Map-Reduce Assignment:
Counting frequency of words in an input text file

Mapper class:Map1

```
import java.io.IOException;
import java.util.*;

import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;


public class Map1 extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
 }
```

Reducer class:Red1

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class Red1 extends Reducer<Text, IntWritable, Text, IntWritable> {

         public void reduce(Text _key, Iterable<IntWritable> values, Context
context)
                           throws IOException, InterruptedException {
                  // process values
```

```
        int count=0;
        for (IntWritable val : values) {
            count += val.get();
        }
        context.write(_key, new IntWritable(count));
    }
}
```

Driver class:Dri1

```
import org.apache.hadoop.conf.Configuration;
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 Dri1 {

    @SuppressWarnings("deprecation")
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = new Job(conf, "wordcount");
        job.setJarByClass(Dri1.class);
        job.setMapperClass(Map1.class);

        job.setReducerClass(Red1.class);

        // TODO: specify output types
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        // TODO: specify input and output DIRECTORIES (not files)
        FileInputFormat.setInputPaths(job, new Path(args[1]));
        FileOutputFormat.setOutputPath(job, new Path(args[2]));

        if (!job.waitForCompletion(true))
            return;
    }
}
```

Input file : input.txt

```
Apache Hadoop (pronunciation: /hÉ™ËˆduË□p/) is an open-source
software framework for distributed storage and distributed processing of
very large data sets on computer clusters built from commodity hardware.
All the modules in Hadoop are designed with a fundamental assumption
that hardware failures are common and should be automatically handled by
the framework.[2]

The core of Apache Hadoop consists of a storage part, known as Hadoop
Distributed File System (HDFS), and a processing part called MapReduce.
Hadoop splits files into large blocks and distributes them across nodes in a
cluster. To process data, Hadoop transfers packaged code for nodes to
process in parallel based on the data that needs to be processed. This
approach takes advantage of data locality[3] â€“ nodes manipulating the
data they have access to â€“ to allow the dataset to be processed faster
and more efficiently than it would be in a more conventional supercomputer
architecture that relies on a parallel file system where computation and
data are distributed via high-speed networking.[4]

The base Apache Hadoop framework is composed of the following
modules:

    Hadoop Common â€“ contains libraries and utilities needed by other
Hadoop modules;
    Hadoop Distributed File System (HDFS) â€“ a distributed file-system that
stores data on commodity machines, providing very high aggregate
bandwidth across the cluster;
```

Output file:part-r-00000

```
(HDFS)      1
(HDFS),     1
(pronunciation:    1
/hÉ™ËˆduË□p/)  1
All     1
Apache      3
Common    1
Distributed 2
File   2
```

```
Hadoop	10
MapReduce.	1
System	2
The	2
This	1
To	1
a	7
access	1
across	2
advantage	1
aggregate	1
allow	1
an	1
and	7
approach	1
architecture	1
are	3
as	1
assumption	1
automatically	1
bandwidth	1
base	1
based	1
be	4
blocks	1
built	1
by	2
called	1
cluster.	1
cluster;	1
clusters	1
code	1
commodity	2
common	1
composed	1
computation	1
computer	1
consists	1
contains	1
conventional	1
```

```
core	1
data	6
data,	1
dataset	1
designed	1
distributed	4
distributes	1
efficiently	1
failures	1
faster	1
file	1
file-system	1
files	1
following	1
for	2
framework	2
framework.[2]	1
from	1
fundamental	1
handled	1
hardware	1
hardware.	1
have	1
high	1
high-speed	1
in	4
into	1
is	2
it	1
known	1
large	2
libraries	1
locality[3]	1
machines,	1
manipulating	1
modules	1
modules:	1
modules;	1
more	2
needed	1
```

```
needs	1
networking.[4]	1
nodes	3
of	5
on	4
open-source	1
other	1
packaged	1
parallel	2
part	1
part,	1
process	2
processed	1
processed.	1
processing	2
providing	1
relies	1
sets	1
should	1
software	1
splits	1
storage	2
stores	1
supercomputer	1
system	1
takes	1
than	1
that	4
the	7
them	1
they	1
to	5
transfers	1
utilities	1
very	2
via	1
where	1
with	1
would	1
â€“	4
```

Terminal output

```
root@ccoew-desktop:/home/ccoew# hdfs dfs -mkdir
/wordcountbatchb

root@ccoew-desktop:/home/ccoew# hdfs dfs -put
/home/ccoew/batchbinput.txt  /wordcountbatchb

root@ccoew-desktop:/home/ccoew# hadoop jar
/home/ccoew/batchbword.jar  Dri1  /home/ccoew/batchbinput.txt
/wordcountbatchb /output

16/09/23 16:53:54 INFO Configuration.deprecation: session.id is
deprecated. Instead, use dfs.metrics.session-id
16/09/23 16:53:54 INFO jvm.JvmMetrics: Initializing JVM Metrics with
processName=JobTracker, sessionId=
16/09/23 16:53:54 WARN mapreduce.JobSubmitter: Hadoop command-
line option parsing not performed. Implement the Tool interface and
execute your application with ToolRunner to remedy this.
16/09/23 16:53:54 INFO input.FileInputFormat: Total input paths to process
: 1
16/09/23 16:53:54 INFO mapreduce.JobSubmitter: number of splits:1
16/09/23 16:53:54 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_local949013430_0001
16/09/23 16:53:55 INFO mapreduce.Job: The url to track the job:
http://localhost:8080/
16/09/23 16:53:55 INFO mapreduce.Job: Running job:
job_local949013430_0001
16/09/23 16:53:55 INFO mapred.LocalJobRunner: OutputCommitter set in
config null
16/09/23 16:53:55 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
16/09/23 16:53:55 INFO mapred.LocalJobRunner: Waiting for map tasks
16/09/23 16:53:55 INFO mapred.LocalJobRunner: Starting task:
attempt_local949013430_0001_m_000000_0
16/09/23 16:53:55 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
16/09/23 16:53:55 INFO mapred.MapTask: Processing split:
hdfs://localhost:9000/ankita/log.txt:0+124829
```

```
16/09/23 16:53:55 INFO mapred.MapTask: (EQUATOR) 0 kvi
26214396(104857584)
16/09/23 16:53:55 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/09/23 16:53:55 INFO mapred.MapTask: soft limit at 83886080
16/09/23 16:53:55 INFO mapred.MapTask: bufstart = 0; bufvoid =
104857600
16/09/23 16:53:55 INFO mapred.MapTask: kvstart = 26214396; length =
6553600
16/09/23 16:53:55 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/09/23 16:53:55 INFO mapred.LocalJobRunner:
16/09/23 16:53:55 INFO mapred.MapTask: Starting flush of map output
16/09/23 16:53:55 INFO mapred.MapTask: Spilling map output
16/09/23 16:53:55 INFO mapred.MapTask: bufstart = 0; bufend = 40241;
bufvoid = 104857600
16/09/23 16:53:55 INFO mapred.MapTask: kvstart =
26214396(104857584); kvend = 26206908(104827632); length =
7489/6553600
16/09/23 16:53:55 INFO mapred.MapTask: Finished spill 0
16/09/23 16:53:55 INFO mapred.Task:
Task:attempt_local949013430_0001_m_000000_0 is done. And is in the
process of committing
16/09/23 16:53:55 INFO mapred.LocalJobRunner: map
16/09/23 16:53:55 INFO mapred.Task: Task
'attempt_local949013430_0001_m_000000_0' done.
16/09/23 16:53:55 INFO mapred.LocalJobRunner: Finishing task:
attempt_local949013430_0001_m_000000_0
16/09/23 16:53:55 INFO mapred.LocalJobRunner: map task executor
complete.
16/09/23 16:53:55 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/09/23 16:53:55 INFO mapred.LocalJobRunner: Starting task:
attempt_local949013430_0001_r_000000_0
16/09/23 16:53:55 INFO mapred.Task:  Using
ResourceCalculatorProcessTree : [ ]
16/09/23 16:53:55 INFO mapred.ReduceTask: Using
ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@7bfbde18
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: MergerManager:
memoryLimit=333971456, maxSingleShuffleLimit=83492864,
mergeThreshold=220421168, ioSortFactor=10,
memToMemMergeOutputsThreshold=10
16/09/23 16:53:55 INFO reduce.EventFetcher:
attempt_local949013430_0001_r_000000_0 Thread started: EventFetcher
for fetching Map Completion Events
16/09/23 16:53:55 INFO reduce.LocalFetcher: localfetcher#1 about to
shuffle output of map attempt_local949013430_0001_m_000000_0
decomp: 43989 len: 43993 to MEMORY
16/09/23 16:53:55 INFO reduce.InMemoryMapOutput: Read 43989 bytes
from map-output for attempt_local949013430_0001_m_000000_0
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: closeInMemoryFile ->
map-output of size: 43989, inMemoryMapOutputs.size() -> 1,
commitMemory -> 0, usedMemory ->43989
16/09/23 16:53:55 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/09/23 16:53:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: finalMerge called with
1 in-memory map-outputs and 0 on-disk map-outputs
16/09/23 16:53:55 INFO mapred.Merger: Merging 1 sorted segments
16/09/23 16:53:55 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 43969 bytes
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: Merged 1 segments,
43989 bytes to disk to satisfy reduce memory limit
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: Merging 1 files, 43993
bytes from disk
16/09/23 16:53:55 INFO reduce.MergeManagerImpl: Merging 0 segments,
0 bytes from memory into reduce
16/09/23 16:53:55 INFO mapred.Merger: Merging 1 sorted segments
16/09/23 16:53:55 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 43969 bytes
16/09/23 16:53:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/09/23 16:53:55 INFO Configuration.deprecation: mapred.skip.on is
deprecated. Instead, use mapreduce.job.skiprecords
16/09/23 16:53:56 INFO mapreduce.Job: Job job_local949013430_0001
running in uber mode : false
16/09/23 16:53:56 INFO mapreduce.Job:  map 100% reduce 0%
16/09/23 16:53:56 INFO mapred.Task:
Task:attempt_local949013430_0001_r_000000_0 is done. And is in the
process of committing
16/09/23 16:53:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
```

```
16/09/23 16:53:56 INFO mapred.Task: Task
attempt_local949013430_0001_r_000000_0 is allowed to commit now
16/09/23 16:53:56 INFO output.FileOutputCommitter: Saved output of task
'attempt_local949013430_0001_r_000000_0' to
hdfs://localhost:9000/ankita/op/_temporary/0/task_local949013430_0001_r
_000000
16/09/23 16:53:56 INFO mapred.LocalJobRunner: reduce > reduce
16/09/23 16:53:56 INFO mapred.Task: Task
'attempt_local949013430_0001_r_000000_0' done.
16/09/23 16:53:56 INFO mapred.LocalJobRunner: Finishing task:
attempt_local949013430_0001_r_000000_0
16/09/23 16:53:56 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/09/23 16:53:57 INFO mapreduce.Job:  map 100% reduce 100%
16/09/23 16:53:57 INFO mapreduce.Job: Job job_local949013430_0001
completed successfully
16/09/23 16:53:57 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=99538
                FILE: Number of bytes written=642043
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=249658
                HDFS: Number of bytes written=1689
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=1873
                Map output records=1873
                Map output bytes=40241
                Map output materialized bytes=43993
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=84
                Reduce shuffle bytes=43993
                Reduce input records=1873
                Reduce output records=84
                Spilled Records=3746
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=28
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=496500736
        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=1405
        File Output Format Counters
                Bytes Written=1240
root@ccoew-desktop:/home/ccoew#
```