Big Data Frameworks CSE3120
Lab – 6 Reduce Side Join Experiment

Name: Naveen Nidadavolu
Roll No: 22MIA1049

Reduce Side Join

Aim: To implement **Reduce-Side Join** in Hadoop using MapReduce, where two datasets (**employees and salaries**) are joined based on a common key (**EmployeeID**) to produce a final output containing **employee names and their corresponding salaries**.

Algorithm:

Mapper Phase

1. Read each record from the input files (**employees.csv** and **salaries.csv**).
2. Extract **EmployeeID** as the key.
3. Tag the record:
    - If from **employees.csv**, tag as `"A,EmployeeName"`.
    - If from **salaries.csv**, tag as `"B,Salary"`.
4. Emit (**EmployeeID, TaggedRecord**).

Shuffle & Sort Phase (Handled by Hadoop)

5. Hadoop groups all records by **EmployeeID** before sending them to the Reducer.

Reducer Phase

6. Initialize **employeeName = null** and **salary = null**.
7. Loop through grouped values:
    - If tagged **"A"**, extract **EmployeeName**.
    - If tagged **"B"**, extract **Salary**.
8. If **both EmployeeName and Salary exist**, emit (**EmployeeID, EmployeeName, Salary**).

Store Output in HDFS

9. Write the final **joined result** to the HDFS output directory.

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.output.FileOutputFormat;

/**
 * Reduce Side Join Example for Employee and Salary Data
 */
public class rsp {

  // Mapper Class
  public static class JoinMapper extends Mapper<Object, Text, Text, Text> {
    public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
      String[] fields = value.toString().split(",");

      // Ensure valid record
      if (fields.length >= 3) {
        String recordType = fields[0].trim(); // "A" for Employee, "B" for Salary
        String joinKey = fields[1].trim();    // Employee ID (Join Key)
        String details = fields[2].trim();    // Employee Name or Salary

        context.write(new Text(joinKey), new Text(recordType + "," + details));
      }
    }
  }

  // Reducer Class
```

```
  public static class JoinReducer extends Reducer<Text, Text, Text, Text> {
    public void reduce(Text key, Iterable<Text> values, Context context) throws IOException,
InterruptedException {
      String employeeName = null;
      String salary = null;

      // Iterate over values
      for (Text val : values) {
        String[] tokens = val.toString().split(",");
        if (tokens.length == 2) {
          if (tokens[0].equals("A")) {
            employeeName = tokens[1];  // Employee Name
          } else if (tokens[0].equals("B")) {
            salary = tokens[1];  // Employee Salary
          }
        }
      }

      // Output only if both values exist
      if (employeeName != null && salary != null) {
        context.write(key, new Text(employeeName + ", " + salary));
      }
    }
  }

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

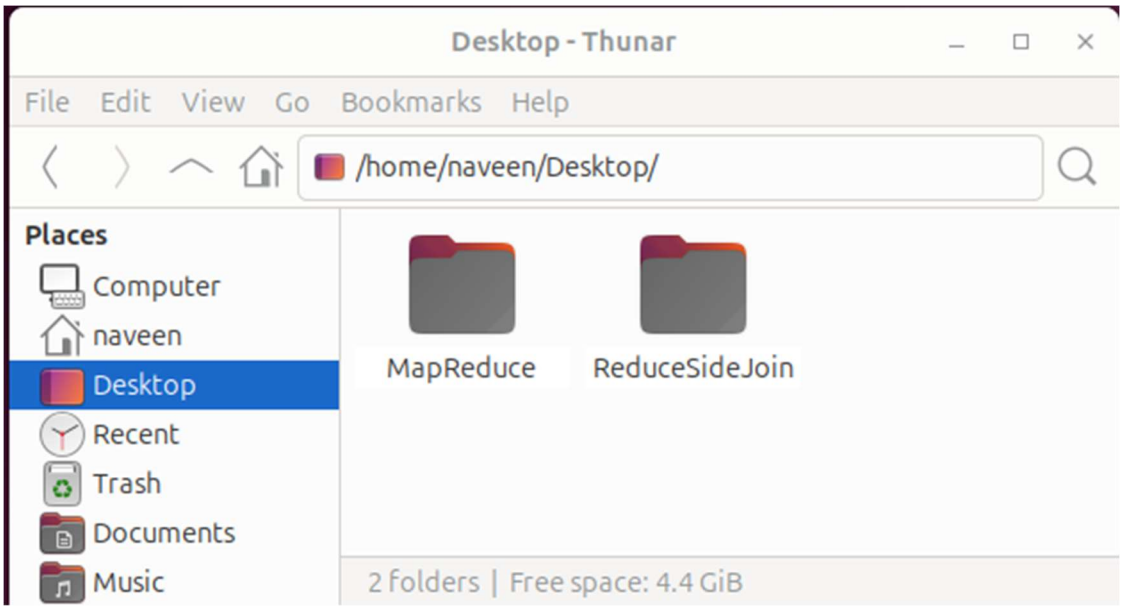
```
        Configuration conf = new Configuration();
        Job job = new Job(conf);
        job.setJarByClass(rsp.class);
        job.setMapperClass(JoinMapper.class);
        job.setReducerClass(JoinReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);

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

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

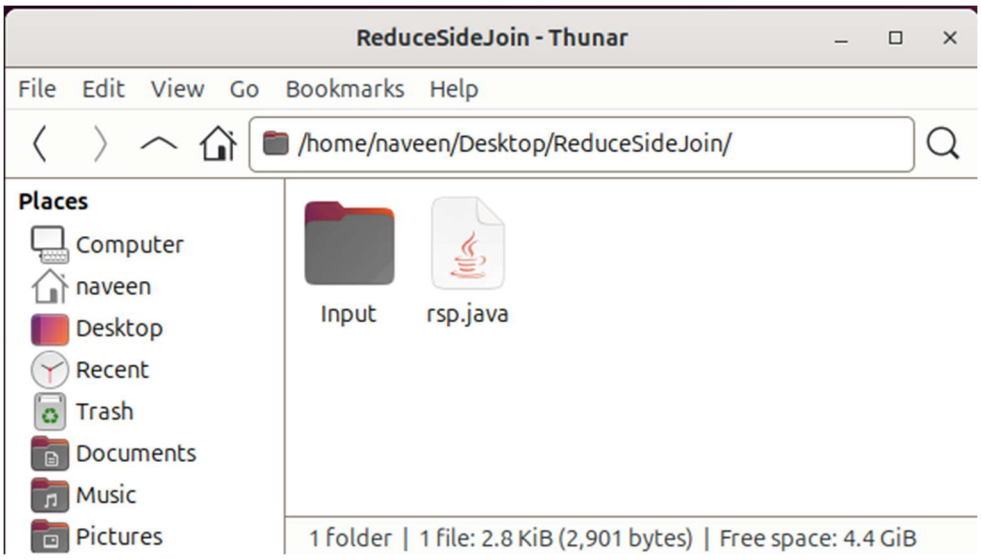
Input files

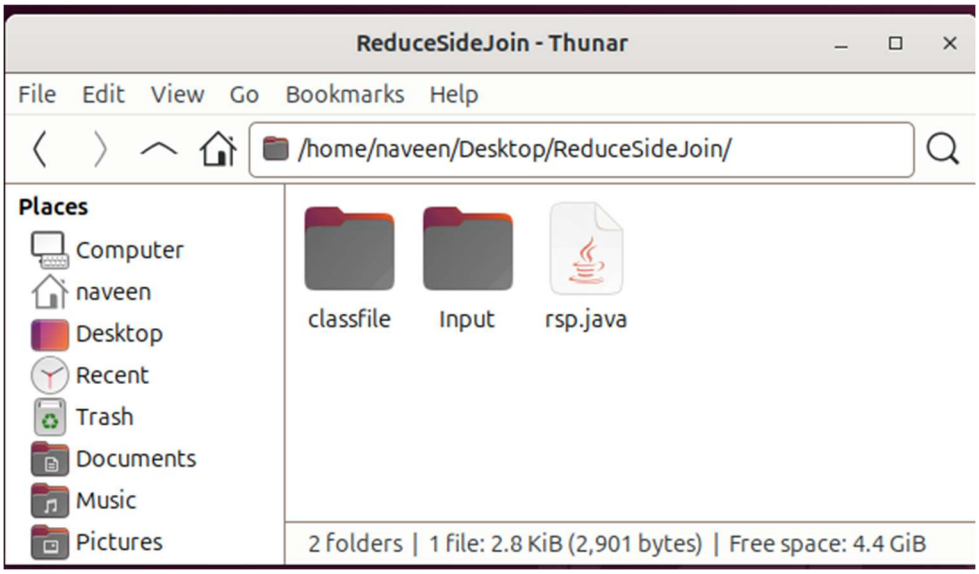
- File 1: employees.csv
- File 2: salaries.csv

Procedure

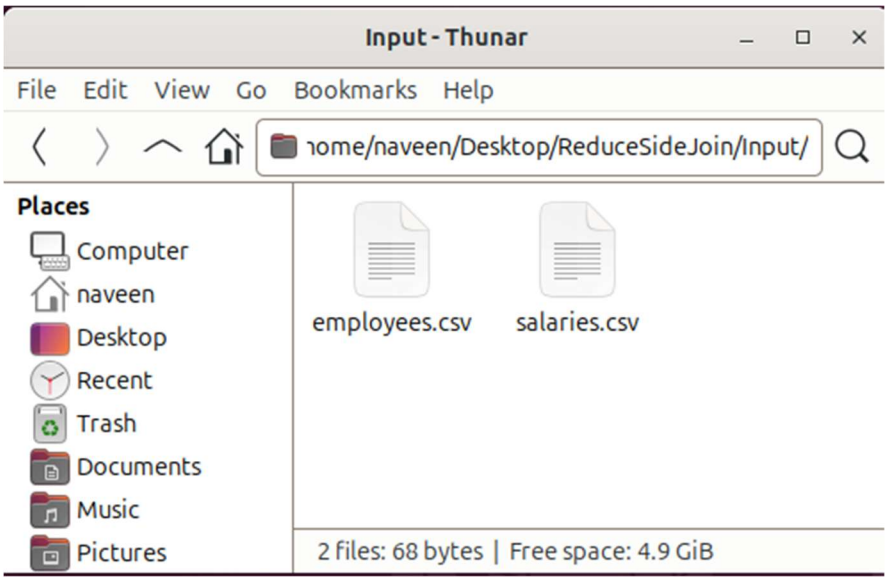
1. Create a folder ReduceSideJoin. Inside the folder create the Input folder and place the input text files and also create an empty folder classfile.



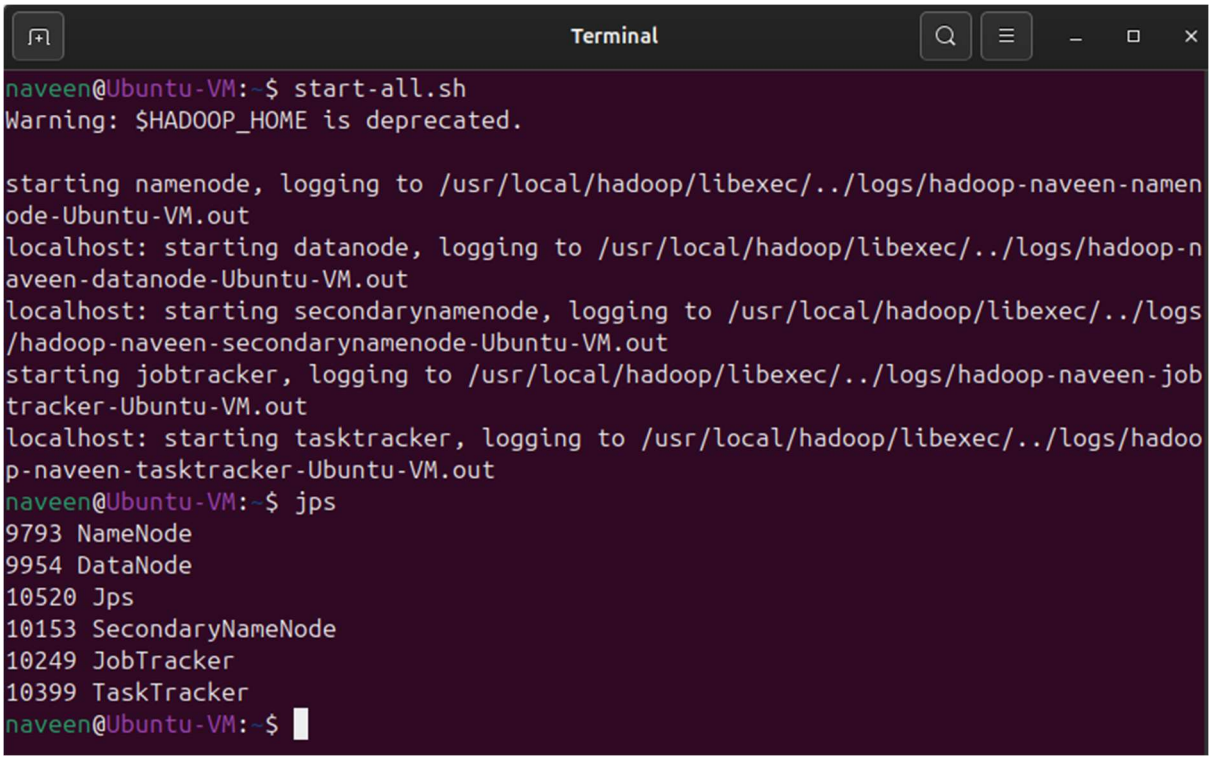




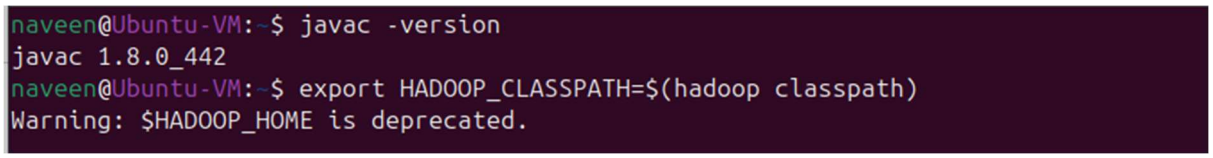
2. Input datsets in Input folder



3. Start the Hadoop services and basic operations



4. Store the Hadoop class path in a HADOOP_PATH variable



```
naveen@Ubuntu-VM:~$ echo $HADOOP_CLASSPATH
/usr/local/hadoop/libexec/../conf:/usr/lib/jvm/java-8-openjdk-amd64/lib/tools.jar:/
usr/local/hadoop/libexec/..:/usr/local/hadoop/libexec/../hadoop-core-1.2.1.jar:/usr
/local/hadoop/libexec/../lib/asm-3.2.jar:/usr/local/hadoop/libexec/../lib/aspectjrt
-1.6.11.jar:/usr/local/hadoop/libexec/../lib/aspectjtools-1.6.11.jar:/usr/local/had
oop/libexec/../lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop/libexec/../lib/com
mons-beanutils-core-1.8.0.jar:/usr/local/hadoop/libexec/../lib/commons-cli-1.2.jar:
/usr/local/hadoop/libexec/../lib/commons-codec-1.4.jar:/usr/local/hadoop/libexec/..
/lib/commons-collections-3.2.1.jar:/usr/local/hadoop/libexec/../lib/commons-configu
ration-1.6.jar:/usr/local/hadoop/libexec/../lib/commons-daemon-1.0.1.jar:/usr/local
/hadoop/libexec/../lib/commons-digester-1.8.jar:/usr/local/hadoop/libexec/../lib/co
mmons-el-1.0.jar:/usr/local/hadoop/libexec/../lib/commons-httpclient-3.0.1.jar:/usr
/local/hadoop/libexec/../lib/commons-io-2.1.jar:/usr/local/hadoop/libexec/../lib/co
mmons-lang-2.4.jar:/usr/local/hadoop/libexec/../lib/commons-logging-1.1.1.jar:/usr/
local/hadoop/libexec/../lib/commons-logging-api-1.0.4.jar:/usr/local/hadoop/libexec
/../lib/commons-math-2.1.jar:/usr/local/hadoop/libexec/../lib/commons-net-3.1.jar:/
usr/local/hadoop/libexec/../lib/core-3.1.1.jar:/usr/local/hadoop/libexec/../lib/had
oop-capacity-scheduler-1.2.1.jar:/usr/local/hadoop/libexec/../lib/hadoop-fairschedu
ler-1.2.1.jar:/usr/local/hadoop/libexec/../lib/hadoop-thriftfs-1.2.1.jar:/usr/local
/hadoop/libexec/../lib/hsqldb-1.8.0.10.jar:/usr/local/hadoop/libexec/../lib/jackson
-core-asl-1.8.8.jar:/usr/local/hadoop/libexec/../lib/jackson-mapper-asl-1.8.8.jar:/
usr/local/hadoop/libexec/../lib/jasper-compiler-5.5.12.jar:/usr/local/hadoop/libexe
c/../lib/jasper-runtime-5.5.12.jar:/usr/local/hadoop/libexec/../lib/jdeb-0.8.jar:/u
sr/local/hadoop/libexec/../lib/jersey-core-1.8.jar:/usr/local/hadoop/libexec/../lib
/jersey-json-1.8.jar:/usr/local/hadoop/libexec/../lib/jersey-server-1.8.jar:/usr/lo
cal/hadoop/libexec/../lib/jets3t-0.6.1.jar:/usr/local/hadoop/libexec/../lib/jetty-6
.1.26.jar:/usr/local/hadoop/libexec/../lib/jetty-util-6.1.26.jar:/usr/local/hadoop/
libexec/../lib/jsch-0.1.42.jar:/usr/local/hadoop/libexec/../lib/junit-4.5.jar:/usr/
local/hadoop/libexec/../lib/kfs-0.2.2.jar:/usr/local/hadoop/libexec/../lib/log4j-1.
2.15.jar:/usr/local/hadoop/libexec/../lib/mockito-all-1.8.5.jar:/usr/local/hadoop/l
ibexec/../lib/oro-2.0.8.jar:/usr/local/hadoop/libexec/../lib/servlet-api-2.5-200812
11.jar:/usr/local/hadoop/libexec/../lib/slf4j-api-1.4.3.jar:/usr/local/hadoop/libex
ec/../lib/slf4j-log4j12-1.4.3.jar:/usr/local/hadoop/libexec/../lib/xmlenc-0.52.jar:
/usr/local/hadoop/libexec/../lib/jsp-2.1/jsp-2.1.jar:/usr/local/hadoop/libexec/../l
ib/jsp-2.1/jsp-api-2.1.jar
naveen@Ubuntu-VM:~$
```

5. Create directory reducesidejoin/Input to store input text files in Hadoop.

```
naveen@Ubuntu-VM:~$ hadoop fs -mkdir /reducesidejoin/Input
Warning: $HADOOP_HOME is deprecated.

naveen@Ubuntu-VM:~$ hadoop fs -ls /
Warning: $HADOOP_HOME is deprecated.

Found 3 items
drwxr-xr-x   - naveen supergroup          0 2025-03-11 23:57 /mapsidejoin
drwxr-xr-x   - naveen supergroup          0 2025-03-22 23:07 /reducesidejoin
drwxr-xr-x   - naveen supergroup          0 2025-03-11 23:57 /tmp
naveen@Ubuntu-VM:~$
```

6. Place the input text files in the created directory.

```
naveen@Ubuntu-VM:~$ hadoop fs -put '/home/naveen/Desktop/ReduceSideJoin/Input/emplo
yees.csv' /reducesidejoin/Input
Warning: $HADOOP_HOME is deprecated.

naveen@Ubuntu-VM:~$ hadoop fs -put '/home/naveen/Desktop/ReduceSideJoin/Input/salar
ies.csv' /reducesidejoin/Input
Warning: $HADOOP_HOME is deprecated.

naveen@Ubuntu-VM:~$ hadoop fs -ls /reducesidejoin/Input
Warning: $HADOOP_HOME is deprecated.

Found 2 items
-rw-r--r--   1 naveen supergroup         34 2025-03-22 23:55 /reducesidejoin/Input/
employees.csv
-rw-r--r--   1 naveen supergroup         34 2025-03-22 23:56 /reducesidejoin/Input/
salaries.csv
naveen@Ubuntu-VM:~$
```

```
naveen@Ubuntu-VM:~/Desktop/ReduceSide/Input$ hadoop fs -cat /reducesidejoin/Input/*
Warning: $HADOOP_HOME is deprecated.

A,101,John
A,102,Alice
A,103,Bob

B,101,5000
B,102,6000
B,104,7000

naveen@Ubuntu-VM:~/Desktop/ReduceSide/Input$
```

7. Check the files by going to HDFS NameNode Web UI using the port 50070.

Contents of directory /

Goto : / go

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
mapsidejoin	dir				2025-03-11 23:57	rwxr-xr-x	naveen	supergroup
reducesidejoin	dir				2025-03-22 23:12	rwxr-xr-x	naveen	supergroup
tmp	dir				2025-03-11 23:57	rwxr-xr-x	naveen	supergroup

Go back to DFS home

Contents of directory /reducesidejoin/Input

Goto : /reducesidejoin/Input go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
employees.csv	file	0.03 KB	1	64 MB	2025-03-22 23:55	rw-r--r--	naveen	supergroup
salaries.csv	file	0.03 KB	1	64 MB	2025-03-22 23:56	rw-r--r--	naveen	supergroup

Go back to DFS home

File: /reducesidejoin/Input/employees.csv

Goto : /reducesidejoin/Input go

Go back to dir listing
Advanced view/download options

```
A,101,John
A,102,Alice
A,103,Bob
```

File: /reducesidejoin/Input/salaries.csv

Goto : /reducesidejoin/Input go

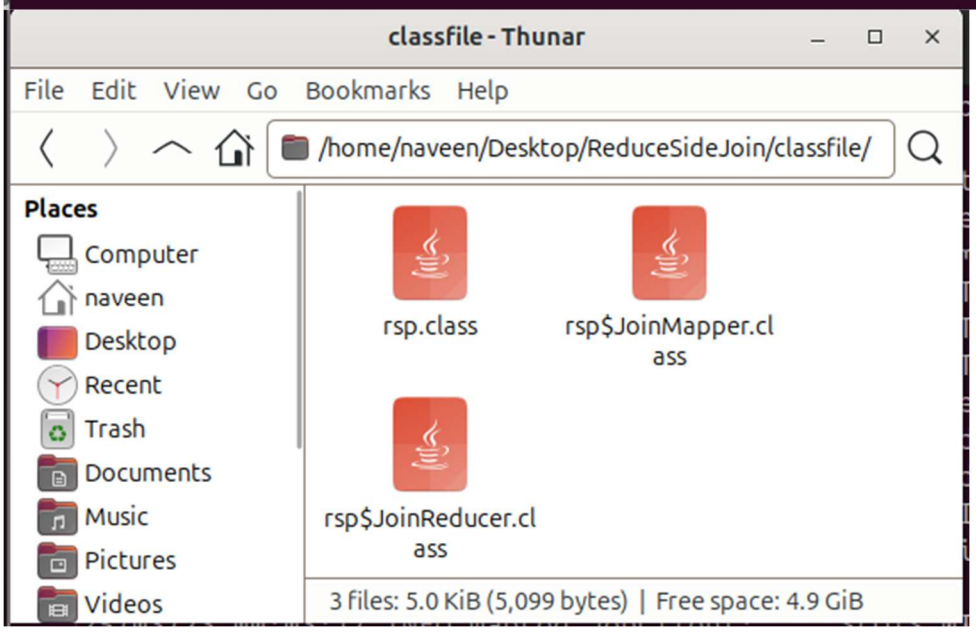
Go back to dir listing
Advanced view/download options

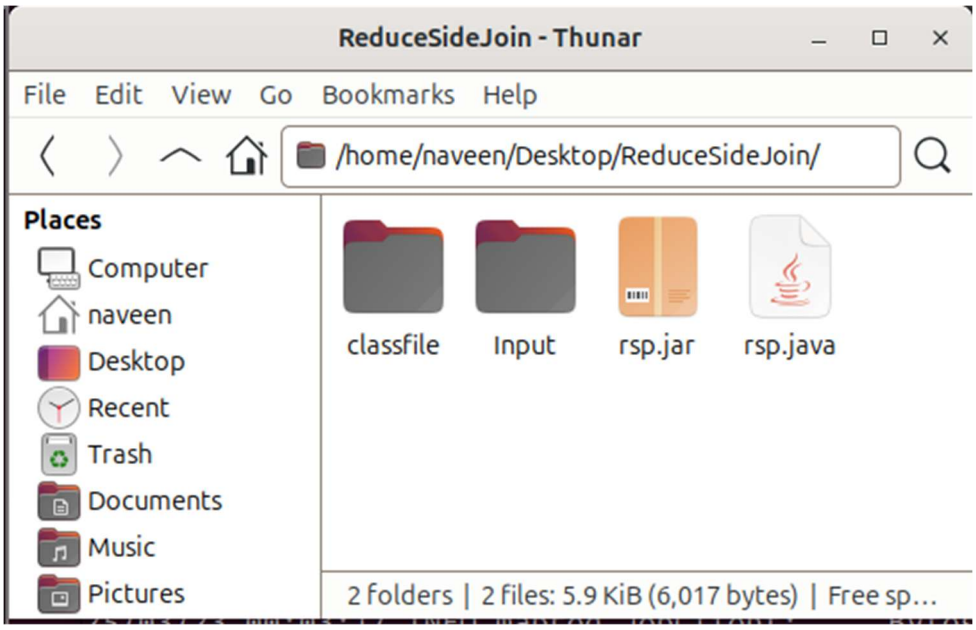
```
B,101,5000
B,102,6000
B,104,7000
```

8. Compile the reduce side join program rsp.java and store it in the classfile and create jar file rsp.jar

```
naveen@Ubuntu-VM:~$ cd Desktop
naveen@Ubuntu-VM:~/Desktop$ cd ReduceSideJoin
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$
```

```
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$ javac -classpath ${HADOOP_CLASSPATH} -d
'/home/naveen/Desktop/ReduceSideJoin/classfile' '/home/naveen/Desktop/ReduceSideJoi
n/rsp.java'
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$ jar -cvf rsp.jar -C '/home/naveen/Deskto
p/ReduceSideJoin/classfile'/ .
added manifest
adding: rsp$JoinMapper.class(in = 1625) (out= 687)(deflated 57%)
adding: rsp$JoinReducer.class(in = 2006) (out= 891)(deflated 55%)
adding: rsp.class(in = 1468) (out= 834)(deflated 43%)
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$
```





9. Run the Hadoop job and store the output

```
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$ hadoop jar '/home/naveen/Desktop/ReduceS
ideJoin/rsp.jar' rsp /reducesidejoin/Input /reducesidejoin/Output
Warning: $HADOOP_HOME is deprecated.

25/03/22 23:26:18 WARN mapred.JobClient: Use GenericOptionsParser for parsing the a
rguments. Applications should implement Tool for the same.
25/03/22 23:26:18 INFO input.FileInputFormat: Total input paths to process : 2
25/03/22 23:26:18 INFO util.NativeCodeLoader: Loaded the native-hadoop library
25/03/22 23:26:18 WARN snappy.LoadSnappy: Snappy native library not loaded
25/03/22 23:26:19 INFO mapred.JobClient: Running job: job_202503222246_0001
25/03/22 23:26:20 INFO mapred.JobClient:  map 0% reduce 0%
```

```
25/03/22 23:26:47 INFO mapred.JobClient:  map 100% reduce 100%
25/03/22 23:26:51 INFO mapred.JobClient: Job complete: job_202503222246_0001
25/03/22 23:26:51 INFO mapred.JobClient: Counters: 28
25/03/22 23:26:51 INFO mapred.JobClient:   Map-Reduce Framework
25/03/22 23:26:51 INFO mapred.JobClient:     Spilled Records=0
25/03/22 23:26:51 INFO mapred.JobClient:     Map output materialized bytes=12
25/03/22 23:26:51 INFO mapred.JobClient:     Reduce input records=0
25/03/22 23:26:51 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=550265
6512
25/03/22 23:26:51 INFO mapred.JobClient:     Map input records=8
25/03/22 23:26:51 INFO mapred.JobClient:     SPLIT_RAW_BYTES=241
25/03/22 23:26:51 INFO mapred.JobClient:     Map output bytes=0
25/03/22 23:26:51 INFO mapred.JobClient:     Reduce shuffle bytes=12
25/03/22 23:26:51 INFO mapred.JobClient:     Physical memory (bytes) snapshot=50419
3024
25/03/22 23:26:51 INFO mapred.JobClient:     Reduce input groups=0
25/03/22 23:26:51 INFO mapred.JobClient:     Combine output records=0
25/03/22 23:26:51 INFO mapred.JobClient:     Reduce output records=0
25/03/22 23:26:51 INFO mapred.JobClient:     Map output records=0
25/03/22 23:26:51 INFO mapred.JobClient:     Combine input records=0
25/03/22 23:26:51 INFO mapred.JobClient:     CPU time spent (ms)=15460
25/03/22 23:26:51 INFO mapred.JobClient:     Total committed heap usage (bytes)=440
926208
25/03/22 23:26:51 INFO mapred.JobClient:   File Input Format Counters
25/03/22 23:26:51 INFO mapred.JobClient:     Bytes Read=68
25/03/22 23:26:51 INFO mapred.JobClient:   FileSystemCounters
25/03/22 23:26:51 INFO mapred.JobClient:     HDFS_BYTES_READ=309
25/03/22 23:26:51 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=165867
25/03/22 23:26:51 INFO mapred.JobClient:     FILE_BYTES_READ=6
25/03/22 23:26:51 INFO mapred.JobClient:   Job Counters
```

```
25/03/22 23:26:51 INFO mapred.JobClient:   File Input Format Counters
25/03/22 23:26:51 INFO mapred.JobClient:     Bytes Read=68
25/03/22 23:26:51 INFO mapred.JobClient:   FileSystemCounters
25/03/22 23:26:51 INFO mapred.JobClient:     HDFS_BYTES_READ=309
25/03/22 23:26:51 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=165867
25/03/22 23:26:51 INFO mapred.JobClient:     FILE_BYTES_READ=6
25/03/22 23:26:51 INFO mapred.JobClient:   Job Counters
25/03/22 23:26:51 INFO mapred.JobClient:     Launched map tasks=2
25/03/22 23:26:51 INFO mapred.JobClient:     Launched reduce tasks=1
25/03/22 23:26:51 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=14002
25/03/22 23:26:51 INFO mapred.JobClient:     Total time spent by all reduces waitin
g after reserving slots (ms)=0
25/03/22 23:26:51 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=24943
25/03/22 23:26:51 INFO mapred.JobClient:     Total time spent by all maps waiting a
fter reserving slots (ms)=0
25/03/22 23:26:51 INFO mapred.JobClient:     Data-local map tasks=2
25/03/22 23:26:51 INFO mapred.JobClient:   File Output Format Counters
25/03/22 23:26:51 INFO mapred.JobClient:     Bytes Written=0
naveen@Ubuntu-VM:~/Desktop/ReduceSideJoin$
```

10. Check the Output files by going to HDFS NameNode Web UI using the port 50070.

Contents of directory /reducesidejoin

Goto : /reducesidejoin go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
Input	dir				2025-03-22 23:13	rwxr-xr-x	naveen	supergroup
Output	dir				2025-03-22 23:26	rwxr-xr-x	naveen	supergroup

Go back to DFS home

Contents of directory /reducesidejoin/Output

Goto : /reducesidejoin/Output go

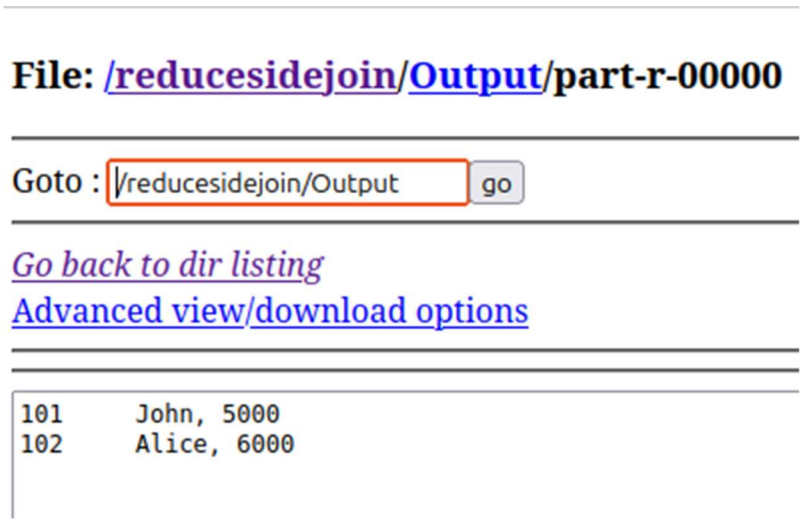
Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	1	64 MB	2025-03-22 23:26	rw-r--r--	naveen	supergroup
_logs	dir				2025-03-22 23:26	rwxr-xr-x	naveen	supergroup
part-r-00000	file	0 KB	1	64 MB	2025-03-22 23:26	rw-r--r--	naveen	supergroup

Go back to DFS home

Output:

```
naveen@Ubuntu-VM:~/Desktop/ReduceSide$ hadoop fs -cat /reducesidejoin/Output/part-*
Warning: $HADOOP_HOME is deprecated.

101     John, 5000
102     Alice, 6000
naveen@Ubuntu-VM:~/Desktop/ReduceSide$
```



Result:

Successfully performed a **Reduce-Side Join** operation using **MapReduce**, merging employee details with their salaries based on **EmployeeID**.