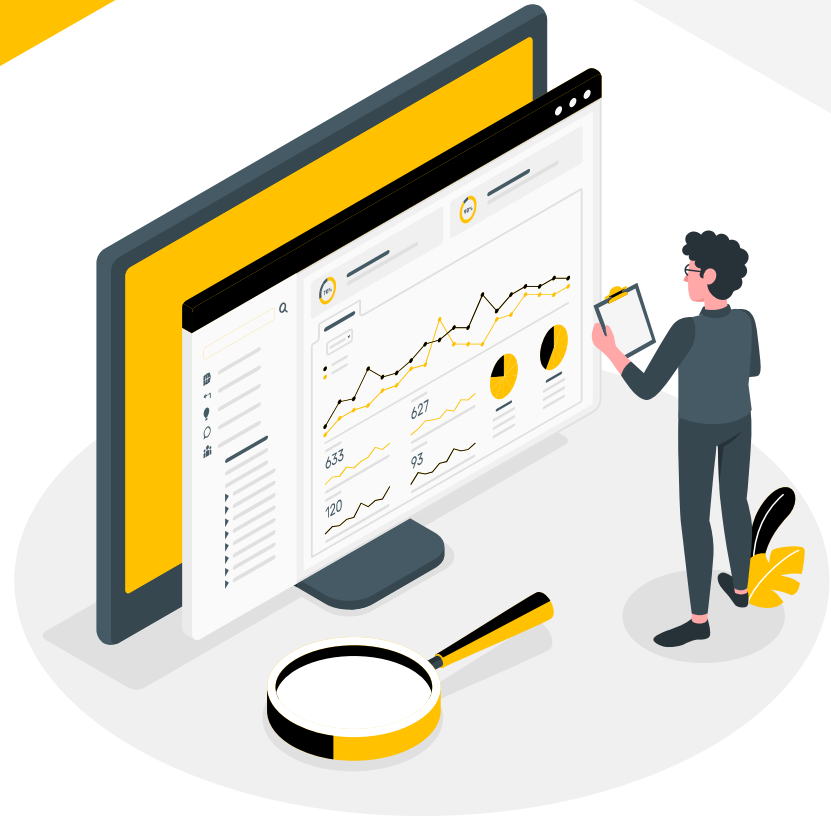
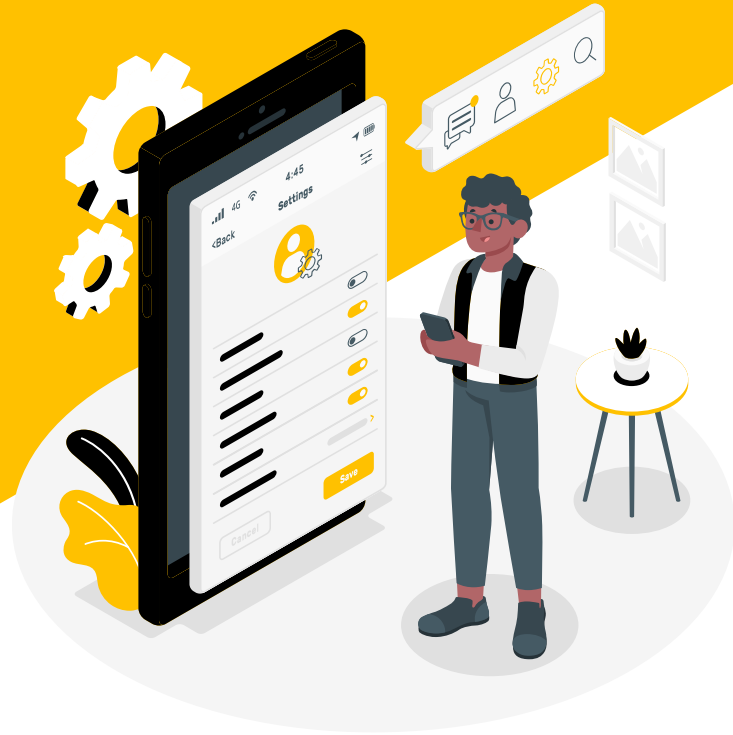


Hadoop

Group 4 - Pagi
Sistem Basis Data





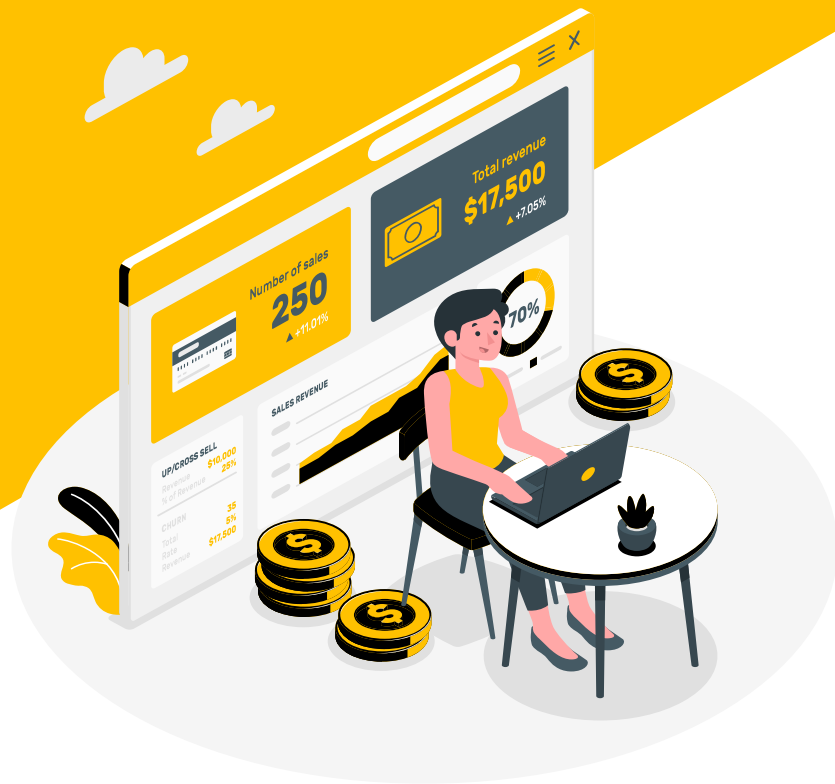
Our Team

Farras Rafi Permana - 2106700990

Zaki Ananda - 2106705474

Andikha Wisanggeni - 2106731503

M Fathan Muhandis - 2106731623



What Is Hadoop?

Introduction to *hadoop*

Hadoop adalah sebuah kerangka kerja (framework) open-source yang digunakan untuk pemrosesan dan penyimpanan data dalam skala besar secara terdistribusi. Hadoop dirancang untuk mengatasi tantangan dalam mengelola dan menganalisis data yang sangat besar (big data) yang tidak dapat ditangani dengan menggunakan sistem tradisional. Framework Hadoop hadir dan memungkinkan pengolahan data lebih banyak, menyimpan data heterogen dan mempercepat proses pengolahannya.

Dengan menggunakan Hadoop, perusahaan dan organisasi dapat memanfaatkan potensi data besar yang mereka miliki untuk mengambil wawasan bisnis yang berharga, melakukan analisis data yang kompleks, dan mengatasi tantangan pemrosesan data dalam skala besar. Hadoop juga memiliki ekosistem yang luas dengan berbagai komponen tambahan, seperti Apache Hive, Apache Pig, Apache HBase, Apache Spark, dan lainnya, yang memperluas kemampuan dan fungsionalitas Hadoop.

How does Hadoop work?

Dalam Hadoop, terdapat empat modul utama yakni **HDFS, YARN, MapReduce, dan Hadoop Common**, berikut penjelasannya:

- **Hadoop Distributed File System (HDFS)** merupakan sistem yang terdistribusi dan beroperasi di hardware standar maupun low-end.
- **Yet Another Resource Negotiator (YARN)** merupakan sistem yang mengatur dan memonitor cluster node dan resource usage.
- **MapReduce** merupakan framework yang membantu program untuk melakukan komputasi data secara paralel
- **Hadoop Common** merupakan penyedia library Java yang dapat digunakan oleh semua modul

Hadoop bekerja dengan mendistribusi dataset dalam jumlah besar ke beberapa mesin berbeda, untuk kemudian data-data ini diproses di waktu yang bersamaan. HDFS digunakan untuk menyimpan data dan MapReduce memproses data tersebut, sementara itu YARN berfungsi untuk membagi tugas. Dalam implementasinya, Hadoop memiliki ekosistem berupa berbagai tool dan aplikasi yang bisa membantu pengumpulan, penyimpanan, analisis, dan pengolahan Big Data.

Tools in Hadoop

1. Spark

Spark merupakan processing system yang terdistribusi dan bersifat open source, dimana tools ini digunakan untuk melakukan batch processing, streaming analytics, machine learning, graph database, dan ad hoc query.



2. Presto

Seperti halnya Spark, Presto juga salah satu software yang bersifat open source. Presto sendiri merupakan SQL query engine terdistribusi yang digunakan untuk analisis data ad hoc low-latency. Dengan Presto inilah, kita dapat memproses data dari sumber yang berbeda-beda, termasuk HDFS dan Amazon S3.



3. Hive

Hive digunakan untuk MapReduce dengan interface SQL, sehingga tool ini cocok untuk analisis data dalam jumlah yang besar.



4. HBase

HBase adalah database yang digunakan Amazon S3 dan HDFS. Tool ini dibuat untuk memproses tabel dengan baris dalam jumlah yang sangat banyak.



Hadoop Advantages

1. **Fleksibel**

Data bisa disimpan dalam format apapun, baik secara structured maupun unstructured. Hal ini memungkinkan pengguna mengakses data dari sumber manapun dengan tipe apapun.

2. **Upgrade kapasitas**

Hadoop merupakan teknologi yang memberikan solusi pada sistem tradisional. Sistem tradisional memiliki data storage yang terbatas, sementara Hadoop bisa ditingkatkan kapasitasnya, sebab framework ini bekerja secara terdistribusi.

3. **Ketahanan tinggi**

HDFS merupakan bagian dari ekosistem Hadoop, yang dikenal memiliki ketahanan tinggi dan meminimalkan risiko kegagalan baik software maupun hardware. Meskipun satu node rusak atau mengalami masalah, HDFS bisa menyediakan backup data untuk melanjutkan proses.



How to Install Hadoop

Prerequisites that must be installed

Getting Started

To get started with Hadoop, these are the prerequisites that must be installed:

1. Java 8 (Recommended) / Java 11

- <https://www.oracle.com/id/java/technologies/javase/javase8-archive-downloads.html>

2. Hadoop

- <https://archive.apache.org/dist/hadoop/common/>

3. Additional Binaries (Sesuaikan versi Hadoop, binary terbaru hanya sampai v3.2.2)

- <https://github.com/styxnanda/winutils>

Prerequisites that must be installed

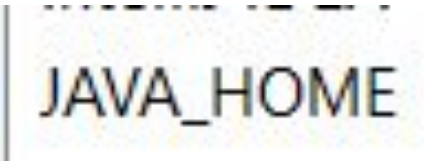
Index of /dist/hadoop/common/hadoop-3.2.2

Name	Last modified	Size	Description
Parent Directory	-	-	-
CHANGELOG.md	2021-01-13 18:48	95K	
CHANGELOG.md.asc	2021-01-13 18:48	833	
CHANGELOG.md.sha512	2021-01-13 18:48	143	
RELEASENOTES.md	2021-01-13 18:48	5.2K	
RELEASENOTES.md.asc	2021-01-13 18:48	833	
RELEASENOTES.md.sha512	2021-01-13 18:48	146	
hadoop-3.2.2-rat.txt	2021-01-13 18:48	1.8M	
hadoop-3.2.2-rat.txt.asc	2021-01-13 18:48	833	
hadoop-3.2.2-rat.txt.sha512	2021-01-13 18:48	151	
hadoop-3.2.2-site.tar.gz	2021-01-13 18:48	43M	
hadoop-3.2.2-site.tar.gz.asc	2021-01-13 18:48	833	
hadoop-3.2.2-site.tar.gz.sha512	2021-01-13 18:48	155	
hadoop-3.2.2-src.tar.gz	2021-01-13 18:48	31M	
hadoop-3.2.2-src.tar.gz.asc	2021-01-13 18:48	833	
hadoop-3.2.2-src.tar.gz.sha512	2021-01-13 18:48	154	
hadoop-3.2.2.tar.gz	2021-01-13 18:48	377M	
hadoop-3.2.2.tar.gz.asc	2021-01-13 18:48	833	
hadoop-3.2.2.tar.gz.sha512	2021-01-13 18:48	150	

Windows x86	201.64 MB	jdk-8u202-windows-i586.exe
Windows x64	211.58 MB	jdk-8u202-windows-x64.exe

hadoop-3.2.1/bin	add 321 winutils
hadoop-3.2.2/bin	compile hadoop-3.2.2

Configure System Variables JAVA_HOME



JAVA_HOME



C:\Program Files\Java\jdk

NOTE:

Tambahkan system variable baru bernama JAVA_HOME
dan arahkan ke instalasi Java

Configure New Path Variable

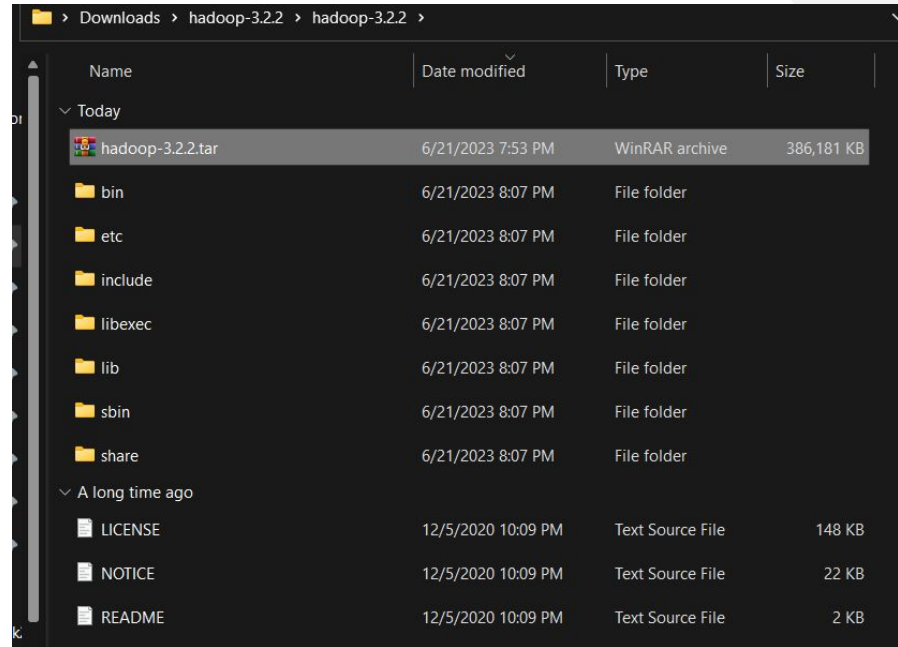
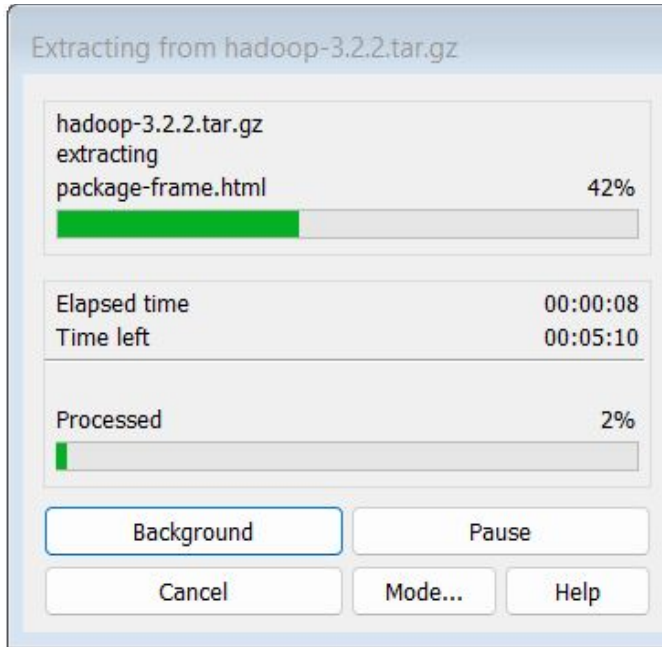
Cek versi java

```
C:\Users\Andikha Wisanggeni>java -version
java version "11.0.16.1" 2022-08-18 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.16.1+1-LTS-1)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.16.1+1-LTS-1, mixed mode)
```

Tambahkan di path, ke directory java dan arahkan ke folder bin

OneDriveConsumer	C:\Users\Andikha Wisanggeni\OneDrive
Path	C:\Users\Andikha Wisanggeni\AppData\Local\Microsoft\Win...
OSVS_ROOTDIR	C:\intelFPGA_lite\21.1\quartus\sonic_builder\bin
	C:\Gradle\gradle-7.5.1\bin
	C:\Program Files\Java\jdk-11.0.16.1\bin

Extract hadoop-3.2.2.tar.gz as admin



Edit \etc\hadoop\core-site.xml

Before

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4   Licensed under the Apache License, Version 2.0 (the "License");
5   you may not use this file except in compliance with the License.
6   You may obtain a copy of the License at
7
8       http://www.apache.org/licenses/LICENSE-2.0
9
10  Unless required by applicable law or agreed to in writing, software
11  distributed under the License is distributed on an "AS IS" BASIS,
12  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13  See the License for the specific language governing permissions and
14  limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 </configuration>
21
```

After

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4   Licensed under the Apache License, Version 2.0 (the "License");
5   you may not use this file except in compliance with the License.
6   You may obtain a copy of the License at
7
8       http://www.apache.org/licenses/LICENSE-2.0
9
10  Unless required by applicable law or agreed to in writing, software
11  distributed under the License is distributed on an "AS IS" BASIS,
12  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13  See the License for the specific language governing permissions and
14  limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>fs.defaultFS</name>
22     <value>hdfs://localhost:9000</value>
23   </property>
24 </configuration>

```

Edit \etc\hadoop\mapred-site.xml

Before

```
1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20
21 </configuration>
```

After

```
1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>mapreduce.framework.name</name>
22     <value>yarn</value>
23   </property>
24 </configuration>
```

Edit \etc\hadoop\yarn-site.xml

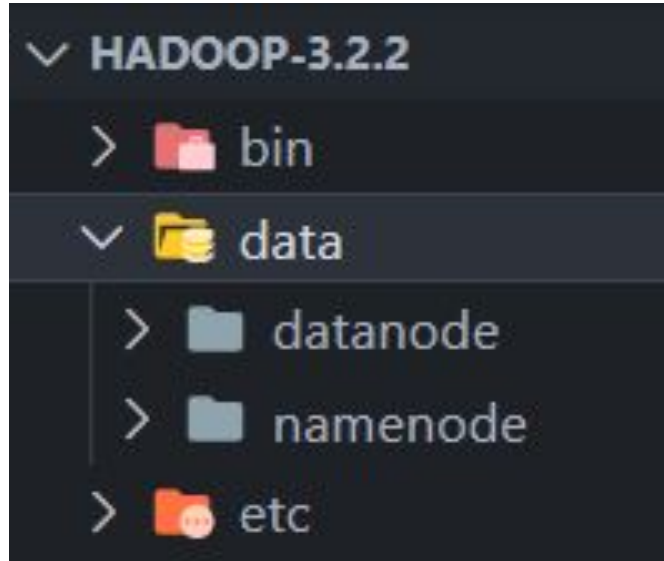
Before

```
1 <?xml version="1.0"?>
2 <!--
3   Licensed under the Apache License, Version 2.0 (the "License");
4   you may not use this file except in compliance with the license.
5   You may obtain a copy of the License at
6
7   http://www.apache.org/licenses/LICENSE-2.0
8
9   Unless required by applicable law or agreed to in writing, software
10  distributed under the License is distributed on an "AS IS" BASIS,
11  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12  See the License for the specific language governing permissions and
13  limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16
17 <!-- Site specific YARN configuration properties -->
18
19 </configuration>
```

After

```
1 <?xml version="1.0"?>
2 <!--
3   Licensed under the Apache License, Version 2.0 (the "License");
4   you may not use this file except in compliance with the license.
5   You may obtain a copy of the License at
6
7   http://www.apache.org/licenses/LICENSE-2.0
8
9   Unless required by applicable law or agreed to in writing, software
10  distributed under the License is distributed on an "AS IS" BASIS,
11  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12  See the License for the specific language governing permissions and
13  limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16
17 <!-- Site specific YARN configuration properties -->
18 <property>
19   <name>yarn.nodemanager.aux-services</name>
20   <value>mapreduce_shuffle</value>
21 </property>
22
23 <property>
24   <name>yarn.nodemanager.mapreduce.shuffle.class</name>
25   <value>org.apache.hadoop.mapred.ShuffleHandler</value>
26 </property>
27
28 </configuration>
```


Create directory and subdirectory



Edit \etc\hadoop\hdfs-site.xml

Before

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20
21 </configuration>
```

After

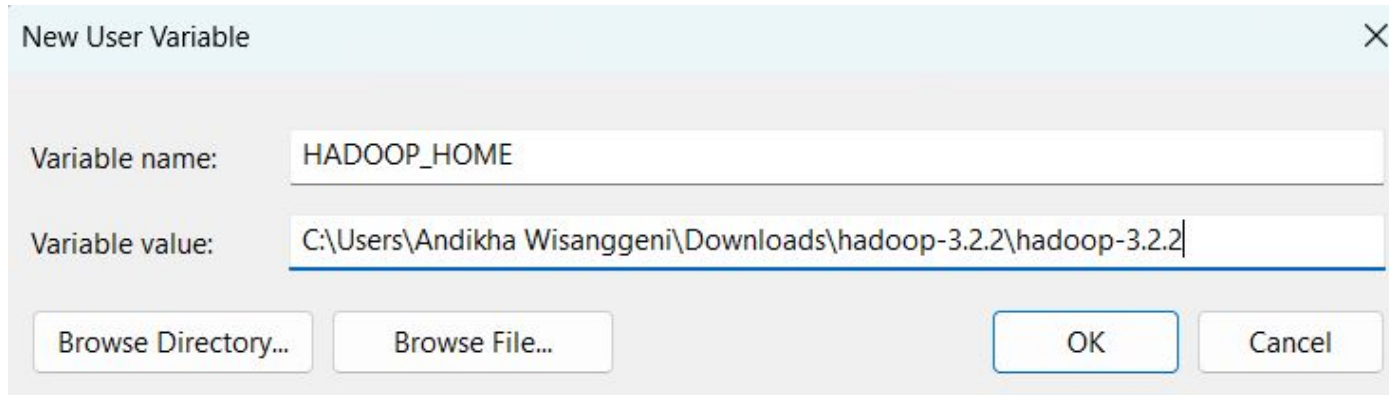
```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20   <property>
21     <name>dfs.replication</name>
22     <value>1</value>
23   </property>
24
25   <property>
26     <name>dfs.namenode.name.dir</name>
27     <value>C:\Users\Andikha Wisanggeni\Downloads\hadoop-3.2.2\hadoop-3.2.2\data\namenode</value>
28   </property>
29
30   <property>
31     <name>dfs.datanode.data.dir</name>
32     <value>C:\Users\Andikha Wisanggeni\Downloads\hadoop-3.2.2\hadoop-3.2.2\data\datanode</value>
33   </property>
34 </configuration>
```

Edit \etc\hadoop\hadoop-env.cmd

Arahkan ke directory Java masing-masing

```
24 @rem The java implementation to use
25 set JAVA_HOME="jdk-11.0.16.1"
26
```

Create HADOOP_HOME and path for /bin dan /sbin



New User Variable

Variable name: HADOOP_HOME

Variable value: C:\Users\Andikha Wisanggeni\Downloads\hadoop-3.2.2\hadoop-3.2.2

Browse Directory... Browse File... OK Cancel

C:\Program Files\...

C:\Users\Andikha Wisanggeni\Downloads\hadoop-3.2.2\hadoop-3.2.2\bin

C:\Users\Andikha Wisanggeni\Downloads\hadoop-3.2.2\hadoop-3.2.2\sbin


Download Additional Binaries

3. Additional Binaries (Sesuaikan versi Hadoop, binary terbaru hanya sampai v3.2.2

- <https://github.com/styxnanda/winutils>

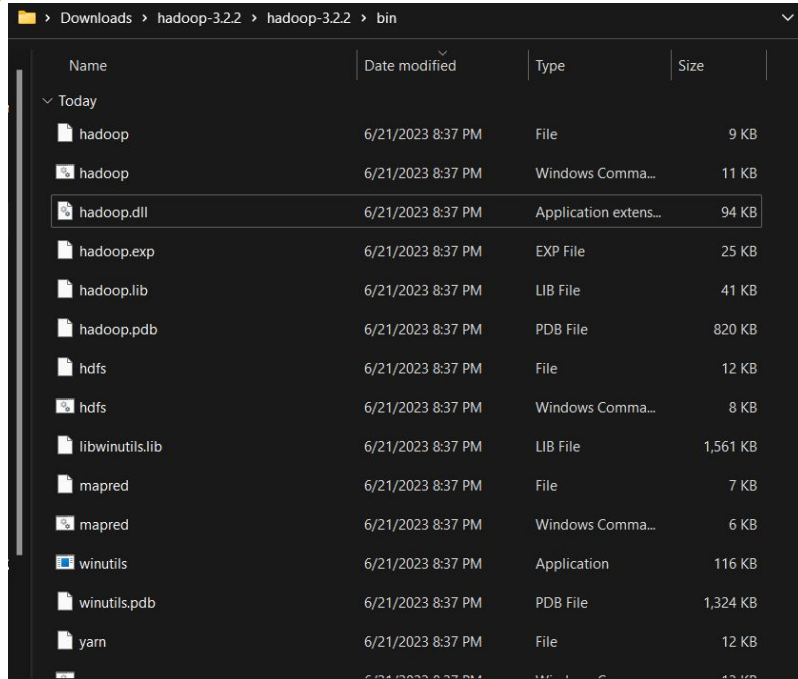
 `hadoop-3.2.1/bin`

`add 321 winutils`

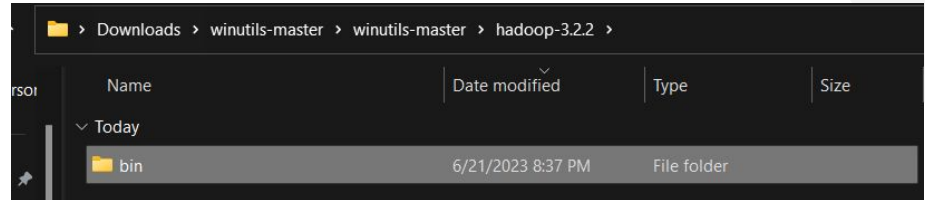
 `hadoop-3.2.2/bin`

`compile hadoop-3.2.2`

Move bin from additional binaries to file hadoop (local) /bin



Name	Date modified	Type	Size
Today			
hadoop	6/21/2023 8:37 PM	File	9 KB
hadoop	6/21/2023 8:37 PM	Windows Comma...	11 KB
hadoop.dll	6/21/2023 8:37 PM	Application extens...	94 KB
hadoop.exp	6/21/2023 8:37 PM	EXP File	25 KB
hadoop.lib	6/21/2023 8:37 PM	LIB File	41 KB
hadoop.pdb	6/21/2023 8:37 PM	PDB File	820 KB
hdfs	6/21/2023 8:37 PM	File	12 KB
hdfs	6/21/2023 8:37 PM	Windows Comma...	8 KB
libwinutils.lib	6/21/2023 8:37 PM	LIB File	1,561 KB
mapred	6/21/2023 8:37 PM	File	7 KB
mapred	6/21/2023 8:37 PM	Windows Comma...	6 KB
winutils	6/21/2023 8:37 PM	Application	116 KB
winutils.pdb	6/21/2023 8:37 PM	PDB File	1,324 KB
yarn	6/21/2023 8:37 PM	File	12 KB

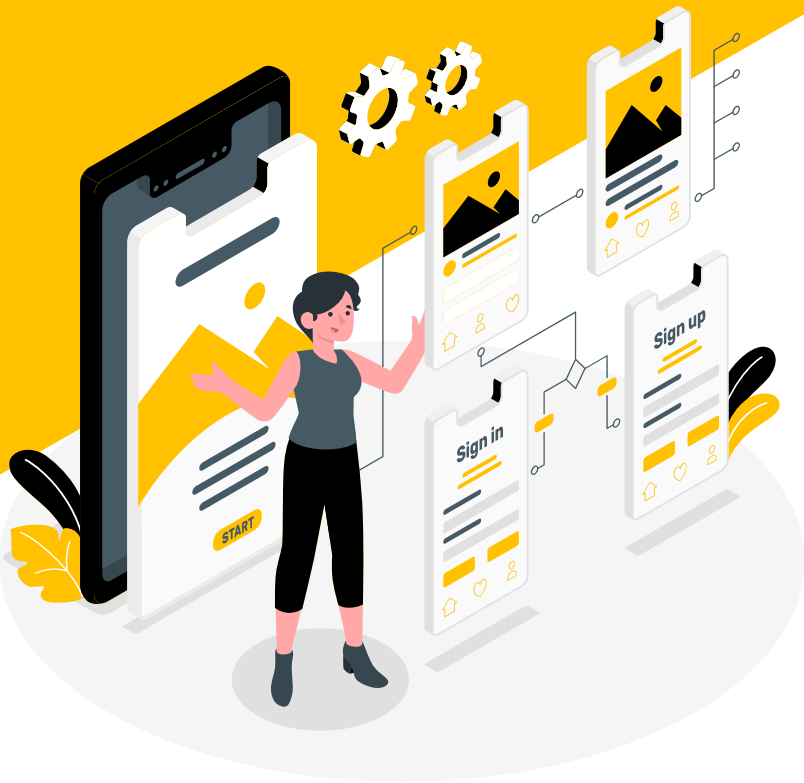


Name	Date modified	Type	Size
Today			
bin	6/21/2023 8:37 PM	File folder	

NOTE:
File yang sama di overwrite

Verify Hadoop

```
C:\Users\fatha>hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /D:/hadoop/share/hadoop/common/hadoop-common-3.2.2.jar
```



How to Run Hadoop

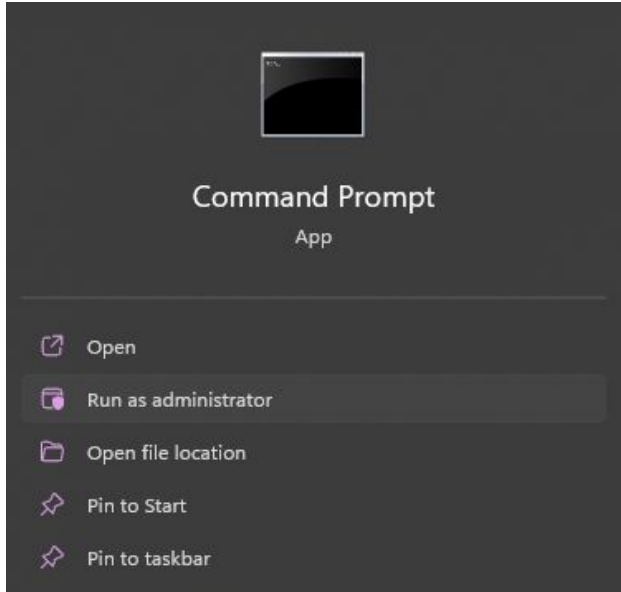
Format the namenode Folder

```
C:\hadoop-3.2.2\sbin>hdfs namenode -format
2023-06-20 19:34:12,020 INFO namenode.NameNode: STARTUP_MSG:
/*****
```

```
2023-06-20 20:16:40,489 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-06-20 20:16:40,494 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-06-20 20:16:40,494 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at noy/192.168.52.1
*****/
```

Berguna untuk menghapus data
sebelumnya pada datanode dan namenode

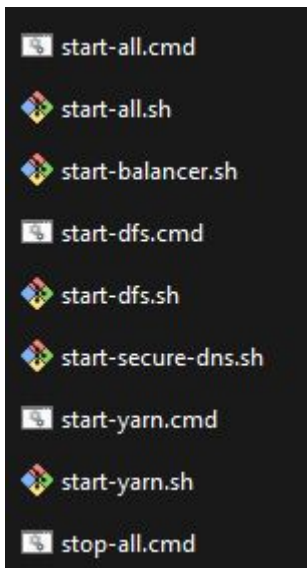
Open CMD



Buka atau jalankan Command Prompt dengan ***Run as Administrator***

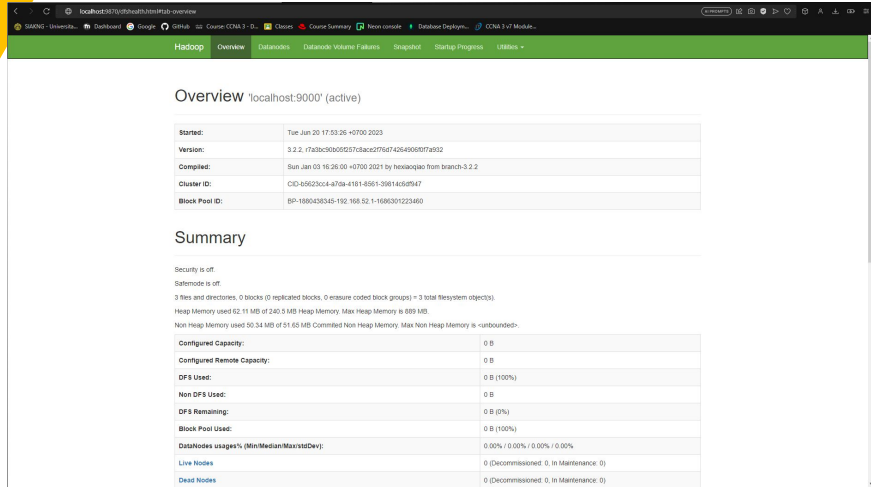
NOTE: Hal ini sangat penting!

Run Hadoop



1. Untuk memulai Hadoop dan daemon-nya, jalankan **start-all.cmd**
2. Untuk memberhentikan Hadoop dan daemon-nya, jalankan **stop-all.cmd**
3. Jika start-all.cmd dan stop-all.cmd sudah deprecated, jalankan **start-dfs.cmd** lalu **start-yarn.cmd**

Check the GUI and Resource Manager



The screenshot shows the Hadoop Overview page for 'localhost:9800' (active). The page has a green header with navigation links: Overview, Datanodes, Datanode Volume Managers, Snapshot, Startup Progress, and Utilities. The main content area is titled 'Overview localhost:9800 (active)' and contains a table with the following information:

Started:	Tue Jun 20 17:53:26 +0700 2023
Version:	3.2.2 - (7ab0c56055703ace27f6d7426490607a932)
Compiled:	Sun Jan 03 16:26:08 +0700 2021 by hexa0qiao from branch-3.2.2
Cluster ID:	CD-65623cc4-470a-4181-8561-39814c8a947
Block Pool ID:	BP-188043345-192-168-52-1-1680301223460

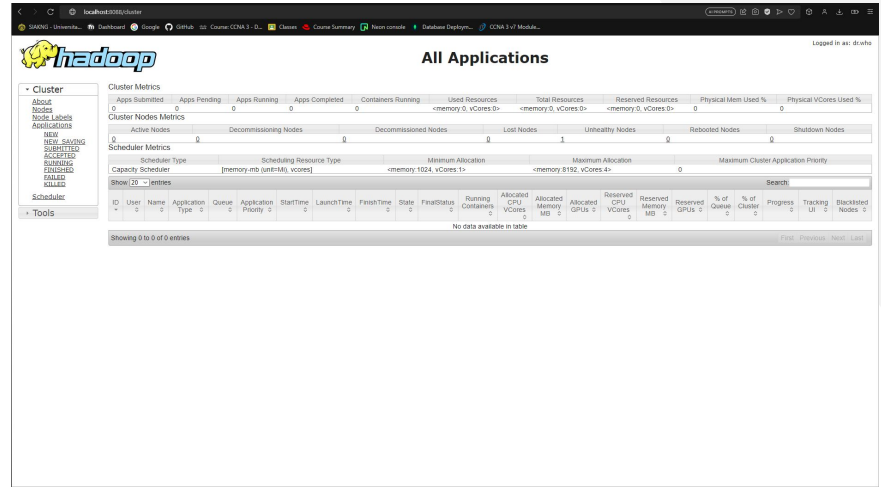
Below the table is a 'Summary' section with the following text:

Security is off.
SafeMode is off.
3 files and directories, 0 blocks (0 replicated blocks, 0 ensure coded block groups) = 3 total filesystem objects.
Heap Memory used 62.11 MB of 240.5 MB Heap Memory. Max Heap Memory is 889 MB.
Non-Heap Memory used 50.34 MB of 61.65 MB Committed Non-Heap Memory. Max Non-Heap Memory is 'unbounded'.

Below this text is a table with the following information:

Configured Capacity:	0 B
Configured Remote Capacity:	0 B
DFS Used:	0 B (100%)
Non-DFS Used:	0 B
DFS Remaining:	0 B (0%)
Block Pool Used:	0 B (100%)
Datanodes usage% (Min/Median/Max/StdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes:	0 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes:	0 (Decommissioned: 0, In Maintenance: 0)

localhost:9870



The screenshot shows the Hadoop All Applications page for 'localhost:8088'. The page has a green header with navigation links: Overview, Datanodes, Datanode Volume Managers, Snapshot, Startup Progress, and Utilities. The main content area is titled 'All Applications' and contains a table with the following information:

Cluster Metrics	Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
	0	0	0	0	0	<memory:0, vCores:0>	<memory:0, vCores:0>	<memory:0, vCores:0>	0	0

Below this table is a 'Cluster Nodes Metrics' section with a table showing the status of nodes. The table has columns for Active Nodes, Decommissioning Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, Rebooted Nodes, and Shutdown Nodes. All values are 0.

Below this table is a 'Scheduler Metrics' section with a table showing the status of the scheduler. The table has columns for Scheduler Type, Scheduling Resource Type, Minimum Allocation, Maximum Allocation, and Maximum Cluster Application Priority. All values are 0.

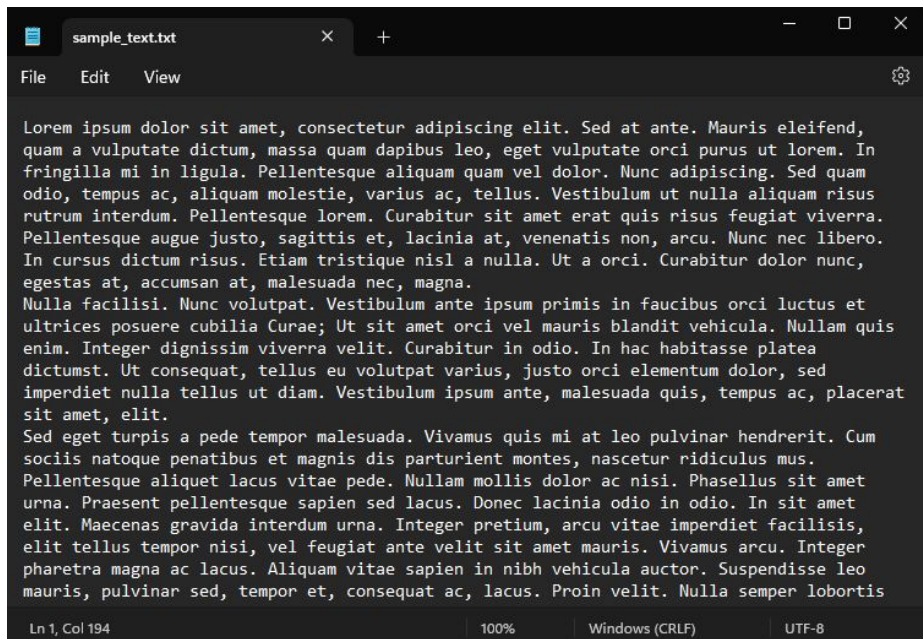
Below this table is a 'Show (20) entries' section with a table showing the status of applications. The table has columns for ID, User, Name, Application, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU, Allocated Memory, Allocated vCores, Reserved CPU, Reserved Memory, Reserved vCores, % of Queue, % of Cluster, Progress, Tracking, and Backstated Nodes. All values are 0.

localhost:8088

Checking the Running Hadoop Daemons

```
D:\hadoop\sbin>JPS
13264 Jps
30192 NameNode
27684 DataNode
30708 NodeManager
27128 ResourceManager
```

Prepare the Input Text File



A screenshot of a text editor window titled "sample_text.txt". The window has a dark theme and a menu bar with "File", "Edit", and "View". The text inside is Lorem Ipsum, starting with "Lorem ipsum dolor sit amet, consectetur adipiscing elit." and ending with "malesuada ac, lacus. Proin vel. Nulla semper lobortis". The status bar at the bottom shows "Ln 1, Col 194", "100%", "Windows (CRLF)", and "UTF-8".

```
sample_text.txt
File Edit View
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed at ante. Mauris eleifend,
quam a vulputate dictum, massa quam dapibus leo, eget vulputate orci purus ut lorem. In
fringilla mi in ligula. Pellentesque aliquam quam vel dolor. Nunc adipiscing. Sed quam
odio, tempus ac, aliquam molestie, varius ac, tellus. Vestibulum ut nulla aliquam risus
rutrum interdum. Pellentesque lorem. Curabitur sit amet erat quis risus feugiat viverra.
Pellentesque augue justo, sagittis et, lacinia at, venenatis non, arcu. Nunc nec libero.
In cursus dictum risus. Etiam tristique nisl a nulla. Ut a orci. Curabitur dolor nunc,
egestas at, accumsan at, malesuada nec, magna.
Nulla facilisi. Nunc volutpat. Vestibulum ante ipsum primis in faucibus orci luctus et
ultrices posuere cubilia Curae; Ut sit amet orci vel mauris blandit vehicula. Nullam quis
enim. Integer dignissim viverra velit. Curabitur in odio. In hac habitasse platea
dictumst. Ut consequat, tellus eu volutpat varius, justo orci elementum dolor, sed
imperdiet nulla tellus ut diam. Vestibulum ipsum ante, malesuada quis, tempus ac, placerat
sit amet, elit.
Sed eget turpis a pede tempor malesuada. Vivamus quis mi at leo pulvinar hendrerit. Cum
sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.
Pellentesque aliquet lacus vitae pede. Nullam mollis dolor ac nisi. Phasellus sit amet
urna. Praesent pellentesque sapien sed lacus. Donec lacinia odio in odio. In sit amet
elit. Maecenas gravida interdum urna. Integer pretium, arcu vitae imperdiet facilisis,
elit tellus tempor nisi, vel feugiat ante velit sit amet mauris. Vivamus arcu. Integer
pharetra magna ac lacus. Aliquam vitae sapien in nibh vehicula auctor. Suspendisse leo
mauris, pulvinar sed, tempor et, consequat ac, lacus. Proin vel. Nulla semper lobortis

Ln 1, Col 194 | 100% | Windows (CRLF) | UTF-8
```

Moving Text File to Input Directory HDFS

```
D:\hadoop\sbin>hadoop fs -mkdir /input_dir  
D:\hadoop\sbin>hadoop fs -put "D:\CoolYeah\Coolyeah - Semester 4\Sistem Basis Data-02\Hadoop\sample_text.txt" /input_dir
```

1 `hadoop fs -mkdir /input_directory`

Membuat folder input pada HDFS

2 `hadoop fs -put "direct to file .txt" /input_directory`

Meletakkan file text ke folder input HDFS

Verify Text File is in HDFS

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

Browse Directory

/input_dir

Go!



Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	-rw-r--r--	fatha	supergroup	157.18 KB	Jun 21 06:41	1	128 MB	sample_text.txt	

Showing 1 to 1 of 1 entries

Previous

1

Next

Hadoop, 2021.

Executing WordCount Program

```
D:\hadoop\sbin>hadoop jar D:/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.2.jar wordcount /input_dir /output_dir
2023-06-21 06:41:07,222 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-06-21 06:41:07,628 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/fatha/.staging/
job_1687304429065_0001
2023-06-21 06:41:07,761 INFO input.FileInputFormat: Total input files to process : 1
2023-06-21 06:41:07,795 INFO mapreduce.JobSubmitter: number of splits:1
2023-06-21 06:41:07,860 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1687304429065_0001
2023-06-21 06:41:07,861 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-06-21 06:41:07,957 INFO conf.Configuration: resource-types.xml not found
2023-06-21 06:41:07,958 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-06-21 06:41:08,112 INFO impl.YarnClientImpl: Submitted application application_1687304429065_0001
2023-06-21 06:41:08,150 INFO mapreduce.Job: The url to track the job: http://noy:8088/proxy/application_1687304429065_0001/
2023-06-21 06:41:08,151 INFO mapreduce.Job: Running job: job_1687304429065_0001
2023-06-21 06:41:14,260 INFO mapreduce.Job: Job job_1687304429065_0001 running in uber mode : false
2023-06-21 06:41:14,260 INFO mapreduce.Job:  map 0% reduce 0%
2023-06-21 06:41:17,307 INFO mapreduce.Job:  map 100% reduce 0%
2023-06-21 06:41:21,344 INFO mapreduce.Job:  map 100% reduce 100%
2023-06-21 06:41:22,358 INFO mapreduce.Job: Job job_1687304429065_0001 completed successfully
2023-06-21 06:41:22,408 INFO mapreduce.Job: Counters: 54
```

NOTE:

Semakin besar ukuran file input, maka akan semakin lama waktu yang dibutuhkan untuk memproses.

Check the Output



All Applications

Cluster

About
Nodes
Node Labels
Applications
NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %
1	0	0	1	0	<memory:0, vCores:0>	<memory:8192, vCores:8>	<memory:0, vCores:0>	67

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Clust
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% o Clust
application_1687304429065_0001	fatha	word count	MAPREDUCE	default	0	Wed Jun 21 06:41:08 +0700 2023	Wed Jun 21 06:41:08 +0700 2023	Wed Jun 21 06:41:20 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0	0.0

Showing 1 to 1 of 1 entries

Check the Output

Browse Directory

Show entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		drwxr-xr-x		fatha		supergroup		0 B		Jun 21 06:41		0		0 B		input_dir	
<input type="checkbox"/>		drwxr-xr-x		fatha		supergroup		0 B		Jun 21 06:41		0		0 B		output_dir	
<input type="checkbox"/>		drwx-----		fatha		supergroup		0 B		Jun 21 06:41		0		0 B		tmp	

Showing 1 to 3 of 3 entries

Browse Directory

Show entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		fatha		supergroup		0 B		Jun 21 06:41		1		128 MB		_SUCCESS	
<input type="checkbox"/>		-rw-r--r--		fatha		supergroup		2.42 KB		Jun 21 06:41		1		128 MB		part-r-00000	

Showing 1 to 2 of 2 entries

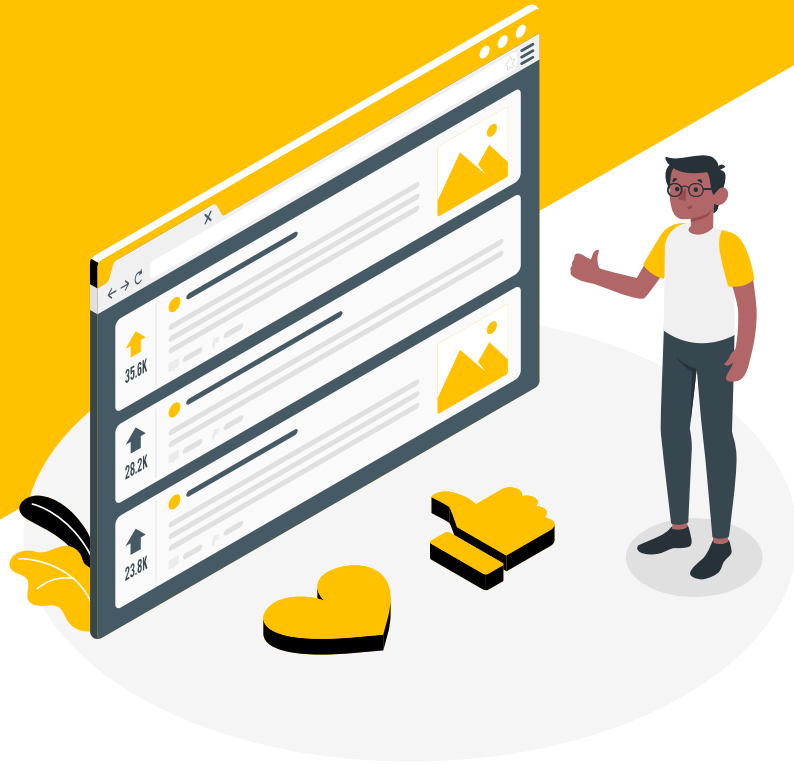
Check the Output

```
D:\hadoop\sbin>hadoop fs -cat /output_dir/part-r-00000
Aliquam 250
Cum 100
Curabitur 200
Curae; 50
Donec 200
Duis 150
Etiam 100
Fusce 100
In 350
Integer 300
Lorem 50
Maecenas 100
Mauris 50
Nulla 200
Nullam 100
Nunc 200
Pellentesque 200
Phasellus 100
Praesent 100
Proin 100
Sed 150
Suspendisse 200
Ut 250
Vestibulum 200
Vivamus 100
a 300
ac 100
ac, 200
```

```
accumsan 100
adipiscing 50
adipiscing, 50
adipiscing. 50
aliquam 200
aliquet 50
amet 300
amet, 150
ante 100
ante, 50
ante. 50
arcu 100
arcu. 100
at 150
at, 250
auctor. 50
augue 200
augue, 50
bibendum 50
blandit 50
commodo 50
condimentum 50
consectetur 50
consequat 50
consequat, 50
cubilia 50
cursus 100
cursus. 50
dapibus 100
diam. 100
```

```
dictum 50
dictum, 50
dictumst. 150
dignissim 100
dis 100
dolor 250
dolor, 50
dolor. 100
dui 50
dui. 50
egestas 100
eget 150
eleifend, 50
elementum 50
elit 150
elit. 250
enim 50
enim. 150
erat 150
erat, 50
eros 50
et 150
et, 150
eu 100
eu, 100
facilisi. 50
facilisis, 50
faucibus 50
faucibus. 50
felis 50
```

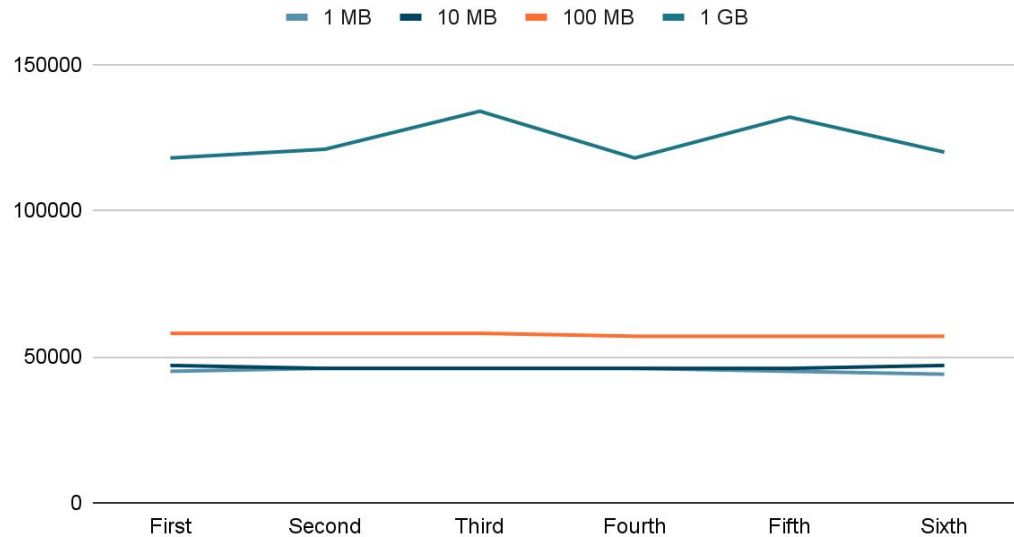
```
suscipit, 50
tellus 200
tellus. 100
tempor 250
tempus 100
tempus, 50
tincidunt 50
tristique 100
turpis 50
turpis. 50
ultrices 150
urna 100
urna. 100
ut 250
varius 150
varius, 50
vehicula 50
vehicula. 50
vel 250
velit 50
velit. 100
venenatis 50
vitae 200
vitae, 50
viverra 100
viverra. 50
volutpat 50
volutpat. 150
vulputate 100
```



Hadoop vs No Hadoop

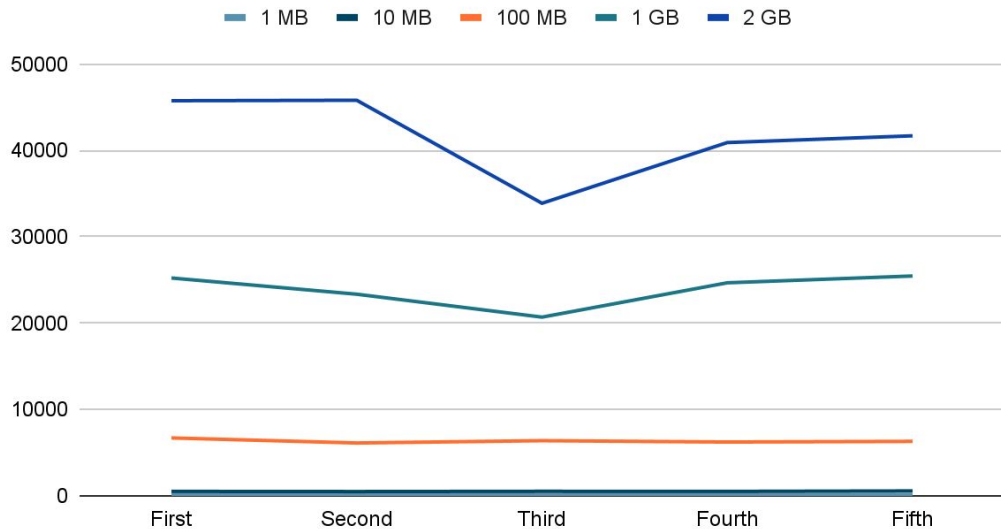
With Hadoop

With Hadoop



Without Hadoop

Without Hadoop



Test on Other Machine

```
[airev@HP bin]$ time ( ./hadoop jar ../share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.5.jar wordcount input out; cat out/part-r-00000; xm -rf out )
```

```
you?      22000
you?!     2000
your      130000
yourjob   2000
yours!    2000
yours?    2000
yourself      4000
```

```
real    0m26,420s
user    0m37,069s
sys     0m0,538s_
```

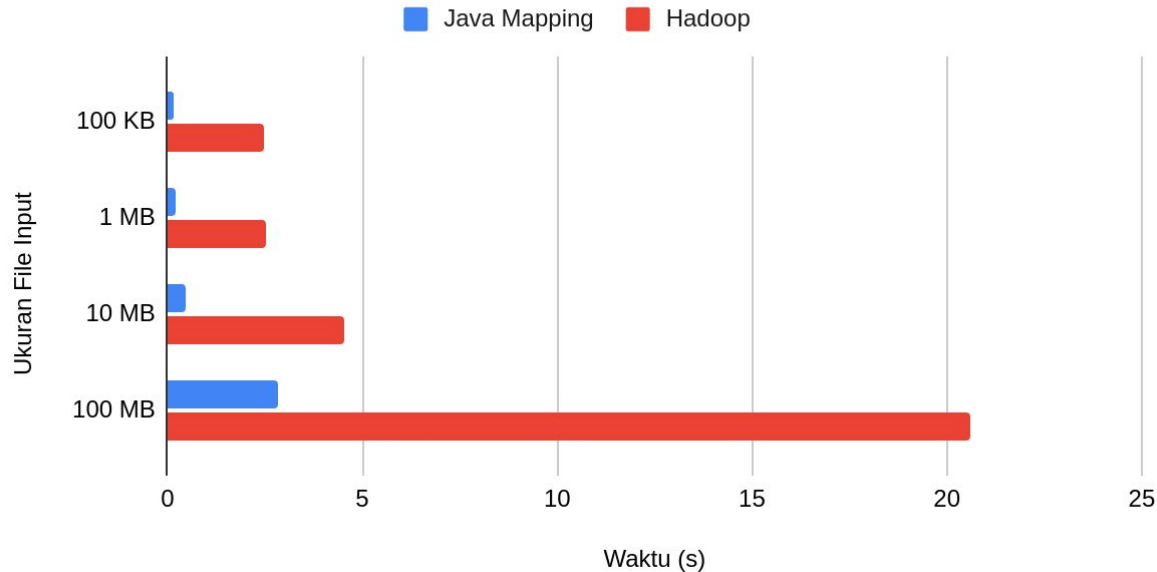
```
[airev@HP bin]$ time java ordinarywordcount
```

```
daisies. = 2000
Oourt = 2000
bees.
Bees?
Specifically, = 2000
Jock.
Yeah. = 2000
```

```
real    0m2,957s
user    0m8,027s
sys     0m0,758s_
```


Test on Other Machine

Perbandingan antara Waktu Pemrosesan File Input
Menggunakan Hadoop terhadap Java Mapping





Analisis Kinerja Hadoop vs Java



Hadoop memiliki keunggulan dalam kinerja pemrosesan data besar (big data) secara terdistribusi. Dengan memanfaatkan cluster Hadoop, tugas-tugas dapat dibagi menjadi unit-unit yang lebih kecil dan didistribusikan di seluruh node, mempercepat waktu pemrosesan secara signifikan.



Penggunaan Java untuk word count cenderung lebih cocok untuk data yang relatif kecil. Kinerja Java tergantung pada implementasi kode yang ditulis, tetapi tidak memiliki kemampuan terdistribusi bawaan seperti Hadoop.



Hadoop dirancang khusus untuk skalabilitas yang tinggi. Dengan penambahan node ke dalam cluster Hadoop, kapasitas pemrosesan dapat ditingkatkan sesuai kebutuhan yang mana memungkinkan Hadoop untuk mengatasi data yang sangat besar dan kompleks.



Skalabilitas Java terbatas pada sumber daya yang tersedia di IDE tempat program dijalankan. Pengolahan data terbatas pada kapasitas mesin tersebut dan tidak secara otomatis terdistribusi ke beberapa node seperti Hadoop.



Implementasi word count dengan Hadoop memerlukan penulisan kode yang melibatkan konfigurasi job, pembuatan fungsi mapper dan reducer, serta penanganan input dan output. Hal ini memerlukan pemahaman yang baik tentang framework Hadoop dan memerlukan waktu dan usaha yang lebih.



Implementasi word count dengan Java lebih sederhana karena hanya perlu menulis kode Java untuk membaca teks, memisahkan kata, dan menghitung kemunculannya. Cara ini lebih mudah dipahami dan memerlukan waktu pengembangan yang lebih singkat.



Penggunaan Hadoop memerlukan infrastruktur cluster yang terdiri dari beberapa node, dimana membutuhkan konfigurasi dan penyiapan yang cermat serta pemahaman tentang pengelolaan cluster Hadoop.



Penggunaan Java tidak memerlukan infrastruktur khusus, dimana pengguna dapat menjalankan program Java pada IDE yang diinginkan.

References

- [1] “Index of /dist/hadoop/common,” *Apache.org*, 2023. Available: <https://archive.apache.org/dist/hadoop/common/>.
- [2] “Java Archive Downloads - Java SE 8 | Oracle Indonesia,” *Oracle.com*, 2019. Available: <https://www.oracle.com/id/java/technologies/javase/javase8-archive-downloads.html>.
- [3] styxnanda, “GitHub - styxnanda/winutils: winutils.exe hadoop.dll and hdfs.dll binaries for hadoop windows,” GitHub, 2023. Available: <https://github.com/styxnanda/winutils>.
- [4] “Apa Itu Hadoop? Tools Yang Banyak Digunakan Dalam Big Data – Inixindo Jogja,” Inixindo Jogja, Mar. 23, 2022. Available: <https://inixindojogja.co.id/apa-itu-hadoop-big-data/>.
- [5] SkillsBuild Training, “🔥 How to Install Hadoop on Windows 11,” YouTube. Apr. 18, 2022. Available: <https://www.youtube.com/watch?v=GNHF0DZK3xQ&t>.
- [6] A. Devkar, “Run Wordcount Program on Hadoop-3.3.0 windows 10,” YouTube. May 13, 2021. Available: <https://www.youtube.com/watch?v=nsi4nVS16lc&t>.

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

