Exercise 3:
Hadoop MapReduce (cont.)

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Concepts and Technologies for Distributed Systems and Big Data Processing – SS 2016

Task 1 Paper Reading

Read the Google File System paper by Ghemawat et al. [1]. You can find the paper at
`http://research.google.com/archive/gfs.html`

Answer the following questions:

a) How does a read work in the Google File System? What are the steps?

b) What are the criteria for positioning chunks when they are created, replicated and rebalanced?

Task 2 Reverse Graph

Complete the following code for ReverseGraph, which should reverse the direction of the edges in a directed graph. The input format is given in Figure 1a, where each line represents a pair which assigns the list of outgoing edges to the nodes in the graph. The expected output is given in Figure 1b. As you can see, for each edge $a \rightarrow b$ in the input there is a corresponding edge $b \rightarrow a$ in the output.

The code snippet already contains a regular expression to retrieve the list of numbers for each input line.

```
1  (3, [1, 2])
2  (1, [2, 3])
```

(a) Input

```
1  (1, [3])
2  (2, [1, 3])
3  (3, [1])
```

(b) Expected output.

Figure 1: ReverseGraph

```
1  public static class TokenizerMapper extends Mapper<LongWritable, Text, Text, Text> {
2    private Text from = new Text();
3    private Text to = new Text();
4
5    private Pattern pattern = Pattern.compile("(\\d+)");
6
7    @Override
8    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
9      Matcher m = pattern.matcher(value.toString());
10
11
12
13
14
15
16
17
18
19
20
21   }
22 }
23
```

```
24 public static class InvertedReducer extends Reducer<Text, Text, Text, Iterable<Text>> {
25   @Override
26   protected void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
27
28
29
30
31
32
33
34
35
36
37   }
38 }
```

Task 3 Relational Join

Complete the following code for RelationalJoin, which should join two tables based on the same numeric identifier. Figure 2a shows the input data for the tables `Department` and `Employee`. The output should indicate the assignments from employees to departments as shown in Figure 2b.

```
1 Department,1234,Sales
2 Employee,Susan,1234
3 Department,1233,Marketing
4 Employee,Joe,1233
5 Department,1233,Accounts
```

(a) Input

```
1 1233,Joe,Accounts
2 1233,Joe,Marketing
3 1234,Susan,Sales
```

(b) Expected output.

Figure 2: ReverseGraph

```
1  public static class JoinMapper extends Mapper<LongWritable, Text, Text, Text> {
2    @Override
3    protected void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
4
5
6
7
8
9
10
11
12
13
14   }
15 }
16
17 public static class JoinReducer extends Reducer<Text, Text, Text, Text> {
18   @Override
19   protected void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
20
21
22
23
24
25
26
27
28
29
30   }
31 }
```

References

[1] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, SOSP '03, pages 29–43, New York, NY, USA, 2003. ACM.