Monitoring People Moving during Covid-19 Pandemic in the Emilia-Romagna Region

Gian Paolo Jesi*, Elisabetta Gori†, Andrea Zucchelli‡, Nicola Spazzoli§ and Gianluca Mazzini¶
Lepida ScpA
Via della Liberazione, 15
40128 Bologna, Italy
Email: *gianpaolo.jesi@lepida.it, †elisabetta.gori@lepida.it, ‡andrea.zucchelli@lepida.it,
§nicola.spazzoli@lepida.it, ¶g.mazzini@ieee.org

2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) | 978-1-6654-1490-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCWC51732.2021.9376065

***Abstract*—This paper is about monitoring people movements in Emilia-Romagna : the region of Italy in which our company, Lepida ScpA, is located. The unfortunate event of the 2020 pandemic triggered by the Covid-19 virus gave us the chance to exploit our BIGDATA infrastructure to provide information to the regional Public Administration (PA) in order to take strategic decisions to face the emergency.**

This goal actually corresponds to the aim with which we started last year to design and implement our infrastructure.

Our monitoring project is grounded on our regional WiFi infrastructure. This WiFi access is provided for free to anyone and it is available in all major cities and municipalities over Emilia-Romagna.

We describe the challenges we faced and the choices we made during the process and the final results we achieved.

***Index Terms*—bigdata, dashboard, covid-19, batch processing**

I. INTRODUCTION

Lepida ScpA is a subsidiary of the Emilia-Romagna Region and it is the main operational instrument regarding the implementation of the Regional ICT Plan. In order to accomplish the Plan, it defines the strategies of broadband networks, ensures and optimizes the delivery of ICT services, develops cloud infrastructure, implements and manages innovative solutions for the modernization of healthcare paths to improve the relationship between citizens and the Regional Health Service in accordance with the provisions of the European, National and Regional Digital Agendas.

Lepida ScpA provides a set of specialized services aimed to local Public Administrations (PA) and citizens that produce a huge amount of unbounded heterogeneous (Big) data, such as: (public) WiFi access locations, Regional healthcare and environmental monitoring data coming from our IoT infrastructure [15], [16]. This flux of information is constantly growing. This BIGDATA creates many opportunities not just for monitoring and managing each single sub-domain, but also for the creation of new business models that foster public and private cooperation. The value of the data corresponds to the chance to be shared and aggregated in a uniform fashion in order to extract *knowledge*, which represents their actual value. On one hand, this value is tangible for PA for its capacity to allocate economic resources (planning) as a function of the perceived needs of the Region and for measuring the effectiveness of the governance; on the other hand, from the private citizen point of view, the availability of reliable and authoritative data, represents the main requirement for having better public services. In addition, this condition represents a potential boost for digital marketing [14].

Since the strategic importance for the PA of being able to exploit these data, Lepida ScpA started the creation of its first implementation [17] of a BIGDATA infrastructure in order to continue its tradition of being the technological reference for the PA in the Emilia-Romagna Region.

The team, working with an Agile approach [3], [6] on the project, achieved the first working infrastructure capable to gather interesting analytics in about six months. During 2020, the infrastructure evolved from a testing to a production environment. Unfortunately, this evolution process happened in correspondence with the worldwide Covid-19 pandemic. In addition to the tragedies generated by the pandemic, the social and economic impact over the region have been dramatic and still they are.

Inline with the original aim - i.e., providing new opportunities - with which we started the BIGDATA project we exploited our data to provide information to the local PA in order to take strategic decisions aimed to fight the pandemic emergency. In this vein, we made the most of the flux of data concerning the WiFi Access Points (AP) of the Region trying to monitor how people is moving. In fact, in the last few years, Lepida ScpA installed thousands of 100% free WiFi AP in public places, such as: train stations, plazas, libraries, schools, and in PA offices, conforming to the guidelines the European Digital Agenda. This asset allows us to monitor if the Regional *restrictions* and/or *lock-down* rules are effective and how they actually limit people movement.

The remainder of this paper is organized as follows. We first discuss the scenario in which we operate in Section II

Section III discusses the choices taken for our architecture and technological details, while Section IV addresses the implementation of the mechanisms required to generate the knowledge for our monitoring system. In Section V we describe our results.

Finally, Section VI we draw our conclusions and discussions for this work.

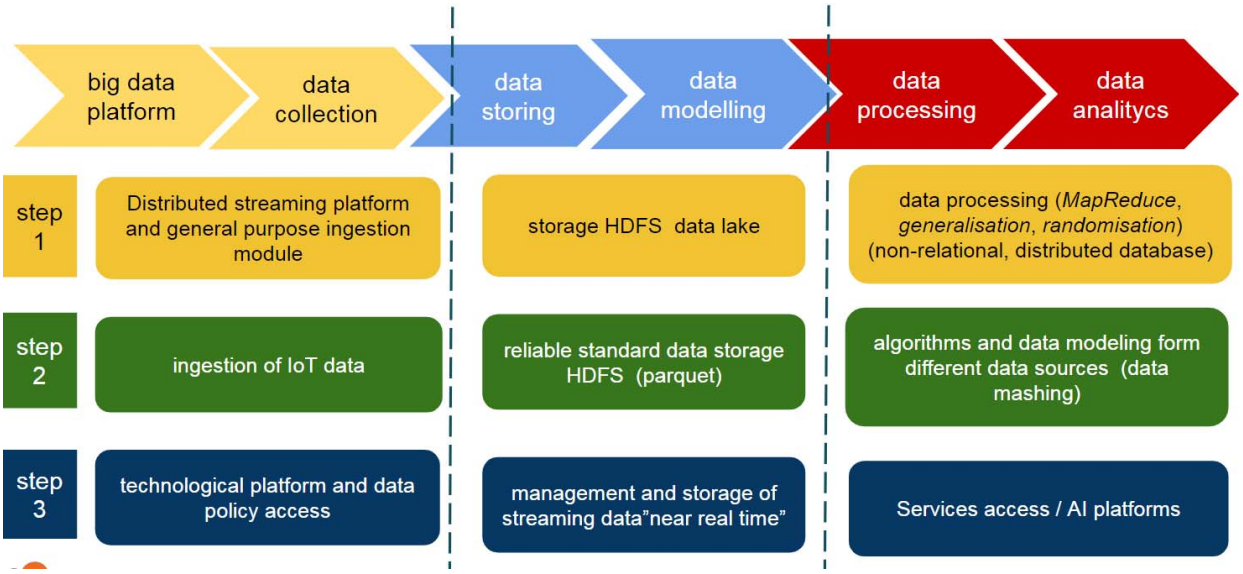


Fig. 1. Logical blocks of a prototypical big data architecture. Columns are referenced as "stages", while rows as "steps".

II. Scenario

The original plan for 2020 before the pandemic strike, was upgrading from the testing to the production environment. This process had to take into account several factors, such as:

- the adoption of a cloud infrastructure provided by Oracle
- the requirement to add the IoT ingestion pipeline of the infrastructure
- provide a secure access the data to partners and customers

In addition to the effort devoted to these factors, when the Covid-19 pandemic started to strike, *we exploited the new infrastructure to monitor people movements in order to provide a feedback to the local lock-down strategies*.

In this work, we focus on our goal to provide insights for the PA and hence on monitoring people movements in our Region and we just briefly touch the other factors.

We reviewed several options before finally choosing the infrastructure for the production deployment. Essentially, since the project is still in its infancy, the economic effort to have all the infrastructure in house and then managing and monitoring it, was way too high. We choose Oracle cloud because we already are their customers and their offer was competitive and their support has always been excellent. The downside of this choice is the need to learn to manage a complex set of cloud services that turned out to be very challenging for our a small team.

The second factor requires to add a novel pipeline in addition to the WiFi AP data. This new pipeline regards the ingestion of our IoT infrastructure [15], [16]. This requirement stems from one of our partners who is interested to visualize the environmental monitoring carried our by the (LoRa) sensors deployed in its municipality. In other words, this is not just a matter of ingesting data from the IoT infrastructure, but also to visualize in (almost) real time the relevant data through a dashboard.

The access to the dashboard opens the road to the last factor: provide a private and secure access to partners and customers. In Europe especially, privacy issues are taken very seriously and even more when a PAis involved. The fact of having to comply to the strict GDPR [7] rules further complicates how entities can eventually share data.

Essentially, the last two factors introduce step two and three of first stage and step three of the second stage depicted in Figure 1.

III. Architecture

We adopted a conservative approach with the architectural design of our infrastructure [17] and we followed a prototypical *lambda architecture* [1].

Essentially, the lambda architecture involves the presence of two parallel pipelines where the first (i) is dedicated to batch processing, while the second (ii) is dedicated to data streaming.

The current trend in the BigData area, especially in those context involving *unbounded* data, such as in IoT, is to build BigData infrastructures on top of just a streaming pipeline, since a *well designed streaming pipeline* can achieve the same result of batch processing with little or no compromise [2].

However, our choice of adopting the lambda architecture is motivated by several factors. The main one is dictated by the experience of our team which is more used to traditional relation database systems and SQL-like queries. In this vein, at the beginning, the majority of our infrastructure in the test environment has been implemented over the batch pipeline. Having both pipelines allows managing latency, throughput and resilience of the platform.

We started to fully design and implement the streaming one during this year directly on the production environment, since the introduction of the IoT ingestion pipeline.

A more practical issue of lambda architectures it to be forced in building and maintaining two distinct code bases to achieve a single goal.

The foundation of our monitoring is based on top of WiFi AP data which are located all over the Emilia-Romagna region.

Since the ingestion of WiFi data is the first pipeline we set up and it is based on the batch layer, our Covid-19 people movements is still mainly batch oriented.

IV. Implementation

Our monitoring system implementation is described as follows. As we previously stated, we exploit the WiFi data collected from the AP logs. Each log record is a semi structured element which contains details such as: AP ip, AP mac, device mac, model, SSID and event (e.g., 'connect' or 'disconnect'). Apache Flume is responsible to read the syslog daemon [9] stream where all APs logs are collected. Here, the user device mac is anonymized by applying a hash function (i.e., SHA256). In this manner, the sensitive information [7] regarding the user disappears from our domain. Each received record is finally encoded into a JSON record.

Flume dispatches the data stream to Kafka which, in turn, dispatches the same stream to the batch oriented pipeline and to the streaming one. The former is represented by the Hive database, while the latter is represented by Elasticsearch. Before being actually dispatched to a pipeline, each record is geo-referenced by merging the record information with an AP location data-set and an AP registry data-set. Both these data-sets are actually stored in a traditional DBMS (Postgres).

On the batch pipeline, all augmented records with locations are stored into a Hive table called '$merge\_table$'.

A time triggered script[1] written in Pyspark is responsible to generate a series of small data-sets. Each data-set is designed to answer a specific question and to draw a specific plot in the dashboard.

In order to be easily accessed by both the Metabase platform and JupyterLab, all these data-sets are stored as Postgres tables.

Table I shows all the knowledge stored in Hive or in Postgres that is relevant for our Covid-19 related monitoring.

The most interesting table is '$corona\_movements$' which is where we calculate all the movements that each user (i.e., each entity represented by its unique mac address which has been hashed) has made over time. The idea is simple and it is iterated day by day. In other words, the movements are calculated every night on just the events of the previous day and the result is appended to the table. As a first step, all records which are 'CONNECT' events and belongs to the previous day, are taken from $merge\_table$. A window is created, partitioned by the hashed mac and ordered by date. A new column 'id' is added to the selected subset leading to the generation of a temporary table. The fresh 'id' column is set to the corresponding row number according to the window partition. A '$source$' data-set is created from the daily subset by collecting all the following fields: hashmac, human_date, apip, latitude, longitude, municipality, province and id. All fields are renamed with a '<current_name>_src' pattern. A '$destination$' data-set is created from the daily subset following the same strategy as before. In this case however, all fields are renames with a '<current_name>_dst' pattern. Both '$source$' and '$destination$' data-sets are temporary tables.

In order to extract movements events, we join $source$ and $destination$ tables as follows:

```
SELECT * FROM source JOIN destination ON
(source.hashmac_src=destination.hashmac_dst
AND
destination.id_dst=(source.id_src + 1))
```

Essentially, we recognize a movement when source and destination have the same hashmac and destination id is the next value of the source. Finally, we just keep only those rows where the IP of the source AP is actually different from the destination AP IP.

This data-set is appended on a day by day basis to the $corona\_movements$ table.

All the tables which are not 'merge_table' and 'corona_movements' are very compact and highly synthetic data sets aimed to answer to a specific question and to produce a single plot in the dashboard. These data sets are all grouped and ordered by date. They are essentially time series data sets and they hold two or three column fields. Using highly synthetic or aggregated data is beneficial for the front end which can produce the dashboard with very little effort in terms of network traffic and computation.

In order to produce map plots showing the geographic location of movements we adopted Elasticsearch and Kibana for the actual plots. As Elasticsearch cannot read from a DBMS, the content of '$corona\_movements$' is also transformed into an elastic index by a script. The script simply converts each data set record into a 'document' generating the required JSON structure.

V. Outcomes

The result of our regional user monitoring during the Covid-19 pandemic is a Metabase dashboard and a commented pdf report which is sent to the regional headquarters by our CEO. During the initial emergency it was generated on a weekly bases, while now is monthly generated.

Here we present the most representative parts of the information presented to the regional headquarters.

The first question we have to answer is how the pandemic and its related lock-down rules have influenced the *number and growth of the users* of the regional network.

The number of unique users detected in 2020 is shown in Figure 2. Two distinct plot are depicted. The bar plot shows the number of unique users per week, while the dotted line shows the sum of all unique users over time. Starting from the end of February (i.e., on February 25th the schools has been closed), the user number started to drop significantly. After the introduction to the restrictions to economical activities and to the real lock-down[2], the growth of new users is basically zero as the line plot is almost horizontal. This is true for the whole lock-down time.

The scenario depicted in Figure 3 is no different. This figure shows the number of connected users per week over time.

[1]It is scheduled by the Unix cron daemon every night ad 0:30 AM.

[2]Respectively on March 8th and March 15th.

TABLE I

KNOWLEDGE REQUIRED TO GENERATE OUR USER MONITORING. THIS KNOWLEDGE IS STORED IN SEVERAL TABLES BOTH IN HIVE AND IN A TRADITIONAL DBMS. WITH THE EXCEPTION OF '*merge_table*' AND '*corona_movements*', ALL THE OTHER ONES REPRESENT A SPECIFIC QUESTION AND PLOT IN OUR DASHBOARD SYSTEM.

Table name	DB	Description
merge_table	Hive	It is generated by (left) joining the *syslog_ap* table with *union* Hive table using the AP IP address as pivot. The *union* table is, in turn, the result of joining AP location with AP registry data.
unique_users	Postgres	Exploits *merge_table*: from all 'CONNECT' events, takes all unique users which are identified by the 'hashmac' field
connections_prov	Postgres	Exploits *merge_table*: counts all 'CONNECT' events and group them by province
connections_data	Postgres	Exploits *merge_table*: counts all 'CONNECT' events and group them by date
connected_users_prov	Postgres	Exploits *merge_table*: counts all distinct users (exploiting the 'hashmac' unique field) in all 'connect' events and group them by province
connected_users_data	Postgres	Exploits *merge_table*: counts all distinct users (exploiting the 'hashmac' unique field) in all 'CONNECT' events and group them by data
corona_movements	Postgres	Exploits *merge_table* in order to collect all movements events. How these events are collected is discussed in Section IV
minicorona	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020
minicorona_extra_municipality	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020, where the source and destination municipality are not the same
minicorona_intra_municipality	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020, where the source and destination municipality are the same
minicorona_extramunicipality_intraprovince	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020, where the source and destination municipality are not the same, while the source and destimation province are the same
minicorona_extraprovince	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020, where the source and destination province are not the same
minicorona_spp	Postgres	Exploit *corona_movements*: counts all movements for each day in the last week (since the current date) and for each distinct province
minicorona_cesenatico	Postgres	Exploit *corona_movements*: counts all movements for each day, since February 2020, where the source or destination municipality is 'Cesenatico'

Fig. 2. Unique users in 2020 collected by the WiFi APs.1st. Data are aggregated per week. The blue bars represents the unique users detected for each week, while the dotted red line represents the cumulative number of distinct users detected over time.

These users are already known by the system. Starting from the end of February the drop of connected users is clear. From March, the number of connected users is about 25% of the average number detected during the pre-Covid era.

Starting from the beginning of May (i.e, the end of the lock-down), the number of unique users and connected ones rises immediately. The slope of the unique users sum curve is pretty similar to the pre-Covid times. The number of connected users instead are lower than expected, but students are still not around and summer vacations are the main reasons.

In fact, since September also the connected users numbers are similar to the pre-Covid times.

Unfortunately, the situation is not going to last long, as the effect of the virus is back. At the end of September, new restriction are introduced [4] and in November [5] these become even more strict with limitation moving to and from distinct regions. The effect is visible in both Figure 2 and 3. The number of unique and connected users is becoming lower and lower every week, closely reaching the levels of the previous lock-down.

In Figure 4, the number of connected users is depicted for each province of the region [3] during the last four weeks starting from the first week of December (inclusive). Each color bar corresponds to a specific week. All provinces exhibit

[3]The provinces correspond to the main cities which are (in the plot order): Bologna, Forlì, Ferrara, Modena, Parma, Ravenna, ReggioEmilia and Rimini.

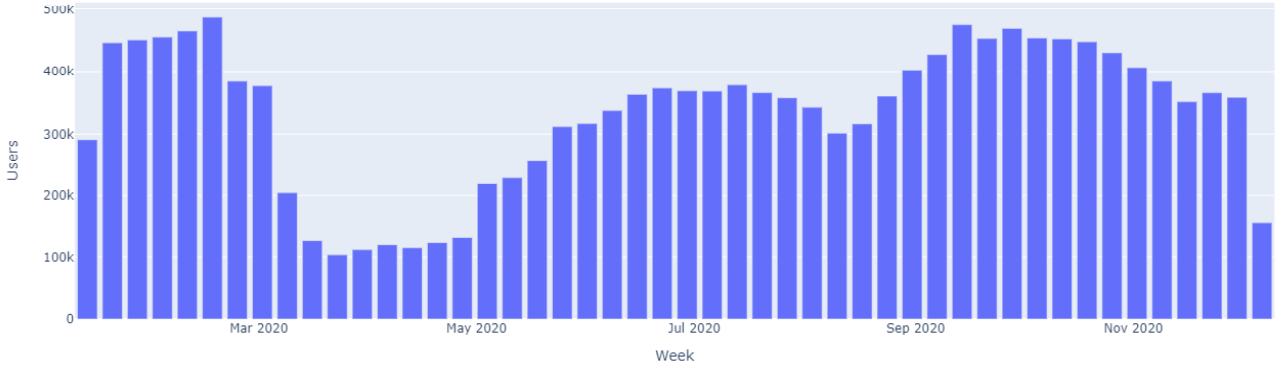


Fig. 3. Users already discovered by the system that connects to any WiFi APs. Data are aggregated per week.

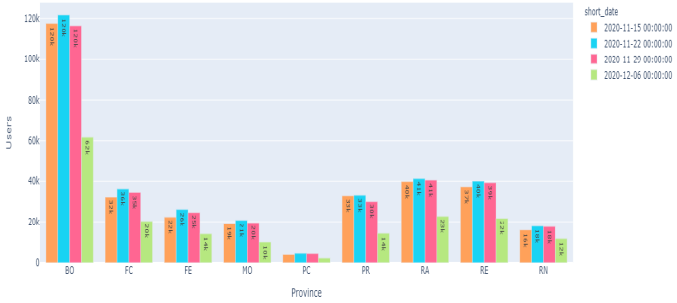


Fig. 4. Connected users during the last four weeks grouped by province. For each province each bar represents a specific week.

a drop in the number of connected users in the last two weeks. In particular, the first week of December counts about 50% less users then the average of the previous weeks. This is a sign of the actual impact of the new restrictions.

Figure 5 shows the daily user movements since March, 1st. User's movements follow a very similar pattern. From March we see a steep drop of the number of movements in the region: during April we have 25% of the value we had during the pre-Covid times.

Starting from September, after vacations and after the end of any restriction, the level of movements reaches high level mimicking the optimism of people. As expected the plot bars follow a pattern in which the working days of the week are the ones with the higher traffic. However, from October, the effects of the virus are back as well as the restrictions to people. The Figure shows a constant decrease of the movement levels at every week. Fortunately, since we are not back into a strict lock-down status, the levels are still higher than in the lock-down of March/April.

It is interesting to understand the type of movements in the region. In fact, we distinguish four kind of movements which are the following: (i) intra municipality, (ii) extra municipality, (iii) intra province and (iv) extra province. Figure 6 focuses on municipality movements in the last 4 weeks starting from the first week of December (inclusive). Essentially, the vast majority of movements are intra municipality ones, while a small proportion regards extra municipality movements. During the weekend, the amount of movements is always lower, especially regarding the extra municipality movements.

The set of extra municipality movements is made by intra and extra province components. Figure 7 highlights their proportion. It depicts the daily province movements during the last 4 weeks starting from the first week of December. The Blue area depicts intra province and hence extra municipality movements, while the red area represents the extra province ones. Four weeks ago the proportion of the extra province set was much higher then the intra province. The introduction of the movements limitations is quite evident especially in the last depicted week, where the proportion of the extra province movements (i.e., red area) is much lower than the other one.

Figure 8 shows the level of movements in the seaside area of Cesenatico on a daily basis. During the lock-down time the movements were very limited in this area. The high peaks are due to the kind of tourism in that area which have rather short period, especially during the weekends.

Finally, the effect of the covid-19 restrictions on the user movements in the EmiliaRomagna region are summarized in Figure 9. Six maps have been generated with Kibana and the snapshots are taken in distinct dates in the first part of the year, which are respectively: January 21st, March 24th, April 14th, May 28th, June 16th, July 14th. All dates are on Tuesday. The blue dots represent the location of a WiFi AP on the territory, while each line represents one or more movements between the two endpoints. The sub-figure 9(a) depict the scenario where the Covid-19 was not perceived a real threat yet. Two months later - sub-figure 9(b) shows the amount of movements at the beginning of the lock-down. The next month (see sub-figure 9(c)) the movements even between province cities are so low that can hardly appear on the map.

Soon after the lock-down some significant movements restart to appear on the map (see sub-figures 9(d,e,f)), but they are still much less intense compared to January. The most realistic reason is that schools and Universities were still

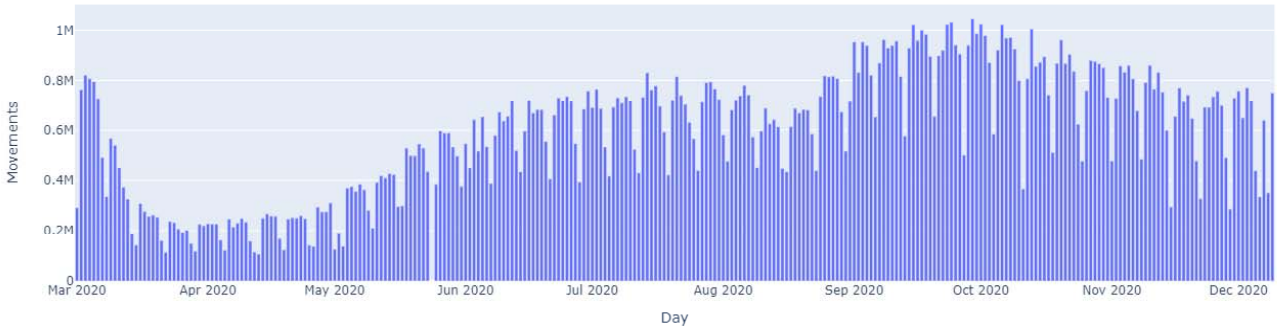


Fig. 5. Daily regional user movements over time; each bar represents a single day.

Fig. 6. Municipality movements during the last 4 weeks starting from the first week of December. The bar plot represents the total amount of movements per day. The red dotted line represents intra municipality movements and the green dotted line represents extra municipality ones.

Fig. 7. Daily province movements during the last 4 weeks starting from the first week of December. The Blue area depicts intra province and hence extra municipality movements, while the red area represents the extra province ones.

closed during these months and, especially this year, many people were on (forced) vacations in June/July.

VI. Conclusions

Lepida ScpA started to design, implement and deploy its own BigData infrastructure in order to provide the foundation for future development for PA and to foster local business opportunities.

The unfortunate event of the Covid-19 gave us the chance to exploit its features and to achieve what it has been designed for: provide new opportunities. In this particular case, we decided to provide information about people movements in our region trying to help our Public Administration to take

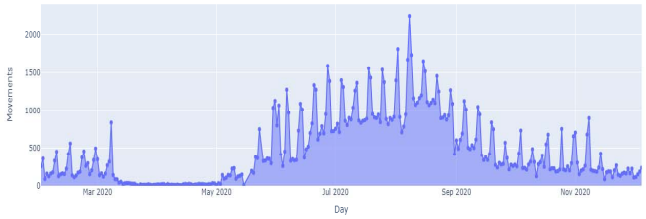


Fig. 8. Daily movements in the Cesenatico seaside zone.

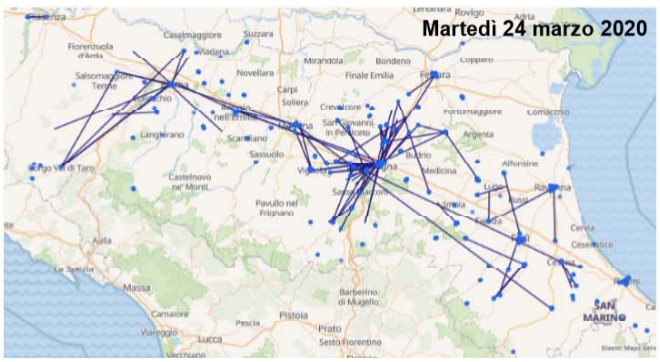
strategic decisions in order to face the emergency. More precisely, our information is suitable to monitor the effect of the restriction rules through the people movements. A statistically reasonable sample of the population is given by the users connect to our regional WiFi facility.

Since April 2020, we started to provide the regional Public Administration with our information in the form of an access link (URL) to our Metabase dashboard and a commented pdf report.

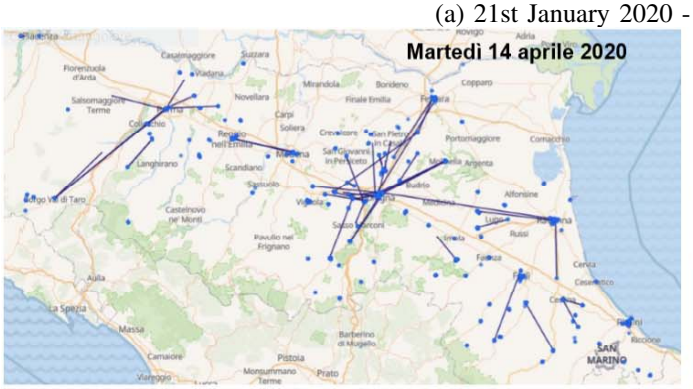
Regarding our future steps, we are still working on the production platform to finalize our architecture which, while it will maintain a batch layer, it is going to be mainly focused on the streaming one.

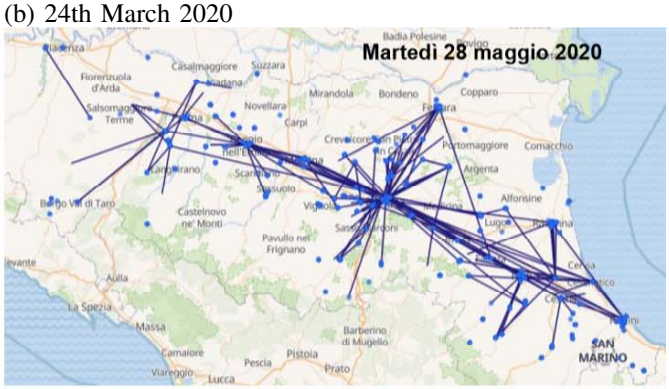
References

[1] Nathan Marz and James Warren, “Big Data - Principles and best practices of scalable realtime data systems”, April 2015, ISBN 9781617290343, pp. 328.
[2] Questioning the lambda architecture, ”https://news.ycombinator.com/item?id=7976785”
[3] Agile Alliance https://www.agilealliance.org/
[4] DECRETO DEL PRESIDENTE DEL CONSIGLIO DEI MINISTRI (DPCM), September 2020, https://www.gazzettaufficiale.it/eli/id/2020/09/07/20A04814/sg
[5] DECRETO DEL PRESIDENTE DEL CONSIGLIO DEI MINISTRI (DPCM), November 2020, https://www.gazzettaufficiale.it/eli/id/2020/11/04/20A06109/sg
[6] Mary Poppendieck and Tom Poppendieck, “Lean Software Development: An Agile Toolkit”, May 2003, pp. 224, Addison-Wesley
[7] The EU General Data Protection Regulation portal, https://eugdpr.org/

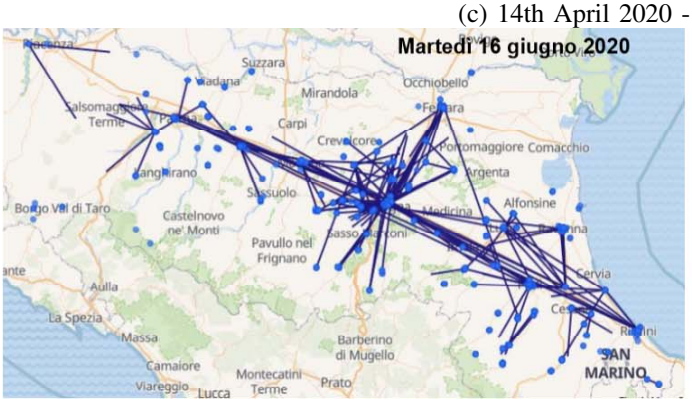


(a) 21st January 2020 - (b) 24th March 2020





(c) 14th April 2020 - (d) 28th May 2020



(e) 16th June 2020 - (f) 14th July 2020

Fig. 9. Movement map snapshots of the EmiliaRomagna region. The blue dots represent the location of a WiFi AP on the territory, while each line represents one or more movements between the two endpoints.

[8] Armbrust M, Xin RS, Lian C, Huai Y, Liu D, Bradley JK, Meng X, Kaftan T, Franklin MJ, GhodsiA, Zaharia M, "Spark SQL: relational data processing in spark", In: Proceedings of the 2015ACM SIGMOD international conference on management of data (SIGMOD15). ACM, New York,pp 1383–1394. doi:10.1145/2723372.2742797

[9] Balliu, A., Olivetti, D., Babaoglu, O. et al. "A Big Data analyzer for large trace logs", Computing (2016) 98(12): 1225-1249, https://doi.org/10.1007/s00607-015-0480-7

[10] Chen Y, Alspaugh S, Katz RH (2012), "Design insights for MapReduce from diverse production work-loads". Tech. Rep. UCB/EECS-2012-17, EECS Department, University of California, Berkeley. http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-17.html

[11] R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

[12] Shvachko K, Kuang H, Radia S, Chansler R, the Hadoop distributed file system. In proceedings of the 2010 IEEE 26th Symposium on mass storage systems and technologies (MSST10). IEEEComputer Society, USA, pp 1–10. doi:10.1109/MSST.2010.5496972

[13] Agile Man Manifesto. https://agilemanifesto.org/iso/en/manifesto.html

[14] AGID Agenzia per l'Italia Digitale. https://www.agid.gov.it/en

[15] Benetti, E., Jesi, G.P, Mazzini, G. "Web interface for managing an Internet of Things Public Network", Software, Telecommunications and Computer Networks (SoftCOM), 2019, 27th International Conference on. IEEE.

[16] Jesi, G.P, Benetti, E., Mazzini, G. "Building an IoT Public Network Infrastructure", In proceeding of the 27th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2019

[17] Jesi, G.P, Gori, E., Micocci, S. and Mazzini, G. "Building Lepida ScpA BigData Infrastructure", In proceeding of the 6th IEEE International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), 21-22 November 2019, Sofia, Bulgaria