v. 3, n. 10, 2023

International Journal of Human Sciences Research

BIG DATA TECHNOLOGY FOR MONITORING ICT SERVICE DATA

Caiafa Marcelo Dante
Engineering and Technological Research Department, Universidad Nacional de La Matanza, Argentina

Aurelio Ariel
Engineering and Technological Research Department, Universidad Nacional de La Matanza, Argentina

Busto Adrian Marcelo
Engineering and Technological Research Department, Universidad Nacional de La Matanza, Argentina

Abstract: Data analysis has become an important source of knowledge for organizations. An adequate treatment allows to obtain valuable information. Its massive processing is possible from Big Data technologies. The work is based on the use of an open source platform for the processing of files generated by the communication systems of a mass service institution with three hundred branches in the country that serves more than two million customers. The research addresses the need to consolidate results based on indicators that add value to decision- making and management processes to improve the operational efficiency of information and communication technology (ICT) services. The objective is the development of a control panel based on measurement of different indicators. This allow the monitoring of its operating costs and the level of quality of customer care. For this, the ELK **TM** (Elasticsearch-Logstash-Kibana) set is used, fed with the call detail records known as CDR (Call Detail Records).

Keywords: Big Data, ICT, CDR, ELK.

INTRODUCTION TO BIG DATA

Big data offers ICT engineers a real opportunity to capture a more comprehensive view of their operations and services [1]. Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones and of a large enormous scale [2].

Particularly within the different use cases based on CDR analysis, one can cite real-time analysis (dynamic monitoring of the communications network), accurate marketing (churn prediction, third party ad personalization), operational efficiency (preventive customer service, service operation optimization) and improvement of the customer experience (monitoring of service quality) [3].

For ICT engineers, the analysis and data collection generates opportunities to execute the analytical process, measurement and diagnosis of incidents. These new technical competencies define a new specialist, known as a data scientist. This profile requires skills in managing information and improving the quality and relationships between data sets [4]. The main skills of the data scientist focus on database design and administration, and the development of data extraction, transformation and loading processes from multiple sources. This process known by ETL (extract, transform and load), allows flexibility in distributed processing workflows [5].

RESEARCH TOPIC

Through this work, it is proposed to respond to the need for analysis of the telecommunications infrastructure of a large organization, its operating cost, its level of availability and the level of quality of the care services it supports.

The result is a dashboard with consolidated information prepared with ELK technology, from the processing of large volume of data contained in the CDRs generated by its telephony servers.

The research proposal consists of CDRs analyzing files generated by the communications solution corresponding to CUCM ™ (Cisco Unified Communication Manager) technology.

This organization has two sets of servers. One called Central Areas that supports the service of the networks of facilities of the metropolitan area. The other called SUC, precisely because it provides service to the branch network exclusively.

The research work focuses on the process of developing a dashboard that consolidates different indicators according to the following objectives:

- Analysis of operating costs based on network traffic flows according to service monitoring, to detect fraudulent behavior when it occurs.
- Monitoring of the quality of care services, resulting from the representation of the mean operating time parameters for each of the interactions.
- Measurement of infrastructure availability levels obtained from branch network traffic.

The aim is to add value to the training of professionals working in the ICT sector, highlighting the skills necessary to achieve a successful and systematic process of data exploitation in a large company in the local market.

ELK TECHNOLOGY PLATFORM

Elasticsearch™ acts as a repository of information that stores the documents it indexes. It does not require a predefined schema, since the same collection of documents may contain a different structure. This makes it flexible and scalable to handle high-volume data.

It is a solution made up of three fundamental components from which it derives its acronym ELK. (Elasticsearch, Logstash and Kibana).

Logstach: it is the pre-processing, it collects the data and processes it before storing it in its databases. It requires a Java Virtual Machine and can run on different operating systems.

Elasticsearch: it is a distributed database. It distributes information and processing across all nodes, is fault tolerant and highly available.

Kibana: it is the presentation tool. The information is displayed and generated from filters and dashboards. The three components define a set of open source tools that combine to achieve the monitoring, consolidation and analysis of text files on multiple servers. It allows to solve three problems:

- Lack of consistency: multiple devices where each one has a different format.
- Time format: each log can have a different time format.
- Decentralization: the files are distributed in multiple routes.

Elasticsearch© is a Lucene-based search engine. It has an open source API (Application Programmable Interface) for information retrieval distributed under the Apache software license developed in Java. It provides a full-text, distributed, parallel-processing search engine with a RESTful web interface. Through HTTP web interfaces and JSON documents, it allows you to perform full-text searches.

The job used ELK™ version 7.10.0 running Windows 10 JVM version 8.0 on a single node.

WORK DEVELOPMENT

The development of the work is structured in 5 fundamental stages, of sequential execution.

In the first stage, the tasks of surveying the network infrastructure, the detail of the IP addressing and the architecture of the communications system were carried out to know the data source.

In stage 2, the CDR records are obtained, the relevant parameters are identified and a data dictionary is prepared for the correct interpretation of the data.

Stage 3 deals with indexing the database. Each CDRs field with relevant information is assigned a specific type of parameter so the search engine can properly process it from text file to JSON document.

Stage 4 runs the ETL process. A text file is prepared from which the Logstash module is configured for data ingestion within the corresponding database.

In the last stage, searches are carried out according to the specific objectives to be achieved, so that the Kibana module allows them to be consolidated into a control panel for presentation.

DATA SOURCE´S CONTEXT AND RELEVANCE

To achieve the proposed objective, the CDR records were used as a data source, generated by the communications systems of an organization with more than two million clients dedicated to mass consumption services. This company has eight thousand employees with three hundred branches located throughout the national territory. Its products are heavily regulated, it is the reason why the competition strategy focuses on differentiation based on the quality of customer service. It enhances the value of this work. The linking process to productive environment was necessary to know the details of service model and the infrastructure that supports the business services:

- the architecture of the different systems that are data sources,
- the dialing plan and the IP addressing plan,
- the structure of the different customer service models.

All these activities were vital to understanding the architecture of the solution technique and the different models of care. This was the key to the data contextualization and its interpretation.

DATA DICTIONARY

The CDR contains per-call information such as calling number, called number, time of call, call duration, etc. These records are generated each time a call is made or received and have different formats according to the technology used by the operator.

In this particular case, the source of data generation is the Cisco technology platform, Call manager version 11.5 (CUCM) whose CDR data structure is partially illustrated in table 1.

Parameter	*Means*
CallingPartyNumber	Call origin number
CalledPartyNumber	Call destination number
Duration	Duration Time (sec)
lastRedirectDn	Last transferred number
origDeviceName	Id source device
destDeviceName	Id target device
origIpv4v6	Source IP address
destIpv4v6	Destination IP address

Table 1. CDR parameters

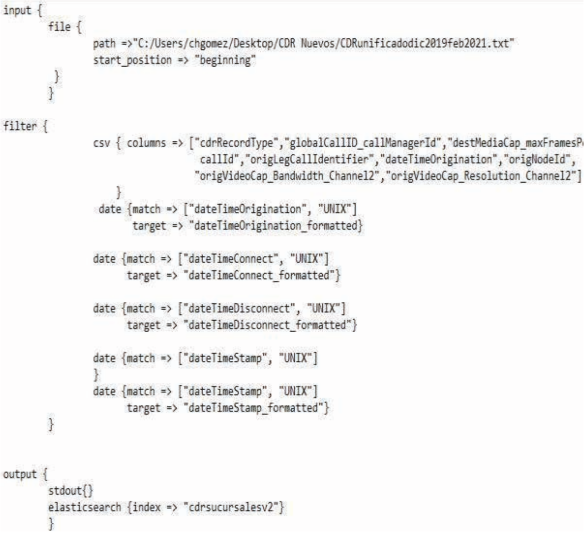
This data can be used for the processes of loading, settlement, billing, network efficiency, fraud detection, value-added services, business intelligence, etc. Additionally, they contribute to improving existing services and processes in different areas [6].

DATABASE INDEXING

To build the database, you must create an index in Kibana. For this, the fields that are needed are defined when formatting the data structure that is expected to be received. This is done through the DevTools section with a PUT request.

ETL PROCESSING

The ETL (extract, transform and load) is the process of collecting data, adapting its fields and loading data to the base. This is done by configuring the logstash with the file showed in fig.1



```
input {
        file {
                path =>"C:/Users/chgomez/Desktop/CDR Nuevos/CDRunificadodic2019feb2021.txt"
                start_position => "beginning"
         }
        }

filter {
                csv { columns => ["cdrRecordType","globalCallID_callManagerId","destMediaCap_maxFramesP
                                   callId","origLegCallIdentifier","dateTimeOrigination","origNodeId",
                                  "origVideoCap_Bandwidth_Channel2","origVideoCap_Resolution_Channel2"]
                    }
                 date {match => ["dateTimeOrigination", "UNIX"]
                       target => "dateTimeOrigination_formatted}

                date {match => ["dateTimeConnect", "UNIX"]
                      target => "dateTimeConnect_formatted"}

                date {match => ["dateTimeDisconnect", "UNIX"]
                      target => "dateTimeDisconnect_formatted"}

                date {match => ["dateTimeStamp", "UNIX"]
                }
                date {match => ["dateTimeStamp", "UNIX"]
                      target => "dateTimeStamp_formatted"}
        }


output {
        stdout{}
        elasticsearch {index => "cdrsucursalesv2"}
        }
```

Fig. 1. ETL process

Three instances are identified. The first is the input. It consists of indicating the file path to extract the data. The next step is the filter where all fields are listed in comma separated format. In the particular case of dates, the data received in UNIX format can be converted to a data type. Finally, the output indicates the name of the index where the data will be loaded.

RESULTS

The final parameter CalledPartyNumber was used to classify voice traffic according to the destination categories: Local, National, Emergency, International and Cellular. It shows telephone traffic distribution and allows to infer providers operating cost.

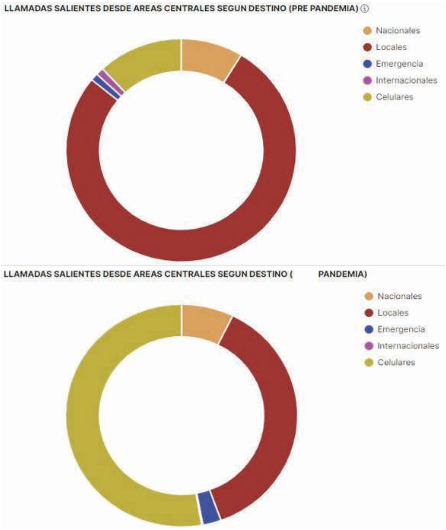


Fig. 2 Outgoing telephone traffic distribution

The filter was applied to different periods to show the changes caused by the restrictions generated by the pandemic. Figure 2 compares the data corresponding to the prepandemic period (Nov / Dec 2019) and the pandemic period (Nov / Dec 2020).

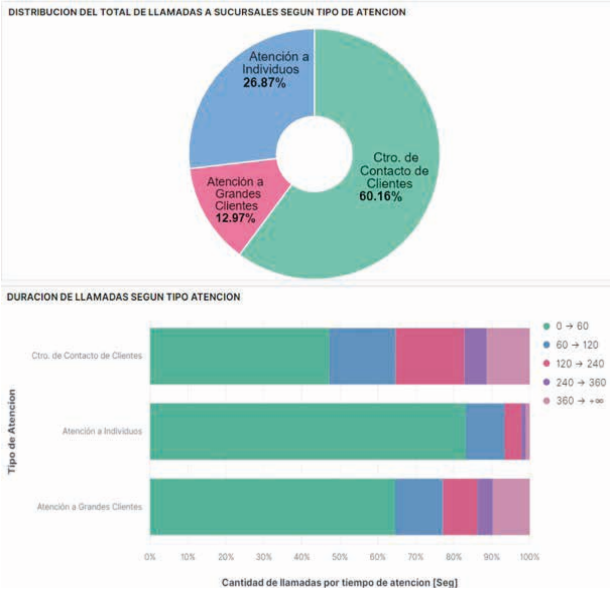


Fig. 3 Customer care level dashboard

Two graphs were used to build the dashboard that measures the level of quality of customer service and is shown in the figure 3. The top graph has the distribution of incoming calls classified by type of attention. The service is made up of three groups: customer contact center (CCC), service to individuals and large customers. The graph below details the duration of the calls in segmented per minute.

CONCLUSIONS

The measurements reflected in the outgoing traffic dashboard reveal the operating costs of the telephone service. Comparing the same months of 2019 (preCovid-19) versus the last 2020 (during Covid- 19), an increase of 320% is observed for the Cellular destination category and a 50% reduction in the Premises destination category. This reflects how the changes imposed by the restrictions of the pandemic that reduced the attendance of personnel to branches in exchange for using cell phones affected, with the increase in cost that it generates.

In the customer service dashboard, it is observed that those attended by the CCC represent 60% of the total calls to branches. The remaining traffic is two for the individual sector and one for large companies. In the CCC the 45% of the calls last less than 1 minute. CC staff usually adjust their behavior to productivity levels.

Comparing the calls answered by branch officers with a duration greater than 5 min, it is observed that large companies have 10%, in individuals it is only 1%.

The tasks of linking with the productive environment, although they require technical skills, highlights the need for soft skills for proper interaction and interpretation of the context.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

AUTHORS’ CONTRIBUTION

“CM conceived the idea and conducted the experiments; CM, AA and BA analyzed the results and revised the manuscript. All authors read and approved the final manuscript.”

REFERENCES

[1] C.-M. Chen, “Use cases and challenges in telecom big data analytics,” APSIPA Transactions on Signal and Information Processing, vol. 5, p. e19, 2016.
Cambridge University Press

[2] J. Verma, S. Agrawal & B. Patel: “Big Data Analytics: challenges and applications for text, audio, video and social media data”. International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, 2016

[3] Elagib, Hashim & Olanrewaju. “CDR Analysis using Big Data Technology”. International Conference on Computing, Networking, Electronics and Embedded Systems Engineering. Available at https://ieeexplore.ieee.org/document/7381414

[4] Morato J., Sanchez Cuadrado, S., & Fernández Bajon, “Trends in the technological profile of information professionals”. *25*(2),169-178. Available at https://doi.org/10.3145/epi.2016.mar.03

[5] Dobre, C., & Xhafa, F. 2014. “Parallel programming paradigms and frameworks in Big Data Era”. International Journal of Parallel Programming, 42(5), 710–738. Available at https://doi.org/10.1007/s10766- 013-0272-7

[6] Agrawal, D, P. Bernstein, E. Bertino, S. Davidson and U. Dayal, “Challenges and Opportunities with Big Data”. USA, Cyber Center Technical Reports, 2011, Available at http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article =1000&context=cctech