Songklanakarin J. Sci. Technol.
44 (4), 1119–1124, Jul. – Aug. 2022

http://www.sjst.psu.ac.th

Original Article

Thailand COVID-19 pandemic data analysis using big data technology

Karma Wangchuk[1*], and Jirarat Ieamsaard[2]

[1] *Department of Information Technology, College of Science and Technology, Royal University of Bhutan, Phuentsholing, 21101 Bhutan*

[2] *Department of Electrical and Computing Engineering, Faculty of Engineering, Naresuan University, Mueang, Phitsanulok, 65000 Thailand*

Received: 7 May 2022; Revised: 28 May 2022; Accepted: 9 June 2022

Abstract

The world has been observing unprecedented circumstances with the outbreak of the COVID-19 pandemic. The upsurge of infection cases has left hospitals overwhelmed, education quality compromised, caused unemployment issues, and affected the international economy, tourism sector, and financial markets. The number of confirmed cases and deaths has been increasing on daily basis. However, a vaccination drive has been conducted globally to reduce the transmission. The purpose of this study was to present the basic system configuration of the Big Data technology called Spark and to performe an analysis of Thailand's COVID-19 pandemic data. Data were collected from an open Thai government data center between January 2020 and 28. May 2021. Bangkok province has the most COVID-19 cases followed by Samut Sakhon province. It has been observed that the ages between 20 and 40 were the most infected by COVID-19 in Thailand.

Keywords: coronavirus, Covid-19, big data technologies, hadoop, spark

1. Introduction

The Coronavirus 2019 (COVID-19) outbreak was first reported in Wuhan on 31 December 2019. On 13 January 2020, the first case outside of China was reported in Thailand and thereafter infections were spreading continuously. According to the World Health Organization (WHO, 2020) as of 29 January 2020, 68 cases were confirmed outside of China in 15 countries and 5,997 cases in China. There were 6,065 cases confirmed globally. COVID-19 was spreading fast and on 30. January 2020, the WHO declared it a pandemic and a Public Health Emergency.

COVID-19 is a contagious pandemic disease caused by a coronavirus (WHO, n.d). According to Sauer (2021), COVID-19 is similar to SARS that started in China in 2003 and spread to other countries. SARS is an acronym for severe acute respiratory syndrome. Both COVID-19 and SARS are types of coronaviruses. In December 2019, a cluster of pneumonia cases was announced in Wuhan city, Hubei province of China. It was linked with the Huanan seafood wholesale market that sells live animals (Hui *et al.*, 2020; Nishiura *et al.*, 2020). Early reports from Wuhan suggested that there were only a few cases and no human-to-human transmission. However, there had been human-to-human transmission. COVID-19 is spread by droplets in air. When an infected person coughs or sneezes, the droplets travel a few feet and fall on surfaces. Therefore, masking and social distancing are effective measures to prevent the virus from spreading (Sauer, 2021). To impede local transmission, WHO suggested washing hands regularly with soap and sanitizing with alcohol-based solutions. Further, people were advised to maintain at least a one-meter distancing, stop touching their face, and cover their mouth and nose while coughing and sneezing. Moreover, unnecessary traveling and mass gatherings were discouraged, and home quarantine was suggested.

The scientific community was alerted to a possible pandemic with human-to-human transmission (Nishiura *et al.*, 2020). The outbreak began in early January and the first case was registered in Thailand outside of China (Hui *et al.*, 2020).

*Corresponding author
Email address: karma.cst@rub.edu.bt

The first case in Thailand reported on 13. January 2020 was a Chinese tourist (Durkee, 2020; Hui *et al.*, 2020). According to the Department of Disease Control (DDC, 2020a), and intensive surveillance protocol and screening of travelers from Wuhan to Thailand international airports had been implemented between 3. And 14. of January. A total of 70 flights comprising 11,163 flight crew members and passengers were tested for respiratory symptoms and febrile illness (DDC, 2020a). However, 15 people had the symptoms and only one case was identified as COVID-19 by laboratory results. The Ministry of Public Health and DDC have strengthened Emergency Operation Centers. The passengers at Suvarnabhumi, Phuket, Chiang Mai, and Don Mueang airports have been closely monitored by using thermal scanners. Subsequently, the cases of COVID-19 in Thailand increased. However, Thailand saw 100 days without COVID-19 cases, but on 24. May, the first case of community transmission was confirmed (Durkee, 2020). By the end of 2020, there were 6,884 confirmed cases in Thailand and 83,135,180 cases globally. The death toll increased to 61 in Thailand and to 1,813,389 internationally (DDC, 2020b).

Similarly, the outbreak rapidly escalated around the world and incurred unprecedented challenges. The rapid spread of coronavirus overwhelmed the world. The precise behavior of the epidemic and prompt health measures against it were unknown, causing unprecedented challenges to WHO and healthcare systems. The available data were evaluated by WHO to study the most effective methods to provide care (Wolkewitz & Puljak, 2020). The researchers, healthcare workers, and decision-makers have been coming together to solve this enormous challenge. Over the months, Covid-19 clinical data were made public for information and research. Artificial Intelligence was applied to analysis of Covid-19 Big Data (Hussain, Bouachir, Al-Turjman, & Aloqaily, 2020).

In recent times, Big Data technologies are used for analysis and research. Different technologies have been evolving to process data seamlessly. According to (DataMites, 2020), data manipulation tools such as Pandas, NumPy, and Dask can process enormous datasets. The open-source Pandas is powerful, flexible, and fast in processing up to 5GB of data. Furthermore, using chunk size, Pandas can process up to 30GB of data, overcoming the physical limitations of hardware memory. However, the data have to be loaded again if an error occurs in between processing and analysis. The Dask supports parallel computing that is developed with Pandas, NumPy, and Scikit-Learn community projects. The Dask processes between 30GB and 200GB of data. However, there are large data beyond what Pandas and Dask could handle. Spark is one of the Big Data technologies trending in the markets for data analysis. Spark can process 1000GB to 1TB and 1000TB to 1PB dataset sizes (DataMites, 2020).

A colossal amount of data is generated every second. The evolution of technology, such as of the telephone to the mobile phone, desktop to cloud computing, car to smart car, and sensors of IoT, have contributed to Big Data generation (Edureka, 2019). Big Data is defined by volume, variety, velocity, value, and veracity (5Vs) of data that have heterogeneous sources consisting of structured, unstructured, and semi-structured formats (Oussous, Benjelloun, Ait Lahcen, & Belfkih, 2018). The frontrunner companies such as Google, Facebook, Samsung, and social media giants Twitter, and Instagram generate a large volume of data every day. According to Edureka (2019), every second Facebook has 695,000+ status updates, and there are 100,000+ tweets, 11,000,000+ instant messages, 698,445+ Google searches, 168,000,000+ emails, and 217+ new mobile users. Similarly, hospitals generate humongous health data (Shilo, Rossman, & Segal, 2020). Advanced technologies are used to process and analyze these data.

The traditional approaches to data analysis are not efficient in processing heterogeneous data from different sources. According to Oussous *et al.,* (2018), traditional tools and platforms lack scalability, have slow responsiveness, performance issues, and poor accuracy. However, there are frameworks and models to deal with Big Data. Technologies such as Hadoop and Spark have more storage and parallel processing to address problems of the traditional tools. Hadoop is a Big Data framework that stores data in a distributed environment and processes it parallelly (Sinha, 2021). It has two fundamental components: Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator (YARN). HDFS handles the storage of heterogeneous data in different clusters, whereas YARN manages resources and allows parallel processing. HDFS has three components: Name Node, Secondary Name Node, and Data Nodes that work based on master-slave architecture. The Name Node is a master daemon that manages and maintains Data Nodes. It stores metadata such as the location of blocks stored, file permissions, size, etc. Data Nodes are slave daemons that store actual data and perform read and write operations. The Secondary Name Node is used in failover situations. It has edit logs and FsImage that have information on changes in the Name Node and Data Nodes. Similarly, YARN has two main components: Resource Manager (master) and Node Managers (slave). The Resource Manager receives processing requests and then passes them to Node Managers. The Node Managers execute tasks in the Data Nodes. There are tools to support HDFS and YARN for storage and processing such as Sqoop, Flume, Kafka, Talent, Spark Streaming, Hive, Pig, Mahout, R Connectors, Ambari, Cloudera, Zookeeper, Oozie, Hbase, Cassandra, MongoDB, and DynamoDB.

However, Hadoop is slow to process data and requires comparatively many lines of code. Moreover, it performs only batch processing. In Hadoop, data are read from disk and written to HDFS. In the next iteration, data are read from HDFS for processing and the final result, after performing all the required iterations, is written back to the disk (Maheshwar & Haritha, 2017). This process requires a lot of disk space and execution time. Industries demand a more powerful processing engine that can process data in both real-time and also batches more quickly. Spark was developed to overcome the problems of Hadoop. It is a powerful open-source engine that provides real-time stream processing, interactive processing, graph processing, in-memory processing as well as batch processing with very fast speed, ease of use, and standard interface (Data Flair, 2021). Apache Spark is 100 times faster than Hadoop (Data Flair, 2021; Databricks, 2013; Maheshwar & Haritha, 2017). The essential components of Apache Spark are Spark Core API, Spark SQL, Spark Streaming, MLlib, GraphX, and SparkR. In Spark, data are executed in memory, contrary to Hadoop. The main unit of data in Spark is Resilient Distributed Dataset (RDD). RDD is a collection of elements distributed across

cluster nodes that is immutable and can perform parallel operations (Data Flair, 2021). There are two RDD operations: Transformations (creates new RDDs) and Actions (execute tasks and return results).

2. Materials and Methods

There are three phases in this study: data acquisition, Big Data system configuration, and the Thailand COVID-19 data analysis. COVID-19 data between January 2020 and 28. May 2021 were collected from the Open Government Data of Thailand website (DDC, 2021a). The Datacenter provides daily COVID-19 reports that are maintained by the Department of Disease Control. A data record consists of patient code, age, sex, nationality, province of isolation, notification date, announcement date, province of onset, district of onset, and quarantine. The patient codes were assigned as integer numbers starting from 1 where 1 was given to the first person who was infected with the COVID-19 and number 2 was given to the second person with the COVID-19 infection. Similarly, the rest of the patients were each assigned a number according to the order in which they were detected positive for COVID-19. There are missing data from the raw dataset. The province of isolation and announcement date had missing data that were replaced with the same values as in cells above and below. These cells had the same province name and date of the announcement. However, variables such as the province of onset and district of onset had many missing data. These variables were not considered for analysis and were deleted. However, the variable nationality had a few missing data. These missing data were either replaced or deleted from the dataset and then analyses were conducted on the preprocessed dataset.

The Spark environment set up in Windows 10 is shown in Figure 1. There are five steps to set up the Spark in the local system. Anaconda individual edition is required for the install, followed by Java and Spark installations. The system and user variables for both Java and Spark are set up as shown in Figure 1. To allow shell command and local files accessibility by the spark, *winutils* is downloaded and placed in the Spark Bin directory. Then open the *anaconda* prompt and install the *findspark* package that allows the Spark to integrate with the Jupyter notebook. Finally, open the Jupyter notebook and import the necessary libraries for running Spark. Spark is based on master-slave architecture. Spark applications can be launched either in local or cluster mode. There are two types of local mode: interactive and batch mode. Similarly, the cluster has an interactive and batch mode. However, the interactive mode is used for development and learning purposes, whereas the batch mode is used for production. Furthermore, cluster mode supports three types of cluster managers such as YARN, Mesos, and standalone.

In this study, apache-spark standalone is used. Figure 2 shows the architecture of Spark Standalone. It consists of a master and workers (slaves) with CPU cores and memory configurations. The master is the cluster manager that has the driver program and Spark context. The Spark application (code) is the driver program that creates Spark context. The Spark context is the entry point for program execution. The basic data object in Spark is the Resilient Distributed Dataset (RDD) that is created in the Spark context. RDD holds data and is partitioned for distributed computing and parallelism. RDD performs two types of operations: transformations and actions. A new RDD is created when a transformation is performed. However, results are generated when actions are performed. Similarly, worker nodes consist of executors and tasks. Worker nodes execute tasks assigned by the master and return to the Spark context.

Overall, data from the external sources are loaded in the Spark and called RDD. The master divides RDD into partitions and distributes them to worker nodes for parallel computation. The number of worker nodes can be increased based on CPU core count and memory amount. Workers execute all the tasks and return the results to the master as shown in Figure 2.

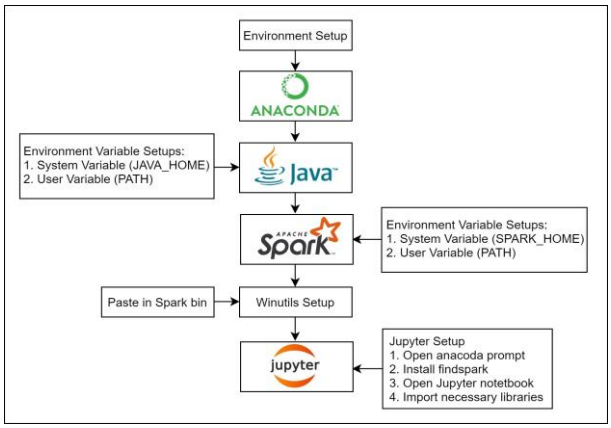

Figure 1. Spark environment setup in Windows

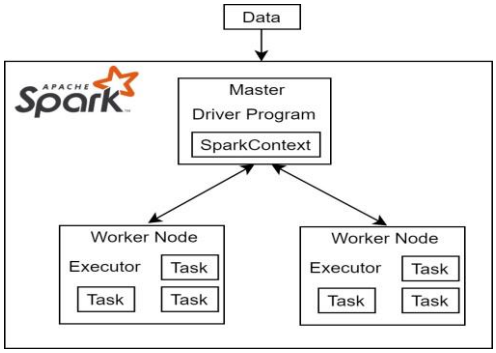

Figure 2. Spark standalone architecture

3. Results

In this study, COVID-19 pandemic data of Thailand between January 2020 and 28. May 2021 were analyzed. On 28. May 2021, 169,671,394 COVID-19 cases were confirmed and 3,526,317 deaths globally (DDC, 2021b). However, Thailand confirmed 144,976 cases and 954 deaths on the same day. Table 1 shows the monthly confirmed and death cases in 2020 and 2021. The confirmed cases increased from 111 in January 2020 to 2,862,048 in May 2021. Similarly, the number of recoveries increased from 54 to 1,819,029. However, there were no death cases in the first two months in Thailand, but the number increased gradually in the following months. The highest number of death cases registered was 15,724 on 28. May 2021.

Table 1. The monthly COVID-19 cases between January 2020 and 28 May 2021

Year	Month	Confirmed	Recovered	Deaths
2020	January	111	54	0
	February	933	439	0
	March	13480	1830	72
	April	77098	47919	1181
	May	93701	88276	1730
	June	93991	90037	1739
	July	100371	95848	1798
	August	104473	98900	1798
	September	100985	96278	1693
	October	113857	107707	1829
	November	116285	111255	1795
	December	150944	123789	1863
2021	January	371766	249611	2163
	February	676844	580825	2267
	March	847265	816312	2737
	April	1243747	886057	3442
	1-28 May	2862048	1819029	15724

In 2020, the total number of confirmed cases, recovered, hospitalized, and death cases were 966,229, 862,332, 88,399, and 15,498, respectively. However, the total confirmed cases in the first five months of 2021 are approximately six fold the total cases confirmed in 2020. Furthermore, cases of hospitalization dramatically increased by approximately 18 fold in 2021. However, the rate of recovery was about five fold larger and death cases increased by one in 2021. Overall, we can expect the number of cases and deaths to increase. Figure 3 illustrates the cases in Thailand over 17 months. In January 2020, the confirmed cases were 111, while 54 were declared as recovered and zero deaths. However, cases sharply increased in the following two months. Nevertheless, the number of people hospitalized dropped below 10,000, whereas confirmed and recovered cases remained over 100,000. In March 2020, 72 deaths were reported and this increased over the month. Furthermore, cases have been rising from October 2020, increasing to 11 fold by May 2021.

The males, females, and others of all the nationalities who tested positive were 47.0%, 46.1%, and 6.9% respectively. Similarly, only Thai males, females, and others were 49.2%, 48.9%, and 1.9% respectively. Overall, the proportion of males who tested positive is slightly higher than that of females. However, "others" gender witnessed a significant rise of 5%. Figure 4 shows the numbers of COVID-19 cases across different age groups. The youngest who tested positive was 12 months old, and the oldest was 100 years. Most of the cases were found between the ages 20 and 40 years. The most cases were observed with age 30 and fewer cases of people above 80 years. The working-age population group is more prone to test positive compared to economically inactive people and children.

Figure 5(a) illustrates 10 provinces with the most COVID-19 cases in Thailand. Bangkok had the highest number of cases followed by Samut Sakhon. On 23. December 2020, 1,202 cases were confirmed in migrant workers and locals residing around the Central Shrimp Market in Samut Sakhon province. The reason for increased cases in the province was due to the commercial hub and migrant workers. Similarly, Figure 5(b) shows the 10 least infected provinces. The least number of cases confirmed is in Bueng Kan province at 27. Uthai Thani and Mukdahan are the second and third infected provinces respectively.

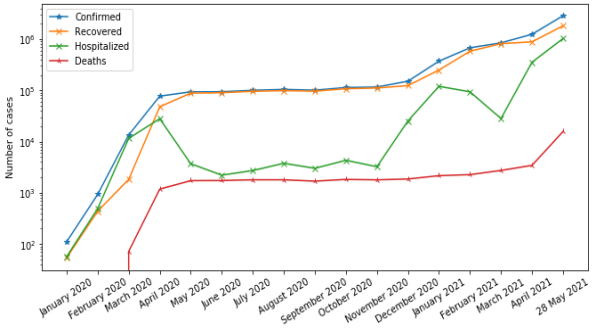

Figure 3. Time profile of COVID-19 cases by month

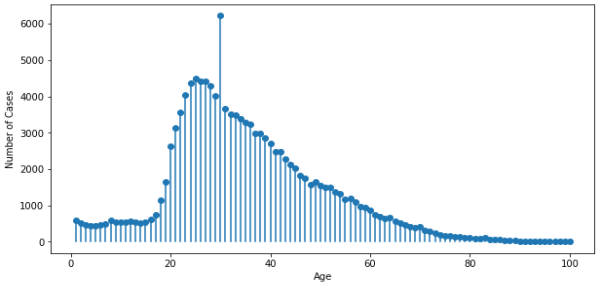

Figure 4. Number of COVID-19 cases by age

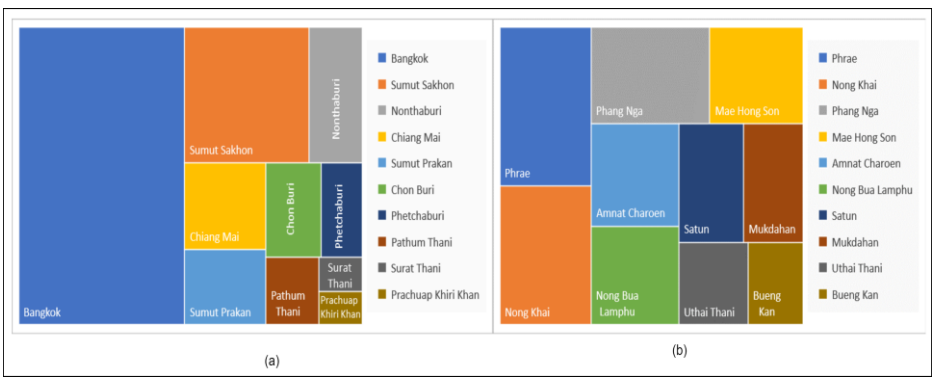

Figure 5. COVID-19 cases: (a) provinces with the most cases, (b) provinces with the least cases

4. Discussion

The adverse scenarios have been observed in the education sector, healthcare system, financial and tourism sector, and within the family. COVID-19 variants have been reported in many countries, including India. An upsurge of confirmed cases and deaths has been reported globally, when experiencing either the second or third wave of COVID-19. However, a vaccination drive has been conducted and the healthcare system understands the measures against COVID-19.

In this study, Thailand COVID-19 pandemic data were analyzed using big data technology called Spark, locally configured in standalone mode. Data consisted of records of COVID-19 cases between January 2020 and 28. May 2021. The first case of COVID-19 outside of China was found in Thailand on 13. January 2020. Despite the measures put in place by the Department of Disease Control, cases have been rising. There was a sharp increase in February and March 2020. However, then the count remained without a significant increase up until October 2020. The number of confirmed cases drastically rose in 2021 overwhelming the healthcare system in each province.

The highest numbers of COVID-19 cases were found in Bangkok and Samut Sakhon provinces. It was found that commercial hubs and entertainment places pose a higher risk of infecting people. Furthermore, people aged between 20 and 40 were more infected. However, there is no difference between males and females in infected counts. Bueng Kan and Mukdahan are the two least infected provinces in Thailand. The analysis suggests that the government needs to prioritize the vaccination drive in the most infected provinces first, as shown in Figure 5, then followed by the rest of the provinces. Mutated variants of the COVID-19 are well known. Therefore, COVID-19 protocols need to be followed to break down the community transmission chain by staying at home, maintaining social distancing, and regularly washing and sanitizing hands.

5. Conclusions

The COVID-19 has changed the daily routine in unprecedented ways. However, with advancements in technology scientists were able to make quick and informed decisions. In record short time, COVID-19 vaccines were pursued and rolled out to the public. Past research has shown that the analysis of huge data has contributed to making clinical decisions. In this study, we configured Spark technology in Windows for Big Data analysis and conducted a COVID-19 data analysis of Thailand. However, obtaining the latest and most accurate data has been an issue. The data updated on the DDC website is not real-time, and deaths that have occurred at homes are not reported. Moreover, having to translate Thailand COVID-19 status to English has been an issue. Furthermore, an analysis of the latest dataset of Thailand COVID-19 and finding cases before and after vaccination would be interesting study topics.

Acknowledgements

We would like to extend our sincere appreciation to Mr. Chatchai Saikhumton, master's student at Naresuan University, Faculty of Engineering, Department of Electrical and Computer Engineering, Computer Vision Lab, for providing us with timely translated data.

References

Data Flair. (2021). *What is spark - Apache spark tutorial for beginners*. Retrieved from https://data-flair.training/blogs/what-is-spark/

Databricks. (2013). *What is apache spark - Benefits of apache spark.* Retrieved from https://databricks.com/spark/about

DataMites. (2020, January 23). *Pandas limitations - Pandas vs Dask vs PySpark - DataMites Courses*. Retrieved from https://www.youtube.com/watch?v=YLg4vuIADnQ

Department of Disease Control. (2020a, January 14). *Novel coronavirus 2019 Pneumonia Situation-Thailand report on January 14, 2020*. Retrieved from https://ddc.moph.go.th/viralpneumonia/eng/file/situation/situation-no11-140163.pdf

Department of Disease Control. (2020b, December 31). *The coronavirus disease 2019 situation-Thailand situation update on 31 December 2020*. Retrieved from https://ddc.moph.go.th/viralpneumonia/eng/file/situation/situation-no357-311263.pdf

Department of Disease Control. (2021a). *COVID-19 daily report Thailand information*. Retrieved from https://data.go.th/en/dataset/covid-19-daily

Department of Disease Control. (2021b). *The coronavirus disease 2019 situation-Thailand situation update on 28 May 2021*.

Durkee, A. (2020). *Thailand Sees first local coronavirus case in 100 days.* Forbes. Retrieved from https://www.forbes.com/sites/alisondurkee/2020/09/03/thailand-first-local-coronavirus-case-in-100-days/?sh=55602edb59a4

Edureka. (2019, September 1). *Big data and hadoop full course*. Retrieved from https://www.youtube.com/watch?v=1vbXmCrkT3Y&t=4589s

Hui, D. S., Azhar, E. I., Madani, T. A., Francine Ntoumi, Kock, R., Dar, O., . . . Petersen, E. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China: Ijidonline.Com. Retrieved from https://www.ijidonline.com/article/S1201-9712(20)30011-4/abstract

Hussain, A. A., Bouachir, O., Al-Turjman, F., & Aloqaily, M. (2020). AI Techniques for COVID-19. *IEEE Access*, *8*, 128776–128795. doi:10.1109/ACCESS.2020.3007939

Maheshwar, R. C., & Haritha, D. (2017). Survey on high performance analytics of bigdata with apache spark. *Proceedings of 2016 International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2016*, 721–725. doi:10.1109/ICACCCT.2016.7831734

Nishiura, H., Jung, S., Linton, N. M., Kinoshita, R., Yang, Y., Hayashi, K., . . . Akhmetzhanov, A. R. (2020). The extent of transmission of novel coronavirus in Wuhan, China. *Journal of Clinical Medicine 2020, 9*(2), 330. doi:10.3390/JCM9020330

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/10.1016/J.JKSUCI.2017.06.001

Sauer, L. (2021, May 19). *What is Coronavirus?* Johns Hopkins Medicine. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

Shilo, S., Rossman, H., & Segal, E. (2020). Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine*, *26*(1), 29–38. doi:10.1038/s41591-019-0727-5

Sinha, S. (2021, July 7). *Introduction to big data and hadoop*. Retrieved from https://www.edureka.co/blog/what-is-hadoop/

Wolkewitz, M., & Puljak, L. (2020). Methodological challenges of analysing COVID-19 data during the pandemic. *BMC Medical Research Methodology*, *20*(1). Retrieved from https://doi.org/10.1186/S12874-020-00972-6

World Health Organization. (August 22, 2021). *Coronavirus*. Retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_1

World Health Organization. (2020). *Novel coronavirus (2019-nCoV) situation report - 9.* Retrieved from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200129-sitrep-9-ncov-v2.pdf?sfvrsn=e2c8915_2