Hadoop streaming

Run hadoop streaming job:

1. copy scripts mapper.py and reducer.py to hadoop node

2. log into that haddop node

3. run job:

   ```
   hadoop jar <path_to_hadoop_streaming_jar> -files <list_of_files_with_mapper_and_reducer> \
     -mapper <file_with_mapper> -reducer <file_with_reducer> -input <input_dir> -output <output_dir>

   hadoop jar /opt/mapr/hadoop/hadoop-2.7.0/share/hadoop/tools/lib/hadoop-streaming-2.7.0-mapr-1808.jar -files mapper.py,reducer.py \
     -mapper mapper.py -reducer reducer.py -input /user/<username>/loremipsum -output /user/<username>/outputs/output
   ```

   Attention: output_dir must not exist!

Tasks

1. count letters in loremipsum

2. sort counted letters by occurrences. Why is it harder than with java api?

3. Count how many incoming transfers were there for each account.

4. find number of unique accounts

Extra task

5. Write mapper and reducer in other technology than python and java, and use it in hadoop-streaming job.

Useful parameters

Enable compression:

```
-D mapreduce.output.fileoutputformat.compress=true \
  -D mapreduce.output.fileoutputformat.compress.codec=org.apache.hadoop.io.compress.GzipCodec
```

Use different input format:

```
-inputformat org.apache.hadoop.mapred.SequenceFileInputFormat
```

Użycie different output:

```
-outputformat org.apache.hadoop.mapred.SequenceFileOutputFormat
```

Identity mapper:

```
-mapper org.apache.hadoop.mapred.lib.IdentityMapper
-mapper /bin/cat
```

Documentation:
`https://hadoop.apache.org/docs/r1.2.1/streaming.html`