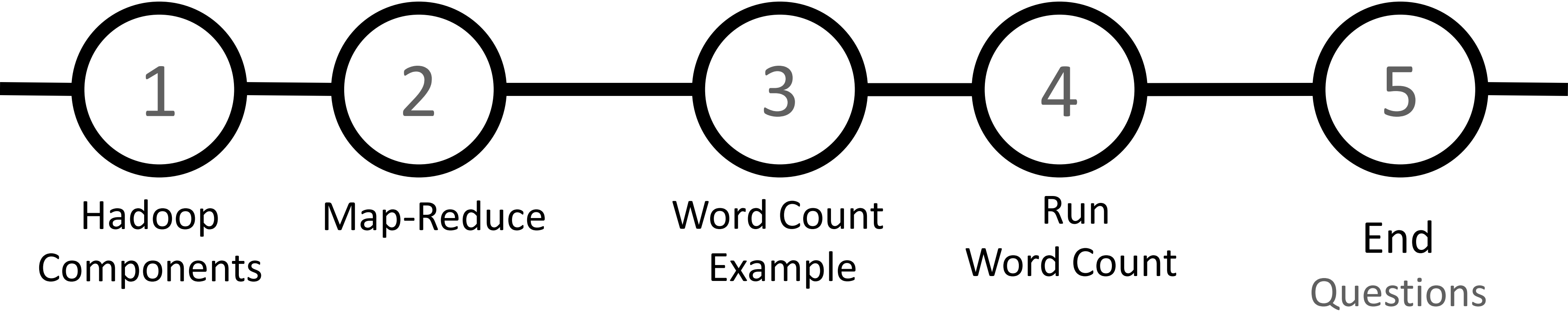
Part 3

AGENDA



1 Hadoop Components

2 Map-Reduce

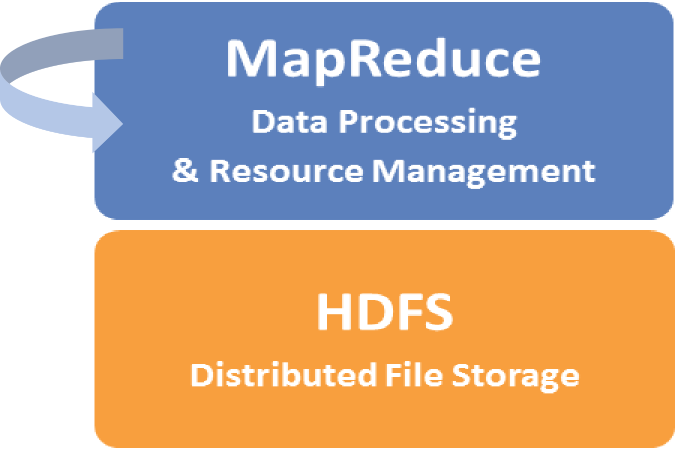
3 Word Count Example

4 Run Word Count

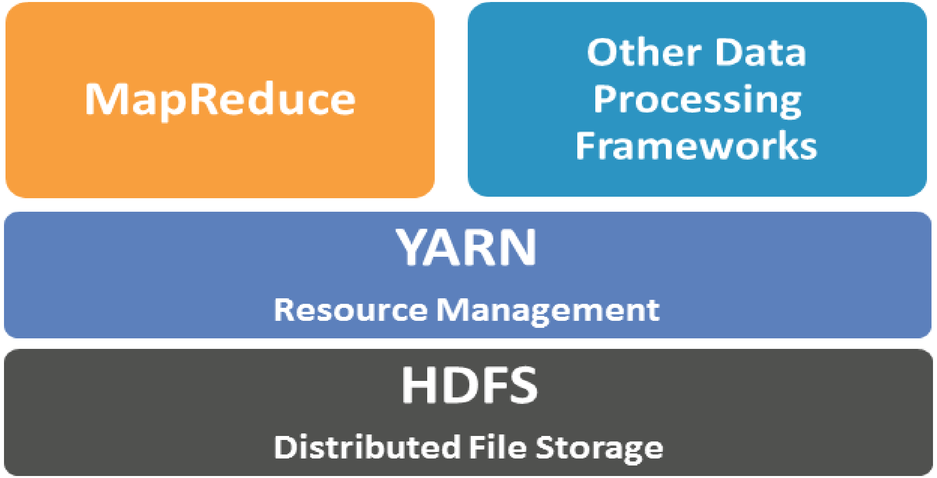
5 End Questions

Hadoop component

Hadoop v1.0



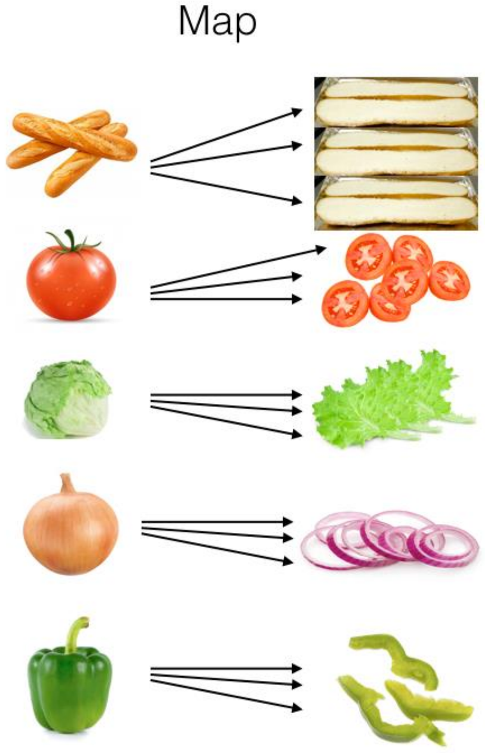
Hadoop v2.0



MapReduce

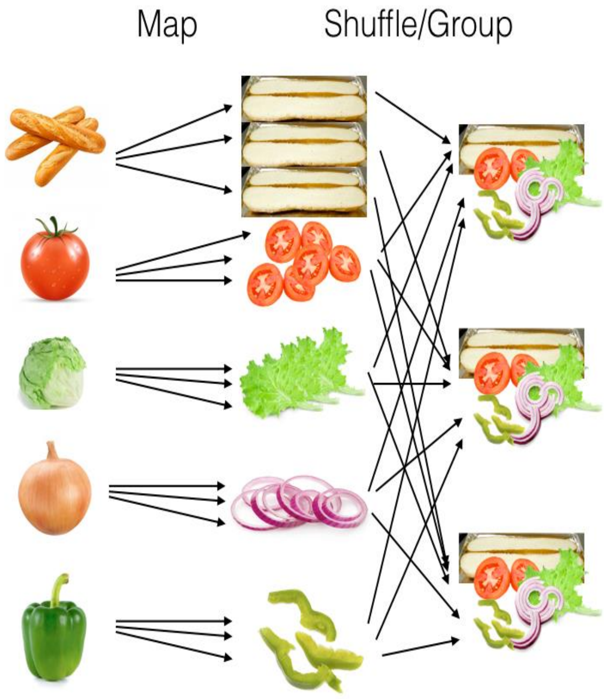
Map Reduce Model

Map



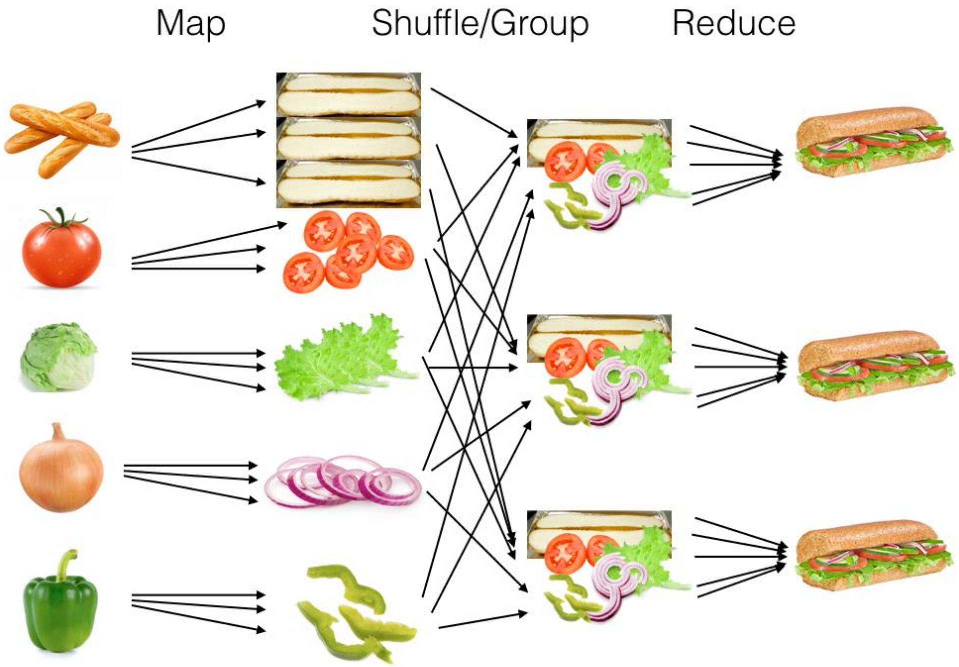
We distribute our raw ingredients amongst the **workers** in our shop. One person takes the tomatoes, one person takes the lettuce, one person takes the onions, and so on. We'll call this the "map" stage.

shuffle/group



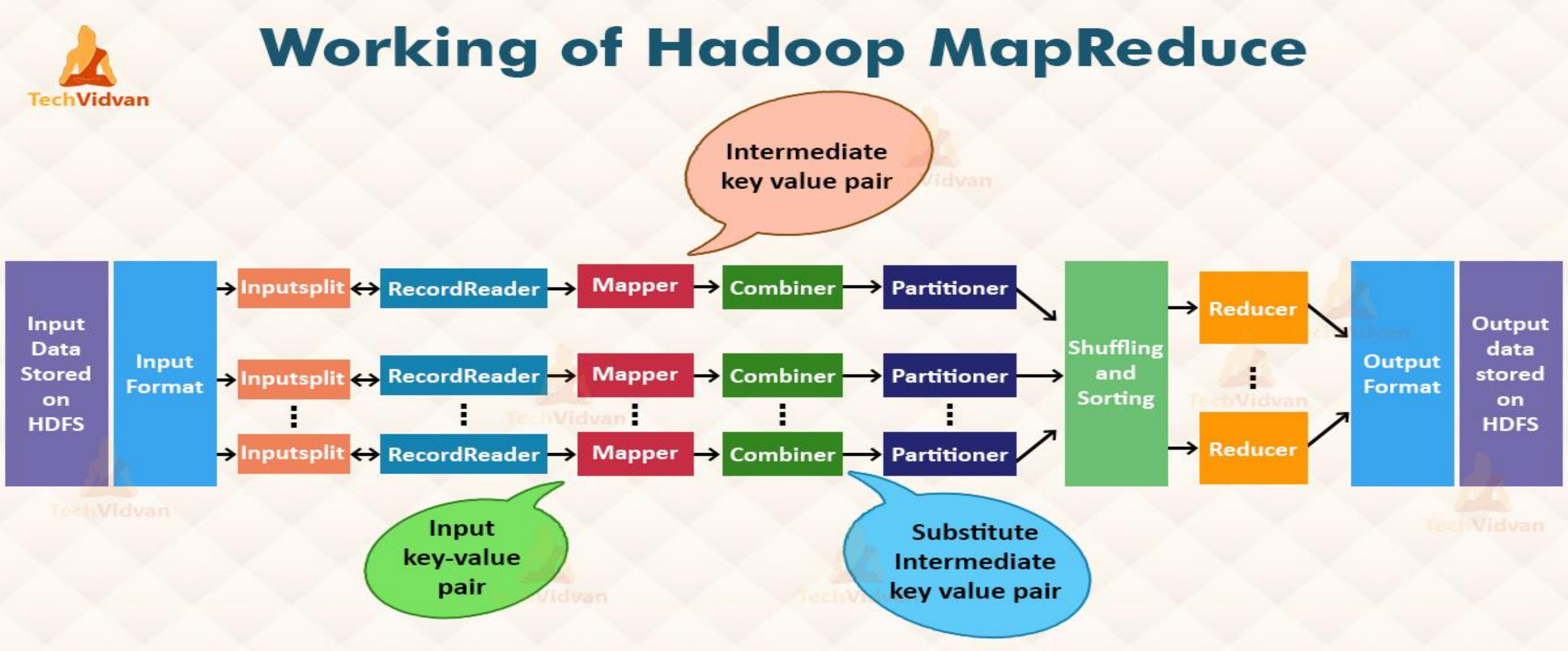
Next, we take these processed ingredients (which we'll call "mapper intermediates") and group them together into piles, so that making a sandwich becomes easy. **We'll call this the "shuffle/group" stage.**

Reduce

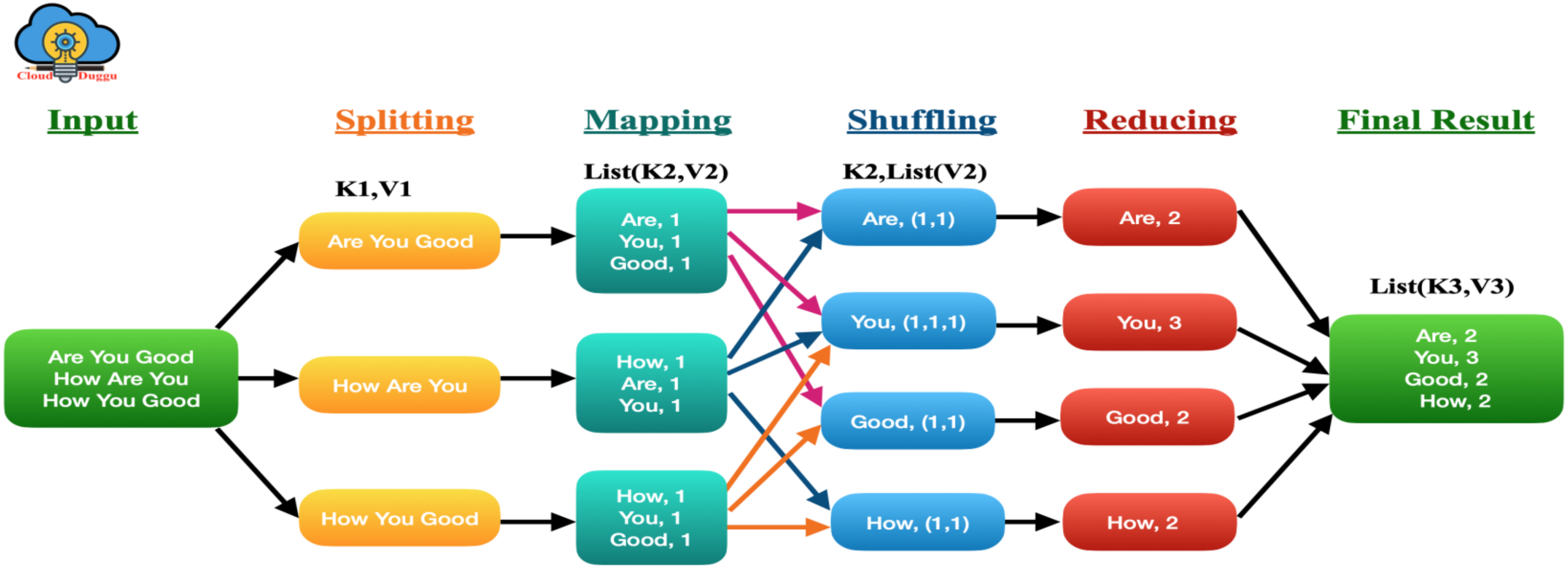


Finally, we'll combine the ingredients into a sandwich. **We'll call this the "reduce" stage.**

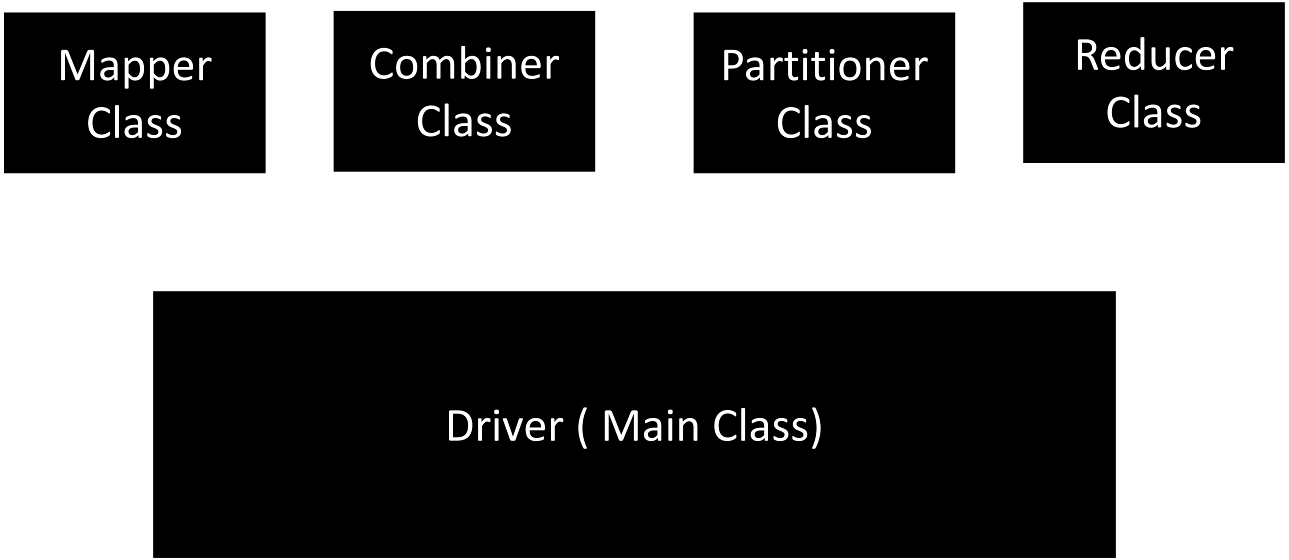
Overall stages



Wordcount Example



Hadoop Client Job



Mapper Class

Combiner Class

Partitioner Class

Reducer Class

Driver (Main Class)

Driver

- ➢ The code that runs on the client machine configures the **job** details by creating an object from the Job **class**, which implements the **JobContext** interface.

```
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, “Job Name");
```

- ➢ It submit the job to cluster.

- ➢ It parse job argument to identify job parameters for example : input & output directory.

Driver Job Configuration:

- It submit the job to cluster. The job object allow you to set configuration for your M/R Job.
- Configure Map per , Combiner ,Partitioner ,Reducer classes.

```
job.setMapperClass(MapperSide.class);
job.setReducerClass(ReducerSide.class);
```

- Set Input /Output [Key- Value] data types for each Mapper & Reducer.

```
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
```

- Configure input & output directory.

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

Keys and Values

- Keys and Values in Hadoop are objects not primitive data types.
- Values are Objects which implement **Writable**.
- Keys are Objects which implement **WritableComparable**. **[Sorting]**
- int in Java Match IntWritable , string java is Text in Hadoop.

Mapper

- The mapper class deals with a single input split(block).
- All mapper classes must extend the Mapper base class.
- All mapper must specify the key and values for input and output.
- All mappers must **override** the map method and pass the key, value, and Context.
- The context is used to **write** the intermediate data and all information about the job conf

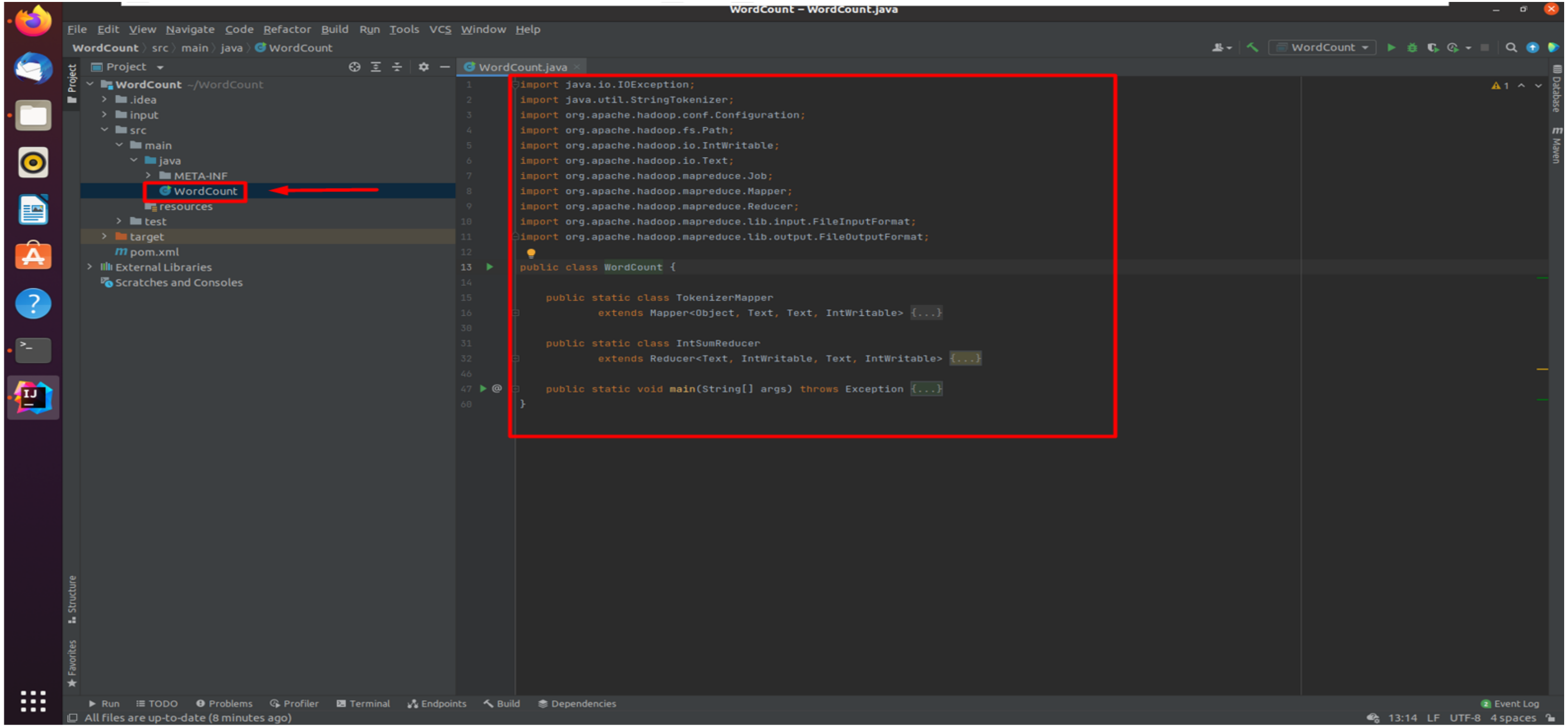
Reducer

- The Reducer receives a **Key and an iterable** collection of Writable objects.
- It also receives a Context object.
- All reducers classes must extend **the Reducer** base class.
- All **Reducer** must specify the key and values for intermediate input and final (or intermediate) output.
- All Reducer must override **reduce** method and pass the key , iterable and context.

MapReduce Demo

Run MapReduce Job Locally

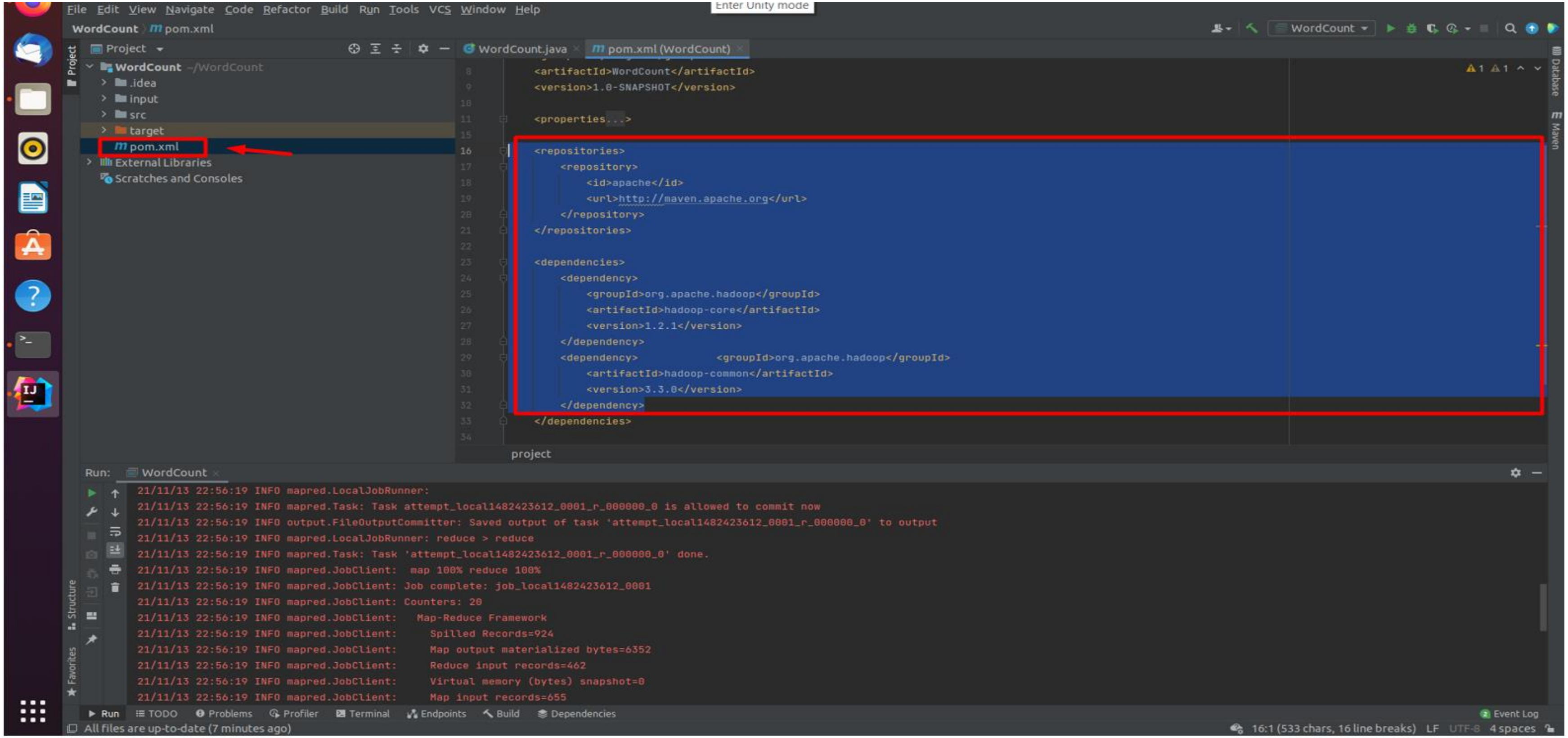
Create Class WordCount.java



```java
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
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;

public class WordCount {

    public static class TokenizerMapper
            extends Mapper<Object, Text, Text, IntWritable> {...}

    public static class IntSumReducer
            extends Reducer<Text, IntWritable, Text, IntWritable> {...}

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

All files are up-to-date (8 minutes ago)

Edit pom.xml File (add repo , dependencies)



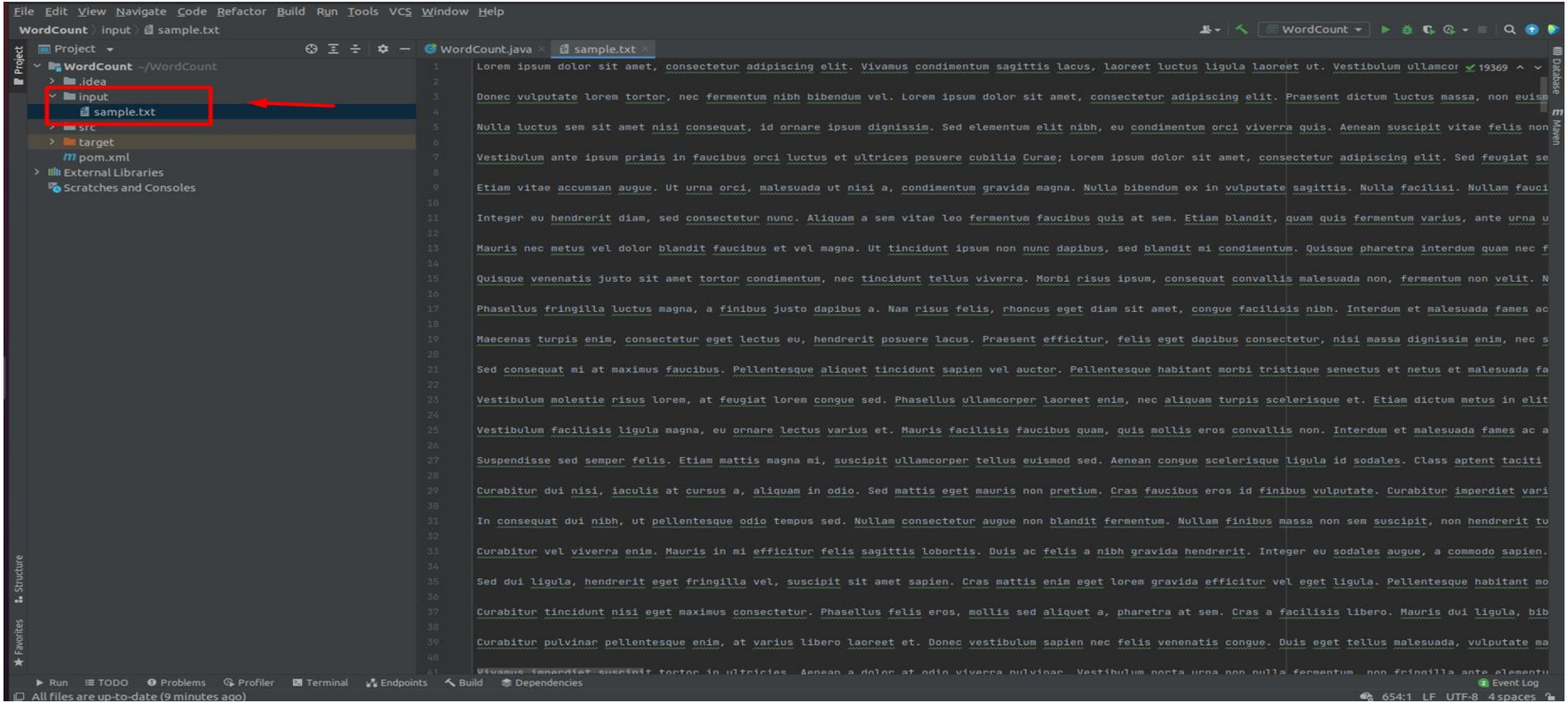
```xml
    <artifactId>WordCount</artifactId>
    <version>1.0-SNAPSHOT</version>

    <properties...>

    <repositories>
        <repository>
            <id>apache</id>
            <url>http://maven.apache.org</url>
        </repository>
    </repositories>

    <dependencies>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-core</artifactId>
            <version>1.2.1</version>
        </dependency>
        <dependency>            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-common</artifactId>
            <version>3.3.0</version>
        </dependency>
    </dependencies>
```

```
21/11/13 22:56:19 INFO mapred.LocalJobRunner:
21/11/13 22:56:19 INFO mapred.Task: Task attempt_local1482423612_0001_r_000000_0 is allowed to commit now
21/11/13 22:56:19 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1482423612_0001_r_000000_0' to output
21/11/13 22:56:19 INFO mapred.LocalJobRunner: reduce > reduce
21/11/13 22:56:19 INFO mapred.Task: Task 'attempt_local1482423612_0001_r_000000_0' done.
21/11/13 22:56:19 INFO mapred.JobClient:  map 100% reduce 100%
21/11/13 22:56:19 INFO mapred.JobClient: Job complete: job_local1482423612_0001
21/11/13 22:56:19 INFO mapred.JobClient: Counters: 20
21/11/13 22:56:19 INFO mapred.JobClient:   Map-Reduce Framework
21/11/13 22:56:19 INFO mapred.JobClient:     Spilled Records=924
21/11/13 22:56:19 INFO mapred.JobClient:     Map output materialized bytes=6352
21/11/13 22:56:19 INFO mapred.JobClient:     Reduce input records=462
21/11/13 22:56:19 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
21/11/13 22:56:19 INFO mapred.JobClient:     Map input records=655
```

Edit pom.xml File (add repo , dependencies)

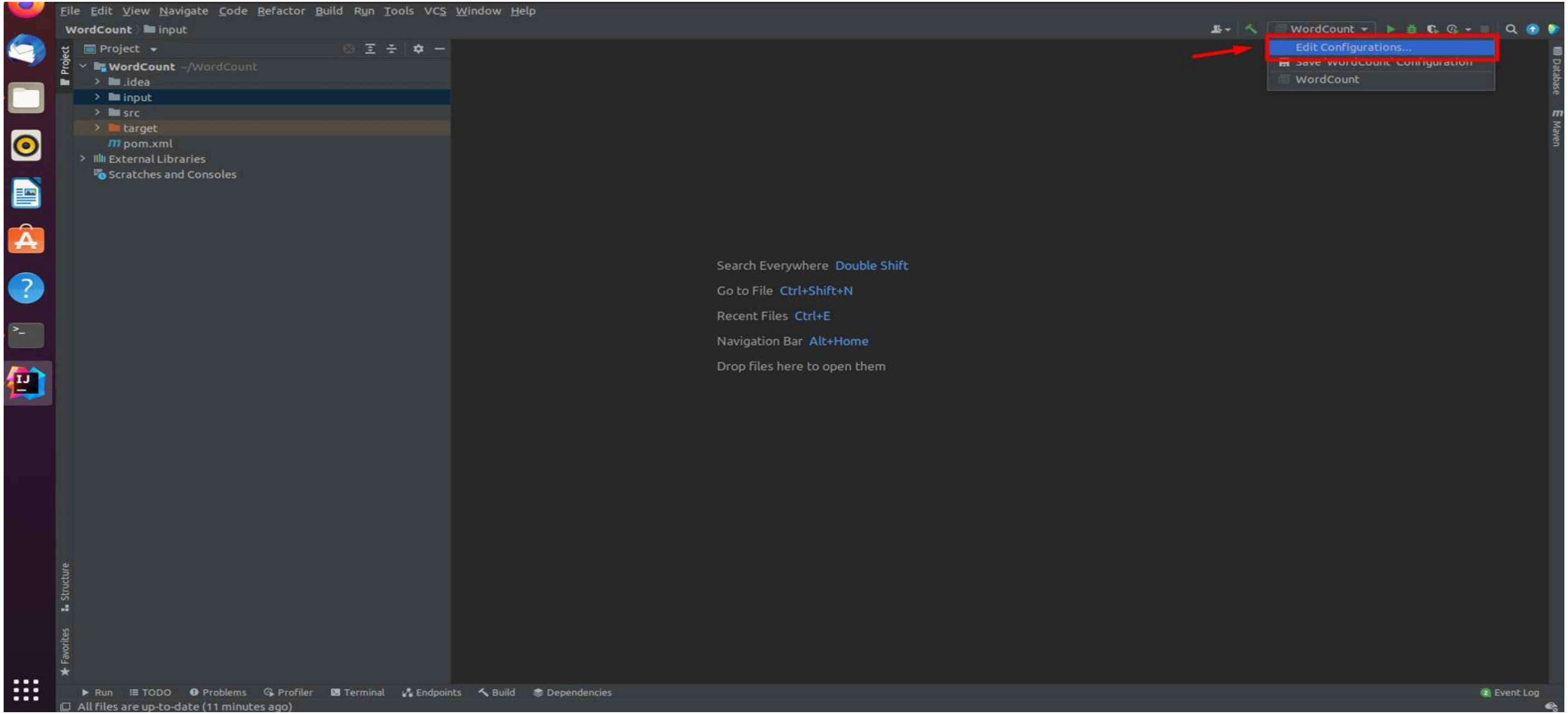
```
<repositories>
   <repository>
     <id>apache</id>
     <url>http://maven.apache.org</url>
   </repository>
 </repositories>

 <dependencies>
   <dependency>
     <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-core</artifactId>
     <version>1.2.1</version>
   </dependency>
   <dependency>      <groupId>org.apache.hadoop</groupId>
     <artifactId>hadoop-common</artifactId>
     <version>3.3.0</version>
   </dependency>
 </dependencies>
```

Create Input directory & Input file

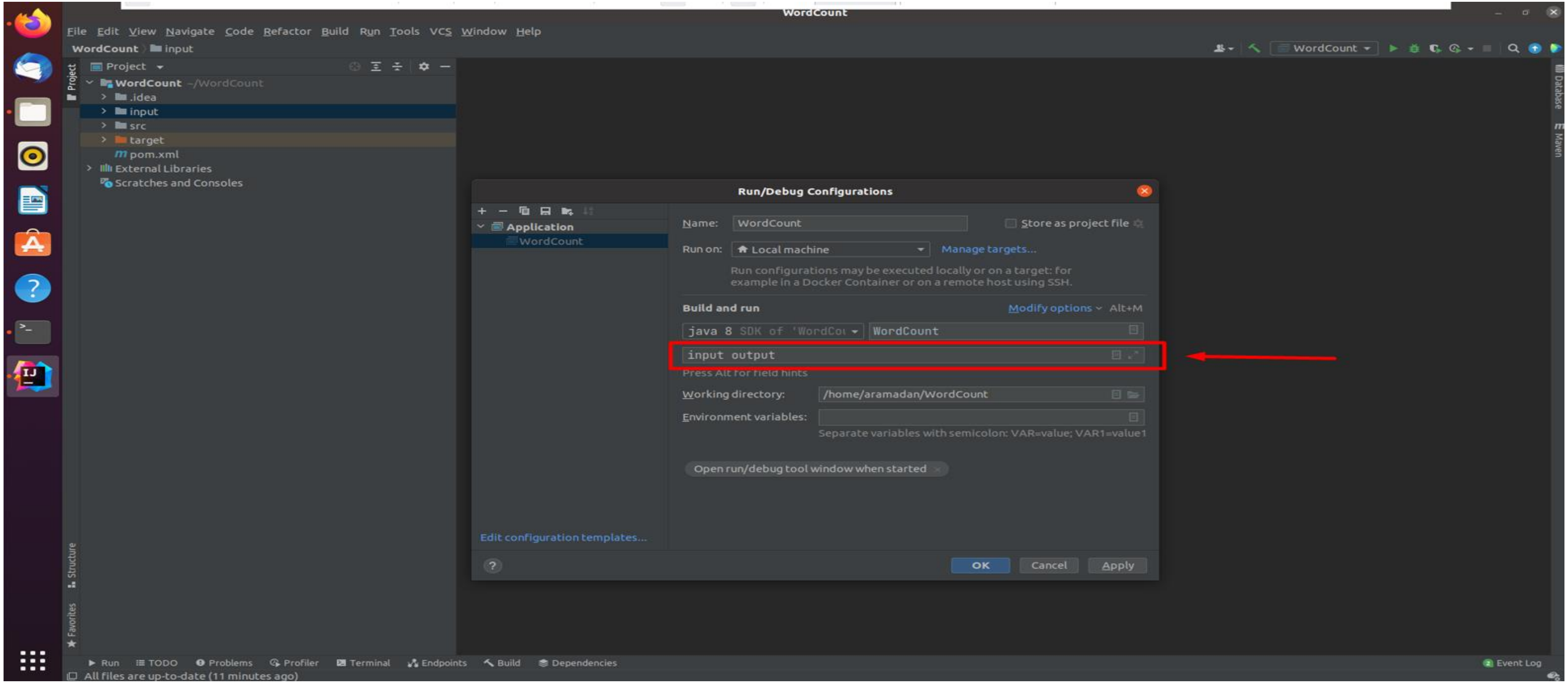


```
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus condimentum sagittis lacus, laoreet luctus ligula laoreet ut. Vestibulum ullamcor

Donec vulputate lorem tortor, nec fermentum nibh bibendum vel. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent dictum luctus massa, non euism

Nulla luctus sem sit amet nisi consequat, id ornare ipsum dignissim. Sed elementum elit nibh, eu condimentum orci viverra quis. Aenean suscipit vitae felis non

Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed feugiat se

Etiam vitae accumsan augue. Ut urna orci, malesuada ut nisi a, condimentum gravida magna. Nulla bibendum ex in vulputate sagittis. Nulla facilisi. Nullam fauci

Integer eu hendrerit diam, sed consectetur nunc. Aliquam a sem vitae leo fermentum faucibus quis at sem. Etiam blandit, quam quis fermentum varius, ante urna u

Mauris nec metus vel dolor blandit faucibus et vel magna. Ut tincidunt ipsum non nunc dapibus, sed blandit mi condimentum. Quisque pharetra interdum quam nec f

Quisque venenatis justo sit amet tortor condimentum, nec tincidunt tellus viverra. Morbi risus ipsum, consequat convallis malesuada non, fermentum non velit. N

Phasellus fringilla luctus magna, a finibus justo dapibus a. Nam risus felis, rhoncus eget diam sit amet, congue facilisis nibh. Interdum et malesuada fames ac

Maecenas turpis enim, consectetur eget lectus eu, hendrerit posuere lacus. Praesent efficitur, felis eget dapibus consectetur, nisi massa dignissim enim, nec s

Sed consequat mi at maximus faucibus. Pellentesque aliquet tincidunt sapien vel auctor. Pellentesque habitant morbi tristique senectus et netus et malesuada fa

Vestibulum molestie risus lorem, at feugiat lorem congue sed. Phasellus ullamcorper laoreet enim, nec aliquam turpis scelerisque et. Etiam dictum metus in elit

Vestibulum facilisis ligula magna, eu ornare lectus varius et. Mauris facilisis faucibus quam, quis mollis eros convallis non. Interdum et malesuada fames ac a

Suspendisse sed semper felis. Etiam mattis magna mi, suscipit ullamcorper tellus euismod sed. Aenean congue scelerisque ligula id sodales. Class aptent taciti

Curabitur dui nisi, iaculis at cursus a, aliquam in odio. Sed mattis eget mauris non pretium. Cras faucibus eros id finibus vulputate. Curabitur imperdiet vari

In consequat dui nibh, ut pellentesque odio tempus sed. Nullam consectetur augue non blandit fermentum. Nullam finibus massa non sem suscipit, non hendrerit tu

Curabitur vel viverra enim. Mauris in mi efficitur felis sagittis lobortis. Duis ac felis a nibh gravida hendrerit. Integer eu sodales augue, a commodo sapien.

Sed dui ligula, hendrerit eget fringilla vel, suscipit sit amet sapien. Cras mattis enim eget lorem gravida efficitur vel eget ligula. Pellentesque habitant mo

Curabitur tincidunt nisi eget maximus consectetur. Phasellus felis eros, mollis sed aliquet a, pharetra at sem. Cras a facilisis libero. Mauris dui ligula, bib

Curabitur pulvinar pellentesque enim, at varius libero laoreet et. Donec vestibulum sapien nec felis venenatis congue. Duis eget tellus malesuada, vulputate ma

Vivamus imperdiet suscipit tortor in ultricies. Aenean a dolor at odio viverra pulvinar. Vestibulum porta urna non nulla fermentum, non fringilla ante elementu
```

Edit Configuration(To add arguments)



Add input & output dirs in arguments



Run/Debug Configurations

Name: WordCount

Run on: Local machine

Run configurations may be executed locally or on a target: for example in a Docker Container or on a remote host using SSH.

Build and run

java 8 SDK of 'WordCou

WordCount

input output

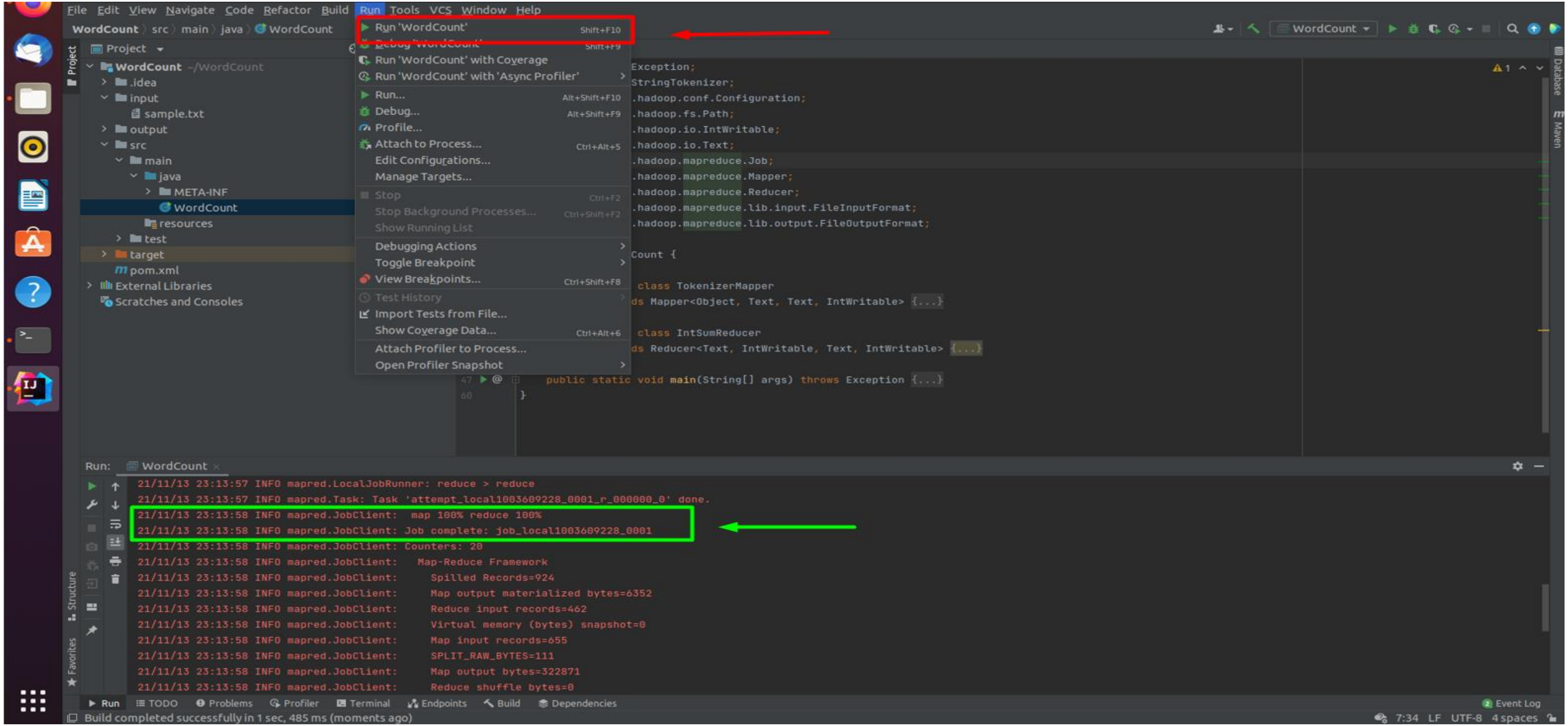
Press Alt for field hints

Working directory: /home/aramadan/WordCount

Environment variables:

Separate variables with semicolon: VAR=value; VAR1=value1

Run Project => Log tell you the Job Completed with 100%Map – 100% Reduce

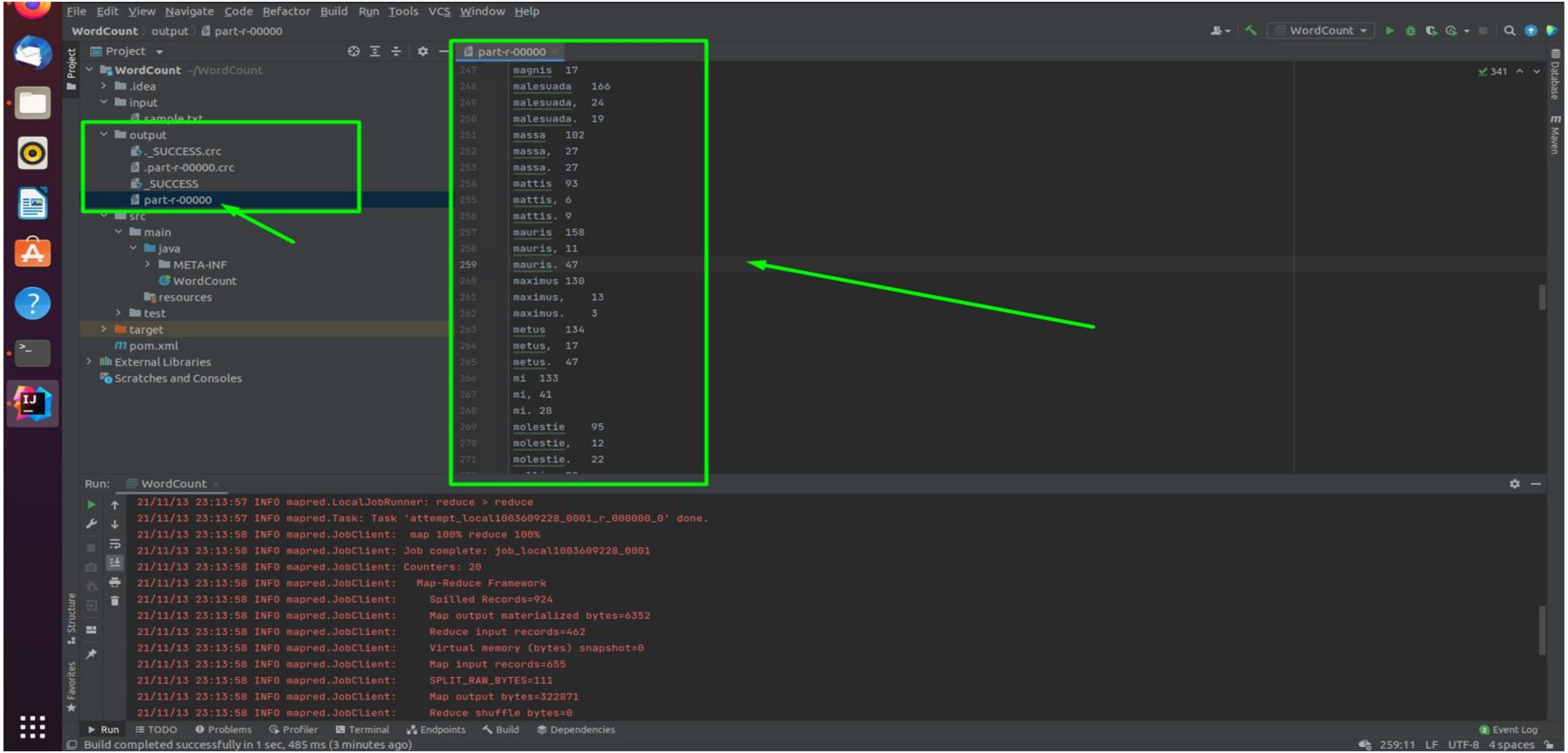


```
Exception;
StringTokenizer;
.hadoop.conf.Configuration;
.hadoop.fs.Path;
.hadoop.io.IntWritable;
.hadoop.io.Text;
.hadoop.mapreduce.Job;
.hadoop.mapreduce.Mapper;
.hadoop.mapreduce.Reducer;
.hadoop.mapreduce.lib.input.FileInputFormat;
.hadoop.mapreduce.lib.output.FileOutputFormat;

Count {

 class TokenizerMapper
ds Mapper<Object, Text, Text, IntWritable> {...}

 class IntSumReducer
ds Reducer<Text, IntWritable, Text, IntWritable> {...}

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

```
21/11/13 23:13:57 INFO mapred.LocalJobRunner: reduce > reduce
21/11/13 23:13:57 INFO mapred.Task: Task 'attempt_local1003609228_0001_r_000000_0' done.
21/11/13 23:13:58 INFO mapred.JobClient:  map 100% reduce 100%
21/11/13 23:13:58 INFO mapred.JobClient: Job complete: job_local1003609228_0001
21/11/13 23:13:58 INFO mapred.JobClient: Counters: 20
21/11/13 23:13:58 INFO mapred.JobClient:   Map-Reduce Framework
21/11/13 23:13:58 INFO mapred.JobClient:     Spilled Records=924
21/11/13 23:13:58 INFO mapred.JobClient:     Map output materialized bytes=6352
21/11/13 23:13:58 INFO mapred.JobClient:     Reduce input records=462
21/11/13 23:13:58 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
21/11/13 23:13:58 INFO mapred.JobClient:     Map input records=655
21/11/13 23:13:58 INFO mapred.JobClient:     SPLIT_RAW_BYTES=111
21/11/13 23:13:58 INFO mapred.JobClient:     Map output bytes=322871
21/11/13 23:13:58 INFO mapred.JobClient:     Reduce shuffle bytes=0
```

Build completed successfully in 1 sec, 485 ms (moments ago)

Output Directory



```
magnis  17
malesuada   166
malesuada,  24
malesuada.  19
massa   102
massa,  27
massa.  27
mattis  93
mattis, 6
mattis. 9
mauris  158
mauris, 11
mauris. 47
maximus 130
maximus,    13
maximus.    3
metus   134
metus,  17
metus.  47
mi  133
mi, 41
mi. 28
molestie    95
molestie,   12
molestie.   22
```

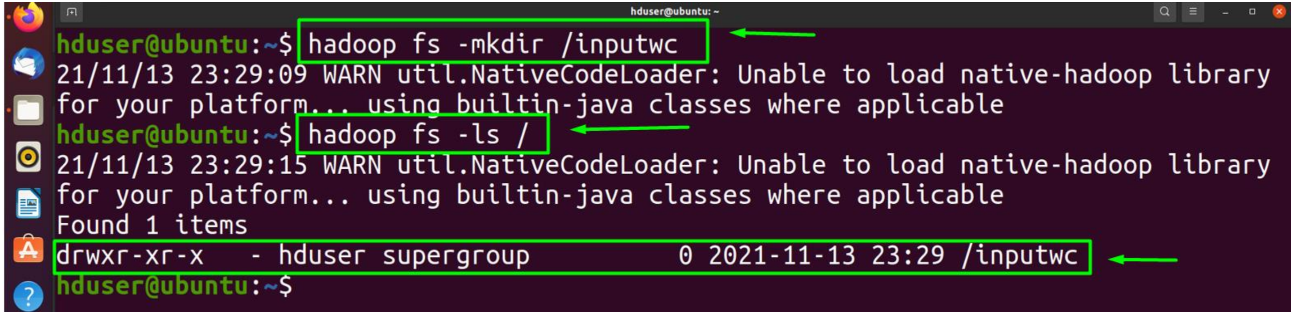
```
21/11/13 23:13:57 INFO mapred.LocalJobRunner: reduce > reduce
21/11/13 23:13:57 INFO mapred.Task: Task 'attempt_local1003609228_0001_r_000000_0' done.
21/11/13 23:13:58 INFO mapred.JobClient:  map 100% reduce 100%
21/11/13 23:13:58 INFO mapred.JobClient: Job complete: job_local1003609228_0001
21/11/13 23:13:58 INFO mapred.JobClient: Counters: 20
21/11/13 23:13:58 INFO mapred.JobClient:   Map-Reduce Framework
21/11/13 23:13:58 INFO mapred.JobClient:     Spilled Records=924
21/11/13 23:13:58 INFO mapred.JobClient:     Map output materialized bytes=6352
21/11/13 23:13:58 INFO mapred.JobClient:     Reduce input records=462
21/11/13 23:13:58 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=0
21/11/13 23:13:58 INFO mapred.JobClient:     Map input records=655
21/11/13 23:13:58 INFO mapred.JobClient:     SPLIT_RAW_BYTES=111
21/11/13 23:13:58 INFO mapred.JobClient:     Map output bytes=322871
21/11/13 23:13:58 INFO mapred.JobClient:     Reduce shuffle bytes=0
```

Build completed successfully in 1 sec, 485 ms (3 minutes ago)

MapReduce on HDFS

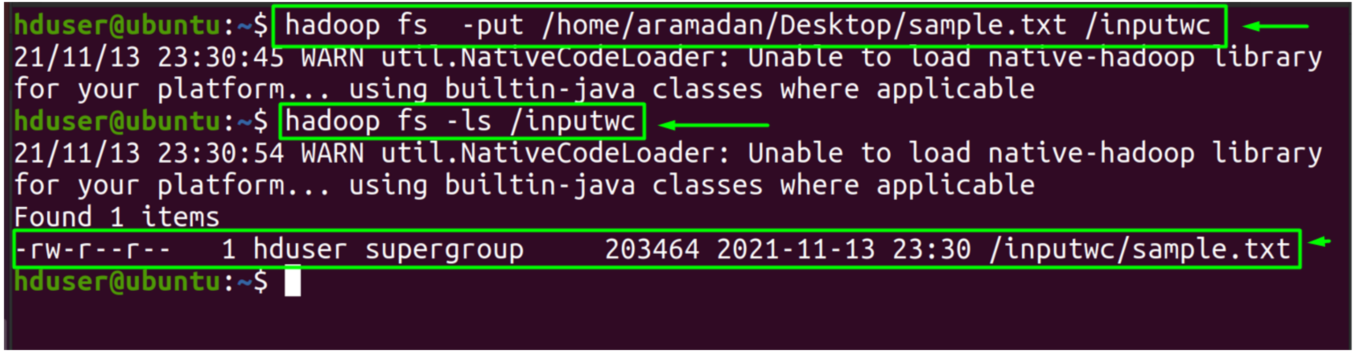
Run MapReduce Job on HDFS

Create Directory on hdfs



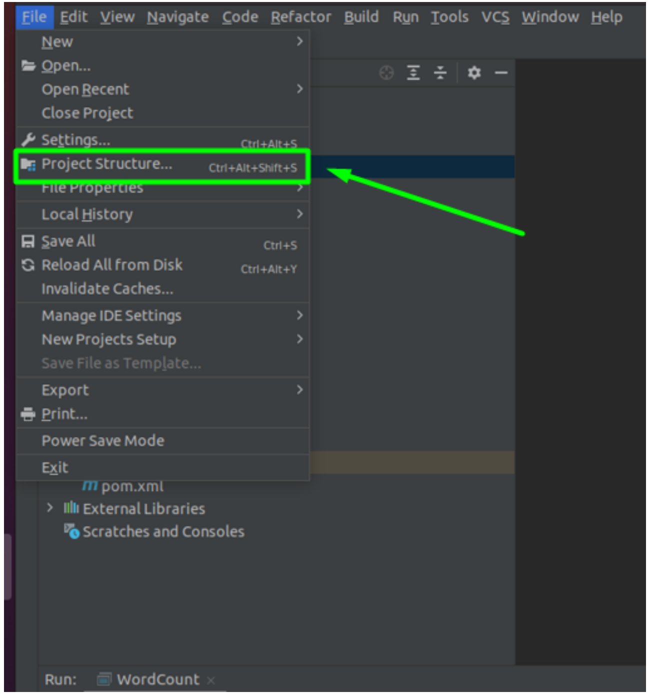
```
hduser@ubuntu:~$ hadoop fs -mkdir /inputwc
21/11/13 23:29:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
hduser@ubuntu:~$ hadoop fs -ls /
21/11/13 23:29:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x   - hduser supergroup          0 2021-11-13 23:29 /inputwc
hduser@ubuntu:~$
```

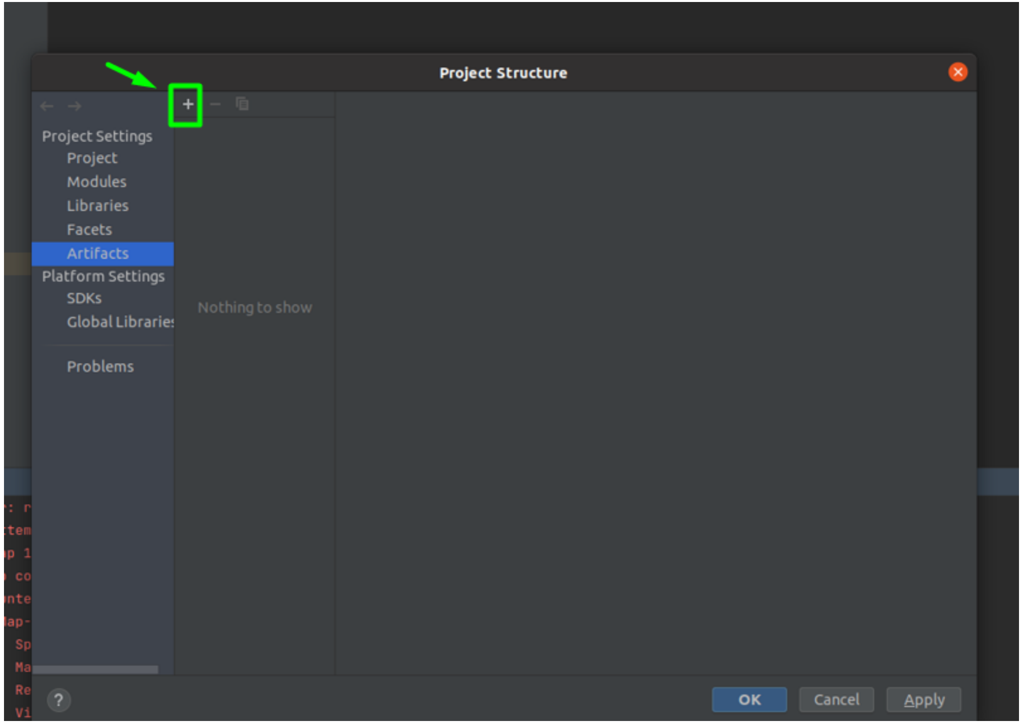
Move File from Linux to hdfs



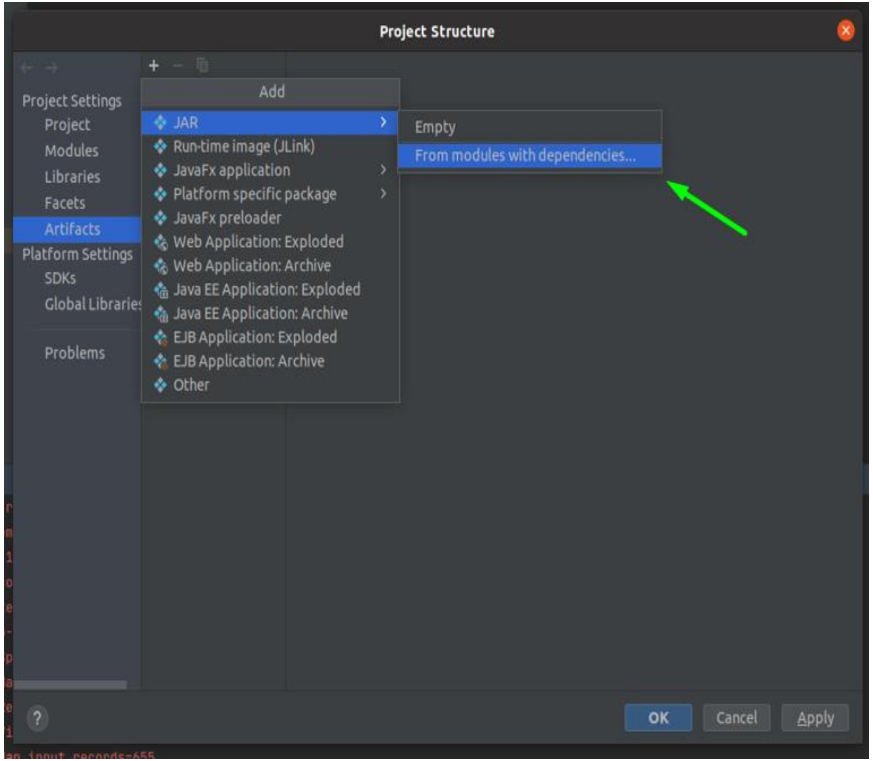
```
hduser@ubuntu:~$ hadoop fs  -put /home/aramadan/Desktop/sample.txt /inputwc
21/11/13 23:30:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
hduser@ubuntu:~$ hadoop fs -ls /inputwc
21/11/13 23:30:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 hduser supergroup     203464 2021-11-13 23:30 /inputwc/sample.txt
hduser@ubuntu:~$ 
```

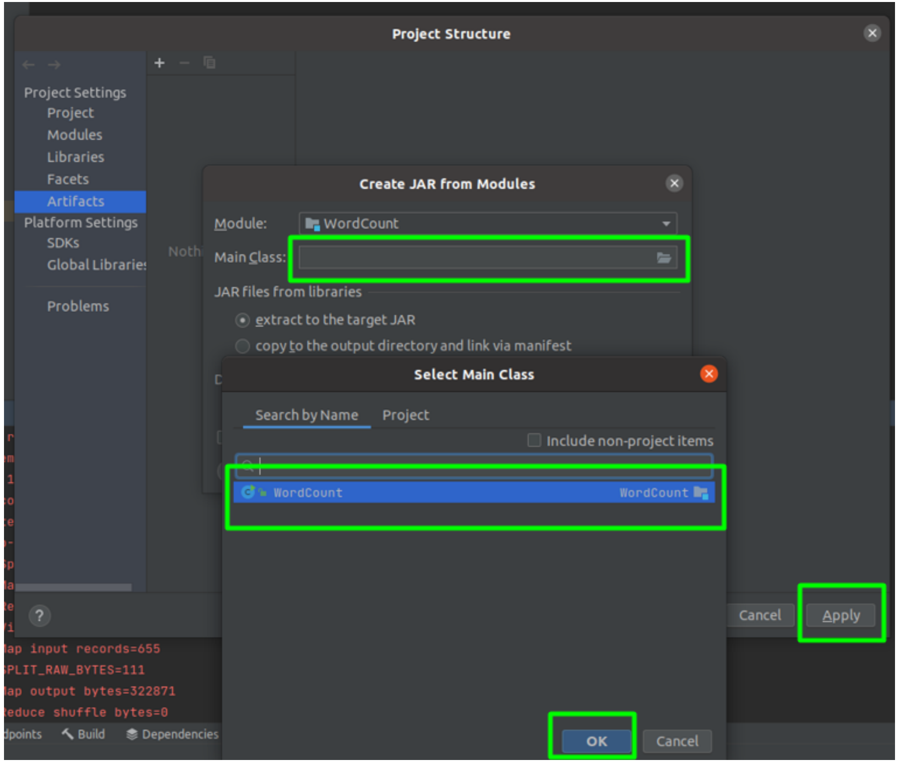
Set IntelliJ to build jar for artifacts





Set IntelliJ to build jar for artifacts





Module: WordCount

Main Class:

JAR files from libraries

● extract to the target JAR

○ copy to the output directory and link via manifest

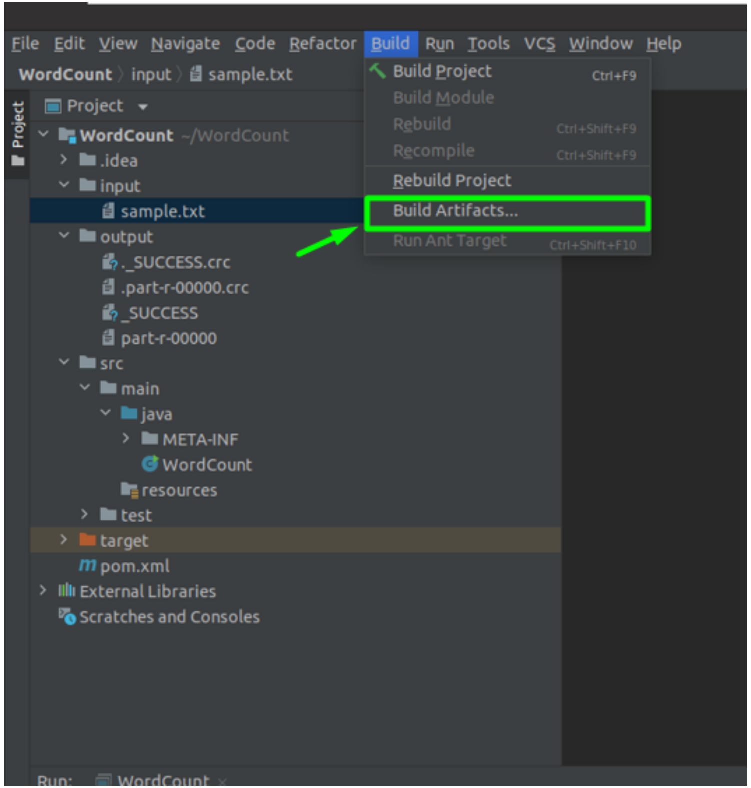
□ Include non-project items

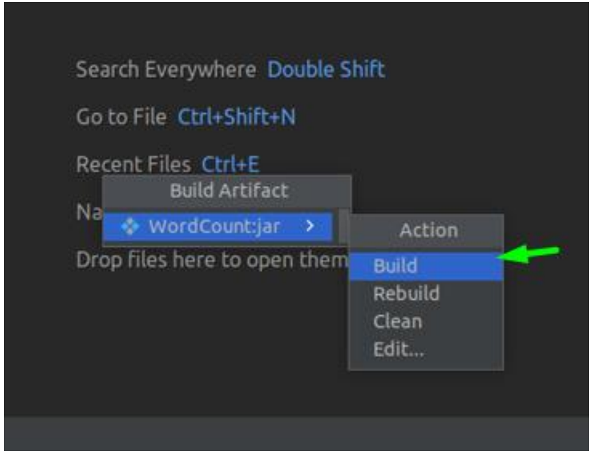
WordCount WordCount

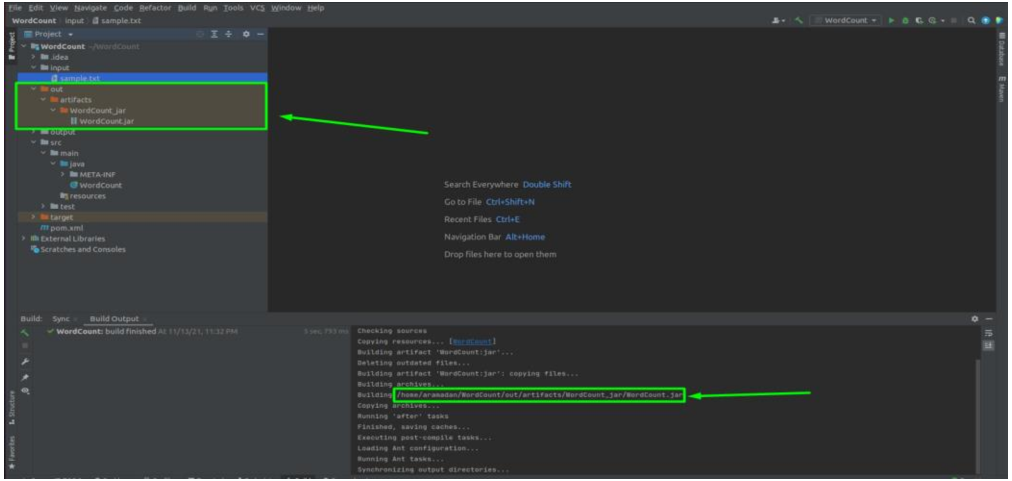
```
ap input records=655
SPLIT_RAW_BYTES=111
ap output bytes=322871
educe shuffle bytes=0
```

Ahmed Ramadan, ary00@fayoum.edu.eg

Build Arificate & Generate Jar File



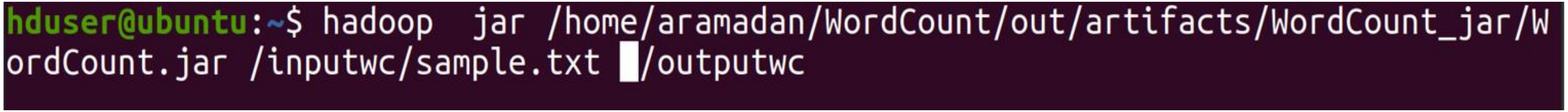




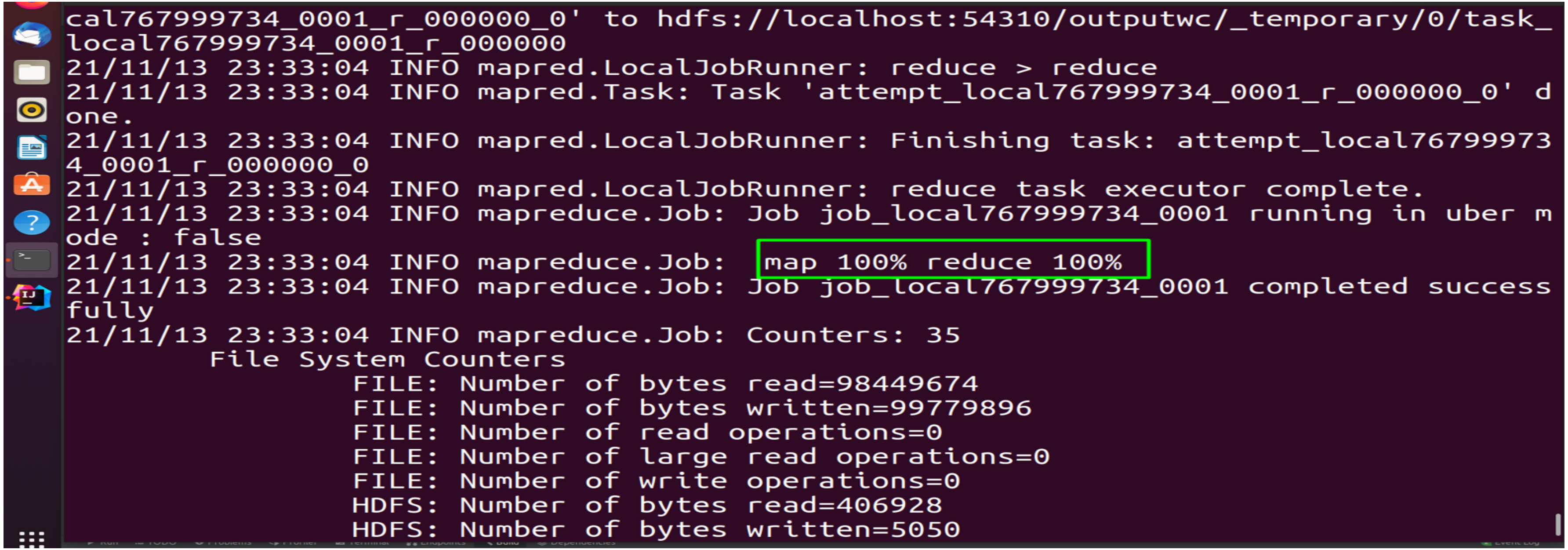
WordCount: build finished At 11/13/21, 11:32 PM 5 sec, 793 ms

```
Checking sources
Copying resources... [WordCount]
Building artifact 'WordCount:jar'...
Deleting outdated files...
Building artifact 'WordCount:jar': copying files...
Building archives...
Building /home/aramadan/WordCount/out/artifacts/WordCount_jar/WordCount.jar
Copying archives...
Running 'after' tasks
Finished, saving caches...
Executing post-compile tasks...
Loading Ant configuration...
Running Ant tasks...
Synchronizing output directories...
```

Run Jar on HDFS

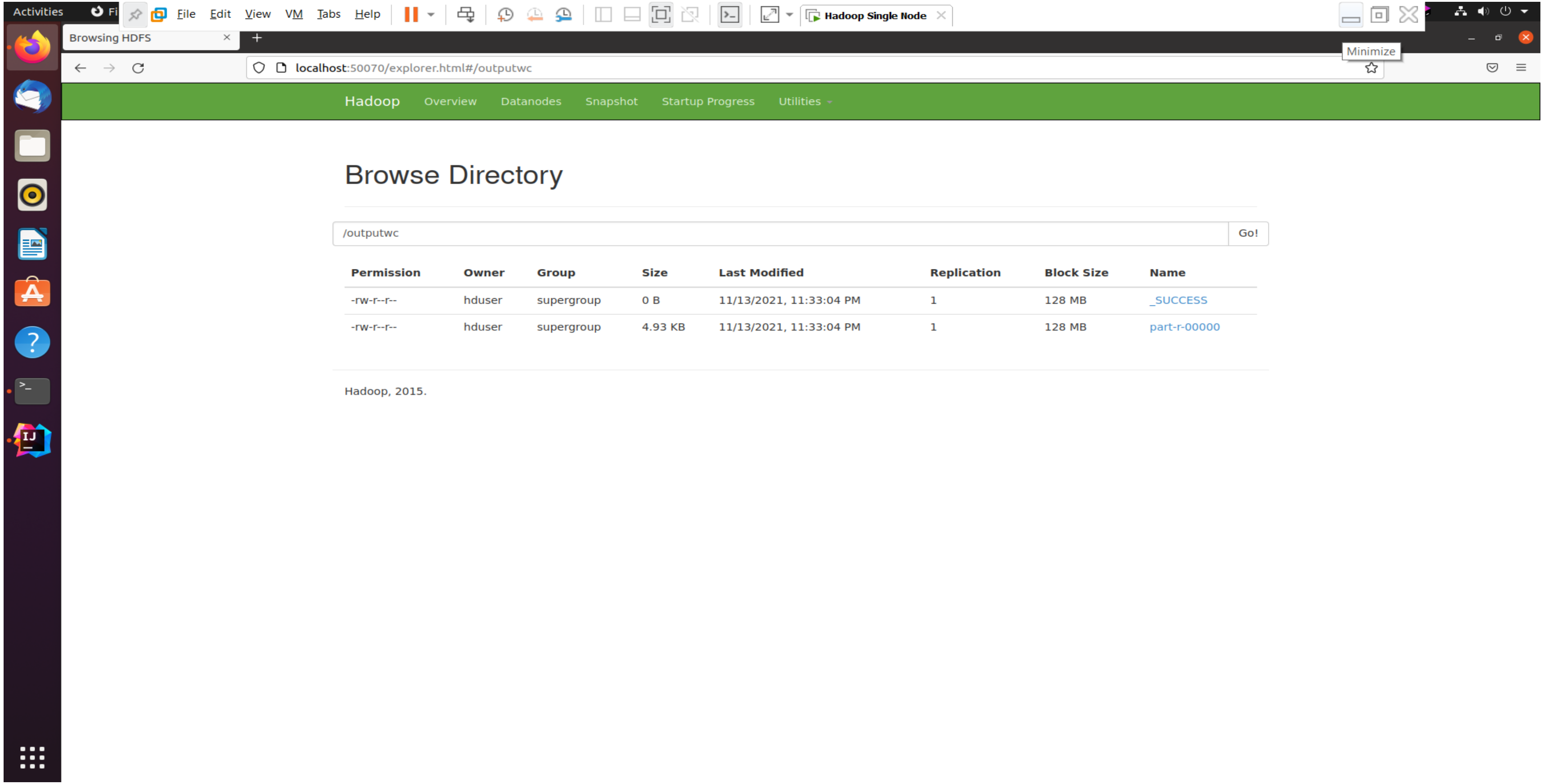


```
hduser@ubuntu:~$ hadoop  jar /home/aramadan/WordCount/out/artifacts/WordCount_jar/W
ordCount.jar /inputwc/sample.txt █/outputwc
```



```
cal767999734_0001_r_000000_0' to hdfs://localhost:54310/outputwc/_temporary/0/task_
local767999734_0001_r_000000
21/11/13 23:33:04 INFO mapred.LocalJobRunner: reduce > reduce
21/11/13 23:33:04 INFO mapred.Task: Task 'attempt_local767999734_0001_r_000000_0' d
one.
21/11/13 23:33:04 INFO mapred.LocalJobRunner: Finishing task: attempt_local76799973
4_0001_r_000000_0
21/11/13 23:33:04 INFO mapred.LocalJobRunner: reduce task executor complete.
21/11/13 23:33:04 INFO mapreduce.Job: Job job_local767999734_0001 running in uber m
ode : false
21/11/13 23:33:04 INFO mapreduce.Job:  map 100% reduce 100%
21/11/13 23:33:04 INFO mapreduce.Job: Job job_local767999734_0001 completed success
fully
21/11/13 23:33:04 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=98449674
                FILE: Number of bytes written=99779896
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=406928
                HDFS: Number of bytes written=5050
```

Results on web portal



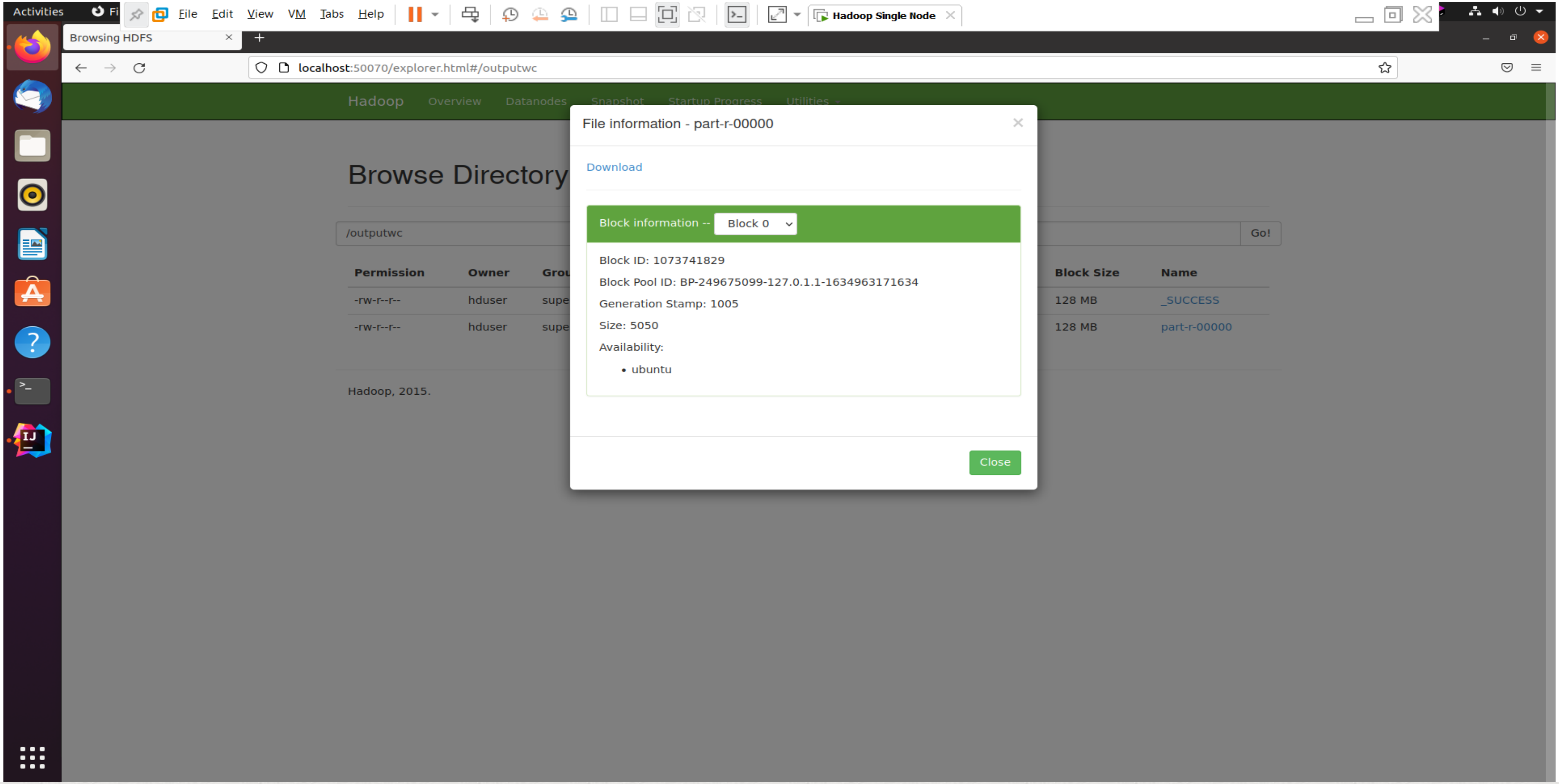
localhost:50070/explorer.html#/outputwc

Browse Directory

/outputwc

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	0 B	11/13/2021, 11:33:04 PM	1	128 MB	_SUCCESS
-rw-r--r--	hduser	supergroup	4.93 KB	11/13/2021, 11:33:04 PM	1	128 MB	part-r-00000

Hadoop, 2015.



File information - part-r-00000

Download

Block information -- Block 0

Block ID: 1073741829

Block Pool ID: BP-249675099-127.0.1.1-1634963171634

Generation Stamp: 1005

Size: 5050

Availability:

- ubuntu

Close

Permission	Owner	Grou	Block Size	Name
-rw-r--r--	hduser	supe	128 MB	_SUCCESS
-rw-r--r--	hduser	supe	128 MB	part-r-00000

Hadoop, 2015.

References

- https://reberhardt.com/cs110/summer-2018/lecture-notes/lecture-14/
- https://techvidvan.com/tutorials/how-mapreduce-works/
- https://www.cloudduggu.com/hadoop/architecture/

?

QUESTIONS

THANK YOU!