A Data Lake and Analytics Platform with Application to COVID-19 Dynamic Analysis

Francinaldo Almeida Pereira
Grad. Prog. in Elec. and Comp. Eng.
Univ. Federal do Rio Grande do Norte
Natal, Brazil
E-mail: falmeida@dca.ufrn.br

Júlio Gustavo S. F. Costa
Grad. Prog. in Elec. and Comp. Eng.
Univ. Federal do Rio Grande do Norte
Natal, Brazil
E-mail: juliogustavocosta@gmail.com

Luiz M. G. Gonçalves
Grad. Prog. in Elec. and Comp. Eng.
Univ. Federal do Rio Grande do Norte
Natal, Brazil
E-mail: lmarcos@dca.ufrn.br

***Abstract*—We propose a platform consisting of a data lake that has been implemented as a web-based service, to specifically solve the Covid-19 data production and processing problem. The main idea is that it can be used by data scientists working on COVID-19-related projects in order to access as much data as possible in one repository and be able not only to analyze that data but also to manage and contribute to new data. Through this platform, it has been possible to dynamically aggregate different data repositories related to the COVID-19 pandemic, in order to provide users, through a web interface, tools for use, transformations, and collaboration of data, as well as analysis and visualization tools integrated to geographic information systems.**

***Index Terms*—Covid-19, Data Lake,**

I. Introduction

Since the beginning of 2020, the world population has been threatened by the pandemic caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1], worldwide known as Covid-19. Officially declared a pandemic by the World Health Organization (WHO) on March 11, 2020, the unfortunate initial expectation regarding the disease has become reality in a few months. Almost two years later, more than 585 million cases have been confirmed worldwide, and more than 5.42 million deaths. In Brazil, one of the current main focuses of the disease, there are more than 34 million cases and 680 thousand deaths according to the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [2]. Such expressive numbers confirm the high rate of virus contamination, and even though this pandemic seems to be nearing its end, its negative legacy justifies the importance of initiatives to prepare for the future.

Since millions of people are affected by COVID-19 in the world, the amount of data being generated and stored is huge. Given this volume, it turns out to be complicated to analyze and establish viable solutions to control the pandemic [3]. Over its course, forecasting the pandemic's behavior was (and still is) one area of significant interest to the academic community. In particular, the use of machine learning to predict the outbreak of the virus, as well as possible deaths in the near future was widespread. Several data-driven approaches can be found in the literature [4]–[6] with development of models based on ARIMA, SEIRD, auto-encoders, among others.

This heavy use of machine learning and data-driven approaches concentrates considerable efforts of the development of the data used. Figure 1 presents the stages of machine learning workflow. It can be noticed that three stages are directly focused on data: data collection, data cleaning, and data labeling.

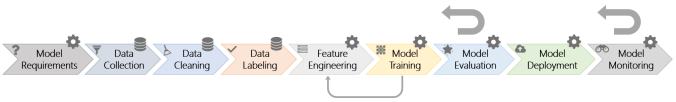

Fig. 1. Stages of Machine Learning Workflow [7]

Nonetheless, dependency on data has some drawbacks. Identified data dependencies in machine learning systems is a key factor that contributes to technical debt [8]. Verma [3] also noticed the lack of standard datasets as one of the major challenges that big data platforms and applications currently meet. To address that, in this work, we aim to contribute to the set of data applications for machine learning by providing a solution that encompasses a repository of data, namely a data lake, enriched with analytical capabilities and focused on COVID-19 data.

The work presented here is part of a collective effort of a group of researchers that compose the N-COVID team. In light of the pandemic unfolding situation, the Natalnet Laboratories of UFRN has created a team to work on the development of studies, procedures, and technological innovations to face the COVID-19 pandemic and contribute to related topics. This project has support from CAPES, Brazil, and its central goal is to develop *methods for predicting the dynamics of viral epidemics and pandemics with clustered data analysis from the perspective of artificial intelligence*.

As part of this project, two integrated systems are being built: one related to the evaluation and application of data-driven models to predict coronavirus outbreaks; another that supports the consumption, production, and management of data relating to the users of the system, and the pandemic and the system as a whole. This is an effort that is being carried out in collaboration with other students from the same graduate

This work is sponsored by CAPES under grant 88887.506889/2020-00.

program working on different aspects. The contributions of the work in this document are limited to the second type of system developed for the endemic and pandemic prediction project.

II. Data Management and Analytics for Covid-19

Based on the developments of other team members, including work that led to previous publications, we identified the lack of a central storage location for data used. The absence of proper analytical solutions to manage and visualize the data was identified as well. In this context, our research can be considered under the theme of data management and analytics. Since the ultimate goal is to have contributions that can be favorable in fighting the current and future epidemics, and not to distance ourselves from the motivation and issues that led to the initial consideration of the current research, throughout its development we try to answer our main reseaarch question stated as *is there an epidemic-related data aggregation, management, and availability tool with geographic information support aimed at data scientists for Covid-19 pandemics?*

To answer this question, we proposed to first study, verify, and analyze existing solutions related to COVID-19, as well as general purpose ones, in order to identify possible gaps and areas of improvement. In addition, based on the aforementioned problem, we formulate our research hypothesis, to be investigated throughout this work, as *it is possible to dynamically aggregate different data repositories related to the COVID-19 pandemic, in order to provide users, through a web interface, tools for use, transformation, and collaboration of data, as well as analysis and visualization tools integrated into geographic information systems*. Thus, the main contribution of this project is the data lake itself, which has been developed as a web-based service, and that can be used by data scientists working on COVID-19-related projects, in order to access as much data as possible in one repository and be able not only to analyze that data but also to manage and contribute with new data.

A. Big Data

One of the most widely used definitions of big data comes from the Information Technology Glossary of Gartner, Inc., a research and advisory firm, as *Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*.

The 3 Vs (volume, velocity, and variety) became for a long time the standard when referring to big data. Later on, veracity and value were also introduced into the definition, forming the 5 Vs of big data [9]. Below are some details on those characteristics:

- **Volume**: the word ”big” in the term makes this perhaps the most associated aspect of big data. In fact, the ability to process large quantities of data is a key factor for the adoption of big data in industry.
- **Velocity**: if the volume of data is huge, the speed with which this data is being generated is just as great. Precisely because of that, the timing of processing this amount of information is also to be considered. The velocity aspect accounts for both the generation and processing of data.
- **Variety**: when dealing with big data, information comes in a wide variety of forms, and this represents a break with traditional ways of storing data. The need to meet this demand for adaptation to the heterogeneity of sources and formats is handled in the various aspects, which also accounts for differences in the importance of these sources of information depending on the nature of the business.
- **Veracity**: dealing with the aspects above causes a need to assess the quality of data. Missing or dirty data or even data that does not correspond with reality may have negative consequences on the process. The goal of the veracity aspect is to make sure the information obtained is authentic and useful.
- **Value**: this is the most important aspect of big data and it refers to the capacity of turning data into something valuable. The simple fact of having data is not sufficient to justify all the work and resources that are generally needed in big data.

Loshin [10] advocates the necessity of looking beyond the definition to understand the big data concept. At its core, applying innovative and cost-effective techniques when dealing with modern high-resources demanding problems is what makes big data. Figure 2 illustrates this concept.

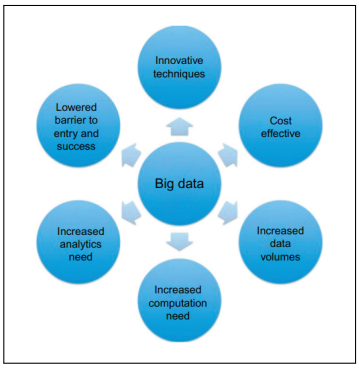

Fig. 2. Big Data Concept [10]

B. Data Warehouse

The data warehousing idea is present in the literature since the late eighties [11]. At that time, IBM was trying to help users concentrate efforts on the information rather than on the means for obtaining it. The original architecture, based only on a relational database, proposed separation between operational and informational systems' data, with the latter periodically supplying the former with updates.

Operational databases keep stored the information needed to enable system transactions. They are used to record and

perform pre-defined operations, and are stored in specific formats according to the needs of the system. Also, control over this source of data is little to none. Informational databases, on the other hand, store analytical data, usually over a large historical period. They are able to preform complex queries and need access to information as quickly as possible. [12]

From the informational databases [13], the recognized father of the data warehousing concept defines a data warehouse as a *subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions*. In contrast with operational systems, the subject-oriented nature of data warehouses turn the focus from application to subject. For instance, in an industry context, subjects could be customers, products, and sales. The process of defining those subjects is very important in the construction of a data warehouse. Integration is one of the main characteristics of a data warehouse. It is responsible for defining a single representation for the data coming from the different systems that will compose the data warehouse's database, thus diminishing inconsistencies when representing the same information. Nonvolatility in a data warehouse restricts changes that can be made to data. Once a set of data is loaded into the warehouse, it can be accessed, but no updates can be done in this set. Only its replacement by a new set is permitted, creating a history of snapshots in the data warehouse.

The last characteristic - time variance - implies the existence of some form of time marking in every unit of data in the data warehouse. Unlike operational databases in which an element may or may not be present in the key structure, for a data warehouse this is a requirement, making data always consistent within a period of time. Data Warehousing promotes the development of applications based on a self-service model [14], in contrast to the traditional report-oriented model. In a data warehouse environment, end users access data and create their own reports directly through user-friendly query tools that implement ad-hoc searches to a consolidated base [15].

The independence of specific, inflexible solutions for processing management information is one of the main attractions behind investments in data warehouses. These characteristics of independence and flexibility, however, hide a complex process of extracting, processing, and loading operational data, which can involve hundreds of databases maintained by dozens of provider systems.

C. Data Lake

A Data Lake can be seen as a collection of data, with origin in different sources, and stored in raw format, without being processed for a specific purpose. Roughly, it is a repository to store data leaving them available to be analyzed and worked on other applications, as needed, and contemplates the storage universe that is defined in the architecture design, acquisition, and storage of data. It can be noticed that a data lake is a concept, not a technology. In fact, several technologies may be needed to implement a data lake. They are designed for data consumption in a process that involves collecting, importing, and processing data for storage or later use. It does not require the creation of a schema before preparing data for storage, since it is based on the schema-on-read principle [16]. Data can simply be consumed and the schema created and applied when the data is used for analysis.

Although data lake and data warehouse concepts may seem interchangeable, they have different purposes. Both of them function as big data repositories and are mainly used for storage, with data lakes being able to store structured, unstructured, and semi-structured raw data whereas data warehouses are limited to the first category. Because this process of transforming data before storing it, in a data warehouse, can be complex and time-consuming, data lakes have the advantage of being able to collect data without an immediate specific intention. Table I highlights the main differences between data warehouses and data lakes.

Characteristic	Data Warehouse	Data Lake
Data	structured	structured, unstructured, and semi-structured
Schema	schema-on-write	schema-on-read
Data Quality	processed data	raw data
Agility	less agile with fixed configuration	highly agile with reconfiguration as needed
Users	business professionals	Data scientists, developers, and business analysts

TABLE I
DATA WAREHOUSE AND DATA LAKE COMPARISON

To leverage the information around raw data, data lakes usually add a layer of context over the data, associating it with some meaningful knowledge [17], but that is not always the case. There are four aspects of a dataset in the context of data lakes [18]: 1) storage systems in which they are hosted may differ; 2) their formats may also be different; 3) useful metadata may either not be present or be described in different formats; 4) they are susceptible to changes over time.

The flexibility of this kind of repository sets data scientists and analysts as the target audience of data lakes. Their need for advanced filters and analysis applied to the data beforehand, aligned with uncertainty over which specific part of data has actual influence in their projects, makes data lakes a very useful feature. For a better understanding, Figure 3 displays the life cycle of data in the context of a data lake.

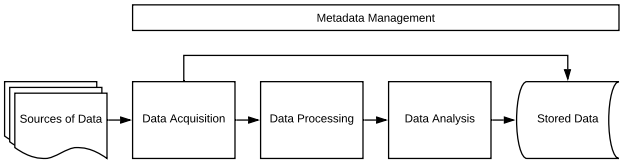

Fig. 3. Life Cycle of a Data Lake [19]

The data may be acquired from multiple sources in a process that may require different techniques. At the time of acquisition, the data, in its raw format, becomes part of the data lake. Further processing may be executed on data to derive meaningful information or to be transformed into alternate

formats. Original data is always kept. The data is further analyzed and stored according to project-specific requirements. While the data flow into a data lake from the point of acquisition, its metadata is captured and managed [19].

D. Big Data Analytics

Big data analytics (BDA) has received considerable attention over the past years. It is closely related to decision-making processes and can be seen as an evolution of the set of decision support technologies popularly known as business intelligence (BI) [20], [21]. As research on data science evolves, combining techniques from the nineties and new research trends concerning big data, machine learning, and computational statistic, big data analytics starts to play an important role in knowledge-intensive organizations [22]. The term is defined as a holistic process that involves the collection, analysis, use, and interpretation of data for various functional divisions with a view to gaining actionable insights and creating business value [23]. In an even broader definition, BDA condenses the process and tools used in order to extract insights from big data. Its scope relates not only the data upon which analysis in performed but also to elements of tools, infrastructure, and means of visualizing and presenting insights [24].

Thus, an important aspect that must be taken into account is that the method used for big data analytics must extract knowledge from the data in an interpretable way, that is, the computational techniques used to perform this task must make the implicit patterns, existing in the data, transparent for those who use them, providing a better understanding. [25]

III. Related Work

One of the most important works on data lake was introduced in 2017 [26], with an open source data lake service that provides a RESTful API for managing data called CoreDB. Being a *database-as-a-service* application, users are able to create datasets in the data lake and perform CRUD operations on them. Both relational databases and NoSQL databases are supported and can be specified at the time of dataset creation. Indexing and querying are also available and are leveraged under the hood by Elasticsearch. Security, access control, and tracing are also handled by CoreDB.

Another system developed is Constance [27], which has as its main focus the management of metadata information with support for discovery, extraction, and summarization of metadata in data sources. A semantic metadata matching was created to annotate incoming data and link semantic equivalent elements across sources. To handle the tasks of data ingestion, metadata management, and query answering, a unified user interface is provided. Idealized as a generic system, the authors claim it can be extended to solve problems in other specific data lake systems.

Specifically, regarding Covid-19, a tool called the COVID-19 Watcher [28] was initially created in 2020 with data for the United States with three sub-levels of geographical organization. The authors developed a methodology to aggregate county-level COVID-19 data into metropolitan areas, with the information provided to users via a public dashboard. Collected data comes from three different sources and is updated automatically at each hour in a process that validates the size and format of files. At the time of this writing, access to the website of the project was not available.

The team at Amazon Web Services (AWS) maintains the COVID-19 Data Lake, centralizing datasets associated with the current pandemics. The data lake provides a vast catalog ranging from tests, cases and deaths to vaccination and availability of hospital beds. Almost all of the information is concentrated either on the U.S. states or globally at the country level. The data is hosted on AWS S3, and is publicly available as is in both CSV and JSON formats. A sample dashboard is provided in order to display some of the capabilities when working with the data in conjunction with other AWS paid services, but it is not maintained alongside the project.

A COVID-WAREHOUSE that integrates COVID-19 data from the Italian Protezione Civile Department in conjunction with pollution and climate data from two Italian regions was also developed [29]. Epidemiological and climate data are updated automatically trough scripts. On the other hand, pollution data are manually collected. Since they are using the data warehouse concept, all data is processed and treated at the moment of integration into the system to comply with a common schema. The authors also present a case study in which the COVID-WAREHOUSE is used to analyze a possible correlation of pollution, climatic events and the evolution of the pandemics.

The C3.ai, a software provider with focus on AI technologies, released the COVID-19 Data Lake with more than 40 sources of data unified into a cloud image and seeking to be the largest source of pandemic-related data. Not only epidemiological tracking data is supplied for several countries but also data on virus, clinical records, mobility trends, demographics and imaging. The team also provides a knowledge graph to show linkages across datasets. The data lake can be accessed free of charge through a RESTful API where information is modeled into types like outbreak location, diagnosis, vaccine coverage, among others.

Work	F1	F2	F3	F4	F5
COVID-19 Watcher	Yes	No	Yes	No	No
AWS COVID-19 Data Lake	Yes	On AWS	No	Yes	No
COVID-WAREHOUSE	Part.	No	Yes	Yes	No
C3.ai COVID-19 Data Lake	Yes	Yes	No	Part.	No
CoreDB	No	Yes	No	No	Yes
Constance	Yes	No	No	No	Yes

TABLE II
RELATED WORKS AND THEIR MAIN FEATURES

Table II summarizes the works most related to ours, with their several features observed: automatic updates, availability of data via API, integrated analytics, geo-information support, and user-provided data.

IV. Data lake Proposal and Implementation

As seen, existing solutions are limited in terms of features and their usage by third-party projects is not straightforward.

Therefore, data scientists still face the problem of not having at their disposal a reliable and unified source of data that could optimize the machine learning workflow. Given this reality and all the issues involved, we see the opportunity to contribute to the scientific community with our proposal. Our proposal consists of developing a data lake distributed in a software as a service (SaaS) fashion. The initial blueprint is composed of three main modules: data collection and management, metadata management, and data analytics. The modules will be seamlessly integrated with all layers abstracted from the main user, who would access the system either via a unified interface on the web or via RESTful API.

The core of the system is the data collection and management module. This module is responsible for gathering data from multiple sources. Due to the dynamic nature of the system, users will be able to add new sources of data from static files and also from online resources, defining the frequency of updates to be later fetched by the system or using webhook endpoints to inform of updates to the data. Next, the metadata management module is responsible for the unified semantics of all data present in the data lake. Each dataset owner will be able to associate community-crafted information terms to add meaning to its own data. For example, only by looking at raw data, the system cannot infer how the number of confirmed positive cases is stored in each source. Even considering only CSV files, header columns may or may not be present, or maybe in different languages. Hence the importance of this module in the collective nature of the solution we are to provide.

Finally, being the last module, data analytics support is very important in this project. We aim to provide data scientists with exploratory data analysis and enhanced business intelligence-powered visualization tools. This module's goal is to save time and resources spent on the initial phases of models' development to help users to better understand the information at their disposal as well as its dirtiness, completeness, and possible existing trends. Currently, a prototype of our proposal has been implemented and is being used by the N-COVID team. Despite being in a proof-of-concept format, so far, experiments had led to a deepening of knowledge on the subject besides functioning as requirements gathering phase so that team's demands could be better understood.

V. Experimental Evaluation of the Platform

Figure 4 shows the initial interface of our platform. It starts with visualization of data from Brazil as a whole and allows the user to see data from the states, just by selecting the map to get data from a specific state. Currently, it is showing predictions using LSTM approach [6]. However, other source codes can be easily deployed following the same process using our proposal. In the following, we show experiments about usability and other issues related to our web data lake platform.

A. Deployment of Code and Data

Our platform allows having data deployed on the website coming from different official sources. At the present, a

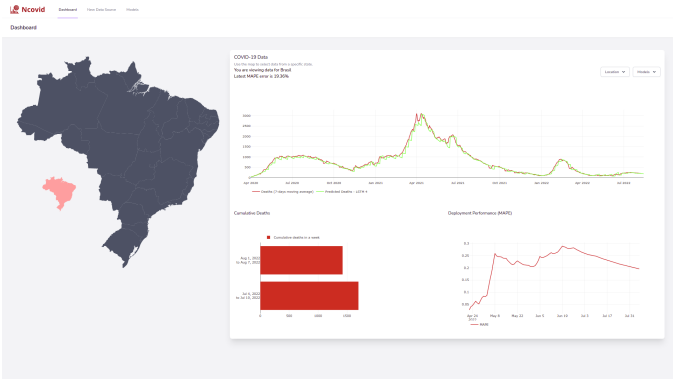

Fig. 4. Initial interface of the N-Covid platform (http://ncovid.natalnet.br/)

currently available data source is the one gathered and made available by Wesley Dash at https://covid19br.wcota.me/. On a daily basis, the system fetches a CSV file from the WCota repository and separates data by each state and nationally, filtering the features of interest and storing the results internally. A microservice called data manager reads from those files and provides the data via an API to be displayed as a time series on the website.

The same data is used separately to train prediction models that are developed by researchers working on our current project. The source code for such models is encompassed in a docker container and deployed to our server. Upon startup, each container communicates with a model registry to inform its availability to train new model instances. This step is necessary in order to provide the training options on the user interface. The container then starts to listen for training requests. Due to the time a container needs to train a new instance, communication between the container and the user is made via asynchronous messaging started by the website. Each container is responsible for a type of prediction model (like the model named Autoregressive SARIMA, for example), and a correspondent unique identification is also present on the request message. A message broker handles incoming messages, forwarding them to the containers which, based on the identification type, decide whether or not to take further steps.

On the occasion that a container receives a request to train a new instance of its type, it communicates with the data manager to fetch the most up-to-date data for a range of dates indicated in the request. The data is returned as a JSON, and the training process starts. At the end of the process, a file with metadata on the new instance is generated along with the model instance itself. The couple is stored by the container, which then sends a new message to be consumed by the instance registry containing the metadata. From this point, the instance is made publicly available on our website and live predictions can be made. Each prediction request goes through a proxy that extracts information from the URL on which model type the prediction refers to. With that information, the request is proxied to be resolved by the right container. Each

container implements a standard API and returns prediction data in a similar fashion, so the website is able to understand the data and couple it with the original time series.

Additional information is displayed, for instance, what is called the deployment performance is calculated based on the forecasts of several days, generating a curve of the mean absolute percentage error (MAPE) updated daily from the day the model instance was deployed. A weekly comparison of cumulative deaths is also displayed. Both of these were created to demonstrate the capabilities of the website to make transformations on original and predicted data.

Regarding usability, the repository is being used by 4 researchers from our associated project. These users basically used the system in all three criteria evaluated in terms of usability: ease of handling and using data (including possible changes in data structure and source); easy understanding of the tool's features; development and deployment of code for new applications developed by researchers. Each of these items was evaluated in only 2 situations: satisfactory or unsatisfactory. Regarding the possibility of uploading and downloading data, and changing its structure, the platform is considered satisfactory by all 4 researchers. As for the understanding of the functions in the interface itself, it is also considered satisfactory by all of them. And, the same for the last criterium, which is the development and deployment of code.

VI. Conclusion

In summary, our proposal refers to developing tools for implementing the data lake related to Covid-19 to be used by data scientists of our research project, which has been developed and is currently available at *https://ncovid.natalnet.br/*. Therefore, as seen, we have developed and tested the main modules and made them operational.

Our main contribution is the platform itself, which integrates a new concept of development, with data acquisition and code deployment done at the same time, in the same tool. An immediate application is in the project that is undergoing, where our data scientists are using our tool in their works on data-driven and mathematical modeling of Covid-19 dynamics [6].

Further work is to populate the data lake with data that our data scientists are using in their works on data-driven and mathematical modeling of Covid-19 dynamics.

References

[1] A. E. Gorbalenya, S. C. Baker, R. S. Baric, R. J. de Groot, C. Drosten, A. A. Gulyaeva, B. L. Haagmans, C. Lauber, A. M. Leontovich, B. W. Neuman, D. Penzar, S. Perlman, L. L. Poon, D. V. Samborskiy, I. A. Sidorov, I. Sola, and J. Ziebuhr, "The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2," *Nature Microbiology*, vol. 5, pp. 536–544, 4 2020.

[2] F. Clement, A. Kaur, M. Sedghi, D. Krishnaswamy, and K. Punithakumar, "Interactive data driven visualization for covid-19 with trends, analytics and forecasting," vol. 2020-September, pp. 593–598, Institute of Electrical and Electronics Engineers Inc., 9 2020.

[3] S. Verma and R. K. Gazara, "Big data analytics for understanding and fighting covid-19," 2021.

[4] I. G. Pereira, J. M. Guerin, A. G. S. Júnior, G. S. Garcia, P. Piscitelli, A. Miani, C. Distante, and L. M. G. Gonçalves, "Forecasting covid-19 dynamics in brazil: A data driven approach," *International Journal of Environmental Research and Public Health*, vol. 17, pp. 1–26, 7 2020.

[5] N. Chintalapudi, G. Battineni, and F. Amenta, "Covid-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in italy: A data driven model approach," *Journal of Microbiology, Immunology and Infection*, vol. 53, pp. 396–403, 6 2020.

[6] D. P. Aragão, D. H. dos Santos, A. Mondini, and L. M. G. Gonçalves, "National holidays and social mobility behaviors: Alternatives for forecasting covid-19 deaths in brazil," *International Journal of Environmental Research and Public Health*, vol. 18, p. 11595, 11 2021.

[7] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," pp. 291–300, IEEE, 5 2019.

[8] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," vol. 28, Curran Associates, Inc., 2015.

[9] Ishwarappa and J. Anuradha, "A brief introduction on big data 5vs characteristics and hadoop technology," vol. 48, pp. 319–324, Elsevier B.V., 2015.

[10] D. Loshin, *Big Data Analytics*. Elsevier, 2013.

[11] B. A. Devlin and P. T. Murphy, "An architecture for a business and information system," *IBM Systems Journal*, vol. 27, pp. 60–80, 1988.

[12] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[13] W. Inmon, *Building the Data Warehouse*. QED Technical Publishing Group, 1992.

[14] K. Corral, D. Schuff, G. Schymik, R. S. Louis, and G. Schymik, "Enabling self-service bi through a dimensional model management warehouse enabling self-service bi enabling self-service bi through a dimensional model management warehouse," 2015.

[15] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," *ACM SIGMOD Record*, vol. 26, pp. 65–74, 3 1997.

[16] A. F. Vermeulen, *Data Science Technology Stack*, pp. 1–13. Berkeley, CA: Apress, 2018.

[17] N. Miloslavskaya and A. Tolstoy, "Big data, fast data and data lake concepts," vol. 88, pp. 300–305, Elsevier B.V., 2016.

[18] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, "Data lake management: Challenges and opportunities," vol. 12, pp. 1986–1989, VLDB Endowment, 2018.

[19] T. John and P. Misra, *Data Lake for Enterprises*. Packt Publishing, 5 2017.

[20] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," 8 2011.

[21] H. J. Watson, "Tutorial: Business intelligence - past, present, and future," *Communications of the Association for Information Systems*, vol. 25, pp. 487–510, 2009.

[22] J. B. de Vasconcelos and Álvaro Rocha, "Business analytics and big data," *International Journal of Information Management*, vol. 46, pp. 320–321, 6 2019.

[23] T. Sakao and A. Neramballi, "A product/service system design schema: Application to big data analytics," 4 2020.

[24] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," *Information Systems and e-Business Management*, vol. 16, pp. 547–578, 8 2018.

[25] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big data analytics: Computational intelligence techniques and application areas," *Technological Forecasting and Social Change*, vol. 153, p. 119253, 2020.

[26] A. Beheshti, B. Benatallah, R. Nouri, V. M. Chhieng, H. Xiong, and X. Zhao, "Coredb: A data lake service," vol. Part F131841, pp. 2451–2454, Association for Computing Machinery, 11 2017.

[27] R. Hai, S. Geisler, and C. Quix, "Constance: An intelligent data lake system," vol. 26-June-2016, pp. 2097–2100, Association for Computing Machinery, 6 2016.

[28] B. D. Wissel, P. J. V. Camp, M. Kouril, C. Weis, T. A. Glauser, P. S. White, I. S. Kohane, and J. W. Dexheimer, "An interactive online dashboard for tracking covid-19 in u.s. counties, cities, and states in real time," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1121–1125, 7 2020.

[29] G. Agapito, C. Zucco, and M. Cannataro, "Covid-warehouse: A data warehouse of italian covid-19, pollution, and climate data," *International Journal of Environmental Research and Public Health*, vol. 17, pp. 1–22, 8 2020.