Big Data Processing
2015-2016 Q2
January 29, 2016
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 17 pages (including this cover page) and 21 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-answer question** has one **or more** correct answers.
- 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 0.

Multiple-answer questions

The following questions may have multiple correct answers. Every question has at least one correct answer. You receive the point if you **check** ✓ all correct answers and only those.

1. (1 point) **[Hadoop]** Which of the following operations require the client to communicate with the NameNode?

   ☐ A client deleting a file from HDFS.

   ☐ A client writing to a new file on HDFS.

   ☐ A client appending data to the end of an existing file on HDFS.

   ☐ A client reading a file from HDFS.

2. (1 point) **[Hadoop]** Which of the following database operations – implemented as Hadoop jobs – require the use of a Mapper **and** a Reducer (instead of only a Mapper)?
   Assume that the dataset(s) to be used do not fit into the main memory of a single node in the cluster.

   ☐ Selection

   ☐ Union (with duplicates removed)

   ☐ Join

   ☐ Intersection

3. (1 point) **[ZooKeeper]** Which of the following statements about ZooKeeper's znodes are correct?

   ☐ Clients are able to create and delete znodes.

   ☐ Znodes are designed for large-scale data storage.

   ☐ The following znode flags can be set: *sequential, transient* and *watch*.

   ☐ Znodes are organised into a hierarchical namespace.

4. (1 point) **[Giraph]** Which of the following statements about Giraph are correct?

   ☐ Giraph employs both Hadoop's map and reduce phases to run computations.

   ☐ Giraph employs ZooKeeper to enforce barrier waits.

   ☐ Computations on data are performed in memory as much as possible.

   ☐ Giraph employs ZooKeeper to distribute the messages sent in superstep S to the correct recipients in superstep $S + 1$.

5. (1 point) **[Streaming]** Which of the following statements about Bloom filters are correct?

   ☐ A Bloom filter has the same properties as a standard Hashmap data structure in Java (`java.util.HashMap`).

   ☐ A Bloom filter is full if no more hash functions can be added to it.

   ☐ A Bloom filter always returns FALSE when testing for an element that was not previously added.

   ☐ A Bloom filter always returns TRUE when testing for a previously added element.

   ☐ An empty Bloom filter (no elements added to it) will always return FALSE when testing for an element.

6. (1 point) **[Streaming]** Which of the following streaming windows show valid bucket representations according to the DGIM rules?

   ☐ [1 0 1 1 1] 0 1 0 [1 1 1 1] 0 [1 0 1]

   ☐ [1 0 1 1 1] 0 0 0 0 [1 1 0 0] 0 [1 0 1] [1 1] 0 0 [1]

   ☐ [1 0 1 1 1] 0 0 0 0 [1 1 0] [1 0 1 0] [1]

   ☐ [1 0 1 1 0 0 0 1] 0 [1 1 1 0 1] [1 0 0 1] 0 [1] [1]

7. (1 point) **[HDFS]** For which of the following data storage scenarios would HDFS be a good choice?

   ☐ The European road sensor system collects the data of billions of road sensors that transmit information on the number of cars crossing the sensor, their speed, the weather conditions, etc. The data is fed into a machine learner to recognize accidents and inform the relevant emergency services to enable faster response times.

   ☐ Albert Heijn uses its stored transaction data (list of items in each shopping transaction) to determine which items to put on offer in their monthly shopping magazine.

   ☐ All UK security cameras are sending their data to a data centre where machine learning and pattern recognition is used to detect currently ongoing thefts and burglaries.

   ☐ Bol.com stores information on all their products (images, description, etc.) on HDFS to serve to customers via their website.

   ☐ High-resolution image and video data sent back to earth by the Mars Curiosity Rover and analyzed by scientists.

8. (1 point) **[Hadoop]** Which of the following statements about Hadoop's partitioner are correct?

   ☐ The partitioner divides the intermediate key space.

   ☐ The partitioner assigns keys to mappers.

   ☐ Within a single job, several different partitioners can be employed.

   ☐ There is no guarantee that a custom partitioner defined for a job is actually being used by the Hadoop framework, the default partitioner may be used instead.

   ☐ The partitioner determines which values are assigned to the same key.

9. (1 point) **[HBase]** Which of the following components exist in HBase?

   ☐ Memstore

   ☐ RegionServer

   ☐ HFile

   ☐ ZFile

   ☐ HSlave

   ☐ ZRegion

10. (1 point) **[Pig]** Which of the following definitions of complex data types in Pig are correct?

    ☐ Tuple: a set of key/value pairs.

    ☐ Tuple: an ordered set of fields.

    ☐ Bag: a collection of tuples.

    ☐ Bag: an ordered set of fields.

    ☐ Map: a set of key/value pairs.

    ☐ Map: a collection of tuples.

11. (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 ordered by the `Unique ID`.
    Data set 2 consists of information about Wikipedia articles written in English: `[Unique ID, Wikipedia URL, Wikipedia Title]`. The lines are unordered.
    Assume that neither data set 1 nor data set 2 fit into the working memory of a DataNode.

    A join should be performed on the field `Wikipedia URL` and the generated data set should look as follows: `[Edit Timestamp, Wikipedia URL, Wikipedia Title]`.

Which of the following specialized joins will lead to an error if chosen as join type for this scenario?

- ☐ sort-merge join
- ☐ skew join
- ☐ fragment-replicate join
- ☐ transient join

12. (1 point) **[Pig]** Consider the following data set:

```
Kermit,1988,7.5,Amsterdam
Gonzo,1987,7.5,Groningen
FozzieBear,1985,9.5,Wageningen
Scooter,1889,7,
Rowlf,,,
Pepe,1988,,Amsterdam
Rizzo,,7.0,
```

which contains the name, year of birth, grade average and birth place for a number of students.
Which of the following Pig scripts will correctly list all those students with a grade average strictly greater than 7?
Note: in the scripts, the `dump filtered;` statement has been left out for brevity.

- ☐
```
data = load 'data' as (name,year,grade,place);
filtered = filter data by grade>7;
```
- ☐
```
data = load 'data' as (name,year,grade,place);
filtered = filter data by grade>7.0;
```
- ☐
```
data = load 'data' as (name,year,grade:int,place);
filtered = filter data by grade>7;
```
- ☐
```
data = load 'data' as (name,year,grade:float,place);
filtered = filter data by grade>7;
```

13. (1 point) **[Streaming]** For which of the following streams is the second-order moment F greater than 45?

- ☐ 10 5 5 10 10 10 1 1 1 10
- ☐ 10 10 10 10 10 5 5 5 5 5
- ☐ 1 1 1 1 1 5 10 10 5 1
- ☐ 10 10 10 10 10 10 10 10 10 10

Free-form questions.

14. **[HDFS]** A large cluster runs HDFS on 1,000 nodes. Each node in the cluster, including the NameNode, has 16 Terabytes of hard disk storage and 64 Gigabytes of main memory available. The cluster uses a block-size of 64 Megabytes and a replication factor of 3. The master maintains 64 bytes of metadata for each 64MB block.

    (a) (1 point) What is the cluster's disk storage capacity? Explain your answer.

    (b) (2 points) A client downloads a 1 Gigabyte file from the cluster: explain precisely how data flows between the client, NameNode and the DataNodes.

    (c) (1 point) A client stores 1 billion (i.e. $1,000,000,000$) tweets as 1 billion small files of about 1.6 Kilobytes each. How much storage space (in memory & on disk) does this take on the NameNode and the DataNodes? Explain your answer.

(d) (1 point) After storing 1 billion files of 1.6 Kilobyte each, what is the cluster's remaining capacity? Explain your answer.

15. (4 points) **[HBase]** Design an HBase schema for two types of Twitter data: (i) relationship data (a user following other users), and, (ii) individual user data (user name, user profile image, user home location as latitude/longitude[1], and user's total number of tweets).

The schema should support the following operations **efficiently** for millions of users:

- `getFollowingList(u)` returns the list of all users a user `u` is following.
- `isFollowing(u,w)` returns `TRUE` if user `u` follows user `w` and `FALSE` otherwise.
- `getFollowedList(u)` returns a list of all users that follow user `u`.
- `getUsersFrom(lat,lon,r)` returns a list of all users whose home location is within a radius of `r` kilometers from the given latitude/longitude (`lat/lon`).

Give a textual and/or visual documentation of the schema, consisting of the table(s) needed, their column families and examples of individual columns with values. Motivate why your design is suitable for this task.

[1] Latitude and longitude are angles that uniquely locate geographic positions on the surfaces of planets such as Earth. Concrete examples of latitude/longitude coordinates are 52.009507/4.360515 (Delft), 52.373801/4.890935 (Amsterdam), 50.8503396/4.3517103 (Brussels) and 52.524268/13.40629 (Berlin).

16. (3 points) **[ZooKeeper]**

A "barrier" separates a process into two logical halves P_1 and P_2. Multiple nodes in the cluster running in coordination with one another will all perform P_1. No node can begin to work on P_2 until every node has completed P_1. The barrier sits thus between P_1 and P_2. As nodes reach the barrier, they all have to wait until the very last node has completed P_1. Then all nodes are released to begin working on P_2.

Describe, how such a barrier can be implemented through ZooKeeper API calls.

To help you with your task, here are the ZooKeeper API calls discussed in class:

```
* String create(path, data, flags)
* void delete(path, version)
* Stat exists(path, watch)
* (data, Stat) getData(path, watch)
* Stat setData(path, data, version)
* String[] getChildren(path, watch)
```

17. (3 points) **[Giraph]** 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. A visualization of a toy example is shown in Figure 1.

```
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
```

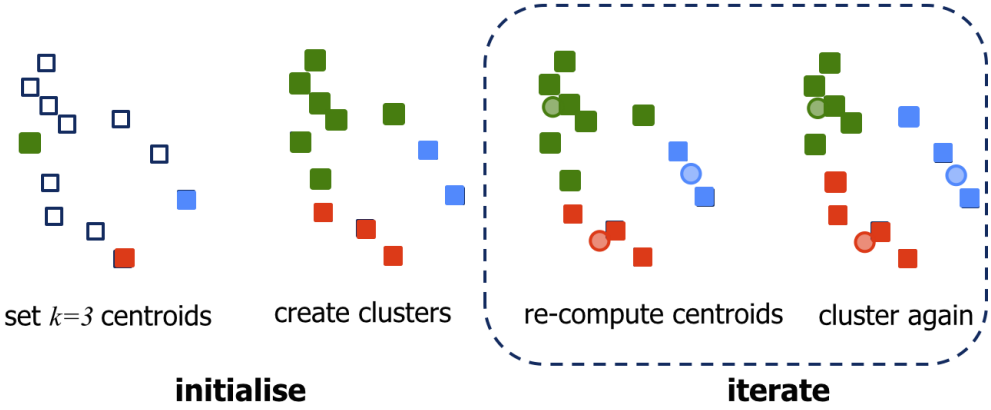


Figure 1: K-means example with $k = 3$ and 12 items to cluster.

Propose how to implement K-means in Giraph. Describe the data contained in the vertices, the location computation, the messages sent/received and the synchronization step.

18. (3 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 on texts (a) to (d). Each text is a DNA sequence in plain text for-

mat[2]. Each DNA sequence is stored in a file of about 1 Gigabyte.
The Hadoop cluster you have available has 25 DataNodes (64 Megabyte block size) and is currently running no other jobs.
Include in your answer the number of map and reduce tasks you anticipate to run for (a) to (d).

(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
...
```

19. **[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.

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

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 (primary key: ID). Note that the dataset is not ordered by ID.

ID	FirstName	LastName	Email	Nationality	Age
5332	John	Walters	john@walters.com	US	27
4499	Klara	Sommer	klaras@yahoo.de	German	39
310	...	...	...	...	...

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 300 million

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

records, the *Log* dataset contains tens of billions of 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 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.

(a) (2 points) Outline how you will solve this task with a single Hadoop job. Name the design patterns that are useful here and explain the core ideas

behind them.

(b) (2 points) Implement this task within a single Hadoop job. Write your answer down in **pseudo-code**.

20. **[Streaming]** We have discussed two approaches in class to solve the following task: *Given a data stream of unknown length, randomly pick k elements from the stream so that each element has the same probability of being chosen.*

    (a) (2 points) Describe how both approaches work.

    (b) (2 points) Which of the two approaches is better suited for *continous distributed sampling*? Explain your answer.

21. **[Pig]** You have access to the meta-data of all images uploaded to the photo-sharing site `Flickr`: `meta.dat`. The dataset is in the following format (one line per photo):

    1. `photo-id`
    2. `user-id`: photo owner
    3. `upload-date`
    4. `photo-title`
    5. `photo-description`
    6. `comment` flag: 1 to indicate users can comment, 0 otherwise
    7. `download` flag: 1 to indicate that the photo can be downloaded, 0 otherwise
    8. `share` flag: 1 to indicate that the photo can be shared, 0 otherwise
    9. `tags`: a list of numerical IDs corresponding to tags (delimited by ";")
    10. `raw-tags`: a list of terms that correspond to the numerical IDs in the `tag` list. All terms are assigned to the photo by the photo owner.
    11. `photo-url`: URL where the photo can be found.

Three example lines of the dataset are shown below:

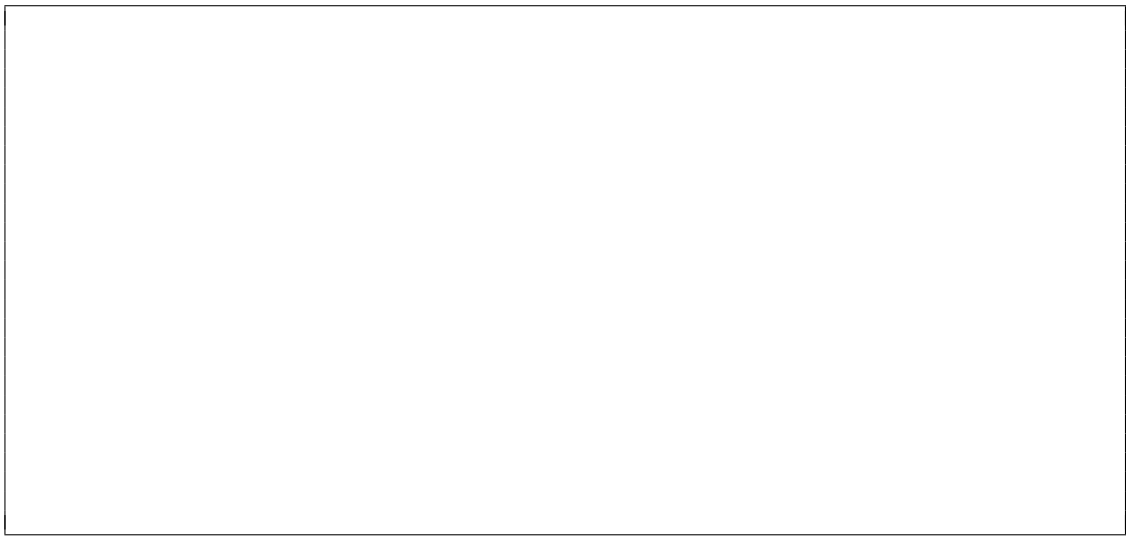
```
1: 862778,8169@N00,1361187927,"Summer","summer in Amsterdam",1,0,1,
     "1279319-8485662778-3637;1279319-8485662778-51205;1279319-8485662778-9279",
     "summer;sunshine;amsterdam",https://www.flickr.com/photos/8169@N00/862778/

2: 43454353,4352@N01,961187333,"Winter","winter in Norway",,,1,
     "43555-34344-3637;5346-4335566-3546466",
     "winter;dark",https://www.flickr.com/photos/4352@N01/43454353/

3: 312324,5322342@N01,1261187455,"Wellness","at a Geysir in Iceland",1,1,1,
     "9545-33-768;544-766-1580",
     "iceland;vacation",https://www.flickr.com/photos/5322342@N01/312324/

4: ...
```

(a) (2 points) Write a Pig script that outputs the ten users who have tagged the largest absolute number of their images with the term (raw-tag) “vacation”. Note that an image also counts if it has been tagged with multiple terms — only one of the tags has to be “vacation”.



(b) (2 points) You have just received a second dataset `users.dat`, which contains additional information about each Flickr user:

- `full-name`
- `date-of-birth`
- `home-location`
- `interests`

Here are three example lines:

```
1: 8169@N00,"John Doe","10-10-1978","USA",,
2: 4352@N01,"Anita Dijk","01-03-1985","Netherlands","horses;polo;traveling"
3: 5322342@N01,,"23-12-1981","Germany","photography"
```

```
4: ...
```

Write a Pig script, that outputs for each distinct country appearing in `users.dat` the user from that country that has taken the largest number of pictures.

(c) (1 point) In general, running a Pig query might take quite some time. Explain how to test and debug your query efficiently.

(d) (1 point) Assume that `meta.dat` and `users.dat` have been loaded into Pig as relations `META` and `USERS`. Describe the outpout of the following two lines:

```
C = cross META, USERS;
dump C;
```

(e) (1 point) We can assume that the majority of users in the dataset come from the USA. How can this knowledge be used to improve the efficiency of the

Pig script you wrote in 21(b)?

(f) (2 points) In general, what is the purpose of the `parallel` keyword in Pig? If the keyword is not supplied, what heuristic does Pig use instead?

THE END.