



**EVALUASI PERFORMA HADOOP DAN SPARK PADA
DIGITALOCEAN MENGGUNAKAN HIBENCH DALAM
KONFIGURASI *PSEUDO DISTRIBUTED***

NASKAH SKRIPSI

**Dimas Wahyu Saputro
NIM 120450081**

**PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN**

2024



**EVALUASI PERFORMA HADOOP DAN SPARK PADA
DIGITALOCEAN MENGGUNAKAN HIBENCH DALAM
KONFIGURASI *PSEUDO DISTRIBUTED***

NASKAH SKRIPSI
Diajukan sebagai syarat maju seminar hasil

Dimas Wahyu Saputro
NIM 120450081

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
LAMPUNG SELATAN

2024

HALAMAN PENGESAHAN

Naskah Tugas Akhir untuk Seminar Hasil dengan judul "**Evaluasi Performa Hadoop dan Spark pada DigitalOcean menggunakan HiBench dalam Konfigurasi Pseudo Distributed**" adalah benar dibuat oleh saya sendiri dan belum pernah dibuat dan diserahkan sebelumnya, baik sebagian ataupun seluruhnya, baik oleh saya ataupun orang lain, baik di Institut Teknologi Sumatera maupun di institusi pendidikan lainnya.

Lampung Selatan, 10 Mei 2024

Penulis,



Dimas Wahyu Saputro
NIM 120450081

Diperiksa dan disetujui oleh,

Pembimbing I

Pembimbing II

Tirta Setiawan, S.Pd., M.Si.
NIP. 199008222022031003

Riksa Meidy Karim, S.Kom., M.Si., M.Sc.

Disahkan oleh,
Koordinator Program Studi Sains Data
Fakultas Sains
Institut Teknologi Sumatera

Tirta Setiawan, S.Pd., M.Si.
NIP. 199102302020012003

HALAMAN PERNYATAAN ORISINALITAS

**Skripsi ini adalah karya saya sendiri dan semua sumber baik yang dikutip
maupun yang dirujuk telah saya nyatakan benar.**

Nama : Dimas Wahyu Saputro

NIM : 120450081

Tanda tangan :

Tanggal : 10 Mei 2024

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Institut Teknologi Sumatera, saya yang bertanda tangan di bawah ini:

Nama : Dimas Wahyu Saputro
NIM : 120450081
Program Studi : Sains Data
Fakultas : Sains
Jenis karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan Hak Bebas Royalti Noneksklusif (*Non-Exclusive Royalty Free Right*) kepada Institut Teknologi Sumatera atas karya ilmiah saya yang berjudul:

Evaluasi Performa Hadoop dan Spark pada DigitalOcean menggunakan Hi-Bench dalam Konfigurasi *Pseudo Distributed*

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Institut Teknologi Sumatera berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Lampung Selatan
Pada tanggal : 10 Mei 2024

Yang menyatakan (Dimas Wahyu Saputro)

ABSTRAK

Evaluasi Performa Hadoop dan Spark pada DigitalOcean menggunakan Hi-Bench dalam Konfigurasi *Pseudo Distributed*

Dimas Wahyu Saputro (120450081)

Pembimbing I: Tirta Setiawan, S.Pd., M.Si.

Pembimbing II: Riksa Meidy Karim, S.Kom., M.Si., M.Sc.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Kata kunci: ini, itu, ini, itu

ABSTRACT

Performance Evaluation of Hadoop and Spark on DigitalOcean using HiBench in a Pseudo-Distributed Configuration

Dimas Wahyu Saputro (120450081)

Advisor I : Tirta Setiawan, S.Pd., M.Si.

Advisor II: Riksa Meidy Karim, S.Kom., M.Si., M.Sc.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt.

Keywords : *this, that, this, thatthis.*

MOTTO

Urip Iku Urup.

HALAMAN PERSEMBAHAN

Untuk diriku, Ibu, dan Bapak.

KATA PENGANTAR

Puji syukur penulis ucapkan ke hadirat Allah SWT atas berkah dan rahmat-Nya sehingga skripsi ini dapat terselesaikan dengan baik. Skripsi ini merupakan karya yang wajib dibuat oleh mahasiswa untuk menyelesaikan pendidikan sarjana di Institut Teknologi Sumatera. Penyusunan skripsi ini banyak mendapat bantuan dan dukungan dari berbagai pihak sehingga dalam kesempatan ini, dengan penuh ke-rendahan hati, penulis mengucapkan terima kasih kepada:

1. Keluarga, Ibu Siti Ervingati dan bapak Kustriyanto, yang selalu memberikan doa, semangat, dukungan, dan motivasi sehingga penulis dapat mencapai tahap ini. Tak lupa pula untuk Andika, Habib, dan Syifa
2. Bapak Tirta Setiawan, S.Pd., M.Si., selaku Koordinator Program Studi Sains Data Fakultas Sains Institut Teknologi Sumatera dan Dosen Pembimbing Utama
3. Bapak Riksa Meidy Karim, S.Kom., M.Si., M.Sc., dan Ibu Amalya Citra S.Kom., M.Si., M.Sc., selaku dosen pembimbing pendamping yang telah memberikan arahan, ilmu, motivasi, serta saran kepada penulis
4. Seluruh dosen dan tenaga kependidikan Sains Data Institut Teknologi Sumatera yang telah memberikan banyak bantuan dan ilmu selama penulis berku-liah
5. Abil, Imam, Sakul, dan sahabat-sahabat yang tidak dapat disebutkan satu per-satu. Terima kasih atas semangat, bantuan dan motivasinya. Semoga kalian selalu dikuatkan.
6. Teman-teman seperbimbingan, dan angkatan 2020 Sains Data Institut Tekno-logi Sumatera

Penulis menyadari bahwa masih terdapat banyak kekurangan pada penulisan skripsi ini. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun dari pembaca demi perbaikan laporan ini. Semoga karya ini dapat bermanfaat bagi para pembaca pada umumnya dan juga bagi penulis pada khususnya.

Lampung Selatan, 10 Mei 2024

Dimas Wahyu Saputro

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGESAHAN	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI	iv
ABSTRAK	v
ABSTRACT	vi
MOTTO	vii
HALAMAN PERSEMBAHAN	viii
KATA PENGANTAR	ix
DAFTAR ISI	x
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xv
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
1.5 Batasan Masalah	4
II LANDASAN TEORI	5
2.1 Tinjauan Pustaka	5
2.2 Konsep <i>Big Data</i>	5
2.3 Ekstraksi Fitur Teks (<i>Text Feature Extraction</i>)	7
2.3.1 <i>Bag of Words</i> (BoW)	7
2.3.2 <i>Term Frequency-Inverse Document Frequency</i> (TF-IDF)	8
2.3.3 Penggunaan <i>Word Count</i> dan <i>Sort</i> pada BoW dan TF-IDF	8
2.4 Komputasi Awan (<i>Cloud Computing</i>)	8
2.5 <i>Shell Script</i>	9
2.6 MapReduce	10
2.6.1 Apache Hadoop	11
2.6.2 Mode Kerja Hadoop	12
2.6.3 Hadoop Distributed File System (HDFS)	13

2.6.4	Hadoop YARN	14
2.7	Apache Spark	15
2.7.1	Arsitektur Spark	15
2.7.2	Integrasi Hadoop dan Spark	16
2.7.3	Keterbatasan <i>Data Sharing</i> pada MapReduce	17
2.7.4	Solusi <i>Data Sharing</i> dengan Spark RDD	17
2.8	HiBench	18
2.8.1	Beban Kerja <i>Micro Benchmark</i> dan Sumber Data	19
2.8.2	<i>Data Generation</i> pada <i>Word Count</i> dan <i>Sort</i>	20
2.8.3	Beban Kerja <i>Word Count</i>	21
2.8.4	Beban Kerja <i>Sort</i>	21
2.9	Data Keluaran HiBench dan Dool	22
III	METODOLOGI PENELITIAN	25
3.1	Alur Penelitian	25
3.2	Penjabaran Langkah Penelitian	26
3.2.1	Identifikasi Masalah dan Studi Literatur	26
3.2.2	Membangun <i>Virtual Machine</i> di DigitalOcean	26
3.2.3	Pemasangan dan Konfigurasi Perangkat Lunak	27
3.2.4	Eksperimen	31
3.2.5	Analisis dan Evaluasi Hasil Eksperimen	35
IV	HASIL DAN PEMBAHASAN	37
4.1	Hasil Penelitian	37
4.1.1	Persebaran Waktu Eksekusi pada Hadoop dan Spark	37
4.1.2	Persebaran <i>Throughput</i> pada Hadoop dan Spark	39
4.1.3	Rata-rata Waktu Eksekusi pada Hadoop dan Spark	40
4.1.4	Rata-rata <i>Throughput</i> pada Hadoop dan Spark	41
4.1.5	Rate of Change	42
4.1.6	Penggunaan CPU	42
4.1.7	Utilisasi Sistem	43
V	KESIMPULAN DAN SARAN	53
5.1	Kesimpulan	53
5.2	Saran	53
DAFTAR PUSTAKA	54	
LAMPIRAN	59	
A Pembuatan <i>Virtual Machine</i> (VM) pada DigitalOcean	60	

B	Instalasi dan Konfigurasi Perangkat Lunak Prasyarat	63
C	Instalasi dan Konfigurasi Hadoop	66
D	Instalasi dan Konfigurasi Spark	71
E	Instalasi dan Konfigurasi HiBench	72
F	Skrip Otomatisasi Eksperimen	75
G	Visualisasi Utilisasi Sistem Sesuai Input Data (<i>Sort</i>)	78
H	Visualisasi Utilisasi Sistem Sesuai Input Data (<i>Word Count</i>)	83

DAFTAR GAMBAR

Gambar 2.1	Beberapa Alat di Dunia Data [23]	6
Gambar 2.2	Contoh Shell Script yang Digunakan pada Penelitian	10
Gambar 2.3	Cara Kerja MapReduce	10
Gambar 2.4	Implementasi MapReduce pada Word Count [31]	11
Gambar 2.5	Arsitektur Hadoop	12
Gambar 2.6	Mode Kerja Hadoop [36]	13
Gambar 2.7	Arsitektur HDFS [38]	13
Gambar 2.8	Arsitektur YARN [40]	14
Gambar 2.9	Komponen Spark	15
Gambar 2.10	Arsitektur Spark	16
Gambar 2.11	Integrasi Spark dan Hadoop	16
Gambar 2.12	<i>Data Sharing</i> pada MapReduce [43]	17
Gambar 2.13	<i>Data Sharing</i> pada RDD [43]	18
Gambar 2.14	Proses yang Terjadi di HiBench [19]	19
Gambar 2.15	Contoh Input dan Output <i>Word Count</i>	22
Gambar 2.16	Contoh Input dan Output <i>Sort</i>	22
Gambar 2.17	Data Keluaran HiBench dan Dool	24
Gambar 3.1	Diagram Alir Penelitian	25
Gambar 3.2	Alur Instalasi Perangkat Lunak	28
Gambar 3.3	Alur Instalasi Perangkat Lunak Prasyarat	29
Gambar 3.4	Alur Instalasi dan Konfigurasi Hadoop	30
Gambar 3.5	Alur Instalasi dan Konfigurasi Spark	31
Gambar 3.6	Alur Instalasi dan Konfigurasi HiBench	31
Gambar 3.7	Total Percobaan	32
Gambar 3.8	<i>End-to-end</i> Penelitian	33
Gambar 3.9	Contoh Percobaan	34
Gambar 4.1	Persebaran Waktu Eksekusi <i>Word Count</i> (Hadoop, Spark) .	37
Gambar 4.2	Persebaran Waktu Eksekusi <i>Sort</i> (Hadoop, Spark)	38
Gambar 4.3	<i>Throughput Word Count</i> (Hadoop, Spark)	39
Gambar 4.4	<i>Throughput Sort</i> (Hadoop, Spark)	40
Gambar 4.5	Rata-rata Waktu Eksekusi (<i>Sort</i>)	41
Gambar 4.6	Rata-rata Waktu Eksekusi (<i>Word Count</i>)	42
Gambar 4.7	Rata-rata <i>Throughput</i> (<i>Sort</i>)	43
Gambar 4.8	Rata-rata <i>Throughput</i> (<i>Word Count</i>)	44

Gambar 4.9 dur	46
Gambar 4.10 th	47
Gambar 4.11 hadoop-spark	48
Gambar 4.12 Penggunaan CPU (Sort)	48
Gambar 4.13 Penggunaan CPU (Word Count)	49
Gambar 4.14 State (Sort)	49
Gambar 4.15 State (Word Count)	50
Gambar 4.16 Utilisasi Sistem (Sort) pada Input Data 100 KB	50
Gambar 4.17 Utilisasi Sistem (Word Count) pada Input Data 100 KB	51
Gambar 4.18 Utilisasi Sistem (Sort) pada Input Data 15 GB	51
Gambar 4.19 Utilisasi Sistem (Word Count) pada Input Data 15 GB	52

DAFTAR TABEL

Tabel 2.1	Penelitian Terdahulu	5
Tabel 2.2	Beban Kerja pada HiBench [19]	20
Tabel 3.1	Konfigurasi Perangkat Keras	27
Tabel 3.2	Perangkat Lunak yang Dibutuhkan	27
Tabel 3.3	Variasi Input Data	34

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perusahaan dan organisasi menghasilkan dan menyimpan data dalam skala besar setiap hari dengan tingkat pertumbuhannya yang dinamis [1]. Pertumbuhan jumlah data diperkirakan akan meningkat hingga 5x lipat pada tahun 2025 dengan *Global Datasphere* diproyeksikan tumbuh dari 33 *Zettabytes* (ZB) pada tahun 2018 menjadi 175 ZB [2] pada tahun 2025. Jumlah data tersebut membutuhkan pengolahan dengan kecepatan tinggi sehingga dapat dimanfaatkan untuk keperluan bisnis dan pengambilan keputusan [3]. Analisis data transaksi nasabah pada perbankan dapat digunakan untuk mendeteksi *fraud* dan meningkatkan keamanan [4]. Data pasien pada bidang kesehatan dapat memantau wabah penyakit dan menemukan pola pengobatan yang optimal [5]. Sementara itu, data interaksi pengguna pada *e-commerce* diolah untuk memberikan rekomendasi produk personal dan merancang strategi peningkatan penjualan [6]. Semakin besar data yang bisa ditangani, semakin banyak peluang analisis dan *value* yang bisa dihasilkan. Namun, semakin besar volume data yang harus diolah, semakin kompleks pula tantangan yang dihadapi dalam mengelolanya secara efisien dan efektif [7].

Tantangan utama dalam mengelola volume data yang besar adalah memastikan tersediaan sumber daya komputasi yang memadai [7]. Pendekatan konvensional pemrosesan data besar seperti memperbanyak jumlah *storage* secara vertikal, dan penggunaan sistem basis data NoSQL dapat memungkinkan pengolahan data yang skalabel dan fleksibel. Namun, ketika skala dan kompleksitas data semakin meningkat, komputasi terdistribusi menjadi pilihan yang lebih tepat karena memiliki sifat *fault-tolerant*[8].

Komputasi terdistribusi adalah cara untuk mencapai paralelisme dengan menggabungkan beberapa mesin independen yang berbeda [9]. Dalam komputasi terdistribusi, data besar dibagi ke dalam sejumlah *node* atau server yang bekerja bersama-sama untuk mengolahnya [10]. Dua teknologi yang umum digunakan dalam komputasi terdistribusi ini adalah Apache Hadoop dan Apache Spark [11]. Hadoop dan Spark adalah dua platform komputasi *big data* yang paling populer dan banyak digunakan di seluruh dunia. Teknologi ini menawarkan berbagai kemampuan untuk mengelola, menyimpan, dan menganalisis data dalam skala besar.

MapReduce adalah alat yang digunakan untuk komputasi terdistribusi, dirancang khusus untuk menulis, membaca, dan memproses jumlah data yang besar [12].

Pemrosesan data dalam MapReduce ini terdiri dari tiga tahap: fase *Map*, fase *Shuffle*, dan fase *Reduce*. Dalam teknik ini, berkas-berkas besar dibagi menjadi beberapa blok kecil dengan ukuran yang sama dan didistribusikan ke seluruh klaster untuk penyimpanan. MapReduce dan sistem file terdistribusi (HDFS) adalah bagian inti dari sistem Hadoop, sehingga komputasi dan penyimpanan bekerja bersama-sama di seluruh *node* yang membentuk klaster komputer [13]. Hadoop MapReduce memerlukan akses ke penyimpanan untuk membaca dan menulis data, sehingga dapat memperlambat proses komputasi, sehingga hadirlah Spark.

Spark, di sisi lain, menawarkan teknologi *Resilient Distributed Datasets* (RDDs) untuk mendukung proses *Map* dan *Reducing* secara lebih efektif dan cepat [14]. Spark bukan hanya alternatif Hadoop, tetapi juga menyediakan berbagai fungsi, misalnya mendukung *MLib*, *GraphX*, dan *Spark streaming* untuk analisis data besar [15]. Spark menggunakan memori untuk menyimpan data sehingga dapat mengurangi siklus baca dan tulis. Perbedaan mendasar ini mengakibatkan menarik untuk melihat perbandingan performa antara keduanya. Salah satu cara untuk membandingkan performa keduanya adalah menggunakan tolok ukur Hibench.

Tolok ukur HiBench adalah salah satu tolok ukur kinerja yang paling sering digunakan. HiBench mencakup sejumlah tugas *benchmarking* yang mencerminkan berbagai jenis pemrosesan data, seperti pengolahan batch, aliran data, *query*, atau pun *machine learning* [16]. Oleh karena itu, HiBench adalah alat yang cocok untuk mengukur dan membandingkan kinerja antara Hadoop dan Spark dalam berbagai skenario penggunaan.

Penelitian tentang evaluasi performa Hadoop dan Spark menggunakan HiBench telah beberapa kali dilakukan. Shi et al. [17] melakukan penelitian dengan dua alat yang dirancang untuk mengukur kinerja MapReduce dan Spark dalam berbagai skenario beban kerja. Penelitian tersebut mengevaluasi kinerja dalam pekerjaan *batch* dan iteratif, dengan fokus pada komponen-komponen penting seperti *shuffle*, dan *caching*. Hasil penelitian menunjukkan bahwa Spark lebih cepat daripada Hadoop dalam beberapa kasus, terutama ketika menangani tugas-tugas pemrosesan data yang lebih kecil. Namun, ketika ukuran data meningkat, Hadoop terbukti lebih efisien. Selanjutnya, perbandingan kinerja antara Hadoop dan Spark juga disorot oleh penelitian Samadi et al. [13], yang menggunakan delapan tolok ukur dari HiBench. Penelitian ini menunjukkan bahwa Spark cenderung lebih efisien ketika menangani data dalam jumlah kecil atau saat memproses tugas dalam memori, sementara Hadoop lebih sukses ketika beban kerja melibatkan operasi I/O penyimpanan yang intensif. Selain itu, penelitian oleh Satish dan Rohan [18] menyoroti perbandingan kinerja antara Hadoop dan Spark khususnya dalam konteks algoritma *K-means*.

Penelitian itu menemukan bahwa Spark dapat mencapai kecepatan hingga tiga kali lipat dibandingkan Hadoop, dengan catatan bahwa performa Spark sangat bergantung pada ukuran memori yang memadai.

Berdasarkan penelitian sebelumnya, penelitian ini bertujuan untuk menyelidiki perbandingan kinerja antara Hadoop dan Spark dengan menggunakan tolok ukur HiBench dengan studi kasus tertentu. Pemahaman mendalam mengenai kekuatan dan kelemahan masing-masing teknologi dalam berbagai konteks pemrosesan data akan membuat organisasi atau peneliti dapat dengan mudah membuat keputusan yang lebih informasional. Selain itu, penelitian ini akan dilakukan dengan memanfaatkan Infrastruktur sebagai Layanan (IaaS) yang disediakan oleh DigitalOcean, memungkinkan penggunaan sumber daya komputasi dalam skala yang fleksibel dan efisien. Dengan demikian, penelitian ini akan memberikan kontribusi berharga dalam membantu pemangku kepentingan dalam pemilihan teknologi pemrosesan data dalam lingkungan komputasi terdistribusi .

1.2 Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana implementasi Hadoop, Spark, dan HiBench di DigitalOcean?
2. Bagaimana kinerja Hadoop dan Spark ketika diuji menggunakan beban kerja *Micro Benchmarks* yang disediakan oleh HiBench?
3. Bagaimana perbandingan kinerja antara Hadoop dan Spark dalam mode *pseudo distributed* dalam konteks pemrosesan data dalam skala besar dengan menggunakan tolok ukur HiBench?

1.3 Tujuan

Penelitian ini memiliki tujuan, yaitu:

1. Untuk mengetahui implementasi Hadoop, Spark, dan HiBench di DigitalOcean.
2. Untuk mengetahui kinerja Hadoop dan Spark ketika diuji menggunakan beban kerja *Micro Benchmarks* yang disediakan oleh HiBench.
3. Untuk mengetahui perbandingan kinerja antara Hadoop dan Spark dalam mode *pseudo distributed* saat memproses data dalam skala besar dengan menggunakan tolok ukur HiBench.

1.4 Manfaat

Hasil dari penelitian ini diharapkan akan memberikan manfaat sebagai berikut:

1. Penelitian ini akan memberikan informasi yang berguna bagi organisasi yang sedang mempertimbangkan pemilihan platform *Big Data*, sehingga *stakeholder* dapat membuat keputusan yang lebih terinformasi.
2. Penelitian ini akan membantu dalam memahami lebih dalam kinerja Hadoop dan Spark dalam berbagai skenario pemrosesan data.
3. Hasil dari penelitian ini dapat menjadi dasar untuk penelitian lebih lanjut dalam pengembangan dan peningkatan platform *Big Data*.

1.5 Batasan Masalah

Penelitian ini memiliki beberapa batasan yang perlu diperhatikan sebagai berikut:

1. Penelitian ini akan fokus pada perbandingan kinerja antara Hadoop dan Spark dalam mode *pseudo-distributed*.
2. Pengujian kinerja akan menggunakan HiBench, sebuah tolok ukur kinerja yang umum digunakan dalam penelitian *Big Data*.
3. Implementasi Hadoop dan Spark akan menggunakan salah satu penyedia layanan awan, yaitu *DigitalOcean*.
4. Penelitian ini akan berfokus pada aspek kinerja. Aspek lain seperti keamanan dan administrasi tidak akan dibahas secara rinci.

BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan beberapa teori dasar supaya memperjelas proses penelitian dan memberikan pemahaman lebih lanjut. Peneltian terdahulu mengenai evaluasi performa Hadoop dan Spark dapat dilihat pada Tabel 2.1.

Tabel 2.1 Penelitian Terdahulu

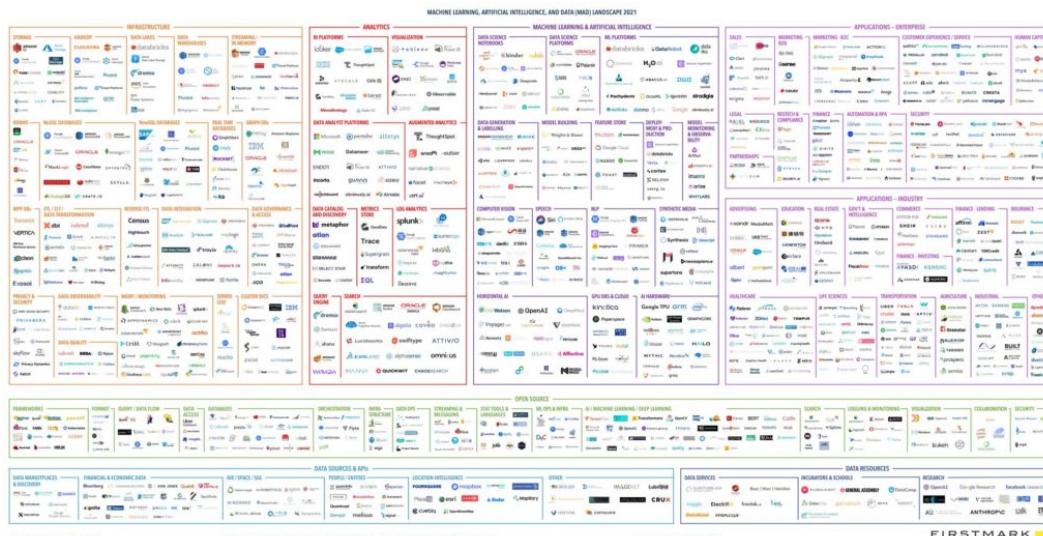
No.	Nama Peneliti	Tahun	Judul	Metode	Hasil Penelitian
1	N. Ahmed, Andre L. C. Barczak, Teo Susnjak, Mohammed A. Rashid [10]	2020	<i>A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench</i>	Penelitian ini menyelidiki parameter-parameter yang paling berdampak, yaitu <i>input splits</i> , dan <i>shuffle</i> , untuk membandingkan kinerja antara Hadoop dan Spark, dengan menggunakan klaster yang diimplementasikan di laboratorium. Guna mengevaluasi kinerja, dua beban kerja dipilih, yakni WordCount dan TeraSort. Metrik kinerja diukur berdasarkan tiga kriteria: waktu eksekusi, <i>throughput</i> , dan <i>speedup</i> .	Kinerja kedua sistem sangat bergantung pada ukuran data masukan dan pemilihan parameter yang tepat. Analisis hasil menunjukkan bahwa Spark memiliki kinerja yang lebih baik dibandingkan dengan Hadoop ketika set data kecil, mencapai peningkatan kecepatan hingga dua kali lipat dalam beban kerja WordCount dan hingga 14 kali lipat dalam beban kerja TeraSort ketika nilai parameter default dikonfigurasi ulang.
2	Rendiyono Wahyu Saputro, Aminuddin, Yuda Munarko [11]	2020	Perbandingan Kinerja Komputasi Hadoop dan Spark untuk Memprediksi Cuaca (Studi Kasus: <i>Storm Event Database</i>)	Mengimplementasikan gugus komputer untuk memproses dataset dengan berbagai ukuran dan dalam jumlah komputer yang berbeda.	Hadoop memerlukan waktu yang lebih sedikit dibandingkan dengan Spark. Hal tersebut karena nilai <i>throughput</i> dan <i>throughput/node</i> Hadoop lebih tinggi daripada Apache Spark.
3	Yassir Samadi, Mostapha Zbakh, Claude Tadonki [1]	2018	<i>Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks</i>	Perbandingan kinerja diimplementasikan pada mesin virtual (VM). Untuk membandingkannya, digunakan HiBench. Perbandingan dilakukan berdasarkan tiga kriteria: waktu eksekusi, <i>throughput</i> , dan <i>speedup</i> . Beban kerja WordCount diuji dengan ukuran data yang berbeda.	Spark lebih efisien dibandingkan Hadoop dalam menangani jumlah data yang besar. Namun, Spark memerlukan alokasi memori yang lebih tinggi, karena memuat data yang akan diproses ke dalam memori dan menyimpannya dalam cache untuk sementara.
4	Satish Gopalani, Rohan Arora [18]	2015	<i>Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means</i>	Hadoop dan Spark dibandingkan menggunakan algoritma pembelajaran mesin (KMeans). Ukuran data yang digunakan adalah sebesar 64MB, 1240MB dengan satu node, dan 1240MB dengan dua node.	Hasil-hasil penelitian dengan jelas menunjukkan bahwa kinerja Spark jauh lebih tinggi dari segi waktu, di mana setiap ukuran dataset mengakibatkan penurunan waktu pemrosesan hingga tiga kali lipat dibandingkan dengan Hadoop.

2.2 Konsep *Big Data*

Big Data biasanya sering didefinisikan bersama dengan kompleksitas suatu data [19]. Berbeda dengan tradisional data, *Big Data* merujuk pada pertumbuhan data dalam berbagai format, baik dari struktur, semi-terstruktur, dan tidak terstruktur [20]. *Big Data* memiliki banyak jenis sehingga membutuhkan teknologi yang lebih

bertenaga serta algoritma yang lebih canggih. Pendekatan teknologi yang sering digunakan oleh *Business Intelligence* biasanya tidak dapat lagi efisien jika digunakan. *Big Data* biasanya didefinisikan menjadi tiga karakteristik (3V), yaitu *Volume*, *Velocity*, dan *Variety* [21]. *Volume* berkaitan dengan jumlah data yang terbentuk atau dibuat secara terus menerus oleh beragam perangkat, seperti telepon genggam dan aplikasi (sosial media, sensor, IoT). Jumlah data diharapkan tumbuh 5x lipat pada tahun 2020 [21]. Selanjutnya, *Velocity* memberikan makna bahwa data bertumbuh secara cepat dan harus diproses secara cepat juga untuk memberikan informasi yang berguna [22]. YouTube adalah ilustrasi yang tepat untuk menggambarkan bagaimana pertumbuhan data begitu cepat. Terakhir, *Variety* berkaitan dengan variasi sumber dan format data.

Penerapan dari *big data* tidak hanya terbatas pada pengumpulan dan penyimpanan data, tetapi juga meliputi analisis dan pengolahan data tersebut untuk menghasilkan wawasan yang berguna. Beberapa sektor yang telah menerapkan *big data* secara luas meliputi kesehatan, keuangan, ritel, dan pemerintahan [20]. Dalam sektor kesehatan, *big data* digunakan untuk menganalisis informasi pasien secara massal guna meningkatkan kualitas perawatan dan menemukan pola-pola penyakit. Se-mentara itu, di sektor keuangan, *big data* membantu dalam analisis risiko, deteksi penipuan, dan personalisasi layanan untuk pelanggan.



Gambar 2.1 Beberapa Alat di Dunia Data [23]

Perkembangan alat dalam *big data* juga sangat pesat seperti pada Gambar 2.1. Salah satu contoh signifikan adalah penggunaan teknologi *machine learning* dan *artificial intelligence* (AI) dalam pengolahan data. AI dan *machine learning* memungkinkan analisis data yang lebih akurat dan cepat, bahkan dengan volume dan variasi data yang sangat besar. Alat seperti Hadoop dan Spark telah menjadi standar dalam

industri untuk mengelola dan memproses data besar. Selain itu, penggunaan *cloud computing* dalam *big data* memungkinkan penyimpanan dan pengolahan data dalam skala yang lebih besar dan lebih fleksibel.

2.3 Ekstraksi Fitur Teks (*Text Feature Extraction*)

Ekstraksi Fitur Teks adalah salah satu proses pada pembelajaran mesin (*machine learning*) dan data analisis yang melibatkan identifikasi dan ekstraksi fitur yang relevan dari data mentah. Fitur-fitur tersebut nantinya akan digunakan untuk membuat data yang lebih informatif, sehingga dapat digunakan untuk klasifikasi, prediksi, dan klasterisasi. Ekstraksi fitur bertujuan untuk mengurangi kompleksitas data (atau yang sering disebut juga *Data Dimensionality*) namun tetap menyimpan sebanyak mungkin informasi yang paling relevan. Hal ini bertujuan untuk meningkatkan performa dan efisiensi algoritma pada pembelajaran mesin dan mempermudah dalam proses analisis. Ekstraksi fitur dapat melibatkan membuat fitur baru (*Feature Engineering*) atau memanipulasi data yang menghasilkan fitur yang berguna. Ekstraksi fitur juga memainkan peran penting dalam banyak penerapan di dunia nyata, misalnya untuk pemrosesan teks dan *Natural Language Processing* (NLP). Dalam skenario ini, data mentah mungkin mengandung banyak fitur yang tidak relevan atau berlebihan. Hal ini menyulitkan algoritme untuk memproses data secara akurat. Dengan melakukan ekstraksi fitur, fitur yang relevan dipisahkan dari fitur yang tidak relevan. Dengan lebih sedikit fitur yang harus diproses, kumpulan data menjadi lebih sederhana dan akurasi serta efisiensi analisis meningkat.

2.3.1 *Bag of Words* (BoW)

BoW adalah teknik sederhana yang mengabaikan urutan dan struktur gramatikal kalimat, dan hanya berfokus pada frekuensi kemunculan kata dalam dokumen. Prosesnya melibatkan langkah-langkah berikut:

1. **Tokenisasi:** Teks dipecah menjadi kata-kata individual (token).
2. **Pembuatan Kosakata:** Daftar unik dari semua token yang ada dalam seluruh kumpulan dokumen dibuat. Ini disebut "kosakata".
3. **Penghitungan Kata (Word Count):** Untuk setiap dokumen, frekuensi kemunculan setiap kata dalam kosakata dihitung.
4. **Representasi Vektor:** Setiap dokumen diwakili sebagai vektor, di mana setiap elemen vektor mewakili frekuensi kata tertentu dalam kosakata.

BoW mudah diimplementasikan dan efisien, namun kelemahannya adalah kehilangan informasi kontekstual dan semantik. Kata-kata dengan frekuensi tinggi, meskipun kurang informatif, dapat mendominasi representasi vektor.

2.3.2 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF mengatasi beberapa kelemahan BoW dengan mempertimbangkan pentingnya kata dalam dokumen dan koleksi dokumen. TF-IDF terdiri dari dua komponen:

1. **Term Frequency (TF):** Mengukur seberapa sering suatu kata muncul dalam dokumen.
2. **Inverse Document Frequency (IDF):** Mengukur seberapa penting suatu kata dalam seluruh koleksi dokumen. Kata-kata yang muncul di banyak dokumen (seperti kata "yang") memiliki IDF rendah, sementara kata-kata yang jarang muncul (dan kemungkinan lebih informatif) memiliki IDF tinggi.

Dengan mengalikan TF dan IDF, kita mendapatkan nilai TF-IDF yang mencerminkan pentingnya kata dalam dokumen dan koleksi dokumen.

2.3.3 Penggunaan Word Count dan Sort pada BoW dan TF-IDF

1. **Word Count:** Digunakan dalam kedua metode (BoW dan TF-IDF) untuk menghitung frekuensi kemunculan kata dalam dokumen.
2. **Sort:** Biasanya tidak digunakan secara langsung dalam proses BoW atau TF-IDF. Namun, pengurutan dapat digunakan untuk:
 - **Memilih fitur:** Memilih fitur dengan nilai TF-IDF tertinggi untuk mengurangi dimensi data dan fokus pada kata-kata yang paling informatif.
 - **Visualisasi:** Mengurutkan kata berdasarkan frekuensi atau nilai TF-IDF dapat membantu dalam visualisasi dan analisis data.

2.4 Komputasi Awan (*Cloud Computing*)

Komputasi awan didefinisikan sebagai sistem informasi yang memungkinkan akses mudah ke sumber daya komputasi atau layanan komputasi sesuai permintaan (*on demand*), misalnya segala sesuatu mulai dari aplikasi (Google Mail, Microsoft One Drive, Siakad Itera) hingga pusat data di seluruh internet dengan sistem bayar sesuai penggunaan. Sistem komputasi awan saat ini menyediakan tiga layanan utama:

1. *Infrastructure as a service* (IaaS), adalah layanan awan yang menawarkan kepada pengguna untuk mengatur dan mengonfigurasikan sumber daya yang dibutuhkan untuk menjalankan aplikasi dan sistem IT. Jenis IaaS biasanya berbentuk komputasi, penyimpanan, dan sumber daya jaringan yang dibuat sebagai layanan.
2. *Platform as a service* (PaaS), adalah layanan awan yang memungkinkan pengguna untuk mengembangkan, mengelola, dan menjalankan aplikasi di lingkungan yang dikontrol oleh penyedia layanan, tanpa harus khawatir dengan infrastruktur yang mendasarinya.

3. *Software as a service* (SaaS), adalah layanan awan yang mengacu pada aplikasi yang berjalan pada infrastruktur awan yang di-hosting oleh vendor atau penyedia layanan dan tersedia untuk pengguna akhir melalui browser web.

Komputasi awan menjadi salah satu aspek terpenting dalam menjalankan komputasi yang kompleks, misalnya untuk menjalankan Hadoop atau Spark. Salah satu komputasi awan yang dapat diandalkan adalah DigitalOcean. DigitalOcean dibentuk pada tahun 2012 untuk memenuhi kebutuhan pengembang untuk mendapatkan akses komputasi awan yang mudah dimengerti dan terjangkau [24]. Salah satu produk DigitalOcean yang sering digunakan adalah Droplet, *easy-to-use* komputer virtual yang siap digunakan dalam hitungan menit. Pengguna dapat memilih lokasi dimana komputer akan dijalankan, bagaimana konfigurasi prosesor serta memori, memilih sistem operasi apa yang akan digunakan, dan banyak hal lainnya.

2.5 *Shell Script*

Shell script merupakan serangkaian perintah yang dieksekusi dalam lingkungan sistem operasi Unix atau Unix-like [25]. *Shell script* memungkinkan pengguna untuk mengotomatiskan tugas-tugas rutin, melakukan pemrosesan file, dan bahkan membangun aplikasi yang kompleks dengan menggunakan perintah-perintah shell. *Shell script* umumnya ditulis menggunakan bahasa pemrograman shell, seperti Bash (Bourne Again Shell), yang merupakan shell standar pada sebagian besar sistem operasi Linux dan MacOS. Sebagai contoh, dalam mengelola pencadangan sistem, seorang administrator dapat membuat *shell script* sederhana yang menjalankan perintah-perintah untuk menyalin file-file penting ke lokasi penyimpanan cadangan secara berkala. Skrip ini dapat dijadwalkan untuk berjalan secara otomatis menggunakan *cron job*, sehingga proses pencadangan dapat dilakukan tanpa campur tangan manusia secara berkala. Dengan menggunakan variabel dan logika sederhana, administrator dapat dengan mudah menyesuaikan skrip ini untuk memenuhi kebutuhan pencadangan spesifik sistem mereka. Dengan demikian, *shell script* tidak hanya menghemat waktu dan tenaga, tetapi juga meningkatkan kehandalan dan konsistensi dalam administrasi sistem.

Gambar 2.2 menampilkan sebuah contoh potongan *shell script* yang digunakan dalam penelitian ini. Berikut adalah penjelasan lebih rinci mengenai skrip tersebut,

1. **Baris 1:** Mendefinisikan interpreter Bash untuk menjalankan skrip.
2. **Baris 3-4:** Mengubah direktori kerja ke direktori HiBench.
3. **Baris 6-7:** Mendefinisikan daftar (*list*) *workLoads* berisi macam-macam beban kerja, yaitu *wordcount* dan *sort*.
4. **Baris 9-15:** Mendefinisikan daftar *scales* berisi skala input data.

```

hadoop@ubuntu-s-4vcpu-8gb-amd-sgp1-01:~$ cat bench_v3.sh
#!/bin/bash

# Ubah direktori kerja ke direktori HiBench
cd /home/hadoop/HiBench-HiBench-7.0

# Daftar workload
workLoads=("wordcount" "sort")

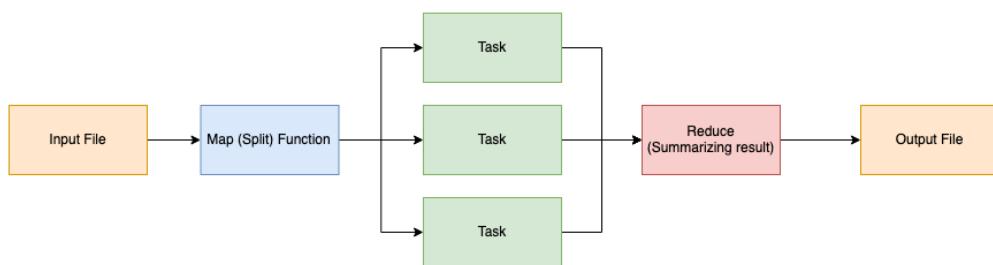
# Daftar skala
scales=(
    "seratuskb"
    "limaratuskb"
    "satumb"
    "Limamb"
    "sepuluhmb"
    "limapuluhmb"
    "seratusmb"
)

```

Gambar 2.2 Contoh Shell Script yang Digunakan pada Penelitian

2.6 MapReduce

MapReduce adalah model pemrograman dan implementasi teknik pemrosesan data berukuran besar yang pertama kali dipopulerkan oleh Google pada tahun 2004[26]. MapReduce menawarkan pemrosesan data yang dapat diandalkan serta *fault-tolerant manner* (tahan terhadap kesalahan). MapReduce berjalan secara paralel dan bera-*da* pada lingkungan komputasi terdistribusi [27]. Model ini mengadopsi arsitektur tersentraliasi, yaitu satu *node* berperan sebagai *master* dan *node* yang lain berperan sebagai *workers* atau *slave* [28], [29]. *Master node* bertanggung jawab untuk melakukan penjadwalan kerja, dan *slave node* berperan untuk menjalankan eksekusi kerja.

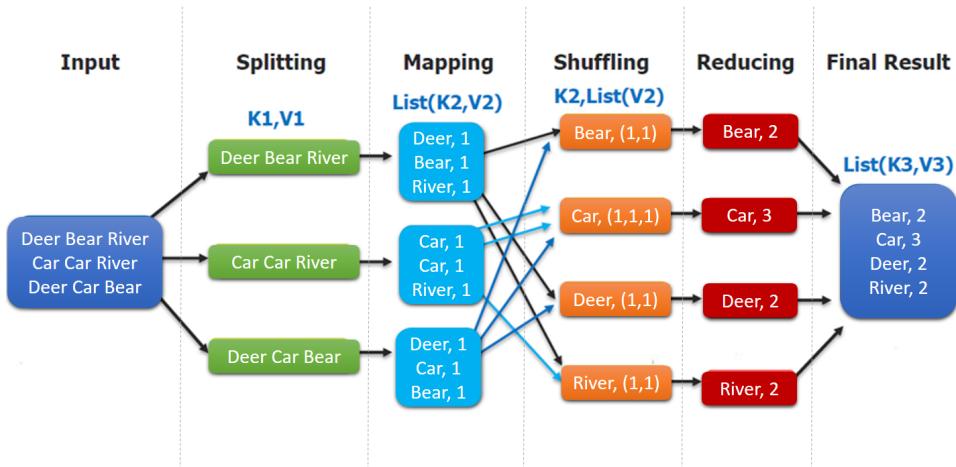


Gambar 2.3 Cara Kerja MapReduce

MapReduce terdiri dari fungsi *Map* dan fungsi *Reduce* [30]. Kedua fungsi ini tersebar di seluruh *slave node* yang terhubung dalam klaster dan berjalan secara paralel. Fungsi *Map* berperan untuk membagi masalah besar menjadi masalah yang lebih kecil dan mendistribusikannya ke *slave node*. Hasil pemrosesan dari *slave node* akan dikumpulkan oleh *master node* melalui fungsi *Reduce*. Sesuai dengan Gambar 2.3, hasil dari proses *Reduce* yang akan dikirimkan sebagai hasil akhir proses

MapReduce.

Implementasi MapReduce pada *Word Count*[7] dapat dilihat pada Gambar 2.4. Pada proses MapReduce, data masukan akan melalui beberapa tahapan pemrosesan. Pertama, data akan dipecah menjadi bagian-bagian yang lebih kecil pada proses pemecahan data masukan (*splitting*). Dalam kasus Hadoop MapReduce, data idealnya akan dipecah menjadi beberapa blok berukuran maksimal 128MB.



Gambar 2.4 Implementasi MapReduce pada Word Count [31]

Kemudian, blok data tersebut akan diproses lebih lanjut pada tahap pemetaan (*mapping*). Pemetaan merupakan salah satu tahapan terpenting dalam MapReduce. Pada tahap ini, blok data yang sudah dipecah akan diproses untuk menghasilkan pasangan kunci-nilai (*key-value pairs*) sementara, seperti pada contoh kasus *wordcount* yang menghasilkan pasangan kunci-nilai *Dear:1*, *Bear:1*, dan *River:1*. Pemetaan dapat melibatkan satu atau beberapa mesin pekerja (*worker*) yang memproses blok data secara paralel.

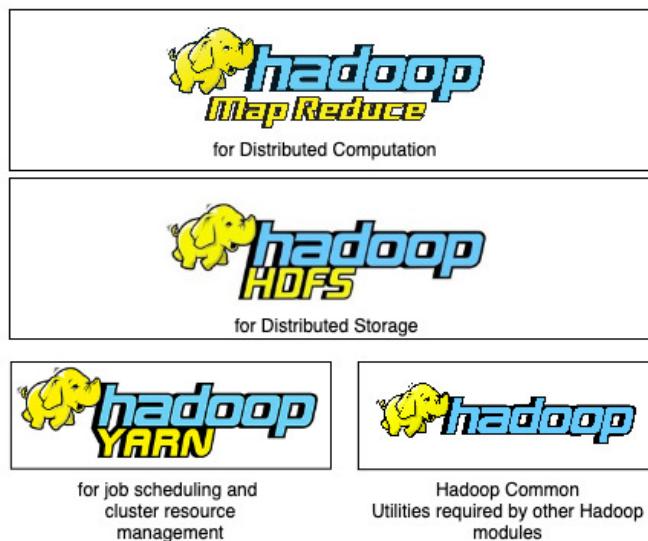
Selanjutnya adalah tahap pengocokan (*shuffling*) di mana pasangan kunci-nilai hasil pemetaan yang tersebar di beberapa mesin akan dikumpulkan berdasarkan kesamaan kuncinya agar bisa diproses lebih lanjut. Misalnya semua pasangan dengan kunci *Bear* dikumpulkan dalam satu mesin.

Pada tahap terakhir yaitu pengurangan (*reducing*), dilakukan agregasi terhadap pasangan kunci-nilai dengan kunci yang sama untuk menghasilkan keluaran akhir. Seperti pada contoh kasus *wordcount*, pasangan *Bear:1* dan *Bear:1* akan dijumlahkan menjadi *Bear:2* oleh proses pengurangan.

2.6.1 Apache Hadoop

Apache Hadoop adalah perangkat lunak sumber terbuka yang ditulis dengan bahasa pemrograman Java untuk pemrosesan dan penyimpanan data menggunakan kom-

putasi terdistribusi [32]. Hadoop dapat diinstalasi pada satu *node* komputer, atau ratusan *node* komputer yang digabungkan dalam sebuah klaster [33]. Berkaitan dengan pemrosesan data, Hadoop mengimplementasikan model MapReduce untuk pemrosesan data secara paralel dan cepat. Selain itu, Hadoop menyediakan sistem penyimpanan data terdistribusi yang dikenal sebagai Hadoop Distributed File System (HDFS) untuk akses data, pemrosesan, dan komputasi [34]. Arsitektur Hadoop secara umum dapat dilihat pada Gambar 2.5.

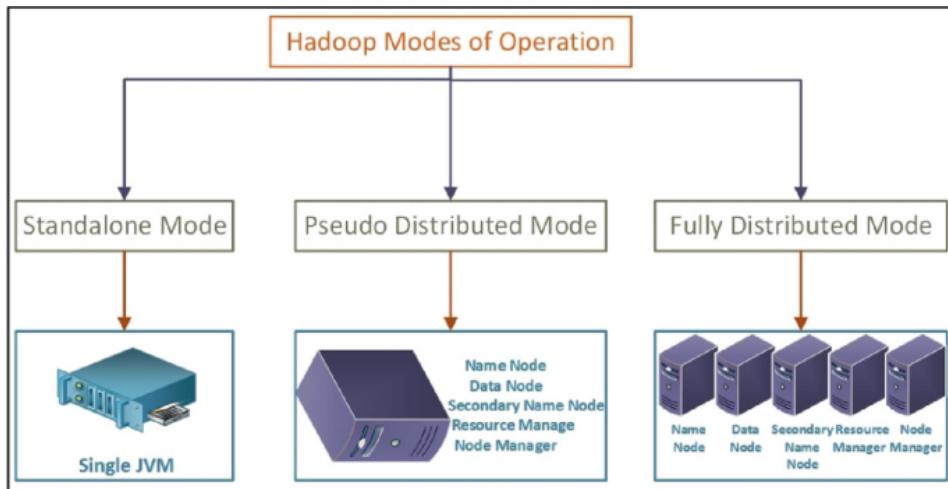


Gambar 2.5 Arsitektur Hadoop

2.6.2 Mode Kerja Hadoop

Hadoop dapat dijalankan dalam tiga mode operasi yang berbeda yaitu *standalone*, *pseudo-distributed*, dan *fully distributed* [35]. Dalam *standalone mode*, semua proses Hadoop berjalan pada satu node tunggal menggunakan sistem berkas lokal tanpa memerlukan konfigurasi kustom pada Hadoop seperti pada Gambar 2.6.

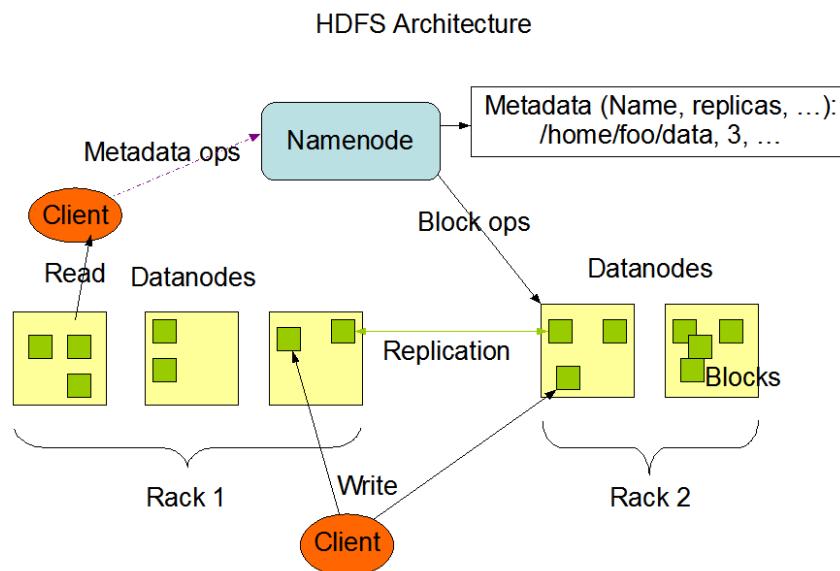
Pseudo-distributed mode menjalankan semua komponen Hadoop pada satu node tunggal tetapi menyimulasikan kluster dengan komunikasi antar proses melalui socket jaringan, sehingga memerlukan konfigurasi pada berkas *core-site*, *mapred-site*, dan *hdfs-site*. Sedangkan *fully distributed mode* menyebarkan proses Hadoop ke beberapa node dalam kluster sebenarnya yang biasanya digunakan untuk tahap produksi. *Fully distributed mode* mendukung skalabilitas, ketersediaan tinggi, dan keamanan dengan memerlukan instalasi Hadoop dan konfigurasi kluster pada setiap node.



Gambar 2.6 Mode Kerja Hadoop [36]

2.6.3 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System adalah sistem file terdistribusi yang dikembangkan sebagai bagian dari Hadoop [37]. HDFS dirancang khusus untuk menyimpan data dalam jumlah besar dan memungkinkan pemrosesan data secara paralel. Beberapa fitur utama dari HDFS antara lain skalabilitas, toleransi kesalahan, *streaming access*, dan cocok untuk aplikasi *batch* seperti MapReduce.



Gambar 2.7 Arsitektur HDFS [38]

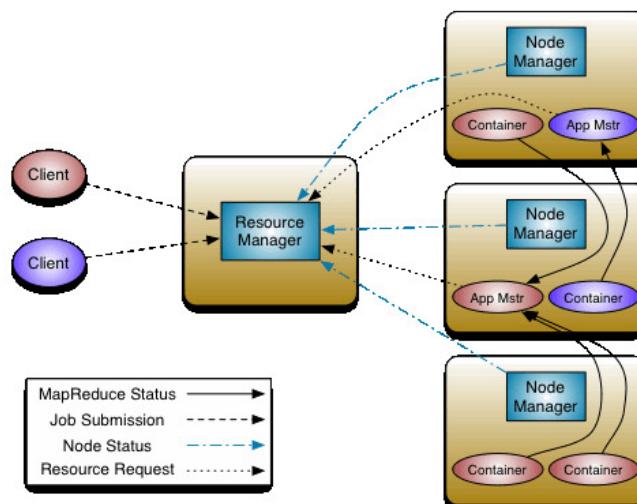
Secara struktur, HDFS terdiri dari NameNode sebagai *node master* yang mengelola *metadata* dan *namespace*, serta DataNode sebagai *node slave* yang bertugas menyimpan data sebenarnya dalam bentuk blok seperti pada Gambar 2.7. Berkas di HDFS dipartisi menjadi satu atau lebih blok berukuran 64MB atau 128MB, kemudian didistribusikan dan disimpan di beberapa DataNode. Setiap block direplikasi

(biasanya 3x) di DataNode yang berbeda untuk toleransi kesalahan. Replikasi blok di *node/rack* yang berbeda juga meningkatkan ketersediaan HDFS.

Dengan desain terdistribusi, HDFS sangat populer digunakan bersama framework Hadoop untuk memproses *big data* [39]. Namun, ketergantungan pada *single* NameNode dan performa akses data acak yang kurang optimal menjadi kelemahan utama HDFS. Secara keseluruhan, HDFS telah terbukti menjadi pilihan matang untuk penyimpanan data massal secara terdistribusi.

2.6.4 Hadoop YARN

Hadoop YARN (Yet Another Resource Negotiator) adalah manajer sumber daya dan sistem penjadwalan untuk kluster Hadoop. Komponen ini diperkenalkan dalam Hadoop 2.x sebagai evolusi dari Hadoop MapReduce 1.x, yang mengintegrasikan manajemen sumber daya dan pemrosesan data dalam satu sistem. YARN memungkinkan kluster untuk menjalankan berbagai aplikasi secara bersamaan dengan efisiensi yang lebih baik. YARN memisahkan fungsi manajemen sumber daya dari mekanisme pemrosesan data, yang sebelumnya keduanya tertanam dalam MapReduce. Dengan demikian, YARN dapat mendukung berbagai paradigma pemrosesan data di atas Hadoop, selain MapReduce, seperti *real-time processing* dan *graph processing*.



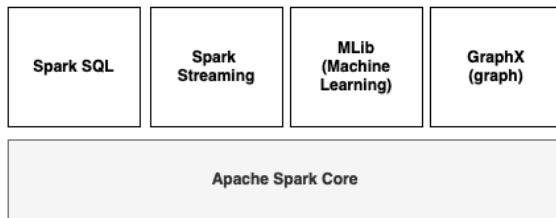
Gambar 2.8 Arsitektur YARN [40]

Struktur utama YARN terdiri dari ResourceManager yang bertugas mengkoordinasikan alokasi sumber daya di seluruh kluster, dan NodeManager yang berjalan di setiap *node* untuk mengawasi penggunaan sumber daya dan mengelola *container* tempat aplikasi dijalankan. ApplicationMaster adalah komponen khusus untuk setiap aplikasi yang bertanggung jawab untuk negosiasi sumber daya dengan Reso-

urceManager dan bekerja dengan NodeManager untuk menjalankan dan memantau *tasks* seperti pada Gambar 2.8.

2.7 Apache Spark

Apache Spark diperkenalkan oleh Apache Software Foundation sebagai *framework* pemrosesan data paralel *open-source* yang dirancang untuk mempercepat pemrosesan *big data* dibandingkan dengan Hadoop MapReduce [41]. Meskipun sama-sama menggunakan model pemrosesan MapReduce, Spark bukanlah hasil modifikasi dari Hadoop MapReduce[7]. Hal ini dikarenakan Spark menggunakan teknologi tersendiri yaitu *Resilient Distributed Datasets* (RDDs) yang memungkinkan Spark memproses data secara *in-memory* sehingga lebih cepat. Selain itu, Spark memiliki klaster pengolahan data tersendiri sehingga dapat berjalan independen tanpa Hadoop. Dengan performa tinggi serta dukungan untuk pemrosesan data secara interaktif, Spark banyak digunakan untuk pemrosesan data skala besar. Komponen yang terdapat pada Spark dapat dilihat pada Gambar 2.9

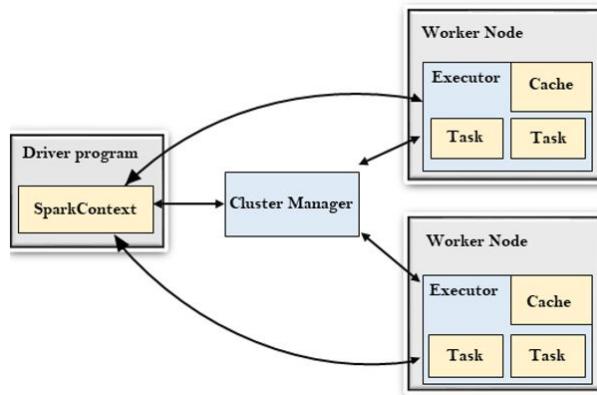


Gambar 2.9 Komponen Spark

2.7.1 Arsitektur Spark

Arsitektur Spark dirancang untuk pemrosesan data terdistribusi yang efisien dan cepat seperti pada Gambar 2.10. Komponen utamanya meliputi *Spark Driver*, *Cluster Manager*, dan *Spark Executor*. *Spark Driver* berperan sebagai otak operasi, bertanggung jawab untuk mengonversi program pengguna menjadi tugas-tugas, menjadwalkan tugas pada *executor*, dan mengelola keseluruhan alur kerja. *Cluster Manager*, yang dapat berupa YARN, Mesos, atau mode *standalone* Spark, menangani alokasi sumber daya dan peluncuran *executor* pada *node-node cluster*. *Spark Executor*, yang berjalan pada *node-node cluster*, menjalankan tugas-tugas pemrosesan data yang diberikan oleh *driver* dan menyediakan penyimpanan dalam memori untuk data yang *di-cache*. Interaksi antara komponen-komponen ini memungkinkan pemrosesan data paralel yang cepat dan toleransi kesalahan yang tinggi. Arsitektur Spark yang fleksibel mendukung berbagai bahasa pemrograman dan sistem pe-

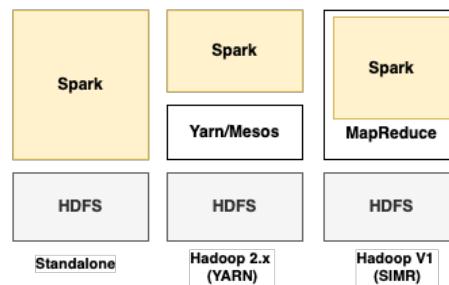
nyimpanan data, menjadikannya solusi ideal untuk berbagai kasus penggunaan data besar.



Gambar 2.10 Arsitektur Spark

2.7.2 Integrasi Hadoop dan Spark

Integrasi Spark dengan Hadoop dapat dilakukan melalui tiga metode berbeda seperti pada Gambar 2.11 [42]. Pertama, metode *Standalone* mengharuskan Spark menempati tempat di atas HDFS (*Hadoop Distributed File System*). Dalam skenario ini, Spark dan MapReduce berjalan berdampingan untuk menangani semua pekerjaan Spark pada kluster. Kedua, metode Hadoop Yarn memungkinkan Spark berjalan pada Yarn tanpa memerlukan instalasi sebelumnya atau akses *root*. Hal ini memfasilitasi integrasi Spark ke dalam ekosistem Hadoop, atau memungkinkan komponen lain berjalan di atas integrasi Hadoop dan Spark. Terakhir, metode *Spark in MapReduce* (SIMR). Dengan SIMR, pengguna dapat memulai Spark dan menggunakan *shell*-nya tanpa memerlukan akses administratif.



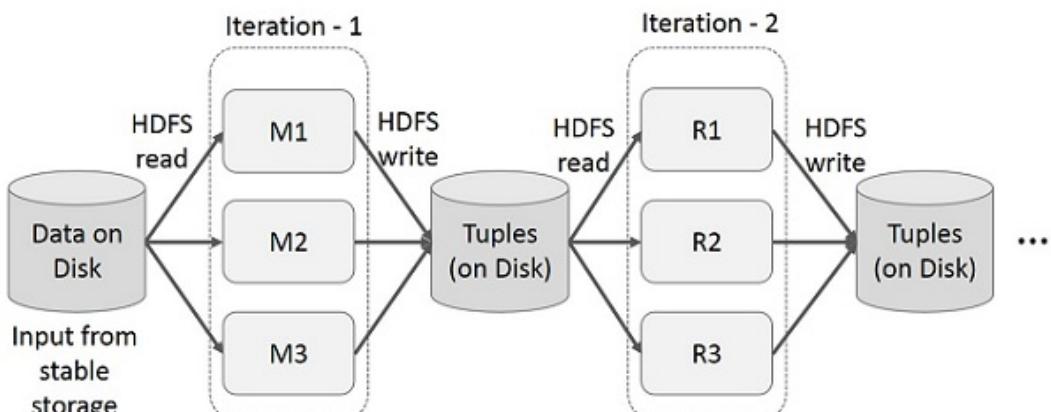
Gambar 2.11 Integrasi Spark dan Hadoop

2.7.3 Keterbatasan *Data Sharing* pada MapReduce

MapReduce, sebagai kerangka kerja pemrosesan data terdistribusi, mengandalkan sistem penyimpanan eksternal yang stabil, seperti HDFS, untuk berbagi data antar tugas (*job*). Hal ini mengakibatkan ineffisiensi karena beberapa alasan, yaitu:

1. **Replikasi Data:** Data perlu direplikasi ke beberapa node untuk toleransi kesalahan dan paralelisme. Replikasi ini memakan waktu dan bandwidth jaringan.
2. **Serialisasi/Deserialisasi:** Data perlu diubah formatnya (serialisasi) sebelum dikirim melalui jaringan dan diubah kembali (deserialisasi) di simpul tujuan. Proses ini menambah beban komputasi.
3. **Disk I/O:** Akses data dari dan ke disk cenderung lambat dibandingkan dengan akses memori. Pada MapReduce, setiap operasi baca-tulis data melibatkan interaksi dengan *disk*, yang memperlambat performa.

Keterbatasan ini terlihat jelas pada aplikasi yang membutuhkan operasi iteratif, di mana hasil antara satu tugas perlu digunakan kembali oleh tugas berikutnya. Pada MapReduce, setiap iterasi memerlukan pembacaan dan penulisan data ke HDFS, seperti yang diilustrasikan pada Gambar 2.12. Akibatnya, aplikasi iteratif pada MapReduce cenderung lambat dan tidak efisien.



Gambar 2.12 *Data Sharing* pada MapReduce [43]

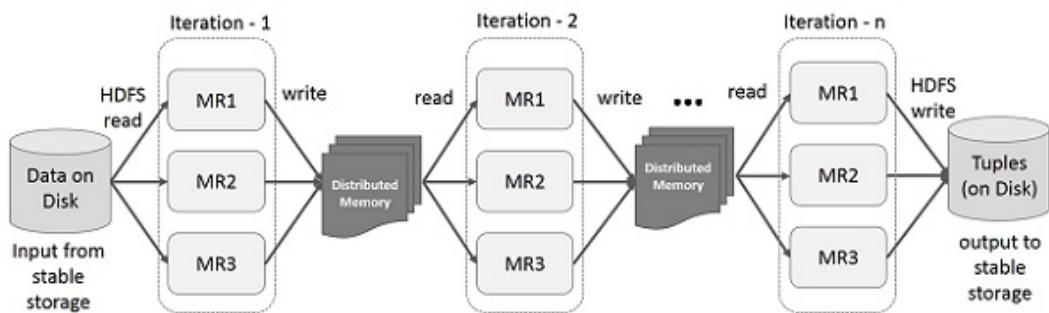
2.7.4 Solusi *Data Sharing* dengan Spark RDD

Spark mengatasi keterbatasan MapReduce dengan memperkenalkan RDD, yaitu koleksi data terdistribusi yang disimpan dalam memori. RDD bersifat *immutable*, artinya data tidak dapat diubah setelah dibuat, dan *fault-tolerant*, artinya data dapat dipulihkan jika terjadi kegagalan node.

Dengan menyimpan data dalam memori, RDD memungkinkan akses data yang jauh lebih cepat dibandingkan dengan akses disk pada MapReduce. Selain itu, RDD

mendukung *lazy evaluation*, di mana operasi pada RDD tidak dieksekusi langsung, melainkan disimpan sebagai *lineage* atau urutan operasi yang perlu dilakukan. Hal ini memungkinkan Spark untuk mengoptimalkan eksekusi tugas dan mengurangi overhead komputasi.

Pada aplikasi iteratif, RDD dapat menyimpan hasil antara dalam memori dan membagikannya antar tugas tanpa perlu mengakses disk, seperti yang ditunjukkan pada Gambar 2.13. Dengan demikian, Spark RDD memungkinkan eksekusi aplikasi iteratif yang jauh lebih cepat dan efisien dibandingkan dengan MapReduce.

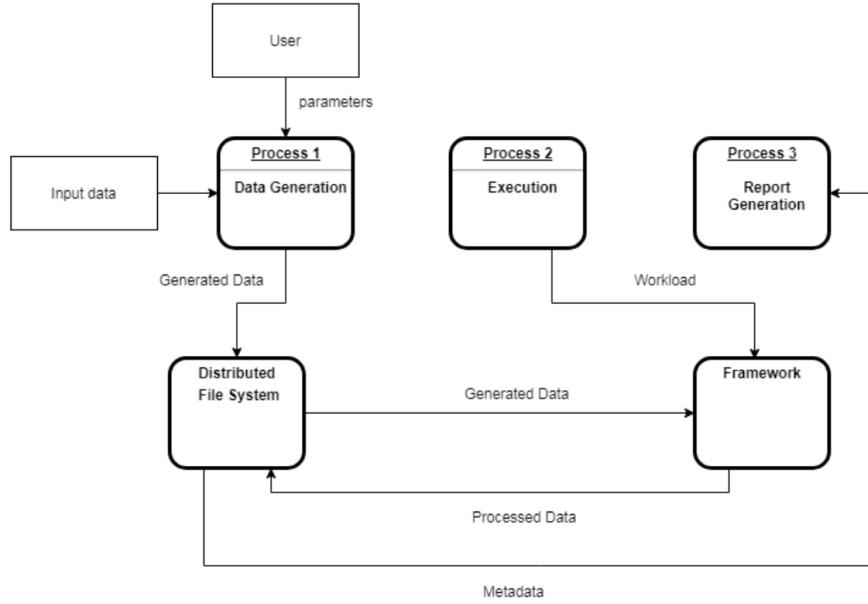


Gambar 2.13 Data Sharing pada RDD [43]

2.8 HiBench

HiBench memudahkan dalam eksekusi pengukuran berbagai beban kerja karena HiBench sudah membungkus sekumpulan perintah dalam bentuk *shell script*[1]. Pengguna hanya perlu menjalankan perintah untuk HiBench melakukan persiapan data. Selanjutnya, pengguna bisa langsung melakukan pengukuran beban kerja. Hasilnya dapat terlihat langsung pada laporan HiBench. Secara umum, alur kerja HiBench terlihat seperti pada Gambar 2.14.

HiBench terdiri dari 3 proses utama. Proses pertama, pengguna melakukan konfigurasi parameter *Data Generation*. Selanjutnya, *Data Generation* akan melakukan pembentukan data yang nantinya akan disimpan pada *Distributed File System* (DFS). Data ini yang akan digunakan pada proses selanjutnya. Proses kedua adalah proses eksekusi. Pengguna akan memicu salah satu beban kerja pada HiBench. Selanjutnya, HiBench akan memberi perintah kepada perangkat lunak (Hadoop/Spark) untuk menjalankan beban kerja tersebut. Setiap melakukan pengukuran, data yang digunakan adalah data dari *Distributed File System* yang sebelumnya sudah dibentuk. Hasil dari eksekusi ini akan disimpan kembali di DFS. Proses terakhir adalah proses pembentukan laporan. Hasil dari proses sebelumnya akan diambil serta akan dibuatkan laporan secara otomatis. Dalam laporan otomatis yang diberikan oleh HiBench, terdapat beberapa metriks yang tersedia, meliputi *Execution*



Gambar 2.14 Proses yang Terjadi di HiBench [19]

Time dan *Throughput*. *Execution Time* memiliki makna seberapa lama suatu kejadian berlangsung. Waktu yang dihitung adalah waktu diantara waktu awal dan waktu terakhir kejadian. Metriks ini dihitung dalam skala detik. Selanjutnya, *Throughput* menghitung berapa banyak unit informasi yang dapat diproses oleh sistem dalam waktu tertentu. Metriks ini dinyatakan dalam *byte*/detik.

2.8.1 Beban Kerja *Micro Benchmark* dan Sumber Data

HiBench versi 7.1 memiliki 29 beban kerja (*workload*) yang dapat diuji [44]. Beban kerja ini dikategorikan menjadi 7 kategori, yaitu *micro*, *ml* (*machine learning*), *sql*, *graph*, *websearch and streaming*. Tabel 2.2 menunjukkan macam-macam beban kerja yang dapat diuji. *Workload name* mengindikasikan algoritma utama atau operasi apa yang dilakukan. *Workload type* merepresentasikan kategori dari beban kerja. *Operation types* menunjukkan klasifikasi jenis operasi yang dilakukan. *Workload Submission Policy* berguna untuk mengetahui bagaimana cara pengguna untuk mengatur atau mengonfigurasikan beban kerja.

Beban kerja *micro benchmarks* merupakan kategori khusus yang dirancang untuk menguji kemampuan *raw processing power* [19]. Dalam kategori ini, terdapat dua beban kerja populer, yaitu Sort, dan WordCount [16]. Beban kerja Sort dan WordCount merepresentasikan pekerjaan MapReduce [12]. Beban kerja Sort akan mengurutkan setiap kata dalam berkas input. Beban kerja WordCount akan melakukan tugas pemetaan (*map task*) dan mengeluarkan output (kata, 1) untuk setiap kata dalam inputnya. Data masukan untuk beban kerja Sort dan WordCount dihasil-

Tabel 2.2 Beban Kerja pada HiBench [19]

Workload Name	Workload Type	Operation Type	Workload Submission Policy	Software Stack
Sort	Micro Benchmark	Algorithm	Pre-Specified Process	Hadoop, Spark
WordCount	Micro Benchmark	Algorithm	Pre-Specified Process	Hadoop, Spark
Terasort	Micro Benchmark	Algorithm	Pre-Specified Process	Hadoop, Spark
Sleep	Micro Benchmark	Algorithm	Pre-Specified Process	Hadoop, Spark
enhanced DFSIO	Micro Benchmark	IO	Parameter Control	Hadoop, Spark
Bayesian Classification	Machine Learning	Algorithm	Parameter Control	Spark
K-means clustering	Machine Learning	Algorithm	Parameter Control	Spark
Logistic Regression	Machine Learning	Algorithm	Parameter Control	Spark
Alternating Least Squares(ALS)	Machine Learning	Algorithm	Parameter Control	Spark
Gradient Boosting Trees (GBT)	Machine Learning	Algorithm	Parameter Control	Spark
Linear Regression	Machine Learning	Algorithm	Parameter Control	Spark
Latent Dirichlet Allocation	Machine Learning	Algorithm	Parameter Control	Spark
Principal Components Analysis (PCA)	Machine Learning	Algorithm	Parameter Control	Spark
Random Forest	Machine Learning	Algorithm	Parameter Control	Spark
Support Vector Machine (SVM)	Machine Learning	Algorithm	Parameter Control	Spark
Support Vector Machine(SVM)	Machine Learning	Algorithm	Parameter Control	Spark
Singular Value Decomposition	Machine Learning	Algorithm	Parameter Control	Spark
Scan, Join, Aggregate	SQL	EO	Pre-Specified Process	Hadoop, Spark
PageRank	Websearch	Algorithm	Parameter Control	Spark
Nutch indexing	Websearch	Algorithm	Parameter Control	Spark, Nutch
NWeight	Graph	Algorithm	Parameter Control	Spark(with GraphX or Pregel)
Identity	Streaming	Algorithm, IO	Parameter Control	Spark Streaming, Flink, Storm and Gearpump
Repartition	Streaming	Algorithm, IO	Parameter Control	Spark Streaming, Flink, Storm and Gearpump
Stateful Wordcount	Streaming	Algorithm, IO	Parameter Control	Spark Streaming, Flink, Storm and Gearpump
Fixwindow	Streaming	Algorithm, IO	Parameter Control	Spark Streaming, Flink, Storm and Gearpump

kan menggunakan program RandomTextWriter yang nantinya akan dibuat melalui proses *Data Generation*.

2.8.2 Data Generation pada Word Count dan Sort

Data generation merupakan tahapan krusial dalam benchmark menggunakan HiBench, khususnya untuk beban kerja *Word Count* dan *Sort*. Tahapan ini bertanggung jawab untuk membentuk data acak yang akan diproses oleh kedua beban kerja tersebut. Tujuannya adalah untuk mensimulasikan skenario nyata dengan input data yang bervolume besar dan beragam.

Pada HiBench, skrip *prepare.sh* seperti pada Algoritma II.1 berperan penting dalam menyiapkan data untuk beban kerja. Skrip ini mengeksekusi program *randomtextwriter* yang terdapat dalam paket Hadoop. *randomtextwriter* menghasilkan sekumpulan data acak yang terdiri dari kata-kata yang diambil dari daftar kata yang telah ditentukan. Jumlah data yang dihasilkan, jumlah *map*, dan jumlah *reduce* dapat dikonfigurasi melalui parameter-parameter yang diberikan kepada skrip *prepare.sh*.

Listing II.1 Skrip yang Digunakan HiBench pada Tahap *Data Generation*

```

1 #!/bin/bash
2
3 current_dir=`dirname "$0"`
4 current_dir=`cd "$current_dir"; pwd`
5 root_dir=${current_dir}/../../../../..
6 workload_config=${root_dir}/conf/workloads/micro/sort.conf
7 . "${root_dir}/bin/functions/load_bench_config.sh"
8
9 enter_bench HadoopPrepareSort ${workload_config} ${current_dir}
10 show_bannar start
11
12 rmr_hdfs ${INPUT_HDFS} || true
13 START_TIME=`timestamp`
14
15 run_hadoop_job ${HADOOP_EXAMPLES_JAR} randomtextwriter \
16     -D mapreduce.randomtextwriter.totalbytes=${DATASIZE} \
17     -D mapreduce.randomtextwriter.bytespermap=$(( ${DATASIZE} / ←
18         ${NUM_MAPS} )) \
19     -D mapreduce.job.maps=${NUM_MAPS} \
20     -D mapreduce.job.reduces=${NUM_RED} \
21     ${INPUT_HDFS}
22 END_TIME=`timestamp`
23 show_bannar finish
24 leave_bench

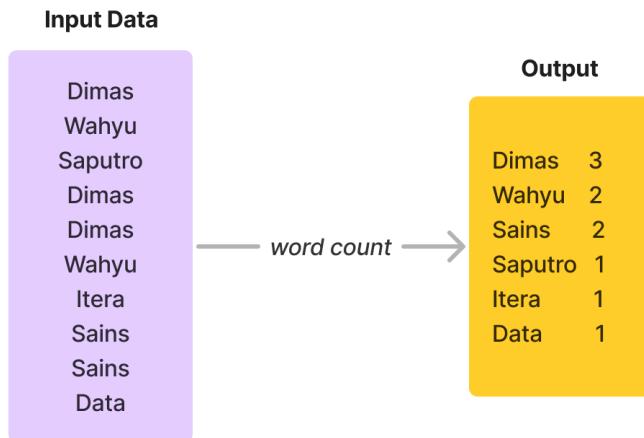
```

2.8.3 Beban Kerja *Word Count*

Word Count adalah algoritma sederhana untuk membaca berkas teks, dan menghitung jumlah kemunculan kata-kata pada file tersebut. Pada algoritma ini, inputnya berupa berkas teks dan outputnya berupa pasangan kata-kata dan jumlah kemunculannya. Beban kerja *word count* akan menghasilkan data keluaran yang lebih kecil dari pada data input. Karena itu, *word count* memiliki sifat *CPU Bound* yang nantinya akan ditandai dengan tingkat penggunaan CPU yang tinggi dan penggunaan I/O ringan. Selain itu, perilakunya diperkirakan akan tetap sama bahkan pada cluster yang lebih besar.

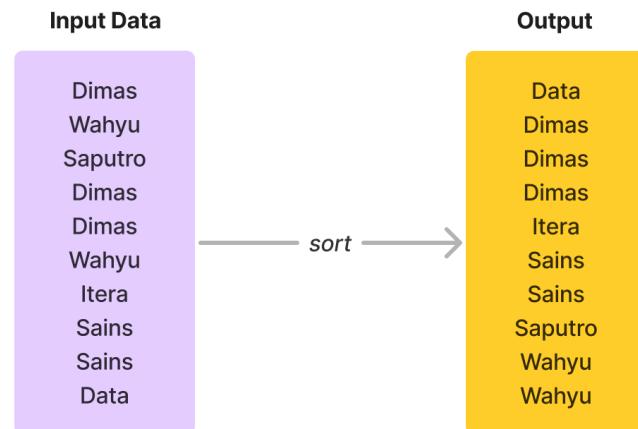
2.8.4 Beban Kerja *Sort*

Sort adalah algoritma yang umum digunakan untuk mengurutkan data berdasarkan kriteria tertentu. Algoritma ini menerima data dalam bentuk acak sebagai input, dan menghasilkan data yang terurut sebagai output. Data input dan output memiliki ukuran yang sama, sehingga beban kerja sort tidak menghasilkan pengurangan



Gambar 2.15 Contoh Input dan Output *Word Count*

data. Kompleksitas algoritma *sort* bervariasi, tetapi umumnya membutuhkan perbandingan dan pertukaran elemen data yang intensif. Oleh karena itu, beban kerja *sort* cenderung bersifat *I/O bound*, dengan pemanfaatan CPU yang rendah dan penggunaan I/O yang tinggi.



Gambar 2.16 Contoh Input dan Output *Sort*

2.9 Data Keluaran HiBench dan Dool

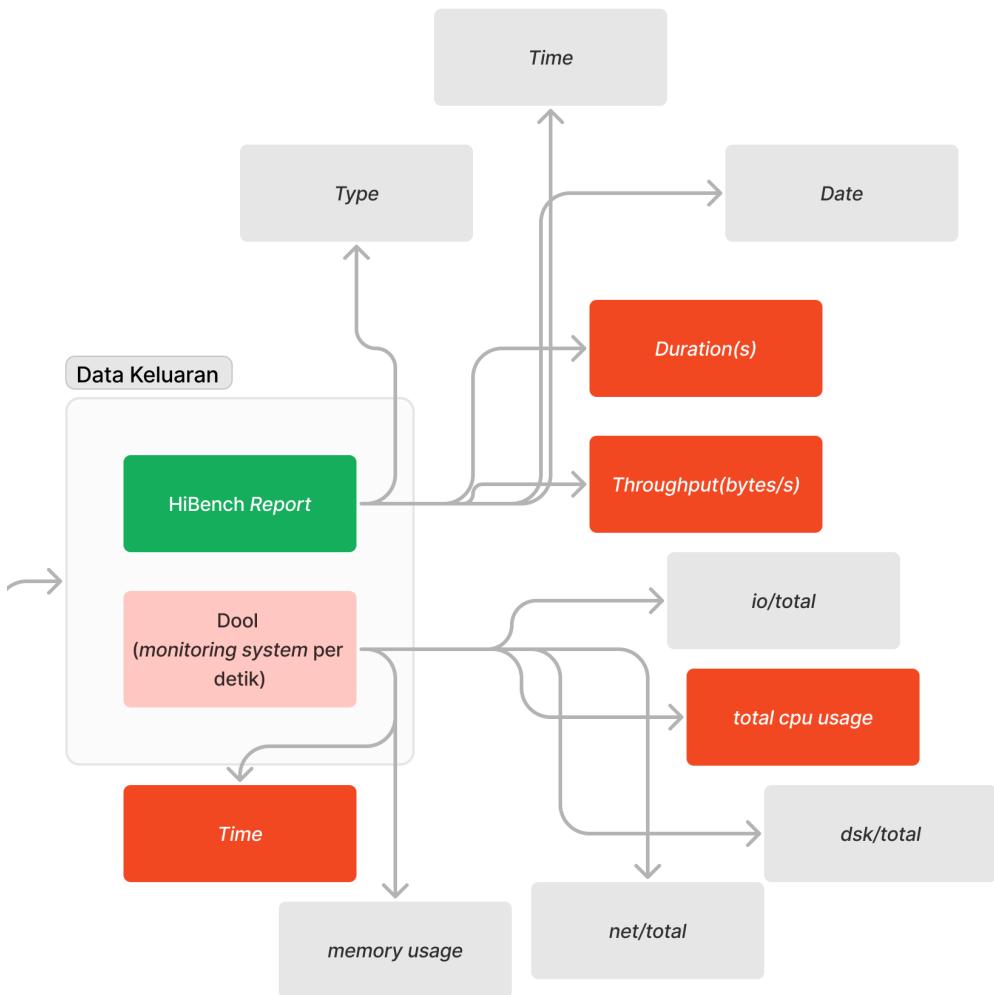
Diagram pada Gambar 2.17 mengilustrasikan data keluaran dari dua alat, yaitu HiBench dan Dool, yang digunakan untuk mengevaluasi kinerja sistem. HiBench berfokus pada pengukuran kinerja keseluruhan dari suatu beban kerja (*benchmark*), sementara Dool memberikan pemantauan sistem yang terperinci secara *real-time*. HiBench menghasilkan laporan yang mencakup dua metrik utama, yaitu

1. **Waktu Eksekusi (Execution Time)**: Menunjukkan total waktu yang dibutuhkan untuk menjalankan beban kerja (*benchmark*) dari awal hingga akhir, diukur dalam detik.
2. **Throughput**: Mengukur jumlah data yang diproses per satuan waktu, biasanya dalam bita (*byte*) per detik. Metrik ini mencerminkan efisiensi sistem dalam menangani beban kerja.

Dool menyediakan pemantauan sistem yang mendetail dan diperbarui setiap detik. Data keluaran Dool meliputi berbagai aspek kinerja sistem, antara lain:

1. **Time**: Timestamp yang menunjukkan waktu pengambilan data.
2. **Date**: Tanggal pengambilan
3. **Type**: Jenis pengukuran yang dilakukan, misalnya CPU, memori, disk, atau jaringan.
4. **io/total**: Total aktivitas input/output (I/O) pada *disk*, diukur dalam jumlah operasi I/O per
5. **total cpu usage**: Persentase penggunaan CPU secara keseluruhan.
6. **dsk/total**: Total aktivitas *disk*, mencakup baca dan tulis, diukur dalam bita per detik.
7. **memory usage**: Jumlah memori yang sedang digunakan oleh sistem.
8. **net/total**: Total aktivitas jaringan, mencakup data yang dikirim dan diterima, diukur dalam bita per detik.

Akhirnya, HiBench memberikan gambaran umum tentang efisiensi sistem dalam menangani beban kerja tertentu, sementara Dool memungkinkan kita untuk memantau berbagai komponen sistem secara *real-time* dan mengidentifikasi potensi *bottleneck* atau masalah kinerja.



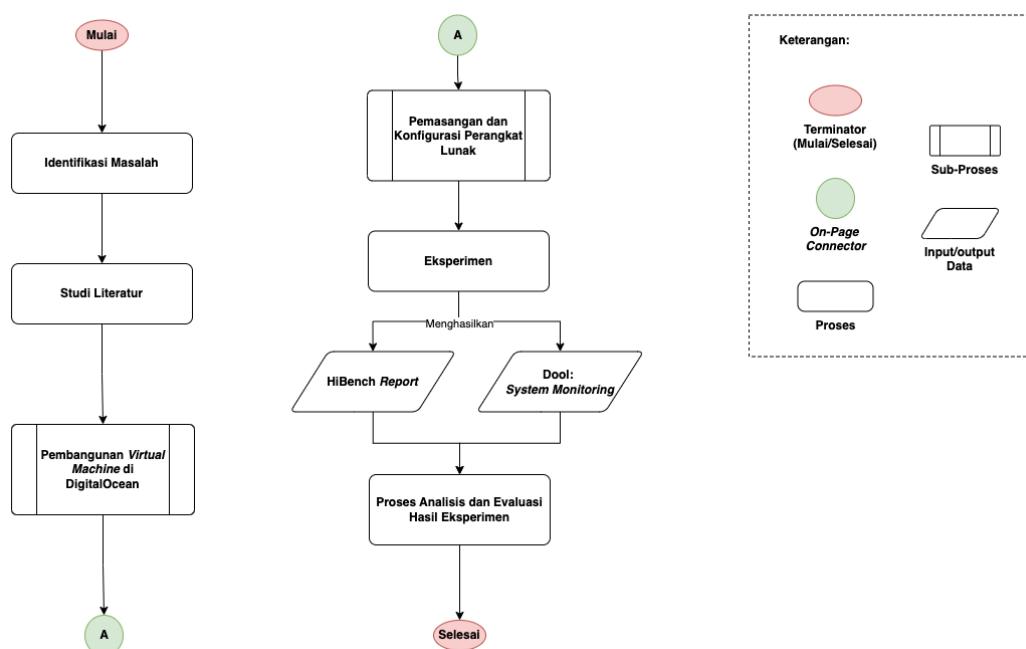
Gambar 2.17 Data Keluaran HiBench dan Dool

BAB III

METODOLOGI PENELITIAN

3.1 Alur Penelitian

Adapun diagram alir penelitian ini ditunjukkan pada Gambar 3.1 terdapat enam tahapan. Langkah awal yang dilakukan pada penelitian ini adalah melakukan identifikasi masalah, yaitu proses mencari, menghimpun, serta menemukan permasalahan yang nantinya akan diselesaikan. Setelah melakukan identifikasi masalah, langkah selanjutnya adalah studi literatur. Studi literatur adalah tahapan untuk mencari solusi dari permasalahan yang sebelumnya sudah kita definisikan. Pencarian solusi ini dapat melalui membaca referensi ilmiah terdahulu, baik melalui jurnal, buku, dokumentasi resmi, tesis, dan lain-lain. Tahapan ini akan memberikan pemahaman mendasar mengenai permasalahan yang sudah didapatkan sebelumnya.



Gambar 3.1 Diagram Alir Penelitian

Kemudian, penelitian ini akan dilanjutkan pada tahap membangun *virtual machine* di DigitalOcean. DigitalOcean adalah perusahaan penyedia layanan awan *Infrastructure as a Service* (IaaS) yang memberikan banyak pilihan kepada pengguna untuk menggunakan berbagai jenis layanan sesuai dengan kebutuhan, salah satunya yaitu *virtual machine*. *Virtual Machine* tersebut dapat dihentikan atau dihapus kapanpun saat tidak lagi diperlukan. Ketika infrastruktur sudah siap digunakan, penelitian dilanjutkan ke tahap pemasangan perangkat lunak, seperti Hadoop, Spark, dan Hi-

Bench. Selanjutnya dilakukan eksperimen pada beban kerja *Micro Benchmarks*. Akhirnya, hasil dari eksperimen akan digunakan untuk proses analisis dan evaluasi.

3.2 Penjabaran Langkah Penelitian

Adapun untuk lebih memperjelas lagi dari setiap langkah yang ada pada Gambar 3.1, dijabarkan secara rinci tahapan-tahapan yang dilakukan pada penelitian ini.

3.2.1 Identifikasi Masalah dan Studi Literatur

Langkah awal penelitian ini adalah identifikasi masalah dan studi literatur. Identifikasi masalah dapat dipahami sebagai tahapan mendefinisikan masalah sehingga masalah tersebut dapat terukur dan jelas untuk dijadikan landasan dalam latar belakang penelitian. Setelah masalah berhasil diidentifikasi, langkah selanjutnya adalah studi literatur yang mana dalam proses ini dilakukan pengumpulan berbagai macam informasi, referensi, dan konsep dasar yang menjadi landasan dasar dari penelitian. Langkah ini dapat dilakukan melalui membaca artikel ilmiah pendukung, buku-buku yang ditulis oleh para ahli, dan jika berkaitan dengan pemrograman dapat melihat dari dokumentasi resmi. Pada tahap ini juga dilakukan analisis terhadap penelitian terdahulu dan dibandingkan dengan identifikasi masalah yang didapatkan untuk membuka celah penelitian baru sehingga penelitian ini dapat bermanfaat.

3.2.2 Membangun *Virtual Machine* di DigitalOcean

Konfigurasi perangkat keras merupakan aspek penting dalam mengevaluasi kinerja aplikasi *big data* berbasis Hadoop dan Spark. DigitalOcean, sebagai penyedia layanan infrastruktur sebagai layanan (IaaS), memberikan pengguna kebebasan penuh untuk membuat, mengkonfigurasi, dan mengelola berbagai infrastruktur yang telah disediakan. Dalam konteks penelitian ini, diperlukan penggunaan mesin virtual, yang dalam DigitalOcean dikenal sebagai "Droplets," yang memungkinkan untuk menyesuaikan berbagai aspek seperti sistem operasi, kapasitas penyimpanan, jumlah prosesor, dan parameter lainnya sesuai dengan kebutuhan spesifik penelitian. Penelitian ini mengadopsi mode *pseudo-distributed* yang memungkinkan penggunaan hanya satu *virtual machine* dalam konfigurasi *single node*. Walaupun hanya menggunakan satu *virtual machine*, mode *pseudo-distributed* memungkinkan setiap proses dalam klaster beroperasi secara independen, menciptakan lingkungan di mana semua proses berjalan mandiri satu sama lain. Hal ini memungkinkan untuk lebih berfokus pada pengumpulan data dan analisis, tanpa perlu melakukan konfigurasi yang rumit terkait dengan pengaturan klaster. Spesifikasi perangkat keras yang digunakan untuk *virtual machine* dalam mode *pseudo-distributed* sesuai pada Tabel

Tabel 3.1 Konfigurasi Perangkat Keras

Nama Parameter	Nilai Parameter
Lokasi Pusat Data	Singapore - Datacenter 1 - SGP1
Sistem Operasi	Ubuntu 20.04 (LTS) x64
Jenis Droplet	Basic
Prosesor	Premium AMD - 4 Core
Memori	8 GB
Penyimpanan	160 GB NVMe SSD

3.1. Penjelasan lengkap tentang pembuatan *virtual machine* (VM) pada *platform* DigitalOcean dan cara mengakses VM tersebut disajikan pada Lampiran A.

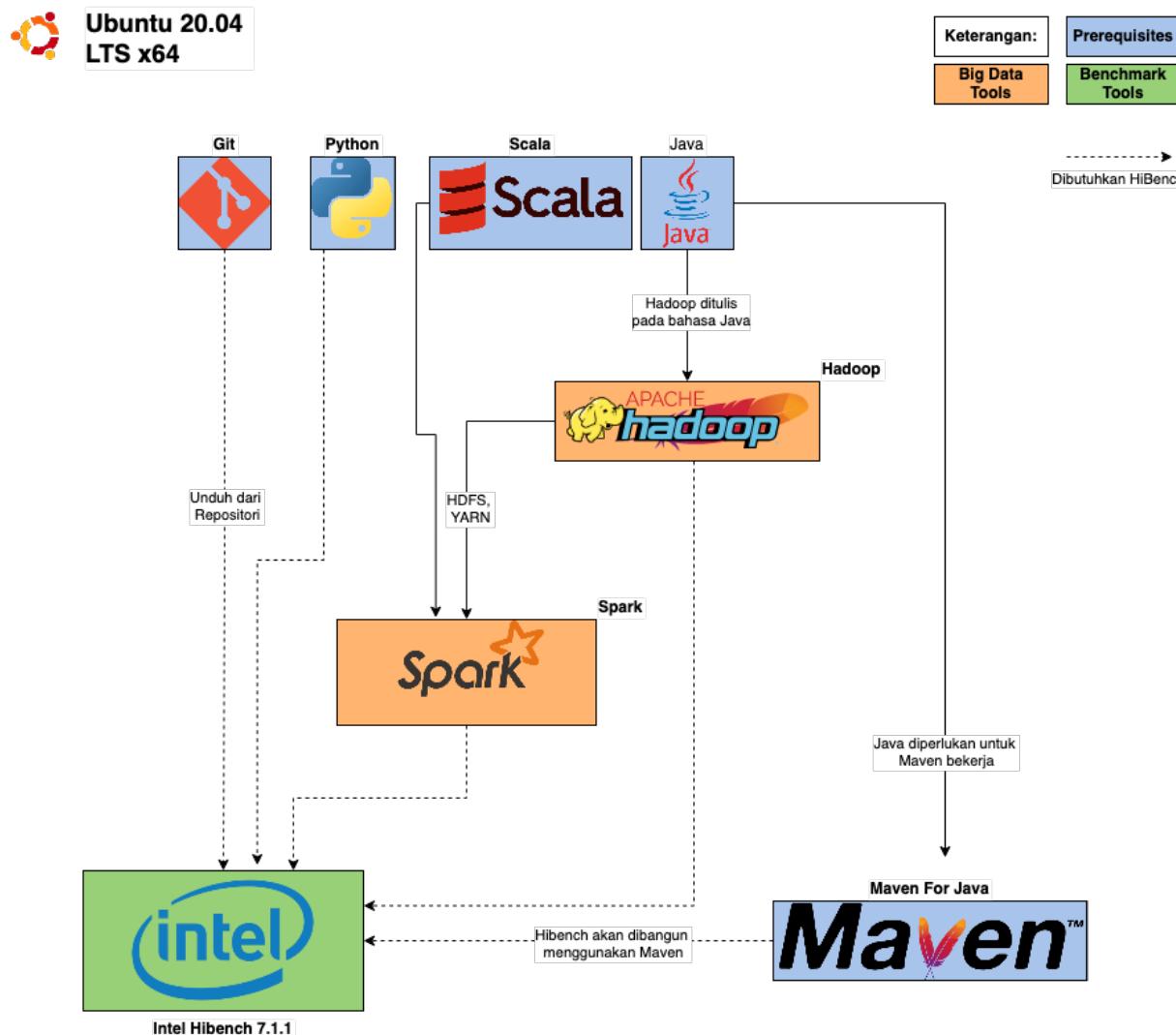
3.2.3 Pemasangan dan Konfigurasi Perangkat Lunak

Pemasangan dan konfigurasi perangkat lunak merupakan hal yang krusial dalam penelitian ini. Perangkat lunak yang diperlukan ditunjukkan pada Tabel 3.2.

Tabel 3.2 Perangkat Lunak yang Dibutuhkan

Perangkat Lunak	Deskripsi
Ubuntu 20.04 LTS x64	Sistem operasi Linux berbasis Ubuntu
Git	Sistem kontrol versi untuk mengelola perubahan dalam kode sumber perangkat lunak
Maven	Perangkat lunak manajemen proyek Java
Java 8	
Python 3.7	Bahasa pemrograman dasar
Scala 2.11.8	
Hadoop 2.4	Perangkat lunak pengolahan data terdistribusi untuk penyimpanan dan manajemen data besar
Spark 2.1.3	Kerangka kerja pemrosesan data terdistribusi yang berjalan di atas Hadoop
Hibench	Alat yang digunakan untuk mengukur kinerja Hadoop dan Spark
Dool	Alat untuk melihat penggunaan <i>resource</i> sistem

Alur kerja instalasi perangkat lunak dalam penelitian ini dapat dilihat pada Gambar 3.2. Pada gambar, terdapat tiga bagian utama, yaitu *prerequisites* (perangkat lunak prasyarat) ditandai dengan warna biru, alat penyimpanan dan pemrosesan *Big Data* ditandai dengan warna oranye, dan alat untuk mengukur kinerja *big data* ditandai dengan warna hijau. Semua perangkat lunak dijalankan pada sistem operasi Ubuntu 20.04 LTS x64.



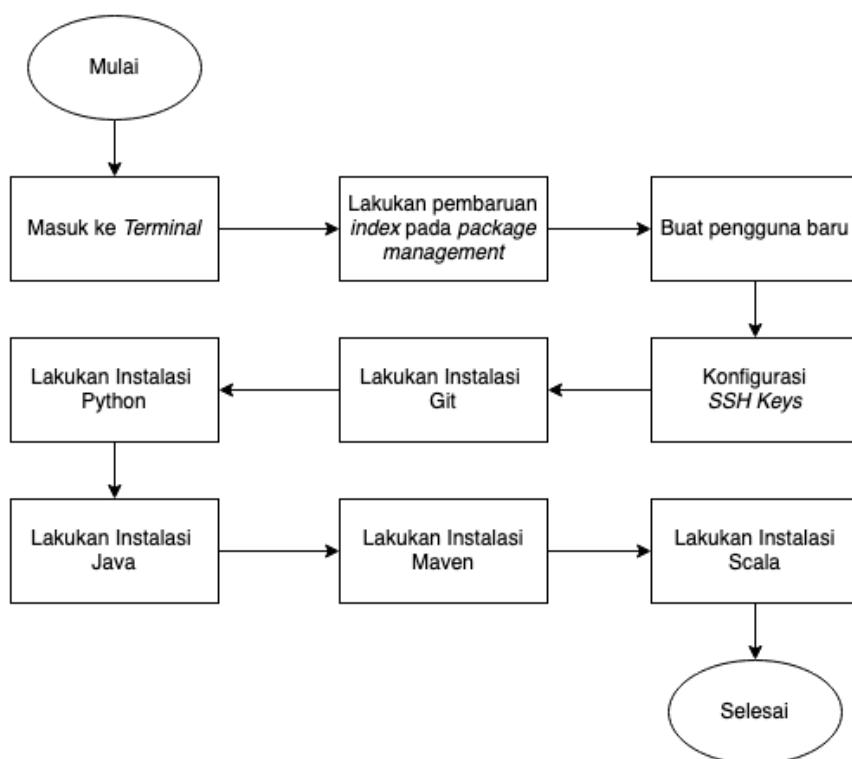
Gambar 3.2 Alur Instalasi Perangkat Lunak

Instalasi Perangkat Lunak Prasyarat

Ada beberapa perangkat lunak yang perlu diimplementasikan sebelum memasang Hadoop, Spark, dan HiBench, yaitu:

1. Ubuntu 20.04 LTS x64
2. Git
3. Java 8 dan Maven
4. Python 3.7
5. Scala 2.11.8

Pemasangan dan konfigurasi perangkat lunak pada tahapan ini tidak membutuhkan urutan. Akan tetapi, pada penelitian ini dibuatkan alur untuk pemasangan dan konfigurasi perangkat lunak prasyarat seperti pada Gambar 3.3. Penjelasan lengkap mengenai tata cara instalasi dan konfigurasi perangkat lunak prasyarat ini disajikan pada Lampiran B.

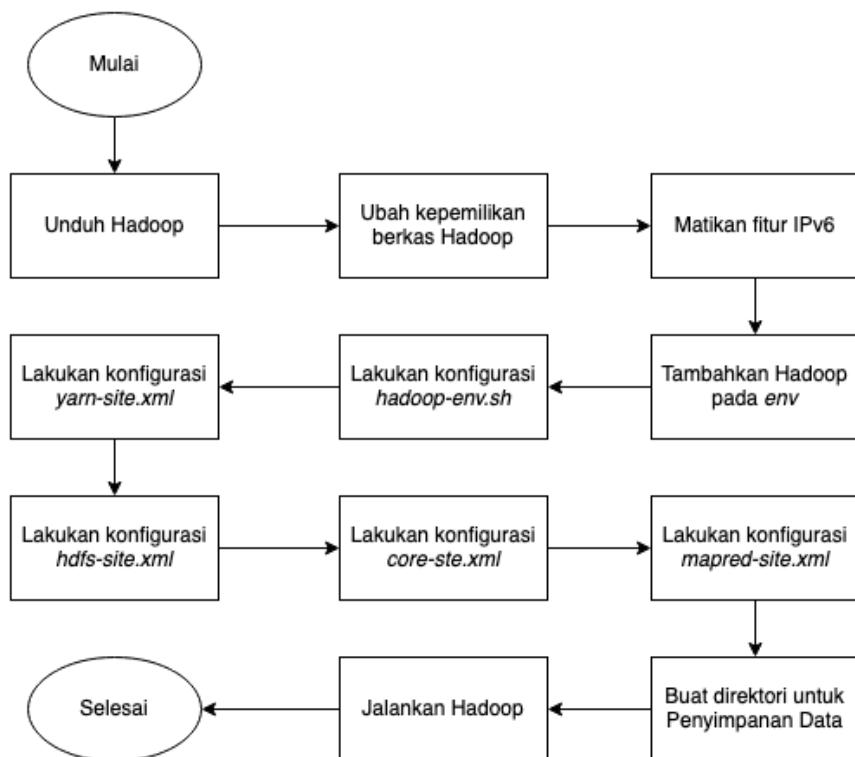


Gambar 3.3 Alur Instalasi Perangkat Lunak Prasyarat

Instalasi dan Konfigurasi Hadoop

Hadoop adalah perangkat lunak *open source* yang efektif dalam menyimpan dan memproses data dalam skala besar. Daripada menggunakan satu komputer besar untuk menyimpan dan memproses data, Hadoop memungkinkan pengklasteran be-

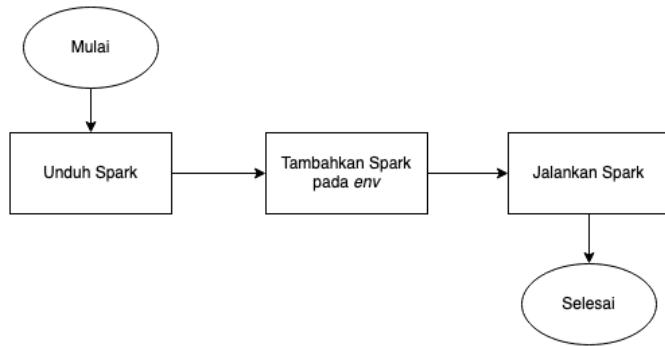
berapa komputer untuk menganalisis set data besar secara paralel dengan lebih cepat. Ada beberapa perangkat lunak prasyarat yang perlu dipasang sebelum menggunakan Hadoop. Setelah perangkat lunak prasyarat berhasil dipasang, Hadoop juga dapat dipasang mengikuti panduan lengkap pada Lampiran C. Secara umum, alur yang harus dilakukan meliputi pengunduhan berkas Hadoop. Selanjutnya akan dilakukan pengubahan kepemilikan berkas ke *user hdfsuser*. Karena Hadoop tidak mendukung IPv6, maka fitur ini perlu dimatikan juga. Alur pemasangan dan konfigurasi Hadoop lebih jelas sesuai dengan Gambar 3.4.



Gambar 3.4 Alur Instalasi dan Konfigurasi Hadoop

Instalasi dan Konfigurasi Spark

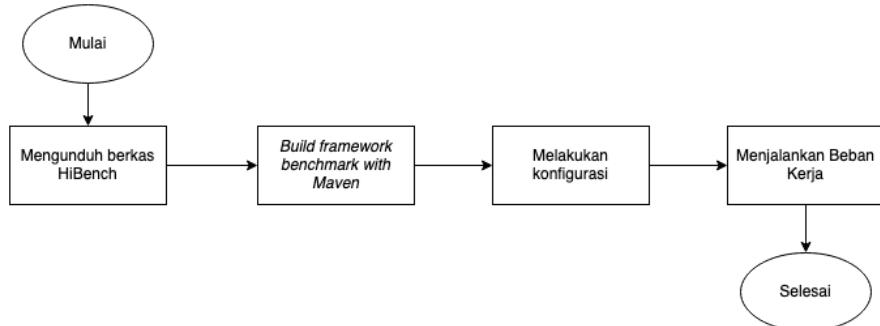
Apache Spark adalah sebuah kerangka kerja pengolahan data terdistribusi yang sangat cepat dan efisien. Spark dan Hadoop memiliki hubungan yang erat. Spark dapat berjalan di atas *Hadoop Distributed File System* (HDFS) dan dapat menggunakan Hadoop YARN sebagai manajer sumber daya. Oleh karena itu, instalasi Spark membutuhkan Hadoop sudah terpasang lebih dahulu. Alur pemasangan dan konfigurasi spark terlihat seperti pada Gambar 3.5. Apabila Hadoop sudah berhasil terpasang, langkah selanjutnya adalah memasang Spark seperti pada Lampiran D.



Gambar 3.5 Alur Instalasi dan Konfigurasi Spark

Instalasi dan Konfigurasi HiBench

Sebelum melakukan eksperimen, diperlukan suatu perangkat lunak pengukuran kinerja sistem *Big Data*, yaitu HiBench. HiBench tidak dapat digunakan secara langsung ketika sudah berhasil diunduh, melainkan harus dilakukan pembangunan beberapa modul yang dibutuhkan dengan Maven dan konfigurasi beberapa parameter.



Gambar 3.6 Alur Instalasi dan Konfigurasi HiBench

Secara umum, alur instalasi dan konfigurasi HiBench sesuai dengan Gambar 3.6. Berkas HiBench diunduh dari repositori, dilanjutkan dengan pembangunan beberapa modul yang nantinya dibutuhkan. Selanjutnya, dilakukan konfigurasi beberapa berkas seperti *hibench.conf*, *hadoop.conf*, dan *spark.conf*. Jika telah dilakukan konfigurasi, dapat dilanjutkan dengan menjalankan beban kerja atau eksperimen. Lebih lanjut, pemasangan dan konfigurasi HiBench dijelaskan pada Lampiran E.

3.2.4 Eksperimen

Setelah instalasi dan konfigurasi perangkat keras dan perangkat lunak berhasil diselesaikan, tahap selanjutnya adalah eksperimen. Tahap ini melibatkan serangkaian pengujian yang terkontrol untuk mengevaluasi kinerja *platform big data* Hadoop

dan Spark dalam menjalankan beban kerja tertentu dengan berbagai ukuran data. Tujuan utama eksperimen ini adalah untuk menjawab pertanyaan penelitian yang telah didefinisikan sebelumnya dan memperoleh pemahaman yang komprehensif tentang karakteristik kinerja masing-masing *platform*.

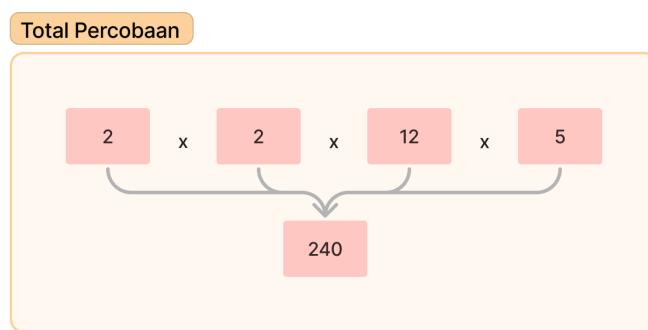
Penelitian ini difokuskan pada pengujian dua beban kerja yang umum dalam pemrosesan big data, yaitu *word count* dan *sort*. Beban kerja ini akan dieksekusi pada dua *platform big data* yang populer, yaitu Hadoop dan Spark. Setiap kombinasi *platform* dan beban kerja akan diuji dengan 12 ukuran data yang berbeda, mulai dari 100 KB hingga 15 GB. Detail ukuran data yang digunakan ditunjukkan pada Tabel 3.3. Untuk memastikan reliabilitas dan konsistensi hasil, setiap kombinasi *platform*, beban kerja, dan ukuran data akan diulang sebanyak 5 kali.

Proses eksperimen menghasilkan dua jenis berkas data:

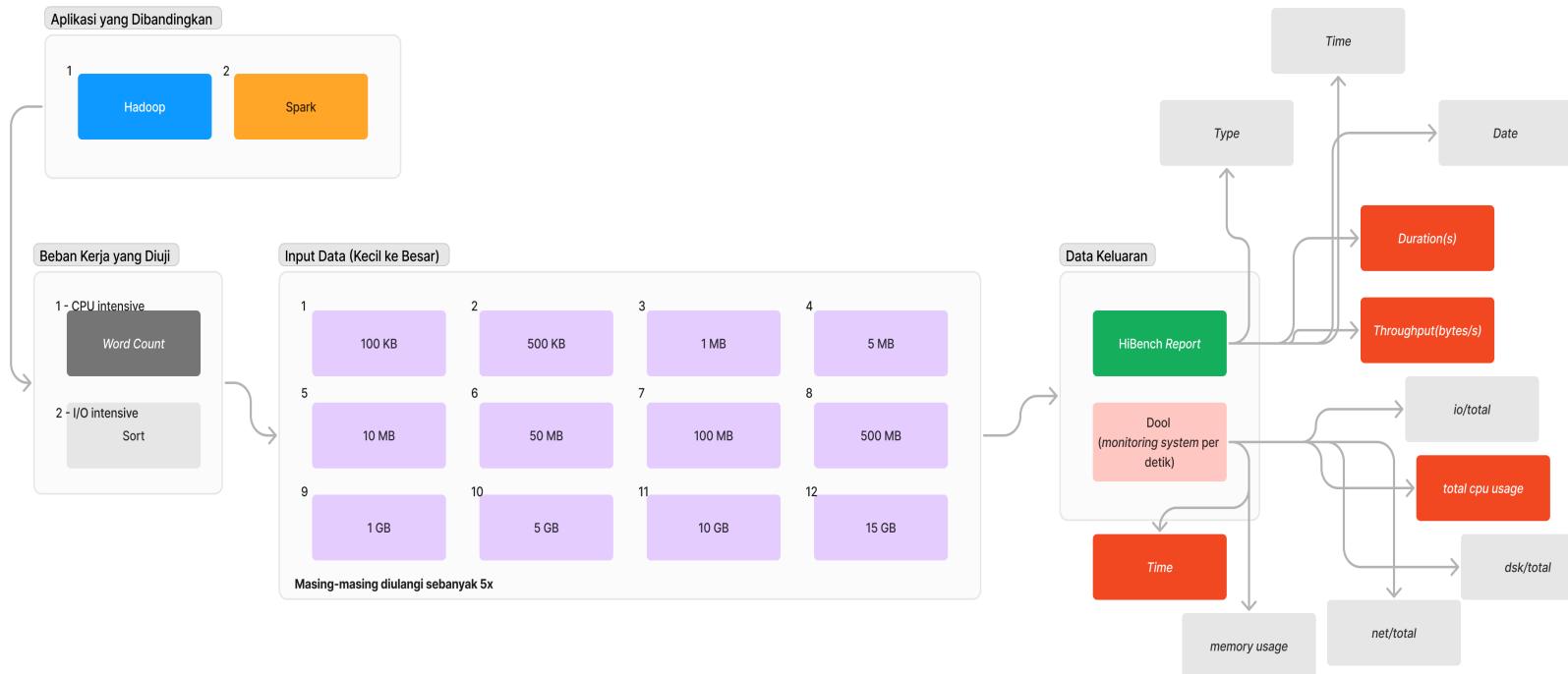
1. *HiBench Report*: Berisi informasi tentang kinerja beban kerja, termasuk waktu eksekusi, dan *throughput*.
2. *Dool System Monitoring*: Berisi informasi detail tentang aktivitas sistem selama eksekusi beban kerja, seperti penggunaan CPU, memori, I/O *disk*, dan jaringan.

Secara keseluruhan, desain eksperimen ini menghasilkan 240 percobaan individu, seperti yang diilustrasikan pada Gambar 3.7. Setiap percobaan mewakili kombinasi unik dari:

1. *Platform big data* (Hadoop atau Spark)
2. Beban kerja (*word count* atau *sort**)
3. Ukuran data (12 variasi)
4. Pengulangan (5 kali)



Gambar 3.7 Total Percobaan

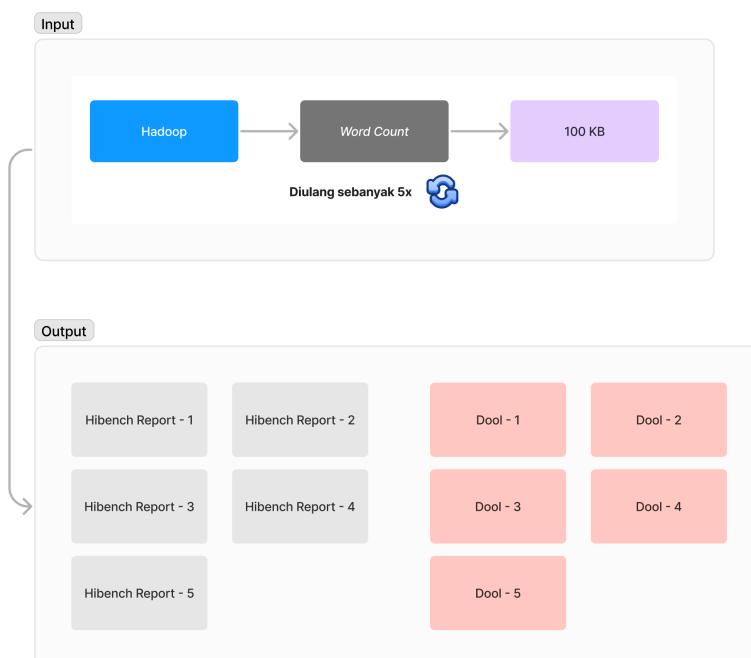


Gambar 3.8 End-to-end Penelitian

Tabel 3.3 Variasi Input Data

No	Label Input Data	Ukuran Input Data (bita)
1	100 KB	$100000 (1 * 10^5)$
2	500 KB	$500000 (5 * 10^5)$
3	1 MB	$1 * 10^6$
4	5 MB	$5 * 10^6$
5	10 MB	$1 * 10^7$
6	50 MB	$5 * 10^7$
7	100 MB	$1 * 10^8$
8	500 MB	$5 * 10^8$
9	1 GB	$1 * 10^9$
10	5 GB	$5 * 10^9$
11	10 GB	$1 * 10^{10}$
12	15 GB	$1.5 * 10^{10}$

Sebagai contoh, untuk *platform* Hadoop dengan beban kerja *word count* dan ukuran data 100 KB, akan menghasilkan 5 HiBench *Report* dan 5 berkas Dool *System Monitoring*, sesuai dengan jumlah pengulangan. Ilustrasi ini dapat dilihat pada Gambar 3.9.



Gambar 3.9 Contoh Percobaan

Karena jumlah percobaan yang banyak, otomatisasi menjadi penting untuk memastikan efisiensi dan akurasi. Skrip khusus dikembangkan untuk mengotomatiskan seluruh proses eksperimen, termasuk konfigurasi HiBench, persiapan data, eksekusi beban kerja, dan pengumpulan data. Detail skrip otomatisasi dapat ditemukan

pada Lampiran F.

Algoritma otomatisasi eksperimen dimulai dengan mengubah direktori kerja ke direktori HiBench. Selanjutnya, algoritma melakukan iterasi untuk setiap beban kerja yang ditentukan. Di dalam setiap iterasi beban kerja, dilakukan iterasi lagi untuk setiap ukuran data. Pada setiap kombinasi beban kerja dan ukuran data, konfigurasi HiBench diubah sesuai dengan ukuran data yang dipilih.

Skrip persiapan data Hadoop dan Spark dijalankan berulang kali hingga proses persiapan berhasil. Setelah data siap, perulangan dilakukan sebanyak jumlah pengulangan yang ditentukan. Dalam setiap perulangan, perangkat lunak "dool" diaktifkan untuk memonitor aktivitas sistem, *benchmark* Hadoop atau Spark dijalankan, dan monitoring sistem dihentikan.

Setelah semua perulangan selesai, algoritma menunggu selama 15 detik sebelum melanjutkan ke ukuran data berikutnya. Proses ini berlanjut hingga semua kombinasi beban kerja dan ukuran data selesai diproses.

3.2.5 Analisis dan Evaluasi Hasil Eksperimen

Setelah menyelesaikan 240 percobaan yang dijelaskan di bagian eksperimen, langkah selanjutnya adalah menganalisis dan mengevaluasi hasil yang diperoleh. Analisis ini bertujuan untuk menjawab pertanyaan penelitian dan memahami bagaimana kinerja Hadoop dan Spark dalam menjalankan beban kerja *word count* dan *sort* dengan berbagai ukuran data. Berikut adalah beberapa aspek yang akan dikaji:

1. Kinerja

- (a) **Persebaran Waktu Eksekusi pada Hadoop dan Spark.** Bagian ini akan menganalisis sebaran waktu eksekusi untuk setiap beban kerja (*word count* dan *sort*) pada kedua aplikasi (Hadoop dan Spark) dengan berbagai ukuran data. Analisis ini akan membantu memahami variabilitas kinerja dan konsistensi hasil pada setiap kombinasi aplikasi, beban kerja, dan ukuran data.
- (b) **Persebaran Throughput pada Hadoop dan Spark.** Mirip dengan analisis waktu eksekusi, persebaran *throughput* juga akan dianalisis untuk setiap kombinasi aplikasi, beban kerja, dan ukuran data. Throughput, yang menunjukkan jumlah data yang diproses per satuan waktu, merupakan metrik penting dalam evaluasi kinerja sistem big data. Visualisasi distribusi throughput akan membantu dalam memahami efisiensi pemrosesan data oleh Hadoop dan Spark.
- (c) **Rata-rata Waktu Eksekusi pada Hadoop dan Spark.** Selain persebaran, rata-rata waktu eksekusi untuk setiap kombinasi akan dihitung

dan dibandingkan. Ini memberikan gambaran umum tentang kinerja relatif setiap aplikasi dalam menyelesaikan beban kerja tertentu dengan ukuran data tertentu. Perbedaan rata-rata waktu eksekusi antara Hadoop dan Spark, serta tren perubahannya seiring dengan peningkatan ukuran data, akan diidentifikasi dan dibahas.

- (d) **Rata-rata Throughput pada Hadoop dan Spark.** Serupa dengan rata-rata waktu eksekusi, rata-rata *throughput* juga akan dihitung dan dibandingkan untuk setiap kombinasi. Analisis ini membantu memahami bagaimana efisiensi pemrosesan data berubah dengan berbagai beban kerja dan ukuran data, serta memberikan wawasan tentang skalabilitas setiap platform.
- (e) **Rate of Change.** *Rate of Change* akan dihitung untuk metrik-metrik kinerja seperti waktu eksekusi dan throughput. Ini akan menunjukkan seberapa besar perubahan kinerja seiring dengan peningkatan ukuran data.

2. Penggunaan Sumber Daya

- (a) **Penggunaan CPU.** Bagian ini akan menganalisis penggunaan CPU oleh Hadoop dan Spark selama menjalankan berbagai beban kerja. Informasi ini dapat diperoleh dari berkas monitoring sistem yang dihasilkan oleh Dool. Analisis penggunaan CPU membantu memahami bagaimana setiap platform memanfaatkan sumber daya komputasi dan mengidentifikasi potensi optimasi.
- (b) **Utilisasi Sistem.** Selain penggunaan CPU, metrik-metrik lain seperti penggunaan memori, dan I/O penyimpanan. Hal ini memberikan gambaran yang lebih komprehensif tentang bagaimana setiap *platform* memanfaatkan sumber daya sistem dan potensi *bottleneck* yang mungkin terjadi selama pemrosesan data besar.

BAB IV

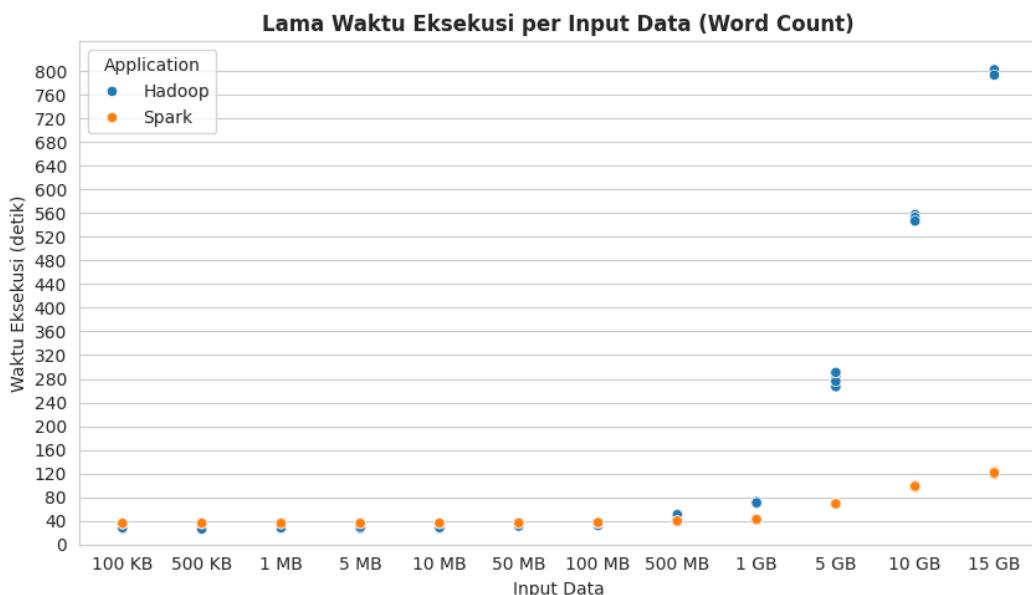
HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Penelitian ini membandingkan kinerja Hadoop dan Spark pada *platform cloud* DigitalOcean menggunakan alat pengujian data besar yang bernama HiBench pada lingkup *Micro Benchmarks*, yaitu *Word Count* dan *Sort* dengan data masukan berupa teks. Setiap pengujian, nilai execution time (waktu eksekusi) memiliki satuan detik dan nilai *throughput* memiliki satuan megabita per detik. Ukuran data input akan diubah secara bertahap untuk merepresentasikan berbagai ukuran data pada dunia nyata.

4.1.1 Persebaran Waktu Eksekusi pada Hadoop dan Spark

Waktu eksekusi adalah waktu yang diperlukan dalam memproses data. Nilai parameter ini didapatkan dengan cara mencari selisih antara waktu awal dan waktu akhir saat Apache Hadoop dan Apache Spark dijalankan atau dihentikan untuk memproses input data dengan beban kerja masing-masing. Satuan pengukuran untuk parameter waktu eksekusi adalah detik atau *seconds*. Setiap beban kerja dijalankan sebanyak 5x untuk mendapatkan hasil yang lebih terukur.



Gambar 4.1 Persebaran Waktu Eksekusi *Word Count* (Hadoop, Spark)

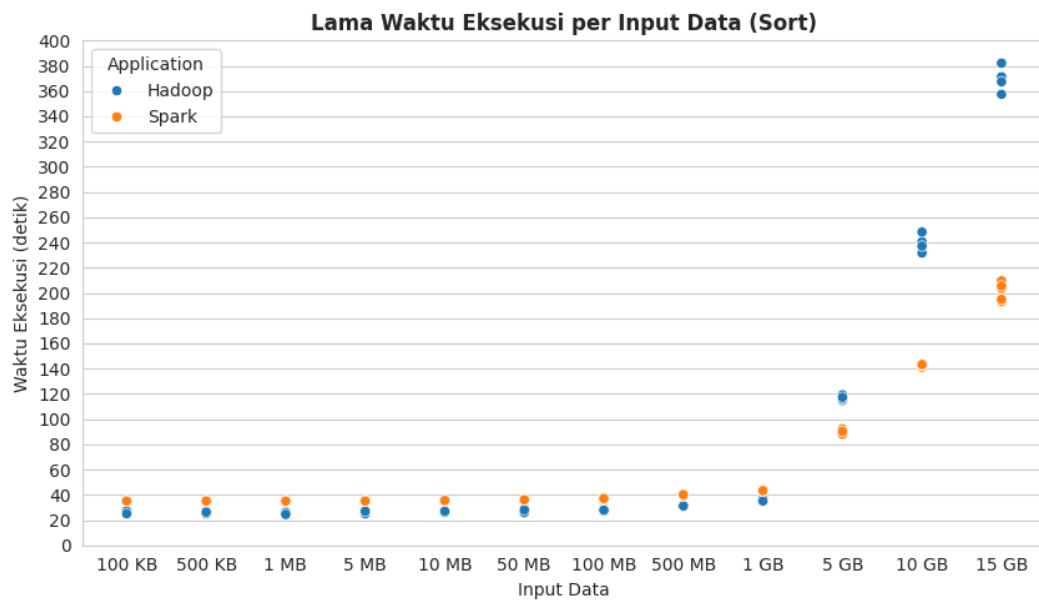
Gambar 4.2 dan 4.1 menyajikan scatter plot yang membandingkan performa Hadoop dan Spark dalam dua tugas pemrosesan data yang berbeda: sorting dan word

count. Sumbu x pada kedua gambar menunjukkan variasi ukuran input data, mulai dari 100 KB hingga 15 GB, sedangkan sumbu y menunjukkan waktu eksekusi dalam detik.

Pada Gambar 4.2, terlihat bahwa Spark secara konsisten mengungguli Hadoop dalam tugas sorting. Titik data Spark selalu berada di bawah titik data Hadoop, menunjukkan waktu eksekusi yang lebih singkat untuk semua ukuran data. Perbedaan performa ini semakin signifikan seiring bertambahnya ukuran data. Sebagai contoh, pada ukuran data 15 GB, Spark menyelesaikan tugas sorting dalam waktu kurang dari 240 detik, sedangkan Hadoop membutuhkan waktu lebih dari 360 detik. Selain itu, titik data Spark lebih rapat, mengindikasikan variabilitas performa yang lebih rendah dan menghasilkan waktu eksekusi yang lebih konsisten.

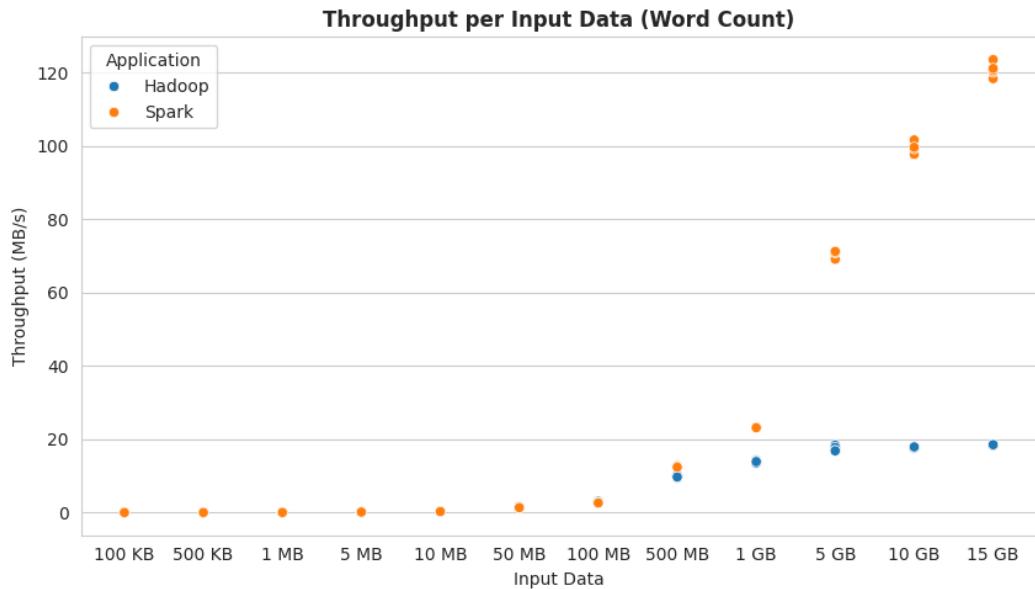
Pada Gambar 4.1, Spark masih menunjukkan performa yang lebih baik daripada Hadoop pada sebagian besar ukuran data. Namun, perbedaan performanya tidak sebesar pada tugas sorting. Pada ukuran data kecil (di bawah 1 GB), kedua framework menunjukkan waktu eksekusi yang relatif sama. Pada ukuran data yang lebih besar, Spark tetap lebih cepat, kecuali pada 15 GB di mana Hadoop menunjukkan waktu eksekusi yang sedikit lebih cepat, namun dengan variabilitas yang lebih tinggi.

Secara keseluruhan, hasil ini menunjukkan bahwa Spark umumnya lebih unggul dan konsisten daripada Hadoop dalam menangani tugas pemrosesan data, terutama untuk sorting. Namun, perbandingan performa dapat bervariasi tergantung pada jenis tugas dan ukuran data.



Gambar 4.2 Persebaran Waktu Eksekusi Sort (Hadoop, Spark)

4.1.2 Persebaran *Throughput* pada Hadoop dan Spark



Gambar 4.3 *Throughput Word Count* (Hadoop, Spark)

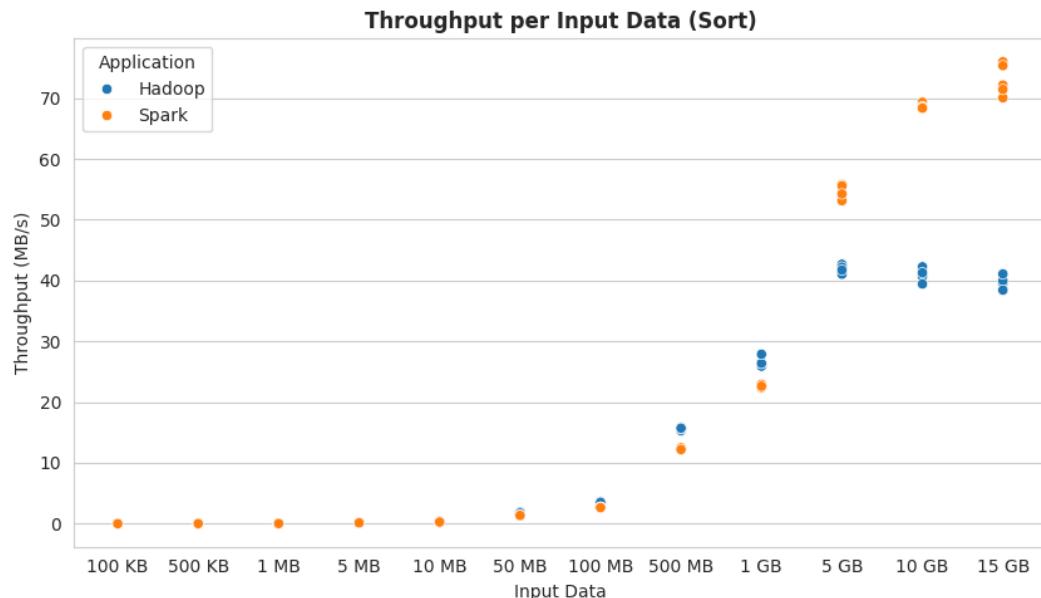
Kedua gambar di atas menyajikan scatter plot yang membandingkan throughput Hadoop dan Spark dalam dua tugas pemrosesan data: sorting (Gambar 4.4) dan word count (Gambar 4.3). Sumbu x pada kedua gambar menunjukkan variasi ukuran input data, sedangkan sumbu y menunjukkan throughput dalam MB/s.

Pada tugas sorting (Gambar 4.4), Spark menunjukkan peningkatan throughput yang signifikan seiring dengan bertambahnya ukuran data. Pada ukuran data terbesar (15 GB), Spark mencapai throughput sekitar 40 MB/s. Sebaliknya, Hadoop menunjukkan peningkatan throughput yang lebih lambat dan hanya mencapai sekitar 35 MB/s pada ukuran data yang sama. Hal ini menunjukkan bahwa Spark mampu memanfaatkan sumber daya secara lebih efisien untuk memproses data dalam jumlah besar.

Pada tugas word count (Gambar 4.3), Spark mencapai throughput yang lebih tinggi daripada Hadoop untuk sebagian besar ukuran data. Perbedaan throughput paling mencolok terlihat pada ukuran data menengah (100 MB - 1 GB), di mana Spark mencapai throughput lebih dari 15 MB/s, sedangkan Hadoop hanya mencapai sekitar 5 MB/s. Meskipun Hadoop menunjukkan peningkatan throughput yang signifikan pada ukuran data terbesar (15 GB), mendekati throughput Spark, Spark tetap memiliki keunggulan dalam efisiensi pemrosesan data, terutama untuk ukuran data yang lebih kecil.

Hasil ini menunjukkan bahwa Spark secara umum lebih efisien dalam memanfaatkan sumber daya untuk memproses data dalam jumlah besar dibandingkan dengan

Hadoop, baik untuk tugas sorting maupun word count.



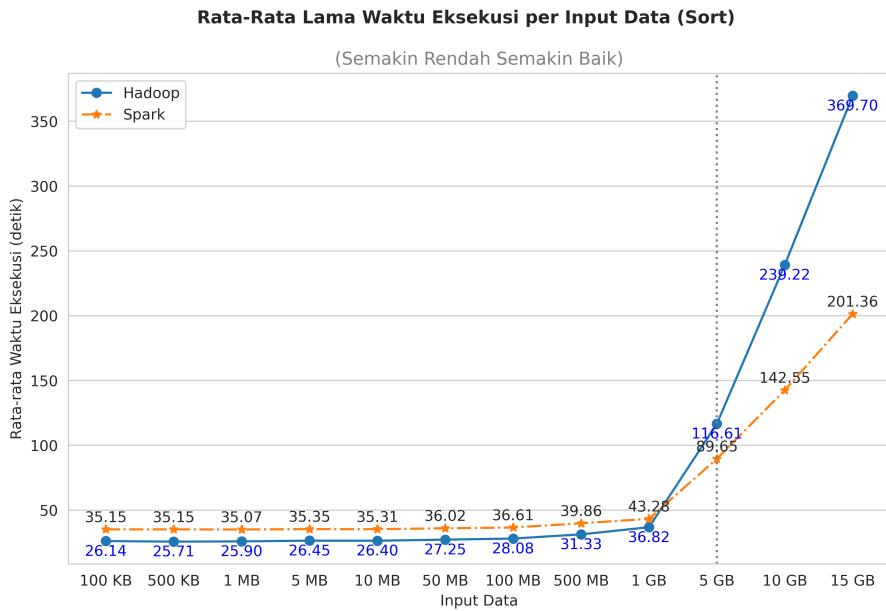
Gambar 4.4 Throughput Sort (Hadoop, Spark)

4.1.3 Rata-rata Waktu Eksekusi pada Hadoop dan Spark

Gambar 4.5 dan 4.6 menyajikan line plot yang menggambarkan rata-rata waktu eksekusi Hadoop dan Spark untuk tugas sorting dan word count dengan berbagai ukuran data. Sumbu x pada kedua gambar menunjukkan ukuran input data, sedangkan sumbu y menunjukkan rata-rata waktu eksekusi dalam detik. Garis vertikal pada kedua gambar menunjukkan titik di mana Spark mulai menunjukkan performa yang lebih cepat dibandingkan Hadoop.

Pada Gambar 4.5, terlihat bahwa Spark secara konsisten lebih cepat daripada Hadoop untuk semua ukuran data pada tugas sorting. Perbedaan performa semakin signifikan seiring dengan bertambahnya ukuran data. Titik di mana Spark mulai mengungguli Hadoop terjadi pada ukuran data 1 GB. Pada titik ini, waktu eksekusi Spark mulai menurun secara signifikan dibandingkan dengan Hadoop yang meningkat secara eksponensial.

Pada Gambar 4.6, Spark juga menunjukkan performa yang lebih baik daripada Hadoop pada sebagian besar ukuran data pada tugas word count. Perbedaan performa mulai terlihat pada ukuran data 100 MB, dan Spark terus unggul hingga ukuran data terbesar (15 GB).

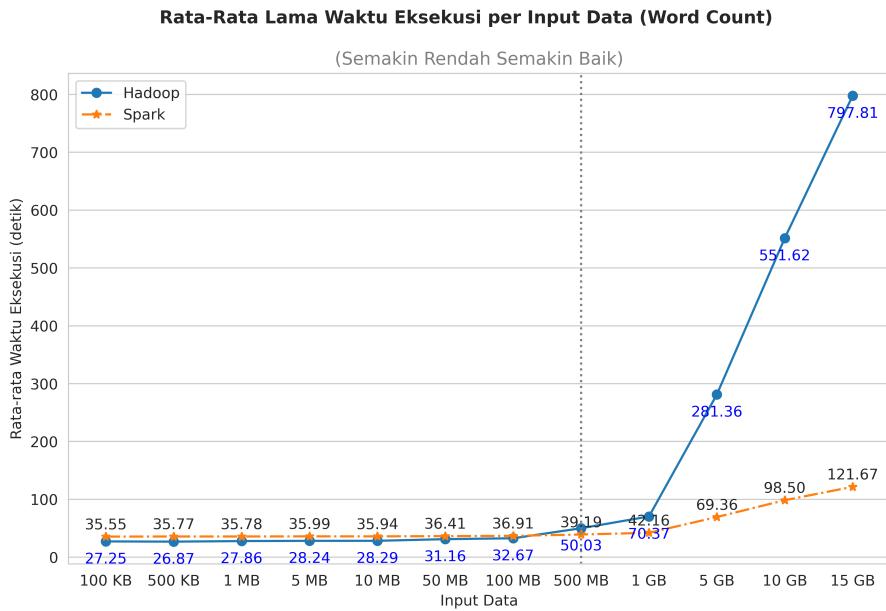


Gambar 4.5 Rata-rata Waktu Eksekusi (*Sort*)

4.1.4 Rata-rata Throughput pada Hadoop dan Spark

Gambar 4.7 dan 4.8 menyajikan line plot yang menggambarkan rata-rata throughput Hadoop dan Spark untuk tugas sorting dan word count dengan berbagai ukuran data. Sumbu x pada kedua gambar menunjukkan ukuran input data, sedangkan sumbu y menunjukkan rata-rata throughput dalam MB/s. Garis vertikal pada kedua gambar menunjukkan titik di mana Spark mulai menunjukkan throughput yang lebih tinggi dibandingkan Hadoop.

Pada tugas sorting, Spark menunjukkan peningkatan throughput yang signifikan seiring dengan meningkatnya ukuran data. Pada ukuran data kecil (di bawah 500 MB), throughput Hadoop dan Spark relatif rendah dan sebanding. Namun, setelah titik 500 MB, Spark secara konsisten menunjukkan throughput yang lebih tinggi, mencapai 68.68 MB/s pada 15 GB dibandingkan dengan 45.39 MB/s untuk Hadoop. Pada tugas word count, Spark juga mengungguli Hadoop dalam hal throughput, meskipun dengan perbedaan yang tidak sebesar pada tugas sorting. Spark mulai menunjukkan throughput yang lebih tinggi pada ukuran data 100 MB. Pada ukuran data 15 GB, Spark mencapai throughput 99.405 MB/s, sedangkan Hadoop hanya mencapai 18.407 MB/s.



Gambar 4.6 Rata-rata Waktu Eksekusi (*Word Count*)

4.1.5 Rate of Change

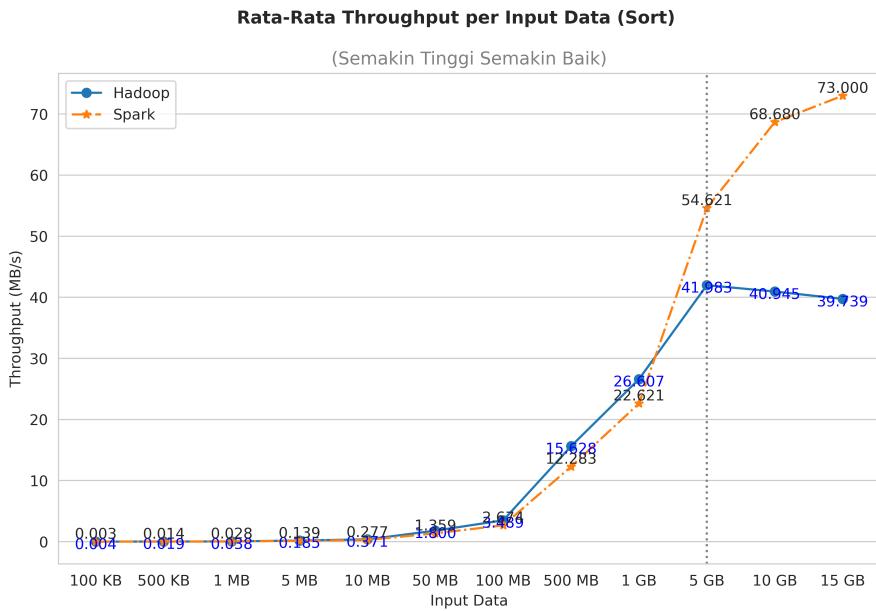
4.1.6 Penggunaan CPU

Gambar 4.12 dan 4.13 menunjukkan pola penggunaan CPU oleh Hadoop dan Spark untuk tugas sorting dan word count pada berbagai ukuran data. Sumbu x mewakili waktu dalam detik, sedangkan sumbu y mewakili persentase penggunaan CPU. Setiap baris grafik menunjukkan ukuran data yang berbeda, mulai dari 100 KB hingga 15 GB.

Pada kedua tugas, terlihat bahwa Spark cenderung menunjukkan pola penggunaan CPU yang lebih stabil dan konsisten dibandingkan dengan Hadoop. Penggunaan CPU Spark umumnya lebih tinggi dan merata di sepanjang waktu eksekusi, mengindikasikan pemanfaatan sumber daya yang lebih efisien dan pemrosesan paralel yang lebih optimal. Di sisi lain, penggunaan CPU Hadoop cenderung lebih fluktuatif, dengan periode lonjakan dan penurunan yang signifikan.

Pada tugas sorting (Gambar 4.12), perbedaan pola penggunaan CPU antara Hadoop dan Spark semakin terlihat pada ukuran data yang lebih besar. Spark mempertahankan penggunaan CPU yang tinggi dan stabil, sementara Hadoop menunjukkan fluktuasi yang lebih besar dan cenderung menurun pada akhir eksekusi. Hal ini menunjukkan bahwa Spark lebih mampu menangani data besar secara efisien dan konsisten.

Pada tugas word count (Gambar 4.13), perbedaan pola penggunaan CPU tidak sejelas pada tugas sorting. Namun, Spark tetap menunjukkan penggunaan CPU yang lebih stabil dan tinggi secara keseluruhan dibandingkan dengan Hadoop.



Gambar 4.7 Rata-rata Throughput (Sort)

Gambar 4.14 dan 4.15 menyajikan bar chart yang menggambarkan perbandingan persentase state U_s (user) antara Hadoop dan Spark untuk tugas sorting dan word count pada berbagai ukuran data dan skenario. State U_s merepresentasikan waktu CPU yang dihabiskan dalam mode user, yang menunjukkan waktu yang dihabiskan untuk menjalankan kode aplikasi.

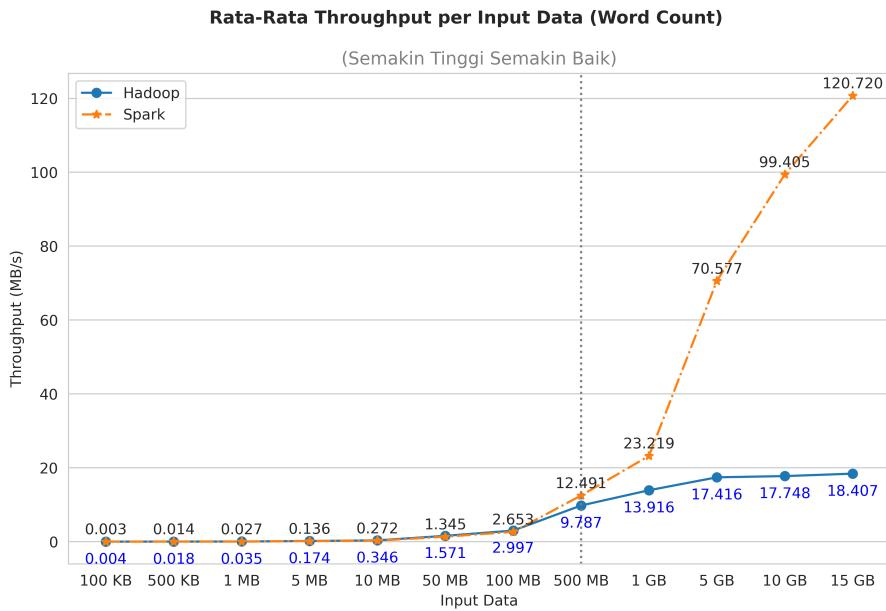
Pada kedua tugas, Spark secara konsisten menunjukkan persentase state U_s yang lebih tinggi dibandingkan dengan Hadoop. Hal ini menunjukkan bahwa Spark lebih efisien dalam memanfaatkan waktu CPU untuk menjalankan kode aplikasi, sembari Hadoop menghabiskan lebih banyak waktu dalam state lain, seperti state sistem (S_s) atau state idle (I_s).

Pada tugas sorting (Gambar 4.14), perbedaan persentase state U_s antara Spark dan Hadoop semakin terlihat pada ukuran data yang lebih besar. Hal ini menunjukkan bahwa Spark lebih mampu memaksimalkan penggunaan CPU untuk tugas komputasi intensif seperti sorting, terutama saat menangani data dalam jumlah besar.

Pada tugas word count (Gambar 4.15), meskipun Spark masih menunjukkan persentase state U_s yang lebih tinggi, perbedaannya tidak sejelas pada tugas sorting. Hal ini mungkin karena tugas word count tidak seintensif sorting secara komputasi.

4.1.7 Utilisasi Sistem

Gambar ?? dan ?? menyajikan informasi pemantauan sistem yang membandingkan penggunaan sumber daya komputasi oleh Hadoop dan Spark selama menjalankan tugas sorting dan word count dengan ukuran data "fiveteengig". Setiap gambar



Gambar 4.8 Rata-rata Throughput (Word Count)

terdiri dari tiga grafik yang menunjukkan penggunaan CPU, Disk I/O, dan memori seiring berjalananya waktu.

Pada kedua tugas, Spark menunjukkan pola penggunaan CPU yang lebih tinggi dan konsisten dibandingkan dengan Hadoop. Grafik penggunaan CPU Spark menunjukkan garis yang cenderung mendatar di dekat tingkat utilisasi maksimum, mengindikasikan bahwa Spark mampu memaksimalkan pemanfaatan CPU untuk pemrosesan data secara terus-menerus. Di sisi lain, grafik penggunaan CPU Hadoop menunjukkan fluktuasi yang lebih besar, dengan periode lonjakan dan penurunan yang signifikan. Hal ini menunjukkan bahwa Hadoop mengalami periode idle yang lebih lama dan tidak memanfaatkan sumber daya CPU seefisien Spark.

Hadoop menunjukkan aktivitas Disk I/O yang jauh lebih tinggi dibandingkan dengan Spark, terutama pada tugas sorting (Gambar ??). Grafik Disk I/O Hadoop menunjukkan lonjakan aktivitas baca dan tulis yang signifikan sepanjang waktu eksekusi. Hal ini sesuai dengan pendekatan berbasis disk Hadoop yang membutuhkan pembacaan dan penulisan data ke disk secara intensif. Sebaliknya, Spark, dengan arsitektur in-memory, meminimalkan operasi Disk I/O. Grafik Disk I/O Spark menunjukkan aktivitas yang jauh lebih rendah dan stabil, yang berkontribusi pada peningkatan performanya.

Spark menunjukkan penggunaan memori yang lebih tinggi dan stabil dibandingkan dengan Hadoop, terutama pada tugas word count (Gambar ??). Grafik penggunaan memori Spark menunjukkan garis yang cenderung mendatar pada tingkat utilisasi yang tinggi, menunjukkan bahwa Spark menyimpan data dalam RAM untuk akses

yang lebih cepat dan pemrosesan yang efisien. Penggunaan memori Hadoop lebih rendah dan fluktuatif, menunjukkan bahwa Hadoop tidak memanfaatkan memori secara optimal.

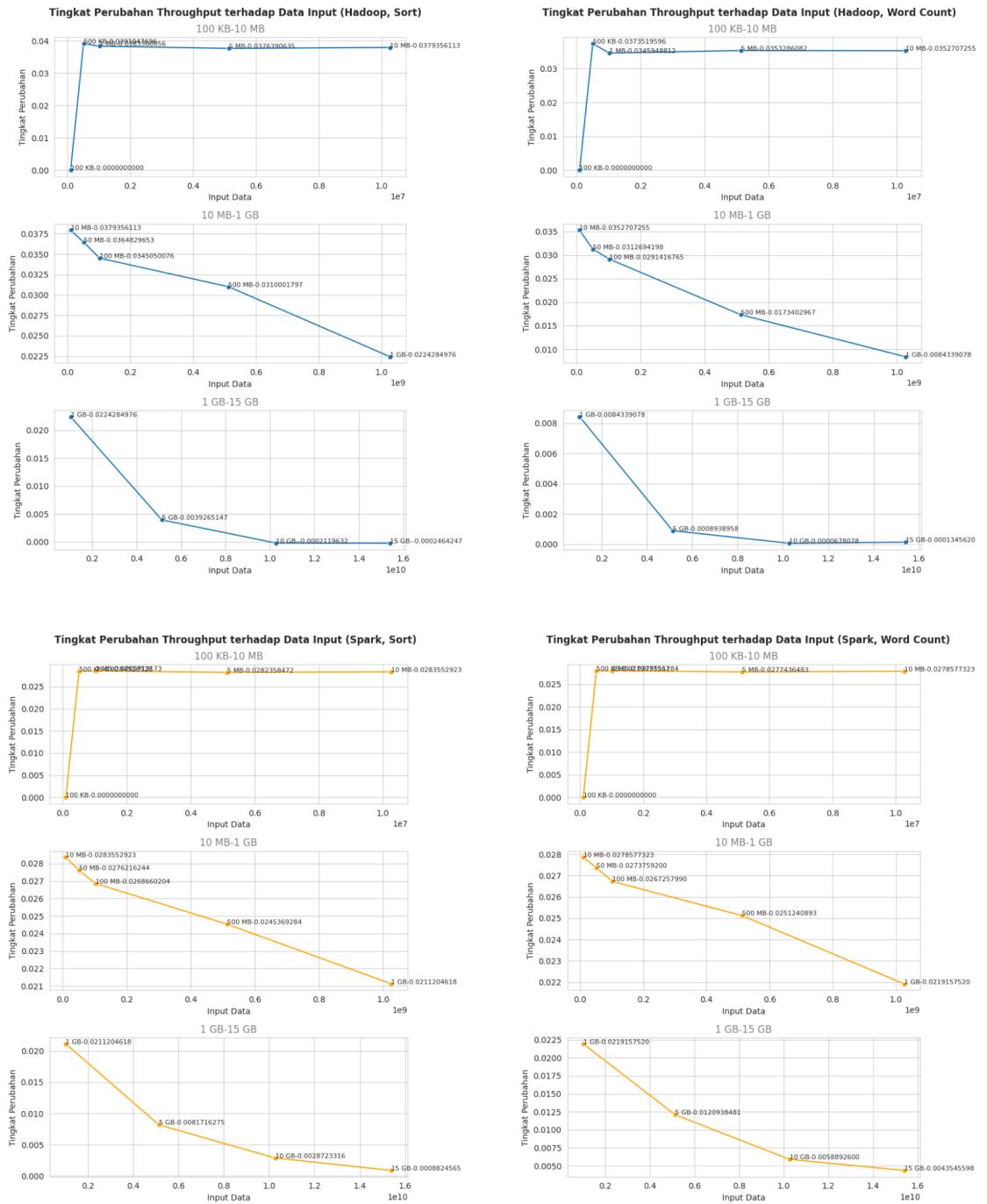
Analisis pemantauan sistem menegaskan keunggulan Spark dalam hal efisiensi dan optimasi penggunaan sumber daya komputasi dibandingkan dengan Hadoop. Spark mampu memaksimalkan penggunaan CPU, meminimalkan operasi Disk I/O, dan memanfaatkan memori secara efisien, yang berkontribusi pada performa dan skalabilitas yang lebih baik dalam tugas-tugas pemrosesan data besar.

Input Data 100 KB

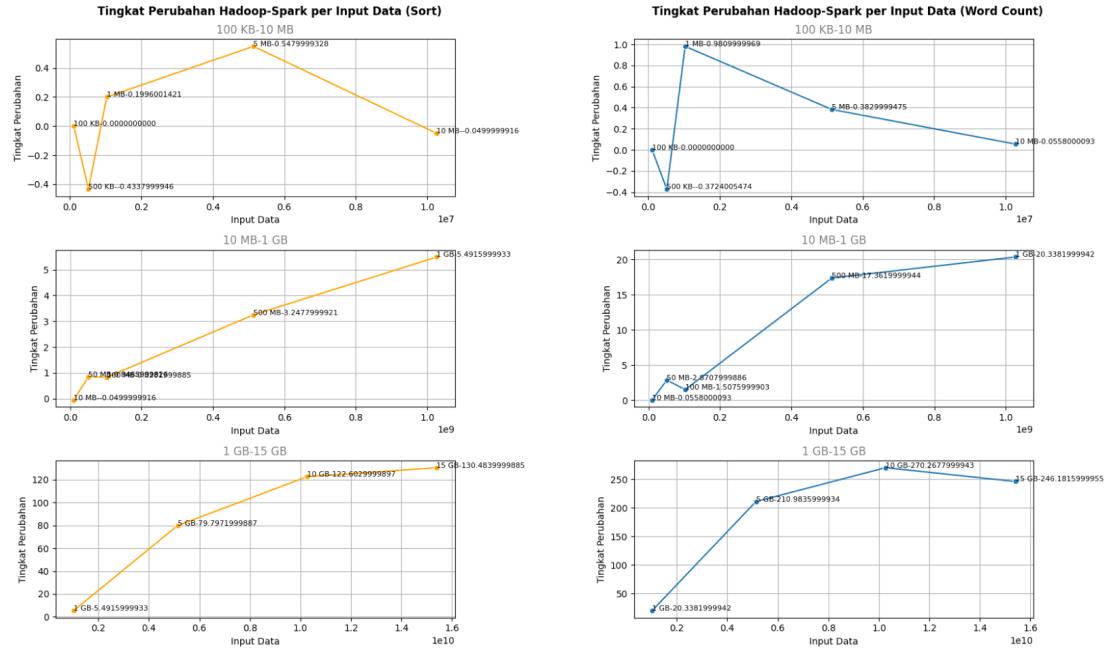
Input Data 15 GB



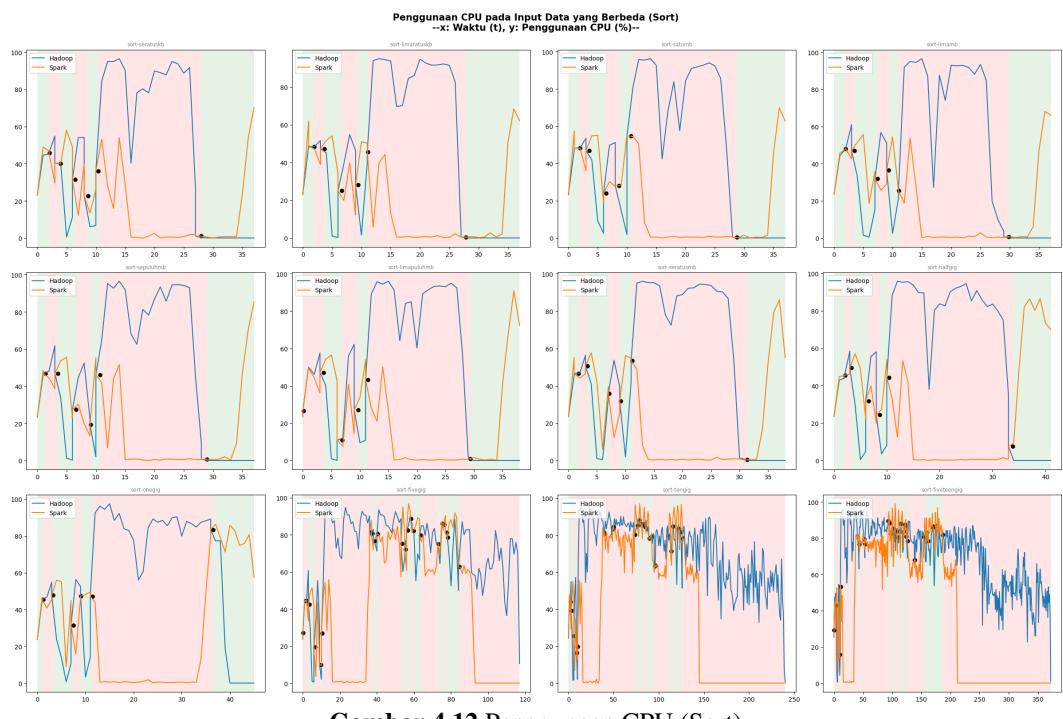
Gambar 4.9 dur



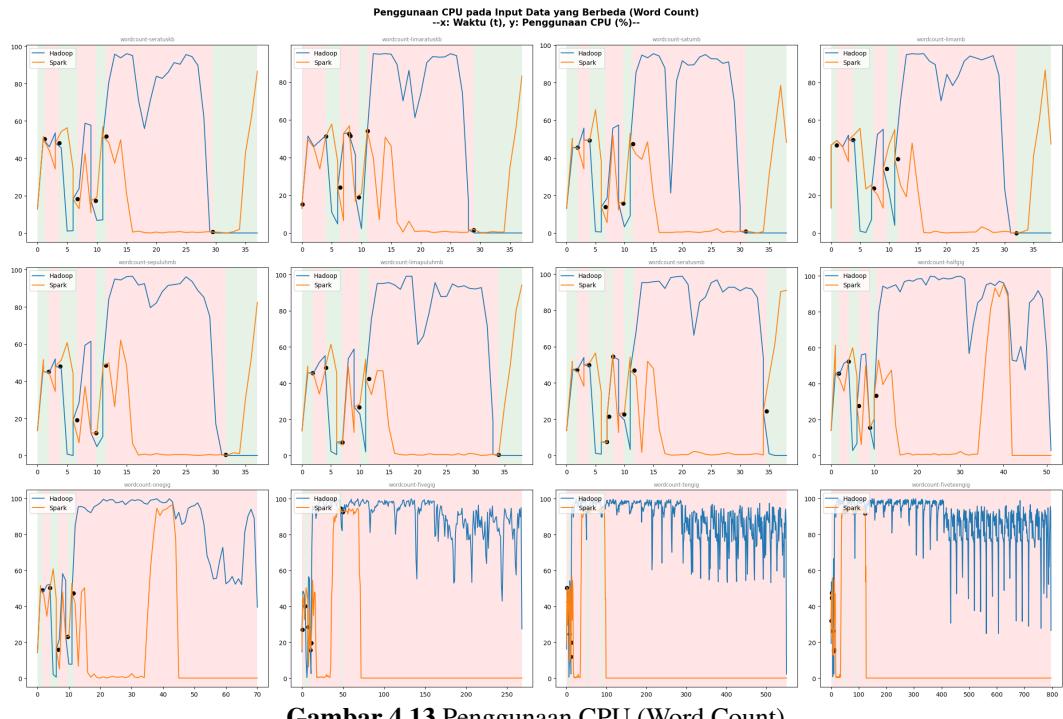
Gambar 4.10 th



Gambar 4.11 hadoop-spark



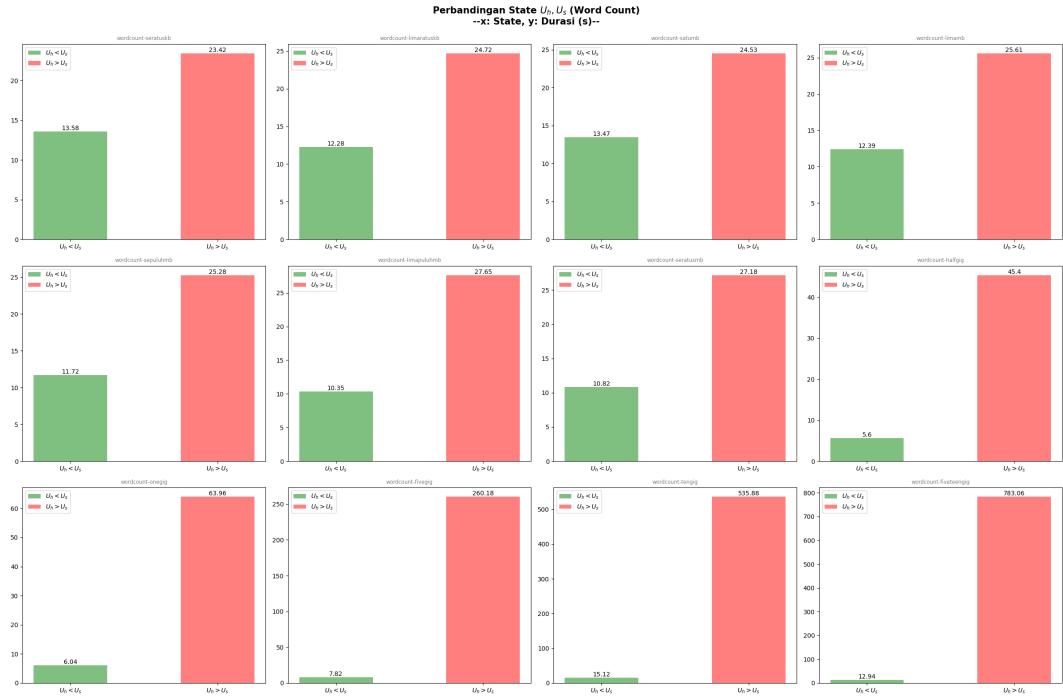
Gambar 4.12 Penggunaan CPU (Sort)



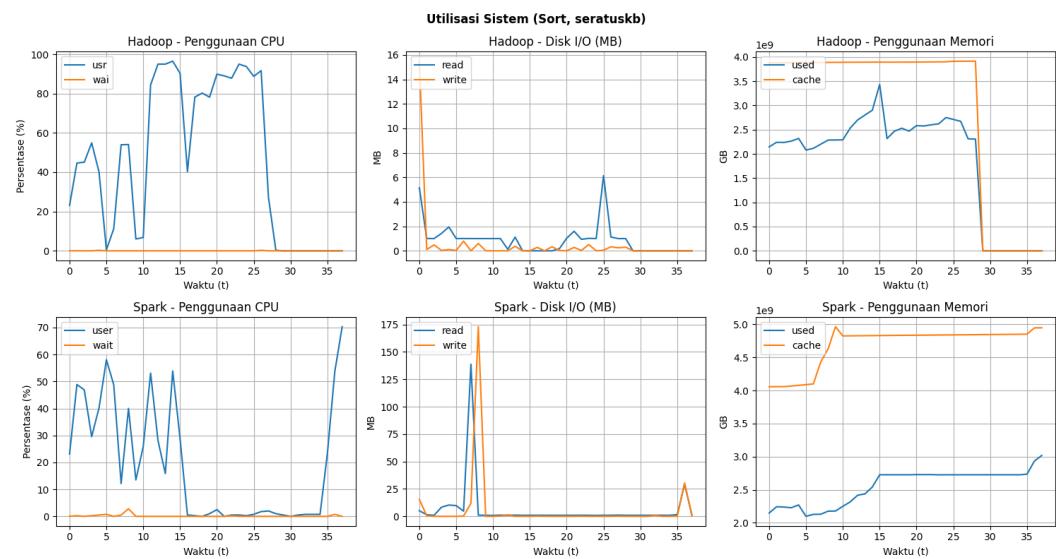
Gambar 4.13 Penggunaan CPU (Word Count)



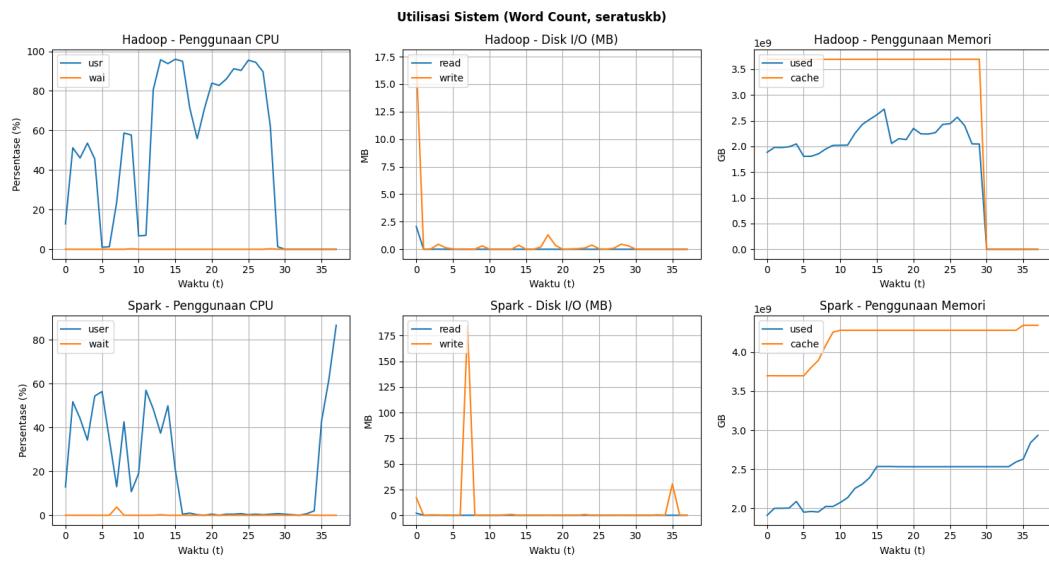
Gambar 4.14 State (Sort)



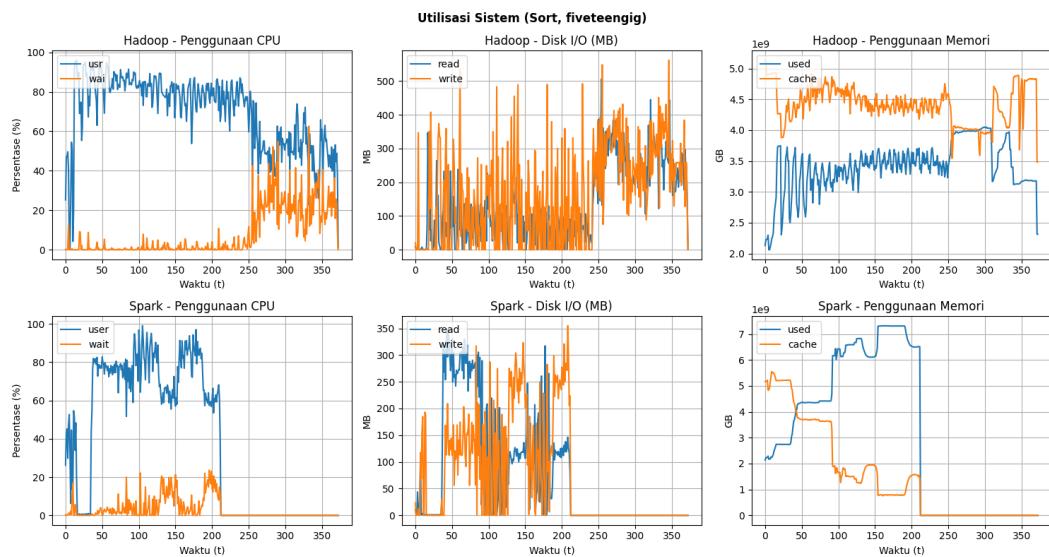
Gambar 4.15 State (Word Count)



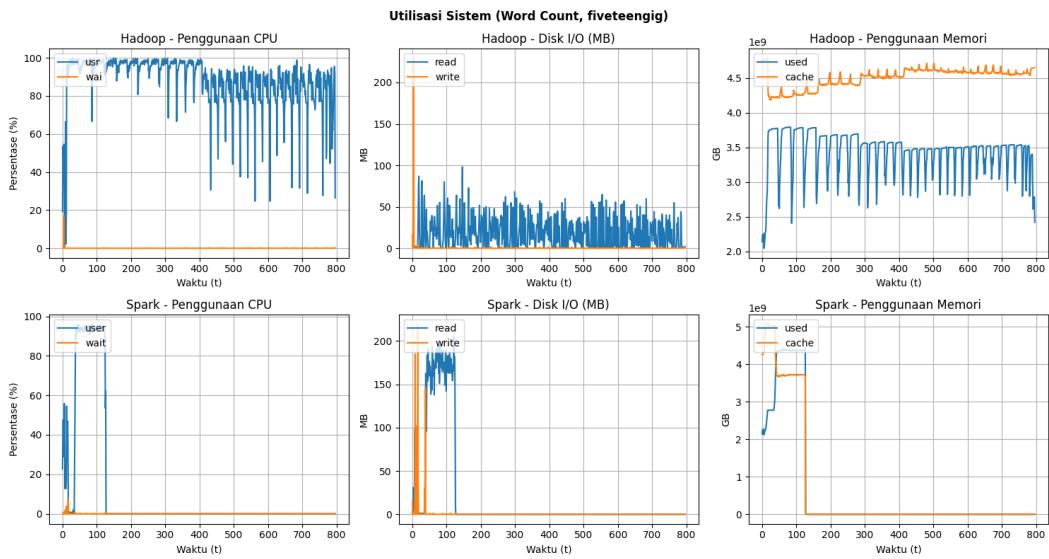
Gambar 4.16 Utilisasi Sistem (Sort) pada Input Data 100 KB



Gambar 4.17 Utilisasi Sistem (Word Count) pada Input Data 100 KB



Gambar 4.18 Utilisasi Sistem (Sort) pada Input Data 15 GB



Gambar 4.19 Utilisasi Sistem (Word Count) pada Input Data 15 GB

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini melakukan evaluasi komprehensif terhadap kinerja Hadoop dan Spark dalam konteks pemrosesan *Big Data*. Dengan menggunakan beban kerja *Micro Benchmarks*, penelitian ini berhasil mengukur dan membandingkan performa kedua platform tersebut dalam mode *pseudo-distributed*. Hasil eksperimen menunjukkan perbedaan signifikan dalam hal efisiensi dan penggunaan sumber daya antara Hadoop dan Spark, dengan Spark menunjukkan keunggulan dalam sejumlah aspek. Analisis ini memberikan wawasan berharga bagi organisasi yang ingin memilih platform Big Data yang tepat, memastikan keputusan mereka didasarkan pada informasi yang akurat dan relevan.

5.2 Saran

Untuk penelitian selanjutnya, disarankan untuk menggali lebih dalam aspek keamanan dan administrasi dari kedua *platform* tersebut. Penelitian yang lebih fokus pada pengukuran aspek skalabilitas dan ketersediaan Hadoop dan Spark dalam berbagai konfigurasi kluster juga akan bermanfaat. Selain itu, penelitian tentang integrasi teknologi baru seperti kontainer dan orkestrasi kluster dalam pemrosesan *Big Data* dapat menjadi topik yang menjanjikan. Akhirnya, implementasi kasus penggunaan nyata dan studi komparatif dalam lingkungan produksi akan memberikan pemahaman yang lebih dalam tentang kinerja dan kegunaan kedua *platform* ini dalam skenario dunia nyata.

DAFTAR PUSTAKA

- [1] Y. Samadi, M. Zbakh, dan C. Tadonki, “Performance comparison between Hadoop and Spark frameworks using HiBench benchmarks”, *Concurrency and Computation: Practice and Experience*, vol. 30, no. 12, e4367, 2018. (dikunjungi pd. 09/21/2023).
- [2] D. Reinsel, J. Gantz, dan J. Rydning, “The Digitization of the World from Edge to Core”, 2018.
- [3] C. Adrian, R. Abdullah, R. Atan, dan Y. Y. Jusoh, “Expert Review on Big Data Analytics Implementation Model in Data-driven Decision-Making”, di dalam *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, Kota Kinabalu, Malaysia: IEEE, Mar. 2018, hlmn. 1–5. (dikunjungi pd. 01/17/2024).
- [4] B. E. Syahputra dan A. Afnan, “Pendeteksian Fraud: Peran Big Data dan Audit Forensik”, *Jurnal ASET (Akuntansi Riset)*, vol. 12, no. 2, hlmn. 301–316, Des. 2020. (dikunjungi pd. 01/18/2024).
- [5] T. W. Sulaiman, R. B. Fitriansyah, A. R. Alaudin, dan M. H. Ratsanjani, “LITERATURE REVIEW: PENERAPAN BIG DATA DALAM KESEHATAN MASYARAKAT”, vol. 1, 2023.
- [6] N. Fernando, M. Mery, J. Jessica, dan J. Andry, “Utilization of Big Data In E-Commerce Business”, *Conference Series*, vol. 3, hlmn. 62–67, Nov. 2020.
- [7] *KOMPARASI KECEPATAN HADOOP MAPREDUCE DAN APACHE SPARK DALAM MENGOLAH DATA TEKS | Jurnal Ilmiah Matrik*, <https://journal.binadarma.ac.id/index.php/jurnalmatrik/article/view/1649>. (dikunjungi pd. 09/21/2023).
- [8] M. Saadoon, S. H. Ab. Hamid, H. Sofian, H. H. M. Altarturi, Z. H. Azizul, dan N. Nasuha, “Fault tolerance in big data storage and processing systems: A review on challenges and solutions”, *Ain Shams Engineering Journal*, vol. 13, no. 2, hlmn. 101 538, Mar. 2022. (dikunjungi pd. 01/18/2024).
- [9] D. Bhattacharya, F. Currim, dan S. Ram, “Evaluating Distributed Computing Infrastructures: An Empirical Study Comparing Hadoop Deployments on Cloud and Local Systems”, *IEEE Transactions on Cloud Computing*, vol. 9, no. 3, hlmn. 1075–1088, Juli 2021. (dikunjungi pd. 10/13/2023).

- [10] N. Ahmed, A. L. C. Barczak, T. Susnjak, dan M. A. Rashid, “A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench”, *J Big Data*, vol. 7, no. 1, hlmn. 110, Des. 2020. (dikunjungi pd. 11/23/2023).
- [11] R. Saputro, A. Aminuddin, dan Y. Munarko, “Perbandingan Kinerja Komputasi Hadoop dan Spark untuk Memprediksi Cuaca (Studi Kasus : Storm Event Database)”, *Jurnal Repotor*, vol. 2, hlmn. 463, Mar. 2020.
- [12] J. Dean dan S. Ghemawat, “MapReduce: Simplified data processing on large clusters”, di dalam *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6*, Ser. OSDI’04, USA: USENIX Association, Des. 2004, hlmn. 10. (dikunjungi pd. 09/28/2023).
- [13] Y. Samadi, M. Zbakh, dan C. Tadonki, “Comparative study between Hadoop and Spark based on Hibench benchmarks”, di dalam *2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech)*, Marrakech, Morocco: IEEE, Mei 2016, hlmn. 267–275. (dikunjungi pd. 09/24/2023).
- [14] H. Ahmadvand, M. Goudarzi, dan F. Foroutan, “Gapprox: Using Gallup approach for approximation in Big Data processing”, *Journal of Big Data*, vol. 6, no. 1, hlmn. 20, Feb. 2019. (dikunjungi pd. 09/29/2023).
- [15] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, dan I. Stoica, “Spark: Cluster computing with working sets”, di dalam *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, Ser. HotCloud’10, USA: USENIX Association, Juni 2010, hlmn. 10. (dikunjungi pd. 09/28/2023).
- [16] S. Huang, J. Huang, J. Dai, T. Xie, dan B. Huang, “The HiBench Benchmark Suite: Characterization of the MapReduce-Based Data Analysis”,
- [17] J. Shi, Y. Qiu, U. F. Minhas, dkk., “Clash of the titans: MapReduce vs. Spark for large scale data analytics”, *Proc. VLDB Endow.*, vol. 8, no. 13, hlmn. 2110–2121, Sept. 2015. (dikunjungi pd. 09/28/2023).
- [18] S. Gopalani dan R. Arora, “Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means”, *IJCA*, vol. 113, no. 1, hlmn. 8–11, Mar. 2015. (dikunjungi pd. 10/13/2023).
- [19] A. Barosen dan S. Dalin, *Analysis and Comparison of Interfacing, Data Generation and Workload Implementation in BigDataBench 4.0 and Intel Hi-Bench 7.0*. 2018. (dikunjungi pd. 12/28/2023).

- [20] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, dan S. Belfkih, “Big Data technologies: A survey”, *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, hlmn. 431–448, Okt. 2018. (dikunjungi pd. 12/28/2023).
- [21] B. Furht dan F. Villanustre, “Introduction to Big Data”, di dalam Sept. 2016, hlmn. 3–11.
- [22] A. K. Sandhu, “Big data with cloud computing: Discussions and challenges”, *Big Data Mining and Analytics*, vol. 5, no. 1, hlmn. 32–40, Mar. 2022. (dikunjungi pd. 12/28/2023).
- [23] *Red Hot: The 2021 Machine Learning, AI and Data (MAD) Landscape*, <https://mattturck.com/data2021/>, Sept. 2021. (dikunjungi pd. 12/28/2023).
- [24] *About | DigitalOcean*, <https://www.digitalocean.com/about>. (dikunjungi pd. 05/07/2024).
- [25] C. Newham, B. Rosenblatt, dan B. Rosenblatt, *Learning the Bash Shell: Unix Shell Programming* (UNIX Shell Programming), 3. ed. Beijing Köln: O'Reilly, 2005.
- [26] K. Kalia dan N. Gupta, “Analysis of hadoop MapReduce scheduling in heterogeneous environment”, *Ain Shams Engineering Journal*, vol. 12, no. 1, hlmn. 1101–1110, Mar. 2021. (dikunjungi pd. 11/08/2023).
- [27] K. C dan A. X, “Task failure resilience technique for improving the performance of MapReduce in Hadoop”, *ETRI Journal*, vol. 42, no. 5, hlmn. 748–760, 2020. (dikunjungi pd. 11/08/2023).
- [28] H. Herodotou, *Hadoop Performance Models*, Juni 2011. arXiv: 1106.0940 [cs]. (dikunjungi pd. 11/08/2023).
- [29] M. Bakratsas, P. Basaras, D. Katsaros, dan L. Tassiulas, “Hadoop MapReduce Performance on SSDs for Analyzing Social Networks”, *Big Data Research*, vol. 11, hlmn. 1–10, Mar. 2018. (dikunjungi pd. 11/08/2023).
- [30] A. Gandomi, M. Reshadi, A. Movaghfar, dan A. Khademzadeh, “HybSMRP: A hybrid scheduling algorithm in Hadoop MapReduce framework”, *Journal of Big Data*, vol. 6, no. 1, hlmn. 106, Nov. 2019. (dikunjungi pd. 11/08/2023).
- [31] *MapReduce - Distributed Computing in Java 9 [Book]*, <https://www.oreilly.com/library/view/distributed-computing-in/9781787126992/5fef6ce5-20d7-4d7c-93eb-7e669d48c2b4.xhtml>. (dikunjungi pd. 11/08/2023).

- [32] *Apache Hadoop*, <https://hadoop.apache.org/>. (dikunjungi pd. 11/08/2023).
- [33] S. Maneas dan B. Schroeder, “The Evolution of the Hadoop Distributed File System”, di dalam *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, Mei 2018, hlmn. 67–74. (dikunjungi pd. 11/08/2023).
- [34] C. Dabas, P. Kaur, N. Gulati, dan M. Tilak, “Analysis of Comments on YouTube Videos using Hadoop”, di dalam *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Nov. 2019, hlmn. 353–358. (dikunjungi pd. 11/08/2023).
- [35] T. John dan P. Misra, *Data Lake for Enterprises: Leveraging Lambda Architecture for Building Enterprise Data Lake*. Birmingham, UK, Mumbai: Packt Publishing, 2017.
- [36] S. Khatai, S. S. Rautaray, S. Sahoo, dan M. Pandey, “An Implementation of Text Mining Decision Feedback Model Using Hadoop MapReduce”, di dalam *Trends of Data Science and Applications: Theory and Practices*, Ser. Studies in Computational Intelligence, S. S. Rautaray, P. Pemmaraju, dan H. Mohanty, timed., Singapore: Springer, 2021, hlmn. 273–306. (dikunjungi pd. 11/09/2023).
- [37] K. Abhishek, Department of CSE, NIT Patna, Ashok Rajpath, Mahendru, Patna - 800005, Bihar, India, M. Kumar Verma, dkk., “Integrated Hadoop Cloud Framework (IHCF)”, *Indian Journal of Science and Technology*, vol. 10, no. 10, hlmn. 1–8, Feb. 2017. (dikunjungi pd. 11/10/2023).
- [38] *Apache Hadoop 3.3.6 – HDFS Architecture*, <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>. (dikunjungi pd. 11/10/2023).
- [39] H. T. Almansouri dan Y. Masmoudi, “Hadoop Distributed File System for Big data analysis”, di dalam *2019 4th World Conference on Complex Systems (WCCS)*, Ouarzazate, Morocco: IEEE, Apr. 2019, hlmn. 1–5. (dikunjungi pd. 11/08/2023).
- [40] *Apache Hadoop 3.3.6 – Apache Hadoop YARN*, <https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html>. (dikunjungi pd. 12/28/2023).
- [41] *Apache Spark™ - Unified Engine for large-scale data analytics*, <https://spark.apache.org/>. (dikunjungi pd. 11/10/2023).

- [42] *Apache Spark - Introduction*, https://www.tutorialspoint.com/apache_spark.htm.
(dikunjungi pd. 12/30/2023).
- [43] *Apache Spark - RDD*, https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm.
(dikunjungi pd. 01/15/2024).
- [44] *Intel-bigdata/HiBench*, Intel-bigdata, Des. 2023. (dikunjungi pd. 12/30/2023).

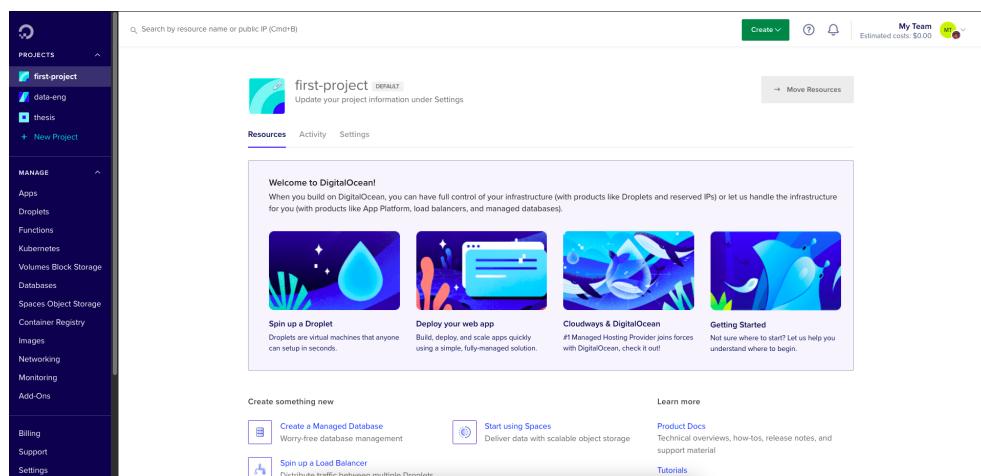
LAMPIRAN

LAMPIRAN A

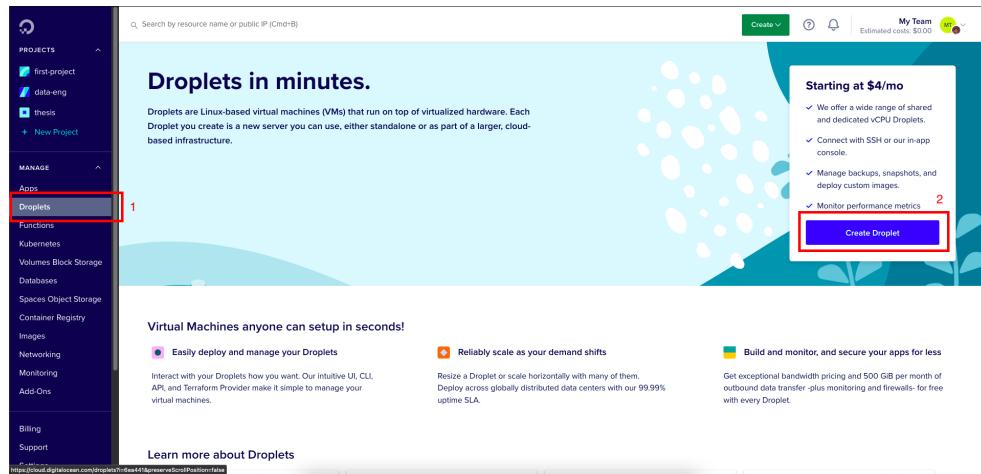
Pembuatan *Virtual Machine* (VM) pada DigitalOcean

Langkah-langkah pembuatan VM pada DigitalOcean dijelaskan seperti berikut,

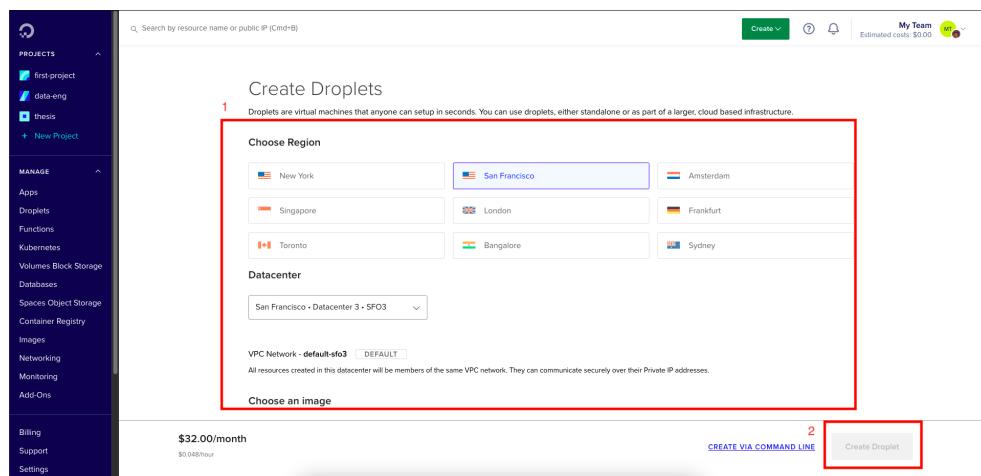
1. Buatlah akun DigitalOcean terlebih dahulu. Jika belum memiliki akun DigitalOcean, disarankan untuk mendaftar melalui *GitHub Student Developer Pack* sehingga nantinya akan diberikan kredit \$200 secara gratis. Jika sudah memiliki akun DigitalOcean, silakan melakukan *login*.
2. Halaman dasbor DigitalOcean akan ditampilkan setelahnya.



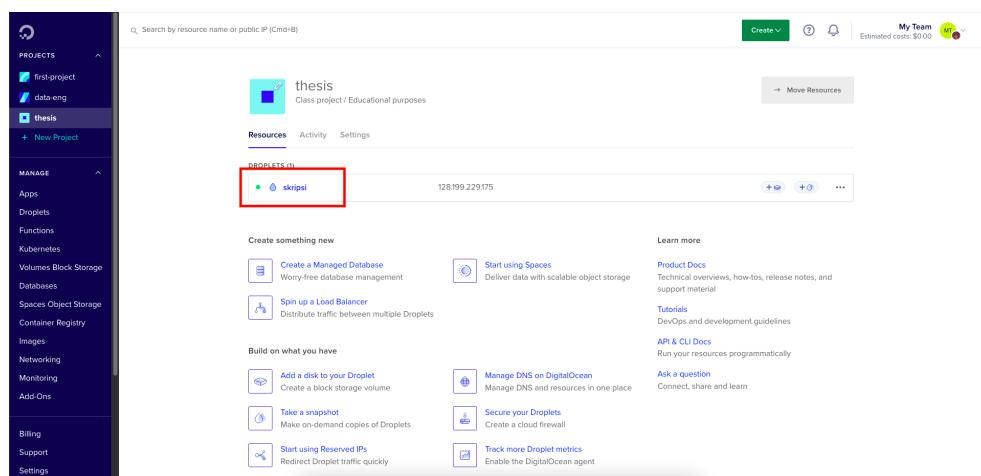
3. Perhatikan menu di sebelah kiri pada laman dasbor DigitalOcean. Tekan Droplets untuk masuk ke laman pembuatan VM. Selanjutnya, tekan *Create Droplets* berwarna biru untuk melakukan konfigurasi VM yang akan dibuat nantinya.



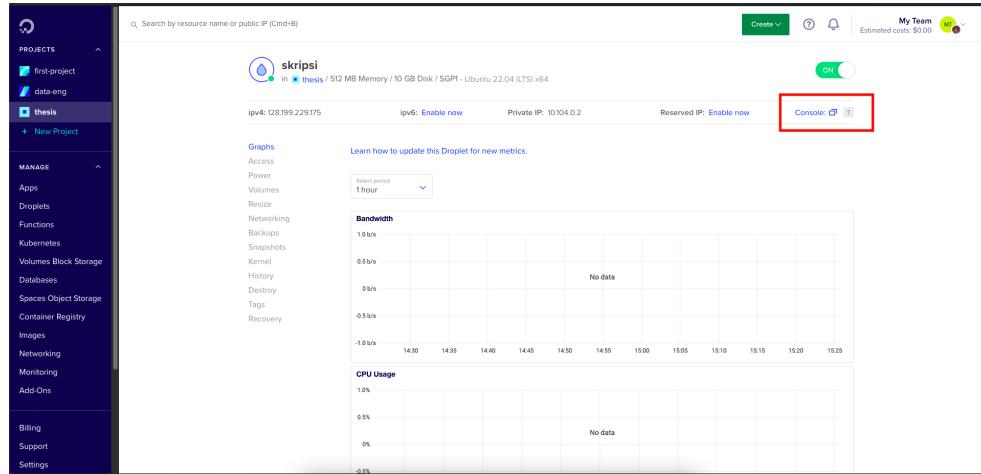
4. Pada laman pembuatan Droplets, lakukan konfigurasi sesuai dengan Tabel 3.1. Jika telah selesai melakukan konfigurasi, tekan *Create Droplets*.



5. Jika pembuatan Droplets berhasil, laman dasbor *Projects* DigitalOcean akan terlihat. Pada bagian Resources akan terlihat Droplets yang baru saja kita buat. Selanjutnya, tekan nama Droplets yang baru saja dibuat.



- Selanjutnya, laman konfigurasi Droplets akan terlihat. Jika diperlukan konfigurasi lanjutan dapat diatur melalui laman ini. Pada tahap ini hanya akan fokus pada konfigurasi perangkat lunak tanpa konfigurasi perangkat keras lebih jauh. Untuk masuk ke VM yang sudah dibuat, tekan Console. Tab baru akan dibuka.



- Akhirnya, lakukan konfigurasi perangkat lunak pada bagian ini.

```

skripsi - DigitalOcean Droplet Web Console
https://cloud.digitalocean.com/droplets/39235116/terminal/ui/
Welcome to Ubuntu 22.04.2 LTS (GNU/Linux 5.15.0-67-generic x86_64)

 * Documentation: https://help.ubuntu.com
 * Management: https://landscape.canonical.com
 * Support: https://ubuntu.com/advantage

System information disabled due to load higher than 1.0
Expanded Security Maintenance for Applications is not enabled.

17 updates can be applied immediately.
13 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The list of available updates is more than a week old.
To check for new updates run: sudo apt update

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

root@skripsi:~# 

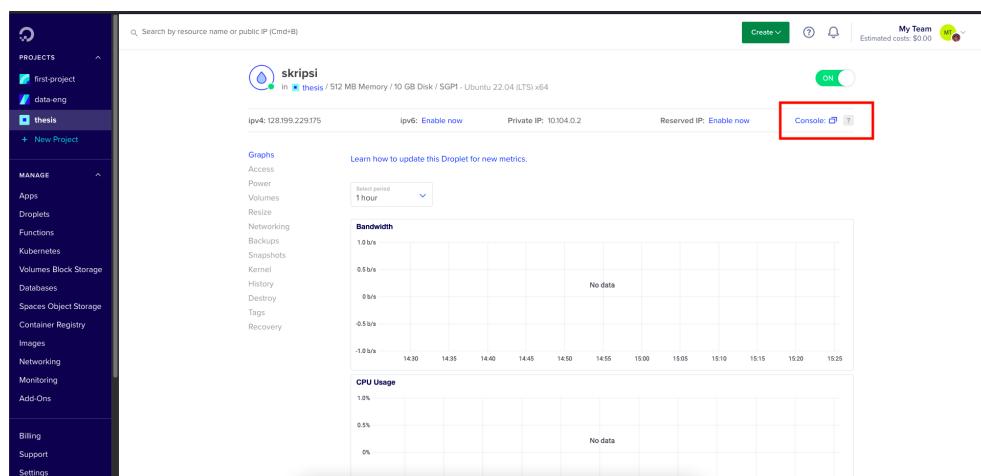
```

LAMPIRAN B

Instalasi dan Konfigurasi Perangkat Lunak Prasyarat

Pemasangan dan konfigurasi perangkat lunak adalah hal yang krusial. Sebelum dilakukan pemasangan perangkat lunak penyimpanan dan pemrosesan *big data*, tentunya perlu disiapkan perangkat lunak prasyarat. Perangkat lunak prasyarat yang dibutuhkan meliputi Git, Java dan Maven, Python, serta Scala. Langkah-langkah pemasangan dan konfigurasi perangkat lunak akan dijelaskan sebagai berikut,

1. Pastikan Droplets pada DigitalOcean sudah dibuat. Masuk ke *Virtual Machine* (VM) yang sebelumnya sudah dibuat melalui *Console* yang berada pada laman konfigurasi Droplets DigitalOcean.



2. Jika Droplets baru saja dibuat, perlu dilakukan pembaruan *index* pada *package management*. *Package management* adalah sistem atau sekumpulan alat yang digunakan untuk mengotomatiskan penginstalan, peningkatan, konfigurasi, dan penggunaan perangkat lunak. Pembaruan *package management* dapat dilakukan dengan `sudo apt update`.
3. Membuat Pengguna Baru
 - (a) Pertama, buatlah grup baru yang bernama *hadoop* dengan perintah `sudo addgroup hadoop`.
 - (b) Kemudian, tambahkan pengguna baru *hdfsuser* dalam grup *hadoop* yang sama dengan perintah `sudo adduser --ingroup hadoop hdfsuser`.
 - (c) Berikan *hdfsuser* izin *root* yang diperlukan untuk pemasangan file. Hak istimewa pengguna *root* dapat diberikan dengan memperbarui file *sudoers*. Buka file *sudoers* dengan menjalankan perintah `sudo visudo`. Tambahkan baris berikut, yaitu `hdfsuser ALL=(ALL:ALL) ALL`.

- (d) Sekarang, simpan perubahan dan tutup editor.
- (e) Selanjutnya, mari beralih ke pengguna baru yang telah dibuat untuk instalasi lebih lanjut menggunakan perintah `su - hdfsuser`.
4. Pengaturan *SSH keys* untuk Hadoop
- (a) Hadoop menggunakan *Secure Shell* (SSH) untuk menjalankan proses antara *master nodes* dan *slave nodes*. Penggunaan SSH akan memberikan banyak keuntungan, salah satunya adalah kecepatan. Jika sebuah klaster aktif dan berjalan, komunikasi antar *nodes* akan berjalan terlalu sering. Begitu pula dengan *job tracker* yang harus sering mengirimkan informasi *task to task* dengan cepat. Lakukan pemasangan ssh dan sshd dengan cara `sudo apt-get install ssh` dan `sudo apt-get install sshd` pada terminal.
- (b) Selanjutnya, lakukan pembuatan *SSH keys* dengan cara `ssh-keygen -t rsa`. Jika pembuatan *SSH keys* sudah dilakukan, jalankan perintah `cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`.
- (c) Ubah perizinan berkas dengan perintah `chmod og-wx ~/.ssh/authorized_keys`.
- (d) Terakhir, untuk memverifikasi koneksi aman sudah terjadi, lakukan `ssh localhost`.
5. Instalasi Git
- (a) Git dapat dipasang menggunakan perintah `sudo apt install git`. Pengguna akan diminta konfirmasi untuk menginstall. Ketik `y` kemudian tekan enter.
- (b) Untuk mengecek versi Git, dapat menggunakan perintah `git --version`.
6. Instalasi Python
- (a) Python dapat dipasang menggunakan perintah `sudo apt-get install python2 && sudo apt install python3.7`. Pengguna akan diminta konfirmasi untuk menginstall. Ketik `y` kemudian tekan enter.
- (b) Untuk mengecek versi Python, dapat menggunakan perintah `python --version`.
7. Instalasi Java 8 dan Maven
- (a) Java 8 dapat dipasang menggunakan perintah `sudo apt install openjdk-8-jre-headless openjdk-8-jdk`. Pengguna akan diminta konfirmasi untuk menginstall. Ketik `y`

kemudian tekan enter.

- (b) Versi dari Java dapat dilihat menggunakan perintah `java -version`.
- (c) Selanjutnya, instalasi Maven dapat dilakukan menggunakan perintah `sudo apt-get -y install maven`.
- (d) Informasi dari Maven beserta Java yang digunakan dapat dilihat menggunakan perintah `mvn -version`.

8. Instalasi Scala

- (a) Scala yang akan dipasang adalah versi 2.12. Jika menggunakan manajer paket, versi yang akan dipasang adalah versi terbaru. Untuk mengunduh versi spesifik dari Scala, dapat menggunakan perintah `sudo wget https://downloads.lightbend.com/scala/2.12.0/scala-2.12.0.deb`.
- (b) Scala dapat dipasang menggunakan perintah `sudo dpkg -i scala-2.12.0.deb`.
- (c) Versi Scala dapat dilihat melalui perintah `scala -version`.

LAMPIRAN C

Instalasi dan Konfigurasi Hadoop

Langkah-langkah pemasangan dan konfigurasi Hadoop akan dijelaskan sebagai berikut,

1. Unduh Hadoop
 - (a) Pastikan perangkat lunak prasyarat sudah berhasil dipasang dan dilakukan konfigurasi. Sebelum dilakukan pemasangan Hadoop, diperlukan untuk mengunduh berkas Hadoop terlebih dahulu dengan perintah `cd /usr/local`, dilanjutkan dengan `sudo wget https://archive.apache.org/dist/hadoop/common/hadoop-3.2.0/hadoop-3.2.0.tar.gz`.
 - (b) Ekstrak berkas Hadoop yang sudah diunduh tadi dengan perintah `sudo tar xvzf hadoop-3.2.0.tar.gz`. Hasil ekstrak berkas Hadoop akan disimpan pada direktori yang sama.
 - (c) Selanjutnya, untuk memudahkan kedepannya, ganti nama folder Hadoop dengan perintah `sudo mv hadoop-3.2.0 hadoop`.
2. Mengubah Kepemilikan Berkas Hadoop
 - (a) Setelah berkas Hadoop sudah berhasil terunduh, selanjutnya ubah kepemilikan berkas Hadoop ke `hdfsuser` yang sebelumnya sudah kita buat dengan perintah `sudo chown -R hdfsuser:hadoop /usr/local/hadoop`.
 - (b) Tambahkan kekuasaan untuk membaca, menulis, dan mengeksekusi pada foler Hadoop dengan perintah `sudo chmod -R 777 /usr/local/hadoop`.
3. Mematikan *IPv6 Networks*
 - (a) Saat ini Hadoop belum mendukung penggunaan *IPv6 Networks*. Hadoop hanya dibangun dan diuji coba pada *IPv4 Networks*. Untuk mematikan IPv6, dapat dimulai dengan menjalankan perintah `cat /proc/sys/net/ipv6/conf/all/disable_ipv6`.
 - (b) Jika hasil yang diberikan bukan angka 1, maka beberapa langkah tambahan harus dijalankan. Jalankan perintah `sudo nano /etc/sysctl.conf`, kemudian tambahkan beberapa baris potongan kode berikut pada akhir berkas,

```
1 # Disable ipv6
2 net.ipv6.conf.all.disable_ipv6=1
3 net.ipv6.conf.default_ipv6=1
```

```
4     net.ipv6.conf.lo.disable_ipv6=1
```

- (c) Simpan berkas. Kemudian jalankan perintah `sudo sysctl -p` untuk mengaktifkan perubahan.

4. Menambahkan Hadoop pada *Environments Variables*

- (a) Hadoop perlu ditambahkan pada *Environments Variables* untuk memudahkan dalam melakukan eksekusi. Untuk menambahkannya, jalankan perintah `sudo nano ~/.bashrc`.
- (b) Tambahkan beberapa baris kode berikut pada akhir berkas *bashrc*.
-

```
1      # HADOOP ENVIRONMENT
2      export HADOOP_HOME=/usr/local/hadoop
3      export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
4      export HADOOP_MAPRED_HOME=/usr/local/hadoop
5      export HADOOP_COMMON_HOME=/usr/local/hadoop
6      export HADOOP_HDFS_HOME=/usr/local/hadoop
7      export YARN_HOME=/usr/local/hadoop
8      export PATH=$PATH:/usr/local/hadoop/bin
9      export PATH=$PATH:/usr/local/hadoop/sbin
10
11     # HADOOP NATIVE PATH
12     export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/←
13         lib/native
13     export HADOOP_OPTS=-Djava.library.path=←
14         $HADOOP_PREFIX/lib
```

- (c) Untuk mendapatkan perubahan dapat dilakukan dengan perintah `source ~/.bashrc`.

5. Konfigurasi Hadoop

- (a) Hadoop menggunakan berkas .xml untuk melakukan konfigurasi pada semua prosesnya. Biasanya, letak direktori untuk melakukan konfigurasi terletak pada `$HADOOP_HOME/etc/hadoop`. Oleh karena itu, jalankan perintah `cd /usr/local/hadoop/etc/hadoop/`.
- (b) Konfigurasi berkas *hadoop-env.sh* dapat dilakukan dengan perintah `sudo nano hadoop-env.sh`, dilanjutkan dengan menambahkan beberapa baris kode seperti di bawah ini,
-

```
1      export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
2      export JAVA_HOME=/usr
3      export HADOOP_HOME_WARN_SUPPRESS="TRUE"
4      export HADOOP_ROOT_LOGGER="WARN,DRFA"
5      export HDFS_NAMENODE_USER="hdfsuser"
6      export HDFS_DATANODE_USER="hdfsuser"
7      export HDFS_SECONDARYNAMENODE_USER="hdfsuser"
```

```
8     export YARN_RESOURCEMANAGER_USER="hdfsuser"
9     export YARN_NODEMANAGER_USER="hdfsuser"
```

- (c) Konfigurasi berkas *yarn-site.xml* dapat dilakukan dengan perintah **sudo nano yarn-site.xml**, dilanjutkan dengan menambahkan beberapa baris kode seperti di bawah ini,
-

```
1 <property>
2   <name>yarn.nodemanager.aux-services</name>
3   <value>mapreduce_shuffle</value>
4 </property>
5 <property>
6   <name>yarn.nodemanager.aux-services.mapreduce.<-
        shuffle.class</name>
7   <value>org.apache.hadoop.mapred.ShuffleHandler</<-
        value>
8 </property>
```

- (d) Konfigurasi berkas *hdfs-site.xml* dapat dilakukan dengan perintah **sudo nano hdfs-site.xml**, dilanjutkan dengan menambahkan beberapa baris kode seperti di bawah ini,
-

```
1 <property>
2   <name>dfs.replication</name>
3   <value>1</value>
4 </property>
5 <property>
6   <name>dfs.namenode.name.dir</name>
7   <value>/usr/local/hadoop/yarn_data/hdfs/namenode</<-
        value>
8 </property>
9 <property>
10  <name>dfs.datanode.data.dir</name>
11  <value>/usr/local/hadoop/yarn_data/hdfs/datanode</<-
        value>
12 </property>
13 <property>
14  <name>dfs.namenode.http-address</name>
15  <value>localhost:50070</value>
16 </property>
```

- (e) Konfigurasi berkas *core-site.xml* dapat dilakukan dengan perintah **sudo nano core-site.xml**, dilanjutkan dengan menambahkan beberapa baris kode seperti di bawah ini,
-

```
1 <property>
```

```
2      <name>hadoop.tmp.dir</name>
3      <value>/bigdata/hadoop/tmp</value>
4      </property>
5      <property>
6          <name>fs.default.name</name>
7          <value>hdfs://localhost:9000</value>
8      </property>
```

- (f) Konfigurasi berkas *mapred-site.xml* dapat dilakukan dengan perintah `sudo nano mapred-site.xml`, dilanjutkan dengan menambahkan beberapa baris kode seperti di bawah ini,

```
1      <property>
2          <name>mapred.framework.name</name>
3          <value>yarn</value>
4      </property>
5      <property>
6          <name>mapreduce.jobhistory.address</name>
7          <value>localhost:10020</value>
8      </property>
```

6. Membuat Direktori Hadoop untuk Menyimpan Data

- (a) Sesuai dengan apa yang ditulis pada *core-site.xml*, langkah pertama yang harus dilakukan adalah membuat direktori sementara untuk dfs menyimpan berkas dengan menjalankan perintah di bawah. Jalankan perintah berikut baris per baris.

```
1      sudo mkdir -p /bigdata/hadoop/tmp
2      sudo chown -R hdfsuser:hadoop /bigdata/hadoop/tmp
3      sudo chmod -R 777 /bigdata/hadoop/tmp
```

- (b) Selanjutnya, jalankan perintah berikut untuk membuat direktori untuk menyimpan berkas data sekaligus mengganti kepemilikan berkas. Jalankan perintah berikut baris per baris.

```
1      sudo mkdir -p /usr/local/hadoop/yarn_data/hdfs/←
        namenode
2      sudo mkdir -p /usr/local/hadoop/yarn_data/hdfs/←
        datanode
3      sudo chmod -R 777 /usr/local/hadoop/yarn_data/hdfs/←
        namenode
4      sudo chmod -R 777 /usr/local/hadoop/yarn_data/hdfs/←
        datanode
5      sudo chown -R hdfsuser:hadoop /usr/local/hadoop/←
        yarn_data/hdfs/namenode
```

```
6     sudo chown -R hdfsuser:hadoop /usr/local/hadoop/←  
      yarn_data/hdfs/datanode
```

- (c) Konfigurasi untuk Hadoop sudah selesai dan dapat dilanjutkan untuk menjalankan *Resource Manager* dan *Node Manager*
7. Menjalankan Hadoop
- (a) Sebelum menjalankan *Hadoop Core Services*, klaster harus dibersihkan dengan cara melakukan *format* pada *namenode*. Jalankan perintah `hdfs namenode -format`.
 - (b) Untuk menjalankan layanan Hadoop, dapat dilakukan dengan perintah `start-all.sh`.
 - (c) Perintah `jps` dapat dilakukan untuk mengecek apakah layanan Hadoop sudah berjalan.
 - (d) Untuk memberhentikan layanan Hadoop, dapat dilakukan dengan perintah `stop-all.sh` pada terminal.

LAMPIRAN D

Instalasi dan Konfigurasi Spark

Langkah-langkah pemasangan dan konfigurasi Spark akan dijelaskan sebagai berikut,

1. Unduh Berkas Spark
 - (a) Pastikan perangkat lunak prasyarat sudah berhasil dipasang dan dilakukan konfigurasi. Sebelum dilakukan pemasangan Spark, diperlukan untuk mengunduh berkas Spark terlebih dahulu dengan perintah `cd /usr/local`, dilanjutkan dengan `sudo wget https://archive.apache.org/dist/spark/spark-3.0.0/spark-3.0.0-bin-hadoop3.2.tgz`.
 - (b) Ekstrak berkas Spark yang sudah diunduh tadi dengan perintah `sudo tar xvf spark-3.0.0-bin-hadoop3.2.tgz`. Hasil ekstrak berkas Spark akan disimpan pada direktori yang sama.
 - (c) Selanjutnya, untuk memudahkan kedepannya, ganti nama folder Spark dengan perintah `sudo mv spark-3.0.0-bin-hadoop3.2 spark`.
2. Menambahkan Spark pada *Environments Variables*
 - (a) Spark perlu ditambahkan pada *Environments Variables* untuk memudahkan dalam melakukan eksekusi. Untuk menambahkannya, jalankan perintah `sudo nano ~/.bashrc`.
 - (b) Tambahkan beberapa baris kode berikut pada akhir berkas *bashrc*.

```
1 # SPARK ENVIRONMENT
2 export PATH=$PATH:/usr/local/spark/bin
3 export YARN_CONF_DIR=$HADOOP_HOME/etc/hadoop
4 export SPARK_HOME=$PATH:/usr/local/spark/bin
```

 - (c) Untuk mendapatkan perubahan dapat dilakukan dengan perintah `source ~/.bashrc`.
3. Menjalankan *Spark Shell*
 - (a) Pastikan bahwa Spark sudah ditambahkan pada *environments variables* dengan perintah `spark-submit --version`.
 - (b) Jalankan layanan Hadoop dengan perintah `start-all.sh`.
 - (c) Jalankan `spark-shell` dengan YARN menggunakan perintah `spark-shell --master yarn`.

LAMPIRAN E

Instalasi dan Konfigurasi HiBench

Langkah-langkah pemasangan dan konfigurasi HiBench akan dijelaskan sebagai berikut,

1. Unduh berkas HiBench
 - (a) Pastikan perangkat lunak prasyarat dan perangkat lunak *Big Data* sebelumnya sudah berhasil dipasang dan dilakukan konfigurasi. Tahap selanjutnya adalah mengunduh berkas HiBench dengan perintah `git clone https://github.com/Intel-bigdata/HiBench.git`. Pastikan berada pada folder `/home/hdfsuser`.
 - (b) Selanjutnya, berikan perizinan ke folder HiBench dengan cara `sudo chmod 755 HiBench`.
2. Membangun *framework benchmark*
 - (a) Sebelum HiBench dapat digunakan, diperlukan pembangunan beberapa modul yang dibutuhkan, misalnya modul *data generation*, modul *hadoopbench*, dan modul *sparkbench*. Langkah awal yang diperlukan adalah masuk ke folder HiBench dengan perintah `cd HiBench`.
 - (b) Selanjutnya, pastikan versi Hadoop, Spark, dan Scala sudah sesuai. Jalankan perintah `mvn -Phadoopbench -Psparkbench -Dhadoop=2.6 -Dspark=1.5 -Dmodules -Pmicro clean package`. Perintah ini akan membangun modul yang dibutuhkan menggunakan Maven. Tidak semua modul akan dipasang. Modul yang akan dipasang salah satunya adalah modul untuk *micro benchmark*.
 - (c) Jika ingin pembangunan modul yang lebih luas cakupannya dapat menggunakan perintah `mvn -Phadoopbench -Psparkbench -Dspark=1.5 -Dscala=2.11 clean package`.
3. Jalankan perintah `sudo apt install bc` supaya berkas Hibench *Report* dapat muncul.
4. Lakukan konfigurasi pada berkas *hibench.conf*. Buka berkas tersebut dengan perintah `sudo nano conf/hibench.conf`.
5. Lakukan beberapa perubahan pada baris sesuai dengan contoh di bawah ini.

<code>1 hibench.masters.hostnames</code>	<code>hdfs://localhost:9000</code>
<code>2 hibench.slaves.hostnames</code>	<code>localhost</code>

6. Menjalankan *Benchmark* Hadoop

- (a) Lakukan konfigurasi pada berkas *hadoop.conf*. Sebelum itu, salin *template* konfigurasinya dengan perintah `cp conf/hadoop.conf.template conf/hadoop.conf`.
- (b) Buka berkas *hadoop.conf* dengan perintah `sudo nano conf/hadoop.conf`. Selanjutnya, lakukan beberapa perubahan seperti pada contoh di bawah.
-

```
1   # Hadoop home
2   hibench.hadoop.home      /usr/local/hadoop
3
4   # The path of hadoop executable
5   hibench.hadoop.executable ${hibench.hadoop.home}-
6   /bin/hadoop
7
8   # Hadoop configraution directory
9   hibench.hadoop.configure.dir ${hibench.hadoop.home}-
10  /etc/hadoop
11
12
13  # The root HDFS path to store HiBench data
14  hibench.hdfs.master      hdfs://localhost:9000
15
16
17  # Hadoop release provider. Supported value: apache
18  hibench.hadoop.release    apache
```

- (c) Simpan perubahan yang sudah dilakukan.
- (d) Untuk menjalankan beban kerja yang sudah dirancang sebelumnya, perintah yang digunakan sebagai berikut. Jalankan baris per baris.
-

```
1   bin/workloads/micro/<nama-beban-kerja>/prepare/-
2   prepare.sh
3   bin/workloads/micro/<nama-beban-kerja>/hadoop/run-
4   .sh
```

- (e) Untuk melihat hasil dari *benchmark* dapat mengakses berkas pada `<HiBench_Root>/report/hibench.report`

7. Menjalankan *Benchmark* Spark

- (a) Lakukan konfigurasi pada berkas *spark.conf*. Sebelum itu, salin *template* konfigurasinya dengan perintah `cp conf/spark.conf.template conf/spark.conf`.
- (b) Buka berkas *spark.conf* dengan perintah `sudo nano conf/spark.conf`. Selanjutnya, lakukan beberapa perubahan seperti pada contoh di bawah.

```
1 # Spark home
2 hibench.spark.home      /usr/local/spark
3
4 # Spark master
5 #   standalone mode: spark://xxx:7077
6 #   YARN mode: yarn-client
7 hibench.spark.master    yarn
```

- (c) Simpan perubahan yang sudah dilakukan.
- (d) Untuk menjalankan beban kerja yang sudah dirancang sebelumnya, perintah yang digunakan sebagai berikut. Jalankan baris per baris.

```
1 bin/workloads/micro/<nama-beban-kerja>/prepare/←
  prepare.sh
2 bin/workloads/micro/<nama-beban-kerja>/spark/run.←
  sh
```

- (e) Untuk melihat hasil dari *benchmark* dapat mengakses berkas pada <HiBench_Root>/report/hibench.report

LAMPIRAN F

Skrip Otomatisasi Eksperimen

```
1 #!/bin/bash
2
3 # Ubah direktori kerja ke direktori HiBench
4 cd /home/hadoop/HiBench-HiBench-7.0
5
6 # Daftar workload
7 workloads=("wordcount" "sort")
8
9 # Daftar skala
10 scales=(
11     "seratuskb"
12     "limaratuskb"
13     "satumb"
14     "limamb"
15     "sepuluhmb"
16     "limapuluuhmb"
17     "seratusmb"
18     "halfgig"
19     "onegig"
20     "fivegig"
21     "tengig"
22     "fiveteengig"
23 )
24
25 # Jumlah pengulangan
26 repetitions=5
27
28 # Looping untuk setiap workload
29 for workload in "${workloads[@]}"; do
30     echo "Menjalankan workload: $workload"
31
32     # Looping untuk setiap skala
33     for scale in "${scales[@]}"; do
34         echo " Skala: $scale"
35
36         # Mengubah konfigurasi HiBench
37         sed -i "s/^hibench.scale.profile.*/hibench.scale.profile<-
38             $scale/" conf/hibench.conf
39
40         # Tahap persiapan Hadoop
41         while true; do
```

```

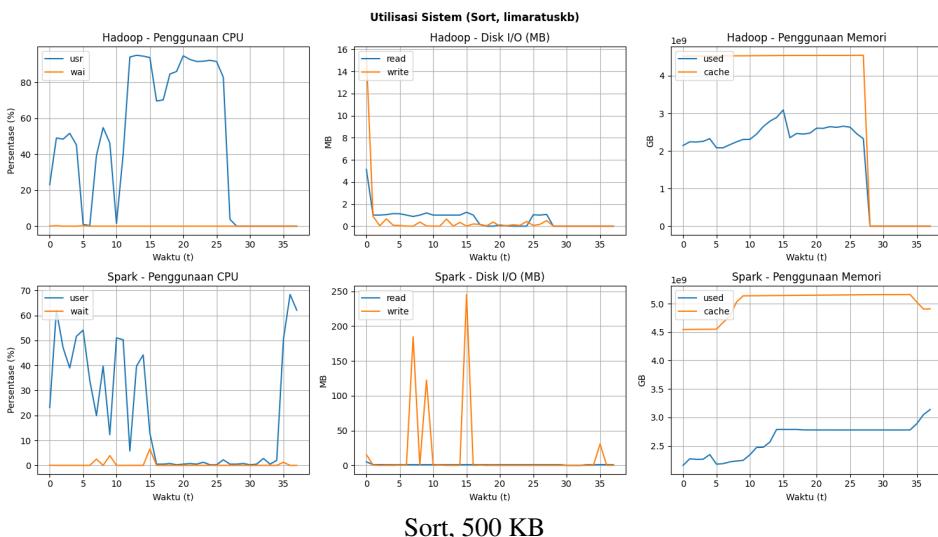
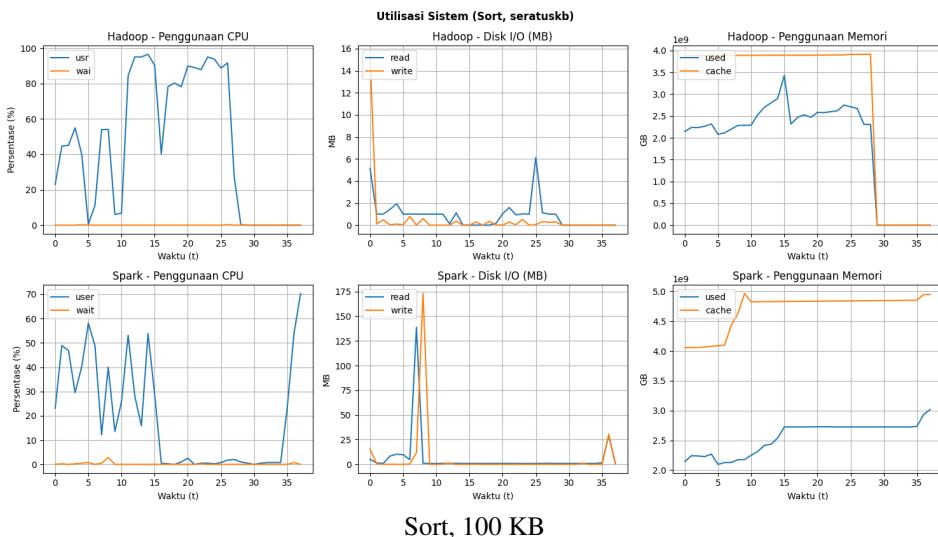
41      bin/workloads/micro/$workload/prepare/prepare.sh
42      if [[ $? -eq 0 ]]; then
43          break
44      fi
45      echo "    Tahap persiapan Hadoop gagal. Mencoba lagi←
46          ..."
47      done
48
49      # Looping untuk setiap pengulangan Hadoop
50      for ((i = 1; i <= repetitions; i++)); do
51          echo "    Percobaan Hadoop $i"
52
53          # Mulai dool di latar belakang
54          nohup /home/hadoop/bin/dool --all --io --output "←
55              $workload-$scale-$i-hadoop.csv" --bytes > /dev/←
56              null 2>&1 &
57          dool_pid=$! # Menyimpan PID proses dool
58
59          # Menjalankan benchmark Hadoop
60          bin/workloads/micro/$workload/hadoop/run.sh
61
62          # Menghentikan dool setelah benchmark selesai
63          kill $dool_pid
64          wait $dool_pid 2>/dev/null
65      done
66
67      # Tahap persiapan Spark
68      while true; do
69          bin/workloads/micro/$workload/prepare/prepare.sh
70          if [[ $? -eq 0 ]]; then
71              break
72          fi
73          echo "    Tahap persiapan Spark gagal. Mencoba lagi←
74              ..."
75      done
76
77      # Looping untuk setiap pengulangan Spark
78      for ((i = 1; i <= repetitions; i++)); do
79          echo "    Percobaan Spark $i"

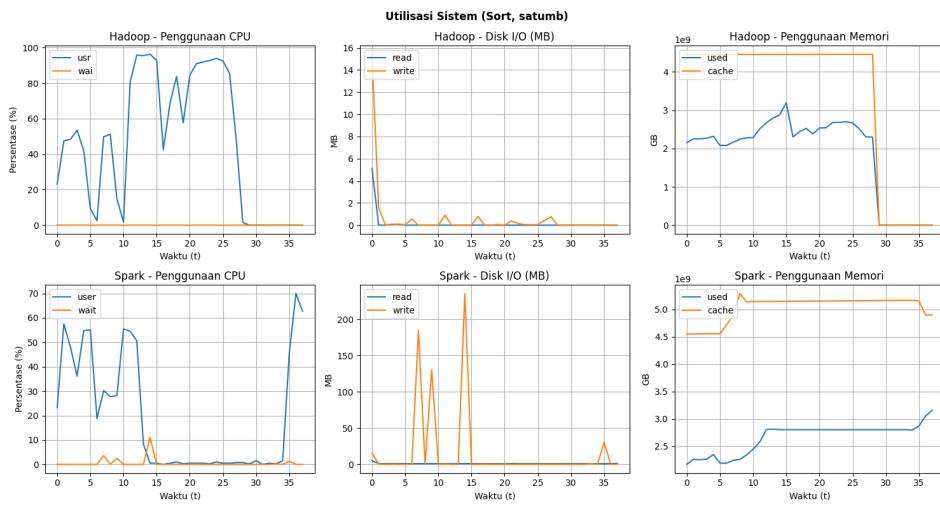
```

```
80
81      # Menjalankan benchmark Spark
82      bin/workloads/micro/$workload/spark/run.sh
83
84      # Menghentikan dool setelah benchmark selesai
85      kill $dool_pid
86      wait $dool_pid 2>/dev/null
87      done
88
89      # Tunggu 15 detik sebelum beralih ke skala berikutnya
90      sleep 15
91      done
92 done
93
94 echo "Selesai."
```

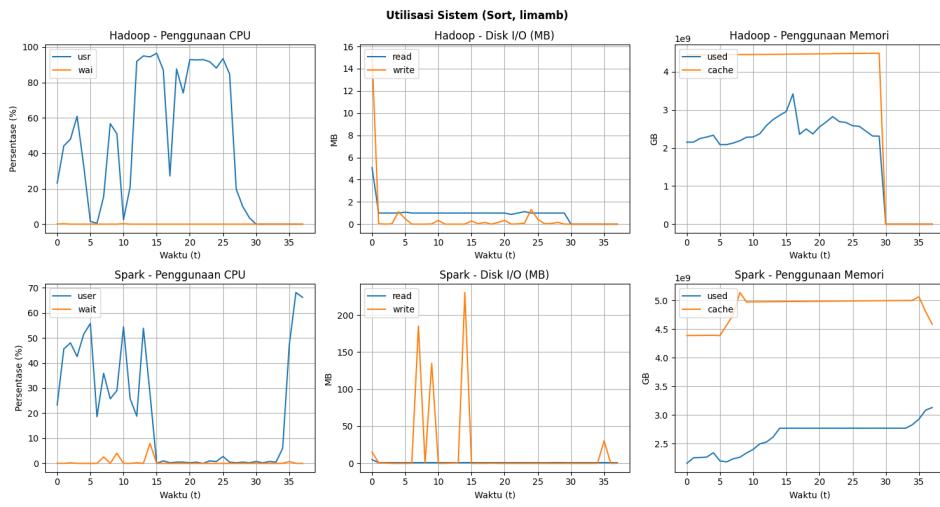
LAMPIRAN G

Visualisasi Utilisasi Sistem Sesuai Input Data (Sort)

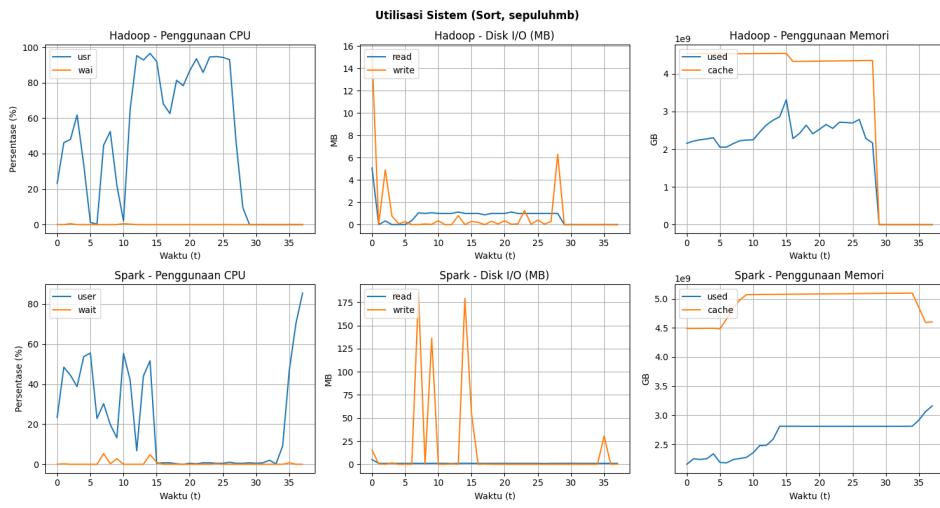




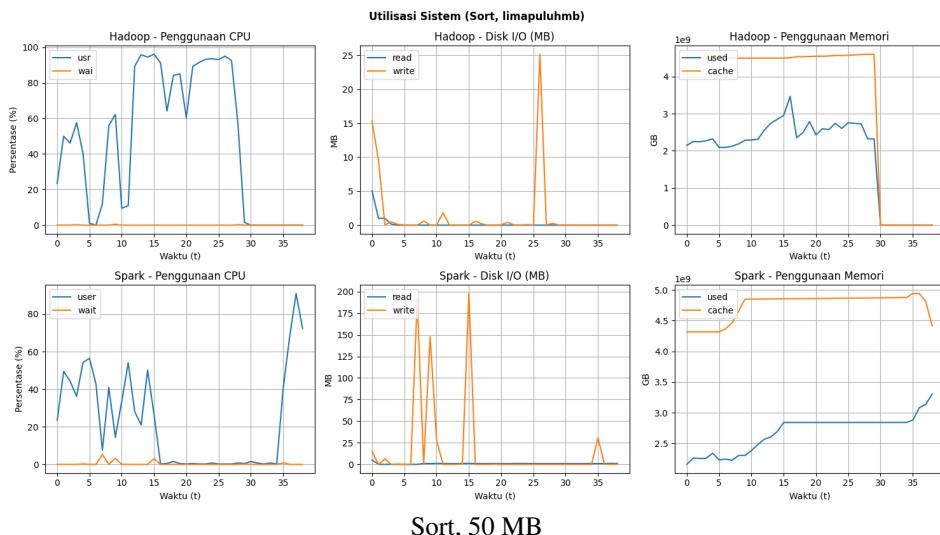
Sort, 1 MB



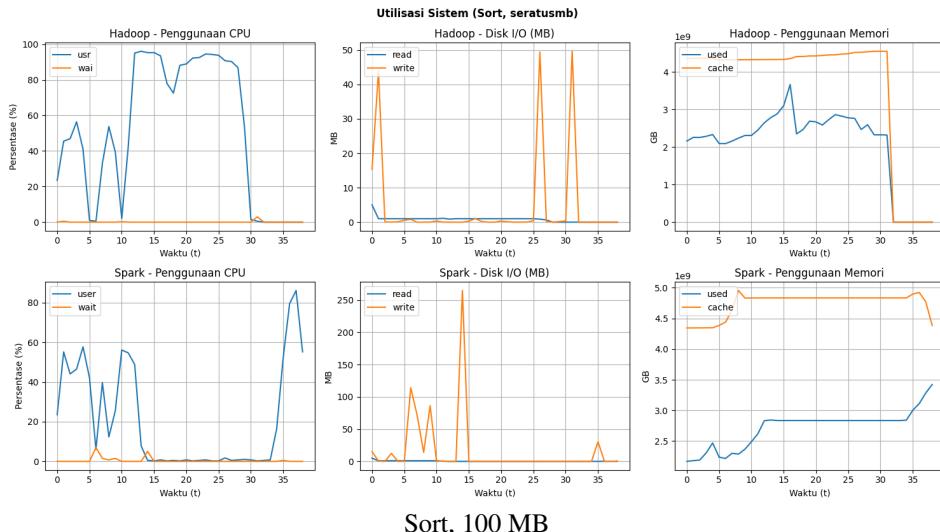
Sort, 5 MB



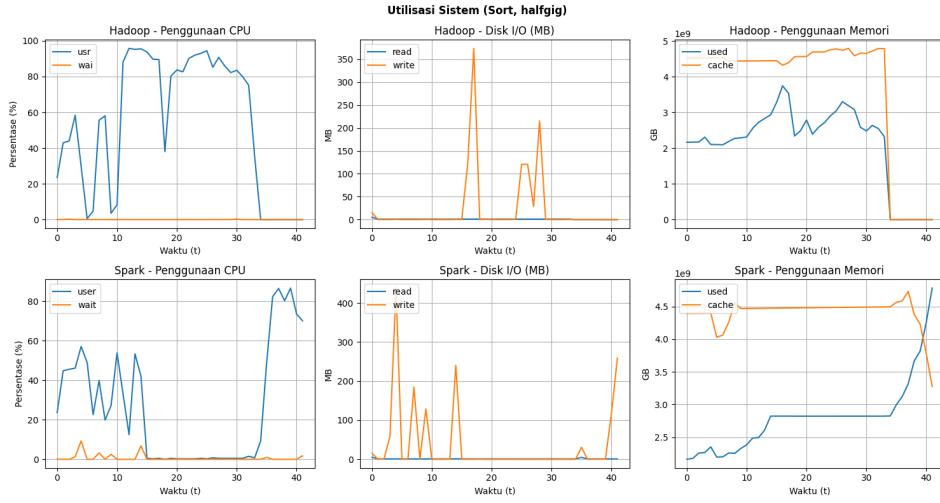
Sort, 10 MB



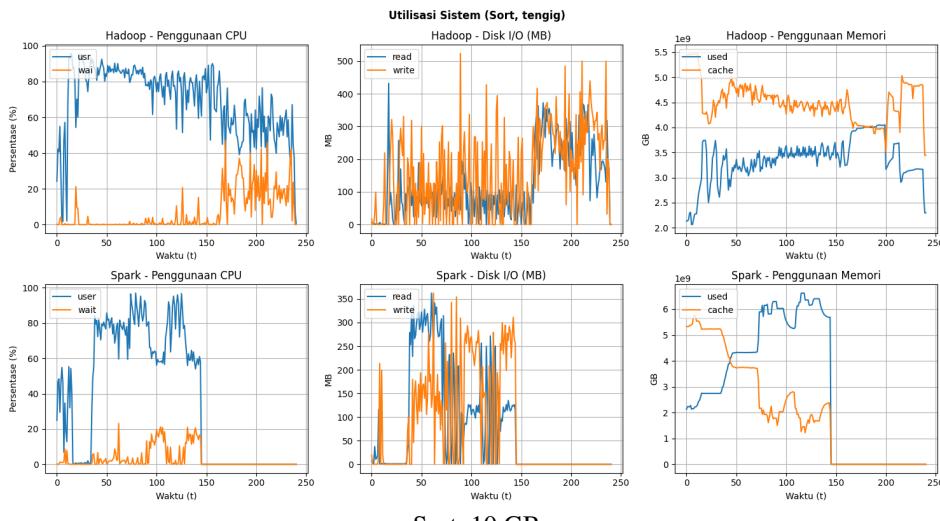
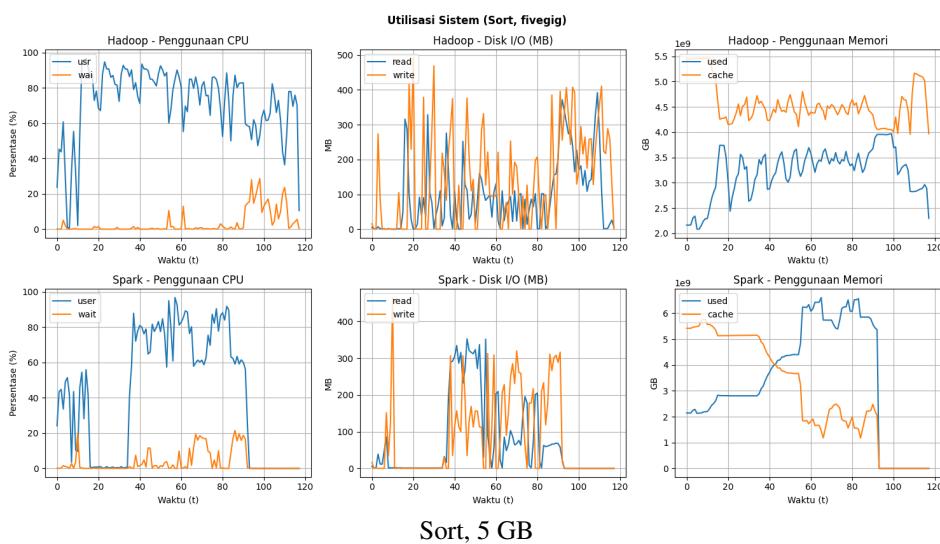
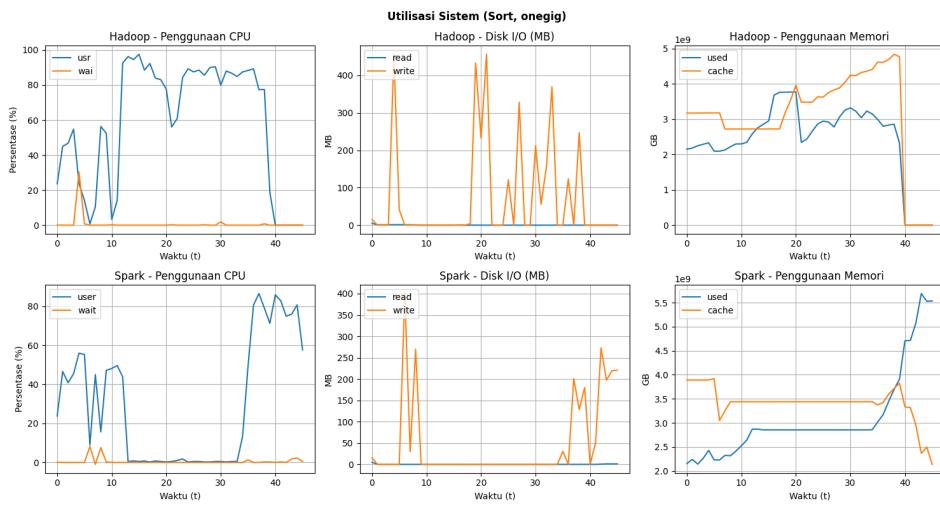
Sort, 50 MB

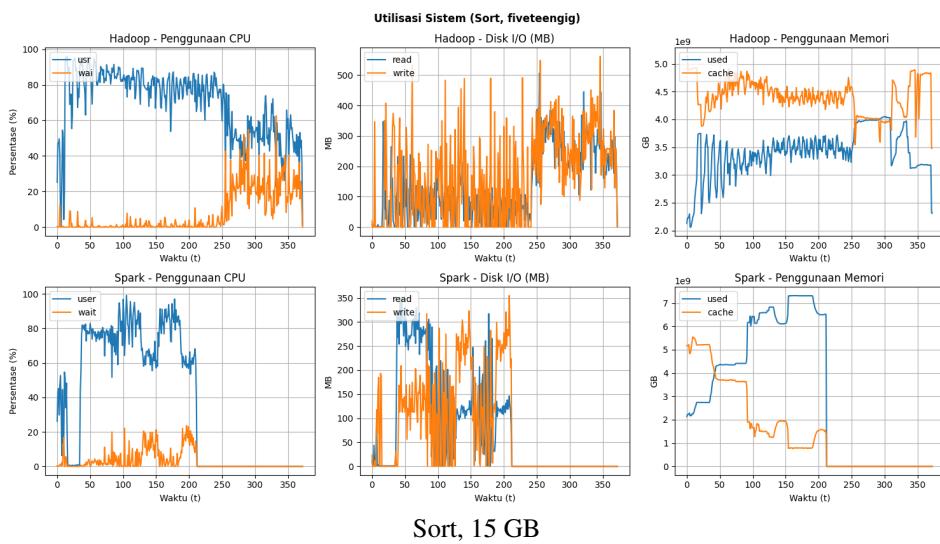


Sort, 100 MB



Sort, 500 MB

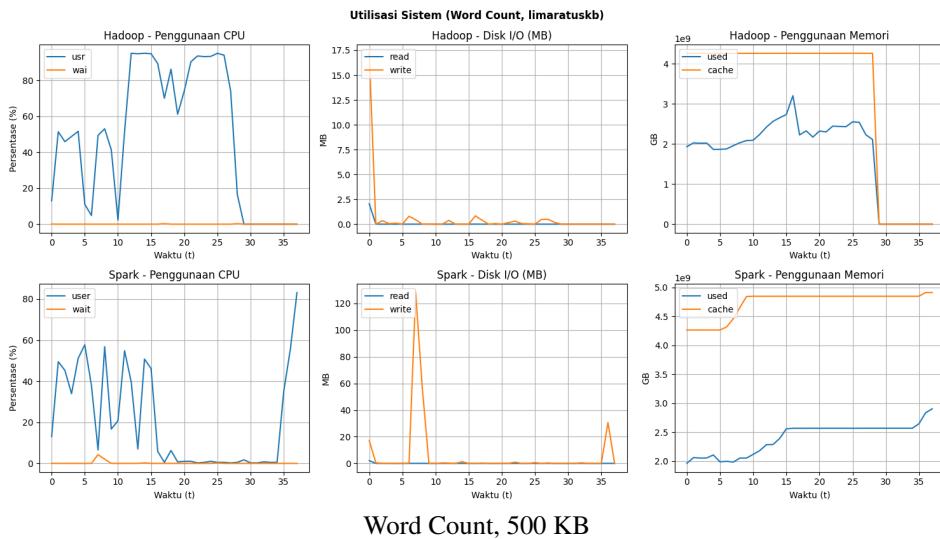
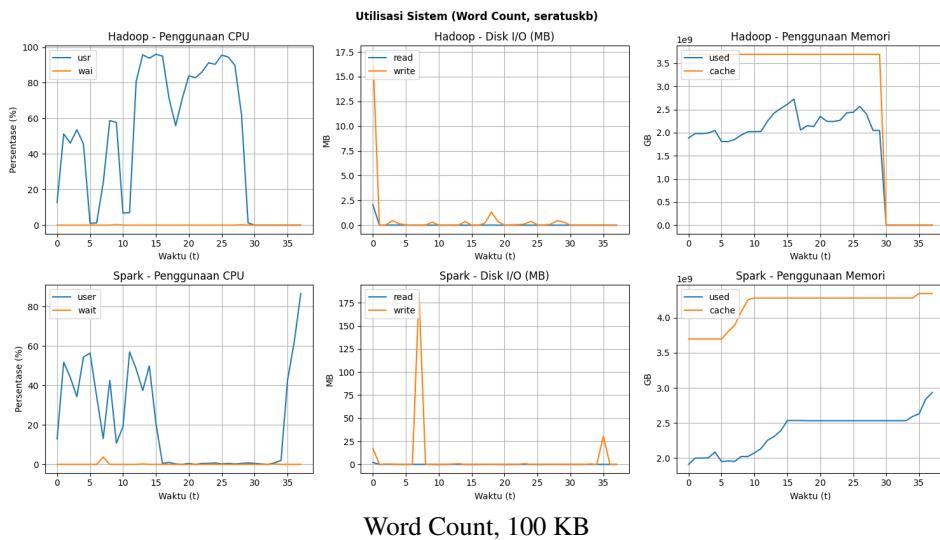


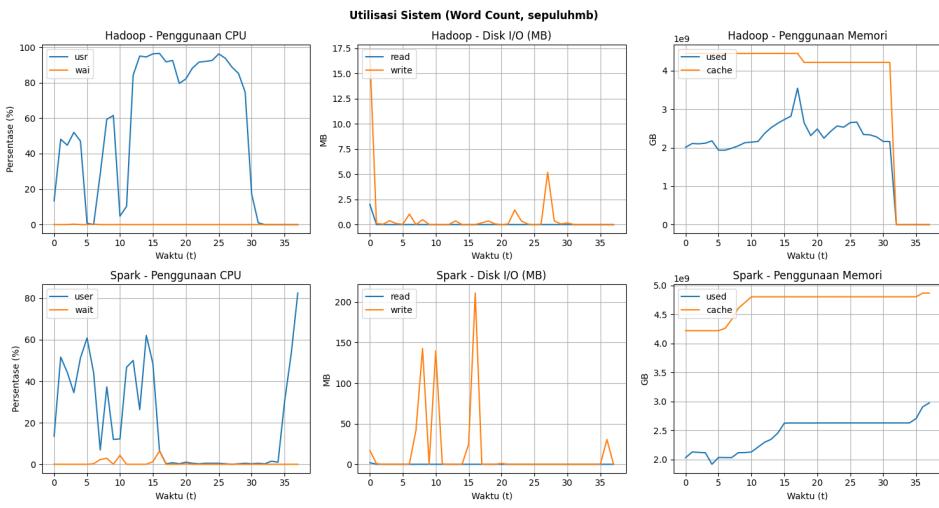
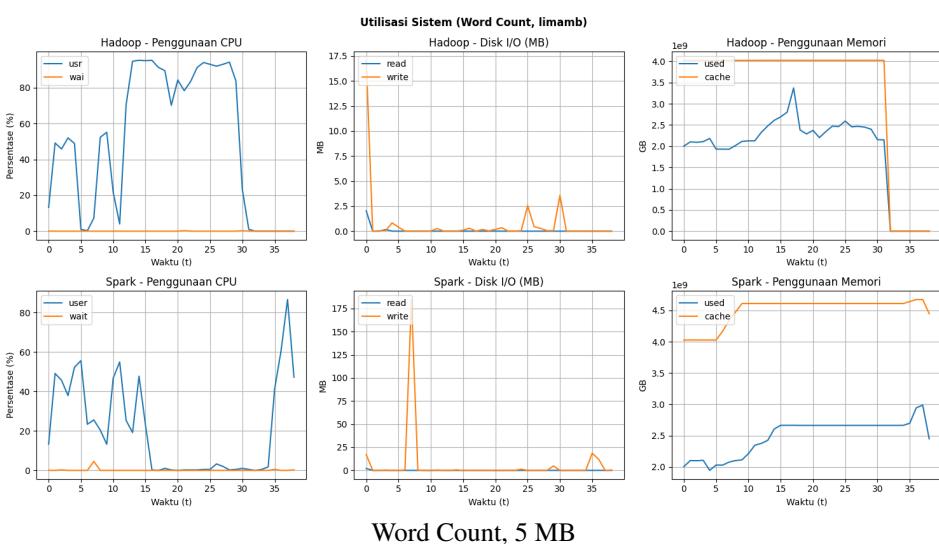
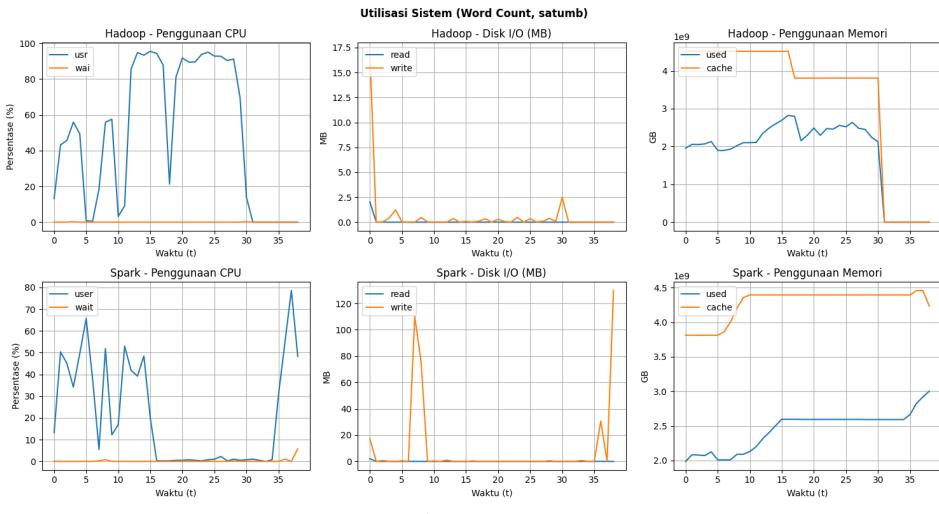


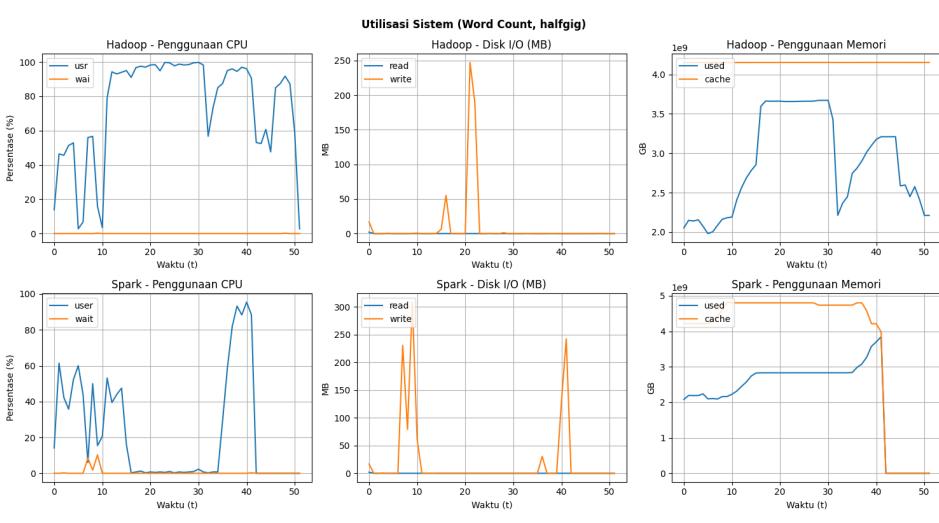
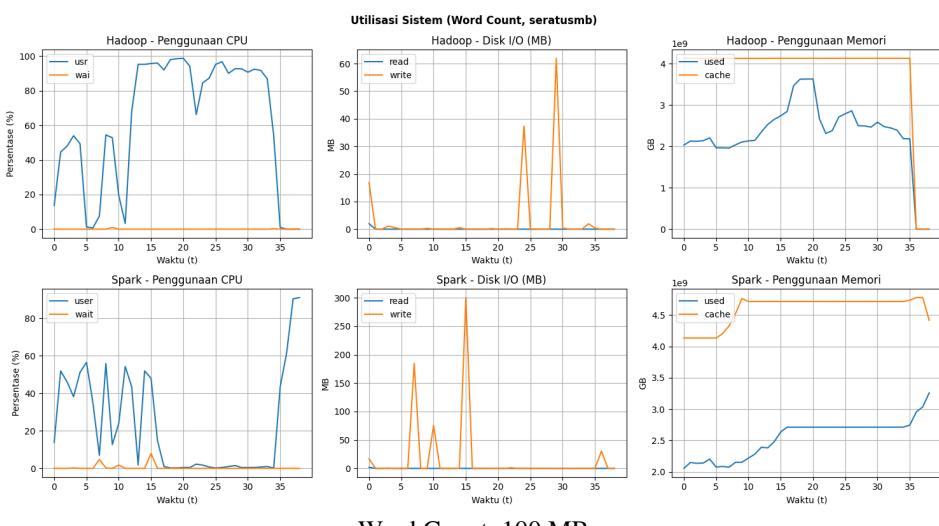
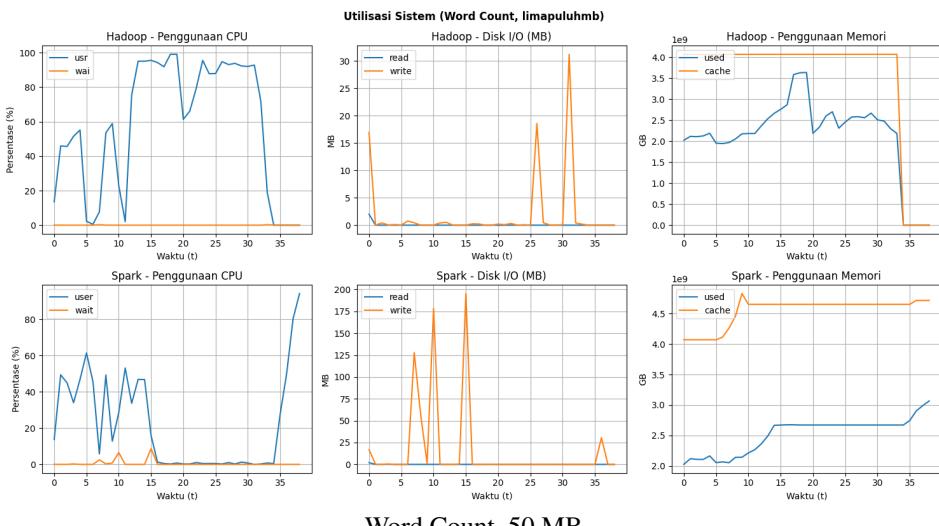
Sort, 15 GB

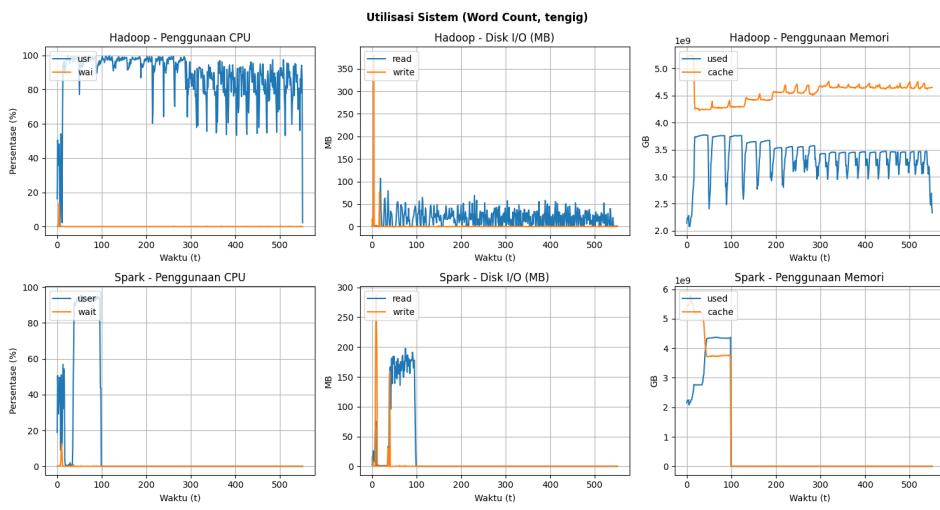
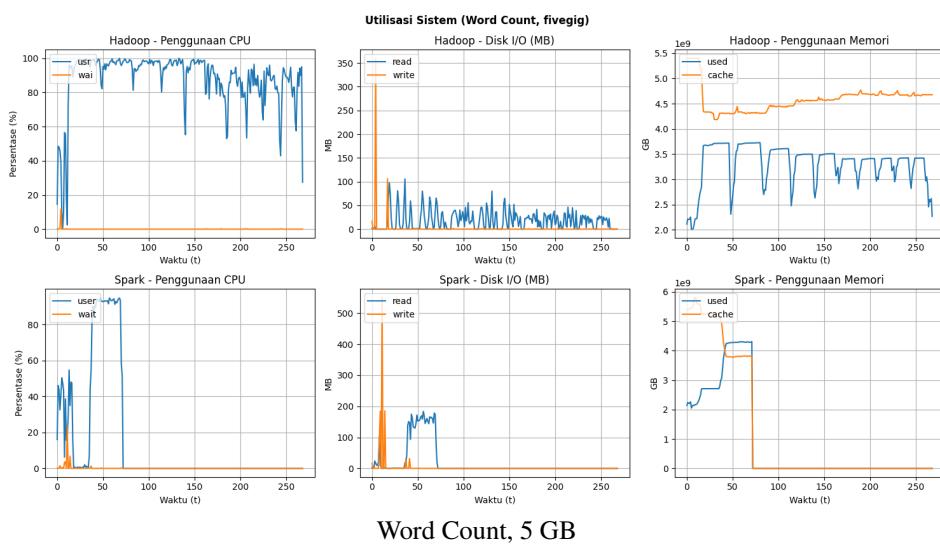
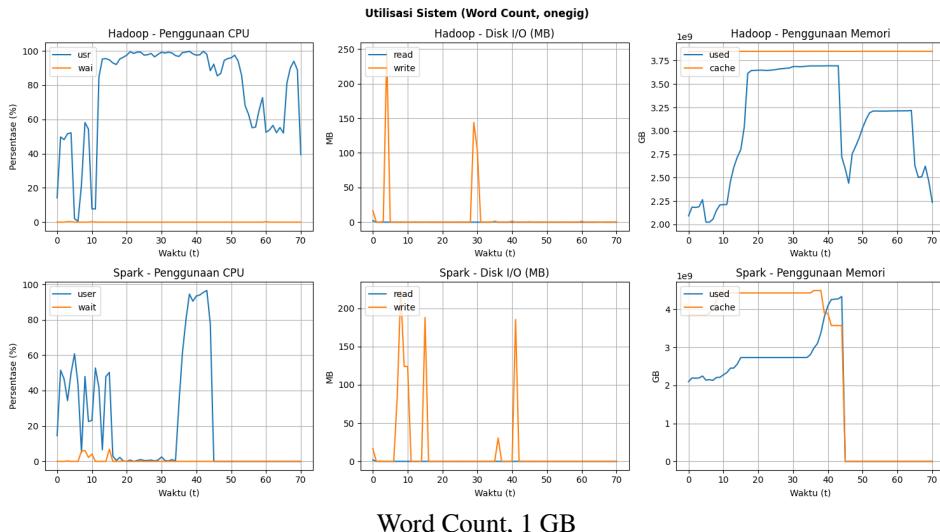
LAMPIRAN H

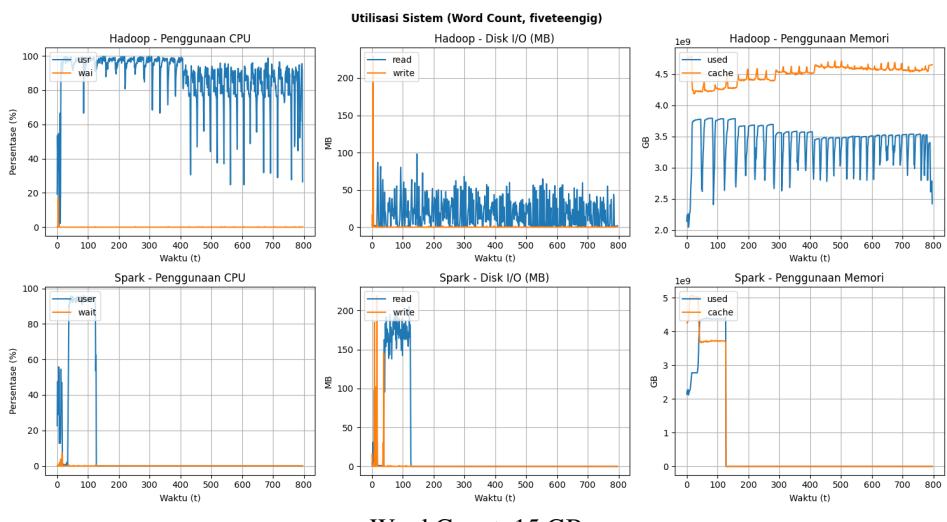
Visualisasi Utilisasi Sistem Sesuai Input Data (Word Count)











Word Count, 15 GB