DELIVERABLE

D4.1 – Pilot Design Smart - Highways

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

Funded by the European Union's H2020 GA - 731932

Document fiche	
Authors:	Marion Sanlaville, Iker Guinea, David Collado, Miguel Carpio, Pablo Ferrando
Internal reviewers:	Marta Galende, Dirk Mayer
Work Package:	WP4
Task:	T4.1
Nature:	R
Dissemination:	PU

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	01/02/2017	CINTRA, INDRA, CI3	Draft
0.2	10/03/2017	CINTRA, INDRA, CI3	Doc for internal review
0.6	14/03/2017	CINTRA, INDRA, CI3	Doc for review
0.8	29/03/2017	CINTRA, INDRA, CI3	Second review
1.0	30/03/2017	CINTRA, INDRA, CI3	Final deliverable

Keywords:	Pilot Design, Smart Highways
Abstract (few lines):	This deliverable reports on the work performed in WP4 T4.1, with respect to the motivation and design of the Smart Highways pilots: the main pilot - to be deployed on AP7 “Ausol” highway, and its replication, on A28 “Norte-Litoral” highway.

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 Nº 731932

Table of Contents

Table of Figures

Definitions, Acronyms and Abbreviations

Acronym	Title
API	Application Programming Interface
CO	Confidential, only for members of the consortium (including Commission Services)
CR	Change Request
D	Demonstrator
DL	Deliverable Leader

DM	Dissemination Manager
DMS	Document Management System
DGT	Dirección General de Tráfico (General Directorate of Traffic in Spain)
DoA	Description of Action
Dx	Deliverable (where x defines the deliverable identification number e.g. D1.1.1)
EIM	Exploitation Innovation Manager
ETL	Extraction–transformation–loading
EU	European Union
FM	Financial Manager
MSx	Project Milestone (where x defines a project milestone e.g. MS3)
Mx	Month (where x defines a project month e.g. M10)
O	Other
P	Prototype
PC	Project Coordinator
PM	partner Project Manager
PO	Project Officer
PP	Restricted to other program participants (including the Commission Services)
PU	Public
QA	Quality Assurance
QAP	Quality Assurance Plan
QFD	Quality Function Deployment
QM	Quality Manager
R	Report
RE	Restricted to a group specified by the consortium (including Commission Services)
RUP	Rational Unified Process
STEP	Standard Technology Evaluation Process
STM	Scientific and Technical Manager
TL	Task Leader
VMS	Variable Message Signs
WP	Work Package
WPL	Work Package Leader
WPS	Work Package Structure

Executive Summary

This deliverable reports on the work performed in WP4 T4.1, with respect to the motivation and design of the Smart Highways pilots. The main pilot to be deployed is in Ausol, and its replication in Norte-Litoral concession.

Starting from the general motivation Intended within the more general Transforming Transports Project, Smart Highways pilots' ambition as described at beginning of this document, is to measure the technical and economic impacts of Big Data applied to the highway markets operated by Cintra. A special focus is set on finding operational efficiencies, better customer experience and generation of new business models.

According to the objectives of both pilots, the present document moves on by giving an overview of the three main goals of the pilots: 1. gathering deeper insights and understandings in road traffic and mobility patterns, 2. Optimize highway operation and 3. Guarantee better use of safer roads. It clearly states how the use cases are related to each objective, and in which way they are addressed by the descriptive/predictive/prescriptive models managed by the Big Data platform team.

The main body of the document is devoted to explain the Big Data platform technologies, techniques and algorithms, as well as the infrastructure needed to support it. It's relevant as well to keep an eye on the Data Assets list retrieved by the pilot team: a large amount of historical data is required to feed the pilot platform, where the availability on demand has to be tracked carefully. Detailed information regarding to Data Assets can be found by following the links on the table (ID Cards - access to Basecamp Files & Docs section required).

Finally a roadmap, following the general three-stages methodology for the project (technology validation, large-scale experimentation & demonstration and in-situ trials) is described for both pilots. Starting on Jan-2017 and lasting for 30 months it covers the whole project period.

Note that Ausol Pilot and its replica in Norte-Litoral will overlap with a gap of 3 months, a situation affordable by the project team due to the fact that both pilots share main objectives and requirements.

1 Motivation and Ambition

The pilot is aimed at measuring the technical and economic impacts of Big Data applied to the highway markets operated by Cintra. There will be two pilots, the first applied in Ausol Highway (Spain) where a relevant data base is available, and a second one applied to Norte-Litoral Highway (Portugal) that will **replicate the Big Data solutions** deployed in Ausol.

The challenge of this pilot is to improve traffic flows (mitigate congestion, reduce accidents,..) along the corridor **and the efficiency of the current infrastructure, while enhancing user experience** on the network. Finally the pilot will contribute to demonstrate the technical and economic viability of the use of Big Data, the common goal of the Project.

The unique constellation and complementarity of Transforming Transport partners allows driving innovation of mobility and logistics processes and services along the whole data value chain. In particular, this will lead to innovation along three key value dimensions for Big Data shown below:

Figure 1: Value Dimensions for Big data Use Cases (source: DHL/DETECON)

The following chapters analyzes different variables that will be taken into account throughout the project:

1.1 Operational Efficiency

The concession will aim to manage the capacity, minimize accidents and road operators will attempt to optimize resources, enhance security, reduce operational costs and mitigate congestion events in highways. The road pilot will focus on the following solutions that aim to optimize the road corridors capacity:

- **Descriptive modelling** of road traffic: Application to increase the understanding of the multivariable nature of highway usage, incidents, and associated phenomena such as bottlenecks, black spots, etc.
- **Predictive and prescriptive modelling** of infrastructure use (from short to medium-long term). Application to optimize the infrastructure resources (tolling lanes, personnel, maintenance tasks, etc.) in a road corridor where there are two alternative roads (highway / conventional road). This solution will model the nature of road traffic, vehicles flow and habit-use patterns to optimize all the highway infrastructure resources. It will also include external variables such as weather forecasts, specific events (concerts, summer holidays) to anticipate their impacts on the motorway and therefore to improve the operations of the infrastructure and the experience of road users.

1.2 Customer Experience

The use of Big Data technologies in the transport sector can be the basis to improve the user's experience on the network. The use of information gathered from different sources and their fusion in a single model is mandatory to improve the understanding of mobility demand patterns in highways. The ideas listed below will help to enhance the customer´s experience:

- **Understanding the mobility patterns in the corridors** and the route choice criteria for both the initial and the replication pilot
- **Develop, test and validate the use of specific tools** aimed to manage the traffic on real time (short term prediction) through the initial pilot
- **Forecasting demand methodologies** using Big Data sources oriented to help public and private decision makers through the initial pilot
- **Demonstrate the replicability** and usability of the previous tools and methodologies in other type of highways applying them in the replication pilot

1.3 New Business model

The fusion and analysis of all the data collections from different sources could generate new business around that data and the algorithms derived. One of the objectives is to **obtain an accurate predictive tool of trip patterns, traffic movements and influx** of people in the corridor area and these forecasts could be very attractive for any economic actors of the region. For instance, it's of high interest for industries as tourism, public transportation, logistics services, , etc.

1.4 Expected results

The main advance over existing solutions is the integration of Big Data from key stakeholders, which will provide **enough capacity to react in less time** for highway´s users benefit. Big Data required by the different players will be available and accessible in order to manage all the information and to have a very fast response.

New technological solutions are required to complete the gaps and uncertainties related to the traditional data collection methods in order to obtain reliable, affordable and real-time information. The use of information gathered from different sources and their fusion in a single model is mandatory to improve the understanding of mobility demand patterns in highways definition and operation and the robustness of the forecasts.

Related to this project, Big Data solutions applied to the mobility patterns will enhance the traffic projections and will improve the operation of existing highways. Cintra has selected Ausol concession to deploy main pilot in Spain, the AP-7 toll highway from Málaga to Guadiaro. As there's a free alternative road, traffic load balance models will be tested more efficiently here than in other concessions. Besides, as part of the E15 European route that connects Europe with Africa, AP-7 supports high traffic peaks during summer due to migration and tourism trips, so deploying the pilot on this tolled section might have a deeper impact on the optimization of highway operations and traffic distribution.

Pilot replication in Portugal will pursue same objectives with a few differences: Norte-Litoral concession is operated as free-flow and there're no relevant alternative roads. Thus, motivation and objectives for the replication pilot are still the same than for main pilot, although conclusions will mainly focus on safety, user experience and operation optimization.

Thanks to both pilots, Cintra will enhance its position as a road operator through qualitative advantages by reducing cost and optimizing services. The target market here is an Intelligent Transport Systems (ITS) for smart highways that allows for better traffic projections and improved traffic control and operations.

- **Improve the traffic distribution** along the road corridor.

- Promote a **better utilization** of safer roads.
- Provide road users with **better information and decision tools.**
- **Optimize** highway operations.

Finally, thanks to the conclusions of this project Cintra will have a better vision for the exploitation of results, among which we find:

- Increase and improve **services to highway users**
- Potential **use of the results for other highways** currently owned by Cintra in different markets:
    - Europe - Spain, Ireland, Greece, Slovakia, Portugal
    - Colombia, Australia, USA

2 Design of Initial Pilot

The initial pilot is the 96-kilometre Ausol highway: this highly-congested semi-urban corridor connects the cities of Málaga, Estepona and Guadiaro in the South of Spain. The AP-7 highway is part of the European route E15, Costa del Sol and the alternative road is the semi-urban N-340/A-7 road.

Figure 2: Influence area for the Initial Pilot

The AUSOL Concessions form the A7 motorway which runs between Malaga and Guadiaro along the southern coast of Spain. There are two concessions included in the package:

- **AUSOL I:** This is the first concession that started operation in June 1999 and runs for 75km between Malaga and Estepona. The concession has two tolled sections with the following toll stations:
  - Calahonda
  - San Pedro
- **AUSOL II:** The second concessions started operation three years later, in August 2002, and runs for 21km between Estepona and Guadiaro. The concession has one toll station
  - Manilva

The complete motorway (Ausol I and II) is made up of three tolled sections separated by free sections at Benalmádena, Marbella and Estepona. The sections of Benalmádena and Marbella have three lanes in each direction, while the rest of the motorway has two. The speed limit on the concession is 120 km/h with sections near toll stations and tunnels having more restrictive speed limits.

The concession runs parallel to the N340/A7, a free alternative dual carriageway with two lanes per direction and grade or signalised intersections at some points along the road.

It is worth pointing out that this corridor is part of a main route for connecting Europe with the North of Africa, thanks to the passage of the Strait of Gibraltar, the final destination of the AP-7 / E-15 highway.

Figure 3: AP-7 / E-15 National Framework

Nowadays the highway AUSOL I presents an Average Annual Daily Traffic of about 14,000 vehicles per day, where the heavy traffic represents 6.2% of the total traffic. After some years of decreasing due to the impact of the economic crisis, the highway is recovering its usual average traffic, that was maximum in 2007 with a total of 20,000 vehicles per day, with a heavy traffic that represents the 6.9% of the total traffic.

In Ausol II, the Average Annual Daily Traffic is over 15,000 vehicles per day with about 11% of heavy vehicles. However, the traffic reached 19,000 vehicles per day in 2007 and heavy vehicles represented nearly 14% of the overall traffic.

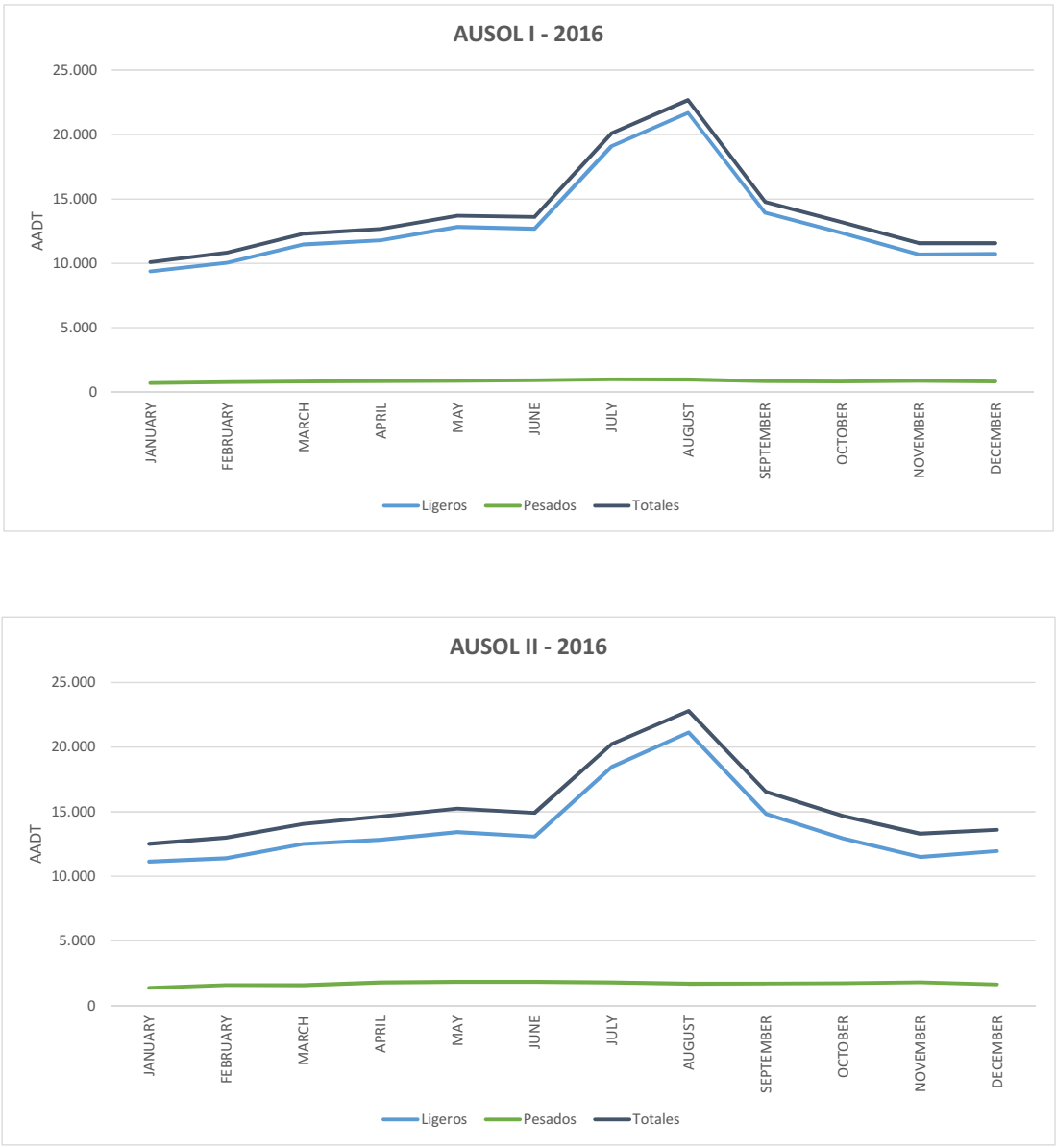


Figure 4: Average Monthly Daily Traffic on the Highway

It is important to note that the highway has a remarkable seasonality. During summer time (July and August), the traffic on the highway has a strong increase, almost double the daily traffic from other weak months such as January or December, due to a high number of tourists in the Málaga area and due to the traffic that crosses the Strait of Gibraltar during the summer time

2.1 Requirements

Pilot requirements and needs are as follows:

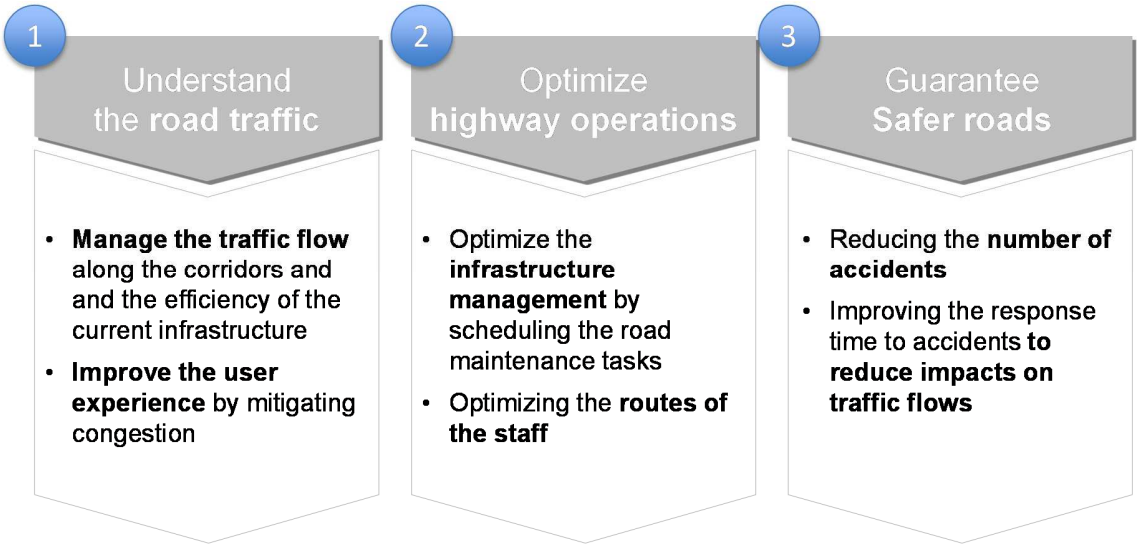


Figure 5: Summary of the requirements

1. **Understand better the road traffic** and mobility patterns

The objective of this pilot is to manage better traffic flows along the corridor and the efficiency of the current infrastructure and to improve the user's experience on the network by mitigating congestion and adjust the supply to their needs. For this purpose, a good understanding of mobility patterns across the corridor is required from all perspectives: origins and destinations, socioeconomic profile of road users, trip purposes, route choice criteria, etc.

2. Optimize **highway operations**

Operations and maintenance tasks can be improved thanks to the analysis of the available data and the optimisation of the routes and scheduling of these activities. This will help improving the efficiency of the infrastructure management.

3. Guarantee **safer roads** and make a better use of these roads

The requirement here consists in reducing the number of accidents, improving the response time to accidents and reducing the impacts on the traffic flows. For this purpose it is necessary to better understand the main causes of accidents and to inform users so they can choose the best route depending on their preferences.

2.2 Objectives

1. The first need to be addressed refers to the **understanding of traffic flows and mobility patterns** in the corridor. This requirement is divided into two main objectives:

   - Develop a **descriptive modelling** of road traffic to analyse:
     - The **mobility patterns** on the road users, i.e., mobility patterns on weekdays, weekends, holidays, especial events, migratory flows (passage of the Strait of Gibraltar) etc.
     - **Impact** of road traffic depending of **the seasonality, type of day, month, hours, etc.**
     - **Highway usage,** taking into account the entries and exits points, routes, etc.
     - Situations of **congestion**

- **Route choice criteria** for different road users
- Traffic **accidents**

- **Forecast traffic flows and the infrastructure use** at short, medium and long term based on external variables (weather, calendar, etc.)

2. As mentioned previously, this pilot aims at **optimizing highway operations** by **prescription of actions to be taken** - with the following objectives:

   - **Staff schedule optimization** on the toll areas based on the mobility patterns knowledge and the traffic forecast. Anticipate changes of operations resources based on short to long-term forecasts**.**
   - **Scheduling of operations and maintenance tasks** along the highway: definition and optimization of maintenance tasks based of the mobility patterns, traffic forecast, weather and other variables.

3. The need for **safer roads** is to be achieved through the objectives below:

   - **Reduce the number of accidents provoked by animals** by predicting, according to probability rules, the risks for this kind of accidents to happen.
   - **Prevent and react to accidents** by analysing their causes and identifying specific circumstances in order to take preventive measures such as specific messages to road users and promoting the use of safer roads. Detect as soon as possible **intrusions** (i.e. animals, landslides, and other non-allowed entities) and **extrusions** (i.e. vehicles that hit safety barriers, get out of the road, etc.

2.3 Use Cases and Scenarios

The technical TT solutions will be designed to answer to several scenarios as described below. Numbers in bullets link objectives defined in previous section with use cases:

	Definition	Justification	Possible use case
1	Describe traffic flows and mobility pattern	The current use of the road infrastructure has to be analyzed to be able to adjust the supply to the demand and make a better use of the corridor.	Analytics of traffic flows and mobility patterns on an average working day
	Predict traffic flows	The anticipation of traffic flows along the corridor allows for a better planning of resources.	Predict traffic flows at each toll plaza at different horizons: at one hour, at one month, at one year
2	Adjust operations resources based on short to long-term forecasts	Traffic flows may vary substantially and this requires a quick adjustment of operations means to avoid any delays at toll stations and any disturbances for road users.	Anticipate a peak in traffic flows and send the appropriate information to the control center
	Optimized routes for operations and maintenance tasks	The current routes do not explicitly take into account external variables such as traffic flows, probability of occurrence of failures, weather statistics, etc.	Define optimized scheduling and routing for different operations and maintenance tasks.
3	Prevent accidents	The analytics of current accidents will bring insights on the main variables that can anticipate and prevent these situations.	Define the situations where a specific message should be addressed to road users
	Analyze and detect intrusion and extrusions	Animals are one of the main causes of accidents along the highway and they are often detected near the main links of the highway. Also it is critical to detect as soon as possible other accidents such as landslides and vehicles that emerge from the pavements	Provide analytics of the intrusions and extrusions

Figure 6: Summary for Use cases and Scenarios

2.3.1 Describe traffic flows and mobility patterns

Definition	Justification	Possible use case
Describe traffic flows and mobility pattern	The current use of the road infrastructure has to be analyzed to be able to adjust the supply to the demand and make a better use of the corridor.	Analytics of traffic flows and mobility patterns on an average working day

- **Definition**

Analytics of traffic flows and mobility patterns at a certain time and at a certain location along the corridor: origins and destinations, country origin, types of vehicles, speed distribution, value of time, level of congestion, accident risk, highway usage

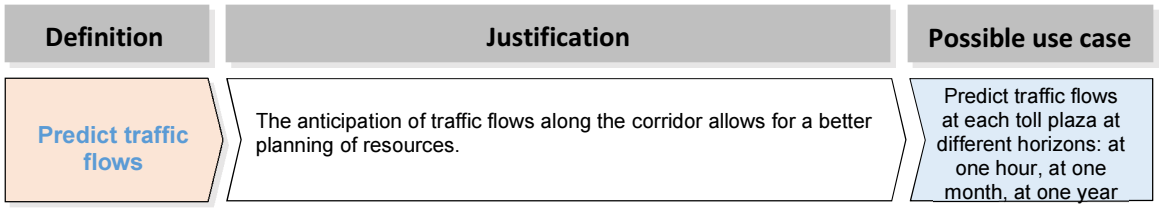
- **Justification**

The current use of the road infrastructure has to be analysed to be able to adjust the supply to the demand and make a better use of the corridor.

- **Possible use case(s)**

Analytics of traffic flows and mobility patterns on an average working day (Monday to Thursday), on Friday, Saturday and Sunday, during the summer time and the rest of the year, per hour and migratory traffic flows. Analytics of route choice criteria, etc.

2.3.2 Predict traffic flows



- **Definition**

Forecast of traffic flows at short, medium and long-term along the corridor: vehicle composition, highway usage, etc. Use all possible explanatory variables such as: toll fares, weather, macroeconomics indices, specific events, calendar, tourism, road connectivity, land use, other transport modes, etc.

- **Justification**

The anticipation of traffic flows along the corridor allows for a better planning of resources.

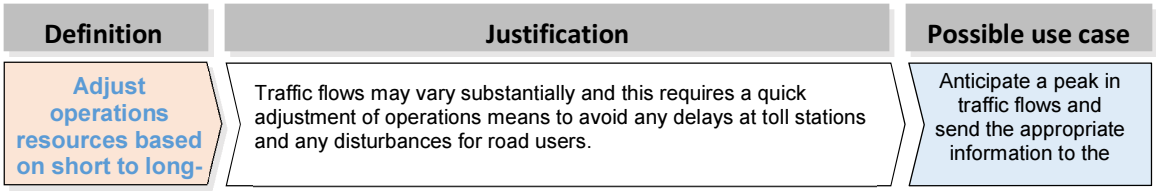
- **Possible use case(s)**

Predict traffic flows at each toll plaza at different horizons: at one hour, at one month, at one year

Predict impacts of specific changes on traffic flows: road connectivity, toll fares, capacity of the road alternative, etc.

Predict migratory flows coming from and/or going to the Passage of the Strait of Gibraltar.

2.3.3 Adjust operations resources based on short to long-term forecasts



- **Definition**

Helping to adjust the operations resources on toll stations (by Full-time Equivalent concepts) by using forecasts over the next time period (hour, week, month, year, etc.): traffic forecasts, risk of accident, etc.

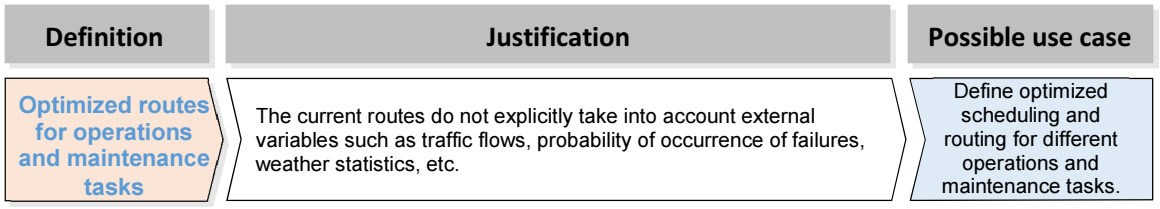
- **Justification**

Traffic flows may vary substantially and this requires a quick adjustment of operations means to avoid any delays at toll stations and any disturbances for road users.

- **Possible use case(s)**

Anticipate a peak in traffic flows and send the appropriate information to the control centre so that appropriate measures can be taken.

2.3.4 Optimised routes for operations and maintenance tasks based on traffic information



- **Definition**

Based on the analysis of current routes followed to undertake current operations and maintenance tasks, define optimised routes that increase efficiency.

- **Justification**

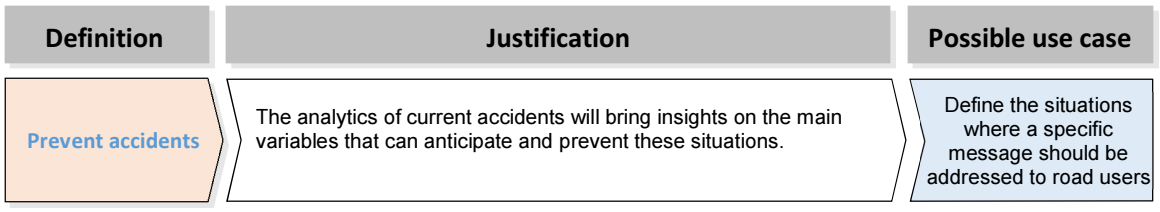
The current routes do not explicitly take into account external variables such as traffic flows, probability of occurrence of failures, weather statistics, etc. As a consequence, they may not be optimised.

- **Possible use case(s)**

Define optimised scheduling for mowing.

Define optimised scheduling and routing for different operations and maintenance task.

2.3.5 Prevent accidents



- **Definition**

Identify situations with a high probability of accidents to inform roads operations teams, users and encourage them to use safer.

- **Justification**

The analytics of current accidents will bring insights on the main variables that can anticipate and prevent these situations.

- **Possible use case(s)**

Define the situations where a specific message on the existing VMS infrastructure should be addressed to road users so that they can be more careful and they can use alternative safer roads.

Analyse the impacts of these measures on the risk of accident, the driver behaviours and on traffic flows.

2.3.6 Analyse and detect intrusion and extrusions

Definition	Justification	Possible use case
Analyze and detect intrusion and extrusions	Animals are one of the main causes of accidents along the highway and they are often detected near the main links of the highway. Also it is critical to detect as soon as possible other accidents such as landslides and vehicles that emerge from the pavements	Provide analytics of the intrusions and extrusions

- **Definition**

Analyse and detect when an animal enters into the highway area and try to predict the potential risk of intrusion in the future, basically based on the historical statistic of accidents caused by animals on the corridor and taking into account the calendar and the meteorological conditions.

Other hazardous effects on the highways will be analysed such us landslides and vehicle that emerge from the pavement.

During the present project we will try to detect and prevent them as soon as possible by means of sensors such as optical fibre, thermal cameras and other systems.

- **Justification**

Animals are one of the main causes of accidents along the highway and they are often detected near the main links of the highway.

- **Possible use case(s)**

Provide analytics of the cases of animal intrusions and emerge of vehicles from the pavement and landslides.

Define preventive measures to avoid and to detect them and act as quickly as possible to avoid any accidents.

2.4 Data Assets

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
Traffic flows - Highway	Traffic volumes at different locations along the highway and at other external locations.	Currently Available	*.csv	https://3.basecamp.com/3320520/buckets/1429164/uploads/427756208
Traffic flows – Outside Highway (surroundings)	Traffic volumes and average speed at different locations along the corridor and at other external locations. Data provided by DGT	Currently Available	*.csv	https://3.basecamp.com/3320520/buckets/1429164/uploads/427756340
Speed Radar	Location of the speed radars (both, fixed and mobile locations) deployed by the Authorities (National, Regional and Local) along the road network.	Currently available	*.xml	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125982
Traffic Events	Traffic events on the road that have an impact on the road traffic, such as roadworks, accidents, traffic congestions, etc.	Currently available	*.xml	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125892
Traffic Events - Highway I	Information related to the events that have happened along the Highway	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427756009
Traffic Events - Highway II	Vehicle queues at toll places along the Highway	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427756106
CCTV Cameras	Last pictures collected by the CCTV cameras deployed on the roads that shows the current status of the traffic flows on real time	Currently available	Location: *.xml Picture: *.jpeg Video: .mp4	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125970
CCTV Cameras-Highway	Videos from CCTV installed in highway's infrastructure	Currently Available	*.mp4	https://3.basecamp.com/3320520/buckets/1429164/uploads/427755800

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
Variable Message Sign	Information showed on the Variable Message Signs which can have an impact on the road traffic	Currently available	*.xml *.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125929
Variable Message Sign - Highway	Information showed on the Variable Message Signs along the highway	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427756463
Population, income and activity data, etc.	This data will be taken at the beginning of the project and it will be introduced as a variable in the Big Data platform	Currently available	*.xls *.cvs	https://3.basecamp.com/3320520/buckets/1429164/uploads/427758439
Meteorological information	Meteorological historic data from the stations located along the corridor	NOT AVAILABLE		
Aemet API	Meteorological data from public stations https://opendata.aemet.es/dist/index.html?#/	NOT AVAILABLE	JSON	
Yahoo Weather API	The Yahoo Weather API allows you to get current weather information for your location.	Currently available	JSON	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125936
Twitter	Information about real time traffic incidents	Currently available	Very diverse depends on the data sought	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125917
Route tracking of vehicles from SoFleet	Additional information provided by other WP in order to enhance tracking vehicles	July 2017	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/429894661
Route tracking of vehicles from maintenance patrols	Additional information provided by Ausol Concession	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427755874
Maintenance Fleet Routes - Highway	Additional information provided by Ausol Concession	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427755874

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
APP mobile	APP for mobile phones with the aim of collecting data about origin and destinations plus travel time and route tracking	2018 (TBC)	JSON	https://3.basecamp.com/3320520/buckets/1429164/uploads/427125965
License Plate Recognition	License Plates recognition records to track migration along European route E15 (DGT)	01/01/2015	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427857223

2.5 Big Data Technology, Techniques and Algorithms

Note: *chapter 2.5 describes the Big Data Technology, Techniques and Algorithms to be used in the project, based on SOFIA2 Platform (by Indra Sistemas, S.A.). SOFIA2 is also the Big Data platform used in "WP6 – Proactive Rail Infrastructures" and "WP8 – Smart Airport Turnaround", so all three packages share platform – but not physical instances.*

2.5.1 Big Data Technology

SOFIA2 (by Indra) is a middleware that allows the interoperability of multiple systems and devices, offering a semantic platform to make real world information available to smart applications (Internet of Things).

It is multi-language and multi-protocol, enabling the interconnection of heterogeneous devices. It provides publishing and subscription mechanism, facilitating the orchestration of sensors and actuators in order to monitor and act on the environment.

Sofia2 is a cross platform that brings an **open source toolset for the Big Data exploit,** as well as multi-device integration through its SDG, APIs and extension mechanisms. Indeed, Sofia 2 community version provides a set of **open APIs** based on the main standards so that any developer can expand the functionality of the platform Sofia for its needs. Using this approach it is not needed any specific skill or training for using Sofia2 beyond the existing open solutions. As communication protocols Sofia2 uses MQTT, RESTful, Ajax Push, Websocket, AMQP and JMS.

Sofia2 supports a wide range of use cases, for example for Smart Cities, Mobility, Energy, Building, Home or Health among others. And using the same approach, the only thing that changes is the semantic definition of the information and the sensors or devices connected.

2.5.1.1 Conceptual model Indra Big Data platform

The main goal for the Sofia2 Big Data Platform is **simplifying the use of all its technologies** and expediting the use and exploitation of data structured - not structured, including also the real time conditions.

Sofia2 is a modular platform that allows to deploy its modules independently according to the needs. All the platform modules are managed from a unified web console, allowing the escalation of the platform depending on the project.

In the picture below shows the whole Sofia's platform – modules in use for pilot highlighted in orange:

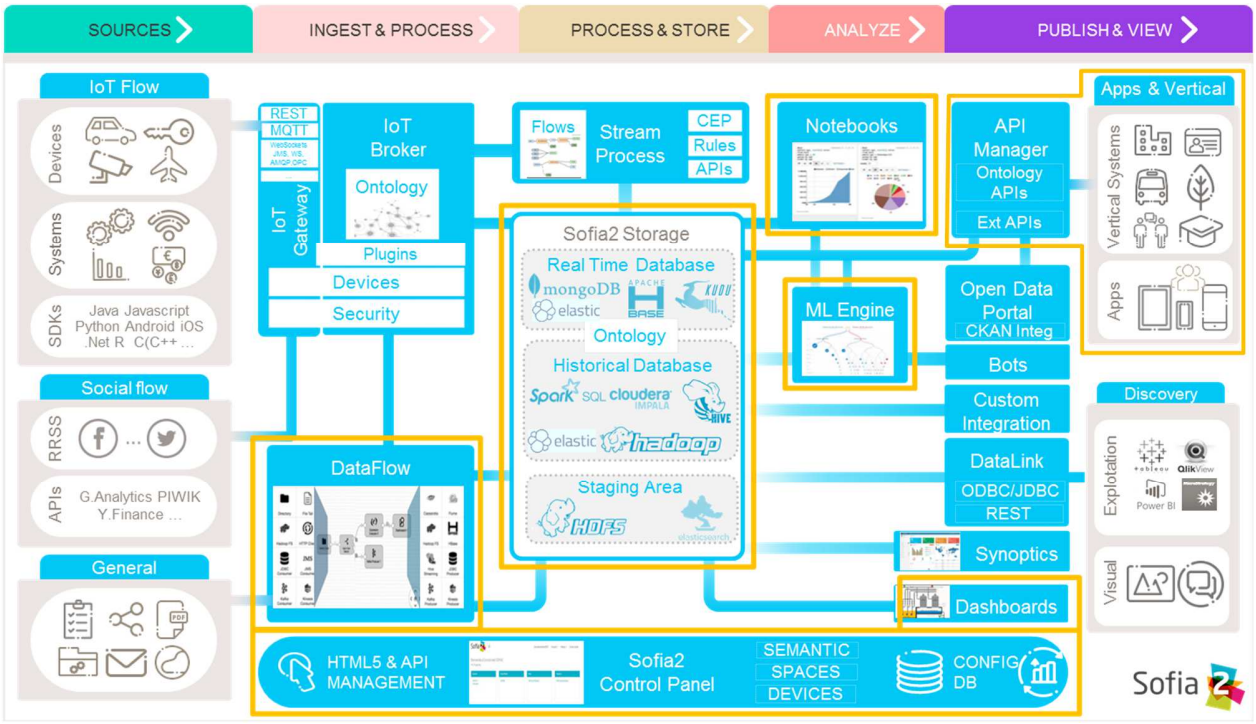


Figure 7: Sofia2 Big Data Platform used modules

2.5.1.2 *Sofia2 Typical workflow*

Workflow and modules needed for WP4 pilots.

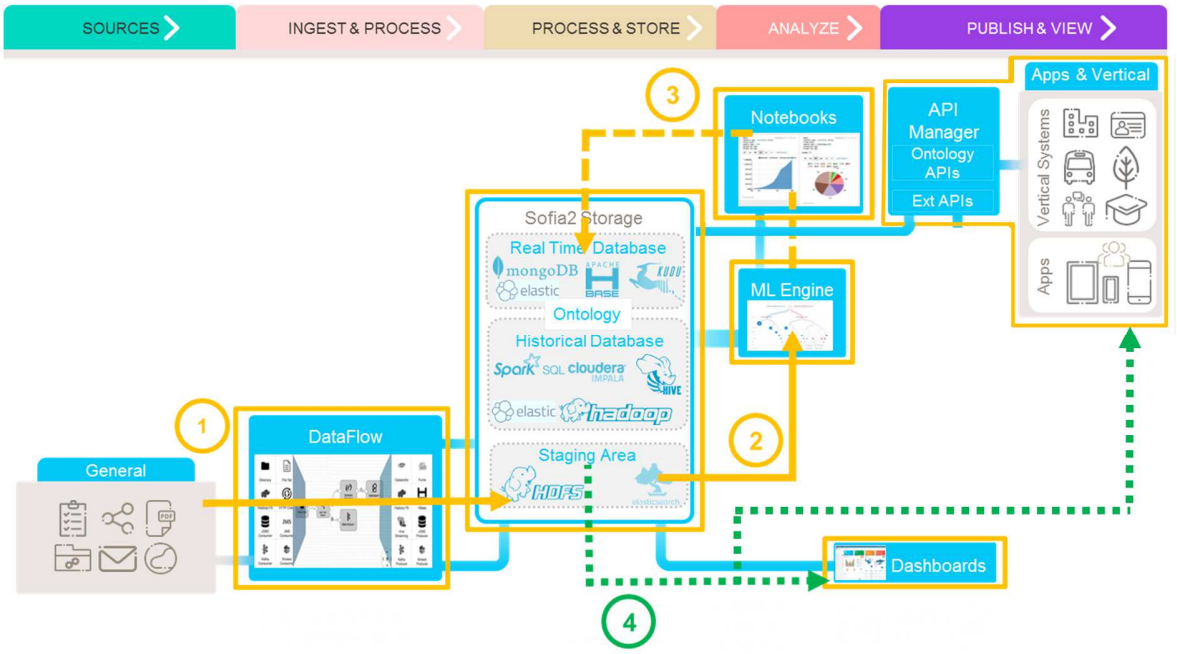


Figure 8: Sofia2 Typical workflow

1- **Upload information**
   Through the DataFlow Module, users define rules for the data processing and data storage in the platform.
2- **Analysis of the information**
   Composed of the data processing (Notebook module) and the Machine Learning Algorithms (ML module)
3- **Storage of models and results**
   Results from the ML Module (Machine Learning Algorithms) are stored on the Storage module
4- **Display & Share information**
   Dashboard module will be in charge of showing the results to the user/operator in an easy an intuitive screen.
   Additionally, through the API Manager, valuable output from the model will feed external systems from the road (i.e. VMS panels, app, etc.) for preventing and reducing congestion, accidents, and others.

2.5.1.3 Sofia2 Storage

The information modeled in the Platform is stored in the Big Data Repository, supported by Hadoop.

Apache Hadoop is an open-source framework that allows the distributed processing of large amounts of data (peta bytes) and working with machine clusters in a distributed way.

Currently Hadoop is synonymous with Big Data for being:

- **Economical:** can run in low-cost equipment, being supported by a cooperative community
- **Scalable:** depending on processing needs, you can add more nodes to cluster system very easily.
- **Efficient** : it can run different processes in parallel.
- **Reliable:** multiple and redundant nodes along the whole cooperative community

The main parts of Hadoop that will be used in the platform are:

- **HDFS (Hadoop Distributed File System):** It is a distributed, scalable and portable file system that allows to store large files across multiple machines
- **Apache Hive:** It is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query and data analysis.
- **Cloudera Impala**: It is a query engine that runs on Apache Hadoop, and brings scalable parallel database technology, enabling users to issue low-latency SQL queries to data stored in HDFS without requiring data movement of transformation.

2.5.1.4 DataFlow (ETL module)

DataFlow module is the main entry of data and information to the platform. This module can be used as an ETL (*Extract, Transform and Load)*, for both purposes: collecting and/or exporting the data in the platform.

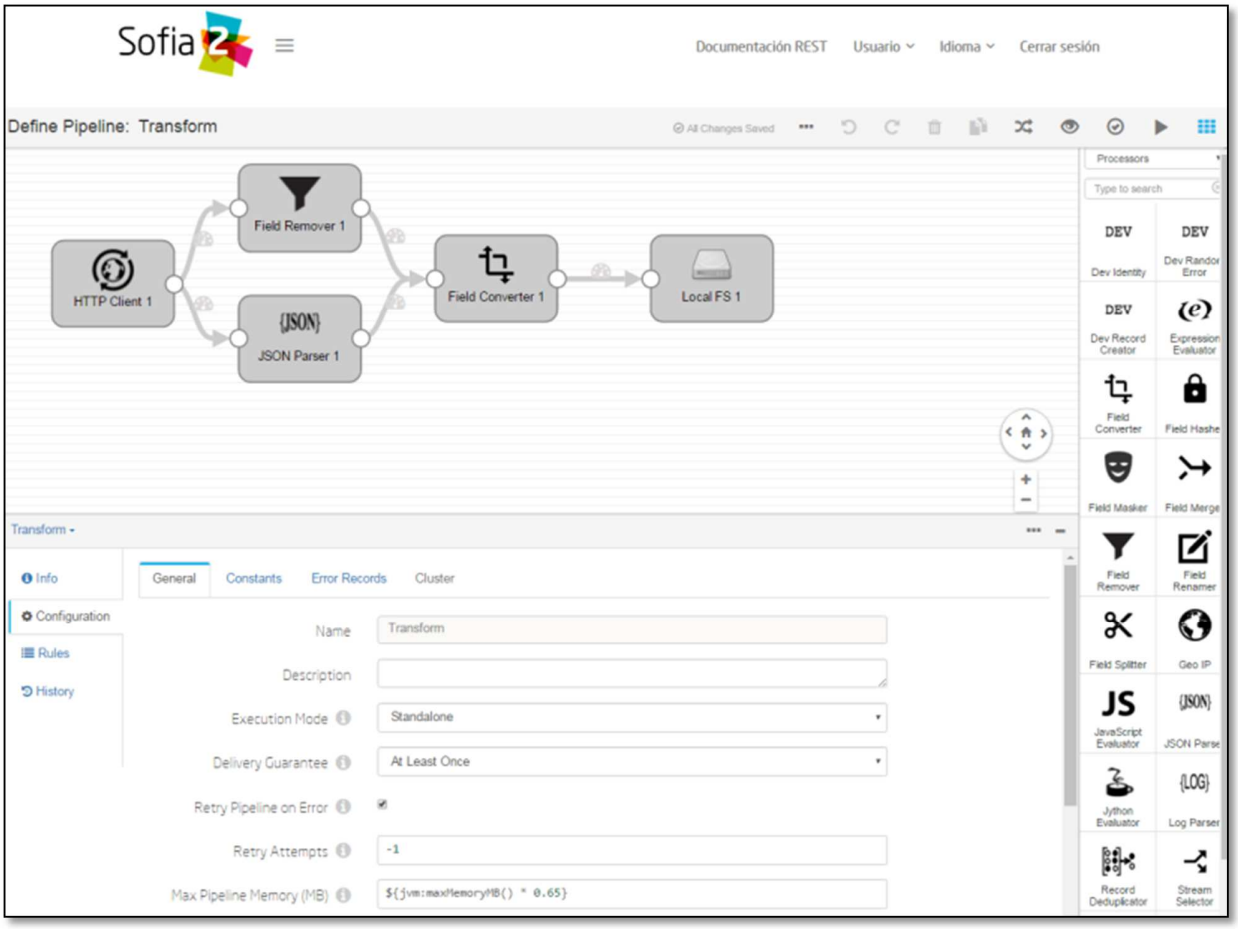


Figure 9: Sofia2 Dataflow module

The main capabilities of the DataFlow module can be summarized as follows:

- **Extraction:** Involves extracting the data from the source system. Sofia2 counts up to 18 different source system where the data extraction process is currently available.

- **Transformation:** It is the process where the data is transformed for being stored in the proper format and/or structure for query and/or analysis purposes .
  It is composed of many different tasks:
  - Evaluation of expressions: performing checks and calculations that can write new fields and/or modify/delete existing fields.

- Actions on fields: data conversion, merge, remove, rename, etc.
- Parsing of JSON, XML and logs
- Flow selector: selection of next steps in the process to be executed
- Evaluators in different programming languages: Python, JavaScript, Jython, etc.

- **Load:** Where the data is loaded into the final target database.
  There are more than twenty possible destinations, to incorporate into the process via Drag & drop from the taskbar. We are highlighting the following components from Sofia2 which lets you to select the ontology, fields, and other additional parameters:
  - AmazonS3
  - Cassandra
  - Hadoop
  - Kafka
  - Flume.

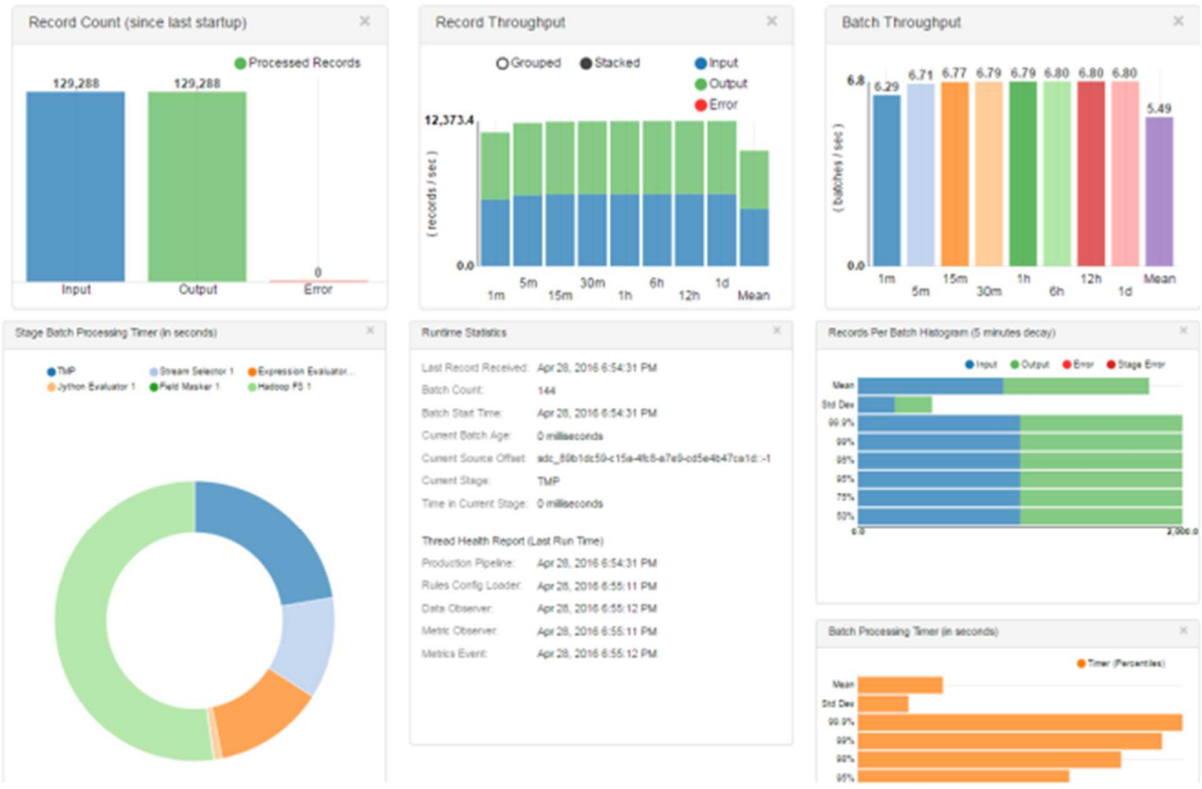


Figure 10: Sofia2 Dataflow monitoring information

2.5.1.5 Notebooks (Collaborative analytical)

The Notebook module is an interactive and intuitive module to display data and to facilitate its analysis. Notebook is basically a collaborative tool that is capable for:

1. Performing complex analysis of information managed by the platform (both, real-time and historical),
2. Combining different programming languages (Spark, R, Python, Hive, SparkSQL and Shell)
3. Generating intuitive figures (such as table, graphs, maps, and others)
4. Planning the execution of procedures (by new notebooks, one per procedure). It mean that you can create different Notebooks with different targets, for example: a Notebook with an specific execution that will be applied in real-time execution, or other Notebook that will be applied based on a specific schedule (hourly, daily, monthly).

As an example, Notebook is able to make a data load from Hadoop Distributed File System (HDFS) to Spark, launching queries and perform complex processes of machine learning through the libraries of MLib, and presenting the data in a map, graph or table.

Figure 11: Sofia2 Notebook

2.5.1.6 Machine Learning

The Machine Learning platform allows users to apply different learning techniques, among which we have to highlight the following:

- **Regression:** Techniques to estimate relationships between variables .

- **Clustering:** Techniques for grouping data by similarities.
- **Classification:** Techniques to identify the membership of an element to a specific group.
- **Recommendation / Prediction:** Techniques for forecasting the value from a new entity based on historical preferences and/or behaviors.

2.5.1.7 Dashboards

Intended to create a simple and visual dashboard with the information managed by the platform. This module offers various types of gadgets (data outputs) which help to generate customized Dashboard

Figure 12: Sofia2 Dashboard types

2.5.1.8 API Manager

SOFIA2 Big Data Platform offers an API Manager module, where valuable data from the platform is available under restful paradigm - and available for subscription from clients.

Additionally, API Manager module connects to external REST services, so it offers an unique and centralized access point to internal and external APIs.

In the scope of this pilot, this module will be used for feeding other systems on the road such as:

- Variable Message Signs.
- Mobile apps based on user's route planning.
- Operation and maintenance teams and route.
- Staff at the toll areas

Those systems will be feed by valuable output from the models that could have a positive impact on the road traffic, for example:

Output from model	Action thru **API Manager**
High risk of road accident	Information on VMS panels about road safety
Congestion	Information on route planner APP for avoiding it
Weather and/or traffic flow pattern	Base restrictions for the O&M tasks scheduler
Peak flow on highway	Increase staff on toll areas

2.5.2 Big Data Techniques and Algorithms

Applying a methodology for data mining processes is an important point to plan and execute such kinds of projects. Some organizations implements KDD (knowledge, discover, datamining) process while others use more specific standards like CRISP-DM (IBM SPSS) or SEMMA (if they are using SAS tools). However, in this project we will use **open software** and mainly we will use **R** and **Python language** and **R Studio tool.**

Data mining or exploitation of information is a process to extract useful, comprehensive and new knowledge with large data volumes. Main objective is to find hidden or implicit information, which cannot be obtained through conventional statistic methods. The inputs for data mining processes are records coming from operational data bases or data warehouses.

We are using a methodology based on **CRISP-DM (*Cross Industry Standard Process for Data Mining)*** with some shortcuts. The major steps are represented in the next figure. From a defined goal where it is implicit the business knowledge it is necessary to prepare data. That data preparation usually includes the data enrichment with Open Data. Afterwards the creation of an advanced model will produce results and require validation. These last three stages (data preparation, creation of advanced models and results validation) constitute a cycle which is iterated until valid results for the business are achieved. You can appreciate the model in a diagram.

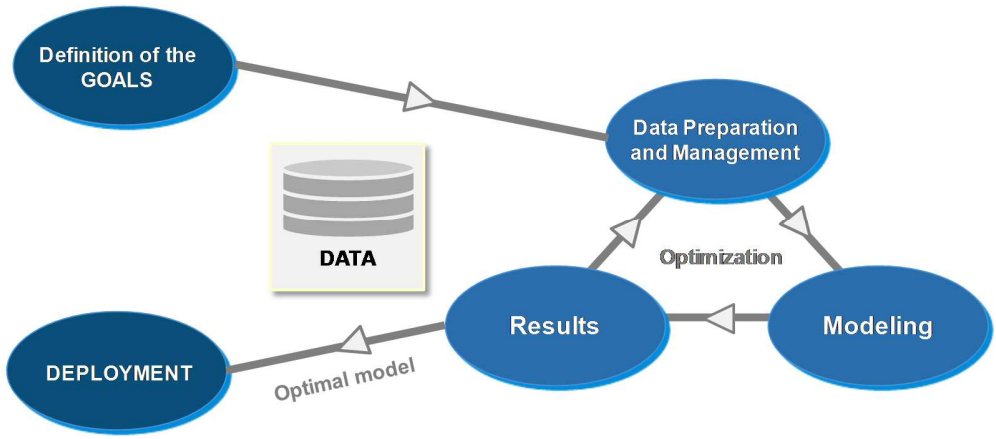


Figure 13: Methodology for advanced data mining

Each stage will be analysed separately so we can provide additional details.

1. Goal definition based on business knowledge
2. Data preparation and management
3. Creating analytic models
4. Validation and conclusions
5. Deployment: Solution integration with the pilot

2.5.2.1 Goal definition based on business knowledge

The first phase is focused on the **understanding what the customer really needs to achieve**. It means that it must be known what are the objectives and requirements to be achieved from a **business point of view**.

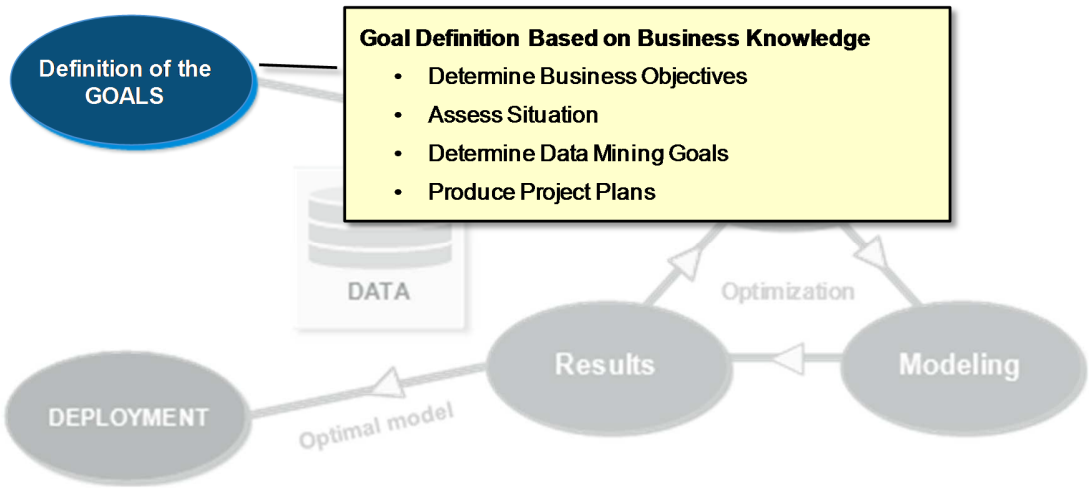


Figure 14: Definition of goals

In this stage, the following steps will be developed:

- **Determine Business Objectives:**
  According to the present pilot, the business objectives have been established in the following 3:

- **Assess situation:**
  This task implies a more detailed research on all the resources, restrictions, assumptions, and other factors that should be considered on the determination of the objectives for the data analysis and the project plan.

- **Determining the Data Mining Goals**
  Definition for the data mining goals and the data mining success criteria.

- **Produce Project Plan**
  Describe the plan for achieving the data mining goals for achieving the Business Objectives.

2.5.2.2 *Data preparation and management*

This stage starts with the collection of the initial data, and continues with the activities that allows to understand the data, identifying the issues and the preparation of the data for building the *useful* data for the model phase

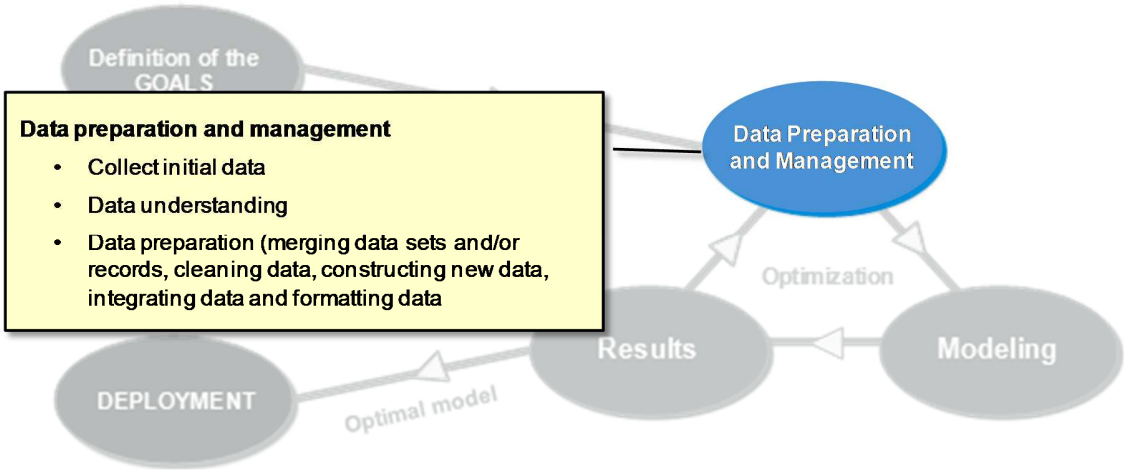


Figure 15: Data preparation and management

So this stage is composed for the following tasks:

- Initial data collection
- Data understandings, which means, describing the data, exploring the data and verifying its quality.
- Data preparation. It is composed of all the tasks and works that will allow the use of the data into the modelling phase: selection of data, data cleaning, data building/creation, integrating, formatting, etc.

2.5.2.3 *Creating analytic models*

In this stage, the modelling techniques (one or more techniques) will be selected to be applied for this specific pilot.

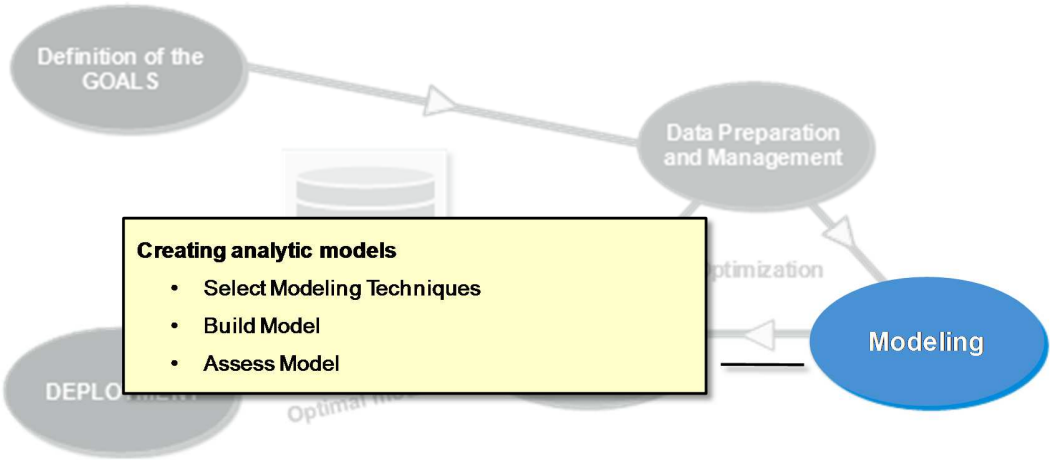


Figure 16: Creating analytic models

Usually there are several model techniques that can be applied for the same issue of data mining, so maybe it is necessary to go back to the Data Preparation and Management stage.

With our methodology we are able to respond to any kind of models: descriptive, diagnostic, predictive and prescriptive. The reader can appreciate the difference in this graphic:

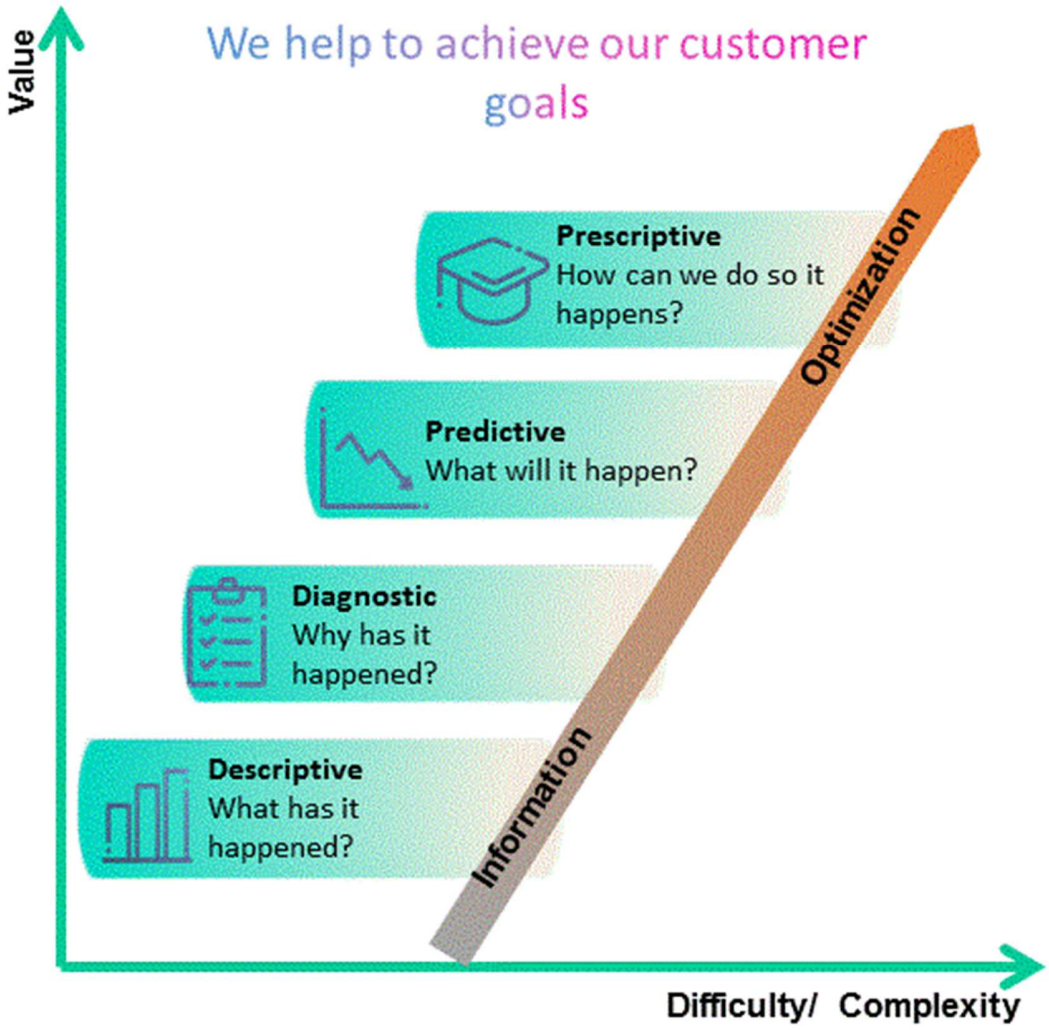


Figure 17: Type of models

As it is shown, the more complex the technique you choose, the more value you can add to your client. In this pilot, it is expected to achieve the **descriptive, predictive and predictive levels.**

One of the main classifications divides machine learning algorithms into two groups:

- Unsupervised algorithms;
- Supervised algorithms.

Unsupervised algorithms are applied when you only have input data and no corresponding output variables. The goal for this technique is to determine the underlying structure or distribution of the data, to organize data by similarity. Examples of application of these

techniques may be customer segmentation, finding hidden patterns, etc... One of the most extended unsupervised algorithms is the K-means algorithm.

On the other hand, **supervised algorithms** try to map a function from the input data to the output variable. In these cases, you know in advance the variable you want to predict. Supervised algorithms are divided into two groups:

- Classification algorithms: the output variable is a categorical one: Fraud not fraud, green red-blue, failure not failure;
- Regression algorithms: the output variable is a real number: A value of a temperature, a pressure, etc.

The next diagram shows the selected type of algorithm for each one of the uses cases and scenarios which have been described earlier in the document:

Use Case	Type of algorithm or model
Describe traffic flows and mobility pattern	• Unsupervised algorithm
Predict traffic flows	• Supervised algorithm. Regression model
Optimised routes for operations and maintenance tasks	• Linear programming model
Adjust operations resources based on short to long-term forecasts	• Supervised algorithm. Regression model
Analyse and detect intrusion and extrusions	• Supervised algorithm. Classification model
Prevent accidents	• Supervised algorithm. Classification model

Figure 18: Definition of the type of model by use case/scenario

2.5.2.4 Model Evaluation

In this stage, a model has been build that seems to have a high quality from a data analysis point of view.

Before starting with the deployment of the model, it is very important to evaluate and review all the executed steps done for the creation of the model, with the aim of comparing with the Business Objectives.

So, the following 3 steps must be deployed:

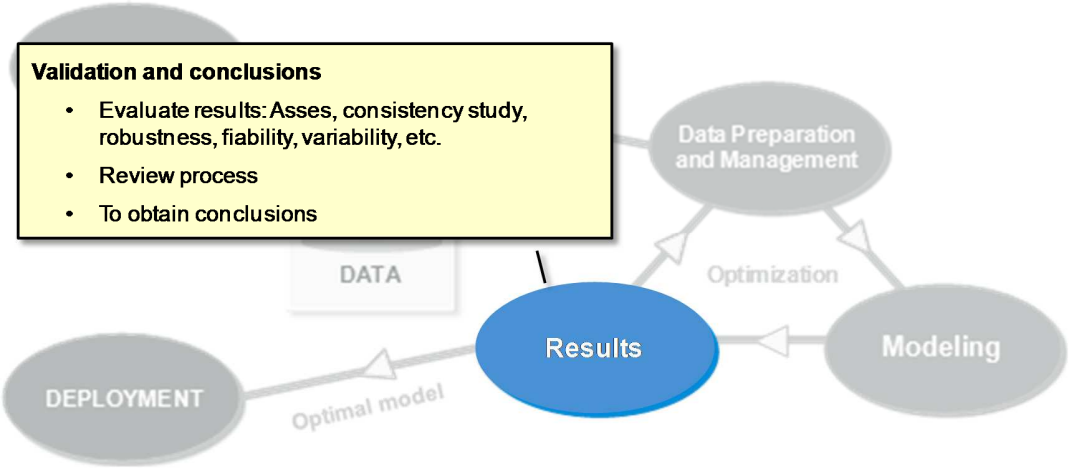


Figure 19: Model validation and conclusions

- **Evaluation results:** Once the more appropriate type of algorithms has been chosen, a procedure to test the model quality and the validity is needed. So data are divided into sampling data for training the model (the algorithm learns from the past) and the other for testing (the accuracy of the algorithm is tested) as the next figure depicts.
- **Review process**. Review the steps looking for mistakes, errors and others
- **Obtaining conclusions** and next steps

2.5.2.5 Deployment: Solution integration with the pilot

Once the model has been created, the last stage is the deployment into the pilot.

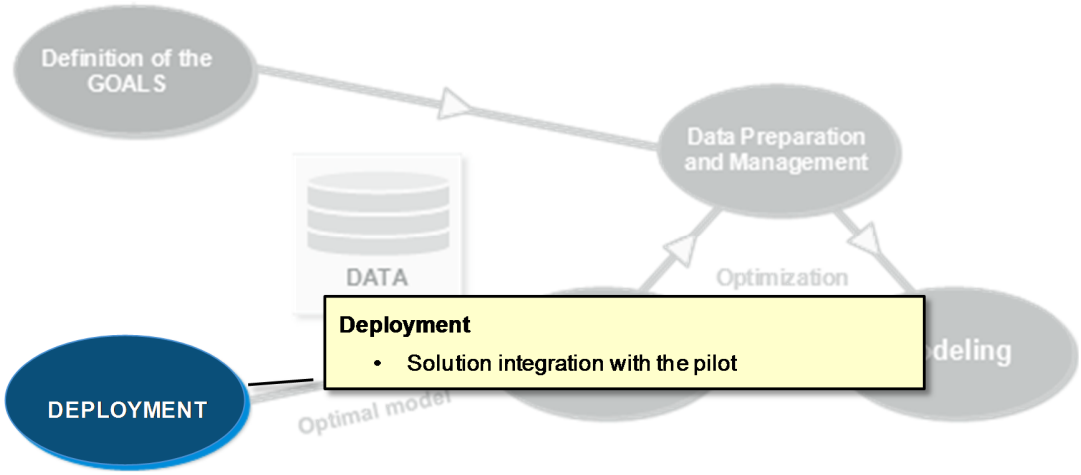


Figure 20: Deployment

The following tasks are recommended to be applied on this stage to guarantee a successful deployment of the model into the pilot:

- Development of a plan. This task takes the results from the evaluation stage and determine an strategy for the deployment.
- The supervision and maintenance of the deployment and results, by the supervision of the steps defined in the previous bullet
- Final report and final revision of the project, evaluating all the steps and results obtained. Here it is important to document both, the success and the fails that will allow for leaning lessons in future projects.

2.6 Positioning of Pilot Solutions in BDVA Reference Model

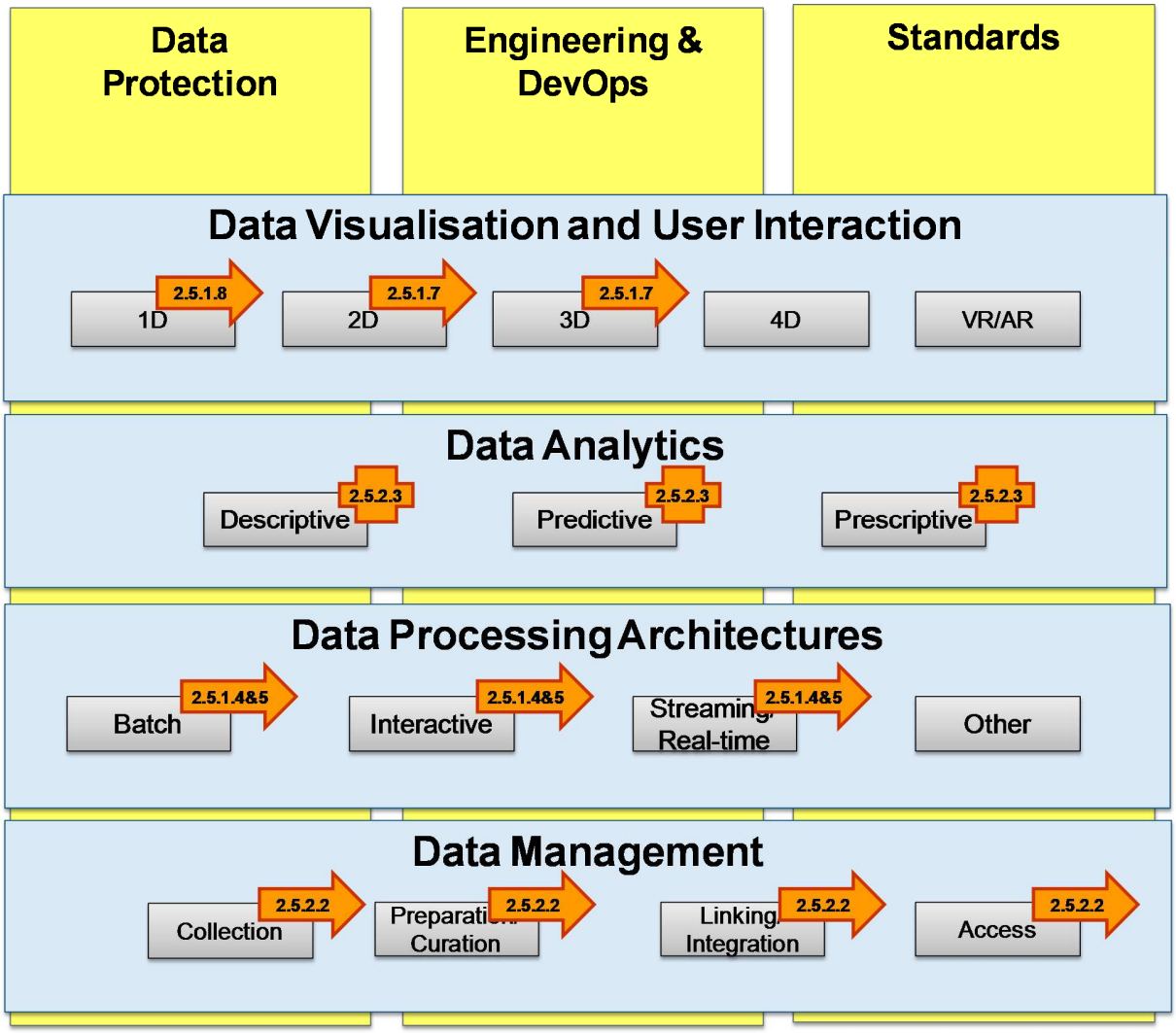


Figure 21: Pilot Big Data components in the BDVA Reference Model

Several of the technical components provide engineering and deployment support. For example, the Indra solutions are part of Sofia2 Big Data Platform.

- ***Data Visualisation and User Interface***
  The pilot will provide a set of reports that will allow the visualization of the information in a readable and useful format on each one of the predictions made so that they can be of help for the decision making.

The output format will be 1D and 2D. Specification and examples are attached within the chapter "*2.5.1.7 Dashboards*".

- ***Data Analytics***
  As part of the present pilot, it will be developed specific algorithms based on predictive data to predict the evolution of each elements of the pilot.

  Firstly a descriptive analysis will be carried out to get a full understanding of the data.

  Secondly, a predictive analysis will be carried out for forecasting the mobility patterns in the short, medium and long term.

  And finally, some valuable information (such as probability of accidents, congestion and others) will feed other devices on the highways (i.e., VMS panels, apps and others), it will have an impact on the traffic, so it means that a prescriptive model also will take place on the project.

  More information you can find them in the chapter "*2.5.2.3 Creating analytic models*".

- ***Data Processing Architectures***
  Processes that allow the feeding of the algorithms with new data collected during the execution of the pilot. The inclusion of new data will be a periodic task given the nature of the data sources.

  Related chapters: "*2.5.1.4 DataFlow (ETL module)*" and "*2.5.1.5 Notebooks (Collaborative analytical)*".

- ***Data Management***
  The identified initial data sources provide information to the pilot in standard formats based on Excel, Pdf and XML files. All these sources will be treated to allow their initial inclusion and the insertion of data progressively throughout the execution of the pilot.

  For more detailed description see "*2.5.2.2 Data preparation and management*"

2.7 Big Data Infrastructure

During Stages 1 and 2, the SOFIA2's infrastructure will be shared between the following 3 pilots:

- WP4 – Smart Highways
- WP6 – Proactive Rail Infrastructures
- WP8 – Smart Airport Turnaround

The following figure shows an overview of the platform, dimensioned for the estimated data volumes in the project for these 2 stages.

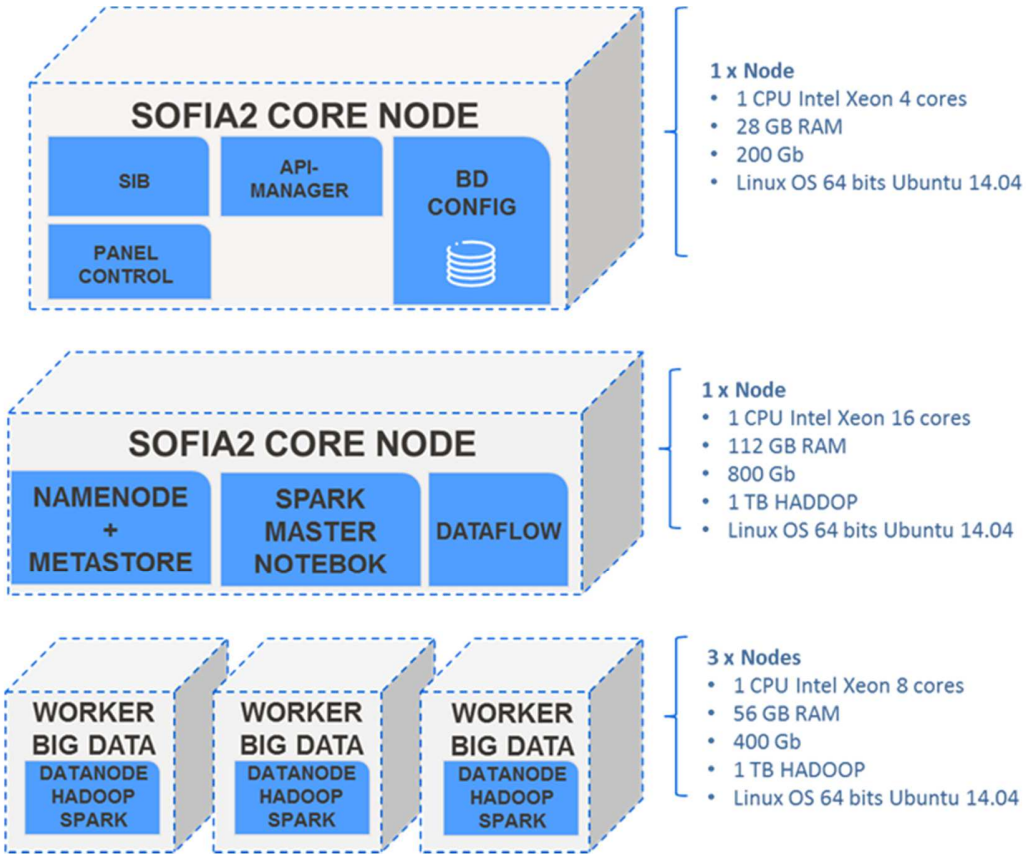


Figure 22: Sofia2 Big Data Platform Infrastructures

During Stages 3, the Big Data Infrastructure will be deployed on the Road Operator's infrastructure: Data hosting will be hosted in San Pedro Datacenter (MALAGA). The architecture consists of a VMware Virtualized Environment supported by one cluster of 4 servers HP (128 GB RAM, RAID 1+HOT SPARE 256GB and 80 cores each) and EMC storage Systems. (80 TB of disk space). The network environment provides redundant connection of fiber channel for storage systems and gigabit Ethernet connections for servers.

2.8 Roadmap

The Ausol Highway pilot follows a **3-stage methodology** for validation and demonstrating the scalability of Big Data solutions. The three stages differ with respect to the embedding of the technology into the operational environment, the form of the Big Data infrastructure being used, as well as the scale of data that is exploited. Starting with a focused technology validation, the 3-stage methodology will ultimately deliver insights from actual in-situ trials, i.e. a prototypical implementation of the solutions in the field, involving real end-user and actual production data. Table below shows the three stages of the methodology along its three main aspects:

Stage	Delivery Date (Project Month)	Features / Objectives Addressed	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1: Technology Validation	M6	**GOALS** -Understand the Business Objectives and determine their feasibility stated during the design stage, based on the existing data. -Definition of the data mining goals on the selected Business Objectives -Development of the Project Plans for achieving the data mining goals **DATA PREPARATION AND MANAGEMENT** -Data understanding -Data assessment (quality concepts: i.e. valuable fields, consistency, hollows, etc.) **RESULTS OBTAINED IN S1** - Acceptance/rejection of the Business Objectives from S0- Data Quality assessment (valuable data for modelling) - Estimation of the scale of data for the Pilot. - Temporary productive environment	Yes Low Scale During Stage 1, a temporary an productive environment will be created for S1 and S2.	Big Data infrastructure during Stage 1 will be from Indra's Infrastructure. More information, please see chapter "**2.7 Big Data Infrastructure**"	The scale of data will be defined in the end of S1

Stage	Delivery Date (Project Month)	Features / Objectives Addressed	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S2: Large-scale experimentation and demonstration	M12	**DATA PREPARATION AND MANAGEMENT** - The valuable data for modelling will be accordingly prepared for the Machine Learning process **MODELING** - The model techniques will be selected and applied for each Business Objective and the 3 models will be built - with historical and simulated data - Evaluation of results and review the model process - First conclusions **RESULTS OBTAINED IN S2** **-** A first descriptive and predictive model will be completed **-** The results of the model will be presented in a draft dashboard and API Manager	Yes Large scale During Stage 2, the temporary and productive environment will include historical data and additionally will include simulated data that will emulate the real-time data from S3	Big Data infrastructure during Stage 2 will be from Indra's Infrastructure. More information, please see chapter "**2.7 Big Data Infrastructure**"	To be defined once the data be processed
S3: In-situ trials	M24	**DEPLOYMENT** **Refining descriptive and predictive models from S2** Development of a Plan for a controlled deployment - Platform connected with real-time data - Supervision and maintenance of the deployment and results **RESULTS OBTAINED IN S3** -- Fully functional descriptive and Predictive model **based on historical and real time data.** - Big Data platform will be deployed and will be running in the road operator's infrastructure	Yes Large scale Road operator's infrastructure	Big Data infrastructure during Stage 3 will be running in the road operator's infrastructure. More information, please see chapter "**2.7 Big Data Infrastructure**"	To be defined once the data be processed

3 Design of Replication Pilot

The replication pilot is the 119-kilometre Norte Litoral highway. This corridor connects the cities of Oporto, Caminha and Ponte da Lima in the North of Portugal. The A-28 highway is operated with an electronic toll system with a total absence of obstacles for users.

Figure 23: Influence area for the Pilot

Euroscut Norte Litoral is a dual carriageway dual lane motorway concession, from Oporto to Caminha (A28, 90km) and from Viana do Castelo to Ponte de Lima (A27, 23km) located in northern Portugal. There are four tolling points and operations started in October 2010.

Nowadays the Norte Litoral highway presents an Average Annual Daily Traffic of about 22,000 vehicles per day, where the heavy traffic represents 6.5% of the total traffic. After some years of decreasing due to the impact of the economic crisis, the highway is recovering its usual average traffic, that was maximum in 2009 with a total of nearly 30,000 vehicles per day, with a heavy traffic that represented the 6.6% of the total traffic.

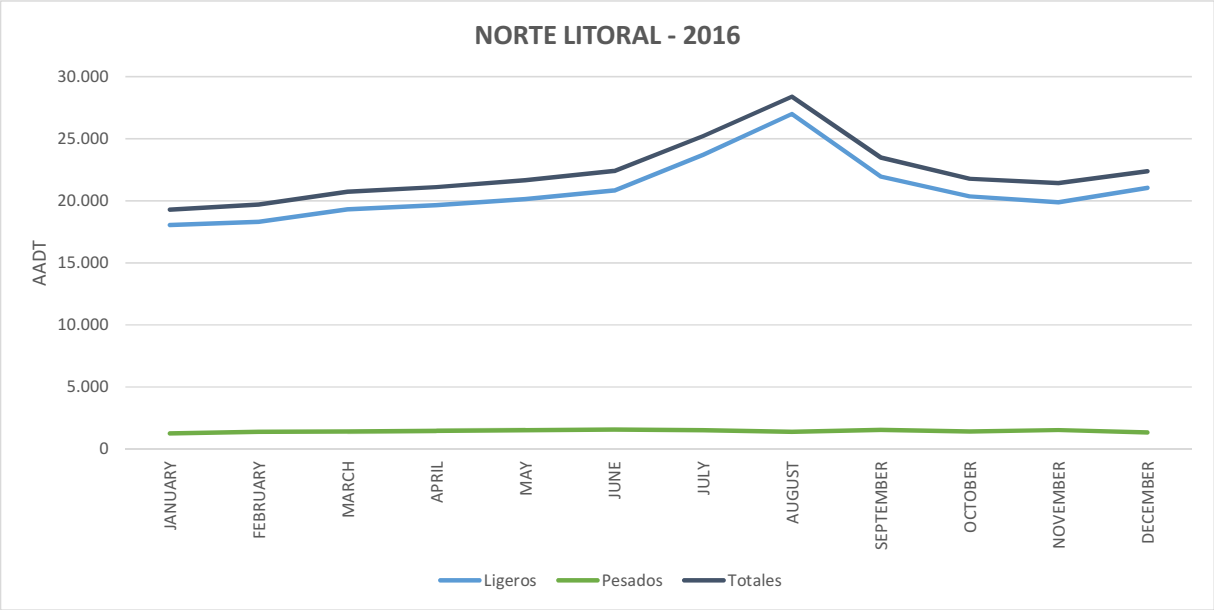


Figure 24: Average Monthly Daily Traffic on the Highway

3.1 Requirements

Detailed in chapter 2.1

3.2 Objectives

Detailed in chapter 2.2

3.3 Use cases and scenarios

Detailed in chapter 2.3

3.4 Data assets

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
Traffic flows - Highway	Traffic volumes at different locations along the corridor and at other external locations.	Currrently Available	*.csv	https://3.basecamp.com/3320520/buckets/1429164/uploads/428921845/download/TT_WP4.3_TrafficFlow_NorteLitoralHighway_v1.0.xlsx
Traffic flows – Outside Highway (surroundings)	Traffic volumes and AVG speed at different locations along the corridor and at other external locations.	Available Soon	TBC	
Speed Radar	Location of the speed radars (both, fixed and mobile locations) deployed by the Authorities (National, Regional and Local) along the road network.	Available without historical	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/428916560/download/TT_WP4.3_SpeedRadar_NorteLitoralHighway_v1.0.xlsx
Traffic Events	Traffic events on the road that have an impact on the road traffic, such as roadworks, accidents, traffic congestions, etc.	Available without historical	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/428917164/download/TT_WP4.3_TrafficEvents_NorteLitoralHighway_v1.0.xlsx

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
Traffic Events - Highway I	Information related to the events that have happened along the Highway	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/428917586/download/TT_WP4.3_Traffic%20events%20-%20Highway%20I_NorteLitoralHighway_v1.0.xlsx
CCTV Cameras-Highway	Videos from CCTV installed in highway's infrastructure	Currently Available	Video: *.mp4	https://3.basecamp.com/3320520/buckets/1429164/uploads/428918365/download/TT_WP4.3_CCTV%20Cameras%20-%20NorteLitoralHighway_v1.0.xlsx
Variable Message Sign - Highway	Information showed on the Variable Message Signs along the highway	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/428923604/download/TT_WP4.3_PMV_NorteLitoralHighway_v1.0.xlsx
Population and socio-economics	This data will be taken at the beginning of the project and it will be introduced as a variable in the Big Data platform	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/427757964/download/TT_WP4.3_SocioEconomic_NorteLitoral_v1.0.xlsx
Meteorological information	Meteorological historic data from the stations located along the corridor	Currently available	*.xls	https://3.basecamp.com/3320520/buckets/1429164/uploads/428924436/download/TT_WP4.3_METEO_NorteLitoralHighway_v1.0.xlsx
Yahoo Weather API	The Yahoo Weather API allows you to get current weather information for your location.	Currently available	JSON	https://3.basecamp.com/3320520/buckets/1429164/uploads/427758016/download/TT_WP4.3_YahooWeatherAPI_NorteLitoral_v1.0.xlsx

Name of Data Asset	Short Description	Initial Availability Date	Data Type	Link to Data ID Card (in basecamp)
Twitter	Information about real time traffic incidents	Currently available	Very diverse depends on the data sought	https://3.basecamp.com/3320520/buckets/1429164/uploads/427757994/download/TT_WP4.3_Twitter4_NorteLitoral_v1.0.XLSX
Route tracking of vehicles From TomTom or SoFleet (TBC)	Additional information provided by other WP in order to enhance tracking vehicles	Not defined yet	TBC	
Maintenance Fleet Routes - Highway	Additional information provided by Norte Litoral Concession	Available Soon	TBC	
License Plate Recognition	License Plates recognition	Not defined yet	TBC	

3.5 Big Data Technology, Techniques and Algorithms

Detailed in chapter 2.5

3.6 Positioning of Pilot Solutions in BDVA Reference Model

Detailed in chapter 2.6

3.7 Big Data Infrastructure

Detailed in chapter 2.7

3.8 Roadmap

Within Norte Litoral´s pilot, it follows the same **3-stage methodology** than for Malaga's pilot for validation and demonstrating the scalability of Big Data solutions. The three stages differ with respect to the embedding of the technology into the operational environment, the form of the Big Data infrastructure being used, as well as the scale of data that is exploited. Starting with a focused technology validation, the 3-stage methodology will ultimately deliver insights from actual in-situ trials, i.e., a prototypical implementation of the solutions in the field, involving

real end-user and actual production data. The three stages of the methodology along its three main aspects are explained below:

Stage	Delivery Date (Project Month)	Features / Objectives Addressed	Embedding in Productive Environment	Big Data Infrastructure Used	Scale of Data
S1: Technology Validation	M9	**DEFINITION OF THE GOALS** - Understand the Business Objectives and determine their feasibility stated during the design stage, based on the existing data. - Definition of the data mining goals on the selected Business Objectives - Development of the Project Plans for achieving the data mining goals **DATA PREPARATION AND MANAGEMENT** - Data understanding - Data assessment (quality concepts: i.e. valuable fields, consistency, hollows, etc.) **RESULTS OBTAINED IN S1** - Acceptance/rejection of the Business Objectives from S0 - Data Quality assessment (valuable data for modelling) - Estimation of the scale of data for the Pilot. - Temporary productive environment	Yes Low Scale During Stage 1, a temporary an productive environment will be created for S1 and S2.	Big Data infrastructure during Stage 1 will be from Indra's Infrastructure.	The scale of data will be defined in the end of S1

S2: Large-scale experimentation and demonstration	M15	**DATA PREPARATION AND MANAGEMENT** - The valuable data for modelling will be accordingly prepared for the Machine Learning process **MODELING** - The model techniques will be selected and applied for each Business Objective and the 3 models will be built - with historical and simulated data. - Evaluation of results and review the model process - First conclusions **RESULTS OBTAINED IN S2** **-** A first descriptive and predictive model will be completed **-** The results of the model will be presented in a draft dashboard and API Manager	Yes Large scale During Stage 2, the temporary and productive environment will include historical data and additionally will include simulated data that will emulate the real-time data from S3	Big Data infrastructure during Stage 2 will be from Indra's Infrastructure.	To be defined once the data be processed
S3: In-situ trials	M27	**DEPLOYMENT** **- Refining descriptive and predictive models from S2** -Development of a Plan for a controlled deployment - Platform connected with real-time data -Supervision and maintenance of the deployment and results **RESULTS OBTAINED IN S3** - Fully functional descriptive and Predictive model based on historical and real time data. Big Data platform will be deployed and will be running in the road operator's infrastructure	Yes Large scale Road operator's infrastructure	Big Data infrastructure during Stage 3 will be running in the road operator's infrastructure.	To be defined once the data be processed

4 Commonalities and Replication

4.1 Common Requirements and Aspects

Firstly it is important to say that in the replication project not only the objectives but also the requirements are very similar.

Related to the common requirements **the data that will be initially provided** in both concessions is as follows:

- Traffic flows
- Speed
- Accidents data
- Videos
- Road messages
- Meteorological information
- Location of each device along the road

In addition we are currently working on the search of additional data sources that can add value to both pilots.

It should also be noted that there is some data that is available in one concession and not another, so in this case data only adds value to the pilot in particular

4.2 Replication

As it is shown below, the pilot domains will start with an initial pilot at the beginning of the project, and then **replicate the solutions by reusing the results** as part of a replication pilot.

The replication pilot considers insights, findings and lessons learned from the initial pilot. This replication approach is one key means to demonstrate the reusability and generic nature of the TT solutions. A replication pilot addresses similar and related objectives as the initial pilot, but typically adds a further level of complexity; e.g. in terms of processes or data assets. In addition, as indicated.The systems and IT architecture have been already deployed in both pilot sites and collection of data over time has been consolidated so data set is increasing constantly.

Specific tests on models and algorithms will be conducted in order to assess the robustness of algorithms and models set up during the stages 1 and 2.

The trial will consist of **running the models during a long period**, enhancing the results of the models by a constant calibration capacity of learning. A critical analysis of results and evaluating the results and benefits.

It is also expected to test the impact of this new tool in the predictability of special events and deployed the solutions in line with the strategies met during the whole project. The consortium

will disseminate results, advertise findings and recommendations of the in situ trial generated by the application of Big data analytics.

All the common data will be managed in order to obtain results related to the objectives of each pilot. However some **technical risks** has been detected:

- Possible lack of some specific datasets to understand mobility patterns (GPS tracks, origin-destination information).
- Difficulties to integrate systems developed within the project with existing road operator system.
- Algorithms developed in the main pilot are not suitable for replication.
- Sensitive information managed in the datasets.
- Algorithms developed need more iterations for refinement than expected.

5 Conclusions

Ausol Pilot and its replica in Norte-Litoral share main objectives and requirements, although the concessionary companies might focus on a different subset of results: by having different business models, they might perceive higher benefits from implementing solutions to optimize infrastructure management, prior to reduce response time to incidents. However, apart from the different data assets needed to feed the Big Data platform, pilot and replica can be considered conceptually as a single extended case.

Hence, it's no coincidence that both pilots share common objectives and requirements – indeed objectives were described to fit into general business model at Cintra, and not only single concession's. Thus, a partial replication of the algorithms developed for main pilot will be easily packaged for replication "as is"– and that's in the core of the more general objective of the project: replicating all methodologies and technologies developed within pilots on new highway concessions.