



HADOOP & JAVA

RADITYA IHSAN DHIAULHAQ 2106733912

RIAN ABRAR MAKARIM LUBIS 2106708242

NAUFAL FEBRIYANTO 2106702674

MICHAEL WINSTON TJAHAJA 2106731270

By K1 Group 7

SPEKIFIKASI LAPTOP



CPU: AMD Ryzen 7 5800H

GPU: NVIDIA GeForce RTX 3060 Mobile 6GB

SSD: SSD M2 PCI-E Gen 3.0 1x1024 GB

RAM: 16 GB DDR 4 3200 MHz

OS: Windows 11 Home 64-bit

TUTORIAL MENGGUNAKAN HADOOP

Melakukan Inisialisasi awal

```
radityadito@LAPTOP-LPITGH4M:~$ cd $HADOOP_HOME
```

```
radityadito@LAPTOP-LPITGH4M:~/hadoop/hadoop-3.3.5$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [LAPTOP-LPITGH4M]
radityadito@LAPTOP-LPITGH4M:~/hadoop/hadoop-3.3.5$ sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

Menambah file input

```
radityadito@LAPTOP-LPITGH4M:~/hadoop/hadoop-3.3.5$ hdfs dfs -put README.txt input
```

TUTORIAL MENGGUNAKAN HADOOP

Menjalankan Wordcount


```
radityadito@LAPTOP-LPITGH4M:~/hadoop/hadoop-3.3.5$ bin/yarn jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.5.jar wordcount input output
2023-06-22 03:47:55,013 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-22 03:47:55,673 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/radityadito/.staging/job_1687358360205_0005
2023-06-22 03:47:55,940 INFO input.FileInputFormat: Total input files to process : 1
2023-06-22 03:47:56,002 INFO mapreduce.JobSubmitter: number of splits:1
2023-06-22 03:47:56,567 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1687358360205_0005
2023-06-22 03:47:56,567 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-22 03:47:56,770 INFO conf.Configuration: resource-types.xml not found
2023-06-22 03:47:56,770 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-22 03:47:57,347 INFO impl.YarnClientImpl: Submitted application application_1687358360205_0005
2023-06-22 03:47:57,437 INFO mapreduce.Job: The url to track the job: http://LAPTOP-LPITGH4M.:8088/proxy/application_1687358360205_0005/
2023-06-22 03:47:57,438 INFO mapreduce.Job: Running job: job_1687358360205_0005
2023-06-22 03:48:04,536 INFO mapreduce.Job: Job job_1687358360205_0005 running in uber mode : false
2023-06-22 03:48:04,537 INFO mapreduce.Job: map 0% reduce 0%
2023-06-22 03:48:08,595 INFO mapreduce.Job: map 100% reduce 0%
2023-06-22 03:48:13,621 INFO mapreduce.Job: map 100% reduce 100%
2023-06-22 03:48:13,627 INFO mapreduce.Job: Job job_1687358360205_0005 completed successfully
2023-06-22 03:48:13,705 INFO mapreduce.Job: Counters: 54
    File System Counters
      FILE: Number of bytes read=254
      FILE: Number of bytes written=552587
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=295
```

TUTORIAL MENGGUNAKAN HADOOP

Menampilkan Output

```
radityadito@LAPTOP-LPITGH4M:~/hadoop/hadoop-3.3.5$ hdfs dfs -cat output/part-r-00000
For      1
Hadoop,  1
about    1
and      1
at:      2
```

Cek status dan time pada localhost:8088



FINISHED Applications

Cluster Metrics											
Apps Submitted		Apps Pending		Apps Running		Apps Completed		Containers Running		Used Resources	
5		0		0		5		0		<memory:0 B, vCores:0>	

Cluster Nodes Metrics					
Active Nodes		Decommissioning Nodes		Decommissioned Nodes	
1		0		0	

Scheduler Metrics											
Scheduler Type		Scheduling Resource Type				Minimum Allocation					
Capacity Scheduler		[memory-mb (unit=Mb), vcores]				<memory:1024, vCores:1>					

Show 20 entries






ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
application_1687358360205_0005	radityadito	word count	MAPREDUCE		default	0	Thu Jun 22 03:47:56 +0700 2023	Thu Jun 22 03:47:58 +0700 2023	Thu Jun 22 03:48:12 +0700 2023	FINISHED	SUCCEEDED

LINK GITHUB

https://github.com/RadityaDito/Hadoop_Vs_Java



TEXT FILES

Name	
 1GB	
 10GB	
 10MB	
 100MB	
 500MB	

JAVA CODE

```
7 public class WordCounter {
8     Run | Debug
9     public static void main(String[] args) {
10         String filePath = "C:\\Users\\Raditya Dito\\Documents\\Universitas Indonesia\\Semester 4\\Sistem Basis Data\\Hadop\\Java Code\\WordCount\\WordCount\\src\\10mb.txt";
11         int iterations = 5;
12         long totalTime = 0;
13
14         for (int i = 0; i < iterations; i++) {
15             long startTime = System.currentTimeMillis();
16
17             try (BufferedReader reader = new BufferedReader(new FileReader(filePath))) {
18                 String line;
19
20                 // Create a word count map
21                 Map<String, Integer> wordCountMap = new HashMap<>();
22
23                 // Read the file line by line
24                 while ((line = reader.readLine()) != null) {
25                     // Clean the text by removing non-alphanumeric characters and converting to lowercase
26                     String cleanedText = cleanText(line);
27
28                     // Split the line into words
29                     String[] words = splitIntoWords(cleanedText);
30
31                     // Update the word count map
32                     updateWordCount(words, wordCountMap);
33                 }
34
35                 // Display the word count results
36                 displayWordCount(wordCountMap);
37
38             } catch (IOException e) {
39                 e.printStackTrace();
40             }
41
42             long endTime = System.currentTimeMillis();
43             long executionTime = endTime - startTime;
44             totalTime += executionTime;
45
46             System.out.println("Iteration " + (i + 1) + " Execution Time: " + executionTime + " milliseconds");
47         }
48
49         double averageTime = (double) totalTime / iterations;
50         System.out.println("Average Execution Time: " + averageTime + " milliseconds");
51     }
52 }
```


JAVA CODE

```
private static String cleanText(String text) {  
    return text.replaceAll(regex:"[^a-zA-Z0-9 ]", replacement:"").toLowerCase();  
}  
  
private static String[] splitIntoWords(String text) {  
    return text.split(regex:" ");  
}  
  
private static void updateWordCount(String[] words, Map<String, Integer> wordCountMap) {  
    for (String word : words) {  
        wordCountMap.put(word, wordCountMap.getOrDefault(word, defaultValue:0) + 1);  
    }  
}  
  
private static void displayWordCount(Map<String, Integer> wordCountMap) {  
    for (Map.Entry<String, Integer> entry : wordCountMap.entrySet()) {  
        String word = entry.getKey();  
        int count = entry.getValue();  
        // System.out.println(word + ": " + count);  
    }  
}
```

SIZE 10MB

```
Iteration 1 Execution Time: 1653 milliseconds
Iteration 2 Execution Time: 1412 milliseconds
Iteration 3 Execution Time: 1403 milliseconds
Iteration 4 Execution Time: 1336 milliseconds
Iteration 5 Execution Time: 1302 milliseconds
Average Execution Time: 1421.2 milliseconds
```

Running Time
Java : 1.42 s
Hadoop : 21 s

Application Overview	
User:	radityadito
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Wed Jun 21 21:50:23 +0700 2023
Launched:	Wed Jun 21 21:50:23 +0700 2023
Finished:	Wed Jun 21 21:50:45 +0700 2023
Elapsed:	21sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

SIZE 100MB

```
Iteration 1 Execution Time: 7415 milliseconds
Iteration 2 Execution Time: 7107 milliseconds
Iteration 3 Execution Time: 7130 milliseconds
Iteration 4 Execution Time: 7075 milliseconds
Iteration 5 Execution Time: 7124 milliseconds
Average Execution Time: 7170.2 milliseconds
```

Running Time

Java : 7.17 s

Hadoop : 24 s

Application Overview	
User:	radityadito
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Fri Jun 09 18:40:53 +0700 2023
Launched:	Fri Jun 09 18:40:54 +0700 2023
Finished:	Fri Jun 09 18:41:17 +0700 2023
Elapsed:	24sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

SIZE 500MB

Iteration 1 Execution Time: 66074 milliseconds
Average Execution Time: 66074.0 milliseconds

Running Time
Java : 66 s
Hadoop : 52 s

Application Overview	
User:	radityadito
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Wed Jun 21 22:10:10 +0700 2023
Launched:	Wed Jun 21 22:10:10 +0700 2023
Finished:	Wed Jun 21 22:11:02 +0700 2023
Elapsed:	52sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

SIZE 1GB

```
"C:\Program Files\Java\jdk-11.0.16.1\bin\java.exe
Iteration 1 Execution Time: 73982 milliseconds
Iteration 2 Execution Time: 75221 milliseconds
Iteration 3 Execution Time: 87175 milliseconds
Iteration 4 Execution Time: 80839 milliseconds
Iteration 5 Execution Time: 74850 milliseconds
Average Execution Time: 78413.4 milliseconds
```

Running Time
Java : 78 s
Hadoop : 76 s

Application Overview

User:	radityadito
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Fri Jun 09 20:15:37 +0700 2023
Launched:	Fri Jun 09 20:15:37 +0700 2023
Finished:	Fri Jun 09 20:16:53 +0700 2023
Elapsed:	1mins, 16sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

SIZE 10GB

```
"C:\Program Files\Java\jdk-11.0.16.1\bin\java.exe
Iteration 1 Execution Time: 716314 milliseconds
Average Execution Time: 716314.0 milliseconds
```

Running Time

Java : 714 s

Hadoop : 354 s

Application Overview	
User:	radityadito
Name:	word count
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Fri Jun 09 21:30:20 +0700 2023
Launched:	Fri Jun 09 21:30:20 +0700 2023
Finished:	Fri Jun 09 21:36:14 +0700 2023
Elapsed:	5mins, 54sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

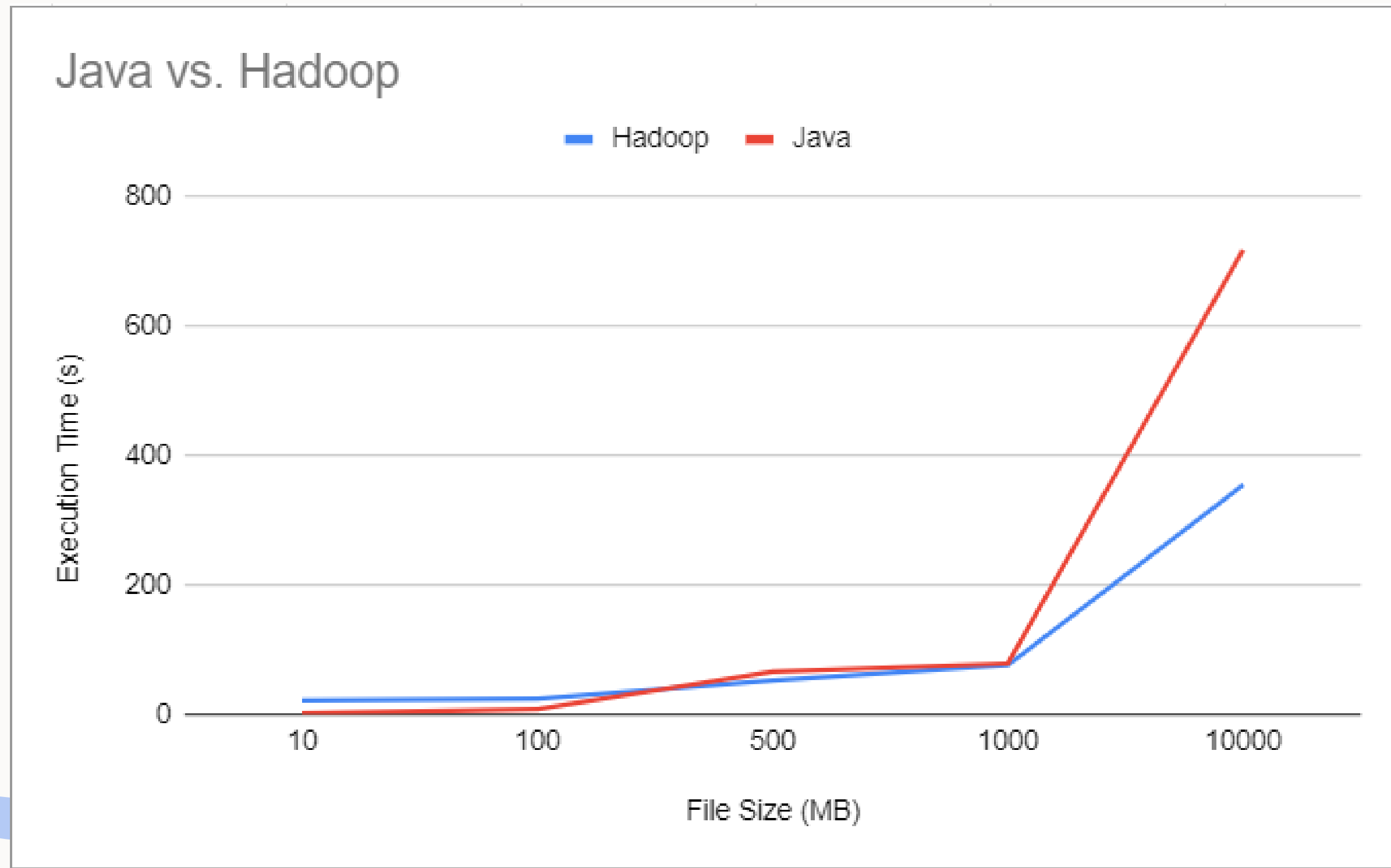
ANALISIS & KESIMPULAN



TABEL DAN GRAFIK

File Size	Time with Framework (seconds)		Ratio
	Java	Hadoop	
10MB	1.42	21	14.78
100MB	7.71	24	3.11
500MB	66	52	0.78
1GB	78	76	0.97
10GB	714	354	0.49

TABEL DAN GRAFIK



ANALISIS

Apabila dibandingkan, wordcount menggunakan Java dan Hadoop dalam hal eksekusi waktu, maka wordcount Java akan lebih cepat karena data yang dihitung disimpan dan diproses di satu mesin atau server tunggal. Sehingga, overhead terkait pengaturan dan komunikasi antar node dalam cluster seperti pada Hadoop hanya sedikit. Selain itu, untuk ukuran file yang relatif kecil (10MB dan 100MB), wordcount dengan Java dapat menyelesaikan tugas lebih cepat karena tidak melibatkan kompleksitas yang signifikan.

Di sisi lain, Hadoop didesain untuk memproses dan menganalisis data dalam skala besar dengan membagi tugas pemrosesan ke beberapa node dalam cluster. Hal ini mengakibatkan terjadinya overhead yang signifikan, meskipun penggunaan Hadoop sangat efektif terutama pada platform terdistribusi dan data dengan ukuran besar.

Pada skala kecil, kinerja Hadoop menjadi lebih lambat daripada wordcount biasa karena membutuhkan waktu untuk menghubungkan node-node dalam cluster dan melakukan komunikasi antar node.

KESIMPULAN

Dalam situasi yang terjadi pada percobaan kali ini, wordcount biasa tidak melibatkan pengaturan dan komunikasi antar node yang rumit sehingga memiliki kinerja yang lebih cepat dibandingkan dengan penggunaan Hadoop. Sehingga, pada skala yang besar dan kebutuhan pemrosesan yang luas, Hadoop menjadi opsi yang lebih optimal.

