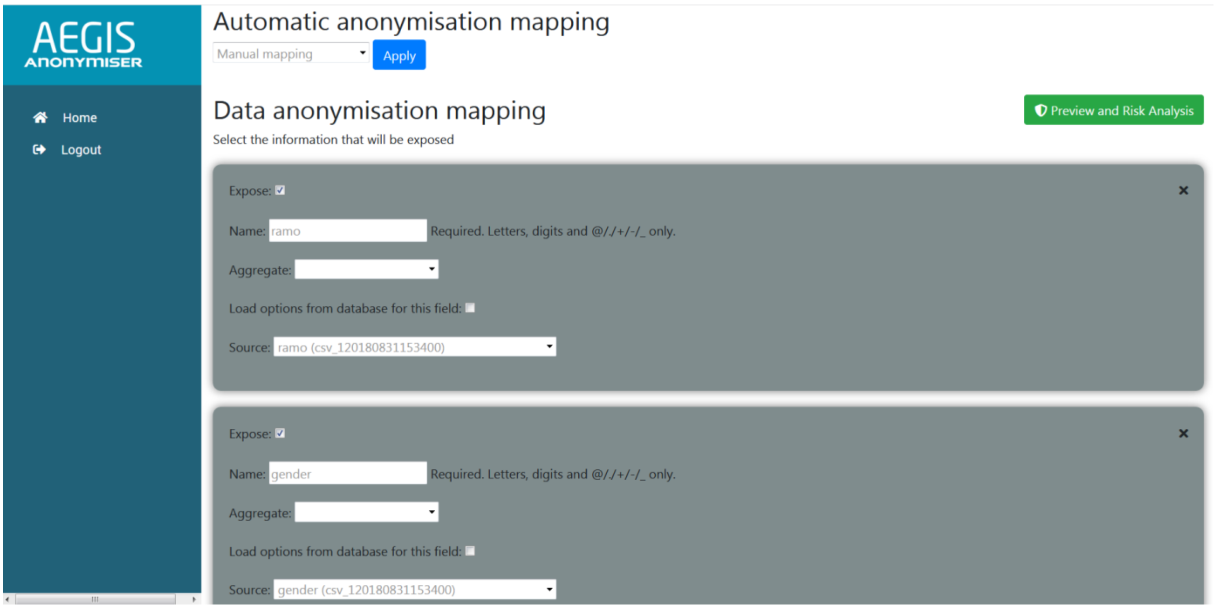
Created	1 year
Status	✓ Active
Access	Keys: 1 Edit
Actions	Deactivate Query Edit Delete

csv_120180802202307

Created	10 months, 2 weeks
Status	Inactive
Access	Keys: 0 Edit
Actions	Activate Query Edit Delete

Figure 4-59: Anonymisation tool - Configuration screen

At the following step, the tool prompts the user to select: a) the fields from the data source that will be exposed to the anonymised set and b) the anonymisation function that will be performed.



Automatic anonymisation mapping

Manual mapping

Data anonymisation mapping

Select the information that will be exposed

Expose: ☑

Name: ramo Required. Letters, digits and @/./+/-/_ only.

Aggregate:

Load options from database for this field: □

Source: ramo (csv_120180831153400)

Expose: ☑

Name: gender Required. Letters, digits and @/./+/-/_ only.

Aggregate:

Load options from database for this field: □

Source: gender (csv_120180831153400)

Figure 4-60: Anonymisation tool - Data selection

The Anonymisation Tool offers a list of predefined anonymisation functions that can be used directly (e.g. city from an exact address, range of values from an integer), as well as a list of aggregation functions (e.g. average).

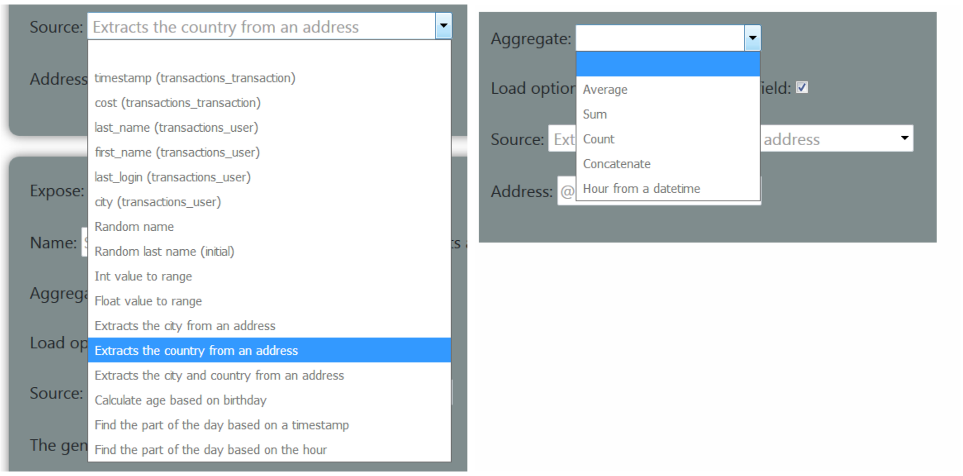
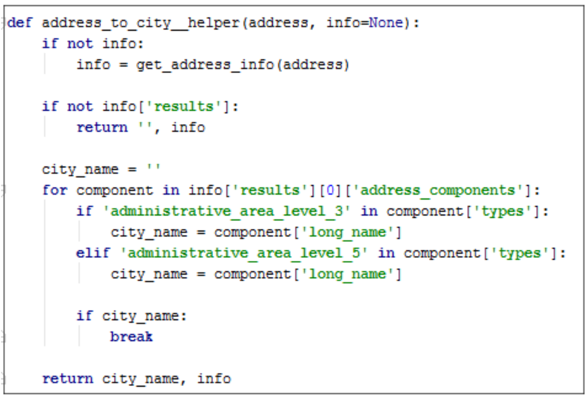

Source:
Extracts the country from an address
Address
timestamp (transactions_transaction)
cost (transactions_transaction)
last_name (transactions_user)
first_name (transactions_user)
last_login (transactions_user)
city (transactions_user)
Random name
Random last name (initial)
Int value to range
Float value to range
Extracts the city from an address
Extracts the country from an address
Extracts the city and country from an address
Calculate age based on birthday
Find the part of the day based on a timestamp
Find the part of the day based on the hour
Expose:
Name:
Aggrega
Load op
Source:
The gen
Aggregate:
Average
Sum
Count
Concatenate
Hour from a datetime
Load option
ield: ☑
Source: Ext
address
Address: @

Figure 4-61: Anonymisation tool - Indicative functions

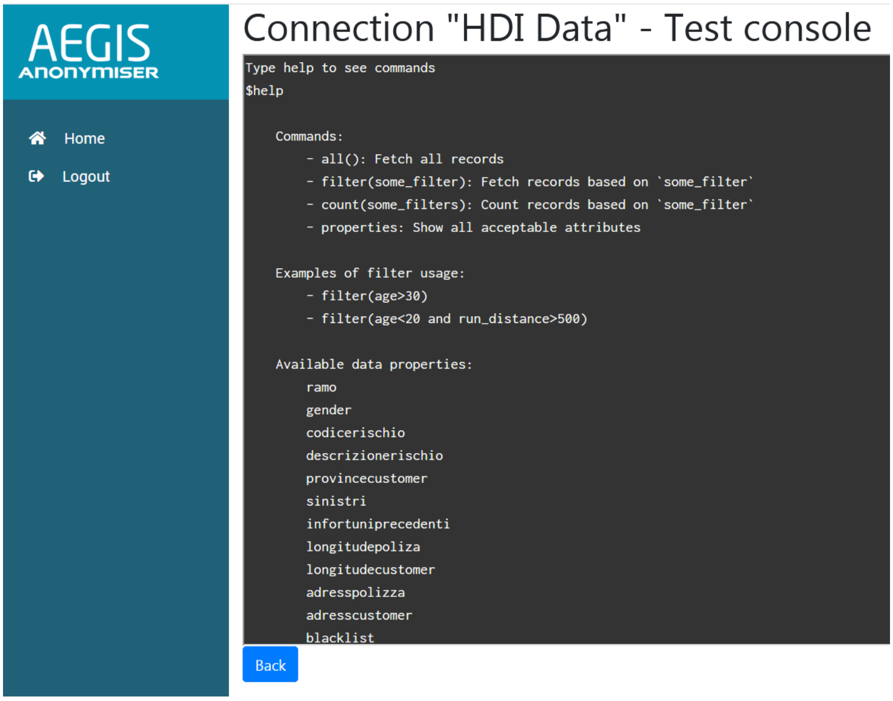
Furthermore, the tool offers the extension of the supported anonymisation functions with custom anonymisation functions. In this case, the user should provide the corresponding python module.



```
def address_to_city__helper(address, info=None):
    if not info:
        info = get_address_info(address)

    if not info['results']:
        return '', info

    city_name = ''
    for component in info['results'][0]['address_components']:
        if 'administrative_area_level_3' in component['types']:
            city_name = component['long_name']
        elif 'administrative_area_level_5' in component['types']:
            city_name = component['long_name']

        if city_name:
            break

    return city_name, info
```

Figure 4-62: Anonymisation tool - Custom user-defined function

Additionally, the tool offers an integrate console to the user in order to perform any queries or test the output of the anonymisation process.

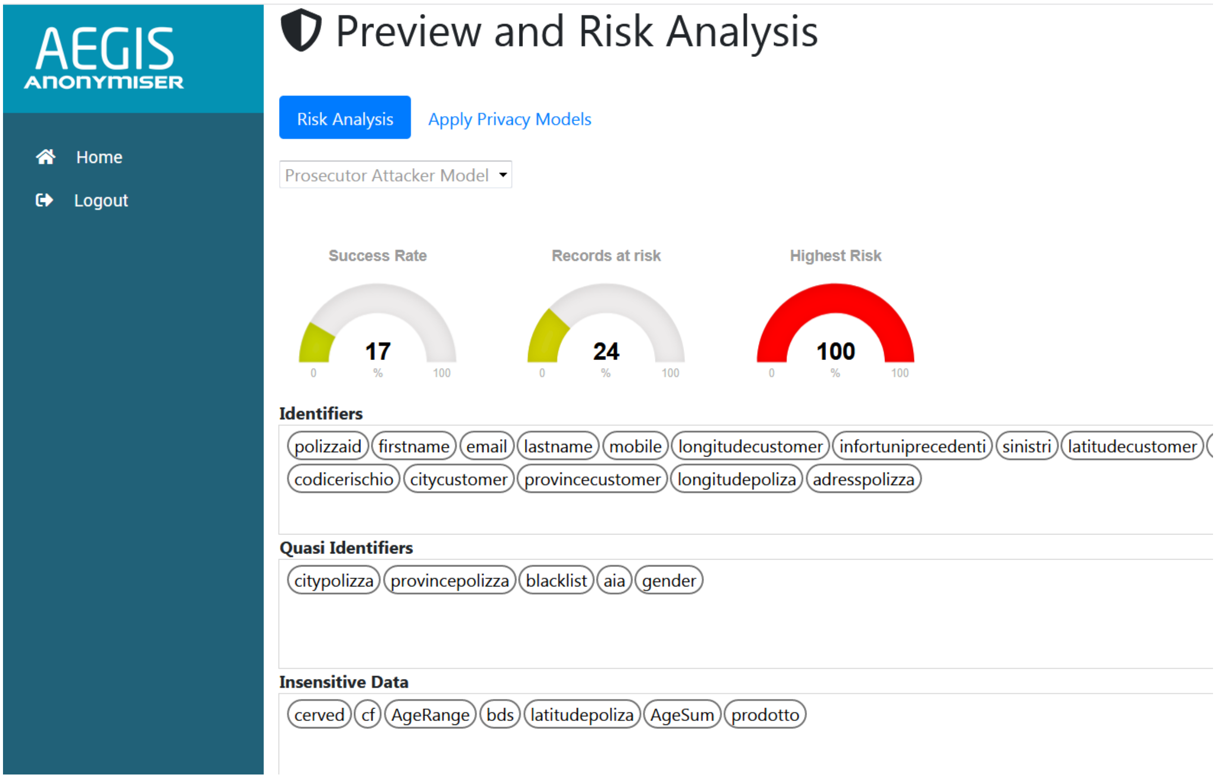


Connection "HDI Data" - Test console

```
Type help to see commands
$help

    Commands:
        - all(): Fetch all records
        - filter(some_filter): Fetch records based on `some_filter`
        - count(some_filters): Count records based on `some_filter`
        - properties: Show all acceptable attributes

    Examples of filter usage:
        - filter(age>30)
        - filter(age<20 and run_distance>500)

    Available data properties:
        ramo
        gender
        codicerischio
        descrizionerischio
        provincecustomer
        sinistri
        infortuniprecedenti
        longitudepoliza
        longitudecustomer
        adresspolizza
        adresscustomer
        blacklist
```

Figure 4-63: Anonymisation tool - Output

Moreover, the user can assess the risk levels of the selected anonymisation mapping via the Preview and Risk analysis, as depicted in the following figure.



Preview and Risk Analysis

Identifiers

polizzaid, firstname, email, lastname, mobile, longitudecustomer, infortuniprecedenti, sinistri, latitudecustomer, codicerischio, citycustomer, provincecustomer, longitudepoliza, adresspolizza

Quasi Identifiers

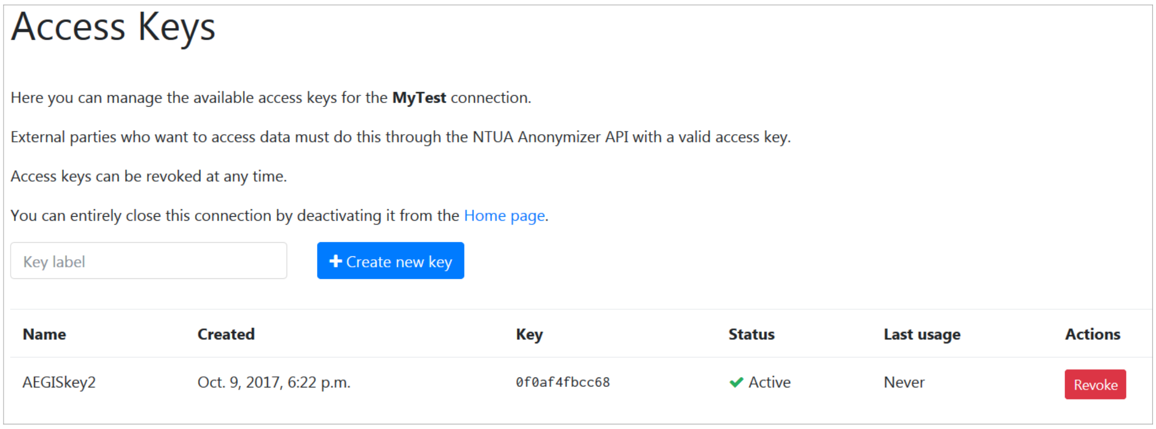
citypolizza, provincepolizza, blacklist, aia, gender

Insensitive Data

cerved, cf, AgeRange, bds, latitudepoliza, AgeSum, prodotto

Figure 4-64: Anonymisation tool - Preview and Risk analysis

As described above, the tool offers to option to expose the output data via a secure API to external parties. To enable this, the provisioning of access keys is mandatory.



Access Keys

Here you can manage the available access keys for the **MyTest** connection.

External parties who want to access data must do this through the NTUA Anonymizer API with a valid access key.

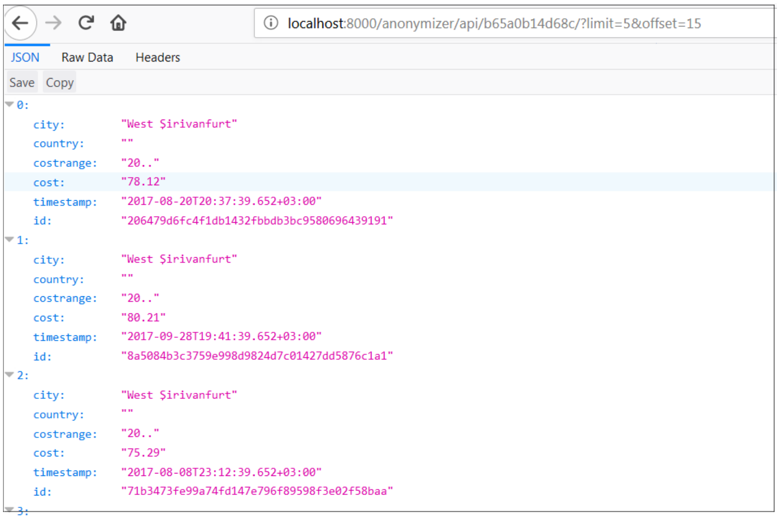
Access keys can be revoked at any time.

You can entirely close this connection by deactivating it from the Home page.

Name	Created	Key	Status	Last usage	Actions
AEGISkey2	Oct. 9, 2017, 6:22 p.m.	0f0af4fbcc68	✓ Active	Never	Revoke

Figure 4-65: Anonymisation tool - Exposed API

In this case, the user can access the produced anonymised data through the secure API in JSON format by utilising the private access keys configured above.



localhost:8000/anonymizer/api/b65a0b14d68c/?limit=5&offset=15
JSON
Raw Data
Headers
Save
Copy

```
▼0:
    city:         "West Şirivanfurt"
    country:      ""
    costrange:    "20.."
    cost:         "78.12"
    timestamp:    "2017-08-20T20:37:39.652+03:00"
    id:           "206479d6fc4f1db1432fbbdb3bc9580696439191"
▼1:
    city:         "West Şirivanfurt"
    country:      ""
    costrange:    "20.."
    cost:         "80.21"
    timestamp:    "2017-09-28T19:41:39.652+03:00"
    id:           "8a5084b3c3759e998d9824d7c01427dd5876c1a1"
▼2:
    city:         "West Şirivanfurt"
    country:      ""
    costrange:    "20.."
    cost:         "75.29"
    timestamp:    "2017-08-08T23:12:39.652+03:00"
    id:           "71b3473fe99a74fd147e796f89598f3e02f58baa"
▼3:
```

Figure 4-66: Anonymisation tool - API response

4.2. AEGIS Platform adoption guidelines

The AEGIS platform has a lot of features to support rapid prototyping and development of data-driven applications for various stakeholders. In general, there are two types of potential users of the AEGIS platform, an experienced user with a data science background who can write and execute data management tasks using notebooks and code generation, or a non-technical user who prefers a simple user interface and uses the application developed by the data scientist on the AEGIS platform to just perform simple, prepared tasks such as loading prepared data into the Visualizer, creating a visualization and interacting with it to support his decision-making process.

This section briefly explains how data scientists can make the most of the potential of the AEGIS platform in own data science projects and how they can use the platform to develop an application meeting the users' needs.

The first step in a data science project is to identify a concrete need of a target group that a data scientist wants to meet through the data-driven application to be developed. This is a complex challenge and data scientists must position themselves into the target audience, which requires also a lot of communication and requirements engineering skills. It is of utmost important that data scientist understand the domain addressed sufficiently. Moreover, it should be examined, if the data (made available by the target group or received by other means) is capable of meeting the needs of the target group. Several approaches can be chosen for this needs-analysis, e.g. interviews with stakeholders, workshops or focus groups. One practical idea is to document – together with representatives of the target group – the as-is situation and the challenge to solve, the to-be situation to be achieved, the technical solution approach to develop (the data-driven application) and the expected impact of using this application for the target group. Achieving a common understanding of the big picture very early is one success factor of data science projects.

In a second step, the data scientist must comprehensively analyse the needs of the target group to elicit concrete requirements for the data-driven application to develop. Therefore, it is important to determine the system boundary and define which aspects should be covered by the envisaged data-driven application, and which should not. Requirements can be documented in full sentences, however, a user story with some mock-ups of the application should be developed, too, describing the interaction with the application from the perspective of the user. There are several methods which can be applied published in the requirements management literature.

In a third step (which is of course also possible to be conducted at the beginning of the process, but probably the motivation of exploring the AEGIS platform is much higher having a concrete project to implement on the platform in mind), the data scientist must familiarise himself with the AEGIS platform and the available possibilities. The AEGIS documentation effectively guides the data scientist through the first steps of getting to know the platform and its modules. A good idea is to start with creating a demo project, uploading some (known) demo data, loading this demo data in a notebook, doing some simple computation and probably visualizing the result to better know how the platform works. The next step is then to examine the AEGIS modules, the Visualiser, the Query Builder, the Algorithm Execution Container, and the Data Harvester. Using these modules can save data scientists a lot of time. In addition, the AEGIS platform supports data-parallel processing services such as MapReduce, Spark and Flink as well as interactive analyses with notebooks such as Jupyter. Real-time analysis is enabled through the use of the integrated Kafka service. Deep learning is made possible by the use of the included Tensorflow service. Support for the above services allows data scientists to develop different types of applications according to their needs and then seamlessly implement and test them on AEGIS platform. But of course, the data scientist must get familiar with all these features first.

Once the data scientist is sufficiently familiar with the platform, he/she can design a concept for the data-driven application to develop and then carefully plan the steps for implementation. One way to be successful is to describe a set of scenarios with a set of test cases to be implemented on the AEGIS platform.

The next step is then to execute these defined steps on the platform. Once the user is satisfied and confident with the current plan of how to implement the application, he can begin the coding phase by developing the application using his framework of interest from the applications supported by AEGIS (MapReduce, Spark, Flink, Tensorflow and Kafka). It is always recommended to test the data-driven application/ service developed closely together with the selected target group already at an early stage to determine whether the application and the way it should be used is suitable for solving the identified needs of the target group. There are several ways to get user feedback, such as thinking aloud, interviewing or focus groups.

Based on the received user feedback and ideas, the data scientist and the target audience can anticipate several phases of application evaluation and application iteration until the data-driven application meets the expectations of both the data scientist and the target audience. This is a usual practice in design-based projects for meeting the requirements of users in a best possible way.

5. CONCLUSION

The scope of deliverable D5.6 as the final deliverable of WP5 was to document the efforts undertaken within the context of the tasks 5.3, 5.4, 5.5, and 5.6 and provide a final evaluation, impact assessment and adoption guideline. It was built on top of the work and outcomes of the deliverables D5.1, D5.2, D5.3, D5.4, and D5.5 towards the aim of providing a summarized assessment of the AEGIS platform and demonstrators, lessons learnt made during demonstrator implementation and finally a how to use the AEGIS platform in own data science projects.

At first, the deliverable provided a summarized result of the AEGIS demonstrators final evaluation, covering both the quantitative and the qualitative evaluation as well as an overview of each implemented demonstrator along with the activities performed.

Following the AEGIS demonstrators' final evaluation, the deliverable presented the AEGIS platform final evaluation, which was conducted quantitatively and qualitatively, too. Additionally, the document presented the results of a security assessment as the consortium decided to perform a security assessment of the AEGIS platform in order to ensure that its security mechanisms are in are operating efficiently and effectively.

Finally, the deliverable presented the AEGIS platform documentation and adoption guidelines featuring a platform usage documentation, as well as the platform adoption guidelines as a how-to best exploit the potential of the AEGIS platform in own data science projects.

The outcomes and knowledge extracted from this deliverable will serve as valuable feedback and information for the exploitation phase, following the AEGIS project.