

Analisis Dataset TOP 1000 IMDb Movies Menggunakan Hadoop

Intan Permatahati, M. Nauval Perdana, Nanda Apriadi, Tiara Putri Amanda, Zahra Maharani
Sistem Komputer, Universitas Sriwijaya, Indonesia

*Korespondensi: hatii3823@gmail.com

ARTICLE INFO

Article History:

- Received 10 January 2023
- Received in revised form 21 March 2023
- Accepted 10 April 2023
- Available online 30 July 2023

ABSTRAK

Dalam studi ini, kami mengeksplorasi analisis dataset TOP 1000 IMDb Movies dengan memanfaatkan keunggulan kerangka kerja Hadoop. Pertama-tama, untuk mengatasi tantangan pemrosesan data yang tidak terstruktur, kami menerapkan strategi khusus melalui infrastruktur yang mampu menangani volume besar dan kompleksitas data, yang umumnya dikenal sebagai Big Data. Kerangka kerja Hadoop telah terbukti efektif dalam menangani tugas-tugas semacam ini dengan menyediakan lingkungan yang terdistribusi untuk pemrosesan data secara paralel. Kemudian, untuk mengoptimalkan kueri data pada dataset IMDb, kami mengadopsi pendekatan dengan menggunakan Python. Keunggulan Python sebagai bahasa pemrograman untuk analisis data memberikan fleksibilitas dan keterbacaan kode, sementara implementasi multinode memungkinkan pemrosesan data yang terdistribusi untuk meningkatkan efisiensi. Gabungan dari kerangka kerja Hadoop, Python, dan pendekatan multinode menghasilkan pendekatan analisis yang efisien dan handal untuk dataset film IMDb TOP 1000.

Kata Kunci: IMDB, Top Movies, Hadoop, Big Data, Multinode

ABSTRACT

In this study, we explore the analysis of the TOP 1000 IMDb Movies dataset by leveraging the advantages of the Hadoop framework. Firstly, to address the challenges of processing unstructured data, we implement a specialized strategy through an infrastructure capable of handling large volumes and the complexity of data, commonly known as Big Data. The Hadoop framework has proven effective in handling such tasks by providing a distributed environment for parallel data processing. Subsequently, to optimize data queries on the IMDb dataset, we adopt an approach using Python. The superiority of Python as a programming language for data analysis provides flexibility and code readability, while the multinode implementation enables distributed data processing to enhance efficiency. The combination of the Hadoop framework, Python, and the multinode approach results in an efficient and reliable analytical approach for the TOP 1000 IMDb Movies dataset.

Keywords: IMDB, Top Movies, Hadoop, Big Data, Multinode

1. PENDAHULUAN

Dalam era revolusi industri 4.0, Big Data bukan lagi menjadi istilah asing. Terlampauinya kemampuan media penyimpanan maupun sistem database yang pertumbuhan datanya terus berlipat ganda dari waktu ke waktu menjadi alasan tercetusnya istilah big data. Big Data adalah istilah yang menggambarkan data yang berukuran besar, berkecepatan tinggi, kompleks, dan variabel data yang memerlukan teknik dan teknologi tingkat tinggi untuk memungkinkan dilakukannya pengumpulan, penyimpanan, distribusi, manajemen, dan analisis

informasi [1] . Big data berguna dalam memprediksi dan menganalisis penyebab suatu masalah pada sistem. Hal ini juga dapat meminimalisir terjadinya kegagalan, apalagi ditambah dengan hasil analisisnya langsung dapat ditampilkan. Selain itu, juga bisa menjadi referensi untuk pengembangan produk juga mengurangi waktu dan biaya. Hal ini yang menjadi alasan penerapan big data di era sekarang hampir di segala bidang. Dalam analisis kali ini, kami menggunakan Hadoop untuk mengolah big data [2] .

Teknologi Big Data di masa sekarang sudah banyak sekali membantu orang atau perusahaan. Di masa sekarang, Big Data memengaruhi banyak pihak dikarenakan ia adalah dasar empirik untuk banyak strategi pemasaran serta keputusan publik. Hal ini dikarenakan Big Data dapat melihat pola perilaku orang. Contoh pemanfaatan Big Data dalam dunia usaha yaitu membantu mendapatkan informasi mengenai feedback masyarakat ke produk-produk yang diluncurkan melalui analisis pendapat masyarakat di media sosial, dan juga membantu perusahaan untuk mengambil keputusan sehingga dapat menghasilkan keputusan terbaik berdasarkan Big Data. Selain membantu menganalisa bisnis, teknologi ini juga dapat digunakan secara luas ranah pemerintahan. Diantaranya, pemanfaatan Big Data di sektor publik dapat memberikan informasi mengenai feedback Masyarakat terhadap layanan pemerintah di masyarakat. Bahkan dapat membantu untuk menemukan solusi terhadap masalah yang ada dengan cara menganalisis masalah tersebut [3]

Dalam analisis ini, kami menggunakan Hadoop sebagai kerangka kerja pemrosesan big data untuk menganalisis dataset besar seperti IMDb Top 1000 Movies. Bertujuan untuk menambah wawasan mengenai rating di berbagai film, jika untuk keperluan pribadi hal ini bisa berguna untuk memutuskan film yang ingin ditonton. Namun, untuk industri film, hal ini dapat berguna untuk pengambilan keputusan di industri film. Mengetahui hubungan antara genre dan klasifikasi film untuk mengidentifikasi tren atau pola spesifik yang mungkin berguna bagi produser, sutradara, dan pemangku kepentingan lainnya di industri film ataupun hanya untuk sekedar pengetahuan umum. Untuk hasil analisis film yang lebih baik, dilakukannya penambahan data mendalam, lalu diolah menggunakan Hadoop, yang dapat digunakan sebagai landasan untuk pengembangan analisis lebih lanjut atau aplikasi bisnis di industri film. Eksperimen ini dapat memberikan wawasan tentang bagaimana rating tiap film yang masuk dalam kategori 1000 terbaik dan bagaimana Hadoop dapat digunakan untuk mengelola dan menganalisis kumpulan data besar secara efisien [4]. Penelitian ini berguna untuk membantu pengguna dalam pemilihan film yang cocok dengan preferensi mereka. Umumnya, pengguna seringkali merasa kebingungan dalam menemukan genre film yang disukai. Bahkan, terkadang untuk mencari film yang ingin dinikmati dapat menghabiskan waktu yang lama. Maka dari itu, analisis rating terhadap genre film ini harapannya akan membantu dalam memahami tren dan preferensi pengguna. Sehingga, tidak terjadinya ekspektasi yang tidak sesuai terhadap sebuah film. Pengurutan berdasarkan rating ini membantu menemukan film-film yang diakui oleh pengguna lain sebagai film yang bagus sesuai preferensi pengguna [5]. Big data adalah istilah yang digunakan untuk menggambarkan jumlah data yang besar, kecepatan dalam pengumpulannya, dan keragaman data yang diperlukan untuk mengatasi tantangan pengelolaan, pemrosesan, dan analisis data. Ada beberapa tools yang sangat mumpuni untuk mengelola dan menganalisis big data, diantaranya:

1. Apache Hadoop : Kerangka kerja sumber terbuka yang populer untuk menyimpan dan memproses big data. Ini termasuk Hadoop Distributed File System (HDFS) dan MapReduce yang memungkinkan analisis terdistribusi [6] .
2. Apache Spark : Kerangka kerja komputasi cepat yang memungkinkan analisis data real-time dan batch. Ini lebih cepat dari pada MapReduce karena menggunakan pemrosesan dalam memori [7] .

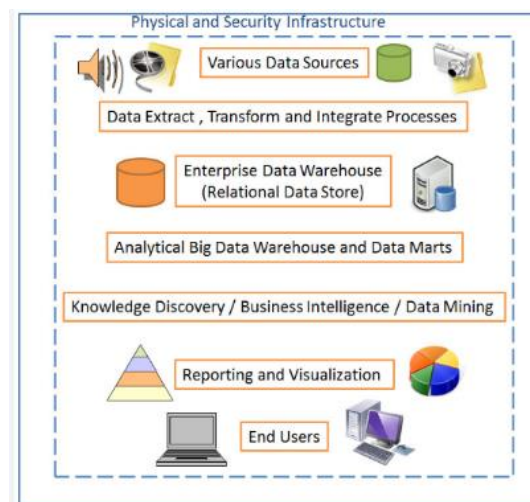
3. Apache Flink: Kerangka kerja pemrosesan aliran data yang sangat baik untuk aplikasi yang memerlukan pemrosesan aliran data waktu nyata.
4. Apache Kafka: Platform streaming data yang dapat digunakan untuk mengumpulkan, mengirim, dan menganalisis aliran data secara real-time [8].
5. Apache HBase: Database NoSQL yang dirancang untuk menyimpan dan mengakses big data dengan cepat.
6. DataBase NoSQL: Beberapa database NoSQL seperti MongoDB, Cassandra dan Couchbase cocok untuk menyimpan dan memproses data semi-terstruktur dan tidak terstruktur [9].
7. SQL-on-Hadoop: Kerangka kerja seperti Apache Hive dan Apache Impala mengaktifkan kueri SQL terhadap data yang disimpan di Hadoop [10].

2. TINJAUAN PUSTAKA

2.1 Big Data

Kata ini berarti kumpulan data yang besar dan terus berkembang, termasuk data heterogen, tidak terorganisir, dan semi-terorganisir. Big data bersifat kompleks sehingga membutuhkan teknologi dan algoritma canggih. Big data didefinisikan menjadi 3 karakteristik utama [11].

- Volume:
Data digital dalam jumlah besar terus dihasilkan dari berbagai hal (TIK, ponsel cerdas, kode produk, jejaring sosial, sensor, log, dll). Intinya, volume berpacu pada data yang terkumpul, tentunya hal ini meliputi triliunan atau lebih baris data.
- Kecepatan:
Melihat dari seberapa cepat pertumbuhan data baru yang harus diolah.
- Varian:
Menggambarkan keberagaman sumber dan tipe data. Salah satu tantangan mengolah big data yaitu kita harus mampu mengelola dan menganalisis data yang ada.

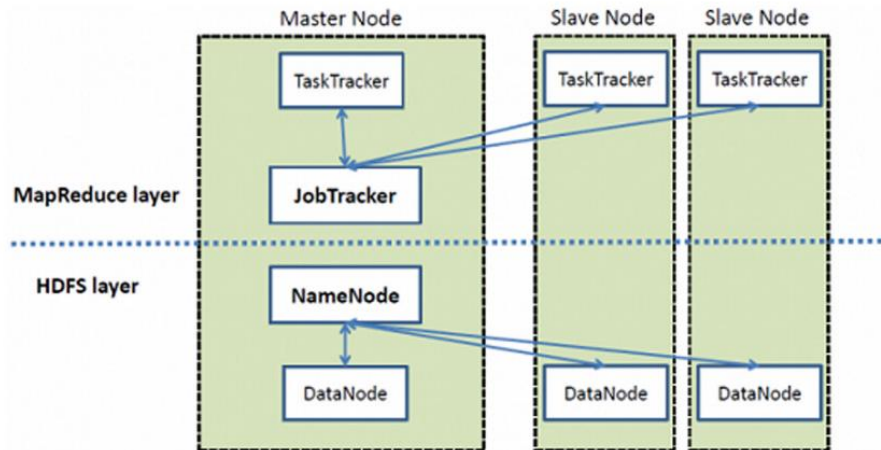


Gambar 1. Arsitektur Big Data

Big Data Analytics membantu mengidentifikasi transformator yang berisiko dan mendeteksi perilaku tidak biasa dari perangkat yang terhubung. Dengan cara ini, Grid Utilities dapat memilih tindakan terbaik. Analisis real-time dari Big Data yang diperoleh membantu memodelkan skenario kegagalan.

2.2 Hadoop

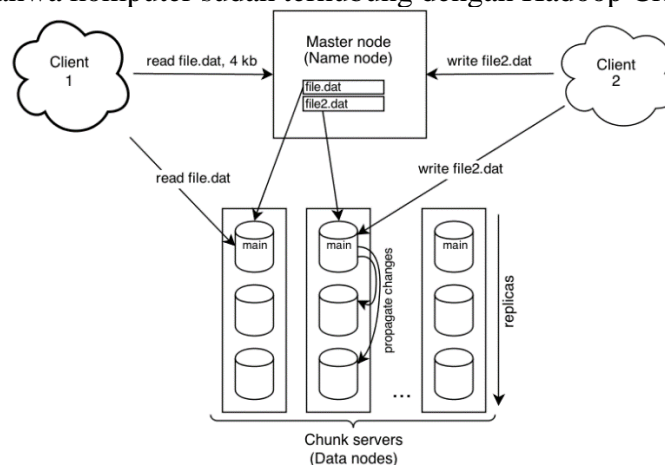
Big Data saat ini menjadi tren teknologi yang signifikan. Salah satu platform perangkat lunak yang dapat digunakan untuk mengelola big data adalah Hadoop. Secara umum, Hadoop adalah perangkat lunak yang memungkinkan penggabungan satu hingga banyak komputer agar dapat bekerja sama dan terhubung untuk menyimpan serta mengelola data secara terpadu [12]. Biasanya Hadoop menggunakan model pemrograman Hive atau MapReduce untuk menyimpan dan memproses Big Data.



Gambar 2. Arsitektur Hadoop

2.3 Hadoop Distributed File System (HDFS)

Hadoop Distributed File System adalah salah satu pengembangan dari Apache, yaitu subtask dari Apache Hadoop. HDFS ini dikembangkan berdasarkan konsep dari GFS (Google File System). Pada Gambar 2 terdapat cara kerja HDFS, yaitu membagi file yang besar menjadi sejumlah bagian kecil sehingga membentuk cluster yang memungkinkan terjadinya pemrosesan secara paralel. Adapun komponen utama dari HDFS yaitu NameNode dan DataNode. NameNode bertugas menjadi master yaitu berkewajiban menyimpan informasi mengenai lokasi penempatan blok data dalam Hadoop Cluster. Sedangkan DataNode bertugas menjadi slave, yaitu mengamankan blok data yang diberikan kepadanya, serta melaporkan kondisinya secara berkala kepada NameNode. Untuk menyimpan data ke dalam HDFS kita perlu memastikan bahwa komputer sudah terhubung dengan Hadoop Cluster [13].



Gambar 3. Arsitektur HDFS

2.4 Dataset

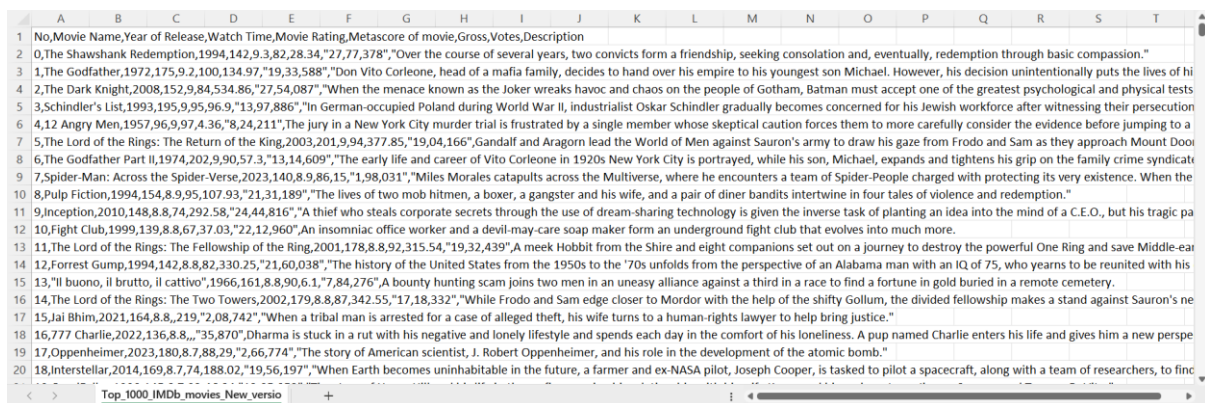
Dataset adalah sekumpulan data. Kumpulan data ini biasanya direpresentasikan dalam bentuk tabel yang dimana setiap baris dan kolom menggambarkan variabel tertentu. Data tersebut diperoleh melalui pengukuran, pengamatan, studi atau analisis. Semakin akurat data yang dikumpulkan, analisis yang dibuat berdasarkan informasi tersebut lebih dapat dipercaya [14]. Terdapat dua jenis dataset :

1. Dataset Privat

Dataset privat merujuk pada kumpulan data yang dapat diambil dari organisasi yang menjadi fokus penelitian. Contohnya mencakup instansi, rumah sakit, dan perusahaan jasa [15].

2. Dataset Publik

Dataset publik merujuk pada kumpulan data yang diambil dari lembaga publik yang telah mendapatkan persetujuan dari para peneliti [16].



No	Movie Name	Year of Release	Watch Time	Movie Rating	Metascore of movie	Gross	Votes	Description	
0	The Shawshank Redemption	1994	142,9	3,82	28,34	27,77	378	"Over the course of several years, two convicts form a friendship, seeking consolation and, eventually, redemption through basic compassion."	
1	The Godfather	1972	175,9	2,100	134,97	19,33	588	"Don Vito Corleone, head of a mafia family, decides to hand over his empire to his youngest son Michael. However, his decision unintentionally puts the lives of his	
2	The Dark Knight	2008	152,9	84,534	86	27,54	087	"When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one of the greatest psychological and physical tests	
3	Schindler's List	1993	195,9	95,96	9	13,97	886	"In German-occupied Poland during World War II, industrialist Oskar Schindler gradually becomes concerned for his Jewish workforce after witnessing their persecution	
4	12 Angry Men	1957	96,9	97	4	36	8,24	211	"The jury in a New York City murder trial is frustrated by a single member whose skeptical caution forces them to more carefully consider the evidence before jumping to a
5	The Lord of the Rings: The Return of the King	2003	201,9	94,377	85	19,04	166	Gandalf and Aragorn lead the World of Men against Sauron's army to draw his gaze from Frodo and Sam as they approach Mount Doom	
6	The Godfather Part II	1974	202,9	90,57	3	13,14	609	"The early life and career of Vito Corleone in 1920s New York City is portrayed, while his son, Michael, expands and tightens his grip on the family crime syndicate	
7	Spider-Man: Across the Spider-Verse	2023	140,8	9,86	15	1,98	031	"Miles Morales catapults across the Multiverse, where he encounters a team of Spider-People charged with protecting its very existence. When the	
8	Pulp Fiction	1994	154,8	9,95	107	93	21,31	189	"The lives of two mob hitmen, a boxer, a gangster and his wife, and a pair of diner bandits intertwine in four tales of violence and redemption."
9	Inception	2010	148,8	74,292	58	24,44	816	"A thief who steals corporate secrets through the use of dream-sharing technology is given the inverse task of planting an idea into the mind of a C.E.O., but his tragic pa	
10	Fight Club	1999	139,8	8,67	37	03	22,12	960	"An insomniac office worker and a devil-may-care soap maker form an underground fight club that evolves into much more."
11	The Lord of the Rings: The Fellowship of the Ring	2001	178,8	8,92	315	54	19,32	439	"A meek Hobbit from the Shire and eight companions set out on a journey to destroy the powerful One Ring and save Middle-earth
12	Forrest Gump	1994	142,8	8,82	330	25	21,60	038	"The history of the United States from the 1950s to the '70s unfolds from the perspective of an Alabama man with an IQ of 75, who years to be reunited with his
13	"Il buono, il brutto, il cattivo"	1966	161,8	8,90	6,1	7,84	276	"A bounty hunting scam joins two men in an uneasy alliance against a third in a race to find a fortune in gold buried in a remote cemetery."	
14	The Lord of the Rings: The Two Towers	2002	179,8	8,87	342	55	17,18	332	"While Frodo and Sam edge closer to Mordor with the help of the shifty Gollum, the divided fellowship makes a stand against Sauron's ne
15	Jai Bhim	2021	164,8	8,219	2,08	742	"When a tribal man is arrested for a case of alleged theft, his wife turns to a human-rights lawyer to help bring justice."		
16	777 Charlie	2022	136,8	8,8	35,870	"Dharma is stuck in a rut with his negative and lonely lifestyle and spends each day in the comfort of his loneliness. A pup named Charlie enters his life and gives him a new perspe			
17	Oppenheimer	2023	180,8	7,88	29	2,66	774	"The story of American scientist, J. Robert Oppenheimer, and his role in the development of the atomic bomb."	
18	Interstellar	2014	169,8	7,74	188	02	19,56	197	"When Earth becomes uninhabitable in the future, a farmer and ex-NASA pilot, Joseph Cooper, is tasked to pilot a spacecraft, along with a team of researchers, to find

Gambar 4. Dataset TOP 1000 IMDb Movies

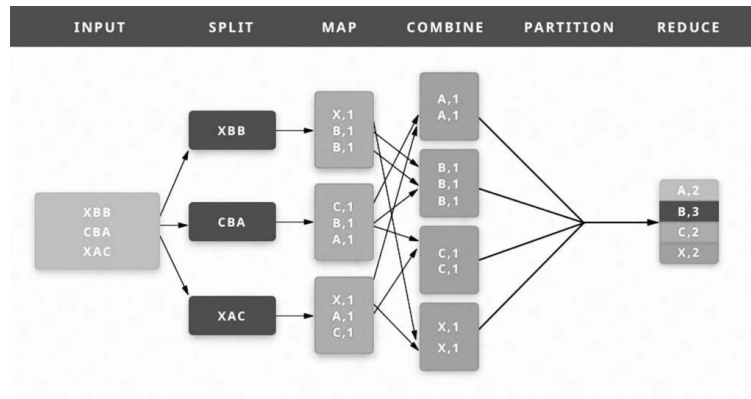
Pada penelitian ini, kami mengambil data melalui Kaggle. Kaggle merupakan platform yang memungkinkan untuk mengeksplorasi berbagai sumber data yang luas dan beragam. Seperti pada Gambar 3. Kami menggunakan dataset TOP 1000 IMDb Movies. Data tersebut merupakan data mentah yang belum diolah menggunakan Hadoop.

2.5 Python

Python merupakan bahasa pemrograman tingkat tinggi yang populer. Python dibuat oleh Guido Van Rossum pada awal 1990-an. Python menyediakan perpustakaan yang komprehensif, memungkinkan pengembang untuk membuat aplikasi canggih dengan menggunakan source code yang tampak sederhana [17]. Python juga dapat dikatakan sebagai bahasa pemrograman yang dapat menyatukan formula kode yang jelas dan juga memiliki sebuah fungsionalitas Pustaka standar yang cukup banyak dan bersifat menyeluruh [18]. Dalam penelitian ini, Python akan menjadi bahasa pemrograman yang digunakan dalam pengurutan 1000 film teratas.

2.6 MapReduce

Metode pengolahan data yang dikenal sebagai MapReduce, dilakukan dengan cara membagi data menjadi potongan-potongan kecil, kemudian hasilnya disatukan kembali. MapReduce dimanfaatkan untuk mengolah data besar yang tersimpan dalam sistem file Hadoop (HDFS) [19]. Gambar 4 di bawah merupakan gambaran dalam pelaksanaan MapReduce yaitu tahap Map dan Reduce. Tahap Map berperan dalam membaca input dalam format pasangan Key/Value. Hasilnya akan dibaca dalam format pasangan Key/Value, kemudian dikelompokkan berdasarkan Key yang sama, proses ini disebut sebagai tahap Reduce [20].



Gambar 5. Arsitektur MapReduce

3. METODELOGI PENELITIAN

3.1. Program yang Digunakan

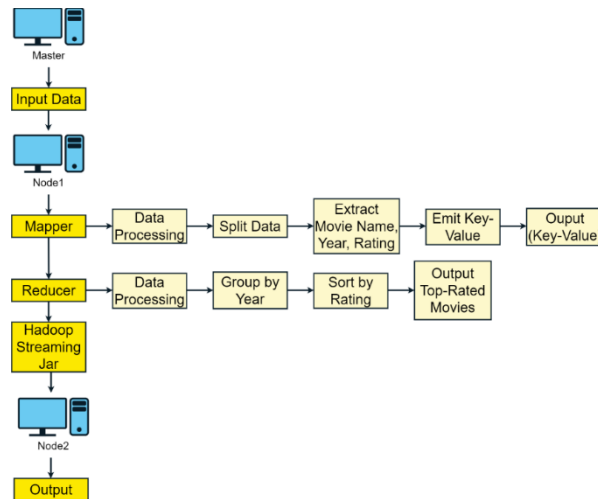
Dalam menganalisis dataset IMDb top 1000 movies ini kami menggunakan dua komponen utama, yakni Apache Hadoop dan bahasa pemrograman Python. Apache Hadoop berfungsi sebagai kerangka kerja pengolahan data besar yang menyediakan sistem penyimpanan terdistribusi melalui Hadoop Distributed File System (HDFS) dan proses pemrosesan data melalui MapReduce. Dalam melakukan MapReduce, Python digunakan untuk menulis kode Mapper dan Reducer, yang mana bertanggung jawab untuk memproses data input dan menghasilkan output.

3.2. Data yang digunakan pada program

Data yang kami gunakan dalam program ini diperoleh melalui sumber yang dapat diandalkan, yaitu <https://www.kaggle.com/datasets/inductiveinks/top-1000-imdb-movies-dataset>. Dataset ini merupakan referensi utama kami dalam analisis ini. Dataset ini mencakup beragam atribut terkait film seperti judul, tahun rilis, durasi, rating IMDb, dan informasi lainnya yang relevan. Sumber data ini dipilih karena keberagaman informasi yang tersedia dan relevansinya dengan tujuan analisis kami. Dengan menggunakan data ini, kami bertujuan untuk melakukan analisis statistik sesuai dengan konteks program yang kami lakukan.

3.3 Cara kerja program

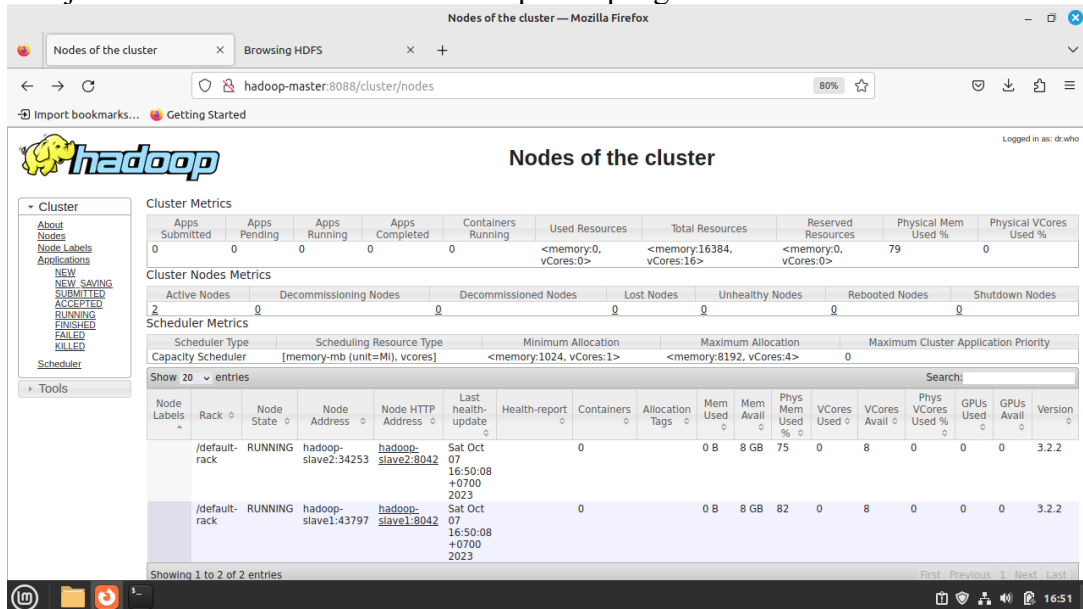
- Proses Data



Gambar 6. Flowchart Proses Data

- Multinode Cluster

Dalam menjalankan proses pengolahan data ini, kami memanfaatkan sebuah multinode cluster yang terdiri dari tiga node, yaitu master, slave1, dan slave2. Ketiga node ini bekerja bersama-sama untuk melakukan proses pengolahan data.



Gambar 7. Multinode Cluster

- Memulai Hadoop

Buka terminal lalu ketik perintah 'start-all.sh', kita dapat melihat status dari semua proses java yang sedang berjalan pada setiap node dengan perintah 'jps'.

```
hadoopuser@hadoop-master: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-master:~$ start-all.sh  
WARNING: Attempting to start all Apache Hadoop daemons as hadoopuser in 10 seconds.  
WARNING: This is not a recommended production deployment configuration.  
WARNING: Use CTRL-C to abort.  
Starting namenodes on [hadoop-master]  
Starting datanodes  
Starting secondary namenodes [hadoop-master]  
Starting resourcemanager  
Starting nodemanagers  
hadoopuser@hadoop-master:~$ jps  
7072 NameNode  
7443 ResourceManager  
7731 Jps  
7261 SecondaryNameNode
```

Gambar 8. Memulai Hadoop & JPS Master

```
hadoopuser@hadoop-slave1: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-slave1:~$ jps  
5650 NodeManager  
5763 Jps  
5547 DataNode
```

Gambar 9. JPS Slave1

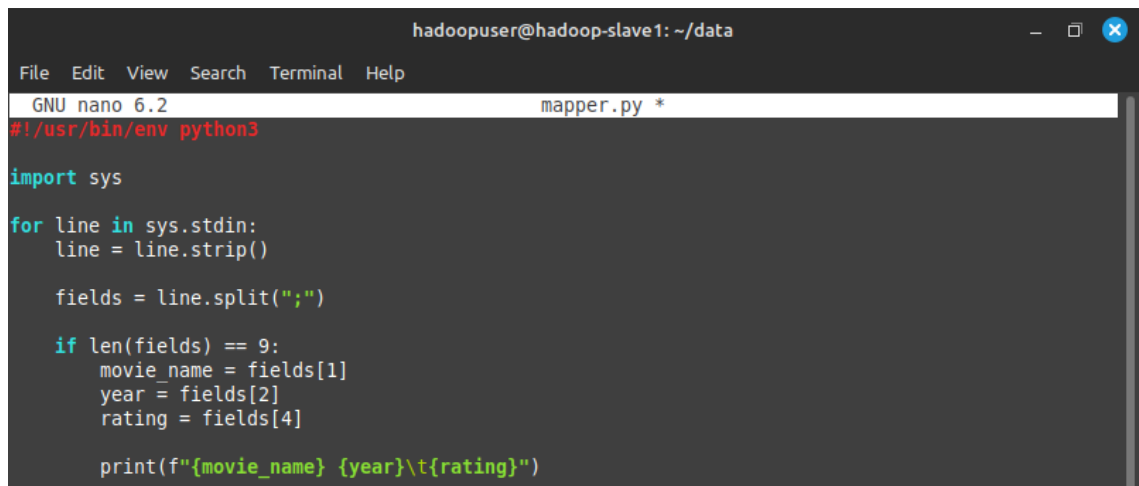
```
hadoopuser@hadoop-slave2: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-slave2:~$ jps  
5441 NodeManager  
5555 Jps  
5337 DataNode
```

Gambar 10. JPS Slave2

- Penginputan Data
Pertama, buatlah sebuah direktori untuk menyimpan data.

```
hadoopuser@hadoop-master:~$ hadoop fs -mkdir /data  
Lalu input data dari local system node master ke direktori Hadoop yang baru dibuat  
dengan perintah '-copyFromLocal'.  
hadoopuser@hadoop-master:~$ hadoop fs -copyFromLocal /home/hadoopuser/data/Top_1000_IMDb_movies_New_version.csv /data  
Kita dapat melihat bahwa data tersebut sudah diinput dengan perintah '-ls /data'  
hadoopuser@hadoop-master:~$ hadoop fs -ls /data  
Found 1 items  
-rw-r--r-- 2 hadoopuser supergroup 204174 2023-10-07 17:02 /data/Top_1000_IMDb_movies_New_version.csv
```

- Proses Mapreduce
Proses mapreduce kami lakukan pada node slave1 dengan tiga proses yaitu membuat kode mapper dan reducer dengan bahasa pemrograman python dan menggunakan tools Hadoop Streaming Jar untuk menjalankan tugas pemrosesan pada kode mapper dan reducer.
 - Mapper



```
hadoopuser@hadoop-slave1: ~/data
File Edit View Search Terminal Help
GNU nano 6.2 mapper.py *
#!/usr/bin/env python3

import sys

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

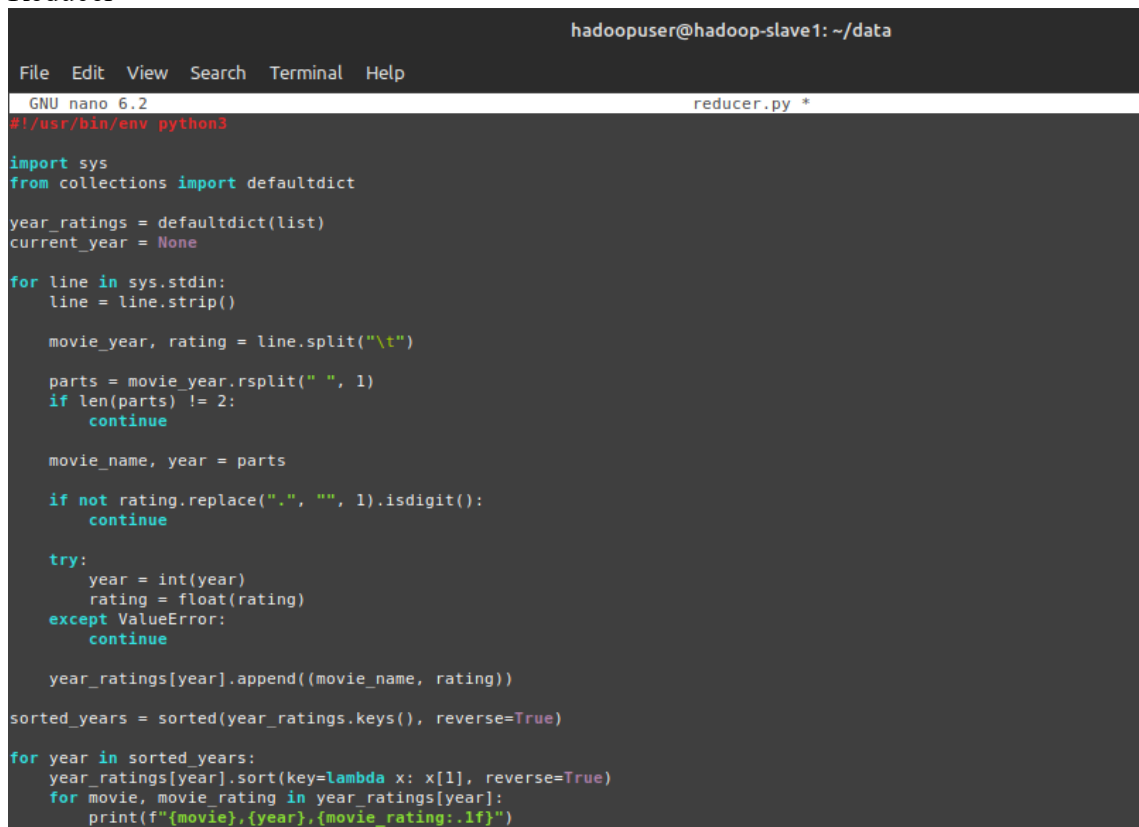
    fields = line.split(";")

    if len(fields) == 9:
        movie_name = fields[1]
        year = fields[2]
        rating = fields[4]

        print(f"{movie_name} {year}\t{rating}")
```

Gambar 11. Proses Mapper

- Reducer



```
hadoopuser@hadoop-slave1: ~/data
File Edit View Search Terminal Help
GNU nano 6.2 reducer.py *
#!/usr/bin/env python3

import sys
from collections import defaultdict

year_ratings = defaultdict(list)
current_year = None

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

    movie_year, rating = line.split("\t")

    parts = movie_year.rsplit(" ", 1)
    if len(parts) != 2:
        continue

    movie_name, year = parts

    if not rating.replace(".", "", 1).isdigit():
        continue

    try:
        year = int(year)
        rating = float(rating)
    except ValueError:
        continue

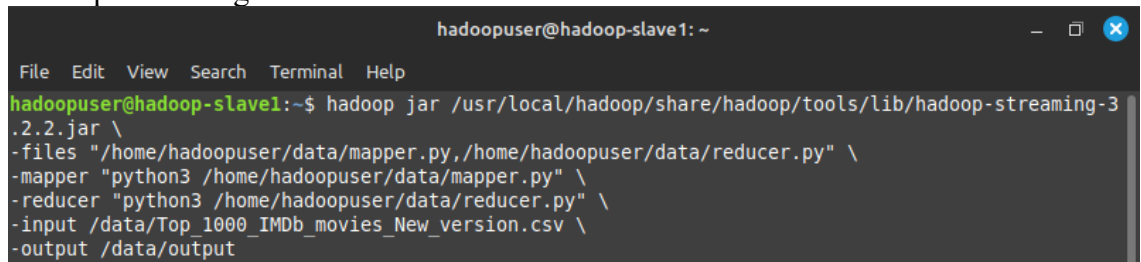
    year_ratings[year].append((movie_name, rating))

sorted_years = sorted(year_ratings.keys(), reverse=True)

for year in sorted_years:
    year_ratings[year].sort(key=lambda x: x[1], reverse=True)
    for movie, movie_rating in year_ratings[year]:
        print(f"{movie},{year},{movie_rating:.1f}")
```

Gambar 12. Proses Reducer

- Hadoop Streaming Jar

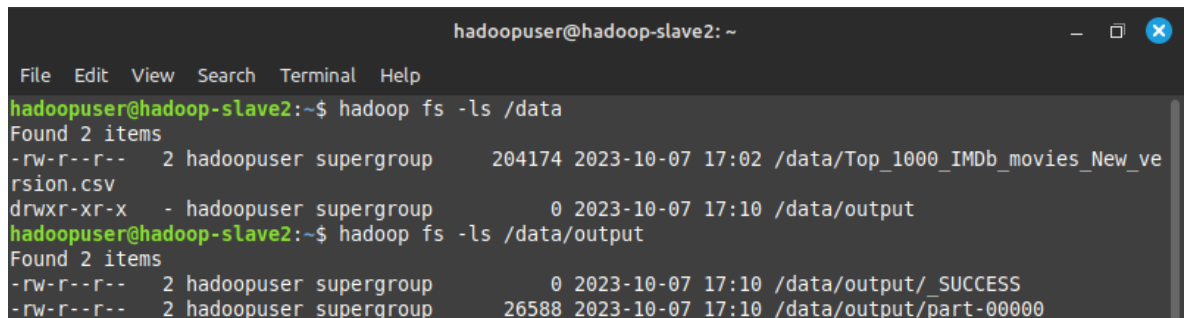


```
hadoopuser@hadoop-slave1: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-slave1:~$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3  
.2.2.jar \  
-files "/home/hadoopuser/data/mapper.py,/home/hadoopuser/data/reducer.py" \  
-mapper "python3 /home/hadoopuser/data/mapper.py" \  
-reducer "python3 /home/hadoopuser/data/reducer.py" \  
-input /data/Top_1000_IMDb_movies_New_version.csv \  
-output /data/output
```

Gambar 13. Hadoop Streaming Jar

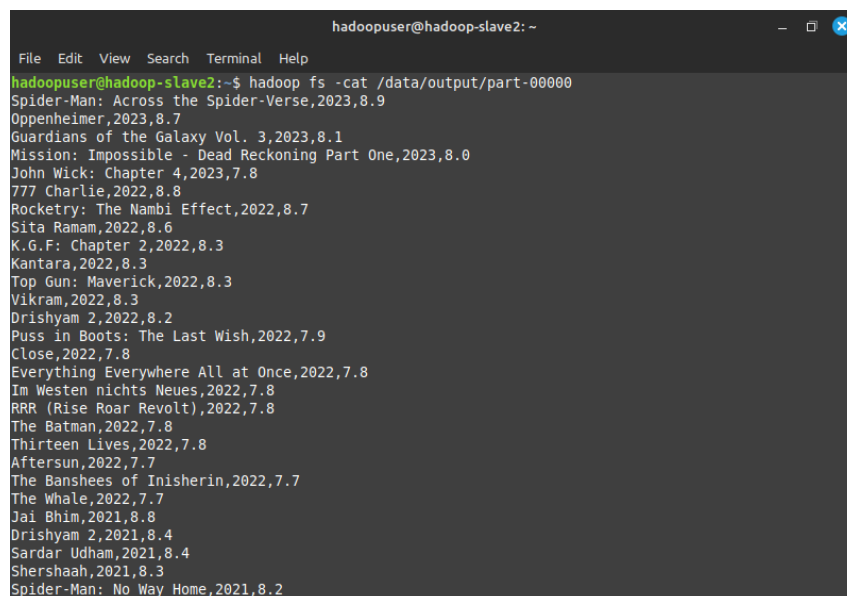
- Output

Setelah proses mapreduce berhasil dijalankan maka akan ada dua file yang muncul pada direktori Hadoop yang telah ditentukan untuk menyimpan output, yaitu file bernama `_SUCCESS` yang menandakan bahwa proses mapreduce telah berhasil selesai tanpa kesalahan dan file `part-xxxxx` yang merupakan output yang dihasilkan oleh mapreduce. Kami menggunakan node slave2 untuk melihat output, menggunakan perintah `'hadoop fs -cat /data/output/part-00000'` untuk melihat isi dari file tersebut.



```
hadoopuser@hadoop-slave2: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-slave2:~$ hadoop fs -ls /data  
Found 2 items  
-rw-r--r--  2 hadoopuser supergroup    204174 2023-10-07 17:02 /data/Top_1000_IMDb_movies_New_ve  
rsion.csv  
drwxr-xr-x  - hadoopuser supergroup      0 2023-10-07 17:10 /data/output  
hadoopuser@hadoop-slave2:~$ hadoop fs -ls /data/output  
Found 2 items  
-rw-r--r--  2 hadoopuser supergroup      0 2023-10-07 17:10 /data/output/_SUCCESS  
-rw-r--r--  2 hadoopuser supergroup    26588 2023-10-07 17:10 /data/output/part-00000
```

Gambar 14. Direktori Output



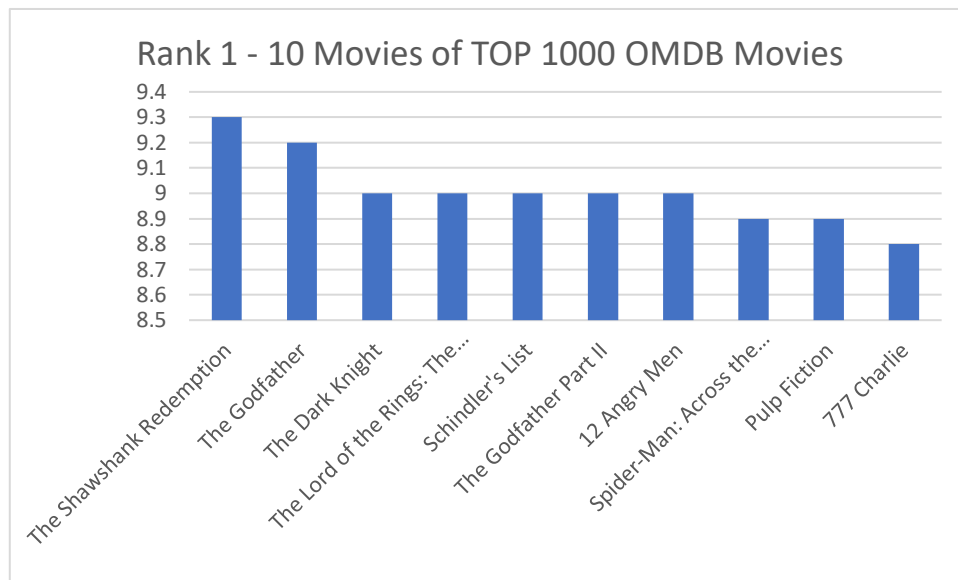
```
hadoopuser@hadoop-slave2: ~  
File Edit View Search Terminal Help  
hadoopuser@hadoop-slave2:~$ hadoop fs -cat /data/output/part-00000  
Spider-Man: Across the Spider-Verse,2023,8.9  
Oppenheimer,2023,8.7  
Guardians of the Galaxy Vol. 3,2023,8.1  
Mission: Impossible - Dead Reckoning Part One,2023,8.0  
John Wick: Chapter 4,2023,7.8  
777 Charlie,2022,8.8  
Rocketry: The Nambi Effect,2022,8.7  
Sita Ramam,2022,8.6  
K.G.F: Chapter 2,2022,8.3  
Kantara,2022,8.3  
Top Gun: Maverick,2022,8.3  
Vikram,2022,8.3  
Drishyam 2,2022,8.2  
Puss in Boots: The Last Wish,2022,7.9  
Close,2022,7.8  
Everything Everywhere All at Once,2022,7.8  
Im Westen nichts Neues,2022,7.8  
RRR (Rise Roar Revolt),2022,7.8  
The Batman,2022,7.8  
Thirteen Lives,2022,7.8  
Aftersun,2022,7.7  
The Banshees of Inisherin,2022,7.7  
The Whale,2022,7.7  
Jai Bhim,2021,8.8  
Drishyam 2,2021,8.4  
Sardar Udham,2021,8.4  
Shershaah,2021,8.3  
Spider-Man: No Way Home,2021,8.2
```

Gambar 15. Hasil Tampilan Output

4. HASIL PENELITIAN

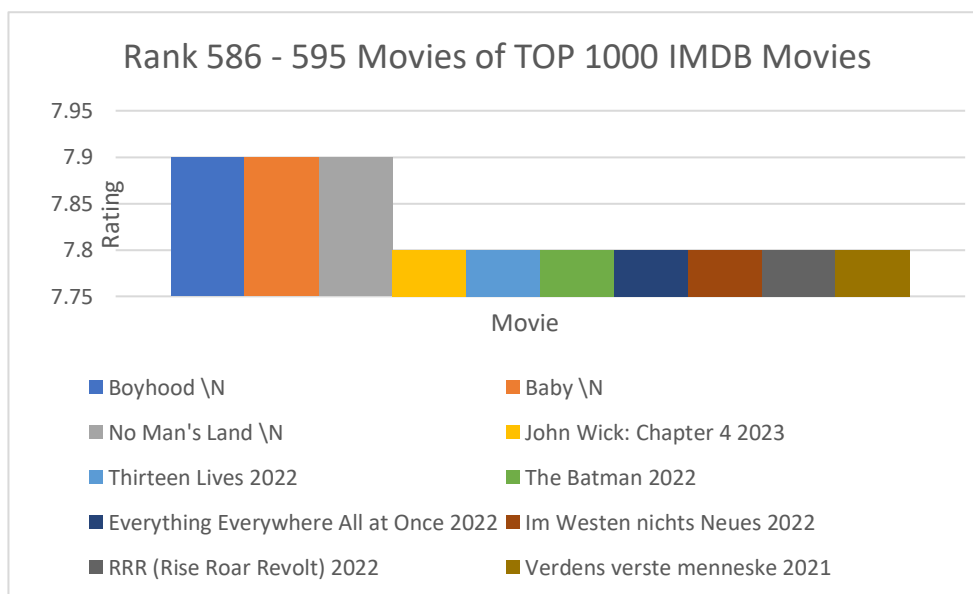
Berdasarkan pengujian yang dilakukan dengan menginput file data yang belum terstruktur ke dalam Apache Hadoop. Isi data berjumlah 1000 dengan besar 179 Kb menghasilkan hasil akhir seperti pada gambar 6. Kami merepresentasikan hasil data tersebut pada chart di bawah ini.

- Diagram batang di bawah menunjukkan 10 film tertinggi dari total keseluruhan 1000 data.



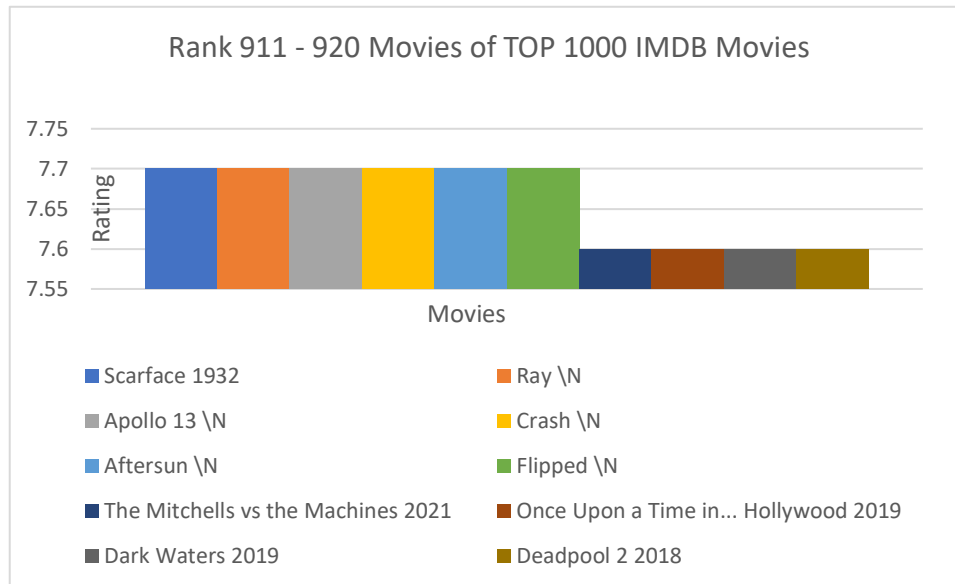
Gambar 16. Top Movie

- Diagram batang di bawah menunjukkan peringkat film dari 586 - 595 dari total keseluruhan 1000 data.



Gambar 17. Rangkaing 586-595

- Diagram batang di bawah menunjukkan peringkat film dari 911 - 920 dari total keseluruhan 1000 data.



Gambar 18. Rangkaian 911-920

5. KESIMPULAN

Apache Hadoop sebagai kerangka kerja pengolahan data besar yang menyediakan sistem penyimpanan terdistribusi melalui Hadoop Distributed File System (HDFS) dan model pemrograman MapReduce untuk pemrosesan secara paralel. Dalam melakukan MapReduce, Python digunakan untuk menulis kode Mapper dan Reducer, yang bertanggung jawab untuk memproses data input dan menghasilkan output. Hasil penelitian ini berfokus pada rekomendasi film, memberikan manfaat kepada pengguna untuk menemukan rekomendasi film berdasarkan urutan ratingnya.

Dari 1000 data yang dianalisis. Ternyata, setelah diolah, terdapat beberapa data yang rusak, kerusakannya bisa terletak pada tahun terbit film ataupun ratingnya. Sehingga data yang diambil untuk divisualisasikan, minimal hanyalah yang tidak ada tahunnya. Kami memvisualisasikan pengelompokan 1000 data tersebut kedalam 3 kluster, yaitu kluster 1 (rank 1 – 10), kluster 2 (rank 586 – 595), dan kluster 3 (rank 911 – 920). Untuk kluster pertama rating tertinggi diraih oleh film “The Shawshank Redemption – 1994” dengan rating 9,3. Sedangkan di kluster kedua ada film “No Man’s Land” dengan rating 7,9 dan “John Wick 4 – 2023” dengan rating 7,8. Dan untuk kluster terakhir ada film “Scarface – 1932” dengan rating 7,7 dan “Deadpool 2 – 2018” dengan rating 7,6.

DAFTAR PUSTAKA

- [1] N. U. Shahid and N. J. Sheikh, “Impact of Big Data on Innovation, Competitive Advantage, Productivity, and Decision Making: Literature Review,” *Open Journal of Business and Management*, vol. 09, no. 02, pp. 586–617, 2021, doi: 10.4236/ojbm.2021.92032.

- [2] E. Riyandani, “‘Big Data vs Big Information vs Big Knowledge’ Oleh: Imam Cholissodin,” 2016. [Online]. Available: <http://bit.ly/2x8ta9S>
- [3] M. Febriansyah Trisnadi, S. Al Faraby, and M. Dwifebri, “Sentiment Analysis pada Movie Review Menggunakan Feature Selection Mutual Information dan K-Nearest Neighbour Classifier.”
- [4] M. Febriansyah Trisnadi, S. Al Faraby, and M. Dwifebri, “Sentiment Analysis pada Movie Review Menggunakan Feature Selection Mutual Information dan K-Nearest Neighbour Classifier.”
- [5] O. Azeroual and R. Fabre, “Processing Big Data with Apache Hadoop in the Current Challenging Era of COVID-19,” *Big Data and Cognitive Computing*, vol. 5, no. 1, p. 12, Mar. 2021, doi: 10.3390/bdcc5010012.
- [6] G. Ravichandran, “Big Data Processing with Hadoop : A Review,” *International Research Journal of Engineering and Technology*, 2017, [Online]. Available: www.irjet.net
- [7] A. Ghaffar Shoro and T. Rahim Soomro, “Big Data Analysis: Apache Spark Perspective,” 2015. [Online]. Available: <https://www.researchgate.net/publication/272825265>
- [8] S. F. Astika, M. Jauhari, N. Isbatuzzin, M. Salman, and K. Ramli, “BUILDING A DYNAMIC SCALABLE PARALLEL CLOUD-BASED SNORT NIDS USING CONTAINERS AND BIG DATA,” *Journal of Southwest Jiaotong University*, vol. 56, no. 5, pp. 317–326, Oct. 2021, doi: 10.35741/issn.0258-2724.56.5.27.
- [9] M. Riza *et al.*, “Relational Data Modeling on the Document-Based NoSQL,” 2022. [Online]. Available: <https://www.researchgate.net/publication/367558756>
- [10] E. R. E. Sirait, “IMPLEMENTASI TEKNOLOGI BIG DATA DI LEMBAGA PEMERINTAHAN INDONESIA,” *Jurnal Penelitian Pos dan informatika*, vol. 6, no. 2, p. 113, Dec. 2016, doi: 10.17933/jppi.2016.060201.
- [11] S. Oliviani, A. B. Osmond, and R. Latuconsina, “IMPLEMENTASI APACHE SPARK PADA BIG DATA BERBASIS HADOOP DISTRIBUTED FILE SYSTEM IMPLEMENTATION APACHE SPARK ON BIG DATA BASED HADOOP DISTRIBUTED FILE SYSTEM.”
- [12] Y. Surahman and H. Saptono, “Implementasi dan Analisis Kinerja HDFS sebagai Infrastruktur Pembangunan Big Data,” *Jurnal Informatika Terpadu*, vol. 4, no. 2, pp. 63–70, Sep. 2018, doi: 10.54914/jit.v4i2.159.
- [13] Y. Surahman and H. Saptono, “Implementasi dan Analisis Kinerja HDFS sebagai Infrastruktur Pembangunan Big Data,” *Jurnal Informatika Terpadu*, vol. 4, no. 2, pp. 63–70, Sep. 2018, doi: 10.54914/jit.v4i2.159.
- [14] A. M. T I Sambu Ua *et al.*, “Penggunaan Bahasa Pemrograman Python Dalam Analisis Faktor Penyebab Kanker Paru-Paru Universitas Bina Nusantara,” *Jurnal Publikasi Teknik Informatika (JUPTI)*, vol. 2, no. 2, 2023, doi: 10.55606/jupti.v2i2.1742.
- [15] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, “‘Everything is data’: towards one big data ecosystem using multiple sources of data on higher education in Indonesia,” *J Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00639-7.
- [16] A. M. T I Sambu Ua *et al.*, “Penggunaan Bahasa Pemrograman Python Dalam Analisis Faktor Penyebab Kanker Paru-Paru Universitas Bina Nusantara,” *Jurnal Publikasi Teknik Informatika (JUPTI)*, vol. 2, no. 2, 2023, doi: 10.55606/jupti.v2i2.1742.
- [17] C. Wibawa, S. Wirawan, M. Mustikasari, and D. T. Anggraeni, “KOMPARASI KECEPATAN HADOOP MAPREDUCE DAN APACHE SPARK DALAM

- MENGOLAH DATA TEKS,” *Jurnal Ilmiah Matrik*, vol. 24, no. 1, pp. 10–20, Apr. 2022, doi: 10.33557/jurnalmatrik.v24i1.1649.
- [18] C. Wibawa, S. Wirawan, M. Mustikasari, and D. T. Anggraeni, “KOMPARASI KECEPATAN HADOOP MAPREDUCE DAN APACHE SPARK DALAM MENGOLAH DATA TEKS,” *Jurnal Ilmiah Matrik*, vol. 24, no. 1, pp. 10–20, Apr. 2022, doi: 10.33557/jurnalmatrik.v24i1.1649.
- [19] Y. Surahman and H. Saptono, “Implementasi dan Analisis Kinerja HDFS sebagai Infrastruktur Pembangunan Big Data,” *Jurnal Informatika Terpadu*, vol. 4, no. 2, pp. 63–70, Sep. 2018, doi: 10.54914/jit.v4i2.159.
- [20] C. Wibawa, S. Wirawan, M. Mustikasari, and D. T. Anggraeni, “KOMPARASI KECEPATAN HADOOP MAPREDUCE DAN APACHE SPARK DALAM MENGOLAH DATA TEKS,” *Jurnal Ilmiah Matrik*, vol. 24, no. 1, pp. 10–20, Apr. 2022, doi: 10.33557/jurnalmatrik.v24i1.1649.