Exercise Sheet 7

Exercise 1 (MapReduce/Hadoop)

1. What is Hadoop?
2. Describe the functioning of the MapReduce programming model.
3. Explain (*in just a few sentences*) two examples, where MapReduce is helpful.
4. Describe the working method of the Google PageRank algorithm.
5. Name an advantage of the 64 MB chunk size of the Hadoop Distributed File System (HDFS)?
6. Name a drawback of the 64 MB chunk size of the HDFS?
7. What kind of data stores the Namenode?
8. What kind of data store the Datanodes?
9. What is Pig?
10. What is Pig Latin?
11. What is Hive?
12. What is HBase?
13. What is Cloudera?

Exercise 2 (PageRank)

In slide set 7 we discussed a page rank example for a network of 3 linked documents (web pages). Invent an example scenatio of a network of 5 linked documents. The network should contain at least 11 links.

Calculate the first 10 iterations of the PageRank algorithm for your example scenatio.

Exercise 3 (Hadoop Cluster)

1. Launch a Hadoop Cluster in an infrastructure service like EC2, Google Compute Engine, or alternatively on your personal computer.

2. Execute the π calculation example, which has been discussed in slide set 7.
3. Find a useful use case for your Hadoop cluster and try it out.
4. Present your use case during the exercise session.

Write down precise instructions with the steps you performed and demonstrate your solution live during the exercise session.