PAPUD

ITEA3 - 16037

Profiling and **A**nalysis **P**latform **U**sing **D**eep Learning

D4.1.2 Use case Demonstrators Report

Release : 1.0
Status : Final
Date : 30th September 2020
Author : Elena Muelas (HI-Iberia)
Visibility : Public

HISTORY OF CHANGES

Version	Date	Changes	Author
V 0.1	10/09/2020	Adding UC5 information	Marc PLATINI
V 0.2	27/09/2020	Adding UC4 information	Laurent GONCALVES, Romain BELMONTE, Marius BILASCO, Hubert NOURTEL
V0.3	7/10/2020	Adding UC3 information	Elena Muelas Inmaculada Luengo
V0.4	13/10/2020	Adding UC1 information	Patrick Constant
V0.5	20/10/2020	Adding UC2 information	Ahmet Servet

Table of contents

Figures

1 Executive summary

This document intends to serve as user manual for each of the use cases develod during the PAPUD project execution. This document will also contain the details for accessing each of the use cases and interact with them from a end user point of view.

2 UC1: E-Commerce

2.1 Use Case short description

The E-commerce use case UC1 purpose is to demonstrate the use of artifial intelligence methods to create a recommendation system for clients of an e-commerce website. This use case was a joint collaboration with :

- Setur
- KocSistem
- IMT
- Pertimm

We demonstrate the use case using data from Setur which consist of 31.000 products sold at Turkish airport duty free shops. We use also purchase data (baskets) to determine popularity of the products and calculate recommendation based upon a Deep Learning approach.

Recommandation are shown to the user in two different ways:

- When showing a product, we propose several recommended products to buy based on this product:

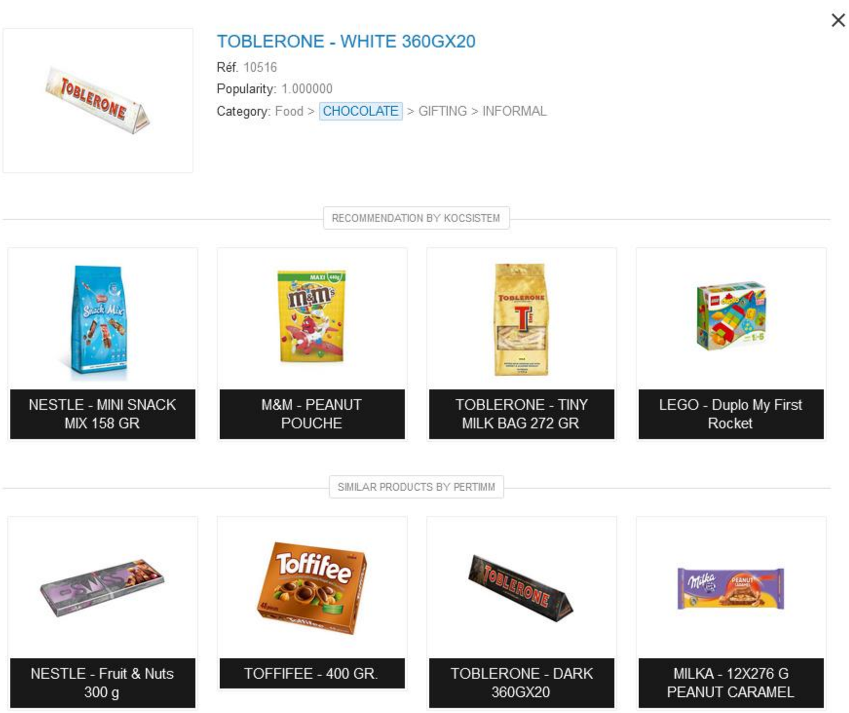


Figure 1 – several recommended products based on a given product

- When showing a result list using a search engine, we propose several recommended products, base on the 5 top products in the results list:

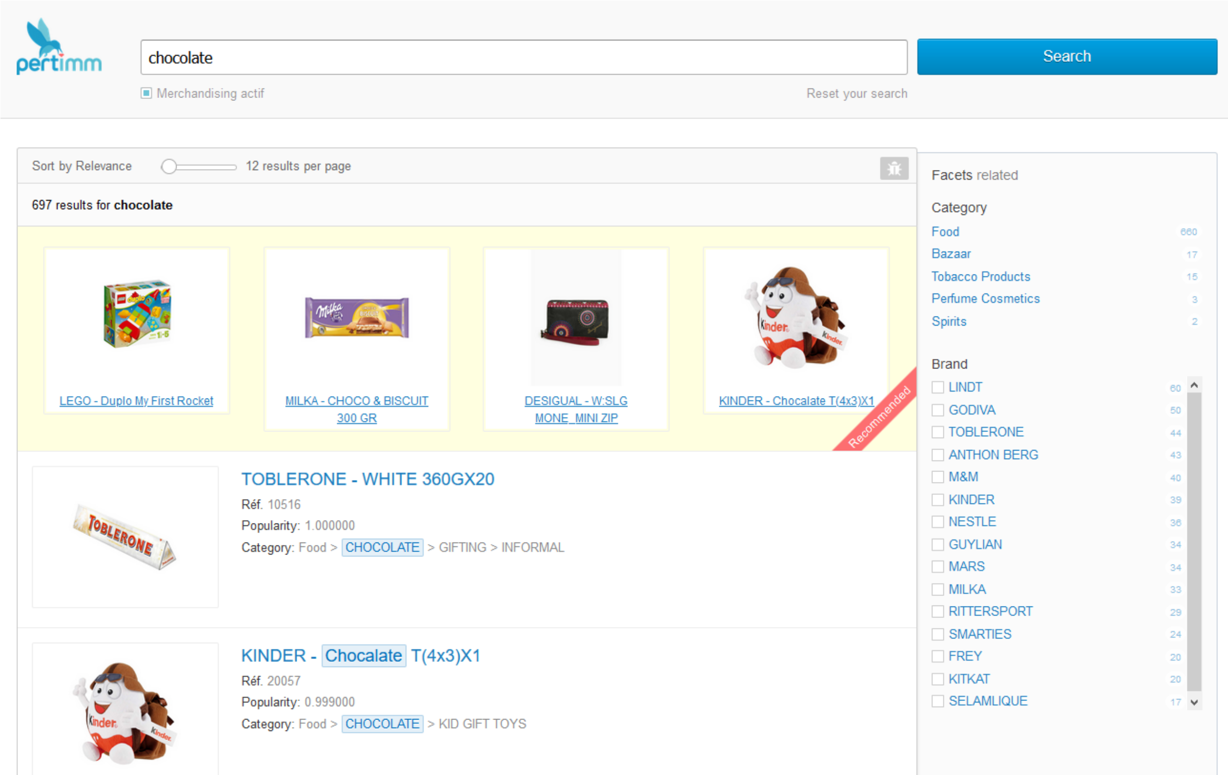


Figure 2 - several recommended products based on an list of products

2.2 Use Case details

2.2.1 Use Case access

The demonstration is accessible at this address:

https://papud-demo.pertimm.net/004-demo-papud-recommendation/search/?user_locale=en

Connexion can be done used the following information:

- User: itea
- Password: itea2020

2.2.2 Use Case final architecture overview

The final architecture overview is the following :

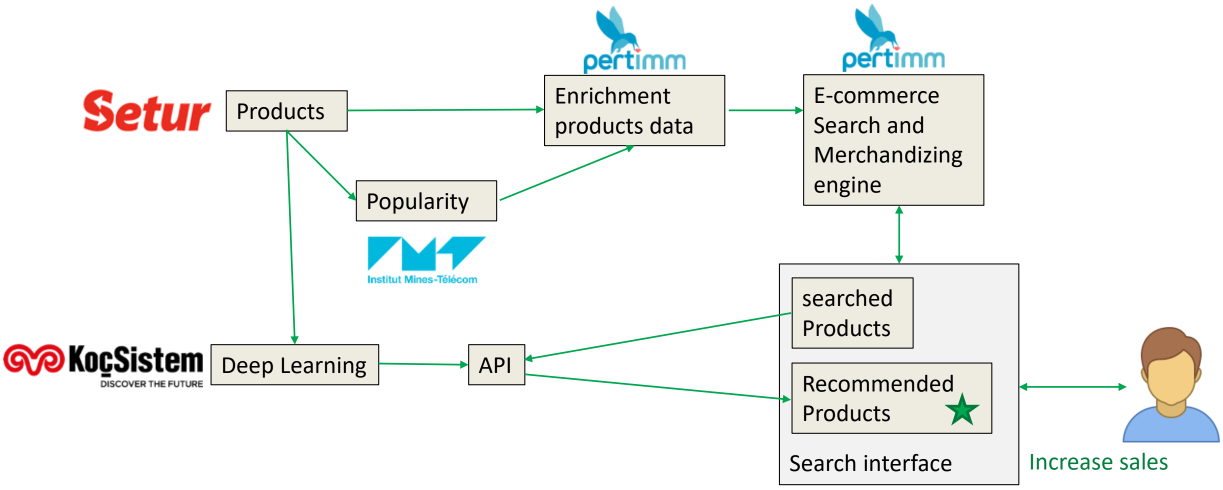


Figure 3 - UC1 architecture

- Setur provides the products,
- IMT calculates a popularity score for each product,
- KocSistem uses a Deep Learning system to calculate recommandations,
- Pertimm indexes all data from Setur enriched with popularity,
- Pertimm creates a search engine interface and integrates KocSistem's Deep Learning mechanism through an API.

2.2.3 Use Case modules and functionalities

When connecting to the demonstration, the customer can search for any product in the setur database. For example, he is looking for "toblerone white". Base on the result list he gets a list of searched products (here two Toblerone White products) as well a a list of recommended products (located on the yellow background part of the demonstration) :

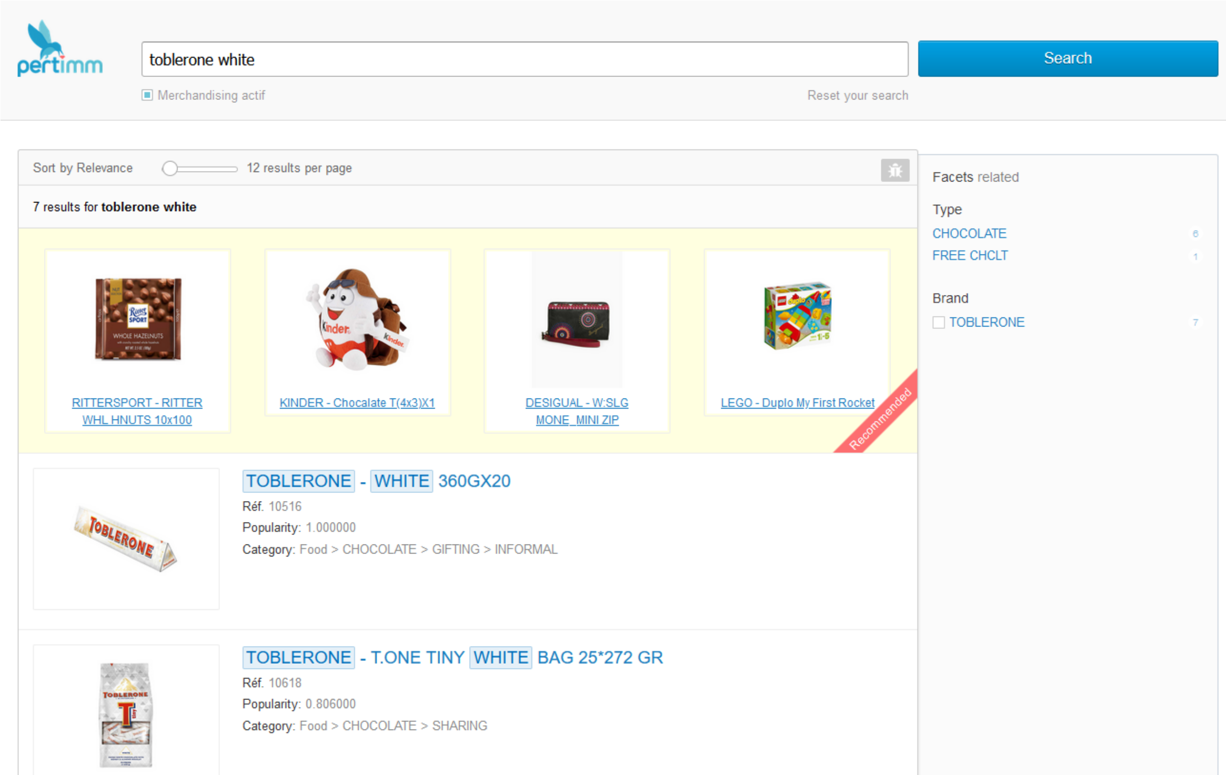


Figure 4 -The customer gets a list of recommended products

The searched product are using the popularity calculations that have been done by the module developed by IMT to sort the products (here with a white background).

The recommand product are calculated by the KocSistem Deep Learning module, and requested by the Pertimm's interface to get the list of product that is shown within the yellow background.

When the customer is clicking on a product, another recommended list of products is given to the customer, those products are calculated by the KocSistem Deep Learning module, based upon this product (here with The toblerone bar) :

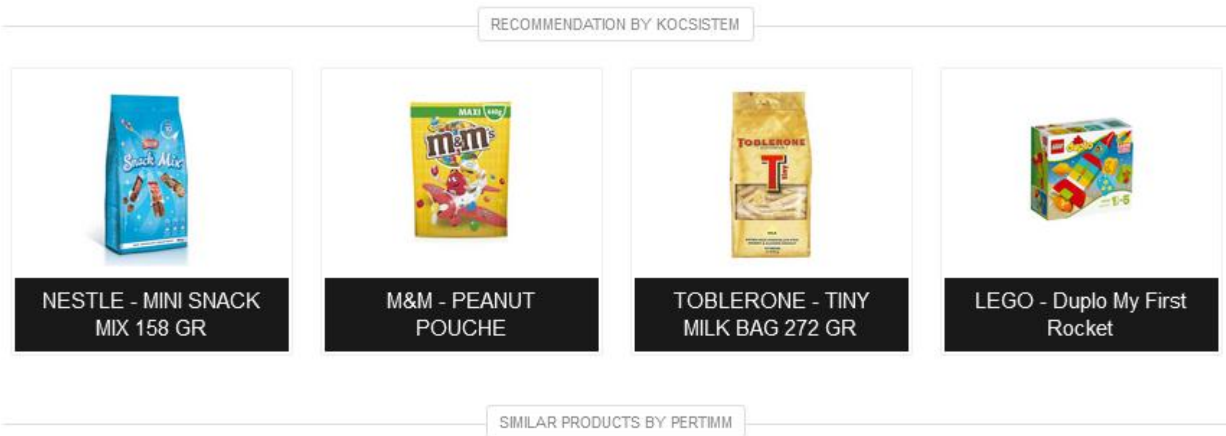


Figure 5 - Kocsistem DL calculated products

2.2.4 Input data and expected results

Input data are coming from Setur's Duty-free:

Stores:

- Sabiha Gökçen, Gaziantep ve Samsun Airports,
- Sea Ports ve
- Land Border Crossing

Products:

- 31K total
- 1 main category and 3 sub-categories
- Used individual products for recommendation
- Grouped the product based on brand and sub categories (~2000)

Current focus: Sabiha Gökçen Departures Store

- 7 stores with seperate cashiers
- Calculations based on unique transaction ID
- Calculations based on unied customer ID from the same day (for cross sell)

Expected result after the processing is a recommendation system integrated in the Pertimm's search engine.

3 UC2: Call Centers

In this use case, we have 2 projects where one is focused on building a solution for multilingual service centres (named TellMi) and the other one on building a solution for the Turkish market (named ChatMe).

3.1 TellMi - Use Case short description

Customers contact companies more and more through text-based channels. The analysis of these texts is often a manual process which makes it very time-consuming and expensive and is therefore not carried out.

Moreover, companies in Belgium face an extra challenge because of the multiple official languages spoken within a relatively small client base. To automate text analysis, it is important to have large data sets. By dividing the group of customers based on language, a lot of potential knowledge is lost.

The objective of this use case is to automate the extraction and collection of the insights from customer interactions such as as the logs provided by call agents in different languages at the same time. Our platform TellMi is able to give an overview of what customers are talking about and how they feel about it across different languages and in real time. Two models have been developed and integrated into the platform:

- Cross-lingual aspect-based sentiment model, which helps annotate topics and corresponding sentiment to customer interactions in different languages
- Cross-lingual fine-grained topic model, which helps reveal the hidden topics in customer interactions in different languages when no labelled data is available (thus topics are unknown).

This is a collaboration between 4C and KU Leuven.

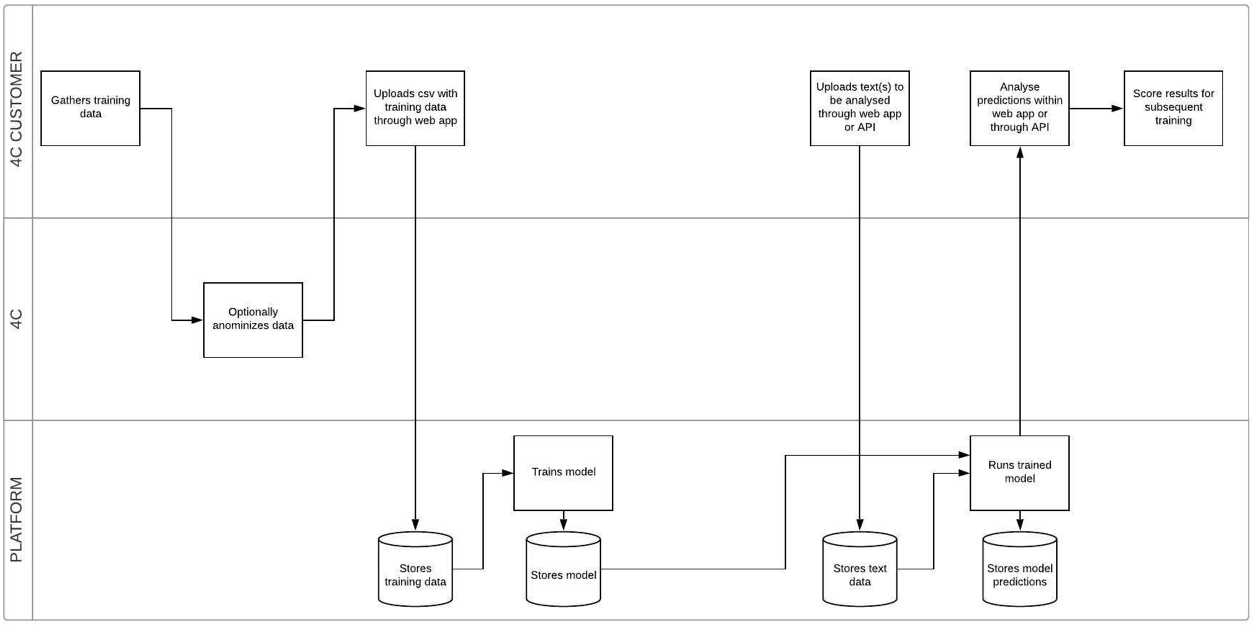
3.2 TellMi - Use Case details

3.2.1 Use Case access

An application for this use case has been built using Heroku and AWS. A development environment is up and running. Once the platform Tellmi goes live, there will be a production environment set up. For testing and demo purposes, it was integrated in a Salesforce development environment.

3.2.2 Use Case final architecture overview

This is the final architecture overview:



Users of the TellMi platform will first need to upload their training data set to train a model specifically for there use case. This can be done through the web interface. Then the trained model is applied to new incoming text to extract insights which are collected eighter through the web interface or by calling the API. The model predictions are then presented in an report or used to automate e.g. picklists, work flows, … .

3.2.3 Use Case modules and functionalities

Once logged into the TellMi platform, projects and teams can be created to monitor the access of users to the different use cases.

To extract insights from text, there are 2 models available on the platform: a cross-lingual aspect-based sentiment model and a cross-lingual fine-grained topic model. A model is trained by uploading a csv file containing labelled text data. It is also possible to set model training parameters such as number of epochs and batch sizes. The status of the training process is indicated.

Once ready, the trained model can be applied to new text documents to extract insights either on the platfom directly or by making an API call. In the first case, simply upload (a set of) new text document(s) onto the platform and it will return an overview of the predictions as well as the predictions per text document, for which the user can indicate whether it was correct or not. This is to monitor the performance of the model over time.

Ideally the predictions are directly integrated in the users work environment. The API keys needed to integrate the platform are also available in the application. This way the predictions can be used to automate and speed up e.g. workflows and present up-to-date overviews of the insights across different languages in real time.

3.2.4 Input data and expected results

The TellMi platform contains two models and each model has its own specific input and output.

- Cross-lingual aspect-based sentiment model:
  This model is trained using topic and sentiment labelled customer interactions in multiple languages such as e.g. call logs in Dutch and French. Once the model is trained, it is applied to similar customer interactions to predict the topics and corresponding sentiment. So that in case of the call agency, new incoming logs can automatically be classified and a complete overview of why customers are calling and how they feel is available real time across all languages.

- Cross-lingual fine-grained topic-model:
  For this model a multilingual set of customer interactions, of which a small sample is keyword annotated, is needed. The keyword labels are used to train a multilingual keyword extraction model, which provides an extra input for the topic model. The output of this model are the hidden fine-grained topics in the text in different languages.

3.3 ChatMe - Use Case short descritpion

Today; Companies are spent huge amount of money for increasing Customer Experience and Customer satisfaction all over the world. Data scientists are working on these hot topics that are analyzing the customers' behaviors and conversations with call center employees to manage customer experience and satisfaction. Turkish consortium is working on analytics of data that is collected from call centers. There are 5 stages that are shown below to reach the valuable results.

1) Intent classification: This is the process of determining the reason for starting the dialogue.

2) Customer satisfaction measurement: It will be determined how the dialogue starts, progresses and ends with the help of sentiment analysis on different parts of the dialogue.

3) Agent performance measurement: Measuring criterion: Agent response times, Wrongly written word ratio, Change of mood values at the beginning and end of dialogues (customer satisfaction)

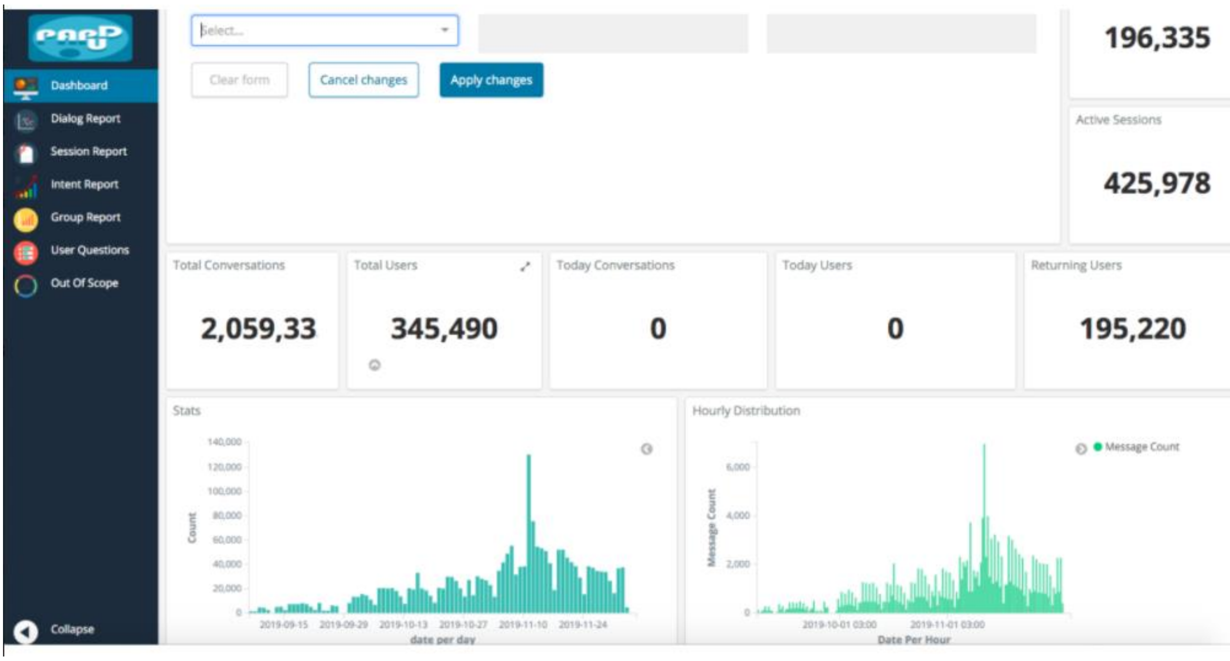
4) Entity Recognition: In order to get rid of unnecessary variations / details in the processing of dialogues, entities such as people, products, companies, brand names, prices, dates will be marked.

5) Correcting spelling mistakes: Word errors are too much. Correction needed to get rid of unnecessary variations. Use Case details

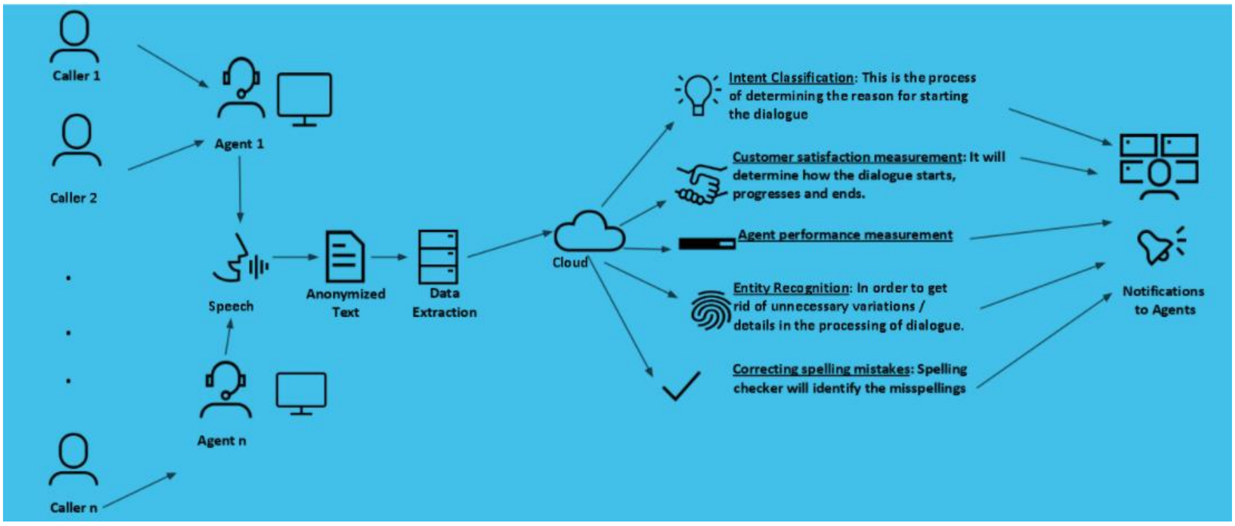
3.4 ChatMe - Use Case details

3.4.1 Use Case access

The ChatMe use-case has been implemented as a proof of concept to one of the e-commerce company to measure the effectiveness of the activities. It has been showed in the study that the performance went up to %80 success by means of understanding the written text in the call centers. Since the proof of concept was implemented in the premises of the e-commerce infrastructure it is not accessible from the outside of the company employees.



3.4.2 Use Case final architecture overview



3.4.3 Use Case modules and functionalities

Each part that is mentioned above of the project will be implemented in Python programming language using the algorithms that are coming from machine and deep learning text analytics literature. The Big Data environments that include MLLib Libraries, document-based NoSQL databases like MongoDB, Real Time Stream Analytic tools like Spark Streaming and infrastructure like Spark cluster used as a platform. In the use-case there are 5 stages that are shown below to reach the valuable results.

1) Intent classification: This is the process of determining the reason for starting the dialogue. Each dialogue can be classified as non-payment for transactions, order change requests, campaigns, product did not come out as desired, my product was broken in cargo, etc. To classify customer intents LSTMs (Long short-term memory), transformers, word / phrase / sentence embeddings are used as methods.

2) Customer satisfaction measurement: It will be determined how the dialogue starts, progresses and ends with the help of sentiment analysis on different parts of the dialogue. LSTMs, transformers, word / phrase / sentence embeddings are used as methods.

3) Agent performance measurement: Measuring criterion:

a. Agent response times,

b. Wrongly written word ratio,

c. Change of mood values at the beginning and end of dialogues (customer satisfaction)

4) Entity Recognition: In order to get rid of unnecessary variations / details in the processing of dialogues, entities such as people, products, companies, brand names, prices, dates are marked. In this way, more reliable results obtained in the classification processes (intent classification, sentiment analysis). Conditional Random Fields and LSTMs are used as a method.

5) Correcting spelling mistakes: Word errors are too much. Correction needed to get rid of unnecessary variations. To do this, a spelling checker identified the misspellings, and then estimate and correct classical methods (edit distance, statistical language models) and neural language models (eg Bidirectional Encoder Representations from Transformers-BERT).

3.4.4 Input data and expected results

Data which is related to ChatMe Turkish use case that is about the call center has been gathered from an e-commerce company in discrete dialogues format. The data type includes customer agent dialogues in the call center as anonymized names of customers and agents, dialogue texts, message sending times.

Customer satisfaction is a term used to measures how products or services supplied by a company meet or surpass a customer's expectation. Measuring the customer satisfaction is one way help managers improve their business, but it is required an extreme analysis of different factors. Results of analytics will be evaluated according to customer behaviors that include Comparison of the metrics that are the frequency of calling the call center, the duration of the customer's dialogue etc. at the end of the project.

In the text preprocessing stage of the project; call center data moves to text documents. After taking text documents, the document is scanned to fix wrong writings using Levenshtein algorithm [1]. Documents are stored in MongoDB database and ElasticSearch. 80 percent of data is used for training the models that includes deep learning algorithms. After training the model, 20 percent of data is used to test model. Created model is used to classify the text data of each call center session. Sentiment analysis is also applied to customer sentences to predict customer satisfaction.

Obtained Results

[1] Kogiso N, Watson LT, Gürdal Z, Haftka RT, Nagendra S. Design of composite laminates by a genetic algorithm with memory. Mech Compos Mater Struct 1994;1:95-117.

- Call center data has been extracted as document
- Wrong spelling mistakes have been fixed
- Models that has been created with DL Algorithms (Word2Vec, Doc2Vec, FastText), have been trained 80 percent of call center data and tested 20 percent of data.
- Sentimental analysis have been realized on the customer’s conversation data.

4 UC3: Human Resources

4.1 Use Case short descritpion

Human Resources Use Case aims to provide a tool for the Human Resources Departments that find the best matches between applicants and job offers in a fast an autonomous manner, helping to find the best talent and optimizing the HR procedures. It would also allow recruiters to check the market evolution and identify salary trends in the IT area.

The main functionalities of the HR Tool offered to the Human Resources Departments are:

- Searching of a specific profile based on key words and filtering by certain criteria as years of experience, employment history, language or salary.
- Uploading and parsing of CVs into a common agreed format to be processed by the recommender. Possibility to
- Obtaining a set of recommended candidates matching a specific job offer included by the HR department through the tool including a a percentage of matching between the CV and the JO.
- Checking the market trends related to salary tendencies in the IT sector for specific profiles which will support the Human Resources Department in the new personnel hiring.

In consequence the expected end-users of the tool will be personnel from the Human Resources Department at first attempt from IT companies willing to optimize the process of hiring new personnel. Based on the tool acceptance by the IT companies, we will evaluate if it is necessary to extend the market and process CVs from any other sectors.

4.2 Use Case details

4.2.1 Use Case access

The Human Resources components are deployed in a distributed manner between the environment hosted by Atos and the HI-Iberia internal servers. As it has been explained along the project, the Deep Learning algorithm have been packaged within a docker and deployed in the Atos server. However the rest of the modules (data bases, time series analytics, etc.) and the interfaces are deployed within the HI-Iberia servers.

The demonstrator is available at the following address: https://papud.hi-iberia.es/

4.2.2 Use Case final architecture overview

The following figure shows the architecture for the Human Resources use case. As it has been described during the project, the platform is divided into 4 blocks: one for data feeds, data acquisition, data analyticis and visualization. In this deliverable we are describing the visualization that enables the interaction with the data analytics and data acquisition layers in the backend. It is important to remark, that the architecture has been enriched with the data feed named “quarterly labour cost survey (Spanish Satatistics Institute) which will be the one used for the time series algorithm in order to obtain the market trends analysis with respect to salaries in the IT sector.

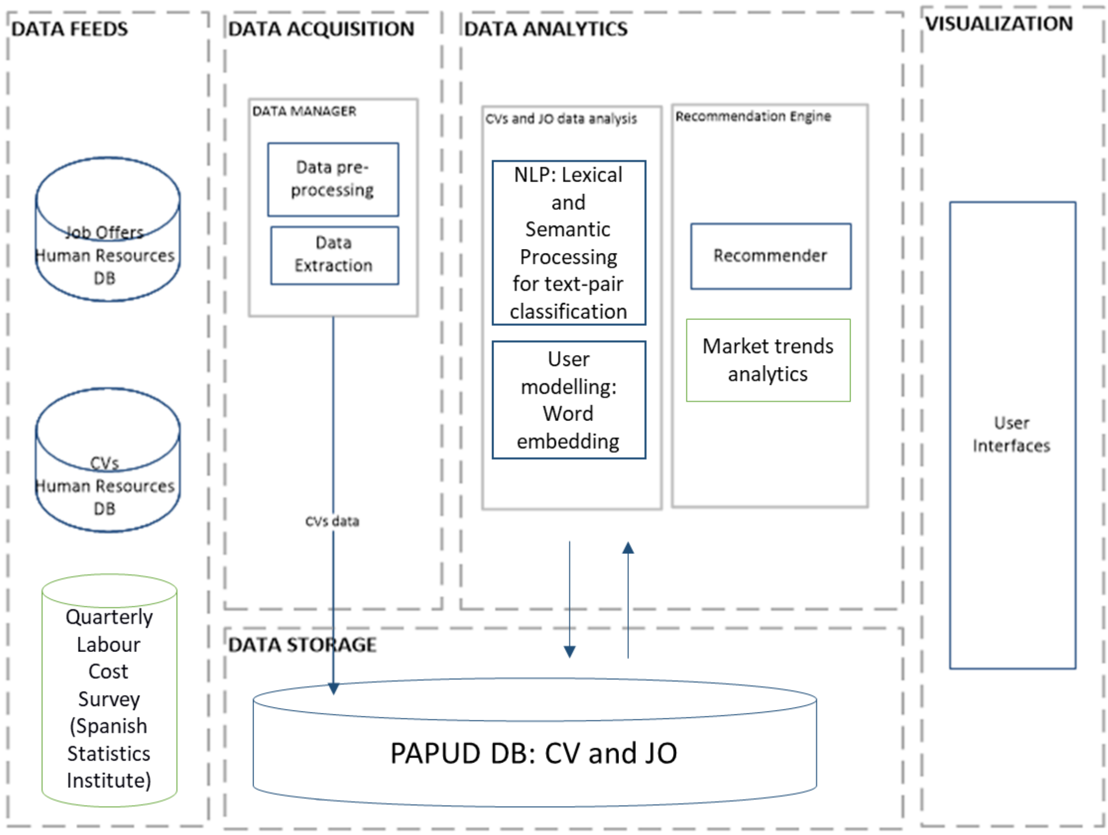


Figure 6 – UC3 Human Resources Architecture

4.2.3 Use Case modules and functionalities

The PAPUD project offers a platform aimed at the human resources department in order to allow them to manage offers and resumes and to be able to obtain the CV that best fits the published job offer. It should be noted that the design of the platform has been carried on with the collaboration of the human resources department staff of the company HI-Iberia in order to achieve a platform adapted to their real needs.

The following lines include some details to clarify the platform working and the different functionalities offered:

- Access to the platform

To access the platform, new users will need to register internally in the company for their workers in the human resources department, creating a username and password that is encrypted using the bcrypt library. Subsequently, the user authenticates on the platform by entering their credentials, and then the backend performs the authentication tasks: the web client sends these credentials to an API located on the application server that performs a read function based on data in order to determine if these credentials match those that the user entered during registration and, therefore, that the user is who they say they are.

It should be noted that in this case no credentials have been provided to access the platform for ITEA personnel since we have worked with resumes and real job offers extracted from the HI-Iberia databases and, in compliance with the Data protection law, we can not provide this information to third parties.

Thus, the following figure shows a screenshot for the log-in page to access to the platform and its functionalities:

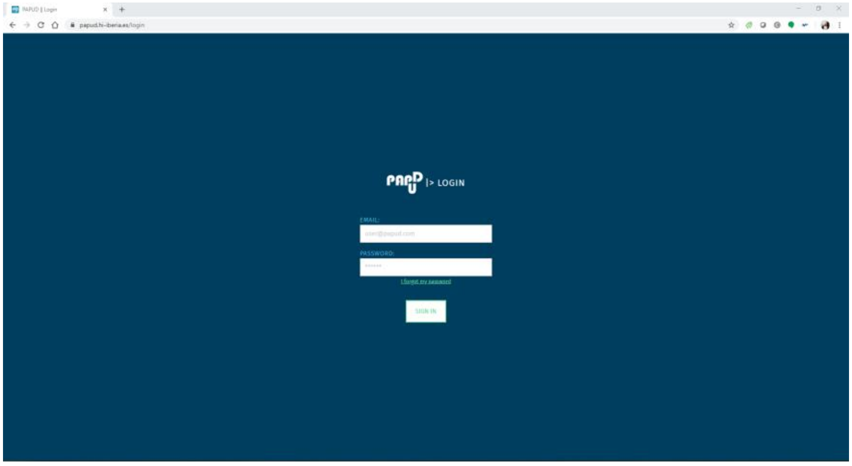


Figure 7 – Log-in page for HR Tool

- **<u>HOME Page</u>**

Once we are logged into the platform, the first page we can see is the *HOME Page.* This page is envisaged for the manual search of candidates based in some criteria selected as relevant by the HR personnel at HI-Iberia. Thus, in addition to a free text field to introduce the main criteria, the search results can be filtered by: years of experience, previous experience in other companies, language and level and salary. HR staff can also choose whether to display all database results that match their search criteria or to filter and show only job postings or resumes.

The following figure shows this HOME Page:

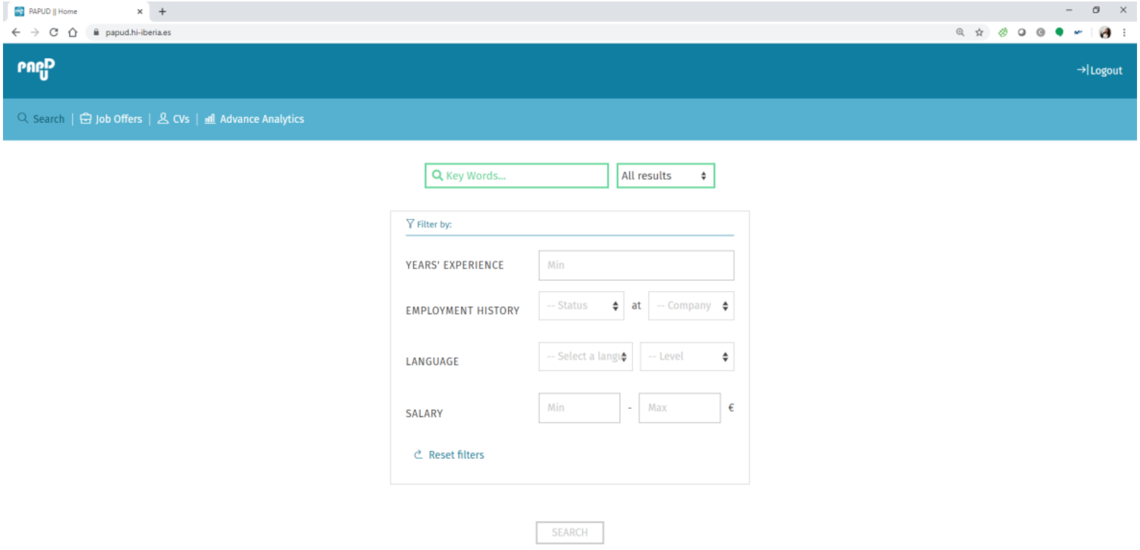


Figure 8 – HR Tool Home Page

- **Search**

Once we have introduced the search criteira (in this case, we have asked to show: CVs with Java and/or Python) the platform shows us the results in the following screen (214 results).

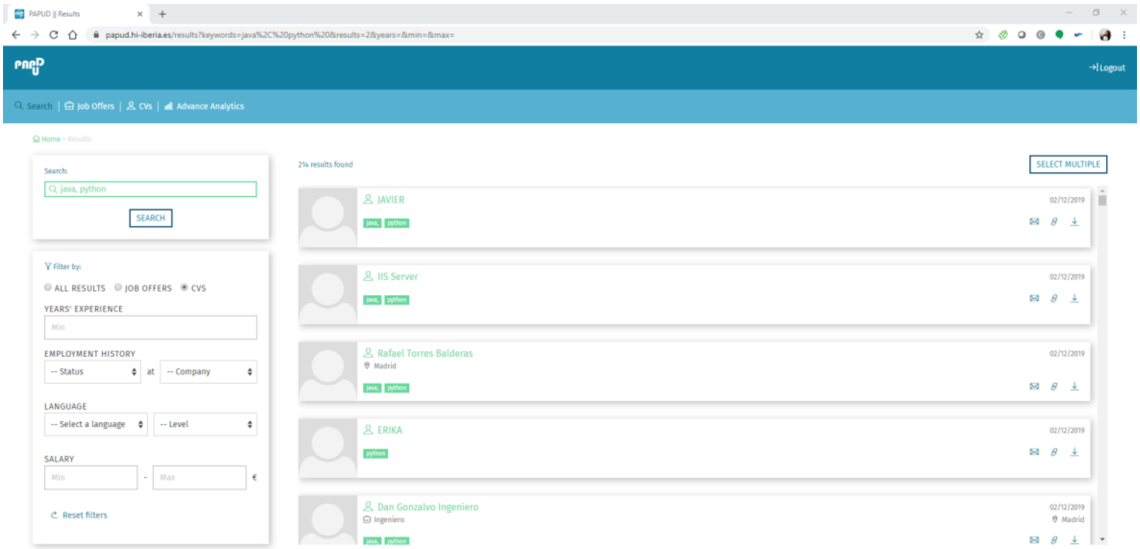


Figure 9 – HR Tool – Search Page and Results (I)

On the left side of the screen, the filtering options are showed again in order to refine the searching results. After applying language filtering, it is verified that 157 CVs appear for Java and/or Python + English skills:

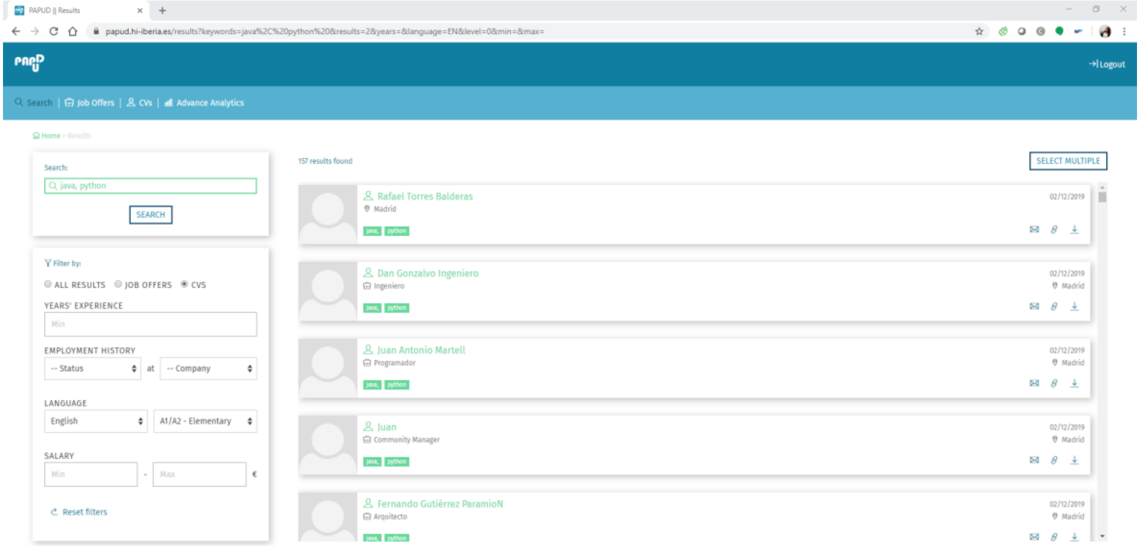


Figure 10 - HR Tool – Search Page and Results (II)

Each of the entries returned within the search shows:

- Name
- Date of CV upload to the database
- Location
- Actual postion
- Tags of the matches between the search and the CV
- Possibility of: sending an email to the candidate; linking the CV to an offer; downloading the CV in .doc format in the project template.

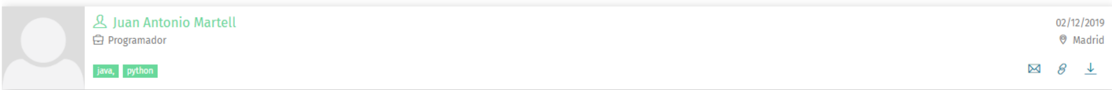


Figure 11 - HR Tool – Search Page and Results (III)

In addition, we can select a certain number of CVs (or even all) of those which appear as a result of the search, and choose an action among the following: send an email to those selected, link them to a job offer or download them in .doc format.

Regarding the technical implementation, the search uses MongoDB's own text search function and, in addition, a score has been assigned to each of the CV and JO fields in order to determine the relevance of a document for a given search query. For a text index, the weight of an indexed field denotes the importance of the field relative to the other indexed fields in terms of the text search score. For each indexed field in the document, MongoDB multiplies the number of matches by the weight and adds the results. With this sum, MongoDB calculates the score for the document. In the case of PAPUD, the following weights have been assigned for each of the fields of the CVs and the offers in order to offer the search results that best meet the criteria:

```
{
  name: "search_index",
  weights: {
    'CV_info.personaldata.name': 8,
    'CV_info.personaldata.contactMails': 7,
    'CV_info.personaldata.location': 10,
    raw_data_CV: 9
  }
```

```
{
  weights: {
    'JO_info.location': 9,
    'JO_info.work_experience.job_title': 10,
    'JO_info.work_experience.position': 10
  }
```

Figure 12 – Search Results MongoDB

- <u>**Job Offers**</u>

In this tab, new offers can be created and all the offers uploaded by human resources personnel can be consulted, deleted and edited.

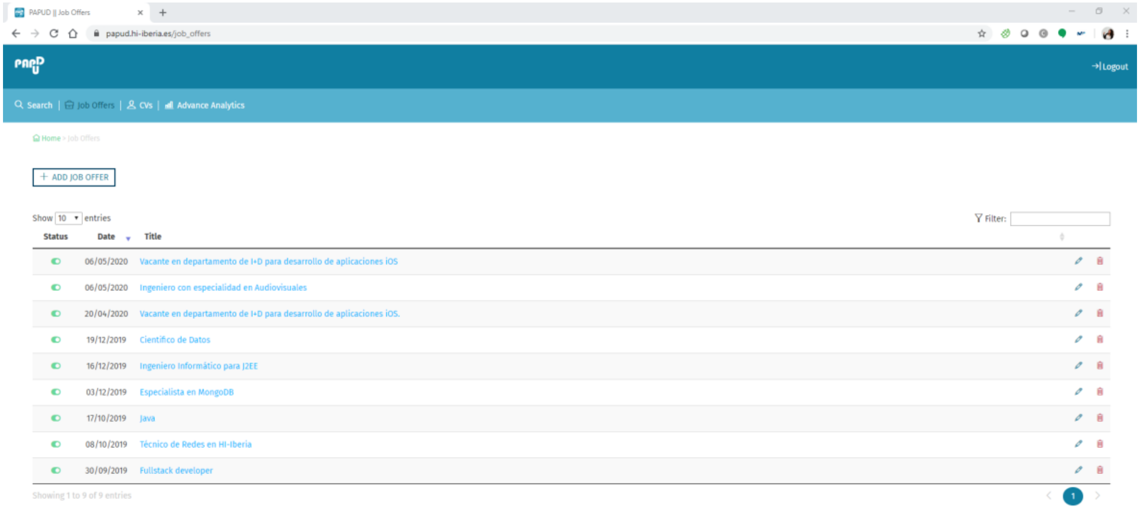


Figure 13 – HR Tool – Job Offers

To create a new offer, just click on the button and the following screen will appear to include the details. The fields included have been decided in cooperation with the human resources department based on their usual search criteria. Thus, the offer contains the following fields:

- Description of the job
    - Offer title
    - Position
    - Details
    - Years of experience required
    - Company that publishes the offer
    - Location
    - Status: active or inactive
- Summary of Technologies
    - OS
    - Programming languages
    - Database languages
    - Other tools

- Languages
  - Language and level
- Salary band
- Other information
  - Driver's license and type
  - Own car
  - Availability to travel

Thus, the view on the platform to create a new job offer is as follows:

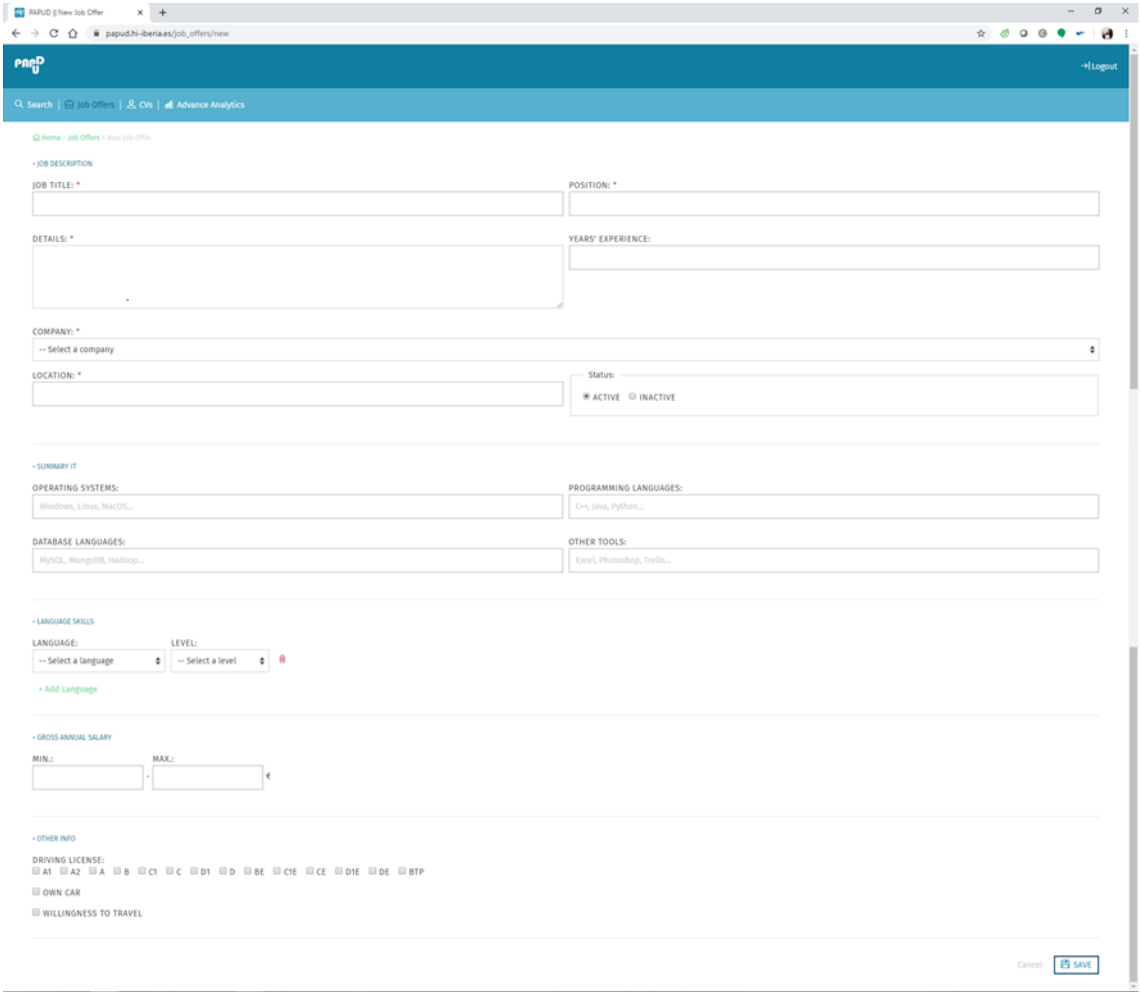


Figure 14 – HR Tool – Job Offer Creation

Once the offer is created, it will appear in the list of available offers and can be consulted at any time.

When opening an offer, the HR staff can not only consult the offer details, edit or delete it, but also:

- On the right side a list with the Candidates Recommended automatically by the recommender. In consequence, each time a new offer is included, the platform shows the candidates that best fit among those stored in the database based on the results offered by the Deep learning model. For each recommended CV, a matching percentage also appears indicating the percentage of similarity between offer-CV. From this screen, the human resources staff can access the CV and review it to link it to the offer if it fits.

Figure 15 – HR Tool – Recommended candidates for a Job Offer

- At the bottom, after the offer details, the list of candidates who have been added to the offer is provided, as well as their current status, that is, in what phase they are (pending contact, in process, interviewed, accepted or rejected). Human resources staff also have a field to add notes where you include details that have emerged in the selection process and that may be relevant in the future. In addition, the user who has added the candidate in the process appears in order to keep track of the department.

Figure 16 – HR Tool – Candidates assigned to a Job Offer

The following figure shows the offers screen:

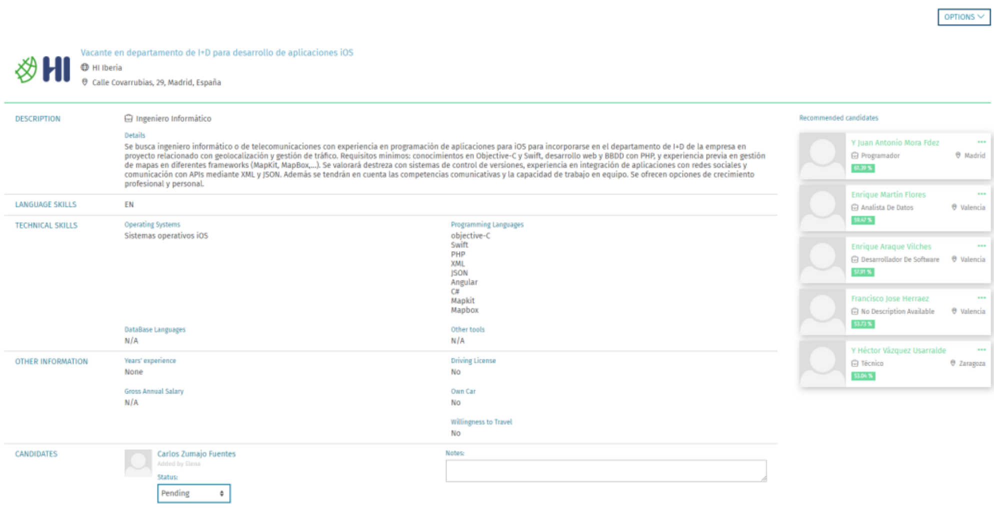


Figure 17 – HR Tool – Job Offer view

- **CVs**

In this tab you can consult all the resumes available on the PAPUD platform. These CVs have been obtained from the HI-Iberia databases, have been parsed to obtain a standardized format as detailed along the project, and have been stored. Manual searches on the platform are performed on this set of resumes and are also used to provide recommendations for the DL model each time a new offer is included.

In the following figure the view of this screen is shown where the resumes appear in alphabetical order, the upload date, the option to edit your personal details (name, location, telephone number and LinkedIn profile) and to delete it from the database.

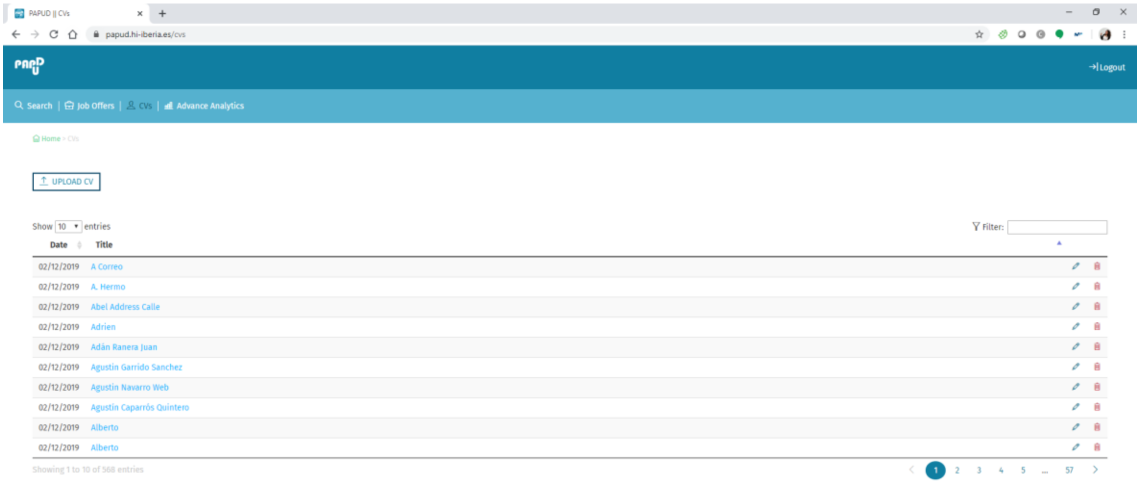


Figure 18 – HR Tool – CVs

In addition, we can upload a new resume if we click on UPLOAD CV. A CV may be uploaded in .doc or .pdf format, which will be parsed and translated into the format defined for the platform and stored in a database.

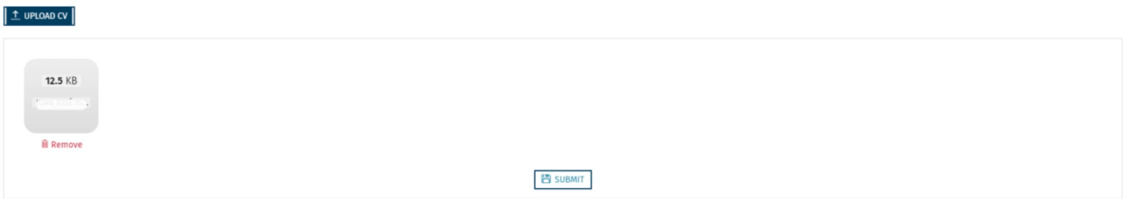


Figure 19 – HR Tool – Uploading a new CV

Once the CV is parsed, it is displayed in the defined format grouping the fields into: professional experience, education, languages, technical knowledge and others. Thus, when we consult a CV, the candidate's information is displayed and, in addition, if it has been associated with an offer, it appears on the right side. The following figure shows the curriculum view:

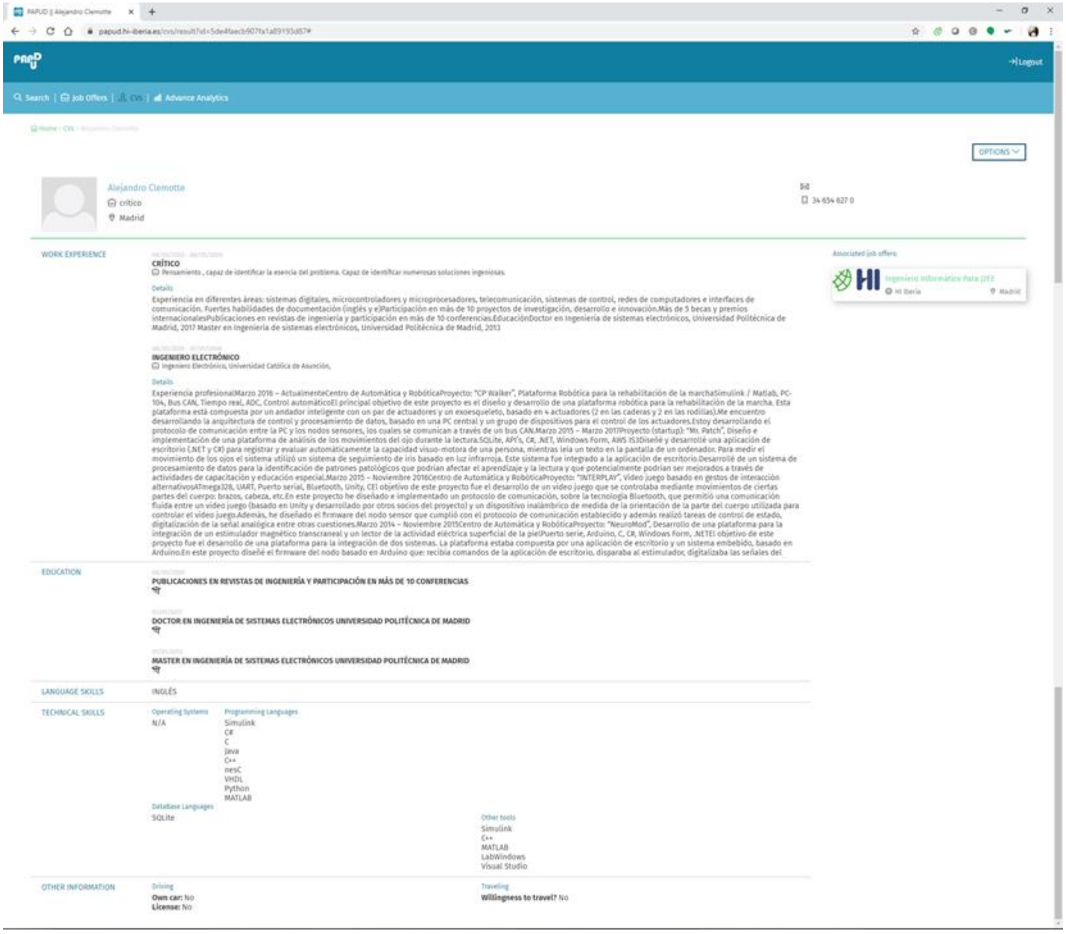


Figure 20 – HR Tool - CV overview

In this screen if we click on the button, you can:

- Send an email
- Associate the CV with an offer
- Download the CV in the standardized format defined internally in HI-Iberia, following the needs of the human resources department.

- **Advance Analytics**

This tab is envisaged for showing the Human Resources staff the salaries tendencies in the following IT sectors based on the current Spanish Statistical Instute collected data:

- Computer Programming, Consultancy and Related Activities
- Professional, Scientific and Technical Activities

The objective of the tab is to support Human Resources department when offering a salary to new hirings following the market trends. The Human Resources staff can visualize the salary evolution through the years quarterly, and also for part and full time job and the information will be shown in the graphic. The following figure shows this tab:

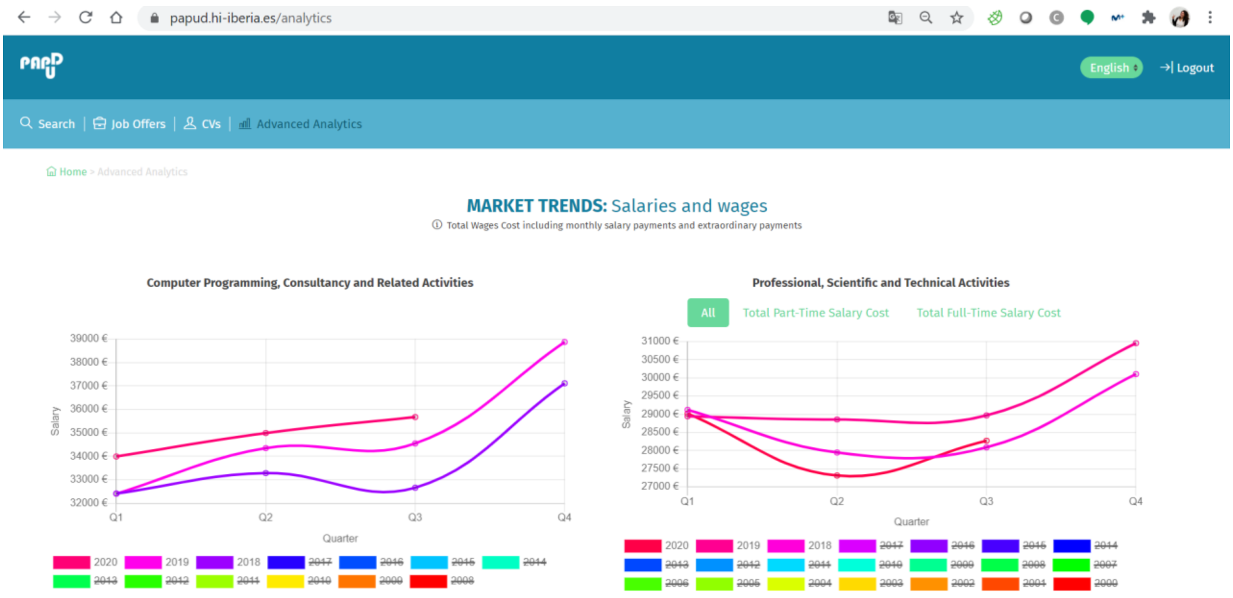


Figure 21 – HR Tool – Advance Analytics screen

4.2.4 Input data and expected results

The inputs for the Human Resources tool are mainly CVs and Job Offers. Particularly, the platform expect a document in any format (.doc, .pdf) containing the CV of the applicant to be uploaded within the CVs secton. For the job offer, the platform has an embedded template to be completed by the HR staff with all the details. Job offers are supposed to be submitted through the relative tab in the HR platform. As result, the expected outputs are a set of recommended CVs that best fists the updated Job Offer.

In addition to this, the HR staff can search manually for specific profiles by providing key words and specific citeria in the HOME Page.

5 UC4: BAREM

5.1 Use Case short descritpion

The use case BAREM stands for "Behaviour Analysis for Reverse Efficient Modeling", the objective is to gather users activity on e-Services thanks to log files and analyze this activity to find mis-using and bad design of the e-service. The final objective is to show this analysis and allow the designer to correct this bugs to achieve an enhancement of the User eXperience.

In this use case, we have selected an existing eService edited and commercialized by SOFTEAM which is the School Transportation Service allowing parent to subscribe for their child to the bus transportation service.

5.2 Use Case details

5.2.1 Use Case access

We achieved to set-up the use case on the Papud environment hosted by Atos. The demonstrator is available at the address : https://papud.e-citiz.com

This demonstrator is initialized for the French region of Rennes, so testers have to act has if they were living in this Region.

5.2.2 Use Case final architecture overview

The technical architecture for the use case is defined in the following diagram:

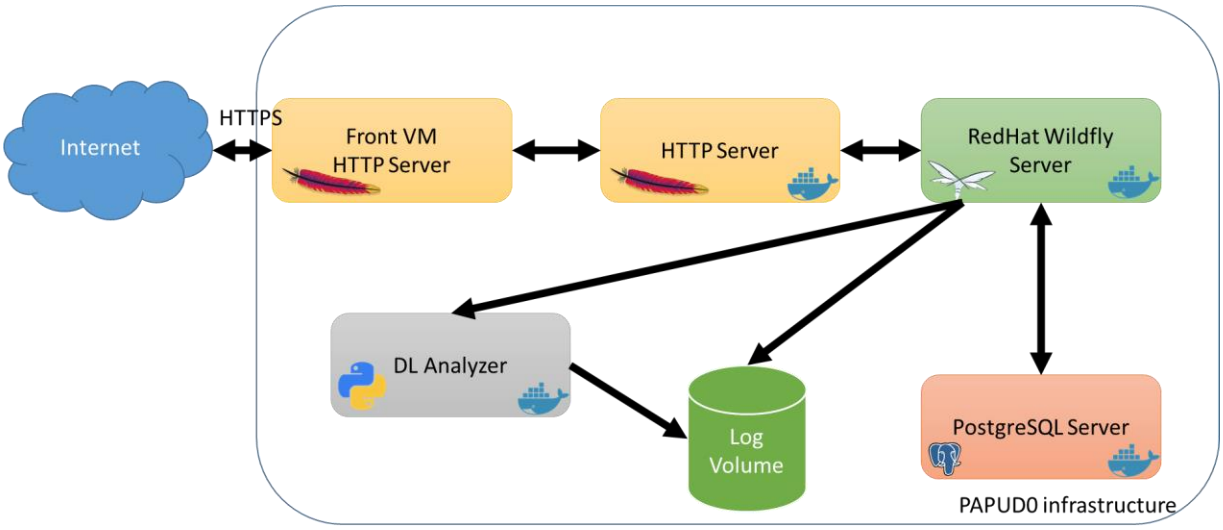


Figure 22 - UC4 architecture overview

This Architecture based on Docker containers is quite usual with a front web server (Apache), an enterprise server (Wildfly) and a Database (PostgreSQL). Due to docker usage, logs files are mapped outside Docker containers to allow their analysis by other containers.

The most important module here is the "DL Analyzer" which contains all the "magical" analysis of our Use case. This "DL Analyzer" can be explained by the following diagram :

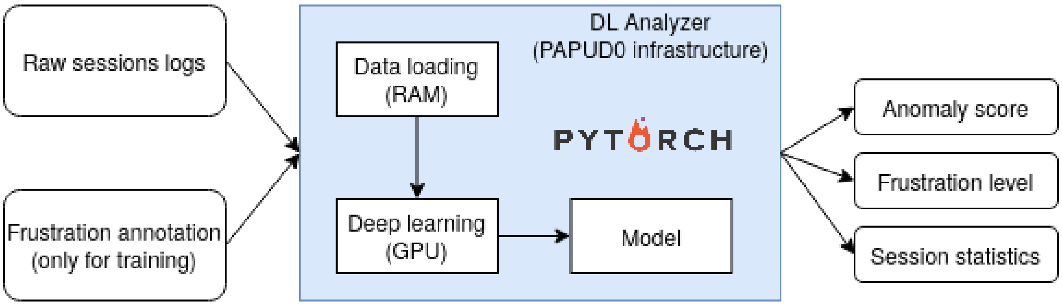


Figure 23 : UC4 DL Analyzer overview

It uses sessions logs provided by the server and the frustration level provided by the previous frustration detection model for training the model. Then, for the using part, it analyze only raw sessions logs in order to output three information : anomaly score in the session, frustration level and several statistics.

5.2.3 Use Case modules and functionalities

The School transportation service is logging navigation records in log files when an end user is using the e-Service. Based on this logs files, several modules are involved:

- The DL Analyzer is in charge of:
  - Extract each user navigation scenario from the whole log file to isolate his specific navigation
  - Based on trained DL models, predict the frustration level of the navigation
  - Based on trained DL models, compute navigation error score, business errors and syntactic errors and compute statistics for all input navigations
  - Generate a report by a json file containing all output statistics with categorization
- The BAREM module (Figure 24) is in charge of displaying a dashboard to the designer (Figure 25) to allow him to analyze the most encountered errors (navigation, business, syntactic) and the frustration level. Thanks to these analyzes the designer will be able to dig into the log file to gather more details if needed and, by the end, correct the running e-Service to enhance the User eXperience.

Figure 24 : BAREM module access for the e-Service Designer

Figure 25 : Part of the dashboard shown to the e-Service Designer

5.2.4 Input data and expected results

The figure below illustrates the different modalities used to train the offline frustration detection tool which is used to annotate logs. There are images with face and landmarks detection, audio, and bio signals including blood volume pulse, electrodermal activity, and electroencephalography. Each of the 19 subjects of the dataset also filled a form with a self-evaluation of their state of frustration during the data capture. The depth and screen images were only used for data annotation and are not used to train deep learning models. Data have been annotated by members of the University of Lille with 2 classes, frustrated and non-frustrated, which corresponds to the output of the offline tool.

To illustrate the kind of input data for the log analysis part, here is an extract of logs files containing end user navigation records:

```
2019-07-29 16:29:34,029 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [***      no.workflow.id     ***] [0x28083E9C] [SESS_START__] [fr_] [WINDOWS_7-FIREFOX] [172.25.0.1] [Citoyen]
2019-07-29 16:29:34,030 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [***      no.workflow.id     ***] [0x28083E9C] [GET_RQ_START] [wildfly-transport-cont]
2019-07-29 16:29:34,030 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_WRKFL_] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl!authentificationcitoyensta
2019-07-29 16:29:34,030 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [AuthentificationCitoyenStartAct] [CitoyenBRWorkflowInstance!authentificationAuthentificatio
2019-07-29 16:29:34,030 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [AuthentificationCitoyenStartAct] [CitoyenBRWorkflowInstance!authentificationAuthentificatio
2019-07-29 16:29:34,030 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [AuthentificationCitoyenReferential] [CitoyenBRWorkflowInstance!authentificationAuthentifica
2019-07-29 16:29:34,034 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [AuthentificationCitoyenReferential] [CitoyenBRWorkflowInstance!authentificationAuthentifica
2019-07-29 16:29:34,034 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [MAD Collectivité et configuration] [CitoyenBRWorkflowInstance!authentificationAuthentificat
2019-07-29 16:29:34,036 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [MAD Collectivité et configuration] [CitoyenBRWorkflowInstance!authentificationAuthentificat
2019-07-29 16:29:34,036 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Initialiser les paramètres url dans la stack] [CitoyenBRWorkflowInstance!authentificationAu
2019-07-29 16:29:34,037 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Initialiser les paramètres url dans la stack] [CitoyenBRWorkflowInstance!authentificationAu
2019-07-29 16:29:34,037 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Initialiser les CSS] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl!init
2019-07-29 16:29:34,037 ERROR [stderr] (default task-77) [SEVERE]  [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [UNDEF_______] [Adding CSS:rpdl/RPDLCitoyenQualif.css]
2019-07-29 16:29:34,038 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Initialiser les CSS] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl!init
2019-07-29 16:29:34,038 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [BP - WS - Set Configuration Authentification] [CitoyenBRWorkflowInstance!authentificationAu
2019-07-29 16:29:34,080 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [BP - WS - Set Configuration Authentification] [CitoyenBRWorkflowInstance!authentificationAu
2019-07-29 16:29:34,080 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [C - Activation de compte] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [C - Activation de compte] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Authentification : Accès direct via url] [CitoyenBRWorkflowInstance!authentificationAuthent
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Authentification : Accès direct via url] [CitoyenBRWorkflowInstance!authentificationAuthent
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [MDP - Détection de l'url] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [MDP - Détection de l'url] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl
2019-07-29 16:29:34,081 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Authentification - Configuration dans session attribute] [CitoyenBRWorkflowInstance!authent
2019-07-29 16:29:34,082 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Authentification - Configuration dans session attribute] [CitoyenBRWorkflowInstance!authent
2019-07-29 16:29:34,082 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Aiguillage - Mode connexion séquentiel] [CitoyenBRWorkflowInstance!authentificationAuthenti
2019-07-29 16:29:34,083 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Aiguillage - Mode connexion séquentiel] [CitoyenBRWorkflowInstance!authentificationAuthenti
2019-07-29 16:29:34,083 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Aiguillage - Mode d'authentification] [CitoyenBRWorkflowInstance!authentificationAuthentifi
2019-07-29 16:29:34,083 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [Aiguillage - Mode d'authentification] [CitoyenBRWorkflowInstance!authentificationAuthentifi
2019-07-29 16:29:34,083 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [BP - Authentification anonyme ?] [CitoyenBRWorkflowInstance!authentificationAuthentificatio
2019-07-29 16:29:34,084 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [END_ACT_____] [BP - Authentification anonyme ?] [CitoyenBRWorkflowInstance!authentificationAuthentificatio
2019-07-29 16:29:34,084 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [START_ACT___] [Identification] [CitoyenBRWorkflowInstance!authentificationAuthentification2_kwwl!identifica
2019-07-29 16:29:34,084 ERROR [stderr] (default task-77) [WARNING] [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [UNDEF_______] [ not found for locale <fr>]
2019-07-29 16:29:34,089 ERROR [stderr] (default task-77) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [8368e7340cf0d6e667764626727f7c1f] [0x28083E9C] [REQ_END_____]
2019-07-29 16:59:34,873 ERROR [stderr] (default task-95) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [***      no.workflow.id     ***] [0x0C1A0442] [SESS_TIMEOUT] [2019/07/29 16:59:34:872]
2019-07-29 16:59:34,874 ERROR [stderr] (default task-95) [INFO]    [CC74.Guichet] [0oFQsAa2FMDks8Uz6pGRWGpy] [***      no.workflow.id     ***] [0x0C1A0442] [SESS_END____]
```

After Deep Learning analysis, here is an extract of the kind of json output:

```
{
    "SessionsAnalyzed": 144,
    "AnomalyDetection": {
        "SessionsWithBusinessError": {
            "count": 96,
            "errorList": [{
                    "error": "[Un compte avec cet identifiant existe d\u00e9j\u00e0.]",
                    "count": 33,
                    "sessionIDs": ["IfwTmPuFLZM9VxPOqb_uPX88.log", "xPN3CXQnlgCK8nDmB8fKzIjl.log", "DiNJB9RNXjdSX0Oj-Cqft8El.log", "AfwDerutsRutlYbZnTL_3N8w.log", "yb-dItP-yPlZMqQewQCFCNAB.log", "oTaYBcfXatOhmomdBMYxNmtL.log",
                }, {
                    "error": "[Vous devez indiquer au moins un num\u00e9ro de t\u00e9l\u00e9phone fixe ou mobile.]",
                    "count": 7,
                    "sessionIDs": ["reIeBTP0y67-gVej48kTeZ2b.log", "i2Mi-xqjRg4NociBwFPlW1OZ.log", "Ltn0CqQveOQVFt4AddPyNoiC.log", "k0nVqr9xbRbbIzgB9d7vcchH.log", "8RkBzjCW9q_cpUjQuOKkg5bU.log", "z9pYwpt_WwqbfVh5jn32r6c_.log",
                }, {
                    "error": "[Type de document non autoris\u00e9]",
                    "count": 17,
                    "sessionIDs": ["reIeBTP0y67-gVej48kTeZ2b.log", "IZ9C4_yUWKRBDvMu75K5LRhP.log", "Ltn0CqQveOQVFt4AddPyNoiC.log", "k0nVqr9xbRbbIzgB9d7vcchH.log", "r3-q73nG_plCepKvSgy9b68I.log", "s770wc2qPSQkmrykjongItg2.log",
                }, {
                    "error": "[Veuillez cocher la case Certification sur l'honneur avant de passer \u00e0 l'\u00e9tape suivante]",
                    "count": 5,
                    "sessionIDs": ["IZ9C4_yUWKRBDvMu75K5LRhP.log", "Ltn0CqQveOQVFt4AddPyNoiC.log", "r3-q73nG_plCepKvSgy9b68I.log", "pX8FAE0NzqiekrHThThajFNR.log", "fKKiIu2INez4nIlItJ4XaPP2.log"]
                }, {
                    "error": "[Veuillez accepter le r\u00e8glement des transports avant de passer \u00e0 l'\u00e9tape suivante]",
                    "count": 5,
                    "sessionIDs": ["IZ9C4_yUWKRBDvMu75K5LRhP.log", "Ltn0CqQveOQVFt4AddPyNoiC.log", "r3-q73nG_plCepKvSgy9b68I.log", "pX8FAE0NzqiekrHThThajFNR.log", "fKKiIu2INez4nIlItJ4XaPP2.log"]
                }, {
                    "error": "[Le mot de passe fait partie des 5 derniers mots de passe]",
                    "count": 1,
                    "sessionIDs": ["YCR9kqlYoTPnc5Zq4MphlVcT.log"]
                }, {
                    "error": "[Nom d'utilisateur ou mot de passe inconnu]",
                    "count": 12,
                    "sessionIDs": ["DiNJB9RNXjdSX0Oj-Cqft8El.log", "V4YLDN0WWPfnTzrK2fCGm0Nr.log", "026jb5i6X5YrvswpaohkKNoL.log", "dAMOG48kXM3L3fEGM3pVds0E.log", "1H5HR_cINqJBGoZDGYph98p0.log", "AZA9RlA7zge4nppzVZuX-Z7A.log",
                }, {
```


All of these technical analysis files are gathered by the BAREM module and displayed inside a web dashboard (see Figure 25allowing the Designer to detect bugs in the design.

6 UC5: HPC

6.1 Use Case short description

This Use Case aims at answer to this question: Can we be able to predict or to understand the failures occurring in a large scale system?

In other words, taking the complexing of such system into account, can we provide the user with information about the failure to speed up its resolution?

In more details, large scale system produces lot of monitoring data and analyzing this data is a tedious task dedicated to the administrators of the system. Our goal is to help the administrators to read this data by detecting automatically the failures and by focusing on the root cause of the failure. The size of this data is this first major issue to address. The second issue is its complexity. Due to the number of different components in HPC, we have also a large variety of type of data (text, numeric, events, etc.).

Our use case is focused on the analysis of the logs (I.e. textual data). We provide the user an overview of the logs related to a specific detected failure among all the logs. This overview can help him to focus on the real cause of the failure and help him to solve it.

6.2 Use Case details

6.2.1 Use Case access

The demonstrator is split into two main parts: the learning part and the web interface part to view the result. The learning part is done offline using a classical command line interface (CLI). The web interface is used by the user to select its logs and to show him the relations between the logs and the detected failure.

The learning part is done using a node with a GPU to accelerate the learning process. The web interface is portable and can be used on any computer. For the demo, the web interface is hosted on the Papud platform. It is required to use a SSH tunnel to access to it to insure security access.

6.2.2 Use Case final architecture overview

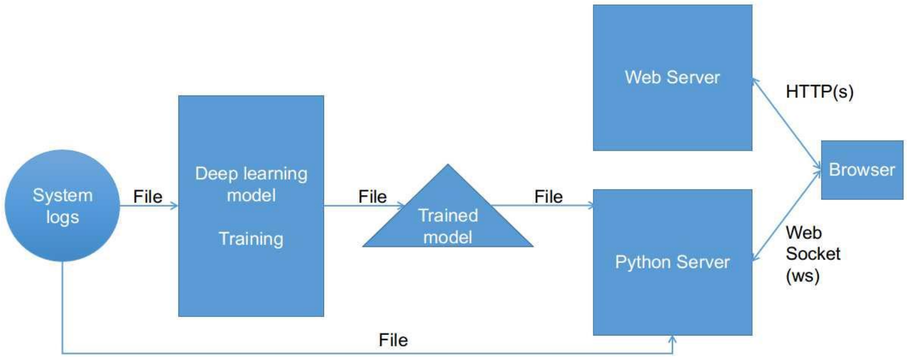


Figure 26 - UC5 final architecture

We follow a classical architecture based on the two main parts. Our first main step is to train our deep learning model to provide a trained model to our python server. Then, our python server is used as a backend by the web server to run the inference using the trained model. User access to the web interface using classical HTTP protocol. In addition, a connection is automatically done between the user's browser and the Python server suing web socket.

6.2.3 Use Case modules and functionalities

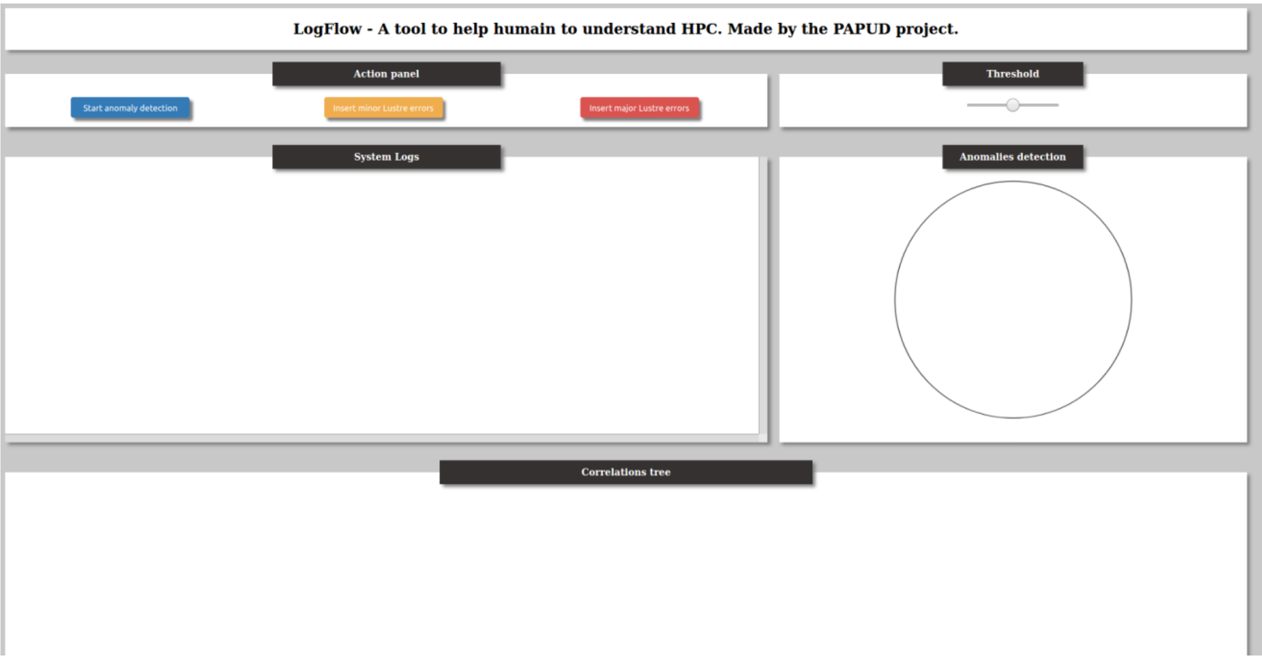


Figure 27 - Logflow demonstrator view

From the user's point of view, it can interact only with the web interface. It is split into 3 main parts: "System logs", "Anomalies Detection", "Correlation Tree". The "System Logs" part shows the logs coming from the system. The "Anomalies Detection" part shows the score of these logs using a circle: inside this circle the logs are normal, outside this circle the log is an anomaly. The last part "Correlation Tree" shows the correlation between the log reported as an anomaly and the previous logs.

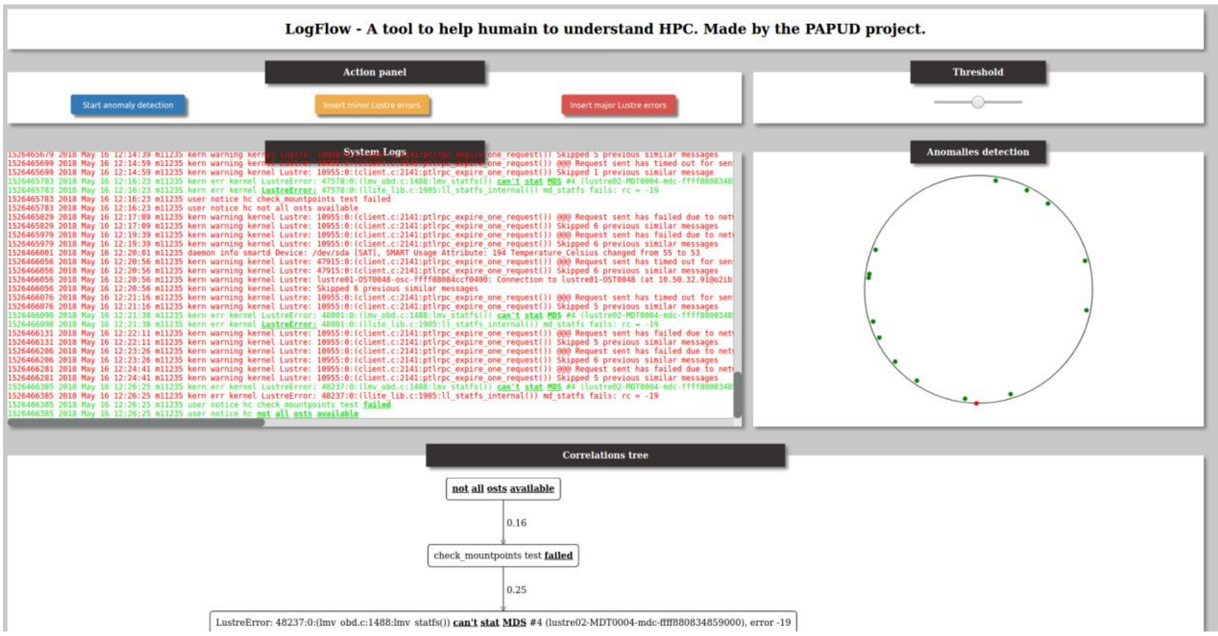


Figure 28 : Logflow demonstrator runnning

In this screenshot, the last log is reported as an anomaly. The correlations detected with previous logs are shown in green (the green logs have a correlation with the last one). Additionally, we can the red point outside the circle corresponding to the log reported as anomaly.

At the bottom, the correlation tree is shown. The complete tree is shown in the following screenshot.

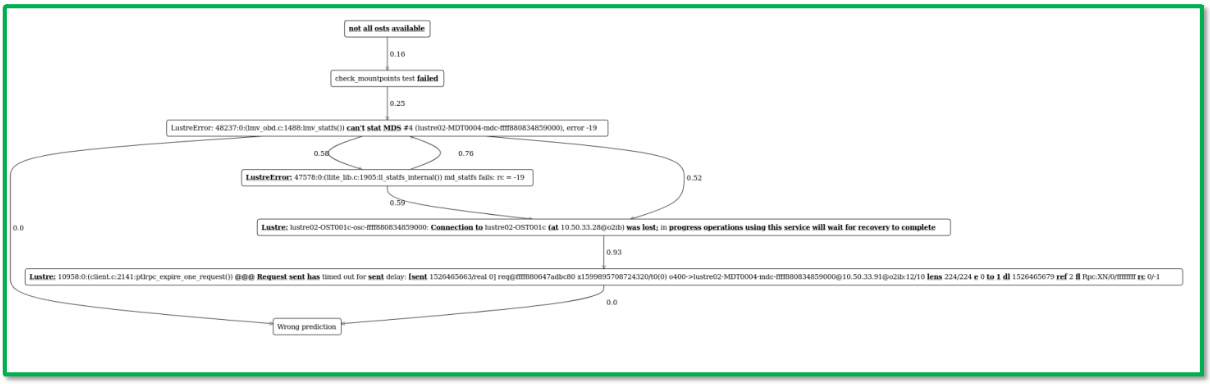


Figure 29 - Logflow correlation tree

Note that the action panel is used to generated a failure for the purpose of the demo. Considering a classical usage, this panel should not be used.

6.2.4 Input data and expected results

The input data of the use case is the system logs provided by the system.

The expected results are two folds:

- The anomaly automatically detected based on the analysis of the log flow.
- The correlations between this anomaly and the previous logs as a tree of correlations.