Exploring the Role of Data Analytics in Combating Future Pandemics: Lessons Learned from the COVID-19 Pandemic in the United Kingdom

Interim Report

Stewart Charles Fisher II
ID: 25020928
25020928@students.lincoln.ac.uk

January 2023

UNIVERSITY OF
LINCOLN

School of Computer Science
University of Lincoln
United Kingdom

Submitted in partial fulfilment of the requirements for the Degree of BSc(Hons) Computer Science

Supervisor Dr. Kamaran Fathulla

Word Count: 926

Abstract

A review of literature pertaining to the essence of my project, detailing techniques that have been used to a similar effect, combined with an update of my current progress on my project.

Keywords— Big Data Analysis, COVID-19

Contents

1 Literature Review

1.1 Background

Since the beginning of the COVID-19 pandemic in early 2020, the world has undergone a monumental change in how we handle and assess our healthcare data. The speed that coronavirus spread across the country was unprecedented and made this a necessity; in the first week of 2022, the number of new coronavirus cases recorded was approximately 1.25 million cases (Mathieu et al., 2020).

As we move forward, we should look at the effectiveness of techniques that have been used and investigate if there are more efficient methods of data analysis and data collection. By ensuring the robustness of data methodology used for this pandemic, we can allow for more accurate and timely analysis.

The aim of this literature review is to explore techniques that have already been employed to analyse the data across the world to highlight high risk demographics, detailing the strengths and weaknesses of the current methods in order to gain an understanding of potential area of improvement for how we handled the data in the United Kingdom, and to identify possible areas of innovation and development.

1.2 Research Objectives

The questions that will guide my research for this project are:

- How has pre-existing data been collected?
- How has pre-existing data been stored and managed?
- What is the current role that data analytics occupies in modelling and tracking the pandemic?
- Have artificial intelligence and machine learning been utilised to analyse data, and how?

1.3 Previous Studies

1.3.1 Concerning the Handling of the Data

A study in 2020, *A Multi-Dimensional Big Data Storing System for Generated COVID-19 Large-Scale Data using Apache Spark*, outlined how Apache Software Foundation implementations, specifically Apache Spark could be integrated to handle the large quantities of data so that it could be processed efficiently.

Researchers from Mansoura University, Egypt (Elmeiligy et al., 2020) conducted research into the the value of using an ASF framework to analyse data; they used the Hadoop Distributed File System to divide the inserted data into a set of Resilient Distributed Datasets, as seen in Figure 1. HDFS is a distributed file system designed to provide fault-tolerant data management. Spark is an analytics engine designed for fast data processing; it provides a programming interface to allow access to data parallelism[1].

Their research found that by implementing the framework to segment the data, they were able to increase the system performance in their modelling.

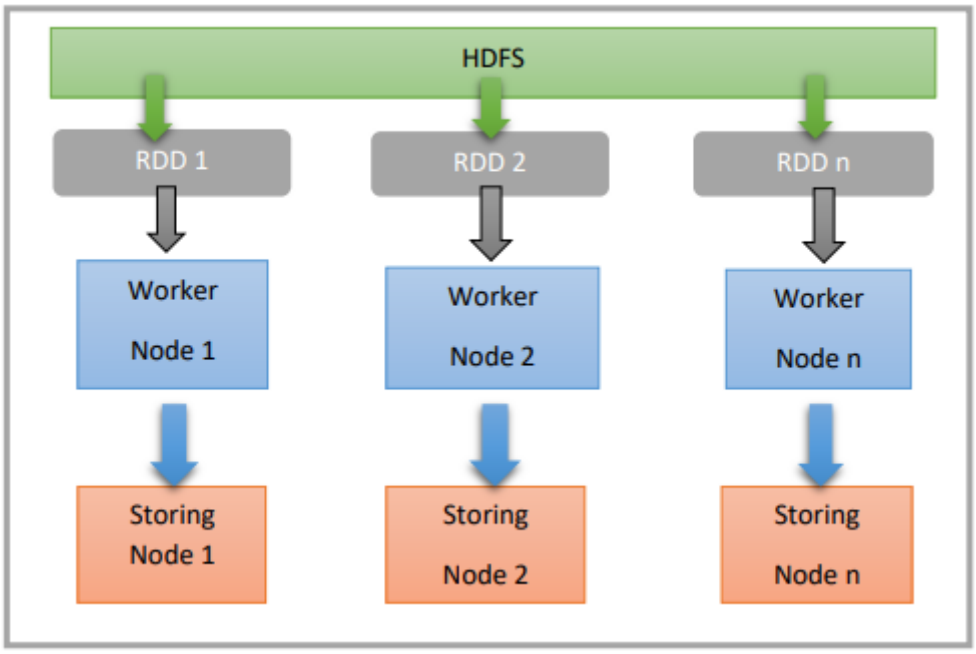

Figure 1: An implementation of an ASF framework (Elmeiligy et al., 2020).

[1]Data parallelism is a technique that distributes a dataset across multiple processors, to allow for faster processing in a smaller time frame.

1.3.2 Concerning the Analysis of the Data

There are a wide variety of methods that have been used to analyse data related to COVID-19; this will likely be due to the variety of the data.

In mid-2020, there was a meta-analysis into the potential of Big Data and Artificial Intelligence in managing the COVID-19 pandemic (Bragazzi et al., 2020), focusing on applications at differing time scales. They determined that for short-term applications, Big Data techniques could be used to monitor the progress of the pandemic outbreak in real-time. The meta-analysis noted a study by a team at the Fogarty International Center (Sun et al., 2020), where they monitored news media and social networks in China to reconstruct the progression of the outbreak in mainland China.

The team cross-referenced reports from DXY.cn[2] with international data sources via media sources (Kyodo News, The Strait Times, and CNN), governments, and official health authorities. They found that while the model worked early on, the overwhelming of healthcare systems and reporting fatigue meant that the data captured from non-traditional sources was becoming diminished over time and couldn't keep up.

Regression analysis is an area of data analysis that has shown some promise; in early 2022, there was a study into the usage of regression analysis for COVID-19 infections and deaths due to issues with access to food, and health (Almalki et al., 2022). The study used a combination of a machine learning regression model and a GIS regression model[3], as seen in Figure 2; the authors used scikit-learn software for the ML regression and ArcGIS-ArcMap software for the GIS analysis. The cases and deaths were taken as the dependent variables and the remaining factors as independent variables. The results showed that while there were instances of strong correlation between the independent variables, there was only weak correlations being presented between the dependent and independent variables, across both methods.

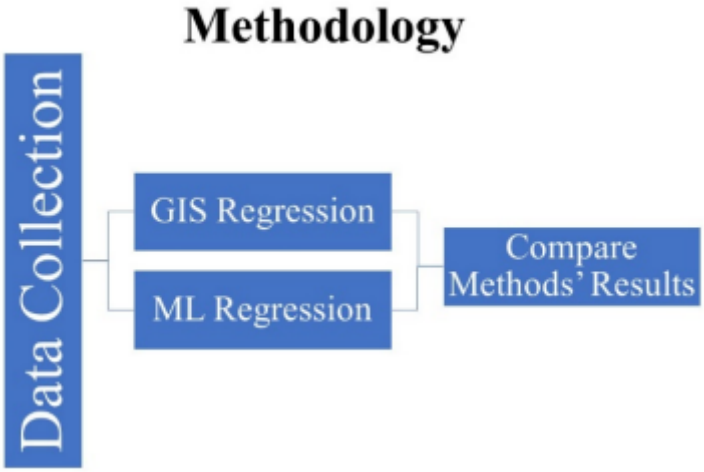

Figure 2: A combination of GIS regression and ML regression (Almalki et al., 2022).

Other forms of regression analysis, beyond linear regression have been found to produce compelling results when identifying relationships within data; in 2020, a study at Yildiz Technical University, Turkey (Qudrat-Ullah & Tsasis, 2017) implemented a negative binomial regression analysis to determine the impact of human mobility on the causality of COVID-19 cases. Unlike the prior study, this study displayed relationships with the dependent variables, showing that there are correlations between COVID-19 cases and the volume of airline traffic, and also the number of airports.

1.4 Conclusion and Future Directions

Upon final summary, it appears that there is a clear potential for Big Data techniques to be integrated into our health crisis management protocols, to observe and track progressions. Regarding the collection of health data, it appears that we could expand the harvesting pool beyond official sources although we should be wary of their limitations. So far, regression models have returned varying qualities in their outputs and so we should aim to compare their efficacy, and aim to observe their performance on different types of data.

Moving forward, the proposed direction of the project will be to complete a multi-discipline investigation into how data can be used in future pandemics:

- The project will aim to build my own model for healthcare data originating from the United Kingdom, using ASF software to manage the data.
- The project will aim to compare different regression analysis models on the harvested healthcare data.
- The project will aim to explore more approaches as to how healthcare data could be collected.

[2]DXY.cn is an online health-focused community for physicians and healthcare professionals in China.

[3]A geographic information system pairs geographic data with software tools to manage and analyse them. (Chang, 2019)

2 References

Almalki, A., Gokaraju, B., Acquaah, Y., & Turlapaty, A. (2022). Regression analysis for covid-19 infections and deaths based on food access and health issues. *Healthcare*, *10*, 324. https://doi.org/10.3390/healthcare10020324

Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., & Wu, J. (2020). How big data and artificial intelligence can help better manage the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, *17*, 3176. https://doi.org/10.3390/ijerph17093176

Chang, K.-T. (2019). *Introduction to geographic information systems*. Mcgraw-Hill Education.

Elmeiligy, M. A., Desouky, A. I. E., & Elghamrawy, S. M. (2020). A multi-dimensional big data storing system for generated covid-19 large-scale data using apache spark. *arXiv:2005.05036 [cs]*. Retrieved January 23, 2023, from https://arxiv.org/abs/2005.05036

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus pandemic (covid-19). *Our World in Data*.

Qudrat-Ullah, H., & Tsasis, P. (2017). *Innovative healthcare systems for the 21st century*. Cham Springer International Publishing.

Sun, K., Chen, J., & Viboud, C. (2020). Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *The Lancet Digital Health*, *2*, e201–e208. https://doi.org/10.1016/s2589-7500(20)30026-1