Penerapan Sorted Wordcount Dengan Mapreduce Hadoop

Tiara Mutia Sari, Vijiantika Fajaria Sastri, M.Iqbal*, Fitri, Rahmadini, Tanti Annastaysa, Febry Afriansyah, M.Rafli Yanda, Fadzin Gumay

Sistem Komputer, Universitas Sriwijaya, Indralaya

*Korespondensi: muhammadiiqbal340357@gmail.com

ARTICLE INFO

Article History:
- Received 11 January 2023
- Received in revised form 21 January 2023
- Accepted 19 February 2023
- Available online 30 March 2023

ABSTRAK

Hadoop memfasilitasi pemrosesan kumpulan data besar secara terdistribusi dan menyediakan fondasi di mana layanan dan aplikasi lain dapat dibangun. MapReduce dan HDFS adalah dua komponen utama Hadoop. Hadoop MapReduce memproses sejumlah besar data terstruktur dan tidak terstruktur yang disimpan dalam HDFS. MapReduce adalah kerangka kerja Hadoop dan model pemrograman untuk memproses data besar menggunakan paralelisme dan distribusi otomatis di ekosistem Hadoop. Salah satu program yang menggunakan konsep MapReduce yang disediakan oleh Hadoop adalah Wordcount. Wordcount merupakan program yang bertujuan untuk menghitung kata pada file plaintext. Proses MapReduce Wordcount ini dibagi menjadi 2 tahap, yaitu proses mapping dan reducing. Wordcount adalah contoh tipikal di mana pengembangan dengan konsep MapReduce dimulai dari Hadoop itu sendiri. Proses reducing pada Wordcount dimasukkan untuk menghitung jumlah kemunculan setiap kata dalam file input yang disediakan. Pada implementasi kali ini, hasil output yang diurutkan tidak hanya berdasarkan kata-kata tetapi juga berdasarkan frekuensi kemunculan kunci.

Kata Kunci: MapReduce, Wordcount, Hadoop, HDFS, Paralel

ABSTRACT

Hadoop facilitates the processing of large datasets in a distributed manner and provides a foundation on which other services and applications can be built. MapReduce and HDFS are the two main components of Hadoop. Hadoop MapReduce processes a substantial amount of structured and unstructured data stored in HDFS. MapReduce is the framework and programming model of Hadoop designed for processing large datasets using automatic parallelism and distribution in the Hadoop ecosystem. One program that utilizes the MapReduce concept provided by Hadoop is Wordcount, a program aimed at counting words in plaintext files. The Wordcount MapReduce process is divided into two stages: mapping and reducing. Wordcount is a typical example where development with the MapReduce concept starts from Hadoop. The reducing process in Wordcount is included to count the occurrences of each word in the provided input file. In this implementation, the sorted output results are based on words and the frequency of critical occurrences.

Keywords: MapReduce, Wordcount, Hadoop, HDFS, Parallel

1. PENDAHULUAN

Perkembangan teknologi yang pesat menyebabkan pertumbuhan data meningkat secara eksponensial. Data yang dihasilkan tersebut jika diolah dapat menghasilkan sebuah informasi yang nantinya dapat digunakan lebih lanjut sesuai kebutuhan. Hadoop menyediakan sistem file terdistribusi dan menyediakan sebuah framework yang bernama MapReduce untuk mengolah dan menganalisis data set yang besar. Karakteristik utama dari Hadoop adalah cara kerjanya yaitu mempartisi data dan melakukan komputasi diberbagaikomputer host yang tergabung didalam clustersehinggamembuat pengolahandata menjadilebih cepat dibandingkan dengan distribusi file secarakonvensional, seperti RDBMS.

Hadoop didesain agardapat berjalan padaperangkat komoditas, yang berarti cluster Hadoop tidakharus dibangun pada perangkat yang mahal dan hanya pada satu vendor saja. Hadoop dapat dibangun pada perangkt standar yang umum dipasaran dan dapat berasal dari berbagai vendor. Solusi untuk menghadapi data dengan skala yang besar tidak lagihanya dengan membangun server yang semakin besar, tetapi dengan menggabungkan low-end/ perangkat komoditas sebagai sebuah sistem terdistribusi yang fungsional seperti Hadoop[1]. MapReduce adalah paradigmapemrograman yang memungkinkan skalabilitas masif di ratusan atau ribuan server di cluster Hadoop. Sebagaikomponenpemrosesan, MapReduce adalah jantung dari Apache Hadoop. Istilah "MapReduce" mengacupada dua tugas terpisah dan berbeda yang dilakukan oleh program Hadoop. Yang pertama adalahpekerjaan peta, yang mengambil satu set data dan mengubahnya menjadi satu set data lain, dimana masing-masing elemen dipecah menjadi tupel. Yang kedua adalah pekerjaan pengurangan output dari peta sebagai output dari peta sebagai input dan menggabungkantupel data tersebut menjadi satu settupelyang lebih kecil. Sesuaiurutan nama MapReduce, pekerjaan pengurangan selalu dilakukan setelah pekerjaan peta[2].

MapReduce WordCount digunakan sebagai program benchmarking (patokan) untuk menganalisa kinerja dari cluster hadoop yang dibuat pada penelitian. Tiga skenario pengujiandilakukan pada single-node cluster dan multi-nide cluster hadoop yang dibuat untuk melihat performa dari MapRedcude hadoop dalam mengolah sebuah file[3][4]. MapReduce Wordcount dipilih sebagai sarana brenchmarking karena program tersebut adalah salah satu program dasar yang terdapat pada Hadoop dan mudah untuk diimplementasikan serta diubah.

2. TINJAUAN PUSTAKA

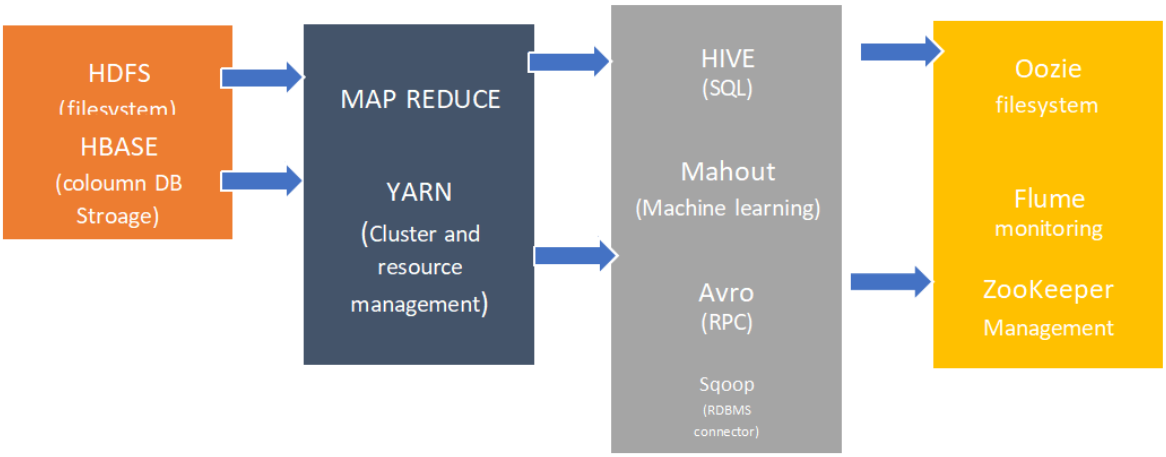
2.1 Wordcount

Word count adalah sebuah fungsi atauperintah yang digunakan untuk menghitung jumlahkata atau karakter dari sebuah dokumen. Kegunaan dari word count adalah untuk membantu Anda mengetahui seberapa panjang dokumenataukonten yang telahdiproduksi. Meskipun bukan faktor peringkat,namun word count cukup penting dalampenerapan SEO. Menulis konten yang lebih panjang membuat algoritma Google lebih mudah dalam menemukan informasi yang disampaikan dalam suatu halaman web.

Gambar 1. Word count

2.2 Hadoop

Hadoop adalah kerangka kerja open source yang dikembangkan oleh Apache dan ditulis dalam Java. Kerangka kerja ini memungkinkan pemrosesan terdistribusi pada kumpulan data besar melibatkan kelompok komputer menggunakan model pemrograman yang sederhana. Hadoop beroperasi di lingkungan yang menyediakan penyimpanan dan komputasi terdistribusi di seluruh kluster komputer. Didesain untuk meningkat dari satu server hingga ribuan mesin, setiap mesin menyediakan komputasi dan penyimpanan lokal.



Gambar 2. Haddop Framework[5][6][7][8]

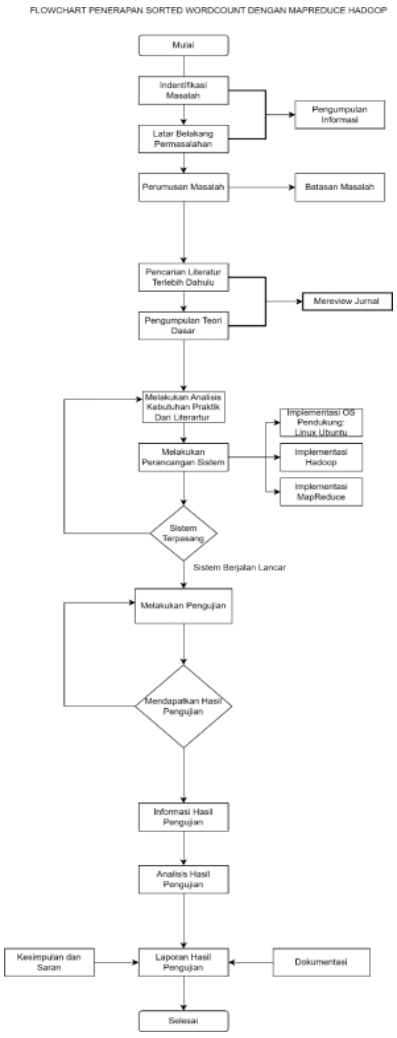
2.3 MapReduce

Hadoop Framework mencakup empat modul, yaitu Hadoop Common, Hadoop Yarn, HDFS, dan MapReduce. MapReduce adalah sebuah framework untuk aplikasi dan pemrograman yang diperkenalkan oleh Google dan digunakan untuk menjalankan pekerjaan komputasi terdistribusi pada sebuah cluster. MapReduce terdiri dari konsep fungsi map dan reduce, yang umumnya digunakan dalam functional programming. Salah satu program yang menggunakan konsep MapReduce yang disediakan oleh Hadoop adalah WordCount. WordCount merupakan program yang bertujuan untuk menghitung kata pada file plaintext. Proses MapReduce pada WordCount dibagi menjadi 2 tahap, yaitu proses mapping dan reducing. Pada tahap map, data input yang umumnya berupa file yang tersimpan dalam HDFS diubah menjadi tuple, yakni pasangan antara key dan value-nya. Pada tahap reduce, data input dari hasil proses map diproses, kemudian dilakukan tahap shuffle dan reduce, dan hasil dataset baru disimpan kembali di HDFS[9]-[15].

3. METODELOGI PENELITIAN

3.1. Kerangka kerja penelitian

Metode yang digunakan untuk menimplementasikan Word counting menggunakan hadoop ecosystem ditunjukan pada gambar berikut ini. Flowchart penerapan Sorted Wordcount dengan MapReduce Hadoop memiliki manfaat signifikan dalam memvisualisasikan langkah-langkah implementasi secara sistematis. Dengan memberikan gambaran yang jelas, flowchart ini tidak hanya membantu pengguna memahami proses dari awal hingga akhir, tetapi juga mendukung perencanaan dan rancangan yang efektif. Keuntungan lainnya mencakup kemampuan flowchart untuk mendeteksi potensi kesalahan atau hambatan dalam proses implementasi, memungkinkan perbaikan proaktif. Flowchart juga berperan sebagai alat pelatihan dan pendidikan, memberikan panduan yang efektif bagi mereka yang ingin memahami atau mempelajari implementasi Sorted Wordcount dengan MapReduce Hadoop. Selain itu, sebagai alat komunikasi tim, flowchart memastikan pemahaman seragam di antara anggota tim proyek, meningkatkan efisiensi dan mengurangi risiko kesalahan selama implementasi. Secara keseluruhan, flowchart ini menjadi aset penting untuk memastikan efisiensi dan keberhasilan penerapan Sorted Wordcount dengan MapReduce Hadoop.



Gambar 3. Flowchart penelitian

3.2. VirtualBox

VirtualBox digunakan sebagai mesin komputer pada percobaan ini. VirtualBox adalah perangkat lunak virtualisasi yang memungkinkan pengguna menginstal sistem operasi tambahan di dalam sistem operasi utama. Virtualisasi mengubah atau mengkonversi sesuatu menjadi bentuk simulasi dari bentuk nyata. Sebagai contoh, jika seseorang telah menginstal sistem operasi Windows pada komputer mereka, mereka dapat menjalankan sistem operasi lain di dalam Windows menggunakan VirtualBox. Manfaatnya termasuk kemampuan untuk mencoba atau belajar menginstal sistem operasi tanpa perlu menginstal ulang PC/laptop. VirtualBox memungkinkan instalasi sistem operasi tambahan tanpa mengganggu sistem operasi utama, memungkinkan pengguna mencoba dan mempelajari sistem operasi baru tanpa merubah data pada hard disk utama. Keuntungan lainnya adalah kemampuan untuk menginstal beberapa sistem operasi secara gratis tanpa memerlukan perangkat keras tambahan.

3.3. Sistem Operasi: Ubuntu 3.3.2

Ubuntu adalah sistem operasi berbasis Linux yang dikembangkan secara reguler dengan rilis update stabil dan berkualitas. Ubuntu 3.3.2 adalah salah satu versi dari sistem operasi ini. Ubuntu merupakan sistem operasi lengkap berbasis Linux, tersedia secara bebas, dan mendapatkan dukungan dari komunitas dan ahli profesional. Ubuntu populer untuk berbagai keperluan, mulai dari menjelajah internet dan tugas kantor hingga pemrograman dan pengelolaan server.

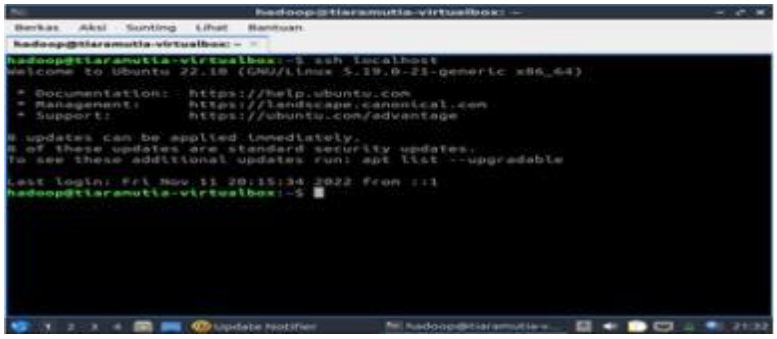
3.4. Bahasa Pemrograman: Python 3

Python 3 adalah bahasa pemrograman tujuan umum yang bersifat interpretatif. Python 3.0, dirilis pada tahun 2008, merupakan revisi utama dari Python sebelumnya yang tidak sepenuhnya kompatibel. Banyak kode Python 2 memerlukan modifikasi agar dapat berjalan di Python 3.

3.5. Proses Instalasi Hadoop

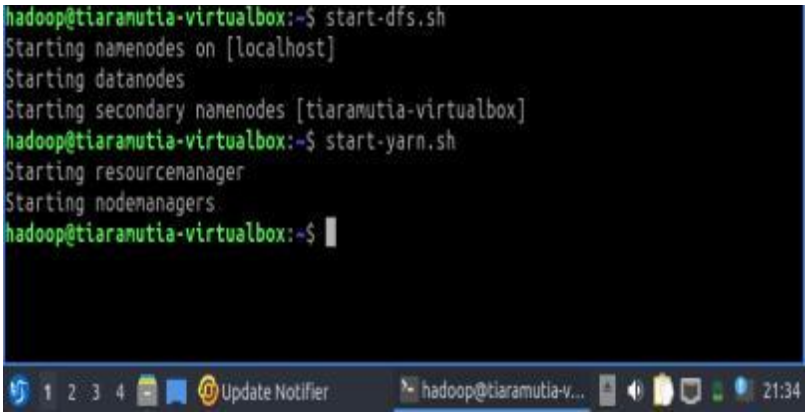
Sebelum melakukan percobaan terhadap program, proses instalasi Hadoop diperlukan. Setelah berhasil diinstal, langkah selanjutnya adalah mencoba membuat dan menjalankan contoh aplikasi MapReduce "WordCount". Langkah-langkahnya adalah sebagai berikut:

1. Langkah pertama Masukke ssh localhost



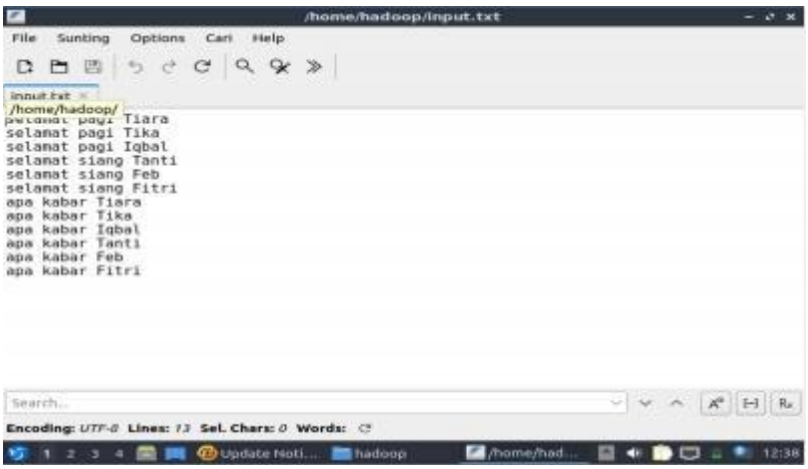
Gambar 4. Akses ssh

2. Kemudian start dfs dan yarn



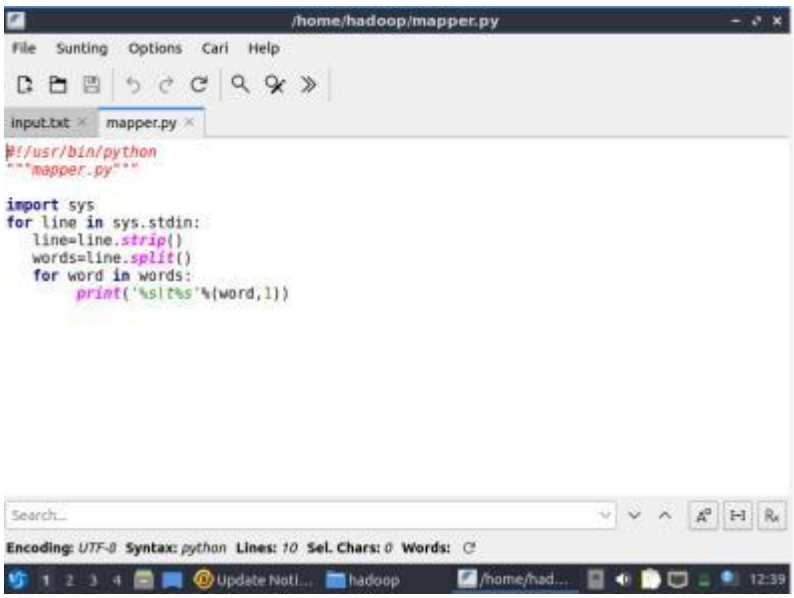
Gambar 5. Menjalankan DFS dan YARN

3. Buat text dinote kemudian save dengan nama input.txt



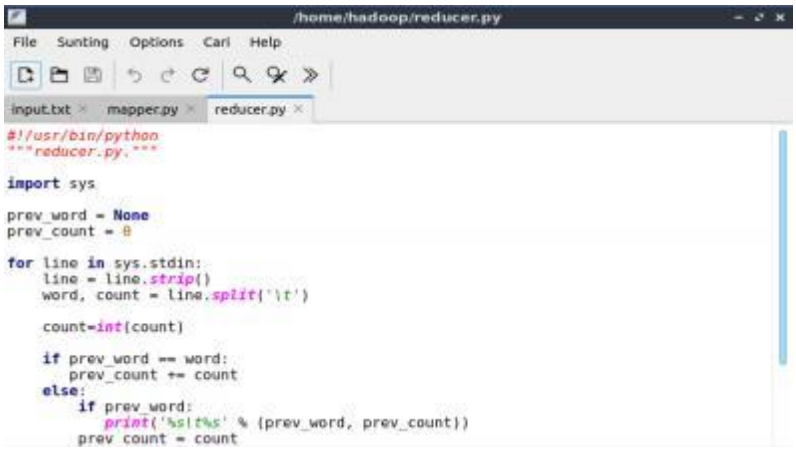
Gambar 6. Input.txt

4. Buat program python3 untuk mapper kemudian save file jadi map



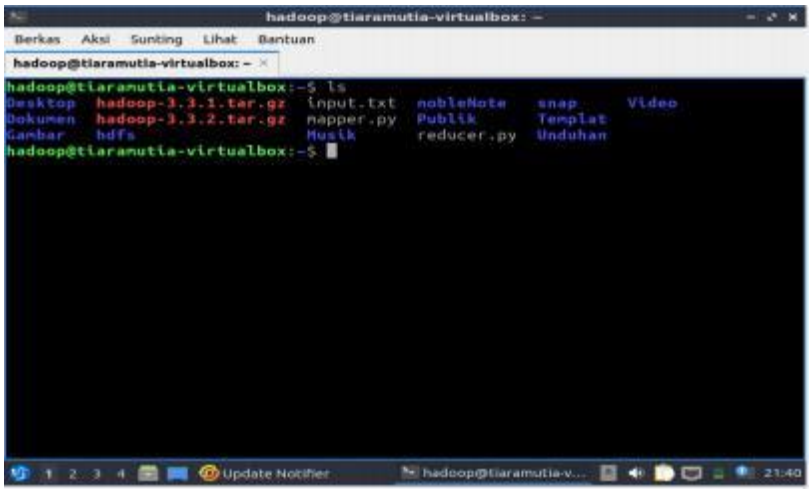
Gambar 7. Fungsi Map

5. Buat program python3 untuk reducer kemudian save file jadi reducer.py



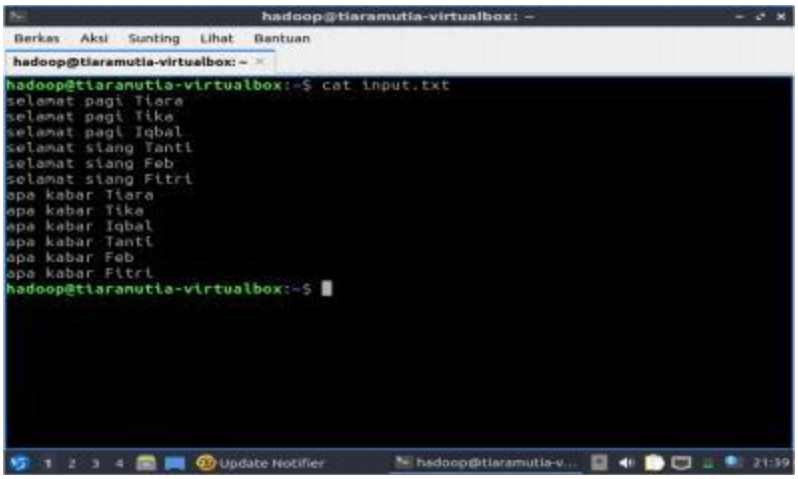
Gambar 8. Fungsi Reduce

6. Check terlebih dahulu dimana kita menyimpan ke 3 file tersebut dan kita menyimpan di /home/Hadoop.



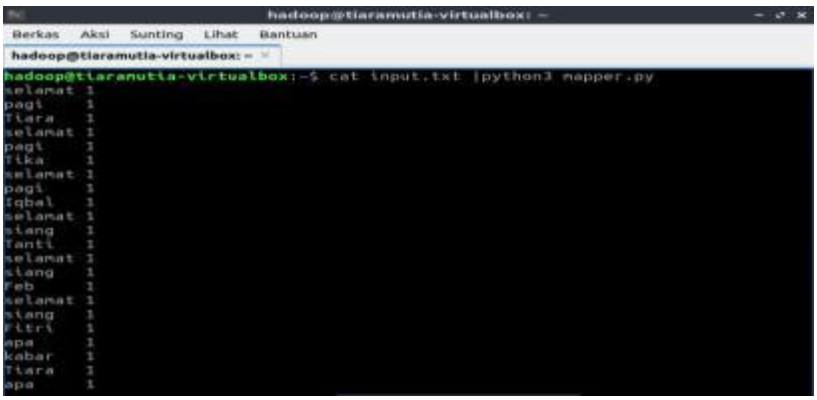
Gambar 9. Home Hadoop

7. Setelah itu Panggil file txt pada terminal Hadoop (cat input.txt)



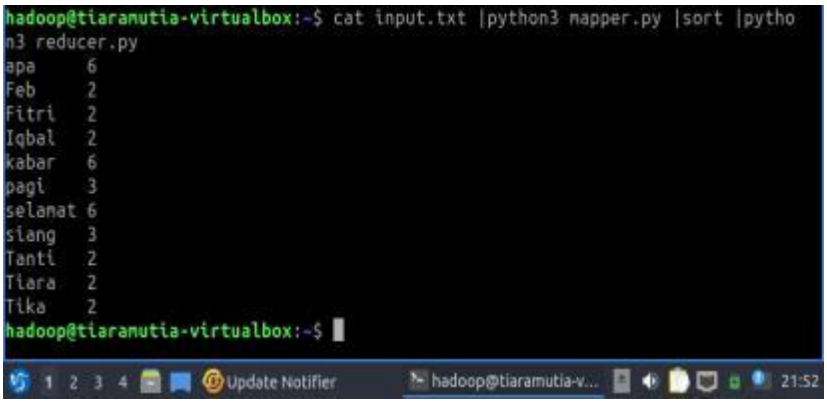
Gambar 10. Melihat isi file input.txt

8. Kemudian mapping dari data input txt tersebut data yang telah dipecah akan di proses untuk menghasilkan intermediary key-value pairs.(cat input.txt |pythton3 mapper.py)



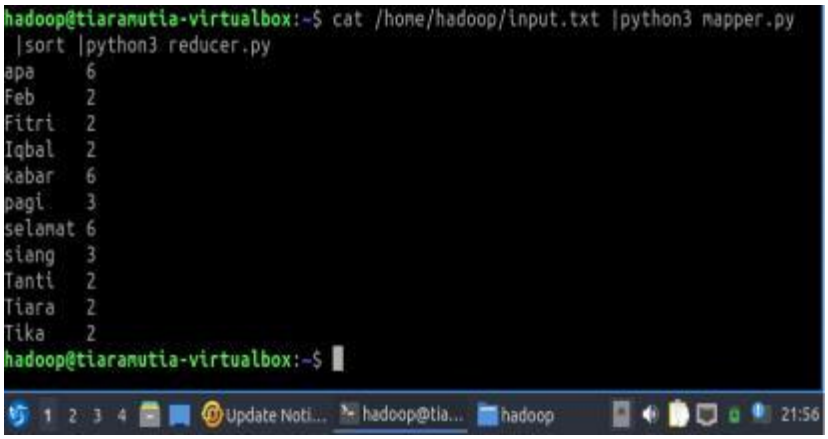
Gambar 11. Proses mapping data

9. Kemudian lakukan masuk ketahapan reducer, yang mana bertugasuntuk melakukan aggregasiterhadap seluruhpasangan intermediary key-value dengan key yang sama.



Gambar 12. Tahapan reducer

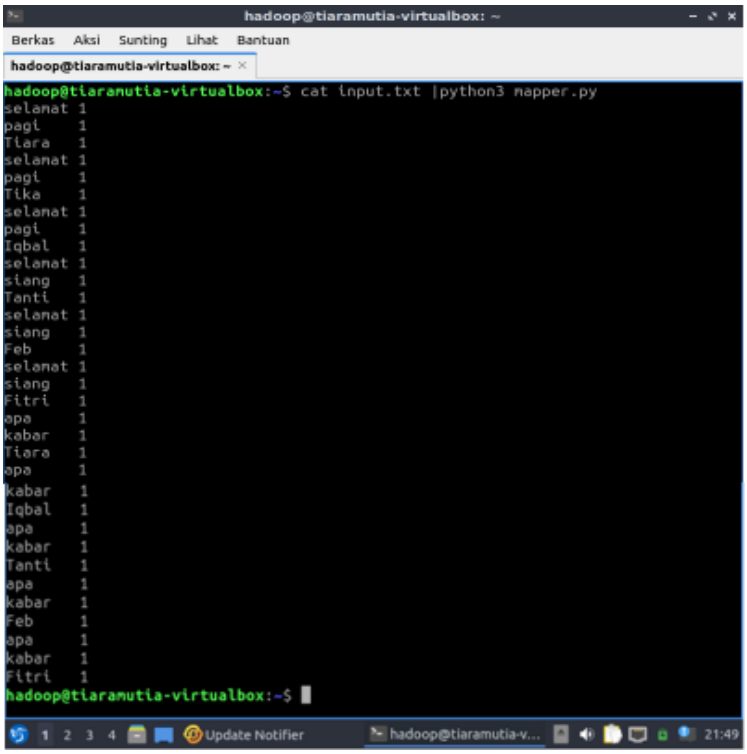
10. Kemudian kita dapat lihat data dari mapreduce yang telah dibuat pada directory HadoopDengan



Gambar 13. Proses Mapreduce

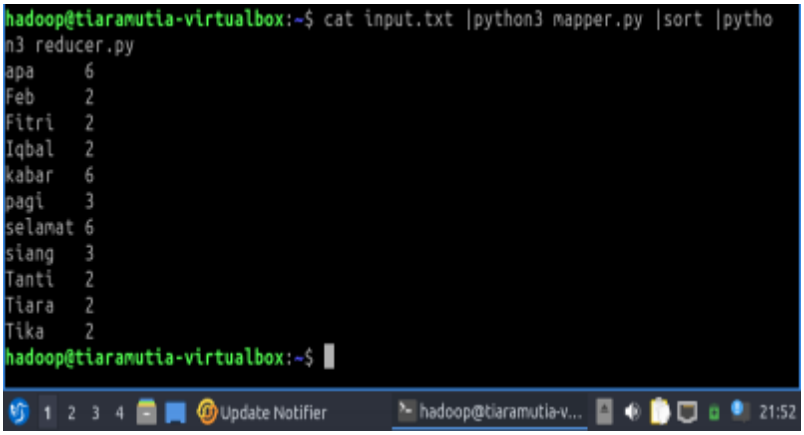
4. HASIL PENELITIAN

Untuk menghasilkan output dari program word count, dilakukan proses mapping dan reducing. Data input diproses dan dibagi menjadi segmen-segmen yang lebih kecil pada fase mapper, di mana jumlah mapper sama dengan jumlah input split. Mapper akan memproses key-value ini menggunakan logika pengkodean untuk menghasilkan output dengan bentuk yang serupa. Proses mapping dapat diilustrasikan melalui gambar di bawah.



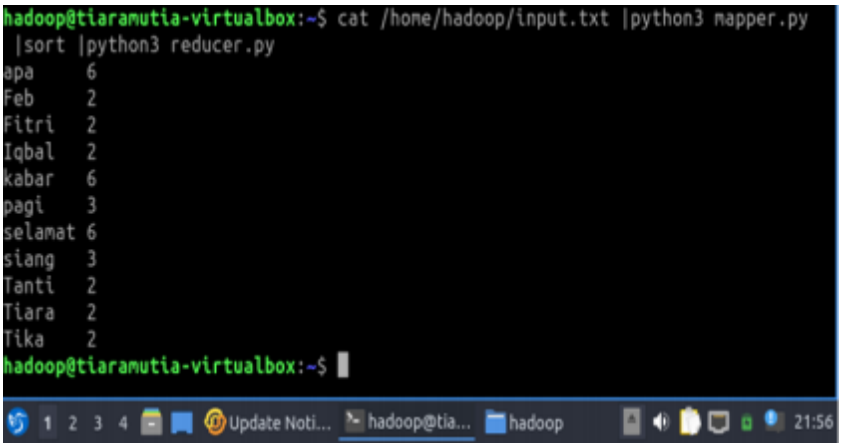
Gambar 14. Output word count

Kemudian dilanjutkan dengan proses reduce. Pada fase reduce, nilai antara dari fase pengocokan direduksi untuk menghasilkan satu nilai output yang merangkum seluruh dataset. Berikut adalah contoh gambaran proses dari reduce.



Gambar 15. Proses reduce

Setelah implementasi program MapReduce untuk word count berhasil dijalankan, maka akan dihasilkan output word count dari MapReduce berdasarkan file data input yang telah dibuat. Berikut adalah hasil dari percobaan tersebut.



Gambar 16. Hasil percobaan

Seperti yang diketahui, setelah melewati tahap mapping dan reduce, program menghasilkan output seperti yang tergambar di atas. Output dari program tersebut adalah word count dari file data input, di mana kata-kata "apa", "kabar", dan "selamat" masing-masing memiliki jumlah 6. Sedangkan untuk kata-kata "Feb", "Fitri", "Iqbal", "Tanti", "Tiara", dan "Tika" masing-masing memiliki jumlah 2. Terakhir, kata-kata "pagi" dan "siang" memiliki jumlah masing-masing 3. Semua kata dalam file data input telah di-mapping, di-reduce, dan di-sortir sehingga masing-masing hitungannya terlihat.

5. KESIMPULAN

Secara definisi, Hadoop MapReduce dapat diartikan sebagai kerangka kerja perangkat lunak untuk menjalankan pekerjaan pemrosesan sejumlah besar data. Data input dibagi menjadi gugus-gugus independen, dan setiap gugus diproses secara paralel di seluruh simpul dalam kluster Anda. Berdasarkan program Penerapan Sorted WordCount dengan MapReduce Hadoop yang dijalankan sebelumnya, dapat disimpulkan bahwa program ini memiliki fungsi perhitungan jumlah kata, di mana MapReduce merupakan salah satu proses yang terintegrasi dalam Hadoop itu sendiri. Sebagai contoh, hasil program terlihat pada gambar terakhir di atas, di mana hasil percobaan menunjukkan bahwa kata-kata "apa", "kabar", dan "selamat" masing-masing memiliki jumlah 6. Sedangkan kata-kata "Feb", "Fitri", "Iqbal", "Tanti", "Tiara", dan "Tika" memiliki jumlah 2. Terakhir, kata-kata "pagi" dan "siang" memiliki jumlah masing-masing 3. Semua kata dalam file data input telah di-mapping, di-reduce, dan di-sortir sehingga masing-masing hitungannya terlihat.

DAFTAR PUSTAKA

[1]. Bangari, K., Meduri, S., Rao, C.: Implementation of word count - Hadoop framework with map reduce algorithm. Int. J. Comput. Trends Technol. 49(3), 179– 182 (2017)
[2]. Zhao, D.: Performance comparison between Hadoop and HAMR under laboratory environment. Procedia Comput. Sci. 111, 223–229 (2017).
[3]. Gohil, P., Garg, D., Panchal, B.: A performance analysis of MapReduce applications on Big Data in cloud based Hadoop. In: International Conference on Information Communication and Embedded Systems, pp. 1 – 6. Institute of Electrical and Electronics Engineers, Chennai (2014)
[4]. Li Y, Zhang H, Kim KH (2011) A power-aware scheduling of MapReduce applications in the cloud. In: 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC). IEEE
[5]. Memishi B, Pérez MS, Antoniu G (2017) Failure detector abstractions for MapReduce-based systems. Inf Sci 379:112– 127
[6]. Fu H et al (2017) FARMS: efficient MapReduce speculation for failure recovery in short jobs. Parallel Comput 61:68–82
[7]. Jiang Y et al (2017) Makespan minimization for MapReduce systems with different servers. Future Gener Comput Syst 67:13–21
[8]. Chen Q, Liu C, Xiao Z (2013) Improving MapReduce performance using smart speculative execution strategy. Parallel Distrib Syst 24:1107
[9]. Nanduri R et al (2011) Job aware scheduling algorithm for MapReduce framework. In: 2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom). IEEE
[10]. Tiwari N et al (2015) Classification framework of MapReduce scheduling algorithms. ACM ComputSurv (CSUR) 47(3):49
[11]. White T., 2012. Hadoop: The Definition Guide, 3rd Edition. 2¶5HLOO
[12]. K. Basuki, H. Palit, and L. Dewi, “Implementasi hadoop: Studi kasus pengolahan data peminjaman perpustakaan universitas kristen petra,” Jurnal Infra, vol. 3, no. 2, pp. 226–232.

A. S. Foundation, “Apache hadoop,” available:. [Online]. Available:https://hadoop.apache.org/

[13]. Mohd Rehan Ghazi, D. G. (2015). Hadoop, MapReduce and HDFS: A Developers Perspective. International Conference on Intelligent Computing, Communication & Convergence.

[14]. Mishra, B. (2020). Big Data Analysis Using Hadoop Map Reduce. International Research Journal of Computer Science, 07(05), 114– 122. https://doi.org/10.26562/irjcs.2020.v0705.005