D7.1 – Port Pilots Design

Project Acronym	TT
Project Title	Transforming Transport
Grant Agreement number	731932
Call and topic identifier	ICT-15-2016
Funding Scheme	Innovation Action (IA)
Project duration	30 Months [1 January 2017 – 30 June 2019]
Coordinator	Mr. Rodrigo Castiñeira (INDRA)
Website	www.transformingtransport.eu

Funded by the European Union's H2020 GA - 731932

Document fiche	
Authors:	Francisco Blanquer[NOATUM], Ignacio Cambrón [NOATUM], Iván Deosdad[NOATUM], Luisa Escamilla[FVAL], Johannes Franke [duisport], Rodrigo Garro[ORB], Pablo Giménez [FVAL], Thomas Jansen [duisport], Andre Litochevski (Software AG), Dirk Mayer [Software AG], Andreas Metzger [UDE], Gerardo Minella[ITI], Ruben Ruíz [ITI], Francisco Sánchez [ITI], Francisco Valverde [ITI]
Internal reviewers:	Arto Luoma [INFO]
Work Package:	WP7
Task:	T7.2 & T7.3
Nature:	R
Dissemination:	PU

Document History			
Version	Date	Contributor(s)	Description
0.1	11/02/2017	ITI, NOATUM	Initial document describing the expected scenarios. Initial contributions to the Big Data technologies and infrastructure sections
0.11	13/02/2017	NOATUM	Some notes corrections in the description
0.2	20/02/2017	FVAL, ITI	Contributions to Motivation and Predictive Maintenance scenario
0.3	23/02/2017	FVAL, ITI	Review on scenarios

0.4	03/03/2017	FVAL, ITI, , NOATUM, ORBITA	Additional reviews. Valencia PORT Description Final Version of Infrastructure & Technological Stack
0.5	08/03/2017	ITI, ORBITA, NOATUM	Final version of scenarios and initial Data assets
0.6	11/03/2017	FVAL, ITI	Review of Motivation and Data Assests
0.7	13/3/2017	ITI, UDE, Duisport, Software AG	Integration with the replication pilot document. Executive Summary and Section 4
0.8	17/3/2017	ITI	General review and final conclusions
0.9	28/3/2017	ITI , UDE	Overall review according to PPC meeting conclusions
1.0	30/3/2017	ITI	Final version

Keywords:	Ports, Pilots, Port Terminal, Yard Crane Scheduling, Predictive Maintenance
Abstract:	This deliverable presents the design of the pilot “Ports as Intelligent Logistics Hubs” developed in WP7 according to the Pilot Coordination Methodology Handbook (D2.1).

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. This document may contain material, which is the copyright of certain TT consortium parties, and may not be reproduced or copied without permission. All TT consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the TT consortium as a whole, nor a certain party of the TT consortium warrant that the information contained in this document is capable of use, nor that use of the information is free

from risk, and does not accept any liability for loss or damage suffered by any person using this information.

ACKNOWLEDGEMENT

This document is a deliverable of TT project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement № 731932

Table of Contents

Table of Figures

Definitions, Acronyms and Abbreviations

Acronym	Title
AIS	Automatic Identification System
ANN	Artificial Neural Network
BDVA	Big Data Value Association
CAPEX	Capital expenditure or capital expense
ECH	Empty Container Handler
ERP	Enterprise resource Planning
IMO	International Maritime Organization
IoT	Internet of Things
NMEA	National Maritime Electronics Association
OPEX	Operating expense, operating expenditure, operational expense, operational expenditure
PCS	Port Community System
RNN	Recurrent Neural Network
RS	Reach Stacker
RTG	Rubber Tyred Gantry crane
SEAMS	Smart, Energy-Efficient and Adaptive Management Platform
STS	Ship to Shore container cranes
TEU	Twenty-foot Equivalent Unit
TermT	Terminal Truck/Tractor
TOS	Terminal Operating system
TL	Task Leader
TT	Transforming Transport

Executive Summary

This deliverable presents the design of the initial pilot and the replication pilot as described in WP7 "Ports as Intelligent Logistics Hubs" of the Transforming Transport project (Tasks T7.1 and T7.2). The deliverable follows the structure and guidelines reported by D2.1 "Pilot Coordination Methodology Handbook" in order to enable the vertical integration between both pilots, and the horizontal integration with the rest of the pilots of the project.

The overall goal of this design is to assess how the application of Big Data technologies could improve and transform some of the most critical operations currently carried on ports ecosystems. Specifically the design considers the application of Big Data technologies in the area of:

(i) Design of new and more advanced cockpits to help getting the insights of day-to-day operations, making it easier to spot possible inefficiencies and also to improve the planning tasks with the introduction of prospective analyses and indicators;
(ii) Predictive maintenance of port equipment with high availability requirements or with high impact in the port operations in case of failure;
(iii) Optimization of port processes, specially those related to containers handling.

Following the TT Methodology, the initial pilot will be developed in Valenciaport to get the initial insights and results will be validated by the replication pilot developed in duisport Inland Port. This deliverable presents the requirements, objectives and use cases of both pilots and a specific roadmap to accomplish them in the expected timeframe. The Big Data technological stack and infrastructure is also presented, and it is also aligned with the BDVA reference model for supporting horizontal integration. Finally, the deliverable presents points in common among pilots and the strategy to perform successfully the replication.

1 Motivation and Ambition

Ports are safeguard places where vessels can stop and keep safe and, furthermore, they are key points in international freight transport around the world. Ports also are a hub where high quantities of goods and containers are concentrated, big volumes of data are generated and a large quantity of different organisations and different transport and handling equipment is required to work as synchronised as possible along the logistic chain. Furthermore, they provide facilities to perform repair, maintenance and supply works, as well as to load and unload operations.

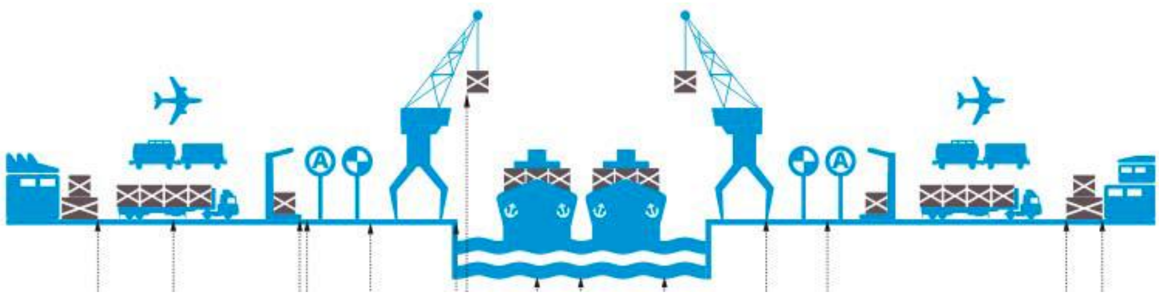


Figure 1. Multi-modal view of Ports

Shipping containers have produced a revolution in the movement of goods driving efficiency throughout the global supply chain and contributing to the market globalisation. In port operations, there are three main modes of transport: sea/river, road and rail modes. At the same time, transported goods present five kinds of typologies: containers, break bulk, ro-ro, solid bulk and liquid bulk.

The industry agrees that next revolution in container handling will be container terminal automation. However, only a small amount of container terminals around the world are today fully automated (< 3% of such terminals exist today). Unfortunately, for the rest of the manned container terminals around the world to become a fully automated terminal is a big challenge and unfeasible today using current approaches. **Transforming** a large manual container terminal into a fully automated one is estimated to require an additional investment of around €250 million. But it is not only this costly investment but also the high risk of labour conflicts with port workers, the main reasons why the quantity of automated container terminals is rather low compared to manual terminals. Achieving a progressive ***automation of port and terminal operations*** is the right answer for these container terminals, **transforming** a full manned operation mode into a semi-automatic operation mode.

However, Ports are not isolated actors, as they need to interact seamlessly with a wide variety of ecosystems in logistics and transport and ports. ***Warehouses, road hauliers, railways, port authorities, customs, border protection agencies, port terminals and vessels are required to interoperate in a connected logistics environment***. In addition, to achieve a complete and success transformation of the sector, different logistic stakeholders that manage and operate their information systems as isolated islands must cooperate.

In the ports of the future, port users, equipment and infrastructures will achieve a zero distance interaction and they will offer more sustainable transport solutions. By using Big Data Analytics together with a wide IoT network of sensors able to locate, monitor, manage and handle any transport equipment and any storage area, logistics service providers will be able to monitor and control in real time all the operations and movements of freight, equipment, vehicles and drivers that are continuously interacting on logistics nodes.

Interacting with large quantity of simultaneous transport movements around big logistics nodes raises many challenges where big data technologies can be applied to build solutions unfeasible otherwise. It is also expected that, in the near future –and thanks to already ongoing projects, container terminals, ports, rail freight stations, airports, warehouses, transport companies and dry ports will produce and store large quantities of sensors, and data, from all the logistics elements. These sensors will control and monitor up to the levels of their vitals the different port operations, and handling all the data generated will be impossible without big data solutions. The huge amount of data expected includes energy consumption, fuel levels, gas emissions, temperatures, locations, working times, idle times, status of machines, security and safety elements, freight/container handling elements, spreaders, height position, drivers' behaviour, drivers' assistance, etc.

The introduction of Big Data Analytics in Ports has potential benefits regarding the exploitation of real-time data for a better supply chain execution and improved management at different levels (See Figure 2 below):

- Transport corridor: Merging the port generated data with the data coming from other transport corridor stakeholders (sea, rail and road)
- Hub level: Merging data generated by different organisations working at the port (i.e. different port terminals, port community stakeholders, port authority, customs, …)
- Operations level: Merging data generated inside a container terminal

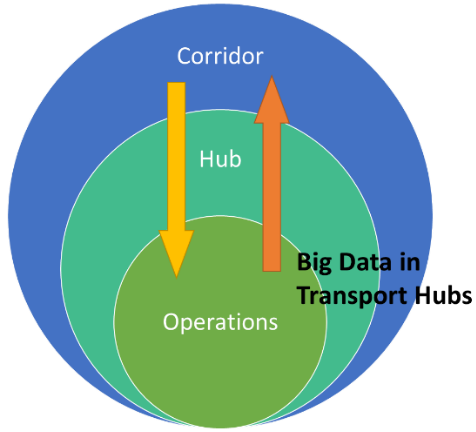


Figure 2. Big Data and Operational Levels

Big Data Analytics techniques, when properly applied, will provide the necessary tools for automating the decision process or for controlling job queues. They will allow dynamically assigning jobs for handling containers, taking into account not only operational data but also global data coming from other actors along the value chain. It is also expected that they will help to improve the planning and execution of maintenance tasks for ensuring a continuity of the operations.

Big data analytics applied to transport hubs will provide more efficient execution of orders; fast, deadlock and collision-free routes; smooth equipment use to ensure equipment longevity and minimisation of downtimes and breakdowns in next generation of manual container terminals.

Large container terminals must handle several thousands of operations daily in the most efficient a cost-effective way. To face this challenge, the first step is to achieve a **smart, dynamic and predictive system** for "job assignment" and "job queuing control" able to concentrate all the real-time data coming from ***hundreds of different sensors distributed among the container yard*** and the handling equipment, as well as from current ***terminal operating systems (TOS)***. This system needs to record and monitor all the movements taking place inside the container terminal.

Merging all data generated and compiled both at port and transport corridor levels with the data generated at the terminal will unveil new opportunities and benefits to transport stakeholders, which have yet to be explored. For instance, some data assets which could be considered are: data generated by different cargo and container terminals; data gathered by automatic gate and traffic control systems located at the port; vessel AIS data; planning, ordering and execution data managed TOS; data coming from formalities, declarations and authorisations handled in Single Window; and data generated by trucks and trains traveling along the transport corridors.

The two pilots developed in the context of Transforming Transport project have the opportunity to assess the potential of merging, processing and accessing data by different stakeholders as they need it. However, in order to really materialize this vision, big data solutions will need to provide trust, and enforce clear confidentiality, access and security rules in the provision of information.

2 Design of Initial Pilot: Valenciaport

The initial pilot in the port of Valencia will be designed and tested in a container terminal, and it will be later on extended towards the port hub and the transport corridor. Two partners are key enablers to conduct the pilot: the Port Authority of Valencia and Noatum Container Terminal Valencia.

The Port Authority of Valencia (PAV), also known as Valenciaport, is the public body responsible for running and managing three state-owned ports along an 80km stretch of the Mediterranean coast in Eastern Spain. The PAV is responsible for managing the ports of Valencia, Sagunto, and Gandia in line with the model implemented in the Spanish state owned port system, in which the port authority provides the areas and infrastructures that support port activity, whilst the private sector is responsible for carrying out operations and providing the equipment and services using the aforementioned infrastructure.

Valenciaport is Spain's leading Mediterranean port in terms of containerised commercial traffic thanks to its dynamic area of influence and an extensive network connecting to major world ports. In specific terms, Valenciaport handled over 71 million tonnes in 2016, which represents an increase of 1.71% over the previous year and constitutes the highest ever throughput figure for the Port Authority of Valencia. In turn, container traffic, which represents the largest share of its throughput, grew by 2.32% to 4.72 million TEUs, thanks to good import-export and transit figures.

Figure 3. Overview of Valencia Port and Container Terminal

Containers are characterized by a high atomization, i.e. goods could arrive at any time, from any origin and can be shipped to any destination. Obviously, if container traffic were completely chaotic, the system would be inefficient and ungovernable. For that reason, there are several mechanisms to predict demand with some anticipation: number of containers to deliver/pick up, origin and destination, and arrival and departure time at the port.

Valenciaport boasts a tightly linked Port Community, due to innovative elements such as its Quality Mark[1] and the ValenciaportPCS technology platform[2], and comprising all public and private economic agents providing services through its ports. ValenciaportPCS has become an essential tool in the modernisation of logistics management for port community companies. ValenciaportPCS handles general port information, along with tracking information from commercial and operational transactions associated with goods transport, giving users easy access to integrated logistics information. This tool helps speeding up the contracting of logistics services and increasing operational management efficiency. Currently, over 500 companies in the port community use ValenciaportPCS on a daily basis.

Figure 4. Valenciaport Control Room

While Noatum Terminal Valencia is the biggest container terminal in the Mediterranean. Noatum is a solid international company that leads the Iberian market with strategic port locations, full logistic services and a reliable and involved professional team. Noatum is present in 20 port terminals located throughout the Iberian Peninsula and Canary Islands, and it has offices in Valencia, Madrid and Barcelona. Noatum is comprised of more than 80 companies with a presence in Spain, Portugal, France, Morocco and Algeria. Noatum greatest asset is its team of over 1,600 experienced professionals, ready to provide its customers with the quality of service

---

1 http://www.valenciaport.com/en/VALENCIAPORT/ValoresCompromisos/Paginas/Calidad.aspx

2 http://www.valenciaportpcs.com/en/

they have come to expect from it. In 2016, it handled more than 2 Million TEUs, from which 40% where dedicated to the Spanish import and export market.

Large container terminals need to handle several thousands of operations daily in the most efficient a cost-effective way. To face this challenge, the first step is to achieve a smart, dynamic and predictive system for "job assignment" and "job queuing control" able to concentrate all the real-time data coming from hundreds of different sensors distributed among the container yard and the handling equipment, as well as from the current TOS. This system needs to record and monitor all the movements taking place inside the container terminal.

In a port like Valenciaport there can be more than 50,000 movements of cranes, trucks and other equipment per day to handle the 4.7 million of TEU movements per year. More than 400 container handling units (e.g. cranes, forklifts, RTGs, internally owned tractors and trailers, etc.); more than 8,000 trucks and other vehicles are visiting the port; and more than 10,000 containers per day are involved in these movements. These figures show the complexity of this environment and the opportunities that the information compiled by the sensors installed on the equipment, trucks, containers and infrastructures, as well as other sources of data could bring through big data solutions. Only considering NOATUM terminal, there are more than 180 container yard equipment equipped with sensors that produce more than 60 data items every second. This real-time monitoring of the equipment needs to be combined with the operations managed by their Terminal Operating System.

In a port there are two types of possible investments: civil and machinery. Civil investment is very rigid (it takes several years to be modified and requires a high budget) and it determines the mooring line and the storage surface, which are a fixed amortisation, whether they are used or not. Machinery is semi-rigid: while fleet size is rigid –takes one year to be modified and requires significant investments, it can be used on demand, in time intervals organized in working periods.

In Valencia, the working periods are of six hours. Consequently, it is possible to change the use of the machinery and, in turn, change the costs incurred by using it at intervals of six hours. It is important to remark that once the machinery is selected for a six hours shift, the costs associated are fixed, whether the machinery is used or not.

In automatic terminals, the overall performance is quite predictable, whereas in manual terminals, performance becomes very difficult to model and calculate. The reason is that the performance in manual terminals depends highly on human factors combined with different machinery types that operate and perform differently. For those reasons, it is difficult to predict how much time and how much resources are needed for a machine to do a specific task.

The optimization of the port system and, thereby, the optimization of any system, has four different approaches: long term, medium term, short term and real time.

- **Long term** helps to determine the civil resources and machinery to be acquired (with at least more than one year in advance). This concept resides under the umbrella of INVESTMENTS-MANAGEMENT, and comes with a lot of uncertainty because of the fact that it is also affected by general trends of the market.
- **Medium term** helps to determine the volumes of activities that will be handled during the year and estimate CAPEX and OPEX to plan actions that cope efficiently with these level of activities.
- **Short term** helps to allocate the machinery and human resources and it is done within 12 to 24 hours in advance. This concept resides under the umbrella of PLANNING. This planning activity is mainly affected by (i) the expected maritime services and (ii) the expected land accesses. At present in Valencia Port, the maritime accesses are well known beforehand, but the land accesses can only be estimated.
- Finally, **real-time** refers to the sequence of orders or OPERATIONS distributed among the resources currently available. Real time processes are affected by unexpected and –most of the time, undesired interruptions. Some factors that influence real time operations are the ones caused by weather, social conditions, machinery breakdowns and external demand (for loading/discharge and reception/delivery operations). These factors can be considered as climate environment, social factors, maintenance work (corrective) and demand profile.

All the cases considered in the pilot for big data processing will always consider the four optimization approaches trying to use the captured data for long, medium, short and real time optimisation. The objective is to make the most of the data available for the three cases considered in the port, namely:

1. The **optimization of port operations**: improve transport and logistic operations by using all the available data to create models for forecasting and algorithms for optimization.
2. The introduction of **predictive maintenance** strategies.
3. The creation of a **predictive dashboard** that gathers the entire information available (IoT platform, accesses, AIS, environmental, TOS) and provides useful indicators and their trends to end-users.

2.1 Requirements

Taking into account the current status of Valenciaport, and evaluating the main areas of improvement that can be addressed under the Transforming Transport project, the following requirements have been selected as the most relevant ones:

- **R1: Improve RTG crane scheduling** by calculating the optimum sequence of crane movements, taking into account the current disposition of the containers and the expected input/output of containers into a container block.
- **R2: Prevent unexpected failures of crane spreaders**, thanks to the application of predictive maintenance techniques. Monitoring for future failure will allow maintenance to be planned before the failure occurs, allowing better planning and extending the life of the devices.
- **R3: Collect and calculate relevant indicators to the port and terminal stakeholders**, both (i) providing real-time observation of the status of port operations and (ii) providing the necessary insights to improve the yard planning by early detection of deviations and bottlenecks and calculation of future demand.

2.2 Objectives

Associated to the requirements expressed in the previous section, the specific objectives of the pilot are the following:

- **O1: Design, implement and deploy and optimization algorithm** that provides the user with the best sequence of crane movements, taking into account (i) the input/outputs to/from the container yard, (ii) the time required per crane movement, and (iii) the current status of the containers block. This algorithm must be fast enough to update its output within the timespan that goes from the arrival of a new truck to the entrance of the terminal, and its eventual arrival at the container yard. The results of the algorithm will be integrated in the web-based dashboard developed to cope with Objective 3.
- **O2: Apply Predictive Maintenance models to cranes' spreaders**, starting with the study and deployment of a set of sensor devices to gather information regarding the spreaders of the STS cranes. The result of the implementation of new predictive maintenance models will be the design and implementation of a new maintenance strategy. The current preventive maintenance strategy will be enriched with a prognosis model that will alert of possible failures as they are detected, triggering a sequence of actions that will derive into a manual inspection and, eventually, to schedule a maintenance task.

- **O3: Develop an advanced Cockpit for better decision-making**, including a predictive decision support system that considers all the historical data available at the Valencia Port Information Systems. Registered and planned accesses, environmental data, and information coming from the TOS and SEAMS Platform will provides the basis to create useful indicators and model their next-future trends. The objective is then to provide relevant indicators to the port and terminal stakeholders in order to visualize trends and enact relevant information for the yard planning, such as behavior patterns of container arrival. This information will pave the way for improving yard utilization and maximizing labor and equipment usage.

2.3 Use Cases and Scenarios

2.3.1 Yard Crane Optimization

Usually the container yard is arranged in blocks of containers. A crane is assigned to one block, and is in charge of managing all containers going in and out the block. The block is a buffer between the land and sea parts of the port. This is the so-called yard crane and it is a crucial resource in a port. Its main objective is to move containers from and to a block in the yard and within the block.

The applicability of an optimization algorithm (or algorithms) to optimize that process is limited by the quality and uncertainty of data. Briefly speaking, there are **two possible approaches**: the first one is known as ***scheduling***, and the objective is to decide the order in which containers should be loaded/unloaded; and the second is known as ***planning*** and the objective is to optimize the assignment of resources, for instance cranes, to cope with these tasks.

Given the current availability of data in the yard crane optimization scenario, we will follow the first approach (scheduling) described above. The main reason is that to get high quality results from the optimization algorithms, a large set of confirmed orders is required. Nowadays the confirmed orders available for scheduling are reduced to 2 or 3, but using this new approach we will be able to estimate the availability of additional 2 or 3 orders. In addition the system will be ready to deal with hundreds of orders as long as a wider time-window of input/output information will be available for the optimization algorithm.

The yard crane optimization scenario consists of finding the best processing order of all containers that need to be moved in, out or within a block in a given time-window. An efficient ordering of the work instructions directly derives into an increase of RTG efficiency and utilization.

There are **four main types of productive movements** considered in this kind of problem: yard to sea, yard to land, sea to yard and land to yard containers. And an additional but unproductive movement: yard to yard. The first two cases refer to containers that are moved out from the yard and then are placed in a transfer zone (a place where the container will be picked by a truck or an automatic vehicle as a straddle carrier). The last two cases refer to situations where the containers move into the yard from sea (import) or from land (export).

The optimization objective for this problem is the minimization of a combination of the idle time of the crane and the earliness-tardiness of each container with respect to its due date or release time date (this depends on the type of container). Usually, containers that go outside from the block (yard to sea or yard to land) have a *due date* and containers that go inside the block (sea to yard or land to yard) have a *release date*. A ***due date*** is a point in time we don't want it to be violated. The reason behind it is that finishing the movement after a due date generates a penalty as the movement is "late", but finishing it before also has an earliness penalization. Conversely, a ***release date*** models the time at which the container has arrived. Therefore, we do not want movements to start before the release date. Starting a movement before this date generates an earliness penalty.

As a first step for addressing this scenario, the "one crane problem" will be studied, and depending on time and resources availability we might be able to apply the same approach to 2 or n more cranes. Whenever there is more than one crane operating simultaneously in a block we consider interference, i.e., two yard cranes cannot move freely independently of each other.

The resolution of the scheduling problem relies in a previous planning. The planning will determine the estimated time of arrival for all containers and the final (or initial) positions inside the block. This means that all input data for the yard crane scheduling algorithm is known, deterministic and fixed, and it is given before the algorithm can be applied.

The problem also considers a single block only (the size of the block is not relevant and is part of the input data), and because of that for scheduling two or more blocks it is necessary to solve two or more scheduling problems that are considered to be independent of each other.

In order to address this scenario, the following data assets are required:

- **Yard layout**: this refers to the definition of the block layout. This includes the number of rows, number of bays and number of tiers. Also, the number of I/O positions in the land and in the sea side of the block should be defined, and also their position respect to the block. They are usually set in the column 0 (sea side) and the column bays+1 (land side) of the block. In the case of the pilot, there is only one I/O position in the land.

- **Yard crane movement data:** the scheduling algorithm needs to quantify the time it takes to perform a movement. To calculate the time needed to perform a movement, it is necessary to know both (i) the initial position (where the yard crane is at the start of the scheduling), and the (ii) the velocity of the crane (in x, y and z axis). When the crane reaches a container, it has an associated *service time*, that represents the time used by the crane to attach or de-attach a container. This time can vary depending on the type of container, on the position (yard or I/O) of the operation, and also on environmental factors, like the wind. So there can be six different values affecting the service time. As a result, given an origin position and a destination position, we can calculate the time needed to go from one to the other. Alternatively, a movement matrix between any possible pairs of positions within the block can be provided. The time units provided are irrelevant for the calculus as long as they are consistent. We consider time to be integers and seconds but it could be minutes or any other approximation.
- **Container data:** The (previously calculated) planning will determine the containers that need to be scheduled (moved) within a block for a given time period. For each container, it is also necessary to define the type of movement (sea to yard, land to yard, yard to sea or yard to land). And depending on the type of container, the initial of final position, in x, y and z coordinates, and the time parameter for each container (due date or release time). This list can be provided in any order, as it is not relevant.
- **Other data:** Reshuffling movements refers to not productive movements that are necessary to reach a container or to move a container into a desired position. These movements should be predefined in the planning, and are processed in the operation as part of the input data. This means that the yard crane scheduling algorithm does not automatically calculate these reshuffling or marshalling movements.

After the application of the yard crane scheduling, the output data is basically an ordered list of the initially provided container data, indicating, for each line, the container moved, the time in which the movement is started, the time at which the movement is finished and the corresponding earliness or tardiness penalty. The order of the containers in this list is the order in which the movements should be performed in order to optimize the objective. Moreover, the input/output assignment for each container is also provided as well as the starting and final occupation time of the assigned input/output.

It is important to remark that the optimization algorithm described above requires a previous planning including: containers to be processed, due and release dates, and machine (crane and maybe other resources) assignment to the block. The Predictive Decision Support Cockpit scenario (2.3.2) will provide indicators regarding the expected status of the yard equipment in future, hence helping Terminal operators in the planning tasks. Therefore, indicators proposed in the aforementioned scenario are designed to allow the future development of planning optimization algorithms, which will feed the scheduling optimization algorithm in an iterative way.

As a complementary approach to be considered in this scenario is the application of premarshalling. In order to increase efficiency of operations inside a port yard, it is possible to reshuffle the container yard, or a part of it, in such a way that the future operations will be carried out more efficiently. Premarshalling problem (PMP) refers to such ordering and moving of containers in a yard block. Movements of containers for premarshalling order are considered as unproductive movements.

Trucks arrive to the port in a pseudo random schedule. The usual is to serve 3 to 5 trucks in each time period. It takes around 45 minutes to the truck to go from the port entrance to the block yard. Because of this, there is some time to prepare the yard for ingoing trucks. This allows to calculate the necessity of premarshalling containers when a truck passes the port door. If there is no idle time to process the premarshalling, the movement of containers will be done in the moment right before the truck arrives. This could cause time penalties like tardiness in container delivery.

The main objective of the PMP is to give certain order to the yard in the minimum quantity possible of container movements. In this case, the velocity of the crane is not relevant. The PMP considers movement of containers only in a block, the list of next trucks to be processed, the origin/destination of their containers and their due dates should be known.

In order to solve the PMP the following data is needed:

- **Yard layout**: Just the same data used in Yard Crane Scheduling Problem
- **Container data**: The (previously calculated) planning will determine the destination /origin of containers that will arrive/departure to/from the yard for a given period of time. In the case of PMP there are two different with slight differences in information needed:
- **Containers going into the yard**: In this case, the information about priority, departure date and destination position inside the yard.
- **Containers going out the yard**: In this case, the relevant information will be its initial position, and the arrival time of the truck picking it up.

- **Truck data**: the relation of each truck entering the port and a container (a truck could pick up or delivery a container) should be known. In addition, the order and arrival time of trucks to the yard is necessary for PMP calculation.

Both approaches, Scheduling and Premarshalling, will be taken into consideration to define the best optimization strategy according to the data available.

2.3.2 Predictive Maintenance for Crane's Spreader

In the context of terminal operative, quay cranes (STS, RTG, etc.) are an essential equipment for supporting transportation tasks (See Figure 5 left). Because of their importance, correct maintenance of this equipment, for instance checking for issues of wire rope or a specific component fatigue, is a key activity. However, maintenance operators usually follow manual procedures and their previous knowledge in order decide when and which part should be replaced.

One of the most critical parts of a crane is the spreader (See Figure 5 right). "Spreader" is a term used in the transport of containers in ports and in railway terminals to refer to the elevator system used to operate the containers that comply with ISO rules. It is usually referred to the telescopic frames that adjust to the length of the container (20', 30', 40' or 45') and attach to its four superior corners, closed with the help of twistlocks. It is, by nature, the most vulnerable element and the one whose breakdowns affect the most the correct functioning of the process. When a spreader is damaged, the subprocess stops, affecting completely the logistical chain: from the mooring crane, through the horizontal transport at the terminal and right up to the yard crane. Knowing beforehand if a spreader will breakdown, means a huge saving and at the same time provides interesting information for establishing a planning.

Current data from Valencia Port, shows that 50% of the maintenance tasks related with STS cranes is a consequence of a spreader breakage. Specifically, 2664 yearly alerts and an aggregated downtime of 910 hours. Maintenance costs are estimated in 172K € per year, but operative costs amount 900K € per year because when a spreader is inoperative also implies that human labour must stop. If we take into account that the yearly working hours of a spreader are 54.520, direct and indirect maintenance costs related with spreaders sum up 21€ per working hour. This figure is by far the highest in absolute terms and, consequently, is essential to improve spreader maintenance in harbour processes. These figures remark the impact of minimizing corrective maintenance tasks by predicting when they should be carried on.

Figure 5. RTG Cranes (left) and Spreader (right)

The goal of Predictive Maintenance is to evaluate the status of an equipment by means of a periodic or continuous monitoring. From the indicators monitored, the outcome of a predictive maintenance analysis is to find out the optimum scheduling, i.e. the most effective one according to downtime cost and performance losses. Predictive maintenance is a step ahead preventive maintenance, as takes into account the real condition of an equipment and not only the knowledge provided by the maintenance staff and a fixed set of rules.

The application of Predictive Maintenance is a promising approach to optimize the terminal operation by means of prognosis models to alert about possible downtime conditions related to a damaged spreader. The pilot will include all the information from the spreader signals, as well as the new sensors such us acceleration and inclination, to relate them with the breakdowns using existing historical data from the maintenance department. Thus, we can predict breakdowns before they happen as well as understanding its causes. By doing so, we can improve the spreaders design. These predictions will be result of applying a prognosis model based on Predictive Maintenance techniques.

Therefore, the first step is to declare the input variables to be monitored and processed and how they will gathered (using sensors or a specific metering device). These variables are gathered from sensoring devices in real time, ideally in the sub-second range. Therefore, data ingestion process will generate a high data throughput, which must be processed using Big Data technologies.

Actually, spreaders are monitored by means of a black box that gathers data regarding its current operative. Currently only data related with the use of the crane by the operator is stored. While this information is useful, is not enough to support an effective Predictive Maintenance approach. Current identified signals to be included are the work cycle, the 20/40 status, current status of flipper, if a twistlock is not working properly, weight (in tons) loaded by the spreader, if there is no movements and its orientation (sea-side or land-side). For this scenario, we will analyse the possibility of include data from the internal PLC provided by the cranes' spreader to get more insight about the inner workings of the operation. As this data is private, we will need approval by the manufacturer. In addition, new sensors will be deployed to get information about the physical operation of the spreader (vibrations, hits, balancing, etc.). Specifically an accelerometer and a inclinometer to understand the how the container has been loaded/unloaded by the crane and if there have been any impact. All this data will be correlated with the historical maintenance reports available, which will be stored using a normalized table.

The next step is to store all data in an efficient repository for analytics processing. Currently, data is stored using a "Big Table" approach but in relational databases (SQL Server and Oracle). In the context of the pilot, we will use a scalable data repository to store the historical and the new gathered data. Then we will apply statistical techniques to the unified view of the data to detect how the different signals influence in the spreader damage. From this analysis, we will deploy at runtime a predictive model to alert the maintenance staff about the need of replacement or fixing of the spreader. This model will be implemented following two main techniques. In a first stage, we will use Statistical Process Controls (SPC) graphs, to find out anomalies in the behaviour of several variables related with the spreader status. SPC graph will help to identify the most relevant signals from the maintenance point of view. In a second stage, we will deploy an Artificial Neural Network (ANN) to check for patterns that suggest the need of a maintenance task. This ANN will be built taking into account the findings from the SPC graphs.

2.3.3 Predictive Decision Support Cockpit for Port and Terminal Stakeholders

With the aim of planning the use and assignment of resources and take decisions about infrastructure investments, it is an essential requirement to know the real status of port processes and the expected use of resources and infrastructures.

This statement are applicable both to container terminals like NOATUM that handle the vessel loading, discharge and the container reception, storage and delivery operations and to the Port Authority in charge of ensuring the smooth movement of vessels, trucks, trains and people in the port area. The goal of this scenario is to develop a visual cockpit to provide predictive indicators to both stakeholders to improve decision support. To enable specific needs, two different views will be designed for the predictive decision support cockpit focusing in the needs of the container terminal and in the needs of the port authority respectively. However, the data sources required for designing these two cockpit views will be common. A relevant success factor in designing the cockpits is that the data handling is performed in such a way that there is not any sensitive or confidential information of one company disclosed to the other. In the pilot case, the processing of the data using the different data sources will be carried out on the big data servers hosted in ITI working under a non-disclosure agreement and the design of the cockpit guarantee this requirement of non-disclosure of sensitive or confidential information.

The port authority cockpit will be focused on the sea and the road traffic given the data sources available (AIS vessel positions, port road traffic and gate activity, terminal gate activity, environmental data and PCS operational events like loading, discharge, reception, delivery and authorisations).

Concerning the case of the container terminal cockpit, one of the main operations in the Terminal is to assign resources, such as operators or equipment, to containers, for a specific period. The main storage facility of a terminal is the container yard, a place where container carriers store temporally their containers when they arrive. The procedure can be summarized as follows: when a vessel arrives at the terminal, containers are unloaded, discharged onto trucks by quay cranes, and then unloaded by RTG cranes at the container yard. Containers are arranged in the yard as a block or stack in such a way that only the containers in top of the stack could be moved or unloaded by RTG cranes. Because of this constrain, in both discharging and loading operations, is critical to know when the container will arrive to the port and when it will be picked up by the carrier to properly assign the resources required. If there is not enough resources assigned, operation will be slower and deadlines will not be met, whereas if more resources than needed are assigned, idle time will increase. This issue could be partially solved by means of a Vehicle Booking System (VBS) to book specific time-slots and know in advance the required availability of a container. However, currently there is not yet any VBS deployed in the port of Valencia. Because of this fact, a properly planning of resources is far from be ideal.

Said that, looking to historical data and to the actual behaviour of the road traffic in the port it is feasible to foresee when containers will arrive to the port terminal. Previously to the arrival of the vessel, the terminal operator and a carrier sign a contract agreement regarding the slots required in the container yard and the storage period. This storage period also includes a "free time" range (for instance one week) in which the carrier is not charged for storing the containers in the terminal. When such "free time" expires, the terminal operator charges an additional fee to the carrier, so they usually pick up the containers in such period. However, they do not usually perform the pick-up in a progressive way but they wait until the last two-three days of its "free-time" window to move all their containers. This common way of working implies peak times of work in the terminal that strongly influence the equipment planning of the terminal. Knowing in advance these expected peak-times will improve the efficiency of the terminal.

Historical data could also provide additional insight to a proper planning of resources. Usually operations in the port are slower when there are adverse meteorological conditions. For example, strong wind affects the loading/unloading of containers using cranes. Knowing the weather forecast and taking into account the resources needed when similar conditions happen in the past, will improve planning operations. However, currently both the historical and port status required for related planning issues is spread among several data sources and applications. Due to the complexity and volume of this information, it is not possible to build manually models and indicators to guide infrastructure investments, resources management and yard planning.

The main goal of this scenario is to analyse both historical and current data available in the terminal and provide relevant indicators to help to the yard planning tasks. Current TOS systems are focused on showing the status of the Terminal. We want to move a step forward and introduce a Big Data Analytics perspective to provide insights and trends. This prediction takes into account all the historical data available, an approach only feasible by means of a scalable Big Data repository. Additionally we will include external information from the Terminal, such as weather conditions, current traffic in nearby roads or strikes. Weather conditions and current traffic datasets are available in systems operated by the Port Authority within the industrial data network and on the port community system. The whole set of data will be the input of a set of predictive models to represent the expected trend in a defined time-range. Such information will be available using user-friendly web-based dashboards with an intuitive interface and a set of indicators designed for the Terminal staff and the Port Authority respectively.

The main components of these dashboards are visual indicators such as graphs, bar charts, percentage gauges. Taking into account historical information related to previous operations, the overall idea is to calculate these indicators according to rules provided by the Port and Terminal staff. Some examples are to find out the performance and capability of both human resources and equipment or to know the resources and time required to perform a container operations. Additionally we will integrate how these indicators are influenced by other parameters, such as the environment, social factors like strikes or holidays, or specific vessel operations. These indicators will give some insights about when the yard space, equipment and cranes are more required (for instance, which are the peak hours/days) and helping to assign properly the required human resources. Predicting efficiency and traffic trends will help to define an informed decision regarding investment in human labour or equipment in real time, short, medium and long term. Alternatively, the dashboard will help to plan a better resource assignment without the need of further investment.

Figure 6. Mockups of KPIs

With the aim of addressing this scenario, the consortium will apply *Big Data* technologies to gather up all the relevant information regarding the planning operations. Currently, data spreads in heterogeneous information systems and a unified view regarding the Terminal status or the Port status is not available. As a first step, we will identify and integrate the most relevant data to perform the expected dashboard views. Next, stakeholders from the Terminal and the Port, will define the business rules that define the required indicators and the subsets of data required, as well as data access restrictions. From that definitions, we will select a predictive analysis technique to evaluate a future trend. Finally, we will show in a web-based dashboard a set of relevant indicators to help the final users in the yard planning tasks. Terminal stakeholder's could access to aggregated data and indicators generated by the port. However, data and indicators generated by a terminal must not be disclosed to other terminals as they are confidential and sensitive. Data will be extracted from the terminal operator TOS and the Port Authority Systems. Next, we select a set of relevant indicators to be considered in this initial design:

- **Arrived vessels:** Number of arrived vessels per day. It is interesting to know the accumulated arrivals and distinguish the type of vessels.

- **Vessels waiting time:** Average waiting time of vessels in the anchorage area.
- **Vessel turnaround time:** Time spent by the vessel at the port.
- **Berth occupancy ratio**: Percentage of the occupancy of the berths by vessels at the port or at one terminal. It will depend on the time berthed, the length of the vessel and the length of the quays.
- **Containers loaded/discharged:** Number of containers loaded/discharged per day at the port or at one particular terminal. It is interesting to know the accumulated containers loaded/discharged. The measurements should be in units and TEUs. It is also relevant to distinguish containers as full or empties and their specific service, export or transhipment.
- **Container movements per hour:** Number of containers loaded or discharged per hour for a given vessel.
- **Unproductive movements:** Number of container movements that do not have to be unloaded but have to be moved.
- **Container dwell time:** Average time that a container is stored in the terminal. Dwell time needs to be calculated separately for outbound (export), inbound (import) and transhipment, as well as to distinguish full and empty containers.
- **Cost per container:** The cost per container can be estimated summing up several variables: the labour cost (depending on time consumed of this labour), the equipment cost (depending also on the time), the energy cost (depending on the energy consumption), the maintenance cost (depending on the cost of repairs), the storage cost (depending on storage time) and indirect costs.
- **Idle Time:** Time that a specific equipment, such a crane or a human resource is not performing an operation in the terminal. It can be represented as a percentage and the ideal value is to be as close as possible to 0%
- **Arrival rates:** Number of trucks and/or vehicles arriving at the port or at one particular terminal each hour. It is interesting to know the arrival rates distributed along the day.
- **Departure rates:** Number of trucks and/or vehicles departing at the port or at one particular terminal each hour. It is interesting to know the departure rates distributed along the day.
- **Truck Turnaround Time (TTT):** Average time inside the port or inside one specific terminal. This indicator represents the time that a truck is waiting to retrieve a container because is not available when the truck arrives to the terminal. The goal is to minimize such time. In the calculation at port level, it is important to note that trucks can stay in the port area parked for several days. It is interesting to visualize the TTT distributed along the day.

- **TTT distribution:** this indicator represents using a bar chart how many transactions have been fulfilled in a specific time range, for instance each hour.
- **Truck waiting time:** the time a specific truck waits since it arrives to the port gates until it is attended. The waiting time can be calculated at different points:
  1. Terminal gates: Terminal arrival time - Port arrival time – estimated transit time (from port gate to terminal gate).
  2. Port gates: Port gate exit time – Terminal gate exit time – estimated transit time (from terminal gate to port gate).
- **Cost:** Overall cost of operation taking into account human resources, equipment amortization, energy costs and maintenance cost. Information should be drill down by time range and specific variables.
- **External factors:** visualization additional information related with the weather forecast (wind speed, rain probability, etc.) and the working schedule.

2.4 Data Assets

This section presents the five main data assets of this the pilot. Following the guidelines of the D2.1 Methodology Handbook, next table summarizes their main features whereas the next subsections provide a detailed description.

Name	Short Description	Available	Data Type	Link to Data ID Card
SEAMS Platform	IoT Database of the sensoring devices currently deployed in the port equipment. Provides values of different sensors each second.	Q1/2017	SQL Server DB	https://public.3.basecamp.com/p/fGC6W4qwHbYi8mjTjtRPeUQe
CATOS	Database from the TOS currently used by Noatum Terminal. Stores information about movements and storage of various types of cargo in and around the container terminal or Port	Q3/2017	Oracle 11G DB	https://public.3.basecamp.com/p/TW4c5cMi4wSvjtXVXt1qbaqH
SCADA	Database that integrates information from PLCs, OPC server and other data sources managed by Valencia Port. Currently provides information about several sensors, access gates and environmental stations.	Q2/2017	Web Services + SQL Server DB	https://public.3.basecamp.com/p/eaNLSkGaB5DVHANJNfudNcPN
Valenciaport PCS	Information system that makes available logistical information among the actors involved in port-related freight distribution.	Q2/2017	Web Services	https://public.3.basecamp.com/p/J5ZUZf5SbFVU822LLeuinygu

Name	Short Description	Available	Data Type	Link to Data ID Card
AIS	Standard maritime system for providing information about the ship to other ships and to coastal authorities automatically.	Q3/2017	Text Logs using NMEA protocol	https://public.3.basecamp.com/p/J7xx585VoxnkDG1q7Cghzgrd

2.4.1 SEAMS Platform

Noatum Container Terminal Valencia has 170 units of Container Handling Equipment (CHE) machinery, namely STS, RTG, RS, ECH and TermT, which are sending data every second into a central server that defines the SEAMS Platform. The figure below shows an example of the different machines and the current architecture. More in detail, each one of the machines has a Siemens 1200 PLC that are collecting data from the sensors like GPS, encoders, inductives, photocells, etc., using an internal polling that stores the values of all the sensors every second, with a timestamp. The database DBBlackbox stores information related to the sensors installed in each Container Handling Equipment (CHE) machines with an average number of 82 variables per machine.

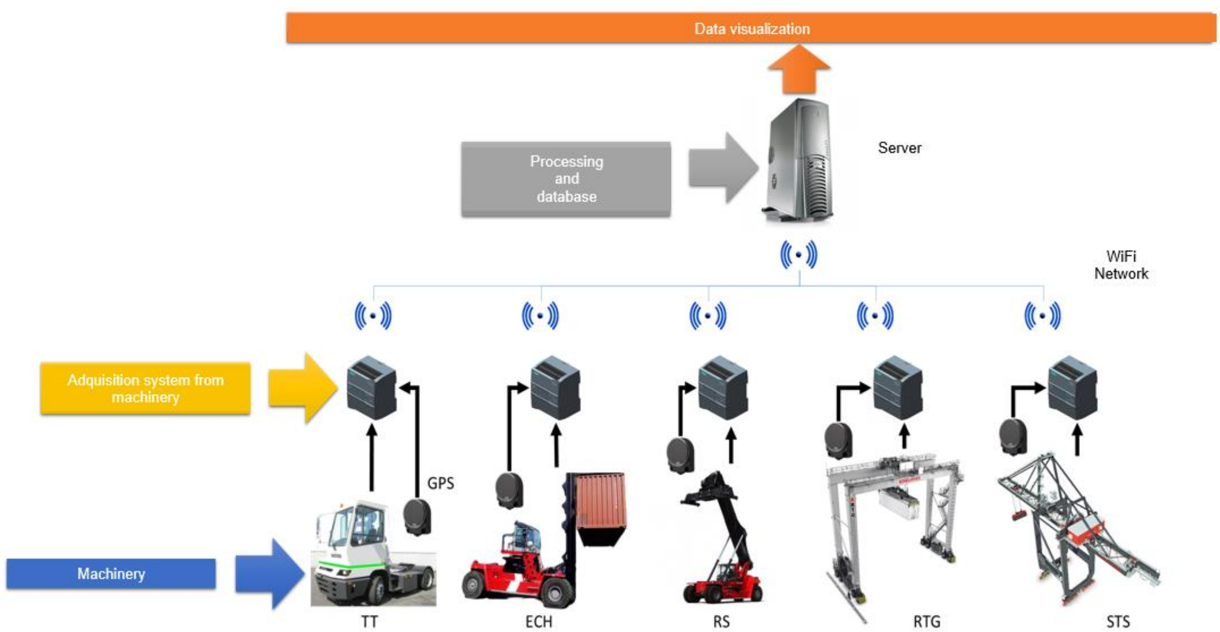


Figure 7. SEAMS Platform Data Architecture

The central server polls all the machines every second asking for updated data. If the machine polled is available, it answers with the last values of the 82 variables and mark then as “sent”. If the machines are not available in the poll moment, each machine keeps the data with timestamp every second and is stored in a SD card in the PLC. When the machine is available again, it will send CSV files with 15 minutes of information each. The total capacity of the SD Card is 6 hours of data for the 12 MB and 3 hours for the 6 MB.

Currently the central database server uses Microsoft SQL Server 2012. This server has 249 GB to store the data obtained from the machines. Up to now, all the data older than one week is being erased. The Database is located on a different server in which the software is being executed. It contains the control tables for the program as well as the final data captured from the machines.

2.4.2 CATOS

CATOS is the TOS currently used by Noatum Terminal. A Terminal Operating System, or TOS, is a key part of the supply chain and primarily aims to control the movement and storage of various types of Cargo in and around a Container terminal or Port. The system also enables a better use of assets, labour and equipment, plan current workload, and get near real-time information (up to a minute) which allows for a timelier and cost-effective decision-making.

CATOS uses other technologies such as internet, EDI processing, mobile computers, wireless LANs and Radio-frequency identification (RFID) to efficiently monitor the flow of products in, out and around the terminal. Data is either a batch synchronization with, or a real-time wireless transmission to a central database. The database can then provide useful reports about the status of goods, locations and machines in the terminal.

The objective of a TOS is to provide a set of computerized procedures to manage cargo, machines and people within the facility to enable a seamless link to efficiently and effectively manage available resources. TOS can be standalone systems, managed as a service or utilize cloud technologies.

CATOS is a comprehensive terminal operating system providing usability, inter-operability, scalability and flexibility across the entire range of terminal’s work processes and decision-making activities. Next figure shows the general workflow managed by the Noatum TOS

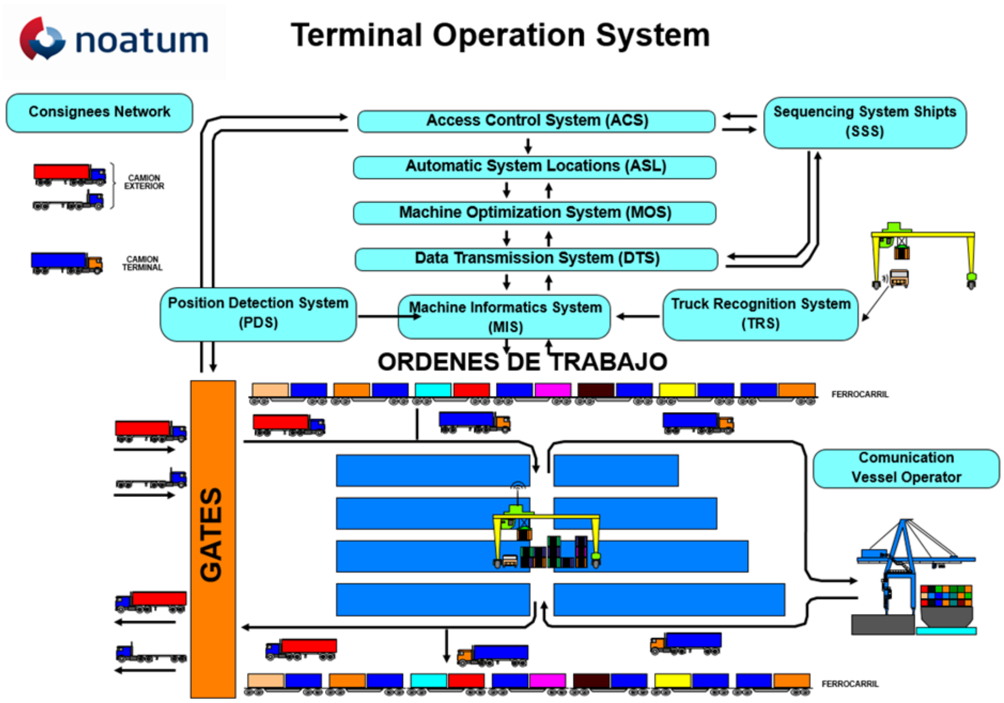


Figure 8. CATOS Data Architecture

As general rule, the terminal has two possible inputs/outputs: Land side (consignees network) and sea side (Vessel Operator). Usually the operations are vessel to vessel, vessel to gate, gate to vessel. The main activity of the process is to manage the staking buffer that is the Yard. So there are several processes:

- ACS is in charge to define the entrance of containers in the terminal (the requests and the confirmation of the arrival at the gate).
- ASL is in charge to define where to store the container
- MOS is in charge to assign the job to the equipment.
- DTS is in charge to move the order to the equipment that is confirmed and verify by the MIS with the PDS and the TRS.

The Database system installed in CATOS is ORACLE 11g r2. Current server has 80 GB stored which sums up 921 million registers allocated in 367 tables. However, the most relevant tables are the next two that store the orders related with container movements

- TB_YARD_JOB: contains the orders of the containers that actually are in the terminal.
- TB_YARDJOB_BACKUP: contains all the historic orders.

2.4.3 SCADA

SCADA system integrates the data to all the sensors deployed in the Port of Valencia: energy sensors, security, doors, lighthouses, fire hydrants, etc. Information from sensors or manual inputs is sent to PLCs (programmable logic controllers) or RTUs (remote terminal units), an then this information is resent to computers with SCADA software. SCADA software analyses and displays the data in order to help operators and other workers to decision making, and finally is stored in a database. Currently the SCADA system uses Microsoft SQL Server 2005.

The common applications between the systems in charge of processing and inserting the information in the different databases of the server are defined below:

- System of Data Capture and Exploitation (SCED): It consists of two services, server / client, that connect via OPC with the PLCs to perform read / write operations on the required PLC variables. The variables to be processed by the application are registered in the SCED database. The reading and writing of variables are done by cyclic subprocesses (hourly, daily, a specific day at a specific time ...) or by value change detection (events).
- Data Dump: Multi-threaded application consisting of several cyclic processes that obtain and process the information generated by various systems and insert it into the corresponding databases of the SQL server. Provides information about: analyzers, gauges, meteorological stations (dataloggers), classifiers, etc.
- Orbita OPC Server (OSO): Multiprocess service with different nodes, each with them relating tags. Communicate by OPC with SCADA servers, transmitting the value / status of their variables. Some of the Nodes will be: Ping Node, BBDD Integrity, Environmental Alarms, Tunnel Alarms
- Access Database: Provides information regarding the accesses to the Port. Information available are events of barriers, gauges, entry/exit steps, identification LPR/OCR, events RPM, second step MEGAPORTS, automatic customs authorities, classification of the type of vehicle.
- Environment database: Information of the environmental network of the Port Authority of Valencia. At this moment, information is stored on the sensoring systems: Meteorological Stations, Pluviometry, Sound Level Meters, Multiparametric sensors, DRAGUER Stations, Particle Monitors and Immunity Booths.

2.4.4 Valenciaport Community System

ValenciaportPCS (http://www.valenciaportpcs.net) is a Port Community System, an information system that makes available logistical information among the actors involved in port-related freight distribution. Today, more than 600 companies use daily ValenciaportPCS, with more than

4,150 users connected that involving 50.000 movements per day (400 container handling units, 8000 trucks, 10.000 containers involved).

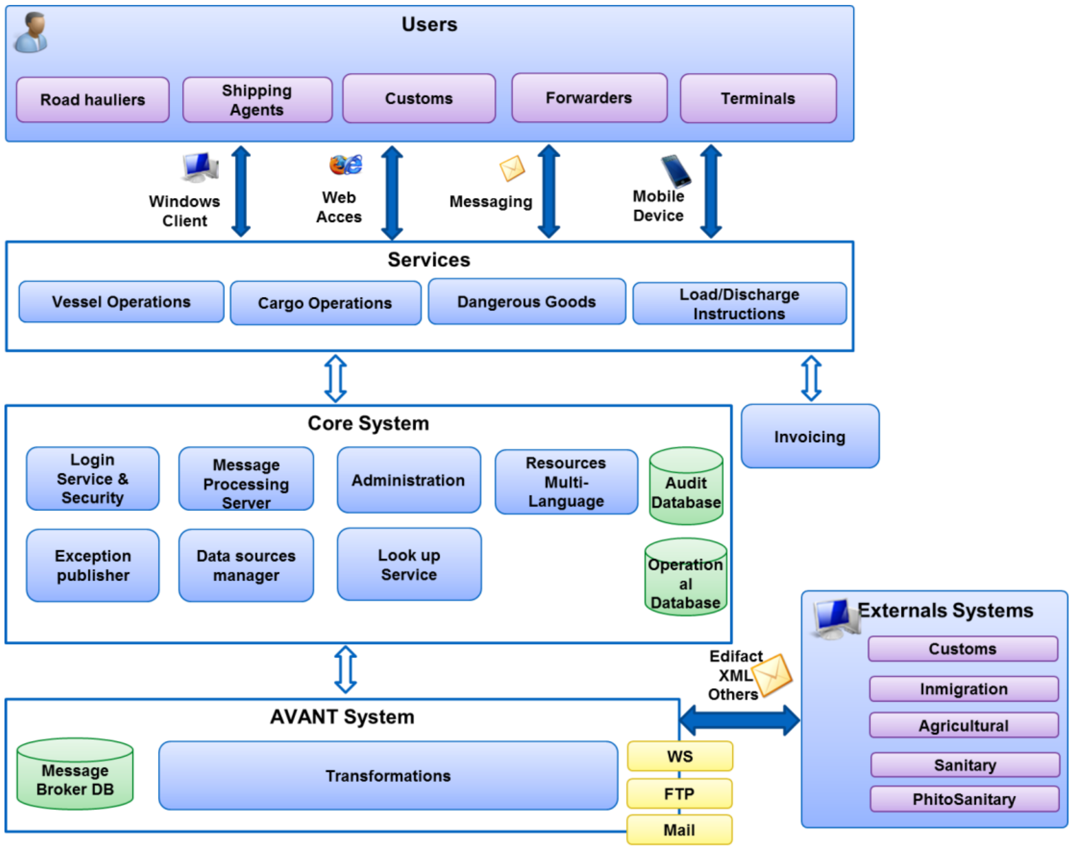


Figure 9. Architecture of ValenciaportPCS

The main actors involved in PCS include freight forwarders that act as intermediaries for importers (consignees) or exporters (consignors), terminal operators that are the interface between the port foreland and hinterland, customs, ocean carriers, inland carriers and the port authority.

The relations between these actors are very complex. The purpose of a PCS is to link effectively existing databases and management systems together, providing format translation services. The outcomes of this system is an improvement in the efficiency and quality of transactions among actors in the logistics chain and correspondingly the efficiency of the regional freight distribution system. The introduction of big data capabilities introduces opportunities to improve performance (reduction of costs and increase of reliability) of the users of a PCS.

ValenciaportPCS has established itself as one of the main e-commerce platforms in terms of number of annual transaction, with an average of more than 23,000 monthly transactions, which makes a total of approximately of 193,000 messages per day. Main information available is terminal inputs / output data, Information on loading and unloading containers from a ship, docking / undocking of a vessel, authorizations, etc.

2.4.5 AIS

Automatic identification systems (AISs) are designed to be capable of providing information about the ship to other ships and to coastal authorities automatically. IMO (International Maritime Organization) requirements established that, starting in 2004, all SOLAS (Safety-of-Life-At-Sea) international traffic ships of gross registered tonnage equal to or greater than 300 tons must have built-in AIS (Automatic Identification System) mobile stations.

The AIS system provides a means to increase the safety of maritime traffic by way of data transmission flow. This system stores relevant navigation data that will definitely contribute to clarify the cause of certain problems and, consequently, contribute to their future resolution.

The coastal AIS base stations receive, process and distribute the ship tracking information that the boats transmit through their AIS terminals under the requirements laid down by the International Maritime Organization (IMO). AIS data is encapsulated within the Marine Standard NMEA protocol (see Figure 10), which defines a standard for communication between wired electronic ship devices.

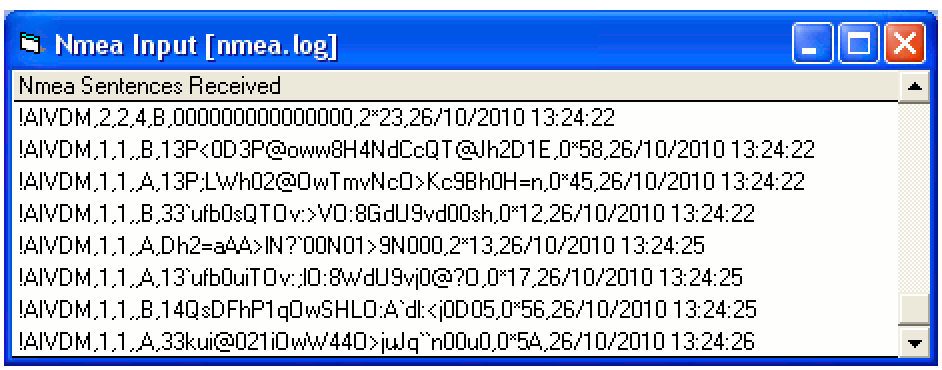


Figure 10. Example of NMEA protocol in AIS

The information is transmitted continuously (the minimum nominal interval is two seconds, increasing according to the lower speed or the manoeuvres of the ship), and the data is updated automatically without any action on the part of the user. The information available is:

- Static information: indicates the basic data of a ship: name, MMSI, call sign, IMO number, dimensions and type of vessel. This information is recorded when the device is installed and only changes if any main feature of the vessel is modified. AIS sends this information every six minutes.
- Dynamic information: shows the data related to the position of the boat required for navigation: position, course, speed and navigational status. It is continuously sent.
- Travel information: These data are entered manually and are required in VTS traffic control devices: dangerous cargo, destination, ETA and route plan. AIS sends this information every six minutes.

2.5 Big Data Technology, Techniques and Algorithms

Next figure summarizes the envisioned architecture to support the pilot. We have defined four functional layers in our *Big Data* processing pipeline and a set of required components to support the pilot. First, we the defined layers and the functional role of each component and, in the next subsections, we describe a set of candidate technologies and algorithms to implement the described architecture.

1. **Data Ingestion**: This block describes the components related on how data will be gathered for the pilot. First, we have identified a Relational ETL component to process the data currently stored in relational engines (Oracle and SQL Server). This data corresponds to information from the TOS (CATOS), gates, environmental data and processed data from sensor devices. Data Ingestion also requires to gather data from sensing devices or industrial equipment. The volume of this data is small but as it is continuously generated (in the order of milliseconds), ingestion requires an extremely fast process. For that reason, we have included a Message Broker component to centralize the data sent from several data sources and distribute it accordingly to the rest of the functional layers. Finally, we have included the Inter-IOT middleware that is currently in development in the context of a European project. When available, we will use this middleware to gather data using a set of data web services.
2. **Data Storage:** Currently historical data is neither aggregated nor stored for a long period because relational databases are not able to scale properly. The main component of this layer is a scalable repository, typically based on a NO-SQL technology, in order to address the storage of all the information generated in the pilot. We also include a distributed file system to stored bulk data required for historical analysis. Both data storage components will be accessible to the rest of the architecture by means of an API using a SQL-like interface.

3. **Data Analytics:** This layer defines the analytics engines made up of the algorithms to implement in the pilot. Optimization algorithms support the first scenario, whereas predictive algorithms support the second scenario and some indicators defined in the third scenario. The corresponding subsections introduce further details about these algorithms. These algorithms are trained or implemented according to the historical data from the Data Storage Layer and, optionally, use the real-time data from the message broker at execution time. Our architecture includes a *Complex Event Processor* to generate alerts according to pre-defined rules that take as inputs events and data gathered in real-time.
4. **Data Visualization:** This layer provides the data to end-users and the analytic results. The data analytics layer runs the implemented algorithms aggregating the data in-memory and, then, providing results to visualize such as the expected trend of an indicator or an optimized sequence of orders. On the other hand, historical data is directly retrieved from the Data Storage using a SQL-like API. These results are represented by a specific KPI or user-friendly chart. All visualizations and KPIs are aggregated in a web-based cockpit as a front-end.

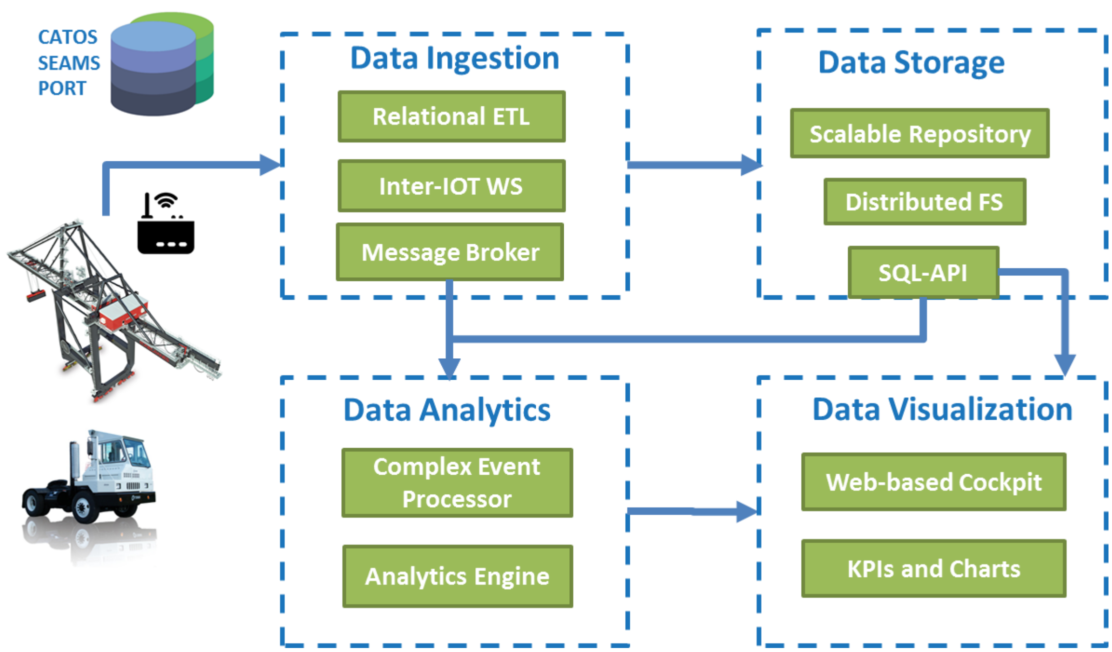


Figure 11. Valenciaport Pilot Architecture

One of the main goals of the pilot is to predict relevant information for the Port stakeholder from the current (and future) available measurements. There are several approaches to perform predictive analytics and in the context of the pilot we have selected two:

- Trend-based evaluation: the goal of this approach is to discover deviation ranges from what it is considered a regular operative. This approach requires a wide network of sensors and devices to precisely define the status of the equipment to monitor. Additionally, conditions and variables related to a failure must be clearly identified. To gain this knowledge, specific conditions when a failure happened should be recorded. Once these conditions have been identified, their current values are monitored during the equipment lifecycle to predict an imminent failure or to show deterioration evidence.
- Advanced Analytics: it is arguably the most popular approach, as no previous knowledge about the equipment is needed to perform an estimation. That knowledge base is replaced by historical information (usually time-series data), which includes the signals that could potentially lead to a failure, or statistical measurements (average, standard deviation, etc.). All this data is the input to pattern matching, regression models or neural networks. The latter is a very popular approach because they are self-adaptive and they do not make assumptions about the problem under study. A drawback of using neural networks is that usually the reasoning behind each decision is not straightforward. Additionally, they required high-quality and representative data to provide and effective prediction. As a rule of thumb, the better the quality and the higher the data volume available, the most accurate the prediction.

To implement both approaches, we have chosen two statistical-based techniques that complement each other: Statistical Process Control charts and Artificial Neural Networks. In order to apply these techniques, we will need a historical collection of data to learn what has happened during a given period of time (the longer this period, the better the models, hence the need of a "big data" approach). This collection will contain measurements from the deployed sensors, information of the performed maintenance actions, information of the failures, and any other useful information. All these data must be timestamped, so that the models can find patterns and correlations between co-occurring events. Once the models are trained from these historical collections, they can be applied to analyse real-time events. To this extent, these events will need to be characterised with similar features to the considered in the initial learning phase (i.e. same type of measurements, information of the maintenance actions, failures, etc.).

Next subsections describe in more detail the technologies and analytical techniques briefly introduced in this section.

2.5.1 INTER-IoT Middleware

A goal of the INTER-IoT H2020 project is to facilitate interoperability of heterogeneous IoT platforms already in place mainly in Port Transport & Logistics sector (i.e. Valencia Port and NOATUM) and other existing components. This interoperability comprises data offered by sensors from Valencia Port NOATUM terminal as well as sensors available on third party equipment (i.e. reefer containers) and vehicles (i.e. external trucks picking up and delivering containers). From a technical perspective the result will be a data broker to retrieve data from all these systems in a seamlessly way. All this data will be made available using a middleware with two main interfaces: a set of Web Services, for retrieving mainly historical data from the integrated data sources, and a message broker following the publish/subscribe approach. When this middleware will be ready for production (expected by the end of 2017), it would be included in the context of the pilot.

2.5.2 Apache Kafka / Mosquitto

In IoT environments the “publish-subscribe” messaging pattern is a common approach to communicate data among devices and services. Briefly, a specific device publishes data messages to a topic. Then, a consumer device or service subscribes to a topic and receives messages, as they are available. Following this communication pattern, producers and consumer are completely decoupled and communicate each other without knowing specific technical details about who generated the data. Both Apache Kafka and Mosquitto are message brokers that implement the aforementioned pub-sub pattern.

In the one hand, Kafka is a message broker completely integrated in the Hadoop technological stack, specifically with HBase and Spark. In the pilot, we will use Kafka for real-time streaming of data an event coming from terminal equipment to be stored into the data cluster. Another advantage is the scalability because Kafka supports distributed deployment.

On the other hand, Mosquitto implements the MQTT protocol and it was designed as a lightweight approach to the publish/subscribe pattern. Mosquitto has a small code implementation and requires less network bandwidth and computational power than similar approaches such as Kafka. Because of these features is an ideal technology for embedded hardware, PLCs and small devices. If hardware limitations arise in the pilot, we will use Mosquitto as a message broker replacement for Kafka. The selection of these technologies is also justified by the architecture exposed by the INTERIoT middleware in order to ensure interface compatibility.

2.5.3 Apache Hadoop HDFS

The Hadoop Distributed File System is a standard choice for managing files in a distributed environment, such as a Big Data cluster. HDFS is not a physical storage system itself, but an intermediate service layer to distribute files physically among different disk and servers and, then, supporting high volumes and fast information retrieval. Instead of supporting constant read-writes, which are very common in other scenarios, HDFS is specifically optimized for reading data from multiple physical sources. This technology has been selected for the pilot because, we expect huge datasets that will be written once, but constantly read by the developed algorithms.

2.5.4 Cassandra

Cassandra is one of the most popular NO-SQL databases to set up a data cluster with low-cost hardware but without compromising performance and volume. The physical model of Cassandra resembles the “Big Table” approach initially defined by Google: a unique table with a random set of columns per each row. As it is not constrained by a specific schema, scalability along multiple nodes is easy to address both at table and row level. Cassandra supports fault-tolerance by means of replicating data among different nodes and promoting transparently any node to master. This distributed nature ensures the scalability requirements demanded by the pilot, even if a node does not reside in the same cluster.

IoT data gathered from current infrastructure is already stored in a relational disperse way: a small set of tables indexed with a pair of id-timestamp values with several *nullable* columns. This specific schema translates to the Cassandra columnar approach in an optimal way. The main structure to store data is a specific file format called SSTable. One interesting feature is that SSTables are immutable: performing inserts or updates in the database does not imply to overwrite or delete data. Cassandra writes newer version of the rows, which could have a different set of columns, identified by a timestamp, then, older rows are deleted periodically using a compaction process. This approach defines Cassandra as an effective database to deal with historical data in a specific range.

In summary, Cassandra meets perfectly the data requirements expected by the pilot in terms of both scalability and availability.

2.5.5 Hive

Hive is a data-warehouse technology in top of Hadoop that uses underlying technologies such as HDFS, MapReduce and HBase. Instead of providing a Map Reduce oriented API, Hive provides HiveQL, a query language that resembles SQL and includes operators to perform joins, unions and aggregations. Specifically designed for batch works, HiveQL queries are translated to a set of mappers and reducers able to process all data available in the cluster. Additionally, a table represents the information stored in the HDFS or another Hadoop data repository, hence providing a relational-like view of the data. Because of these features, Hive is commonly used in data mining and analytics processes that not require a real-time response.

2.5.6 Spark

Spark is an open-source framework to implement data processing applications using the main memory as intermediate storage, then reducing disk I/O and improving performance. When data to be processed fits in memory, is a processing approach several times faster than using equivalent Map/Reduce technologies. Its role in the pilot is to act as an analytics engine for implementing the different algorithms. One advantage is that Spark manages data from different data sources in a unified way and supports data ingestion either in real-time or in batch. As data will come from different information systems, the use of Spark guarantees a seamlessly approach for data processing.

Spark extends its core functionality with additional libraries. One of the most relevant for the pilot is Spark Mlib: a set of ready-to-use machine learning algorithms in top of Spark. MLib provides a wide range of functions related with clustering, regression or dimensionality reduction among others. We expect to implement the required ML algorithms, specifically the ones to address the predictive maintenance scenario, using this library as technological foundation.

2.5.7 WSO2 middleware stack

WSO2 provides an open-source service-oriented architecture (SOA) middleware for enterprise applications. This middleware is made up of independent components, such as an Enterprise Service Bus, an API Manager or an Identity Server, so it can be adapted for developing a tailored solution. The entire WSO2 middleware stack works seamlessly across private, public, WSO2 managed and hybrid clouds, as well as on premise.

WSO2 *Data Analytics Server* (DAS) is a platform to aggregate, analyse and monitor data flows in that happens in a business process. DAS receive a continuous stream of events from different datasource and process them using real-time analytics. This technology also includes a *Complex Event Processor* (formerly known as WSO2 CEP) that enables to define alerts from events that happen in real time. Currently WSO2 DAS is considered one of the most efficient open-source solutions in this domain: a single DAS Server node with 4 cores and 4GB of RAM is able to process up to 100K event per second or up to 6M if the execution engine itself generates events. The execution engine (Siddhi) provides a SQL-like language to specify conditions and check for event occurrences, aggregate several event flows and detect logical or temporal patterns. Another advantage is its integration capabilities with event flows with different formats (XML, JSON) or following different transmission protocols (HTTP Rest, JMS, SOAP, Kafka, MQTT). In the context of the pilot, we will use WSO2 DAS in order to implement rules from data gathered from the IoT devices deployed in the yard. Results of these rules will enable indicators and trends to be shown in the Dashboard.

2.5.8 Tensorflow

TensorFlow is an open source software library for numerical computation based on data flow graphs developed by Google. Its flexible architecture allows to deploy computation to one or more CPUs or GPUs. TensorFlow was originally developed for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well. We will explore this technology to implement the predictive maintenance approach using ANN or RNN.

2.5.9 Optimization Algorithms

In the last 10 years, port scheduling problems have received more and more attention in operations research literature. However, the development of algorithms (outside mathematical models) is yet an incipient matter. Next we briefly present some works in this area

The Yard Crane Scheduling problem is studied in (Zyngiridis, 2005). In this thesis, the author proposes an integer linear programming solution to the problem to study the effect of adding a second crane to a block. The same year (W.C. Ng; K.L. Mak, 2005) study the yard crane scheduling for loading and unloading containers with different ready times with the objective of minimising the sum of waiting times. They propose a branch and bound algorithm and some lower and upper bounds.

A complete literature review on yard crane scheduling is carried out by (Özge Nalan Alp;Hayri Baraçlı, 2009). In (Iris F. A. Vis; Kees Jan Roodbergen, 2009), the authors study the problem by reformulating it as an asymmetric Steiner traveling salesman problem which can be solved optimally by means of combination of bipartite networks and dynamic programming.

Linear programming is used to obtain optimal solutions of small problems in (Iris F. A. Vis; Kees Jan Roodbergen, 2009), where the authors demonstrate how adding a second crane to a yard decreases the time needed to process containers by a 30%.

In (Iris F. A. Vis; Hector J. Carlo, 2010), the problem of sequencing two automated cranes is studied. This paper proposes a mathematical model to obtain a lower bound for the makespan and a simulated annealing heuristic method to solve large instances.

Finally, the article (Héctor J. Carlo; Iris F.A. Vis; Kees Jan Roodbergen, 2014) presents a complete overview of storage yard operations and proposes a classification scheme of scientific papers.

We could conclude that these works are focused on finding mathematical models and dynamic or linear programming. Those powerful tools are usually able to find optimal solutions to the studied problems. The main drawback of these techniques is the amount of time needed to solve large problems, as they are usually bound to solve small instance problems.

In order to fulfil the real-life requirements of the port pilot, we need to obtain good scheduling solutions in real time (less than 60 seconds). This means that we should apply fast algorithms and that a good solution is preferred to an optimal solution if it can accomplish with the time restriction. Our approach will develop a simple and fast heuristic to solve the first scenario, maybe based on assignation rules. In addition, we will need the development and test of one or more assignation rules.

2.5.10 Anomaly detection: Statistical Process Control

Statistical Process Control (SPC) charts analyse the performance of one or several variables (measures) over time by plotting their data points with respect to upper and lower control limits. Such control limits are established according to historical data, when available, or to a hypothesis regarding what is considered as normal behaviour. Therefore, it is possible to separate common caused variation from an anomaly behaviour. One advantage of using SPC is that they could be generalized in order to show how different variables influence each other. In the context of Predictive Maintenance, the application of SPC charts lead to the discovery of behaviour patterns linked to a breakdown. SPC charts are an interesting initial analysis to find out which potential variables to monitor, since they detect anomalies which usually are associated with undesired equipment behaviour. These anomalies and their causes, i.e. the current values of the variables, are particularly useful information for the maintenance staff.

In general, statistical control applies statistical inference techniques to confirm an initial hypothesis, which considers that the process of interest is under control and working as expected. These techniques take as a reference a probability distribution, which depends on the statistic to evaluate, to decide how probable a given measurement or its derived statistics are.

Depending on the accuracy requirements given by the stakeholders, we can establish the SPCs limits according to the probability values considered as suitable. These boundaries will decide whether a hypothesis should be rejected.

SPCs learn from historical measurements taken from normal situations. Usually, the longer these measured periods are, the more reliable the SPCs will be. However, as a side effect, there will also be more anomalous measurements, which need to be cleaned before the estimation. In addition, SPCs should be re-learned as the conditions of the modelled process change.

One-dimensional SPCs (Figure 12) try to predict significant increments or decrements of a measure's position (i.e. its working levels) or variability (i.e. its dispersion) from a unique signal from one sensor. This natural variability will be influenced by external factors, which, in turn, should be modelled with their specific one-dimensional SPCs.

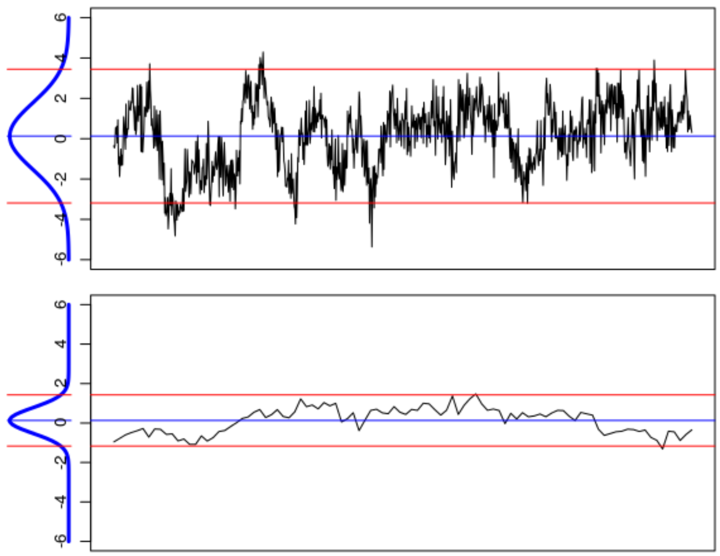


Figure 12. One-dimensional Statistical Process Control charts.

On the other hand, multi-dimensional SPCs (Figure 13) try to capture the relationships between several co-occurring sensor signals. In this way, they are able to detect when their relationship structure breaks down or when their representing point on the multi-dimensional space is far from the rest of points. In order to evaluate these multi-dimensional values, we can employ Hotelling's T-squared distribution, which can represent them and their relationships via a unique one-dimensional measure.

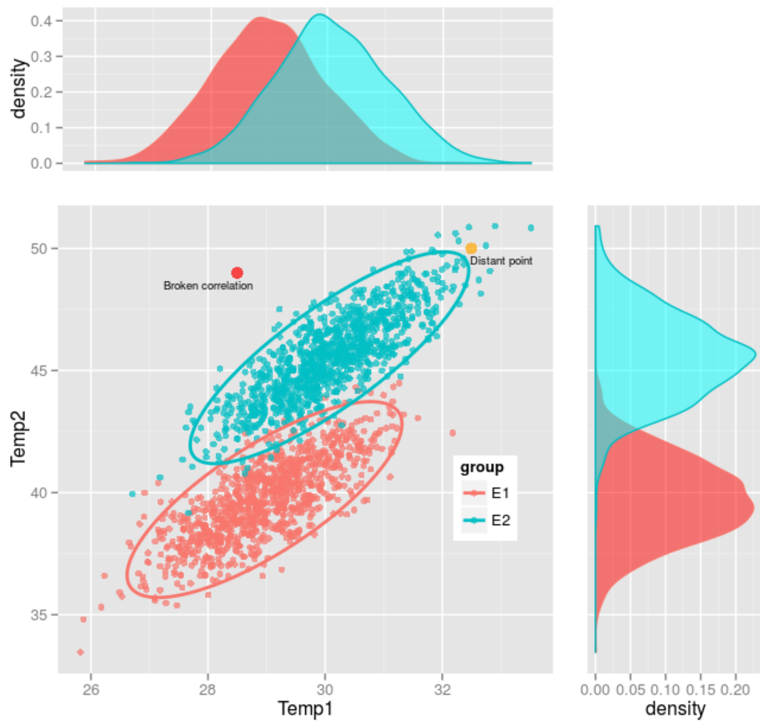


Figure 13. Multi-dimensional Statistical Process Control charts.

2.5.11 Predictive maintenance: Artificial Neural Networks

Artificial Neural Networks (ANNs) are a data reasoning approach that resembles how the human brain works. Following that approach, a neural network is made up of processing nodes, each one with a small subset of the whole knowledge and a set of rules to process it. Then, the network is organized in tiers creating interconnections between all the available nodes. The first tier receives the input information, for instance the whole set of measurements by several sensors, processes the information according to the implemented rules, and sends the output to a node from another tier. The last tier is the output layer from which the one or several answers are delivered. Initially a neural network is trained using a large amount of input data and their expected output. Then the network is deployed using the actual data to provide novel knowledge.

In the context of this project, we propose to use ANNs to predict the remaining time until the next failure. In order to maximize their accuracy, we will pre-process the data, both in the case of the historical collections employed in the training phase, and in the case of the real-time data to be analysed in the production phase. This pre-processing stage would consider techniques like:

- Temporary alignment and interpolation of missing data. This step deals with different sampling frequencies, sensor failures, etc. in order to ease the co-occurring events pattern analysis.
- Aggregated statistics computation (e.g. average, standard deviation, percentiles, etc.), to extract information from the measure's probability distributions.
- Alternative representations (e.g. variable projections, variable selections, Gaussian Mixture Models with Maximum Likelihood Estimation, etc.), to identify significant patterns for each process and its aggregated statistics by means of more suitable projections.
- Information from previous instants, in order to find temporary patterns.
- Additional information (SPCs, external data sources, time until next failure, etc.).

As a result, raw data will be transformed into feature vectors describing what is happening at a given moment and when will the next failure be. During the training phase, ANNs will take each feature vector from the pre-processed historical data collection to learn to predict the remaining time until the next failure. Once trained, the ANN models will be able to analyse real-time data from the sensors, pre-processed in the same way as before-described, in order to predict when the next failure will occur.

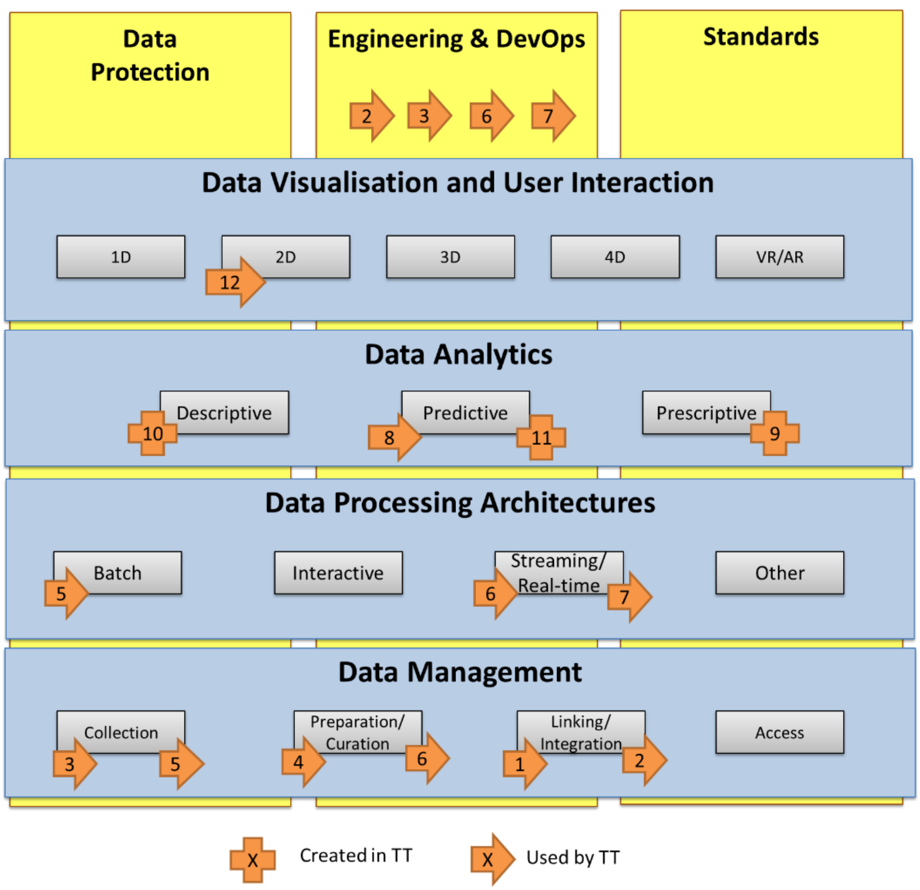
In addition, we propose to explore recurrent neural networks (RNNs), which can better capture dynamic temporal behaviour by implementing an internal memory. In particular, long short-term memory (LSTM) are deep learning RNNs that can handle signals with a mix of low and high frequency components.

2.5.12 Javascript chart libraries

The management of Big Data requires both a highly interactive interface to navigate and understand the provided information. In order to build user-friendly and responsive visualizations of the proposed KPIs and analytic results, we will use Javascript-based libraries according to the pilot needs. Two candidate libraries are Amcharts, which provide a set of ready-to-use charts for data visualization, and D3.js, which provides a powerful framework to build our own visualizations. Then, the implemented charts/widgets will be deployed as a component of a ReactJS or AngularJS application that will connect with the services of the Analytics and Storage Layer.

2.6 Positioning of Pilot Solutions in BDVA Reference Model

Next figure shows how the technologies and algorithms explained in the previous section (2.5) align with the BDVA Reference Model. Previously, each technology has a tag according to the last number of their corresponding subsection (2.5.x). Some technologies have several roles in the proposed referenced model. For instance, we will use Spark framework for processing both raw and real-time data. Regarding Hive, we will use as a collection technology to store historical data which will be processed as batch jobs. Regarding Data protection, no specific technologies have been defined as data will be confidential and only available using VPN. Contributions to standards are not foreseen at this point of the project.



1. INTER-IoT
2. Kafka / Mosquitto
3. Hadoop
4. Cassandra
5. Hive
6. Spark
7. WSO2
8. Tensorflow
9. Optimization Algorithm
10. SPCs
11. ANN
12. JS libraries

Figure 14: Big Data Value Reference Model

2.7 Big Data Infrastructure

Currently Valencia Port Terminal gathers IoT data from different equipment involved in terminal infrastructure. This IoT infrastructure was initially developed in the context of the SEA Terminal project, and is based on a set of “black boxes” connected to the terminal equipment to be sensorized, namely RTGs, STSs, TTs, RS and ECHs. All data from the corresponding black boxes are transferred to a central server using a WLAN deployed at the terminal. Additionally all black boxes are equipped with a GPS transmitter, therefore coordinates are transmitted to the central server. Then, the Smart, Energy-Efficient and Adaptive Management Platform (SEAMS) platform aggregates all received data. The SEAMS system gathers in real-time information about the status of the container terminal operations to an end-user. In the context of the pilot this data will be transferred to the data storage provided by ITI for further analysis.

ITI will provide the computing infrastructure for Big Data storage and processing required for the three scenarios deployed in the pilot. This infrastructure is made up of two clusters with different capabilities. First we will use a data storage cluster, which currently provides, 400+ CPUs, 4+ TB RAM and more than 100TB available for storage. In the context of the pilot, we will use this cluster as main storage repository and for deploying database engines. Additionally, we will use an analytics cluster exclusively designed to perform high-intensive computing tasks. This cluster provides 48 CPUs, 168 GB RAM and a storage layer based on a SSD disks array to dramatically increase I/O throughput. Both clusters are interconnected using a 10GB Ethernet network.

The physical resources of both clusters are virtualized, so computational nodes will be added to the pilot by demand. Specific Big Data services and tools are managed using the distribution Cloudera for Hadoop Systems. This distribution simplifies the deployment of several of the Hadoop-related technologies introduce in section 2.5, namely HDFS, Hive, Sqoop and Spark.

To ensure the scalability and reliability required by the pilot, the infrastructure deployment follows the Lambda architecture. This architecture consists of three layers as depicted in Figure 15:

- **Batch Layer**: This layer will process historical data available from Valencia Port information systems and raw data coming from the current IoT infrastructure. This layer will be deployed in the storage cluster in order to support the data requirements of the pilot. This layer includes two components, a scalable data storage layer to store all data available and a set of database engines for data management. These engines will use the data storage on demand and will provide data access to the other layers.
- **Real-time Layer**: this layer will process the sensorized data coming from the port equipment (TT, RTGs, trucks, etc.) in “near” real-time. This layer will be deployed in the analytics cluster. The components to deploy in this layer are the optimization algorithms to support the Yard Crane Scheduling scenario, a set of machine learning algorithms to support the predictive maintenance tasks and a set of programming rules to enact

relevant indicators or yard planning and monitoring activities. Each component will define a computational master node and a set of slave nodes according to their specific computing requirements. The master node will connect to the db engines provided by the batch layer and distribute the data accordingly for processing.

- **Server layer**: This layer acts as a middleware to access to both the batch and real-time layers in a seamlessly way. Server layer will provide a unified query mechanism to provide structured data to the pilot applications to be developed, specifically the analytics algorithms. Also this layer will include a Web server for supporting the deployment of the dashboard.

By default, each layer will be supported by a predefined number of computing nodes or virtual instances, according to their expected requirements in terms of CPU, memory and disk storage. However, this initial configuration is not fixed, so resources will be elastically assigned at execution time.

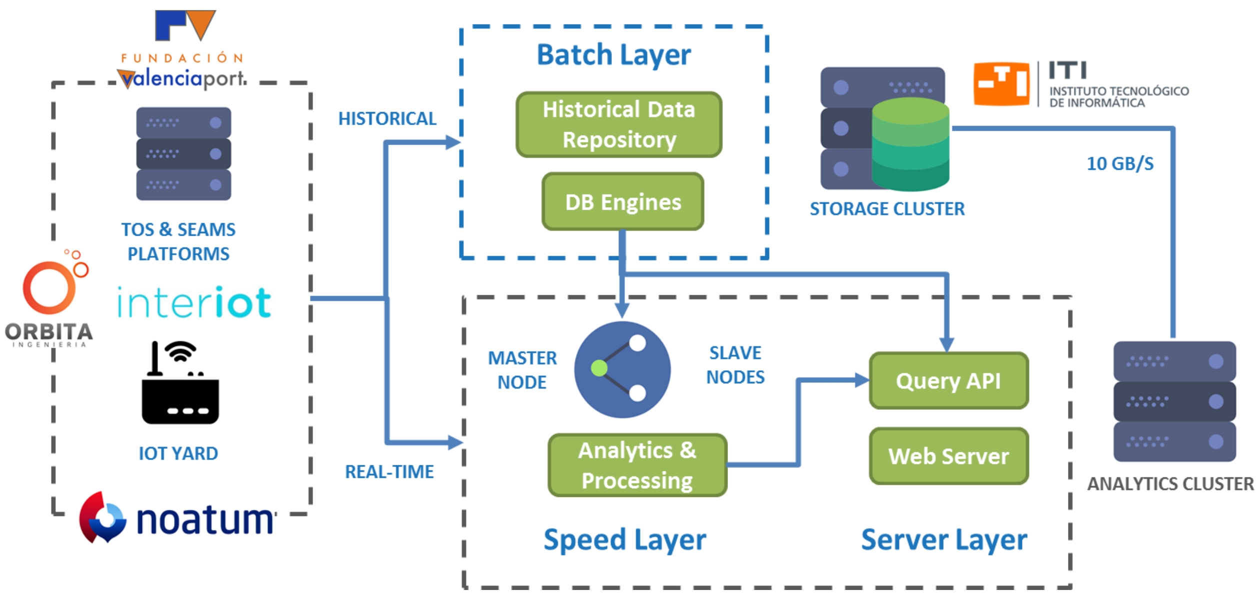


Figure 15. Valencia Port Pilot Infrastructure

NOATUM and Valencia Port Authority will transfer real data from their datasets and information systems (CATOS, SEAMS) to the ITI infrastructure for applying the analytics required by the three scenarios. With the aim of meeting the network bandwidth requirements, the infrastructure will use a fibre connection (which provides up to 4x200 Mb/s in the case of Valencia Port Authority) between the Big Data cluster deployed in the ITI and their information systems. Historical data will be transferred using a data dump from current relational databases whereas, in later stages of the pilot, current data will be transferred in "near" real-time. If this network is not available or bandwidth is not enough, periodic dumps of the data will be transferred to the batch layer. Finally, to be compliant with the privacy data requirements, partners involved in the pilot will sign a non-disclosure agreement in order to access the infrastructure.

2.8 Roadmap

We have established three objectives, each one related with a specific use case, which will be developed incrementally according to the TT Methodology. Next tables detail the roadmap for each objective according to defined stages, S1 (Technology Validation), S2 (Large-scale experimentation and demonstration) and S3 (In-situ trials).

Table 1: Roadmap for delivering pilot results (along TT stages) – Objective 1: "Yard Crane Optimization"

Stage	Date	Features / Objectives	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1	M6	Analysis of the current data sources to enact relevant data for the optimization algorithms. Report about candidate techniques for solving optimization problems	Historical data collected from TOS	Not needed. Data will be exported from relational databases	Small scale to check to check correctness of data and initial results. If not enough data is available synthetic data from literature will be used
S2	M12	Validation of the optimization results against current operation processes deployed in Valenciaport	Integration with TOS data	Analytics cluster to test the first release of the optimization algorithm	Medium scale. If not enough data is available synthetic data from literature will be used
S3	M24	Deployment of the optimization algorithms using real data from current operations	Real-time data from the TOS systems	Analytics cluster connected to Valenciaport data assets	Large scale (actual operations)

Table 2: Roadmap for delivering pilot results (along TT stages) – Objective 2: "Predictive Maintenance for Crane's Spreaders"

Stage	Date	Features / Objectives	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1	M6	Analysis and integration of historical data and data sources related with the spreader operative. Deployment of new sensoring devices and gathering of new signals from the spreader	Historical data and backups from current data on a weekly basis	Data Storage cluster for importing data from relational databases.	Medium scale (approx. 150GB) to perform an initial exploratory analysis of data
S2	M12	Training of models for SPC using the data gathered in S1. Initial validation of SPCs with maintenance staff to understand root causes of detected anomalies.	Integration of all historical data from defined data assets and new data coming from the spreader	Analytics cluster to evaluate SPC using the information available in the Data Storage cluster	Medium scale (approx. 300GB). Same data that in M6 plus six additional months.
S3	M24	Deployment of the ANN to perform predictive maintenance in real-time. Deployment of the alert system regarding detected anomalies and the need of performing maintenance tasks.	Real-time data from all the systems/devices plus historical data generated along the project.	Full resources of the Analytics and Data clusters connected to Valencia Port and Terminal systems. IoT network from Valencia Port.	Large scale (approx. 1TB and beyond) Historical data and current data generated in the operation

Table 3: Roadmap for delivering pilot results (along TT stages) – Objective 3: "Predictive Decision Support Cockpit for Port and Terminal Stakeholders"

Stage	Date	Features / Objectives	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1	M6	Analysis of historical data from TOS according to the indicators defined. Initial design of the Predective Cockpit Interface	Historical data manually extracted from the TOS systemTOS	Not needed. Initial set of data will be exported in Excel	Small scale

S2	M12	Initial deployment of the web interface using a set of the historical data. Validation of the current indicators implementation. Definition and evaluation of predictive models for indicators.	Integration of the TOS data and data coming from Valencia Port systems	Data Storage cluster deployed in Objective 2. Analytics cluster to train and evaluate predictive models for indicators	Medium scale
S3	M27	Final deployment of the Predictive DSS including future values and predictive trends of indicators according to the models defined.	Real-time data from the TOS systems and historical data	Data Storage cluster to store additional historical information	Medium scale

3 Design of duisport Pilot (Replication Pilot)

The Duisburger Hafen AG, **duisport**, owns and manages the Port of Duisburg. With a total handling of 3.7 million TEU[3], duisport is the world's largest inland port. For this port and logistics location, the duisport Group offers full service packages in the area of infra- and supra-structure, including relocation management. In addition, the subsidiaries also provide logistics services, such as the development and optimization of transport and logistics chains, rail freight services, building management and packaging logistics. The approximately 300 logistics-oriented enterprises located at the Port of Duisburg generate value of about 3 billion euros per year. Eight multi-modal container terminals, over 400 weekly combined transports to over 80 direct destinations in Europe and Asia as well as comprehensive warehousing and storage capacities are tied in locally with market- and customer-oriented services.

The duisport Group (approximately 1,100 employees) together with the companies located in the port area, along with the logistics companies in the region employ more than 45,000 people today – 20,000 more than 18 years ago.

The duisport pilot will demonstrate the use of big data solutions for the proactive management of bi-modal terminal operations as well as for predictive maintenance of terminal equipment. This pilot will thereby assess in how far solutions developed in the Valencia pilot may be replicated and reused for the more challenging setting of the duisport inland port. Compared to Valencia port, the added complexity in duisport stems from the fact that the port is situated in the middle of a large city (with close to ½ million inhabitants) and at the center of Germany's largest metropolitan area, the Rhine-Ruhr metropolitan region (with close to 10 million inhabitants). This means that duisport has a multitude of roads, tracks and water ways that serve as entry and exit points for containers to and from the actual terminals and ports (see Figure 16). In addition, roads need to be shared with many other cars within the metropolitan area.

[3] Twenty-foot equivalent unit, a measure used for capacity in container transportation, based on the volume of a 20-foot-long (6.1 m) intermodal container; see https://en.wikipedia.org/wiki/Twenty-foot_equivalent_unit

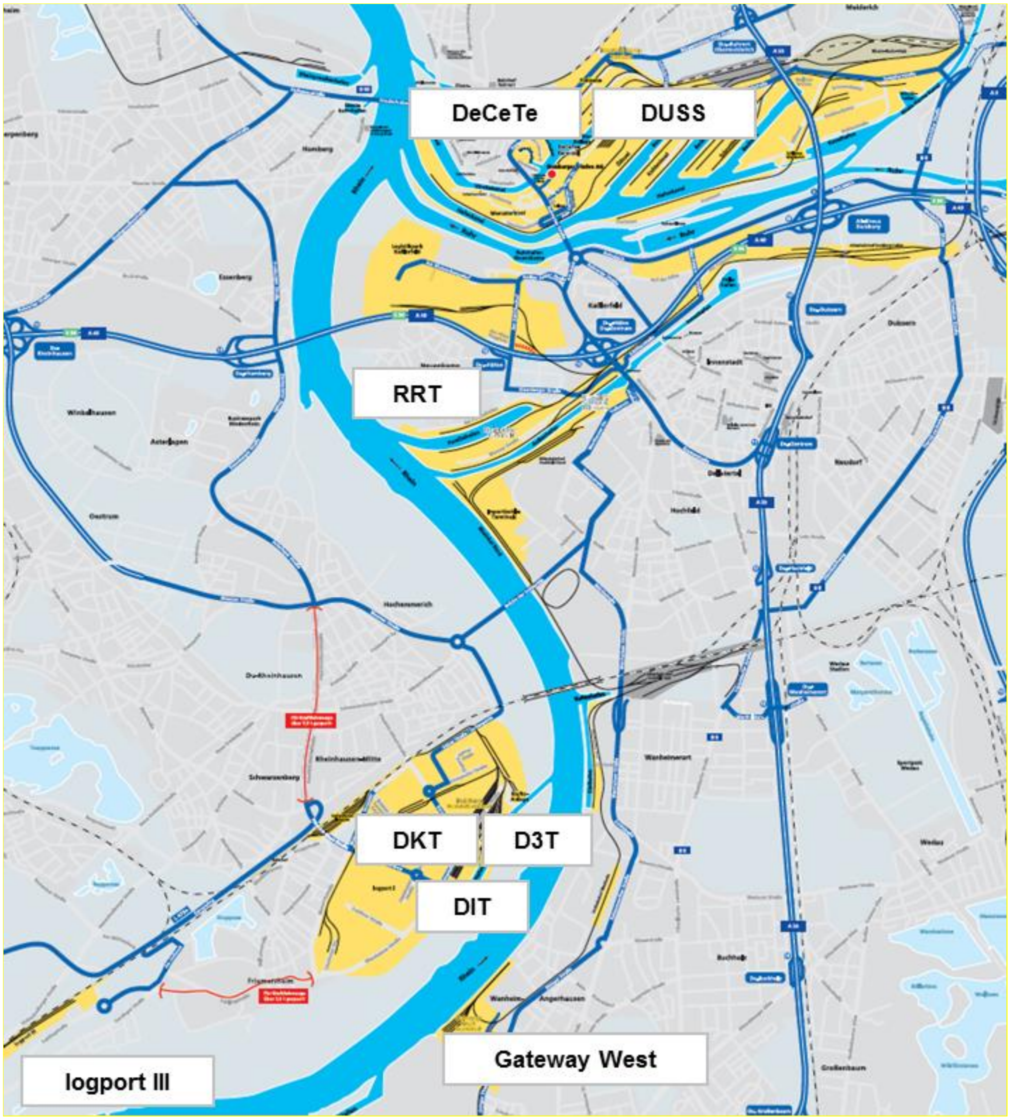


Figure 16: Map of duisport and how it is situated within the metropolitan area (DeCeTe, DUSS, … = Container Terminals)

The duisport pilot focus on use cases, processes and data assets that are linked to the "logport III" bi-modal terminal (see bottom left corner in Figure 16). The "logport III" terminal has the following characteristics:

- Covering an area of 15 hectares;
- Offering 9 rail connections;
- Running 7 transhipment tracks and 2 shunting tracks;
- Operating 2 gantry cranes;

- (among other services) Intermodal transportation services to Scandinavia managed by Samskip van Dieren Multimodal
- Interconnection to other duisport port areas by rail ("logport shuttle")

"logport III" is connected to other port areas via its own rail shuttles operated by duisport rail. Through this connection, logport III is ideally connected to the intermodal network of duisport which offers connections to more than 80 destinations in Europe and Asia. This includes daily rail and barge shuttles to the seaports of Antwerp and Rotterdam.

3.1 Requirements

The duisport pilot will address two main requirements.

- Req. 1: Improved operational management of terminal operations, which includes
    - Increased overall terminal efficiency by enabling terminal operators / dispatchers to proactively manage terminal and port operations based on real-time, predictive monitoring and analytics;
    - Increased robustness of terminal processes by ensuring timely response and mitigation of (potential / anticipated) problems.
- Req. 2: Cost-optimized maintenance of port equipment, which includes
    - Predictive maintenance of terminal equipment (e.g. cranes, reach stackers) to reduce the number of failures during operation;
    - Combination of planned and predicted maintenance to reduce outage times of equipment;
    - Strategic optimization of equipment usage, configuration (e.g., which types of components showed better performance in the field).

3.2 Objectives

The requirements identified in Section 3.1 are addressed by the following two main objectives pursued in the duisport pilot:

3.2.1 Objective 1: Terminal productivity cockpit

The terminal productivity cockpit will exploit advanced data processing and predictive analytics to facilitate proactive decision making and process adaptation. In particular, the productivity cockpit will leverage cutting-edge predictive business process monitoring solutions, i.e., real-time predictive big data analytics for terminal processes.

Predictive business process monitoring aims at anticipating potential process performance violations during the execution of a business process instance (Jalonen; Lönnqvist, 2009). To this end, predictive business process monitoring predicts how an ongoing process instance will unfold up to its completion (Metzger; Leitner; et al., 2015). If a deviation from the planned or acceptable situation, i.e. a violation, is predicted, a process instance may be proactively adapted to prevent the occurrence of such potential violations (see Figure 17). As an example, if a delay in delivery time is predicted for an ongoing freight transport process, faster means of transport or alternative transport routes may be scheduled before the delay actually occurs.

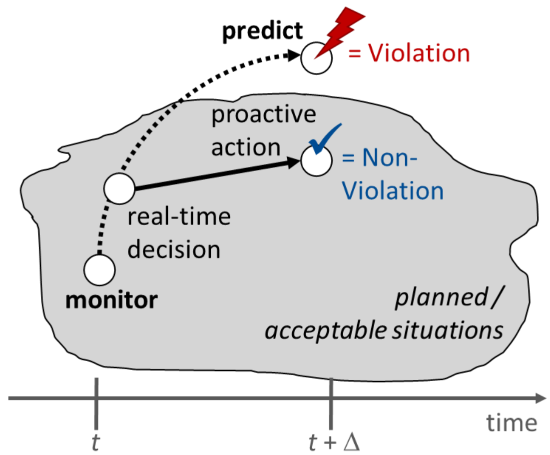


Figure 17: Proactive process adaptation based on predictive process monitoring

The results of predictive analytics will be visualized as part of the terminal productivity cockpit and thereby will be presented to terminal operators / dispatchers for decision support. **¡Error! No se encuentra el origen de la referencia.** shows an example of a terminal productivity cockpit jointly developed by UDE and duisport in the German LoFIP-project (Metzger; Schmidt; et al., 2014). The LoFIP-cockpit is only capable of showing facts (monitoring information) and does not yet provide predictive information. TT cockpits will enhance this with predictions concerning the terminal processes.

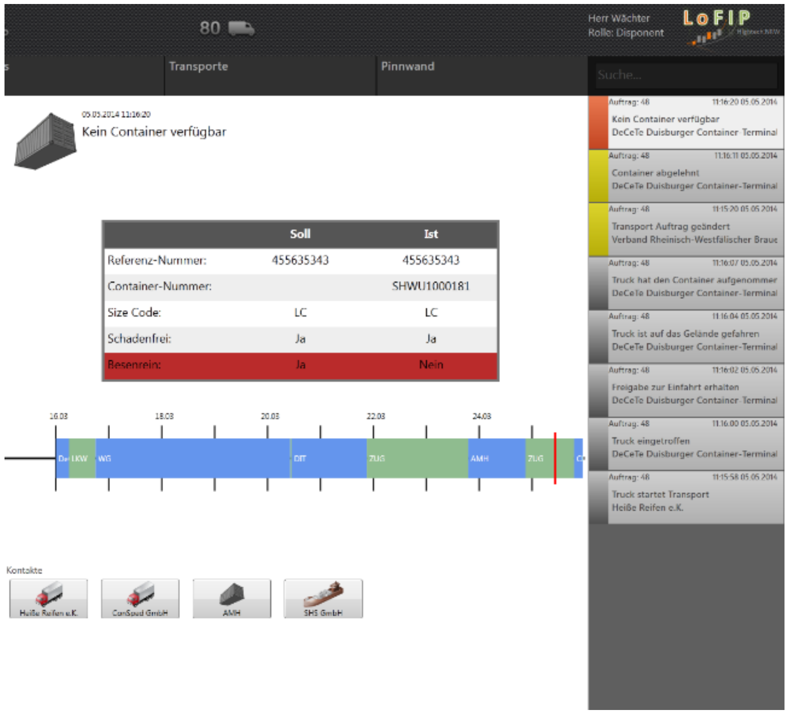


Figure 18: Example for a terminal productivity cockpit (from lofip.de project)

In TT, to provide additional information for taking decisions, the predictions will further be enhanced with additional information. In particular we aim to computing and visualizing how reliable a given prediction may be considered (Metzger; Föcker, 2017), rather like in the case of weather prediction ("80% chance of rain").

3.2.2 Objective 2: Predictive Maintenance System

Repairing a broken piece of hardware, such as a crane or fork-lifter, only after it has encountered a debilitating issue or during a prescribed maintenance interval may waste valuable time and resources. Therefore, the aim is to prevent or minimize proactively the occurrence of equipment and system failures by using available data to predict the need for maintenance and replacement of equipment based on facts and equipment usage rather than on estimates.

In the setting of the duisport pilot, predictive maintenance serves an important role to reduce the changes for equipment failures and thus delays and reduced effectiveness of terminal operations. The Predictive Maintenance System thus complements the Terminal Productivity Cockpit in increasing overall terminal productivity.

With availability of data about machine usage and performance a reactive orientation to maintenance is not only no longer necessary, it is also an indicator of competitive disadvantage. Predictive maintenance overcomes the drawbacks of such preventive maintenance by constantly monitoring actual equipment condition and using the information to predict when a problem is likely to occur. The condition of a component, measured when the equipment is in operation, governs planned and scheduled maintenance. When something does go wrong, descriptive and predictive analytics provide the right information at the right time to guide their actions. Within TT, the available data of equipment usage will be tracked and analyzed to exploit this information as part of the predictive maintenance system facilitating decision making and scheduling.

Within the duisport pilot, we plan exploiting a model-based and trend-based approach towards predictive maintenance (replicating the concepts and solutions from the Valencia port pilot). The goal of the trend-based approach is to discover deviation ranges from what is considered regular operative ranges. This approach requires a wide network of sensors and devices to precisely define the status of the equipment to monitor. Conditions and variables related to a failure are identified based on models. To gain this knowledge, specific conditions (failure patterns) when a failure happened will be recorded. Once these failure patterns have been identified, their current values are monitored during the equipment lifecycle to predict an imminent failure or to show deterioration evidence. The model-based approach that we will explore here allows identifying the reasons for why certain failures have been identified and thus facilitates identifying potential root causes for failures.

3.3 Use Cases and Scenarios

The use cases that will be address in the duisport pilot are connected to two main terminal processes: truck handling and train handling. We thus first describe these processes in Section 3.3.1 and then the use cases in Sections 3.3.2 – 3.3.3.

3.3.1 Processes

The processes at the terminal are differentiated on the basis of the mode of transport. The main steps and activities of the two relevant processes together with the respective data sources at the terminal logport III are shown in Figure 19 for truck handling and in Figure 20 for train handling respectively. The steps and activities of the processes are explained below, while the data sources are further explained in Section 3.4.

Truck Handling. Figure 19 depicts the process of handling a truck within the terminal. The depicted process begins with the a-priory notification of the truck at the terminal, i.e., before the truck actually arrives at the logport III terminal. This is done via an app that is connected to the traffic management system of the port. When arriving at the terminal's in-gate, an OCR-Gate captures the license plate of the truck and the state of the container (such as damages for liability issues). After checking all required documents, the truck is passed through and is transferred to the unloading point. The driver of the truck receives the information where to park the truck for unloading the container. After unloading the container via reach stacker or crane, the container is transferred to the temporary container stock or to a free slot of a train waggon on the railway infrastructure of the terminal. On the other side, another container (from temporary stock, train, another truck) will be transported to the previous truck for loading. The truck then departs the terminal and is again captured by an OCR-Gate at the terminal's out-gate.

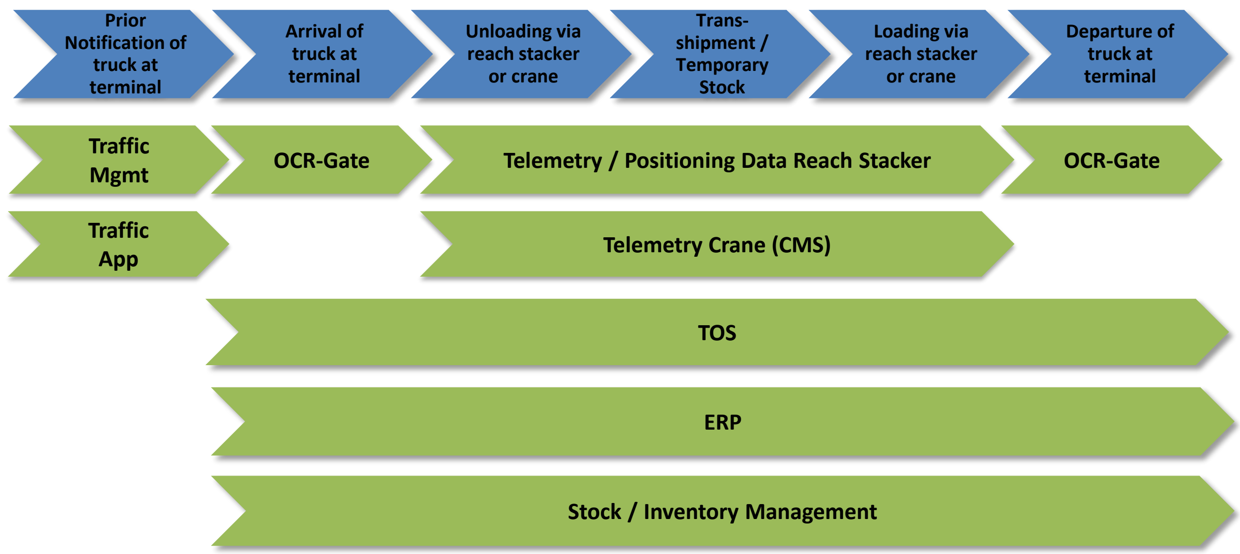


Figure 19: Process and data description logport III (truck handling)

Train Handling. Figure 20 depicts the process of handling a train in the terminal. Before arriving in a shunting yard on the infrastructure of duisport, the train is announced via the network dispatching system of DB Netz. The duisport subsidiary duisport rail transfers the train to the logport III terminal. Before arriving on the infrastructure of the terminal, the train is captured via a Rail Gate. The Rail Gate captures the waggons and containers (such as waggon and container number, or damages to waggons and containers). As soon as the train is placed on the terminal infrastructure, the unloading process is done via a crane or a reach stacker. Similar to the truck unloading process, the container is transferred to the temporary container stock, to a free slot of another train or to a free slot on a waiting truck. On the other side, another container (from temporary stock, train, another truck) is transported to the previous train for loading. After completing the loading process of all planned/scheduled containers, the train departs from the terminal and is captured again by the Rail Gate.

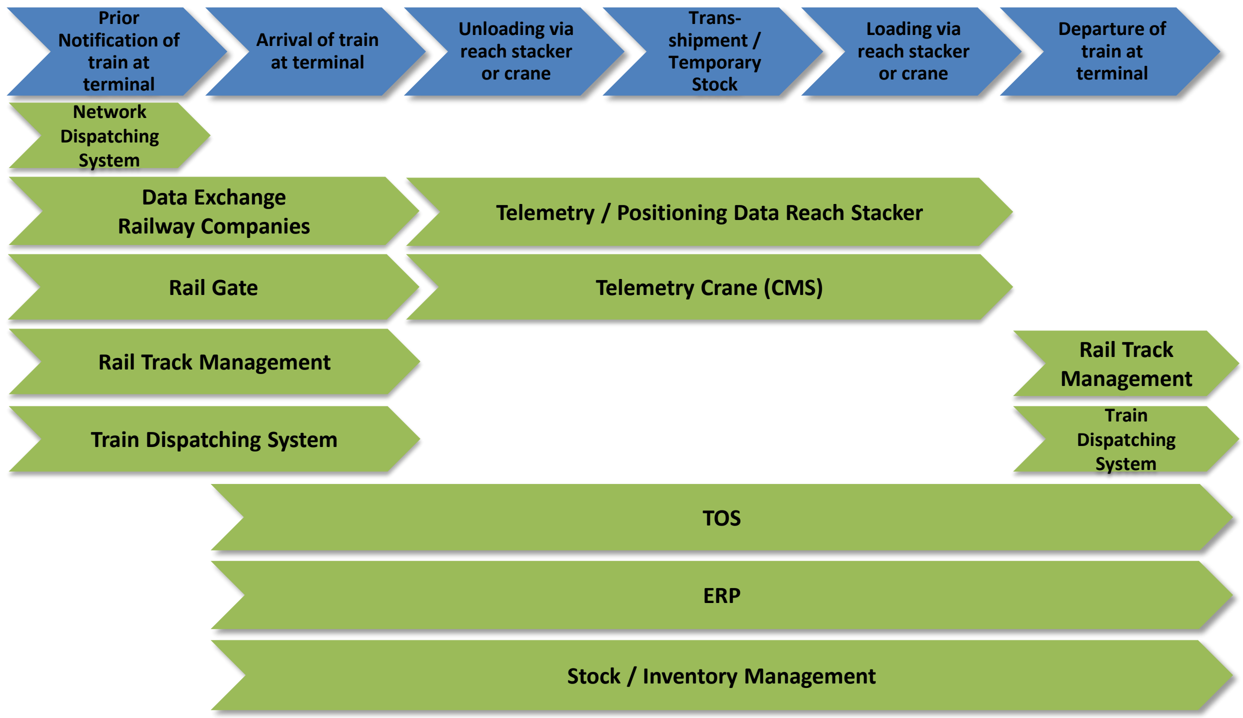


Figure 20: Process and data description logport III (train handling)

3.3.2 Use Case 1: Prediction of Departure Windows: Train to Seaport

As explained in in the introduction to Section 0, the logport III terminal is connected to seaports (Antwerp or Rotterdam) via train connections. In order for a container to reach the planned closing of a vessel in the seaports, it is thus essential that the train together with the planned/schedules containers leave logport III such as to arrive in time at the respective seaport.

The aim of Use Case 1 is to predict the earliest departure time of a given train, based on the predicted completion of the loading of the containers planned to be transported with the given train. The predicted earliest departure time will be compared with the acceptable earliest departure time sufficient to reach the ship. If the predicted time will be too late, a violation will be reported to the terminal managers (cf. Figure 17 in Section 3.2.1) so they can consider mitigation actions, such as re-planning or rescheduling terminal operations if needed.

The prediction has to take into account the bi-modal nature of containers reaching log-port III, via trucks and trains, as well as internal terminal operations.

3.3.3 Use Case 2: Predictive Maintenance of Cranes

Besides the mobile reach stackers, the stationary cranes are the most relevant equipment for container loading and unloading from the different modes of transport. As described above, the logport III terminal is a bimodal terminal with handling of trucks and trains.

As described in Use Case 1, the delayed handling of a train may imply that a closing of a vessel in the seaports Antwerp or Rotterdam cannot be reached. In the case of a breakdown or malfunctioning of a crane, this has severe implications on container handling and thus usually means significant delays in container handling. Due to this delayed handling, a train might not reach the closing of a vessel in the seaports. Therefore, a breakdown of a crane can cause significant loss of time and money. Regular maintenance of cranes is therefore essential to guarantee robust terminal processes.

However, maintenance also implies costs and in particular means that during actual maintenance the equipment is not available for operations. Therefore, unnecessary maintenance activities or too frequent maintenance windows may imply avoidable costs and inefficiencies in terminal operations. Combining planned, less-frequent maintenance windows with the predictive maintenance of components will offer improvement opportunities for terminal efficiency. In particular, exploiting advanced data analytics to determine which components may soon reach the end of their life cycle will offer improvement opportunities for terminal efficiency. In this use case, higher maintenance costs may even be compensated by shortened total maintenance times.

The most parts of a crane under stress during operations are the steel rope and the bumpers. The ropes’ wear depends on weight of the containers, wind forces, etc. whereas the wear of the bumpers depend on the speed of drivers, weight of containers, etc.

In this use case we aim to predict the failure of the most relevant components of a crane to reduce breakdown times as well as to reduce costs of maintenance (e.g., by making better use of the maintenance time windows) and further to improve the quality of maintenance time prediction.

3.4 Data Assets

The duisport pilot will explore and exploit the following data sources and data assets. How the information from these data assets relates to the terminal processes is explained in Section 3.3.1 above.

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
ERP	Enterprise resource planning system (ERP system) to support all business processes running at the terminal – especially for accounting.	Q1/2017	Windows Service / SQL DB	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Telemetry crane	The on-board system of the crane collects various telemetry data.	Q2/2017	Windows Service / SQL DB	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Network Dispatching System	Public dispatching systems (Leidis NK, system of DB Netz and RNE TIS, European System) for train management by delivering real-time train data.	Q2/2017	Tbd	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Rail Gate	OCR gate for trains arriving at / departing from the terminal. The rail gate captures the waggons and containers (e.g. waggon and container number, damages to waggons and containers).	Q2/2017	Xml	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Telemetry / Positioning Data Reach Stacker	The on-board system of the reach stackers collects various telemetry and positioning data.	Q3/2017	Csv	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Train Dispatching System	Dispatching software for the the subsidiary duisport rail (public railway company).	Q3/2017	Windows Service / SQL DB	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Traffic Management	The system collects all truck data, anonymizes and bundles them with available traffic data such as travel time, traffic situation and congestions before forwarding it to mobile devices and LED signs. Arriving truck drivers receive all relevant traffic information as they approach and are quickly routed to the next available loading area or terminal.	Tbd	Xml	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
OCR Gate	OCR gate for trucks at the terminal in- and out-gate. The OCR gate captures the	Tbd	Xml	https://3.basecamp.com/3320

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
	license plate of the truck and the state of the container (e.g. damages for liability issues). Feasibility of capturing of the container number via the OCR gate is currently being examined.			520/buckets/1429164/vaults/449193238
TOS	Terminal operating system to control the movement and storage of various types of cargo (container and trailer) at terminal.	Tbd	Xml	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Stock- / Inventory Management	System to support spare part management for different terminal equipment.	Tbd	Tbd	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Data Exchange Railway Companies	Information (e.g. number of containers, container type) provided by the railway companies.	Tbd	Csv	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238
Rail Track Management	Rail track management for managing the infrastructure within the port.	Tbd	Windows Service / SQL DB	https://3.basecamp.com/3320520/buckets/1429164/vaults/449193238

3.5 Big Data Technology, Techniques and Algorithms

The big data technology stack employed for the duisport pilot combines commercial big data technology from Software AG with state of the art research results of UDE on advanced predictive analytics. Figure 21 shows the main components of the duisport pilot big data stack.

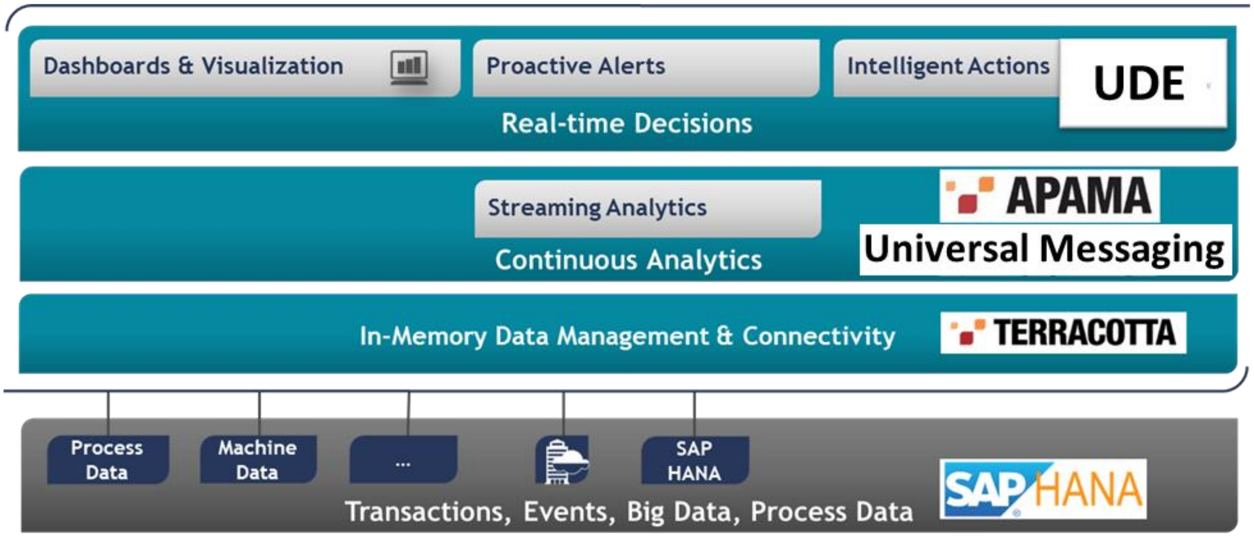


Figure 21: Big data stack of duisport pilot

The big data technology employed in the duisport pilot will include SAP HANA (for persistent data storage) in combination with Software AG Terracotta (as in-memory database), enhanced by Software AG Apama for real-time processing of event streams. This will be complemented by UDE's predictive analytics techniques, building on ensemble machine learning-based prediction to provide enhanced decision support. A significant advantage of ensemble prediction is its increased forecast accuracy, as well as the ability to provide local accuracy indicators for each individual forecast, thereby providing additional support for the logistics operator / dispatchers in decision making.

The individual building blocks of the duisport pilot big data stack are explained in the following sub-sections.

3.5.1 HANA (Persistent Storage)

SAP HANA is used for persistent storage of pilot data. SAP HANA is readily offered by the pilot's big data infrastructure (see Section 3.7) and is available as a complete, encapsulated, secure environment. The SAP HANA database will be used for storing historic data, thereby facilitating training the machine learning based analytics components (see Section 3.5.5).

In addition, the database will serve facilitate integrating data from the different data sources (see Section 3.4). A particular benefit of SAP HANA in this regards is that it offers interfaces to other SAP products (in particular ERP systems) and thus appears a more general approach for data access and integration.

Other functionalities of SAP HANA are not required and will not be used (e.g., In-Memory Processing will be done by Software AG solutions, see Section 3.5.2).

3.5.2 Terracotta (In-Memory Data Management)

As persistent storage is much to slow for real-time access and analytics, the Software AG In-Memory Data Management solution Terracotta will be employed. In particular, BigMemory of Terracotta is used to store massive amounts of data in memory in order to enable fast access. Massive datasets are instantly available in ultra-fast RAM distributed across any size server array. If needed, hundreds of terabytes of heterogeneous data can be maintained in-memory, with guaranteed latency of low milliseconds.

Teracotta will facilitate real-time streaming of data and events to the Universal Messaging event broker / event bus (see Section 3.5.3). These data items may come from real-time data sources (such as process monitoring or machine data), or may be streamed from the SAP HANA database, thereby simulating the real-time arrival of events. Depending on the stage of development of the pilot (see Section 3.8), either historic data may be streamed (Stages 1 and 2) or real-time data from the field may be used (Stage 3).

3.5.3 Universal Messaging (Event Broker / Event Bus)

Universal Messaging is a message-oriented middleware product that guarantees message delivery across public, private and wireless infrastructures. Universal Messaging has been built from the ground up to overcome the challenges of delivering data across different networks. It provides its guaranteed messaging functionality without the use of a web server or modifications to firewall policy. It is a single end-to-end solution for the delivery of real time data. It will be used for streaming all relevant data such as complex events for streaming analytics to different consumers.

The data analytics components (such as UDE’s predictive monitoring techniques, see Section 3.5.5, or dashboards for data visualization) can thereby subscribe to streams of relevant events. The benefit is a loosely couple and flexible architecture which allows plugging and playing different solution components. In particular, this will be used to build the pilot’s Terminal Productivity Cockpit (see Section 3.2.1) and the Predictive Maintenance System (see Section 3.2.2).

3.5.4 Apama (Complex Event Processing / Event Stream Processing)

Apama Streaming Analytics is Software AG’s platform for analyzing high-velocity streaming data to detect business risks and opportunities in real-time. It can filter, aggregate, enrich and analyze a high throughput of data from multiple disparate live data sources and in any data format to identify simple and complex patterns to visualize business in real-time, detect urgent situations, and automate immediate actions.

3.5.5 UDE's Ensemble Predictive Process Monitoring

The predictive business process monitoring solution used as part of the Terminal Productivity Cockpit will be based on machine learning. Machine learning deals with (semi-)automated learning of relationships or behavior from data. Figure 22 outlines the overall procedure of the machine learning-based predictive monitoring technique. Essentially, there are two separate steps. First, in the model learning step, a predictor is trained from an existing set of training data, e.g., a data set containing historical process execution traces. In this first step, we also need to decide which data is most important for prediction. This can be done manually by a human or semi-automatically using techniques such as principal component analysis. Second, in the run-time prediction step, data collected from the running process instance is used as input to the trained machine learning model to generate a concrete prediction for this process instance (Jalonen; Lönnqvist, 2009).

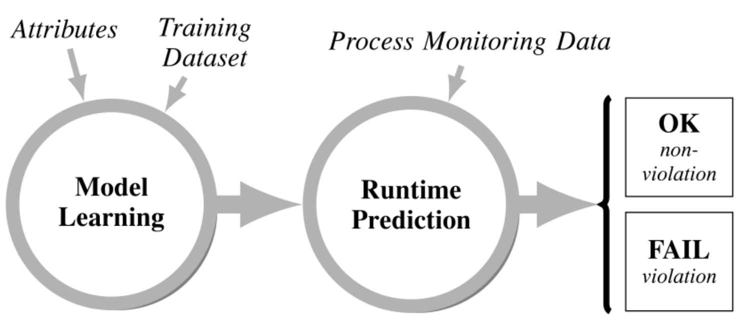


Figure 22: Machine-learning based Predictive Monitoring

A key requirement for the applicability of any predictive business process monitoring technique is that the technique delivers accurate predictions. Informally, prediction accuracy characterizes the ability of a prediction technique to forecast as many true violations as possible, while – at the same time – generating as few false alarms as possible. Prediction accuracy is important to avoid the execution of unnecessary process adaptations, as well as not to miss required process adaptations (Metzger; Sammodi; Pohl, 2013).

Research on predictive business process monitoring focused on improving a prediction technique's aggregate accuracy. Aggregate accuracy takes into account the results of a set of predictions. Examples for aggregate accuracy metrics are precision, recall or mean average prediction error. When compared with aggregate accuracy metrics, prediction reliability estimates provide additional information about the error of an individual prediction for a given business process instance. Reliability estimates provide more information about the individual prediction error than aggregate accuracy metrics. Reliability estimates thus facilitate distinguishing between more and less reliable predictions. Reliability estimates can help deciding whether to trust a prediction and consequently whether to perform a proactive adaptation of the given process instance. As an example, a process manager may only trust predictions with a reliability greater than 80%. Only then, the process manager would proactively schedule faster – and therefore often more expensive – means of transport.

We will use ensemble prediction techniques on top of the machine-learning based prediction (see Section 3.2.1) to compute reliability for predictive process monitoring (Metzger; Föcker, 2017). The principle approach to compute these reliability indicators is illustrated in Figure 23. The majority of the predictions decides on the overall prediction that is delivered, while the number of predictions that agree is used to compute the reliability.

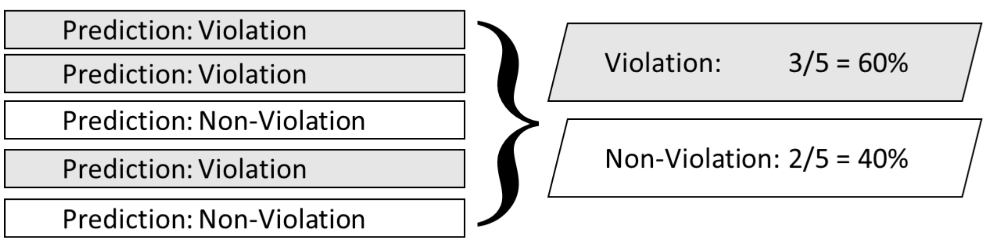


Figure 23: Computing prediction reliability from ensembles

3.5.6 Dashboards & Visualization

Apama comes along with an integrated dashboard builder. Depending on the specific requirements of the pilot, Software AG can also bring in the MashZone NextGen self-service analytics tools for business users. Thereby, even non-technical users can create dashboards directly from live information sources. Combining data from any source — such as data warehouses, ARIS, XML files, databases and streaming big data. The dashboards can be viewed on any device, anywhere, for right-time decision-making.

Additional concepts for dashboards and visualization, jointly developed by UDE and duisport in the German LoFIP-project (Metzger; Schmidt; et al., 2014), will complement the visualization solutions from Software AG.

3.6 Positioning of Pilot Solutions in BDVA Reference Model

Figure 24 depicts the components of the duisport pilot big data stack in the BDVA reference model. The components are numbered in the order of the sections in Section 3.5 (i.e., #1 = Section 3.5.1), where these are explained in more details.

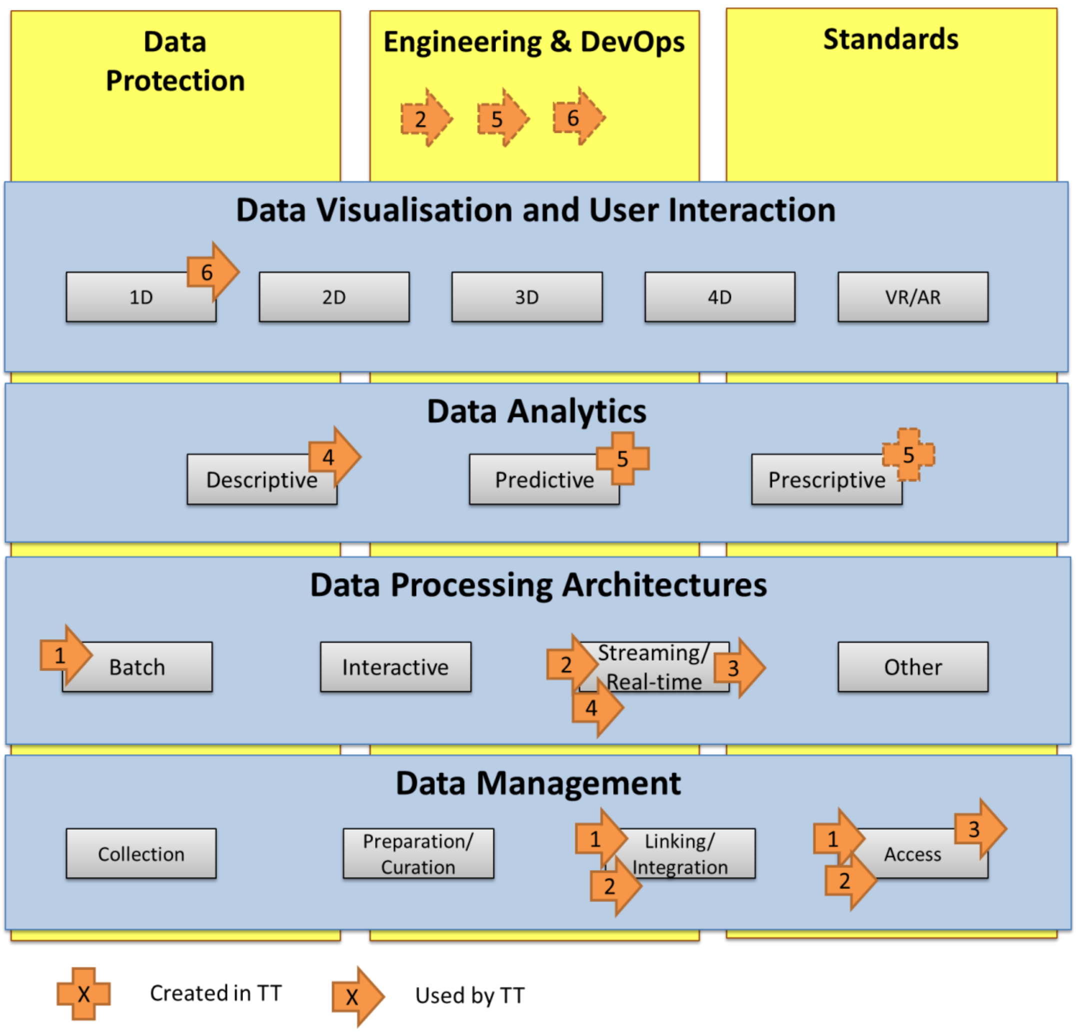


Figure 24: Pilot big data components in the BDVA Reference Model

Concerning the vertical elements of the BDVA Reference Model, at this point in time no dedicated use and contributions to **standards** are envisioned. **Data protection** is ensured through the deployment of the pilot solutions on the secure SDIL data center, which is part of the pilot's big data infrastructure (see Section 3.7). Several of the technical components provide **engineering and deployment** support. For example, each of the Software AG solutions offers administrative and monitoring support. Universal Messaging comes along with Enterprise manager. A graphical Tool developed using the Universal Messaging Administration API. Every aspect of the tool has corresponding API functionality and it is possible to monitor, manage and configure multiple realms and clusters from anywhere. Apama provides different tools for business analysts and IT-developers. Business analysts can use the design view in Apama Studio to build their Apama queries, whereas IT-developers have a source view to create queries or to monitor the EPL. Apama also have a visualization and dashboard builder aboard. The UDE predictive monitoring solutions are implemented on top of the WEKA open Source Machine Learning toolkit, which provides – among others – support for training and validating models.

3.7 Big Data Infrastructure

Data hosting and actual servers will be in the country of the duisport pilot, i.e., in Germany. This will involve the use of the German Big Data innovation space (Smart Data Innovation Lab - SDIL[4]) through TT partner Software AG. SDIL offers to TT a development environment with SAP HANA as well as components of the Digital Business Platform from Software AG, including Terracotta In-Memory Data Management, APAMA Complex Event Processing, as well as Universal Messaging streaming analytics. Figure 25 shows an overview of the SDIL technical platform.

[4] http://www.sdil.de/en/

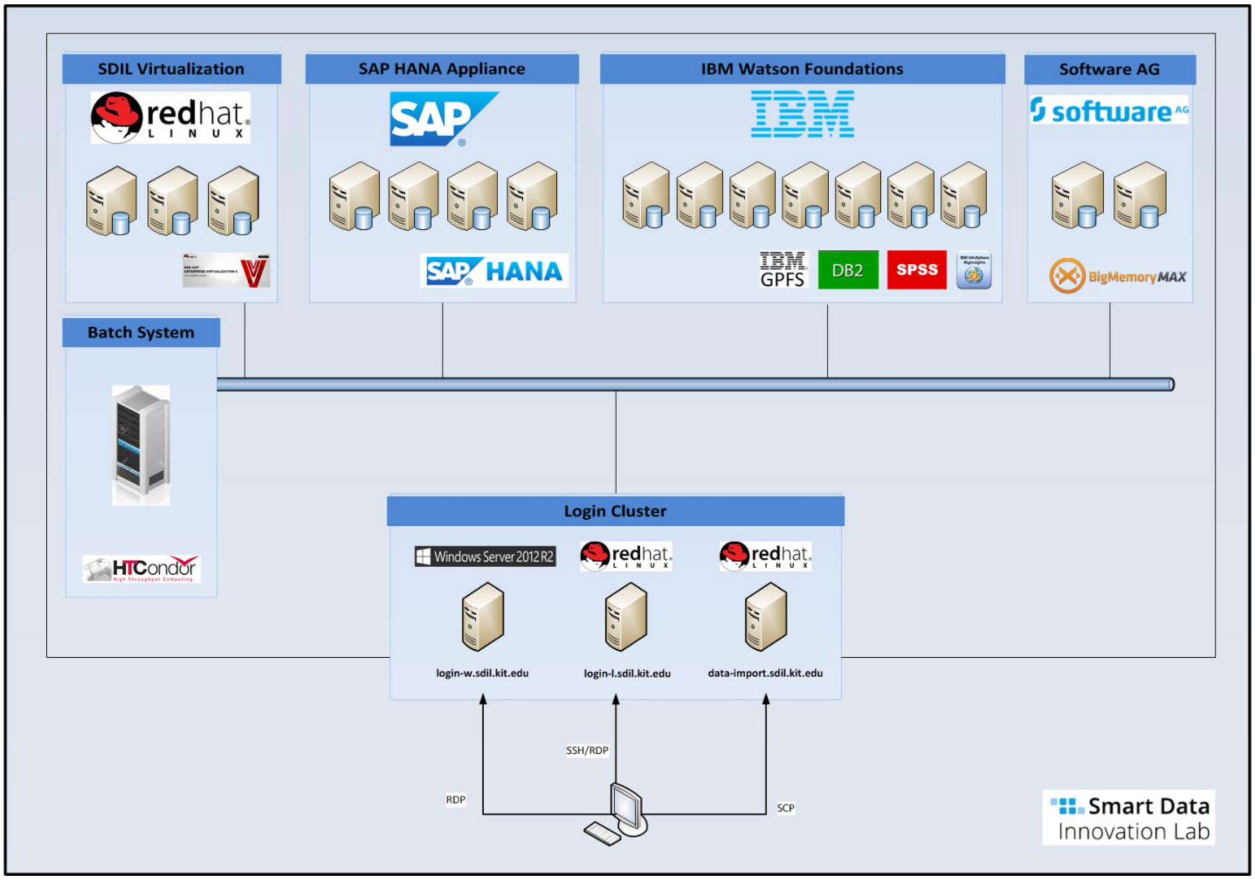


Figure 25: Overview of SDIL Platform (http://www.sdil.de/en/platform)

Depending on the concrete processing requirements that will be determined for stages S2 and S3 of the project, SAP HANA will run on up to 320 Cores (4 servers with 80 cores each) with up to 4TB RAM (each server hosts 1TB of RAM) and disk space of 80TB (each server hosts 20TB of disk space).

The TT duisport pilot has been approved as an SDIL project by SDIL's board and thus the use of the SDIL for the purposes of TT are ensured.

3.8 Roadmap

The pilot solutions will be developed in an incremental and iterative fashion. The release of pilot solutions along the three main stages (following the overall TT methodology as defined in deliverable D2.1) are depicted in the below two tables. These two tables are presented for Objective 1 and Objective 2 respectively. It should be noted that we estimate that for Objective 2, only two stages will be feasible to be performed during the course of TT.

Table 1: Roadmap for delivering pilot results (along TT stages) – Objective 1: "Terminal Productivity Cockpit"

Stage	Delivery Date	Features / Objectives Addressed	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1: Technology Validation	M9	Gathering first insights (project internal) about the potential of the data for the purpose of **predictive process monitoring** by means of a release 1 prototypical implementation of the Terminal Productivity Cockpit (data processing and prediction core components).	Existing, historic data collected and prepared for technology validation.	Historic data stored in SAP HANA and amenable for analysis via Teracotta and Apama	Small scale (sufficient size for training machine learning models and their cross-validation)
S2: Large-scale experimentation and demonstration	M15	Solution validation of **predictive process monitoring** capabilities of Terminal Productivity Cockpit, involving the participation of relevant stakeholders (dispatchers, etc.) along the port value chain, including from hub level and transport corridor level based on release 2 prototypical implementation of the Terminal Productivity Cockpit (including dashboards and visualization).	Historic data enriched with additional data sources and data sets	Ditto + use of Mashzone NG for visualization	Large scale (more diverse, larger data sets to evaluate scalability); larger size of data may be produced via simulation
S3: In-situ trials	M27	Deployment of Terminal Productivity Cockpit in parallel to the actual port operations.	Use of real-time	SDIL data processing infrastructure will receive a copy of daily generated data for experimental purposes	Large scale (actual operations)

Table 2: Roadmap for delivering pilot results (along TT stages) – Objective 2: "Predictive Maintenance System"

Stage	Delivery Date	Features / Objectives Addressed	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1: Technology Validation	M15	Gathering first insights (project internal) about the potential of the data for the purpose of **predictive maintenance** by means of a release 1 prototypical implementation of the Predictive Maintenance System; Analysis of replicability of solutions from Valencia pilot.	Existing, historic data collected and prepared for technology validation.	Historic data stored in SAP HANA and amenable for analysis via Teracotta and Apama	Small scale (sufficient size for validation of trend-based approach)
S2: Large-scale experimentation and demonstration	M27	Solution validation **predictive maintenance** capabilities, involving the participation of relevant stakeholders (maintenance personnel, etc.) along the port value chain, including from hub level and transport corridor level based on release 2 prototypical implementation of the Predictive Maintenance System.	Historic data enriched with additional data sources and data sets.	Ditto + use of Mashzone NG for visualization	Large scale (more diverse, larger data sets to evaluate scalability); larger size of data may be produced via simulation
S3: In-situ trials	-	No in-situ trials planned for predictive maintenance.	-	-	-

4 Commonalities and Replication

4.1 Common Requirements and Aspects

This section presents common points from both pilots according to the previously presented objectives. Using as reference the requirements defined by the initial pilot, described in section 2.1, and the replication pilot, presented in section 3.1, the following table shows how they align:

Table 1 Commonalities between pilot requirements

Valencia Port Pilot (Initial)	duisport Pilot (Replication
Improve RTG crane scheduling aiming to increase its productivity, utilization and efficiency.	Not considered
Predict when specific equipment from cranes will fail or needs to be replaced ***reducing maintenance costs.***	***Predictive maintenance*** of terminal equipment to ***reduce the number of failures*** during operation Combination of planned and predicted maintenance to ***reduce outage times of equipment***;
Provide ***relevant indicators*** in order to ***visualize trends*** and enact relevant information for the yard planning, hence, ***improving yard utilization*** and maximizing labour and equipment usage	***Increased overall terminal efficiency*** by enabling terminal operators / dispatchers to proactively manage terminal and port operations based on ***real-time, predictive monitoring and analytics***;
Not considered	Strategic optimization of equipment usage, configuration
Not considered	Increased robustness of terminal processes by ensuring timely response and mitigation of problems.

First requirement in common is that both pilots foresee to reduce dramatically maintenance costs related with an equipment failure and to predict when this issue could happen. In both pilots, a predictive maintenance approach is described to fulfil such goal, taking into account a continuous monitoring of the equipment. Specifically, both pilots will focus on specific parts of cranes, a popular piece of equipment in ports (and in other logistics sectors) whose right operative is essential. We foresee that the initial pilot will provide insights about root causes or variables related to crane failures, specifically the spreader. For instance, the risk of a breakage monitoring the number of container impacts or the average lifted weight. In addition, the initial pilot will characterize how the current operative processes, i.e. current preventive maintenance tasks, could evolve using the initial insights. Then, the replication pilot will take into account in its initial deployment these detected variables to validate if they are specific or not to the initial pilot. In addition, the replication pilot could provide additional data to verify variables not considered by the initial one. From the analytics perspective, the initial predictive approach will use SPC charts (approach described in section 2.5.5) to detected anomalies and behavioural patterns that lead to a crane failure. We will use and extend accordingly equivalent techniques in the replication pilot to build predictive models more accurate.

The second common requirement is to improve the efficiency of the terminal/port/yard by means of predictive indicators that use simultaneously historical and real-time data. Both pilots propose to present such information in a web-based cockpit in order to enable a user-friendly interaction. On the one hand, the replication pilot will present an initial demonstrator of their predictive cockpit. On the other hand, initial pilot will propose an overall design of the cockpit and an additional set of useful indicators. Indicators will provide both an approach to gather the aggregated values and a corresponding visualization. Still we must to analyse in more detail the required indicators for both cockpits to find similarities, but there is a high probability of coincidence as both pilots work with similar processes.

Regarding data assets, several commonalities arise in both pilots. For instance the use of an ERP and a TOS as main applications to manage and support their operational processes. Though each pilot uses a specific technology/software vendor, common information is expected. Positioning (using GPS or telemetry) is other common data source in both pilots that will help to define similar approaches to describe the status of the port/terminal/yard. Finally, current traffic in the terminal/port is another data asset available in both pilots. The use of similar concepts helps to identify common data requirements and potentially reuse indicators or predictive techniques.

Regarding the technological stack, there is no specific alignment between the employed technologies, but they are some similarities in the architecture these technologies enable. In both pilots a main data storage technology is proposed (SAP Hana in duisport, and Hive-Cassandra for Valencia Port) to store the raw data coming from different data assets. In both pilots, a highly scalable data repository is presented. Another common point is the introduction of a Complex Event Processor (Apama Streaming Analytics in duisport and WSO2 Data Analytics Server in Valencia Port). Finally, both pilots presents results using a web-based application.

4.2 Replication

The replication in WP7 will proceed along the two main objectives of the pilot domain:

- **Predictive Maintenance**: The initial pilot in Valencia will lead the way and explore the applicability of predictive maintenance techniques for cranes. In particular, the initial pilot will investigate into the use of statistical process control, as well as artificial neural networks. The replication pilot will replicate the statistical process control solutions, while adding the challenge of aiming to predict the maintenance of cables (which are more difficult to instrument and monitor).
- **Terminal Productivity Cockpit with Predictive Process Analytics**: For this objective, the "order" of replication will reverse. The replication pilot, duisport, starts with investigating into predictive process monitoring and using reliability estimates to improve decision making in cockpits. The lessons learned and concepts employed as part of these predictive productivity cockpits will then be replicated by the Valencia pilot. The decision-making approaches defined in the duisport pilot are validated according to the specific processes and indicators of the Valenciaport pilot.

5 Conclusions

This deliverable reports on the initial design of the pilot according to the TT methodology presented in D2.1. Current status of both pilots is summarized as follows:

- At this stage of the project, both pilots have clearly defined their objectives and use cases. The need of applying Big Data is clearly justified according to the heterogeneity of the involved data and the complexity of the underlying processes.
- The application of predictive maintenance and predictive indicators is a key selling point in both pilots. The aggregation of historical data will help to understand current processes whereas real-time data will guide the decision making process. In the port context, several processes will benefit from this approach not possible without the introduction of Big Data technologies. According to the initial design, we expect improvement in the overall efficiency of both port equipment (as a consequence of less breakdowns) and human assets (because of a better planning)
- The pilot shows a wide array of data assets coming from different data sources, namely text files, relational databases, web services or data streaming. Providing a unified view of this heterogeneity will highly improve data understanding. As first step towards such direction, this deliverable identifies and describes precisely the available data assets following the guidelines of the TT project.
- Technological stack and infrastructure are presented and aligned according to the BDVA reference model. This exercise has helped us to understand better, how pilots contribute to the Big Data ecosystem. Though there are no common technologies in both pilots, we have detected a clear architectural alignment in how data will processed, analysed and visualized.

6 Appendix 1. Bibliography

Héctor J. Carlo; Iris F.A. Vis; Kees Jan Roodbergen, 2014. Storage Yard Operations in Container Terminals: Literature Overview. European Journal of Operational Research, p. 412–430.

Iris F. A. Vis; Hector J. Carlo, 2010. Sequencing Two Cooperating Automated Stacking. Transportation Science, pp. 168-182.

Iris F. A. Vis; Kees Jan Roodbergen, 2009. Scheduling of Container Storage and Retrieval. OPERATIONS RESEARCH, pp. 456-467.

Özge Nalan Alp;Hayri Baraçlı, 2009. YARD CRANE SCHEDULING IN CONTAINER TERMINALS - A LITERATURE REVIEW. Hammamet, Tunisia, s.n., pp. 229-232.

Robert F. Dell; Johannes O. Royset, 2009. OPTIMIZING CONTAINER MOVEMENTS USING ONE AND. JOURNAL OF INDUSTRIAL AND MANAGEMENT OPTIMIZATION, p. 285–302.

W.C. Ng; K.L. Mak, 2005. Yard crane scheduling in port container terminals. Applied Mathematical Modeling, pp. 263-276.

Zyngiridis, I., 2005. Optimizing Container Movements Using One and Two Automated Stacking Cranes. Monterye, California: Naval PostGraduate School.

Jalonen, H., Lönnqvist, A.: Predictive business–fresh initiative or old wine in a new bottle. Management Decision 47(10), 1595–1609 (2009)

Metzger, A., Leitner, P., Ivanovic, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S., Pohl, K.: Comparing and combining predictive business process monitoring techniques. IEEE Trans. on Systems Man Cybernetics: Systems 45(2), 276–290 (2015)

A. Metzger, P. Schmidt, C. Reinartz, and K. Pohl, "Management operativer Logistikprozesse mit Future-Internet-Leitständen: Erfahrungen aus dem LoFIP-Projekt (Industrietransfer-Beitrag)," in Software Engineering 2014, Fachtagung des GI-Fachbereichs Softwaretechnik, February 25-28, 2014, Kiel, Germany, ser. LNI. GI, 2014.

A. Metzger and F. Föcker, "Predictive business process monitoring considering reliability estimates,"in Advanced Information Systems Engineering - 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Springer, 2017.

A. Metzger, O. Sammodi, and K. Pohl, “Accurate proactive adaptation of service-oriented systems,” in Assurances for Self-Adaptive Systems, J. Camara, R. de Lemos, C. Ghezzi, and A. Lopes, Eds., vol. LNCS 7740. Springer, 2013, pp. 240–265.