D8.7

Big Data Analytics Tools and Platforms - First Iteration (M12)

Document Owner:	NTUA
Contributors:	TXT, IBM
Dissemination:	Public
Contributing to:	WP8 - Manufacturing Intelligence Methods and Tools for Prod- Serv design
Date:	04/03/2016
Revision:	1.00

VERSION HISTORY

NBR	DATE	NOTES AND COMMENTS
0.10	19/12/2016	Initial table of contents
0.20	23/12/2016	Initial inputs to section 2.1, 2.2.1
0.30	09/02/2016	Demonstration of Social-Big Data Analytics Platform
0.40	10/02/2016	Initial inputs for section 3
0.50	15/02/2016	Updated section 2, incorporation of section 1 and 4
0.60	19/02/2016	Release of PSY- InfluenCial platform in GitHub
0.70	22/02/2016	Updated inputs for section 3
0.80	24/02/2016	Final draft circulated for internal review
0.90	02/03/2016	Revised draft including updates in Section 3
1.00	04/03/2016	Final draft for submission to the EC

DELIVERABLE PEER REVIEW SUMMARY

ID	Comments	Addressed (X) Answered (A)

Table of Contents

Executive Summary

In PSYMBIOSYS, the role of manufacturing intelligence in product-service design is instrumental. To this end, the purpose of the first iteration of the deliverable at hand (D8.7 "Big Data Analytics Tools and Platforms - First Iteration") has been to define and document the PSYMBIOSYS big data analytics approach, planning and progress of implementation.

The big data analytics approach adopted in PSYMBIOSYS is two-fold depending on the type of data (e.g. origin, meaning, structure) as it aims at retrieving and analyzing:

1. "Social" data (i.e. indicating interactions between stakeholders) in order to identify influencers and point the design team towards the most popular/trending ideas and proposals.
2. Production data (e.g. the errors, cycle times, downtime intervals and percentage of scrap of machineries) in order to identify the need for improvements or re-engineering of the product if the design can be improved.

With regard to social data, the first version of the Social-Big Data Analytics Platform (PSY-InfluenCial) has been released in February 2016. Building on state-of-the art big data technologies (like Hadoop and Spark) and leveraging sentiment analysis engines (i.e. the FITMAN Unstructured and Social Data Analytics Specific Enabler), the PSY- InfluenCial platform calculates the most influencing persons, the topics they mostly promote and their correlations. Next steps along the PSY- InfluenCial platform indicatively include: completion of experimentation with the use cases (that is already ongoing), design of more advanced Big Data algorithms (complementary to the PageRank algorithm), the deployment of admin panels and the incorporation of additional social big data sources.

As far as the Manufacturing-Big Data Analytics Platform is concerned, its full release is expected with the final version of this deliverable (D8.8 "Big Data Analytics Tools and Platforms - Final Iteration").

1 Introduction

PSYMBIOSYS aims at improving the competitiveness of European Manufacturing industries by developing an innovative product-service engineering environment, symbolized by a five-pointed symbiosis star – design-production, product-service, knowledge-sentiment, EDA-SOA, business-innovation – and able to dramatically reduce the time-to-market of more attractive and sustainable product-service solutions.

The scope of WP8.4 "Big Data Analytics Methods and Tools for Manufacturing Intelligence" is:

- To study and analyse tools and platforms for the analysis of big data coming from the Product-Service value chain.
- To explore the big data technologies which are relevant for manufacturing and may contribute to product-service design.
- To identify big data sources for extracting knowledge related to the P-S value chain.
- To develop and deploy big data analytics infrastructures which are in line with the PSYMBIOSYS use case needs.

The present deliverable D8.7 "Big Data Analytics Tools and Platforms – First Iteration" presents and documents the big data analytics tools developed in the first iteration of the PSYMBIOSYS project. The PSYMBIOSYS big data analytics tools will be supported and updated to incorporate experiences and feedback gathered from the initial, exploratory trials' application, in their final release which is expected to be delivered in the third year of the project implementation.

1.1 Introduction to the software release

Today, big data is listed among the core technological megatrends which are expected to disrupt practically any industry. Big Data are instrumental in order to make sense of the data deluge, the tons of data available in the manufacturers' systems and in the web, in general.

Typically, Big Data are data sets that are too large or complex for traditional data processing applications and are characterized by a set of "V" features; whenever Volume, Velocity, Variety or Veracity of data is present, Volatility, Complexity and Variability may be also present as depicted in the following figure, leading to the need for applying Big Data technologies and practices.

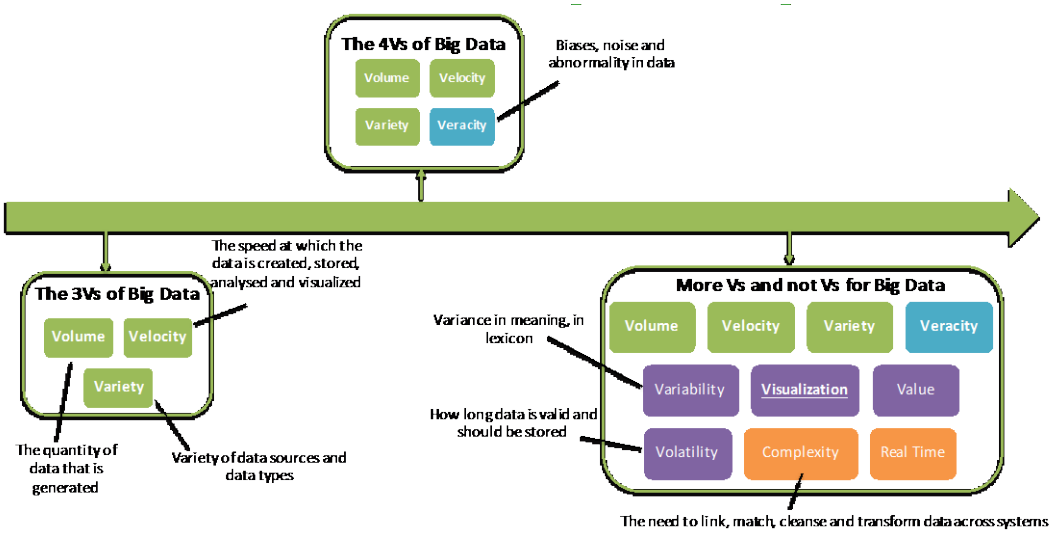

Figure 1-1: Big Data Concept

In manufacturing, big data analytics can be leveraged in numerous ways, e.g.

- *Plant operations and production*: Product quality tracking; Remote diagnostics/condition-based monitoring of products/field assets; Analysis of geo-distributed sensor data; Supply planning; Output forecasting; Simulation & testing of new manufacturing processes; Mass-customization of P-S.
- *Logistics-Shipments*: Product shipments monitoring; Inventory shrinkage locations detection; Appropriate inventory levels prediction; Supply chain bottlenecks identification.
- *Customer service and warranties*: Trends identification in customer inquiries; Monitoring of P-S usage to detect manufacturing/ design problems; Real-time sensor-based monitoring, diagnostics, and maintenance.
- *Sales and Marketing*: Profitable customers' identification; Optimal sales approaches/offers decision support; Campaign planning and optimal spending.

Today, the ecosystem of Big Data tools has grown at a very vast rate and is thriving, with the MapReduce framework, along with the Apache Hadoop, Spark and Flink platforms, being considered among the core big data technologies.

In PSYMBIOSYS, a dual approach in big data analytics has been followed in order to allow for effective processing of big data depending on their source and translate them to actionable information for the product-service design team. In particular, the purpose of the PSYMBIOSYS big data analytics is two-fold depending on the type of data (e.g. origin, meaning, structure):

1. To retrieve and analyze "social" data (i.e. indicating interactions between stakeholders) in order to identify influencers and point the design team towards the most popular/trending proposals.

2. To retrieve and analyze production data in alignment with the industry 4.0 paradigm, in order to identify the need for improvements or re-engineering of the product if the

design is not compatible or can be improved (upon mining the errors, cycle times, downtime intervals and percentage of scrap of machineries).

1.2 Positioning of the deliverable in PSYMBIOSYS

In alignment with the global architecture reflected in deliverable D9.1 "Functional and Modular Architecture", the big data analytics outcomes are positioned in the overall PSYMBIOSYS picture in the following way:

- The Big Data analysis concerning social data interacts with the Sentiment part of the Knowledge-Sentiment tussle in WP5 "Symbiotic knowledge-sentiment cooperation" in terms of retrieving social data (from the "Crowd Innovation Platform") and providing insights who are the (online) influencers on a specific domain. In this way, the more influential and reliable perspectives for a product-service will be highlighted (in WP5) in order to guide the product-service design team accordingly, to real customer trends and needs.
- The Big Data analysis concerning production data internal to the company (e.g. coming from machineries) interacts with WP6 "SOA-EDA Secure IT Platform for Manufacturing Intelligence" in order to discover difficulties in the production to be fed back to the product-service design team in order to design a product easier to be manufactured in WP3 "Symbiotic engineering- manufacturing collaboration" and WP4 "Symbiotic product-service lifecycle concurrency".

1.3 Structure of the document

The present deliverable is structured as follows:

- Section 2 documents the functionalities, the architecture, the technology stack and the manuals of the Social-Big Data Analytics Platform (PSY- InfluenCial).
- Section 3 documents the functionalities, the architecture, the technology stack and the manuals of the Manufacturing-Big Data Analytics Platform.
- In Section 4, the conclusions deriving from the work performed and documented in the deliverable at hand, as well as directions and recommendations for the next steps are reported.

2 Social-Big Data Analytics Platform (PSY- InfluenCial)

In this section, the description of the released PSY- InfluenCial platform is provided. The section starts summarising the overall information about the software released (description, overall data, functionalities and architecture), after that technical information are reported about architectural stack, technical manual for installation and licensing (including third party components). Finally, the user manual and conclusions and future steps conclude this section.

2.1 Software Description

2.1.1 Overall Data

Item	Value
Component Name	Social-Big Data Analytics Platform (PSY- InfluenCial)
Software version	v1.0
Reference workpackage	WP8.4
Responsible Partner	NTUA
Contact person	Fenareti Lampathaki, flamp@epu.ntua.gr
Source control	https://github.com/epu-ntua/PSYMBIOSYS-Influencial
Short Description	The PSYMBIOSYS InfluenCial Platform aims at leveraging big social data to identify influencers per manufacturing domain and reveal (potentially) "hidden" topics promoted by such influencers, indirectly contributing to the design phase of symbiotic product/services.

2.1.2 Purpose of the tool

The PSYMBIOSYS InfluenCial Platform builds on the premise that ideas expressed by influential people are more probable to turn out into new trends that affect the product-service design phase. PSY-InfluenCial is a big data infrastructure to analyze "social" data (i.e. indicating interactions between stakeholders) in order to identify influencers and point the product-service design team towards the most popular/trending ideas.

It needs to be noted that the definition of an influencer may vary depending on the context to include:

- People who are often mentioned in other people's discussions, e.g. in social media.
- People whose expressed opinions are commonly referenced by others, e.g. in social media (retweets, fb share…), in ideation platforms, in blogs.
- People whose expressed opinions are widely adopted and liked, e.g. through upvotes or likes.
- People who are popular in a network, e.g. number of followers or number of blog views.
- Friends of influent people for a specific topic should also be influent and typically talk about the same or similar topics, e.g. in Twitter, Instagram, Blogs.

Acknowledging different definitions of social influencers, the PSYMBIOSYS InfluenCial Platform shall implement different algorithms for Link Analysis, Association Rule Learning,

Correlation Computation, Unusual Voting Pattern Detection, Link Propagation Analysis, Clustering. In the first iteration, the PSYMBIOSYS InfluenCial Platform ranks influencers according to the times they are mentioned in other people's discussion taking also into consideration how "important" those people are, according to the PageRank algorithm.

2.1.3 Summary of Functionalities

The PSYMBIOSYS InfluenCial Platform provides data analysts, product-service managers and generally the product-service design team with the following core functionalities:

- *Detect influencers*: Identify and rank influencing behaviour per industry, topic and in time with the help of the PageRank algorithm.
- *Track and cluster interactions*: Visualize the correlation between different influencers on specific topics of interest.

In general, the procedure followed included the following steps: (a) Import data that contain or imply the importance of a user in a certain network/communication channel; (b) Extract the interactions that indicate influencing behaviour and model them in a predefined format; and (c) Apply importance measuring algorithms and identify influencers graphs (clusters of influencers or independent graphs). When the information retrieved is in the form of text, mentions of specific predefined entities are extracted and algorithms to detect entities commonly appearing together are applied. When the influencers along with the associated topics are calculated, the PSYMBIOSYS InfluenCial Platform starts investigating what is the correlation between trends and influencers to identify who is talking about what in the specified time period. Such an extensive calculation also classifies hashtags to topics and explores link propagation in time, in order to result into the daily impact of each influencer on each topic category.

In the first release of the PSYMBIOSYS InfluenCial Platform, the Big Data sources from social networks include: Twitter and Instagram, that were considered as most relevant for the PSYMBIOSYS application in the fashion and furniture domains (by the PIACENZA and AIDIMA use cases).

Selected functionalities of the PSYMBIOSYS InfluenCial Platform are also exposed as a RESTful API as presented in the following table.

Table 2-1: PSYMBIOSYS InfluenCial Platform – API Calls

Big Data Results API Calls			
VERB	**PATH**	**DESCRIPTION**	**Parameters**
GET	/[hostname]/spark/influencers/	Get a list of all influencers.	Page: int Network: string
GET	/[hostname]/spark/influencers/social/	Get a list of all influencers based on a specific social network.	Page: int Network: string
GET	/[hostname]/spark/influencers/{influencerID}/	Get information about an	

		influencer with the given id.	
GET	/[hostname]/spark/graphs/	Get all the discussion graphs based on specific topics.	Page: int Network: string Topics: list of strings
GET	/[hostname]/spark/recommendations/	Get a list of recommended topics based on a list of input topics.	Topics: list of strings

Web Interface Internal DB API Calls			
VERB	**PATH**	**DESCRIPTION**	**Parameters**
POST	/[hostname]/api/topic/	Add a new topic.	name: string twitter_hashtag: string
POST	/[hostname]/api/influencer/	Add a new influencer	first_name: string last_name: string twitter_username: string image_link: string rank: float stars: float updated: Date
POST	/[hostname]/api/week/	Add a new week of results	start_date: Date score: float
POST	/[hostname]/api/correlation/	Add a new correlation of results.	influencer: int, influencer_id topic: int, topic_id average_score: float updated: Date

2.2 Technical Information

2.2.1 Internal Architecture

The architecture upon which the PSYMBIOSYS InfluenCial Platform is built relies upon a set of state-of-the-art, open technologies (Figure 2-1):

- *Pre-Processing Layer* which is based on ElasticSearch as an indexing repository that allows fast and complex queries on unstructured data, allowing real-time data and analytics. Note: the ElasticSearch infrastructure is indirectly included through the FITMAN Anlzer Specific Enabler.
- *Cluster Computing Engine* built on top of Hadoop and Spark in order to conduct the necessary big data processing activities, running the algorithms for the identification of social influencers and their graph interrelations.

- *Results Storage Layer* to store the structured results of the big data analysis conducted on the selected social data.
- *Post-Processing, Export and Visualization Engine*: JavaScript-based visualization frameworks have been used to present in a meaningful and user-friendly way the results of the analysis.

It needs to be noted that the Data Connectors in the present first release include only the FITMAN Anlzer Specific Enabler that has already implemented RESTful based connectors in order to retrieve data according to the users' preferences, yet in future releases further data sources are expected to be incorporated.

The PSYMBIOSYS InfluenCial Platform is built on the Django Web Framework, written in Python, which follows the model–view–controller (MVC) architectural pattern.

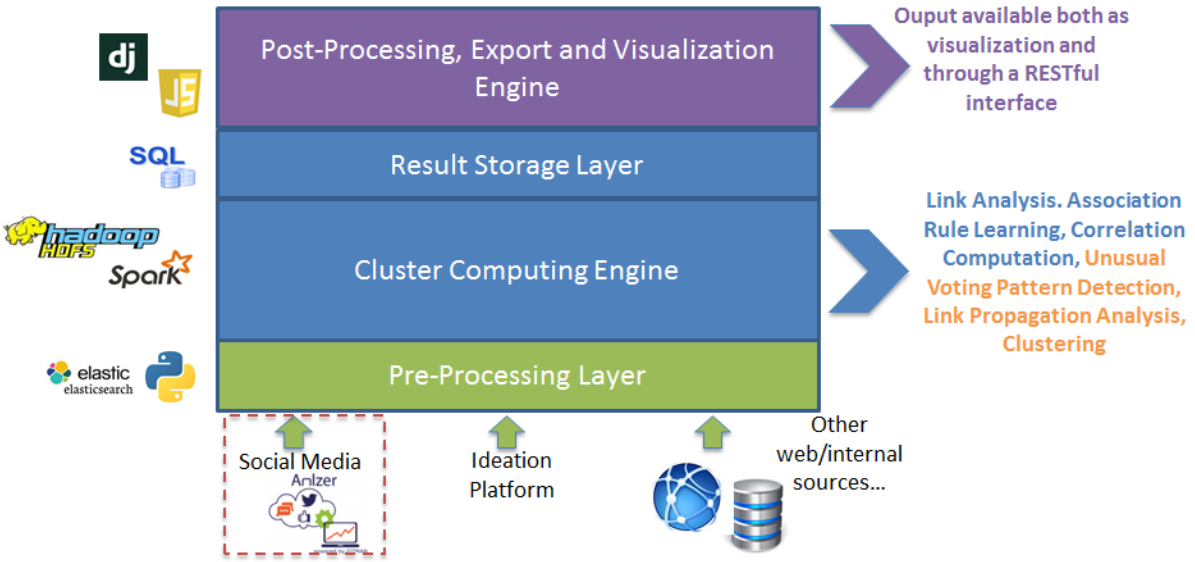

Figure 2-1: PSYMBIOSYS InfluenCial Platform: High-level Architecture

2.2.2 Technological stack

Item	Value
Nature	Web Application
System requirements	• Big Data Infrastructure: 8 CPU, 16 gb RAM • Web App: 1 CPU, 512 mb RAM • Anlzer Infrastructure: 2 CPU, 4 gb RAM
Programming Language	Python
Development Tools	Any Python IDE (e.g. Eclipse or Pycharm)
Additional Libraries / Tools / Frameworks	• FITMAN Anlzer SE, v2.0 • Apache Hadoop, v2.5.1 • Apache Spark, v1.6.0 • Django, v1.8 • Python, v2.7.9 • REQUESTS

	• PYTHON-TWITTER • (python) MARKDOWN • psycopg2, v2.5.1 • (python) white-noise • python-instagram • (python-django) static • (python-django) REST FRAMEWORK • (python) unirest • WIDGET-TWEAKS • gunicorn • Docker
Application Server	Nginx and gunicorn
Databases	HDFS for Big Data and SQL for the Web Interface (i.e. SQLite)
I/O formats	JSON serializations

2.2.3 Technical Manual

Since there is a significant complexity to deploy the tools and frameworks related PSY-InfluenCer platform, an open source Docker container was created and is available at: https://github.com/epu-ntua/pyspark-docker. This Docker image helps running Spark (on Docker) with the following installed: pySpark (Spark 1.6.0) on Hadoop 2.5.1; python 2.6.6; numpy 1.9.0; scipy 0.14.0; scikit-learn 0.15.2.

Further instructions on how to set up the PSY-InfluenCer platform are available at: https://github.com/epu-ntua/PSYMBIOSYS-Influencial/blob/master/README.md

In order to set up the FITMAN Anlzer Specific Enabler, detailed guidelines are available at: https://github.com/epu-ntua/FITMAN-Anlzer/blob/master/InstallationGuide.md and https://goo.gl/TbcIyq

2.2.4 Licensing

The PSY-InfluenCer platform is released under the MIT Licence.

In case of the development of a visual interface based on the Software, the above copyright notice shall be placed in a clear and visible position also.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

The 3rd party tools that were used for the development of the PSY-InfluenCer platform are listed in the following table along with their licences.

Software	**License**
FITMAN Anlzer SE, v2.0	MIT
Apache Hadoop, v2.5.1	Apache 2.0
Apache Spark, v1.6.0	Apache 2.0
Django, v1.8	BSD
Python, v2.7.9	Python Open Source License
REQUESTS	Apache 2.0
PYTHON-TWITTER	Apache 2.0
(python) MARKDOWN	MIT License
psycopg2, v2.5.1	GNU Lesser General Public License
(python) white-noise	MIT licence
python-instagram	Include Copyright (c) 2014, Facebook, Inc. All rights reserved.
(python-django) static	MIT License
(python-django) REST FRAMEWORK	Include Copyright (c) 2011-2016, Tom Christie All rights reserved.
(python) unirest	Deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software
WIDGET-TWEAKS	
gunicorn	
Docker	Apache 2.0

2.3 User Manual

In its final release, the PSY- InfluenCial platform will provide flexibility to the product-service design team to decide which sources should be considered and which algorithms should be applied before the influencer identification process starts as indicated in the following figure.

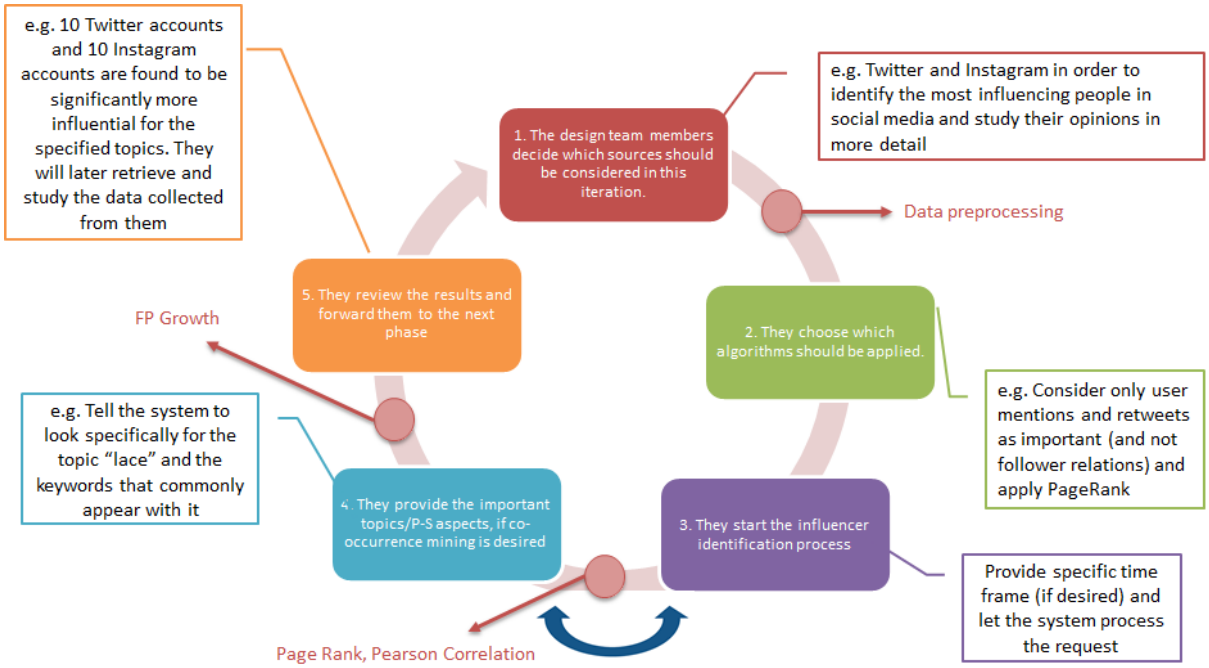

Figure 2-2: PSY- InfluenCial platform (end-user view)

In the first release, the functionality developed includes only steps 3-5 and is explored in the next paragraphs through an illustrative example of a data analyst or a product-service manager in the fashion industry who looks for potential influencers on the domain.

Figure 2-3: PSY- InfluenCial platform home page

Upon logging in in the PSY- InfluenCial platform (in figure 2-3), the user views the landing page (figure 2-4), presenting various topics for which he may retrieve the top influencers.

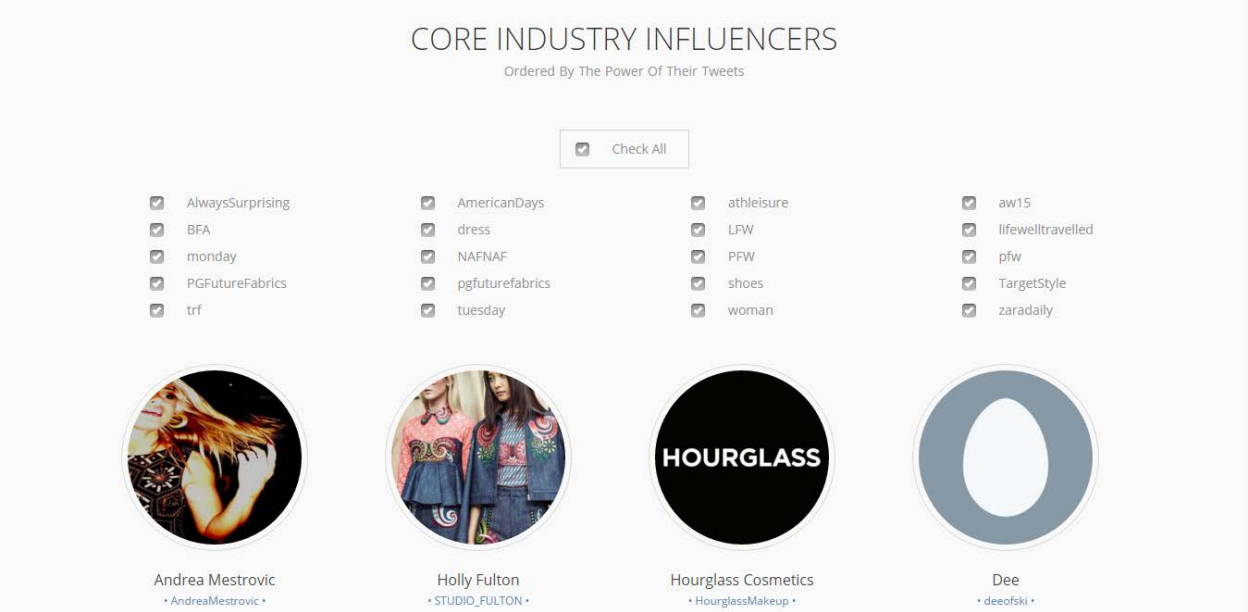

Figure 2-4: PSY- InfluenCial platform: landing page for topics selection

By scrolling further down the user is presented with the core influencers for the fashion industry as a whole (figure 2-5). To further reduce the amount of influencers on the page, the user is able to filter them through a series of checkboxes (as depicted in figure 2-4), which are in fact the topics that these influencers are mainly talking about.

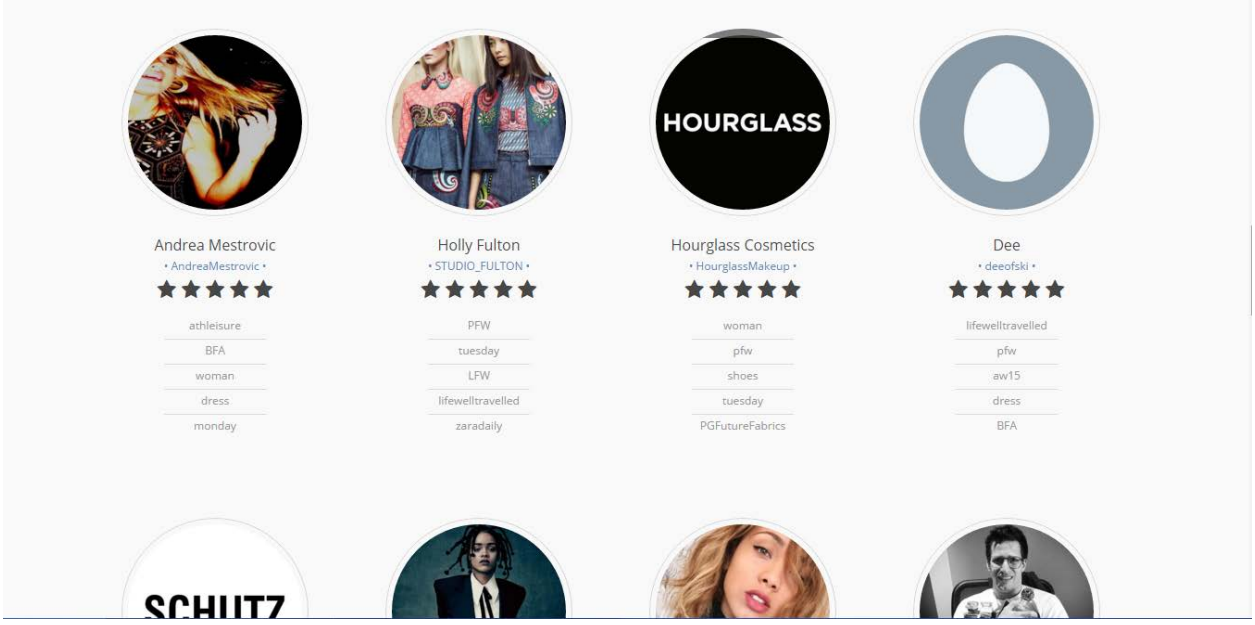

Figure 2-5: PSY- InfluenCial platform: influencers' overview

Each influencer has a profile picture which matches the one that he has on Twitter, a name, a twitter name that is a clickable link for his profile on Twitter, a PageRank score and a list of the five (5) topics mostly associated with him. The PageRank score is a simple metric derived from the PSYMBIOSYS InfluenCial algorithms in order to denote the social importance of the user by analyzing the traffic in his social accounts, specifically his posts and the

sharing/retweeting of his posts by others. This score is normalized in a scale of zero (0) to five (5) and is presented through a star system. Thus, the higher stars a user has the more socially important he is. Below the stars, the main topics the user is posting/tweeting about are presented. They are the same topics as the ones we encountered as filters in Figure 2-4, enabling the user to find out the specific influencers per topic.

By clicking on an influencer, additional information appears right underneath. The top three (3) topics of influencer, followed by two types of charts, a timeline chart (in Figure 2-6) and a mean-average chart (in Figure 2-7).

Figure 2-6: PSY- InfluenCial platform: influencers' details - timeline chart

The timeline chart depicts the intensity of the user's social activity throughout the past weeks for each of the topics that he influences, while the mean-average chart refers to the whole time span for which the PSYMBIOSYS InfluenCial algorithms has collected data.

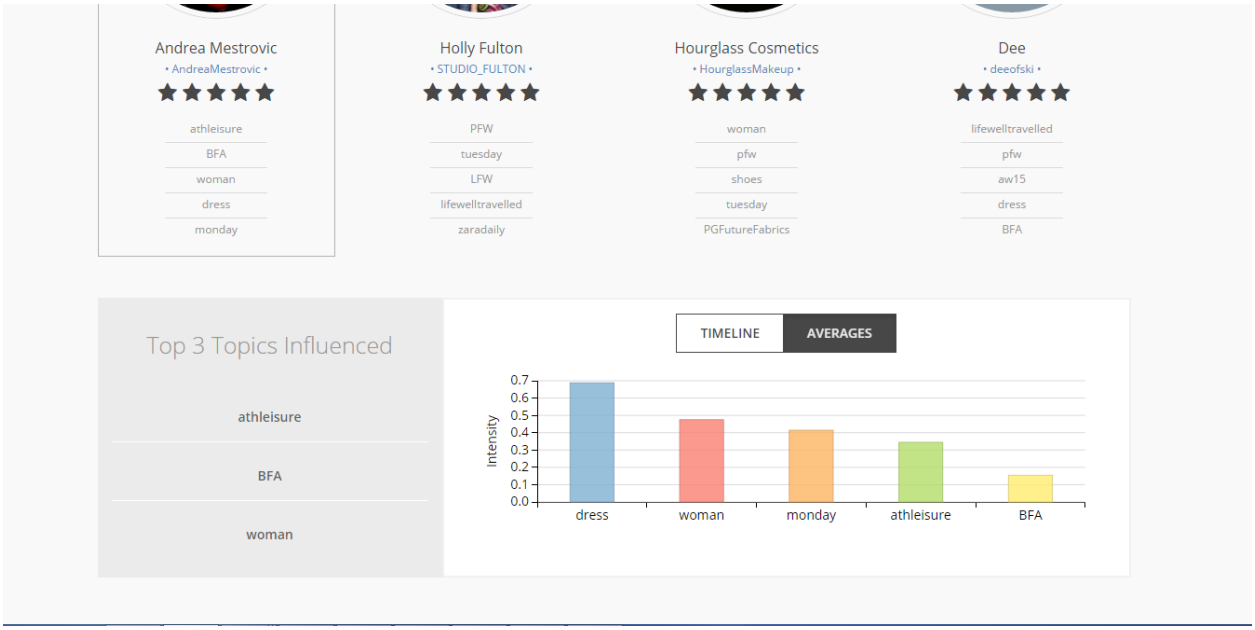

Figure 2-7: PSY- InfluenCial platform: influencers' details - averages chart

The influencers page generally allows for navigation to the most socially dominant people, filtered out by the topics that they influence. It is accompanied by a dedicated discussions page that provides an easy way to view the correlation between different influencers on the same topics.

As depicted in Figure 2-8, the user is prompted to enter a list of topics that he/she is interested in, and a graph will denote the correlations between the influencers on the topics that the user has selected. The system also recommends topics based on the topics that the user has entered.

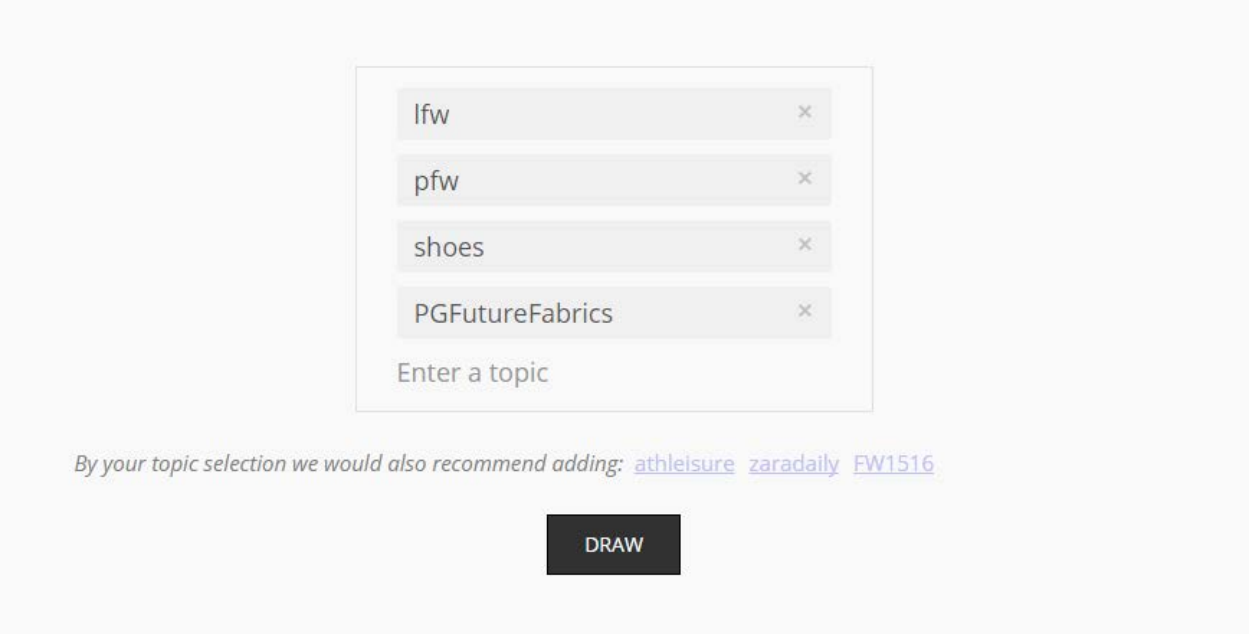

Figure 2-8: PSY- InfluenCial platform: discussion page – topic selection

Finally, the output is a graph where the nodes represent the influencers and the edges signify different topics (each topic has its own unique colour). Thus, an edge between two nodes is in fact a correlation between these two influencers on the specific topic. In other words, these two users have mentioned each other (or are someway related to each other) on their social activities. It is easily understandable that the more edges derive from a node, the more the retweets/reposts the specific node (influencer) had by others. Lastly, the topic to edge-colour mapping is viewable on the right side of the page, while the details of any node that is hovered at any time are available on the left side of the page.

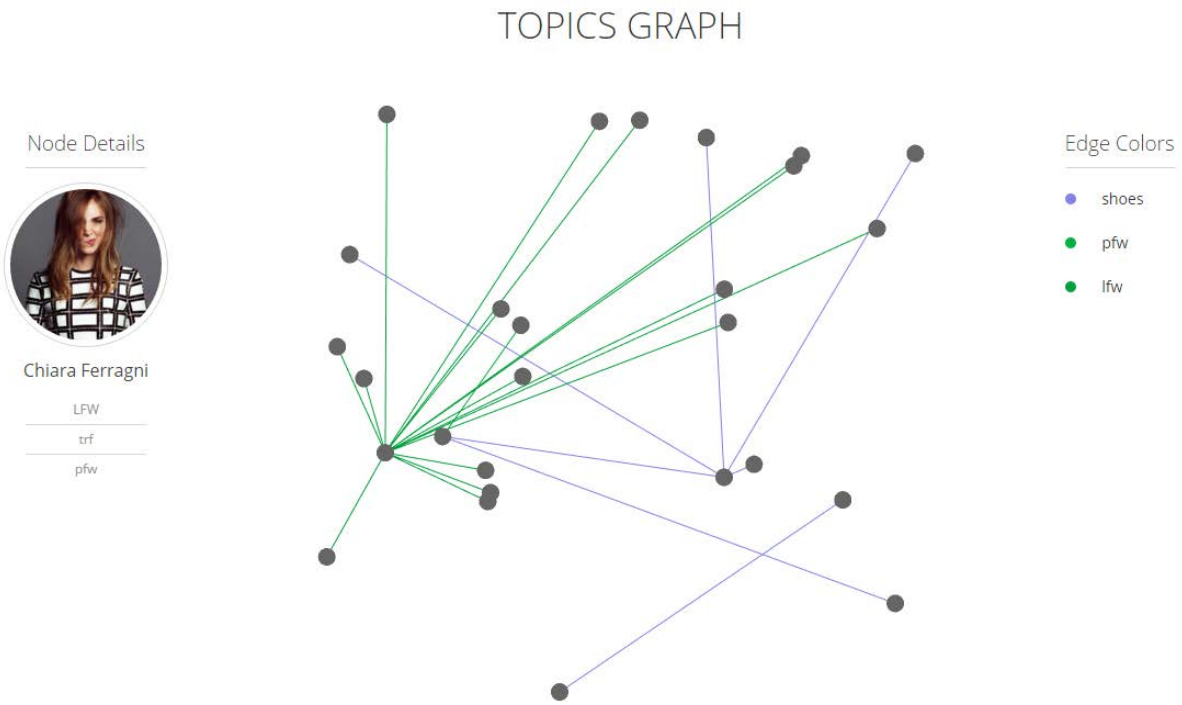

Figure 2-9: PSY- InfluenCial platform: discussion page – graph visualization

2.4 Conclusions and Future plans

In summary, the present first release of the PSYMBIOSYS InfluenCial Platform features the following:

- The influencer identification is based on a PageRank implementation for social data retrieved by Twitter and Instagram.
- Topic-user correlation and frequent co-occurrence mining have been incorporated in the platform.
- A web user interface for navigating the results is deployed and an initial version of the API for exporting results is available.

- The testing phase including experimentation with end users' data (i.e. fashion data for the PIACENZA use case and furniture data for the AIDIMA use case) has already started.

The next steps for the PSYMBIOSYS InfluenCial Platform include:

- Creation of more results' visualisations or API calls to be fed into Crowd Innovation Platform which is under development in WP5.
- Implementation of an admin panel in order to provide better control on the data sources and algorithms.
- Incorporation of additional algorithms for social influencers' identification.
- Further experimentation with trials in order to validate results and improve the algorithms for social influencers.
- Additional connectors and pre-processors for additional platforms-social sources (e.g. the PSYMBIOSYS Ideation Platform in WP5).

3 Manufacturing-Big Data Analytics Platform (PSY-AP)

3.1 Software Description

3.1.1 Overall Data

Item	Value
Component Name	Assets Platform (PSY-AP)
Software version	v0.50
Reference workpackage	WP8.4
Responsible Partner	TXT
Contact person	Marco Gallazzi, marco.gallazzi@txtgroup.com
Source control	N.A.
Short Description	The platform allows to collect a very big amount of data coming from the production line, in order to show in a graphical view important information that will allow the management team to take critical decision in different business areas.

3.1.2 Purpose of the tool

The need of verifying data that comes from production and machinery is strategic, but can be difficult. Very big amounts of data are difficult to analyse without the right platform. The PSYMBIOSYS Big data platform (PSY-AP) will allow the organization to collect all information, production parameters, signals and events come from production, when needed, to analyse very fast such information collected and take right decisions to improve production.

Another strategic point achieved with this platform is to predict production values and, on based on these values, to create specific alerts to redesign the product.

3.1.3 Summary of Functionalities

The platform provides data collector, data analysis, prediction algorithms, threshold definition and data graphical views with the following core functionalities:

- *Data prediction*: On the basis of historical data, identify possible production quality defects for each signal type with the help of the right algorithm.
- *Threshold view*: Show and define for each signal all thresholds.
- *Data visualization*: Show for each signal the trend and signal cross the threshold.

In general, the procedure included the following steps: (a) Acquire data from the production line that contain or imply the importance of quality; (b) Apply importance-measuring algorithms and identify defects; (c) show to the end user the results of analysis.

3.2 Technical Information

3.2.1 Internal Architecture

In this paragraph the PSYMBIOSYS architecture for big data is presented, focusing on the different purposes of the tools analysed and used.

CLUSTER DATA STORING

The first step in a big data architecture is to choose how to store data in an efficient way in order to store large amounts of data using fast and scalable techniques and tools in order to build on the top the analysis that has to be performed.

For the PSYMBIOSYS purposes in big data, the selected tool is Hadoop, which permits to store large amount of data (TB or even PB) in cluster that can be composed by more than 200 machines.

Hadoop is also able to store data without having to pre-process them, meaning that data can be stored leaving them unstructured and without knowing from the beginning how they will be used in the future analysis.

As data are stored using a distributed file system (HDFS) handled directly by Hadoop, this methodology is scalable (new nodes can be added at any time) and data replication is asynchronous in order to have a small impact on performance.

CLUSTER MANAGING

Although Hadoop has high performance when dealing with large amounts of data, it has comparable performances when working with smaller datasets. Just after creating the Hadoop architecture, suppose with the minimum infrastructure constituted by one master and 3 slaves, the amount of data is quite small, but having the complete map reduce algorithm can have a bad impact on the performances.

For this reason, Spark has been developed and introduced also in PSYMBIOSYS; Spark is a free and open source framework that can be mounted on top of Hadoop in order to provide a set of functionalities to manage the HDFS and the way to interrogate it.

With Spark, it is possible to configure a simpler map-reduce algorithm that can be used when data are not big enough for Hadoop (starting of the architecture or for some use cases where data are unstructured but not so big in volume). In those scenarios, Spark strongly increases, the performances of Hadoop simplifying the map reduce which is a complex and time consuming algorithm.

Spark also provides other features to extract data from Hadoop such as streaming data, machine learning, graph data processing and SQL queries.

DATA PROCESSING

Depending on the type of analysis to be performed, different tools can be used. In this section, two different tools have been analysed.

R is both a programming language and a software environment that provides for statistical analysis, time series analysis, simple graphic representation and reporting. It is used to find statistical models such as linear and non-linear regressions, clustering and classification. R can be also integrated with Weka (if needed) in order to get more machine learning and data mining algorithms. The basic algorithms are written in R languages, which is object oriented, while other additional libraries are written in C++.

In case traditional business intelligence is also required, Kylin provides a distributed analytic engine for SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting large datasets.

Kylin has been created to cover the missing of a business intelligence tool for Hadoop and big data in general, decreasing the latency to execute queries on a distributed architecture. The OLAP cube in Kylin in pre-built speeding up the queries that are already calculated and cached.

RESULT VISUALIZATION

About the data visualization, one of the most important technology is Vaadin. Vaadin is an open source java server-side framework for building single page web application. It uses Java as the programming language for creating web script. It is designed to create and maintain of high quality web-based user interface easy.

Another import technology is Kendo UI. It is an HTML5 user interface framework to build interactive and high-performance websites and applications. It is used to build HTML5 apps that look native on any device, build desktop and mobile application for any browser. The Kendo UI have a particular component named Kendo UI PivotGrid to perform operations over multidimensional/pivot data. The widget uses the OLAP approach to present the result of multidimensional queries.

Vaadin and Kendo UI are used to visualize the data and both the tools create reports in a web environment. They can be used depending on the circumstances and one of these two tools can be chosen to build reports in HTML5.

3.2.2 Technological stack

Item	Value
Nature	Webapp
System requirements	The architecture described previously is implemented using four Hadoop servers, one master and three slave machines. In order to maximize the performances of the relatively small amount of nodes, a Spark client machine performs in-memory data processing operations, while Hadoop is used only as a distributed file system, HDFS. The Hadoop server machines mounts Ubuntu Server 14.04 64 bit. Machines detail: • No. 1 Master Hadoop Server o Name: vm-hadoop-master o Operating System: Ubuntu Server 14.04 LTS 64 bit ▪ Ubuntu Basic

	▪ OpenSSL Server o CPU 8-core o RAM: 12 GB o Memory Support: Hard drive 120 GB ▪ Root partition: 8 GB ▪ Swap partition: 32 GB • No. 3 Slave Hadoop Servers o Name: vm-hadoop-slave<*n*> o Operating System: Ubuntu Server 14.04 LTS 64 bit ▪ Ubuntu Basic ▪ OpenSSL Server o CPU 4-core o RAM: 8 GB o Memory Support: Hard drive 120 GB ▪ Root partition: 8 GB ▪ Swap partition: 32 GB
Programming Language	HTML5, EDMX, C#, Javascript
Development Tools	Microsoft visual studio 2013, Kylin, Pig, Hive
Additional Libraries	Vaadin, telerik kendo UI
Application Server	Linux
Databases	HDFS for Big Data and SQL for the Web Interface (i.e. SQLite)
I/O formats	JSON serializations

3.2.3 Technical Manual

To be available in the 2nd iteration (in D8.8).

3.2.4 Licensing

To be available in the 2nd iteration (in D8.8).

3.3 User Manual

To be available in the 2nd iteration (in D8.8).

3.4 Conclusions and Future plans

In summary, the present first release of the PSYMBIOSYS Assets Platform features the following:

- The data collection retrieves information directly from the production line.
- Prediction algorithms have been incorporated in the platform.
- The threshold user interface is developed and an initial version threshold management and graphical data visualization is available.

The next steps for the PSYMBIOSYS Assets Platform include:

- Incorporation of additional algorithms to predict in less time production failures.
- Improve graphical data view with tablet portable view.
- Further experimentation with others markets with same production problems.

4 Conclusions

Today, big data architectures do not have to satisfy only the requirement of handling large Volumes of data (GB or even TB), but also need to address aspects, such as:

- Variety: data coming from different sources (relational databases, sensors, social networks, etc.), having as result that they are also unstructured.
- Velocity: data are continuously stored and their volume is constantly increasing needing a scalable and flexible way to store them.
- Veracity: data are often dirty, meaning that outliers or not useful data has to be considered and they need to be treated before starting the analysis.

The purpose of the first iteration of the deliverable at hand (D8.7 “Big Data Analytics Tools and Platforms - First Iteration”) has been to document the PSYMBIOSYS dual big data analytics approach, planning and progress of implementation:

- With regard to social data, the first version of the Social-Big Data Analytics Platform (PSY- InfluenCial) has been released. Its scope is to identify social influencers whose ideas and proposals for product-service design should be weighted higher than the general public. Building on a Hadoop-Spark architecture and implementing the PageRank algorithm, the PSY- InfluenCial platform allows for targeted search for specific topics to identify the graphs of influencers and understand how such graphs overlap with each other. In the next release of the platform, the main indicative updates that will be planned (depending on the outcomes of the experimentation with the PSYMBIOSYS use cases) are: more advanced Big Data algorithms, admin panels and additional social big data sources.
- With regard to production data, the Manufacturing-Big Data Analytics Platform (PSY-AP) processes production data internal to the company (e.g. from machineries) in order to discover difficulties in the production and provide feedback to the product-service design team. The official release of the Manufacturing-Big Data Analytics Platform is planned for the final version of this deliverable (D8.8).

Annex I: References

Apache Hadoop (2016). Accessed on January 15th, 2016 from: http://hadoop.apache.org/.

Apache Spark (2016). Accessed on January 15th, 2016 from: http://spark.apache.org/

Django Web Framework (2016). Accessed on January 15th, 2016 from: https://www.djangoproject.com/

Docker (2016). Accessed on January 15th, 2016 from: http://www.docker.com/

Eclipse Open Source Community (2016). Accessed on January 15th, 2016 from: https://eclipse.org/

FITMAN Unstructured and Social Data Analytics (Anlzer) Specific Enabler Documentation (2016). Accessed on January 15th, 2016 from: http://catalogue.fitman.atosresearch.eu/enablers/documentation

FITMAN Unstructured and Social Data Analytics (Anlzer) Source Code (2016). Accessed on January 15th, 2016 from: https://github.com/epu-ntua/FITMAN-Anlzer

Instagram API Platform (2016). Accessed on January 15th, 2016 from: https://www.instagram.com/developer/

PostgreSQL (2016). Accessed on January 15th, 2016 from: http://www.postgresql.org/

PyCharm: Python IDE (2016). Accessed on January 15th, 2016 from: https://www.jetbrains.com/pycharm

Python Programming Language (2016). Accessed on January 15th, 2016 from: https://www.python.org/

Spark Python Programming Guide (2016). Accessed on January 15th, 2016 from: https://spark.apache.org/docs/0.9.0/python-programming-guide.html

SQLite (2016). Accessed on January 15th, 2016 from: https://www.sqlite.org/

Twitter REST APIs (2016). Accessed on January 15th, 2016 from: https://dev.twitter.com/rest/public