Solution of Exercise Sheet 7

Exercise 1 (MapReduce/Hadoop)

1. What is Hadoop?

   Hadoop is a MapReduce framework for parallel data processing in Clusters. It operates according to the MapReduce programming model.

2. Describe the functioning of the MapReduce programming model.

   The MapReduce programming model splits tasks into smaller parts and distributes them for parallel processing to different compute nodes. The final result is created by merging the partial results. Steps of MapReduce are:

   - *Partitioning of the initial data.*
   - *Mapping (map) the data to a data structure which consists of a key-value pair.*
   - *Grouping the key-value pairs of identical keys.*
   - *Distributing (shuffle) and sorting (sort) the key-value pairs.*
   - *Reducing (reduce) the key-value pairs to obtain the result.*

3. Explain (*in just a few sentences*) two examples, where MapReduce is helpful.

   Some examples are:

   - *Distributed frequency analysis. This counts how many times words exist in a long text.*
   - *Distributed grep. This greps the lines of text that contain a search pattern.*
   - *Calculation of website requests by analyzing web access log data.*
   - *PageRank algorithm for calculating the importance of a web page in the internet.*

4. Describe the working method of the Google PageRank algorithm.

   It rates linked documents (web pages). The working principle is that the numerical weight (PageRank) PR_p of a web page p depends of the number and the numerical weight of the web pages, which refer to p.

5. Name an advantage of the 64 MB chunk size of the Hadoop Distributed File System (HDFS)?

Lesser network overhead.

6. Name a drawback of the 64 MB chunk size of the HDFS?

   More internal fragmentation, compared with file systems with smaller chunks.

7. What kind of data stores the Namenode?

   It stores only the metadata. It knows all files and directories, which exist in the HDFS Cluster and stores the numbers of chunks of the files, the number of copies and their locations (Datanodes).

8. What kind of data store the Datanodes?

   They store the user data.

9. What is Pig?

   It is an extension for Hadoop which can be used for the analysis of very large amounts of semi-structured, structured or relational data. Pig includes a programming language and a compiler for queries on data.

10. What is Pig Latin?

    It is the programming language of Pig. It is used to specify sequences of individual transformations on data.

11. What is Hive?

    It is a data warehouse system on the basis of Hadoop.

12. What is HBase?

    It is a column-oriented database to manage very large amounts of data in Hadoop Clusters.

13. What is Cloudera?

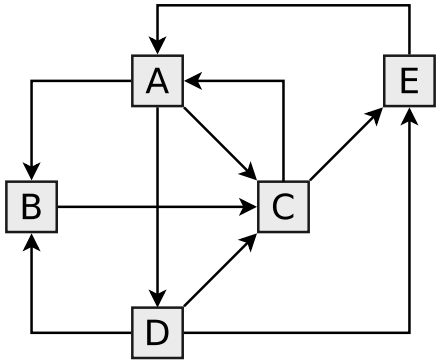
    Cloudera offers a Hadoop distribution which is easy to deploy and use.

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.

$$PR(p) = (1-d) + d * \sum_{p_i \in L_{IN}(p)} \frac{PR(p_i)}{\text{amount } L_{OUT}(p_i)}$$



We consider for this example $d = 0.5$

- $PR_{n+1}(A) = 0.5 + 0.5 * (\frac{PR(C)}{2} + PR(E))$
- $PR_{n+1}(B) = 0.5 + 0.5 * (\frac{PR(A)}{3} + \frac{PR(D)}{3})$
- $PR_{n+1}(C) = 0.5 + 0.5 * (\frac{PR(A)}{3} + PR(B) + \frac{PR(D)}{3})$
- $PR_{n+1}(D) = 0.5 + 0.5 * \frac{PR(A)}{3}$
- $PR_{n+1}(E) = 0.5 + 0.5 * (\frac{PR(C)}{3} + \frac{PR(D)}{3})$

	0	1	2	3	4	5
A	1	1,25000	1,29167	1,28125	1,26563	1,26461
B	1	0,83333	0,81944	0,82176	0,81964	0,81710
C	1	1,33333	1,23611	1,23148	1,23052	1,22692
D	1	0,66667	0,63889	0,63657	0,63696	0,63661
E	1	0,91667	0,94444	0,91551	0,91397	0,91379

	6	7	8	9	10	PR
A	1,26362	1,26277	1,26246	1,26235	1,26228	1,26225
B	0,81687	0,81663	0,81649	0,81643	0,81640	0,81639
C	1,22542	1,22507	1,22480	1,22467	1,22462	1,22459
D	0,63618	0,63614	0,63611	0,63608	0,63607	0,63607
E	0,91283	0,91239	0,91229	0,91222	0,91218	0,91217

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.

TBD