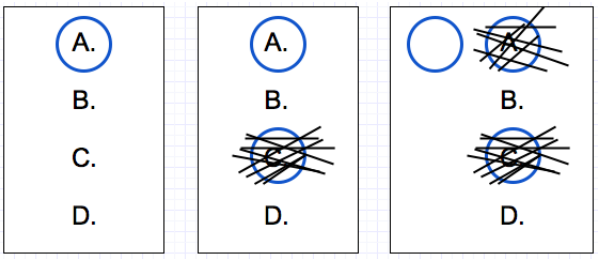
Big Data Processing
2013-2014 Q2
April 7, 2014 (Resit)
Lecturer: Claudia Hauff
Time Limit: 180 Minutes

Name: _____

Student number: _____

Answer the questions in the spaces provided on this exam. If you run out of room for an answer, continue on the back of the page. Mark clearly which question the answer refers to.

- Before you start, write down your name and student number on this page. On all the following pages, write your student number at the top of the page.
- This exam contains 13 pages (including this cover page) and 18 questions. Check to see if any pages are missing.
- The use of material (book, slides, laptop, etc.) during the exam is not allowed.
- The amount of points each question is worth is indicated.
- Write clearly! If your writing cannot be deciphered, it will not be considered for grading.
- Every multiple-choice question has just **one** correct answer. To select an answer, circle the letter.
  In the examples below, answer A. is considered for grading in all cases.



- The total number of available points is 43.

Good luck!

Multiple-choice questions.

1. (1 point) **[Hadoop]** Which of the following statements is correct?
   In Hadoop there exists ...
   A. one JobTracker per Hadoop job
   B. one JobTracker per Mapper
   C. one JobTracker per node
   D. one JobTracker per cluster

2. (1 point) **[Hadoop]** Bob has a Hadoop cluster with 50 machines under default setup (replication factor 3, 128MB input split size). Each machine has 100GB of HDFS disk space. The cluster is currently empty (no job, no data). Bob intends to upload 1 Terabyte of plain text (in 5 files of approximately 200GB each), followed by running Hadoop's standard `WordCount`[1] job. What is going to happen?
   A. The data upload fails at the first file: it is too large to fit onto a node
   B. The data upload fails at the last file: due to replication, all disks are full
   C. `WordCount` fails: too many input splits to process
   D. `WordCount` runs successfully

3. (1 point) **[HDFS]** How does HDFS ensure the integrity of the stored data?
   A. through checksums
   B. by comparing the replicated data blocks with each other (majority vote)
   C. through error logs
   D. by comparing the replicated blocks to the master copy

4. (1 point) **[General]** Large graphs (with billions of nodes and edges) are typically stored in files formatted as
   A. adjacency matrix
   B. adjacency list
   C. adjacency pairs
   D. adjacency chart

5. (1 point) **[Hadoop]** When the primary NameNode crashes, the secondary NameNode takes over. NameNodes do not persistently store (i.e. write to disk) the location of blocks. How does the secondary NameNode learn about the blocks' locations in the cluster?

[1] WordCount is a simple Hadoop job which counts the number of occurences of each word in a given input.

A. DataNodes send regular heartbeat messages, which include information about blocks they maintain

B. The secondary NameNode sends a special message to all DataNodes, asking for their block information

C. Before a crash, the primary NameNode always copies its memory content to the secondary NameNode

D. The secondary NameNode replays the edit log, which contains the blocks' locations

6. (1 point) **[Hadoop]** The time it takes for a Hadoop job's Map task to finish mostly depends on

   A. the placement of the blocks required for the Map task

   B. the duration of the job's shuffle & sort phase

   C. the placement of the NameNode in the cluster

   D. the duration of the job's Reduce task

7. (1 point) **[Hadoop]** Hadoop's job scheduling can be *speculative*: if a task has not yet finished, an identical copy of the task can be executed on a second DataNode. The output of whichever node finishes first is used. Speculative execution is possible both for map and reduce tasks. In practice, speculative execution is mostly restricted to map tasks. What is the main reason?

   A. Map tasks are more prone to stragglers (slowly executed tasks), since DataNodes always execute queued Reduce tasks before any Map tasks.

   B. Speculative execution of reduce tasks leads to erroneous values in Counters at the end of a job.

   C. Speculative execution of reduce tasks can create a large amount of additional network traffic.

   D. All of the above.

8. (1 point) **[HBase]** In HBase, table cells are indexed by

   A. row key + column key + timestamp

   B. row key + timestamp

   C. row key

   D. None of the above

Free-form questions.

9. (2 points) **[Sampling]** What is the goal of *Reservoir Sampling*? What is its advantage over *Min-wise Sampling?*

10. (2 points) **[General]** In light of the big data technologies presented in this course, briefly explain the meaning of the statement: *Move code to data!*

11. (3 points) **[General]** The *Internet Archive* is a digital archive that makes snapshots (a local copy) of a huge number of Web pages, archiving the Internet for future generations. It currently has collected more than 400 billion snapshots of Web pages; very dynamic pages such as the front page of the New York Times website or the Volkskrant website are crawled several times per day. Other Web pages are crawled only once a year.

    Assume you are tasked by the Internet Archive to provide the following services for their data:

    1. All snapshots crawled within the past 30 days should be directly accessible to users: it should take merely seconds between a user submitting a URL and the Internet Archive returning the list of snapshots taken during the past 30 days.
    2. Requesting older snapshots requires more patience: a user can submit a list of URLs and a time frame of interest (e.g. `http://www.nytimes.com/index.`

`html` between March 1, 2009 and March 15, 2009) and within a few hours or at most days the Internet Archive should return the requested snapshots.

3. On the website of the Internet Archive live statistics should be shown, indicating the number of URL requests issued by users today and the past month, the number of snapshots crawled today and the number of older snapshot requests currently being processed.

Given your knowledge of big data (and small-data) technologies, discuss which technologies you would use to provide these three services.

12. (2 points) **[General]** Enterprise routers are the backbone of large-scale computer networks. They can forward millions of data packets per second based on the destination address of each packet. You are asked to design an algorithm that counts with *reasonable accuracy* and a *low memory footprint* the *unique* number of destination addresses of data packets passing through a router.

13. **[Hadoop]** K-means clustering

Given a number of items (e.g. points in a 2D space), the goal of the K-means clustering algorithm is to assign each item to one of k clusters (the number k is fixed in advance).
Below you find the pseudo-code of K-means as well as a visualization of a toy example:

```
1  Input: items to be clustered, number k (#clusters)
2  Output: cluster label of each item

3  Initialise:
4    - Pick k items randomly (the initial cluster centroids)
5    - For each item:
6       - Compute distance to all centroids
7       - Assign item to the cluster with minimum distance

8  Repeat until no more label changes or 1000 iterations reached:
9    - Re-compute cluster centroids (the mean of assigned items)
10   - For each item:
11      - Compute distance to all centroids
12      - Assign item to the cluster with minimum distance
```

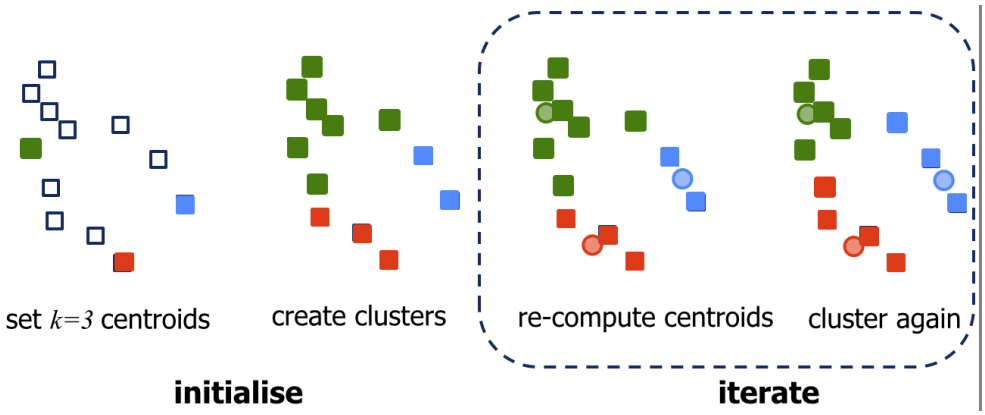


Figure 1: K-means example with $k = 3$ and 12 items to cluster.

All questions refer to the pseudo-code given above.

(a) ($\frac{1}{2}$ point) To implement the K-means algorithm in Hadoop, several jobs need to be run. What is the maximum number of jobs that need to be run?

(b) ($\frac{1}{2}$ point) Indicate which lines of the pseudo-code contain the work of a single job (give the line numbers).

(c) (2 points) List all information that needs to be maintained between subsequent jobs.

(d) (1 point) How can the maintained information be made accessible to the subsequent job?

(e) (1 point) What is the Mapper's input and output in each job?

(f) (1 point) What is the Reducer's input and output in each job?

(g) (1 point) Write out **in pseudo-code** the steps taken in Hadoop's map and reduce phases *for a single job*. If your approach requires Counters, Partitioners

or Combiners indicate this as well.

(h) (1 point) What is the main disadvantage of implementing K-means in plain Hadoop?

(i) (3 points) Lets now consider the implementation of K-means via the BSP model of computation (e.g. Pregel or Giraph). Describe the data contained in the vertices, the location computation, the messages sent/received and the synchronization step.

14. (3 points) **[Hadoop]** Assume you have a cluster of machines with disks mounted to them. All files throughout the cluster have a unique path. This setup is sufficient to run Hadoop on the cluster, HDFS does not need to be installed. Discuss the advantages and/or disadvantages of running Hadoop without HDFS in such a setting.

15. (2 points) **[MapReduce]** *MapReduce is a batch query processor*. Explain this statement with respect to processing time and data usage.

16. (2 points) **[Hadoop]** In Hadoop, the default Partitioner has the following `getPartition()` method:

```
public int getPartition(K key, V value, int numReduceTasks) {
   return (key.hashCode() & Integer.MAX_VALUE) % numReduceTasks;
 }
```

Discuss the suitability of the Partitioner when running Hadoop's standard `WordCount` implementation (with 25 available reduce tasks) on texts (a) to (d). Each text is a DNA sequence in plain text format[2].

[2] An element of a DNA sequence (a nucleobase) can either be an **A**, **C**, **G** or **T**. No other elements are possible.

(a)

```
ACAAGATGCC ATTGTCCCCC GGCCTCCTGC TGCT
CTCCGGGGCC ACGGCCACCG CTGCCCTGCC CCT
 ...
```

(b)

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCT
GCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTG
...
```

(c)

```
A G C A T A T G C A G G A A G C G G C A G G A A
T A A G G A A A A G C A G C C T C C T G A C T T
...
```

(d)

```
AA GG CC AA TT A TT A TT GG G C C AA TT G C
GG AA CC AA TT T T T AA GG AA AA CC A TT TT
...
```

17. **[Hadoop]** On the next page, a data set is described. For both queries (a) and (b), write out **in pseudo-code** the steps taken in Hadoop's map and reduce phases to answer these queries. If your approach requires Counters, Partitioners or Combiners indicate this as well.

    A researcher has crawled a large part of the Web (billions of Web pages) and stored on HDFS. He has written a Hadoop job which computes for a number of search requests (millions of search requests by millions of users retrieved from a log of search requests) and each Web page how well the page answers the search request. A larger score indicates a better fit of the Web page to the search request. The scores then determine the ranking of the Web pages (the highest score has rank 1) for each search request. For each search request, the top ranked 1 million Web pages are written to result files. These files have the following format per line:

1. search request ID
2. Web page ID (ranked once per search request)
3. rank of the Web page for the search request
4. score of the Web page for the search request
5. algorithm used for the computation (a single alg. per search request)
6. user who had the search request (a single user per search request)
7. time user originally had the search request in the format day-month-year;hours-minutes-seconds; (a single timestamp per search request)

An example:

```
1 clueweb09-en0000-58-29364 1 5.72600747826061 algA user1 12-3-2009;05:03:55
1 clueweb09-en0007-36-22844 2 5.23147899253392 algA user1 12-3-2009;05:03:55
2 clueweb09-en0000-58-29364 1 5.10239837552381 algA user1 12-3-2009;05:04:29
4 clueweb09-en0011-29-06509 1 1.44401629962568 algF user83 3-5-2009;07:25:18
2 clueweb09-en0008-41-27093 2 3.82252405092304 algB user1 12-3-2009;05:04:29
2 clueweb09-en0007-82-00416 3 3.80670840271686 algB user1 12-3-2009;05:04:29
3 clueweb09-en0009-50-40915 2 4.78512025979586 algC user12 25-4-2007;15:44:06
3 clueweb09-en0005-62-36831 3 4.53200534510577 algC user12 25-4-2007;15:44:06
4 clueweb09-en0006-42-01715 2 1.36352294296361 algF user83 3-5-2009;07:25:18
3 clueweb09-en0009-19-00252 1 4.79349119865143algC user12 25-4-2007;15:44:06
1 clueweb09-en0007-82-00416 3 4.96730544492365 algA user1 12-3-2009:05:03:55
.....
```

Here, for search request 1 (by user `user1`) algorithm `algA` was used. Web page `clueweb09-en0000-58-29364` has the highest score and is thus at rank position 1.

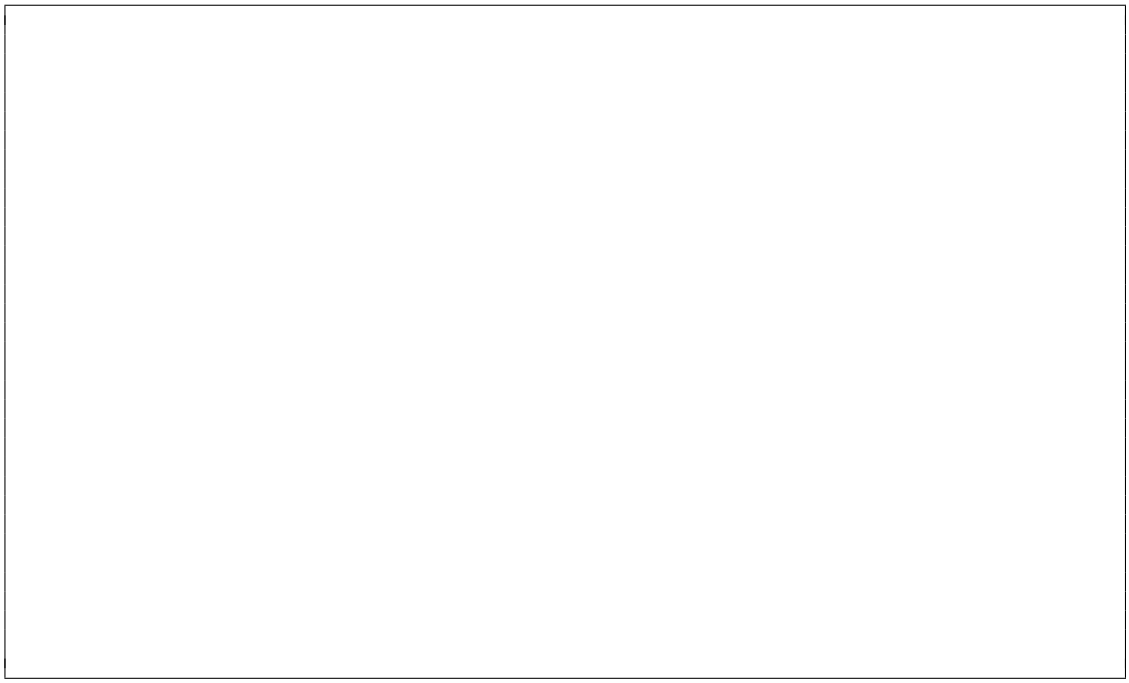
(a) (2 points) For each Web page p appearing in the result rankings of more than 100 search requests, compute the minimum and maximum rank of p achieved.

(b) (2 points) Compute the number of distinct users for whom algorithm `algF` was employed.

18. **[Pig]** This question also considers the data set described in question 17. Now you are asked to write **Pig** scripts to answer the queries (a) and (b).

(a) (2 points) Retrieve for each month (i.e. January to December, across all years) the absolute number of search requests issued.

(b) (2 points) Output the number of distinct users who issued a single search request only.



THE END.