Track & Know

PROJECT

NEWSLETTER

Track & Know | Issue 2 | Date: April 2019

WELCOME TO THE SECOND TRACK & KNOW NEWSLETTER!

In this newsletter, you can find news on:

- Updates about the big data platform of the project
- The first results based on our preliminary data analysis
- Publications, so far in 2019

ABOUT THE NEWSLETTER

This newsletter informs you about the results and activities of the EU H2020 research project Track& Know. The aim is to keep all relevant actors interested in managing big data, more specifically on the type of big data we focus on in the project and the tools/methods we develop to handle, analyse and visualize these datasets. T&K focuses on resolving key business cases for 3 test pilots, namely transport/mobility, insurance and health care. Business cases which will be explored in these pilots are as follows but not limited to: minimizing patients travel, carpooling and electric mobility potential, driver behaviour profiling etc.

Project coordinator
Dr. Ibad Kureshi
Senior Research Scientist – Inlecom Systems
Square de Meeus 38/40
1000 Brussels
Belgium
Ibad.kureshi@inlecomsystems.com

THE T&K BIG DATA PLATFORM

Updates from INTRASOFT International SA

One of the major objectives of the Track & Know project is to **integrate online data streams, heterogeneous, contextual and archival data on one big data platform**. This enables big data experts and stakeholders to advance their operational, processing and decision making activities.

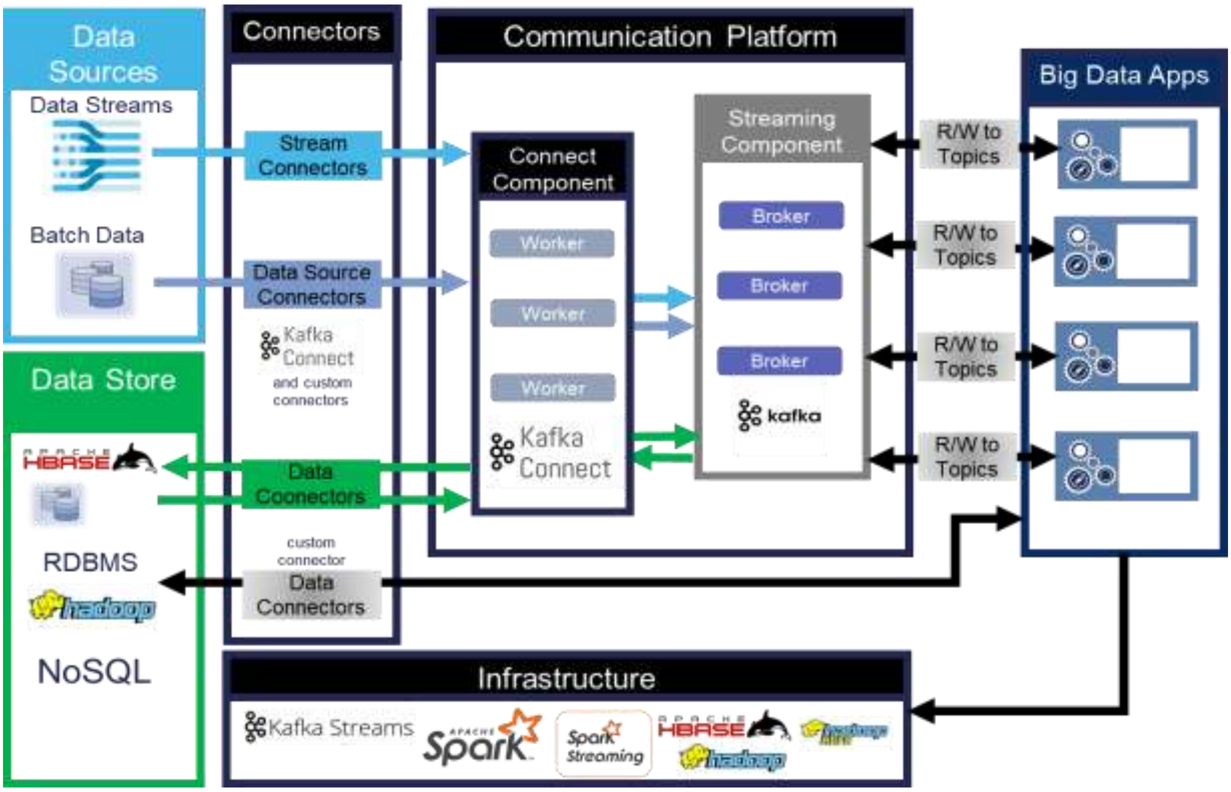
To this end, **the big mobility data integrator (BMDI) has been developed**. BMDI is a fully featured industrial grade solution:

- that is able to scale out and accommodate various big data from different domains, interoperating with all modern data storage technologies as well as other persistence approaches.
- that can support all important big data languages including Python, Java, R and Scala as well as other traditional programming approaches.

The big data platform consists of:

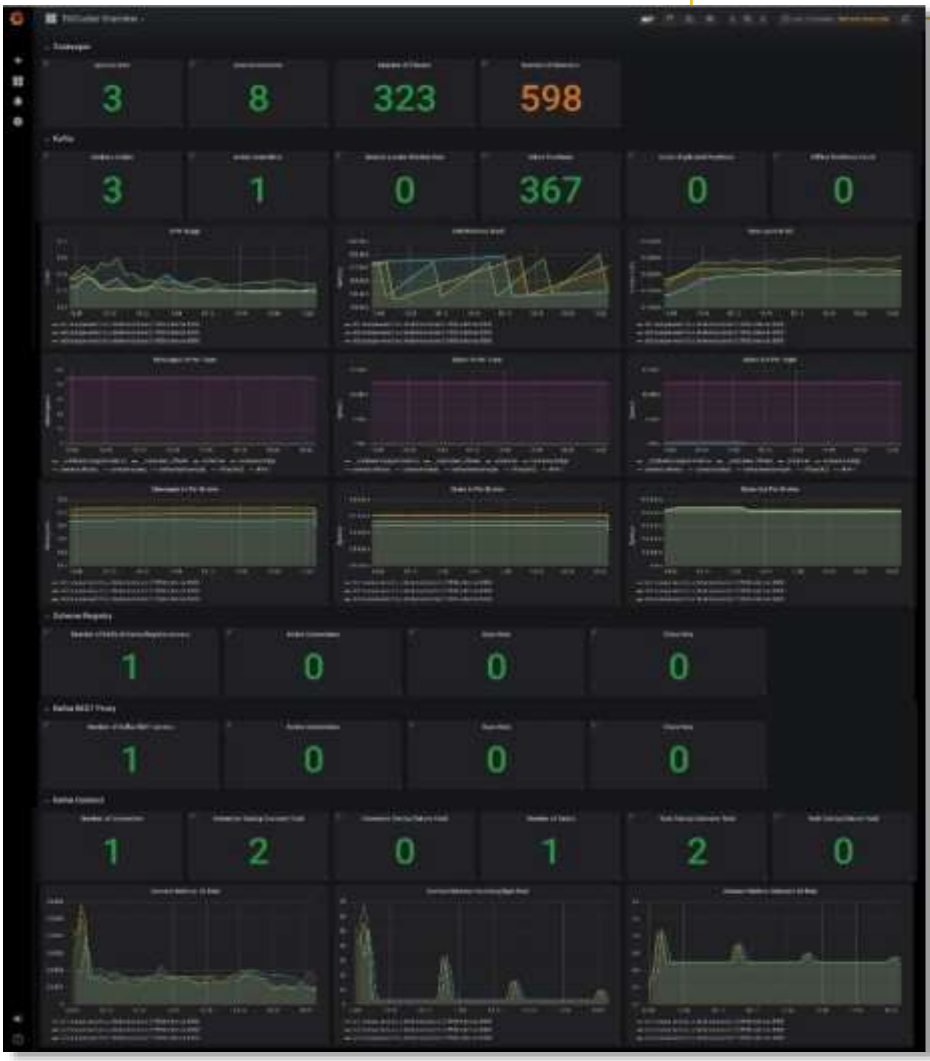
1. **Data sources and data store components** both in a structured or unstructured format that can be made available and potentially be connected to the big data platform. **The big data platform can efficiently interoperate with all the modern data storage technologies of a big data ecosystem such as** RDBMS, NoSQL, HDFS Hadoop, Apache HBASE, etc. as well as other persistence approaches such as Mongo, MySQL, JDBC, etc.

2. **Connectors together with the Communication platform**, that connect external data sources and make them available to the big data platform. External data sources are connected by means of “Stream Connectors” and “Data Source Connectors”. The Communication Platform (i.e. Apache Kafka Cluster) allows to publish and subscribe to streams of records (Topics). These streams of records are stored and consumers (Big Data Apps) can process them as they occur.



3. **Underlying infrastructure**. The underlying infrastructure spans multiple virtual machines (VMs) and provides all the necessary technologies and components that enable the storage and analysis of the data involved. It further allows the usage of algorithms developed using any technology by providing a distributed computing environment. Some of them include among others Apache Spark, Hadoop, Kafka Streams, Spark Streaming, etc.

4. **Big data apps.** Big data applications can be implemented in all important big data languages including Python, Java, R and Scala. Traditional programming approaches (C/C++, Ruby, Perl, PHP) can also be supported and efficiently interoperate with the big data platform. This is also referring to a variety of big data toolboxes which are being developed within the project such as the **big data processing (BDP) toolbox, the big data analytics (BDA) toolbox, the complex event recognition (CER) toolbox and the visual analytics (VA) toolbox**.



The platform also provides the **graphical dashboard** for administration and monitoring. It visualises selected metrics of interest. The available dashboards allow not only a very good overview of the cluster performance and health status but also provide significant information when performance tuning is necessary.

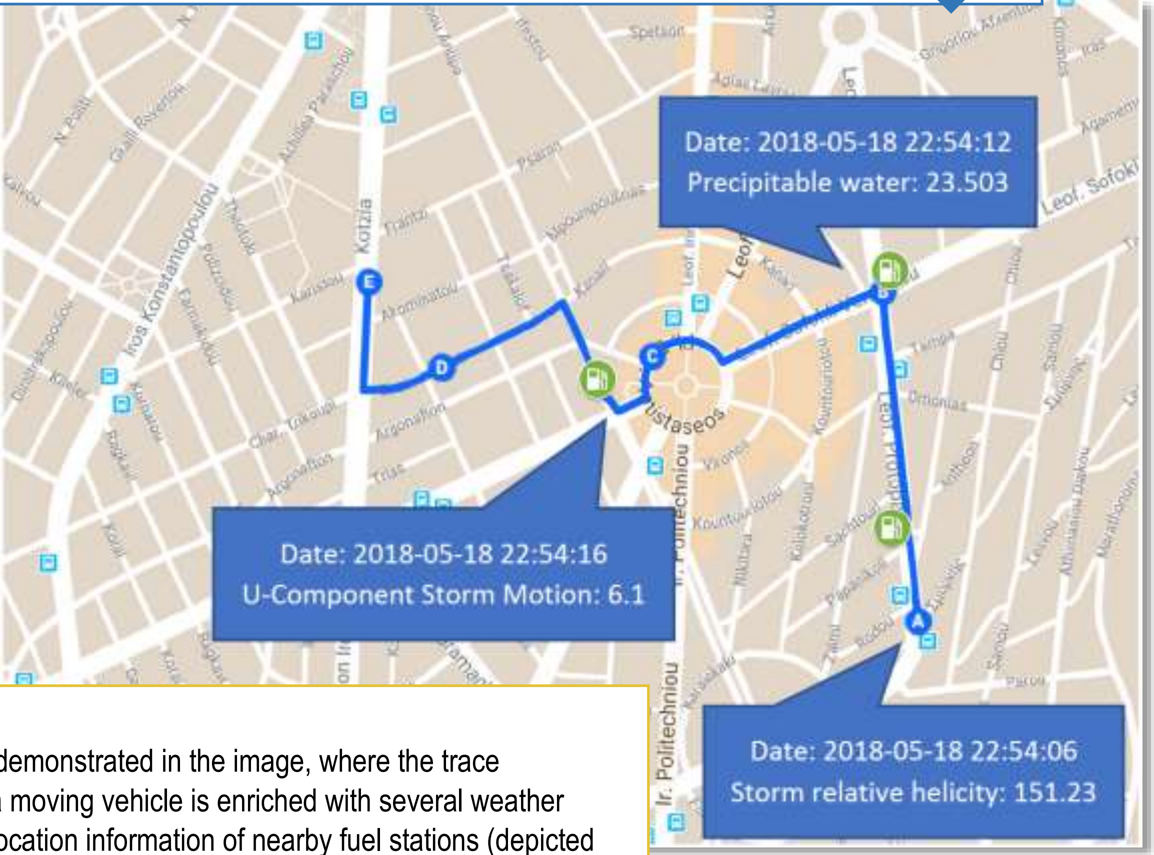
Currently, several components of the big data platform are in a testing phase with online data streams and other data sources for three pilots of the Track & Know project.

THE FIRST RESULTS

Integrating contextual information into streaming positional data

Updates from University of Piraeus Research Centre (UPRC) on the VFI dataset

UPRC has built an online tool that takes as input streaming data and enriches them with selected weathers attributes indicating road conditions (e.g. wind, rain or ice) and nearby fuel stations. This enrichment of positional data is useful to improve mobility data analysis tasks in several use case scenarios in pilots related to mobility and insurance. It will help improve the accidents risks estimation, hot spot analysis and with certain other assumption also improve the electric mobility analysis. This tool will be integrated in the big data processing toolbox.



Date: 2018-05-18 22:54:12
Precipitable water: 23.503

Date: 2018-05-18 22:54:16
U-Component Storm Motion: 6.1

Date: 2018-05-18 22:54:06
Storm relative helicity: 151.23

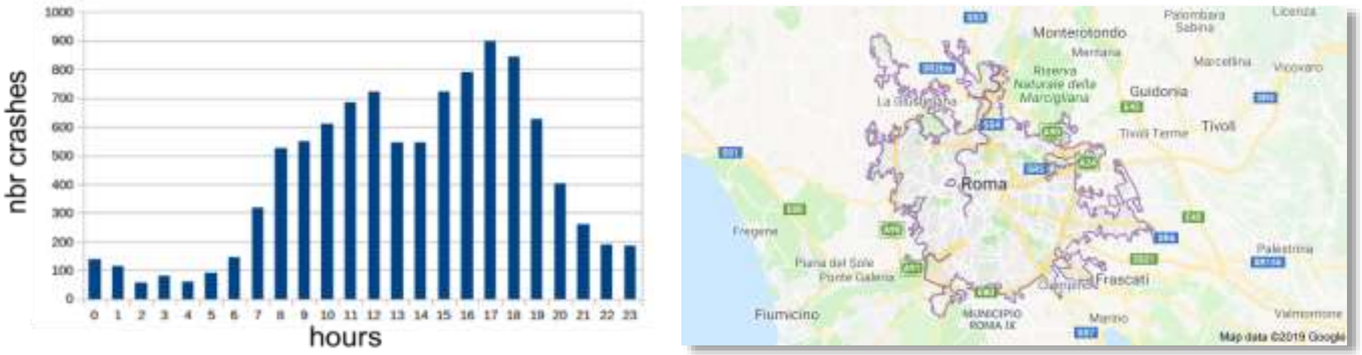
An example is demonstrated in the image, where the trace information of a moving vehicle is enriched with several weather attributes and location information of nearby fuel stations (depicted in a green colour). The dataset made available by Vodafone Innovus (VFI) is currently used to implement this tool. The data is of spatiotemporal nature, but also includes engine data from sensors, as well as data indicating the driver's behaviour (e.g., harsh cornering, harsh breaking, etc.). Weather data is obtained from the National Oceanic and Atmospheric Administration (NOAA), US, based on their global forecast system. This open source data consist more than one hundred weather attributes, which are related to temperature, rain, snow, ice, wind and humidity etc. Location information of fuel stations is obtained from OpenstreetMap that also offers a wide variety of Points of Interest (POIs) information on locations (e.g. fuel stations).

Crash prediction

Updates from Consiglio Nazionale delle Ricerche (CNR) based on the Sistematica dataset

The risk of an individual accident is predicted through **machine learning models** based on various mobility indicators such as features of mobility distribution, driving behavior events (accelerations, etc.), car model. This is relevant for the insurance pilot and helps estimate risk scores. This method will be integrated within the Big Data Analytics (BDA) toolbox.

Area of experiment (Rome) and hourly crash distribution



The three types of features considered

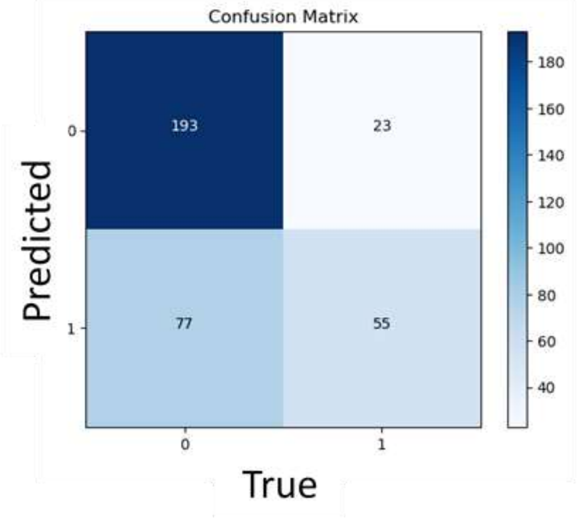
trips

events

model

Recall Crash: 0.71

Precision Crash: 0.41



Preliminary results show reasonable precision and recall with various models, the best being the Random Forests technique, and the most discriminant features are those based on events, together with the frequency of traveling.

A more thorough increase in performance is expected as soon as we acquire environmental data through a fully functioning trip enriching pipeline. This pipeline contains more contextual information about the trip (such as weather and other information by using other tool boxes) than is available now.

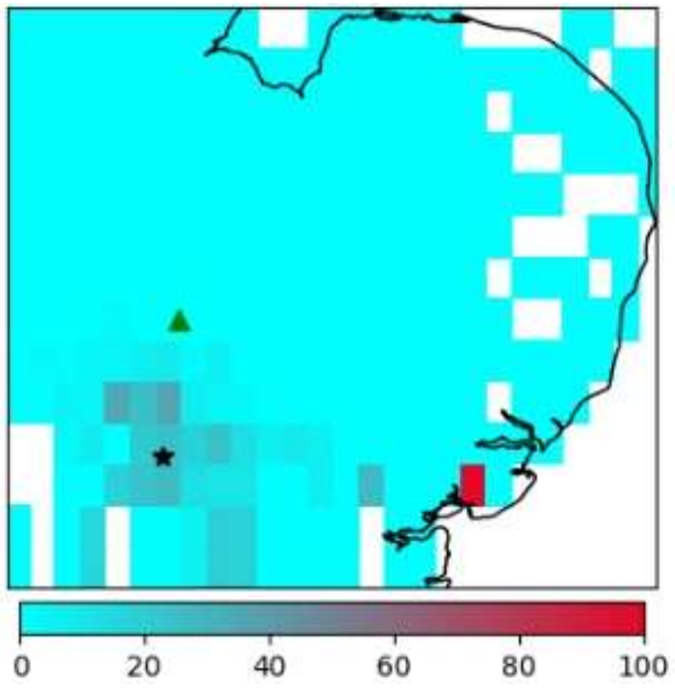
Patients mobility:
Relative dependence of patients and outreach clinics

Updates from Inlecom Systems (ILS) on the Royal Papworth dataset

Royal Papworth Hospital (RPH) operates a Respiratory Support and Sleep Centre covering patients within the East Anglia region of England. East Anglia covers a large rural and urban catchment area comprising of Norfolk, Suffolk, Cambridgeshire and Essex. Travel distances from the edge of the catchment area to RPH (pictured as the green triangle) can be in excess of 100 miles in one direction with driving times exceeding 2 hours in the off-peak. Due to the rural nature of the population, travel times using public transport can be in excess of 4 hours.

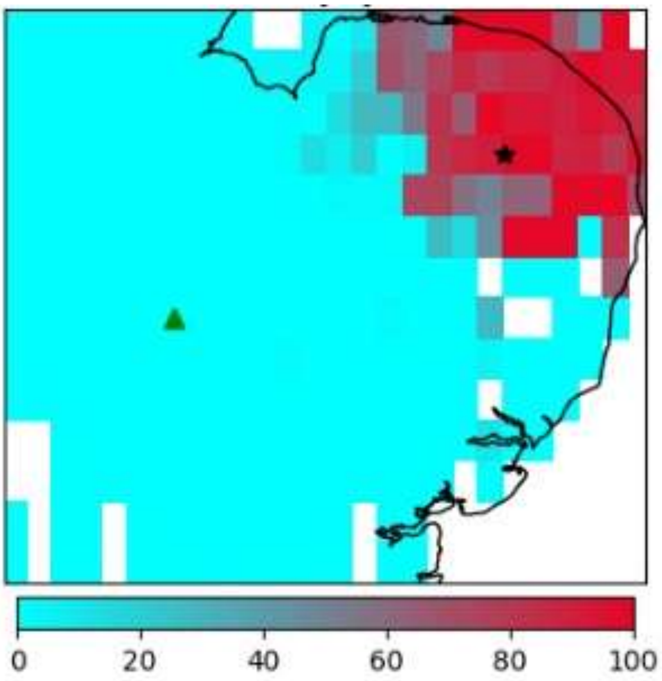
To ensure that all patients within the catchment have adequate access to the service, RPH has established seven outreach clinics that operate once-a-month. An analysis (attendance, no-show rates, distances travelled) of patients using these outreach clinics has revealed many interesting patterns. One such result of the early analysis has shown the relative importance of certain outreach clinics. In the figure two outreach clinics, one located in Stevenage (RSSCSTEA) and the other in Norwich (CSSNOR), are analysed with respect to patients choosing to attend the outreach clinic instead of travelling to RPH. These two outreach clinics on average see the same number of patients per year, approx. 250-300, and have similar no-show rates, approx. 17%.

(a)



Patient Dependance RSSCSTEA vs RPH

(b)



Patient Dependance CSSNOR vs RPH

Splitting the area into tiles of 0.1 latitude by 0.1 longitude, in the map (a) we can see that, over the last 5 years, less than 50% of the patients living in tiles around Stevenage have used the outreach clinics. Patients are more likely to visit RPH instead. This is in stark contrast to Norwich where in many tiles all the patients of the service utilised the outreach clinic and did not travel to RPH. This variance between the two sites can be attributed to distance between the tile and RPH, and the capacity of the outreach clinic. While Norwich is further away from RPH and it can be assumed that it is the distance that factors into patient's decision making. But Stevenage also has a significantly higher population density and it is perhaps the capacity of the one-day a month clinic that forces patients to opt to travel to RPH.

Two other interesting results from the above images are the unshaded tiles. While a blue cell means patients from that tile went exclusively to RPH instead of the specific outreach clinic being analysed, and the red tiles are those that exclusively chose the outreach clinic, the white or unshaded cells are regions where patients went to neither location. From image (b) we can see several spaces in the bottom right quadrant where patients never went to RPH and did not use the two outreach clinics shown here. When analysing all 7 clinics most of these unshaded cells are accounted for and by and large there are no gaps in RPH's service. However, the 5 tiles that make up an inverted 'P' in the bottom left quadrant are never shaded. This implies that there is a competing service that is recruiting patients for within RPH's catchment area.

PUBLICATIONS 2019

- Nikitopoulos P., Sfyris G. A., Vlachou, A., Doulkeridis C., Telelis O. (2019) Parallel and Distributed Processing of Reverse Top-k Queries, *In Proceedings of the 35th IEEE International Conference on Data Engineering* (ICDE 2019).
  Download here
- Andrienko N., Andrienko G., Garcia J. M. C., Scarlatti D. (2019) Analysis of Flight Variability: a Systematic Approach, *IEEE Transactions on Visualization and Computer Graphics*, vol. 25(1), pp.54-64.
  Download here
- Liua S., Andrienko G., Wu Y., Cao N., Jiang L., Shi C., Wang Y. S., Hong S. (2019) Steering Data Quality with Visual Analytics: the Complexity Challenge, *Visual Informatics*, vol 2(4), pp.191-197.
  Download here
- Koutroumanis N., Santipantakis G., Glenis A., Doulkeridis C.,Vouros G. (2019) Integration of Mobility Data with Weather Information*, In proceedings of the International Conference on Extending Database Technology (EDBT) and International Conference on Data Theory (ICDT) Workshops 2019*, Lisbon, Portugal.
  Download here

Find & Follow us
Website: https:// trackandknowproject.eu
Twitter: @TrackandKnow
LinkedIn: Track & Know
Researchgate

The project has received funding from the European Union's
Horizon 2020 Research and Innovation Programme under Grant Agreement No 780754.