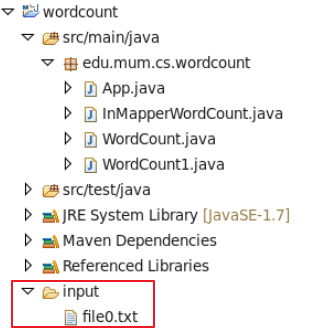
1. Install VMWare or VirtualBox
2. Download Cloudera QuickStarts from the following link
   1. <https://www.cloudera.com/downloads.html>
   2. Select a corresponding Platform
   3. Give the virtual machine 2 CPU core and 8G memory and Start the virtual machine
   4. Launch Cloudera Express on desktop, then Hadoop will start.
3. Find the Example: WordCount v1.0
   1. <https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0>
   2. <https://cwiki.apache.org/confluence/display/HADOOP2/WordCount>
4. Create a java project using eclipse inside the virtual machine
5. Copy Example: WordCount v1.0 into the project
6. We will find many errors in the project, to solve these errors, add external jars into the project from the following locations in the virtual machine.
   1. File system/usr/lib/hadoop/client-0.20
   2. File system/usr/lib/hadoop
   3. File system/usr/lib/hadoop/lib
7. Add input folder and test file into the project class path
   1. 
   2. The content of file0.txt: Hadoop is an elephant Hadoop is an elephant Hadoop is an elephant
8. We can run or debug the project in Hadoop local mode it’s convenient to develop a map reduce program. Add program arguments “input output”

The output:

19/11/10 21:35:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

19/11/10 21:35:35 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id

19/11/10 21:35:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=

19/11/10 21:35:35 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.

19/11/10 21:35:35 WARN mapred.JobClient: No job jar file set. User classes may not be found. See JobConf(Class) or JobConf#setJar(String).

19/11/10 21:35:35 INFO input.FileInputFormat: Total input paths to process : 1

19/11/10 21:35:36 INFO mapred.LocalJobRunner: OutputCommitter set in config null

19/11/10 21:35:36 INFO mapred.JobClient: Running job: job\_local1989419825\_0001

19/11/10 21:35:36 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter

19/11/10 21:35:36 INFO mapred.LocalJobRunner: Waiting for map tasks

19/11/10 21:35:36 INFO mapred.LocalJobRunner: Starting task: attempt\_local1989419825\_0001\_m\_000000\_0

19/11/10 21:35:36 WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counter is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead

19/11/10 21:35:36 INFO util.ProcessTree: setsid exited with exit code 0

19/11/10 21:35:36 INFO mapred.Task: Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin@297d1d8b

19/11/10 21:35:36 INFO mapred.MapTask: Processing split: file:/home/cloudera/workspace/wordcount/input/file0.txt:0+65

19/11/10 21:35:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer

19/11/10 21:35:36 INFO mapred.MapTask: io.sort.mb = 100

19/11/10 21:35:36 INFO mapred.MapTask: data buffer = 79691776/99614720

19/11/10 21:35:36 INFO mapred.MapTask: record buffer = 262144/327680

19/11/10 21:35:36 INFO mapred.LocalJobRunner:

19/11/10 21:35:36 INFO mapred.MapTask: Starting flush of map output

19/11/10 21:35:36 INFO mapred.MapTask: Finished spill 0

19/11/10 21:35:36 INFO mapred.Task: Task:attempt\_local1989419825\_0001\_m\_000000\_0 is done. And is in the process of commiting

19/11/10 21:35:36 INFO mapred.LocalJobRunner:

19/11/10 21:35:36 INFO mapred.Task: Task 'attempt\_local1989419825\_0001\_m\_000000\_0' done.

19/11/10 21:35:36 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1989419825\_0001\_m\_000000\_0

19/11/10 21:35:36 INFO mapred.LocalJobRunner: Map task executor complete.

19/11/10 21:35:36 WARN mapreduce.Counters: Group org.apache.hadoop.mapred.Task$Counter is deprecated. Use org.apache.hadoop.mapreduce.TaskCounter instead

19/11/10 21:35:36 INFO mapred.Task: Using ResourceCalculatorPlugin : org.apache.hadoop.util.LinuxResourceCalculatorPlugin@1b7d8bfa

19/11/10 21:35:36 INFO mapred.LocalJobRunner:

19/11/10 21:35:36 INFO mapred.Merger: Merging 1 sorted segments

19/11/10 21:35:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 140 bytes

19/11/10 21:35:36 INFO mapred.LocalJobRunner:

19/11/10 21:35:36 INFO mapred.Task: Task:attempt\_local1989419825\_0001\_r\_000000\_0 is done. And is in the process of commiting

19/11/10 21:35:36 INFO mapred.LocalJobRunner:

19/11/10 21:35:36 INFO mapred.Task: Task attempt\_local1989419825\_0001\_r\_000000\_0 is allowed to commit now

19/11/10 21:35:36 INFO output.FileOutputCommitter: Saved output of task 'attempt\_local1989419825\_0001\_r\_000000\_0' to output

19/11/10 21:35:36 INFO mapred.LocalJobRunner: reduce > reduce

19/11/10 21:35:36 INFO mapred.Task: Task 'attempt\_local1989419825\_0001\_r\_000000\_0' done.

19/11/10 21:35:37 INFO mapred.JobClient: map 100% reduce 100%

19/11/10 21:35:37 INFO mapred.JobClient: Job complete: job\_local1989419825\_0001

19/11/10 21:35:37 INFO mapred.JobClient: Counters: 20

19/11/10 21:35:37 INFO mapred.JobClient: File System Counters

19/11/10 21:35:37 INFO mapred.JobClient: FILE: Number of bytes read=622

19/11/10 21:35:37 INFO mapred.JobClient: FILE: Number of bytes written=331634

19/11/10 21:35:37 INFO mapred.JobClient: FILE: Number of read operations=0

19/11/10 21:35:37 INFO mapred.JobClient: FILE: Number of large read operations=0

19/11/10 21:35:37 INFO mapred.JobClient: FILE: Number of write operations=0

19/11/10 21:35:37 INFO mapred.JobClient: Map-Reduce Framework

19/11/10 21:35:37 INFO mapred.JobClient: Map input records=3

19/11/10 21:35:37 INFO mapred.JobClient: Map output records=12

19/11/10 21:35:37 INFO mapred.JobClient: Map output bytes=114

19/11/10 21:35:37 INFO mapred.JobClient: Input split bytes=120

19/11/10 21:35:37 INFO mapred.JobClient: Combine input records=0

19/11/10 21:35:37 INFO mapred.JobClient: Combine output records=0

19/11/10 21:35:37 INFO mapred.JobClient: Reduce input groups=4

19/11/10 21:35:37 INFO mapred.JobClient: Reduce shuffle bytes=0

19/11/10 21:35:37 INFO mapred.JobClient: Reduce input records=12

19/11/10 21:35:37 INFO mapred.JobClient: Reduce output records=4

19/11/10 21:35:37 INFO mapred.JobClient: Spilled Records=24

19/11/10 21:35:37 INFO mapred.JobClient: CPU time spent (ms)=0

19/11/10 21:35:37 INFO mapred.JobClient: Physical memory (bytes) snapshot=0

19/11/10 21:35:37 INFO mapred.JobClient: Virtual memory (bytes) snapshot=0

19/11/10 21:35:37 INFO mapred.JobClient: Total committed heap usage (bytes)=486539264

1. Or we can export the project to a jar file, and run it using the following step:
   1. Make necessary directory in HDFS
      1. hadoop fs -mkdir /user/cloudera /user/cloudera/wordcount /user/cloudera/wordcount/input
   2. Prepare the input file
      1. echo "Hadoop is an elephant" > file0
      2. echo "Hadoop is as yellow as can be" > file1
      3. echo "Oh what a yellow fellow is Hadoop" > file2
      4. hadoop fs -put file\* /user/cloudera/wordcount/input
   3. run the jar file
      1. hadoop jar wordcount.jar /user/cloudera/wordcount/input /user/cloudera/wordcount/output
   4. check the output
      1. hadoop fs -cat /user/cloudera/wordcount/output/\*
   5. if we want to run again, we must delete the output and go to step “run the jar file”
      1. hadoop fs -rm -r /user/cloudera/wordcount/output