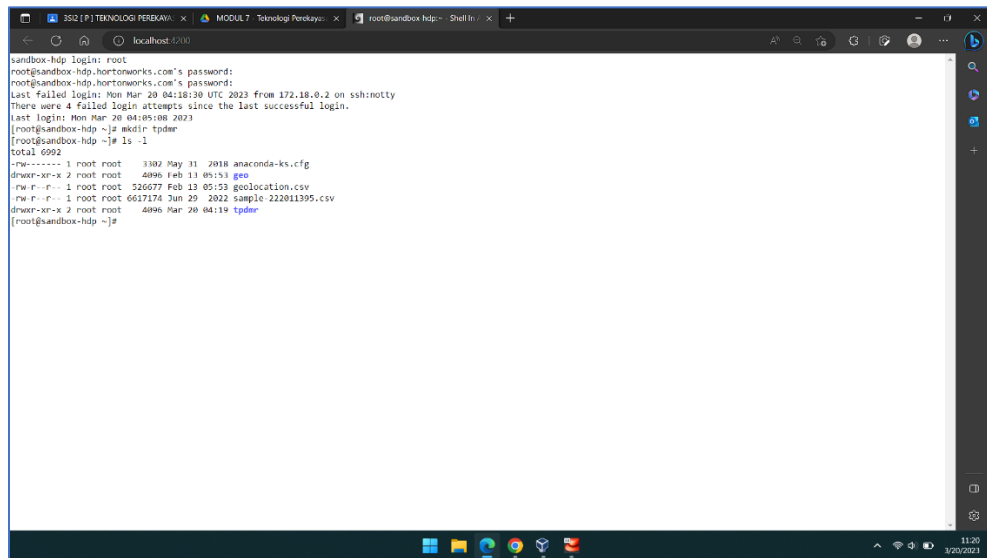


Praktikum Teknologi Perekayasa Data

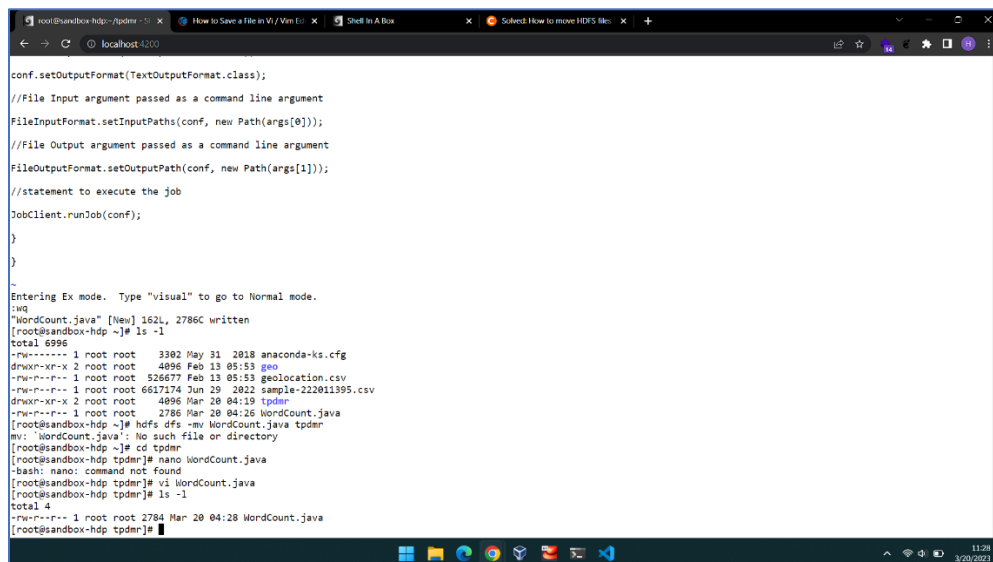
Tanggal Praktikum : Monday, March 20, 2023

1. Buat Direktori



```
sandbox-hdp login: root
root@sandbox-hdp:~$ ssh root@172.18.0.2
root@sandbox-hdp:~$ ssh root@172.18.0.2
Last failed login: Mon Mar 20 04:18:30 UTC 2023 from 172.18.0.2 on ssh:root@172.18.0.2
There were 4 failed login attempts since the last successful login.
Last login: Mon Mar 20 04:05:08 2023
[root@sandbox-hdp ~]# mkdir tpdmer
[root@sandbox-hdp ~]# ls -l
total 6992
-rw-r--r-- 1 root root 3302 May 31 2018 anaconda-ks.cfg
drwxr-xr-x 2 root root 4096 Feb 13 05:53 geo
-rw-r--r-- 1 root root 526677 Feb 13 05:53 geolocation.csv
-rw-r--r-- 1 root root 6617174 Jun 29 2022 sample-222011395.csv
drwxr-xr-x 2 root root 4096 Mar 20 04:19 tpdmer
[root@sandbox-hdp ~]#
```

2. Buat File Java



```
conf.setOutputFormat(TextOutputFormat.class);
//File Input argument passed as a command line argument
FileInputFormat.setInputPaths(conf, new Path(args[0]));
//File Output argument passed as a command line argument
FileOutputFormat.setOutputPath(conf, new Path(args[1]));
//statement to execute the job
JobClient.runJob(conf);
}
}
~
Entering Ex mode. Type "visual" to go to Normal mode.
~wq
"WordCount.java" [New] 162L, 2786C written
[root@sandbox-hdp ~]# ls -l
total 6996
-rw-r--r-- 1 root root 3302 May 31 2018 anaconda-ks.cfg
drwxr-xr-x 2 root root 4096 Feb 13 05:53 geo
-rw-r--r-- 1 root root 526677 Feb 13 05:53 geolocation.csv
-rw-r--r-- 1 root root 6617174 Jun 29 2022 sample-222011395.csv
drwxr-xr-x 2 root root 4096 Mar 20 04:19 tpdmer
-rw-r--r-- 1 root root 2786 Mar 20 04:26 WordCount.java
[root@sandbox-hdp ~]# hdfs dfs -mv WordCount.java tpdmer
mv: 'WordCount.java': no such file or directory
[root@sandbox-hdp ~]# cd tpdmer
[root@sandbox-hdp tpdmer]# nano WordCount.java
~bash: nano: command not found
[root@sandbox-hdp tpdmer]# vi WordCount.java
~
[root@sandbox-hdp tpdmer]# ls -l
total 4
-rw-r--r-- 1 root root 2784 Mar 20 04:28 WordCount.java
[root@sandbox-hdp tpdmer]#
```

3. Compile Java

```
< > localhost:4200
```

```
[root@nsandbox-hdp tpdr]#  
[root@nsandbox-hdp tpdr]#  
[root@nsandbox-hdp tpdr]# hls7.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-  
bash: hls7.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop: No such file or directory  
[root@nsandbox-hdp tpdr]#  
[root@nsandbox-hdp tpdr]# napreduce-client-common-3.1.1.3.0.1.0-187.jar WordCount.java  
bash: napreduce-client-common-3.1.1.3.0.1.0-187.jar: command not found  
[root@nsandbox-hdp tpdr]# java -classpath /usr/hdp/3.0.1.0-187/hadoop-common-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-  
com-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-common-3.1.1.3.0.1.0-187.jar WordCount.java  
WordCount.java:2: error: class, interface, or enum expected  
rt java.io.IOException;  
  
1 error  
[root@nsandbox-hdp tpdr]# java -classpath /usr/hdp/3.0.1.0-187/hadoop/hadoop-common-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-  
com-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-common-3.1.1.3.0.1.0-187.jar WordCount.java  
WordCount.java:2: error: class, interface, or enum expected  
rt java.io.IOException;
```

```
1 error  
[root@nsandbox-hdp tpdr]# vi WordCount.java  
[root@nsandbox-hdp tpdr]# java -classpath /usr/hdp/3.0.1.0-187/hadoop/hadoop-common-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-  
com-3.1.1.3.0.1.0-187.jar:/usr/hdp/3.0.1.0-187/hadoop-napreduce/hadoop-napreduce-client-common-3.1.1.3.0.1.0-187.jar WordCount.java  
[root@nsandbox-hdp tpdr]# ls -l  
total 16  
-rw-r--r-- 1 root root 1417 Mar 20 04:33 WordCount.class  
-rw-r--r-- 1 root root 2787 Mar 20 04:33 WordCount.java  
-rw-r--r-- 1 root root 1926 Mar 20 04:33 WordCountMap.class  
-rw-r--r-- 1 root root 2087 Mar 20 04:33 WordCountReduce.class  
[root@nsandbox-hdp tpdr]# jar -cf wordcount.jar *.class  
added manifest  
adding: WordCount.class(in = 1417) (out= 783)(deflated 50%)  
adding: WordCountMap.class(in = 1926) (out= 883)(deflated 58%)  
adding: WordCountReduce.class(in = 2087) (out= 817)(deflated 59%)  
[root@nsandbox-hdp tpdr]# ls -l  
total 20  
-rw-r--r-- 1 root root 1417 Mar 20 04:33 WordCount.class  
-rw-r--r-- 1 root root 3853 Mar 20 04:34 wordcount.jar  
-rw-r--r-- 1 root root 2787 Mar 20 04:33 WordCount.java  
-rw-r--r-- 1 root root 1916 Mar 20 04:33 WordCountMap.class  
-rw-r--r-- 1 root root 2087 Mar 20 04:33 WordCountReduce.class  
[root@nsandbox-hdp tpdr]#
```

4. MapReduce

```
Map output materialized bytes=461
Input split bytes=294
Combine input records=0
Combine output records=0
Reduce input groups=0
Reduce shuffle bytes=461
Reduce input records=0
Reduce output records=0
Spilled Records=118
Shuffled Map <=
Failed Shuffle=0
Merged Map outputs=2
GC Time elapsed (ms)=55
CPU time spent (ms)=700
Physical memory (bytes) snapshot=1155198976
Virtual memory (bytes) snapshot=15049600
Total committed heap usage (bytes)=4545650924
Peak Map Physical memory (bytes)=22242468
Peak Map Virtual memory (bytes)=363807162
Peak Reduce Physical memory (bytes)=263153400
Peak Reduce Virtual memory (bytes)=2648148016
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
[root@sandbox-hdp tpdr]# hdfs fs -cat output/part-00000
Jolo 12
J11 11
J1n1n 10
Jolo 10
J1n 10
[root@sandbox-hdp tpdr]# hdfs fs -rm -r output
21/03/29 09:22:26 INFO fs.TrashOp$CleanupTask: Moved 'hdfs://sandbox-hdp.hortonworks.com:8020/user/root/output' to trash at hdfs://sandbox-hdp.hortonworks.com:8020/user/root/.Trash
[root@sandbox-hdp tpdr]#
```

The screenshot shows a Kali Linux desktop environment with a terminal window open. The terminal displays the output of a Hadoop MapReduce job. The left sidebar shows the file manager with the path /home/. The main window shows the following output:

```
Map input records=99  
Map output records=99  
Map output bytes=801  
Map output materialized bytes=801  
Input split bytes=24  
Combine input records=0  
Combine output records=0  
Reduce input groups=5  
Reduce shuffle bytes=801  
Reduce input records=99  
Reduce output records=5  
Spilled Records=116  
Shuffled Maps =2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=1516  
CPU time spent (s)=30.0  
Physical memory (bytes) snapshot=1155100976  
Virtual memory (bytes) snapshot=102426988  
Total committed heap usage (bytes)=942669824  
Peak Map Physical memory (bytes)=26264560  
Peak Map Virtual memory (bytes)=2839867392  
Peak Reduce Physical memory (bytes)=1400131840  
Peak Reduce Virtual memory (bytes)=2640150016  
Shuffle Progress  
BAD_Tier  
COMMITTED=0  
IN_PROGRESS  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=240  
File Output Format Counters  
Bytes Written=82  
  
[root@sandbox-hdp tpdrj]# hdfs dfs -cat output/port-00000  
Deja 12  
JSL 11  
JSL 10  
JSL 10  
Deja 10  
Deja 10  
JSL 10  
[root@sandbox-hdp tpdrj]#
```

At the bottom of the terminal window, there is a footer for "UNGUISHED VERSION" with contact information for support.

hadoop

Cluster

Nodes

Node Labels

Applications

NEW

SAVING

IMPORTING

ACCEPTED

REJECTING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used
1	0	0	1	0	0 B	4.0 B	0 B	0

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	Max Capacity Scheduler
Capacity Scheduler	[memory MB, vCores]	<memory 256, vCores 1>	<memory 4096, vCores 4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	Start Time	Finish Time	State	Final Status	Running Containers	Allocated CPU V-Cores	Allocated Memory MB	Reserved CPU V-Cores	Reserved Memory MB	% of Queue
application_1679285302016_0001	hive	HIVE_6400c81abb1-4c16-5cac-8748e98374de	TEZ	default	0	Mon Mar 20 11:15:27 +0700 2023	Mon Mar 20 11:27:22 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_167960733281_0001	hive	HIVE_85621ee9f-2ab5-40ba-ba41-1d7167b3219b	TEZ	default	0	Mon Mar 13 13:52:14 +0700 2023	Mon Mar 13 13:57:30 +0700 2023	FAILED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0
application_167963617209_0001	hive	HIVE_363b3c4a-7011-4962-9927-6229a5080950	TEZ	default	0	Sun Mar 12 22:45:29 +0700 2023	Sun Mar 12 23:04:07 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_167963620954_0001	hive	HIVE_a91f1845c-7b6a-405b-a987c-78e6d980507b	TEZ	default	0	Sun Mar 12 22:20:13 +0700 2023	Sun Mar 12 22:52:38 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1679620118824_0002	hive	HIVE_8732d331-4b79-4963-	TEZ	default	0	Mon Feb 13 11:28:42	Mon Feb 13 11:39:07	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0

11:58 3/20/2023

Penugasan

1. File Java : <https://pastebin.com/VpVwfWAj>

Kode Untuk Split data menjadi key (country) and value (salary)


```
// Split into key and value with comma separated
String[] line = value.toString().split(",");
// get the salary
int salary = Integer.parseInt(line[2]);
// get the country
String country = line[1];
// emit the country and salary
collector.collect(new Text(country), new IntWritable(salary));
```

Kode untuk mendapatkan nilai salary maksimum untuk setiap key (negara).

```
int max = 0;
while (values.hasNext()) {
    max = Math.max(max, values.next().get());
}
System.out.println("Max salary for " + key + " is " + max);
collector.collect(key, new IntWritable(max));
```

3. Map Reduce untuk mencari maximum salary tiap country

```
change absen text selector - nub: x  tipsen2 | Railway x root@sandbox-hdp:~/hdpmr - S x (M/V) Rapids - JCT40 - YouTube x +
localhost:2000
23/03/20 06:21:20 INFO mapreduce.Job: map 100% reduce 0%
23/03/20 06:21:35 INFO mapreduce.Job: map 100% reduce 100%
23/03/20 06:21:37 INFO mapreduce.Job: Job job_1679285362016_0012 completed successfully
23/03/20 06:21:37 INFO mapreduce.Job: Counters: 53
  File System Counters
    FILE: Number of bytes read=18006
    FILE: Number of bytes written=740228
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=43891
    HDFS: Number of bytes written=2187
    HDFS: Number of read operations=11
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=2
    Launched reduce tasks=1
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=1192084
    Total time spent by all reduces in occupied slots (ms)=36392
    Total time spent by all map tasks (ms)=298021
    Total time spent by all reduce tasks (ms)=9098
    Total vcore-milliseconds taken by all map tasks=298021
    Total vcore-milliseconds taken by all reduce tasks=9098
    Total megabyte-milliseconds taken by all map tasks=305173504
    Total megabyte-milliseconds taken by all reduce tasks=9316352
  Map-Reduce Framework
    Map input records=2000
    Map output records=2000
    Map output bytes=14000
    Map output materialized bytes=18012
    Input split bytes=244
    Combine input records=0
    Combine output records=0
    Reduce input groups=243
    Reduce shuffle bytes=18012
    Reduce input records=2000
    Reduce output records=243
    Spilled Records=4000
    Shuffled Maps =2
    Failed Shuffles=0
```



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	Memory Used	Memory Total	Memory Reserved	VCores Used
3	0	0	3	0	0 B	4 GB	0 B	0

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
Capacity Scheduler	[memory-mb (unit-M), vcores]	<memory 256, vcores 1>	<memory 4096, vcores 4>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue
application_1679285362016_0012	root	maxsalary.jar	MAPREDUCE	default	0	Mon Mar 20 13:18:09 +0700 2023	Mon Mar 20 13:21:35 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1679285362016_0010	root	wordcount.jar	MAPREDUCE	default	0	Mon Mar 20 12:14:41 +0700 2023	Mon Mar 20 12:19:24 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1679285362016_0001	hive	HIVE-6400c9bf1ab1-4c1d-8ca4-a7449da8744e	TEZ	default	0	Mon Mar 20 11:15:27 +0700 2023	Mon Mar 20 11:27:22 +0700 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0
application_1678680733281_0001	hive	HIVE-85d21ee6-2c05-46ba-bef1-1d71d7b3218b	TEZ	default	0	Mon Mar 13 13:52:14 +0700 2023	Mon Mar 13 13:57:20 +0700 2023	FAILED	FAILED	N/A	N/A	N/A	N/A	N/A	0.0
application_1678635517059_0001	hive	HIVE-38c53b4a-7011-40d2-	TEZ	default	0	Sun Mar 12 22:46:29	Sun Mar 12 23:04:07	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0

4. Result

```

[root@sandbox-hdp tpdmr]# hdfs dfs -cat output/part-00000
Bytes Written=2187
AD 48991
AE 47361
AF 49699
AG 48991
AI 46823
AL 49153
AM 40530
AN 48698
AO 44383
AQ 44854
AR 47320
AS 33192
AT 46708
AU 49962
AW 47652
AX 46480
AZ 38710
BA 49415
BB 41976
BD 43241
BE 49392
BF 49741
BG 49842
BH 49097
BI 35932
BJ 44400
BM 46669
BN 49418
BO 47851
BR 49568
BS 44043
BT 39670
BV 47810
BW 48975
BY 37689
BZ 45601
CA 46716
CC 48854

```