Big Data Frameworks CSE3120

Lab – 5 Map Reduce Join Experiment

Name: Naveen Nidadavolu
Roll No: 22MIA1049

Map Reduce Join

Aim: To perform a join operation using MapReduce on two datasets—students.txt and marks.txt—to combine student names with their respective marks based on Student_ID.

Algorithm

1. **Initialize MapReduce Job**

   Set job configuration, input paths, and output path.

   Specify mapper, reducer, and output key-value classes.

2. **StudentMapper (Map Phase)**

   For each line:

   a. Split by space.

   b. Emit (Student_ID, "STUDENT:Name").

3. **MarksMapper (Map Phase)**

   For each line:

   a. Split by space.

   b. Emit (Student_ID, "MARKS:Marks").

4. **Shuffle and Sort Phase**

   Group records by Student_ID.

5. **JoinReducer (Reduce Phase)**

   For each key:

   a. Initialize name and marks.

   b. For each value:

      i. If prefix is STUDENT:, assign to name.

      ii. If prefix is MARKS:, assign to marks.

   c. Emit (Student_ID, "Name Marks") if both present.

6. **Write Output**

   Output the joined records to the specified directory.

7. **Complete Job**

8. Wait for job completion and exit.

Procedure

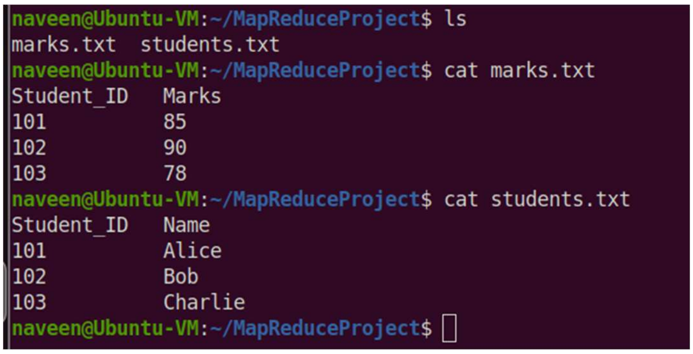
1. Create input text files (students.txt and marks.txt) and upload them to HDFS.
2. Run the MapReduce join program to process the input files.
3. The Mapper reads and tags records from both files based on Student_ID.
4. The Reducer merges records with the same Student_ID to produce the final joined output.
5. The output is stored in HDFS.
6. Retrieve and display the results.

Program

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class MapReduceJoin {

   public static class StudentMapper extends Mapper<Object, Text, Text, Text> {
      public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
         String[] tokens = value.toString().split("\\s+");
         if (tokens.length == 2 && !tokens[0].equals("Student_ID")) { // Ignore header
            context.write(new Text(tokens[0]), new Text("STUDENT:" + tokens[1]));
            }
        }
    }

    public static class MarksMapper extends Mapper<Object, Text, Text, Text> {
        public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
            String[] tokens = value.toString().split("\\s+");
            if (tokens.length == 2 && !tokens[0].equals("Student_ID")) { // Ignore header
                context.write(new Text(tokens[0]), new Text("MARKS:" + tokens[1]));
            }
        }
    }

    public static class JoinReducer extends Reducer<Text, Text, Text, Text> {
        public void reduce(Text key, Iterable<Text> values, Context context) throws
IOException, InterruptedException {
            String name = "";
            String marks = "";

            for (Text val : values) {
                String value = val.toString();
                if (value.startsWith("STUDENT:")) {
                    name = value.split(":")[1];
                } else if (value.startsWith("MARKS:")) {
                    marks = value.split(":")[1];
                }
            }

            if (!name.isEmpty() && !marks.isEmpty()) {
                context.write(key, new Text(name + "\t" + marks));
            }
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "MapReduce Join");

        job.setJarByClass(MapReduceJoin.class);
        job.setReducerClass(JoinReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);

        MultipleInputs.addInputPath(job, new Path(args[0]), TextInputFormat.class,
StudentMapper.class);
        MultipleInputs.addInputPath(job, new Path(args[1]), TextInputFormat.class,
MarksMapper.class);
        FileOutputFormat.setOutputPath(job, new Path(args[2]));

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

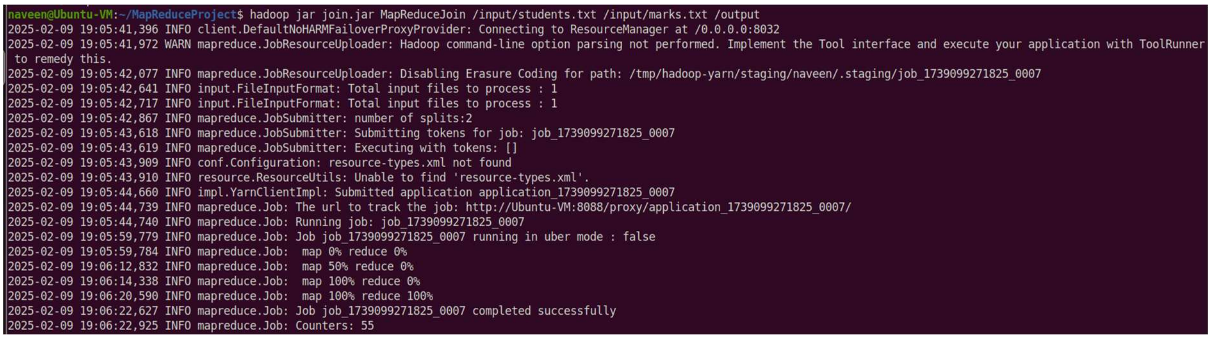
1.Input files

- File 1 (Students): Student_ID Name
- File 2 (Marks): Student_ID Marks

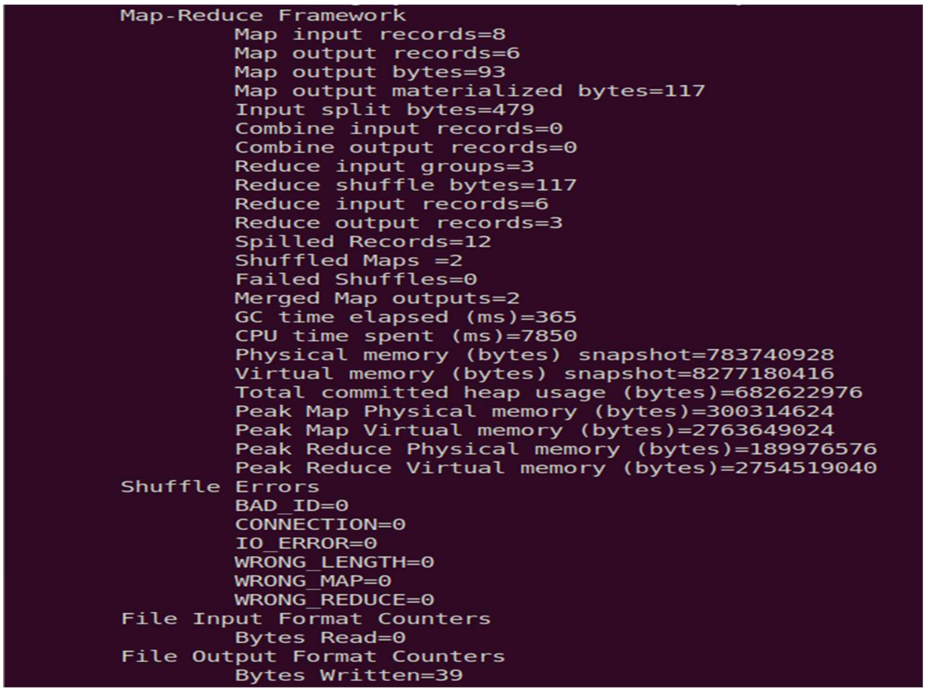


```
naveen@Ubuntu-VM:~/MapReduceProject$ ls
marks.txt  students.txt
naveen@Ubuntu-VM:~/MapReduceProject$ cat marks.txt
Student_ID    Marks
101           85
102           90
103           78
naveen@Ubuntu-VM:~/MapReduceProject$ cat students.txt
Student_ID    Name
101           Alice
102           Bob
103           Charlie
naveen@Ubuntu-VM:~/MapReduceProject$ 
```

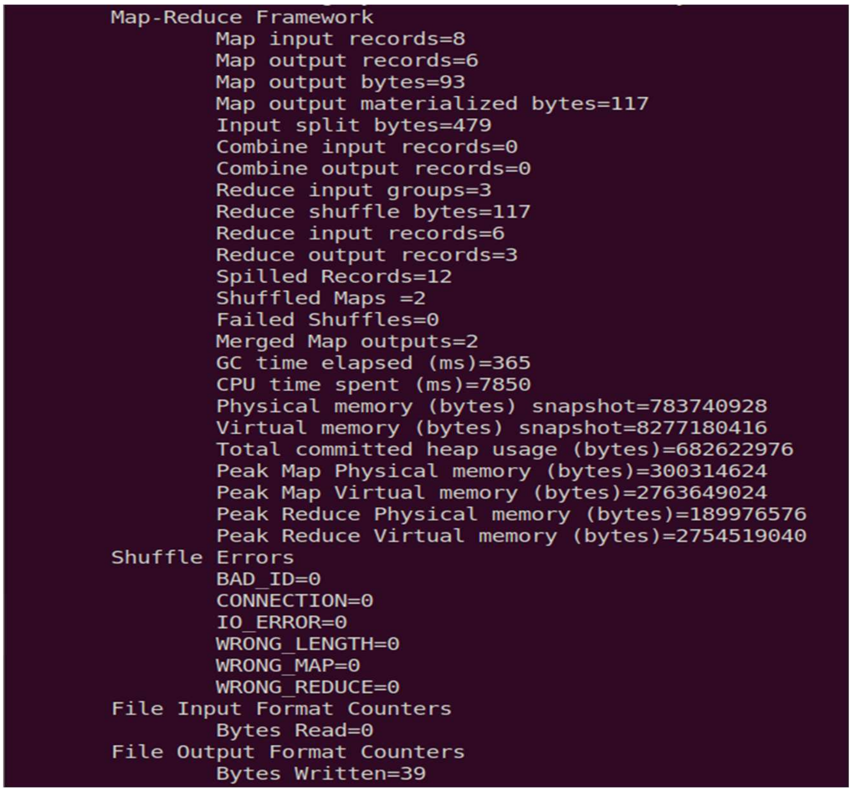
2.Running MapReduceJoin Program



```
naveen@Ubuntu-VM:~/MapReduceProject$ hadoop jar join.jar MapReduceJoin /input/students.txt /input/marks.txt /output
2025-02-09 19:05:41,396 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-02-09 19:05:41,972 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner
 to remedy this.
2025-02-09 19:05:42,077 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/naveen/.staging/job_1739099271825_0007
2025-02-09 19:05:42,641 INFO input.FileInputFormat: Total input files to process : 1
2025-02-09 19:05:42,717 INFO input.FileInputFormat: Total input files to process : 1
2025-02-09 19:05:42,867 INFO mapreduce.JobSubmitter: number of splits:2
2025-02-09 19:05:43,618 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1739099271825_0007
2025-02-09 19:05:43,619 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-02-09 19:05:43,909 INFO conf.Configuration: resource-types.xml not found
2025-02-09 19:05:43,910 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-02-09 19:05:44,660 INFO impl.YarnClientImpl: Submitted application application_1739099271825_0007
2025-02-09 19:05:44,739 INFO mapreduce.Job: The url to track the job: http://Ubuntu-VM:8088/proxy/application_1739099271825_0007/
2025-02-09 19:05:44,740 INFO mapreduce.Job: Running job: job_1739099271825_0007
2025-02-09 19:05:59,779 INFO mapreduce.Job: Job job_1739099271825_0007 running in uber mode : false
2025-02-09 19:05:59,784 INFO mapreduce.Job:  map 0% reduce 0%
2025-02-09 19:06:12,832 INFO mapreduce.Job:  map 50% reduce 0%
2025-02-09 19:06:14,338 INFO mapreduce.Job:  map 100% reduce 0%
2025-02-09 19:06:20,590 INFO mapreduce.Job:  map 100% reduce 100%
2025-02-09 19:06:22,627 INFO mapreduce.Job: Job job_1739099271825_0007 completed successfully
2025-02-09 19:06:22,925 INFO mapreduce.Job: Counters: 55
```

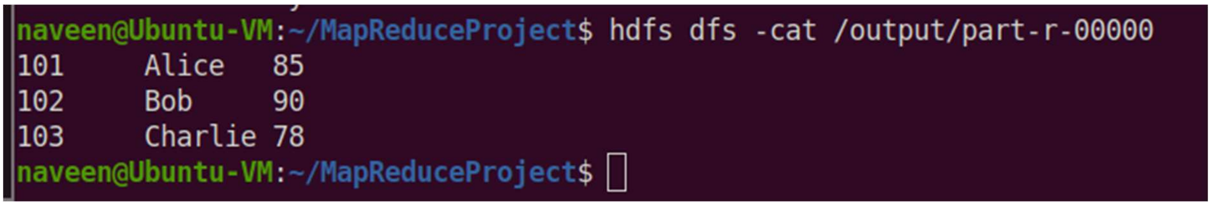


```
        Map-Reduce Framework
                Map input records=8
                Map output records=6
                Map output bytes=93
                Map output materialized bytes=117
                Input split bytes=479
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=117
                Reduce input records=6
                Reduce output records=3
                Spilled Records=12
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=365
                CPU time spent (ms)=7850
                Physical memory (bytes) snapshot=783740928
                Virtual memory (bytes) snapshot=8277180416
                Total committed heap usage (bytes)=682622976
                Peak Map Physical memory (bytes)=300314624
                Peak Map Virtual memory (bytes)=2763649024
                Peak Reduce Physical memory (bytes)=189976576
                Peak Reduce Virtual memory (bytes)=2754519040
        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=39
```



```
        Map-Reduce Framework
                Map input records=8
                Map output records=6
                Map output bytes=93
                Map output materialized bytes=117
                Input split bytes=479
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=117
                Reduce input records=6
                Reduce output records=3
                Spilled Records=12
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=365
                CPU time spent (ms)=7850
                Physical memory (bytes) snapshot=783740928
                Virtual memory (bytes) snapshot=8277180416
                Total committed heap usage (bytes)=682622976
                Peak Map Physical memory (bytes)=300314624
                Peak Map Virtual memory (bytes)=2763649024
                Peak Reduce Physical memory (bytes)=189976576
                Peak Reduce Virtual memory (bytes)=2754519040
        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=39
```

Output:



```
naveen@Ubuntu-VM:~/MapReduceProject$ hdfs dfs -cat /output/part-r-00000
101     Alice   85
102     Bob     90
103     Charlie 78
naveen@Ubuntu-VM:~/MapReduceProject$ 
```

Result: Successfully performed a join operation using MapReduce, merging student details with their marks based on Student_ID.