ONTIC D2.1: Requirement Strategy

Chatzi Sotiria, Maravitsas Nikolaos, Elena Baralis, Daniele Apiletti, Luigi Grimaudo, Paolo Garza, Silvia Chiusano, Tania Cerquitelli, Philippe Owezarski, Maria Salazar, et al.

Big Data Network Traffic Summary Dataset

Deliverable
Requirements Strategy

ONTIC Project
(GA number 619633)

Deliverable D2.1
Dissemination Level: RESTRICTED

Authors

Chatzi Sotiria, Maravitsas Nikolaos
ADAPTIT

Elena Baralis, Daniele Apiletti, Luigi Grimaudo,
Paolo Garza, Silvia Chiusano, Tania Cerquitelli
POLITECNICO DI TORINO

Philippe Owezarski
CNRS

Maria Fernanda Salazar
ERICSSON

Alberto Mozo, Alex Martínez
UPM

Version

ONTIC_D2.1.2014.05.28.1.1

Version History

Previous version	Modification date	Modified by	Summary
2014.04.03.1.0	2014-4-3	ADAPTIT	Initial ToC
2014.05.01.1.0	2014-5-1	ADAPTIT	Draft 1
2014.05.16.1.0	2014-16-5	ADAPTIT	Draft 2
2014.05.23.1.5	2014-23-5	UPM	General review
2014.05.26.1.1	2014-26-5	ADAPTIT	Review
2014.05.28.1.0	2014-28-5	EMC ADAPTIT	Review
2014.05.28.1.1	2014-29-5	UPM	Quality assurance

Quality Assurance:

	Name
Quality Assurance Manager	**Sandra Gómez (UPM)**
Reviewer #1	**Spyros Sakellariou (ADAPTIT)**
Reviewer #2	**Fernando Arias (EMC)**

Table of Contents

List of figures

List of tables

1. Acronyms and Definitions

Acronyms

Table 1: D2.1 Acronyms

Acronym	Defined as
BIG	Big Data Public Private Forum
EP	Enforcement Point
FP	Framework Program
Gbps	Gigabits per second
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
ISP	Internet Service Provider
NOSQL	Not-only SQL
ONTIC	Online Network Traffic Characterization
ONTS	ONTIC Network Traffic Characterization Dataset
QoS	Quality of Service
RDBMS	Relational Database Management System
SQL	Structured Query Language
PC	Policy Controller
UI	User Interface
WP	Work Package

2. Executive Summary

ONTIC is a project that aims at creating a framework which will enable accurate traffic characterization over vast networks. The typical amount of data that goes through the core network of an ISP during a day, is at least in the magnitude of Terabytes. Thus, ONTIC's role is to design and implement:

- Massively scalable online architectures to characterize network traffic streams and identify traffic patterns, develop models that predict forthcoming traffic, so as to detect deviations from normal behavior in real time.
- Massively scalable offline algorithms to perform traffic analysis over a vast collected data set. This will be achieved by new and innovative scalable data mining algorithms that exploit Big Data and distributed computing techniques and can run over multi-node clusters.

The outcome of the above, will be a generic and autonomous platform that can be used as the core engine for developing scalable and performance sensitive network analysis applications.

Deliverable D2.1 is responsible for introducing a Requirements strategy for the ONTIC project. The Requirements strategy, consists of a roadmap and a group of guidelines on how the requirements of the ONTIC network traffic dataset, the Big Data Architecture, the use cases and the offline and online algorithms will be defined. Additionally, relations and implication among requirements from different work packages will be detailed.

Firstly, the objectives of this deliverable will be stated, as well as its scope. The intended audience will be addressed and useful previous readings will be suggested. Following, an outline of the ONTIC project as well as the relation between the technical WPs will be presented. Subsequently, the ONTS Big Data Architecture requirements strategy will be described. Then, a first estimation of the requirements of offline and online analytics algorithms will be made. Finally, the use cases, their requirements and the methodology to extract them will be displayed.

3. Scope

Deliverable D2.1 is describing part of the work implemented in the framework of the Task T2.1 entitled: “Requirements and Functionalities of a Big Data Architecture for network traffic summary dataset”

Deliverable D2.1 deals with the strategy that will be followed to extract the requirements for (a) ONTIC network traffic dataset, (b) the Big Data Architecture, (c) the analytic algorithms to be used in ONTIC, and (d) the Use Cases representing real life scenarios of telecom networks.

The main objective of deliverable D2.1 is to give a clear idea of the way of thinking and the guidelines that will help the ONTIC project to define the ONTIC network traffic dataset, the Big Data Architecture, the offline and online analytics algorithms and the use cases. Additionally, deliverable D2.1 will highlight and address relations and implication among requirements from different work packages.

Deliverable D2.1 gives a short description of the main ideas of ONTIC project and its technical aspects. After covering shortly the main technical lines, an effort is made to interconnect the details of each technical WP, explaining how the evolution of one WP affects the evolution of the others and vice versa. Yet again, the project on its whole, even the non-technical WPs, are presented, in an effort to portray in a brief and concise manner the mindset behind the conception of ONTIC.

This deliverable describes how to define both the requirements of the ONTIC network traffic dataset and the Big Data Architecture in WP2, using a Data Value Chain model adopted by relevant industry and academic institutions.

Additionally, the approach on the way the requirements for the offline (WP3) and online (WP4) algorithms are envisaged, is deployed.

Deliverable D2.1 outlines how Use Case requirements are going to be developed. All three WP5 Use Cases will play a crucial role on the development of both online and offline algorithms because these algorithms will be integrated in the prototypes implementing Use Cases. According to these, specific offline or online algorithms can be rendered more or less appropriate.

The final target of deliverable D2.1 is to examine, explain and visualize the methodology that will be used for the extraction of WP5 Use Case requirements. Deliverable D2.1 will outline the process to define Use Case requirements using Agile paradigms, like Scrum.

4. Intended audience

This document is oriented to partners participating in WP2, WP3, WP4 and WP5.

5. Suggested previous readings

It is expected that partners' background is sufficient to address the contents of this document, however, some previous readings are suggested.

1. FP7 Big Data Public Private Forum (BIG) project, delivered a public document titled "D2.2.1 First Draft of Technical White Papers" that is a good introduction to Big Data state-of-the-art.
2. A survey about the application of machine learning techniques to traffic classification can be found in "A survey of techniques for Internet Traffic Classification using Machine Learning", Nguyen and Armitage, IEEE Communication Surveys & Tutorials, 2008.
3. A broad overview of Agile methodology from one of its initiators can be found in "Agile software development (Vol. 2006). Cockburn, A. , Boston: Addison-Wesley".

6. Outline of ONTIC project and WP relationships

Accurate identification and categorization of network traffic according to application type is an important element of many network management and engineering tasks related with QoS, capacity planning and detection of network attacks. Terabytes of data may be transferred through the core network of a typical ISP every day. Moreover, an exponential growth, of more than 50 billion of connected devices to Internet, is expected. Therefore, this scenario hampers network data capture and analysis. Proactive and dynamic QoS Management, Network detection intrusion and Early detection of congestion network problems, among other applications in the context of network management and engineering, can benefit from the existence of an accurate and scalable mechanism for online characterization of network traffic patterns evolution.

To this end ONTIC project will investigate, implement and test:

1. A novel architecture of scalable mechanisms and techniques to be able to a) characterize online network traffic data streams, identifying traffic patterns evolution, and b) proactively detect anomalies in real time when hundreds of thousands of packets per second are processed.
2. A completely new set of scalable offline data mining mechanisms and techniques to characterize network traffic, applying a big data analytics approach and using distributed computation paradigms in the cloud, on extremely large network traffic summary datasets, consisting on trillions of records.

ONTIC project will integrate offline and online mechanisms and techniques into an autonomous network traffic characterization system to be used as cornerstone, of a new generation of scalable and proactive network management and engineering applications.

Additionally, ONTIC project will generate a petabyte size dataset composed of real network traffic summaries obtained during 24 months from a set of data flows (1.5 Gbps on average) that cross the core network of a medium size ISP that participates in the ONTIC consortium. The contents of this dataset will be anonymized and made publicly available at the end of the project to foster new research initiatives in the field of big data analytics.

Figure 1 and Figure 2 illustrate a technical overview of the ONTIC project and the relations among ONTIC work packages.

WP2 (Big Data Network Traffic Characterization) deals with network packet extraction and creation of the ONTIC Big Data Network Traffic Summary Dataset to be processed later by batch/offline (WP3) and stream/online (WP4) analytics algorithms and mechanisms. As previously indicated, this dataset will be composed of a collection of packet headers captured at a ratio of 1.5 Gbps, what implies that about 200,000 packets per second will have to be processed. Storing only packet headers, 1 Tera Byte of information per day will be generated, and so, the ONTIC dataset will reach a size of about 720 Tera Bytes. Consequently, a Big Data Architecture will be required to be designed in order to cope with velocity (200,000 packets per second) and volume (720 Tera Bytes) Big Data characteristics of the ONTIC dataset.

WP3 (Scalable Offline Network Traffic Characterization) and WP4 (Scalable Online Network Traffic Characterization) will test their analytics mechanisms and algorithms using concrete pieces of the ONTIC dataset (e.g. the collection of packet headers extracted during a couple of months) in their own local laboratory deployments. Nevertheless, during the last six months, final integration and validation tests will be done in EMC’s Analytics Workbench composed by a

1000 Hadoop cluster with 50 Peta bytes of total disk space. To this end, the ONTIC dataset will be previously uploaded to EMC's Analytics Workbench.

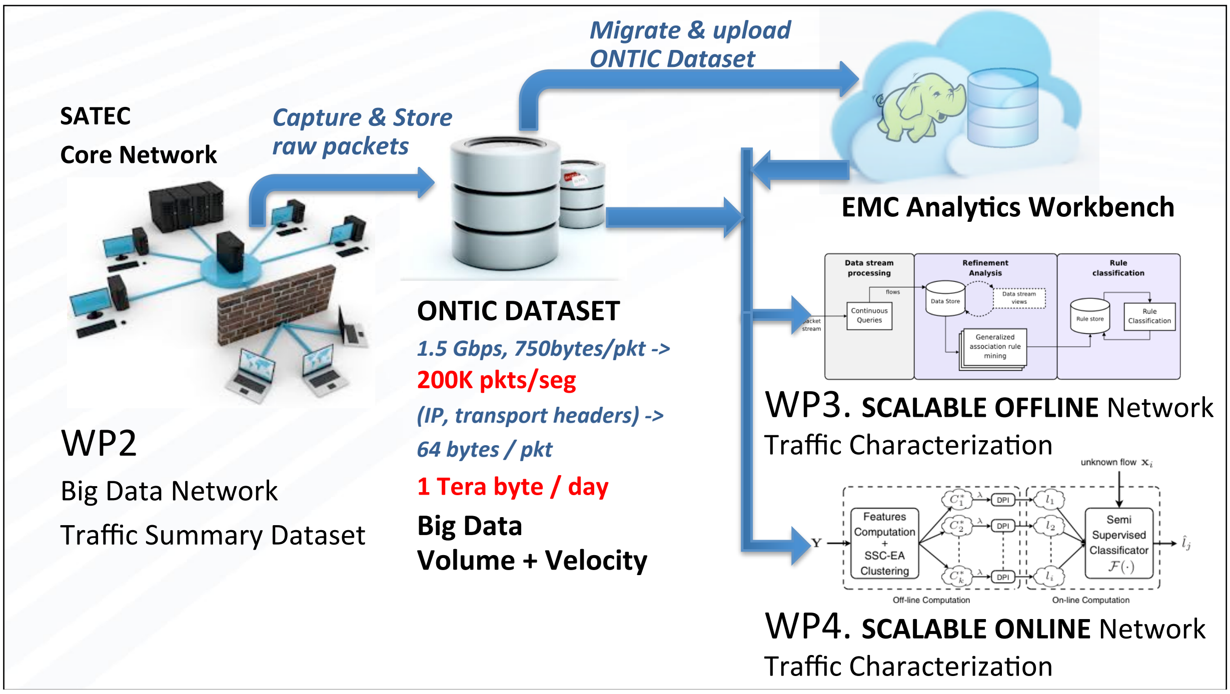

Figure 1: ONTIC Project Overview and WP relationships.

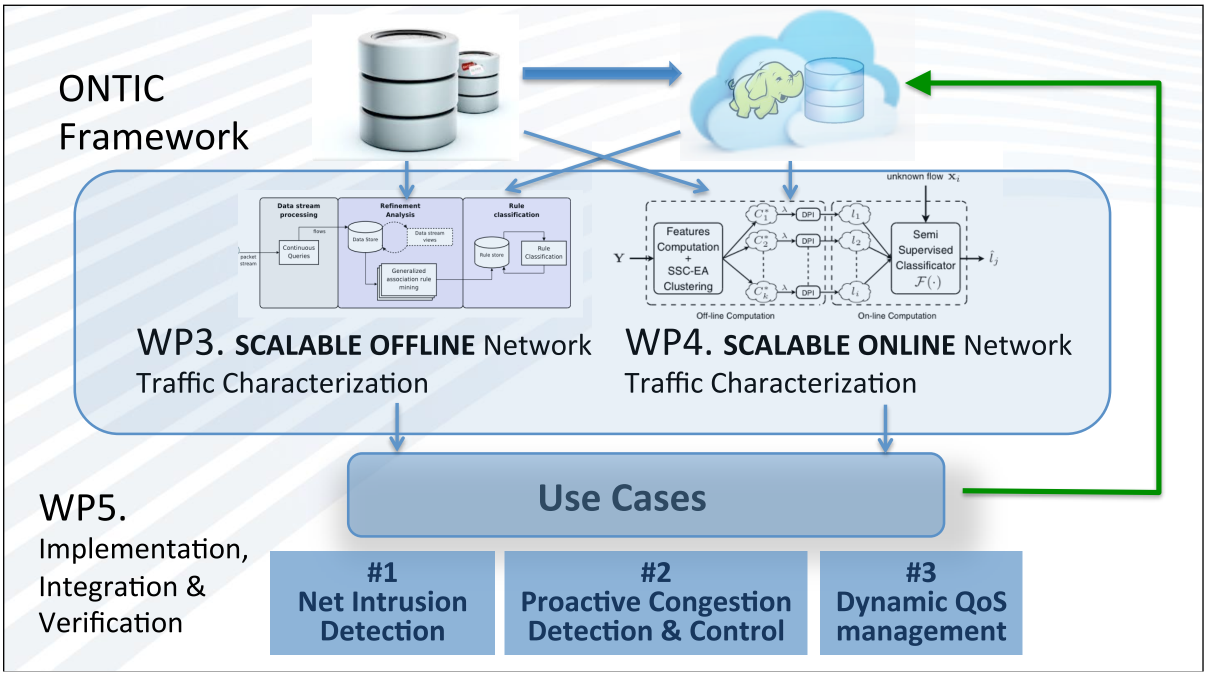

Figure 2: ONTIC Project Overview and WP relationships (cont.).

Specifically, WP3 will design and develop a scalable analysis system for batch/offline characterization network traffic traces. Algorithms developed in the ONTIC framework will run on top of the de-facto Big Data standard Hadoop platform using the stored packet headers previously in the ONTIC dataset. Furthermore, the Spark framework will be explored, thanks to the promising results of the latest reports. On the other side, WP4 will address, with the design of a framework composed of techniques that can mostly be qualified as semi supervised, traffic characterization that is able (1) to catch traffic patterns evolution and (2) to detect network anomalies and intrusions. This system will have to run online with a reactive level close to real-time and to cope with the huge amount of data and their high dimensionality level.

WP5 (Implementation, Integration and Verification) will mainly deal with (a) the integration of offline and online mechanisms into an autonomous network traffic characterization system and (b) the specification, designing, implementation and validation of three prototypes based on the following Use Cases: (1) Network Intrusion Detection, (2) Proactive Congestion Detection and Control Systems, and (3) Dynamic QoS Management. These Use Cases will determine which algorithms in WP3 and WP4 will have to be investigated, mainly from a scalability perspective. Specifically, clustering, classification and prediction algorithms will be considered as candidates for WP3 and WP4 research.

It must be noted that technical work packages (WP2, WP3, WP4 and WP5) are connected among them, and therefore, the requirements strategy for these work packages will have to address the concrete implications among them. In particular, WP3 and WP4 requirements could affect to WP2 ones, and WP5 requirements can imply to define specific requirements in WP3 and WP4.

Finally, WP1 (Project Management) covers the aspects concerned with the overall project management and coordination, and WP6 (Dissemination and Exploitation) aims to ensure the maximum awareness and visibility of project results.

7. ONTIC Dataset Big Data architecture requirements strategy for WP2

WP2 deals with the designing, deploying, managing and provisioning of ONTIC Network Traffic dataSet (ONTS). ONTS dataset will be provisioned with real ISP network flows captured from Interhost (SATEC subsidiary) core network during 24 months. The ONTS dataset, containing trillions of records, will be stored in a Big Data platform to be subsequently used by ONTIC batch and streaming analytics algorithms, which will be designed and implemented respectively in WP3 and WP4. Such a Big Data Platform will be implemented using ONTIC Big Data Architecture for storing network traffic datasets as its base design. As previously described, WP3 and WP4 will test their analytics mechanisms and algorithms using concrete pieces of the ONTIC dataset, and therefore, WP3 and WP4 input requirements will possibly affect to WP2 requirements. Hence, this relation will be taken into account when defining WP2 requirements.

The requirements definition process will be based on iterative and incremental development processes, where the requirements will evolve throughout the duration of WP2. In order to address in a systematic way the requirements definition process, ONTIC Big Data Architecture will adopt the Data Value Chain model proposed by FP7 Big Data Public Private Forum project (http://www.big-project.eu) in its deliverable “D2.2.1 First Draft of Technical White Papers”. Then, the following phases will be considered when defining ONTIC Big Data Architecture requirements:

1) Data acquisition

   This phase will deal with the process of gathering, filtering and cleaning data (e.g., getting rid of irrelevant data) before the data is put in a data warehouse or any other scalable storage solution on which data analysis can be carried out. Requirements in this phase will address the extraction of structured and/or unstructured data from the network, and the deployment of scalable and distributed batch and/or stream processing architectures to manage the extracted data (e.g. Hadoop, Spark and Storm platforms among others). Accordingly, WP2 tasks will address the specification of requirements described in this phase.

2) Data Analysis

   This phase is concerned with making raw data, which has been acquired, amenable to use in domain specific usage. This will entail processing data into a richer representation to be used by analytics components. Moreover, the techniques associated with Big Data Analysis will encompass those related to data mining and machine learning techniques, to information extraction and new forms of data processing including, for example, stream data processing. Therefore, Big Data Analysis requirements specification will be specifically addressed by the tasks included in WP3 and WP4.

3) Data Curation

   This phase deals with how information is managed, preserved and reused. Data quality, cleaning, augmentation and validation are some of the various aspects that this phase needs to address. Since these topics are addressed in WP2 tasks, they will be taken into account when specifying Big Data Architecture requirements.

4) Data Storage

This phase is responsible for different aspects of storage, organization and manipulation of data. In particular, Data Storage phase will address how to deal with RDBMS limitations using NOSQL or Cloud Storage options. Since ONTS dataset will reach about 1 Petabyte size, a Big Data storage solution will be required to store their contents. Therefore, Hadoop or Spark platforms jointly with NOSQL database instances should be considered as candidates for ONTS storage solution. Additionally, the concrete analytics designed in WP3 and WP4 will condition the specific ONTS structure, and so, in a first stage, packet headers can be stored in flat files ordered by the time the packets were captured. Later, WP3 and WP4 analytics could establish new storage structures (e.g. network flow summaries ordered by their starting or completion time) to be considered in the Big Data Architecture. More complex SQL relational structures are not expected to be required. WP2 tasks will address the specification of requirements described in this phase.

5) Data Usage

This phase deals with the advanced data usage for supporting smart decisions. Requirements in Data Usage phase are related with decision support, decision making and automatic steps in a domain-specific usage. ONTIC integrates this phase in WP5 Use Cases, and so, the proposed smart orchestration of the results generated by the analytics components in each Use Case will address the specific requirements to be defined in Data Usage phase.

8. Requirements strategy for WP3

WP3 goal is to design and develop a scalable analysis system for offline characterization of network traffic traces. As previously stated, WP3 requirements definition task will take into account that WP5 use cases will have to integrate specific mechanisms developed in WP3. To this aim, the requirement strategy is described below.

8.1 Input requirements

The input requirements will be defined by applying the following strategy.

A. State of the art survey to determine the most promising approaches in the project context and the type of data they have been applied to (e.g., which attributes and domains have been considered).

B. WP5 use case analysis to identify primary goals of the analysis algorithms.

8.1.1 Example

WP3 analysis algorithm data inputs are traffic flow records, not raw packet headers. In particular, each record is a bidirectional flow (i.e., the same flow identifies both client-server and server-client traffic) in the following form, with an arbitrary number of attributes.

Table 2: Input requirements for offline analytics

Source IP	Dest IP	Source Port	Dest Port	Upload	Download	Attribute_X

A more detailed example is available at http://tstat.tlc.polito.it/measure.shtml#log_tcp_complete

Other WP3 requirements are inputs from WP5 use cases, which will drive the algorithm outcomes and applications.

8.2 Content and format requirements

The content and format requirements will be identified by applying the following strategy:

A. Exploration of public datasets of network traffic traces to identify the most popular storage formats, their advantages and drawbacks.

B. WP2 dataset definition to determine the content and the format of the data to be fed to the WP3 algorithms.

Some additional information is required by WP3 algorithms and should be provided by domain experts of the project partners:

- Supervised classification needs labeled data.
- Both unsupervised clustering and association rule mining results need domain expert interpretation.

- Traffic flow feature selection is a task to be performed on input data by domain experts.

Focusing on the latter requirement, the number of features (attributes) of the flows can be very high, but typically only a subset of them is passed as input to the algorithms at any given run. This selection should be driven by domain experts.

8.2.1 Example

Input data can be stored as structured flat text files (e.g., a record per row, tab separated columns), possibly compressed to reduce storage space.

The full dataset should be split into chronologically homogeneous subsets, each stored into a different file (e.g., a file every 60 minutes of captured data, with splits at each hour of the day).

Each split of data should consist of two files, a data file and a metadata file, the latter providing a structured text description of the former: the columns (attributes, in detail), the time and probe that captured the data, and any other relevant information regarding the network data themselves, in a machine-readable format.

8.3 Technological requirements

WP3 algorithms will run on top of the Hadoop platform and its current evolutions, such as Spark and the most recent advances in the Big Data processing frameworks.

The technological requirements will be defined by the Hadoop platform evolution.

9. Requirements strategy for WP4

This section aims at defining the requirement strategy that will be conducted in the WP4 of ONTIC project. WP4 aims at defining a scalable online network traffic characterization system.

As previously stated, WP4 requirements definition task will take into account that WP5 Use Cases will have to integrate specific mechanisms developed in WP4.

One of the key functions to be designed and developed, is related to scalable online flow classification. Based on this function, the two objectives are (1) to design a traffic pattern evolution sub-system, and (2) to design a network anomalies and intrusion detection sub-system (use case #1).

Given the online and scalable nature of the system and sub-systems, the key issues to be addressed are related to:

1. The traffic capture system and the way the capture traffic stream is afterwards sent to the flow classification system. This capture system has to be fast and needs to process data to transform it in the appropriate time series whose format has to be defined during the first months of the WP;
2. The computing of this large amount of data of high dimensions by the data-mining and Machine-Learning algorithms. These algorithms have to be able to handle several traffic traces from several capture devices in real time and then in a scalable way. This implies, to define the performance parameters to be used in the rest of the project.

1) Traffic capture

Given the research aspect of this work package, it is also required to provide to the researchers some data for designing, testing, and validating the proposed algorithms. This part of the work has of course, first, to be performed off-line.

Last, task 4.4 deals with integrating and testing the full architecture with all its sub-systems.

There will then be two stages with regard to the work to be done in this work package.

Stage 1:

For the research work in the domain of (1) flow classification, (2) traffic pattern evolution, and (3) anomalies and intrusion detection, raw traffic traces are required. Such raw traffic data is essential for starting any research activity in the three quoted domains. Traffic traces covering one week on very few network links would be enough for issuing new algorithms, with significant tests and validations for entering the on-line mode.

Stage 2:

It is related to the online traffic characterization for the 3 domains quoted just before. This is this second stage that is of importance for this deliverable.

At this time, it is not possible to state what kind of traffic data will be required. We then start with the idea that we will need a copy of all raw packets flowing on one or several links of the network. The schedule of WP4 states that:

1. We need to provide a first version of the traffic data format our algorithms will work on at t_0 + 12. It will be the result of the first year of research in the 3 research tasks quoted previously (D4.1). We expect to be able to simplify the type of data our algorithms are requiring, for instance by aggregating some features, for instance. Such work could be more efficiently performed by the capture devices. Anyway, they could perform this task much faster than the machines running the machine learning and data mining algorithms.
2. But research work will continue for almost all the duration of the project. An important milestone is at t_0 + 24, and a final version of the traffic format (D4.2).
3. After t_0 + 24, we expect that the traffic trace format required by the algorithm will not be modified. We nevertheless prefer to let this possibility open, in case an evidence for changing the format appears. Nevertheless, at this end of the project, we will focus mainly on the integration of the sub-systems, and optimization tasks.

Figure 3 illustrates what has to be done to transform raw packet traces into aggregated traffic traces to be presented at the different data-mining and machine-learning algorithms that will be designed / used for flow classification, traffic pattern evolution, and anomaly detection.

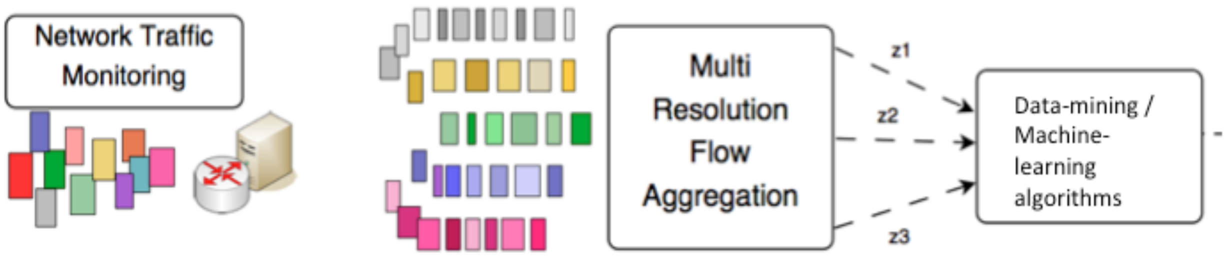

Figure 3: Traffic pre-processing for scalable online characterization

2) Computing of big data by data-mining and machine learning algorithms

Given the real-time and scalability objectives for the classification, traffic pattern study, and anomaly detection functions, the related algorithms need to be fast and efficient. It is then required to define the performance parameters to be assessed based on the constraints associated to the traffic, as well as based on the use cases requirements. During the first months of this WP, the different performance parameter of importance will be defined together with the expected performance values, i.e. fast computing on significantly huge amount of traffic data, with very high accuracy.

10. Requirements strategy for WP5

This section is intended to provide a requirements strategy for WP5, regarding the use cases definition process. This definition process is contained within the Task 5.2: Verification, Use Cases and Field Trials.

For this purpose, the strategy to be followed is described as follows, divided in several main objectives:

1. Provide an initial description of ONTIC use cases.

   This description is provided in the current deliverable *D2.1 – Requirements Strategy*, in Section 11.1. This preliminary use cases description provides a first input for WP3 and WP4, which are in charge of the design and development of the offline and online network traffic characterization systems. With those definitions, WP3 and WP4 contributors will have a start point to set the goals for the analytics algorithms.

2. Provide a basic set of requirements.

   A basic set of requirements will be provided for month 12 (January 2015) as part of the deliverable *D5.1 - Use Case requirements*, which is one of the contributions of the Task 5.2 outcomes. These requirements will evolve and will be augmented along the Work Package duration.

3. Provide progress on use cases requirements.

   A more detailed and completed set of requirements will be provided for month 24 (January 2016) as part of the deliverable *D5.2 – Progress on Use Cases*, which is one of the contributions of the Task 5.2 outcomes.

4. Provide final set or requirements

   The final set of requirements for Use Case 1 will be provided as part of the deliverable *D5.4 – Use Case #1 Network Intrusion Detection*, whose delivery is planned by month 36 (January 2017).

   The final set of requirements for Use Case 2 will be provided as part of the deliverable *D5.5 – Use Case #2 Proactive Congestion Detection and Control Systems*, whose delivery is planned by month 36 (January 2017).

   The final set of requirements for Use Case 3 will be provided as part of the deliverable *D5.6 – Use Case #3 Dynamic QoS management*, whose delivery is planned by month 36 (January 2017).

   Deliverables D5.4, D5.5 and D5.6 are part of the contributions of the Task 5.2 outcomes.

5. Use Agile methodologies.

   WP5 way of working is based on Agile principles, so a simplified version of the Agile methodology will be used in WP5 to the set of requirements gathering. The base methodology to be used will be SCRUM, and to obtain the requirements is planned to use the User Stories procedure. The agile bases to work with are briefly explained in subsection 11.2. The deliverable *D5.1 - Use Case requirements* will contain a detailed description of this methodology to obtain the requirements.

6. Interaction with related tasks and Work Packages

As the involved deliverables in the use cases definition and use cases requirements gathering are contributions of the Task 5.2 outcomes, it is important to analyze the relation of such task with the rest of tasks and work packages.

Task 5.2 takes inputs from Task 5.1, whose main goal is to integrate the offline and online network traffic characterization systems. As Task 5.1 lasts from month 20 to month 33 (September 2015 to October 2016), and the current deliverable is due to month 4 (May 2014), it is out of the scope of this section and documents the relation between Task 5.2 and Task 5.1.

On the other hand, task 5.2 relations with the other work packages comprises Work Packages WP3 and WP4. WP3 and WP4 expect the use cases requirements in order to set the goals to start analyzing the algorithms to be used.

10.1 Initial use cases description

The following subsections show a high-level description of the ONTIC use cases.

10.1.1 Use Case 1. Network intrusion detection

The result of this use case cannot be stated at this point of the project. However, we rely on previous work to draw the line of what we expect this use case to be at the end of the project.

It is well admitted now, that network anomaly detection is a critical aspect of network management for instance for QoS, security, etc. The continuous arising of new anomalies and attacks create a continuous challenge to cope with events that put the network integrity at risk. Most network anomaly detection systems proposed so far, employ a supervised strategy to accomplish the task, using either signature-based detection methods or supervised-learning techniques. Yet, both approaches present major limitations: the former fails to detect and characterize unknown anomalies (letting the network unprotected for long periods), the latter requires training and labeled traffic, which is difficult and expensive to produce. Such limitations impose a serious bottleneck to the previously presented problem.

At this stage the directions we will follow for this use case are:

- To take advantage of an unsupervised clustering approach to detect and characterize network anomalies, without relying on signatures, statistical training, or labeled traffic, which represents a significant step towards the autonomy of networks;
- To propose for accomplishing unsupervised detection some robust data-clustering techniques to avoid general clustering lacks as sensitivity to initial conditions, course of dimensionality, cluster correlation, etc.
- To use the clustering results for issuing traffic characteristics and especially the rules characterizing the anomalies, and that could be used as filtering rules in security devices, for instance.

10.1.2 Use Case 2. Proactive Congestion Detection and Control Systems

The congestion detection Use Case will be implemented in two phases, as part of an incremental process. In phase 1, a new control system will detect congestion in real time, showing analytics results in a basic user interface (UI). In that way an expert can take the proper corrective actions. In phase 2 the solution evolves, and the control system is able to apply in a self-managed way corrective actions, always in real time.

Summarizing, the implementation of use case 2 is conceived within an incremental process with two phases: phase 1 is focused on big data **analysis**, and phase 2 is focused in the **actuation** of the big data algorithms.

10.1.2.1 System model

A network operator wants to detect congestion problems in its network, as the traffic is growing more and more. In order to avoid service delivery to be compromised, the network operator needs to incorporate a new function to control and analyze the traffic of its own network. As efficiency is a must for a network operator, the new function is enhanced with an analytic subsystem. With this new integrated subsystem, the network operator will proactively detect congestion situations. When congestion is detected, the system will inform an expert to analyze the situation or take self-managed corrective actions.

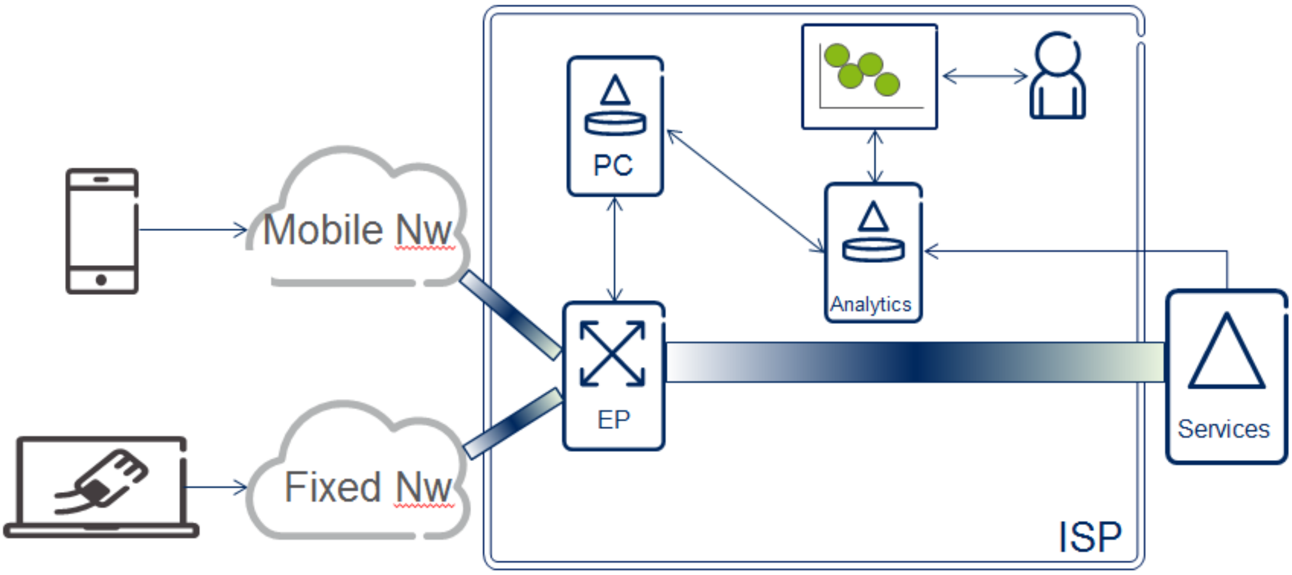

Figure 4: Architecture for Use Case 2

Figure 4 shows a high-level description of the architecture. According to this figure the network operator receives traffic from fixed and mobile networks through the Enforcement Point (EP), a node that (a) interconnects fixed and mobile networks with ISP's internal services, and (b) applies congestion control and prevention policies to the traffic that crosses it. The Overload Controller, represented by the PC node (PC comes from Policy Controller), will analyze network traffic and deploy specific traffic policies on EP node. With the help of an online analytic subsystem (represented by the Analytics node), the Overload Controller will be able to process in real time all the data, and analyze the network traffic patterns. To this end, the Analytics node will run prediction, classification and clustering algorithms. The result of the analysis will be shown in the UI (User Interface), represented by a green chart in Figure 4. Therefore, the UI will present analytics results in nearly real-time related with the congestion situation of the ISP and the characterization of the network traffic that is crossing the EP node.

Due to the complexity of deploying the congestion control system described in Figure 4 in a real production network, a simulated scenario will be used instead, and synthetic traffic (based on the ONTIC dataset) will be generated and injected in the simulated network.

10.1.3 **Use Case 3. Dynamic QoS management**

The use case related to the QoS scenario will be implemented in two phases, as part of an incremental process. Phase 1 is focused on big data analysis. Once completed, the second phase will be focus on the actuation of the big data algorithms developed in phase 1. Next subsections describe each phase.

10.1.3.1 System model

The administrator of a network would like to characterize the traffic of the applications crossing its network, being able to setup specific QoS parameters for these applications by taking into account user profiles and network preferences and resources.

Nowadays all these operations are done manually by having a prior knowledge about the concrete type of applications, source and destination IPs and related ports, etc. Currently, due to the exponential growth of network traffic, this task cannot be manually afforded in most situations. Therefore, network operators want to have a set of tools that can make recommendations to them about the best policies to be injected in the system, in order to improve the QoS of the network applications.

Figure 4 shows the architecture for use case 3. Notice that it is exactly the same one for use case 2 (see section 10.1.2 The Analytics node will process and characterize the traffic that crosses the network in nearly real-time allowing the network administrator to be aware of the different types of services and applications running on its network and their quantitative composition. To this end, the Analytics node will run prediction, classification and clustering algorithms. Based on the overall view of key network parameters (e.g. free available bandwidth and congestion level in each network link) the analytics node will also generate a proposal of a new set of QoS policy rules that optimizes the overall QoS for every service, taking into account the set-up provided by the network administrator and the users service level. Later on, these new rules can be executed by the enforcement points in a self-managing way, optimizing the QoS for the different services and users.

As in the previous use case, due to the complexity of deploying the dynamic QoS management system described in Figure 4 in a real production network, a simulated scenario will be used instead, and synthetic traffic (based on the ONTIC dataset) will be generated and injected in the simulated network.

10.2 Methodology

This section describes the methodology to be used in WP5 to coordinate the efforts of every task in WP5.

ONTIC is a Big Data -use case driven- research project that, in the end, shows the power of a new generation of online/offline distributed and scalable big data algorithms, developed along the three years of duration of the project. To have a clear picture of how to manage priorities and requirements in this WP5 to be close to the market needs, it is quite important to implement a clear and simple process.

The proposed way of working, will help managing end to end process of every task in WP5, from setting priorities on what research field should be addressed first, selecting and specifying use cases close to the market needs, and validating them. This way of working will be inspired in the Agile methodology, adapted to the specific constraints of a Research and Innovation project.

The Agile methodology is a lightweight project management framework with broad applicability for managing all types of iterative and incremental projects. Roles (e.g. product owners and actors) and artifacts (e.g. backlogs, user stories and sprints) defined in well-known Agile methodologies (e.g. SCRUM) will be used. In WP5 the **"Product Owner"** will be in charge of **identifying** and **prioritizing user stories** (requirements) and of updating and prioritize the **"Backlog"** following **the "Actors" (customer) needs**. In this WP5 the WP Leader will take the role of such **"Product Owner"**. As said, the prioritized requirements to be implemented are identified by the **"Actors"** (The market) and described in the **"User Stories"**. The priorities in the Backlog will be set-up by the Work Package leader (Product Owner) and will be implemented in the defined **"Sprints"** that will be flexible enough to adapt it to the pace of the project.

Following a more detailed description of the different roles and tools are presented.

10.2.1 Actors

Firstly, the identification of the different actors is needed in the methodology; they could be the potential customers of the outputs of the project. In this case, as it is related to the work on WP2 the potential customers are WP3, WP4 and WP5 in the end they will be the receivers of the outputs generated by it. Once actors are defined, a prioritized list of user stories is required. Actors will define their priorities and therefore will drive the research and development of the big data components.

10.2.2 User Stories

Actors will define the user stories that, later on, will be prioritized in the WP5. User story is the proposed tool used to summarize the requirements coming from the different WPs. A definition of done will be also linked to the user story to assure that WP Leader/ Task Leaders will work on it, and knows what it is expected to be shown as a result. Once the Sprint ends, WP5 Leader/ Task leaders will show to the rest of the consortium the outputs, which should be aligned with the user story definition and the definition of done.

The standardized way to define a user story is as follows:

As <user>

I want <what>

So that <why>

10.2.3 Task

When the User Stories have been prioritized, the WP5 Leader/Task Leaders selects those ones that they are going to work on. Once the User Stories are selected the WP5 Leader/ Task Leaders splits them in tasks for their internal work. In the end, the WP5 Leader/ Task Leaders will present to the rest of the consortium the outputs of the ongoing sprint.

10.2.4 Backlogs

There will be two backlogs:

- A prioritized backlog of user stories
- A new backlog (Sprint Backlog) will be created at the beginning of each sprint, in the Sprint Planning. Will consider the new needs identified in every sprint review.

The WP5 Leader/ Task Leaders are responsible to select the user stories that are going to work based on the prioritization and do it the same process in the following sprints.

10.2.5 Tools

An excel file stored in the common repository is proposed as a way to manage the prioritized user stories backlog and the related definition of done, tasks, etc. This simple tool has the advantage of providing an easy way of accessing the information without administrative overloads.