Pengolahan Dataset Diagnosa Kesehatan Mata Menggunakan Hadoop Pada YARN Framework

Aldi Hoirul Fatih*, Edo Pratama, Febiyona Melista Br Tarigan, M. Azrell Samudra, Putri Resti Ningsih
Sistem Komputer, Universitas Negeri Sriwijaya, Indonesia

*Korespondensi: putrirestiningsih@gmail.com

ARTICLE INFO

Article History:
- Received 13 July 2023
- Received in revised form 28 August 2023
- Accepted 20 September 2023
- Available online 30 October 2023

ABSTRAK

Big data merujuk pada dataset dengan volume yang besar dan kompleksitas tinggi, mencakup aspek ukuran, jumlah, jaringan, dan lainnya. Karena variasi dan volume yang signifikan, pengelolaan big data memerlukan teknik dan alat khusus. Dalam penelitian ini, kami akan menggunakan dataset diagnostik kesehatan untuk melakukan analisis wordcount dengan menggunakan alat Hadoop pada YARN dan MapReduce. Selain itu, kami akan memanfaatkan Apache Flink sebagai platform untuk memproses data secara paralel.

Kata Kunci: Big Data, Hadoop, MapReduce, Wordcount, YARN, Apache, Flink.

ABSTRACT

Big data refers to datasets that are large and complex in terms of size, quantity, networks, and other aspects. Due to the significant variations and volume, big data requires special techniques and tools for processing. In this paper, we will use a diagnostic health dataset to perform word count analysis using Hadoop tools on YARN and MapReduce. Additionally, we will utilize Apache Flink as one of the platforms for processing data in parallel.

Keywords: Big Data, Hadoop, MapReduce, Wordcount, YARN, Apache, Flink.

1. PENDAHULUAN

Kata-kata “Big Data” sudah tidak asing lagi di masa sekarang ini. Segala bentuk informasi dengan mudah didapatkan saat ini hanya dengan melalui akses internet. Bahkan, ketertarikan mengenai big data terus berkembang secara eksponensial sejak 2011. Namun informasi dari sebuah big data tidak langsung serta-merta didapatkan. Sumber data banyak sekali didapatkan melalui survey-survey singkat yang didapatkan di Internet, Postingan Facebook, Twitter, Youtube, Blog Pribadi dan lainnya yang selalu bertambah setiap hari. Postingan tersebut dapat dikumpulkan dalam suatu kueri yang disebut sebagai ‘big data’. Big Data memiliki ciri-ciri: memiliki volume yang besar, variasi yang banyak, dan datanya bersifat dinamis (cepat berubah).

Big data dengan segala ciri-cirinya itu dapat menjadi sebuah potensi ataupun tantangan bagi yang ingin memperoleh informasi dari dalamnya. Salah satunya ialah tools, perangkat, atau teknologi yang digunakan untuk mengolah data tersebut. Semakin tepat dan efektif teknologi yang digunakan, semakin baik output yang akan dihasilkan dari data tersebut. Selain itu, keterampilan seorang individu juga menjadi faktor penentu seberapa kualitas output yang dihasilkan agar dapat menjadi sebuah informasi dan wawasan yang berguna. Pengelelolaan sebuah big data tentunya harus dapat dilakukan dengan cepat tanpa mengesampingkan kualitas yang dihasilkan. Dengan tuntutan demikianlah, maka big data akan lebih baik diolah dengan

menggunakan metode terdistribusi atau paralel. Dengan demikian lahirlah istilah pemrosesan paralel.

Dengan terus berkembangnya teknologi hingga masa sekarang, tentunya ada banyak sekali tools yang dapat digunakan untuk mengolah data seperti MapReduce, ZohoAnalytics, MongoDB, Cassandra, dan lainnya. Tujuan tools tadi sama, yaitu untuk dapat menghasilkan insight yang actionable dari sebuah data. MapReduce merupakan salah satu model pemrograman populer yang digunakan untuk memproses suatu data secara paralel pada sebuah cluster. Model pemrograman MapReduce dapat ditemui pada software Hadoop. Dalam kasus kami, kami mencoba melakukan pengujian dataset Diagnosa Kesehatan Mata dengan menggunakan multimode pada dua sistem operasi dan menggunakan Hadoop-Mapreduce pada Framework YARN.

2. TINJAUAN PUSTAKA

2.1 Big Data

Big Data merupakan istilah yang merujuk pada sebuah kumpulan informasi yang tidak dapat diolah atau dianalisis hanya dnegan menggunakan perangkat yang konvensional. Istilah ini diberikan pada kumpulan data yang sangat besar dan kompleks, sehingga tidak memungkinkan diproses dengan cara yang konvensional. Dalam Gartner IT Glossary, Big Data didefinisikan sebagai berikut: Big Data is *high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enchance isight, decision making, and process automation.*

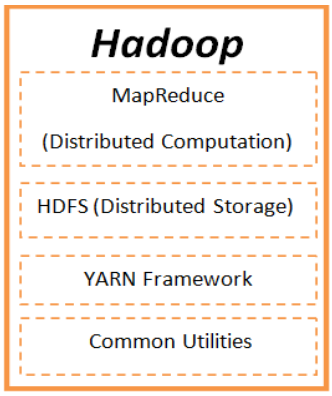
Dari beberapa definisi di atas, dapat diketahui bahwa Big Data memiliki ciri-ciri *high-volume (*volume besar), *high-velocity* (kecepatan yang tinggi) dan/atau *high-variety* (variasi yang banyak). Big data dapat menjadi sebuah tantangan bagi yang ingin mengolahnya, tantangan yang didapatkan biasanya berbentuk:

a. Ukuran: Volume yang dimiliki oleh suatu dataset tentu sangat menjadi faktor kritikal mengapa ia harus diolah secara khusus.
b. Kompleksitas: Perlu diketahui, struktur, permutasi, dan pola yang dimiliki oleh big data sangat banyak dan besar.
c. Teknologi: Big data membutuhkan sebuah tools dan metode yang tepat untuk dapat diproses sesuai dengan kebutuhan dataset yang digunakan.[1]

2.2 Hadoop

Hadoop merupakan sebuah framework yang bersifat open source yang dapat ditulis dalam bahasa pemrograman Java yang memungkinkan pemrosesan terdistribusi dari sebuah kumpulan data besar kelompok dengan menggunakan model pemrograman sederhana. Hadoop memiliki empat modul yang berkerja di dalamnya yaitu:

a. Hadoop Common: Merupakan sebuah Pustaka Java yang diperlukan oleh modul Hadoop yang lainnya. Pustaka-pustaka ini menyediakan file sistem dan abstraksi pada tingkat Sistem Operasi dan skrip yang diperlukan untuk memulai Hadoop.
b. Hadoop YARN: Merupakan sebuah framework yang berkerja pada tingkat penjadwalan pekerjaan dan pengelolaan sumber daya cluster.[2]
c. HDFS: Merupakan sebuah sistem file terdistribusi yang menyediakan akses throughput tinggi ke suatu data pada aplikasi.[3]
d. MapReduce: Merupakan sistem berbasis YARN untuk memproses data secara paralel dari kumpulan data yang besar.

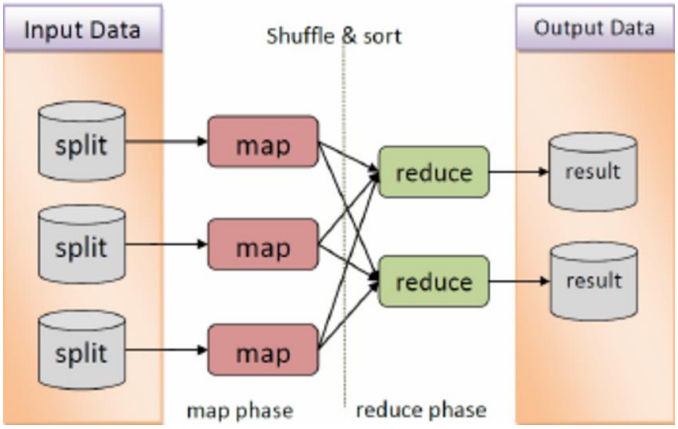


Gambar 1: Modul yang terdapat pada Hadoop

2.3 MapReduce

Adalah sebuah model pemrograman yang menjalankan sebuah komputasi yang digunakan untuk melakukan komputasi paralel pada dataset berukuran besar.[4]. MapReduce dirancang untuk dijalankan pada sebuah cluster. Komputasi MapReduce memiliki dua tahap inti yaitu *Map* dan *Reduce*. Input yang digunakan ke dalamnya dinotasikan dalam bentuk *key* dan *value*.

a. Proses 'Map': Sebuah proses dimana node master menerima sebuah input data, kemudian input tersebut dibagi-bagi menjadi beberapa bagian yang didistribusikan ke *node worker*. *Node worker* inilah yang akan memperoses inputan tersebut yang kemudian akan dikembalikan lagi ke *node master*.
b. Proses 'Reduce': Merupakan sebuah proses ketika *node master* menerima respon balik dari *node worker*. Respon tersebut akan digabungkan menjadi satu jawaban besar untuk mendapatkan solusi dari permasalahan utama.[5]



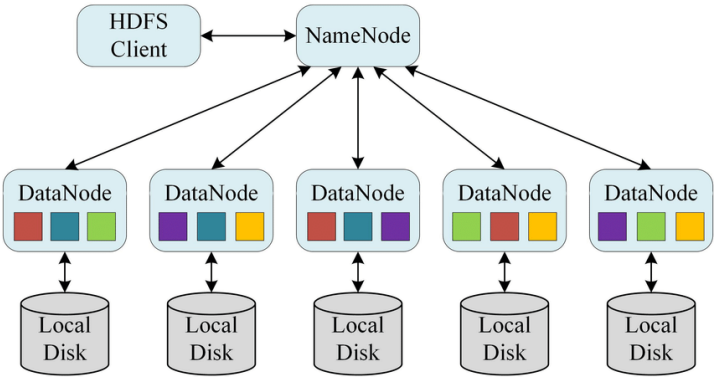
Gambar 2: Peta sistem cara kerja MapReduce
(sumber: ResearchGate)

2.4 YARN

YARN (*Yet Another Resource Negotiator)* merupakan sebuah framework yang mengatur jadwal pengerjaan yang diterapkan pada MapReduce serta *resource management* pada sebuah cluster.[6]

2.5 HDFS

HDFS (*Hadoop Distributed File System*) merupakan sebuah komponen penyimpanan di dalam Hadoop. HDFS memiliki model yang hampir serupa dengan *Google File System* (GFS). Untuk menampung sejumlah data yang memiliki ukuran yang besar, maka HDFS menggunakan blok file yang lebih besar dari *file system* pada umumnya yang dioptimalkan untuk mengurangi beban input/output pada sebuah jaringan. [4]



Gambar 3: Peta environment HDFS[4]

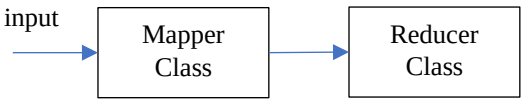
3. METODELOGI PENELITIAN

3.1 Model Pemrograman

Model pemrograman yang digunakan pada pengujian dataset diagnosa kesehatan mata kali ini ialah dengan menggunakan MapReduce. MapReduce sendiri memiliki 3 tahapan yaitu tahap map, shuffle, dan yang terakhir ialah tahap *reduce*. Untuk tahap *shuffle* dan *reduce* digabung menjadi satu tahapan utama yaitu tahapan *reduce*. Model pemrograman ini dilakukan dengan 2 fungsi yaitu fungsiMap dan fungsi Reduce. Kedua fungsi ini memiliki keyword <key, value> sebagai gambaran antara input dan outputnya. Proses tahapan MapReduce ialah sebagai berikut:

a. Pada tahap map, sistem akan memproses data inputan yang biasanya merupakan sebuah file yang tersimpan di dalam HDFS. Oleh sistem, maka inputan tersebut akan dipasangkan antara key dan value-nya <key, value> dalam sebuah tuple.
b. Pada tahap reduce, tuple-tuple yang telah dihasilkan akan melalui proses *shuffle* dan *reduce* yang outputnya akan kembali disimpan dalam HDFS.

Kedua tahapan tersebut, jika divisualisasikan secara sederhana akan menampilkan gambar berikut:



Gambar 4: Kelas-kelas pada MapReduce

3.2 Perancangan dan Implementasi Sistem

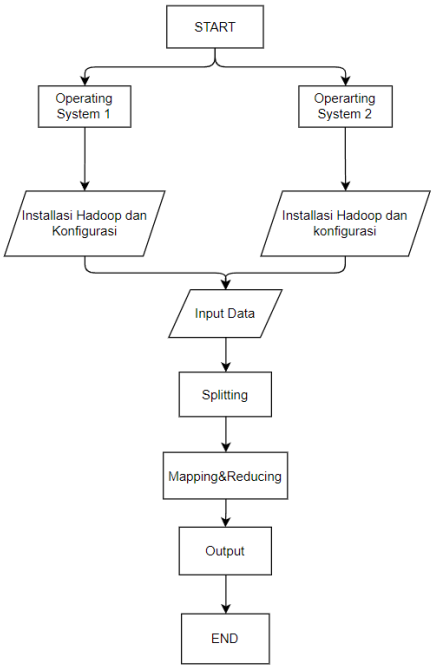
a. Rancangan Umum

Pada penelitian kali ini, perancangan sistem akan dimulai dengan melakukan installasi dan konfigurasi software Hadoop pada dua sistem operasi Ubuntu. Setelah melakukan installasi Hadoop, maka dilanjutkan dengan

melakukan konfigurasi pada ssh-Localhost, HDFS, MapReduce, dan YARN dengan menggunakan terminal. Setelah melakukan konfigurasi dan Hadoop dapat terhubung dengan localhost, maka selanjutnya akan dilakukan percobaan untuk memasukkan file csv.

Pada saat pengujian, kami juga menggunakan bahasa pemrograman Python sebagai pendukung yang akan membantu sistem untuk melakukan word-count. Data selanjutnya akan dimasukkan dan melalui 2 proses yaitu proses mapper dan reducer. MapReduce akan berkerja sesuai dengan bantuan framework YARN yang bertugas untuk memanajemen resources yang tersedia. Sebelum diproses oleh MapReduce-YARN, data tersebut kami masukkan ke dalam HDFS.

Proses pengolahan data secara paralel dapat dilihat pada diagram alir berikut ini:



Gambar 5: Flowchart gambaran umum sistem

b. Tentang dataset yang digunakan.

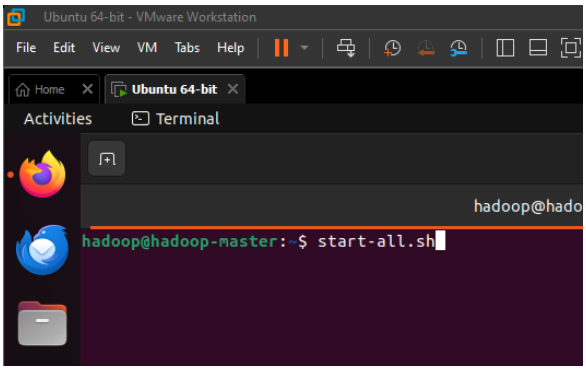
Pengujian kali ini akan dilakukan pada dataset Diagnosis kesehatan mata yang kami peroleh dari website Kaggle. Dataset yang kami gunakan memiliki karakteristik sebagai berikut:

1. Memiliki ukuran 2GB (csv+gambar)
2. Memiliki atribut ID, usia pasien, jenis kelamin pasien, Fundus mata kiri, Fundus mata kanan, Diagnosa mata kiri, Diagnosa mata kanan, Normal, Diabetes, Glaukoma, Katarak, Penyakit yang berhubungan dengan usia, Hipertensi, Myopia, dan Gangguan Lainnya.
3. Memiliki total 6392 baris dataset.

.

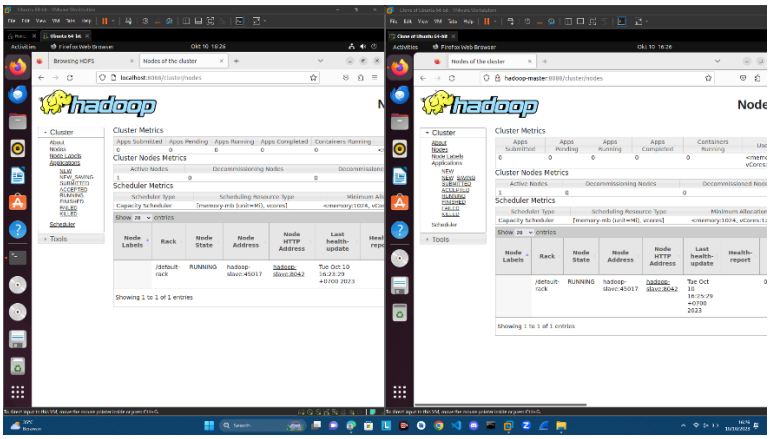
4. HASIL PENELITIAN

Sebelum menguji data, terlebih dahulu kami mengaktifkan sshlocalhost, dfs, dan yarn dengan menggunakan perintah start-all.sh. perhatikan gambar di bawah ini:



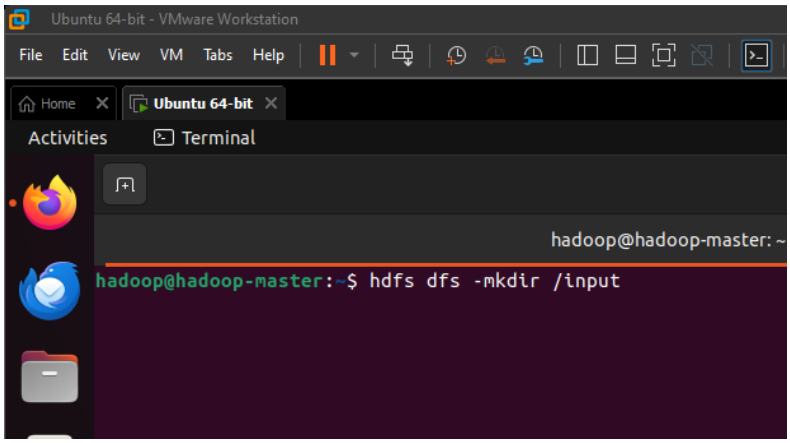
Gambar 6: Memulai YARN dan HDFS

Karena kami melakukan pemorsesan data pada 2 OS, maka akan muncul tampilan interface yang menunjukkan bahwa fungsi multinode pada kedua sistem operasi telah menyala. Perhatikan pada gambar 7.

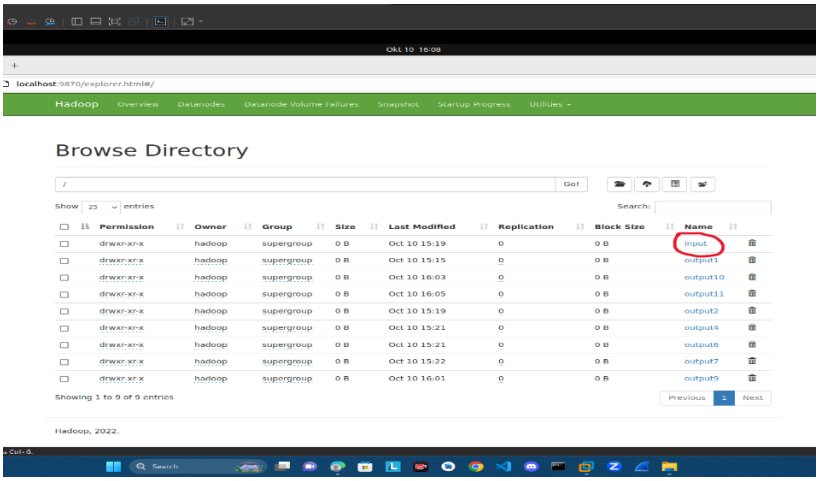


Gambar 7: Multinode pada Hadoop

Setelah masuk ke dalam sistem localhost Hadoop, selanjutnya kami akan membuat direktori input dengan menggunakan perintah hdfs dfs -mkdir /input. Gambar 9 merupakan gambar yang menujukkan bahwa direktori input telah dibuat.

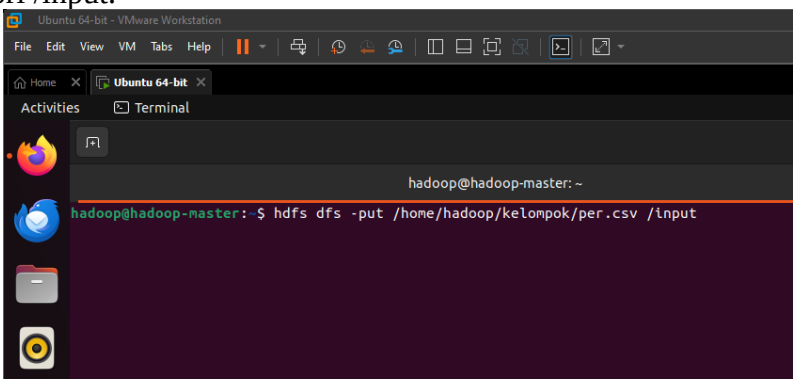


Gambar 8: Perintah membuat direktori

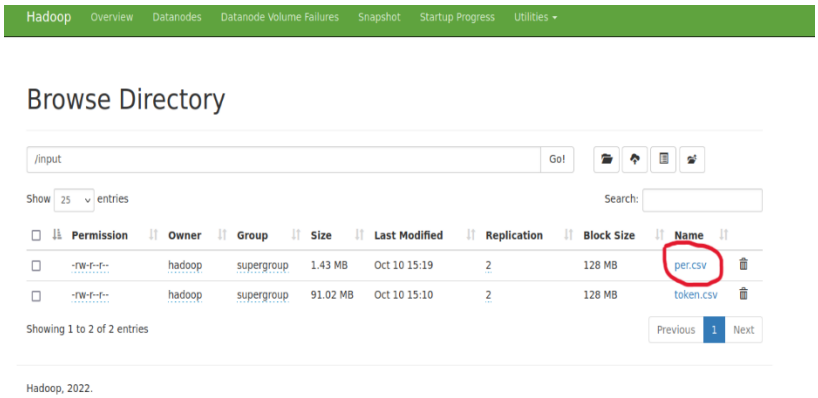


Gambar 9: Direktori yang dihasilkan

Setelah membuat direktori, langkah selanjutnya ialah memasukkan folder csv ke dalam direktori /input yang telah dibuat. Disini kami menggunakan perintah hdfs dfs -put /home/hadoop/kelompok/per.csv /input. Gambar 10 menunjukkan bahwa file csv telah berada di dalam direktori /input.

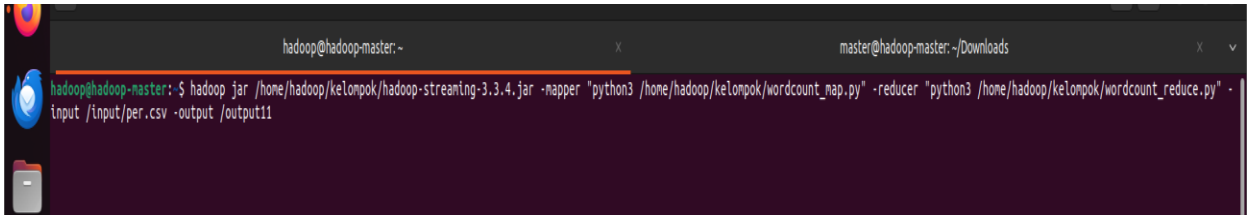


Gambar 10: Perintah memindahkan folder .csv ke dalam direktori input



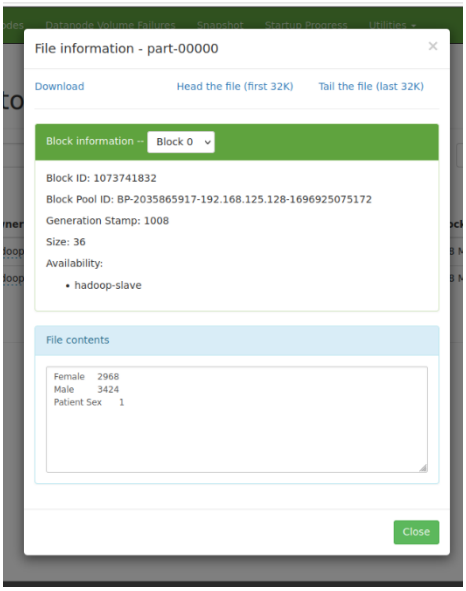
Gambar 11: Folder csv yang telah masuk direktori

Kami menggunakan bahasa pemrograman python yang membantu sistem untuk menentukan kata-kata seperti apakah yang dapat dihitung (misal, jika terdapat beberapa kata yang sama, maka ia akan dihitung sebagai satu kelompok kata yang sama, jika tidak, maka sistem akan menghitung kata tersebut dalam satu kelompok yang baru).



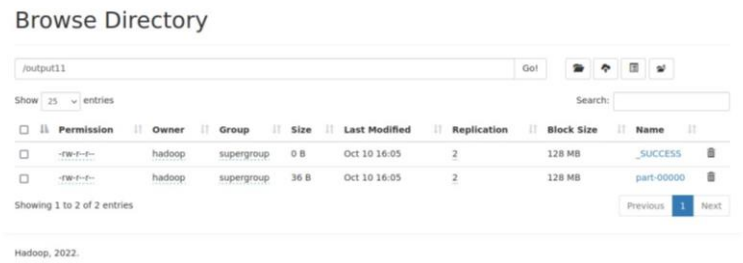
Gambar 12: Perintah mapper reducer

Setelah itu, MapReduce akan melakukan pemrosesan file secara paralel dan sesuai dengan yang telah ditentukan oleh pemrograman python yang telah dibuat sebelumnya, maka setiap kata yang ada di dalam file .csv akan dihitung jumlahnya dan menghasilkan output map-reduce seperti berikut:



Gambar 13: Hasil Mapper-Reduce yang memiliki output WordCount

Setelah keluar hasil wordcount-nya, maka selanjutnya output tersebut disimpan kembali ke dalam direktori input yang telah dibuat sebelumnya:



Gambar 14: Output yang disimpan di dalam direktori

5. KESIMPULAN

Big data merupakan suatu data yang memiliki volume yang besar, kecepatan perkembangan data yang tinggi, dan variasi yang besar. Pengolahan untuk big data tidak dapat dilakukan secara konvensional lagi, ia membutuhkan tools dan teknik yang khusus untuk dapat diekstraksi datanya. Salah satu tools yang dapat digunakan untuk mengekstraksi data yang besar adalah Hadoop. Hadoop akan menggunakan HDFS untuk membantu sistem agar dapat membagi file tersebut dalam sistem penyimpanan yang terdistribusi. Pengelolaan jadwal yang dilakukan di dalam sistem dilakukan oleh kerangka kerja YARN. Sistem penyortir suatu data dikerjakan oleh MapReduce.

Pada percobaan yang kami selesaikan, kami menerapkan modul YARN untuk melakukan pemrograman wordcount yang dilakukan untuk menghitung jumlah kata yang terdapat dalam file .csv yang telah disiapkan. Untuk melakukan wordcount, kami juga memerlukan bantuan multinode pada dua sistem operasi Ubuntu dan menerapkan MapReduce untuk melakukan penyortiran dan penjumlahan pada kata-kata yang serupa untuk selanjutnya dihitung.

DAFTAR PUSTAKA

[1] J. S. Ward and A. Barker, “Undefined By Data: A Survey of Big Data Definitions,” Sep. 2013, [Online]. Available: http://arxiv.org/abs/1309.5821

[2] Y. Surahman and H. Saptono, “Jurnal Informatika Terpadu EVALUASI KINERJA HDFS SEBAGAI INFRASTRUKTUR PEMBANGUNAN BIG DATA,” *Jurnal Informatika Terpadu*, vol. 4, no. 2, pp. 63–70, 2018, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/JIT

[3] C. Kurniawan, “A Survey on Big Data Analytics Model,” *ITEJ Juli-2019*, vol. 4, no. 1.

[4] H. Judul, “ANALISIS KINERJA HADOOP PADA CLUSTER RASPBERRY PI.”

[5] Y. Eko Testiono, “PENERAPAN HADOOP FRAMEWORK PADA PENGELOLAAN BIG DATA ARSIP NASIONAL RI.” [Online]. Available: http://journal.unpar.ac.id/../rekayasa/article/viewFile/1602/1530,

[6] P. Akas, T. Taqwin, A. B. Osmond, and R. Latuconsina, “IMPLEMENTASI METODE MAPREDUCE PADA BIG DATA BERBASIS HADOOP DISTRIBUTED FILE

SYSTEM IMPLEMENT OF MAPREDUCE METHOD ON BIG DATA BASED ON HADOOP DISTRIBUTED FILE SYSTEM."