D8.1 – Smart Airport Turnaround Pilot Design

<table>
<thead>
<tr>
<th>Project Acronym</th>
<th>TT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>Transforming Transport</td>
</tr>
<tr>
<td>Grant Agreement number</td>
<td>731932</td>
</tr>
<tr>
<td>Call and topic identifier</td>
<td>ICT-15-2016-2017</td>
</tr>
<tr>
<td>Funding Scheme</td>
<td>Innovation Action (IA)</td>
</tr>
<tr>
<td>Project duration</td>
<td>30 Months [1 January 2017 – 30 June 2019]</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Mr. Rodrigo Castiñeira (INDRA)</td>
</tr>
<tr>
<td>Website</td>
<td><a href="http://www.transformingtransport.eu">www.transformingtransport.eu</a></td>
</tr>
<tr>
<td>Project Acronym</td>
<td>TT</td>
</tr>
</tbody>
</table>
## Document fiche

| Authors: | Julen Oguiza (AG), Benjamin Moreno (AG), Juan Antonio Ubeda (Indra), Nikolaos Papagianopoulos (AIA), Harris Markopoulos (AEGEAN), Juan Francisco Garcia Lopez (INDRA), Anna-Lisa Mautes (JEPP), Niels Stark (JEPP), David Scarlatti (BRTE) |
| Internal reviewers: | Daniel Clavero (Lince)  
George Dimitrakopoulos (Intra) |
| Work Package: | WP8 |
| Task: | T8.1-T8.3 |
| Nature: | R |
| Dissemination: | PU |

## Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Contributor(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>01/03/2017</td>
<td>Juan Francisco Garcia Lopez (INDRA), Anna-Lisa Mautes (JEPP)</td>
<td>Initial draft with Initial Pilot information</td>
</tr>
<tr>
<td>0.2</td>
<td>05/03/2017</td>
<td>Juan Francisco Garcia Lopez (INDRA), Anna-Lisa Mautes (JEPP)</td>
<td>Revision of initial pilot scope and completing pending sections</td>
</tr>
<tr>
<td>0.3</td>
<td>10/03/2017</td>
<td>Juan Francisco Garcia Lopez (INDRA)</td>
<td>Completing sections 2.5, 2.6, 2.7 and 2.8</td>
</tr>
<tr>
<td>0.4</td>
<td>27/03/2017</td>
<td>Juan Francisco Garcia Lopez (INDRA)</td>
<td>Consolidated comments by Andreas Metzger</td>
</tr>
<tr>
<td>0.5</td>
<td>30/03/2017</td>
<td>Juan Francisco Garcia Lopez (INDRA), David Scarlatti (BRTE)</td>
<td>Consolidated comments by Daniel Clavero</td>
</tr>
<tr>
<td>1.0</td>
<td>31/03/2017</td>
<td>Niels Stark (JEPP)</td>
<td>Consolidated inputs for both pilots. Added overall motivation section. Added roadmap table to replication pilot. Input to common section. Final formatting.</td>
</tr>
</tbody>
</table>
Keywords: Design, Big Data Platform, Requirements, Objectives, Use Cases/Scenarios

Abstract (few lines): This deliverable reports on the work planned in WP8. It describes the objectives, the design, the use cases, the data assets and the technologies to be applied and used in the Initial Pilot and the Replication Pilot.

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 Nº 731932
Table of Contents

LIST OF TABLES ................................................................................................................................. 7
DEFINITIONS, ACRONYMS AND ABBREVIATIONS .............................................................................. 8
EXECUTIVE SUMMARY ............................................................................................................................ 9
1 OVERALL MOTIVATION AND AMBITIONS FOR PILOT DOMAIN .................................................... 10
2 INITIAL PILOT: SMART PASSENGER FLOW ...................................................................................... 12
  2.1 REQUIREMENTS ............................................................................................................................... 13
    2.1.1 Functional requirements: .......................................................................................................... 13
    2.1.2 Non-functional requirements: .................................................................................................... 14
      2.1.2.1 Security ................................................................................................................................ 14
      2.1.2.2 Operation ................................................................................................................................ 14
      2.1.2.3 Interfacing .............................................................................................................................. 14
  2.2 OBJECTIVES .................................................................................................................................... 15
  2.3 USE CASES / SCENARIOS .............................................................................................................. 20
    2.3.1 Scenario 1: Identification of passengers’ arrival time patterns to the airport terminal ........ 20
      2.3.1.1 Objectives .............................................................................................................................. 20
      2.3.1.2 Contribution/benefits: ........................................................................................................ 21
      2.3.1.3 Input data ............................................................................................................................... 21
      2.3.1.4 Example: ................................................................................................................................. 21
    2.3.2 Scenario 2: Extraction of passengers’ movement patterns along the airport terminal ........ 21
      2.3.2.1 Objectives .............................................................................................................................. 21
      2.3.2.2 Contribution/benefits: ........................................................................................................ 22
      2.3.2.3 Input data ............................................................................................................................... 22
      2.3.2.4 Example: ................................................................................................................................. 22
    2.3.3 Scenario 3: Anticipate detection of transfer passengers with short time to take their departure flights. 22
      2.3.3.1 Objectives .............................................................................................................................. 23
      2.3.3.2 Contribution/benefits: ........................................................................................................ 23
      2.3.3.3 Input data ............................................................................................................................... 23
      2.3.3.4 Example: ................................................................................................................................. 24
    2.3.4 Scenario 4: Extraction of passengers’ behaviour features and patterns regarding airport services (retailing, food and beverage...)* ........................................................................ 24
      2.3.4.1 Objectives .............................................................................................................................. 24
      2.3.4.2 Contribution/benefits: ........................................................................................................ 24
      2.3.4.3 Input data ............................................................................................................................... 25
      2.3.4.4 Example: ................................................................................................................................. 25
  2.4 DATA ASSETS ................................................................................................................................... 25
  2.5 BIG DATA TECHNOLOGY, TECHNIQUES AND ALGORITHMS .................................................... 26
    2.5.1 Detailed Explanation of Big Data Technology ........................................................................... 26
      2.5.1.1 Conceptual model of Athens platform .................................................................................. 26
      2.5.1.2 Conceptual model Indra Big Data platform ........................................................................ 28
    2.5.2 Detailed Explanation of Technology, Techniques and Algorithms ........................................... 41
      2.5.2.1 Goal definition based on business knowledge ...................................................................... 42
      2.5.2.2 Data preparation and management ...................................................................................... 43
Table of Figures

Figure 1: Evolution of Passengers in AIA ................................................................. 12
Figure 2: Passenger flow process and check points .................................................. 16
Figure 3: Transfer Passengers .............................................................................. 17
Figure 4: Airport Infrastructure ............................................................................ 18
Figure 5: Turnaround process coordination ........................................................... 19
Figure 6: Retailing on passenger flow process ....................................................... 20
Figure 7: Sofia2 Big Data Platform ........................................................................ 28
Figure 8: Sofia2 Big Data Platform used modules .................................................. 29
Figure 9: Control Panel Sofia2 ............................................................................. 30
Figure 10: Typical Workflow .................................................................................. 31
Figure 11: Sofia2 DataFlow module ....................................................................... 33
Figure 12: Sofia2 DataFlow example diagram .......................................................... 34
List of tables

Table 1: Data Assets ................................................................. 26
Table 2: Supervised Algorithms ................................................. 45
Table 3: Characteristics of Athens’ Server ..................................... 51
Table 4: Roadmap for Objective 1 ............................................... 52
Table 5: Roadmap for Objective 2 ............................................... 53
# Definitions, Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-BPC</td>
<td>Automated Boarding Pass Control</td>
</tr>
<tr>
<td>AIA</td>
<td>Athens International Airport</td>
</tr>
<tr>
<td>AODB</td>
<td>Airport Operational Data Base</td>
</tr>
<tr>
<td>CO</td>
<td>Confidential, only for members of the consortium (including Commission Services)</td>
</tr>
<tr>
<td>CR</td>
<td>Change Request</td>
</tr>
<tr>
<td>D</td>
<td>Demonstrator</td>
</tr>
<tr>
<td>DCS</td>
<td>Departure Control System</td>
</tr>
<tr>
<td>DL</td>
<td>Deliverable Leader</td>
</tr>
<tr>
<td>DM</td>
<td>Dissemination Manager</td>
</tr>
<tr>
<td>DMS</td>
<td>Document Management System</td>
</tr>
<tr>
<td>DoA</td>
<td>Description of Action</td>
</tr>
<tr>
<td>Dx</td>
<td>Deliverable (where x defines the deliverable identification number e.g. D1.1.1)</td>
</tr>
<tr>
<td>EIM</td>
<td>Exploitation Innovation Manager</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>FM</td>
<td>Financial Manager</td>
</tr>
<tr>
<td>LoS</td>
<td>Level of Service</td>
</tr>
<tr>
<td>MIS</td>
<td>Management Information System</td>
</tr>
<tr>
<td>MSx</td>
<td>project Milestone (where x defines a project milestone e.g. MS3)</td>
</tr>
<tr>
<td>MTB</td>
<td>Main Terminal Building</td>
</tr>
<tr>
<td>Mx</td>
<td>Month (where x defines a project month e.g. M10)</td>
</tr>
<tr>
<td>NMS</td>
<td>Network Management System</td>
</tr>
<tr>
<td>O</td>
<td>Other</td>
</tr>
<tr>
<td>P</td>
<td>Prototype</td>
</tr>
<tr>
<td>PC</td>
<td>Project Coordinator</td>
</tr>
<tr>
<td>PM</td>
<td>partner Project Manager</td>
</tr>
<tr>
<td>PO</td>
<td>Project Officer</td>
</tr>
<tr>
<td>PP</td>
<td>Restricted to other programme participants (including the Commission Services)</td>
</tr>
<tr>
<td>PU</td>
<td>Public</td>
</tr>
<tr>
<td>QA</td>
<td>Quality Assurance</td>
</tr>
<tr>
<td>QAP</td>
<td>Quality Assurance Plan</td>
</tr>
<tr>
<td>QFD</td>
<td>Quality Function Deployment</td>
</tr>
<tr>
<td>QM</td>
<td>Quality Manager</td>
</tr>
<tr>
<td>R</td>
<td>Report</td>
</tr>
</tbody>
</table>
Executive Summary

This deliverable reports on the work performed in WP8/T8.1 “Smart passenger flow pilot design” with respect to the design of the Initial Pilot “Smart Passenger Flow” as well as the Replication Pilot “Smart Passenger Flows and Airport Turn Around”. It describes the requirements, objectives, use cases, scenarios, the big data technologies, techniques and algorithms used. It also positions the pilot solutions within the BDVA reference model. The big data infrastructure to be applied in the different pilots is described and a roadmap for the developments and integration is given. The last sections give an overview about commonalities between the initial and the replication pilot.
1 Overall Motivation and Ambitions for Pilot Domain

In the aviation transport domain, significant, double digit savings potentials cannot be found anymore in aerodynamic improvements or more efficient engines. Instead, the use of Integrated Intelligent Information will allow to free up significant savings potentials. This is mainly achieved through two dogmas: 1) pro-active disruption management by analysis of historical data and comparison with actual real time data. This allows the prediction of unfavourable situations and the early deployment of counter-measures to avoid them. 2) Optimizing operations across all involved domains through processing available data according to business goals.

Significant gains in operational efficiency will be achieved by turning available data into Integrated Intelligent Information. This is realized through analysis, integration, networking and implementation of smart applications and services across domains, to optimize operations on a holistic level. Instead of optimizing single stovepiped services, smart algorithms optimize processes according to selected criteria: best economy, best performance, maximum safety/security, maximum availability and more. Industry can define, prioritize and realize performance criteria according to their business goals. By adding new resource allocation rules, this pilot potentially will: (1) decrease the number of passenger losing the connecting flight, (2) facilitate a better scheduling of daily operation and resources required, taking into account parameters not considered nowadays detected through the application of big data exploitation principles, (3) enable a better understanding of the impact of every single process step on the overall airport operation performance.

On the airport side, new business models will include ones, such as (1) oriented retailing, knowing the expected preferences of passengers before their arrival, enabling the implementation of itinerant or mobile retailers, (2) qualified rating of the commercial spaces, with the information of passenger tracking along the terminal, (3) data sell to companies or other interested parties about passenger preferences or any other big data outcome.

Passenger satisfaction will be significantly increased through better use of resources, which will allow to plan for peak hours, assign systems and workforce where needed based on real time situation analysis. Delays will in many times be avoided or at least reduced. The result will be less missed connections and decreased passenger waiting times. Airlines can better connect with their customers and tailor the services offered according to customer segmentation. KPI are: reductions in magnitude of delays, missed connections, lost baggage, passenger complaints, station costs, stable ticket prices. Customer experience and satisfaction of all
stakeholders will include (1) optimization and better monitoring of airport operation providing a better service to all the stakeholders, (2) Reduction of connecting passengers losing their flights, reducing the number of complains and enabling a better scheduling of its internal resource, (3) Losing less connections, and receiving more adequate retailing offers on their trip, optimising their time before the flight, (4) Reception of more qualified passengers with clearer preferences about their expectations, (5) improve passenger experience by offering personalized services depending on passenger roles.
2 Initial Pilot: Smart Passenger Flow

Athens International Airport S.A. (AIA) was established in 1996 as a Public-Private Partnership with a 30-year concession agreement, with a corporate goal to create sustainable value to all stakeholders by offering value-for-money services. Thanks to these services, Athens International Airport has earned numerous international awards and distinctions.

Despite of the global economic crisis, Athens International Airport is one with the highest rates of growing in Europe during the last years. The following graph shows the evolution of annual passengers for this airport:

![Passengers in AIA](image)

**Figure 1: Evolution of Passengers in AIA**

Due to this continue growing rate, one of the main challenges for AIA is to optimize the management of its infrastructure to receive more flights and passengers.

Aegean is the main greek airline and has earned recently the award as Best Regional Airline in Europe in 2016. Having AIA has its hub airport, Aegean manages around the 50% of AIA’s passengers, so the efficiency and quality of AIA services have a strong impact in the passenger experience of AEGEAN’s travellers.

The Athens pilot will use big data technologies to analyze passenger flows, detecting patterns of passenger behaviors to provide a better understanding of it, and helping airport and airline staff to improve the predictability and efficiency of services offered.
2.1 Requirements

2.1.1 Functional requirements:
The smart passenger flow pilot should comply the following functional requirements that have been defined considering AIA and Aegean/Goldair needs regarding the current situation of passenger flows through the airport. The compliance of such requirements may depend on data availability and their alignment to achieve pilot objectives.

Req.1: Predict time of passenger arrival to the terminal
The pilot should be able to predict how long before the time of departure of their flight will the passengers arrive to the airport terminal. This prediction may depend on many factors such as the mean of transport they arrive by, the traffic situation, flight time, passenger preferences, demographics... that have to be considered in the prediction.

Req.2: Predict time of passenger arrival to the processing stations and their demand
As for the previous requirement the pilot should predict the time each passenger is expected to go through the airport processing stations such as security screening, passport control, check-in desks...considering as well every factor (passenger groups, passenger class, etc.) that can affect that time to be changed. Together with the arrival time, the pilot should also predict for each processing station the amount of passengers that will receive in a certain period of time.

Req.3: Elaborate passenger movement heat maps based on demographics
The smart passenger flow pilot should provide the enough information to elaborate passenger movement heat maps mainly based on demographics but considering other factors that can be relevant for shops and restaurants as well.

Req.4: Predict processing times of stations
The pilot should assess and predict how long takes to process a passenger in each station considering passenger profile, current airport status and other relevant information.

Req.5: Assess and predict time to reach the gate
The pilot should assess how long currently take to passengers to reach the gate once they have entered the Main Terminal Building. Every relevant factor, such as checking and security queues, etc... should be considered.

Req.6: Identify transfer passenger late arrivals
The pilot should anticipate the identification of passengers that have connecting flights and are expected to arrive late.
**Req.7: Asses the efficiency of airport signage**

The pilot should assess the efficiency of the airport signage looking for areas where passengers do not behave as expected. In case of detecting areas where passengers get confused, the pilot should identify the reasons for that to happen. For instance, as mentioned by AIA passenger service department, for the situation of arrival passengers waiting for lifts while there are other free lifts behind the busy ones.

**Req.8: Asses the use of MTB entrances**

As for the previous requirement, and facing other need stated by AIA passenger services department, the pilot should assess the use of MTB entrances and identify reasons for why passenger use some entrances or others.

**Req.9: Identify passenger arrival gates (regarding ground handling buses)**

To be able to identify in advance which arrival gate will the buses use for arrival passengers would be an added value for the AIA passenger services department.

### 2.1.2 Non-functional requirements:

**2.1.2.1 Security**

- **NF Req.1:** No personal data should be extracted and stored out of data sources.
- **NF Req.2:** Data should be stored whether in AIA or INDRA facilities but never shared with third parties.

**2.1.2.2 Operation**

- **NF Req.3:** The pilot should be continuously fed during its operation in order to improve the prediction accuracy.

**2.1.2.3 Interfacing**

- **NF Req.4:** The pilot shall not affect any other interfaced system
- **NF Req.5:** The pilot should be interfaced with the following systems
  - MIS
  - AODB/RMS
  - Wifi location system
  - E-parking
  - DCS
  - A-BPC
  - NM
2.2 Objectives

The two main objectives of the Smart Passenger Flow pilot are:

**Obj.1: Operation Management Predictive Optimization Module**

By means of a real time operational module, the pilot will exploit predictive analytics with Passenger Flow data obtained in real time from airport and airline systems. This module will facilitate proactive decision making in real time whenever there is any disruption over the initial plan. The initial targets to be covered with this module are explained in the following sub-objectives:

- **Sub Obj. 1.1:** Reduce delays in departure flights caused by late passengers.

  Along the transit through the airport terminal many facts can affect the passengers causing them to reach the boarding gate later than the scheduled boarding time. Currently, airlines and airports spend many resources trying to locate and warn those passengers trying them to proceed immediately to the boarding gate in order to complete the boarding process without delays.

  Such delays cause not only a reduction of revenue for the stakeholders involved, but also a reduction of the expected level of service as well as a negative perception of the airline and airport.

  Is this why airport and airlines need to anticipate those disruptions and be able to carry out collaborative decision making in order to reduce flight delays caused by late passengers. Late passenger trending features could be identified and preventive actions could be carried out to inform and promote passengers to reach the gate on time. Thereby the passenger flow would be more efficient and it could support the information sharing between the airport and airlines.

  To achieve this objective following requirements should be considered:

  - Req.1: Predict time of arrival to the terminal
  - Req.2: Predict time of arrival to the processing stations and their demand
  - Req.4: Predict processing times of stations
Sub Obj. 1.2: Reduce the number of passenger missing connections and lost baggage

Flights waiting to passengers coming from other transfer flights are another cause of delays that affect airports and airlines operation. Due to the airlines hub and spoke operation, dependency between arriving and departing flights has arisen and therefore any disruption on the incoming flights affect directly to the outgoing ones.

Therefrom emerges the necessity from the airlines to acquire insight on how those late transfers happen and how to solve them.

The automatization on the identification of late transfer passengers will enable an anticipation on the airline process triggering reducing the probability of human error and permitting ground handlers to be aware of the situation once the problem has been identified.

To achieve this objective following requirements should be considered:

- Req.4: Predict processing times of stations
- Req.5: Assess and predict time to reach the gate
- Req.6: Identify transfer passenger late arrivals
• **Sub Obj. 1.3:** Improve the efficiency of passenger processing stations (security screening, passport control, check in desks...)

The airport processing stations can be a limiting factor of the airport capacity. Airports devote great effort to allocate resources depending on the demand and that is one of the main issues they have to deal with. Long waiting queues make the passenger feel uncomfortable and reduces passengers’ travelling experience, but too many agents in processing stations when there is not enough demand, increase airport operation costs without necessity.

In this case, the airport should have a tool that enables an early decision making so that an optimized resources allocation can be performed.

To achieve this objective following requirements should be considered:

• Req.1: Predict time of arrival to the terminal
• Req.2: Predict time of arrival to the processing stations and their demand
• Req.4: Predict processing times of stations
Sub Obj.1.4: Reduce overall turnaround times

During the aircraft turnaround process most of the stakeholders come together while an extremely complex coordination is required. This process is supported by very restrictive Service Level Agreements between the airlines and ground handlers. The breach of those SLAs cause many disruptions within the turnaround process and become in delays, penalties and frustrated passengers.

The end of the turnaround process comes defined by the TOBT (Target Off Block Time). This is the milestone where the aircraft is ready for pushback (in case it is required) or to start taxing to the runway. The uncertainty when estimating the TOBT, in case of any delay during the turnaround process, causes corrective measures to be applied with the consequent loss of efficiency.

The TOBT can be affected by many parameters and passengers are one of those factors that can cause a late departure. Obtaining insight on how passengers move through the airport, regarding checkpoint cross times, and other factors will provide AIA and Aegean useful information to anticipate decision making and increase the efficiency of the turnaround process.

To achieve this objective following requirements should be considered:

- Req.1: Predict time of arrival to the terminal
- Req.2: Predict time of arrival to the processing stations and their demand
- Req.4: Predict processing times of stations
- Req.5: Assess and predict time to reach the gate
- Req.6: Identify transfer passenger late arrivals
**Obj. 2: Descriptive passenger behaviour system**

**Sub Obj.2.1:** Obtain insight on how passenger behave along their journey, especially within the airport terminal to enable customized services/offers, increasing passenger satisfaction and non-aeronautical revenue. As in many other areas, in airport retailing as further you know your customer more accurate will be the offers you serve them. This customization of the offer makes the customer to reduce the wasted time looking for a product. Moreover, if you offer it right when the customer needs it, the probability of making a deal will be almost absolute.

This idea can be extrapolated to the rest of the services that airports offer to the passengers increasing airport non-aeronautical revenue and improving passenger satisfaction.

To achieve this objective following requirements should be considered:

- Req.3: Elaborate passenger movement heat maps based on demographics
- Req.7: Asses the efficiency of airport signage
- Req.8: Asses the use of MTB entrances
- Req.9: Identify passenger arrival gates (regarding ground handling buses)
2.3 Use cases / scenarios

Within the Smart Passenger Flow pilot the following scenarios have been identified:

2.3.1 Scenario 1: Identification of passengers’ arrival time patterns to the airport terminal

Passengers have many different options to get to the airport when they are going to travel. Furthermore, each passenger arrives to the airport earlier or later depending on many circumstances. For airport and airlines, these circumstances that make a passenger arrive earlier or later to the airport are mainly unknown. Having this information before the arrival of passengers happens can make the difference between an efficient flow along the airport, or a completely collapsed process. Acquiring the insight on how and when passengers are going to arrive, enables airport and airlines to anticipate decisions in order to prevent disruptions before they take place.

2.3.1.1 Objectives

This scenario meets with the following objectives:

- Sub Obj.1.1: Reduce delays in departure flights caused by late passengers.
- Sub Obj.1.3: Improve the efficiency of passenger processing stations (security screening, passport control, check in desks...)

Once the airport is aware of passengers arrival patterns the following actions may be carried out:

- Optimize staff allocation at passenger processing stations (screening, check-in...)
- Identify bottlenecks and underused infrastructure
- Early prediction of passengers arrival time to the airport terminal
- Prevent long queueing time and collapses
- Enable early mitigation actions
- Enable the information sharing between stakeholders

2.3.1.2 Contribution/benefits:
- Reduce the loss of revenue caused by late departures
- Improve the airport efficiency and LoS
- Increase the airport capacity
- Increase the passenger travel experience

2.3.1.3 Input data
The following would be the ideal data required to perform Big Data analytics but its availability cannot be guaranteed (data availability assessed in section 2.4):

- Flight plan
- Airport slot
- Passenger
- Baggage
- Boarding Pass Reading
- Mobile Phone Location

2.3.1.4 Example:
Evaluating trends of passenger arrivals, peaks can be identified and airport can increase the security screening resources to enable a smooth passenger processing

For instance, depending on flight and passenger characteristics, could be predicted that they tend to arrive to the terminal 50 minutes prior to departure, therefore specific actions could be carried out to permit on time arrival to the departure gate

2.3.2 Scenario 2: Extraction of passengers’ movement patterns along the airport terminal
Passengers behave in a different way across the airport terminal mainly due to the many things the airport offers and to their own circumstances as well. For example, some passengers arriving early tend to go to a restaurant before crossing the security screening and they cross it close to the flight departure. This situation, in case the security screening becomes suddenly collapsed can make this passenger to lose the flight, or induce the airline to wait for him causing several disruptions and unexpected costs.

Identifying passengers’ patterns on their movements along the terminal can help to prevent this situation keeping the airport aware and enabling an anticipated passenger notification.

2.3.2.1 Objectives
This scenario meets with the following objectives:

- Sub Obj. 1.1: Reduce delays in departure flights caused by late passengers.
• Sub Obj.1.3: Improve the efficiency of passenger processing stations (security screening, passport control, check in desks...)
• Sub Obj.1.4: Reduce overall turnaround times

Once the airport is aware of passengers movement patterns the following actions may be carried out:

• Anticipate passenger last calls
• Assign terminal agents to escort passengers to the gate
• Inform passengers about the situation and promote them according to airport and airline needs
• Enable the information sharing between the airport and airline

2.3.2.2 Contribution/benefits:
• Reduce the loss of revenue caused by late departures
• Improve the airport efficiency and LoS
• Increase the airport capacity
• Increase the passenger travel experience

2.3.2.3 Input data
The following would be the ideal data required to perform Big Data analytics but its availability cannot be guaranteed (data availability assessed in section 2.4):

• Flight plan
• Airport slot
• Passenger
• Baggage
• Boarding Pass Reading
• Mobile Phone Location

2.3.2.4 Example:
Passengers travelling on holidays to European destinations tend to spend 20 minutes buying in the duty free

2.3.3 Scenario 3: Anticipate detection of transfer passengers with short time to take their departure flights.

When an arriving aircraft is delayed and onboard are passengers that have to be transferred to a connecting flight, the airline supervisor has to look for the list of those passengers, analyze their available time for doing the transfer and launch internal procedures to help these passengers to reach their gates on time.
Automatize and anticipate the detection of these problematic flights and the passengers involved on them provides the airlines a highly powerful tool to anticipate the decision making to avoid possible disruptions.

The real time identification of late passengers carrying out an automatization of current methods permits to apply airline procedures efficiently and to reduce the possibility of human error.

This process is usually done by the Airline itself, but sharing this information with Airport Operational staff may allow to find a coordinated solution to simplify the process executed by the airline.

2.3.3.1 Objectives
This scenario meets with the following objective:

- Sub Obj. 1.2: Reduce the number of passenger missing connections and lost baggage

Once the airline is aware of transfer passengers patterns the following actions may be carried out:

- Reallocate aircraft stands to enable on time transfers
- Assign terminal agents to escort passengers to the gate
- Inform passengers about the situation and induce them not to waste time in other activities while in transfer
- Enable the information sharing between the airport and airline

2.3.3.2 Contribution/benefits:
- Reduce the loss of revenue caused by late departures
- Improve the airport efficiency and LoS
- Increase the airport capacity
- Increase the passenger travel experience

2.3.3.3 Input data
The following would be the ideal data required to perform Big Data analytics but its availability cannot be guaranteed (data availability assessed in section 2.4):

- Flight plan
- Airport slot
- Passenger
- Baggage
- Boarding Pass Reading
2.3.3.4 Example:
Anticipating the identification of a late arrival of transfer passengers and considering that have not enough time to reach the departure gate, airline and airport are advised to carry out corrective measures such as changing gate allocation or assigning terminal agents to escort them to the gate.

2.3.4 Scenario 4: Extraction of passengers’ behaviour features and patterns regarding airport services (retailing, food and beverage…)*

(* This scenario depends on the availability of data assets from retail department in Athens International Airport. During the period in which this document has been writing, there is no confirmation we will have that information)

Passengers spend many time and money in airport amenities. Those are considered the main airport non-aeronautical revenue. Passengers depending on certain features and circumstances tend to buy souvenirs or go to have a coffee or many other things. The airport amenities as any other businesses have their own target costumer and try to attract it to the extent. However, considering they are within an airport process, there are several external factor that can make a passenger not to buy a souvenir even if initially was expecting it.

Acquiring insight on why passengers that are supposed to buy something do not do it, can make airport, airline and retailers to adapt their offers to engage that market share that is escaping to their business.

2.3.4.1 Objectives
This scenario meets with the following objectives:

- Obj. 2: Descriptive passenger behaviour system
  - Sub Obj.2.1: Obtain insight on how passenger behave along their journey, especially within the airport terminal to enable customized services/offers, increasing passenger satisfaction and non-aeronautical revenue.

Once the stakeholders are aware of passengers behavior patterns the following actions may be carried out:

- Customize passenger focused services and offers
- Make passengers feel comfortable
- Anticipate passengers notifications
- Enable the information sharing between the airport and airline

2.3.4.2 Contribution/benefits:
- Increase non-aeronautical revenue
- Improve the airport efficiency and the LoS
- Increase the airport capacity
2.3.4.3 Input data

The following would be the ideal data required to perform Big Data analytics but its availability cannot be currently guaranteed (data availability assessed in section 2.4):

- Flight plan
- Airport slot
- Passenger
- Boarding Pass Reading
- Mobile Phone Location
- Passenger’s purchases

2.3.4.4 Example:

Passengers travelling on holidays to London usually enter the terminal from the central entrance and buy chocolate bars when they have more than one hour to take their flights

2.4 Data assets

The following data assets have been identified as sources of information for the Initial Pilot.

Due to the airport is changing its layout and installing new systems, some of which are going to be sources of information for this project, the availability and details of some of these data assets is under analysis at the moment of writing this document, hence the feasibility of use them will be validated in the following stages.

<table>
<thead>
<tr>
<th>Name of Data Asset</th>
<th>Short Description</th>
<th>Initial Availability Date</th>
<th>Data Type</th>
<th>Link to Data ID Card (in basecamp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport Slot</td>
<td>Arrival and departure flights information according to flight schedule at the airport</td>
<td>02/02/2017</td>
<td>XML</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/424792930">https://3.basecamp.com/3320520/buckets/1429164/uploads/424792930</a></td>
</tr>
</tbody>
</table>
### Table 1: Data Assets

<table>
<thead>
<tr>
<th>Name of Data Asset</th>
<th>Short Description</th>
<th>Initial Availability Date</th>
<th>Data Type</th>
<th>Link to Data ID Card (in basecamp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baggage</td>
<td>Baggage checked by passengers</td>
<td>15/03/2017</td>
<td>Plain Text</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/424792937">https://3.basecamp.com/3320520/buckets/1429164/uploads/424792937</a></td>
</tr>
<tr>
<td>Mobile Phone Location</td>
<td>Location at the airport of mobile phones connected to the airport WiFi</td>
<td>TBD</td>
<td>JSON</td>
<td><a href="https://3.basecamp.com/3320520/buckets/1429164/uploads/424792987">https://3.basecamp.com/3320520/buckets/1429164/uploads/424792987</a></td>
</tr>
</tbody>
</table>

2.5 Big data technology, techniques and algorithms

2.5.1 Detailed Explanation of Big Data Technology

Indra will use two different platforms to apply the big data algorithms. One of them will be hosted in Athens and it will be used to validate in real time the big data algorithms. The second one is a cloud platform to analyze and define the big data algorithms. In the following sections we will explain each one of them. With these two platforms, we will design algorithms based in historical data and we will test and validate them in real time.

2.5.1.1 Conceptual model of Athens platform

Indra has a product, called InPlan, which is an implementation of A-CDM (Airport Collaborative Decision Making) concept, designed to collect operational information and predict and evaluate in real time, the impact of some disruptions in turn around process, coming from different airport processes.

InPLAN offers the following capabilities:

2.5.1.1.1 Common Situation Awareness (CSA) of several turnaround processes.

By means of Information Sharing Platform, the integration of information from and to external systems allows system to consolidate, store and share the most updated situation with external partners (hereafter stakeholders).

InPLAN guarantees a common airport operation plan for the local stakeholders by gathering and supplying information about the current situation: Common situation awareness.
Therefore, CSA allows the system to predict some disruptions in the initial plan, to contribute to an early decision making and to build a pre-departure sequence that meets the airport operations plan.

The system allows to show the operational information in different ways, such as tables, dashboards, alerts, etc.

2.5.1.1.2 Monitor Operational Turnaround Processes.
The system collects and shows information of different turnaround processes:

- **Aircraft process.** Thanks to the information received from external sources, InPLAN does tracking of the aircraft process of a flight from the initial planning from origin (Inbound) to the takeoff (Outbound) including the turn-round phase.

- **Passenger process monitoring** so it can be followed how the passengers can affect to the aircraft process. Besides, it supports a procedure to increase the security filter throughput to reduce any possible delay in the departure phase.

- **Ground Handling process,** allowing Ground Handling agents to handle the turn-round activities through direct intervention, using a specific tablet devices application. The turn-round activities monitoring allows to follow the critical path activities and their impact in the aircraft process.

2.5.1.1.3 Performance of Turnaround Processes.
InPLAN contains support functional modules to increase the supervision and the efficiency in the issue resolution and facilitate the decision making. These support tools are:

- **Performance Monitoring Module:** Offering an airport global situation through the dashboards. Also it provides the capability to know the cause of the problem making drill-down of the KPIs predefined.

- **Performance Management Module:** Support the collaborative management process for issues (from KPI alerts or Airport Events) allowing the partners involved in the collaborative management to assess the problem and detect the best solution for the issue that suit with all partners.

As KPI (Key Performance Indicator) system, InPLAN is capable to calculate and visualize in real time the values of the indicators predefined for the Capacity, Efficiency and
Predictability Key Performance Areas based on the specific steering configuration defined by the airport.

2.5.1.1.4 Integration Layer.
InPLAN has an integration module to integrate information from different external systems. This module centralizes the management and transformation of data coming from external systems in a valid information model, which is used in a common way for the rest of functional modules. The information received and calculated by the systems is stored in a centralized and dedicated database.

2.5.1.2 Conceptual model Indra Big Data platform
Indra has a platform of its own of Big Data, designed developed on frameworks Open Source that enables an easy integration of commercial products of third manufacturers. Sofia2 Big Data Platform's main objective is simplifying the use of all its technologies and expediting the use and exploitation of data structured and not structured, even in real time.

Figure 7: Sofia2 Big Data Platform

Sofia2 It is a Modular Platform that allows to deploy its modules independently according to the needs. All the concepts of the platform are managed from a unified web console allows scaling per the needs based on proven technologies.
Built on widely-tested Open Source software: like Hadoop, (default distribution Cloudera CDH), Spark, HIVE, .... It supports real-time scenarios, batch, ML, visualization, ... It is extensible and adaptable, integrates security at the data modeling level, offering validations in the data exploitation and semantics.

The modules of the platform Sofia 2 used in the construction of the pilot are explained in the following subsections.

All interaction with the Sofia2 platform can be done from the Control Panel Sofia2.

- This module is a Web application developed using HTML5 technology that offers a complete web administration/configuration that allows you to manage all the concepts that it handles the platform. All these steps can be done through the API Rest, which allows to exploit all the platform configuration capabilities from other UI or third-party solutions.
- All the features of the platform are operated / configured from this module,
- This console also includes what is known as Base data of configuration or ConfigDB, which is a database that stores all the configuration of the platform.
The control panel is a tool that, through standardized interfaces, allows a representation of structured information and operation design intended to make the most of the users of the system, facilitating learning and reducing response time.

It provides users (depending on their role privileges) features like these:

- **Wizards guided for beginners**
- **Modeling entities so guided**
- **Management of users and roles**: assignment of roles to users, assignment of permissions on information stored on the platform.
- **Visual development of dashboards and synoptic**
- **Visual creation of rules**
- **Visual modeling of workflows ETL**
- **Query about the BDTR and BDH tool**: allows access to the data inserted into the platform via a motor integrated queries
- **Monitoring of processes**: through the console web is possible carry the control on the planning of processes (as by ej. the load of files, the step of data to the BDH, etc.), allowing also its monitoring in time real.
2.5.1.2.1 Sofia2 Typical workflow

1. **Upload information**
   Ingestion of information (this information can be treated in this phase to homogenize it and ensure the quality of the data) through DataFlow Module for the definition of treatment and storage in Stagging Area of Module Storage.

2. **Analysis of the information**
   Processing of information and execution of machine learning algorithms through the Notebook and Machine Learning modules to obtain business value data.

3. **Storage of models and results**
   The values resulting from the executions of the machine learning algorithms are stored on the platform.

4. **Display information**
   Through the Dashboard module, all the information stored in the platform, both inserted in point 1 and inserted in point 3 can be accessed and consulted.

2.5.1.2.2 Sofia2 Storage

The information modeled in the Platform is stored in the Big Data Repository included in the platform.
The reference implementation of this repository that is supported on Hadoop is used. **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:

- **Economic**: runs on low-cost equipment forming clusters.
- **Scalable**: If you need more processing power or single storage capacity there is to add more nodes to the cluster very easily.
- **Efficient**: Hadoop distributed data and processes it in parallel on the nodes
- **Reliable**: Hadoop moves processing (Tasks) to data.

The main parts of Hadoop that uses the solution are:

- **HDFS** is the Hadoop distributed file system:
  - System of files distributed that abstracts of the storage physical and offers a vision only of all the resources of storage from the cluster.
  - To the store a file, it part in blocks and stores each block in node different from the cluster. It also replicates each block in at least three nodes.
  - It is possible to store files larger than the maximum size of any of the machines in the cluster disk.
  - If a node of the cluster is fault, the system continues running while is repaired using the information replicated in other nodes.
- **Hive**: Infrastructure data warehouse on Hadoop, which allows SQL queries data stored in Hadoop.
- **Impala** that allow the access via SQL online to the data stored in HDFS:

2.5.1.2.3 DataFlow (ETL module)

It is one of the points of entry as possible, information on the platform. This module can be used as ETL, either to intake data as for complex transformations within the platform or export of data involving intermediate transformations.
Making focus in them capabilities ETL of the module, we can highlight the following capabilities by each phase of the process:

- **Extraction**: Have 18 the origins of integrated data, among which are as available sources: Excel, AmazonS3, HadoopFS, Sofia2 (which lets you select the ontology, fields, query...), Kafka...

- **Transformation**: Will concatenate successive transformations and actions on the data until the entire process. To do this it has 20 possible tasks:
  
  - Evaluation of expressions: performs checks and calculations that can write fields new or existing.
  - Actions on fields: different actions available on the fields as: Converter, Merger, Masker, Hasher, remove, rename...
  - Parser of JSON, XML and logs: parses information valid per the different types of format of logs, and schema XML and JSON.
  - Flow selector: to select the next activity to execute on the dataset, depending on conditions of execution.
- Evaluators in different languages: different specific actions on the data available for the coding languages (Python, JavaScript, Jython...)
- Other components such as the Replicator registry or the replacement of values

- **Load:** There are more than twenty possible destinations, to incorporate into the process via Drag & drop from the taskbar. Of them we can highlight the Sofia2 component (which lets you select the ontology, fields, and other additional parameters), AmazonS3, Cassandra, Hadoop, Kafka, Flume...

This module includes tools of monitoring both for expedite the development as to make monitoring of the execution of the process a time activated and published.
2.5.1.2.4 Notebooks (Collaborative analytical)

Allows make of way very simple and interactive, analytical on data of sources very varied, including the sources of data of Sofia2.

Of this way is could, for example, make loads of files from HDFS to spark, load of data in tables Hive, launch consultations or perform a process complex of machine learning through the libraries of MLlib of Spark.

Also the possible the use of code R as well as the numerous libraries of the language, allowing by examples display maps of leaflet.

Sofia2 Notebooks can combine Scala code, Spark, SparkSQL, Hive, R, Shell, or many others with html content or reactive policy angle, allowing interactions in real time with a powerful interface, and all in a shared environment, multi-user.

Each supported language is managed by an interpreter, so it always that you want to write code for a certain language should be write an own marker in the paragraph.

In addition it allows instant visualization of data, being able to easily configure graphics and quickly change the display of the same type. Also is possible the creation of graphics advanced thanks to libraries own of each language.
SparkSQL:

Figure 15: Sofia2 Notebook Spark graphics

HIVE:

Figure 16: Sofia2 Notebook HIVE graphics
Python:

```python
N = 200
Ndata = [Ellipse(x=rand(2), width=rand(), height=rand(), angle=rand()*360)
            for i in range(N)]

fig = figure()
sr = fig.set_subplots(331, aspect='equal')
for a in sr.items():
    a.set_artist(e)
    a.set_clip_box(ax.bbox)
    a.set_visible(True)
    a.set_facecolor(rand(33))

ax.set_xlim(0, 10)
ax.set_ylim(0, 10)
show()
```

Figure 17: Sofia2 Notebook Python graphics
Each Notebook consists of paragraphs, which may have different languages, and can run individually the paragraphs and viewing the output of the same, as well as the State of execution.

Both paragraphs, and the full notebook can outsource via url, seeing in real time in all cases, the executions of notebooks or paragraph.
Another feature important is the possibility of plan the execution of them notebooks through an expression CRON, and can run notebook repeatedly and without loss of context, and can select an interval of execution of them predesigned or write one custom.

With all these features have a tool web collaborative, that is capable of perform analysis complex of the information managed by the platform (both in time real as historical), combining different languages and generating views graphic (u others actions), that is can plan for their execution periodic, cooling automatically the result of the analytical that is exposed in a URL.

2.5.1.2.5 Machine Learning
Module Machine Learning platform-allows you to apply and shape easily for different learning techniques, among which we can highlight the following:

- **Regression**: Techniques to estimate relationships between variables and determine the relative importance of these in the prediction of new values.
- **Clustering**: Techniques for segmenting the data in similar groups.
- **Classification**: Techniques to identify the membership of an element to a specific group.
- **Recommendation / Prediction**: Techniques for predicting the value or preference of a new entity based on historical preferences or behaviour.

Through the interpreter Sofia2 allows:

- Store models created on the platform. From this it will be possible to manage them from the web console, from which we can also invoke them based on parameters and give them permissions.
• Publish Scripts Sofia2Models that provides methods to retrieve the model, save it, invoke it, assess its quality.
• Generate REST APIs allowing to evaluate input data sets through the generated models. This facilitates its invocation through standard mechanisms that also have integrated security platform.

This module allows you to define workflows visually, so that it is only necessary to introduce the configuration parameters and input data to define analytic processes.

2.5.1.2.6 Dashboards
This module allows you to create a simple and visual dashboard on the information managed by the platform.

This module allows to create simple and visual dashboards from platform control Panel.

For that it offers various types of gadgets

![Figure 20: Sofia2 Dashboard types](image)

who can join to generate a full Dashboard, on the information that is added to the platform, or on historical information.
2.5.2 Detailed Explanation of Technology, 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 being its main goal to find hidden or implicit information, which cannot be obtained through conventional statistics 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 with some shortcuts. The major steps are represented in the next diagram. 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.
Methodology of Advanced Data Mining

1. Goal Definition Based On Business Knowledge
   - Start gathering background information about the current business situation.
   - Describe Problem Area.
   - Describe Current Solution

2. Data preparation & Management
   - Data Understanding
   - Data Preparation (merging data sets and/or records, cleaning data, constructing New Data, integrating data and formatting Data)
   - Sample
   - Results: Initial data table

3. Creating analytic models
   - Segmentation
   - Predictive models
   - Prescriptive models

4. Validation and conclusions
   - Assess: Consistency study, robustness, reliability, variability, etc.
   - To obtain conclusions

5. Deployment
   - Solution integration with the pilot

Figure 21: Methodology of 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 goal for a data analyst is to understand what the customer really needs to achieve. Very often, the customer has many goals which compete among them and requirements which should be balanced. It is important to discover which is the primary objective, and the relations with the rest of objectives.

The analyst should describe the criteria which are useful from the business perspective so they can easily understand the situation. Afterwards, it is necessary a more detailed research about all the resources, restrictions, presumptions and other factors which should be considered to determine the objective of data analysis and project plan.
Afterwards, the business goals are converted into data mining goals, so the goals are translated into technical issues. However, it is important to determine criteria for business success. The tool to be used is also selected in this stage.

2.5.2.2 Data preparation and management

First of all, from an initial data collection, it is possible to identify the data quality, discover the first knowledge and identify interesting data subsets to make hypothesis regarding to hidden information.

Secondly the final data set to be used in the analysis is built and it includes tasks such as table selection, records and attributes, as well as transformation, new specific variables and data cleaning.

The data cleaning can include the substitution of data with defects to the data estimation through modelling. Other operations include production of derived variables or creation of new variables.

Other common operation consists of combining data with open sources, especially when there are relationships between the initial data and the Open Data, for instance, combining data with socio-economic variables in EUROSTAT.

The combined data also cover aggregations, as new values calculated as summary information from multiple records. For instance, a table with customer shopping new fields could be number of shopping, average in the shopping quantity, percentage of articles in promotion, etc..

2.5.2.3 Creating analytic models

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:
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 predictive level.

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...

The next table summarizes the most common algorithms in supervised learning in both categories:
### Table 2: Supervised Algorithms

<table>
<thead>
<tr>
<th>Classification problem</th>
<th>Regression problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>Simple regression</td>
</tr>
<tr>
<td>Decision tree</td>
<td>Ridge regression</td>
</tr>
<tr>
<td>Random forest</td>
<td>Lasso regression</td>
</tr>
<tr>
<td>Gradient boosted trees</td>
<td>Elastic net (Ridge+Lasso)</td>
</tr>
<tr>
<td>Neural network, Deep learning</td>
<td>Regression tree</td>
</tr>
<tr>
<td>Adaboost</td>
<td>K-neighbors regression</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>SVR</td>
</tr>
<tr>
<td>K-neighbors</td>
<td>Random forest regression</td>
</tr>
<tr>
<td>SVM</td>
<td>Gradient boosted tree regression</td>
</tr>
<tr>
<td></td>
<td>AFT</td>
</tr>
</tbody>
</table>

In both types of problems, many different algorithms from the listed above are tested and the most accurate is chosen.

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:

![Figure 23: Relationship between scenarios and algorithms](image)
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.

![Model creation Process](image)

**Figure 24: Model creation Process**

### 2.5.2.4 Model Evaluation

The data scientist is able to interpret the models according to his domain knowledge, the success criteria in data mining and the desired test design. Later, he discusses with the business analysts the results in the business context. Depending on the model evaluation, the adjust parameters are reviewed and adjusted for a new model evaluation until the best model has been achieved until the model can answer the business goals in a better way. It is even possible to encounter business decisions which make the model deficient. So according to the evaluation results and the process review, the project team decides how to proceed. The equipment decides if the project has to end, if it should continue by modifying the development so more iterations are necessary either a new data mining process should start.

A good way to define the total outputs of data mining is OUTPUTS=MODELS+CONCLUSIONS

### 2.5.2.5 Deployment: Solution integration with the pilot

Supervision and maintenance are important issues if data mining results are part of the daily business.

Generally, data mining processes are not running independently in an IT environment but they have to interrelate with other applications or be incorporated into the business processes. So, we think this stage is crucial to assure the success of the data mining algorithms.
2.6 Positioning of Pilot Solutions in BDVA Reference Model

- **Data Visualization and User Interaction:** The pilot will provide a set of specific reports/dashboards 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 in the optimization of the maintenance works. (See section 2.5.1.1.1 and 2.5.1.2.6 for more details)

- **Data analytics - Algorithms developed for the pilot:** A descriptive analysis will be carried out in the first place to get a full understanding of the data and unsupervised algorithms will be used to discover the hidden patterns already depicted. Finally, a predictive algorithm will anticipate the number of passengers that will arrive to the airport on every time interval. (See section 2.5.1.2.4 and 2.5.1.2.5 for more details).

- **Data Processing Architectures:**
Batch process: Processes that allow the feeding of the algorithms with new data collected during the execution of the pilot will be specified for each data source. The inclusion of new data will be a periodic task given the nature of the data sources. (See section 2.5.1.2.4 and 2.5.1.2.5 for more details).

Interactive: There are identified some data sources that has an unstructured or semi-structured format, for this reason will be necessary to process this information by interactive methods in order to be able to use and include them at the pilot. (See section 2.5.1.1.4, 2.5.1.2.4 and 2.5.1.2.5 for more details).

Streaming Real-Time: Some systems will provide information in real time to validate how the predictive algorithms fit with the real-time information. (See section 2.5.1.1.4)

• 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 chapter 2.4. The techniques used to manage the data will be collection, preparation, and linking:

  o Collection: techniques and tools for gathering and storing data in its original form (i.e., raw data.).
  o Preparation/Curation: techniques and tools for converting raw data into cleansed, organized information.
  o Linking/Integration: techniques and tools for matching, aligning, and integrating information.
  o Access: techniques, tools, and interfaces for accessing information (incl. access rights management)

  For more detailed description see chapter 2.5.1.1.4, 2.5.1.2.3, 2.5.1.2.4 and 2.5.1.2.5.

2.7 Big data infrastructure

2.7.1 Platform’s Architecture:
The project will use two different platforms, one hosted in Athens for in-situ trials validations and another hosted in cloud.

The following figure shows an overview of platforms’ architecture.
At pilot stage 1 and 2 we will use a Cloud infrastructure in order to execute the first steps of our methodology (data preparation and management, creating analytic models and model evaluation). This first steps and the first iterations of our methodology cycle will be executed at the Cloud infrastructure but at the end of stage 2 and the stage 3, the pilot will be deployed at Athens, in Athens International Airport’s data management center.

Big Data Algorithms will be focus on managing big set of data and designing descriptive and predictive algorithms.

Athens’ Platform will focus on real time integrations and presentation of real time alerts based on Big Data Algorithms.

2.7.2 Indra’s Big Data Infrastructure:

The following figure shows an overview of the Big Data platform. This platform is common to three Indra’s pilots in three different domains, Aviation, Railway and Highway, so it is dimensioned for the estimated volumetry in the project. Each domain will use this common platform including the specific developments needed to validate each pilot objectives.
Figure 27: Sofia2 Big Data Platform Infrastructure

**SOFIA2 CORE NODE**
- **1 x Node**
  - 1 CPU Intel Xeon 4 cores
  - 28 GB RAM
  - 200 Gb
  - Linux OS 64 bits Ubuntu 14.04

**SOFIA2 CORE NODE**
- **1 x Node**
  - 1 CPU Intel Xeon 16 cores
  - 112 GB RAM
  - 800 Gb
  - 1 TB HADOOP
  - Linux OS 64 bits Ubuntu 14.04

**WORKER BIG DATA**
- **3 x Nodes**
  - 1 CPU Intel Xeon 8 cores
  - 56 GB RAM
  - 400 Gb
  - 1 TB HADOOP
  - Linux OS 64 bits Ubuntu 14.04

**Figure 27: Sofia2 Big Data Platform Infrastructure**
2.7.3 Athens’s infrastructure

The hardware available for in-situ trials is:

<table>
<thead>
<tr>
<th>Oracle Server X5-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>RAM</td>
</tr>
<tr>
<td>Memory</td>
</tr>
</tbody>
</table>

Table 3: Characteristics of Athens’ Server

Figure 28: Oracle Server X5-2
2.8 Roadmap

The two main objectives for this pilot will be implemented and validated in different stages, as it is explained in the table below:

**Objective 1: Optimize operation management regarding passenger flow**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Delivery Date (Project Month)</th>
<th>Features / Objectives Addressed</th>
<th>Embedding in Productive Environment</th>
<th>Big Data Infrastructure Used</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Technology Validation</td>
<td>M6</td>
<td>Gathering Initial insight about <strong>Objective 1</strong></td>
<td>Data collection and initial data analytics algorithm validation</td>
<td>Indra’s Big Data Platform</td>
<td>Small scale (sufficient size for training machine learning models and their cross-validation)</td>
</tr>
<tr>
<td>S2: Large-scale experimentation and demonstration</td>
<td>M14-15</td>
<td>Solution validation of <strong>Objective 1</strong></td>
<td>On site trials in Athens using real-data</td>
<td>Indra’s Big Data Platform + Athens platform</td>
<td>Large scale (actual operations)</td>
</tr>
<tr>
<td>S3: In-situ trials</td>
<td>M26</td>
<td>Refinement and Solution validation of <strong>Objective 1</strong></td>
<td>On site trials</td>
<td>Indra’s Big Data Platform + Athens platform</td>
<td>Large scale (actual operations)</td>
</tr>
</tbody>
</table>

Table 4: Roadmap for Objective 1
# Objective 2: Improve the understanding of passenger behaviour

<table>
<thead>
<tr>
<th>Stage</th>
<th>Delivery Date (Project Month)</th>
<th>Features / Objectives Addressed</th>
<th>Embedding in Productive Environment</th>
<th>Big Data Infrastructure Used</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Technology Validation</td>
<td>M6</td>
<td>Gathering Initial data sets for <strong>Objective 2</strong></td>
<td>Data collection and initial data analytics</td>
<td>Indra’s Big Data Platform</td>
<td>Small scale (initial collection of data)</td>
</tr>
<tr>
<td>S2: Large-scale experimentation and demonstration</td>
<td>M14-15</td>
<td>Analysis and Design of models to face the <strong>Objective 2</strong></td>
<td>Data collection and data analytics algorithm validation</td>
<td>Indra’s Big Data Platform</td>
<td>Large scale (sufficient size for training machine learning models and their cross-validation));</td>
</tr>
<tr>
<td>S3: In-situ trials</td>
<td>M26</td>
<td>Solution validation of <strong>Objective 2</strong></td>
<td>No-in-situ trials planned for this objective.</td>
<td>Indra’s Big Data Platform</td>
<td>Large scale (more diverse, larger data sets to evaluate scalability);</td>
</tr>
</tbody>
</table>

**Table 5: Roadmap for Objective 2**
3 Replication pilot

The replication pilot will be installed in Malpensa Airport managed by SEA. SEA joined Transforming Transport project at the beginning of March 2017 (gracefully covering the non-show of previously committed partners), so this pilot maturity is not expected to be the same as other pilots started by January 2017. The plan is to catch-up along 2017.

SEA Società per Azioni Esercizi Aeroportuali was born in 1948 as “Aeroporto di Busto” joint-stock company with the aim of including Milan and the Lombardia region in the first international air routes. The first airport dates back to 1952 (Malpensa), the second (Linate) opened in 1960. Only in 1955 the Company Shareholders’ Meeting resolved to change the company name to SEA (Società per Azioni Esercizi Aeroportuali). This new company took over direct responsibility for providing assistance services to Airline Companies and Aircrafts.

The first convention between the Government and SEA was signed in 1960, recognizing the private status of the Milan airports. Another 40 years concession agreement was signed in 2001, prolonging the private status of the Milan airports until 2041.

The SEA Group mission is to create value for all parties directly involved in the Group business. This objective is pursued through the offer of services and solutions to meet the growing demands of a market comprising multiple entities including passengers, airlines, airport operators and commercial partners, operating on Malpensa and Linate airports.

The airport infrastructures managed by SEA guarantee aircraft access towards major international destinations to a multitude of users operating in a catchment area that is amongst the most developed in Europe. To do so it is essential to create an effective synergy between the Air Side Operations and the Land Side Operations, in order to be able to predict any possible delay, avoid congestions, and minimize any negative effect for passengers, airlines and the Airport itself.

Malpensa Airport operates as a Airport Collaborative Decision Making (A-CDM) enabled airport. Airport Collaborative Decision Making (A-CDM) is an operational procedure to improve air traffic management through a better information sharing among all the stakeholders.

This new procedure increases efficiency and punctuality by improving air traffic flow and airport capacity management, reducing delays by improving events predictability and optimizing resources utilization. In other words, it represents a change in the operational methodology in
the turnaround management, from “first come – first served” to “first ready – first served”. One of the main objectives of A-CDM is to estimate the “Target Take Off Time” (TTOT) as thoroughly as possible in order to improve the “en route” and “sector” planning by the European ATM; this can be reached by implementing a series of “DPI” (Departure Planning Information) and “FUM” (Flight Update Messages) sent to Network Management Operations Centre (NMOC/CFMU). Therefore Airport CDM can be considered as a basis for connecting the airport to ATM system.

A-CDM is therefore a procedure that foresee a better collaboration between all the stakeholders using more updated, better quality and unique meaning data; all the procedures and communication policies will be standardized in order to reduce any possible error source.

Transforming Trasport project will bring new techniques and technologies originated in the Big Data ecosystem that will allow the use of big historical datasets not well exploited to beat current predictions capabilities. This better predictions will play a key role in enhancing the turnaround process.

### 3.1 Requirements

During the phase of the replication pilot following requirements should be fulfilled:

**Requirement 1: Enable fleetwide turnaround optimization**

A holistic turnaround optimization enables the synchronization of the turnaround processes and flight scheduling for all aircraft of an airline at one of their hubs. Furthermore, based on the holistic view of the overall fleet, priorities can be defined for incoming and outgoing flights to better manage disruptions and irregularities.

**Requirement 2: Enable extension of A-CDM data**

The integration of ground operations into the A-CDM is one step in the direction of an holistically connected A-CDM. Connecting all actors and sharing all information commonly increases situational awareness and enables the improvement of managing disruption.

**Requirement 3: Integrate feedback of initial pilot scope passenger flow**

One scope of the replication pilot is to integrate the outcomes and feedback from the initial pilot which goal is to analyse the passenger flow and passenger behaviour at the airport. The replication pilot will use the knowledge about the impact of passenger flow information in the turnaround process in the new enhanced turnaround optimization. Current predictions do not take into account passenger flows related data, since this may be one of the sources of inaccuracy the holistic view of the pilot will benefit from this additional information.
Functional requirements:

The smart aircraft and passengers turn-around pilot should comply the following functional requirements that have been defined considering SEA’s clients’ needs regarding the current situation of passengers and flights flows through our airports. The compliance of such requirements may depend on many daily variables: the data availability and their alignment will be essential to achieve pilot objectives.

Two Airlines and their clients will be taken into account for the pilot.

3.2 Objectives

To fulfill the above mentioned requirements a set of objectives are defined to be achieved during the replication pilot phase.

Objective 1: Improve prediction of ETA based on machine learning

To improve the prediction of ETA historical airport and airline data like ETA, ATA, date, weather, traffic... will be analysed in a first step. Learning algorithms will be used to develop a more accurate prediction of ETA. The objective of improving the prediction of ETA supports the fulfilment of requirement 1, as a more accurate ETA will improve the turnaround process and leads to a more robust turnaround plan.

Objective 2: Upscale turnaround optimization tool and develop prioritization for turnaround processes

To fulfill requirement 1 it is necessary to develop a tool, which enables a holistic view on the airline fleet. To achieve Objective 2 an existing turnaround optimization tool will be analysed and expanded. Based on the upscaling effort, algorithms are developed to prioritize flights for turnaround taking into account data like current delay, ferry/feeding flight, passenger connections, passenger status, turnaround time, gate assignment...

SubObj 2.1: Track and predict Aircraft sequence once landed

The pilot should track aircrafts movements after landing concerning the stand/gate assignment.

SubObj 2.2: Track and predict Aircraft turn-around timing process
The pilot should track all the aircrafts’ turn-around operations starting from the block-on in the assigned stand and ending with the block-off.

SubObj 2.3: Track and predict Aircraft sequence before departure

The pilot should track ETD delays in accordance to the eventual assigned slot, with particular attention to any new slot request.

SubObj 2.4: Track and predict Aircraft de-icing sequence before departure

The pilot should track, and extract a prediction in accordance to the results, aircrafts sequences and consequent delays in accordance to the de-icing operations.

Objective 3: Feed optimized turnaround data into A-CDM platform

In order to fulfill requirement 2 it is necessary to feed at least the improved TOBT into the A-CDM platform. Depending on the required amount of information, additional data can be fed into the A-CDM. The improved TOBT increases situational awareness of all relevant actors and enables them to quickly react on disruptions and irregularities.

Objective 4: Feed new ETA into Airport A-CDM platform based on machine learning

In order to allow better optimization of the turnaround process is key to have more accurate estimations on the arrival time of the incoming flights. Currently the A-CDM platform publishes an Estimated Time of arrival which is updated each time a new FUM is generated by ECTL. A-CDM focuses on the principle that a departing flight is fundamentally a continuation and reidentification of an arrival flight that transitions through a ‘ground trajectory’ phase. The receipt of Flight Update Messages (FUM) provides a more accurate estimated landing time (ELDT) as early as 3 hours from touchdown. This information provides airport stakeholders with the information they need to best allocate resources should the ELDT of an arrival flight shift significantly.

FUM are distributed via the EUROCONTROL B2B web service or ATFM Fixed Telecommunications Network (AFTN). The distribution of reliable arrival updates between airport partners has demonstrated improvements in the following areas:

- Stand planning
- Ground handling resource allocation
- Fleet planning
departure punctuality

However all estimations are triggered by messages which are not predicted, so, whenever new information is available, new calculations are triggered based on flight plans, plus standard taxi times and minimum turnaround times. No predictability based on previous operations is applied.

This method has brought many advantages to the airports using A-CDM (see “A-CDM Impact Assessment, Final Report, Eurcontrol, March 2016”), and the expectation is that enhanced estimations based on leveraging historical data can bring even more.

Objective 5: Monitor performance of the Airport A-CDM platform estimations

In order to achieve objective 5, it is needed to compile and analyze historical data referent to Actual Arrival times; in parallel it is possible to monitor the Estimated Times to understand better how the precession evolves with time and detect the “blind” points in the process, the ones generating more unpredictability. Just this information will be very valuable for designing better procedures or detecting the most impacting incompliances.

Objective 6: Coordinate prioritization with ATC

In order to achieve an improvement of turnaround processes a connection to ATC will be established to communicate the preferences of prioritization of aircraft calculated by the turnaround optimization to ATC.

Objective 7: Track and predict time of arrival to the terminal and to the processing stations

To achieve Requirement 3, the pilot should track passengers footprints, i.e. each passenger arrival to the terminal and permanence in all the airport processing stations such as check-in desks, security screening, passport control, shops, etc. Footprints will be used to predict, for each of the processing station, the permanence and the amount of passengers in every hour of the day.

Objective 8: Assess and predict time to reach the gate

To achieve Requirement 3, the pilot should track how long currently takes to passengers to reach the gate once they have entered the Main Terminal Building. Every relevant factor, such as check-in and security queues, etc... should be considered to predict any possible delay.
Following Table maps Requirements to Objectives:

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement 1: Enable fleetwide turnaround optimization</td>
<td>Objective 1: Improve prediction of ETA based on machine learning</td>
</tr>
<tr>
<td></td>
<td>Objective 2: Upscale turnaround optimization tool and develop prioritization for turnaround processes</td>
</tr>
<tr>
<td>Requirement 2: Enable extension of A-CDM data</td>
<td>Objective 3: Feed optimized turnaround data into A-CDM platform</td>
</tr>
<tr>
<td></td>
<td>Objective 4: Feed new ETA into Airport A-CDM platform based on machine learning</td>
</tr>
<tr>
<td></td>
<td>Objective 5: Monitor performance of the Airport A-CDM platform estimations</td>
</tr>
<tr>
<td></td>
<td>Objective 6: Coordinate prioritization with ATC</td>
</tr>
<tr>
<td>Requirement 3: Integrate feedback of initial pilot scope passenger flow</td>
<td>Objective 7: Track and predict time of arrival to the terminal and to the processing stations</td>
</tr>
<tr>
<td></td>
<td>Objective 8: Assess and predict time to reach the gate</td>
</tr>
</tbody>
</table>

3.3 Use Cases / Scenarios

Use Case 1: Detection of patterns in Arrival Time biggest deviations

Since accuracy of the arrival time is critical for optimizing the turnaround process, identifying the cases for which the deviation of the ETA from the final TA is too big and finding possible patterns will help to find enhancements in the process.

The application of algorithms based on the data can help to identify patterns non obvious to the naked eye and identify thus the root causes, i.e. some periodic events like weather extreme conditions or systemic sectors regulations can lead to repetitive wrong ETA, but this is something difficult to detect just looking at the day of operations real time vision offered by the A-CDM platform.

The collection of historical data, with the addition of extra data linked will be used in this use case to discover this hidden patterns.

The advantage of using the A-CDM data is that there will be fine granularity on the different slices of the total time period estimated to finally find the ETA. This will allow to detect the patterns for specific intervals (Landing – In Block Time – Off Block Time – Take Off)
Use Case 2: Winter Season Operations

Aircraft flow in Malpensa and Linate is strongly influenced by winter weather. The climate in Milano and, in general, in Lombardy, strongly affects every day airports’ life due to some extreme, frequent, highly variable, conditions such as: fog, freezing fog, freezing rain, snow, and ice.

To anticipate and, eventually, face winter adversities SEA has structured the Winter Season Operations. The “Emergenza Neve” season, begins early in November and finishes by the end of March, thus occupying almost ½ year. To face “Emergenza Neve”, SEA has trained a huge number of people, the so called Winter Team, both in Air Side and Land Side. Besides the de-icing operations in Air Side, in fact, SEA has trained a Contingency Team that have to manage passengers in case of airport closure.

In this usecase special focus will be placed to the turnaround in the winter season, looking for specific seasonal patterns.

3.4 Data Assets
(detailed description of the assets, their characteristics and availability)

<table>
<thead>
<tr>
<th>Name of Data Asset</th>
<th>Short Description</th>
<th>Initial Availability Date</th>
<th>Data Type</th>
<th>Link to Data ID Card (in basecamp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flight Plan</td>
<td>Flight Plans created for 59+1 airline simulation</td>
<td>01/03/2017</td>
<td>txt</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Schedule</td>
<td>Airline schedule for 59+1 airline simulation</td>
<td>01/03/2017</td>
<td>txt</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Fuel Consumption</td>
<td>Fuel results for each flight in 59+1 airline simulation</td>
<td>01/03/2017</td>
<td>csv</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Ground tasks</td>
<td>Names and duration of ground tasks for 59+1 airline simulation</td>
<td>01/03/2017</td>
<td>csv</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Flight results</td>
<td>Flight time results for 59+1 airline simulation</td>
<td>01/03/2017</td>
<td>csv</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Weather forecast</td>
<td>The Global Forecast System (GFS) is a weather forecast model produced by the National Centers for Environmental Prediction (NCEP).</td>
<td>10Oct2006–Present (approx. two years to present online)</td>
<td>weather variables at several pressure levels</td>
<td>To be added 2nd half 2017</td>
</tr>
<tr>
<td>Real time flight</td>
<td>The NM B2B Web Services is an</td>
<td>2017</td>
<td>FlightPlan</td>
<td>To be added</td>
</tr>
<tr>
<td>plans</td>
<td>interface provided by the EUROCONTROL Network Manager (NM) for system-to-system access to its services and data, allowing NM customers to retrieve and use the NM information in their own systems, according to their business needs.</td>
<td>Message</td>
<td>2nd half 2017</td>
<td></td>
</tr>
<tr>
<td>Real time surveillance ADSB</td>
<td>Flightradar24 offer a live feed service that provides partners with near realtime updates of flights positions.</td>
<td>2017</td>
<td>Aircraft Positions</td>
<td>To be added 2nd half 2017</td>
</tr>
</tbody>
</table>

Real time surveillance: Commercial available ADSB Based surveillance will be used. The intention is to favor EU based provides (i.e. FR24) vs USA based ones (i.e. Flightaware), but technical feasibility must be guaranteed. This commercial providers usually offers an ETA which will be used as a baseline for benchmarking the predictive capabilities developed in TT.

Weather forecasts: Currently the most accessible source of weather data openly available is the NOAA “gfs”. This source offers several daily updates of global coverage of different weather related metrics and variables.

Real time flight plans: Eurcontrol B2B service has recently implemented the publish/subscribe offer for flight plans which will allow TT to know the real time flight plans of arriving flights to a given airports. Access to this source may require extend license agreement current in place for BR&TE.

3.5 Big Data Technology, Techniques and Algorithms

3.5.1 Data Lake
The pilot will need the creation of a Data Lake to store and link all the dataset available for the project. The Data Lake will offer enough space and scalability to cope with the incoming data during pilot operation. Data Lineage and/or Governance will be desirable features of the Data Lake. As a minimum an HDFS store will be needed for the raw data. Non-SQL databases may be needed for specific datasets. The Data Lake may be supplemented by a traditional relational database for specific ad-hoc query purposes.
3.5.2 Data processing:
For the data preparation needed for the learning algorithms, batch technology may suffice and Map-Reduce or Spark over YARN will be enough. The processing pipeline could benefit from HIVE for providing SQL support. For the live processing of the incoming data some streaming near-real time capability will be needed, here Spark Streaming is a good candidate (Apache Hadoop may be considered if needed).

3.5.3 Learning Algorithms:
Standard well known algorithms will be evaluated for computing Time of Arrival based on known data in real/time. The historical data loaded into the Data Lake will be used for the training and the result applied to the new data received in streaming to predict a Time of Arrival (enhanced ETA). This prediction will be used by the optimizer to prescribe specific action to the parties involved in the turnaround process.

3.5.4 Optimization Algorithms

3.5.5 Visualizations
Current visualizer of Arrivals can be used to show this new enhanced ETA, so no new visualizations need to be developed. For benchmarking purposes a simple visualization of the comparison of current market available ETA and the new one developed could be implemented, other option will be just to compute the differences to obtain a metric.
3.6 Positioning of Pilot Solutions in BDV Reference Model

3.7 Big Data Infrastructure

Section needs to be added at a later point in time. A final decision could not be made due to the late completeness of all partners in WP8.
# 3.8 Roadmap

Definition of what will be available as results of each Stage (1-3)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Delivery Date (Project Month)</th>
<th>Features / Objectives Addressed</th>
<th>Embedding in Productive Environment</th>
<th>Big Data Infrastructure Used</th>
<th>Scale of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Technology Validation</td>
<td>M9</td>
<td>• collection &amp; processing of data around the Turn Around Process</td>
<td>• Virtual Airline, connected to operational AOC systems, running in shadow mode</td>
<td>• Scalable, cloud based due to high variety of data sources.</td>
<td>• 3-4 Tb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Validate flow models from initial pilot</td>
<td>• On site installation at MXP, connection to AODB</td>
<td>• Services provided by platforms with multitude of different data sets streaming</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• integration within the operational environment of an airline</td>
<td>• Scalable, cloud based due to high variety of data sources.</td>
<td>• analytics for real-time data or batch analytics for historical data</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Improve ETA predictions</td>
<td>• 3-4 Tb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2: Large-scale experimentation and demonstration</td>
<td>M15</td>
<td>• use the developed optimization algorithms to determine Airline preferences for all Turn Around events</td>
<td>• Two environments</td>
<td>• SWIM</td>
<td>*see above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• link the fleet wide planning to the local situation at FAR and MXP</td>
<td>• AOC@FRA and MXP</td>
<td>• AODB</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Calculate ETA with probability</td>
<td>• physically decoupled from any production systems, but will operate in a quasi-real environment</td>
<td>• FPL, NOTAM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Connection to MXP AODB, ATC, Ground Handling</td>
<td>• ATC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Traffic &amp; Radar simulation</td>
<td>• Wx</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Ground Handling</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Trials on selected scenarios (e.g. Deicing) will examine the dependencies of the Hub and Spoke operations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>between Frankfurt and Milano</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3: In-situ trials</td>
<td>M27</td>
<td>• add optimization functions to ground ops tools (value add)</td>
<td>• Trials on selected scenarios (e.g. Deicing) will examine the dependencies of the Hub and Spoke operations</td>
<td>•</td>
<td>*see above</td>
</tr>
</tbody>
</table>
4 Commonalities and Replication

4.1 Common requirements and aspects
As explained in the previous sections, both pilots will be focus on optimizing the turnaround. The initial pilot will do it, analysing the passenger flows and its impact on the airport services and aircraft departures. While the initial pilot is putting the focus on the so called “landside” of the airport, the replication pilot will take the airside as a central element of the investigations. It will analyse to what extent it is possible to coordinate the turn around process and its resources on the apron to dynamic preferences of the flight schedule of the airline. It will be a coordination between airport operational control and airline operations center.

From the Big Data techniques point of view, both pilots will exploit predictive models, learning from historical data and using the pattern in the data to do better estimations, used later for the optimization of the turnaround process. In particular, regression models are shared by both prototypes.

At technology level, both pilot will use common products, in particular HDFS, Hive, Spark, SparkSQL.

Some datasets are common, the more relevant being “FlightPlans”.

4.2 Aspects of Replication
One of the expected outputs from the initial pilot will be the measurement (KPI) of how the passenger flows impact on the delay of aircraft within the airport. This KPI will be taking into account in the replication pilot in order to asses in how far this KPI developed in the Athens pilot may be replicated and reused for the more challenging setting of the Milano airport.