Syntax Admiration: p-ISSN 2722-7782 | e-ISSN 2722-5356 Vol. 5, No. 11, November 2024

Perbandingan Uji Performa Impala dan Hive-Hadoop

Dede Kusuma[1*], Aviarini Indrati[2]
[1,2] Universitas Gunadarma, Jakarta, Indonesia
Email: deltakusuma@gmail.com, avi@staff.gunadarma.ac.id

Abstrak

Jumlah data yang berkembang pesat saat ini juga membutuhkan penyimpanan yang cepat. Hal ini dikarenakan kebutuhan akan data juga sangat penting, dan untuk mengakses data tersebut juga membutuhkan waktu yang cepat. Oleh karena itu, perlu nya mengetahui tools yang mendukung pemrosesan data dalam jumlah yang besar dan waktu yang cepat. Kehadiran Impala dan Hive-Hadoop membantu dalam mengambil keputusan tools mana yang akan digunakan untuk menyimpan data dan dengan cepat mendapatkan data yang dibutuhkan. Dalam analisis dan perbandingan ini ingin mengetahui bagaimana kinerja kedua tools tersebut, yaitu apakah *Impala* dan *Hive-Hadoop* dapat mengakses data terstruktur. Penelitian ini dilakukan dengan menggunakan metode eksperimen dan memanipulasi lingkungan eksperimen menggunakan virtualisasi komputer. Hasil analisis yang didapat adalah, *Impala* lebih cepat daripada *Hive-Hadoop* karena mengurangi latensi dan *Impala* tidak didasarkan pada algoritma *MapReduce*.

Kata Kunci: Perbandingan, Uji Performa, Impala, Hive, Hadoop

Abstract

The amount of data that is growing rapidly today also requires fast storage. This is because the need for data is also very important, and to access the data also requires fast time. Therefore, it is necessary to know the tools that support the processing of large amounts of data and fast time. The presence of Impala and Hive-Hadoop helps in making decisions about which tools to use to store data and quickly get the data needed. In this analysis and comparison, we want to know how the performance of these two tools is, namely whether Impala and Hive-Hadoop can access structured data. This research was conducted using experimental methods and manipulating the experimental environment using computer virtualization. The analysis results obtained are, Impala is faster than Hive-Hadoop because it reduces latency and Impala is not based on the MapReduce algorithm.

Keywords: *Comparison, Performance Test, Impala, Hive, Hadoop*

Pendahuluan

Dalam beberapa tahun terakhir, perkembangan teknologi informasi berkembang sangat kuat dan pesat (Wahyono, 2019);(Arfah & Harbi, 2019). Dengan perkembangan tersebut, tidak dapat dipungkiri bahwa pertumbuhan jumlah data juga sangat dipengaruhi oleh pengolahan data dalam suatu operasi teknologi (Azizah & Saptono, 2020). Jumlah data menjadi semakin besar dan sulit untuk dimasukkan ke dalam sistem basis data apa

pun, sehingga sulit untuk dianalisis dan diproses dengan cepat (Yusuf, 2021). Jumlah data yang terus bertambah ini mulai disebut sebagai big data (Siti Nurmiati, 2020).

Big data dapat memproses pencarian, penyimpanan, dan analisis data dengan cepat, berkembangnya teknologi informasi yang bertujuan untuk mempermudah dalam melakukan penelitian, penyimpanan data dan sejenisnya, maka tidak memakan banyak waktu untuk melakukan pekerjaannya (Juditha, 2020);(Priyatna et al., 2020);(Karim et al., 2020).

Big Data menjadi perhatian akademisi dan industri IT. Dalam dunia digital dan komputasi, informasi dapat dikumpulkan dan diproses dengan sangat cepat menggunakan teknologi ini (Bhadani & Jothimani, 2016). MapReduce, yang pertama kali diusulkan oleh Google pada tahun 2004 dan memiliki kemampuan untuk menangani data yang sangat besar secara bersamaan dan terdistribusi, telah terbukti menjadi metode yang efektif untuk menangani volume data yang besar dan tidak terstruktur ini. Ada banyak framework untuk melakukan operasi pada Big Data menggunakan model *MapReduce*, salah satunya adalah *framework Apache Hadoop*, Hadoop memproses data secara terdistribusi ke banyak komputer bahkan ribuan komputer dengan HDFS dan MapReduce Hadoop (Jankatti et al., 2020).

Jika semakin banyak data yang disimpan di suatu server, maka memori server penyimpanan akan penuh. Jika hadoop hanya memiliki satu server, mungkin saja server tersebut terhenti dan kehabisan memori (Cholik, 2021). Teknologi yang sama dapat memproses data tetapi memiliki cakupan yang berbeda. Hadoop bukan database tetapi kerangka kerja untuk memproses data besar. di dalam hadoop terdapat dua layanan dalam membaca hadoop file system, terdapat hive sebagai query tools dengan algoritma mapreduce dan impala sebagai *query tools* tanpa menggunakan algoritma mapreduce (Yuspiani & Wahyuddin, 2021).

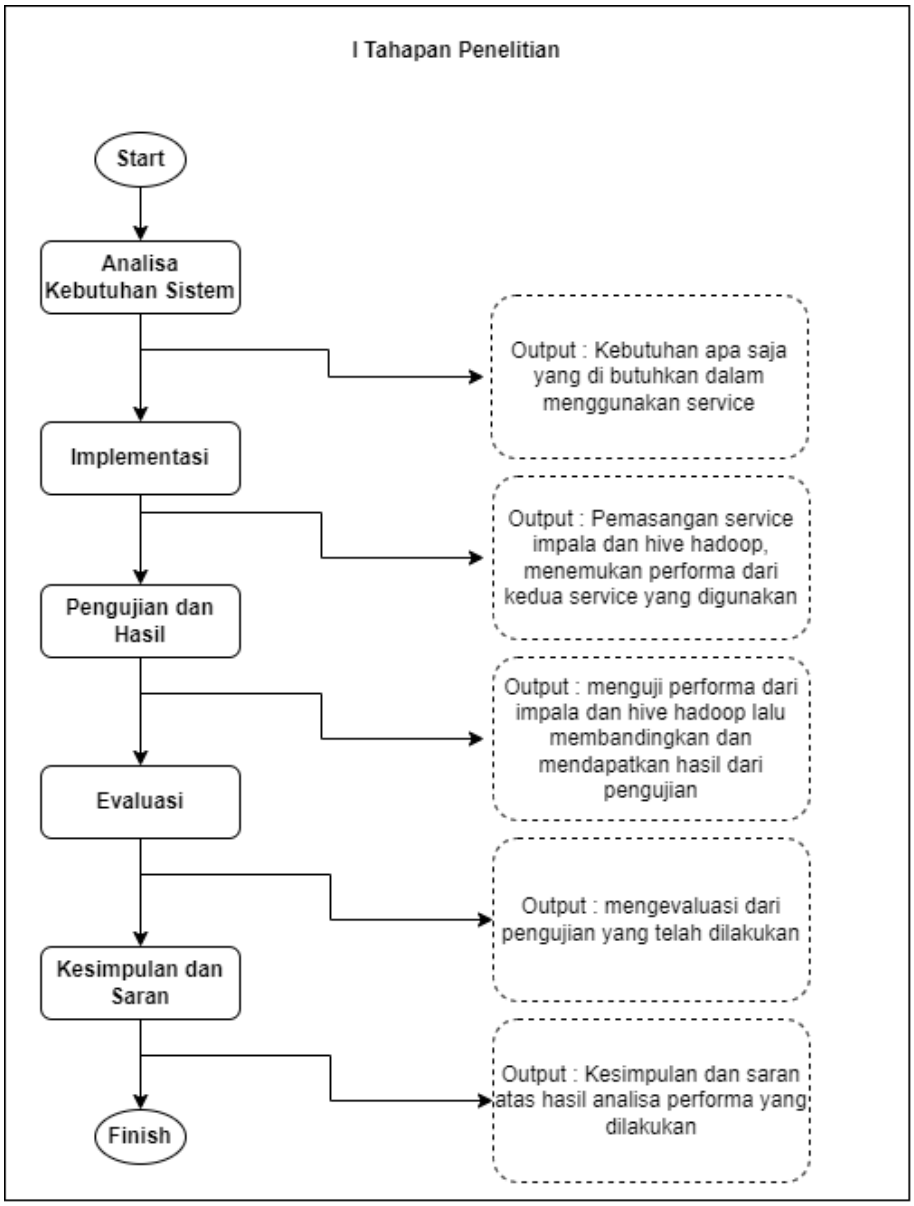
Hive dan impala memiliki perbedaan pemrosesan olah data di dalam big data, dalam fungsional yang sama pada sistem hadoop kecepatan menjadi faktor yang penting dalam kinerja pengolahan data besar, semakin cepat kinerja pengolahan data maka semakin efisien untuk pengguna pengolahan data tersebut (Muhammad Wali et al., 2023). Untuk mengukur kecepatan dari masing-masing layanan di dalam sistem yang sama, waktu dalam pemprosesan *query* dan ukuran data menjadi parameter pengujian yang akan dilakukan (Camacho-Rodríguez et al., 2019). Berdasarkan uraian diatas, penulis melakukan analisis kinerja antara kedua layanan tersebut pada hadoop sehingga dapat diketahui bagaimana kinerja antara keduanya dalam pemrosesan big data dan dapat memilih layanan manakah yang lebih efisien untuk melakukan pengolahan data serta pengambilan keputusan.

Berdasarkan dari latar belakang diatas terdapat beberapa masalah yang dapat diidentifikasikan sebagai berikut: 1) Bagaimana menganalisis kebutuhan sistem yang digunakan dalam pengujian kinerja impala dan hive? 2) Bagaimana analisis kinerja impala dan hive dengan parameter ukuran serta waktu untuk mengukur kecepatan dari layanan tersebut?

Penelitian ini bertujuan untuk menganalisis kebutuhan sistem dan membandingkan kecepatan kinerja Impala dan Hive dalam memproses data besar berdasarkan parameter ukuran data dan waktu pemrosesan, guna memberikan panduan pemilihan teknologi yang efisien dalam pengolahan data besar. Secara teoritis, penelitian ini berkontribusi pada pengembangan ilmu pengetahuan terkait performa alat pengolahan data besar dalam sistem Hadoop, sedangkan secara praktis, hasilnya dapat menjadi acuan bagi praktisi IT dan organisasi dalam memilih tools yang lebih efektif untuk mendukung pengambilan keputusan berbasis data secara cepat dan akurat.

Metode Penelitian

Peneliti akan memanipulasi atau bereksperimen dengan Impala dan Hadoop agar peneliti mengetahui kinerja keduanya.



Gambar 1. Tahapan Penelitian

Diagram ini menggambarkan tahapan penelitian yang dimulai dari analisa kebutuhan sistem untuk menentukan kebutuhan yang diperlukan dalam penggunaan layanan. Setelah itu, dilakukan implementasi dengan memasang layanan Impala dan Hive Hadoop serta mengidentifikasi performa masing-masing (Dianta et al., 2022). Tahap selanjutnya adalah pengujian untuk mengukur dan membandingkan performa kedua layanan, yang kemudian dilanjutkan dengan evaluasi terhadap hasil pengujian. Akhirnya, penelitian ini ditutup dengan tahap kesimpulan dan saran, di mana dihasilkan rekomendasi berdasarkan analisa performa layanan yang telah diuji, memberikan panduan untuk perbaikan atau optimasi ke depan. Dalam penelitian ini metode yang digunakan untuk mengumpulkan data, yaitu melakukan pengujian atau evaluasi untuk menganalisis kinerja komponen Impala dan Hive Hadoop

Hasil dan Pembahasan

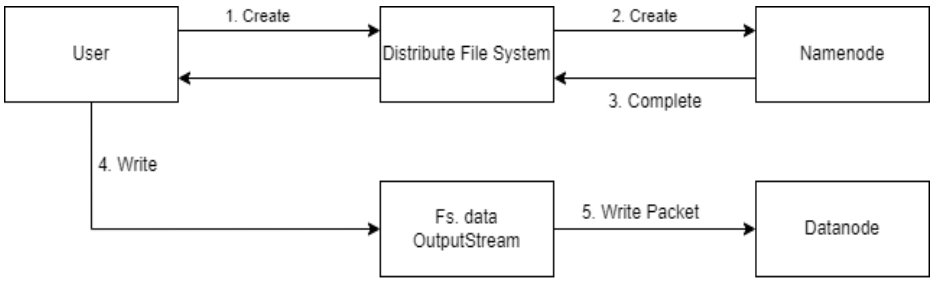
Analisa Kebutuhan *Software*

Software yang akan digunakan pada penelitian ini adalah diantaranya : VMware, karena lingkungan percobaannya adalah *virtual*, maka diperlukan salah satu aplikasi untuk virtualisasi yakni VMware. Os Linux CentOS-7, merupakan sistem operasi yang akan digunakan untuk penelitian.

Hadoop Versi 7.1.7, platform berbasis java untuk mendukung aplikasi yang berjalan pada *bigdata*. *Hive* Versi 3.1.3000.7.1.7.0-551, perangkat lunak *data warehouse* yang digunakan untuk query dan mengolah data yang didistribusikan secara SQL dan Hadoop *MapReduce*, model komputasi berbasis java pada sistem terdistribusi dalam rangka mendukung aplikasi *Big Data*. JDK 1.8.0_232, perangkat lunak yang digunakan untuk melakukan proses kompilasi dari kode Java ke *bytecode*. *Impala version* mesin pemrosesan SQL *query* paralel besar yang digunakan untuk memproses *volume* data yang tinggi yang disimpan dalam *cluster Hadoop*. MobaXterm merupakan suatu terminal yang memiliki kinerja yang ditingkatkan pada X *server* dan satu *set* perintah Unix (GNU / Cygwin) yang dikemas dalam sebuah *file* .exe tunggal dan *portable* (tanpa proses instalasi).

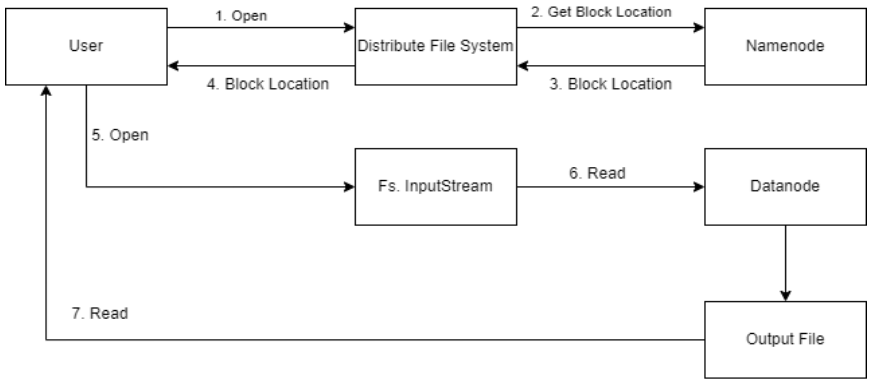
Rancangan Sistem

Peneliti akan menggunakan satu server untuk memasang Hadoop dan pendukung Hive dan Impala, yang akan digunakan untuk memproses data SQL. Untuk proses *write* yang terjadi pada *Hadoop* akan dijalankan di *HDFS* dengan skema berikut:



Gambar 2. Proses Write yang Terjadi di HDFS

Jika user memberi perintah untuk menyimpan data, namenode akan berkomunikasi dengan datanode untuk memberi tahu bahwa ada file yang akan disimpan di HDFS dan menanyakan lokasi datanode yang dapat diakses. Setelah mendapatkan daftar nama dan alamat datanode, komputer user akan langsung mentransfer data ke datanode yang ada, seperti yang ditunjukkan pada gambar 2. Data yang dikirim sudah dibelah otomatis menjadi blok atau kepingan data yang disimpan dalam datanode. Setelah kepingan data diterima, datanode akan mengirimkan laporan ke namenode tentang penerimaan dan penyimpanan data. Namun, proses pembacaan HDFS adalah sebagai berikut:



Gambar 3. Proses Read yang terjadi di HDFS

Gambar 3 menunjukkan alur kerja proses pemrosesan data atau file di HDFS. Saat pengguna memberikan perintah pemrosesan pada komputernya, pesan akan dikirim ke namenode untuk menanyakan nama dan alamat datanode yang harus diakses untuk mendapatkan data. Selanjutnya, setelah user memperoleh nama dan alamat datanode, datanode akan diakses secara langsung, dan data akan ditampilkan sesuai perintah user.

Rancangan Pengujian

Pengujian pada *hadoop* akan menggunakan *hive* dan *impala* yang merupakan data *warehouse* yang ada di *hadoop* itu sendiri, dimana dua *query tools* tersebut memungkinkan untuk melakukan *load* dan *select* data, sehingga bisa dibandingkan kinerja keduanya berdasarkan percobaan pada dua *tools* tersebut. *Query* yang akan digunakan adalah *query select* dan *load* untuk diujikan pada kedua perangkat.

Pengujian Lama Waktu *Load* Data

Untuk menguji waktu load data yang lama, berikut adalah langkah-langkah yang harus dilakukan: 1) Memastikan topologi sistem yang diukur. 2) Menentukan data warehouse dan database Hive dan Impala yang akan digunakan. 3) Menentukan data untuk pengujian. 4) Menentukan *query* yang akan digunakan yaitu *insert*. 5) Menentukan berapa kali pengujian akan dilakukan. 6) Menjalankan *Hive* atau *impala* yang telah ter *install* di komputer untuk menguji kinerja *Hadoop*. 7) Melihat *output* yang dihasilkan dari alat ukur yaitu waktu *load* data pada *Hadoop* yang telah ditentukan datanya. 8) Ulangi langkah-langkah di atas untuk penambahan data 2GB-5GB. dan selanjutnya dicatat waktu pengukuran saat pengujian.

Pengujian Lama Waktu *Select* Data

Untuk melakukan pengujian lama waktu pemilihan data, langkah-langkah berikut harus dilakukan: 1) Memastikan topologi sistem yang diukur. 2) Menentukan data *warehouse* dan *database* yang akan digunakan yaitu *Hive* dan *Impala*. 3) Menentukan data untuk pengujian. 4) Menentukan *query* yang akan digunakan yaitu *select*. 5) Menentukan berapa kali pengujian akan dilakukan. 6) Menjalankan *Hive* atau *impala* yang telah terinstal di komputer untuk menguji kinerja *Hadoop*. 7) Melihat *output* yang dihasilkan dari alat ukur yaitu waktu *load* data pada *Hadoop* yang telah ditentukan datanya. 8) Untuk menambah 2GB hingga 5GB, ulangi langkah-langkah di atas. Kemudian, catat waktu pengujian.

Data Penelitian

Data yang akan digunakan dalam penelitian ini adalah data dengan format CSV yang berukuran 1GB sampai 5GB. Data didapat dari *generate script* python terlampir pada lampiran. Peneliti menggunakan data *generate random* data dengan ukuran 1GB sampai 5GB. Data ini merupakan data yang terstruktur dalam 1 tabel sehingga menggunakan format CSV agar memudahkan dimasukkan ke dalam format *database* terstruktur. Untuk memudahkan memasukkan data ke dalam tabel sistem *SQL*, maka data dibuat dengan format *delimiter* (pembatas) koma, sehingga perintah atau format yang dijalankan bisa disesuaikan dengan format *SQL*.

Pengujian

Peneliti sudah memastikan bahwa parameter yang akan digunakan untuk penelitian ini adalah di antaranya: a) Besarnya data yakni antara 1GB sampai dengan 5GB. *b) Query* eksekusi pada *Hive* dan *Impala* yakni *Load* data dan *select* data. Parameter yang digunakan diatas adalah uji coba yang dilakukan oleh peneliti sehingga bisa mengetahui bagaimana performa dari *Impala* dan *Hive* yang digunakan untuk bahan penelitian. Untuk *select* data ada tiga waktu yang akan dihasilkan dan dalam penelitian ini mengambil hasil waktu yang di keluarkan dari sistem.

Pengujian Hive

Tabel 1. Pengujian Load Data pada Hive

Pengujian Ke	Ukuran Data				
	1 GB (s)	2 GB (s)	3 GB (s)	4 GB (s)	5 GB (s)
1	0.313	0.290	0.366	0.372	0.405
2	0.416	0.341	0.352	0.428	0.438
3	0.241	0.266	0.302	0.404	0.424
Rata -rata	0.323	0.299	0.340	0.401	0.422

Tabel 2 menunjukkan hasil keseluruhan pengujian data pilihan.

Tabel 2. Pengujian Select Data pada Hive

Pengujian Ke	Ukuran Data				
	1 GB (s)	2 GB (s)	3 GB (s)	4 GB (s)	5 GB (s)
1	60.394	94.220	128.975	204.407	225.813
2	44.416	88.440	130.819	219.115	300.883

3	69.593	153.069	219.050	207.633	202.809
Rata -rata	58.134	111.909	159.614	210.385	243.168

Pengujian Impala

Tabel 3. Pengujian Load Data pada Impala

Pengujian Ke	Ukuran Data				
	1 GB (s)	2 GB (s)	3 GB (s)	4 GB (s)	5 GB (s)
1	0.15	0.18	0.09	0.10	0.11
2	0.12	0.10	0.08	0.08	0.10
3	0.11	0.08	0.09	0.08	0.11
Rata -rata	0.13	0.12	0.08	0.09	0.11

Tabel 4 menunjukkan hasil keseluruhan pengujian dari pilihan data.

Tabel 4. Pengujian Select Data pada Impala

Pengujian Ke	Ukuran Data				
	1 GB (s)	2 GB (s)	3 GB (s)	4 GB (s)	5 GB (s)
1	10.543	57.332	88.159	116.740	138.783
2	31.307	49.644	78.234	83.894	136.111
3	26.461	52.483	81.557	82.918	123.725
Rata -rata	22.770	53.153	82.650	94.517	132.873

Analisa Pengujian Load Data Hive vs Impala

Pengujian menunjukkan bahwa mengisi data di hive dan impala membutuhkan waktu yang lebih lama, tetapi mengisi data di impala lebih efektif daripada di hive tabel perbandingan waktu load data antara hive dan impala dapat dilihat pada tabel 5.

Tabel 5. Rata-rata Waktu Load Data di Hive dan Impala

Tabel Rata - Rata		
Ukuran Data	Load Data Hive (s)	Load Data Impala (s)
1 GB	0.323	0.130
2 GB	0.299	0.120
3 GB	0.340	0.080
4 GB	0.401	0.090
5 GB	0.422	0.110

Berdasarkan tabel 5, waktu yang butuhkan oleh *hive* lebih lambat dibandingkan impala, dalam hal ini karena *hive* menggunakan *mapreduce* untuk memproses *load* data berbeda dengan *impala* yang tidak menggunakan *mapreduce* dalam proses *load* data. Sehingga waktu yang diperlukan untuk *load* data jauh lebih cepat dibandingkan dengan *hive*.

Analisa Pengujian Select Data Hive vs Impala

Tabel 6 Rata-rata Waktu Select Data di Hive dan Impala

Tabel Rata - Rata		
Ukuran Data	**Select Data Hive (s)**	**Select Data Impala (s)**
1 GB	58.134	22.770
2 GB	111.909	53.153
3 GB	159.614	82.650
4 GB	210.385	94.517
5 GB	243.168	132.873

Tabel 6 menunjukkan bahwa waktu yang dibutuhkan hive untuk membaca data semakin lama seiring dengan jumlah data yang dimasukkan. Untuk *Impala* yang tidak menggunakan *logic MapReduce*, data yang telah dipetakan akan secara cepat didapatkan dan ditampilkan oleh sistem sehingga waktu yang diperlukan lebih cepat dibandingkan dengan *hive*.

Kesimpulan

Kesimpulan dari pengujian yang telah dilakukan oleh peneliti terkait performa Hive dan Impala adalah sebagai berikut: Pada penelitian ini, peneliti ingin mengetahui seberapa baik masing-masing sistem menulis dan membaca data. Besar data, yang berkisar antara 1 GB dan 5 GB, adalah indikator atau parameter. Parameter eksekusi lainnya adalah query load dan select data, yang digunakan.

Hasil dari performa yang ditampilkan oleh masing - masing query tools pada hadoop sangat signifikan yang dapat dilihat dari lama waktu query. Berdasarkan parameter ukuran dan waktu sebagai uji kecepatan query yang di jalankan dapat di ambil kesimpulan query menggunakan impala lebih cepat dibandingkan dengan menggunakan hive yang dapat membuat suatu pengambilan keputusan menjadi lebih cepat dan tepat dengan hasil performa yang sangat baik bisa dikatakan impala bisa menjadi query tools yang lebih cepat dibandingkan dengan hive.

BIBLIOGRAFI

Arfah, M., & Harbi, I. Y. (2019). Studi Penentuan Tempat Pembuangan Akhir Sampah di Kota Tebing Tinggi dengan Metode Proses Hirarki Analitik. *Talenta Conference Series: Energy and Engineering (EE)*, *2*(3).

Azizah, N., & Saptono, H. (2020). Uji Performa Dan Perbandingan Rdbms Mysql Dan Hive-Hadoop. *Jurnal Informatika Terpadu*, *6*(1), 20–28.

Bhadani, A. K., & Jothimani, D. (2016). Big data: challenges, opportunities, and realities. *Effective Big Data Management and Opportunities for Implementation*, 1–24.

Camacho-Rodríguez, J., Chauhan, A., Gates, A., Koifman, E., O'Malley, O., Garg, V., Haindrich, Z., Shelukhin, S., Jayachandran, P., & Seth, S. (2019). Apache hive: From mapreduce to enterprise-grade big data warehousing. *Proceedings of the 2019 International Conference on Management of Data*, 1773–1786. https://doi.org/10.1145/3299869.3314045.

Cholik, C. A. (2021). Perkembangan Teknologi Informasi Komunikasi/ICT dalam Berbagai Bidang. *Jurnal Fakultas Teknik Kuningan*, *2*(2), 39–46.

Dianta, I. A., Aqham, A. A., & Setiawan, D. (2022). Penerapan big data untuk mengatur sistem analisis data. *Jurnal Ilmiah Teknik Mesin, Elektro Dan Komputer*, *2*(1), 40–46.

Jankatti, S., Raghavendra, B. K., Raghavendra, S., & Meenakshi, M. (2020). Performance evaluation of Map-reduce jar pig hive and spark with machine learning using big data. *International Journal of Electrical and Computer Engineering*, *10*(4), 3811.

Juditha, C. (2020). Dampak Penggunaan Teknologi Informasi Komunikasi Terhadap Pola Komunikasi Masyarakat Desa. *Jurnal PIKOM; Penelitian Komunikasi Dan Pembangunan*, *21*(2), 131–144.

Karim, A., Bangun, B., Purnama, I., Harahap, S. Z., Irmayani, D., Nasution, M., Haris, M., & Munthe, I. R. (2020). *Pengantar teknologi informasi*. Yayasan Labuhanbatu Berbagi Gemilang.

Muhammad Wali, S. T., Efitra, S., Kom, M., Sudipa, I. G. I., Kom, S., Heryani, A., Sos, S., Hendriyani, C., Rakhmadi Rahman, S. T., & Kom, M. (2023). *Penerapan & Implementasi Big Data di Berbagai Sektor (Pembangunan Berkelanjutan Era Industri 4.0 dan Society 5.0)*. PT. Sonpedia Publishing Indonesia.

Priyatna, C. C., Prastowo, F. X. A. A., Syuderajat, F., & Sani, A. (2020). Optimalisasi teknologi informasi oleh lembaga pemerintah dalam aktivitas komunikasi publik. *Jurnal Kajian Komunikasi*, *8*(1), 114–127.

Siti Nurmiati, S. N. (2020). Sistem Informasi Administrasi Dan Pembayaran Pada Smkn 1 Ciomas Bogor Berbasis Web. *Incomtech*, *9*(2), 57–61.

Wahyono, H. (2019). Pemanfaatan teknologi informasi dalam penilaian hasil belajar pada generasi milenial di era revolusi industri 4 . 0. *Proceeding of Biology Education*, *3*(1), 192–201.

Yuspiani, Y., & Wahyuddin, W. (2021). TRANSFORMASI ARSIP DI ERA BIG DATA. *Idaarah: Jurnal Manajemen Pendidikan*, *5*(1), 73–82.

Yusuf, R. (2021). Analisis metode evaluasi koleksi sebagai acuan kegiatan pengembangan koleksi. *Pustaka Karya: Jurnal Ilmiah Ilmu Perpustakaan Dan Informasi*, *9*(2), 85–94.