

Big Data Processing
2014-2015 Q2
January 28, 2015
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 14 pages (including this cover page) and 23 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.
- When you are asked to write pseudo-code, **do not forget** to also add specific configurations that your code might require to run correctly (Partitioner or Reducer setting, etc.).
- The total number of available points is 46.

Good luck!

Multiple-choice questions.

1. (1 point) **[Filtering]** A Bloom filter guarantees no
 - A. false positives
 - B. false negatives
 - C. false positives and false negatives
 - D. false positives or false negatives, depending on the Bloom filter type

2. (1 point) **[Hadoop]** Bob has a Hadoop cluster with 20 machines under default setup (replication 3, 128MB input split size). Each machine has 500GB of HDFS disk space. The cluster is currently empty (no job, no data). Bob intends to upload 5 Terabyte of plain text (in 10 files of approximately 500GB each), followed by running Hadoop's standard WordCount¹ job. What is going to happen?
 - A. The data upload fails at the first file: it is too large to fit onto a DataNode
 - B. The data upload fails at a later stage: the disks are full
 - C. WordCount fails: too many input splits to process
 - D. WordCount runs successfully

3. (1 point) **[Streaming]** The following stream is processed using the DGIM algorithm: 1 0 1 1 1 0 0 0 0 1 1 0 1 0 1 0 1 (the most recent bit appears to the right). Applying DGIM to this stream results in:
 - A.

1	0	1	1	1	0	0	0	0	1	1
---	---	---	---	---	---	---	---	---	---	---

0	1	0	1
---	---	---	---

0	1
---	---
 - B.

1	0	1	1	1
---	---	---	---	---

0	0	0	0
---	---	---	---

1	1	0
---	---	---

1	0	1	0
---	---	---	---

1

 - C.

1	0	1	1	1
---	---	---	---	---

0	0	0	0
---	---	---	---

1	1
---	---

0

1	0	1
---	---	---

0

1

 - D. None of the above

4. (1 point) **[Streaming]** What are DGIM's maximum error boundaries?
 - A. DGIM always underestimates the true count; at most by 25%
 - B. DGIM either underestimates or overestimates the true count; at most by 50%
 - C. DGIM always overestimates the count; at most by 50%
 - D. DGIM either underestimates or overestimates the true count; at most by 25%

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

-
5. (1 point) **[Streaming]** Which algorithm should be used to approximate the number of distinct elements in a data stream?
- A. Misra-Gries
 - B. Alon-Matias-Szegedy
 - C. Flajolet-Martin
 - D. None of the above are algorithms that approximate the number of distinct elements
6. (1 point) **[Pig]** Which Hadoop guarantee does Pig break?
- A. A mapper can output an arbitrary number of key/value pairs.
 - B. A combiner's input and output types have to match.
 - C. A reducer receives all values corresponding to a key.
 - D. A custom partitioner can be implemented.
7. (1 point) **[HBase]** An HBase table may contain many different kinds of data elements, with varying access patterns and size characteristics. Through which mechanism can be ensured that data with similar access patterns is semantically grouped together?
- A. Record family
 - B. Column family
 - C. Table family
 - D. Row family
8. (1 point) **[Hadoop]** In Hadoop, the optimal input split size is the same as the
- A. block size
 - B. average file size in the cluster
 - C. minimum hard disk size in the cluster
 - D. number of DataNodes
9. (1 point) **[ZooKeeper]** Using the ZooKeeper API, the following call is made:
`create(/my_app/p, data, SEQUENTIAL | EPHEMERAL).`
What is (i) the name of the created znode and (ii) when will it be destroyed?
- A. (i) my_app1/p, (ii) at the end of the client's session
 - B. (i) my_app1/p, (ii) after a failure of the client machine
 - C. (i) my_app/p1, (ii) after a failure of the client machine
 - D. (i) my_app/p1, (ii) at the end of the client's session

10. (1 point) **[Pig]** Assume you want to join two datasets within a Pig script:
Data set 1 consists of all Wikipedia edits captured for all languages in one log file; one line contains the fields [Unique ID, Wikipedia URL, Edit Timestamp, Editing UserID]. The lines are unordered.
Data set 2 consists of information about Wikipedia articles written in Frisian (less than 70,000 articles overall): [Unique ID, Wikipedia URL, Wikipedia Title]. The lines are ordered alphanumerically by the Unique ID field.
The join should be performed on Wikipedia URL and the generated data set should look as follows: [Edit Timestamp, Wikipedia URL, Wikipedia Title].
Which join is the most efficient one to use assuming a Hadoop cluster with 15 machines, each one with about 4GB of memory and 1TB of disk space?
- A. sort-merge join
 - B. skew join
 - C. fragment-replicate join
 - D. default join
11. (1 point) **[Hadoop]** Which of the following statements is correct about Heartbeat messages in a Hadoop cluster?
- A. A group of DataNodes together send a single heartbeat message to save network bandwidth.
 - B. A heartbeat message is sent at most once a day by each active DataNode.
 - C. A heartbeat message is sent every 5 to 30 seconds by every active DataNode.
 - D. None of the above.
12. (1 point) **[Hadoop]** Which two elements in a standard Hadoop cluster form the main bottlenecks when the cluster grows in size (i.e. more machines are added to the cluster)?
- A. NameNode & JobTracker
 - B. DataNode & TaskTracker
 - C. ChunkServer & JobTracker
 - D. TaskTracker & JobTracker
13. (1 point) **[HDFS]** Which of the following operations requires NO communication with the NameNode?
- A. A client deleting a file from HDFS.
 - B. A client reading a block of data from HDFS.
 - C. A client requesting the filename of a given block of data.
 - D. None of the above.

18. (2 points) **[Hadoop]** In the shuffle & sort phase, a job with m mappers and r reducers may involve up to $m \times r$ distinct copy operations. In which scenario are exactly $m \times r$ copy operations necessary?

19. (3 points) **[Design Patterns]** Consider the Hadoop pseudo code shown below. It computes for each key, the *log average* across all values associated with that key. The formula for the log average is: $\text{logAvg} = \exp\left(\frac{1}{n} \sum_{i=1}^n \log a_i\right)$.

```
map(string t, int r):
    EmitIntermediate(string t, int r)

reduce(string t, ints [r1, r2, r3, ..])
    float sum = 0.0;
    int count = 0;

    foreach int r in ints {
        sum += log(r);
        count++;
    }
    float logAvg = exp( 1/(float)count * sum );
    Emit(string t, float logAvg);
```

Rewrite the pseudo-code and **include a Combiner** to speed up the job. Make sure that both versions of the code (i.e. with and without the Combiner) achieve the same final result.



20. (4 points) **[Design Patterns]** You are given two large-scale datasets containing information about users and their actions on the EdX platform - a platform where massive open online courses (MOOCs) are offered to anyone interested to join. The first dataset *Users* contains the profiles of all users having registered on the platform. A number of example records are shown in Table 1 (ID is the primary key).

ID	FirstName	LastName	Email	Nationality	Age
4332	John	Walters	john@walters.com	US	27
4499	Klara	Sommer	klaras@yahoo.de	German	39
..

Table 1: Example records for dataset *Users*. The header (top row) is not included in the *Users* file.

The second dataset (*Log*) is a log of user actions recorded by the EdX platform. A number of example records are shown in Table 2. The ID uniquely identifies the user performing the action. The *Users* dataset contains less than 10 million records, the *Log* dataset contains several billion records. In extreme cases, for a single user millions of actions may have been recorded (e.g. a bot or a registered Web crawler). Due to the large-scale nature of these files, the data does not reside

ID	Activity	Target	Duration	Timestamp
4332	C	E4322		3432433224
4332	Q	E9800	12	3432433299
4499	C	V00233		3432435933
4332	Q	E345	78	3432430775
4499	P	P343	2390	3432434349
...

Table 2: Example rows for dataset *Log*. The header (top row) is not included in the *Log* file

in a database but on a Hadoop cluster; the files are stored as plain text files on HDFS in the format of one record per line.

You are asked to **join the two datasets**, based on the ID column. The final result is a single dataset which should have the format shown in Table 3.

ID	FirstName	LastName	Email	Nationality	Age	Activity	Target	Duration	Timestamp
4332	John	Walters	john@walters.com	US	27	C	E4322		3432433224
4332	John	Walters	john@walters.com	US	27	Q	E9800	12	3432433299
4499	Klara	Sommer	klaras@yahoo.de	German	39	C	V00233		3432435933
4332	John	Walters	john@walters.com	US	27	Q	E345	78	3432430775
4499	Klara	Sommer	klaras@yahoo.de	German	39	P	P343	2390	3432434349
...

Table 3: Example output. The header (top row) should not be included in the output.

How would you implement this task within a single Hadoop job? Write your answer down in **pseudo-code**.

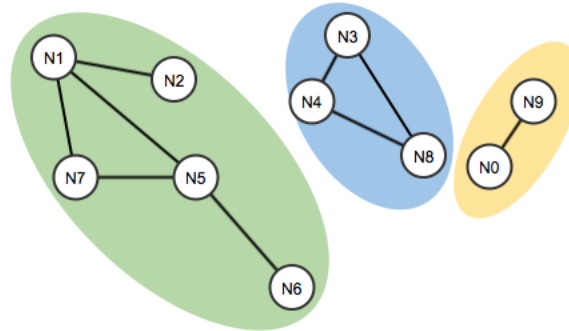


Figure 1: Connected components example

21. **[Giraph]** Twitter has released a large-scale “friendship” graph: in this graph, every Twitter user is a vertex and two vertices are connected via an undirected edge if the two respective users have exchanged private messages in the past. The graph contains several hundred million vertices and billions of edges.

One important property of large graphs is the number of connected components and their respective sizes. A connected component is a set of vertices that are linked to each other by paths (and no other vertices are contained in the set). As a concrete example, the undirected graph shown in Figure 1 has 10 vertices and 9 edges. The graph consists of three connected components.

- (a) (1 point) Name two advantages of the BSP-inspired Giraph framework over “plain” Hadoop when implementing graph algorithms.

- (b) (4 points) How can the connected components of a graph be computed in Giraph (or Pregel)? Explain superstep by superstep the local computation(s), messages send and votes to halt. You can assume that the graph is encoded in adjacency list format, i.e. each vertex “knows” with which other vertices it

- (c) (1 point) For the graph shown in Figure 1, draw the BSP-principle inspired diagram according to your devised algorithm. Indicate the supersteps, messages passed and votes to halt.



22. **[Hadoop]** You have developed your own search engine and implemented a number of algorithms to bring the best search results to your users. Every time a user clicks on a search result, you log that event in the following format:


1. search request the user submitted
2. URL clicked by the user (a URL from the result ranking)
3. Rank of the URL that was clicked in the result ranking
4. algorithm used to compute the result ranking
5. user who issued the search request and clicked a result
6. time of the click in the format day-month-year;hours-minutes-seconds

An example log is shown below:

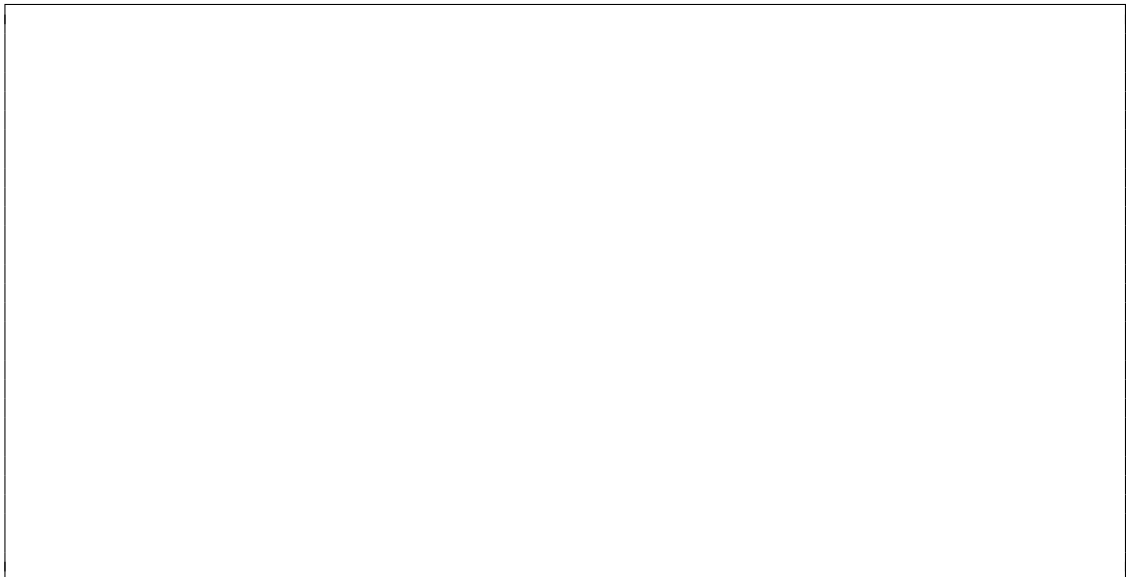
```
jaguar http://www.jaguar.com/login 1 algA user1 12-3-2009;05:03:55
jaguar http://autoscout24.nl/jag 2 algA user1 12-3-2009;05:04:07
lottery ticket http://www.imdb.com/title/tt0979434/ 1 algB user243 12-3-2009;05:04:29
buy a car http://www.autotrader.com/index.html 7 algF user83 3-5-2009;07:25:18
.....
```

In the example log, when user1 issued the query “jaguar”, algorithm A was used to determine a ranking of search results and the user clicked first on the search result at rank 1 (line 1) and 12 seconds later on the search result shown at rank 2 (line 2). Note that you cannot assume any particular ordering of the lines in the log file.

- (a) (2 points) For every search request issued by more than 1000 unique users, output the ten most clicked URLs.



- (b) (2 points) Compute the number of distinct users that have never clicked twice on the same URL, regardless of the issued search request.



23. **[Pig]** This question also considers the data set described in question 22. Now you are asked to write **Pig** scripts to answer the following two questions.

(a) (2 points) For each URL and year retrieve the number of unique search requests that led to a click (if that number is above zero). The output should look similar to this:

```
http://www.autotrader.com/ 2009 12
http://www.autotrader.com/ 2013 4555
http://www.imdb.com/title/tt0979434/ 2010 244
http://www.imdb.com/title/tt0979434/ 2013 4343
.....
```

(b) (2 points) For each user with more than 100 days of activity (does not have to be consecutive), determine the lowest rank he ever clicked a result on.

THE END.