Stockholm University
DSV dept.
Academic year 2017/18

Big Data with NoSQL

Final report of the first assignment

Presented by:

- Giorgos Ntymenos
- Giacomo Bartoli

Part 1:

Data	**Step 0** - Store Data in HDFS	**Step 1** - Map	**Step 2** - Shuffle and Sort	**Step 3** - Reduce
● ★ ■ ▲ ■ ⬠ ◆ ★ ● ● ⬠ ▲ ▲ ▲ ■ ★ ● ★ ■ ★ ● ⬠ ▲ ● ■ ★ ■ ● ▲ ● ◆ ▲ ◆ ★ ◆ ■	● ★ ■ ▲ ■ ⬠ ◆ ★ ● ——— ● ⬠ ▲ ▲ ▲ ■ ★ ● ★ ——— ■ ★ ● ⬠ ▲ ● ■ ★ ■ ——— ● ▲ ● ◆ ▲ ◆ ★ ◆ ■ ———	■,1 ●,1 ⬠,1 ★,1 ▲,1 ●,1 ■,1 ◆,1 ★,1 ——— ■,1 ●,1 ⬠,1 ★,1 ▲,1 ●,1 ▲,1 ▲,1 ★,1 ——— ■,1 ●,1 ⬠,1 ★,1 ■,1 ●,1 ■,1 ▲,1 ★,1 ——— ■,1 ●,1 ▲,1 ★,1 ◆,1 ●,1 ◆,1 ▲,1 ◆,1 ———	■,1 ■,1 ■,1 ■,1 ■,1 ■,1 ■,1 ●,1 ●,1 ●,1 ●,1 ●,1 ●,1 ●,1 ●,1 ——— ⬠,1 ⬠,1 ⬠,1 ★,1 ★,1 ★,1 ★,1 ★,1 ★,1 ★,1 ——— ▲,1 ▲,1 ▲,1 ▲,1 ▲,1 ▲,1 ▲,1 ◆,1 ◆,1 ◆,1 ◆,1 ———	■,7 ——— ●,8 ——— ⬠,3 ——— ★,7 ——— ▲,7 ——— ◆,4 ———

Part 2:

After running Map Reduce given that dataset, we got the following results:
<shape, frequency>

```
circle, 6973
diamond, 10153
pentagon, 25145
square, 12966
star, 4963
triangle, 39800
```

The code we added for implementing Mapper.py and Reducer.py is the following:

Mapper.py line 12:
```
#12    print "%s\t%s" % (word,1)
```

Reducer.py lines 17,34,42:
```
#17    word,count = line.split('\t',1)
#34    print '%s\t%s' % (current_word,current_count)
#42    print '%s\t%s' % (current_word,current_count)
```