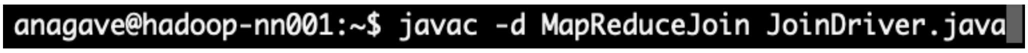
To Join Two Datasets:

Step 1:

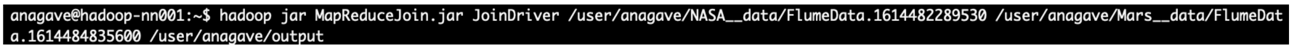
To compile the the code



```
anagave@hadoop-nn001:~$ javac -d MapReduceJoin JoinDriver.java
```

Step 2:

To Execute the code



```
anagave@hadoop-nn001:~$ hadoop jar MapReduceJoin.jar JoinDriver /user/anagave/NASA__data/FlumeData.1614482289530 /user/anagave/Mars__data/FlumeDat
a.1614484835600 /user/anagave/output
```

```
2021-03-12 20:25:25,120 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
2021-03-12 20:25:26,259 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2021-03-12 20:25:26,508 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException:
Output directory hdfs://hadoop-nn001.cs.okstate.edu:9000/user/anagave/
LineCount_output_part4 already exists
	at
org.apache.hadoop.mapred.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:
131)
	at
org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:279)
	at
org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
	at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1576)
	at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1573)
	at java.security.AccessController.doPrivileged(Native Method)
	at javax.security.auth.Subject.doAs(Subject.java:422)
	at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1845
)
	at org.apache.hadoop.mapreduce.Job.submit(Job.java:1573)
	at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:576)
	at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:571)
	at java.security.AccessController.doPrivileged(Native Method)
	at javax.security.auth.Subject.doAs(Subject.java:422)
	at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1845
)
	at org.apache.hadoop.mapred.JobClient.submitJobInternal(JobClient.java:571)
	at org.apache.hadoop.mapred.JobClient.submitJob(JobClient.java:562)
	at org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:873)
	at LineCount.main(LineCount.java:60)
	at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
```

```
        at
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:4
3)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
anagave@hadoop-nn001:~$ hadoop jar lc.jar LineCount /user/anagave/Mars__data/
2021/02/27/20/FlumeData.* /user/anagave/LineCount_output_part
2021-03-12 20:25:38,382 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
2021-03-12 20:25:39,281 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2021-03-12 20:25:39,478 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at hadoop-nn001.cs.okstate.edu/192.168.122.2:8032
2021-03-12 20:25:39,786 WARN mapreduce.JobResourceUploader: Hadoop command-
line option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
2021-03-12 20:25:39,803 INFO mapreduce.JobResourceUploader: Disabling Erasure
Coding for path: /tmp/hadoop-yarn/staging/anagave/.staging/job_1615570959549_0114
2021-03-12 20:25:40,186 INFO mapred.FileInputFormat: Total input files to process : 49
2021-03-12 20:25:40,503 INFO mapreduce.JobSubmitter: number of splits:49
2021-03-12 20:25:40,736 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1615570959549_0114
2021-03-12 20:25:40,736 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-12 20:25:41,041 INFO conf.Configuration: resource-types.xml not found
2021-03-12 20:25:41,041 INFO resource.ResourceUtils: Unable to find 'resource-
types.xml'.
2021-03-12 20:25:41,186 INFO impl.YarnClientImpl: Submitted application
application_1615570959549_0114
2021-03-12 20:25:41,299 INFO mapreduce.Job: The url to track the job: http://hadoop-
nn001.cs.okstate.edu:8088/proxy/application_1615570959549_0114/
2021-03-12 20:25:41,303 INFO mapreduce.Job: Running job: job_1615570959549_0114
2021-03-12 20:25:46,416 INFO mapreduce.Job: Job job_1615570959549_0114 running
in uber mode : false
2021-03-12 20:25:46,419 INFO mapreduce.Job:  map 0% reduce 0%
2021-03-12 20:25:51,555 INFO mapreduce.Job:  map 33% reduce 0%
2021-03-12 20:25:55,603 INFO mapreduce.Job:  map 59% reduce 0%
2021-03-12 20:25:56,615 INFO mapreduce.Job:  map 65% reduce 0%
2021-03-12 20:25:59,650 INFO mapreduce.Job:  map 86% reduce 0%
2021-03-12 20:26:00,660 INFO mapreduce.Job:  map 88% reduce 0%
2021-03-12 20:26:02,682 INFO mapreduce.Job:  map 100% reduce 0%
2021-03-12 20:26:03,704 INFO mapreduce.Job:  map 100% reduce 100%
2021-03-12 20:26:03,720 INFO mapreduce.Job: Job job_1615570959549_0114
completed successfully
2021-03-12 20:26:03,872 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=894
                FILE: Number of bytes written=14305870
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=9449983
        HDFS: Number of bytes written=17
        HDFS: Number of read operations=172
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=10
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=49
        Launched reduce tasks=5
        Other local map tasks=1
        Data-local map tasks=46
        Rack-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=718275
        Total time spent by all reduces in occupied slots (ms)=175190
        Total time spent by all map tasks (ms)=143655
        Total time spent by all reduce tasks (ms)=35038
        Total vcore-milliseconds taken by all map tasks=143655
        Total vcore-milliseconds taken by all reduce tasks=35038
        Total megabyte-milliseconds taken by all map tasks=735513600
        Total megabyte-milliseconds taken by all reduce tasks=179394560
    Map-Reduce Framework
        Map input records=1729
        Map output records=1729
        Map output bytes=27664
        Map output materialized bytes=2334
        Input split bytes=7501
        Combine input records=1729
        Combine output records=48
        Reduce input groups=1
        Reduce shuffle bytes=2334
        Reduce input records=48
        Reduce output records=1
        Spilled Records=96
        Shuffled Maps =245
        Failed Shuffles=0
        Merged Map outputs=245
        GC time elapsed (ms)=1137
        CPU time spent (ms)=32880
        Physical memory (bytes) snapshot=18634072064
        Virtual memory (bytes) snapshot=344526479360
        Total committed heap usage (bytes)=42599448576
        Peak Map Physical memory (bytes)=366059520
        Peak Map Virtual memory (bytes)=6386380800
        Peak Reduce Physical memory (bytes)=272928768
        Peak Reduce Virtual memory (bytes)=6400557056
    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=9442482
	File Output Format Counters
		Bytes Written=17
```

Step 3:

To view the contents in output file

```
anagave@hadoop-nn001:~$ hdfs dfs -cat /user/anagave/output/part*
2021-03-12 20:51:30,307 WARN util.NativeCodeLoader: Unable to load
native-hadoop library for your platform... using builtin-java
classes where applicable
```

- Since there are no common fields so it doesn’t display anything.