A Big Data Architecture to a Multiple Purpose in Healthcare Surveillance: The Brazilian Syphilis Case

Rodrigo Dantas da Silva
Federal University of Rio Grande do Norte
Natal, RN, Brazil
rodrigo.silva@lais.huol.ufrn.br

Jean Jar Pereira de Araújo
Federal University of Rio Grande do Norte
Natal, RN, Brazil
jean.jar@lais.huol.ufrn.br

Álvaro Ferreira Pires de Paiva
Federal University of Rio Grande do Norte
Natal, RN, Brazil
alvaro.ferreira@lais.huol.ufrn.br

Ricardo Alexsandro de Medeiros Valentim
Federal University of Rio Grande do Norte
Natal, RN, Brazil
ricardo_valentim@ufrnet.br

Karilany Dantas Coutinho
Federal University of Rio Grande do Norte
Natal, RN, Brazil
karilany@ufrnet.br

Jailton Carlos de Paiva
Federal Institute of Education, Science and Technology of Rio Grande do Norte
Natal, RN, Brazil
jailton.paiva@lais.huol.ufrn.br

Azim Roussanaly
University of Lorraine
Nancy, Lorraine, France
azim.roussanaly@loria.fr

Anne Boyer
University of Lorraine
Nancy, Lorraine, France
anne.boyer@loria.fr

ABSTRACT

For many decades society did need to monitor and assess the standard of living of the population. In the 1950s, the United Nations (UN) saw this need and proposed 12 areas that should be evaluated, the first of which is listed under "Health and Demography", which focuses on what is expressed as the level of a population's health. Decades have passed and great results have been gained from similar initiatives such as reducing mortality from infectious diseases and even eradicating some others. In the age of the digital society, needs have grown. Monitoring demands that once perished from data to become concrete now suffer from the opposite effect, the excess of data from everywhere. Healthcare systems around the world use many different information systems, collecting and generating hundreds of data at unimaginable speed. We are billions of people on the planet and most of us are connected to the virtual world, sharing information, experiences and events with some kind of cloud. In this information age, the ability to aggregate and process this data is a major factor in raising public health to a new level. The development of tools capable of analyzing a large volume of data in seconds and producing knowledge for targeted decision making can help in the fight against specific diseases, in the process of continuing education of professionals, in the formation of new professionals, in the elaboration of new policies. with the specific locoregional look, in the analysis of hidden trends in front of so much information faced in everyday life and other possibilities. The present work proposes an architecture capable of storing and manipulating seeking to standardize the variables in order to allow to correlate this large amount of data in a systematic way, providing to several services and researchers the possibility of consuming health, social, economic and educational data for the promotion of public health.

CCS CONCEPTS

• **Information systems** → **Mediators and data integration**; **Enterprise application integration tools**.

KEYWORDS

big data, healthcare surveillance, syphilis, epidemiology

ACM Reference Format:
Rodrigo Dantas da Silva, Jean Jar Pereira de Araújo, Álvaro Ferreira Pires de Paiva, Ricardo Alexsandro de Medeiros Valentim, Karilany Dantas Coutinho, Jailton Carlos de Paiva, Azim Roussanaly, and Anne Boyer. 2020. A Big Data Architecture to a Multiple Purpose in Healthcare Surveillance: The Brazilian Syphilis Case. In *EATIS2020: 10th Euro American Conference on Telematics and Information Systems, May 13-15, 2020, Coimbra, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3401895.3402092

1 INTRODUCTION

The use of technologies for analyzing big data is not new, but it has been drawing attention especially in recent years. With the arrival of the 2000s, the strengthening of social networks, mobile devices and web 2.0, some players from the technological world had to adapt to meet the demands. Conservative studies estimate that enterprise server systems in the world have processed 9.57 zettabytes of data in 2008, this number is expected to have doubled every two years from that point, these examples provide a small glimpse into the rapidly expanding ecosystem of diverse sources of massive datasets

EATIS2020, May 13-15, 2020, Coimbra, Portugal

ACM ISBN 978-1-4503-7711-9/20/05…$15.00
https://doi.org/10.1145/3401895.3402092

currently in existence [5]. In 2004, the world introduced the big data processing model called MapReduce, developed by Google. The major contributions of this work is a simple and powerful interface that enables automatic parallelization and distribution of large-scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs [1] [2]. This processing model not only made room For its creator, other companies took advantage of this new technology and embraced the idea, developing new applications. Apache Hadoop is an open source software framework that dramatically simplifies writing distributed data intensive applications since the primary primitive of map / reduce is a distributed sort, most of the custom code is glue to get the desired behavior [15].

This kind of approach did an effect not only within the virtual world, on social networks, but also on people's real life, in their daily lives. At the same time as technological evolution is already directing us to generate much more data. In a few years we will experience a significant increase in connected devices, thanks to the advent of IPv6 and the 5G connection standard. Big Data, like Big Oil, is big precisely because it can control access to data as well as extraction of information and knowledge, and we need to pay attention to that fact because knowledge, like energy, is not just a passive, inert resource. It is fuel: fuel for our ideas, our actions, everything. And the power that comes with control over that fuel is therefore formidable [10].

One of the most prominent areas in the adoption of this technology, which has already presented results, is the health area. This is an area that is basic to anyone, where a single medical procedure can contain structured and unstructured data, images, audio and digital signals. By scaling this scenario to the reality of a country where thousands of people use health services a day, we then bring dialogue into the scope of big data technologies. Using health data from a population or even an individual can help reduce costs, fight endemics, treat rare diseases, and so on. By analyzing a patient's medical data it is possible to see the patient's clinical picture broadly.

One subarea of health that proves to be a major investment field is epidemiology. It is the branch of medicine that studies the different factors that influence the spread of diseases, their frequency, their mode of distribution, their evolution and the means necessary for their prevention. In other words, it is the area of science responsible for analyzing the health-disease relationship with social, environmental and biological dimensions. In the 1950s the World Health Organization (WHO) convened a committee that could propose a method to define and assess the standard of living of a population. However, it was concluded that it would be impossible to construct a single index. It was then suggested 12 items that should be evaluated separately, the first of which being "Health and Demography", focused on what is expressed as the health level of a population [7]. Since then, the assessment of the standard of living of human populations has been a matter of interest to public and private institutions for a long time. We currently monitor every small step of the patient within the system from birth to death, detailed by his or her passage in the public health system.

For epidemiology, the variety of sources and diversity of data types is nothing new. Equally unsurprisingly for epidemiologists is the work of uniting datasets of different natures to obtain answers. However, the large increase in data volume presents new challenges, such as data validation, to avoid, for example, the reduction in data quality. The process of data surveillance is essential for the development of analyzes, as they, if low quality, can bias the whole results. In our current society, one person is capable of generating hundreds of data per minute. We make use of our smartphones as essential for living, we increasingly use wearable devices that monitor our health [17] [12]. We are increasingly adopting an automated lifestyle with smart devices in our homes, cars and workplaces. Many of us are addicted to social networks and we open the door to our virtual world.

In Brazil, in 2016, the Ministry of Health created an agenda of strategic actions for the qualification of health professionals in the diagnosis and treatment of sexually transmitted infections. The following year, the Union Court of Auditors realized after an extensive audit that the number of syphilis in the country had skyrocketed, growing by over 5000% in less than a decade. Given this scenario, the Ministry of Health has launched a project to combat and rapid response to syphilis, a campaign called "Syphilis No!"

Given these facts, the present work proposes an architecture for data storage and processing that, in its conformation, allows the standardized availability of data from the areas of health, social data and social network. It also allows researchers from different institutions to use this data to study theories and revalidate studies with the objective of meeting public health needs, such as epidemiological surveillance and the current situation of syphilis in the country.

The next sections will better present the project "Syphilis No!", The architectural models for big data solutions and data manipulation, the methodology adopted in this work for its development and what results have already been achieved.

2 THE PROJECT "SYPHILIS NO!"

Syphilis is a secular disease, present in various passages of history, which is once again becoming endemic. But not only in Brazil, case numbers have also grown in countries in North America, Europe, Asia and Oceania. Despite being a disease of simple treatment and diagnosis, it is having room to spread in society. Therefore, the Ministry of Health understood that there was no space for traditional methods, a communication campaign that reached all ages was necessary, a campaign to strengthen the capacities of health professionals, as well as the development of new techniques of monitoring and technologies for the most accurate and effective diagnosis. There was an indispensable need to create a horizontal process between the various areas of modern society so that we could indeed overcome the barrier of casual habit and indeed create technical and technological capacities to break the primacy of a disease that could already have been eradicated. The Ministry of Health Rapid Response to Syphilis in the Healthcare project that received the popular name "Syphilis No!" It is an action that goes beyond the classic campaigns of testing, prevention and treatment.

This project, started in 2018, is a partnership between the Brazilian Ministry of Health, the Pan American Health Organization (PAHO / WHO) and the Federal University of Rio Grande do Norte

(UFRN). The project already operates in all states of Brazil through the work of supporters hired to create an interface between the project and the locations with the worst indicators. The project also provides for ongoing training activities for health professionals, public awareness campaigns, in addition to promoting academic and medical research focused on new technologies to support and combat the disease.

3 THE ARCHITECTURE FOR BIG DATA

When we talk about data we need to go back to several basic concepts. Data is the smallest unit of information, and knowledge is the aggregate set of information that promotes state change. In the computing world the most traditional way of storing data is through databases. The CAP Theorem, or Brewer's Theorem, tells us that it is impossible for distributed data storage to simultaneously provide more than two of the three guarantees that are:

- `Consistency`: Each partition receives the most recent write or an error.
- `Availability`: The system continues to function as expected even if nodes fail.
- `Fault-Tolerant Partition`: The system continues to function despite an arbitrary number of messages dropped (or delayed) by the network between nodes.

In other words the theorem states that for a fault tolerant partitioned system, it would be necessary to choose between consistency or availability. In terms of large data processing can take a long time to execute a simple query. These queries could not be executed in real time and the processing delay would eventually return a result from hours ago.

3.1 Lambda Architecture

Seeking to solve these problems Nathan Marz published in his personal blog in 2011 a possible solution that allowed all three guarantees for distributed data storage systems, he called Lambda Architecture [4] [6] [3]. Marz's proposal solves this problem by creating two paths to the data flow, as shown in the Figure 1. All data received by the system goes through both data streams.

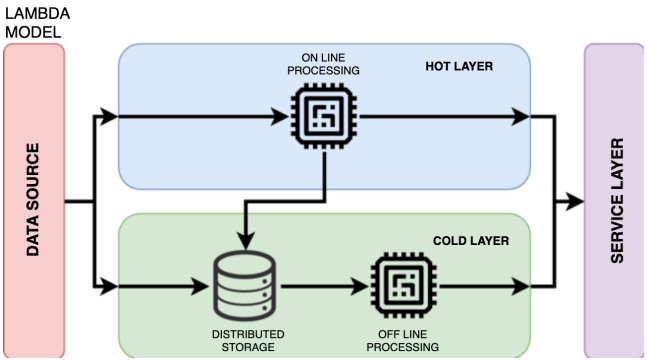

Figure 1: Representation of Lambda Architecture. Author's image.

- `Batch Layer or Cold Layer`: stores all input data in its raw form and makes the data available for batch processing. Results are stored in another area for consumption.
- `Quick Layer or Hot Layer`: analyzes data in real time. This data is constrained by latency requirements so that it can be processed as quickly as possible. This usually requires some disadvantage, which in this case is the level of accuracy.

Both layers serve a service layer, which indexes the batch data and receives incremental updates of the hot data. Raw data stored in the batch layer is unchanging. That is, new data is always added and old data is never overwritten. Any change to a specific data is taken as a new record with a new timestamp. These rules allow recalculation at any point in time in the data history. The ability to recalculate raw data is important as it allows new views to be created as the system evolves.

A disadvantage of Lambda architecture is its complexity. Data processing logic appears in two paths (hot and cold) using different structures. The complexity of administration and the risk of duplication in calculations can be a serious problem.

3.2 Kappa Architecture

Kappa architecture comes as a simplification of Lambda. The proposal is similar, just delete the cold path and to process batch data simply send it as a stream and process it by the hot path [9], as shown in the Figure 2.

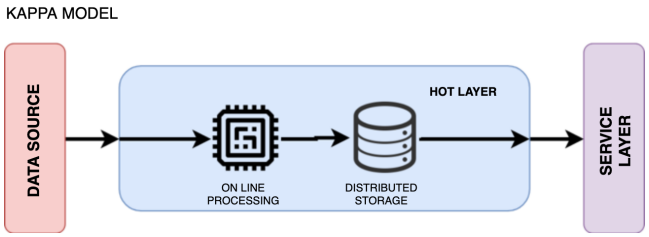

Figure 2: Representation of Kappa Architecture. Author's image.

The Kappa architecture comes with the proposal to make data canonical, that is, to standardize data so that it is possible to select and dynamically shape the execution of one or more business services. Simply put, the way Kappa works is as if all data were logged as a log of a system or service, and this log is provided for the architecture to process and store through a continuous stream of ingestion [8]. This architecture was designed by Jay Kreps in 2014. An evaluation by [14] to compare the architectures showed that Lambda uses 2.2 times longer than Kappa architecture, and that Lambda architecture needs about 10 - 20% more CPU and 0.5 GB of RAM.

3.3 ETL vs ELT

Since the emergence of Business Intelligence solutions and the entire area of Analytics, some techniques and concepts have been created, developed and gained their space, especially when we talk about the main source: data. One of the main structures in these areas is called Data Warehouses (DW), large data warehouses that

seek to structure data in order to answer the operational questions of a business area. With the advent of DWs a process has evolved and become essential, the Extract-Transform-Load (ETL).

ETL is a three-step data processing technique which in practice means extracting data from a source, transforming it to correct any abnormalities and suit business needs, and ultimately carries it in a new framework that supports the queries you want. Although ETL provides great solutions to many problems, it does generate some other problems for itself. When we talk about large volumes and data diversity the ETL technique no longer meets the needs. Suppose you need to process thousands of files with many GB of data and that this data needs to be made available to countless applications. According to [16] and [13] the ETL process can be easily complex, underperforming and facing data availability issues. Also according to the authors, the ETL process was conceived in such a way that the Extraction and Loading steps must occur at times when the source and destination systems of the data are in their maintenance windows, so that it does not occur completely. of services.

For authors [11] while the database technologies used for DW have evolved in performance and scalability in recent years, the ETL process has not evolved at the same pace. As a result, most BI infrastructures are facing bottlenecks: they can't easily get data on demand. The authors further state that, to eliminate the disadvantages of ETL, the adoption of new storage media. They suggest that the Extract-Load-Transform (ELT) approach may overcome these needs. The basic idea behind the process is to extract the data, store it as captured, and transform it only at the time of use, tailoring the transformations to suit different needs. The authors suggest four advantages of ELT over ETL:

- Flexibility to add new data (EL part of the process).
- Aggregation can be applied countless times on the same data (part T of the process).
- The transformation can be retrofitted even on legacy data.
- Accelerated process of implementation.

4 METHODOLOGY

The main objective of this work is to build a data architecture that allows other researchers to have access to data, concentrating not only data regarding the health of a population, but also data from areas that indirectly influence the shaping of a society and thus reflect how public health policies are implemented. The work is thus divided in the construction of a cluster for storage and processing of this data, a parallel step of curating the historical data of society in the areas of public health, demographic, economic and education. We chose to use the ELT process because the ultimate purpose of this architecture is to provide data from multiple sources to various systems and researchers, that is allowing the end user to decide which structure is best for the data they want to consume. It is also guaranteed that the data will be exactly the same for any purpose.

4.1 The Cluster

A cluster capable of storing and processing 8TB (terabytes) of data was built. This cluster was structured according to Lambda architecture, using Apache Hadoop version 3.0.3 as its development and management platform. For data repository was implemented Apache Hive version 3.1.2 on Hadoop. For cluster administration, Apache Ambari 2.7.4 is used. There are also two separate servers, one for storing the original data and one for data manipulation. The latter is used only for conformation of variables, such as gender normalization, as well as file format conversion. The choice for this structure was due to the scalability of the storage and processing infrastructure at a viable project cost.

Over the cluster was implemented an inverted indexing service, in order to obtain greater efficiency in the process of searching and processing data. Reverse indexing works like a book's index, which shows which pages occur in a given term, unlike traditional ways of storing data where an id is associated with a given data set. This service is used as a "gateway" by the data ingestor, thus allowing a prior analysis of the data before storing it in the repository. The pre-parse process allows you to map all values that dataset has, after this analysis the system creates a unique hash code for this dataset by inserting it into the repository. The service also provides a RESTfull service so that other applications can consume data as desired. An abstraction about the architecture can be found in the Figure 3 below.

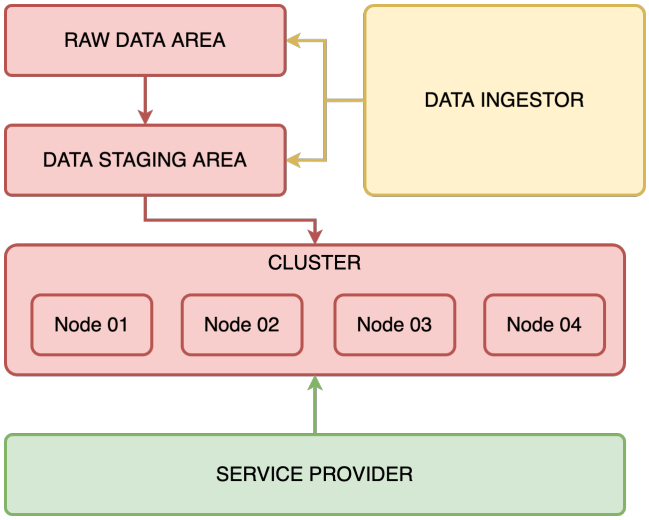

Figure 3: Abstraction of the implemented architecture focused on layers of interaction from the process of data acquisition, indexing, storage and retrieval to provide service. Author's image.

4.2 Data Curation

Data from Ministry of Health systems are available in monthly batches, representing all consolidated data of the corresponding competence. Data regarding outpatient production, hospital production, mortality, disease notification, health facility infrastructure, health teams, health professionals and health financing are captured. Data on basic education, higher education and education funding are also being collected in batch format. Other annual frequency data were also collected, such as municipalities' Gini indicator, municipalities' GDP and population.

Table 1: Sample codes found used for gender

Samples	Standardization
Man, Woman, Male, Female, M, F, 1, 2	Man, Woman

All data are previously studied by the development team of this project, with the support of public health experts. The databases that have data dictionary, go through a parameter check, in order to see if the collected data actually correspond to what theoretically should, according to the dictionary. Other data, such as notifiable diseases, do not have a dictionary. In this case the notification forms are used to construct a dictionary so that the parameter matches can be checked. In this last scenario, the construction is accompanied by nurses and epidemiologists for technical support. None of the bases are discarded if they fail to match variables and the dictionary. This routine is only for mapping out what needs a particular source can meet.

The curation process also guides how data should be normalized before it is inserted into the cluster and made available to researchers. Variables corresponding to gender, education, locality, procedures and others, change their insertion according to the origin system. Thus, an organizational pattern of the data is created so that all data can correspond to the same universe of options, as shown in the table 1.

5 RESULTS

The project is still in its first months of development, however some experiments have already been performed and some previous results have been obtained. Approximately 2TB (terabyte) of public data on health, education and economics, commensurate with the past decade, has already been collected.

One of the first experiments performed using the built data architecture was to understand who are the syphilis patients in Brazil. For this, we used data corresponding to the last 10 years of syphilis notifications provided by the Ministry of Health. A simple clustering technique was used observing some social variables of the patients, such as ethnicity, education and age group. We found 130 distinct groups of characteristics using the k-means algorithm with 97.41% accuracy. A representation can be found in the Figure 4 which shows in the first column the gender, at the second one the ethnicity, at the third the education level and last but not least, the fourth column shows the age range.

This result reinforces the need and objectives of these projects. First, we must observe the dimensionality of the different groups found. Brazil is a country with just over 200 million inhabitants and there are 130 different representations of syphilis patients. Second, understanding the behavior of this population in their social contexts will help to see how to create actions for these 130 groups. We realized, for example, that a portion of young people between 20 and 29 years old had complete medical education, what was missing in the training of these young people to create awareness about STIs? In other words, what actions are necessary to reach this segment of the population?

6 DISCUSSION

For many years, some governments have been developing public health policies without seeing the causal facts and consequences of these actions in a given society. Another point is that some public policies do not consider the country's dimensionality, as is the case in Brazil, which often presents geographical, cultural, financial, educational and social diversity. In view of these points, it is essential that we understand the various locoregional structures so that we can build targeted strategies, adapted to the real needs of a city, state or region. This allows for much more efficient responses and better use of the budget. But this whole vision had never been taken from paper.

The project proposed here brings to light new possibilities for public health, going beyond the traditional clinical view. The ability to look at data from different areas and understand the social behavior of a disease enables health managers to act in a preventive way, but not in a generic and standardized way for the whole country, but observing the peculiarities of each location.

Some of the studies in progress will also result in the statistical mapping of the health care network, such as the relationship between the technical capacity of health professionals and the capacity to respond to a certain condition. Allowing, for example, to understand the network of hospital beds in a given region and allowing to assist in the best order of use.

It is worth mentioning that a web platform for researchers' access management is under development, which can find in this space the database catalog, with the description of its variables, relationship between databases, as well as the option to download a sample for develop your research work according to the filtering options. This platform is really a project concern with the rules of the General Data Protection Law (LGPD / Brazil) and the General Data Protection Regulation (GDPR / Euro).

ACKNOWLEDGMENTS

I would like to acknowledge the research team at the Laboratory of Technological Innovation in Health and the Federal University of Rio Grande do Norte as well as the Brazilian Ministry of Health for funding this research.

REFERENCES

[1] Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51, 137–150. https://doi.org/10.1145/1327452.1327492

[2] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google File System. *ACM SIGOPS Operating Systems Review* 37, 29–43. https://doi.org/10.1145/945445.945450

[3] Zirije Hasani, Margita Kon-Popovska, and Goran Velinov. 2014. Lambda architecture for real time big data analytic. *ICT Innovations* (2014), 133–143.

[4] Michael Hausenblas and Nathan Bijnens. 2015. Lambda architecture. *URL: http://lambda-architecture. net/. Luettu* 6 (2015), 2014.

[5] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. 2014. Trends in big data analytics. *J. Parallel and Distrib. Comput.* 74, 7 (2014), 2561 – 2573. https://doi.org/10.1016/j.jpdc.2014.01.003 Special Issue on Perspectives on Parallel and Distributed Processing.

[6] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja. 2015. Lambda architecture for cost-effective batch and speed big data processing. In *2015 IEEE International Conference on Big Data (Big Data)*. 2785–2792. https://doi.org/10.1109/BigData.2015.7364082

[7] Ruy Laurenti, Maria Helena Prado de Mello Jorge, Maria Lúcia Lebrão, and Sabina Léa Davidson Gotlieb. 2005. *Estatísticas de saúde.* EPU.

[8] Adam Leadbetter, Damian Smyth, Robert Fuller, Eoin O'Grady, and Adam Shepherd. 2016. Where big data meets linked data: applying standard data models to

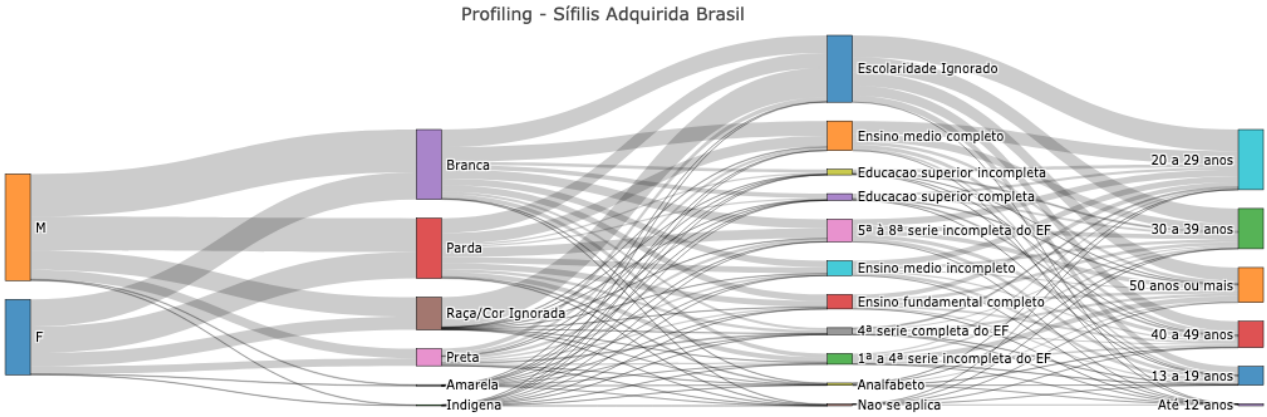

Figure 4: Sankey diagram depicting the profiles of people who oppose syphilis between 2007 and 2017. Image of the author..

environmental data streams. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2929–2937.

[9] Jimmy Lin. 2017. The lambda and the kappa. *IEEE Internet Computing* 21, 5 (2017), 60–66.

[10] M.P. Lynch. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. Liveright. https://books.google.com.br/books?id=v4b8CQAAQBAJ

[11] Pablo Michel Marín-Ortega, Viktor Dmitriyev, Marat Abilov, and Jorge Marx Gómez. 2014. ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data. *Procedia Technology* 16 (2014), 667 – 674. https://doi.org/10.1016/j.protcy.2014.10.015 CENTERIS 2014 - Conference on ENTERprise Information Systems / ProjMAN 2014 - International Conference on Project MANagement / HCIST 2014 - International Conference on Health and Social Care Information Systems and Technologies.

[12] Stephen J. Mooney, Daniel J. Westreich, and Abdulrahman M. El-Sayed. 2015. Commentary: Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)* 26, 3 (May 2015), 390–394. https://doi.org/10.1097/EDE.0000000000000274 25756221[pmid].

[13] P. Petrova, V. Jotsov, and V. Sgurev. 2018. Puzzle Methods for Automatic Selection of Data Cleansing Techniques. In *2018 International Conference on Intelligent Systems (IS)*. 820–826. https://doi.org/10.1109/IS.2018.8710580

[14] A. Sanla and T. Numnonda. 2019. A Comparative Performance of Real-time Big Data Analytic Architectures. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. 1–5. https://doi.org/10.1109/ICEIEC.2019.8784580

[15] T. White. 2009. *Hadoop: The Definitive Guide: The Definitive Guide*. O'Reilly Media. https://books.google.com.br/books?id=bKPEwR-Pt6EC

[16] A. Wibowo. 2015. Problems and available solutions on the stage of Extract, Transform, and Loading in near real-time data warehousing (a literature study). In *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. 345–350. https://doi.org/10.1109/ISITIA.2015.7220004

[17] Sean D. Young. 2015. A "big data" approach to HIV epidemiology and prevention. *Preventive Medicine* 70 (2015), 17 – 18. https://doi.org/10.1016/j.ypmed.2014.11.002