Virtual ITS European Congress, 9-10 November 2020

Paper number ITS-TP19681

Hybrid Big Data Platform for Intelligent Road Transportation Services

Nikolaos Peppes[1], Theodoros Alexakis[1], Ioannis Loumiotis[1], Konstantinos Demestichas[1], Evgenia Adamopoulou[1*]

1. Institute of Communication and Computer Systems, Athens, Greece, contact e-mail: eadam@cn.ntua.gr

Abstract

The ever-increasing demand for transportation as well as the massive accumulation of population in urban centers arise the need for infrastructure and system development in order to manage efficiently the highways and vehicles. Moreover, given the rapid growth and the evolution of the information and technology area, it is now possible to develop Intelligent Transportation Systems (ITS) that go beyond traditional approaches. Furthermore, nowadays, highways and vehicles have a vast variety of sensors installed that collect data such as speed, acceleration, direction, and so on. The vast volume and variety of collected data emerges the need for Big Data techniques and analytics to be employed in state-of-the-art ITS. The MANTIS project aims to provide an overall solution consisting of heterogeneous applications as well as a driver assistance system for improved transportation. In this paper, we present an innovative Hybrid Big Data Platform which is implemented in the context of this project.

Keywords: Big Data, Intelligent Transportation Systems (ITS), Transportation

Introduction

The ever-evolving needs for data retrieval, storage and processing have made it necessary to study and implement Big Data management systems. In addition, the rapid growth of the telecommunications technology has created the conditions for infrastructure development which can achieve real-time interconnection and data transmission from and to heterogeneous applications and systems. Big Data techniques, including data mining, machine learning, artificial intelligence, data fusion, social networks and so on have been used by many organizations and companies with great results [1], [2]. Among the many applications and systems using Big Data technologies are the Intelligent Transportation Systems (ITS). ITS incorporate state-of-the-art technologies and tools including advanced sensor technologies, data acquisition and transmission technologies as well as smart control and actuators technologies. MANTIS (Multiservice cApable iNtelligent Transportation Systems) is an ITS research project which integrates a comprehensive framework of heterogeneous applications and demonstrates driver assistance systems so to achieve improved road transport.

In the context of the MANTIS framework a Hybrid Big Data Platform (HBDP) is implemented which exploits the advantages of Big Data and cloud technologies. MANTIS HBDP utilizes the advantages of Big Data technologies as well as cloud computing and benefits the whole project in the following key aspects: i) Vast amounts of heterogeneous and complex data that are created by vehicles and highway can now be handled efficiently by HBDP. ii) HBDP can improve the whole MANTIS operation efficiency by analyzing the current and historical massive vehicle and road data and last but not least iii) Through Big Data analytics the safety level of the road and the drivers can be improved. For this purpose, data collected from the vehicles are transmitted to the HBDP of MANTIS where, after suitable analysis, dangerous road conditions can be predicted and spotted in order to alarm the drivers and warn operational management centers for accidents, closed lanes, etc. Also, MANTIS analytics make possible to classify anonymously drivers' behavior using statistical methods and tools applied to collected driving data.

The MANTIS Hybrid Big Data Platform creates a common place for all stakeholders of the framework to send and extract data. Furthermore, it enables vehicle-to-vehicle (V2V) communication as well as

vehicle-to-operation centers (V2X) communication both in real time and historically. Through data exchange and analysis, it is possible to reach useful conclusions that will improve road safety and efficiency as well as the drivers' behavior.

The remainder of the paper is organized as follows: First, the paper presents an overview of Big Data technologies and related work, and then provides a presentation of the MANTIS Hybrid Big Data Platform. Subsequently, it presents the first experimental results of the platform's operation together with a relevant discussion. Lastly, a summary is provided and useful conclusions are drawn.

Big Data Overview and Related Work

Big Data Concept and Definition

The well-known concept of the so-called Big Data Analytics is based on the vast volume of information and data in general that can be processed and analyzed by lifting the limitations of conventional data technologies and computers. One of the very first implicit references to the Big Data idea was made by Laney in 2001 who stated that businesses should accommodate new frameworks and technologies in order to handle the rapid increase of data Volume, Velocity and Variety [3]. Through the following years the 3 V's model was expanded with some extra V-features such as Veracity, Validity, Volatility, Value, etc. which all of them are applied in the data of ITS domain.

Andrea De Mauro et al. made an elaborative study about Big Data definition and key research topics and proposed the following elaborative formal definition which concludes all the aspects of Big Data definitions [4]:

"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value"

Related Work and MANTIS Framework

Due to their envisioned capabilities, enabling the development of a plethora of new services upon them, Big Data Analytics are gaining more and more ground also in ITS applications since transportation systems produce huge amounts of heterogeneous data at an enormous speed which must be processed in real-time. It is undoubtable that the exploitation of Big Data analytics in ITS is highly promising, leading to a great volume on research on this topic. Despite, the large number of studies about Big Data in ITS only a few aim to proposing a holistic solution in this specific domain. Most of the studies focus on the prediction of the traffic conditions using different methods such as ANN [5], deep learning [6] approaches and so on. Also, other approaches and applications in the ITS domain focus on a road condition warning system platform, like the one proposed by Teke and Duran [7] or a dangerous driving events modelling platform by Alvarez-Coello et al. [8].

A very interesting study for a more holistic solution was conducted by Liu et al. who described in [9] an integrated data exchanging platform for a highway in China. Additionally to [9], Mian et al. proposed a multi-engine platform to support various types for traffic data analysis [10]. Moreover, in [2] a more generic idea is described for a common architecture considering a Big Data Analytics platform for an Intelligent Transportation System. Zhu et al. stated that Big Data Analytics for ITS can be divided into three layers, the data gathering layer, the data processing and analytics layer and the application layer [2]. In this light, the MANTIS platform aims to offer an overall solution in ITS domain by expanding the current studies and their results. The Hybrid Big Data Platform implements B2B and C2B interfaces for data exchange as well as services and tools for data storage, processing and analysis. The described Platform collects vast amount of data from heterogeneous sources and through its hybrid integration can provide both real-time and historical processing and analysis in order to extract valuable information both for the drivers and the highway operation centers. Following a similar approach, the MANTIS platform has three basic systems; the storage, the analyzer and the monitoring units which consist of state-of-the-art tools and components which are presented in the next section of this study.

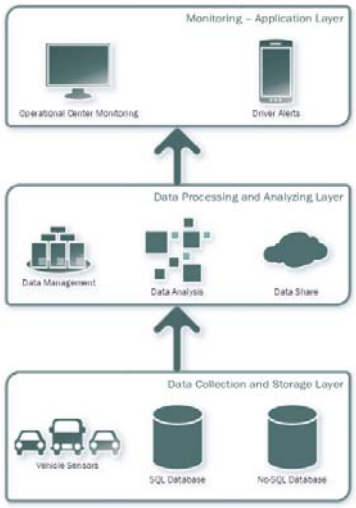

Figure 1. 3 Layer architecture

MANTIS Hybrid Big Data Platform Architecture

The Hybrid Big Data Platform in MANTIS was designed in order to be: (1) Reliable: The platform is designed to operate without any external intervention, so it is of outmost importance that it runs reliably and handle unexpected situations without downtime. (2) User-friendly: HBDP is deployed for both the operational and administration centers as well as the drivers. The platform is practically a black box for the users and therefore what is required is just "one-click" initialization. (3) Expandable: the MANTIS HBDP is designed in such way so to be expandable and suitable for other application beyond ITS domain.

The Hybrid Big Data Platform consists of multiple components which interact between each other and provide an integrated solution in the limits of MANTIS project. The current platform was completely developed in Python language which is considered as one of the most widespread and famous computer languages while it is the most compatible for the development of Big Data and Cloud applications nowadays. Apart from the aforementioned, the platform is actually implemented into an open source environment alongside with the usage of Flask, a Pythonic microweb framework (microframework) which renders the process of web application design and deployment much easier, by using the appropriate modules and libraries. Due to the platform's necessities a client-server architecture is used. Flask is used to initialize a server which consists the backbone and will cover many requirements of the platform, for instance the data transfer between client(s) and server and the implementation of other important back-end operations.

Besides of the framework (flask) that is responsible for the server-side development, the platform also comprises other components for the implementation and the fulfillment of additional demands and requirements:

The **Hybrid Data Management System** is composed of a Relational and a Non-Relational (NoSQL) Database and is responsible for Big Data management, manipulation and analysis operations and/or purposes. The aforementioned operations are applied in data that are collected from heterogeneous sources. For real time purposes and data manipulation is chosen MySQL a commonly known relational database, while for historical data processing is used MongoDB, the most popular NoSQL database.

A **Communication interface with external applications and sensors**. The current system consists of interfaces which operate as REST APIs concerning the data or message streaming between the main platform's components. By using REST architecture is feasible to execute basic communication operations with various types of systems. REST API interfaces receive and send data and/or messages in

a two-way communication from and to the Hybrid's Data Management System databases as well as to vehicles and external sources.

The **Big Data Operational Framework** is in charge of integrating, processing and manipulating the structures that outcome from big data sources. The chosen Big Data Operational Framework is Hadoop. Through the Hadoop Ecosystem is implemented an interface which is connected to the NoSQL Database and is responsible for any operation that is related to the Big Data structures.

An **Intelligent Data Analysis Framework API** is responsible for the data process and the extraction of the analysis results with regard to the users driving behavior as well as the routes assessment. Spark is a framework suitable for implementing complex computational and analysis processes.

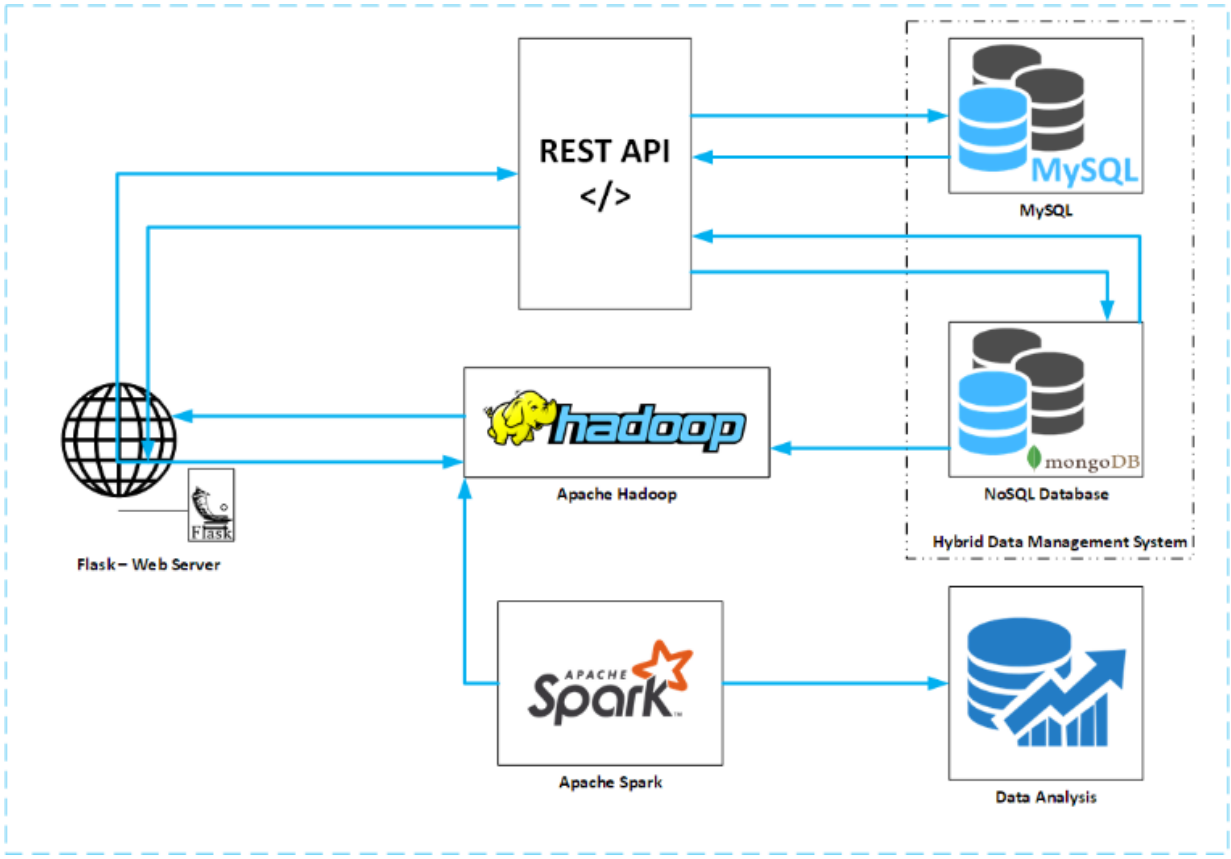

Figure 2. MANTIS Hybrid Big Data Platform architecture

Interfaces of Hybrid Big Data Platform

REST applications (APIs) are actually developed and used for the hybrid big data platform with regard to establish communication with vehicles and third-party systems. The term 'REST' actually indicates the way that the data are represented through our platform. Rest architecture is one of the most widespread architectures for web applications and web server development by implementing all the commonly known HTTP methods. The API is deployed into the server and includes the main characteristics of the REST architecture: uniform interface, client-server interaction, stateliness, cacheable resources, layered system and code on demand. Through REST API any approved user or third-party side can regularly update not only the relational but also the NoSQL database into the platform's big data environment.

Vehicles data are transferred through REST API by the appropriate data format usage. In a continuous session requests send from the vehicles and reach towards the platform's server including a particular

URL and an endpoint together with the aforementioned data in a specific format. In this way the API is feasible to apply all the possible CRUD (Create Read Update Delete) methods by using the appropriate query in order to achieve a specific operation each time.

By the same token, a REST API interface is used to establish communication between third party systems and platform in order to receive data. Each application can retrieve data either by the relational database for real time analysis or by the NoSQL database for batch data analysis.

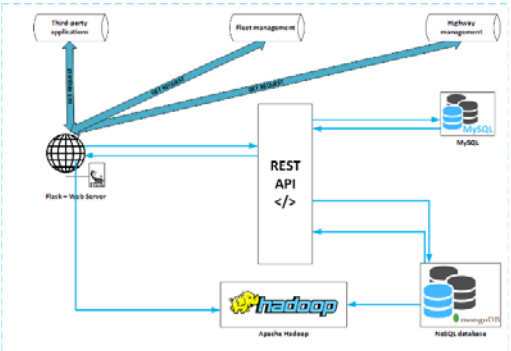

Figure 3. MANTIS B2B communication

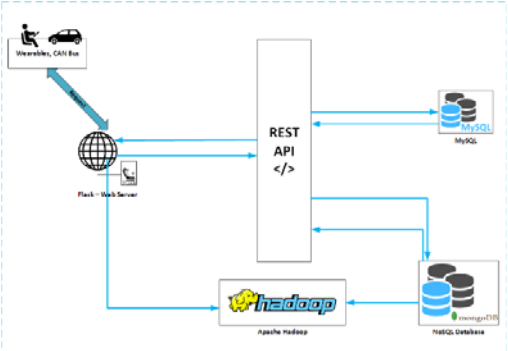

Figure 4. MANTIS C2B communication

Data Analysis and Results of the Hybrid Big Data Platform Using Machine Learning

The current section presents the results obtained from the implementation of the cluster analysis (clustering) on a dataset that was created from the collection of raw data retrieved from eight properly equipped vehicles, which were used for field trials on a test period of 18 days with different sampling intervals based on the drivers' and vehicles' shifts. The aforementioned vehicles are equipped with a variety of sensors. The described Hybrid Big Data Platform is the destination endpoint for storing and processing purposes. The stored data consists of a dataset of approximately fifty thousand diverse records in terms of speed, acceleration, location and vehicle identifier, which will be used as the input during the execution of the K-Means algorithm (or Clustering), in order to classify the data points into a number of consequent clusters.

Clustering is one of the most widely used machine learning methods for unsupervised learning with the purpose of determining the number of similar object or behavior groups that are more related to each other than to objects of other groups from a dataset of unlabeled input data. The K-Means Algorithm is commonly indicated as a prototype-based clustering algorithm, where each cluster is depicted by a prototype, which can be either the average of similar points (centroids) or the most frequently occurring points (medoid). In our case, the representation of each cluster is denoted by the Euclidean distances between the centroids and the existing vehicles' data points. The squared Euclidean distance is a widespread method approach for calculating the similarity as the opposite of the distance between the data objects. Based on the aforesaid, the K-Means algorithm could be described as an optimization approach for minimizing the inside cluster Sum of Squared Errors (SSE), known as cluster inertia. The equation for the SSE is shown below where $\mu^{(j)}$ is the centroid for cluster j:

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{n} w^{(i,j)} \left\| x^{(i)} - \mu^{(j)} \right\|_2^2$$

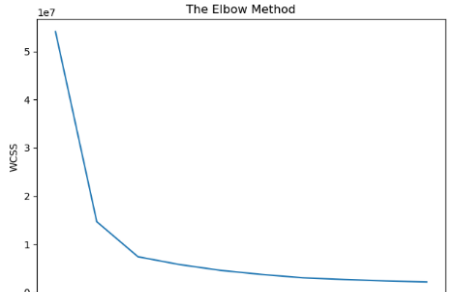

Figure 5. Elbow rule chart

One of the fundamentals issues during the implementation of the K-Means algorithm, that needs to be emphasized, is the determination and the definition of the optimal number of clusters (K). There is a variety of methods applied in order to identify this optimal number. The elbow method is such a valid

and usable graphical approach to calculate the optimal value of K for a given task. Particularly, the elbow method computes the K-Means algorithm for a defined range of values of K, estimates the Within-Cluster Sum of Square (WCSS) or distortion, and visualizes the plot of WCSS related to the number of K. The value of K where the WCSS begins to increase in a more rapid way or otherwise the location of the elbow (bend) is indicated as the optimal number of K clusters. During our implementation of the elbow method estimation process, we defined a range value of K=1-10, which resulted to an optimal number of K=3 clusters for our analysis dataset, as shown in Figure 5.

Regarding our vehicle dataset analysis process in conjunction with the implementation of the aforementioned described K-Means Algorithm, we investigated new possible correlations between the eight different vehicles in terms of the speed and speed difference during discrete time moments, which were extracted from our vehicles' dataset. Under these circumstances, we set the number of cluster K=3, as it was previously substantiated, the maximum number of iterations equal to three hundred whilst we defined the algorithm to run fifteen times with different random centroids in order to select the one with the lowest SSE. In addition to the above configurations, a tolerance threshold equal to 1e-04 (0.0001) was selected for the purpose of declaring possible convergences regarding changes in the WCSS.

Consequently, the outcome of the implementation of K-Means Algorithm to our vehicles' dataset input is illustrated in figures 6a and 6b.

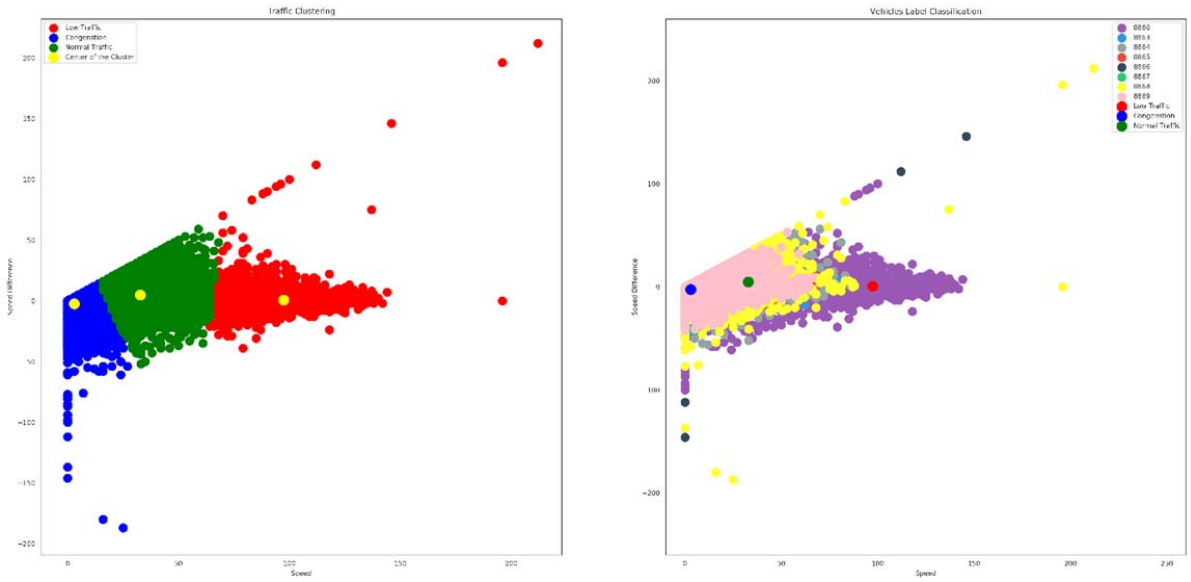

Figure 6. a) Traffic clustering and b) vehicles label classification charts

On the left-hand side (Figure 6a) the clustering resulted in three different recognizable data group clusters. The centroids (or the average centers) of each of the three clusters are represented with yellow marked labels. The declared clusters are considered as three different possible traffic conditions: blue color depicts a possible condition of congestion as a combination of low speed and negative speed differentiation values range; which indicates that, due to the specific traffic situation, decelerations often occur; green color indicates the combination of medium speed values range and a mixed sample of speed differentiations that suggests normal traffic conditions; and eventually low traffic condition is depicted with red color as a result of high speed values and speed differentiations close to zero value implicating more steady and easy driving conditions. Figure 6b presents another viewpoint of the described analysis with different label classification of the eight distinct vehicles ID's in terms of the current clustering results.

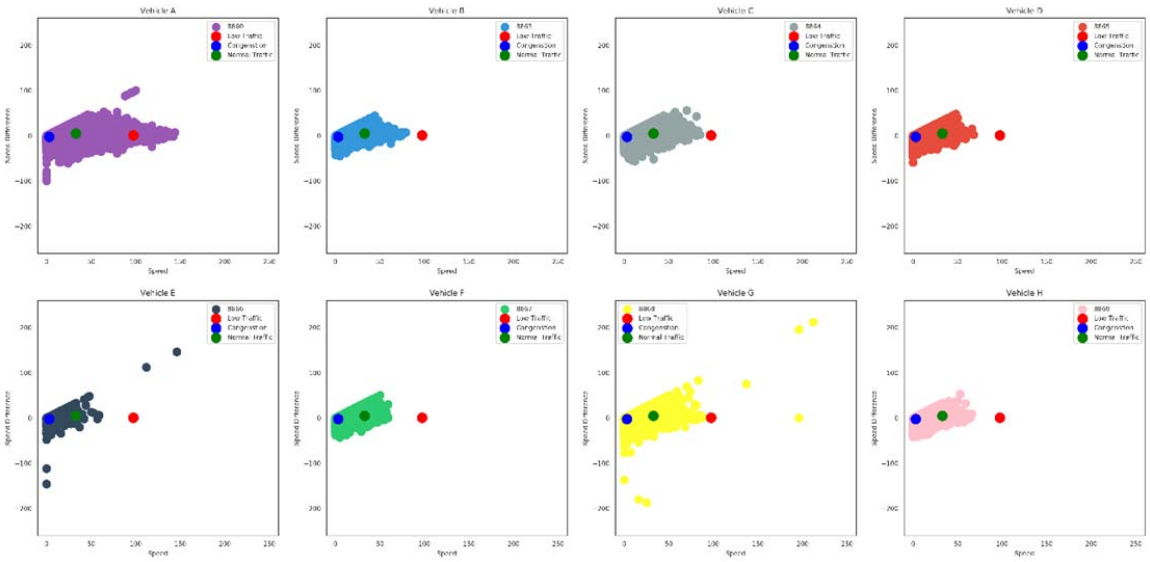

Figure 7. Individual vehicle clustering indication

Figure 7 depicts the samples taken for each and every of eight vehicles alongside with the centroids of the clusters. It is obvious that the majority of the drivers usually move between the blue and the green centroid which implies that in most cases there is a certain degree of traffic. More specifically, vehicle A has the most samples on the third (red) cluster compared to other vehicles. This fact declares that the driver of Vehicle A drives at high speed in low traffic conditions. In addition, drivers of Vehicle E and Vehicle G provide some extreme negative or positive measurements compared to the majority of the records of this specific dataset. These measurements show either a fault on the measuring equipment or some deviant and dangerous driving behavior of the drivers which must be studied further in order to accurately identify the real cause.

The results presented above are the very first steps of Big Data Analytics from the HBDP which is described in the previous section. It is obvious that the platform can handle efficiently the storage, process and analysis of vast amounts of data. The results are indicative of the platform's capabilities at this early stage and pave the way for more detailed analysis to produce Key Performance Indicators (KPIs). The Intelligent Data Analysis Framework API aims to integrate both real-time and non-real-time data processing in order to extract results about driving behavior and route evaluation. Furthermore, through the analysis of massive data streams in real time using machine learning algorithms the platform will create alerts and messages for the drivers in case of possible dangerous or emergency situations.

Conclusion

The aim of the current work is to demonstrate the architecture of an innovative Hybrid Big Data Platform in the context of MANTIS framework. The ongoing development of this platform as well as the MANTIS framework in general, aim to expand the boundaries and lift the limitations of current Intelligent Transportation Systems. Furthermore, the very first set of results that came from the Intelligent Data Analysis Framework API using Big Data analytics and Machine Learning algorithms were presented. These first results are a milestone for the platform integration and deployment and can lead to further analysis so as to produce useful indicators concerning the drivers', vehicles' and road's safety. The results and the data are anonymous and encrypted in order to avoid security and privacy breaches.

A future prospect of the MANTIS project is to exploit the capabilities of Big Data analytics firstly to alert and warn drivers for emergency situations and then to predict and prevent them. Moreover, future projects

could include Distributed Ledger Technologies such as Blockchain and smart contracts so to provide smarter, trustful and more personalized applications in the context of the ITS domain.

Acknowledgment
This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-04610 "Multiservice cApable iNtelligent TransportatIon Systems" - MANTIS). This paper reflects only the authors' views, and the Operational Programme is not liable for any use that may be made of the information contained therein.

References

1. Bello Orgaz, G., Jung, J., Camacho, D. (2015). Social Big Data: Recent achievements and new challenges, *Information Fusion*, vol. 28
2. Zhu, L., Yu, F., Wang, Y., Ning, B., Tang, T. (2018). Big Data Analytics in Intelligent Transportation Systems: A Survey, *IEEE Trancactions on Intelligent Transportation Systems*, vol. 20, pp. 1-16.
3. Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety, META Group Research Note, no. 6.
4. De Mauro, A., Greco, M., Grimaldi, M. (2014). What is Big Data? A Consensual Definition and a Review of Key Research Topics. In Proceedings *4th International Conference on Integrated Information*, Madrid.
5. Loumiotis, I., Demestichas, K., Adamopoulou, E., Kosmides, P., Asthenopoulos, V., Sykas, E. (2018). Road Traffic Prediction Using Artificial Neural Networks. 2018 South-Eastern European Design Automation, *Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)*, Kastoria, Greece pp. 1-5.
6. Yisheng, L., Duan, Y., Kang, W., Li, Z., Wang, F., Y. (2014). Traffic Flow Prediction with Big Data: A Deep Learning Approach, *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 865-873.
7. Teke M., Duran, F. (2019). The design and implementation of road condition warning system for drivers, *Measurement and Control*.
8. Alvarez-Coello, D., Klotz, B., Wilms, D., Fejji, S., Gomez, J., M., Troncy, R. (2019). Modeling dangerous driving events based on in-vehicle data using Random Forest and Recurrent Neural Network, *2019 IEEE Intelligent Vehicles Symposium* (IV), pp. 165-170.
9. Liu, W., Zhou, Y., Long, K., Luo, H., Xu, P. (2019). Design of Data Interchange Platform for Digital Highway. In Proceedings *International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha, China.
10. Mian, R., Ghanbari, H., Zareian, S., Shtern, M., Litoiu, M. (2014). A Data Platform for the Highway Traffic Data, 2014 *IEEE 8th International Symposium on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA)*, Victoria, BC, pp. 47-52.