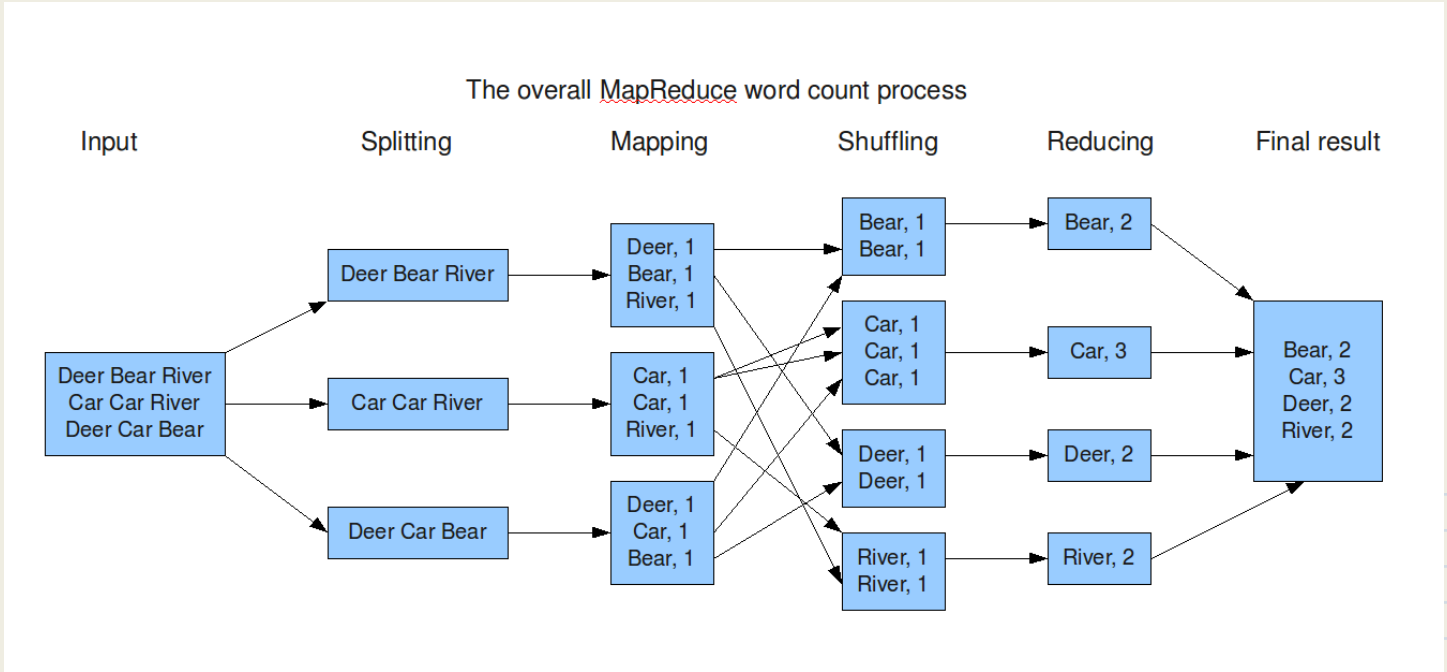
Spark Examples

Ing. Mario Alberto Giraldo
Estudiante Maestria en Ingenieria

Word-Count


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

```java
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().replaceAll("[^A-Za-z0-9]", " "));
                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);
        job.setNumReduceTasks(1);
        FileInputFormat.setMaxInputSplitSize(job, 20000000); //para config
        FileInputFormat.addInputPath(job, new Path(args[0]+"/*"));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

```scala
val f = sc.textFile(inputPath)
val w = f.flatMap(l => l.split(" ")).map(word => (word, 1)).cache()
w.reduceByKey(_ + _).saveAsText(outputPath)
```

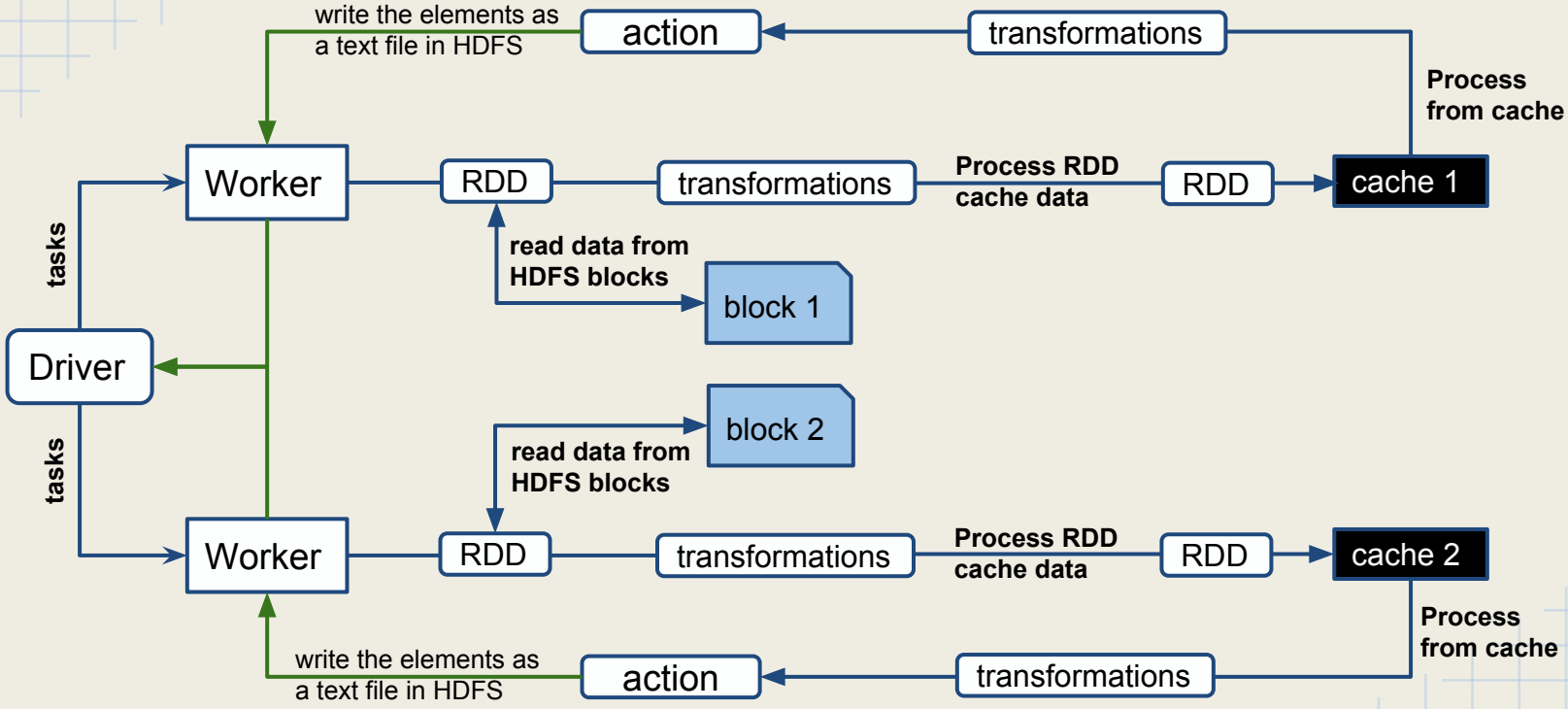
spark-submit --class com.cloudera.sparkwordcount.SparkWordCount --master yarn sparkwordcount-0.0.1-SNAPSHOT.jar hdfs://10.10.0.11/user/admin/liboswordcount

word count	SUCCEEDED	root	100%	100%	root.root	N/D	2m:33s
Spark Count	SUCCEEDED	root	100%	100%	root.root	N/D	50s

spark-submit --master local[*]
spark-submit --master yarn

MapReducer vs Spark

```
1  val f = sc.textFile(inputPath)
2  val w = f.flatMap(l => l.split(" ")).map(word => (word, 1)).cache()
3  w.reduceByKey(_ + _).saveAsText(outputPath)
```



Interactive shell

```
text_rdd = sc.textFile("hdfs://localhost:9000/user/maat/a3a.t")
text_rdd = sc.textFile("hdfs://localhost:9000/user/maat/a3a.t",6)

text_rdd.getNumPartitions()

text_rdd = sc.textFile("hdfs://localhost:9000/user/maat/a3a.t",6).cache()
ext_rdd.first
text_rdd.take(5)
text_rdd.count
```

Program run "Movie Rank" with shell pyspark

Spark UI:
pyspark = http://127.0.0.1:4040/jobs/
spark-shell=http://127.0.0.1:4041/jobs/

References

- Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia-Learning Spark_ Lightning-Fast Big Data Analysis-O'Reilly Media (2015)
- Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills-Advanced Analytics with Spark_ Patterns for Learning from Data at Scale-O'Reilly Media (2015)
- https://github.com/praveensripati/spark-examples
- https://www.youtube.com/watch?v=sTpzQdPnDmI - Video instalación Spark.