

## BDA LAB ASSIGNMENT NO.5

### MAP REDUCE 2

NAME : KETAKI PATIL  
ROLL NO : PA-17  
BATCH : A1

---

### CODE :

#### WORD MAPPER

```
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class w_Mapper
    extends Mapper<LongWritable, Text, Text, IntWritable> {

    private static final int MISSING = 9999;

    @Override
    public void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException {

        String line = value.toString();
        String year = line.substring(15, 19);
        int airTemperature;
        if (line.charAt(87) == '+') { // parseInt doesn't like leading
plus signs
            airTemperature = Integer.parseInt(line.substring(88, 92));
        } else {
            airTemperature = Integer.parseInt(line.substring(87, 92));
        }
        String quality = line.substring(92, 93);
        if (airTemperature != MISSING && quality.matches("[01459]")) {
            context.write(new Text(year), new IntWritable(airTemperature));
        }
    }
}
```

## WORD DRIVER

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class w_Driver {

    public static void main(String[] args) throws Exception {
        if (args.length != 2) {
            System.err.println("Usage: MaxTemperature <input path> <output path>");
            System.exit(-1);
        }

        Job job = new Job();
        job.setJarByClass(w_Driver.class);
        job.setJobName("w_Driver");

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        job.setMapperClass(w_Mapper.class);
        job.setReducerClass(w_Reducer.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

## WORD REDUCER

```
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class w_Reducer
    extends Reducer<Text, IntWritable, Text, IntWritable> {

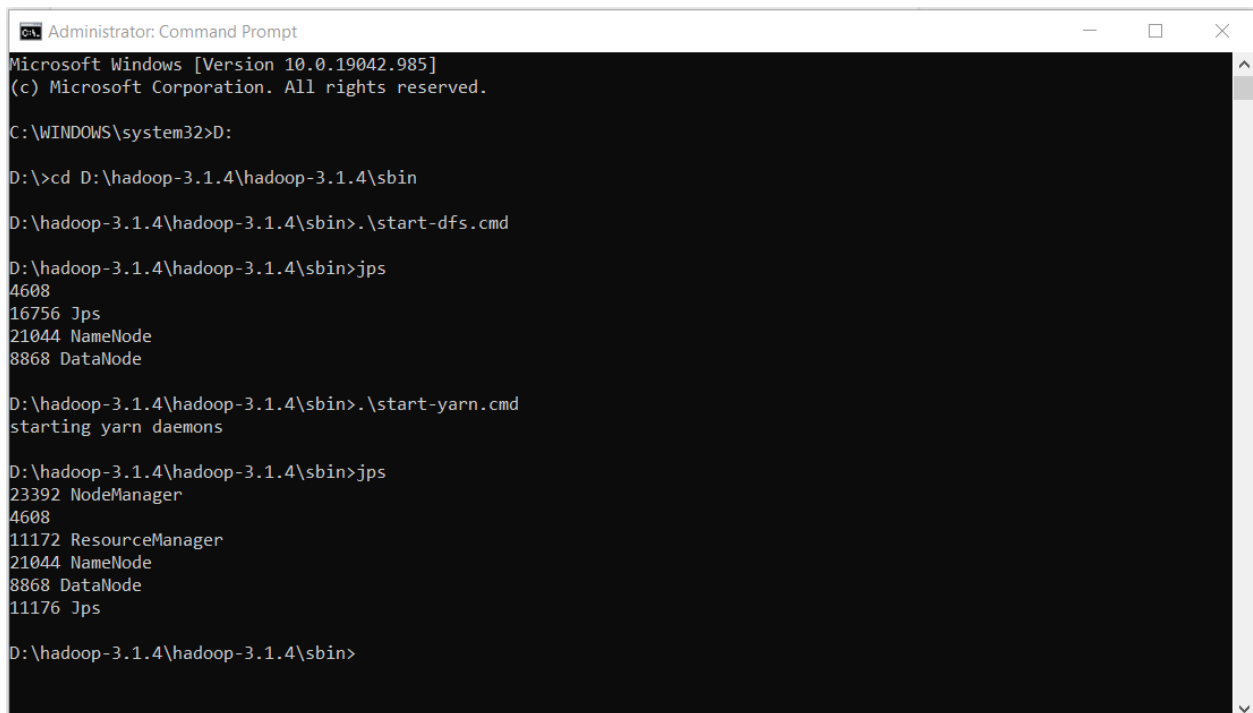
    @Override
    public void reduce(Text key, Iterable<IntWritable> values,
```

```
Context context)
throws IOException, InterruptedException {

    int maxValue = Integer.MIN_VALUE;
    for (IntWritable value : values) {
        maxValue = Math.max(maxValue, value.get());
    }
    context.write(key, new IntWritable(maxValue/10));
}
}
```

## OUTPUT :

**Starting datanode, namenode, node manager and resource manager through start dfs and yarn commands.**



```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19042.985]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>D:

D:\>cd D:\hadoop-3.1.4\hadoop-3.1.4\sbin

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>.\start-dfs.cmd

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>jps
4608
16756 Jps
21044 NameNode
8868 DataNode

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>.\start-yarn.cmd
starting yarn daemons

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>jps
23392 NodeManager
4608
11172 ResourceManager
21044 NameNode
8868 DataNode
11176 Jps

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>
```

# Datanode and Namenode

```
Apache Hadoop Distribution - hadoop datanode
2021-06-01 18:45:42,524 INFO impl.FsDatasetImpl: Time00 milliseconds
n D:\Hadoop\data\datanode: 106ms
2021-06-01 18:45:42,524 INFO impl.FsDatasetImpl: Total00 milliseconds
9ee263a-33a1-4e03-94a7-0d9824b7d599;nsid=1653617406;c=1620916509044) storage fea80054-3ea8-445a-bc4a-4961dbd7cb13
7-1620916509044: 108ms
2021-06-01 18:45:42,528 INFO impl.FsDatasetImpl: Addi2021-06-01 18:45:42,722 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
509044 on volume D:\Hadoop\data\datanode...
2021-06-01 18:45:42,722 INFO blockmanagement.BlockReportLeaseManager: Registered DW fea80054-3ea8-445a-bc4a-4961dbd7cb13
2021-06-01 18:45:42,529 INFO impl.BlockPoolSlice: Rep (127.0.0.1:9866).
68.1.27-1620916509044)current/replicas doesn't exist 2021-06-01 18:45:42,818 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-c5e150a7-f692-4180-9d7b-ad3edd
2021-06-01 18:45:42,565 INFO impl.FsDatasetImpl: Timead220d for DW 127.0.0.1:9866
20916509044 on volume D:\Hadoop\data\datanode: 36ms 2021-06-01 18:45:42,922 INFO BlockStateChange: BLOCK* processReport 0x6b97955f735f4a8b: Processing first storage report
2021-06-01 18:45:42,568 INFO impl.FsDatasetImpl: Totalfor DS-c5e150a7-f692-4180-9d7b-ad3eddad220d from datanode fea80054-3ea8-445a-bc4a-4961dbd7cb13
68.1.27-1620916509044: 41ms 2021-06-01 18:45:42,930 INFO blockmanagement.BlockManager: Initializing replication queues
2021-06-01 18:45:42,568 INFO checker.ThrottledAsyncCh2021-06-01 18:45:42,931 INFO hdfs.StateChange: STATE* Safe mode extension entered.
2021-06-01 18:45:42,582 INFO checker.DatasetVolumeCheThe reported blocks 26 has reached the threshold 0.9990 of total blocks 27. The minimum number of live datanodes is not
2021-06-01 18:45:42,636 INFO datanode.VolumeScanner: required. In safe mode extension. Safe mode will be turned off automatically in 29 seconds.
2021-06-01 18:45:42,646 INFO datanode.DirectoryScanne80-9d7b-ad3eddad220d node DatanodeRegistration(127.0.0.1:9866, datanodeUId=fea80054-3ea8-445a-bc4a-4961dbd7cb13, infoPo
8 PM with interval of 21600000ms rt=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57,cid=CID-39ee263a-33a1-4e03-94a7-0d9824b7d599;nsid=1653617406
2021-06-01 18:45:42,654 INFO datanode.DataNode: Block;c=1620916509044), blocks: 27, hasStateStorage: false, processing time: 10 msec, invalidatedBlocks: 0
54-3ea8-445a-bc4a-4961dbd7cb13) service to localhost/2021-06-01 18:45:42,935 INFO blockmanagement.BlockManager: Total number of blocks = 27
2021-06-01 18:45:42,738 INFO datanode.DataNode: Block2021-06-01 18:45:42,935 INFO blockmanagement.BlockManager: Number of invalid blocks = 0
UId fea80054-3ea8-445a-bc4a-4961dbd7cb13) service to2021-06-01 18:45:42,936 INFO blockmanagement.BlockManager: Number of under-replicated blocks = 8
2021-06-01 18:45:42,739 INFO datanode.DataNode: For r2021-06-01 18:45:42,936 INFO blockmanagement.BlockManager: Number of over-replicated blocks = 0
0000msec CACHEREPORT_INTERVAL of 10000msec Initial de2021-06-01 18:45:42,937 INFO blockmanagement.BlockManager: Number of blocks being written = 0
2021-06-01 18:45:42,980 INFO datanode.DataNode: Succ2021-06-01 18:45:42,937 INFO hdfs.StateChange: STATE* Replication Queue initialization scan for invalid, over- and under
report(s), of which we sent 1. The reports had 27 to-replicated blocks completed in 6 msec
msecs for RPC and NN processing. Got back one command2021-06-01 18:46:03,055 INFO hdfs.StateChange: STATE* Safe mode ON, in safe mode extension.
2021-06-01 18:45:42,981 INFO datanode.DataNode: Got tThe reported blocks 27 has reached the threshold 0.9990 of total blocks 27. The minimum number of live datanodes is not
044 required. In safe mode extension. Safe mode will be turned off automatically in 9 seconds.
```

# Resource Manager and Node Manager

```
Apache Hadoop Distribution - yarn resourcemanager
StoreEventType for class org.apache.hadoop.yarn.nodelabels.CommandJun 01, 2021 6:47:11 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
2021-06-01 18:46:59,752 INFO ipc.CallQueueManager: Using callQueueINFO: Initiating Jersey application, version 'Jersey: 1.19.02/11/2015 03:25 AM'
Capacity: 5000 scheduler: class org.apache.hadoop.ipc.DefaultRpcJun 01, 2021 6:47:11 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
2021-06-01 18:46:59,755 INFO ipc.Server: Starting Socket Reader INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver to GuiceManagedComponentProvider with
2021-06-01 18:46:59,825 INFO pb.RpcServerFactoryPBImpl: Adding p the scope "Singleton"
rPB to the server
2021-06-01 18:46:59,832 INFO ipc.Server: IPC Server Responder: Jun 01, 2021 6:47:12 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
2021-06-01 18:46:59,832 INFO ipc.Server: IPC Server listener on gleton"
2021-06-01 18:46:59,840 INFO util.JvmPauseMonitor: Starting JVM Jun 01, 2021 6:47:12 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
2021-06-01 18:47:00,074 INFO ipc.CallQueueManager: Using callQueueINFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the s
Capacity: 5000 scheduler: class org.apache.hadoop.ipc.DefaultRpcJun 01, 2021 6:47:12 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
2021-06-01 18:47:00,085 INFO ipc.Server: Starting Socket Reader 2021-06-01 18:47:13,023 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@15dc339f(node/,file:///C:/Users/keta
2021-06-01 18:47:00,101 INFO pb.RpcServerFactoryPBImpl: Adding rk/AppData/Local/Temp/jetty-0_0_0-8042-_-any-8896537751553535376.dir/webapp/,AVAILABLE){jar:file:D:/hadoop-3.1.4/hadoo
ocolPB to the server
2021-06-01 18:47:00,102 INFO ipc.Server: IPC Server Responder: p-3.1.4/share/hadoop/yarn/hadoop-yarn-common-3.1.4.jar/webapps/node)
2021-06-01 18:47:00,102 INFO ipc.Server: IPC Server listener on p-3.1.4/share/hadoop/yarn/hadoop-yarn-common-3.1.4.jar/webapps/node)
2021-06-01 18:47:00,581 INFO ipc.CallQueueManager: Using callQue2021-06-01 18:47:13,045 INFO server.Server: Started @23162ms
Capacity: 5000 scheduler: class org.apache.hadoop.ipc.DefaultRpc2021-06-01 18:47:13,046 INFO webapp.WebApps: Web app node started at 8042
2021-06-01 18:47:00,584 INFO ipc.Server: Starting Socket Reader 2021-06-01 18:47:13,755 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : LAPTOP-CGT0UDGT:49389
2021-06-01 18:47:00,587 INFO pb.RpcServerFactoryPBImpl: Adding r2021-06-01 18:47:13,758 INFO util.JvmPauseMonitor: Starting JVM pause monitor
ocolPB to the server
2021-06-01 18:47:00,588 INFO ipc.Server: IPC Server Responder: 2021-06-01 18:47:14,113 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
2021-06-01 18:47:00,593 INFO resource.ResourceTracker: Tr2021-06-01 18:47:15,152 INFO security.NMContainerTokenSecretManager: Registering with RM using containers: []
2021-06-01 18:47:15,024 INFO resource.ResourceTrackerServth id 228435363
httpPort: 8042) registered with capability: <memory:8192, vCore2021-06-01 18:47:15,159 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key wi id
2021-06-01 18:47:15,098 INFO rmnode.RMNodeImpl: LAPTOP-CGT0UDGT: -1629386179
2021-06-01 18:47:15,175 INFO capacity.CapacityScheduler: Added r2021-06-01 18:47:15,160 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as LAPTOP-CGT0UDGT:49389
vCores:8> with total resource of <memory:8192, vCores:8>
```

# Making a new directory in hadoop

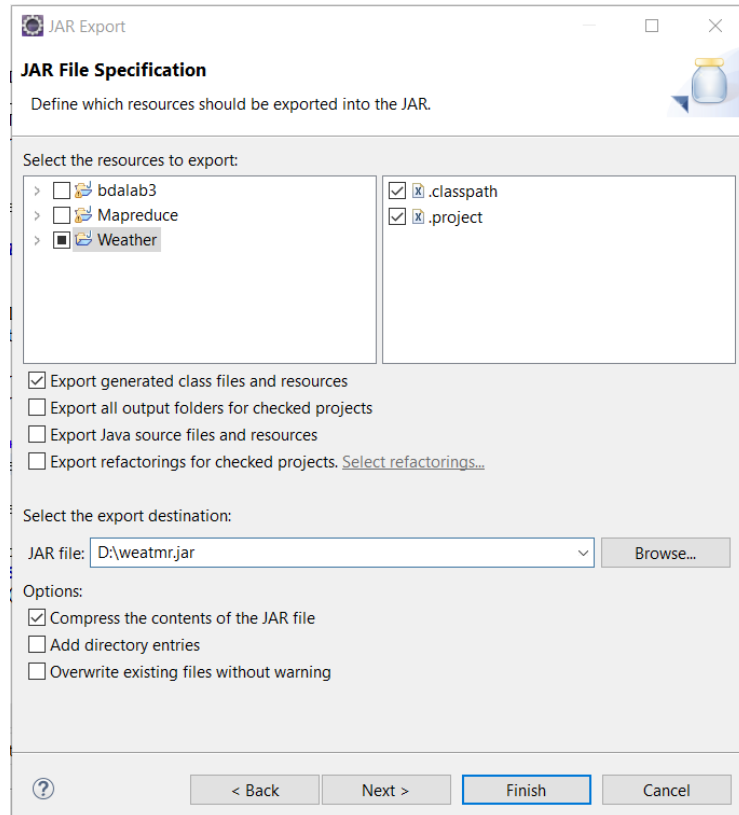
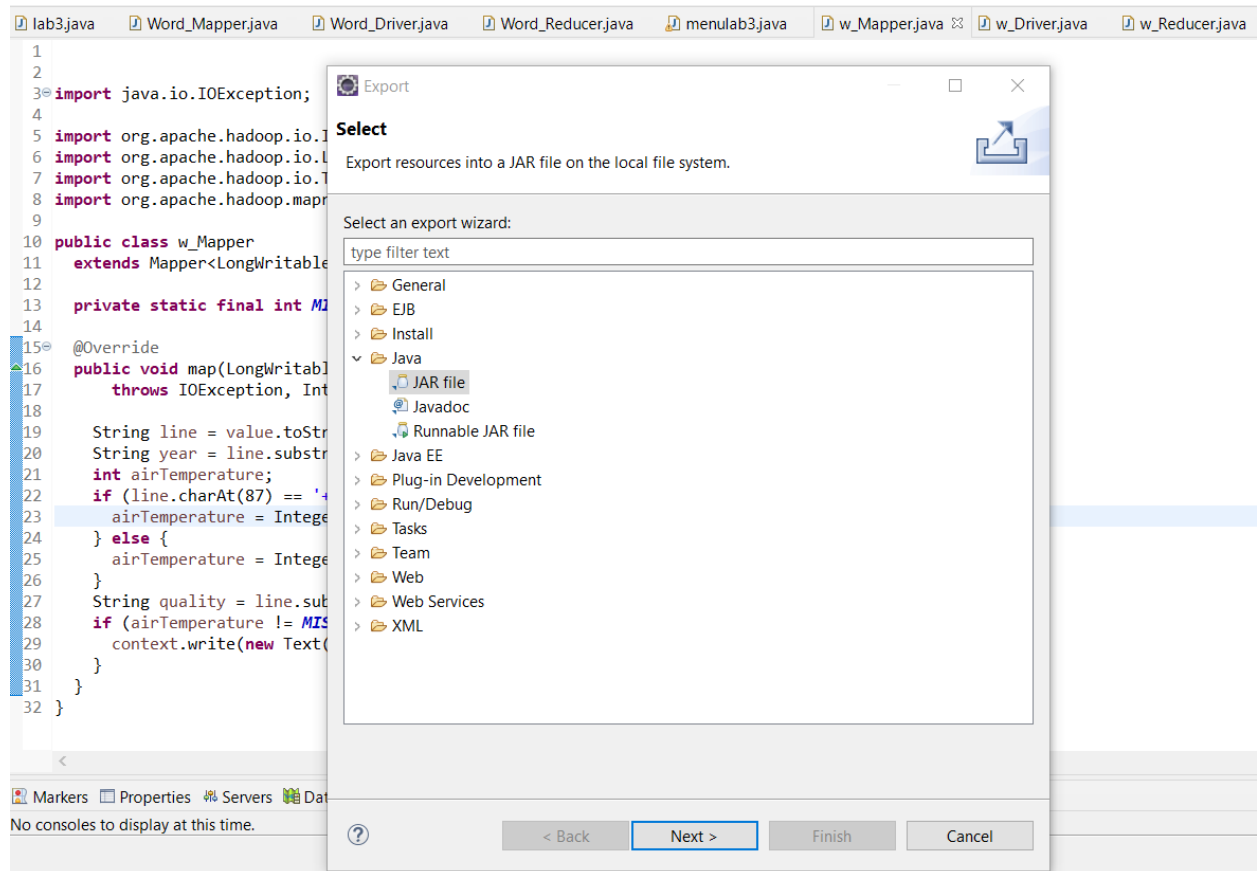
```
D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hdfs dfs -ls /
Found 9 items
drwxr-xr-x - ketak supergroup 0 2021-05-24 12:12 /Weather
drwxr-xr-x - ketak supergroup 0 2021-06-01 18:58 /WordCount
drwxr-xr-x - ketak supergroup 0 2021-05-18 22:20 /WordMap
-rw-r--r-- 1 ketak supergroup 60 2021-05-18 22:06 /hdfs
-rw-r--r-- 1 ketak supergroup 60 2021-05-18 22:07 /ketaki
-rw-r--r-- 1 ketak supergroup 60 2021-05-18 22:13 /mapper
drwx----- - ketak supergroup 0 2021-05-21 13:02 /tmp
drwxr-xr-x - ketak supergroup 0 2021-05-20 11:19 /user
-rw-r--r-- 1 ketak supergroup 123 2021-05-18 22:16 /wordmapper

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>hdfs dfs -ls /Weather/
Found 2 items
-rw-r--r-- 1 ketak supergroup 888190 2021-05-24 12:12 /Weather/1901.txt
-rw-r--r-- 1 ketak supergroup 888978 2021-05-24 12:12 /Weather/1902.txt

D:\hadoop-3.1.4\hadoop-3.1.4\sbin>cd ..

D:\hadoop-3.1.4\hadoop-3.1.4>cd bin
```

## Jar file of project is exported from java program and stored :



## Jar file stored :



weatmr  
Executable Jar File  
3.87 KB

## Execution of project :

```
Administrator: Command Prompt
C:\hadoop-3.1.4\hadoop-3.1.4\bin>hadoop jar D:\weatmr.jar w_Driver /Weather/ weatmr_out
021-06-03 19:42:40,262 INFO client.RPCProxy: Connecting to ResourceManager at /0.0.0.0:8032
021-06-03 19:42:40,839 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
021-06-03 19:42:40,926 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ketak/.staging/job_1622726027131_0004
021-06-03 19:42:41,622 INFO input.FileInputFormat: Total input files to process : 2
021-06-03 19:42:42,138 INFO mapreduce.JobSubmitter: number of splits:2
021-06-03 19:42:42,574 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622726027131_0004
021-06-03 19:42:42,577 INFO mapreduce.JobSubmitter: Executing with tokens: []
021-06-03 19:42:42,992 INFO conf.Configuration: resource-types.xml not found
021-06-03 19:42:42,993 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
021-06-03 19:42:43,272 INFO impl.YarnClientImpl: Submitted application application_1622726027131_0004
021-06-03 19:42:43,411 INFO mapreduce.Job: The url to track the job: http://LAPTOP-CGT0UDGT:8088/proxy/application_1622726027131_0004/
021-06-03 19:42:43,414 INFO mapreduce.Job: Running job: job_1622726027131_0004
021-06-03 19:42:55,701 INFO mapreduce.Job: Job job_1622726027131_0004 running in uber mode : false
021-06-03 19:42:55,702 INFO mapreduce.Job: map 0% reduce 0%
021-06-03 19:43:04,054 INFO mapreduce.Job: map 50% reduce 0%
021-06-03 19:43:05,868 INFO mapreduce.Job: map 100% reduce 0%
021-06-03 19:43:14,996 INFO mapreduce.Job: map 100% reduce 100%
021-06-03 19:43:19,058 INFO mapreduce.Job: Job job_1622726027131_0004 completed successfully
021-06-03 19:43:19,141 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=144425
  FILE: Number of bytes written=956019
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1777374
  HDFS: Number of bytes written=16
  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)=14590
  Total time spent by all reduces in occupied slots (ms)=4903
  Total time spent by all map tasks (ms)=14590
  Total time spent by all reduce tasks (ms)=4903
  Total vcore-milliseconds taken by all map tasks=14590
  Total vcore-milliseconds taken by all reduce tasks=4903
  Total megabyte-milliseconds taken by all map tasks=14948160
  Total megabyte-milliseconds taken by all reduce tasks=5020672
Map-Reduce Framework
  Map input records=13130
  Map output records=13129
  Map output bytes=118161
  Map output materialized bytes=144431
  Input split bytes=206
```

## Active Nodes in Hadoop Web interface at localhost:8088 :

localhost:8088/cluster

### All Applications

Cluster

About

Nodes

Node Labels

Applications

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
4	0	0	4	0	0 B	8 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	Maximum Memory
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	Reserved CPU Vcores	Reserved Memory MB	% of Vcores
application_1622726027131_0004	ketak	w_Driver	MAPREDUCE	default	0	Thu Jun 3 19:42:43 +0550 2021	Thu Jun 3 19:43:17 +0550 2021	Thu Jun 3 19:43:17 +0550 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0%

## Max Temp on cmd :

```
D:\hadoop-3.1.4\hadoop-3.1.4\bin>hadoop fs -cat /user/ketak/weatmr_out/part-r-00000
1901    31
1902    24

D:\hadoop-3.1.4\hadoop-3.1.4\bin>
```

## Downloading part-r-00000 from localhost:9870 :

localhost:9870/explorer.html#/user/ketak/weatmr\_out

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

### Browse Directory

/user/ketak/weatmr\_out

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	ketak	supergroup	0 B	Jun 03 19:43	1	128 MB	_SUCCESS
<input type="checkbox"/>	-rw-r--r--	ketak	supergroup	16 B	Jun 03 19:43	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2020.

## Max Temp file :

```
part-r-00000 (2) - Notepad
File Edit Format View Help
1901    31
1902    24
```