IT Journal Research and Development (ITJRD)
Vol. 4, No. 2, Maret 2020, pp. 164 – 171, E-ISSN: 2528-4053 |P-ISSN: 2528-4061
DOI : 10.25299/itjrd.2020.vol4(2).4099

Implementasi Hadoop Dan Spark Untuk Analisis Penyebaran Demam Berdarah Dengue Berdasarkan Data Twitter

Irfan Rizqi Prabaswara[1] **and Ragil Saputra**[2]
Departemen Ilmu Komputer/ Informatika, Fakultas Sains dan Matematika, Universitas Diponegoro[1,2]
irfanprabaswara@gmail.com[1], ragil.saputra@live.undip.ac.id[2]

Article Info

History :
Dikirim 19 November 2019
Direvisi 26 November 2019
Diterima 01 Maret 2020

Kata Kunci:
Big Data
Hadoop
Spark
Twitter
Plotting Tren

Ringkasan

Big data merupakan sumber data yang memiliki volume yang besar, variasi yang banyak, dan aliran data yang sangat cepat. Contoh big data antara lain data dari media sosial dan query pencarian Google. Data tersebut mampu melacak aktivitas penyakit dan data yang ada tersedia setiap saat. Pengolahan big data bukanlah suatu hal yang mudah, sehingga diperlukan suatu tools yang dapat membantu proses pengolahan terhadap big data. Salah satu tools tersebut adalah hadoop. Meskipun kinerja hadoop lebih unggul daripada RDBMS tradisional, akan tetapi pengolahan data menggunakan hadoop belum maksimal. Sehingga, diperlukan pengolahan data yang lebih cepat. Salah satu cara untuk meningkatkan kecepatan pengolahan data ialah menerapkan spark untuk proses pengolahan data yang ada di HDFS (Hadoop Distributed File System). Pada penelitian ini dilakukan plotting tren dan pemetaan pada data Demam Berdarah Dengue (DBD) yang berasal dari media sosial twitter. Penelitian ini bertujuan untuk membuat visualisasi data yang diperoleh dari twitter dengan menggunakan hadoop dan spark dalam memantau perkembangan DBD di wilayah Asia Tenggara. Hasil dari plotting tren menunjukkan adanya hubungan yang kuat antara data twitter, data asli kejadian DBD yang diperoleh dari WHO. Penelitian ini juga melakukan pengujian performa hadoop dan spark. Semakin besar alokasi memory executor yang diterapkan serta semakin besar dan serupa alokasi maksimal memory scheduler yang diterapkan pada tiap node, maka waktu yang dibutuhkan untuk menyelesaikan task semakin singkat. Akan tetapi, pada titik tertentu konfigurasi hadoop dan spark menemui titik puncaknya, sehingga jika alokasi diperbesar menghasilkan hasil yang sama.

Koresponden:
Ragil Saputra
Departemen Ilmu Komputer/ Informatika, Fakultas Sains dan Matematika
Universitas Diponegoro
Jl. Prof. Soedarto, S.H. Tembalang Semarang, Indonesia, 50275
Email : ragil.saputra@live.undip.ac.id

1. PENDAHULUAN

Big data menjadi tren dalam dunia teknologi informasi saat ini. Big data merupakan sumber data yang memiliki volume yang besar, variasi yang banyak, dan aliran data yang sangat cepat [1]. Menurut Statistical Analysis System (SAS), big data adalah suatu kondisi populer yang digunakan untuk mendefinisikan perkembangan eksponensial serta ketersediaan dari data terstruktur maupun tidak [2].

Big data dapat digunakan untuk menggambarkan fenomena yang sedang terjadi saat ini [3]. Contoh big data yang dapat digunakan untuk menggambarkan fenomena saat ini adalah data dari media sosial twitter. Data tersebut mampu melacak aktivitas demam berdarah dengue dan data yang ada tersedia setiap saat [4]. Selain itu, data twitter sebagai salah satu big data media sosial juga dapat digunakan untuk mengetahui persebaran dengue di Brazil [5].

Pengolahan big data bukanlah suatu hal yang mudah [2]. Pengolahan big data tidak dapat disamakan dengan pengolahan data dengan ukuran yang relatif kecil. Single computer akan terhambat kinerjanya atau juga tidak akan dapat mengolah data jika ukurannya melebihi kapasitas memori pada komputer tersebut [6]. Oleh karena itu diperlukanlah suatu tool atau kerangka kerja yang dapat membantu proses pengolahan terhadap big data.

Salah satu tools atau kerangka kerja yang dapat digunakan untuk proses pengolahan big data adalah hadoop. Hadoop merupakan kerangka kerja yang dapat diimplementasikan pada single computer ataupun multiple computer dalam suatu jaringan tertentu [2]. Hadoop memiliki MapReduce sebagai model pemrograman untuk analisis big data dan hadoop distributed file system (HDFS) sebagai sistem file yang digunakan untuk menyimpan data yang tidak terstruktur. Selain itu, hadoop juga memiliki yet another resource negotiator (YARN) yang berfungsi sebagai pengatur resource pada seluruh aplikasi di dalam sistem [7].

Penelitian yang dilakukan oleh Ranjan (2017) menyatakan bahwa hadoop lebih scalable dan efisien daripada RDBMS tradisional. Selain itu, penelitian Ranjan (2017) yang lain menyatakan bahwa hadoop mempermudah untuk proses pengolahan data dalam waktu yang singkat. Namun, pengolahan data menggunakan hadoop belum maksimal. Saat ini, diperlukan pengolahan data yang lebih cepat untuk memenuhi kebutuhan pengolahan data [8]. Salah satu cara untuk meningkatkan kecepatan dalam pengolahan data ialah menerapkan spark untuk proses pengolahan data yang ada di HDFS. Spark muncul pada tahun 2012 dengan mengembangkan model MapReduce yang ada pada hadoop untuk mendukung lebih banyak komputasi secara efektif, seperti interactive queries dan stream processing [6]. Kecepatan komputasi yang dilakukan oleh spark 100 kali lebih cepat daripada MapReduce yang terdapat pada hadoop [9].

Penelitian yang dilakukan oleh Ryanto (2017) menyatakan bahwa spark menunjukkan kinerja komputasi yang lebih cepat hingga 5 kali lipat daripada hadoop pada cluster tervirtualisasi. Selain itu, spark juga memberikan throughput yang lebih tinggi pada kinerja I/O cluster daripada MapReduce. Penelitian yang dilakukan oleh Oliviandi (2018) menyatakan bahwa penggunaan spark untuk memproses big data sangatlah tepat karena dapat menurunkan response time rata-rata 50% hingga 70% dari MapReduce.

Berdasarkan pemaparan yang telah dijelaskan sebelumnya dapat disimpulkan bahwa data internet sebagai salah satu big data dapat digunakan untuk menganalisa, memantau, dan memprediksi terjadinya penyakit. Selain itu, dapat disimpulkan bahwa penggunaan Hadoop MapReduce, HDFS, dan Hive lebih efisien dibandingkan dengan penggunaan RDBMS. Namun, permasalahan yang muncul yaitu apakah penggunaan Hadoop dan Spark secara bersamaan dapat digunakan untuk mengelola big data serta apakah pengembangan tersebut menghasilkan performa yang baik dan efisien? Oleh karena itu, pada penelitian ini dilakukan penerapan hadoop dan spark untuk plotting tren dan pemetaan kejadian penyakit. Plotting tren dan pemetaan dilakukan pada kasus demam berdarah dengue (DBD) di beberapa negara di Asia Tenggara dengan data yang didapatkan dari media sosial twitter.

2. METODE PENELITIAN

2.1. Perencanaan Implementasi Hadoop dan Spark

Dibutuhkan beberapa langkah perencanaan sebelum implementasi hadoop dan spark sebagai lingkungan big data dalam mengelola data twitter untuk plotting tren dan pemetaan persebaran demam berdarah dengue ini benar-benar bisa dilakukan. Langkah-langkah tersebut adalah sebagai berikut :

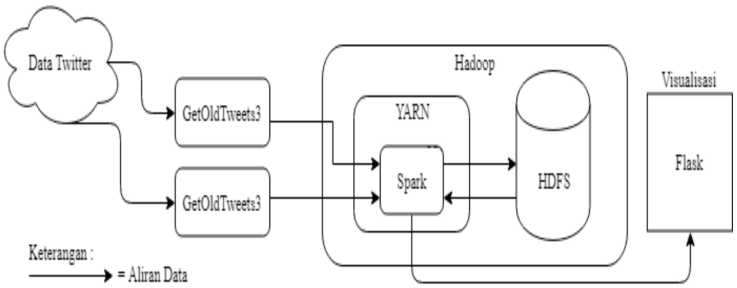
1. Menginstal hadoop dan spark sebagai lingkungan big data penelitian pada masternode dan slavenode.

2. Melakukan konfigurasi untuk tiap node agar dapat saling berinteraksi. Setelah kedua langkah tersebut dilakukan, maka hadoop dan spark sudah bisa dijalankan. Pada penelitian ini akan dilakukan beberapa perbandingan konfigurasi untuk mengetahui performa optimal hadoop dan spark untuk mengelola data twitter.

2.2. Desain Arsitektur Sistem

Berisi tentang teori yang digunakan dalam penelitian. Bisa saja terdiri dari beberapa subbab seperti yang ditunjukkan section berikut ini. Cara citasi studi kepustakaan lihat aturan penulisan jurnal [4].

Penelitian ini mengimplementasikan hadoop dan spark pada dua laptop yang selanjutnya disebut sebagai masternode dan slavenode. Arsitektur sistem yang dikembangkan pada penelitian ini dapat pada gambar 1.



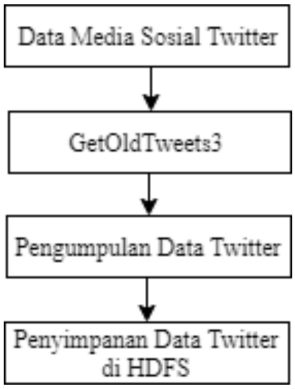
Gambar 1. Arsitektur Sistem

Data twitter diambil menggunakan spark dengan memanfaatkan library GetOldTweets3. Kemudian data diubah kedalam bentuk dataframe dan disimpan pada HDFS. Data yang tersimpan di HDFS dikelola kembali menggunakan spark dan hasil visualisasi data ditampilkan menggunakan flask. Flask digunakan untuk menampilkan data pada tampilan versi website.

Spark berjalan di atas hadoop. Hal ini bertujuan agar spark dapat mengakses data yang terdapat pada HDFS. Selain itu, spark juga berjalan di atas YARN. Hal ini bertujuan agar spark dapat menggunakan kemampuan YARN sebagai cluster manager dalam membagi job selama proses berjalan.

2.3. Pengumpulan dan Penyimpanan Data Twitter

Alur dalam pengumpulan dan penyimpanan data twitter pada penelitian ini dapat dilihat pada gambar 2.



Gambar 2. Alur Pengumpulan dan Penyimpanan Data Twitter

Data media sosial twitter diambil menggunakan library python GetOldTweets3. Proses ini dimulai dengan memasukkan lokasi dan kata kunci yang dicari. Lokasi yang digunakan adalah Indonesia, Thailand,

Singapura, Malaysia, Kamboja, Filipina, dan Brunei Darussalam. Kata kunci yang digunakan adalah demam berdarah, *aedes, dengue fever, fogging*, bệnh sốt xuất huyết dengue , dan *dengue rashes.*

Setelah memasukkan lokasi dan kata kunci, dilanjutkan dengan proses mengaktifkan sambungan TCP. Hal ini dilakukan agar aplikasi pengambilan tweet dapat terhubung dengan aplikasi spark, sehingga data twitter dapat langsung disimpan di HDFS melalui spark. Proses pengambilan tweet menghasilkan string tweet dengan delimiter tab (/t). String tweet ini dikirim ke spark melalui socket yang sudah dibuat sebelumnya. Setelah mendapatkan string tweet, proses dilanjutkan dengan mengubah string tweet menjadi dataframe. Pada dataframe yang dibuat, kolom yang dipilih yaitu tahun, username, text, dan date. Setelah itu, data disimpan di HDFS dalam format (*.csv).

Dalam proses penyimpanan di HDFS, dibuat juga checkpoint. Checkpoint ini berguna untuk melakukan backup jika ada kegagalan dalam suatu proses, sehingga pengulangan proses yang gagal cukup mulai dari yang sudah dicapai hingga checkpoint, tidak perlu mengulang proses sejak awal.

2.4. Pengelolaan Data Twitter

Proses pengelolaan data twitter terdiri dari proses plotting tren dan pemetaan kejadian DBD di Asia Tenggara. Proses plotting tren dan pemetaan kejadian DBD berdasarkan data twitter dimulai dengan mengambil data twitter dari HDFS. Selanjutnya, data ini diubah ke dalam bentuk dataframe untuk diproses lebih lanjut. Dataframe yang ada kemudian disatukan agar menjadi satu dan mudah untuk diolah. Pada proses ini dibuat juga alias untuk setiap kolom. Setelah dataframe selesai disatukan, dilakukan penghitungan data (count tweet) berdasarkan negara, tahun, dan bulan. Setelah proses count tweet selesai, dilakukan plotting tren dan pemetaan DBD menggunakan plotly sehingga menghasilkan tren dan peta persebaran DBD berdasarkan data twitter.

2.5. Skenario Pengujian Hadoop dan Spark

Pengujian dilakukan dengan mengeksekusi proses pengolahan data twitter hingga menghasilkan tren dan peta persebaran DBD berdasarkan data twitter. Pengujian yang dilakukan adalah dengan mengubah konfigurasi alokasi maksimal memory scheduler pada YARN. Alokasi maksimal memory scheduler berarti Resource Manager tidak dapat mengalokasikan memori ke kontainer melebihi alokasi maksimalnya. Selain itu, pengujian juga dilakukan dengan mengubah konfigurasi alokasi memory executor pada spark. Hal ini bertujuan untuk melihat waktu yang dibutuhkan hadoop dan spark dalam memproses data. Pengujian dilakukan sebanyak 11 kali dengan konfigurasi yang dapat dilihat pada tabel 1.

Tabel 1. Konfigurasi Pengujian.

Pengujian Ke-	Alokasi Maksimal Memory Scheduler		Alokasi Memory Executor
	Masternode (GB)	Slavenode (GB)	Memory (GB)
1	2	2	1
2	4	2	1
3	4	4	1
4	4	4	2
5	4	4	3
6	8	4	1
7	8	4	2
8	8	4	3
9	8	6	1
10	8	6	2
11	8	6	3

Tiap pengujian dianalisa hasilnya pada sebuah grafik evaluasi pengujian. Dari grafik ini dapat disimpulkan konfigurasi seperti apa yang optimal pada sebuah arsitektur big data dengan spesifikasi perangkat yang berbeda antara satu dengan yang lainnya.

3. HASIL DAN PEMBAHASAN

3.1. Implementasi Sistem

3.1.1. Implementasi Perangkat Keras

Sistem yang dibangun pada penelitian ini menggunakan dua perangkat keras sebagai master-node dan slave-node. Spesifikasi perangkat keras yang digunakan pada penelitian ini dapat dilihat pada tabel 2.

Tabel 2. Spesifikasi Perangkat Keras.

Aspek Teknologi	Master-node	Slave-node
Merk	Acer Nitro 5	MSI GF63 8RD
Processor	Intel Core i7-7700HQ	Intel Core i7-8750H
RAM	16 GB	8 GB
Graphic Card	NVidia Geforce GTX 1050 2GB	NVidia Geforce GTX 1050Ti 4GB

Kedua perangkat saling terhubung pada cluster multinode pada satu jaringan internet yang sama dengan konfigurasi perangkat dapat dilihat pada tabel 3.

Tabel 3. Konfigurasi Perangkat pada Cluster Multinode.

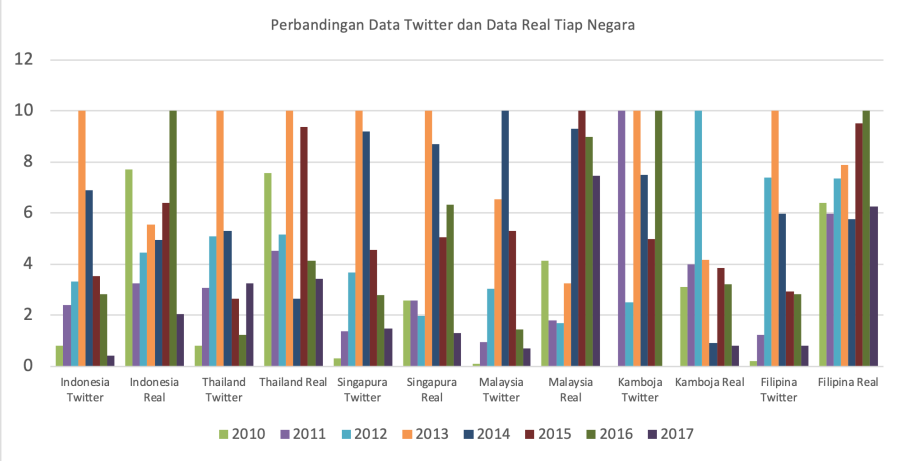
Aspek Teknologi	Master	Slave/Worker
Sistem Operasi	Lubuntu	Lubuntu
RAM	16 GB	8 GB
Jumlah Core	4	6

3.1.2. Implementasi Perangkat Keras

Perangkat lunak yang digunakan pada penelitian ini adalah hadoop versi 3.1.2 dan spark versi 2.4.0. Instalasi hadoop dan spark dilakukan di atas sistem Lubuntu versi 18.0.4.

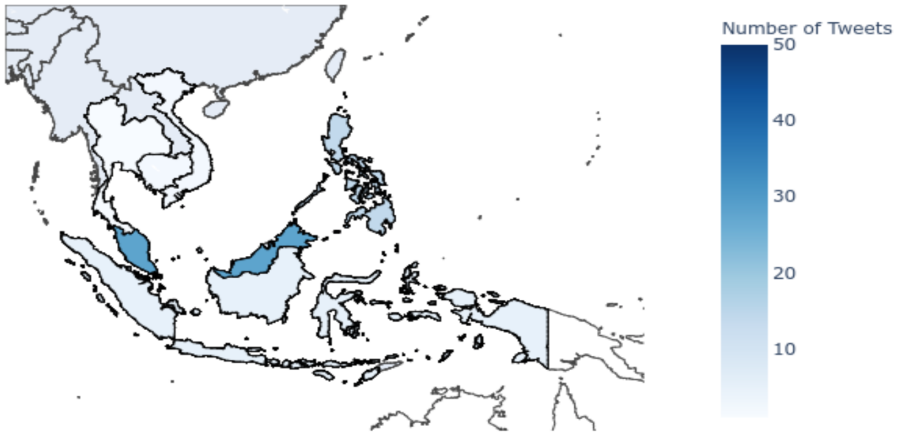
3.1.3. Hasil Tren dan Peta Persebaran DBD

Data twitter yang berhasil dikumpulkan berjumlah 4.056.690 tweet. Tren dan peta persebaran DBD berdasarkan data twitter dibandingkan dengan data asli yang didapatkan dari WHO. Grafik perbandingan data twitter dengan data real di Asia Tenggara dapat dilihat pada gambar 3.



Gambar 3. Perbandingan Tren Data Twitter dengan data WHO

Gambar 3 menggambarkan perubahan tren yang terjadi pada data twitter untuk masing-masing negara di Asia Tenggara memiliki kecenderungan perubahan yang sama dengan data asli kejadian DBD yang diperoleh dari WHO, meskipun jumlahnya tidak sama. Keselarasan antara data twitter dengan data kejadian DBD dari WHO juga ditunjukkan pada peta persebaran DBD berdasarkan data twitter yang dihasilkan pada penelitian ini. Peta persebaran kejadian DBD berdasarkan data twitter dan data kejadian DBD dari WHO secara berturut-turut dapat dilihat pada gambar 4.



Gambar 4. Peta Persebaran DBD Data Twitter Tahun 2017

Berdasarkan data twitter yang ditunjukkan pada gambar 4 dapat dilihat bahwa negara yang memiliki kasus demam berdarah paling tinggi pada tahun 2017 adalah Malaysia dan Filipina yang ditandai dengan warna biru yang lebih tua daripada negara lain. Hal ini sesuai dengan data real WHO yang menunjukkan bahwa pada tahun 2017, kasus terbanyak terjadi di negara Malaysia dan Filipina.

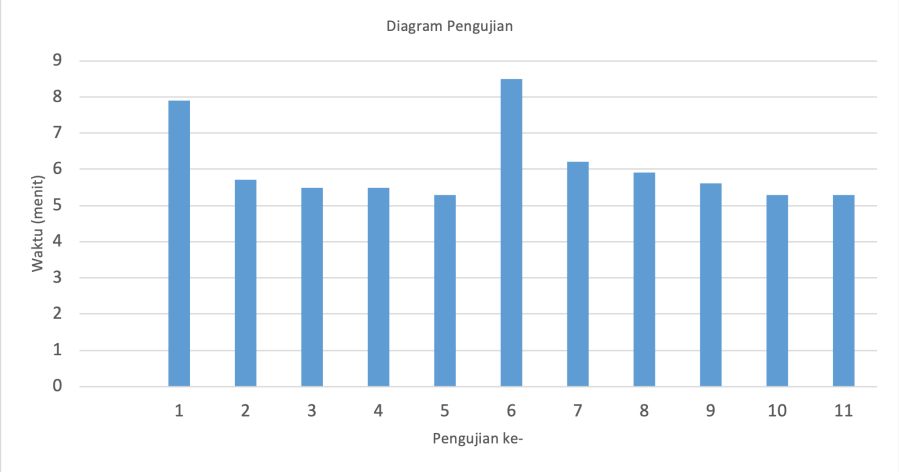
3.2. Hasil Pengujian Infrastruktur Hadoop dan Spark

Pengujian dilakukan dengan mengimplementasikan 11 jenis konfigurasi yang sudah direncanakan sebelumnya dan mengambil waktu eksekusi yang dibutuhkan dalam memproses data twitter. Jumlah jobs yang berjalan pada tiap pengujian sejumlah 367 jobs. Scheduling mode yang digunakan pada penelitian ini yaitu FIFO, sehingga tiap jobs yang datang pertama didistribusi dan dieksekusi terlebih dahulu. Hasil pengujian dapat dilihat pada tabel 4 dan gambar 5.

Tabel 4. Hasil Pengujian Sistem.

Pengujian Ke-	Alokasi Memory Executor	Alokasi Memory Maksimum Container		Jumlah executor		Waktu (Menit)
	Memory (GB)	Master Node (GB)	Slave Node (GB)	Master Node	Slave Node	
1	1	2	2	4	4	7.9
2	1	4	2	4	4	5.7
3	1	4	4	4	4	5.5
4	2	4	4	4	2	5.5
5	3	4	4	4	2	5.3
6	1	8	4	4	4	8.5
7	2	8	4	4	2	6.2
8	3	8	4	4	2	5.9
9	1	8	6	4	4	5.6
10	2	8	6	4	2	5.3
11	3	8	6	4	1	5.3

Dari penelitian yang dilakukan, waktu terlama yang dibutuhkan untuk mengolah data twitter hingga menghasilkan tren dan peta persebaran DBD adalah 8,5 menit dan waktu tercepat yang dibutuhkan adalah 5,3 menit. Waktu terlama terjadi pada pengujian ke 6 dengan konfigurasi alokasi memory executor 1 GB dan alokasi memory maksimum container pada masternode dan slavenode masing-masing 8 GB dan 4 GB. Waktu



Gambar 5. Diagram Hasil Pengujian Sistem

tercepat terjadi pada pengujian kelima, kesepuluh, dan kesebelas. Pengujian kelima menggunakan konfigurasi alokasi memoru executor 3 GB dan alokasi memory maksimum container pada masternode dan slavenode masing-masing 4 GB. Pengujian kesepuluh dan kesebelas masing-masing menggunakan konfigurasi alokasi memory maksimum container pada masternode dan slavenode masing-masing 8 GB dan 6 GB. Pengujian kesepuluh menggunakan konfigurasi alokasi memory executor sebesar 2 GB sedangkan pengujian kesebelas menggunakan konfigurasi alokasi memory executor sebesar 3 GB.

4. KESIMPULAN

Plotting tren data twitter menunjukkan hasil yang baik ketika dibandingkan dengan data real yang diperoleh dari WHO. Performa terbaik yang didapatkan pada penggunaan hadoop dan spark untuk plotting dan pemetaan kejadian DBD berdasarkan data twitter pada penelitian ini adalah dengan waktu eksekusi 5,3 menit. Performa terbaik ini dapat dicapai dengan mengalokasikan memory executor 3GB dan memory scheduler maximum 4GB untuk masing-masing node. Alokasi yang lebih besar dengan menggunakan data yang sama pada penelitian ini menghasilkan hasil yang sama.

Pustaka

[1] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, *Big Data for Dummies*. New Jersey: John Wiley & Sons, Inc.
[2] K. Basuki, H. Palit, and L. Dewi, "Implementasi hadoop: Studi kasus pengolahan data peminjaman perpustakaan universitas kristen petra," *Jurnal Infra*, vol. 3, no. 2, pp. 226–232,.
[3] B. RA, O. MJ, and B. WA, *Mapping collective behavior in the big-data era*. Cambridge University.
[4] C. A. M. Toledo, C. Degener, L. Vinhal, G. Coelho, W. Meira, C. Codeco, and M. Teixeira, "Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting dengue at country and city level," *PLOS*, vol. 11, no. 7, pp. 1–13,.
[5] M. Carlos, M. Nogueira, and R. Machado, "Analysis of dengue outbreaks using big data analytics and social networks," in *4th International Conference on Systems and Informatics (ICSAI*, Hangzhou.
[6] A. Ryanto, "Analisis kinerja framework big data pada cluster tervirtualisasi : Hadoop mapreduce dan apache spark," *Makassar*.
[7] A. S. Foundation, "Apache hadoop," available:. [Online]. Available: https://hadoop.apache.org/.
[8] S. Oliviandi, A. Osmond, and R. Latuconsina, "Implementation apache spark on big data based hadoop distributed file system," *e-Proceeding of Engineering*, vol. 5, no. 1, pp. 1005–1012,.
[9] A. S. Foundation, "Apache spark," available:. [Online]. Available: https://spark.apache.org/.

BIOGRAFI PENULIS

Irfan Rizqi Prabaswara obtained Bachelor Degree in Computer Science from Diponegoro University in 2019. His current research interests include big data, data mining, and machine learning.

Ragil Saputra obtained Bachelor Degree in Mathematics from Universitas Diponegoro in 2003, obtained Master Degree in Computer Science from Universitas Gadjah Mada in 2011. He has been a Lecturer with the Department of Informatics, Universitas Diponegoro, since 2005. His current research interests include include big data, technology adoption, and information system.