ENTER2015
22nd International Conference on Information
Technology and Travel & Tourism
February 3-6, 2015 – Lugano (CH)

A practical approach to big data in tourism: a low cost Raspberry Pi cluster

Mariano d'Amore[a], Rodolfo Baggio[b], and Enrico Valdani[a]

[a]Center for Research on Marketing and Services(CERMES),
Bocconi University, Italy
(mariano.damore, enrico.valdani)@unibocconi.it

[b] Master in Economics and Tourism
Bocconi University, Italy
rodolfo.baggio@unibocconi.it

Abstract

Big Data is the contemporary hype. However, not many companies or organisations have the resources or the capabilities to collect the huge amounts of data needed for a significant and reliable analysis. The recent introduction of the Raspberry Pi, a low-cost, low-power single-board computer gives an affordable alternative to traditional workstations for a task that requires little computing power but immobilises a machine for long elapsed times. Here we present a flexible solution, devised for small and medium sized organisations based on the Raspberry Pi hardware and open source software which can be employed with relatively little effort by companies and organisations for their specific objectives. A cluster of six machines has been put together and successfully used for accessing and downloading the data available on a number of social media platforms.

Keywords: Raspberry Pi; Big Data, online social networks; tourism organisations

1 Introduction and background

Objective of this paper is to describe a system based on scalable open source technology that makes the collection of online social networks data manageable by academics and practitioners for a better understanding of the tourism phenomenon. A simple mapping 'exercise' demonstrates the functionality of the system.

Online social networks (OSNs) are not only a powerful tool for promoting and marketing tourism products and destinations (Leung et al., 2013), but, given their incredible diffusion, are also (and probably more importantly) an extraordinary source of information on a wide range of topics: from the preferences of tourists to their activities at destination, to the way they behave, or to how the value what offered to them. And this can be seen explicitly, through their comments and discussions, or implicitly by tracking the trails they leave while moving and visiting places (Hays et al., 2013; Jungherr & Jürgens, 2013; Wood et al., 2013). Moreover, the information available online can provide much better capabilities to assess the real consistency of tourist flows by analysing their activity in the social media space (Heerschap et al., 2014). This is a crucial task, because today we still rely on 'traditional' counting of visitors coming to accommodation establishments, while the phenomenon of *alternative accommodations* (private houses, couchsurfing, farms, religious establishments etc.) is developing fast, bringing tourists that are not fully accounted for, but must be serviced and satisfied for a good reputation of companies and destinations.

Obviously, there are many issues in this practice and many highlight the need for a rigorous approach to the analysis of social media data in order to avoid misinterpretations and pitfalls (the extensive review of Bonchi et al. (2011) provides a good discussion on these issues).

Big Data is the buzzword that in present times identifies the massive volume of both structured and unstructured data reputed to be easily available on the Web and difficult to process using traditional database and software techniques or by using traditional statistical methods. Actually, the problem is not much in the volume, as many other fields have faced this issue well before (astrophysics, particle physics, genetics, meteorology etc.), but rather in their fragmentation and variability, and in the need to combine structured and unstructured analysis techniques to extract meaningful outcomes. And, what is more, all these operations are often intended to be performed in a business and operational environment and not in an academic 'protected' laboratory, which adds an issue of speed in assessing situations that change quite rapidly.

Big Data is considered by many an incredible opportunity for its supposed capacity to provide answers to practically any question that could be asked about people's behaviours, views and feelings. As a matter of fact, it is rather surprising to see that a phenomenon once considered engendering disorientation and confusion, the so called *information overload*, once changed name into *Big Data* is now believed by many a kind of panacea, able to provide a wealth of useful and undeniable insights into many aspects of the modern life of individuals, organisations and markets (Mayer-Schönberger & Cukier, 2013; McAfee et al., 2012).

Many of these claims are more than justified and, actually, the capability to well unravel complex phenomena by tapping and combining all the available sources of information is an extraordinary advantage for those who can exploit fully the possibilities that are available today (Bedeley & Nemati, 2014).

However, besides the marketing buzz, a more careful and neutral analysis of the *Big Data* phenomenon highlights a number of issues, some of whom are well known in the academic environment, but may be not fully familiar to industry and practitioners. On these, the paper by Boyd and Crawford (2012) gives a good summary. As the authors note, many allegations of objectivity and accuracy may be misleading. Large data sets from online sources are often unreliable, and their dynamicity frequently prevents any attempt of replication of a study for confirmatory purposes. Moreover, errors and gaps can be magnified when multiple data sets are used together. The difficulty and cost for gaining access to *Big Data*, then, risks producing a new kind of digital divide between those who can afford the endeavour and gain the advantages and those who cannot avail themselves of the indications that can derive from such studies.

The large quantities of data available put under stress our conventional analysis methods. In absence of a very clear research objective and a likewise rigorous data collection plan, the risk of discovering meaningless effects or deceptive outcomes is quite high. One example is the recent revision to the Google Flu Trends indicator that has been found to be contaminated by a number of additional factors (Lazer et al., 2014). Large quantities of data also need a reconsideration of the statistical methods used for the analysis (Fan et al., 2014). When thousands of series of data are available, in fact, the probability to find a significant correlation between any two series, even if made of completely random numbers, can be as high as 90% (see for example Granville, 2013).

In summary, scholar and practitioners agree that there are remarkable benefits in having access to a vast amount of data that cover practically any aspect of human life, and in them, obviously, those related to their travels. But in order to achieve these benefits there is a need for a particularly high care in the handling of any investigation that uses information coming from the vastly populated online world. The first important point is, as well known, a clear setting of the objectives (the research questions), the second one, a direct consequence, the decision on the methods to be used for collecting the data needed.

1.1 Issues in data collection

Although many applications exist that provide some kind of controlled access to the wealth of data online social networks (OSNs) collect, when a specific investigation is required the main issue is in the possibility to gather a quantity of elements that can provide significant results. In these cases, collecting tweets, Facebook posts, reviews or similar elements can be a rather complicated task.

One possibility is to resort to a data provider such as Gnip (gnip.com), Datasift (datasift.com) or Topsy (topsy.com). This can be, however, a rather expensive solution, at least for a small company or destination, as prices are in the range of some tens of thousands euro. The second possibility is to use some applications developed for other purposes that however provide a plugin for downloading such data. Examples are NodeXL (nodexl.codeplex.com), a free network analysis add-in for Excel, or Gephi (gephi.github.io) an open source platform for visualisation and analysis of complex networks which comes with this type of additions.

Nonetheless, as also pointed out by many scholars and practitioners in informal conversations, all these solutions are not fully satisfactory for their intrinsic limitations in the amount of data they can handle, or in the type of information they can collect, that are functional to the providers' main scope, and not necessarily fully in line with what a *user* might need (Hansen et al., 2010). Therefore, a personalised approach may be required when some specific objective is to be accomplished.

An 'independent' set of programs for collecting the data needed can be put in place with not too much difficulty by using the widely available open source programs that access the APIs (application program interfaces) offered by practically all OSNs. These scripts are generally well documented and relatively easy to tailor for a specific need, and involve only a limited knowledge of computer programming (at the reach of a high school student in computer science). Their use, however, needs to take into account the constraints set by the APIs, that typically put a limit in the quantity of data that can be retrieved in a certain period of time (requests issued per hour), or on the topics that are available. When some of these limitations are exceeded, the IP address of the machine is blocked and the download terminates. Therefore, it can be advisable to have multiple machines working in parallel with different addresses. One more consequence of the time limits is that it may result in the immobilisation of hardware resources for long periods of time (days or weeks) if significant amounts of data must be collected, which is, obviously, a practice that many small organisations cannot afford.

2 The system

The system presented in this work is a possible answer to some of the issues described above. The objective is to make available a low-cost solution which is simple and flexible enough and does not require too specialised skills, so that a small organisation or company can set a personalised analysis or research program by deciding autonomously what questions to investigate and what data best fit the study objectives.

The system is a cluster of six Raspberry Pi machines, connected between them and with the Internet via an Ethernet switch and running a Linux-based operating system with a series of Python programs for accessing OSNs' APIs. This section describes the system and its hardware and software components.

2.1 The Raspberry Pi computer

Raspberry Pi (Fig. 1) is a credit card sized single-board computer developed in the UK by the Raspberry Pi Foundation (www.raspberrypi.org). Originally designed as an educational low-cost system (Severance, 2013), the Raspberry Pi, for its low cost, simplicity and flexibility has spurred an incredible global interest so that in the (little more than) two years of its life, more than three million machines have been sold, making this one of most diffused computers of all times.

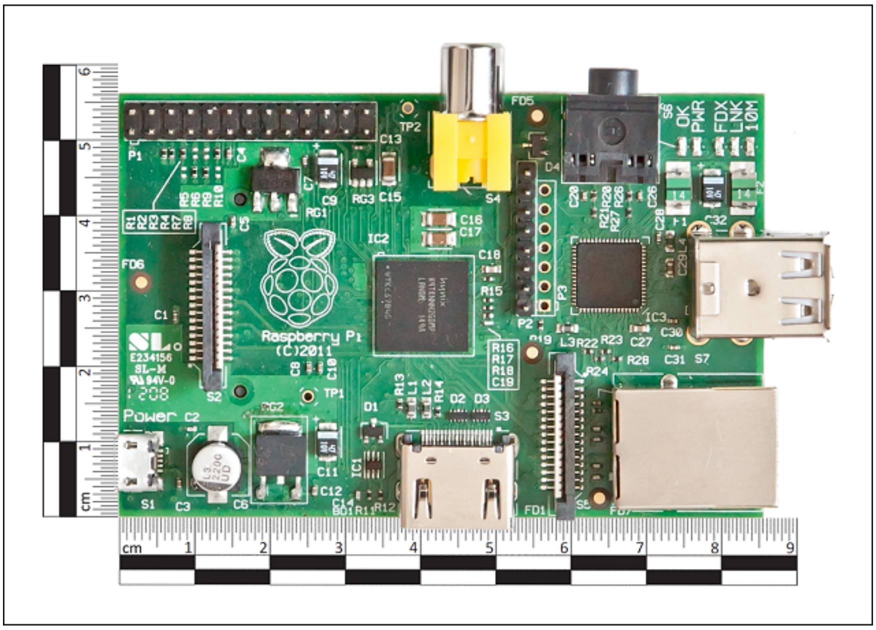

Fig. 1. The Raspberry Pi computer

The machine (Fig. 1) is powered by a Broadcom BCM2835 system-on-a-chip (SoC) multimedia processor with an ARM 1176JZF-S 700MHz processor and a VideoCore IV 24GFLOPS GPU. For storage it uses an SD Card (holding the operating system and all the software applications needed). The Raspberry Pi is available in two models: model A, with 256MB RAM, 1 USB port, and model B, equipped with 512MB RAM, 2 USB ports, and 10/100 Ethernet port. The device can be interfaced with external components via a set of GPIO (General-purpose Input/Output) and UART (Universal Asynchronous Receiver/Transmitter) connectors. Also available are an RCA and an HDMI ports (for external video), and a 3mm audio jack

The computer is powered by a DC 5V 1A input. The USB ports (model B) cannot provide more than a 100mA current, so devices using more than that are incompatible and for them a self-powered device or USB hub is needed (Upton & Halfacree, 2013). A Raspberry Pi model B costs $35 (about €30). A working installation (computer, SD card, case and power supply) can be purchased for less than €50. When energy consumption is considered, a single computer uses (average activity) about 2.5W, while the switch uses about 4W and the hard disk about 10W. For the whole cluster we therefore estimate a consumption of about 30W. The low heat production allows using it without any special air conditioning or ventilation requirements (it can also be closed in a drawer, if needed for security reasons).

On the software side, a number of popular Linux distributions have Raspberry Pi specific versions. The Foundation provides on its website a number of operating system images that can be freely downloaded. Raspbian is the default Linux OS (operating system) provided. It is based on a Debian ‘wheezy’ distribution, optimised for the specific hardware. It contains a large number of packages among which the most important piece is the Python programming language. In fact, Python is the core language around which the Raspberry Pi was built (actually that is the meaning of Pi in the name).

2.2 Cluster architecture

The cluster is composed of six Raspberry Pi model B machines connected via an Ethernet switch (Fig. 2).

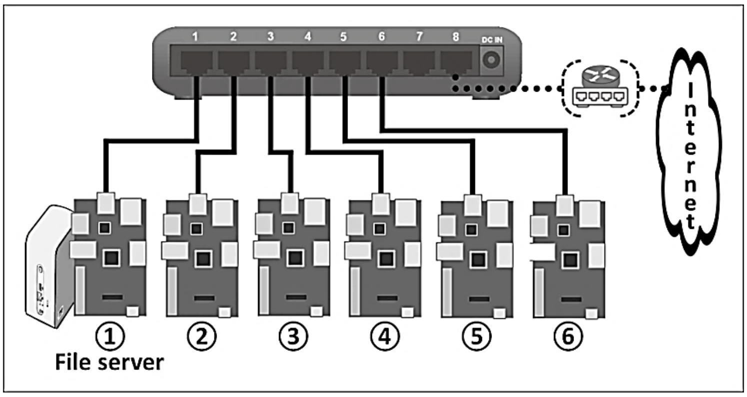

Fig. 2. The Raspberry Pi cluster schematic diagram (the router shown is optional and dependent on the specific network setting)

The rack holding all the components together was assembled by using Lego bricks (Fig. 3), loosely following what done by the Glasgow team for their PiCloud system (Tso et al., 2013).

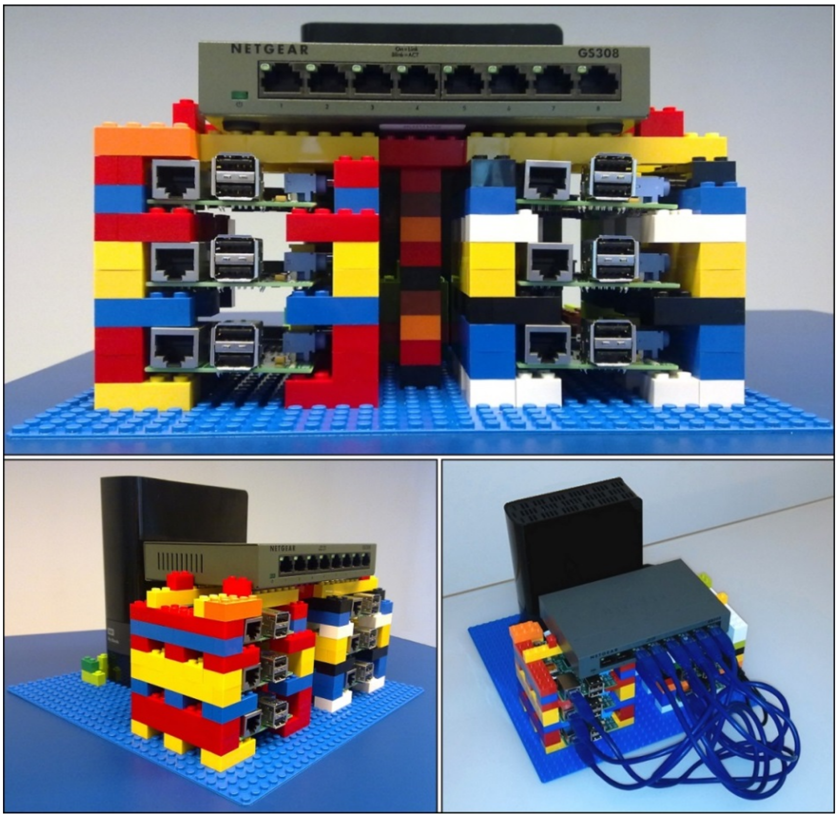

Fig. 2. The Raspberry Pi cluster

One of the machines (see Fig. 2) is equipped with a 2TB USB powered hard disk and acts as a file server for the cluster, accumulating into a series of SQLite databases all the data collected. The databases are specialised by OSN in order to allow the maximum flexibility for the analysis and reduce the load on the different machines.

All machines are given a static IP address and, in our installation, access the Internet via the main University gateway. In cases in which a similar configuration is not possible, a router is needed to connect the cluster to the Net.

2.3 Software components

All the machines have a similar configuration. Besides the Raspbian distribution, a Samba system was installed. Samba (www.samba.org) is an open source suite that provides file and print services, the clients, and the file server, and is interoperable across different operating systems so that the files can be seamlessly transferred to a Windows or a Mac machine for more complex processing.

The software programs needed for downloading data from the OSNs (online social networks) were written using the Python programming language. Python is a general-purpose, high-level programming language whose design philosophy emphasises code readability. It has a relatively simple and compact syntax, thus allowing fast development of applications. Despite being an interpreted language, Python is highly efficient and the execution speed of its programs is, in many cases, comparable with that of a compiled language (Cai et al., 2005). Its open source philosophy has made possible to count on a great number of packages and libraries. Python installers are available for all environments (see www.python.org).

For our purposes, a quite large number of programs is available that allow using the APIs of many social media platforms (see a list at: www.pythonapi.com). This makes easier the task of accessing a specific platform (Russell, 2013). Moreover, where textual analysis would be needed, the Natural Language Toolkit (NLTK) provides a set of highly efficient functions (Bird et al., 2009).

In general our scripts for downloading data from an OSN were built by following the diagram shown in Fig. 4.

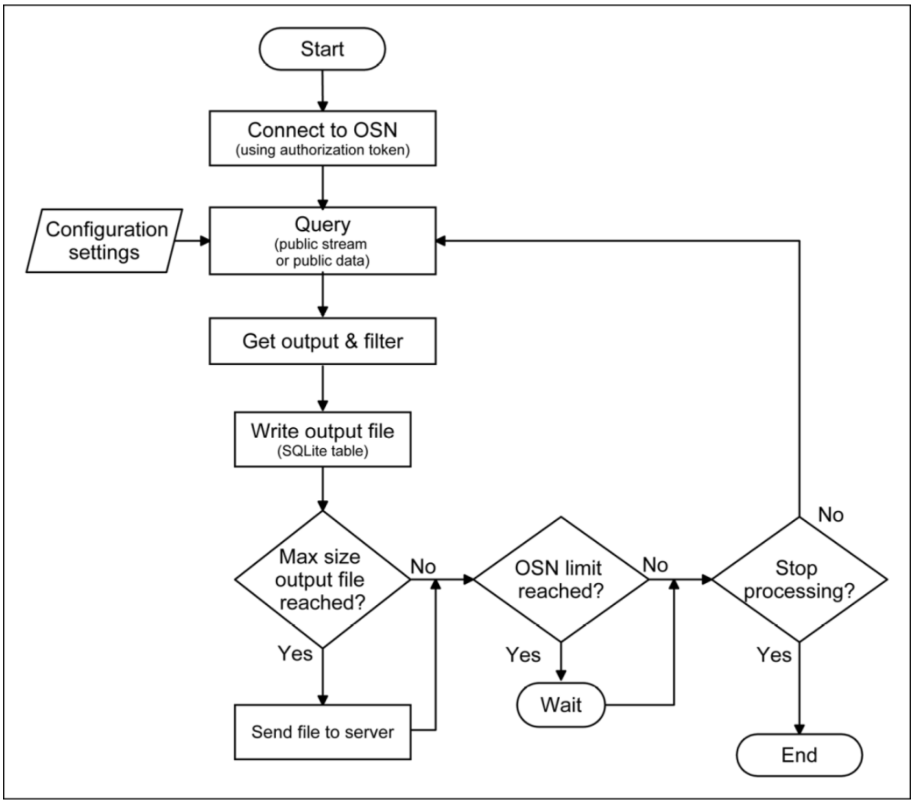

Fig. 4. The Raspberry Pi cluster schematic diagram (the router shown is optional and dependent on the specific network setting)

For all the cases, a preliminary step consists of acquiring an authorization token that is used to access the platform and its API (instructions are given in each set of Python libraries or on the developers section of the OSN considered).

All the scripts take into account the different restrictions imposed by the OSNs, that typically consist of a limit of the number of requests accepted in a certain time interval. Moreover, in order not to overload the single machines, a size limit was set for the output files. Once reached this size the files are transferred to the file server. Periodically (e.g. once a week in a ‘long run’ task) all machines are stopped, and a script on the file server merges the different files into a single database.

The file server machine also provides a PHP web application (presently under development) that allows managing the different jobs on the other computers in the cluster and the configuration files needed by them (e.g. the OSN to be used, or lists of objects, tags etc. to be examined).

It is important to note here that the whole system (all hardware components, cables and power supplies) has a total cost of about € 400.00, with operating costs (energy requirements) in the range of a couple of euro per week, which is a reasonable investment even for a very small company.

3 A demonstration: geolocating people in a tourism destination

The whole system was tested by executing a simple task: the geolocation of people in a destination. The city of Lugano (CH) was chosen as centre of a 5 km radius area. The cluster was employed for downloading data from Facebook, Twitter, Foursquare and Instagram. For each source all posts (tweets, check-ins, pictures) carrying a geocoded tag (in the area of interest) were selected. The resulting items were then aggregated into a list of positions (latitude and longitude) and weighted by the number of elements referring to each position. The list was then used as input for the Heatmap.js script by Patrick Wied (www.patrick-wied.at/static/heatmapjs). Heatmap.js is an open source JavaScript library that can be used for visualising geocoded data in real time by using an HTML5 canvas element to draw heatmaps.

In a heatmap three dimensional data are used where two dimensions represent Cartesian coordinates (x and y values, latitude and longitude in our case) and the third one is used for showing the intensity of a data point. The intensity is rendered as a colour, usually red (hot) for the maximum and blue (cold) for the minimum. The script produces a layer which can be superimposed on a map. OpenStreetMap (www.openstreetmap.org), the collaborative free editable map of the world was used as a basis.

On another attempt we put coloured markers (each marker represents the source of data: round solid = Facebook, round light = Twitter; square light = Instagram; square solid = Foursquare) on a Google map (using a simple javascript code calling its API). The results are shown in Fig. 5 and 6.

In a time period of two weeks a total of 2288 points of interest (POIs, points referenced by at least one source) were collected, with a total of 19378 mentions, split as shown in Table 1.

Table 1. Points of interest (POIs) and mentions collected for the different OSNs

OSN	POIs	Mentions
Facebook	291	12923
Foursquare	15	57
Instagram	1958	4739
Twitter	24	1659
Total	*2288*	*19378*

Looking at Table 1 it is interesting to note how (in the period under consideration) Instagram seems to be the most used tool, followed by Facebook. Twitter shows a lower utilization, which is probably due to the fact that most tweets do not carry geocoded information. Foursquare is used very little.

Although quite simple and straightforward, this exercise can be of great interest for a destination manager or a tourism stakeholder as, if carried on at regular intervals, is able to provide a faithful representation of the number (at least a significant sample) and the presence of active social media users in an area. This can be used to better inform plans, strategies and actions, for example by helping in the choice of the platform to use more extensively in order to engage users.

A number of other parameters can be recorded (when available due to privacy settings) about the individuals sampled, such as place of origin (which allows distinguishing between visitors and locals), age, preferences, places visited before the trip to the destination or even intentions to go somewhere else or to perform some activity, and so on. Moreover, a semantic analysis of the text collected can provide, as known (Cambria et al., 2013; Ko et al., 2013), useful indications about the attitude of the users with respect to the different POIs. The whole set of analyses, as said, can be performed on data specifically collected for a particular objective.

One more consideration is in order. The number of useful points recovered may seem small, if compared with the claims of hugeness of the data available. This is a demonstration of the fact that when specific objectives for an investigation are set, things can be not as easy or fast as it might be inferred by the buzz around *Big Data*.

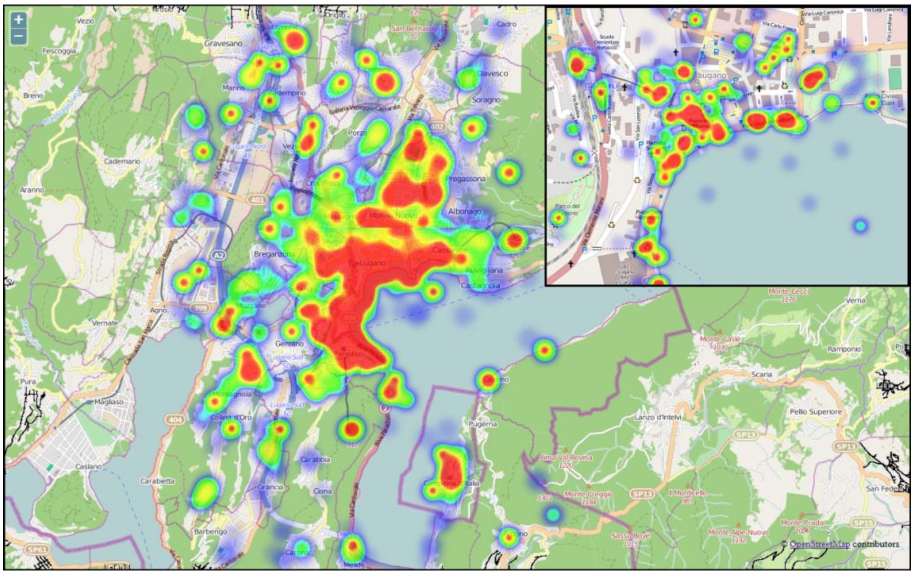

Fig. 5. Heatmap of social media users in Lugano (CH). Inset shows the most central area of the city

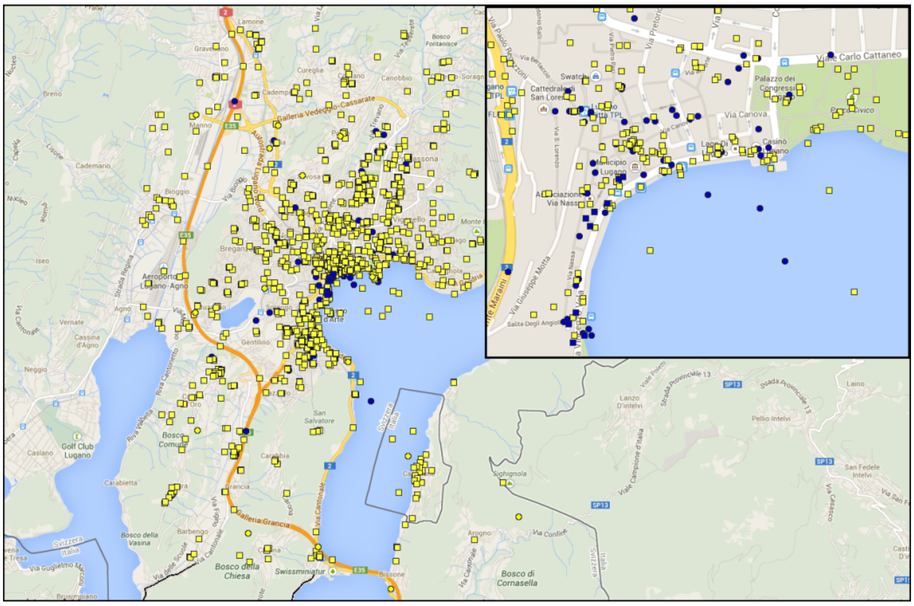

Fig. 6. Markers map of social media users in Lugano (CH). Markers represent the source of data: round solid = Facebook, round light = Twitter; square light = Instagram; square solid = Foursquare. Inset shows the most central area of the city

4 Concluding remarks

When considered critically and rationally, *Big Data* is a worthy way to better explore the highly dynamic tourism phenomenon. By accessing the huge quantity of trails left behind the online activities of billions of individuals, it is possible to distil valuable information on practically every aspects of interest for researchers, practitioners and managers. Coupled with more traditional investigation methods, the analysis of such information can turn out to be of fundamental importance for a better understanding of the different phenomena connected with the tourism world or as an aid in informing more effective choices (Jungherr & Jürgens, 2013). However, despite the apparent universal availability of the data in question, actually collecting those needed for a specific objective can be, from a practical point of view, a tricky task. As discussed, issues of time cost and competence may be difficult to overcome by the smallest companies and organisations.

The Raspberry Pi cluster presented here is an affordable solution. Its architecture, combined with the usage of open source software makes it an ideal candidate for the job of retrieving data from the wide array of online social platforms. The cluster has a number of advantages with respect to conventional arrays of computers. It is a low-cost solution, with a low total power consumption, easy portability (due to its small size and weight) and easy scalability. The only specialist skill needed is the one required for the personalisation of the programs used. But this issue can be easily solved even by a small organisation by taking advantage of external programmers or computer science students.

The cluster is now operational and we look forward to employing it in a series of research programs on the most important issues in the usage and the value of the contemporary online technologies and in the opinions of users on several issues.

References

Bedeley, R., & Nemati, H. (2014). *Big Data Analytics: A Key Capability for Competitive Advantage.* Paper presented at the 20th Americas Conference on Information Systems (AMCIS), Savannah, GA, 7-9 August.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media Inc.

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology, 2*(3), art.22.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662-679.

Cai, X., Langtangen, H. P., & Moe, H. (2005). On the performance of the Python programming language for serial and parallel scientific computations. *Scientific Programming, 13*(1), 31-56.

Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Big social data analysis. In R. Akerkar (Ed.), *Big Data Computing* (pp. 401-414). Boca Raton, FL: Chapman and Hall/CRC.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review, 1*(2), 293-314.

Granville, V. (2013). *The curse of big data*. Retrieved June, 2014, from http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, MA: Morgan Kaufmann.

Hays, S., Page, S. J., & Buhalis, D. (2013). Social media as a destination marketing tool: its use by national tourism organisations. *Current Issues in Tourism, 16*(3), 211-239.

Heerschap, N., Ortega, S., Priem, A., & Offermans, M. (2014). *Innovation of tourism statistics through the use of new big data sources.* Paper presented at the 12th Global Forum on Tourism Statistics, Prague, CZ, 15-16 May. Retrieved July 2014 from http://www.tsf2014prague.cz/assets/downloads/Paper%201.2_Nicolaes%20Heerschap_NL.pdf.

Jungherr, A., & Jürgens, P. (2013). Forecasting the pulse. How deviations from regular patterns in online data can identify offline phenomena. *Internet Research, 23*(5), 589-607.

Ko, H. G., Ko, I. Y., Kim, T., Lee, D., & Hyun, S. J. (2013). Identifying User Interests from Online Social Networks by Using Semantic Clusters Generated from Linked Data. In Q. Z. Sheng & J. Kjeldskov (Eds.), *Current Trends in Web Engineering* (pp. 302-309). Berlin: Springer

Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(14), 1203-1205.

Leung, D., Law, R., van Hoof, H., & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing, 30*(1-2), 3-22.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data. The management revolution. *Harvard Business Review, 90*(10), 61-67.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Sebastopol, CA: O'Reilly Media Inc.

Severance, C. (2013). Eben Upton: Raspberry Pi. *Computer, 46*(10), 14-16.

Tso, F. P., White, D. R., Jouet, S., Singer, J., & Pezaros, D. P. (2013). The Glasgow Raspberry Pi Cloud: A Scale Model for Cloud Computing Infrastructures. *Proceedings of the IEEE 33rd International Conference on Distributed Computing Systems Workshops (ICDCSW), Philadelphia, PA, USA, 8-11 July*, 108-112.

Upton, E., & Halfacree, G. (2013). *Raspberry Pi user guide*. Chichester, UK: John Wiley & Sons.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific reports, 3,* art.2976.

Acknowledgements

The authors wish to thank the staff of the CERMES Center and of the Bocconi IT services for their help and support during the project.