REPLICATE	Project no. 691735 REPLICATE PROJECT Renaissance of Places with Innovative Citizenship And Technology	This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

REPLICATE PROJECT

Renaissance of PLaces with Innovative Citizenship And Technology

Project no. 691735

H2020-SCC-2015 Smart Cities and Communities

Innovation Action (IA)

D3.9 Use of Big Data for mobility services

Due date of deliverable: 31/01/2019

Actual submission date: 14/01/2019

Start date of project: 01/02/2016 Duration: 60 months

Organisation name of lead contractor for this deliverable: Euskaltel

Status (*Draft/Proposal/Accepted/Submitted*): *Submitted*

Project co-funded by the European Commission within the 7th Framework Programme		
Dissemination Level		
PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Editor/Lead beneficiary:	Euskaltel
Internally reviewed by:	Telecom

Index of contents

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

1. EXECUTIVE SUMMARY

This document has the purpose to describe how the use of Big Data can serve to produce services of interest for cities and telecommunication companies.

The urban mobility patterns are a critical task for transportation planning and management in the cities. Traditional methods, such as household or street surveys, although they provide detailed information, are slow and costly processes for the acquisition and analysis of information, preventing the possibility of generating transport demand models with the desired frequency and quality. Different sources have been explored in the last decade, providing more and more complementary advantages: Floating Car Data (FCD), allows to obtain information on travel times, data of public transport cards, which is multimodal information in the metropolitan area, applications mobile. The work reflected in this document incorporates the operational local scale applications information from mobile telephony as a source, accessing to a large amount of geolocated data over time with high representativeness. In addition, the traces associated with a private WiFi service are also integrated to improve the accuracy and success of the algorithms.

This document refers to the technical approach, including architecture, processes and algorithms, the context and the case of the pilot of San Sebastian.

2. REPLICATE

The main objective of REPLICATE project is the development and validation in three lighthouse cities (**San Sebastián** – Spain, **Florence** – Italy and **Bristol** – UK) of a comprehensive and sustainable City Business Model to enhance the transition process to a smart city in the areas of the energy efficiency, sustainable mobility and ICT/Infrastructure. This will accelerate the deployment of innovative technologies, organizational and economic solutions to significantly increase resource and energy efficiency improve the sustainability of urban transport and drastically reduce greenhouse gas emissions in urban areas.

REPLICATE project aims to increase the quality of life for citizens across Europe by demonstrating the impact of innovative technologies used to co-create smart city services with citizens, and prove the optimal process for replicating successes within cities and across cities.

The Business Models that are being tested through large scale demonstrators at the three cities are approached with an integrated planning through a co-productive vision, involving citizens and cities' stakeholders, providing integrated viable solutions to existing challenges in urban areas and to procure sustainable services. Sustainability of the solutions is fostered in three areas: economic and environmental and finally, fostering transparency in the public management.

In addition, the Model features the replicability of the solutions and their scale up in the entire city and in follower cities, particularly in three follower cities (**Essen** – Germany, **Laussane** – Switzerland and **Nilüfer**-Turkey) that are involved in the project and therefore, have access to know-how and results achieved on the project so they can apply the developed model. At the moment, there are 2 observer cities, Guanzhou (China) and Bogota (Colombia).

3. INTRODUCTION

3.1 Relation to Other Project Documents

This document describes the work carried out around the Aggregate characterization of urban mobility based on operational information from mobile network companies. The steps and work have been performed as stated in the DoA. The work is compliant with the general picture of the San Sebastian ICT Architecture, being the O/D matrices an additional data source for the centralization, processing and exploitation of mobility data, in order planners, operators and transport authorities to have a real view of the current state of mobility both at urban level and in its surrounding territory.

3.2 Reference documents

This document is based in the following projects level documents:

Ref.	Title	Description
REPLICATE Grant Agreement signed 240713.pdf	Grant Agreement	Grant Agreement no. 691735
DoA REPLICATE (691735)	REPLICATE Annex 1 – DoA to the GA	Description of the Action
REPLICATE Consortium agreement signed December 2015 (7th December version)	Consortium Agreement	REPLICATE project – Consortium Agreement
REPLICATE Project Management Plan	D1.1 Project Management Plan (v.1) (29/04/2016)	REPLICATE Project Management Plan

3.3 Abbrevations list

ANS1	Abstract Syntax Notation One
API	Application Program Interface
CA	Consortium Agreement
CDR	Call detail record
CGNAT	Carrier-grade NAT
DNS	Domain Name System

DoA	Annex I–Description of the Action
DOCSIS	Data Over Cable Service Interface Specification
DPI	Deep Packet Inspection
EC	European Commission
EKT	Euskaltel
GA	Grant Agreement
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switching Center
GSM	Global System for Mobile communications
H2020	Horizon 2020
HSPA	High-Speed Packet Access
HSS	Home Subscriber Server
ICT	Information and Communications Technology
LTE	Long Term Evolution
MMe	MME Communication
MVNO	Mobile Virtual Network Operator
O/D	Origin/Destination (matrix)
PC	Project Coordinator
PL	Pilot Leader
PMP	Project Management Plan
SGSN	Serving GPRS Support Node
SIM	Subscriber identity module
TC	Technical Coordinator
UMTS	Universal Mobile Telecommunications System
WiFi	Technology for radio wireless local area networking
WP	Work Package
WPL	Work Package Leader

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

4. DELIVERABLE DESCRIPTION

The deliverable D3.9 Use of Big Data for Mobility Services describes the work carried out in the context of a private-public partnership that can show how the use of Big Data can serve to produce services of interest for cities and telecommunication companies. Based on the data managed by the biggest operator in the Basque Country, EUSKALTEL, advanced data analytics methods can provide very exact aggregated information about mobility of people in the city as mobility heat points, origin-destiny matrix, etc.

This deliverable describes the adopted technical approach, including details on architecture, processes and algorithms as well as an enumeration of available interfaces.

The document is divided in the next structure:

- *Section 5. San Sebastian mobility use case, describes the use case proposed for San Sebastian pilot, identifying the context, scenario and factors to be considered.*

The next three chapters, provide details of the technical approach.

- *Section 6. Mobility Analytics Algorithms,* describes the different processes, data structures and algorithms defined to extract aggregated information, O/D matrices, from a priori data sources (CDRs and WiFi Kalean traces).
- *Section 7. Communications networks for Smart Cities* describes the network infrastructures available in a city to provide connectivity to citizens.
- *Section 8.* Data lake describes the Big data infrastructure built to store and transform the data coming from the network into people movement information.

Following sections include main impacts and conclusions of the work carried out:

- *Section 9. Innovations, impacts and scalability,* identifies all the impacts obtained, including social, environmental, economic and innovation.
- *Section 10. Conclusions.*

Finally, an annex has been included in the document *ANNEX I – Application Programming Interfaces*. It contains updated information about the APIs for the access to the information for 3rd parties (in this case, San Sebastian Municipality).

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

5. SAN SEBASTIAN MOBILITY USE CASE

There are different situations and factors that affect the mobility in the city of San Sebastián that should be taken in consideration.

During the weekend, the demand for mobility is not so pigeonholed and this is where the weather can influence more appreciably especially in terms of internal displacement. In the city of Donostia it is necessary to take into account the visits coming from France by proximity and that it is manifested mainly on the weekends with reasons for the different events that can be celebrated in the city and that are mentioned below.

It should be noted that one of the aspects that most influences mobility in both the engine and the cyclist is the weather. With the climate alternatives presented in Donostia with changes in short periods of time, motorized mobility can undergo changes from one day to the next simply by passing one rainy day to another without rain. The presence of rain also translates into cyclist mobility that falls appreciably when there is precipitation. For the same reason it is needed to distinguish the summer season from the rest of the year. Another aspect to consider in a general way is the calendar. It is necessary to differentiate within the weekly period, the working days of the weekend. During the working days the demand for mobility is greater and it differs during the day. The uses are closely linked to the schedules of entry and exit to the work centers and schools.

With regard to events that have impact in the mobility of the city, three main types can be established: sports, cultural and specific events. Within sporting events, it is necessary to differentiate if the event itself involves an occupation of the public way (e.g. popular races) in which case it is necessary to take into account the affection that supposes for the urban network and for public transport, in addition to the demand for mobility that generates access to the meeting point of the event for both participants and spectators. The other type of sporting event to be considered is that which takes place in a delimited area and in which one only has to foresee the demand for mobility to access and exit the site. The same differentiation must be foreseen for cultural events, although in this case, public road occupation situations are not as notable and are more linked to closed venues. In any case, events such as the San Sebastian Day, the Santo Tomas Fair or the Carnival, among others, cannot be avoided, in which the condition of mobility of the city is obvious. Finally, there are the specific events that can be generators of mobility, such as those linked to commercial activities such as the seasons of both winter and summer sales, or other external events that close by are also attractors of mobility in the city (period of cider houses, San Fermines, ...).

For the management of this demand of the city, different instruments are available that can alter the flows of the modes of transport and it is necessary to highlight: parking (underground and on the surface), public space management (pedestrian spaces, coexistence zones, areas, ...) and access control to specific areas.

Finally, it is important to mention the need to differentiate external from internal mobility, since the way of resolving the displacements coming from outside the city is logically very different from the way internal displacements are resolved since being Donostia a medium-sized city, a large part of its internal displacements is at pedestrian distances. As an example, the following indicative graph of distances linked to time to make them on foot.

Figure 1: Distances and time to make on foot

The most common methods for gathering information on mobility behavior were interviews, questionnaires and surveys, all of them facing a number of limitations relating to its accuracy, representativeness and reliability. Replicate proposes to obtain an aggregated characterization of urban mobility based on operational information from mobile network companies. Aligned with the pilot use case, the following decisions have been adopted:

- Donostialdea, the region around the city pilot location, is selected as the study area, collecting relevant geolocated data over time.
- Flexibility on the geometries of the sectors in which the study area is divided, according the different decision-making needs.
- Configurable period to perform the aggregation (set for Replicate to 30 min).
- Query capabilities to obtain origin/destination matrices, with all the data added, applying any filter of days or hours. These services provide enough temporal accuracy to cross calculated mobility patterns with additional information as weather or event calendar to understand the relevance of such factors.

This document describes the work that took place to obtain such mobility aggregated information, from algorithms to the underlying architecture to carry out the different calculation and processes.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

6. MOBILITY ANALYTICS ALGORITHMS

In this section, the analytics algorithms to compute aggregated Origin-Destination (O/D) Matrixes is described. These matrixes are an aggregated representation of the mobility patterns within the city of San Sebastian and basically store the number of displacements that have occurred within different sectors of the city. The matrixes are computed every 30 minutes which allows representing the mobility taking in to account the time of the day.

6.1 Data

Different types of data are considered to generate the O/D matrixes. Specifically, the following information is considered:

- The Call Detail Records (CDRs) which stores details of the calls as the type of a call and its duration. In the case of Replicate the user id, the serving base station id and the hour of the day at which the connection is performed are used. These data constitute the largest part of the data used.
- Euskaltel "WIFI Kalean" data. WIFI Kalean is a service that allows Euskaltel clients to share their private WIFI spots. The clients who use this service share the usage of the routers with the rest of the users in such a way that a user can use a router placed in some other user home in a transparent way. In this case we use the MAC of the WIFI, the user id and the hour of the connection.
- Localization of the base stations and WIFI routers. These data allow locating the users to be able to estimate their trips.

All these data are store in HDFS filesystem using and HBASE database in such a way that the information can be process in parallel which is key in order to be able to analyze the high amount of data needed to process.

6.2 Processes

The processing of the data is performed using the Map-Reduce paradigm which follows the diagram shown in the figure below. In this Figure a big pool of data, labelled as Big Data stores all the records to be processed. In the case of the Replicate project this is realized using an HBASE HDFS distributed data base.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

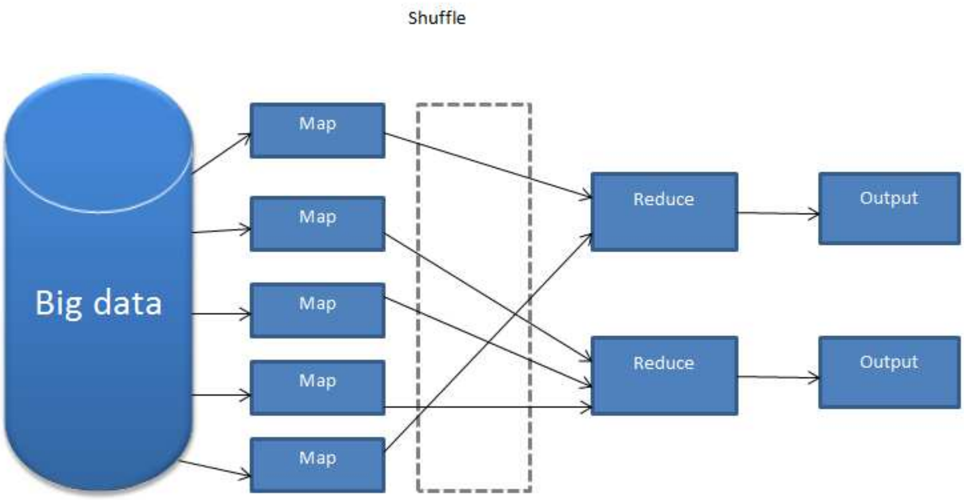

Figure 2: Structure of the Map-Reduce Procedures

In order to deal with these data, there exists many Map procedures which are in charge of reading each of the entries and passing them to the proper Reduce procedure. One of the many possible parallelization key in this case, and probably the most adequate one, corresponds to divide the entries in sets corresponding to the same mobile device.

This way, each of the Reduce procedures obtains the data for the same mobile and therefore the problem simplifies inside this procedure to a single mobile problem. It is in this procedure that algorithms described in the Figure 3 are implemented.

Although it is not shown in the previous figure, the overall architecture also considers other procedures which have the same internal structure of the Reduce procedures, but they run locally.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

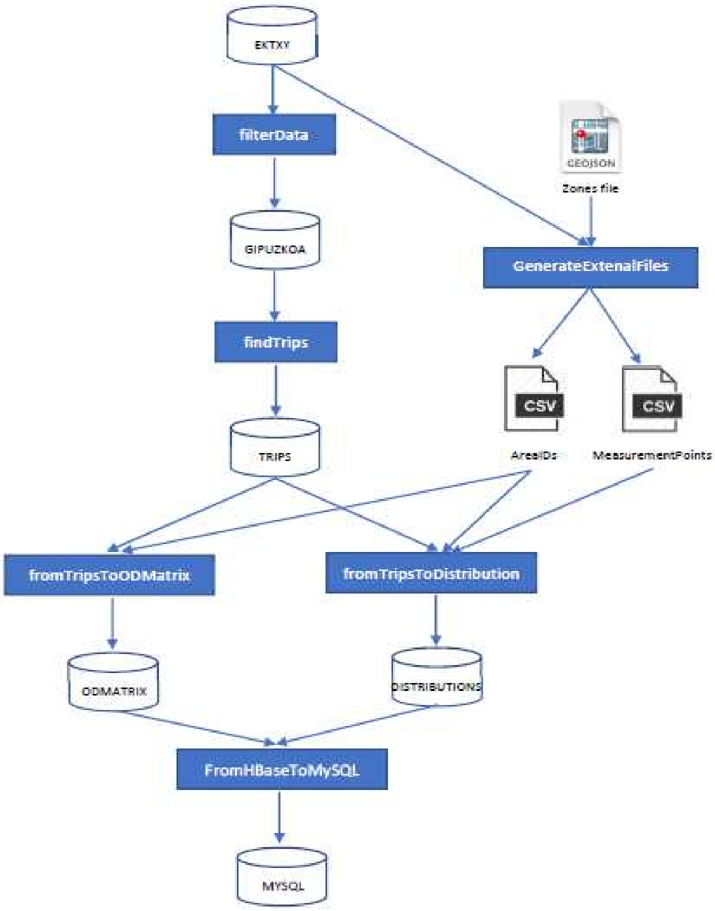

Figure 3: Summary of Map-Reduce Processes performed

6.2.1 Computation of Antennae Tessellation

A user that is connected to a base station is located in the zone of influence of its antenna. The specific zone of influence is difficult to determine because the exact extension covered by a base station depends on many different factors. Orography, buildings in the area, the actual power of the antenna and even the weather have a crucial impact on the exact delimitation of the zone of influence. It is customary to consider, as an approximation, the Voronoi diagram also known as the Thiessen polygons. This is a tessellation of the map in polygons in such a way that all the points within a polygon are closest to a specific antenna. It has been considered that a user that is connecting to a particular antenna is located within the Voronoi polygon associated to that antenna.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

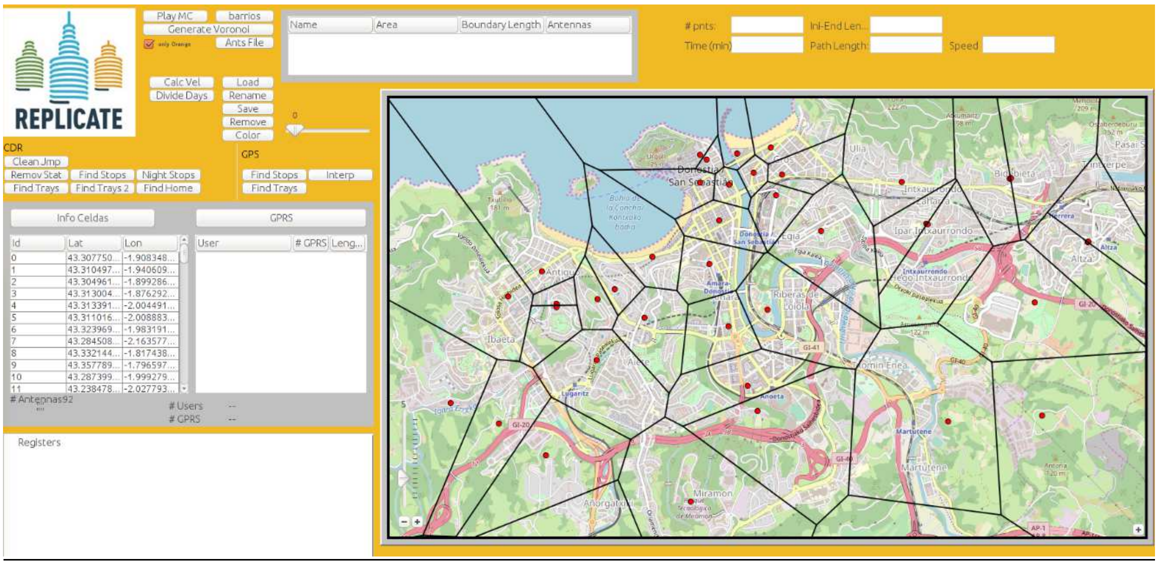

Figure 4: Tool developed to compute the Voronoi Areas around the city of Donostia.

There are many ways of computing what are the polygons that delimit the Voronoi tessellation, in the case of the Replicate project, the result shown at Figure 4: Tool developed to compute the Voronoi Areas around the city of Donostia., an implementation of the Fortune Algorithm [1] has been used. In the case of the Fortune algorithm, the computation is performed by sweeping a line through the plane and considering the set of parabolas that conform the points equidistant to the antennae, wherever the parabolas cross other parabolas are the vertices of the result tessellation polygons, see *Figure 5*. The advantage of using this algorithm is that the computation can be performed in *O* (N log N) time, for N antennae.

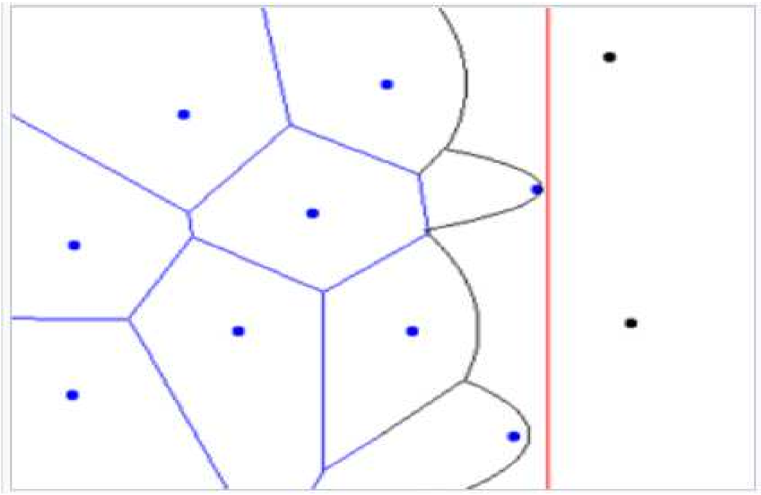

Figure 5: Example of the sweeping line, the parabolas and the resulting polygons behind (Image obtained from[2].

A similar structure but with a much finer resolution could be developed in order to locate the users that are connected to the WIFI Kalean. Now the nodes to consider are the locations of the routers of the users associated to the service. It is important in that case to notice that this infrastructure was not directly developed for providing this service but for providing WIFI within the users' homes. This implies that there are even more factors that influence the validity of the approximation. In this case, factors as the height (floor number) where the router is located, the furniture around the router, the construction material and even if the device is turn off or not should be considered.

For the purposes of the Replicate project and considering that the maximum radius of influence for WIFI is around 100 meters, it was considered that the position of the user connected to a router is the position of the router itself. Furthermore, to combine these measurements with the CDRs, It will further assume that a user is located at the closest antenna to the location of the router.

In conclusion the tessellation obtained considering the base stations constitute the initial and finer resolution that it is consider in Replicate project for the computation of the O/D matrixes.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

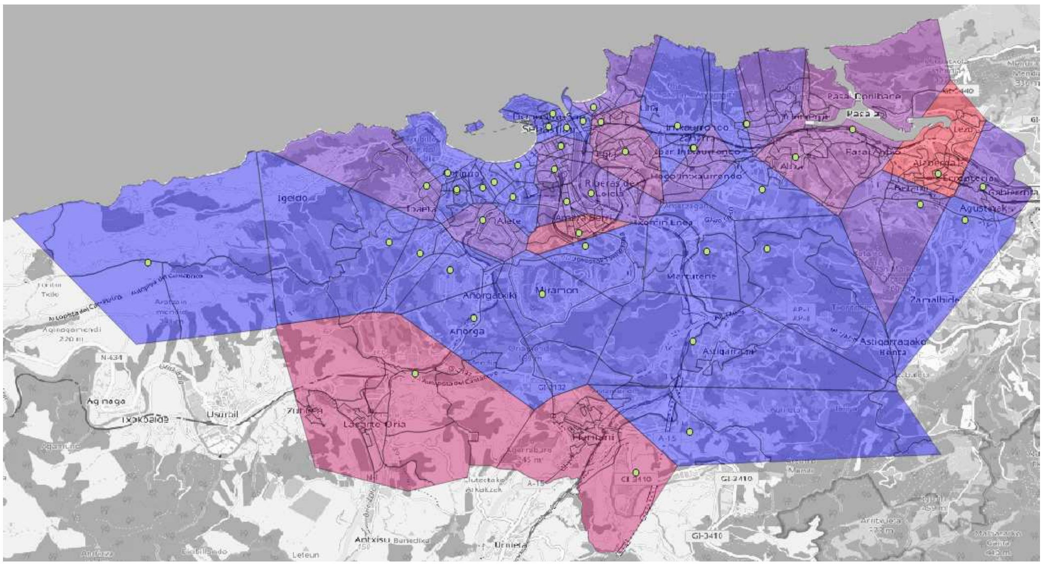

Figure 6: Diagram of the resulting tessellation around the city of Donostia (the colors are not relevant).

The computation of the Voronoi tessellation is relevant for locating the WIFI Kalean measurements to its closest base station and visualization purposes. This specific procedure is performed off-line and outside of the Map-Reduce engine.

6.2.2 Cleaning the data

The real data obtained from the CDRs tend to have spurious jumps, not representing real transitions from different zones. This effect occurs even when a user, actual a cell phone, is static at the same exact position. There exist multiple reasons why these jumps take place, among the reasons changes of atmospheric conditions, changes of power, multipath or singular shielding are of special importance.

After carefully analyzing the obtained data it was considered that these jumps correspond to the following characteristics:

- A trip composed by only two measurements each of the measurements in different neighbor zones in which the origin and destination do not agree.
- At least one of the transitions occurs in an abrupt manner in which the inferred speed of the user is anomaly high, in many occasion supersonics.

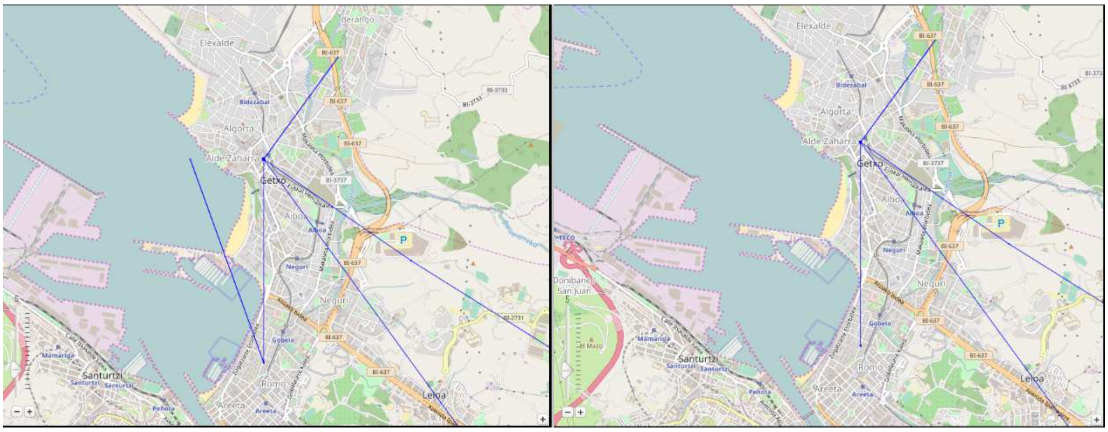

Figure 7: Example of the cleaning of the data. In the left side map a set of measurements is shown, there is an artificial jump that is corrected on the set on the right map

6.2.2.1 Jump elimination

These characteristics allow the jumps to be filtered such that the resulting O/D matrix is not affected by this noise. Several algorithms have been developed and assessed to detect and eliminate the jumps, making use of different trajectory parameters.

6.2.2.2 Elimination of "same location" data

This process consists in the elimination of measurements the produce the same exact position. The elimination of these data does not affect the evaluation of the O/D matrix in any way while reducing the overall data considered lightening the computational cost of the further processes.

6.2.3 Computing Individual Trips

As a previous step to computing the O/D matrix each of the individual trips needs to be computed. A trip is considered as a set of connections with base stations or WIFI Kalean routers of the same mobile device which are temporally ordered, a displacement is produced, and consecutive measurements are not distant apart in time.

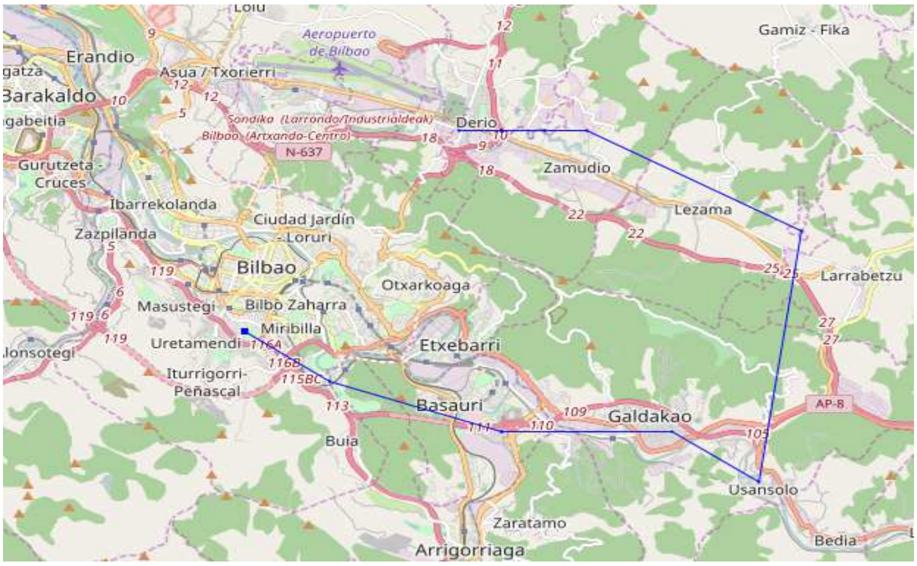

Figure 8: Example of a trip identified by the algorithm

To obtain the individual trips is was observed that it is more accurate to compute first the stops, periods of time while a mobile is in the same location and to compute the trips as the measurements between consecutive stops latter.

The result of the Individual Trip identification consists in a series of timestamp and location data which is stored again in the HBase data base and it is labelled with its corresponding mobile Id.

6.2.3.1 Stop Generation

The generation of stops follows the following procedure:

- Computation of cluster of points with zero velocity Each of the measurement points within the trajectory determines the velocity. This velocity is a one-sided finite different approximation which only considers the current points and its immediately previous one. Once this speed is computed it is considered that a value less than 0.2 km/h corresponds to mobiles that are stopped.
- (Optional) Expansion of initial clusters: The adjacent points, in time, to the zero velocity clusters are added to them. This addition is performed while the distance to the corresponding cluster is not too far away (< 1 km) and more importantly these points do not correspond with a straight-line displacement. To measure this feature, we introduce a parameter called Walk Randomness which measures how random is the path of the points being added.
- Joint Cluster of Points that are two close both temporal and spatially.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

- Erasing of Small clusters: Specifically, clusters that contain localized measurements during less than 4 minutes.

Once the stops are computed the individual trips are computed but considering the measures between the previously computed stops.

6.2.4 Aggregating the Data

The process and aggregation of data is realized via the Map-Reduce paradigm one more time. In this case the Map procedures read the data information for each of the trips realized by every user and key the data in such a way that trips with the same Origin and the same Destination are sent to the same Reduce procedure. The objective of the Reduce procedure is to count the number of different individual trips with the same origin and the same destination.

It is important to note that this aggregation is performed in time windows, all the trips which begin in the same 30 minutes are aggregated together to contribute to the same matrix element.

In addition to aggregate the trips in the O/D Matrix divided by time windows, different distributions have been collected, as for example in terms of:

- Length of trips
- Day of the week.

6.2.5 Changing Tessellation

It is important to note that the tessellation used to compute the initial O/D matrix, the tessellation shown in *Figure 6*, does not necessarily agree with any useful tessellation for the city of hall and the people responsible to take decisions with the obtained O/D matrix. Therefore, a way to transform the computed O/D matrix from the initial tessellation is needed.

In the upper half of *Figure 9* the original tessellation is shown, within this tessellation any two zones denoted by i, j are considered and the O/D matrix element T_{ij} storing the number of trips performed with their origin starting in zone i and their destination in zone j.

On the bottom half of the same figure we consider a different tessellation. The goal now is to obtain the elements of the O/D matrix $\tilde{T}_{ab}$ for the new tessellation where a, b are any two zones of the new tessellation considered.

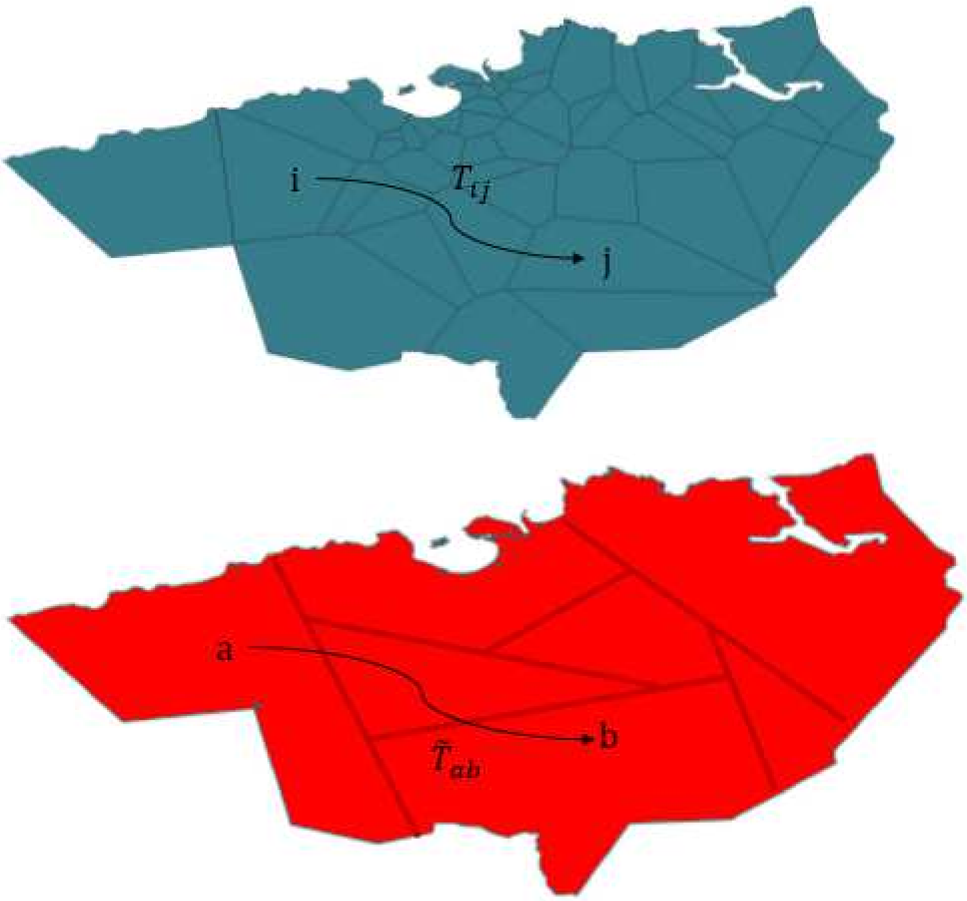

Figure 9: On the top of the figure the original Tessellation resulting from the Voronoi polygons obtained from the locations of the base station; On the bottom another tessellation within the area of interest where the O/D matrix is required to be computed.

Assuming that density of trips is homogenous along the areas of both tessellations it is straight forward to prove that the new matrix elements can be computed by the following equation:

$$\tilde{T}_{ab} = \sum_j \sum_i T_{ij} \frac{d_{i|a}}{A_i} \frac{d_{j|b}}{A_j}$$

Where $d_{i|a} \equiv A(a \cap i)$ is the area of the intersection of the zone a (in the new tessellation) with zone i (in the initial one, analogously with $d_{j|b}$, and A_i is the area of the zone i, analogously with A_j.

6.2.6 Computation of Population

One interesting procedure corresponds to computing the current population in the region around Donostia that it is considered in the Project. The result of this procedure provides a means to compare with current information of the census are allow us to gauge the penetration of Euskaltel cell provider.

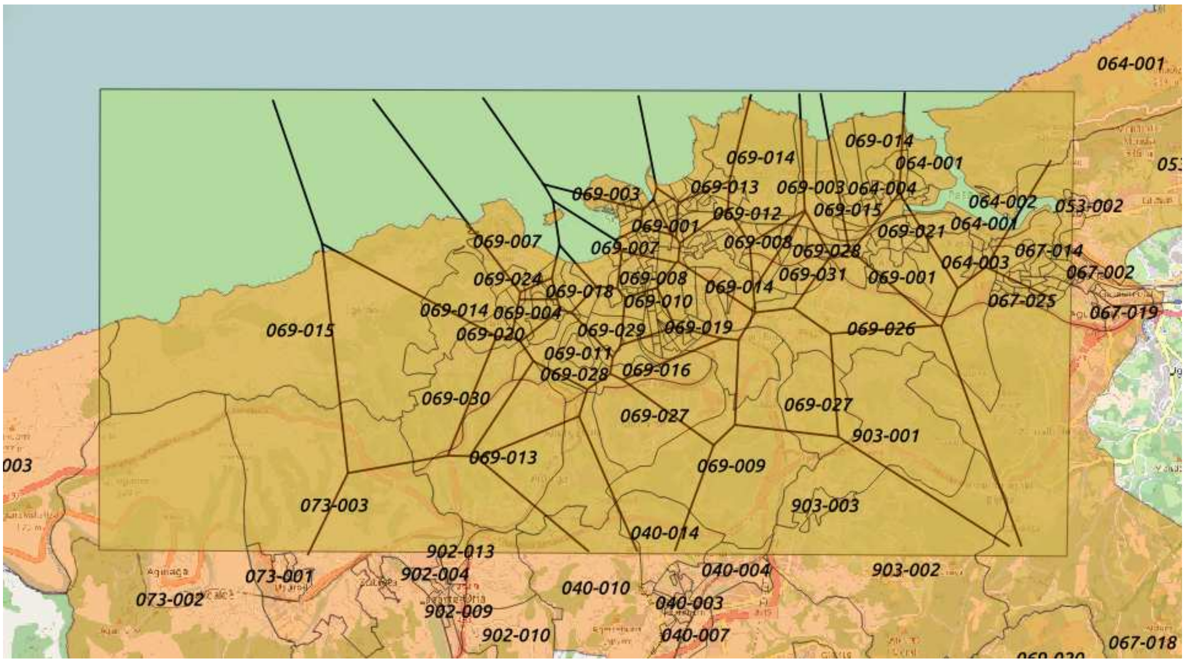

Figure 10: Voronoi zones and the census section in the same map round the considered region around Donostia.

Firstly, the census data is considered. The census data has been obtained from [7] while the actual geographical section corresponding to the census data has been obtained from [8]. The main interest is to assign a population value to each of the Voronoi zones which constitutes the initial and computation geographical division. In order to extrapolate this information, It was assumed, that the population in each of the census section is homogenously distributed and use a GIS program, [9], in order to perform the calculation.

With the aim to compute the population of each of the Voronoi zones using the CDRs it is considered that a user lives in a particular Voronoi zone if most of the nights within a week sleeps in that area. This assumption does not hold true for night workers or mobile devices that sleep at the work place. The procedure is performed again using a Map-Reduce process in which the aim is to compute the number of stops that start and end on different calendar days or they start and end very early on a single day. The results are accessible through the demo web site very amounts to one third of the population given by the census. In *Figure 11* the color code corresponds to the measured population following this procedure.

6.3 Visualization

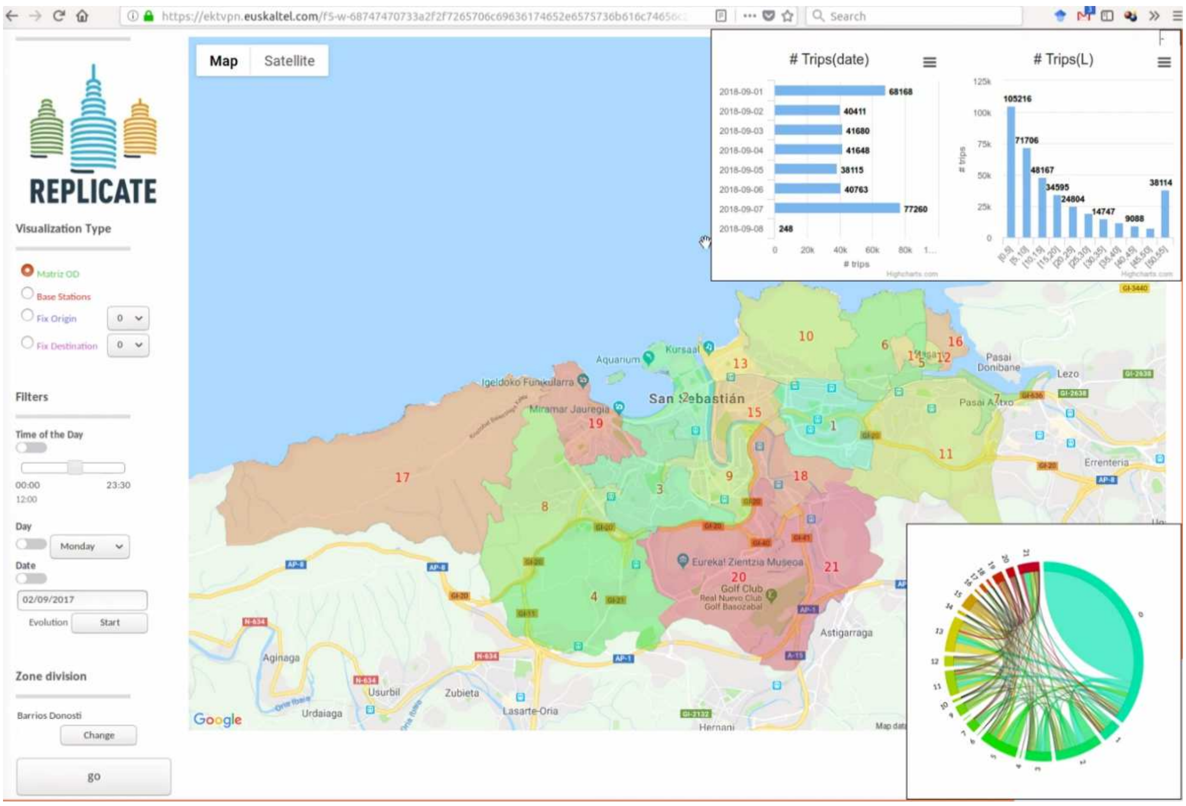

Figure 11: Chord Diagram visualization of the O/D matrix

There exist multiple ways to visualize the obtained results. In the project two different methods has been considered: via chord diagrams or via colored maps.

A chord diagram consists in a circle divided in sectors, where each of the sectors denotes one of the Voronoi zones (the region of points that are served by the same base station). The length of the circular sector in the diagram is directly related to the total number of trips originated from that zone. When these sectors are related to others, the width of the connecting chords corresponds to the number of trips from the original sector to the destination one. In right lower side of the *Figure 12* it can be seen an example of such diagram, with 0 being the region outside the Voronoi zones. In this Figure the colors of the different zones do not contain information and are only to distinguish better the different zones.

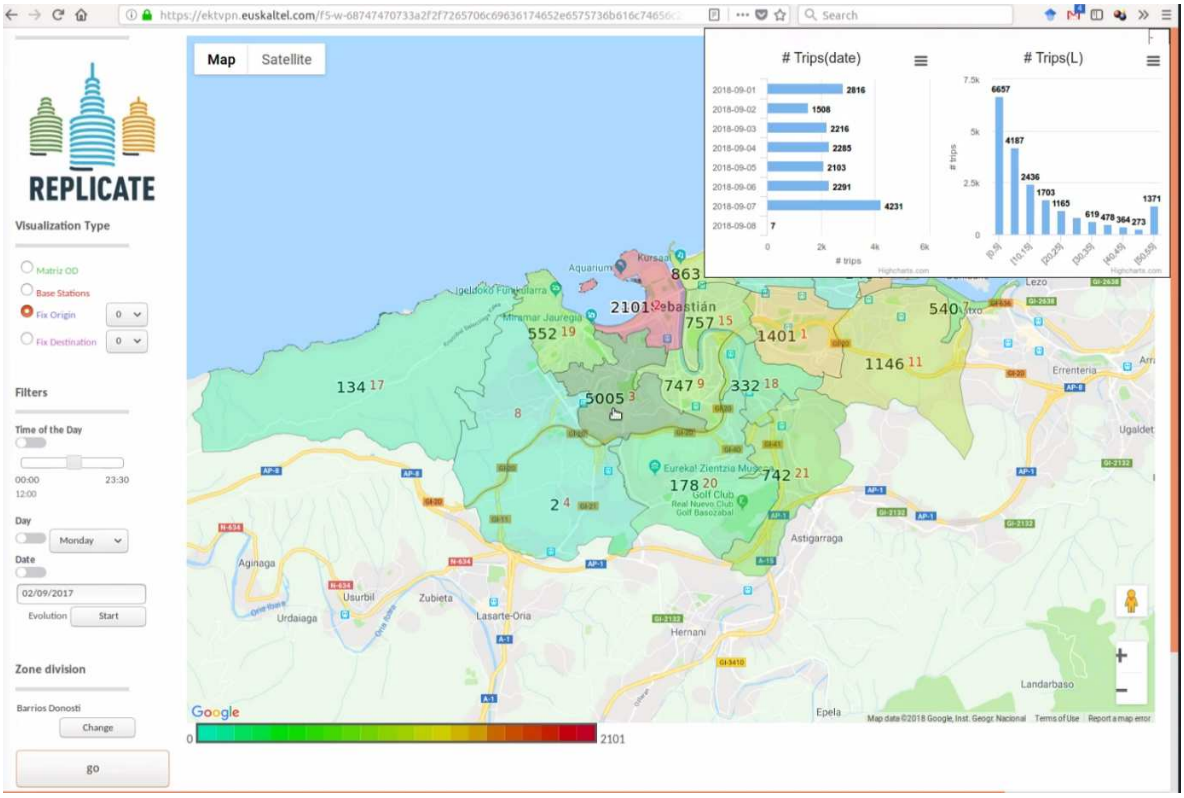

Figure 12: Colored Map for visualizing O/D matrix Row or Column.

A different visualization can be obtained by fixing the origin (row of the O/D matrix) or the destination (column). In this way it is different zones are painted with different colors considering the number of trips that start/end with the selected zones.

6.4 Integration

An Application Programming Interface has been defined and agreed with the Municipality, to export calculated diagrams and reports. In Annex I there is defined in detail the different interfaces.

7. COMMUNICATIONS NETWORK DEVELOPMENTS FOR SMART CITIES

7.1 Mobile Network

Thanks to a full MVNO (Mobile Virtual Network Operator) agreement with Orange, Euskaltel is able to provide mobile services for both its residential and business customers. Euskaltel is the largest MVNO operator in Spain. The core network of the company currently supports GSM, UMTS, HSPA and LTE services. Services are offered under the full MVNO model for both current technologies.

In order to implement the Full MVNO model, because of the nature of these networks, Euskaltel has been provided with the entire necessary infrastructure to allow for a complete control of the customers, except for radio access.

This network enables the phone services offered to all mobile customers in a virtual mobile network operator according to the "Full MVNO" model.

Figure 13: "Full MVNO" model

Customer control goes from the device, whose SIM is provided by Euskaltel, through controlling commuting stages, routing and signaling, as well as voice traffic and data traffic and all the applications, platforms and infrastructure associated to business support. This is part of Euskaltel's intention to improve and enriching its mobile services offer, improving its integration with other Euskaltel's services in a landline-mobile convergence scenario.

The circuit-switch mobile network comprises two mobile commuting plants (GMSC) and two control nodes with three functions included (HLR/FNR/AUC) which, together, service all mobile voice Euskaltel's customers.

The packet-switch mobile network comprises HSS, MMe, GGSN and DNS for establishing data connections. All nodes are deployed in a redundant structure.

The GGSN transports data information towards the national or roaming host operators´s SGSN. This part facilitates IP connections towards customers' networks, Internet and the Service Network, managing the routing and allocation of IP addresses, creates, keeps and erases packages connections from mobile customers and generates billing registers.

The need to implement services of higher value, controlling data applications used by the customer, with download quota management and bandwidth or application restriction policies has let to Euskaltel including a DPI solution (Deep Packet Inspection) in which PCRF controls traffic and GGSN applies policies provided by DPI.

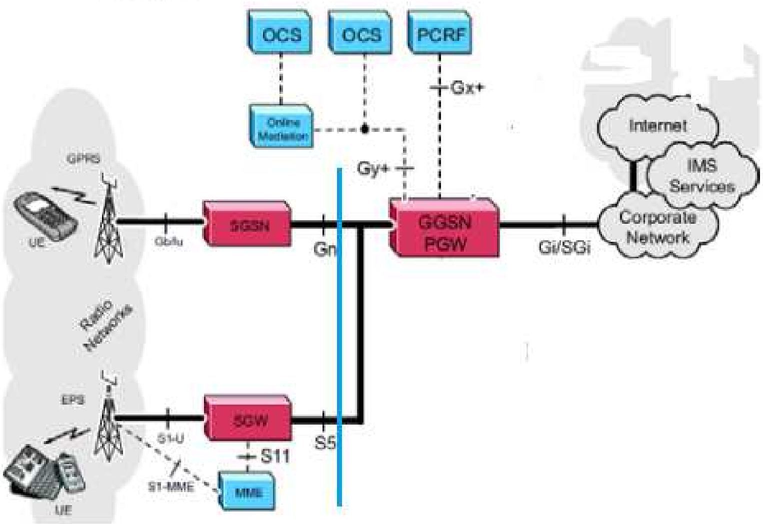

Figure 14: Mobile network architecture

The GGSN registered all the activity log into the CDR (Call Detail Record) files. It is a feature of the system that takes the details of data traffic, such as type, time, duration, amount of Bytes and, identification of the antenna where the mobile is connected. CDRs can be used for network control, accounting and billing purposes. In particular, in our case, the used data is the user identifier, the base station identifier (telephony antenna) and the day and time when the connection is made. These data make up most of the information processed.

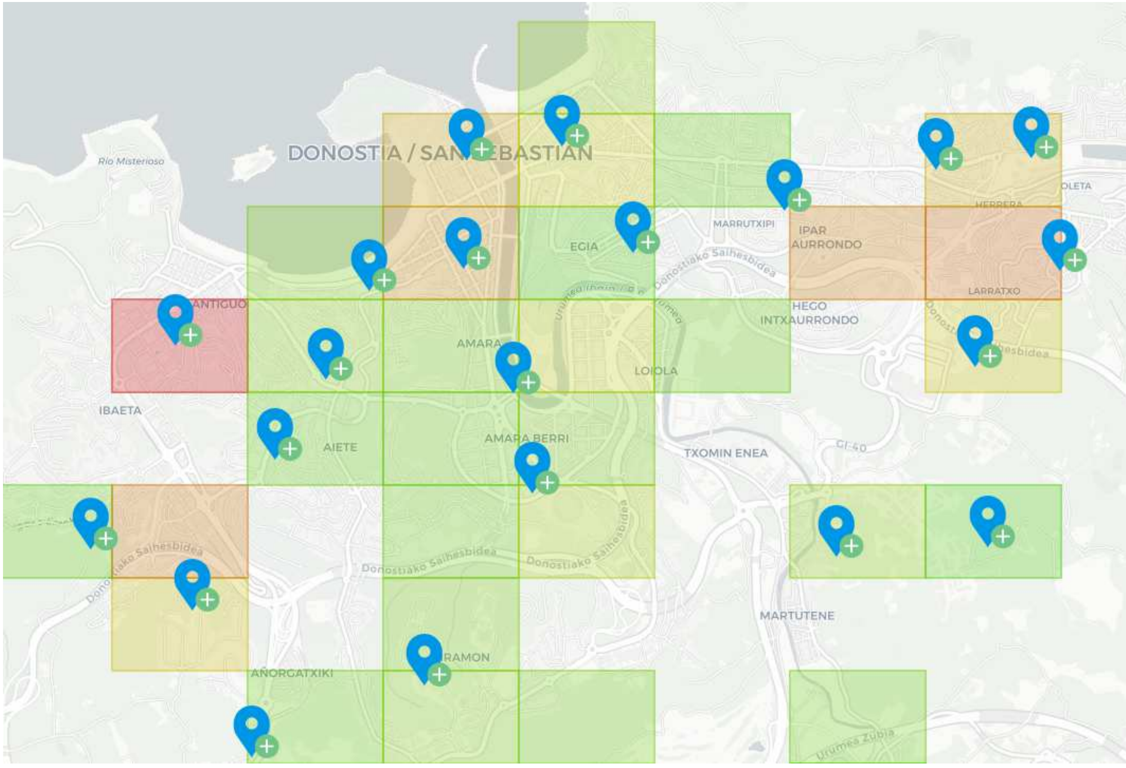

Figure 15: Areas covered by antennas

This graph determines the areas covered by antennas and the color chart shows the volume of users connected to them.

- Green zones: Areas with low population density.
- Yellow zones: Areas with low/medium population density
- Orange zones: Areas with medium/high population density
- Red zones: Areas with high population density

7.2 WiFi-Kalean

WIFI Kalean is a system that allows WIFI to be shared among the different Euskaltel customers. The clients assigned to this service share the WIFI routers that they have contracted with the rest of the users that connect via WIFI in a transparent manner.

Figure 16: Areas covered by wifi

In this graph the distribution of Wifi points available in the area of Donostia is shown.

The coverage of this Operator SP-Wi-Fi service, has been completed by Access Points installed in relevant places of the Basque country.

The following diagram shows the different blocks involved in Euskaltel SP-Wi-Fi solution:

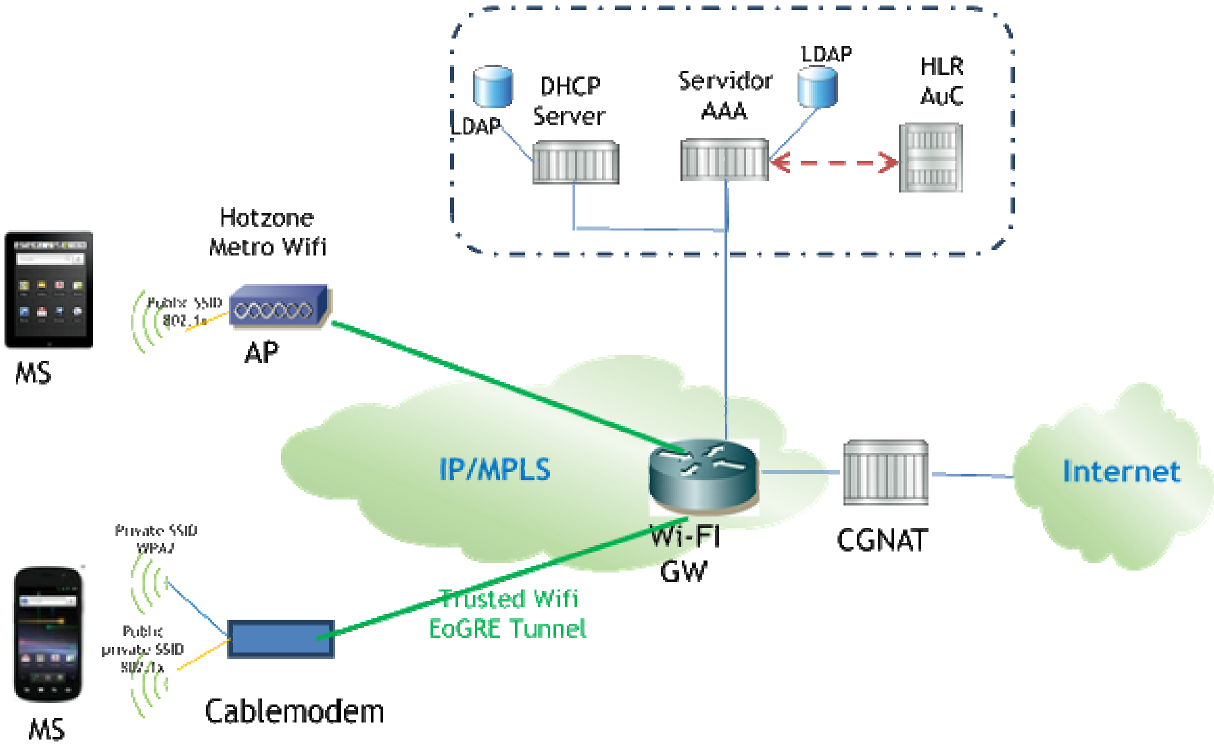

Figure 17: SP-WiFi network architecture

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

This service allows the users to access Internet once the user has been identified on the system through credentials. Credentials can be obtained during a previous registration process or in case of smartphones that support EAP-SIM, they can be included in the device (smarphone SIM card).

The service is supported on a specific SSID available to customers that have the appropriate credentials.

Users receive a private IP address and have access to Internet through the CGNAT platform. The Internet service characteristics will depend on the customer type.

Users need to go through a registration process to obtain credentials. This process allows the following points:

1. Users to be identified for legal purposes
2. Determine users' relation to Euskaltel, as service user will depend on their link with Euskaltel.

Service access is provided through two types of equipment, that give place to two types of architectures. The access connection to the Core is always through the HFC Euskaltel network, using Docsis 3.0 CMs. Customer CMs will usually radiate SP-Wi-Fi SSID service. The Hotspot coverage has been extended using dedicated APs that will connect to Euskaltel network through a DOCSIS 3.0 CM.

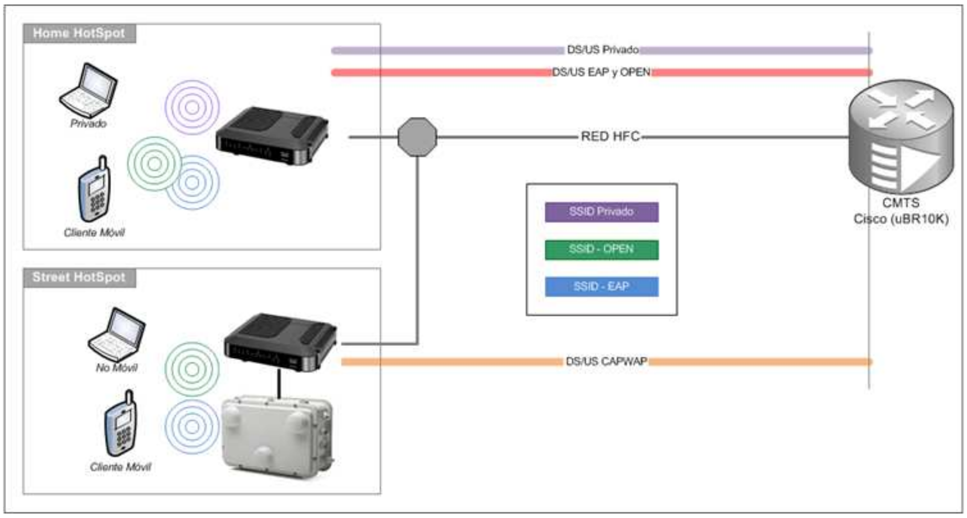

Figure 18: Access points connectivity

Accounting information of authenticated customer sessions generated by AAA Server is stored to know Access Points the customer connected to in every session. This information is mandatory for legal purposes on data retention and will additionally allow Euskaltel identify"Hot Spots" of the Wifi access network that concentrate most activity.

This accounting information generated by customers' smartphone connection to SP-Wi-Fi service, will be processed by a Big Data platform - REPLICATE Project-, and will allow Euskaltel to geolocalize customers through Access Points geo-localization.

8. DATA LAKE

8.1 Infrastructure/architecture

In order to be able to face the challenge of storage and treatment in pseudo real time, that is from logs, HortonWorks HDP 2.4 has been installed in the Euskaltel DataCenter.

Due to consume the events directly from the network would be highly intrusive, the process loads the logs. These logs are generated continuously with no affection on the equipment. This is why the process is called pseudo real time

The platform infrastructure consists of 26 nodes, 13 storage nodes and 13 process nodes.

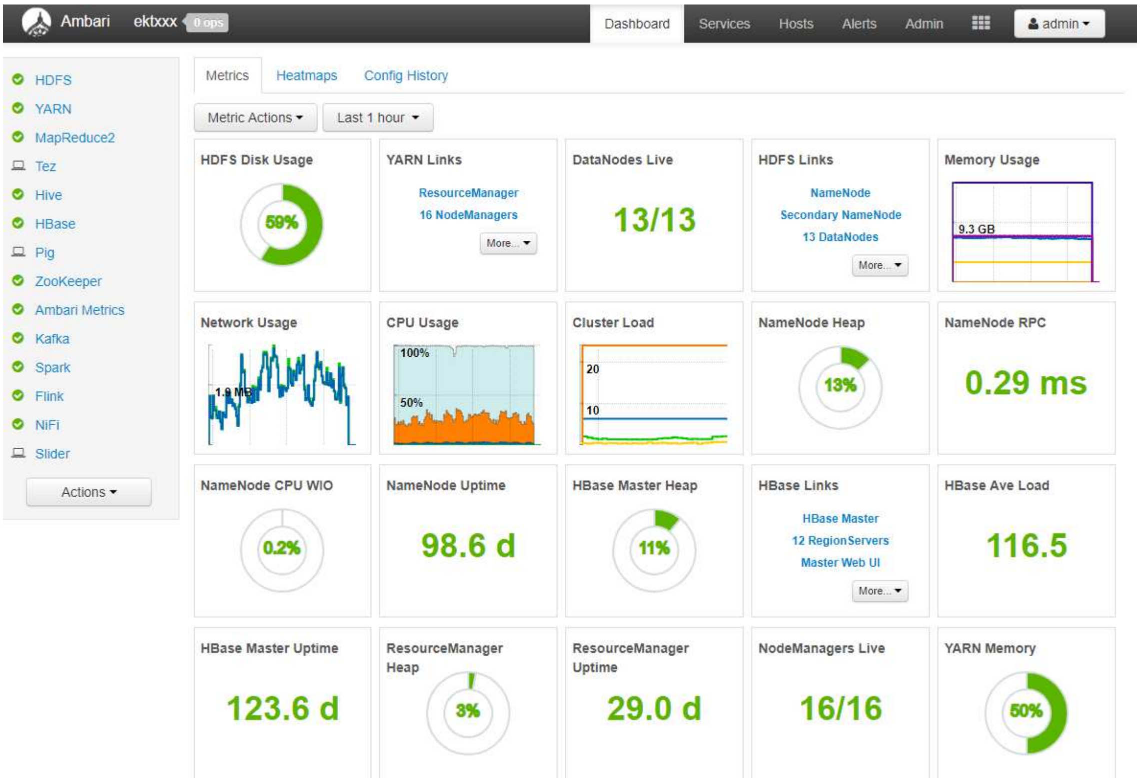

Figure 19: Ambari control panel of HDP platform

The information of the network elements is introduced in the platform through NIFI.

Once the network team has sent the log files to the bigdata SFTP a NIFI controller enters it into the platform. log to log the NIFI controllers break down the information until they get all the internal information encoded in the log.

Once the information is decoded, it is packaged in AVRO objects and sent to the Kafka tails where it persists until the next step.

As final point the processes in flink streaming read the queue and register to register add the location information with GeoHash attributes and stores the result in the database HBASE.

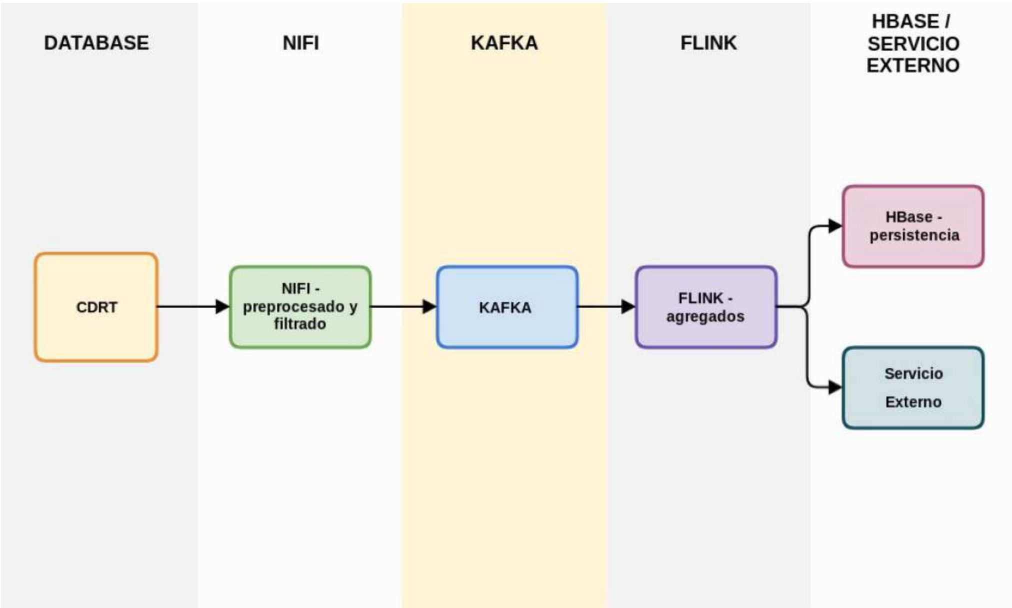

Figure 20: Information flow and intake

- *DATABASE*: Records arrive at the SFTP
- *NIFI*: First phase of ingestion and treatment.
- *KAFKA*: Intermediate persistence.
- *FLINK*: Aggregated information.
- *HBASE/SERVICIO EXTERNO*: Final persistence and consultation of geohases for the heat map.

8.1.1 HortonWorks

HortonWorks Data Platform (HDP) is the HortonWorton distribution for Bigdata environments. It is a framework for distributed storage and processing of large datasets from multiple sources. HDP enables agile application deployment, automatic learning and deep learning workloads, real-time data storage, and security and governance.

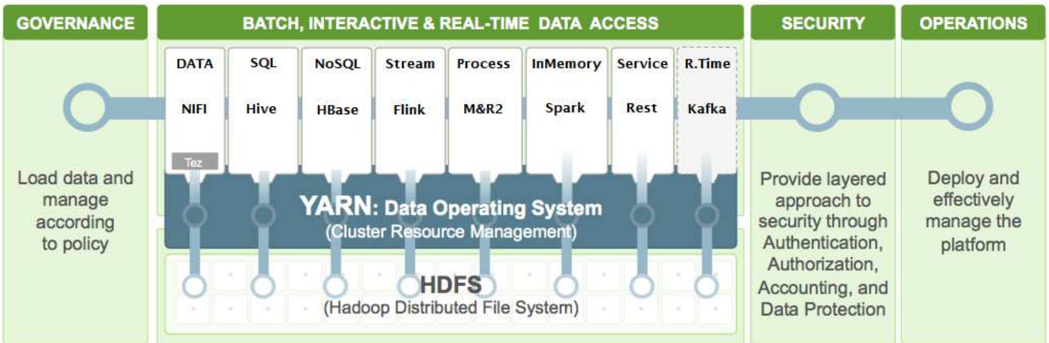

Figure 21: HDP Global Diagram

8.1.1.1 Apache Nifi – Data Routing and Transformation

Apache Nifi is an integrated data logistics and simple event processing platform, that provides an end-to-end platform that can collect, curate, analyze and act on data in real-time. This platform can be deploted on-premise, or in the cloud.

The basic capabilities of NIFI are reading and transforming information with high fault tolerance, flow control and process parallelization.

Some characteristics:

- a Web-based product
    - Seamless experience between design, control, feedback, and monitoring
- Highly configurable
    - Loss tolerant vs guaranteed delivery
    - Low latency vs high throughput
    - Dynamic prioritization
    - Flow can be modified at runtime
    - Back pressure
- Data Provenance
    - Track dataflow from beginning to end
- Designed for extension
    - Build your own processors and more
    - Enables rapid development and effective testing
- Secure

- SSL, SSH, HTTPS, encrypted content, etc...
- Multi-tenant authorization and internal authorization/policy management

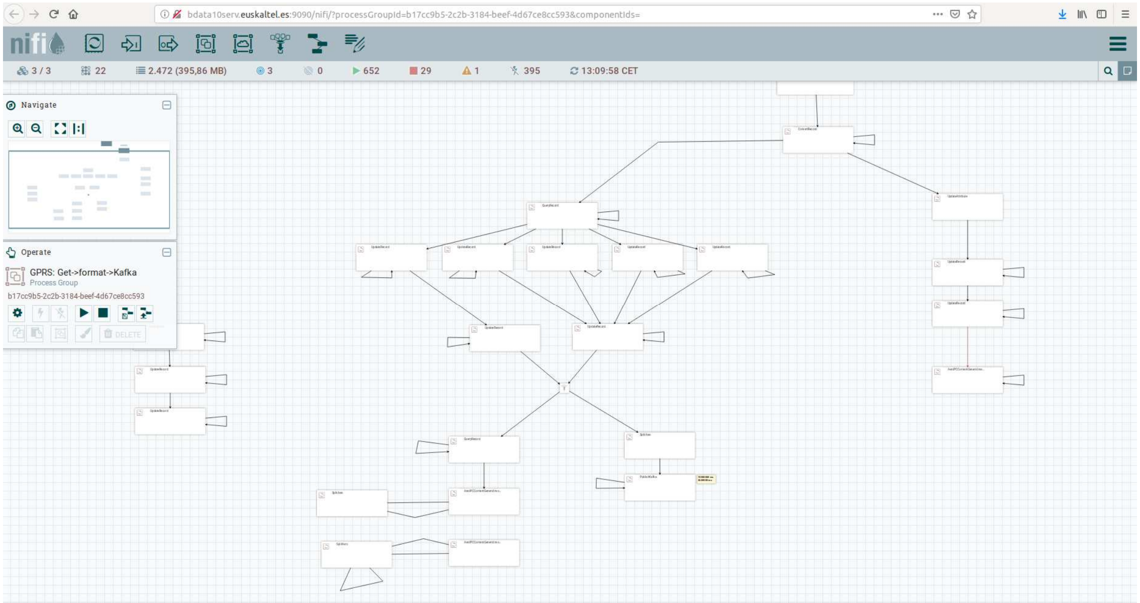

Figure 22: NIFI Dashboard

8.1.1.2 Apache Kafka – Distributed Streaming Platform

As streaming platform has three key capabilities:

- Publish and subscribe to streams of records, similar to a message queue or enterprise messaging system.
- Store streams of records in a fault-tolerant durable way.
- Process streams of records as they occur.

Kafka is generally used for two broad classes of applications:

- Building real-time streaming data pipelines that reliably get data between systems or applications
- Building real-time streaming applications that transform or react to the streams of data

Concepts

- Kafka is run as a cluster on one or more servers that can span multiple datacenters.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

- The Kafka cluster stores streams of *records* in categories called *topics*.
- Each record consists of a key, a value, and a timestamp.

8.1.1.3 Apache Flink – Stream Procesing

Distributed processing engine for stateful computations over *unbounded and bounded* data streams.

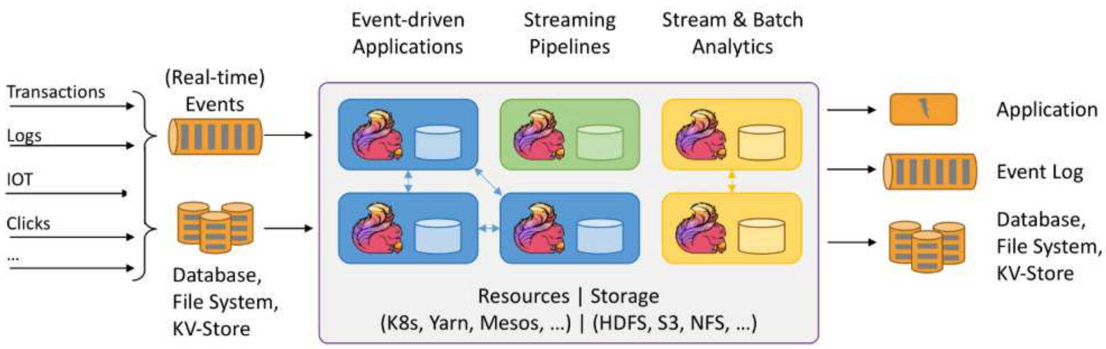

Figure 23: Flink Architecture

Any kind of data is produced as a stream of events. credit card transactional information, sensor measurements, machine logs, or user interactions on a website or mobile application, all these data are generated as a stream.

Data can be processed as *unbounded* or *bounded* streams.

1. **Unbounded streams** have a start but no defined end. They do not terminate and provide data as it is generated. Unbounded streams must be continuously processed, i.e., events must be promptly handled after they have been ingested. It is not possible to wait for all input data to arrive because the input is unbounded and will not be complete at any point in time. Processing unbounded data often requires that events are ingested in a specific order, such as the order in which events occurred, to be able to reason about result completeness.
2. **Bounded streams** have a defined start and end. Bounded streams can be processed by ingesting all data before performing any computations. Ordered ingestion is not required to process bounded streams because a bounded data set can always be sorted. Processing of bounded streams is also known as batch processing.

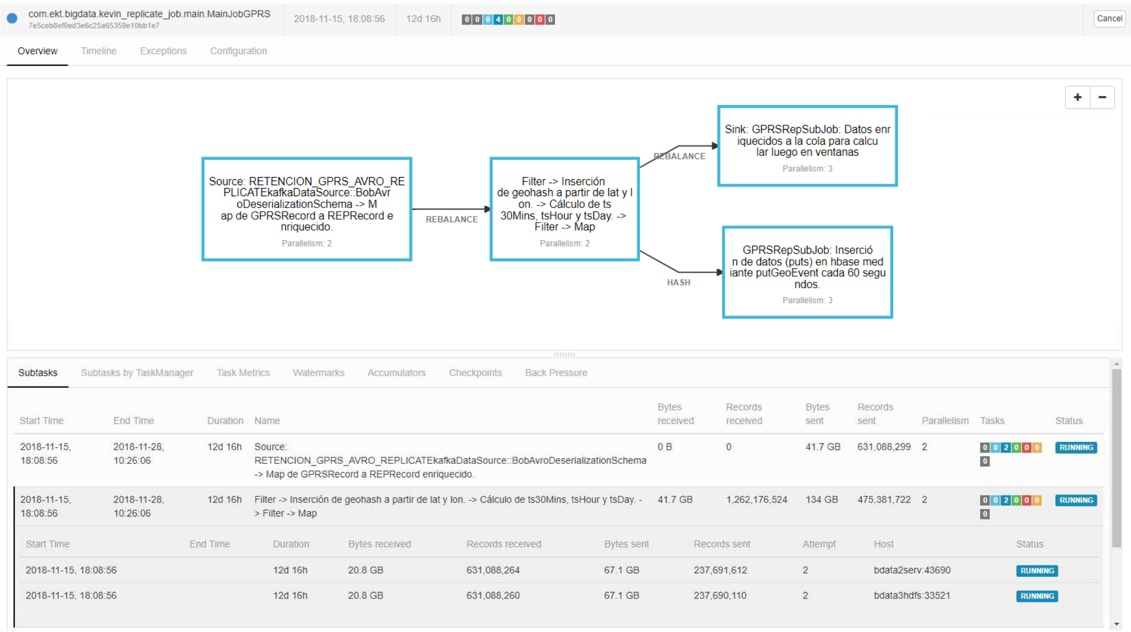

Figure 24: Flink Dashboard extract

8.1.1.4 Apache HBASE – Database for BigData Storage

Apache HBase provide random, realtime read/write access to Big Data, hosting of very large tables, billions of rows X millions of columns, atop clusters of commodity hardware.

- Linear and modular scalability.
- Strictly consistent reads and writes.
- Automatic and configurable sharding of tables
- Automatic failover support between RegionServers.
- Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.
- Easy to use Java API for client access.
- Block cache and Bloom Filters for real-time queries.
- Query predicate push down via server side Filters
- Thrift gateway and a REST-ful Web service that supports XML, Protobuf, and binary data encoding options

8.2 Data Input

8.2.1 Mobile Data Source

The mobile data sources used are the CDR files written by the network elements. These CDRs are encoded under the ANS1 standard and in Ericsson's proprietary format.

In this format the data is encoded in binary so that the space occupied by the information is the minimum.

The charging standard 3GPP TS 32.240 specifies field categories. The following scheme is used:

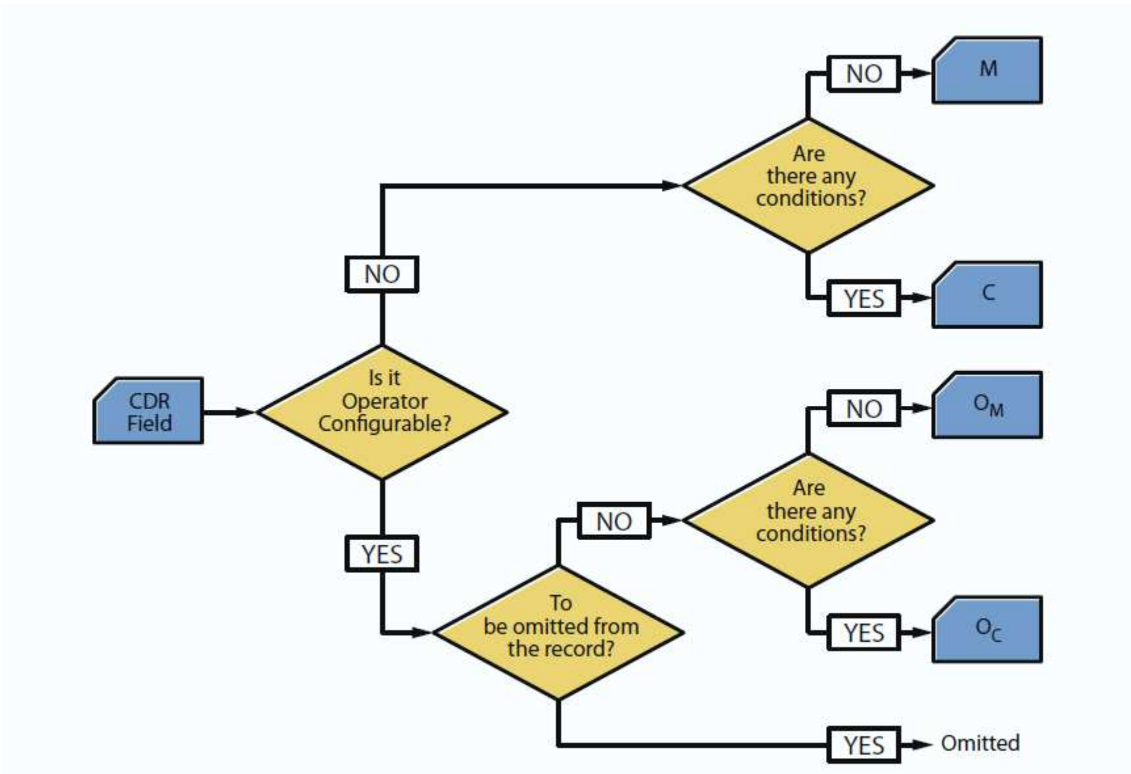

Figure 25: Flow that determines the occurrence of events in the CDRs

- Mandatory field (M): The field is always present in a CDR.
- Conditional field (C): The field is present if a required condition is fulfilled.
- Mandatory and operator-configurable field (OM): The field is always present if not excluded by configuration.
- Conditional and operator-configurable field (OC): The field is present if a

required condition is fulfilled and the field is not excluded by configuration.

Lists the fields that can be present in a G-CDR, eG-CDR, or PGW-CDR together with the field categories for each supported release of the charging standard and for the special closing CDR (Special).

The records are stored in ASN R7 format. That is:

ggsnPDPRecord [21] GGSNPDPRecord

GGSNPDPRecord ::= SET

{

recordType [0] RecordType,

servedIMSI [3] IMSI,

ggsnAddress [4] GSNAddress,

chargingID [5] ChargingID,

sgsnAddress [6] SEQUENCE OF GSNAddress OPTIO

accessPointNameNI [7] AccessPointNameNI OPTIONAL,

pdpType [8] PDPType OPTIONAL,

servedPDPAddress [9] PDPAddress OPTIONAL,

dynamicAddressFlag [11] DynamicAddressFlag OPTIONA

listOfTrafficVolumes [12] SEQUENCE OF ChangeOfCharCo

recordOpeningTime [13] TimeStamp,

duration [14] CallDuration,

causeForRecClosing [15] CauseForRecClosing,

recordSequenceNumber [17] INTEGER,

nodeID [18] NodeID OPTIONAL,

recordExtensions [19] ManagementExtensions OPTIO

localSequenceNumber [20] LocalSequenceNumber,

apnSelectionMode [21] APNSelectionMode OPTIONAL,

servedMSISDN [22] MSISDN OPTIONAL,

chargingCharacteristics [23] ChargingCharacteristics,

chChSelectionMode [24] ChChSelectionMode OPTIONAL

iMSsignalingContext [25] NULL OPTIONAL,

sgsnPLMNIdentifier [27] PLMN-Id OPTIONAL,

servedIMEISV [29] IMEI OPTIONAL,

rATType [30] RATType OPTIONAL,

mSTimeZone [31] MSTimeZone OPTIONAL,

userLocationInformation [32] OCTET STRING OPTIONAL

}

CDR normalization example

The following information is an example of how the binary CDR is decoded in the NIFI step to get the basic information from antennas. In the record there is information regarding the telephone numbers and their location as well as other data such as data consumed in the time between the opening and closing of the CRD.

1934656711880|0|0|0|0|0|0|0|0|0|0|?.?.?.?|i??????.euskaltel.mobi|0800|09/08/2018 20:26:29 CEST|1787|22|37|0|0|2245725040|0121|70|NUMBER??????|6|21403|031B931F7396B7FE7483FFFF004A00|0|1809082226302B0200|0112F4300C363C15|01.21403.800316.21|031B931F7396FEFE764BFFFF0085005300|0|1809082256162B0200|0112F4300C363C15|01.21403.800316.21||0|||||0|||||0||||8663860322715612|?.?.?.?

The meaning of the data is:

ServedMSISDN|DataVolumeGPRSUplink0|DataVolumeGPRSDownlink0|DataVolumeGPRSUplink1|DataVolumeGPRSDownlink1|DataVolumeGPRSUplink2|DataVolumeGPRSDownlink2|DataVolumeGPRSUplink3|DataVolumeGPRSDownlink3|DataVolumeGPRSUplink4|DataVolumeGPRSDownlink4|**SgsnAddressV4Bin**|**AccessPointName**|ChargingCharacteristics|**RecordOpeningTime**|**Duration**|CauseForRecClosing|RecordSequenceNumber|CreditControlEnabled|ResultCode|LocalSequenceNumber|PdpType|**RecordType**|ChargingID|RATType|extractedPLMNId|QosNegotiated0|ChangeCondition0|ChangeTime0|**UserLocationInformation0**|extractedULI0|QosNegotiated1|ChangeCondition1|ChangeTime1|**UserLocationInformation1**|extractedULI1|QosNegotiated2|ChangeCondition2|ChangeTime2|UserLocationInformation2|extractedULI2|QosNegotiated3ChangeCondition3|ChangeTime3|**UserLocationInformation3**|extractedULI3|QosNegotiated4|ChangeCondition4|ChangeTime4|**UserLocationInformation4**|extractedULI4|**ServedIMEI**|**ggsnServedPDPAdressV4Bin**

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

8.2.2 Wifi Data Source

The accounting information that will be sent to the Big Data platform will have an entry for every Start ticket and another entry for every Stop ticket.

Here it is an example of the accounting log:

```
2014-06-
04_09:52:49;00002E65;680xxxxxx;100.64.103.11;442a.xxxx.xxxx;4c00.8216.xxxx;iWAG-
MKT-1.mkt.euskaltel.net;;;;Start
2014-06-
04_10:24:02;00002E64;680xxxxxxx;100.64.103.10;6848.xxxx.xxxx;4c00.xxxx.xxxx;iWAG-
MKT-1.mkt.euskaltel.net;570;0;9551;Stop
```

Fields are separated by ";". Position and meaning of every field is the following:

Table 1. Content of the WiFi accounting CDR

Pos.	Field	Comment	RADIUS Attribute	Start	Stop
1	Date	AAAA-MM-DD_HH:MM:SS format	N/A	✓	✓
2	SessionID	Session identifier. Acct-Session-Id RADIUS (44) Attribute.	(44) Acct-Session-Id	✓	✓
3	Username	User identifier. User-Name Attribute.	(1) User-Name	✓	✓
4	IP-Addres	IP address used by user	(8) Framed-IP-Address	✓	✓
5	User Mac	User's device MAC. MMMM.MMMM.MMM format.	(31) Calling-Station-Id	✓	✓
6	AP/CM MAC	Access point MAC. MMMM.MMMM.MMM format.	N/A	✓	✓
7	iWAG	iWAG hostname	(32) NAS-Identifier	✓	✓
8	Session time	Session time in seconds	(46) Acct-Session-Time		✓
9	DS traffic	Received Bytes by user	(43) Acct-Output-Octets		✓
10	US traffic	Sent Bytes by user	(42) Acct-Input-Octets		✓
11	Type	Start or Stop ticket	N/A		✓

8.3 Data Output

Once the data is processed and enriched, the platform leaves them stored in HBASE, available for the calculation modules and for the heat map painting service.

Example of data for calculation modules

ANI_CODE????????????????????000 **column**=a:cell, **timestamp**=1533701407000, **value**=???-???-????-???

ANI_CODE????????????????????000 **column**=a:type, **timestamp**=1533701407000, value=\x00\x00\x00\x00

ANI_CODE????????????????????000 **column**=b:ani, **timestamp**=1533701407000, value=NUMBER??????

ANI_CODE????????????????????000 **column**=g:geohash, **timestamp**=1533701407000, value=eztvzghdhy9b

ANI_CODE????????????????????000 **column**=g:lat, **timestamp**=1533701407000, value=@E\x9B\x88e\x94\xAFO

ANI_CODE????????????????????000 **column**=g:lon, **timestamp**=1533701407000, value=\xC0\x06\x89k\xB9\x8C~(

MAC_??:??:??:??:??:??:?? **column**=l:A, **timestamp**=1530846705478, **value**=ADDRESS_??????????????

MAC_??:??:??:??:??:??:?? **column**=l:L, **timestamp**=1530846705478, **value**=43.29621222

MAC_??:??:??:??:??:??:?? **column**=l:M, **timestamp**=1530846705478, **value**=MAC_??:??:??:??:??:??:??

MAC_??:??:??:??:??:??:?? **column**=l:X, **timestamp**=1530846705478, **value**=-2.25747338

Example of data for heat map

01530181800000ezwq5pu71sb8 **column**=S:COUNT, **timestamp**=1537375950918, **value**=\x00\x00\x00\x00\x00\x00\x00\x01

01530181800000ezwq5pu71sb8 **column**=S:COUNTG, **timestamp**=1537375950918, **value**=\x00\x00\x00\x00\x00\x00\x00\x01

01529427600000eztyhw5f1q7g **column**=B:L, **timestamp**=1529506269566, **value**=@E\xA3D^\xD4\xA1\xAD

01529427600000eztyhw5f1q7g **column**=B:X, **timestamp**=1529506269566, **value**=\xC0\x07\xB0U2a|\x1C

eztugtpc5v4p **column**=C:LAT, **timestamp**=1542683766024, **value**=@E\x86j\x8Bd!\xBE

eztugtpc5v4p **column**=C:LON, **timestamp**=1542683766024, **value**=\xC0\x07\xFE\xDE'\x12=\x9A

eztugtpc5v4p **column**=C:MAC, **timestamp**=1542683766024, **value**=??:??:??:??:??:??:??

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

9. INNOVATIONS, IMPACTS AND SCALABILITY

9.1 Innovation solution

The use of CDR data for mobility analysis is of increasing interest to the transport planning agents; here, the potential to use CDR data to underpin mobility analysis significantly advances the trend away from time consuming and/or resource intensive techniques. Repurposing CDR data provides several advantages: unparalleled scale, coverage, spatial granularity and temporal accuracy.

Although, the processes stages and supporting architecture needed in the exploitation of operational are known, the main innovation and success of the proposed work are due the data richness and large coverage of Euskaltel in the pilot geographical scope. Specifically, the composition of trajectories manages the fusion of complementary CDR and WiFi Kalean data. Additionally, another innovation relies in the fact that without deploying additional infrastructure for sensing, the information provided included both vehicles and pedestrian, those last, very difficult to be monetarized with other sensing techniques.

9.2 Social impacts

In this case, the social impact is indirect, a good understanding of the demand and mobility needs in the city, contrasted with the current transport offer, will identify neighborhoods or areas with particular needs.

9.3 Environmental impacts

The possibility of aggregating mobility information, a demand model, will allow a continuous monitoring of the Urban Sustainable Mobility Plan of the city, constituting one of its indicators. By definition, the USMP constitutes a framework of objectives and planning in the short, medium and long term, encompassing actions and proposals in order to improve urban sustainability, in environmental, social coexistence and economic terms.

9.4 Replication and scalability potential

The deployed infrastructure and algorithms will be able to manage, from the ingestion, calculation and aggregated information provision of a more extensive region or higher temporal interval of analysis.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

9.5 Economic feasibility

The availability of data extracted from telecommunication operation logs about the mobility of people in the city, both vehicles and pedestrians, could minimize the development and installation of additional infrastructures in the cities for mobility management. Specific, mobility analytics algorithms applied to this data can provide very exact information about mobility heat points, origin-destiny matrix, etc. As result, an efficient business model can be created, where the cities can save money in deployment and management of infrastructures, and network operators can monetize the data of the network.

9.6 Impact on SME´s

A lot of SMEs of this area depends on tourism. The small commerce and tourism that will have information of the surroundings and with it to realize performances that allow to catch more visitors, offering promotions and measuring the impact of their marketing campaigns.

9.7 Other

The city of San Sebastian in order to promote a green mobility and the decision making, identifies as a key issue the knowledge of the mobility in the city. Currently, an annual data consolidation is carried out, with insufficient frequency to assist in the operation. The information is not exploited in a coordinated way, data sources are not merged, they are treated separately, so it is even difficult to assess whether the information that is already available is sufficient or not.

Additionally, only vehicular information, conventional traffic sensory information and transactional public transport information were addressed. The availability of aggregate characterization of urban mobility based on operational information from mobile network companies, will Improve the support for decision making be means of a deeper knowledge of the mobility in the city.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

10. CONCLUSIONS

The deliverable D3.9 Use of Big Data for Mobility Services that described the work carried out in the context of the aggregate characterization of urban mobility based on operational information from mobile network companies has some advantages compared to traditional methods. These are the advantages: a non-intrusive capture of information, there is no active participation of citizens, the data collected are not subject to subjective interpretations, a diagnosis of global mobility regardless of the mode of transport used and the availability to obtain a large amount of geolocated data over time with a high representativeness in the study area, DonostiaIdea. In the context of the project, in addition to the Call Detail Records (CDRs), the connectivity information of the Kalean WIFI service has been included, complementing the first source in local detail in the urban area.

In summary, the value of the information has been contrasted, to improve real-time knowledge of urban mobility in San Sebastian pilot. To achieve this goal, a Data Lake, the Big data infrastructure built to store and transform the data coming from the network into people movement information, and a set of mobility analytics algorithms, have been developed, defined to extract aggregated information. Additionally, an Application Program Interface for the information of the third parties (in this case, San Sebastian Municipality) is now available.

This development can show how the use of Big Data can serve to produce services of interest for cities and telecommunication companies. The new context of a private-public partnership that arises as result of this project, enables benefit for all the partners involved, and as result, for the citizen.

11. References (Bibliography)

[1] https://en.wikipedia.org/wiki/Voronoi_diagram

[2] Steven Fortune. "A sweepline algorithm for Voronoi diagrams". *Proceedings of the second annual symposium on Computational geometry*. Yorktown Heights, New York, United States, pp.313–322. 1986. ISBN 0-89791-194-6. ACM Digital LibrarySpringerLink

[3] https://en.wikipedia.org/wiki/Fortune%27s_algorithm

[4] http://static.googleusercontent.com/media/research.google.com/es/us/archive/mapreduce-osdi04.pdf

[5] http://hadoop.apache.org/

[6] https://hbase.apache.org/

[7] http://www.eustat.eus/movil/elementos/ele0010400/ti_Poblacion_por_distritos_y_secciones_censales_de_Gipuzkoa_segun_sexo_grupos_de_edad_y_nacionalidad_2011/tbl0010406_c.html#axzz4hEsO7YyX

[8] http://www.geo.euskadi.net/cartografia/DatosDescarga/Limites/Unidades_estadisticas/SECCIONES_EUSTAT_5000_ETRS89.zip

[9] https://www.qgis.org/es/site/

[10] http://www.ingeniovirtual.com/tipos-de-graficos-y-diagramas-para-la-visualizacion-de-datos/

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 691735

12. ANNEX I. Web Services

12.1 Origin-Destination Matrices

Web service responsible for returning the information of the Origin / Destination matrices calculated.

The name of the service is *getMatriz*, with the following access points:

➢ *getTotalMatrix*

Obtains the complete Origin / Destination matrix, with all the data added, without applying any filter of days or hours.

- Url: server/webresources/getODMatrix/getTotalMatrix
- Method: GET
- Input parameters: None or QueryParam (inline parameter) field

Name	Data type	Parameter type	Description
zones	string	Optional	Name of the file that contains the zoning (GEOJSON format) to be used. This file must have been previously uploaded with the geometryManager service, or manually to the server ("data" folder of the service).

In case of not receiving parameters, the zone division by default will be used.

.

- Output data:

JSONArray with the matrix Origin Destination: two-dimensional array (N + 1) x (N + 1), where:

- N: number of zones of the division. The dimensions of the array are of size (N + 1) to include those trips that are outside the study areas (indexes 0 of the array).
- Element [i] [j]: number of trips originating in zone "i" and destination in zone "j"

- Element [0] [j]: number of trips originating outside the study area and destination in zone "j".
- Element [i] [0]: number of trips originating in zone "i" and destination outside the study area.

12.2 O/D Matrices with distribution by travel length

Gets the Origin / Destination matrix offering the possibility of including filters through the Input parameters.

- Url: server/webresources/getODMatrix/getODMatrix
- Method: GET
- Input parameters: QueryParam structure (online parameter) with the fields corresponding to the filters that are to be applied

Name	Data type	Parameter type	Description
Date	String	Optional	Date in m / d / a format with which to filter the data.
Day	Int	Optional	Day of the week, from 0 to 6 (0 = Monday, 6 = Sunday) with which to filter the data. Keep in mind that if you include the date field and the day field. In the probable case that the selected date does not correspond to the corresponding day of the week, the server returns an error 400 when making an inconsistent request.

hour	String	Optional	Time in hh: mm format with which to filter the data
startDate	String "yyyy-MM-dd HH:mm"	Mandatory if endDate is not null	Date and Time of the beginning of the query to be performed. If present, the date, day and hour parameters are not taken into account.
endDate	String "yyyy-MM-dd HH:mm"	Mandatory if startDate is not null	Date and Time of the end of the query to be made
interval	String "XXa"	Optional	Denotes the period to perform the aggregation. "XX" is the numerical value and "a" denotes the units that can be: 'm': minutes, 'h': hours, 'd': days, 'S': 7 days, 'M': 30 days. If this parameter is not included, the result is the matrix from the beginning to the end added. If included, aggregation is done. The minimum value that can be considered is 30 minutes, which is the smallest unit that is calculated in the project.

			A period less than 30 minutes returns all matrices calculated in the service, ie every 30 mins.
zones	String	Optional	Name of the file that contains the zoning (GEOJSON format) to be used. This file must have been previously uploaded with the geometryManager service, or manually to the server ("data" folder of the service).

- Output data (in case interval does not appear):
  JSONArray with the Origin / Destination matrix filtered according to the parameters: two-dimensional array (N + 1) x (N + 1), where:
  - N: number of zones of the division. The dimensions of the array are of size (N + 1) to include those trips that are outside the study areas (indexes 0 of the array).
  - Element [i] [j]: number of trips originating in zone "i" and destination in zone "j"
  - Element [0] [j]: number of trips originating outside the study area and destination in zone "j".
  - Element [i] [0]: number of trips originating in the "i" zone and destination outside the study area.

- Output data (with a non-null Interval):
  JSONArray whose elements have the following format:
  {"date": "2017-02-01 13:00", "mat": [[21849,121,110,159,105,85,73,165, ...]], "num": 40}

- "date": denotes the date of the first element of the aggregation, it does not have to match startDate because it corresponds to the aggregation of 30 minutes stored in the server.
- "mat": the Origin / Destination matrix added in the period specified by the interval parameter.
- "num": number of aggregations of 30 minutes that are part of the aggregate in the period given by the Interval parameter.

12.3 O/D Matrices with distribution by date

Web service in charge of returning the information of the Origin / Destination Matrices with the segmented data based on the length of the trips.

The name of the service is *getDistributions*, with the following accesses:

➢ *getTotalLengthDistribution*

Obtains the Origin / Destination matrix, segmented by travel length, without applying any filter of days or hours.

- URL: server / webresources / getDistributions / getTotalLengthDistribution
- Method: GET
- Input parameters: QueryParam structure (inline parameter) with the fields:

Name	Data type	Parameter type	Description
zones	string	Optional	Name of the file that contains the zoning (GEOJSON format) to be used. This file must have been previously uploaded with the geometryManager service, or manually to the server ("data" folder of the service).

In case of not receiving parameters, the zone division by default will be used.

- Output data:

JSONArray with the origin / destination matrix segmented by the length of the trips: three-dimensional array (N + 1) x (N + 1) x 11, where:

- N: number of zones of the division. The dimensions of the array are of size (N + 1) to include those trips that are outside the study areas (indexes 0 of the array).
- Third dimension (0..10): different lengths of trips, which have been divided into 5km sections:
  - Section 0 : length trips between 0 and 5km
  - Section 1 : length trips between 5km and 10km
  - Section 2 : long trips between 10km and 15km
  - ...
  - Section 9 : long trips between 45km and 50 km
  - Section 10 : trips longer than 50km.
- Element [i] [j] [t]: number of trips originating in zone "i" and destination in zone "j", and with the length corresponding to section "t"
- Element [0] [j] [t]: number of trips originating outside the study area and destination in zone "j", and with the length corresponding to section "t"
- Element [i] [0]: number of trips originating in area "i" and destination outside the study area, and with the length corresponding to section "t".

➢ *getLengthDistribution*

Gets the matrix Source / Destination complete, segmented by travel length, offering the possibility of including filters through the Input parameters.

- URL: server / webresources / getDistributions / getLengthDistribution
- Method: GET
- Input parameters: QueryParam structure (inline parameter) with the fields corresponding to the filters that are to be applied:

Name	Data type	Parameter type	Description
date	string	Optional	Date in m / d / a format with which to filter the data.
day	int	Optional	Day of the week, from 0 to 6 (0 = Monday, 6 = Sunday) with which to filter the data. In the probable case that the selected

			date does not correspond to the corresponding day of the week, the server returns an error 400 when making an inconsistent request.
hour	String	Optional	Time in hh: mm format with which to filter the data
zones	string	Optional	Name of the file that contains the zoning (GEOJSON format) to be used. This file must have been previously uploaded with the geometryManager service, or manually to the server ("data" folder of the service).

- Output data:
  JSONArray with the Origin / Destination matrix segmented by the length of the trips, and filtered according to the Input parameters: three-dimensional array (N + 1) x (N + 1) x 11, where:
  - N: number of zones of the division. The dimensions of the array are of size (N + 1) to include those trips that are outside the study areas (indexes 0 of the array).
  - Third dimension (0..10): different lengths of trips, which have been divided into 5km sections:
    - Section 0 : length trips between 0 and 5km
    - Section 1 : length trips between 5km and 10km
    - Section 2 : long trips between 10km and 15km
    - ...
    - Section 9 : long trips between 45km and 50 km
    - Section 10 : trips longer than 50km.
  - Element [i] [j] [t]: number of trips originating in zone "i" and destination in zone "j", and with the length corresponding to section "t"
  - Element [0] [j] [t]: number of trips originating outside the study area and destination in zone "j", and with the length corresponding to section "t"

- Element [i] [0] [t]: number of trips originating in zone "i" and destination outside the study area, and with the length corresponding to section "t"

12.4 Zone Configuration Management

Web service in charge of managing everything related to the geometries of the sectors in which the study area is divided.

This new service obeys to the new functionality of selection of different division zones, since the data are calculated and stored for a specific division. For this it is necessary to upload the GEOJSON file to the server with the geometry of the new zones.

The new geometries are stored in the server ("data" folder) and its associated information in the database.

The name of the service is geometryManager, with the following access points:

➢ *getGeometries*

Returns a list of the geometries that are stored in the database.

- URL: server / webresources / geometryManager / getGeometries
- Method: GET
- Input parameters: none.
- Output: JSON structure with the fields:

Name	Data		Description
status	"ok" / "error"		Result of the query. Only in case of "ok", the following fields will also appear.
geometries	Array of elements		Array with the geometries stored in the database.
	geomid	int	Identifier of the geometry in the DB (key of the table)
	geomfile	string	name of the geometry file created on the server
	geomname	string	name of the geometry

current	int	Identificador de la geometría que se está utilizando en la visualización.

➢ *uploadGeometry*

In charge of uploading the geometry file on the server, performing the following operations:

- Save the file uploaded in the "data" folder of the server, giving it the name the timestamp + "_" + original name of the file. Example: if the original file was "barrios_donosti.geojson", the file will be stored as "1532675979112_barrios_donosti.geojson"
- It generates a new file with the geometric centers of the zones, with the same name of the file uploaded + "areaCenters" (in the previous example, it would be "1532675979112_barrios_donosti.geojson_areaCenters"). This file is used for the "Base Stations" option of the web interface.
- Perform the intersections of the new zones with the previous ones, which will be used to adapt the data to the new zones. This is done here so that it is only done once, because they are complex calculations and it takes some time.
- Add the information to the database where the information of the new geometries is stored.
    - URL: server / webresources / geometryManager / uploadGeometry
    - Method: PUT
    - Input parameters: when receiving a file from the HTML form, it receives the parameters in the MULTIPART_FORM_DATA format. These parameters are:
        - @FormDataParam ("newgeomname") String newgeomname
        - @FormDataParam ("file") InputStream uploadedInputStream
        - @FormDataParam ("file") FormDataContentDisposition fileDetail
        - @FormDataParam ("newgeomidfield") String IDfield
    - Output: JSON structure with the fields:

Name	Data	Description
status	"ok" / "error"	Result of the query. Only in case of "ok", the following fields will also appear.
name	string	Name of the geometry, the same that is passed by parameter in the call to the service

filename	string	Name of the file uploaded, in the previous example it would be "1532675979112_barrios_donosti.geojson"
id	int	Identifier of the geometry created in the DB (key).