Conference Paper · June 2018

DOI: 10.1109/ISCC.2018.8538622

CITATIONS
2

READS
100

5 authors, including:

Eliza Gomes
Federal University of Santa Catarina
18 PUBLICATIONS 57 CITATIONS
SEE PROFILE

Daniel Penz
SENAC SC; CESUSC
26 PUBLICATIONS 25 CITATIONS
SEE PROFILE

Viviane Etges Gomes
Universidade do Estado de Santa Catarina
2 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Carlos Roberto De Rolt
Universidade do Estado de Santa Catarina
48 PUBLICATIONS 135 CITATIONS
SEE PROFILE

Evaluating the tools to analyze the data from the ParticipACT Brazil Project: A test with Elasticseach Tools Ecosystem with Twitter data

Eliza Gomes
Department of Informatics and Statistic (INE)
Federal University of Santa Catarina (UFSC),
Florianopolis, Brazil
E-mail: eliza.gomes@posgrad.ufsc.br

Daniel Penz, Viviane Etges Gomes and Carlos Roberto De Rolt
Faculty of Administration and Economic Science (ESAG)
State University of Santa Catarina (UDESC),
Florianopolis, Brazil
Email: daniel.penz2017@edu.udesc.br
Email: viviane.gomes@edu.udesc.br
Email: rolt@udesc.br

Mario Dantas
Department of Computer Science (DCC),
Federal University of Juiz de Fora (UFJF),
Juiz de Fora, MG, Brazil
E-mail: dantas@ice.ufjf.br

***Abstract*—This article aims to present tests of one data analytics engine in order to evaluate the efficiency and practicality of the chosen tools (Elasticsearch ecosystem - ELK), based on data collected from Twitter. The study is motivated by the need to choose data analysis tools appropriate to the Participact Brasil project where scientific research is carried out on urban problems in smart cities. We can conclude that ELK ecosystem provides a set of efficient and fast tools. However, despite the large amount of technical documentation, the installation and use of the tools want the monitoring of specialists, besides requiring a more robust computing environment than the conventional one.**

I. Introduction

The intensification of urbanization and the intense dissemination of communication and sensing technologies create new opportunities for the application of knowledge in order to improve the quality of life in the cities. Voluntary and collaborative data collection using citizens smartphones in a crowdsensing system associated with a big data platform characterizes the computing environment used at this work. It is composed of data collection, cataloging, storage, organization, analysis and publication tools.

At the last months we are trying to use a large amount of data, and during the process we faced some challenges who are related to scalability, indexation, functionality, cost reduction, performance and some others some of the them faced by other researchers [1], [2], [3], [4], [5], [6]. Therefore, this article reports our experience in the definition and tests of computational tools to implement a data analytics platform to study urban problems in smart cities in scope of ParticipACT Brazil project.

Langi et. al [2] define Elasticseach, Logstash and Kibana (ELK) as a reliable data storage, search engine, analyses and visualization tool set. On the other hand, we have some social networks worldwide like twitter that holds an important role in information distribution.

Therefore, this article focus is in data analytics step of Participact computational plataform and involves the evaluation of the feasibility of using ELK - Elasticsearch, Logstash, and Kibana, since its installation, ease of use, researcher training and integration with others tools by a group of researchers on the social sciences area. The tests were carried out from data collected from Twitter, of a specific region and period in order to know the subjects of interest of those users. The focus is on understanding how the ELK works and not on the behavior of social network users.

Experience has shown that installing and using the ELK requires expert monitoring and computational power not present on most desktops. It is recommended to combine in the data analysis laboratory, complex and powerful tools with others that have more friendly features and interfaces so that the culture of data analysis can be disseminated, especially among researchers in the social sciences area.

In Section 2, we describe the ParticipACT Brazil project and the data analysis module step. We present related works in Section 3. Section 4 presents our article proposal. In Section 5, we present a case study carried out to test the ELK. Our conclusions and directions for future works are presented in Section 6.

II. ParticipACT Brazil

ParticipACT Brazil [7] is an extension of ParticipACT [8] crowdsensing platform of the University of Bologna. This project has been developed by State University of Santa Catarina (UDESC) in partnership with Federal University of Santa Catarina (UFSC) and has financial and data support of public and private entities. The project is being developed in Florianopolis city, capital of the state of Santa Catarina, Brazil. The aim of ParticipACT Brazil project is to use data from utilities companies and crowdsensing campaigns to conduct

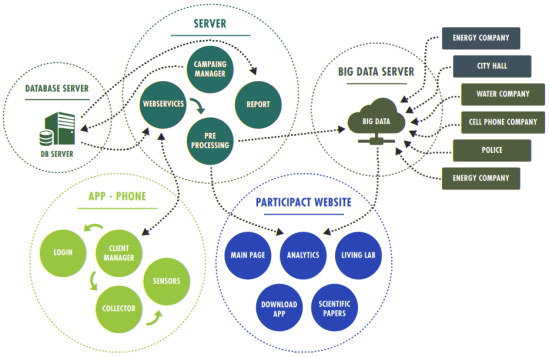

Fig. 1. ParticipACT Brazil Infrastructure [10]

researches and studies and eventually to direct managers to resolve urban problems. Also we make possible the creation a open data portal [9] which can be used by the anyone who want to help improve and discover more about the Florianopolis people habits and make inferences on how we can live better on society.

ParticipACT Brazil infrastructure, as shown in Figure 1, is divided into three main components: Big Data, Crowdsensing and Website.

- *Big Data*: consists of database server and big data server. Database server is responsible for storing the information captured by crowdsensing. Big data server integrating different databases from crowdsensing cam-paigns and from different government agencies and re-search institutes to provide a detailed study of urban problems. The idea is provide information to legitimize and justify the decision making of public and private managers directing them to an intelligent management based on scientific research.
- *Crowdsensing*: consists of server and app - phone. Server is responsible for controlling the crowdsensing campaigns. App - Phone collects data through campaigns with the cooperative and voluntary participation. The crowdsensing is accomplished through smartphones with access to a ParticipACT Brazil App for data collection. The goal of managing a crowdsensing campaign is coordinating a group of people to collect a certain type, and maybe complex, of data. Crowdsensing participants can access information collected from the campaigns by a website.
- *Living lab website*: presents a smart city portal that is to the user to interact with the ParticipACT Brazil system. In this website it is possible that the user views the results of analysis, download the application, and others.

The Figure 2 shows the operation project as a continuous cycle. We can start this cycle receiving data from external sources such as public or private companies and crowdsensing campaigns (Data Sources and Crowdsensing Platform). This data is stored without processing or manipulation in a data repository (Data Gathering) and presented to the researcher public through the data portal platform [9]. Then, the data is consumed by Big Data platform for processing and storage

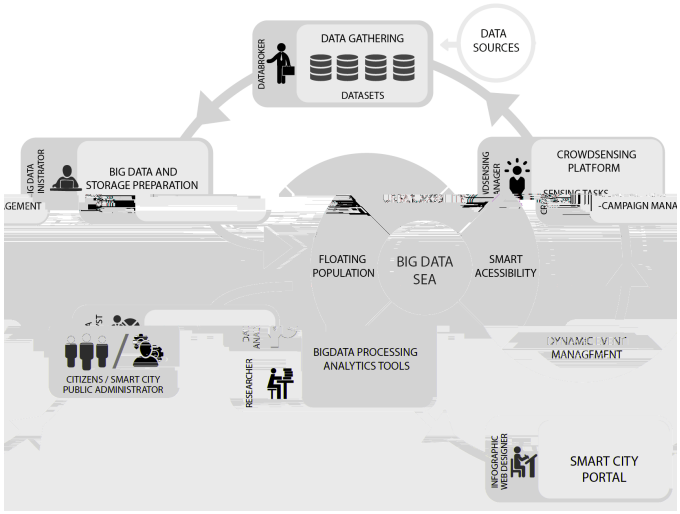

Fig. 2. ParticipACT Brazil Operating Cycle [11]

(Big Data and Storage Preparation). The data can be processed and manipulated in order to solve the problems proposed by

problems.

- *Crowdsensing Manager*: responsible for setting up and publishing of crowdsensing campaigns.

All the described cycle of the ParticipACT Brazil is performed for the purpose of, initially, resolve four challenges, which are:

- *Urban Mobility*: traffic jam is a problem present in big cities. The goal of the project is to present evidence to explain the reason of this problem in order that public managers can resolve it intelligently.
- *Floating Population*: in tourist cities such as Florianopolis, have a wide variety of the amount of people present in the city throughout the year. This variety can lead to problems such as urban mobility, supply of water and energy. For this, the project aims to estimate more accurately the number of tourists that visit the city, so it is possible prevent problems that increasing population can generate.
- *Dynamic Event Management*: the goal is to develop a system to dynamically manage emergency events in a city, identifying the presence on the spot of people having skills that can be applied to specific tasks. For example, if a person is hit by a car is possible to know that a doctor is a few meters from the incident and request, via smartphone, the help of this expert.
- *Smart Accessibility*: the objective of this project is to create an interactive map able to inform which places have accessibility conditions for people with physical disabilities.

III. Related Works

Some related works were found about the implementation of the Elasticsearch ecosystem.

Zheng et. al [1] emphasize the perspective of cloud computing. The article goal is on the progress of storing and searching for space-time data for this, it has been used a platform based in Proxmox VE tool and the open search engine: Elasticsearch (both in the bias of virtualization). For the attainment of conclusions, it was made an experiment, realized using AIS data, which lead us to the powerful combination among virtualization and Elasticsearch in indexing and storing the time and location data with efficiency and high reliability.

The research of Langi et. al [2] claims that social media analysis can show the rating of a service or a product from the user's perspective, in this case Twitter customer's. The article gives a view of the Elasticsearch tool, that analyze big data. The goal is to compare two ways of inputting twitter data to Elasticsearch (Twitter River and Logstash), which is important in the output of the system once accuracy and efficiency of input and output are major features to support systems of big data. The discoveries of this research highlight the average of CPU process, average of RAM and average of Disk used, among Twitter River and Logstash.

Bai [3] research has as center point the log events generated by modern enterprises, the pattern of the this data, can mine business value. The article present a method of data search in real time using the Flume as the log collect events, the Elasticsearch and the Hbase, as a collecting data tool, highlighting the second one as a viable instrument of search for large data logs.

The focus of Bhadane et. al [4] research, is to optimize an intelligent algorithm and add it to a software platform that tab medications, intending to retrieve a list of similar medicines according to a keyword. To this end, it has been used the Elasticsearch as an indexing tool. And however it is a efficient non conventional database, that works in real time, it would need some improvements in the user's side.

The article of Bagnasco et. al [5] aims to monitoring system to inspect the site activities both in terms of IaaS and applications running on the hosted virtual instances. Using the Elasticsearch, Logstash and Kibana stack.

Kononeko et. al [6] relate in their research the problems of generate, process, and retain data and seek ways to analyze it effectively. The insights derived from these large data sets and aim to better understand how to solve hard problems and gain competitive advantage. The research conclude this data is so fast-moving and voluminous, it is increasingly impractical to analyze using traditional offline to make analysis, read-only relational databases and make suggestions, to new big data technologies and architectures, including Hadoop and NoSQL databases, who have evolved to better support the needs of organizations analyzing such data. In particular suggests, Elasticsearch as a distributed full-text search engine. As conclusion, found positive relation between Elasticsearch and scalability issues, big data search and performance. Also concludes that the relational databases were simply never designed to support this kind of research.

Summarizing positive points and limitations we can consider:

- Among these works for big data analysis, the Elasticsearch ecosystem works as an integrated set of tools who can generate scalability (uniform and automatic distribution of indexes shards across the cluster nodes), corroborate the findings of [3], [6], [2] and agility (refreshes independently with constant indexes actualizations) as we can find in the works [3], [6], [1], [2] and [4].
- Within the works we can find applicability of the ecosystem in any interface, the self-configuration as a complete database [6] and the functionality of searcher and of indexer [3], who we can consider also good aspects.
- Some limitation has been found, like the necessity to improve the capacity of the system process information, performance and the slowly learning curve like describes [6], in addiction have been found some issues about the search mechanism who were found by [3] and issues about regarding the memory issues [4].
- The cost reduction through equipments, time, licenses, and some others, was the main focus of the use of the Elasticsearch Ecosystem, but after a evaluation some economies but less than we really had in mind [4].
- The several studies are conclusive, remains issues, but the researches [3], [1], [6], [2], [5] findings evidence with

the ones found in our study when relating the excellent performance of Elasticsearch Ecosystem in analysis of big data.
- Other important points are the open source solutions and the applications who are built to handle data that needs to be processed and analyzed in a rapid personalized manner. Through them we gave the possibility, to each developer, the capacity to work with a set of personalized views [3], [1], [6], [5] and [4].

IV. Big Data Platform

Gomes et. al [11] paper propose a model and an architecture of big data platform for ParticipACT Brazil project, composed of five layers, as can be seen in the Figure 3.

- *Layer One*: consists of data provided by utilities companies to be stored in the project database. Different sources have different data types and formats.
- *Layer Two*: is responsible for the integration and conversion of received data into a single format. To carry out this work we chose the Pentaho data integration (PDI) tool [12]. PDI provides the extraction, transformation and loading (ETL) engine to capture, clean and store data using a uniform format.
- *Layer Three*: represents the data storage. The storage process is performed using Hadoop Distributed File System (HDFS) [13]. HDFS is open-source and has the capacity to carry gigabytes per second and has petabytes storage capacity. Splits the files into multiple chunks that are stored on different servers, which increases the fault tolerance due to this replication [14].
- *Layer Four*: is responsible for the data processing. Apache Spark [15] is used to perform the work of processing the data in a fast and continuous way, since it is a data stream. Spark suports in memory computing which enables fast computation.
- *Layer Five*: consists of the search and analysis of data. With this layer is possible to create dashboards to graphically view the studies carried out with the received data. Elasticsearch ecosystem, also known as ELK Stack (Elasticsearch, Logstash and Kibana) [16].

Thereby, this article aims to present tests of the data analytics engine in order to evaluate the efficiency and practicality of the chosen tools (Elasticsearch ecosystem), based on the environment and the proposal of the ParticipACT Brazil project.

In the next subsection we detail the concepts and characteristics of the Elasticsearch ecosystem chosen for performing the data analysis.

A. Elastisearch Ecosystem

For analysis and search of data we used Elasticsearch ecosystem (Elasticsearch, Logstash and Kibana). As we can see in Figure 4, Logstash collects and transforms logs, Elasticsearch searches and analyzes and Kibana enables the visualization and management [17].

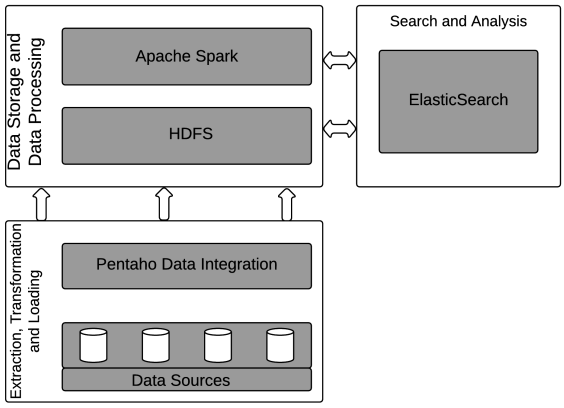

Fig. 3. Architecture for Big Data Platform of ParticipACT Brazil [11]

Fig. 4. ELK Stack

1) Elasticsearch: is an open-source distributed, RESTful search and analytics engine whose goal is to store, search and analyze a great amount of data with a lower latency. Elasticsearch is used for full-text search, structured search, analytics [18]. It is written in Java and uses Apache Lucene [19], a full-text search engine library, for indexing and searching.

2) Logstash: is an open source data collection engine with fast pipelining capabilities [20]. It is used to collect, parse and analyze a great quantity of unstructured and structured events and data to a central service. The key features of Logstash are: centralized data processing, support for custom log formats and plugin development.

3) Kibana: is an open source Apache 2.0 licensed data visualization platform which analyzes and views of structured and unstructured data stored in Elasticsearch, that is designed to work with ElasticSearch. Kibana is written in HTML and JavaScript and uses the search and indexing capabilities of Elasticsearch to display powerful graphics for the end users. It is possible to perform advanced analytic and view data through a variety of histograms, geomaps, pie charts, graphs and tables [21].

V. Implementation and Experimental Results

To test the Elasticsearch ecosystem we performed a case study with the objective of evaluating the functioning of the tools. In the next subsections we detailed the case study, an analysis of the data captured and pointed some challenges encountered and learning during the installation process and use of the tools.

We installed Elasticsearch ecosystem in a machine with operation system Windows 7, 8GB of memory and AMD Phenom(tm) II X4 B93 processor.

TABLE I
GEOGRAPHIC COORDINATES USED AS A PARAMETER TO CAPTURE TWEETS

Florianopolis Neighborhood	Latitude	Longitude
Saco dos Limoes	-27.604255	-48.524764
Itacorubi	-27.579362	-48.504569

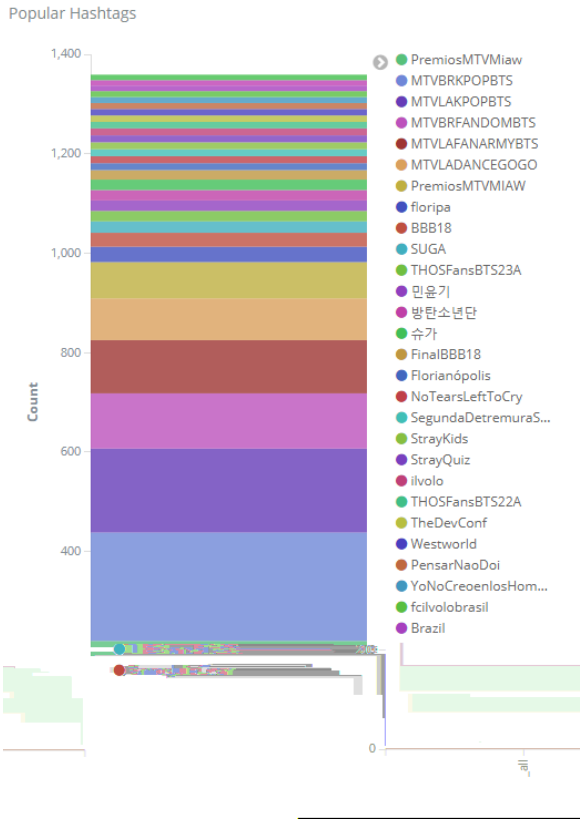

Fig. 5. Hashtags Most Used

A. Case Study

For the tests, we chose to capture twitter data from a specific geographic region in the city of Florianopolis. The search period was from April 20 to 23, 2018. During this period, 16,446 tweets were captured. The Table I presents the geographic coordinates used as parameter to capture tweets. We did not used any keywords, in other words, we did a free search, only with territorial restriction.

The choice of geographic area for data collection was random since the objective was to test the ELK stack.

B. Data Analysis

With Kibana tool, we performed some analysis with the data that we obtained. Figure 5 shows the 30 hashtags most used during the analyzed period. We can see that the first 7 hashtags most cited relate to a MTV award.

On the other hand, Figure 6 presents the 25 keywords most popular. We can verify that the keyword *Bom dia* (good morning) and its variations (*B dia*, *Bom diaa*, *Bom Dia*, *bom dia*) are the most appear.

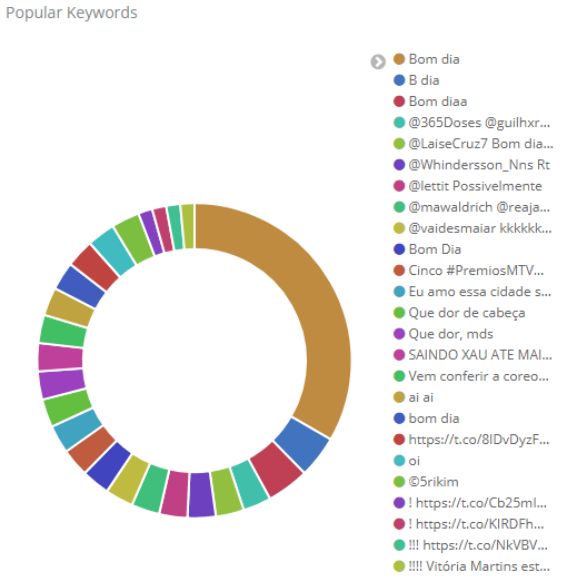

Fig. 6. Keywords Most Used

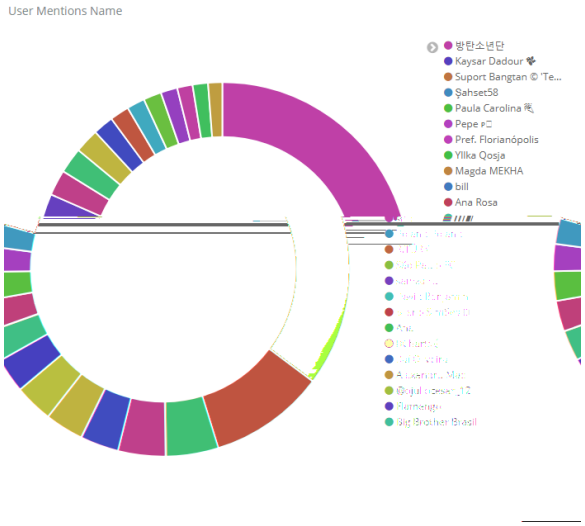

Fig. 7. Most Mentioned Users

Finally, Figure 7 shows the 25 users most mentioned. The first user most mentioned is BTS, a South Korean musical group.

We can see a relationship between the hashtags most used and users most mentioned, since it is a music award and a music group, respectively.

The results of the data collected and analyzed only confirm that we succeeded in the installation of the computational tools for the collection, analysis of data and generation of graphs. The data were classified in order to allow the counting of subjects preferred by twitter users in that period.

C. Experience and Lessons Learned

During the installation and use of the Elasticsearch ecosystem, we can find some advantages like: Elasticsearch ecosystem is a multiplatform system and can be installed in Windows, Linux and MacOS; there is a wealth of technical documentation that explains how to install and use the tools, considering that these are open source tools; and no costs of licenses.

In other side, the large amount of tutorials, videos and blogs that explain how to install and configure ELK tools makes this process easier, but still complex. This complexity is due to the need to install Java and configure it as environment variable, the creation of configuration files for the correct operation of Logstash, as well as the need to use the command terminal of the operating system to the installation of tools, an uncommon procedure for users with basic knowledge in computing.

The activation of the social data collection process requires a series of steps for its configuration. From the registration and the accreditation in Twitter, to use Google Maps to find the coordinates of the geographic area to be monitored. In addition, we must create configuration files and templates in JSON format for these geographic coordinates to be considered at the time of the search.

All these steps and specificities in the installation of the tools require a good understanding in technical documents and computational skills, which makes the process of installing and configuring the tools difficult for researchers with few experience in the computational area.

In addition to software and human issues, the Elasticsearch ecosystem consumes a lot of memory that makes it difficult to install on simple and common machines with 4GB of RAM.

In academics environments, instructors spend a lot of time preparing the tools and teaching students how to use them. This inertial period only compensated later after the learning curve overcome when a reasonable level of productivity is obtained.

VI. Conclusion

In this article, we presented tests carried out with Elasticsearch ecosystem (Logstash, Elasticsearch and Kibana) with the objective of evaluating the tools and corroborating our chose. The tests consisted in analysis of data captured from Twitter during a period and geographic region pre-established.

As result, we can to verify that Elasticsearch ecosystem provides efficient tools. Since Logstash and Elasticsearch provide fast and quality search and analysis of data and Kibana offer robust view with graphs, maps and tables. These features found in the Elasticsearch ecosystem meet the expectations of search and analysis of data of the ParticipACT Brazil project, in particular with regard to licensing costs.

However, a considerable effort will have to be made to prepare teaching material and training of students and researchers in more advanced stages. The data analysis laboratory should be pre-prepared and receive intensive care of technological upgrading of the tools, increase processing capacity, RAM of the workstations and storage in the cloud. The use of more user-friendly tools, such as Tableau, for beginners in the discipline of data analysis is an alternative that is being considered because it makes the training process more productive.

Our next work involves the analysis of fake news in social networks to define in a function its life cycle, since the first posting, identification of its falseness and resilience.

References

[1] Y. Zheng, F. Deng, Q. Zhu, and Y. Deng, "Cloud storage and search for mass spatio-temporal data through Proxmox VE and Elasticsearch cluster," *CCIS 2014 - Proceedings of 2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, pp. 470–474, 2014.

[2] P. P. Langi, Widyawan, W. Najib, and T. B. Aji, "An evaluation of Twitter river and Logstash performances as elasticsearch inputs for social media analysis of Twitter," *Proceedings of 2015 International Conference on Information and Communication Technology and Systems, ICTS 2015*, pp. 181–186, 2016.

[3] J. Bai, "Feasibility analysis of big log data real time search based on Hbase and ElasticSearch," *Proceedings - International Conference on Natural Computation*, pp. 1166–1170, 2013.

[4] C. Bhadane, H. A. Mody, D. U. Shah, and P. R. Sheth, "Use of Elastic Search for Intelligent Algorithms to Ease the Healthcare Industry," *International Journal of Soft Computing and Engineering*, vol. 3, no. 6, pp. 222–225, 2014.

[5] S. Bagnasco, D. Berzano, A. Guarise, S. Lusso, M. Masera, and S. Vallero, "Monitoring of IaaS and scientific applications on the Cloud using the Elasticsearch ecosystem," *Journal of Physics: Conference Series*, vol. 608, no. 1, 2015.

[6] O. Kononenko, O. Baysal, R. Holmes, and M. W. Godfrey, "Mining modern repositories with elasticsearch," *Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014*, pp. 328–331, 2014.

[7] P. Brasil, "Participact brasil," April 2018. [Online]. Available: http://labges.esag.udesc.br/participact/

[8] ParticipACT, "Participact," April 2018. [Online]. Available: http://participact.unibo.it/infoen/

[9] Catalog, "Participact brazil data catalog," April 2018. [Online]. Available: https://dados.esag.udesc.br

[10] E. Gomes, M. A. Dantas, D. D. de Macedo, C. De Rolt, M. L. Brocardo, and L. Foschini, "Towards an infrastructure to support big data for a smart city project," in *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2016 IEEE 25th International Conference on*. IEEE, 2016, pp. 107–112.

[11] E. Gomes, M. A. Dantas, D. D. de Macedo, C. De Rolt, J. Dias, and L. Foschini, "An infrastructure model for smart cities based on big data," *International Journal of Grid and Utility Computing*, in Press.

[12] Pentaho, "Pentaho data integration," April 2018. [Online]. Available: http://www.pentaho.com

[13] HDFS, "Hadoop distributed file system," April 2018. [Online]. Available: https://hadoop.apache.org

[14] W. Tantisiriroj, S. W. Son, S. Patil, S. J. Lang, G. Gibson, and R. B. Ross, "On the duality of data-intensive file system design: reconciling HDFS and PVFS," *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 67:1—-67:12, 2011. [Online]. Available: http://doi.acm.org/10.1145/2063384.2063474

[15] Spark, "Apache spark," April 2018. [Online]. Available: https://spark.apache.org/

[16] Elasticsearch, "Elasticsearch ecosystem," April 2018. [Online]. Available: https://www.elastic.co/

[17] B. Dixit, R. Kuc, M. Rogozinski, and S. Chhajed, *Elasticsearch: A Complete Guide*. Packt Publishing, 2017.

[18] Elasticsearch, "Elasticsearch," April 2018. [Online]. Available: https://www.elastic.co/products/elasticsearch

[19] Lucene, "Apache lucene," April 2018. [Online]. Available: https://lucene.apache.org/core/

[20] Logstash, "Logstash," April 2018. [Online]. Available: https://www.elastic.co/products/logstash

[21] Kibana, "Kibana," April 2018. [Online]. Available: https://www.elastic.co/products/kibana