SoSe 2014: M-TANI: Big Data Analytics

Lecture 2 – 30/04/2014

Sead Izberovic
Dr. Nikolaos Korfiatis
What is MapReduce?

• “A programming model for large-scale distributed data processing” [1]

• Inspired by the map and reduce primitives from functional programming languages [2]

• Used by:
  - Google
  - Yahoo!
  - Facebook
  - …
Map in functional programming

• *map* takes following arguments:
  ▪ A function *func*
  ▪ A set of values *val*

• *map* returns the result of the computation as a set *results*

\[ map(func, val) \rightarrow results \]
Map example in Python

```python
>>> map(len, ['Hallo', 'Big', 'Data', '!'])
[5, 3, 4, 1]
```
Reduce in functional programming

• `reduce` takes following arguments:
  ▪ A binary function `bi_func`
  ▪ A set of values `red_val`

• `reduce` returns the value of the computation `result`

\[
reduce(bi\_func, red\_val) \rightarrow result
\]
Reduce example in Python

```python
>> reduce(+, [5,3,4,1])
13
```
MapReduce execution overview

Input Data

Split 1 → Map 1 → Reducer 1 → Output 1

Split 2 → Map 2

... → ... → ... → ...

Split M-1 → Map M-1 → Reducer R → Output R

Split M → Map M

Figure adapted from [2]
Map in MapReduce

• *map* takes following arguments:
  - A key $K$
  - A value $V$

• *map* creates a set of intermediate key-values pairs

\[
map(K,V) \rightarrow [(K_1,V_1),\ldots,(K_n,V_n)]
\]
Reduce in MapReduce

• *reduce* takes following arguments:
  - A key $K_{\text{red}}$
  - A set of values $[V]$

• *map* creates a set of intermediate key-values pairs

$$\text{map}(K_{\text{red}}, [V]) \rightarrow (K_{\text{red}}, V_{\text{red}})$$
MapReduce example

Word count

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Splits</th>
<th>Map phase</th>
<th>Reduce phase</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello Hello</td>
<td>Hello Hello</td>
<td>Hello : 2</td>
<td>Hello : 2</td>
<td>Hello : 5</td>
</tr>
<tr>
<td>World!</td>
<td>World!</td>
<td>World : 1</td>
<td>Hello : 2</td>
<td></td>
</tr>
<tr>
<td>Hello World,</td>
<td>Hello World,</td>
<td>Hello : 2</td>
<td>World : 1</td>
<td></td>
</tr>
<tr>
<td>Hello World!</td>
<td>Hello World!</td>
<td>World : 2</td>
<td>World : 1</td>
<td>World : 4</td>
</tr>
<tr>
<td>Hello World!</td>
<td>Hello World!</td>
<td>Hello : 1</td>
<td>World : 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>World : 1</td>
<td>World : 1</td>
<td></td>
</tr>
</tbody>
</table>

Figure adapted from [3]
Word count example in Hadoop

• Writing the Mapper

```java
public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
    private static final IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            output.collect(word, one);
        }
    }
}
```

source code from [4]
Word count example in Hadoop

• Writing the Reducer

```java
public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
        int sum = 0;
        while (values.hasNext()) {
            sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
    }
}
```

source code from [4]
Word count example in Hadoop

• Create the necessary directories

[cloudera@localhost ~]$ hadoop fs -mkdir /user/cloudera/wordcount

[cloudera@localhost ~]$ hadoop fs -mkdir /user/cloudera/wordcount/input

[cloudera@localhost ~]$ hadoop fs -mkdir /user/cloudera/wordcount/output

[cloudera@localhost ~]$ hadoop fs -mkdir /user/cloudera/lib
Word count example in Hadoop

• Copying a file from the local file system to the HDFS

   [cloudera@localhost ~]$ hadoop fs -put input_file target_directory_in_HDFS

• Copying a text file from the local file system to the HDFS

   [cloudera@localhost ~]$ hadoop -put ./the_tragedie_of_hamlet.txt /user/cloudera/wordcount/input/

• Copying a jar file from the local file system to the HDFS

   [cloudera@localhost ~]$ hadoop fs -put ./wordcount.jar /user/cloudera/lib/
Word count example in Hadoop

• Executing the word count example

```
[cloudera@localhost big_data]$ hadoop jar ./wordcount.jar org.myorg.WordCount /user/cloudera/wordcount/input/ /user/cloudera/wordcount/output/out
```

```
INFO mapreduce.Job: map 0% reduce 0%
INFO mapreduce.Job: map 100% reduce 0%
INFO mapreduce.Job: map 100% reduce 100%
```
Hue - Hadoop User Experience

- Web application for interacting with Hadoop
Word count example with Hue

• Creating the workflow
Word count example with Hue

• Creating the workflow
Word count example with Hue

• Creating the workflow
Word count example with Hue

• Creating the workflow
Word count example with Hue

• Creating the workflow

Edit Node: Word_count_example

- Name: Word_count_example
- Description: Counting words
- Action type: Java

Jar name: /user/cloudera/lib/wordcount.jar
Main class: org.myorg.WordCount

All the paths are relative to the deployment directory. They can be absolute but this is not recommended. You can parameterize values using case sensitive $[parameter]$. HDFS
Word count example with Hue

• Creating the workflow
Word count example with Hue

- Executing the word count example
Word count example with Hue

• Executing the word count example
Word count example with Hue

• The results
Literature

**MapReduce: The programming model and practice**, SIGMETRICS 
URL: http://research.google.com/archive/papers/mapreduce-sigmetrics09-tutorial.pdf

**MapReduce: simplified data processing on large clusters**. In *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6 (OSDI'04)*, Vol. 6. USENIX Association, Berkeley, CA, USA, 10-10

**An Inside Look at Google BigQuery**, White paper, 
URL: https://cloud.google.com/files/BigQueryTechnicalWP.pdf

4. Cloudera Inc. 
**Example: WordCount v1.0** 
URL: http://www.cloudera.com/content/cloudera-content/cloudera-docs/HadoopTutorial/CDH4/Hadoop-Tutorial/ht_wordcount1_source.html