Journal of Computer Science & Computational Mathematics, *Volume 13, Issue 3, September 2023*
DOI: 10.20967/jcscm.2023.03.003

Performance Analysis of Apache Hadoop Using Hive on COVID19 Datasets

Mohamed Faris Laham[1], Shafinah Kamarudin[2,*], Nor Asilah Wati Abdul Hamid[1,2], Zurita Ismail[1], Siti Nur Fathiah Ainuddin[2]

[1]Laboratory of Computational Sciences & Mathematical Physics, Institute for Mathematical Research, Universiti Putra Malaysia
[2]Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
[*]Corresponding author email: shafinah@upm.edu.my

Abstract**:** This study aimed to investigate the performance of Apache Hadoop for use in analysing COVID-19 data from the Vaccine Adverse Event Reporting System 2021 in the United States. Apache Hadoop and Apache Hive were employed to analyse the performance of Hadoop for processing COVID-19 health-related information. Furthermore, a Hive script was created using Derby's meta store to measure the Apache Hadoop's execution time, time reduction, and speedup. Different sizes of COVID-19 datasets and various virtual core numbers were used to ascertain Apache Hadoop's best performance, with the datasets being of gigabyte size. The findings show that for all dataset sizes, Apache Hadoop's execution time lowers while increasing the number of cores used. When working with larger datasets, Apache Hadoop's ability to reduce processing time compared to a serial approach was apparent. In addition, when the number of cores increased, Apache Hadoop's speedup on Hive increased for nearly every input. This study revealed that 86,6406 persons in the United States were recorded in the COVID-19 dataset, of whom 46% received the Moderna vaccine and 45% received the Pfizer vaccine. 1.14 percent of male patients and 2.65 percent of female patients perished. Additionally, 48,524 people reported experiencing chills as a symptom. Overall, this study proves that Apache Hadoop and Hive are efficient in both execution time and the ability to manage higher data volumes while also providing a significant speedup. For future studies, other performance metrics can be explored.

Keywords**:** Apache Hadoop, Apache Hive, COVID-19, health information, vaccine.

1. Introduction

The fast development of digital technology and the internet has increased the significance of big data in recent years. The demand for efficient data analysis tools has increased due to the exponential increase in data created. As a result, a number of big data technologies have been developed, including Apache Hadoop and Apache Spark, which can effectively handle and analyze enormous amounts of data [1].

According to [2] and [3], Hadoop MapReduce is a well-known distributed processing system which is applied to examine massive datasets. It separates the data into manageable chunks and disperses it across a group of processors which can perform concurrent data analysis [4]. On the other hand, the in-memory processing engine Spark is far quicker than MapReduce in performing calculations; for more information, see [5]. A data warehousing tool, Apache Hive, offers a SQL-like interface for searching and analyzing massive datasets kept in Hadoop; for further information, see [6]. It is built on top of Hadoop and enables SQL queries for data access and manipulation. Metadata is stored in the Derby meta store, an integrated meta store that is utilized by the Hive services.

Big data technologies must function well in practical contexts, such as the COVID-19-related worldwide disaster [7]. Therefore, it is crucial to assess how well these technologies perform in various circumstances, such as those with diverse dataset sizes and processing loads [8]. In the future, this can assist in improving the performance of the technologies by identifying their shortcomings.

The SARS-CoV-2 virus, which causes the infection known as Coronavirus Disease (COVID-19), is a popular subject matter which is relentlessly discussed over social media [9]. COVID-19 is a highly contagious disease which can cause severe illness or death, regardless of age, although the majority of those infected with the virus will only develop mild to moderate respiratory symptoms and recover without need for any specific treatment [10]. Common symptoms of the COVID-19 virus are cough, tiredness, flu, and loss of smell or taste, and some of the less common symptoms include are sore throat, diarrhoea, headache, and irritated eyes [11].

The present study evaluated the performance of Apache Hadoop and conducted data analysis on COVID-19 datasets using two methods: Apache Hadoop and Apache Hive. Apache Hive was selected due to its faster execution and high-quality results for small and large datasets compared to Apache Pig [12]. We used Derby as the embedded meta store to run alongside Hive services for metadata storage. Four different sizes of COVID-19-related datasets, each with a varying number of cores - 1, 2, 4, and 8 were used to evaluate the performance of Apache Hadoop. The Vaccination Adverse Event Reporting System (VAERS) dataset, a global warning system for vaccination safety issues in the United States was also used.

The evaluation of Apache Hadoop's functionality and the analysis of COVID-19 data are two of the larger research goals related to studying the VAERS dataset. The project intends to illustrate the potential of these technologies for evaluating large-scale health-related data, such as COVID-19 data, by studying the VAERS dataset using Apache Hadoop and Apache Hive. The examination of the VAERS dataset also aids in the discovery of patterns and trends in the use of vaccines, the occurrence of symptoms, and other health-related data that can assist in guiding future healthcare policies and procedures.

This paper is divided into a number of sections to facilitate the readers' understanding. In Section 2, the study's methodology is briefly explained. Section 3 presents the experiment's findings and analyses, and Section 4 presents the conclusions of the study.

2. Methodology

The flowchart in Figure 1 provides a detailed description of the experimental methodology applied. Ubuntu, Apache Hadoop, and Apache Hive are the three software packages used in the present study. The Malaysian Ministry of Health's GitHub repository and Kaggle were used to collect datasets pertinent to COVID-19 for analysis. The performance analysis was performed using Apache Hive and HiveQL, a simple SQL-like language. To evaluate Apache Hadoop's performance with various dataset sizes and core counts, three performance indicators were employed. The first indicator was execution time, this measured the total time it took Hadoop to perform data processing. The second indicator was time reduction, which measures the time it takes to process a dataset using Hadoop as opposed to other techniques. The third indicator was speedup, which measures how quickly Hadoop processes a dataset in comparison to other techniques. It is possible to evaluate Hadoop's efficiency in handling massive datasets and assess how it performs in comparison to other approaches through the analysis of these indicators.

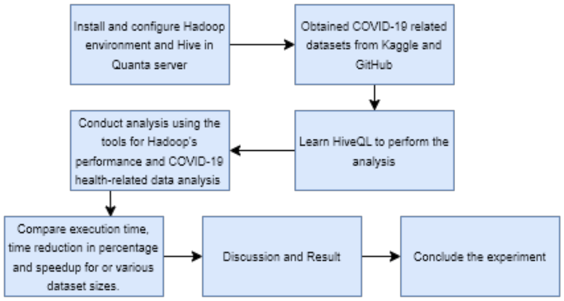
Install and configure Hadoop environment and Hive in Quanta server
Obtained COVID-19 related datasets from Kaggle and GitHub
Learn HiveQL to perform the analysis
Conduct analysis using the tools for Hadoop's performance and COVID-19 health-related data analysis
Compare execution time, time reduction in percentage and speedup for or various dataset sizes.
Discussion and Result
Conclude the experiment

Figure 1. Flowchart

2.1 Setting Up Software

Ubuntu	Ubuntu version 20.04.5 in the Quanta server is used to support Apache Hadoop 3.3.4 and Apache Hive 3.1.3 to ensure all the software runs smoothly.
Apache Hadoop	To use the newest feature of Apache Hadoop, version 3.3.4 of Hadoop is installed.
Apache Hive	For Apache Hive to function properly on Apache Hadoop 3.3.4, a compatible version of Apache Hive, which is 3.1.3, is installed.

2.2 Performing Analysis

Three performance metrics were chosen as analysis tools, namely execution time, time reduction, and speedup. The formula used to calculate the time reduction and speedup of the tool is as follows:

$$\textbf{Time reduction (\%)} = \frac{T_{serial} - T_{parallel}}{T_{parallel}} \tag{1}$$

$$\textbf{speedup} = \frac{T_{serial}}{T_{parallel}} \tag{2}$$

where T is execution time.

3. Results & Discussion

The execution time as defined within the present experimental parameters is the amount of time needed for each process to complete its job. As previously indicated, the experiment consisted of four different input sizes: 1.1 MB, 1.07 GB, 3.03 GB, and 5.07 GB. These input sizes are executed on four virtual cores: 1, 2, 4, and 8.

Table 1 and Figure 2 shows Apache Hadoop's execution time. The results demonstrated that the execution time decreases when moving from 1 core to 2 cores for all dataset sizes. Notably, the graph indicates a significantly higher reduction in larger dataset sizes (1.07 GB, 3.03 GB, and 5.07 GB) compared to the small 1.1 MB dataset.

Table 1. Execution Time Using Hive in seconds.

Input Size	1	2	4	8
1.1 MB	19.748	19.178	19.145	19.080
1.07 GB	22.842	21.901	21.357	21.120
3.03 GB	29.949	28.666	28.897	27.980
5.07 GB	68.107	66.246	66.378	65.198

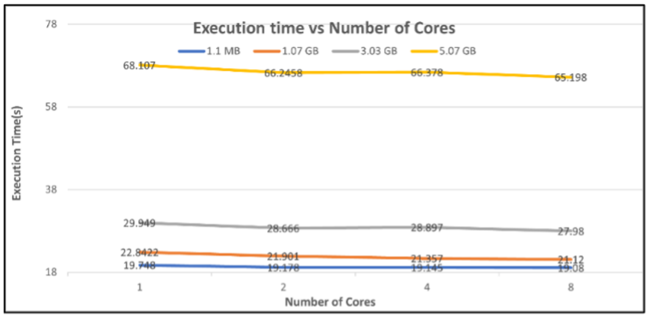

Figure 2. Execution Time vs Number of Cores

Meanwhile, Table 2 and Figure 3 provides a more accurate illustration of the performance differences when using different datasets sizes. Time reduction is defined in this study as the difference in execution time between parallel and serial processes. The time reduction does increase from 2.89% to 3.38%, for the 1.1 MB dataset size. However, for gigabyte-sized datasets, the time reduction was far greater, reaching up

Table 2. Time Reduction Using Hive in percentage.

Input Size	2	4	8
1.1 MB	2.89	3.05	3.38
1.07 GB	4.12	6.50	7.54
3.03 GB	4.28	3.51	6.57
5.07 GB	2.73	2.54	4.27

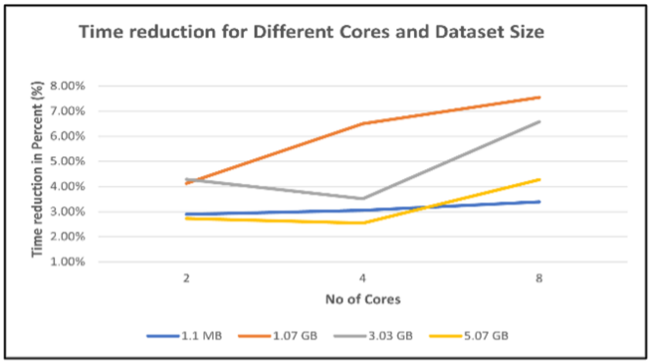

Figure 3. Time reduction vs Number of Cores

to 7.54%. Thus in general, larger datasets showed improved time reduction.

Finally, Table 3 and Figure 4 shows that as the number of cores increases, the speedup of Apache Hadoop employing Hive for nearly all input sizes also showed an increase. However, in some cases, when several programmes are running on the processors simultaneously, the speedup sometimes reduced. For instance, the speedup with two cores drops from 1.045 to 1.036 when a process with four cores was 3.07 GB. Compared to datasets of 1.1 MB, which showed a flat rate of speedup improvement, whereas datasets in the gigabyte range of size have better speedup, as illustrated by the graph.

Table 3. Speedup of Apache Hive in seconds.

Input Size	1	2	4	8
1.1 MB	1	1.029	1.031	1.035
1.07 GB	1	1.043	1.07	1.08
3.03 GB	1	1.045	1.036	1.07
5.07 GB	1	1.03	1.028	1.045

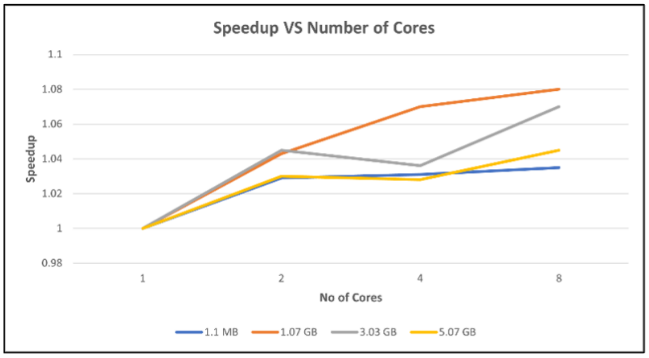

Figure 4. Speedup of Apache Hadoop

4. Data Analysis

Some data analysis was performed based on the data from VAERS USA 2021. Figure 5 shows that out of 89,0837 people in the VAERS data, Moderna and Pfizer-Biontech were the two most inoculated COVID-19 vaccines, at 46% and 45%, respectively. Chills were the most common COVID-19 symptom affecting 48,524 patients as seen in Figure 6.

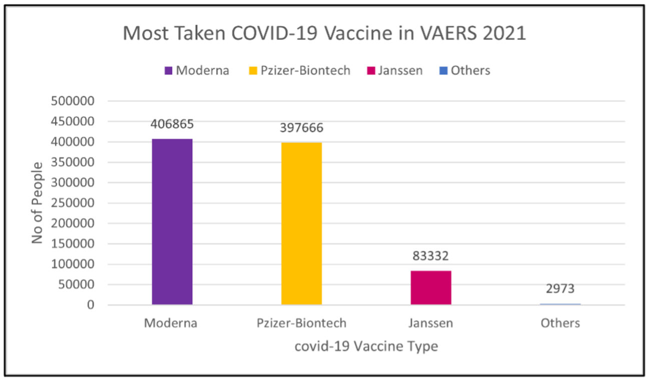

Figure 5. Most Taken COVID-19 Vaccine in US based on VAERS 2021.

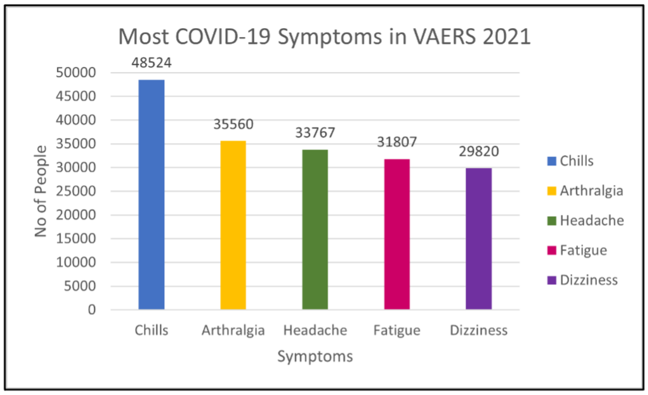

Figure 6. Most COVID-19 symptoms is based on VAERS 2021.

There were 61,4016 male and 25,2390 female COVID-19 patients in the VAERS dataset, and the percentage of victims who died was 1.14 percent for male patients and 2.65 percent for female patients. As shown in Figures 7 and 8, hypertension was a regularly occurring medical condition for 2419 male and 1932 female distinct patient medical medical histories for the male and female patients, respectively.

5. Conclusion

In conclusion, the current study has shown evidence that the performance of Apache Hadoop is enhanced when gigabyte-sized datasets were used. Based on the performance analysis, using just two cores is sufficient for datasets up to 5 GB in size, with the percentage decrease in execution time being more prominent as the dataset size increases. The data analysis of the VAERS 2021 dataset revealed that Moderna (46%) and Pfizer (45%) were the two top vaccines in the United States. The top symptoms of COVID-19 infection reported

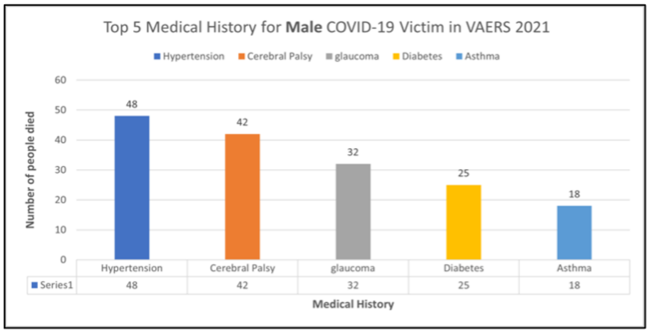

Figure 7. Top 5 Medical History for Male COVID-19 victim in US based VAERS 2021.

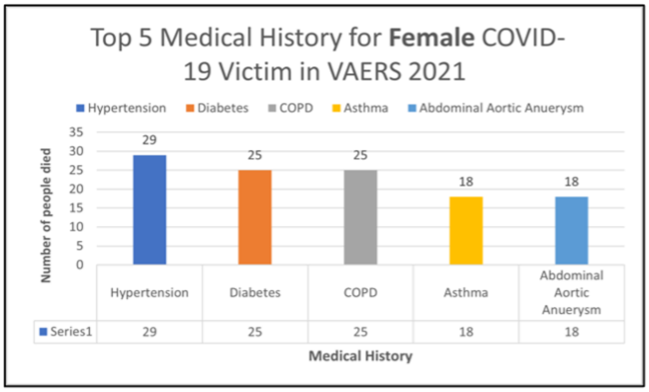

Figure 8. Top 5 Medical History for Female COVID-19 victim in US based VAERS 2021.

was chills, affecting 48,524 people, and the highest fatalities were among patients with a history of hypertension.

Any future studies along these lines should include utilizing MySQL meta store since it allows for multiple Hive sessions, thereby saving more time. Studies could also focus on using larger datasets exceeding 50 GB with a higher allocation of virtual cores to observe more evident differences in performance metrics between multicore processors.

6. Acknowledgement

This research was supported by Universiti Putra Malaysia through Research Grant Scheme, GP/2020/9693400 and partially supported by Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

References

[1] A. Fuad, A. Erwin, and H. P. Ipung, “Processing performance on apache pig, apache hive and mysql cluster,” in *Proceedings of International Conference on Information, Communication Technology and System (ICTS) 2014*. IEEE, 2014, pp. 297–302.

[2] A. P. Rodrigues and N. N. Chiplunkar, “Real-time twitter data analysis using hadoop ecosystem,” *Cogent Engineering*, vol. 5, no. 1, p. 1534519, 2018.

[3] R. Singh and P. J. Kaur, “Analyzing performance of apache tez and mapreduce with hadoop multinode cluster on amazon cloud,” *Journal of Big Data*, vol. 3, no. 1, pp. 1–10, 2016.

[4] Y. Li, “Performance analysis of scheduling algorithms in apache hadoop,” in *2020 16th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2020, pp. 149–154.

[5] M. M. Rathore, H. Son, A. Ahmad, A. Paul, and G. Jeon, “Real-time big data stream processing using gpu with spark over hadoop ecosystem,” *International Journal of Parallel Programming*, vol. 46, pp. 630–646, 2018.

[6] X. Chen, L. Hu, L. Liu, J. Chang, and D. L. Bone, “Breaking down hadoop distributed file systems data analytics tools: Apache hive vs. apache pig vs. pivotal hwaq,” in *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*. IEEE, 2017, pp. 794–797.

[7] J. Sheng, J. Amankwah-Amoah, Z. Khan, and X. Wang, “Covid-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions,” *British Journal of Management*, vol. 32, no. 4, pp. 1164–1183, 2021.

[8] I. Stančin and A. Jović, “An overview and comparison of free python libraries for data mining and big data analysis,” in *2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2019, pp. 977–982.

[9] S. Anitha and M. Metilda, “Apache hadoop based effective sentiment analysis on demonetization and covid-19 tweets,” *Global transitions proceedings*, vol. 3, no. 1, pp. 338–342, 2022.

[10] A. Çalıca Utku, G. Budak, O. Karabay, E. Güçlü, H. D. Okan, and A. Vatan, “Main symptoms in patients presenting in the covid-19 period,” *Scottish medical journal*, vol. 65, no. 4, pp. 127–132, 2020.

[11] W. H. Organization. (2023). [Online]. Available: "https://www.who.int/health-topics/coronavirus/coronavirus"

[12] M. F. Laham, N. A. W. A. Hamid, S. Zainorin, and Z. Ismail, “Performance evaluation between apache pig and hive on covid-19 vaccination progress,” in *3rd International Conference on Applied & Industrial Mathematics and Statistics (ICoAIMS2022)*. AIP, 2023.

7. Author’s Contributions

Shafinah Kamarudin and **Siti Nur Fathiah Ainuddin** conceived the study and performed the computation. **Nor Asilah Wati Abdul Hamid** critically reviewed and verified the results. **Mohamed Faris Laham** and **Zurita Ismail** draft manuscript preparation. All authors provided feedback throughout the work and contributed to the final manuscript.