MAPPER

```
package MapReduceJoin;

import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FSDataInputStream;
import org.apache.hadoop.fs.FSDataOutputStream;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

import org.apache.hadoop.io.IntWritable;

public class DeptEmpStrengthMapper extends MapReduceBase implements
Mapper<LongWritable, Text, TextPair, Text> {

	@Override
	public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reporter reporter)
				throws IOException
	{

		String valueString = value.toString();
		String[] SingleNodeData = valueString.split("\t");
		output.collect(new TextPair(SingleNodeData[0], "1"), new
Text(SingleNodeData[1]));
	}
}
```

```
package MapReduceJoin;

import java.io.IOException;

import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
```

```
public class DeptNameMapper extends MapReduceBase implements Mapper<LongWritable,
Text, TextPair, Text>  {

	@Override
	public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output,
Reporter reporter)
				throws IOException
	{
		String valueString = value.toString();
		String[] SingleNodeData = valueString.split("\t");
		output.collect(new TextPair(SingleNodeData[0], "0"), new
Text(SingleNodeData[1]));
	}
}
```

DRIVER

```
package MapReduceJoin;

import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.mapred.lib.MultipleInputs;
import org.apache.hadoop.util.*;


public class JoinDriver extends Configured implements Tool {

	public static class KeyPartitioner implements Partitioner<TextPair, Text> {
		@Override
		public void configure(JobConf job) {}

		@Override
		public int getPartition(TextPair key, Text value, int numPartitions) {
			return (key.getFirst().hashCode() & Integer.MAX_VALUE) %
numPartitions;
		}
	}

	@Override
	public int run(String[] args) throws Exception {
```

```
		if (args.length != 3) {
			System.out.println("Usage: <Department Emp Strength input>
<Department Name input> <output>");
			return -1;
		}

		JobConf conf = new JobConf(getConf(), getClass());
		conf.setJobName("Join 'Department Emp Strength input' with 'Department Name
input'");

		Path AInputPath = new Path(args[0]);
		Path BInputPath = new Path(args[1]);
		Path outputPath = new Path(args[2]);

		MultipleInputs.addInputPath(conf, AInputPath, TextInputFormat.class,
DeptNameMapper.class);
		MultipleInputs.addInputPath(conf, BInputPath, TextInputFormat.class,
DeptEmpStrengthMapper.class);

		FileOutputFormat.setOutputPath(conf, outputPath);

		conf.setPartitionerClass(KeyPartitioner.class);
		conf.setOutputValueGroupingComparator(TextPair.FirstComparator.class);

		conf.setMapOutputKeyClass(TextPair.class);

		conf.setReducerClass(JoinReducer.class);

		conf.setOutputKeyClass(Text.class);

		JobClient.runJob(conf);

		return 0;
	}

	public static void main(String[] args) throws Exception {

		int exitCode = ToolRunner.run(new JoinDriver(), args);
		System.exit(exitCode);
	}
}
```

REDUCER

```
package MapReduceJoin;

import java.io.IOException;
import java.util.Iterator;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class JoinReducer extends MapReduceBase implements Reducer<TextPair, Text, Text,
Text> {

	@Override
	public void reduce (TextPair key, Iterator<Text> values, OutputCollector<Text, Text>
output, Reporter reporter)
			throws IOException
	{

		Text nodeId = new Text(values.next());
		while (values.hasNext()) {
			Text node = values.next();
			Text outValue = new Text(nodeId.toString() + "\t\t" + node.toString());
			output.collect(key.getFirst(), outValue);
		}
	}
}
```

```
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -mkdir /cs228/join
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -put ~/eclipse-
workspace/cs228/MapReduceJoin/DeptStrength.txt /cs228/join/
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -put ~/eclipse-
workspace/cs228/MapReduceJoin/DeptName.txt /cs228/join/
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ ls
AvgTemp      MapReduceJoin      WordCount
AvgTemp.jar  MapReduceJoin.jar  WordCount.jar
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop jar MapReduceJoin.jar
MapReduceJoin.JoinDriver /cs228/join/DeptStrength.txt /cs228/join/Deptname.txt /cs228/join/output
Usage: <Department Emp Strength input> <Department Name input> <output>
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -put ~/eclipse-
workspace/cs228/MapReduceJoin/DeptEmpStrength.txt /cs228/join/
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop jar MapReduceJoin.jar
MapReduceJoin.JoinDriver /cs228/join/DeptEmpStrength.txt /cs228/join/Deptname.txt /cs228/join/output
Usage: <Department Emp Strength input> <Department Name input> <output>
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs /cs228/join
/cs228/join: Unknown command
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hdfs dfs -ls /cs228/joinFound 3 items
-rw-r--r--   1 hduser supergroup         50 2022-06-22 15:19 /cs228/join/DeptEmpStrength.txt
-rw-r--r--   1 hduser supergroup         59 2022-06-22 15:16 /cs228/join/DeptName.txt
-rw-r--r--   1 hduser supergroup         50 2022-06-22 15:16 /cs228/join/DeptStrength.txt
```

```
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop jar MapReduceJoin.jar
/cs228/join/DeptEmpStrength.txt /cs228/join/DeptName.txt /cs228/join/output1
22/06/22 15:26:53 INFO Configuration.deprecation: session.id is deprecated. Instead, use
dfs.metrics.session-id
22/06/22 15:26:53 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
22/06/22 15:26:53 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker,
sessionId= - already initialized
22/06/22 15:26:53 INFO mapred.FileInputFormat: Total input paths to process : 1
22/06/22 15:26:53 INFO mapred.FileInputFormat: Total input paths to process : 1
22/06/22 15:26:53 INFO mapreduce.JobSubmitter: number of splits:2
22/06/22 15:26:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local910966636_0001
22/06/22 15:26:53 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
22/06/22 15:26:53 INFO mapred.LocalJobRunner: OutputCommitter set in config null
22/06/22 15:26:53 INFO mapreduce.Job: Running job: job_local910966636_0001
22/06/22 15:26:53 INFO mapred.LocalJobRunner: OutputCommitter is
org.apache.hadoop.mapred.FileOutputCommitter
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Waiting for map tasks
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Starting task: attempt_local910966636_0001_m_000000_0
22/06/22 15:26:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/06/22 15:26:53 INFO mapred.MapTask: Processing split:
hdfs://localhost:54310/cs228/join/DeptName.txt:0+59
22/06/22 15:26:53 INFO mapred.MapTask: numReduceTasks: 1
22/06/22 15:26:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/22 15:26:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
22/06/22 15:26:53 INFO mapred.MapTask: soft limit at 83886080
22/06/22 15:26:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/22 15:26:53 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
22/06/22 15:26:53 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/22 15:26:53 INFO mapred.LocalJobRunner:
22/06/22 15:26:53 INFO mapred.MapTask: Starting flush of map output
22/06/22 15:26:53 INFO mapred.MapTask: Spilling map output
22/06/22 15:26:53 INFO mapred.MapTask: bufstart = 0; bufend = 63; bufvoid = 104857600
22/06/22 15:26:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/22 15:26:53 INFO mapred.MapTask: Finished spill 0
22/06/22 15:26:53 INFO mapred.Task: Task:attempt_local910966636_0001_m_000000_0 is done. And is in the
process of committing
22/06/22 15:26:53 INFO mapred.LocalJobRunner: hdfs://localhost:54310/cs228/join/DeptName.txt:0+59
22/06/22 15:26:53 INFO mapred.Task: Task 'attempt_local910966636_0001_m_000000_0' done.
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local910966636_0001_m_000000_0
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Starting task: attempt_local910966636_0001_m_000001_0
22/06/22 15:26:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/06/22 15:26:53 INFO mapred.MapTask: Processing split:
```

```
hdfs://localhost:54310/cs228/join/DeptEmpStrength.txt:0+50
22/06/22 15:26:53 INFO mapred.MapTask: numReduceTasks: 1
22/06/22 15:26:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
22/06/22 15:26:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
22/06/22 15:26:53 INFO mapred.MapTask: soft limit at 83886080
22/06/22 15:26:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
22/06/22 15:26:53 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
22/06/22 15:26:53 INFO mapred.MapTask: Map output collector class =
org.apache.hadoop.mapred.MapTask$MapOutputBuffer
22/06/22 15:26:53 INFO mapred.LocalJobRunner:
22/06/22 15:26:53 INFO mapred.MapTask: Starting flush of map output
22/06/22 15:26:53 INFO mapred.MapTask: Spilling map output
22/06/22 15:26:53 INFO mapred.MapTask: bufstart = 0; bufend = 54; bufvoid = 104857600
22/06/22 15:26:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536);
length = 13/6553600
22/06/22 15:26:53 INFO mapred.MapTask: Finished spill 0
22/06/22 15:26:53 INFO mapred.Task: Task:attempt_local910966636_0001_m_000001_0 is done. And is in the
process of committing
22/06/22 15:26:53 INFO mapred.LocalJobRunner:
hdfs://localhost:54310/cs228/join/DeptEmpStrength.txt:0+50
22/06/22 15:26:53 INFO mapred.Task: Task 'attempt_local910966636_0001_m_000001_0' done.
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local910966636_0001_m_000001_0
22/06/22 15:26:53 INFO mapred.LocalJobRunner: map task executor complete.
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Waiting for reduce tasks
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Starting task: attempt_local910966636_0001_r_000000_0
22/06/22 15:26:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
22/06/22 15:26:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin:
org.apache.hadoop.mapreduce.task.reduce.Shuffle@6d1f64b0
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464,
maxSingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10,
memToMemMergeOutputsThreshold=10
22/06/22 15:26:53 INFO reduce.EventFetcher: attempt_local910966636_0001_r_000000_0 Thread started:
EventFetcher for fetching Map Completion Events
22/06/22 15:26:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local910966636_0001_m_000000_0 decomp: 73 len: 77 to MEMORY
22/06/22 15:26:53 INFO reduce.InMemoryMapOutput: Read 73 bytes from map-output for
attempt_local910966636_0001_m_000000_0
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 73,
inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->73
22/06/22 15:26:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local910966636_0001_m_000001_0 decomp: 64 len: 68 to MEMORY
22/06/22 15:26:53 INFO reduce.InMemoryMapOutput: Read 64 bytes from map-output for
attempt_local910966636_0001_m_000001_0
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 64,
```

```
org.apache.hadoop.io.nativeio.NativeIO$POSIX$CacheManipulator.posixFadviseIfPossible(NativeIO.java:146)
    at org.apache.hadoop.io.ReadaheadPool$ReadaheadRequestImpl.run(ReadaheadPool.java:206)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
22/06/22 15:26:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: finalMerge called with 2 in-memory map-outputs and 0
on-disk map-outputs
22/06/22 15:26:53 INFO mapred.Merger: Merging 2 sorted segments
22/06/22 15:26:53 INFO mapred.Merger: Down to the last merge-pass, with 2 segments left of total size:
121 bytes
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: Merged 2 segments, 137 bytes to disk to satisfy reduce
memory limit
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: Merging 1 files, 139 bytes from disk
22/06/22 15:26:53 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
22/06/22 15:26:53 INFO mapred.Merger: Merging 1 sorted segments
22/06/22 15:26:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
127 bytes
22/06/22 15:26:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/22 15:26:53 INFO mapred.Task: Task:attempt_local910966636_0001_r_000000_0 is done. And is in the
process of committing
22/06/22 15:26:53 INFO mapred.LocalJobRunner: 2 / 2 copied.
22/06/22 15:26:53 INFO mapred.Task: Task attempt_local910966636_0001_r_000000_0 is allowed to commit
now
22/06/22 15:26:53 INFO output.FileOutputCommitter: Saved output of task
'attempt_local910966636_0001_r_000000_0' to
hdfs://localhost:54310/cs228/join/output1/_temporary/0/task_local910966636_0001_r_000000
22/06/22 15:26:53 INFO mapred.LocalJobRunner: reduce > reduce
22/06/22 15:26:53 INFO mapred.Task: Task 'attempt_local910966636_0001_r_000000_0' done.
22/06/22 15:26:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local910966636_0001_r_000000_0
22/06/22 15:26:53 INFO mapred.LocalJobRunner: reduce task executor complete.
22/06/22 15:26:54 INFO mapreduce.Job: Job job_local910966636_0001 running in uber mode : false
22/06/22 15:26:54 INFO mapreduce.Job:  map 100% reduce 100%
22/06/22 15:26:54 INFO mapreduce.Job: Job job_local910966636_0001 completed successfully
22/06/22 15:26:54 INFO mapreduce.Job: Counters: 38
    File System Counters
        FILE: Number of bytes read=26365
        FILE: Number of bytes written=778758
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=277
        HDFS: Number of bytes written=85
        HDFS: Number of read operations=28
```

```
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=5
    Map-Reduce Framework
        Map input records=8
        Map output records=8
        Map output bytes=117
        Map output materialized bytes=145
        Input split bytes=442
        Combine input records=0
        Combine output records=0
        Reduce input groups=4
        Reduce shuffle bytes=145
        Reduce input records=8
        Reduce output records=4
        Spilled Records=16
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=2
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=916979712
    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=85
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop -ls  /cs228/join/output1
Error: No command named `-ls' was found. Perhaps you meant `hadoop ls'
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop fs -ls  /cs228/join/output1
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2022-06-22 15:26 /cs228/join/output1/_SUCCESS
-rw-r--r--   1 hduser supergroup         85 2022-06-22 15:26 /cs228/join/output1/part-00000
hduser@bmsce-Precision-T1700:~/eclipse-workspace/cs228$ hadoop fs -cat  /cs228/join/output1/part-00000
A11 50      Finance
B12 100     HR
C13 250     Manufacturing
Dept ID Total Employee      Dept Name
```