Implementasi MapReduce Pada Dataset Spotify Top Music Untuk Mengetahui Artist yang paling banyak Didengar Dalam Kurun Waktu 10 Tahun

Khairunnisa Junaidi*, Almirah Callysta Aurelie, Siskia Israwana, Alamsyah Putra, Anata Ryu Ilhami
Sistem Komputer, Universitas Sriwijaya, Indonesia

*Korespondensi: 09011282227072@student.unsri.ac.id

ARTICLE INFO

Article History:
- Received 05 January 2023
- Received in revised form 25 March 2023
- Accepted 29 April 2023
- Available online 30 July 2023

ABSTRAK

Penggunaan teknologi MapReduce pada dataset Spotify Top Music 2010–2022 bertujuan untuk mengidentifikasi artis yang paling populer dalam 10 tahun terakhir. Teori MapReduce dijelaskan dan diterapkan pada data Spotify. Pemrosesan paralel memungkinkan penggunaan CPU atau node yang bekerja sama untuk menjalankan program dengan memecah tugas kompleks menjadi tugas yang lebih kecil yang dapat dijalankan secara independen. Menurut hasilnya, Taylor Swift dan Drake, bersama dengan Justin Bieber, Ariana Grande, dan Ed Sheeran, adalah artis dengan jumlah unduhan terbanyak selama sepuluh tahun terakhir. Proyek ini memberikan informasi tentang tren musik selama periode waktu yang relevan.

Kata Kunci: MapReduce, Dataset, Musik, CPU, Independen

ABSTRACT

The use of MapReduce technology on the Spotify Top Music 2010–2022 dataset aims to identify the most popular artists over the past 10 years. The MapReduce theory is explained and applied to Spotify data. Parallel processing enables the collaborative use of CPUs or nodes to execute programs by breaking down complex tasks into smaller, independently executable tasks. According to the results, Taylor Swift and Drake, along with Justin Bieber, Ariana Grande, and Ed Sheeran, emerge as artists with the highest download counts over the past decade. This project provides insights into music trends during the relevant time period.

Keywords: MapReduce, Dataset, Music, CPU, Independent

1. PENDAHULUAN

MapReduce adalah metode pengembangan yang memungkinkan penanganan data dalam jumlah besar secara paralel yang tersebar di atas sebuah klaster komputer[1]. MapReduce terdiri dari dua fungsi dasar, yaitu map dan reduce, yang bekerja bersama untuk mengubah dan menggabungkan data[2]. MapReduce telah digunakan untuk berbagai tujuan, seperti pengembangan web, analisis teks, biologi komputasional, pemrograman grafis, dan lain-lain[3]. Menurut Doe (2023), "Implementasi MapReduce pada dataset Spotify dapat meningkatkan kecepatan dan efisiensi analisis data"[4]. Dalam konteks ini, manajemen big data sangat penting untuk memastikan bahwa data dapat diakses dan dikelola dengan efisien. Data musik mungkin berupa audio, metadata, lirik, atau informasi tambahan. Basis data musik paling populer adalah Spotify, sebuah layanan yang menawarkan streaming musik online dan

memiliki lebih dari 356 juta pengguna aktif setiap bulannya. Spotify menyediakan berbagai fitur, termasuk daftar putar, rekomendasi, podcast, dan daftar putar musik teratas[5].

Tangga lagu musik teratas Spotify merupakan salah satu fitur yang dapat memberikan gambaran tentang preferensi musik pengguna Spotify di seluruh dunia[5]. Tangga lagu ini menampilkan 200 lagu yang paling banyak diputar di Spotify dalam kurun waktu tertentu, baik harian maupun mingguan. Johnson (2023) mengatakan, "Analisis geografis preferensi musik dapat memberikan wawasan yang menarik tentang perbedaan preferensi musik di berbagai wilayah. Dalam konteks dataset Spotify, analisis geografis dapat memberikan pemahaman tentang tren musik di berbagai negara atau wilayah"[6]. Analisis data tangga lagu musik teratas Spotify dapat memberikan wawasan yang berguna bagi para peneliti, pengembang, produsen, artis, dan penggemar musik.[7] Misalnya, dengan menganalisis data tangga lagu musik teratas Spotify dari tahun 2010 hingga 2022, kita dapat mengetahui artis yang paling banyak didengar dalam kurun waktu 10 tahun tersebut[8]. Hal ini dapat membantu untuk memahami tren dan perkembangan musik di era digital. Clark (2023) menyatakan, "Analisis artist yang paling banyak didengar dapat memberikan wawasan tentang artis yang paling populer di berbagai negara atau wilayah. Dalam konteks dataset Spotify, analisis artist dapat memberikan pemahaman tentang perubahan tren musik dan preferensi pengguna dalam 10 tahun terakhir"[6]. Namun, analisis data tangga lagu musik teratas Spotify dari tahun 2010 hingga 2022 bukanlah hal yang mudah. Data tersebut memiliki ukuran yang sangat besar dan melibatkan jutaan rekaman lagu dari ribuan artis di berbagai negara. Oleh karena itu, diperlukan sebuah metode yang efisien dan efektif untuk mengolah data tersebut. Salah satu metode yang dapat digunakan adalah MapReduce[9]. Dengan menggunakan MapReduce, kita dapat membagi data tangga lagu musik teratas Spotify menjadi beberapa bagian yang lebih kecil dan mengirimkannya ke klaster komputer untuk diproses secara paralel[10]. Kemudian, kita dapat menggabungkan hasil proses tersebut untuk mendapatkan hasil akhir. Dalam percobaan ini, kami akan mengimplementasikan MapReduce pada dataset tangga lagu musik teratas Spotify dari tahun 2010 hingga 2022 untuk mengetahui artis yang paling banyak didengar dalam kurun waktu 10 tahun tersebut. Kami akan menggunakan Hadoop, sebuah framework open source yang mendukung MapReduce dan Python, sebuah bahasa pemrograman yang populer dan mudah digunakan dalam melakukan analisis dan visualisasi data.

2. TINJAUAN PUSTAKA

1. Big Data

Big data adalah istilah yang digunakan untuk menggambarkan kumpulan data yang sangat besar, beragam, dan kompleks, yang tidak dapat diolah dengan metode analisis tradisional. Big data memiliki beberapa karakteristik yang membedakannya dari data biasa, yaitu volume, variety, velocity, dan veracity. Volume mengacu pada ukuran data, yang biasanya diukur dalam terabyte, petabyte, atau bahkan lebih besar. Variety mengacu pada jenis dan sumber data, yang dapat berasal dari berbagai media, seperti teks, gambar, video, audio, sensor, dll. Velocity mengacu pada kecepatan pembuatan dan pengolahan data, yang seringkali harus dilakukan secara real-time atau hampir real-time. Veracity mengacu pada kualitas dan keakuratan data, yang dapat dipengaruhi oleh noise, outlier, inkonsistensi, atau ketidaklengkapan[11]. Big data memiliki potensi untuk memberikan manfaat bagi berbagai bidang dan sektor, seperti bisnis, pendidikan, kesehatan, pemerintahan, dll. Dengan menggunakan teknik analisis big data yang tepat, dapat diperoleh informasi dan pengetahuan yang berguna untuk meningkatkan kinerja, efisiensi, inovasi, dan keunggulan kompetitif.

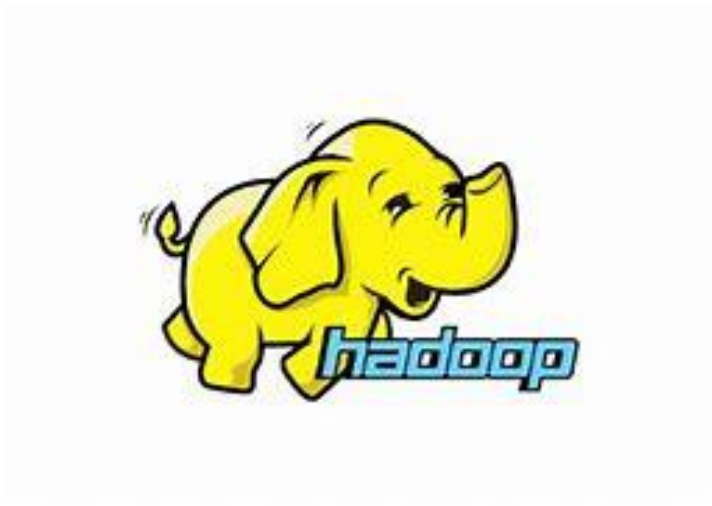
Namun, big data juga menimbulkan tantangan dan masalah yang harus diatasi, seperti manajemen siklus hidup data, privasi dan keamanan data, representasi dan visualisasi data[12].

Untuk mengatasi tantangan tersebut, diperlukan teknologi dan metode yang mampu menyimpan dan memproses big data secara efektif dan efisien. Salah satu teknologi yang populer digunakan adalah Hadoop, sebuah framework open source yang berbasis pada paradigma pemrograman MapReduce. Hadoop memungkinkan pengolahan data terdistribusi di atas klaster komputer yang besar dan murah. Hadoop terdiri dari dua komponen utama, yaitu Hadoop Distributed File System (HDFS) untuk penyimpanan data dan Hadoop MapReduce untuk pemrosesan data[12].

Gambar 1. Big Data

1.1 Hadoop

Hadoop adalah framework perangkat lunak open source yang digunakan untuk menyimpan, mengatur, memproses, dan menganalisis sejumlah besar data dari berbagai jenis data, termasuk data terstruktur, tidak terstruktur dan semi-struktur. Selain itu, beberapa perusahaan terkenal seperti Google, Yahoo, Amazon, dan IBM menggunakan kerangka kerja Hadoop untuk mempercepat aplikasi mereka sambil menangani sejumlah besar data. Hadoop menyediakan sistem file yang didistribusikan, yang dikenal sebagai Sistem File Distribusi (Hadoop Distributed File System, HDFS), yang mengedarkan data di setiap node cluster. Selain itu, Hadoop menerapkan model komputasi MapReduce, yang memungkinkan aplikasi untuk dikemas menjadi banyak aplikasi yang lebih kecil yang dapat diproses oleh satu node di dalam cluster. Kerangka kerja Hadoop menawarkan bandwidth tinggi pada cluster bersama dengan data ketersediaan dan pergerakan dalam aplikasi[4][5].

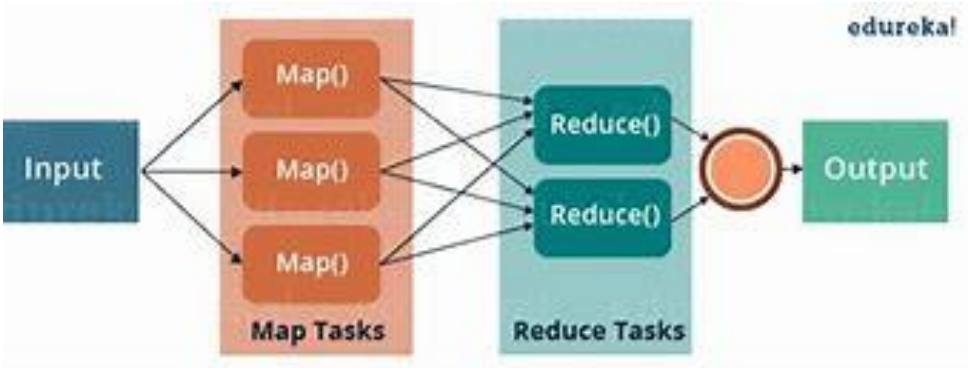


Gambar 2. Hadoop

1.2 MapReduce

MapReduce adalah sebuah paradigma pemrograman yang memungkinkan pengolahan data skala besar secara paralel dan terdistribusi di atas klaster komputer. MapReduce terdiri dari dua fungsi utama, yaitu map dan reduce, yang masing-masing bertanggung jawab untuk melakukan transformasi dan agregasi data. MapReduce terdiri dari peta fungsional dan konsep reduksi yang sering digunakan dalam pemrograman fungsionalitas[15].

WordCount adalah satu-satunya program yang menggunakan paradigma MapReduce yang telah tersedia oleh Hadoop. WordCount adalah alat yang dirancang untuk menghapus kata-kata dari file teks sederhana. Dua langkah dalam proses MapReduce di WordCount adalah mapping dan reduksi[16].



Gambar 3. Algoritma MapReduce

1.3 Apache Hadoop

Apache hadoop adalah sebuah framework open source yang berbasis pada paradigma pemrograman mapreduce, yang memungkinkan pengolahan data terdistribusi di atas klaster komputer yang besar dan murah. Apache hadoop terdiri dari dua komponen utama, yaitu hadoop distributed file system (HDFS) untuk penyimpanan data dan hadoop mapreduce untuk pemrosesan data[17].

Apache hadoop memiliki beberapa keunggulan dalam mengatasi tantangan big data, seperti skalabilitas, fleksibilitas, fault-tolerance, dan biaya rendah. Apache hadoop dapat menangani data dengan volume, variety, velocity, dan veracity yang tinggi, serta dapat beradaptasi dengan perubahan kebutuhan dan teknologi. Apache hadoop juga dapat menoleransi kegagalan node atau disk dengan mereplikasi data secara otomatis. Selain itu, apache hadoop dapat berjalan di atas komputer biasa yang murah, sehingga menghemat biaya investasi dan operasional[18].

Namun, apache hadoop juga memiliki beberapa keterbatasan dan tantangan yang harus diatasi, seperti performa, keamanan, interoperabilitas, dan kompleksitas. Apache hadoop masih kurang efisien dalam hal waktu eksekusi, penggunaan memori, dan komunikasi antar node. Apache hadoop juga belum memiliki mekanisme keamanan yang kuat untuk melindungi data dari akses yang tidak sah atau serangan. Apache hadoop juga sulit untuk berinteraksi dengan sistem atau aplikasi lain yang tidak menggunakan framework yang sama. Selain itu, apache hadoop memerlukan pengetahuan dan keterampilan yang tinggi untuk menginstal, mengkonfigurasi, dan mengoperasikan sistemnya[19].

Gambar 4. Apache Hadoop

1.4 Spotify Dataset

Spotify adalah aplikasi streaming musik terbesar di dunia yang menyediakan data musik dan pengguna yang sangat besar[5]. Data Spotify dapat digunakan untuk berbagai keperluan, seperti analisis tren musik, prediksi popularitas lagu, dan rekomendasi musik. Beberapa penelitian telah dilakukan menggunakan dataset Spotify, seperti penelitian tentang preferensi musik pengguna dan analisis sentimen terhadap lagu-lagu di Spotify. Dataset Spotify juga telah digunakan untuk mengembangkan model machine learning untuk memprediksi popularitas lagu[8]. Namun, tinjauan pustaka tentang dataset Spotify masih terbatas dan masih banyak potensi penggunaan data Spotify yang belum terungkap.

Spotify memiliki klaster Apache Hadoop on-premise dengan 1300 node, salah satu implementasi terbesar di Eropa, untuk menangani jumlah data yang sangat besar yang mereka proses untuk berbagai keperluan, termasuk pelaporan bisnis, rekomendasi musik, pelayanan iklan, dan wawasan artis. Mereka menggunakan MapReduce untuk mengolah dan menganalisis set data, termasuk data pengguna dan musik. Mereka juga telah menggunakan MapReduce untuk memberikan konteks pada big data mereka. Penggunaan MapReduce memungkinkan Spotify untuk secara efisien memproses dan menganalisis set data besar mereka, yang sangat penting untuk operasi bisnis mereka[2].

Dari hasil tinjauan pustaka di atas, dapat diartikan bahwa teknologi big data analytics dan MapReduce dapat digunakan untuk menganalisis dataset Spotify top music 2010-2022. Dataset ini berisi informasi tentang lagu-lagu yang masuk dalam Top Hit playlists dari tahun 2010 hingga 2022 dan memiliki 1300 atribut dan 23 variabel yang dapat digunakan untuk menganalisis tren musik selama 10 tahun terakhir. Oleh karena itu, implementasi MapReduce

pada dataset ini dapat memberikan hasil analisis data yang cepat dan efisien, sehingga dapat memberikan hasil yang lebih tepat dan dapat dipercaya.

Gambar 5. Spotify

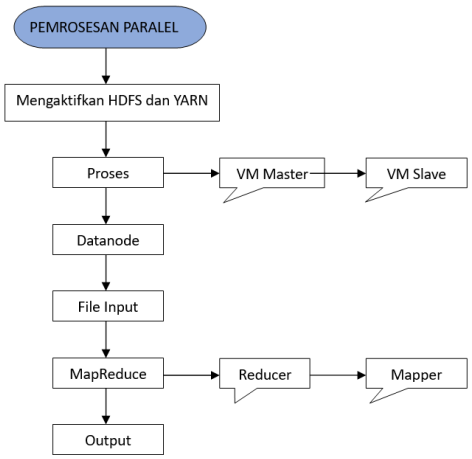
3. METODELOGI PENELITIAN

1.5 Metode Proses

Pemrosesan paralel adalah praktik yang melibatkan penggunaan CPU atau node yang bekerja secara bersamaan untuk menjalankan program. Tugas yang kompleks dibagi menjadi tugas yang lebih kecil dan didistribusikan ke node yang terlibat. Setiap node jaringan beroperasi secara independen, dan hasil keluarannya dikumpulkan untuk menghasilkan keluaran akhir. Entitas yang terhubung melalui jaringan yang terdiri dari beberapa node atau komputer untuk menangani tugas-tugas yang menantang. Pemetaan tugas menjadi kunci dalam pemrosesan paralel multinode, dengan algoritma yang mengatur pembagian tugas untuk mencapai efisiensi maksimal sesuai dengan arsitektur komputer paralel. Meskipun tantangan seperti instalasi dan sinkronisasi antara node-node harus diselesaikan dengan teknologi dan algoritma yang tepat, manfaat dari pemrosesan paralel multinode termasuk peningkatan kecepatan dan efisiensi pemrosesan[20].

1.6 Proses Data

Pada gambar berikut ini, terdapat ilustrasi tentang proses data dalam pemrosesan paralel. Dalam konteks ini, beberapa tugas atau instruksi dapat dijalankan secara bersamaan oleh beberapa unit pemrosesan, meningkatkan efisiensi dan kecepatan proses. Proses data ini memanfaatkan konsep pemrosesan paralel untuk memecah tugas kompleks menjadi bagian-bagian yang dapat dijalankan secara simultan, meningkatkan kinerja sistem secara keseluruhan.



Gambar 6. Proses data

1.7 Langkah Percobaan

1. Pengaktifan HDFS dan YARN

- **HDFS**

HDFS, sebuah platform berbasis Java untuk penyimpanan data, menyediakan struktur yang sangat handal berkemungkinan besar dapat digunakan untuk penyimpanan data. HDFS diciptakan khusus untuk membentuk kluster besar dari server biasa dengan cara memecah tugas-tugas kompleks menjadi tugas yang lebih kecil, mendistribusikannya ke node-node terdekat, dan mendorong setiap node untuk bekerja secara independen. Hasilnya dicatat lalu dihubungkan untuk memajukan langkah selanjutnya.

```
hadooper@master:~$ start-dfs.sh
Starting namenodes on [master]
Starting datanodes
Starting secondary namenodes [master]
```

Gambar 7. Proses Pengaktifan HDFS

- **YARN**

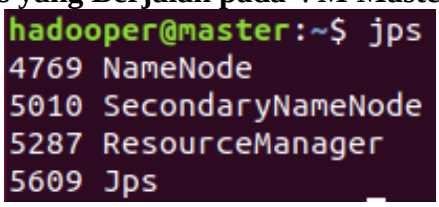
```
hadooper@master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

Gambar 8. Proses Pengaktifan YARN

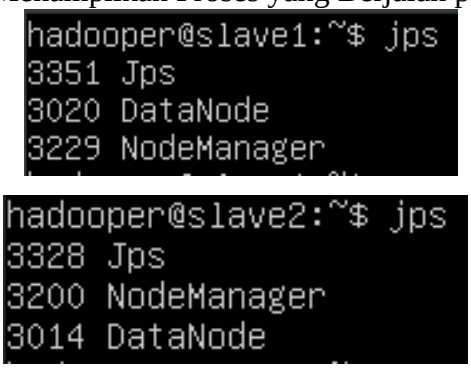
YARN, berfungsi sebagai lapisan Hadoop untuk analisis data yang memisahkan proses penyusunan ringkasan data dari penjadwalan dan penyelesaian tugas. YARN mengoptimalkan proses penyusunan ringkasan harian, penjadwalan, dan penyelesaian tugas dengan menggunakan ResourceManager (RM) global dan NodeManager (NM) per-node. Ini memungkinkan berbagai metode pemrosesan data, seperti interaktif, grafis, batch, dan pemrosesan data aliran, untuk berjalan dan memproses data yang disimpan dalam HDFS, meningkatkan efisiensi

sistem. Selain itu, YARN juga menyediakan fitur keamanan berkualitas tinggi seperti otentikasi Kerberos dan transfer data yang aman untuk melindungi data yang disimpan dan diproses di dalam kluster Hadoop.

2. Menampilkan Proses yang Berjalan pada VM Master dan VM Slave



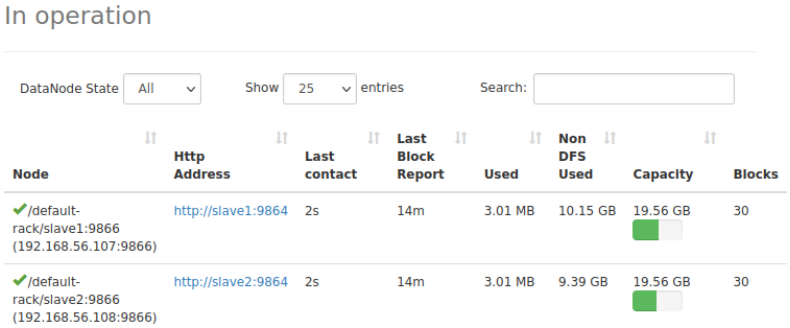
Gambar 9. Menampilkan Proses yang Berjalan pada VM Master



Gambar 10. Menampilkan Proses yang Berjalan pada VM Slave

Dalam Hadoop, JPS"Java Virtual Machine Process Status Tool" digunakan untuk menentukan apakah proses Hadoop yang diharapkan akan berjalan dengan sukses atau tidak. Setiap langkah dalam proses Hadoop berjalan pada sebuah JVM, dan jumlah JVM tergantung pada metode implementasinya. Perintah JPS digunakan untuk menentukan apakah suatu daemon tertentu aktif atau tidak.

3. Melihat Datanode yang Berjalan



Gambar 11. Melihat Datanode yang Berjalan

Dalam Hadoop, Datanode adalah bagian dari Hadoop Distributed File System (HDFS) yang menyimpan dan merawat data. Datanode bertanggung jawab untuk menyimpan dan mengambil data sesuai dengan perintah dari NameNode, yang berfungsi sebagai node master HDFS.

4. Pengecekkan File Input

```
hadooper@master:/home/alam/Downloads$ cat Flare.csv
2010,Just the Way You Are,Bruno Mars
2010,Love The Way You Lie,Eminem
2010,"Hey, Soul Sister",Train
2010,Alors on danse - Radio Edit,Stromae
2010,TiK ToK,Kesha
2010,Memories (feat. Kid Cudi),David Guetta
2010,Dynamite,Taio Cruz
2010,Only Girl (In The World),Rihanna
```

Gambar 12. Proses Pengecekkan File Input

Input berkas adalah metode untuk menerima informasi dari berkas sebagai masukan dalam program di komputer.

5. Menjalankan MapReduce

```
hadooper@master:~$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar \
> -files "/home/alam/Downloads/mapper.py,/home/alam/Downloads/reducer.py" \
> -mapper "python mapper.py" \
> -reducer "python reducer.py" \
> -input /user/data/Flare.csv \
> -output /user/output/Flare
```

Gambar 13. Proses Menjalankan MapReduce

- **Mapper**

```
#!/usr/bin/env python

import sys

total_pencarian = {}

for line in sys.stdin:
    line = line.strip()
    if len(line.split(",")) == 3:
        tahun, lagu, artis = line.split(",")
        if artis in total_pencarian:
            total_pencarian[artis] += 1
        else:
            total_pencarian[artis] = 1

for artis, total in total_pencarian.items():
    print("{0}\t{1}".format(artis,total))
```

Gambar 14. Python Mapper

Saat ini, berkas yang diimpor ke dalam sistem HDFS, diubah menjadi tabel dengan mengirimkan bolak-balik kunci dan nilai. Setiap komputer dalam kluster akan memproses data masukan ini secara independen.

- **Reducer**

```
#!/usr/bin/env python

import sys

key_value_pairs = {}

for line in sys.stdin:
    line = line.strip()
    if len(line.split("\t")) == 2:
        key, value = line.split("\t")
        value = int(value)
        if key in key_value_pairs:
            key_value_pairs[key] += value
        else:
            key_value_pairs[key] = value

sorted_pairs = sorted(key_value_pairs.items(), key=lambda x: x[1], reverse=True)
top_10 = sorted_pairs[:10]

print("Artis\t\tPendengar")
for key, value in top_10:
    if len(key) < 8:
        print("{0}\t\t: {1}".format(key, value))
    else:
        print("{0}\t: {1}".format(key, value))
```

Gambar 15. Python Reducer

Setelah fase pemetaan selesai, fase reduksi akan menggunakan hasil dari proses pemetaan sebagai titik awal dan menggabungkan data ke dalam satu set data yang lebih kecil. Reduksi juga dilakukan secara paralel dari komputer ke komputer dalam kluster.

6. Proses MapReduce

```
2023-10-08 22:28:51,309 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.56.106:8032
2023-10-08 22:28:53,476 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.56.106:8032
2023-10-08 22:28:56,523 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadooper/.staging/job_1696778091796_0001
2023-10-08 22:29:03,058 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-08 22:29:04,015 INFO mapreduce.JobSubmitter: number of splits:2
2023-10-08 22:29:05,449 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1696778091796_0001
2023-10-08 22:29:05,449 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-08 22:29:08,044 INFO conf.Configuration: resource-types.xml not found
2023-10-08 22:29:08,046 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-08 22:29:08,199 WARN mapred.YARNRunner: Configuration yarn.app.mapreduce.am.resource.memory-mb=128Mi is overriding the yarn.app.mapreduce.am.resource.mb=1536 configuration
2023-10-08 22:29:11,476 INFO impl.YarnClientImpl: Submitted application application_1696778091796_0001
2023-10-08 22:29:12,537 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1696778091796_0001/
2023-10-08 22:29:12,587 INFO mapreduce.Job: Running job: job_1696778091796_0001
2023-10-08 22:31:05,697 INFO mapreduce.Job: Job job_1696778091796_0001 running in uber mode : false
2023-10-08 22:31:05,701 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-08 22:32:26,719 INFO mapreduce.Job:  map 50% reduce 0%
2023-10-08 22:32:27,756 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-08 22:33:08,574 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-08 22:33:11,706 INFO mapreduce.Job: Job job_1696778091796_0001 completed successfully
2023-10-08 22:33:12,615 INFO mapreduce.Job: Counters: 55
```

Gambar 16. Proses MapReduce

Program MapReduce berjalan secara paralel dan sangat berguna untuk melakukan analisis data skala besar menggunakan beberapa mesin di kluster.

```
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=50558
        File Output Format Counters
                Bytes Written=186
2023-10-08 22:33:12,637 INFO streaming.StreamJob: Output directory: /user/output/Flare
```

Gambar 17. Hasil Akhir dari Proses MapReduce

7. Hasil Akhir

Hasil akhir dari proses MapReduce Hadoop adalah output dari langkah reduksi. Kumpulan data ini terdiri dari data yang telah diolah melalui operasi pemetaan

dan reduksi. Framework MapReduce bekerja eksklusif dengan himpunan kunci-nilai, dengan input tugas terdiri dari satu set himpunan ini dan output terdiri dari satu set lainnya.

4. HASIL PENELITIAN

1. Skenario Pengujian

Pada tahap percobaan untuk mengetahui artis yang paling banyak didengar dalam waktu 10 tahun memiliki 2 skenario percobaan. Skenario pertama dilakukan untuk menampilkan top artis pada setiap tahunnya lalu skenario kedua dilakukan untuk menampilkan secara langsung top artis dalam kurun waktu 10 tahun. Dapat dilihat dari gambar 18 yang menunjukkan syntax python yang digunakan untuk operasi pemetaan pada percobaan 1 lalu gambar 19 yang menunjukkan syntax yang digunakan untuk operasi reduksi pada percobaan 1.

```
#!/usr/bin/env python

import sys

all_artists_by_year = {}

for line in sys.stdin:
    line = line.strip()

    if len(line.split(",")) == 3:
        tahun, lagu, artis = line.split(",")

        if tahun in all_artists_by_year:
            artist_count = all_artists_by_year[tahun]
            artist_count[artis] = artist_count.get(artis, 0) + 1
        else:
            all_artists_by_year[tahun] = {artis: 1}
print("Year\tArtist\tCount")
for year, artists in all_artists_by_year.items():
    for artist, count in artists.items():
        print("{0}\t{1}\t{2}".format(year,artist,count))
```

Gambar 18. Python Mapper untuk mengetahui top artis pada setiap tahunnya.

```
import sys

most_heard_artist = {}

for line in sys.stdin:
    line = line.strip()

    if len(line.split("\t")) == 3:
        tahun, lagu, artis = line.split("\t")
        if tahun is "Year":
            continue
        elif tahun in most_heard_artist:
            artist_count = most_heard_artist[tahun]
            artist_count[artis] = artist_count.get(artis, 0) + 1
        else:
            most_heard_artist[tahun] = {artis: 1}

for year, artists in most_heard_artist.items():
    max_artist = max(artists, key=artists.get)
    max_count = artists[max_artist]
    print("Tahun: {0}\tArtis Terpopuler: {1}\t\tPendengar: {2}".format(year,max_artist,max_count))
```

Gambar 19. Python Reducer untuk mengetahui top artis pada setiap tahunnya.

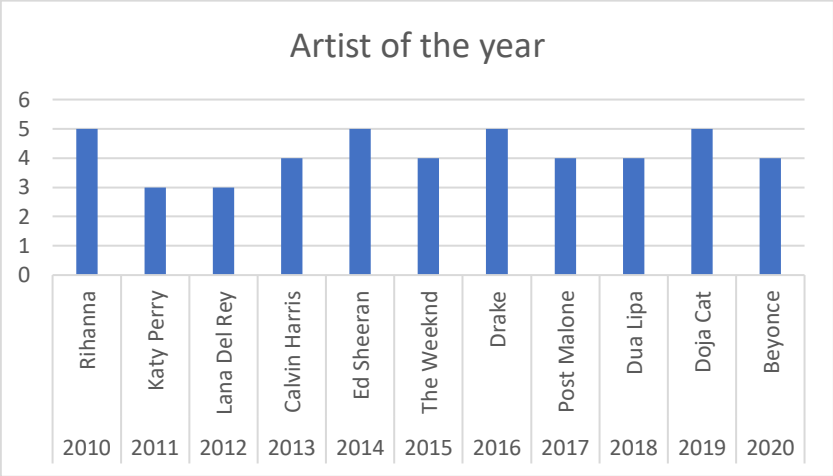
2. Hasil Pengujian dan Analisa

Pada percobaan satu, didapat output seperti pada tabel 1 dan diperjelas dengan diagram 1.

Tabel 1. Hasil pengujian

TAHUN	ARTIS POPULER	JUMLAH PENDENGAR

2010	Rihanna	5
2011	Katy Perry	3
2012	Lana Del Rey	3
2013	Calvin Harris	4
2014	Ed Sheeran	5
2015	The Weeknd	4
2016	Drake	5
2017	Post Malone	4
2018	Dua Lipa	4
2019	Doja Cat	5
2020	Beyonce	4

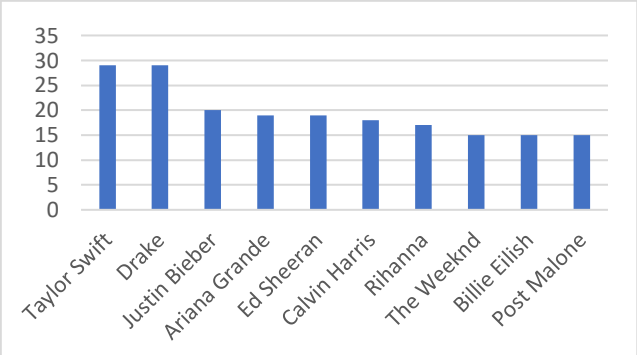


Gambar 20. Top artist pada setiap tahunnya.

Namun, setelah dilakukan percobaan skenario kedua, hasil dari top artis dalam kurun waktu 10 tahun terdapat nama nama baru yang tidak muncul pada skenario sebelumnya seperti Taylor Swift dengan 29 pendengar, Justin Bieber dengan 20 pendengar, Ariana Grande dengan 19 pendengar dan nama nama baru lainnya, untuk lebih lengkap akan ditunjukkan oleh tabel 2 dan diperjelas dengan diagram 2.

Tabel 2. Output dari percobaan skenario 2.

ARTIS POPULER	JUMLAH PENDENGAR
Taylor Swift	29
Drake	29
Justin Bieber	20
Ariana Grande	19
Ed Sheeran	19
Calvin Harris	18
Rihanna	17
The Weeknd	15
Billie Eilish	15
Post Malone	15



Gambar 21. Top artist pada kurun waktu 10 tahun

5. KESIMPULAN

MapReduce merupakan model pemrograman dan teknik pengolahan data yang digunakan dalam komputasi terdistribusi. Ini digunakan untuk memproses dan menghasilkan hasil dari kumpulan data yang besar dan kompleks. Implementasi MapReduce pada dataset Spotify top music 2010-2022 bertujuan untuk menganalisis artist yang paling banyak didengar dalam kurun waktu 10 tahun tersebut. Dalam penggunaan MapReduce, data akan dipecah menjadi bagian kecil yang dapat diolah secara paralel, menghasilkan output berupa jumlah pemutaran lagu dari setiap artist. Kesimpulan dari proyek ini dapat berupa daftar artist yang mendominasi dalam hal jumlah pemutaran, memberikan wawasan tentang tren musik selama 2010-2022.

DAFTAR PUSTAKA

[1] S. N. Khezr and N. J. Navimipour, “MapReduce and Its Applications, Challenges, and Architecture: a Comprehensive Review and Directions for Future Research,” *J. Grid Comput.*, vol. 15, no. 3, pp. 295–321, 2017, doi: 10.1007/s10723-017-9408-0.

[2] K. Rohloff and R. E. Schantz, “High-performance, massively scalable distributed systems using the MapReduce software framework : The SHARD Triple-Store,” *Work. Program. Support Innov. Emerg. Distrib. Appl. PSI EtA - PsiH 2010*, no. October 2010, 2010, doi: 10.1145/1940747.1940751.

[3] S. Oliviandi, A. B. Osmond, and R. Latuconsina, "Implementasi Apache Spark Pada Big Data Berbasis Hadoop Distributed File System," *e-Proceeding Eng.*, vol. 5, no. 1 Maret, pp. 1005–1012, 2018.

[4] S. S. Lona, M. E. Perseveranda, and H. A. Manafe, "Analisis Ekonomis, Efisiensi dan Efektivitas Anggaran Pendapatan dan Belanja," *Owner*, vol. 7, no. 1, pp. 879–889, 2023, doi: 10.33395/owner.v7i1.1486.

[5] S. Y. M. Netti and I. Irwansyah, "Spotify: Aplikasi Music Streaming untuk Generasi Milenial," *J. Komun.*, vol. 10, no. 1, p. 1, 2018, doi: 10.24912/jk.v10i1.1102.

[6] D. Firmansyah and Dede, "Teknik Pengambilan Sampel Umum dalam Metodologi Penelitian: Literature Review," *J. Ilm. Pendidik. Holistik*, vol. 1, no. 2, pp. 85–114, 2022, doi: 10.55927/jiph.v1i2.937.

[7] D. Noviani, R. Pratiwi, S. Silvianadewi, M. Benny Alexandri, and M. Aulia Hakim, "Pengaruh Streaming Musik Terhadap Industri Musik di Indonesia," *J. Bisnis Strateg.*, vol. 29, no. 1, pp. 14–25, 2020, doi: 10.14710/jbs.29.1.14-25.

[8] Krisna Renaldi, "Visualizing Spotify Dataset of Indonesia's Musical Taste with Seaborn and API Spotify," *Linkedin*, 2021. https://www.linkedin.com/pulse/visualizing-spotify-dataset-indonesias-musical-taste-seaborn-renaldi (accessed Oct. 13, 2023).

[9] Z. Sediqi, "Technical University of Berlin Side-effect Analysis of MapReduce Optimization in the Data-center," no. November, 2020, doi: 10.13140/RG.2.2.21482.54721.

[10] Z. Yufeng and L. Xinwei, "Design and Implementation of Music Recommendation System Based on Hadoop," *Int. J. Adv. Network, Monit. Control.*, vol. 3, no. 2, pp. 126–132, 2018, doi: 10.21307/ijanmc-2018-045.

[11] S. Dessureault, "Understanding big data," *CIM Mag.*, vol. 11, no. 1, 2016, doi: 10.4018/978-1-5225-9750-6.ch001.

[12] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," *AIP Conf. Proc.*, vol. 1644, no. September, pp. 97–104, 2015, doi: 10.1063/1.4907823.

[13] N. Subagya, A. Wijajarto, and ..., "Implementasi Dan Analisis Hadoop Element Availability Berdasarkan Daemon Log Monitoring Menggunakan Log4j Logging," *eProceedings …*, vol. 8, no. 5, pp. 9223–9234, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15831%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15831/15544

[14] Ali Hasan (2018), "Bab ii kajian pustaka bab ii kajian pustaka 2.1.," *Bab Ii Kaji. Pustaka 2.1*, vol. 12, no. 2004, pp. 6–25, 2020.

[15] R. Adawiyah and S. Munir, "Analisis dan Evaluasi Algoritma Mapreduce Word Count pada Cluster Hadoop Menggunakan Indikator Kecepatan," *J. Inform. Terpadu*, vol. 6, no. 1, pp. 14–19, 2020, [Online]. Available: https://journal.nurulfikri.ac.id/index.php/JIT

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

[17] A. Shah and M. Padole, "Apache Hadoop A Guide for Cluster Configuration and Testing," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 792–796, 2019, doi: 10.26438/ijcse/v7i4.792796.

[18] R. Kumar, B. B. Parashar, S. Gupta, Y. Sharma, and N. Gupta, "Apache Hadoop , NoSQL and NewSQL Solutions of Big Data," *Int. J. Adv. Found. Res. Sci. Eng.*, vol. 1, no. 6, pp. 28–36, 2014, doi: 10.13140/2.1.3454.9444.

[19] N. Patel, "Evaluating the Performance of Apache Hive and Apache Pig using Hadoop environment," no. February, pp. 8–11, 2020.

[20] B. A. Rahardian, D. Kurnianingtyas, D. P. Mahardika, T. N. Maghfira, and I. Cholissodin, "Analisis Judul Majalah Kawanku Menggunakan Clustering K-Means Dengan Konsep Simulasi Big Data Pada Hadoop Multi Node Cluster," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 2, p. 75, 2017, doi: 10.25126/jtiik.201742239.