A Software Architecture for Extreme-ScaLe
Big-Data AnalyticS in Fog CompuTIng Ecosystems

# D2.5 Distributed Data Analytic Platform Requirements

## Version 1.0

### Document Information

| | |
|---|---|
| **Contract Number** | 825473 |
| **Project Website** | https://elastic-project.bsc.es/ |
| **Contractual Deadline** | M6, May 2019 |
| **Dissemination Level** | PU |
| **Nature** | R |
| **Author(s)** | Usman Wajid, John Beattie (ICE) |
| **Contributor(s)** | Jurgen Assfalg (FLO), Giacomo Codecasa (FLO), Marco Merlini (THALIT) |
| **Reviewer(s)** | Marco Merlini (Thalit) |
| **Keywords** | ELASTIC, Data Analytics, Platform |

***Notices:*** *The ELASTIC project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº 825473.*

# Change Log

| Version | Author | Description of Change |
|---|---|---|
| V0.1 | Usman Wajid, John Beattie (ICE) | Initial Draft with contribution in all sections |
| V0.2 | Usman Wajid, John Beattie (ICE) | Further contributions under various sections for the completion of the deliverable |
| V0.3 | Marco Merlini (THALIT) | Internal peer-review process |
| V0.4 | Usman Wajid, John Beattie (ICE) | Updated in light of comments received from internal reviewer |
| V0.5 | Usman Wajid | Review and final updates |
| V0.6 | Usman Wajid | Packaging for submission to the EC |
| V1.0 | Eduardo Quiñones (BSC) | Ready to be submitted to EC |
| | | *(Final Change Log entries reserved for releases to the EC)* |
| | | |
| | | |
| | | |

# Table of contents

Change Log ..... 2

1. Executive Summary ..... 4

2. Overview of Distributed Data Analytic Platform ..... 4

3. Data Landscape and Analytic Needs ..... 7

4. Data Analytic Platform Requirements ..... 8

4.1 Requirements Generated from The Use-Case Requirements ..... 9
4.2 Requirements Generated from the Use Case Data Sets ..... 16

5. Dependencies on ELASTIC Ecosystem ..... 20

5.1 Requirements on the ELASTIC Architecture ..... 21
5.1.1 Communication Requirements ..... 21
5.1.2 Security Requirements ..... 21
5.1.3 Real Time Requirements ..... 21
5.1.4 Low Power Requirements ..... 21

6. Summary ..... 21

7. Acronyms and Abbreviations ..... 22

8. References ..... 22

# 1. Executive Summary

The purpose of this ELASTIC deliverable D2.5 (Distributed Data Analytic Platform Requirements) is to gather and analyse the requirements concerning the design and development of a Data Analytic Platform that can realise extreme scale data analytics in the ELASTIC architecture. The requirements gathering and analysis activity takes into account the Description of Action (DoA) and the consultation with the use-case partners regarding the definition of the use-case scenario in the ELASTIC project. The data analytic needs in the use-case scenario are identified and translated into the requirement of the distributed data analytic platform. The analysis and prioritisation of the gathered requirements kicks off the relevant development and integration activities in the ELASTIC project.

The requirements gathering and analysis activity reported in this deliverable involved interaction with and within pilot partners, who had the main role to specify their requirements on the data analytic platform. The technical partners assisted the pilot partners in the translation and documentation of these requirements. The technical partners also assisted the pilot partners in analysing and prioritising the requirements for implement (i.e. development and integration) activities in the project.

This document captures the data analytic platform requirements that will be used in conjunction with the Use-case Requirement Specification and Definition (D1.1), Initial Sensing and Dataset Collected deliverable (D1.2), General Requirements of the Fog Architecture (D5.1) and Software Architecture Requirements and Integration Plan (D3.1) to provide specific implementation and integration guidelines to the technical partners in the ELASTIC project.

# 2. Overview of Distributed Data Analytic Platform

The distributed data analytic platform in ELASTIC is developed to cater for domain specific as well as generic needs of analysing data across the compute continuum. Such needs for distributed data analytics are important in certain domains, such as smart city where the combination of different data sources and the means to analyse data at different levels (from Edge to the Cloud) has an unprecedented impact on energy management, environment, traffic and transport services.

The need to analyse data in a distributed way is also important to deal with the following factors:

- **Distributed Systems** are largely built by interlinking multiple components (tools and services) that are often provided by different providers. While the analysis of data flowing through different components is efficient and cost effective, if performed on the Cloud, the end-to-end communication delays and the performance of the system (in terms of latency) increases and becomes unpredictable, making distributed analytics a more viable option to derive robust and real-time guarantees
- **The transfer** of data from mobile data sources is largely constrained by the availability of energy, as well as unstable communications, which causes communication delays, unstable data throughput, loss of data and temporal unavailability. This, coupled with the widely distributed fixed (i.e. not moving) data sources, means there is a need to analyse data in a distributed approach to address the aforementioned issues without constraining the mobility of the data sources

- **Security** is a continuously growing priority for organisation of all sizes, as it affects data integrity, confidentiality and potentially impacting on safety. However, the extraction of all IoT data, also including sensitive data, to the Cloud may pose security risks. Therefore, local or edge analysis of sensitive data for extraction of events or for performing local control actions is considered more secure approach

In order to address the need for distributed data analytics, there are several frameworks and technological solutions are being developed.

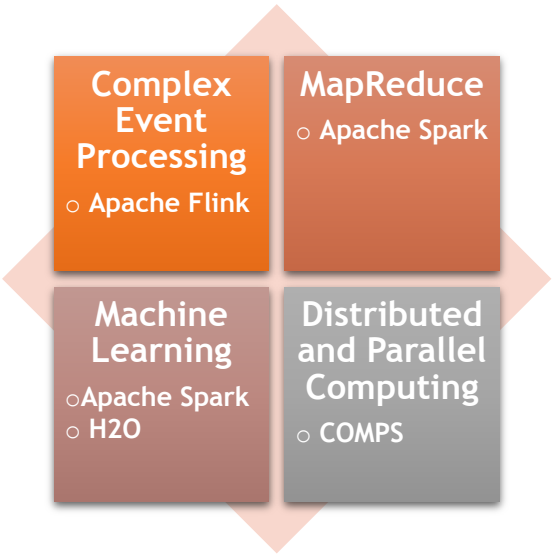


*Figure 1: Core components of the distributed data analytic platform*

- **Complex Event Processing** is a method of tracking and analysing (processing) streams of information and deriving a conclusion from them. Complex event processing (CEP) is event processing that combines data from multiple sources to infer events or patterns that suggest more complicated circumstances. The goal of complex event processing is to identify meaningful events (such as opportunities or threats) and respond to them as quickly as possible. CEP solutions are built on stream processing technologies to monitor a change of state i.e. when a measurement exceeds a predefined threshold of time, temperature, or other value. CEP provides insight into running operations by querying live data feeds, and correlating identified events against historical data to provide insight and analysis. In the ELASTIC domain, multiple sources of data (e.g. from video cameras, traffic signals, proximity sensors etc) can be combined in a CEP solution to provide a holistic picture of the operating transport management systems or the environment
  Relevant technologies to be investigated/implemented in ELASTIC include:
  - **Apache Flink** is an open-source and distributed stream processing framework that provides support for real-time complex event processing. Flink has been designed to run in all common cluster environments, perform computations at in-memory speed and at any scale
- **MapReduce** is a framework for processing parallelizable problems across large datasets using a large number of nodes. Processing can occur on data stored either in a filesystem or in a database. MapReduce can take advantage of the locality of data, processing it near the place it is stored in order to minimize communication overhead. In the ELASTIC Domain, MapReduce can be used for distributed pattern-based searching and distributed sorting in a distributed computing environment.

- **Apache Spark** is an open-source project for fast data processing and cluster computing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. Spark facilitates the implementation MapReduce functionality. It also supports both iterative algorithms, which visit their data set multiple times in a loop, and interactive/exploratory data analysis, i.e., the repeated database-style querying of data. Apache Spark uses memory and can use a disk for processing. Spark is able to execute batch-processing jobs between 10 to 100 times faster than the Hadoop-based MapReduce framework

- **Machine Learning** includes software that can learn from historic data. Machine Learning techniques give systems the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of training data. In the ELASTIC domain, Machine learning can be used for prediction of events (e.g. probability of pedestrians crossing the streets upon train arrival at the stop in certain periods of the day; probability of crowds at stops in certain periods of the days; etc.)
  Relevant technologies to be investigated/implemented in ELASTIC include:
  - **Apache Spark** provides a library (MLlib) that contains many algorithms and utilities for Machine Learning. The library includes popular algorithms such as Classification (logistic regression), Regression (generalized linear regression, survival regression), Decision Trees (random forests, gradient-boosted trees), Recommendation, Clustering (K-means, Gaussian mixtures (GMMs). Moreover, workflow utilities include Feature transformations: standardization, normalization, hashing, Machine Learning pipeline construction, model evaluation and hyper-parameter tuning and saving and loading models and pipelines
  - **H2O** is an open-source platform for distributed in-memory machine learning with linear scalability. H2O supports the most widely used statistical and machine learning algorithms including gradient boosted machines, generalised linear models and deep learning. The platform provides fast and parallel data processing from distributed sources
- **Distributed and Parallel Computing** frameworks enable the simultaneous use of multiple compute resources to execute software applications. Typically, the parallel computing frameworks offer programming models where the users specify the functions to be executed as asynchronous parallel tasks. At runtime the system exploits the concurrency of the code, automatically detecting and enforcing the data dependencies between tasks and spawning these tasks to the available resources, which can be nodes in a cluster environment.
  - **COMPS** provides a complete framework, composed by a programming model and a runtime system, which enables the development of parallel applications for distributed infrastructures at a very low cost because of two main reasons: (1) the model is based on sequential programming on top of popular programming languages (e.g. Java, Python), meaning that users do not have to deal with the typical duties of parallelization and distribution; and (2) the model abstracts the application from the underlying distributed infrastructure, hence COMPSs programs do not include any detail that could tie them to a particular platform (see Deliverable D3.1 [3] for further information).

# 3. Data Landscape and Analytic Needs

Based on the analysis of use-case scenario, as defined by the use-case partners, the ELASTIC platform is expected to collect and process data from different sources in order to provide a comprehensive representation of the transportation network dynamics. In terms of high-level requirements for the Distributed Data Analytic Platform, there are two envisioned scenarios:

- In the online scenario the platform is expected to provide users (pedestrians, car drivers, tram drivers) with information on relevant situations and/or with alerts on potential hazards
- In the offline scenario the platform is expected to deliver users (traffic engineers) with data useful to assess the transportation network performance so as to enable further optimisation.

The realisation of the above scenarios through the Distributed Data Analytic Platform will enable the use-case partners to investigate the interaction between the public and the private transport, from two perspectives, namely:

- Safety in the transport infrastructure
  - Avoidance of hazards for pedestrians, drivers, tram passengers
- Improvements in the network performance
  - Analyse and improve service levels for public transport
  - Analyse and improve service levels for private transport

The realisation of the Distributed Data Analytic Platform and addressing the above perspectives will require the ELASTIC platform to rely on a number of data inputs, including the data from following sources from the use-case domain i.e. the public transport infrastructure in the City of Florence. The data will be ported through the ELASTIC fog computing architecture for analysis in the distributed analytic platform. Examples of data sources include:

- Traffic light plans and status/transitions, through an API that is expected to be published by the Urban Traffic Control (UTC) centre and or by probes installed on the traffic lights or control units
- Traffic flow data, through an API published by the mobility supervisor. The mobility supervisor acts as a data collector for traffic data acquired by different sub-systems. Some dedicated traffic sensors might also be installed at the pilot site(s)
- Video cameras, to be installed at the pilot site(s)
- Tram position, as an output of the NGAP sub-system to be developed within the ELASTIC project
- Obstacle detection events, as an output of the ADAS sub-system to be developed within the ELASTIC project
- Traffic events (e.g. accidents, road restrictions etc.), with position, timestamp and severity through an API published by the mobility supervisor
- Urban public transportation data (bus stop with timetable, nearly real-time position of buses, expected/forecast at bus stops) through an API published by the mobility supervisor – real-time data might be available for a subset of buses only

The Distributed Data Analytic Platform is expected to perform analytics over heterogeneous data types to extract meaningful information that can be used for raising alerts or for decision making. The data type used by the system include:

- Live video streams recorded by the cameras, in order to detect higher-level events as presence of people (i.e. crowds in specific areas) and some behaviours (i.e. pedestrians crossing the street). The video streams are unlikely to be transmitted to the cloud due to bandwidth concerns. The video data will be processed at the edge and the output from that processing will be made available to the ELASTIC continuum, where the Distributed Data Analytic platform can perform any further processing if required. In case the video stream is required to be transmitted to the cloud, a private network featuring security mechanisms to avoid cyber-attacks will be provided.
- Tram position, in order to detect higher-level events as approaching/occupying/leaving an intersection or stop

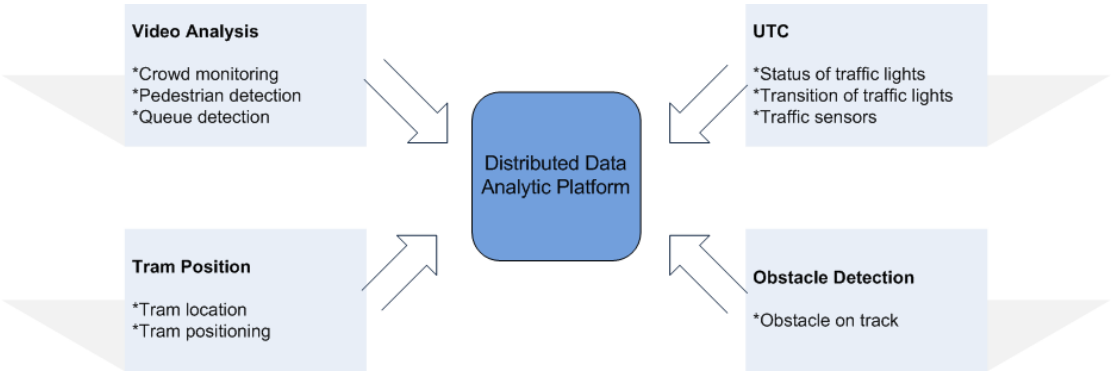


*Figure 2: Data sources from the use-case domain*

The outcome of the Distributed Data Analytic platform is expected to close the control loop by delivering information and alerts to the users of public transport infrastructure through the use of relevant devices and protocols e.g. through the use of vehicle-to-infrastructure and/or vehicle-to-vehicle communication.

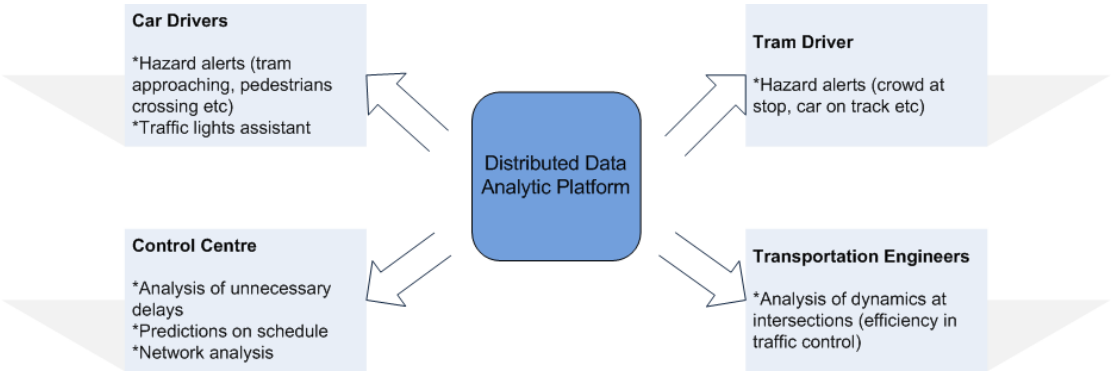


*Figure 3: Outcomes of the Distributed Analytic Platform*

Moreover, the analysis of historic traffic data from the use-case domain can reveal recurrent patterns in transportation network dynamics, as well as support traffic engineers in the evaluation of KPIs useful to assess performance of the overall transportation network, from different perspectives

# 4. Data Analytic Platform Requirements

As the data analytic services operate on the user's data and supply the user's applications, the requirements on the Distributed Data Analytics platform stem from the user's requirements and from the user's data sets.

## 4.1 Requirements Generated from The Use-Case Requirements

The use case requirements are specified in D1.1 [1]. The following table lists the Use Case Requirements distilled from D1.1 and various earlier documents, which have an impact on Data Analytics. Some of the requirements specified in D1.1, particularly concerning NGAP and ADAS systems, are considered strictly CONFIDENTIAL by use-case partners. The confidential requirements cannot be disclosed outside the ELASTIC consortium and therefore only a pointer to those requirements is provided in this document e.g. *[SYS-ELA-ADAS-REQ-097]*.

The following table uses the MoSCoW method to identify the priority of the requirement.

The MoSCoW method is a prioritisation technique used in software development to reach a common understanding with stakeholders on the importance of each requirement. In the ELASTIC project, MoSCoW method is used to prioritise the user-requirements for implementation activities according based on the following priorities:

- *Must*: Requirements marked as Must are critical for the user partners and also critical to highlight the value-proposition of ELASTIC platform.
- *Should*: Requirements marked as Should are important but not critical for the user partners.
- *Could*: Requirements marked as Could are desirable. They focus on some advance functionalities that can be offered by the Distributed Analytic platform but will require much long delivery
- *Won't*: Requirements marked as Won't (or will not) have been agreed by user partners as least critical and would require too long to develop. These will be addressed if time and resource permits during the project lifetime.

It is important to note that these are the priorities as they apply to Data Analytics not to the Elastic System as a whole. In particular, a priority of Wo't may mean that we think that Data Analytics plays no part in that requirement and that we expect it to be handled by other parts of the system.

Table 1: Requirements Extracted from Use-Case Partners

| Num. | Use-Case Scenario | Requirements | MoSCoW |
|---|---|---|---|
| UC1 | Data needs to be transferred to the Cloud from the tram | Real-time data analysis needs to be performed on the edge or in the Cloud to extract meaningful events or trigger relevant actions | M - MUST |
| UC2 | Data needs to be transferred to the Cloud from the trackside.<br>[SYS-ELA-FLO-REQ-002] | Same as above.<br>The source of the data may be on-board tram, track-side or central. | M - MUST |

| | | | |
|---|---|---|---|
| UC3 | Data needs to be transferred wirelessly to the Cloud from the trackside | Same as above.<br>The destination of the processed data may be on-board tram, trackside or central. | M - MUST |
| UC4 | Data needs to be transferred wirelessly from the tram to other edge devices | Elasticity of data analytic services across multiple edge nodes | S - SHOULD |
| UC5 | Data needs to be transferred to central data store for subsequent, off-line processing<br>[SYS-ELA-GEN-REQ-009] | Historic or combined data analysis in the Cloud | M - MUST |
| UC6 | A cloud application shall use positioning information from NGAP system to show train position along the configured rail track.<br>[SYS-ELA-GEN-REQ-010] | The analytic services may need to supply a GUI application with positioning information.<br>**At this stage it is not clear if analytics would need to be involved.* | C - COULD |
| UC7 | Analysed data (events and alarms) needs to be transferred to the Cloud from the tram<br>[SYS-ELA-ADAS-REQ-017] | Elasticity of analytic services to be able to process events and alarms from edge to the Cloud | M - MUST |
| UC8 | A cloud application shall use information from ADAS system to implement statistical analytics about relevant events along the line.<br>[SYS-ELA-GEN-REQ-018] | The analytic services must provide statistical information on relevant events. | M - MUST |
| UC9 | Both Rail Track Status and Electric Energy consumption systems shall process data from on board sensors, so the outputs of these systems do not necessarily need cloud processing.<br>[SYS-PredMaint-001] | The analytic services could process track status and electrical energy consumption. | C - COULD |
| UC10 | Output from Cloud analytic services should be available to multiple applications/nodes | The analytic services (running on the Cloud) must offer appropriate interfaces to provide analytic outcomes to different applications/ nodes | M - MUST |

| | | | |
|---|---|---|---|
| UC11 | Output from data analytics needs to be transferred to tram drivers [SYS-ELA-FLO-REQ-29] | Same as above | M - MUST |
| UC12 | System shall distribute information/alerts to other systems (e.g. mobility supervisor) [SYS-ELA-FLO-REQ-30] | Same as above | M - MUST |
| UC13 | System shall deliver information/ alerts to car drivers to provide them with road conditions [SYS-ELA-FLO-REQ-27] | Same as above | M - MUST |
| UC14 | System shall deliver information/alerts to vulnerable road users (VRUs - e.g. pedestrians, cyclists) to provide them with information/alerts on hazards [SYS-ELA-FLO-REQ-28] | Same as above | M - MUST |
| UC15 | Data transfer mechanisms needs to be able to handle raw data | Analytic services are able to handle raw data from heterogeneous sources | M - MUST |
| UC16 | Data transfer needs to be able to handle images | Analytic services are able to analyse images | M - MUST |
| UC17 | Data transfer needs to be able to handle streaming video | Analytic services are able to analyse video. It is unlikely that will be streamed to the Cloud. | C - COULD |
| UC18 | Data transferred will be used for monitoring [SYS-ELA-ADAS-REQ-019] | Real-time stream analysis to provide monitoring interface | M - MUST |
| UC19 | Data transferred will be used for control [SYS-ELA-ADAS-REQ-019] | Real-time stream analysis to support monitoring as well as control | S - SHOULD |
| UC20 | Data transferred will be used for alerts | Real-time stream analysis to provide alerts through appropriate interfaces | M - MUST |
| UC21 | System should be capable of transferring broadcast/multicast messages | Analytic services are able to deal with multicast messages in order to avoid | S – SHOULD |

| | | | |
|---|---|---|---|
| | | unnecessary processing | |
| UC22 | System should be capable of transferring (& generating?) acknowledgement messages | The analytic services are able to generate appropriate response | S – SHOULD |
| UC23 | System should be capable of prioritising messages | The analytic services are able to recognise the priority associated with different messages and adjust their processing accordingly | C – COULD |
| UC24 | NGAP system output data shall be linked to the date and time they refer to.<br>[SYS-ELA-GEN-REQ-011]<br>[SYS-ELA-ADAS-REQ-020] | Data analytic services must be able to process time stamped information and maintain this information in their outcome | M - MUST |
| UC25 | [SYS-ELA-ADAS-REQ-020] | Data analytics must be able to process geo-localisation information and maintain this information in their outcome | M - MUST |
| UC26 | [SYS-ELA-ADAS-REQ-020] | Analytic services must be able to process source identifier and maintain source identifier in their outcomes | M - MUST |
| UC27 | [SYS-ELA-ADAS-REQ-023] | Analytic services implement standardised security, authentication and authorisation protocols | M - MUST |
| UC28 | [SYS-ELA-ADAS-REQ-029 to 34] | Data analytics services are capable of processing high definition camera data and be able to support the CAM_PROT protocol | M - MUST |
| UC29 | [SYS-ELA-ADAS-REQ-035 to 37] | Data analytics services are capable of | M – MUST |

| | | | |
|---|---|---|---|
| | | processing high resolution LiDAR data | |
| UC30 | [SYS-ELA-ADAS-REQ-038 to 45] | Data analytic services are capable of processing the output from the specified radar device. | M – MUST |
| UC31 | [SYS-ELA-ADAS-REQ-047] | Probably a hardware requirement and not a requirement on the data analytic platform.<br>**If this suggests pulling data from an Ethernet interface then the prioritisation will change* | W – WON'T |
| UC32 | [SYS-ELA-ADAS-REQ-089] | More of a requirement on the platform rather than analytics. However, Data Analytics must be able to receive data from the tram via a base station i.e. not directly from the tram. | M - MUST |
| UC33 | [SYS-ELA-ADAS-REQ-090] | Same as above.<br>The analytic services will process the data coming in the ELASTIC platform. | W – WON'T |
| UC34 | [SYS-ELA-ADAS-REQ-096] | Data analytic services dealing with the ADAS system data are compliance with the specified cybersecurity standard | M – MUST |
| UC35 | [SYS-ELA-ADAS-REQ-097] | This is not a requirement on the analytic platform. The ELASTIC platform will cover the security of the data. However, the analytics services may need access to | W – WON'T |

| | | | |
|---|---|---|---|
| | | secure data in unencrypted form | |
| UC36 | [SYS-ELA-ADAS-REQ-098] | The data analytic services satisfy GDPR for managing personal data | M – MUST |
| UC37 | Implementation of I2V communications protocols.<br>[SYS-ELA-FLO-REQ-15] | The analytic services need to able to convert to/from I2V communications protocols | M - MUST |
| UC38 | Analysis of live video feed to detect relevant event (e.g. pedestrians crossing street)<br>[SYS-ELA-FLO-REQ-016] | The analytic services need to be able to analyse live video feeds. | M – MUST |
| UC39 | System shall feature storage at the edge for temporary storage of big data.<br>[SYS-ELA-FLO-REQ-017]<br>[SYS-ELA-GEN-REQ-009] | The analytic services need to be able to send processed data to local storage system | M – MUST |
| UC40 | System shall acquire positional data from NGAP.<br>[SYS-ELA-FLO-REQ-021] | The analytics services must be able to process NGAP data | M - MUST |
| UC41 | System shall acquire obstacle detection data from ADAS. This will include relevant information about detected or possible obstacles and whether the track is free or obstructed.<br>[SYS-ELA-FLO-REQ-22]<br>[SYS-ELA-GEN-REQ-012]<br>[SYS-ELA-GEN-REQ-013]<br>[SYS-ELA-GEN-REQ-014] | The analytics services must be able to process ADAS data relating to possible obstacles. This information will not include video. | M – MUST |
| UC42 | System shall manage data about presence/behaviour of people (pedestrians crossing the street, crowd at tram stop)<br>[SYS-ELA-FLO-REQ-23] | The analytics services must be able to process data relating to the presence/behaviour of pedestrians. | M – MUST |
| UC43 | System shall acquire data about state/state transitions of traffic lights controlled by the urban traffic control system.<br>[SYS-ELA-FLO-REQ-24] | The analytics services must be able to process data from Urban Traffic Control System | M - MUST |

| | | | |
|---|---|---|---|
| UC44 | System shall produce information/alerts on traffic events to provide road users and operation with (nearly) real-time information and alerts.<br>[SYS-ELA-FLO-REQ-26] | The analytics services must be able to generate information/ alerts in near real-time. | M - MUST |
| UC45 | System shall support traffic optimization<br>[SYS-ELA-FLO-REQ-31] | The analytics services must be able to process traffic optimization. | S - SHOULD |
| UC46 | System shall provide traffic engineers with KPIs on transportation network performance.<br>[SYS-ELA-FLO-REQ-032] | The analytics services must be able to generate transportation network performance information. | S - SHOULD |
| UC47 | System shall provide traffic engineers with KPIs on interactions between public transport (tramway), private transport (cars) and other factors (pedestrians)<br>[SYS-ELA-FLO-REQ-033] | The analytics services must be able to generate information on interaction between various types of road users | M - MUST |
| UC48 | [SYS-ELA-FLO-REQ-048] | The analytics services must be able to process CAM_PROT data containing the specified information. | M - MUST |
| UC49 | [SYS-ELA-FLO-REQ-049] | The analytics services must be able to process CAM_PROT data containing relative information. | M - MUST |
| UC50 | [SYS-ELA-FLO-REQ-050] | The analytics services must be able to process CAM_PROT data containing data indicating which objects are obstacles. | M - MUST |
| UC51 | [SYS-ELA-FLO-REQ-051] | The analytics services must be able to process CAM_PROT data containing various status information of the device. | M - MUST |

| | | | |
|---|---|---|---|
| UC52 | [ SYS-ELA-FLO-REQ-51 to 55] | The analytics services must be able to process CAM_PROT data in the required formats | M - MUST |
| UC53 | System shall compute various track geometry parameters.<br>[SYS-PredMaint-RailTrack-10 to 60] | The analytics services are not required to process this information at this stage. | C - Could |
| UC54 | System shall associate tram power consumption with speed, position, kilometric chainage, and vehicle weight.<br>[SYS-PredMaint-EnergyConsupt-10 to 70] | The analytics services are not required to process this information at this stage. | C – Could |
| UC55 | Track measurement data might be sent to the cloud to associate them to successive rail track status monitoring, thus associating possible electric power energy peaks (in turn due to large tram weight, acceleration, speed, etc.) with next resulting rail track consumed status. In fact, such correlation might help determining a possible source cause of such rail track consumed status.<br>[SYS- PredMaint-EnergyConsumpt-74] | The Analytics services could process track geometry data. | C - COULD |
| UC56 | Upon termination of tram route, the system shall accurately (time/space) synchronize the collected data in order to let them be displayed as overlaid to a route map, as associated to the progressive registered positions of the tram vehicle.<br>[SYS-PredMaint-EnergyConsumpt-90] | The Analytics Services must synchronize stored energy consumption data to allow it to be used by a GUI application. | |

## 4.2 Requirements Generated from the Use Case Data Sets

Data sets defining the format of the data for Predictive Maintenance, NGAP and ADAS are defined D1.2 [2].

The follow table list the requirements that these data sets places on the Distributed Data Analytic platform.

| | | | |
|---|---|---|---|
| UC57 | System shall process laser-generated rail coordinates at a rate of 1000 measures/s. [DATA-PredMaint-Laser] | Analytic services must process rail coordinates and produce rail track profile; horizontal, vertical and 45˚orientation track wear status, and track gauge. This information will probably be supplied by NGAP. | C - COULD |
| UC58 | System shall process acceleration and 3-D angular speed from inertial platform at rates of up to 100 measures/s. [DATA-PredMaint-Laser] | Analytic services must process data from inertial platform and produce 3-D accelerations and angular speeds; vehicle elevation level; and track curvature radius. | M - MUST |
| UC59 | System shall process HD, coloured video in H.264 format at 25 fps from cameras pointing at each rail. [DATA-PredMaint-Video] | Analytic services must process video data and produce videos of both rails. This information is likely to be processed at the edge. | C - COULD |
| UC60 | System shall process triaxial acceleration from both left and right wheels at 400 measures/s. [DATA-PredMaint-Accel] | Analytic services must process acceleration data and triaxial acceleration from both left and right wheels at 400 measures/s. | M – MUST |
| UC61 | System shall process kilometric chainage and tram speed at a rate of 500 measures/s [DATA-PredMaint-Chain] | Analytic services must process kilometric chainage and produce positioning information. | M – MUST |
| UC62 | System shall process GPS lat-long data at a rate of 5 measures/s [DATA-PredMaint-GPS] | Analytic services must process GPS data to record vehicle position | M – MUST |
| UC63 | System shall process HD, coloured, video in H.264 format at a rate of 25 fps from tram's forward-facing camera. [DATA-PredMaint-Video] | Analytic services must process H.264 video to produce anonymized video. This information is likely to be processed at the edge | C - COULD |
| UC64 | System shall process tram's power consumption at a rate of 10 measures/s. [DATA-PredMaint-Power] | Analytic services must time-space correlates tram's power consumption | M – MUST |

| | | | |
|---|---|---|---|
| UC65 | System shall process the tram's weight at a rate of 10 measures/s.<br>[DATA-PredMaint-Weight] | Analytic services must time-space correlates tram's weight. | M – MUST |
| UC66 | System shall provide acceleration and 3-D angular speed.<br>[DATA-NGAP-Accel] | This data is not available to the Analytic services. However, the variance of the accelerations and 3-D angular speeds is supplied, at the number of unprocessed IMU packets, and the variance in speed. These can be used to give an indication of the reliability of the on-board data. | C - Could |
| UC67 | System shall process GPS coordinates and time synchronization data.<br>[DATA-NGAP-GPS]<br>[SYS-ELA-GEN-REQ-003] | This data is not available to the Analytic services. However, the number of unprocessed GPS reading and the variance of the calculated position are supplied and can be used to give an indication of the reliability of the data. | C - Could |
| UC68 | System shall process radar and time synchronization.<br>[DATA-NGAP-GPS]<br>[SYS-ELA-GEN-REQ-003] | This data is not available to the Analytic services. However, the number of incorrectly processed radar reading is supplied and can be used to give an indication of the reliability of the data. | C - Could |
| UC69 | System shall provide various ADAS inputs.<br>[DATA-NGAP-ADAS-Inputs]<br>[SYS-ELA-GEN-REQ-003] | Considering requirement [SYS-ELA-GEN-REQ-004], these inputs will not be processed by the analytic services.<br>**If this requirement changes then the prioritisation will also change* | W – Won't |
| UC70 | System shall process various ADAS data sets (processed video and binary/structured-text) from the ADAS output, which can provide information on obstacle typology, distance, speed, | Analytic services must process ADAS data sets.<br>The analytics can be used to perform statistical analysis on the numbers of: | M – MUST |

| | | | |
|---|---|---|---|
| | criticality level, position and time synchronization.<br>[DATA-NGAP-ADAS-Outputs]<br>[SYS-ELA-GEN-REQ-005] | - Accidents;<br>- False alarms;<br>- Pedestrians crossing the line;<br>- Obstructions of each of the various types<br>- The traffic on the crossing path between the tracks and the road. | |
| UC71 | Possible analytics relating to public transport:<br>- Frequency of service (absolute value, regularity);<br>- Occupation of intersections;<br>  ○ Duration;<br>  ○ Frequency;<br>  ○ Pattern (tram in one direction, two trams in opposite directions, two trams in the same direction);<br>- Number of outages due to external factors;<br>- Presence/entity of crowds at tram stops;<br>[D1.1-Section-7.2.4] | The analytic services should be able to assist in the collection of the data of this data. However the requirement may also involve some central or "back-office" processing which may be outside of the scope of Elastic. | S - SHOULD |
| UC72 | Possible analytics relating to private transport:<br>- Occupation of intersections;<br>  ○ (Average) Green phase duration;<br>  ○ Number of occurrences of pre-emption due to tram priority;<br>- Pedestrians crossing while traffic light is red; | As above | S - SHOULD |
| UC73 | Possible analytics relating to safety:<br>- Accidents;<br>  ○ Quantity;<br>  ○ Temporal and spatial distribution. | As above | S - SHOULD |

# 5. Dependencies on ELASTIC Ecosystem

ELASTIC aims to develop an ecosystem incorporating software components from multiple computing areas, including distributed data analytics, embedded computing, IoT, CPS, software engineering, HPC, and edge and cloud technologies.

The ELASTIC software development ecosystem [3] (in Figure 4) describes the positioning, relationships and/or dependencies of the Distributed Data Analytic Platform on other components in the ecosystem.

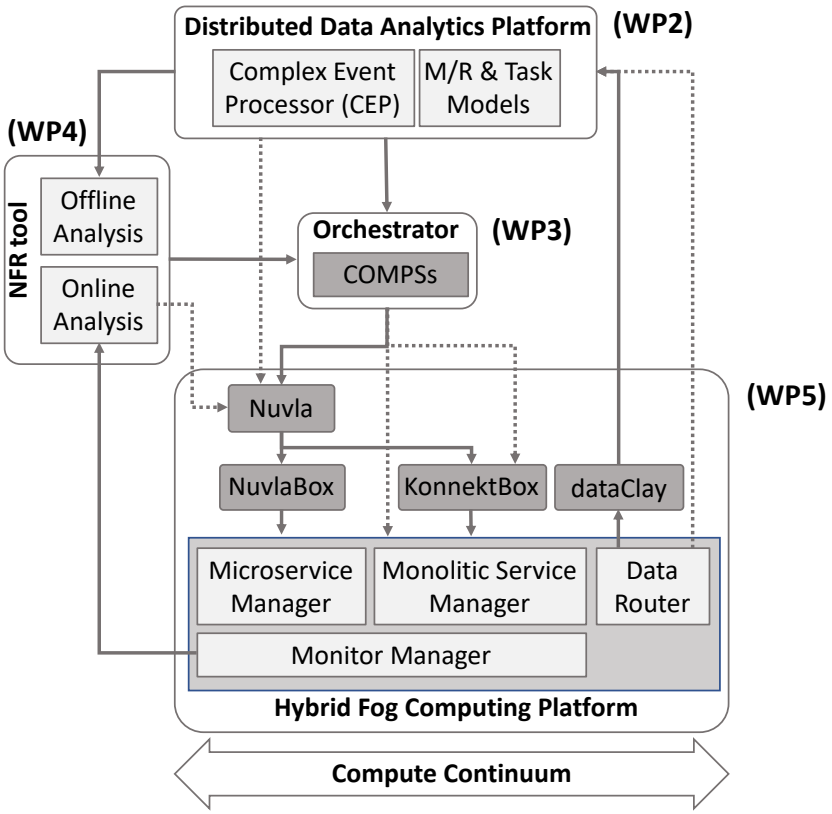


*Figure 4: ELATIC Software Development Ecosystem*

The Distributed Analytic platform will be composed of a number of loosely coupled Data Analytic Services that will perform application-specific analytic tasks. An overview of the technologies used by Data Analytic Services is provided in Section 2. Each Data Analytic service can be split into a *main program* and a number of *remote methods*.

The main program will be centrally located and may contain the API's used by any user applications. It will contain calls to the remote methods.

The remote methods will be elastically deployed to the most appropriate processing units and will contain the actual data analytical operations. They run as docker containers under the control of the Microservice Manager and receive the data to be processed from the Data Router, a component in the Hybrid Fog Computing Platform.

## 5.1 Requirements on the ELASTIC Architecture

### 5.1.1 Communication Requirements

The data analytic services rely on the ELASTIC platform for all their communication needs. Individual remote methods will need access to specific data sources some of which will be mobile (tram based) and others static (track side and possibly even from control centres). Over the course of a tram journey, it is quite likely that the source of the data and/or the route taken by the data may change. Such a change could also be considered a criterion for redeploying the analytic services. Similarly, consumers of any output generated by the analytics may be widely distributed and may be upstream (away from edge) or downstream (towards edge). The may also change over time.

### 5.1.2 Security Requirements

The use-case requirements place security requirements on the system as a whole. The analytics services needs to work within this secure area. The analytics MUST be able to trust the data supplied by the platform (to the tram should not be stopped because of a false event generated by a hacker). Some of the analytics will need access to encrypted file store for “local” storage of data for off-line processing.

### 5.1.3 Real Time Requirements

The use-case requirements also place real time constraints on the whole ELASTIC system and the analytic services need to work within these constraints. The analytic services do not place any additional requirements on the system.

### 5.1.4 Low Power Requirements

Should the underlying computing infrastructure enter a low power state, the analytic services may need to be redeployed to another suitable infrastructure. The remote methods will not be monitoring the platform and so rely on the Monitor Manager to trigger the redeployment. Ideally, any redeployment should be achieved with no loss of data and no repetition of data.

# 6. Summary

This deliverable describes the requirements on and from the Distributed Data Analytic Platform in the ELASTIC architecture. As the analytic services operate on the user’s data and supply the user’s applications, these requirements are largely based on the use-case requirements and hence are likely to change with changes to the use-cases.

The general requirements placed on the rest of the ELASTIC platform are largely the same as those required by the use cases namely security, reliable data transfer to and from disparate sources and targets, near real time response and low power requirements. The main difference being that the Data Analytic services are “in the middle”, so for example, if there is a requirement for data to be encrypted, then the analytics will need to have access to the decrypted data in order to process it and be able to encrypt any results generated. In addition, there are requirements on the system to monitor the non-functional requirements of the system and to be able

to generate a trigger to enable the analytics to be redeployed if those requirements are not met.

The requirements of the Distributed Data Analytic platform described in this deliverable set the foundation for the development of analytic functionality in the ELASTIC ecosystem. The next step would be to analyse these requirements for translation into technical implementation tasks that will be carried out by distributed development teams in the ELASTIC project.

# 7. Acronyms and Abbreviations

A list of widely used acronyms is provided below:

- ADAS – Advanced Driving Assistant System
- CA – Consortium Agreement
- D – deliverable
- DoA – Description of Action (Annex 1 of the Grant Agreement)
- EB – Executive Board
- EC – European Commission
- FCW – Forward Collision Warning
- FoV – Field of View
- GA – General Assembly / Grant Agreement
- GDPR – General Data Protection and Regulation
- GPS – Global Positioning System
- GUI – Graphical User Interface
- HPC – High Performance Computing
- IMU – Inertial Measurement Unit
- IPR – Intellectual Property Right
- KPI – Key Performance Indicator
- M – Month
- MS – Milestones
- NGAP – Next Generation Autonomous Positioning
- PM – Person month / Project manager
- VRU – Vulnerable Road Users
- WP – Work Package
- WPL – Work Package Leader

# 8. References

[1] ELASTIC, “D1.1. Use case requirement specification and definition”. 2019.

[2] ELASTIC, “D1.2. Initial sensing and datasets collected”. 2019.

[3] ELASTIC, “D3.1. Software architecture requirements and integration plan”. 2019