PROTEUS

Scalable online machine learning for predictive analytics and real-time interactive visualization

687691

D2.7 Prototypes setup and deployment

Lead Author: Daniel Toimil

With contributions from: David Piris, Marta García, Guillermo Barquero

Reviewer: Álvaro Agea

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Consortium (CO)
Contractual delivery date:	30/11/2016
Actual delivery date:	26/11/2016
Version:	1.0
Total number of pages:	18
Keywords:	

Abstract

The early detection of defects appearing in the Hot Strip Mill is of high importance for ArcelorMittal. In the previous deliverable the objectives pursued, the requirements about the available data and the Data Management Plan were described in order to achieve this goal.

The current deliverable goes one step beyond and details some requirements and restrictions to integrate the new technology in ArcelorMittal facility, including data accessibility, interoperability and security issues. Moreover, the continuous validation of the stages of the project is an essential tasks and to this extent several specific, measurable, achievable, relevant and time phased Key Performance Indicators (KPIs) will be proposed, although further information can be found in deliverable D2.6.

Therefore, this document will serve as a guideline for our researcher partners to obtain valuable feedback about the infrastructure required to obtain and deal with the data and the way to make an objective measurement about the progress achieved. This will establish the foundations to be followed by the remaining deliverables, i.e., the three prototypes and the final integrated version.

[End of abstract]

Executive summary

In the previous deliverable, ArcelorMittal detailed the requirements, objectives, and Data Management Plan in order to address the early-stage detection of defects in the Hot Strip Mill. This second document contains the setup of the prototype execution to ensure a correct evaluation of the project advances, including detailed benchmark and KPIs for impact assessment.

ArcelorMittal, as the end-user industrial partner, provides the actual factory facilities for testing the new technology developed in the scope of the PROTEUS project in realistic conditions. Thus, this document will define how the software prototypes will be deployed and integrated in the ArcelorMittal factory, describing the infrastructure to get actual data in real-time from the factory sensors and historical registers from the systems without affecting to the daily functioning steelmaking process. Hardware and software requirements and restrictions to the integration with existing systems will be also specified here.

The contribution of ArcelorMittal, as data provider and as prototype tester, and the definition of clear functional, scientific, technical and verifiable requirements are key issues. This will demonstrate the outputs and ensure the exploitation of PROTEUS into the steelmaking industry and other domains with equivalent requirements. To achieve these objectives, a continuous validation process in the ArcelorMittal facilities will be followed**.** In this first deliverable, certain specific, measurable, achievable, relevant and time phased Key Performance Indicators (KPIs) for each evaluation/validation test is proposed. Nonetheless, deliverable D2.6 has further information about the KPIs to be used. Detailed description of how to execute the new technology in the factory system, including data accessibility, interoperability and security, is attached.

To sum up, making use of specific scenario methodology and concrete KPIs, researchers will obtain valuable feedback about their advances to provide guidance for the next steps, the three prototypes and the final integrated version.

Document Information

IST Project Number	687691	**Acronym**	PROTEUS
Full Title	Scalable online machine learning for predictive analytics and real-time interactive visualization		
Project URL	http://www.proteus-bigdata.com/		
EU Project Officer	Martina EYDNER		

Deliverable	**Number**	D2.7	**Title**	Prototypes setup and deployment	
Work Package	**Number**	WP2	**Title**	Industrial case: requirements, challenges, validation and demonstration	

Date of Delivery	**Contractual**	30/11/2016	**Actual**	26/11/2016
Status	version 1.0		final □	
Nature	report x demonstrator □ other □			
Dissemination level	public □ restricted x			

Authors (Partner)	Daniel Toimil, Marta García, Guillermo Barquero (AMIII), David Piris (TREE)			
Responsible Author	**Name**	**Responsible Author**	**Name**	**Responsible Author**
	Partner		**Partner**	

Abstract (for dissemination)	The early detection of defects appearing in the Hot Strip Mill is of high importance for ArcelorMittal. In the previous deliverable the objectives pursued, the requirements about the available data and the Data Management Plan were described in order to achieve this goal. The current deliverable goes one step beyond and details some requirements and restrictions to integrate the new technology in ArcelorMittal facility, including data accessibility, interoperability and security issues. Moreover, the continuous validation of the stages of the project is an essential tasks and to this extent, several specific, measurable, achievable, relevant and time phased Key Performance Indicators (KPIs) will be proposed, although further information can be found in deliverable D2.6. Therefore, this document will serve as a guideline for our researcher partners to obtain valuable feedback about the infrastructure required to obtain and deal with the data and the way to make an objective measurement about the progress achieved. This will establish the foundations to be followed by the remaining deliverables, i.e., the three prototypes and the final integrated version.
Keywords	Requirements, KPI

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
13/10/2016	1	Daniel Toimil	First draft
17/10/2016	2	Marta García, Guillermo Barquero	Review copy
31/10/2016	3	David Piris Valenzuela	Review copy
26/11/2016	4	Daniel Toimil	Final version

Table of Contents

Executive summary 3
Document Information 4
Table of Contents 5
Abbreviations. 6
1 Introduction 7
2 Hardware Infrastructure 8
2.1 Hardware Specifications 9
2.2 Cluster Network Topology 10
3 Software Deployment 11
3.1 Operating System 11
3.2 List of software to be installed 11
4 Connectivity with ArcelorMittal Data Sources 13
4.1 Data Sources definition 13
4.2 Interface with Data Sources 14
4.3 Topology 15
5 Evaluation of Objectives during Iterative Prototyping 17
6 Conclusions 18

Abbreviations.

KPI: Key Performance Indicators

1 Introduction

In deliverable D2.1, ArcelorMittal detailed the requirements, objectives, and Data Management Plan in order to address the early-stage detection of defects in the Hot Strip Mill. In fact, one of the main defects affecting the coils is their flatness, which determines the quality of the products. Predicting this kind of dimensional defects is the main target in this scenario due to its impact on the global steel process. Nonetheless, as it was mentioned, the steel production phase is a complex operation divided in different processes. A key process is the coil production because the defects introduced in that early stage have a great economic impact afterwards. For this reason, the Hot Strip Mill takes on an important role in this steel making system, as it is the facility where the slabs coming from the continuous casting facility are transformed into coils.

The four phases of the Hot Strip Mill process, described in detail in deliverable D2.1, involve multiple and diverse parameters that affect the final dimensional properties of the obtained coil. Some examples of these parameters are the temperature, the tension in the rollers or the speed of the plate when entering the coiler. All the data that register these parameters is obtained from a sensor network installed across the facility. Thus, PROTEUS will address the challenge of providing scalable online machine learning and real-time interactive visual analytics capabilities to solve the defects problem by exploiting these massive streaming real-time data generated during the Hot Strip Mill.

Once the context of the problem is recalled, the current deliverable goes one step beyond and details some of the requirements that ArcelorMittal, as the end-user industrial partner, needs to test the new technology developed in realistic conditions. We refer to the prototypes, one of the mainstays of PROTEUS, where the result of the research developed will be tested. Nonetheless, and once again, it will be essential to impose some requirements in order to make feasible and profitable the prototype execution. Thus, the main goal of this deliverable is to describe the correct installation of the prototypes so that the former requirements are addressed.

To be more precise, this document will define how the software prototypes will be deployed and integrated in the ArcelorMittal factory, describing the infrastructure to get actual data in real-time from the factory sensors and historical registers from the systems without affecting to the daily functioning steelmaking process. Hardware and software requirements and restrictions to the integration with existing systems will be also specified, such as data accessibility, interoperability and security.

Finally, together with the technical aspects derived from the correct framework set-up, several Key Performance Indicators (KPIs), that will be used to measure the performance of each prototype solution, will be introduced.

The performance metrics and KPIs will be defined by using existing benchmarks as a baseline. Nonetheless these KPIs will have to be specific, measurable, achievable, relevant and time phased for each evaluation/validation test proposed. A further detailed list of KPIs will be given in deliverable D2.6.

In order to cover the former mentioned points, the document is structured as follows: in Section 2 the Hardware Infrastructure are determined, both the hardware Specifications required and the Cluster Network Topology; Section 3 defines the Software deployment, specifying the operating system needed, the list of software to be installed in order to make the prototypes work and the installation process of the PROTEUS proposed solution; Section 4 is devoted to determine the connectivity with ArcelorMittal Data Sources , describing the interface to deal with the Data Sources, the communication topology and the communication protocol. The evaluation of objectives during Iterative Prototyping, by means of realistic KPIs, is introduced in Section 5. Finally, the document ends with a conclusion section where the key aspects treated in the document are summarized.

2 Hardware Infrastructure

In this section, necessary hardware infrastructure will be defined for the proper functioning of Proteus project. Infrastructure described in this document is focused to resolve real time process paradigm.

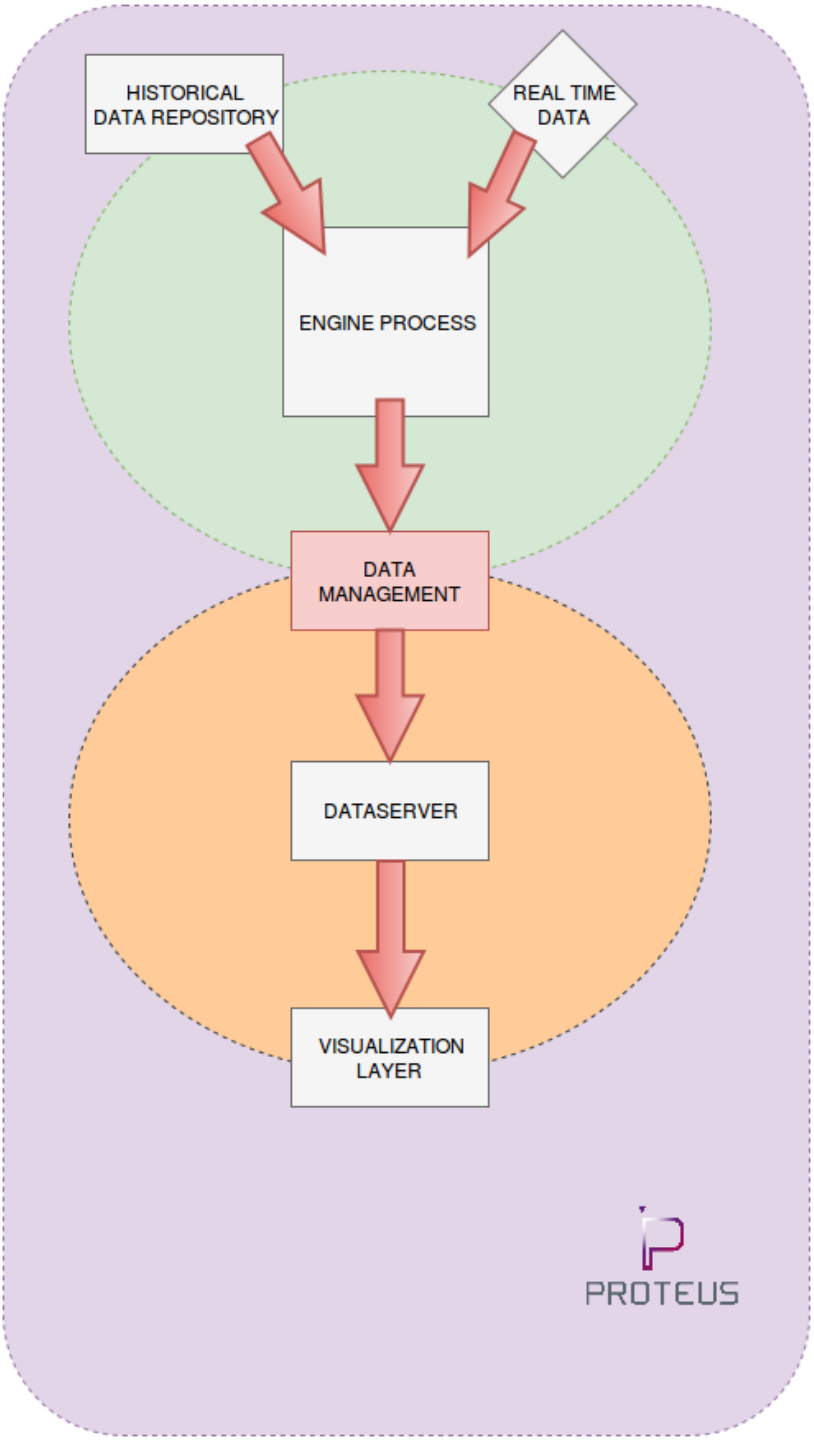


Figure 1 Project Infrastructure

Dataserver layer (aggregated data) could be integrated as a future improvement in Proteus architecture. If response time from retrieve, manipulate, and visualize historical data doesn't fits in initial proposal under 1 second, a dataserver to make incremental operations would be a great improvement choice.

It would be explained in two differents topics: hardware specifications and cluster network topology.

2.1 Hardware Specifications

2.1.1 Historical Data Repository

Historical Data Repository requires to be under a machine with the next specifications:

- 2 x Xeon E5 2650v2
- 64GB RAM
- 4TB HDD

To manage a Big Data solution, this server would be distributed into 4 virtualized machines in order to make a suitable parallel computing system.

- Master Node:
  - 8Gb Ram
  - 235 Gb HDD
- 3 Slaves Node (each one):
  - 8Gb RAM
  - 635 Gb HDD

2.1.2 Engine Process

Engine process requires to be deployed in a cluster with those specifications:

- Master Node:
  - 8Gb Ram
  - 235 Gb HDD
- 3 Slaves Node (each one):
  - 8Gb RAM
  - 635 Gb HDD

2.1.3 Data Management

Data management layer would be integrated inside engine process described previously:

- Master Node:
  - 8Gb Ram
  - 235 Gb HDD
- 3 Slaves Node (each one):
  - 8Gb RAM
  - 635 Gb HDD

2.1.4 Dataserver

Dataserver layer could be deployed in a machine with the next hardware specifications:

- 8 Gb Ram
- 1 Tb HDD

2.1.5 Visualization Layer

Hardware is under client side.

2.2 Cluster Network Topology

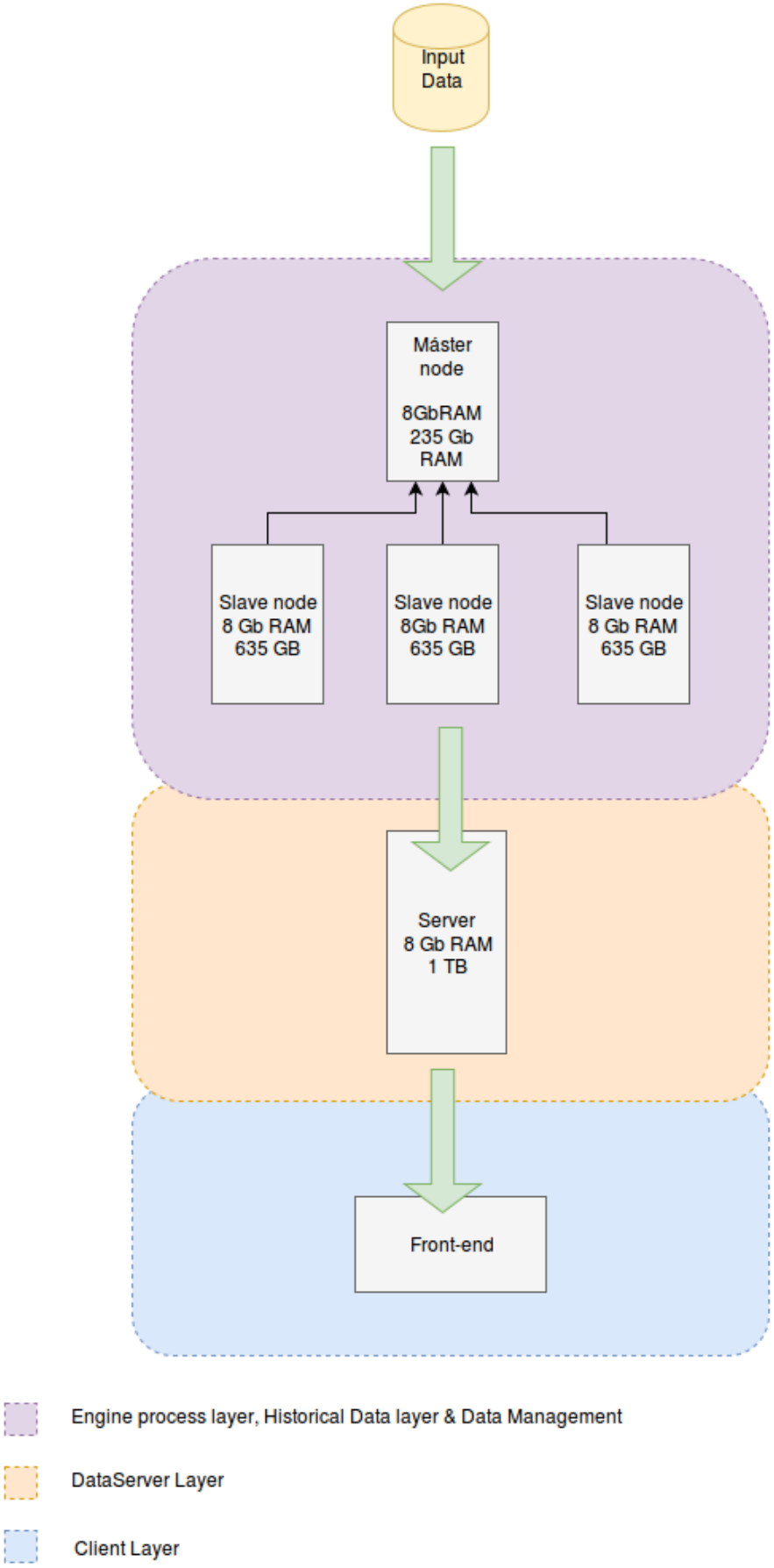


Figure 2 Cluster Network Topology

3 Software Deployment

3.1 Operating System

Every solution from each partner will be deployed and merged in an unique platform so it may support all the software necessary for those solutions. Thus, the operating system is highly important and it should be as reliable and stable as possible in order to reduce the risk of appearance of incompatibilities and crashes. For this reason, Linux distribution CentOS has been chosen. It is an open-source, stable, manageable and reproducible platform derived from Red Hat Enterprise Linux (RHEL). CentOS runs only the most stables versions of packaged software so it helps to reduce the risk of errors and crashes. In addition, thanks to its close link to Red Hat, it is equipped with a wide array of impressive security features such as a powerful firewall and the SELinux policy mechanism. To sum up, CentOS is chosen because it is a free, lightweight, fast and reliable operating system which fits perfectly the purpose of this project. In this deployment, version decided to use in this solution is 6.7.

3.2 List of software to be installed

Data streaming processing, machine learning and Big Data analysis require an appropriate framework that could manage the tough specifications in terms of speed and scalability. A lot of research has been carried out and many software have been developed to cover specific requirements and solve particular problems. Among all the available software, the other members of the consortium have specified a list of required software that need to be installed in the platform. All these software requirements are gathered in Table 1. The software listed covers from server storage and Big Data managing, in both real time and offline way, to modelling tools for creating machine learning models.

Table 1 Software development requirements from each partner

	Software Needed
DFKI	• **Apache Flink 1.1.2** • **Redis Server 3.2** • **Apache Kafka 0.10.0.1** • **JDK 8** • **ssh-server , ssh-client**
Lambdoop	• Redis Server 3.2 • JDK 8 • **Apache Cassandra 3.7** • Internet Connection (INPUT/OUTPUT)
BU	• Apache Flink 1.1.2 • Redis Server 3.2 • Apache Kafka 0.10.0.1 • JDK 8 • ssh-server , ssh-client • **Matlab** • **R**

TREE	• Apache Kafka 0.10.0.1 • **Apache ZooKeeper 3.4.6.2** • **D3 Java Library**

4 Connectivity with ArcelorMittal Data Sources

4.1 Data Sources definition

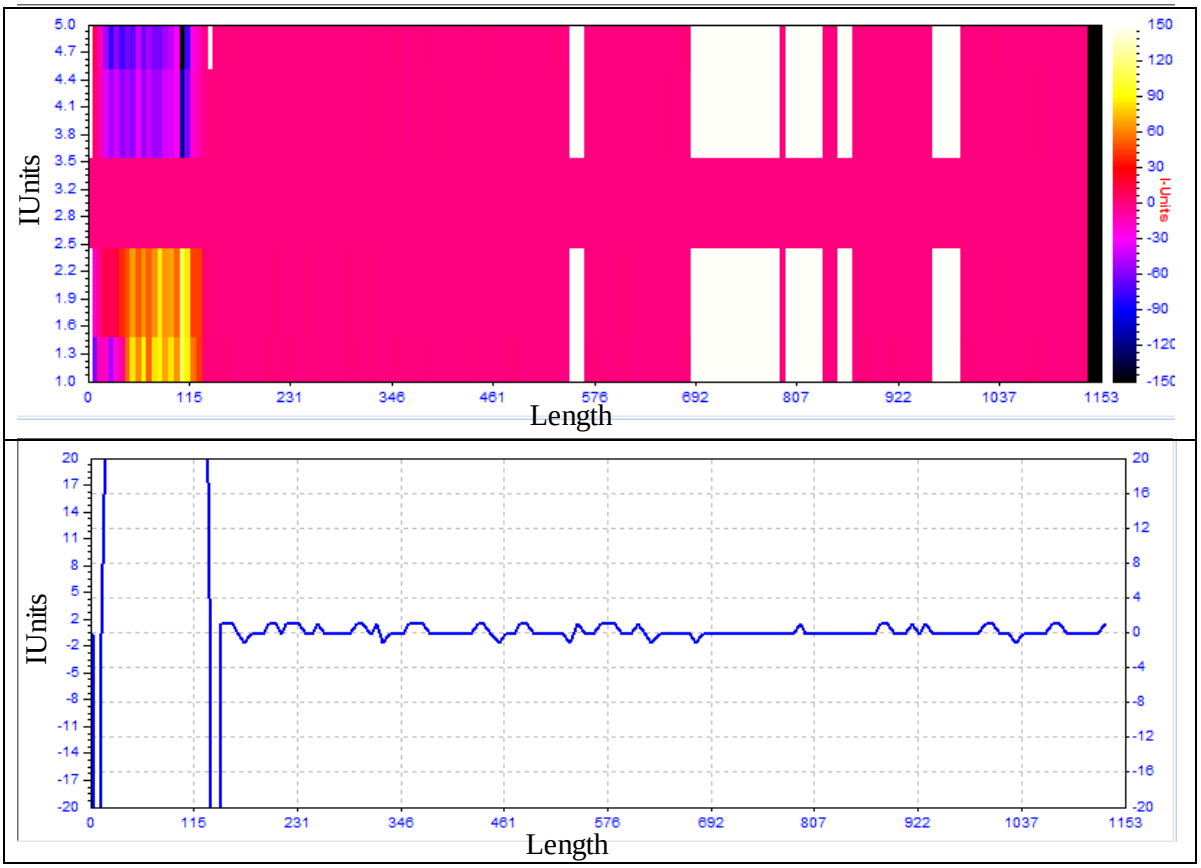
The data collected and intended to be analysed is presented in two different datasets:

1. **Process dataset**: This dataset is the result of joining several tables from the database of the Hot Strip Mill. This data is composed of the measurements of sensors, which are distributed along the facility measuring several process variables. In this dataset, each coil has associated a single value for each variable, which, in many cases, represents a summary (an average) of the measurements collected along the time.

   The data source from which this dataset is constructed is the Hot Strip Mill process database. This database includes both quantitative and qualitative data, stored in 42 different tables. All the tables share the same key variable, which allows us to join and relate all this information with certain specific coils. The tables in the Hot Strip Mill database store a total of 7475 variables related to the coil production process since 2010 with ~840000 records for each variable. It contains mostly numerical and categorical values and its size increase as new coils are produced. The size of each of the 42 tables is around 300-700 MB.

   These tables are updated continuously, as new coils are processed. As a general guideline, the generation rate is usually between 32 and 500 milliseconds. This variation makes the system have misunderstandings with data caught at different time instants, since there is not a general generation rate. Our system has mixed all this data together in order to have them in a manageable way to analyse it.

   It is important to recall that, data containing information about detected defects are not updated at the same rate as the Hot Strip Mill database, since the processed coils are not evaluated for defects until later when the coil is processed in the flatness measurement system.



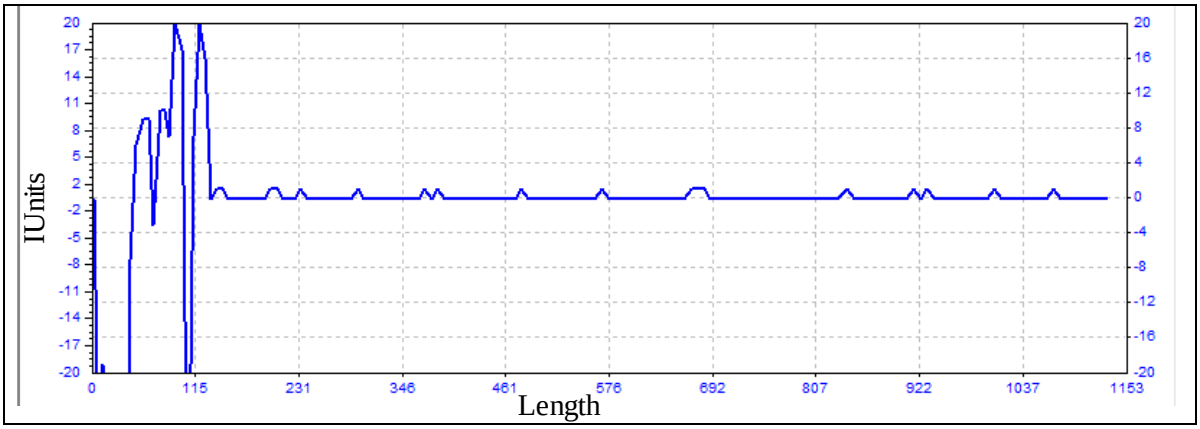


Figure 3 Top to bottom: Flatness map, asymmetrical and symmetrical flatness

2. **Continuous data**: This data comes from the system that measures flatness and from different sensors installed in the Hot Strip Mill. It consists of a set of time series variables, such as temperature or flatness, that have been measured continuously along the coil length. It also contains the target variables (flatness) that may be employed in the construction of the prediction model. The interval in which each variable is measured varies depending on the system capabilities to acquire and store the data and the speed of the coil being processed. For example, if a coil is produced in approximately two minutes and we usually have 100-200 values of each signal, we can approximately say that the generation rate of these variables is around 1 second. Thus, the analysis of this data requires online capabilities in order to obtain the results as soon as possible. The historical values of these variables have a current size of ~300 GB.

   Additionally, flatness maps are also generated once the entire coil is processed and measured by the flatness measurement system. These maps are unstructured data (images) that contains useful information about the flatness of the coil. One flatness map is available for each coil.

   This real time data includes 29 numeric variables and images (maps) from the processed coils. Examples of the target variables can be found in Figure .

4.2 Interface with Data Sources

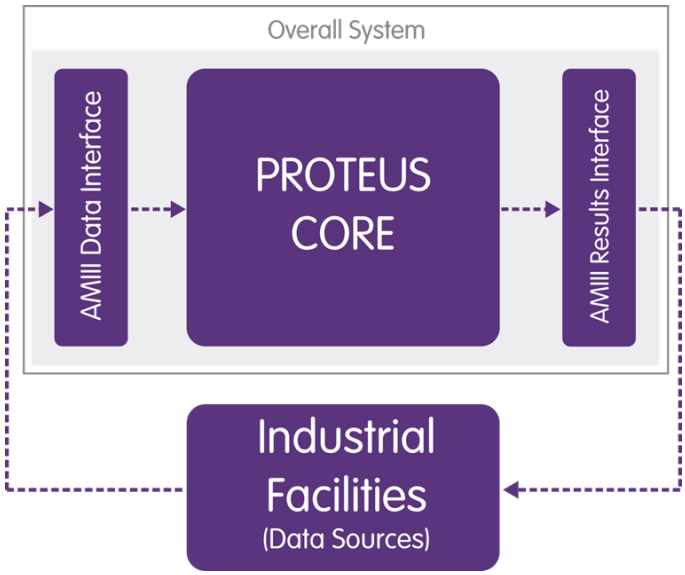


Figure 4 Scheme of the system.

1. Initially the data is collected from the Industrial Facilities, which in our case are the process and flatness databases that collect information measured by sensors at the Hot Strip Mill.

2. All the stream of data from our facilities is pre-processed in the AMIII Data Interface. This interface is an internal tool where all the data will enter compulsory and which has two main goals. The first one is to integrate the data from the different data sources into the appropriate format to send to our partners. Mainly, this process collects the tables arriving from the different sensors and gathers them in a unique table containing all the information with the structure outlined in previous sections. And the second goal is to anonymize/codify the data, so any confidential information is not displayed. To do so, some of the standard procedures are:

    - To replace the name of the variables to generic ones.
    - To replace the value of the classes in the factor variables so that any private internal class name is replaced by a generic one. For example, the type of steel could have two classes, BH or IF, and we would replace them by Class_1 and Class_2.

3. In the PROTEUS Core the data is analyzed by our partners and the results are obtained.
4. Finally, the results provided by our partners are introduced in the AMIII Results Interface, where they are adapted for its application and interpretation by the operators in our facilities.

4.3 Topology

Real-time high frequency streaming processing presents huge technical challenges that cannot be overcome in a traditional way. PROTEUS aims to fill the technological gap between the state of the art and a real industrial solution. Actually, it is not only a technological issue but also a problem of topology and architecture. Different subsystems in charge of specific complex functions must be integrated in a platform where no bottlenecks, in terms of latency, are allowed. In this section the topology of the overall system is defined. The objective is to clarify what different systems will have to be deployed and how to connect them in order to apply the solution to the industrial use case.

As it can be seen in Figure 4, the system adopts a quite complex topology to be able to apply machine learning models to high frequency real time data streams. Low latency must be guaranteed when updating the models with new information coming from the data sources. Additionally, the final representation and display should be agile enough to keep all the information that is being generated in the platform.

There are two kinds of data sources: historical and streaming. In order to prove the capabilities of the platform, both of them are going to be fed by historical data for an easier implementation. Both data sources are connected, on the one hand, to a machine learning block. This module will need to divide the input stream of data into N different sub-streams to be able to process that huge amount of data in real time. The models generated in the N branches will be different since each one will use a different part of the input data. Then, the information generated by all the models will be merged into a final model in a cache memory. Since the input streaming is changing in real time, the final model stored in cache should be updated at the same rate in order to have the last update of the model ready to make predictions. It is a huge technical challenge to process Bid Data streams of data in real time, so this topology aims to guarantee low latency with high predicting performance.

The other key point of the system is the online representation of the information within the system. The module in charge of managing the data that may be displayed in front end is the Data Management Service. It will receive data from historical and streaming data sources, with an intermediate step to generate a buffer. It will also be fed with the information from the machine learning plus the information of the models stored in cache memory. As it is computationally expensive to bring historical data needed by the visualizations, a traditional database can be used to temporally storage this information. Thus, the expensive process of sending a lot of data through the networks is only made once.

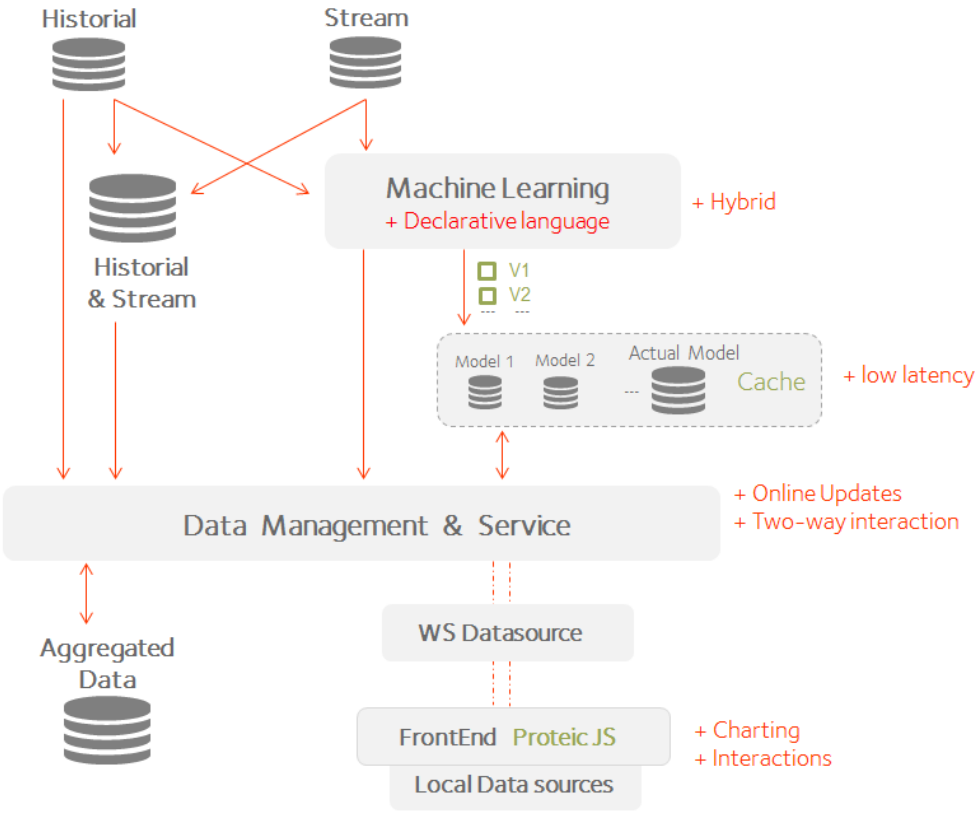


Figure 5 Scheme of the topology

Eventually, the FrontEnd module is in charge of the visualization itself including charting displays and an interaction display for the end user. The visualizations can be generated in the Data Management Service if it is reasonable to send them to the FrontEnd. Otherwise, the FrontEnd will receive the information needed to generate the visualizations from the Data management Service and generate them itself.

5 Evaluation of Objectives during Iterative Prototyping

One key aspect while iterative prototyping is to accurately evaluate the rate of achievement of the objectives. To that end, key performance indicators (KPI) are defined in deliverable D2.6. Here we provide a brief overview of some aspects associated only to the industrial use case. The KPIs defined in D2.6 will be used to evaluate the success of the use case and they should be specific, measurable, achievable, relevant and time phased.

The objectives of this project can be merged in three main groups: Scalable architecture for batch data and data streams processing, scalable online Machine Learning and real-time visualization.

The starting point is to develop a basis system that enables the implementation of the online machine learning algorithms and visual analytics demanded by the project. We have to deal with historical (sensor measurements of flatness) as well as online data (sensor measurements that arrive in real-time). For that reason, PROTEUS will design and implement a processing engine to analyse both batch data and data streams in an hybrid way.

Online Machine Learning area aims to develop algorithms that will cover all steps of real-time stream processing from pre-processing to advanced predictive and tracking analytics. PROTEUS will contribute to the maturity of this field by designing and developing a library of Scalable Online Machine Learning and Data Mining Algorithms (named SOLMA). In particular PROTEUS will not address only scalability, but also complexity. While for scalability various computational concepts will be applied to develop highly scalable streaming algorithms, for complexity we will consider distributed multivariate streams where data is potentially complex to reflect on the third element of big data which is variety. All algorithms developed will be theoretically analysed to determine their bounds.

Regarding online learning, two KPIs has been selected: accuracy and recall. Accuracy is defined as the sum of all true positives and true negatives divided by the total population. It is a simple and widely used performance indicator that provides a first insight of how the classification model is working. However, accuracy does not fit unbalance problems since it does not consider the relative error of each class, and so recall indicator is introduced.. Recall indicator is defined as the total number of true positives divided by all the condition positive, so it is independent from the total number of negative samples. Both KPIs can be calculated using both the output of PROTEUS and the output of the flatness measurement system.

Advanced visualization of data analytics in real-time, user experience and usability is still an open issue in the context of Big Data. A key challenge here is to meet the requirements in supporting real-time interaction while considering the challenges of volume, velocity and variety of Big Data. To deal with these challenges, this working package will develop innovative solutions based on incremental visual methods that allow end-users to explore both batch data and streams efficiently to make well-informed decisions in real time. Two KPIs has been defined to measure the performance of visualization tasks.

An important requirement of the visualization tool is to be able to show in real time whether the actual coil is exceeding the flatness boundary or not. Hence, the first KPI defined is the capability of real-time visual identification of the flatness status of the coil. Due to high frequency real-time requirements the visualization output must be able to update the information fast enough to capture all the information provided by the machine learning models. The KPI selected to this end is the response time of the tool. It is measured as the time difference between the end-user request and the response of the tool.

6 Conclusions

To solve the Big Data challenges identified in the Hot Strip Mill process, PROTEUS will investigate three main areas: Scalable architecture for batch data and data streams processing, online Machine Learning and real-time visualization. Each of this areas have their own hardware, software and connectivity requirements that need to be fulfilled.

The current deliverable details the requirements that ArcelorMittal needs to test the developed prototypes in realistic conditions. It is essential to define the global context and impose some requirements in order to make feasible and profitable the prototype execution. Thus, the main goal of this deliverable is to describe the correct installation of the prototypes so that the former requirements are addressed.

Hardware and software requirements are defined in Section 3. These specifications may fulfill all the computational and storage requirements, as well as the compatibility between subsystems. Each member of the consortium has defined the hardware and software required to deploy the areas it is in charge of. Hardware defined in this document fits with an real time engine architecture: distributed cluster based in master-slave model and an input streaming tool like Kafka. CentOS has been chosen as operating system for the platform due to its unique capabilities regarding speed, stability and reliability. Apart from the operating system, a list with the required software by each member of the project is attached.

The integration of the system with ArcelorMittal data sources is an important aspect of the final project. In this deliverable we define the data sources, the connectivity gateway and the communication topology of PROTEUS. There are two datasets available for this project. The first one has each coil associated to a single value for each variable, which, in many cases, represents an average of the measurements collected along the time. The second one consists of multivariate time series, such as temperature, that have been measured continuously along the coil length. It also contains the target variables (flatness) that may be employed in the construction of the prediction model. An interface called AMIII Data Interface is also defined. It will be in charge of merging the data from industrial data sources and anonymize them to preserve ArcelorMittal data confidentiality. Finally, the communication topology is clearly defined for both proposed stages. In the first one a simpler communication between process data and a Docker is proposed. Nonetheless, in the second stage a stream-ready broker as Kafka is proposed to manage the communication between AMIII Data Interface and PROTEUS Core.

Finally, KPIs defined in deliverable D2.6 has been reviewed and it has been defined how they are going to be measured in the process. Regarding online learning, two KPIs has been selected: accuracy and recall. Those KPIs are going to be measured comparting the output of PROTEUS and the one from flatness measure system. With these KPIs we can evaluate the performance of the machine learning algorithms for both balanced and unbalanced datasets. As regards visualization, the first KPI that has been defined is the capability of real-time visual identification of the flatness status of the coil. The second KPI is the response time of the system that will be measured as the time difference between an end-user request and the response of the visualization tool. This KPI will indicate if the tool is fast enough to adapt to real time, fast-evolving application.

[end of document]