EKONOMICKÁ UNIVERZITA

Fakulta hospodárskej informatiky

Katedra aplikovanej informatiky

Map Reduce príklady

**Predmet:** Big Data

**Meno študentov:** Martin Jankech, Patrik Hajdučík

**Stupeň/ročník:** druhý/2.ročník

**Školský rok:** 2022/2023

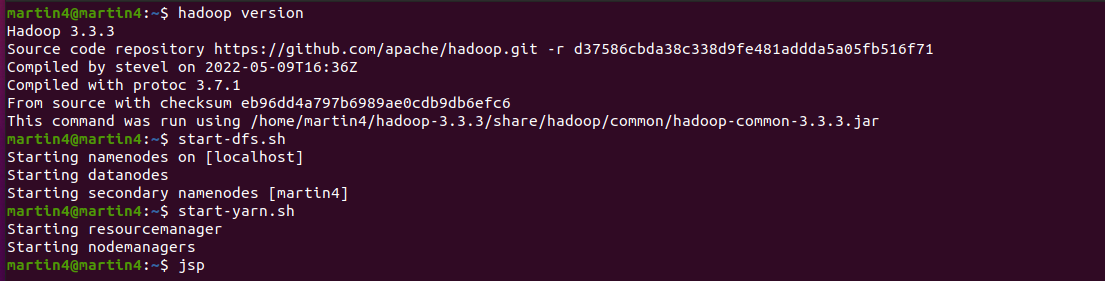
**Meno prednášajúceho:** Ing. Jaroslav Kultán PhD.

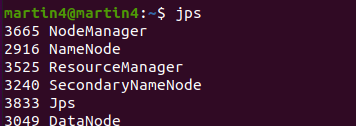
**Meno cvičiaceho:** SCHMIDT, Peter, Ing. Mgr., PhD.

# Úvod

V tomto zadaní si predstavíme 2 map reduce príklady vytvorené v programovacom jazyku Java. Ako vývojové prostredie sme využili InteliJ a ako operačný systém sme použili Linux Distribúciu Ubumtu 20.04

# WordCount-cez príkazový riadok

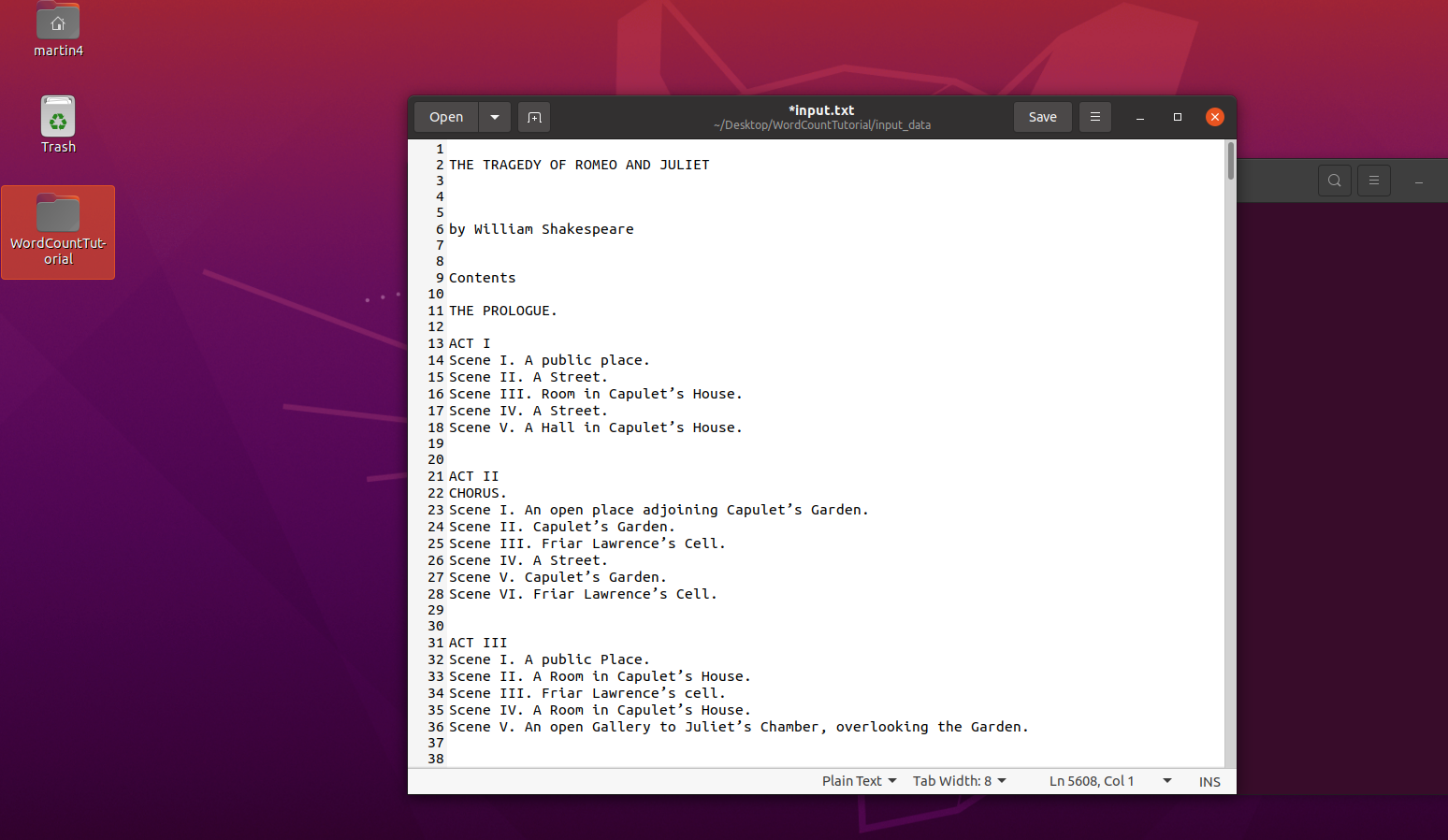
1. Najpr sa uistíme že máme nainštalovaný a spustený Hadoop



1. Uistíme sa že nám funguje java



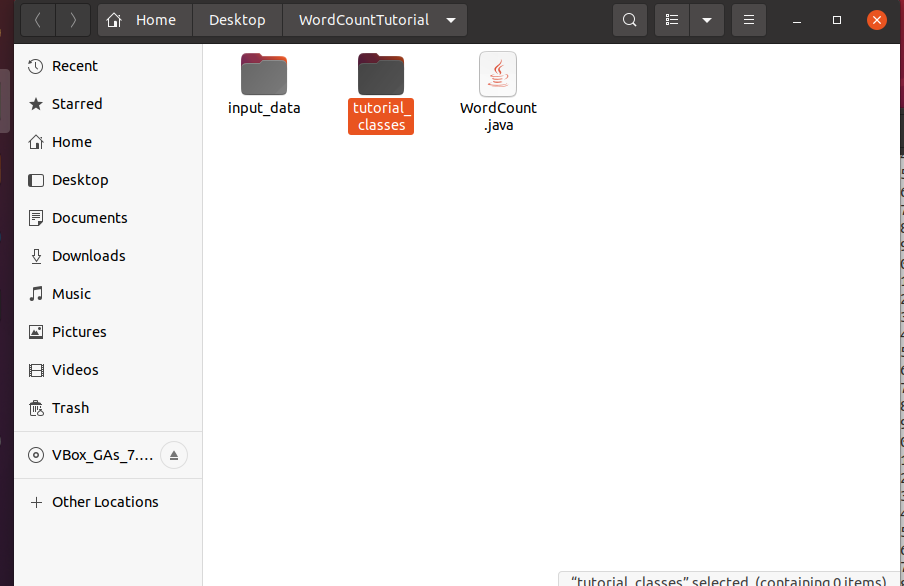
1. Vytvoríme si na ploche alebo ľubovoľnom mieste priečinok WordCountTutorial
2. V tomto priečinku vytvoríme ďalší prečinoch s názvom input-data
3. V priečinku input-data vytvoríme textový dokument s názvom input.txt do ktorého vložíme text na ktorom budeme robiť map reduce. My sme si zvoli textové dielo Romeo a Julia.



1. Pridáme java file s map reducom kódom ktorý bol získaný z apache dokumentácie

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>{  
  
 private final static IntWritable one = new IntWritable(1);  
 private Text word = new Text();  
  
 public void map(Object key, Text value, Context context  
 ) throws IOException, InterruptedException {  
 StringTokenizer itr = new StringTokenizer(value.toString());  
 while (itr.hasMoreTokens()) {  
 word.set(itr.nextToken());  
 context.write(word, one);  
 }  
 }  
 }  
  
 public static class IntSumReducer  
 extends Reducer<Text,IntWritable,Text,IntWritable> {  
 private IntWritable result = new IntWritable();  
  
 public void reduce(Text key, Iterable<IntWritable> values,  
 Context context  
 ) throws IOException, InterruptedException {  
 int sum = 0;  
 for (IntWritable val : values) {  
 sum += val.get();  
 }  
 result.set(sum);  
 context.write(key, result);  
 }  
 }  
  
 public static void main(String[] args) throws Exception {  
 Configuration conf = new Configuration();  
 Job job = Job.getInstance(conf, "word count");  
 job.setJarByClass(WordCount.class);  
 job.setMapperClass(TokenizerMapper.class);  
 job.setCombinerClass(IntSumReducer.class);  
 job.setReducerClass(IntSumReducer.class);  
 job.setOutputKeyClass(Text.class);  
 job.setOutputValueClass(IntWritable.class);  
 FileInputFormat.addInputPath(job, new Path(args[0]));  
 FileOutputFormat.setOutputPath(job, new Path(args[1]));  
 System.exit(job.waitForCompletion(true) ? 0 : 1);  
 }  
}

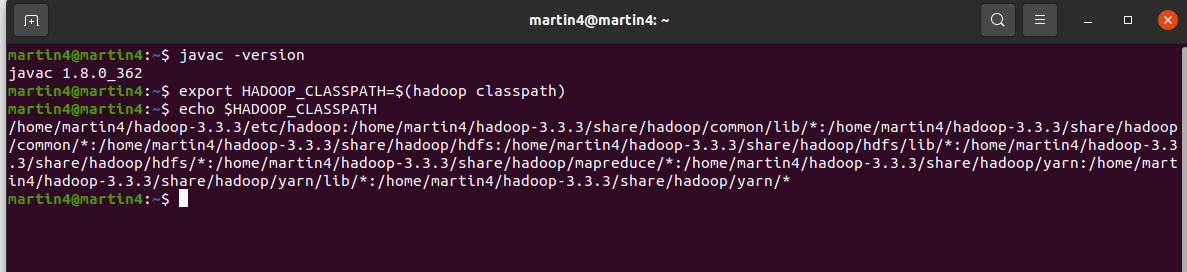
1. V priečinku WordCountTutorial vytvoríme dalši priečinok tutorial\_classes



1. V dalšom kroku nastavíme HADOOP\_CLASSPATH premenné prostredia

export HADOOP\_CLASSPATH=$(hadoop classpath)

echo $HADOOP\_CLASSPATH

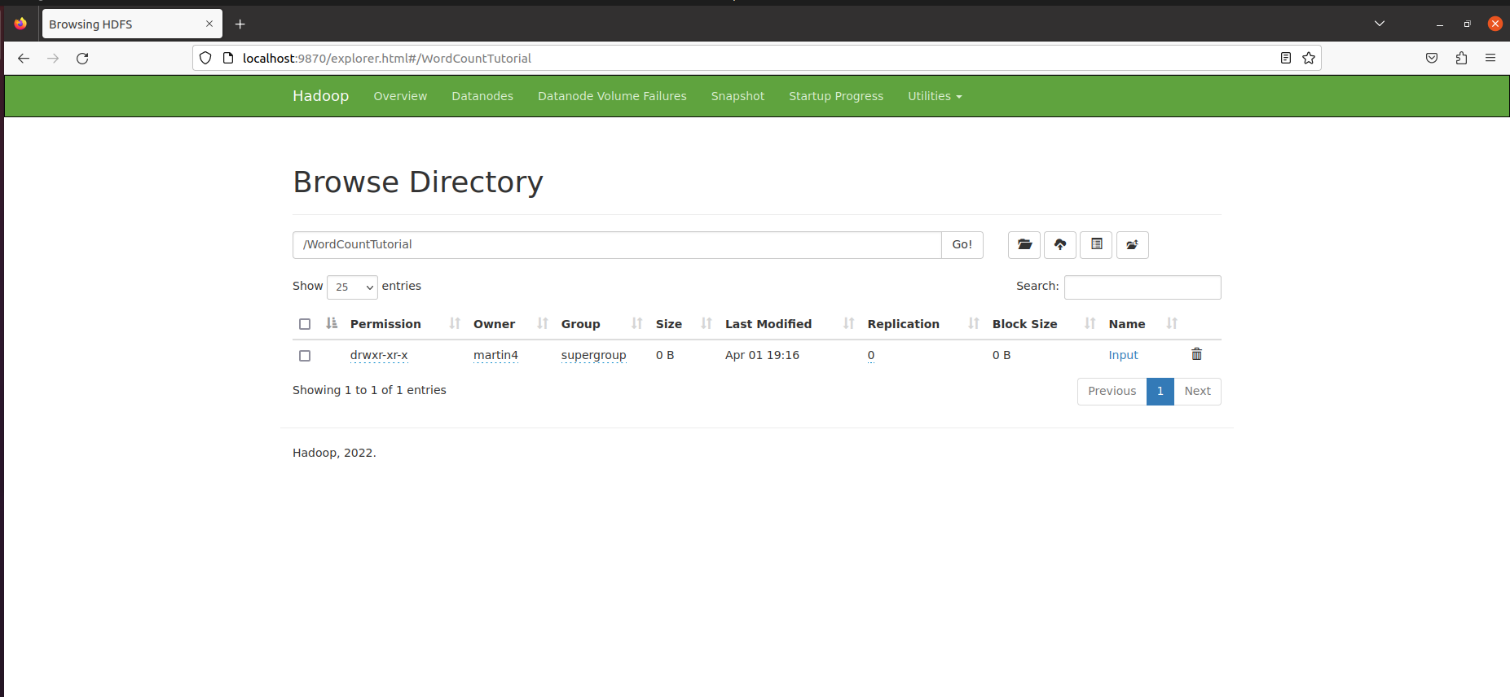


1. Taktiež vytvoríme priečinok v HDFS

hadoop fs -mkdir /WordCountTutorial

1. A taktiež vytvoríme priečinok pre Input

hadoop fs -mkdir /WordCountTutorial/Input

1. Otvorime localhost:9870 a skontrolujeme č isa nám vytvorili priečinky
2. dalej pridame do priečínka HDFS input náš textovy dokument input.txt

hadoop fs -put

'/home/martin4/Desktop/WordCountTutorial/input\_data/input.txt' /WordCountTutorial/Input

1. Zmenime cestu pomocou príkazu cd

cd /home/martin4/Desktop/WordCountTutorial

1. Skompilujeme java kód

javac -classpath ${HADOOP\_CLASSPATH} -d

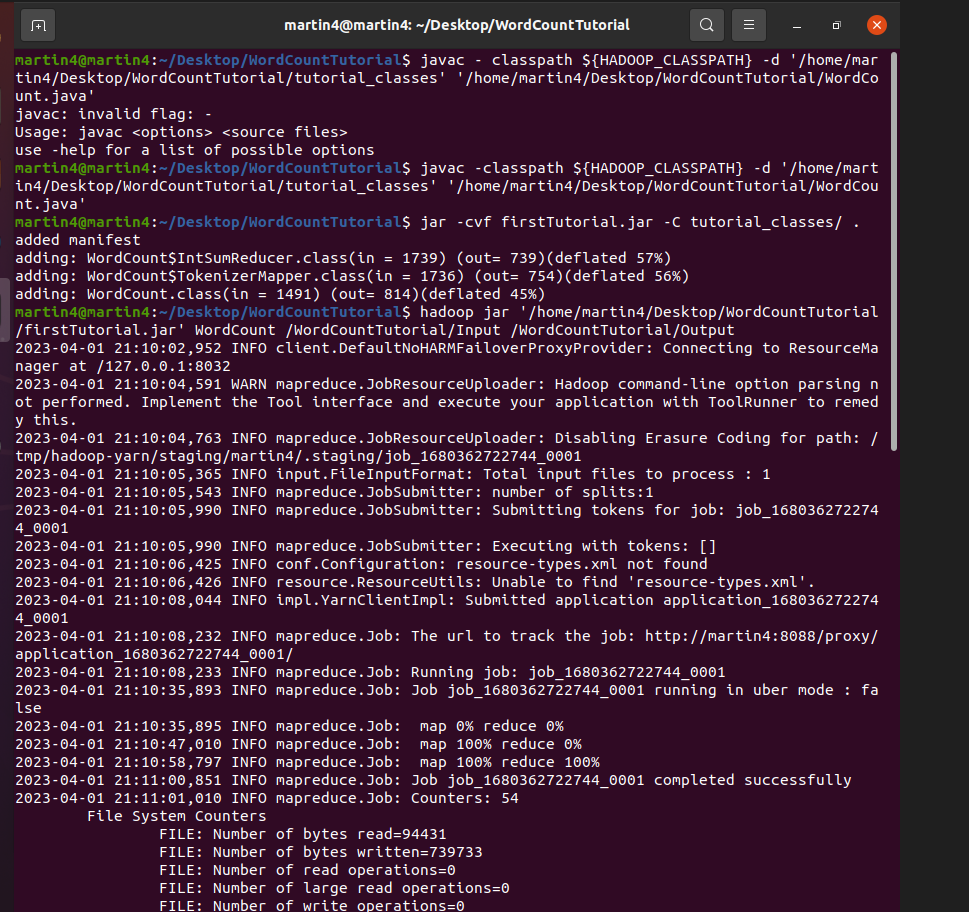
'/home/martin4/Desktop/WordCountTutorial/tutorial\_classes' '/home/martin4/Desktop/WordCountTutorial/WordCount.java'

1. Vystupne subory vložíme do jedného súboru

jar -cvf firstTutorial.jar -C tutorial\_classes/ .

1. Teraz spustime jar na hadoope

hadoop jar '/home/martin4/Desktop/WordCountTutorial/firstTutorial.jar' WordCount /WordCountTutorial/Input /WordCountTutorial/Output



1. Pre zobrazenie výstupu zadáme nasledujpci príkaz
2. hadoop dfs -cat /WordCountTutorial/Output/\*

