

Big Data Processing
2016-2017 Q2
January 30, 2017
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 15 pages (including this cover page) and 22 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.
- This exam contains open questions and two types of closed questions (**multiple-choice** and **multiple-answer**).
- 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 52.

Multiple-choice questions

The following questions have a single correct answer. You receive the point if you **check** ✓ the correct answer.

- (1 point) **[Spark]** What is a Resilient Distributed Dataset?
 - An immutable distributed collection of elements.
 - A mutable distributed collection of elements.
 - A write-enabled distributed collection of elements.
 - A spilled distributed collection of elements.
- (1 point) **[Streaming]** What is the space complexity of the FREQUENT algorithm? Recall that it aims to find all elements in a sequence whose frequency exceeds $\frac{1}{k}$ of the total count. In the equations below, n is the maximum value of each key and m is the maximum value of each counter.
 - $O(k(\log m + \log n))$
 - $o(k(\log m + \log n))$
 - $O(\log k(m + n))$
 - $o(\log k(m + n))$
- (1 point) **[Streaming]** What are DGIM's maximum error boundaries?
 - DGIM always underestimates the true count; at most by 25%.
 - DGIM always overestimates the count; at most by 50%.
 - DGIM either underestimates or overestimates the true count; at most by 50%.
 - DGIM either underestimates or overestimates the true count; at most by 25%.
- (1 point) **[Streaming]** Given is the following stream: A B $\overset{\uparrow}{C}$ B B $\overset{\uparrow}{A}$ C D A A.
What is the estimated second order moment according to the AMS algorithm? The arrows show the two randomly picked positions.
 - 20
 - 30
 - 40
 - 50

5. (1 point) **[General]** What is referred to as a "deadlock" in distributed systems?
- A condition where a set of processes have requests for resources that can never be satisfied.
 - A process that is successfully shut down after having acquired a pre-emption service.
 - A set of processes that experiences resource starvation due to failing hardware.
 - A lock of a shared resource that is only released once the process that gained the lock is no longer active.

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.

6. (1 point) **[Spark]** Which of the following statements about Spark Streaming are correct?
- Spark Streaming receives live input data streams and divides the data into small batches which are processed by the Spark engine.
 - Spark Streaming receives live input data streams and processes each data element as soon as it arrives.
 - Spark Streaming does not make use of the core Spark engine.
 - Spark Streaming is a data generator process to simulate high-velocity data streams.
7. (1 point) **[Spark]** Which of the following statements about actions and transformations are correct?
- All transformations in Spark are lazy.
 - By default, each transformed RDD may be recomputed each time you run an action on it.
 - Some actions in Spark are lazy.
 - When an RDD is persisted, each node stores any partitions of it that it computes in memory and reuses them in other actions on that dataset.

8. (1 point) **[ZooKeeper]** Which of the following statements about ZooKeeper's znodes are correct?
- Znodes are designed for large-scale data storage.
 - The following znode flags can be set: *sequential* and *watch*.
 - Clients are able to create and delete znodes.
 - Znodes are organised into a hierarchical namespace.
9. (1 point) **[Hadoop]** Which of the following statements about Hadoop's partitioner are correct?
- The partitioner assigns keys to mappers.
 - Multiple partitioners cannot be employed within a single Hadoop job.
 - 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 keys are sent to the same reducer.
10. (1 point) **[Streaming]** Which of the following streaming windows show valid bucket representations according to the DGIM rules?
- | | | | | |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 |
|---|---|---|---|---|

 0 0 0 0

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

 0

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

1	1
---	---

 0 0

1

 - | | | | | |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|

 0 0 0 0

1	1
---	---

 0 0

1	1
---	---

 0

1

 - | | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
|---|---|---|---|---|---|---|---|

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

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

 0

1

 - | | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|

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

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

 0

1

 0

1

11. (1 point) **[Streaming]** For which of the following streams is the second-order moment $F_2 \geq 30$?
- 10 2 10 2 2 1 1 1
 - 5 7 7 2 5 5 2 2
 - 1 1 1 1 1 1 1 0
 - 10 10 1 1 10 10 1 1
12. (1 point) **[General]** Which of the following are fallacies of distributed computing?
- The latency is zero.
 - The bandwidth is infinite.
 - The network topology does not change.
 - The network is reliable.

13. (1 point) **[Giraph]** Which of the following statements about Giraph are correct?

- Giraph employs only Hadoop's map phase to run computations.
- Giraph employs ZooKeeper to distribute the messages sent in superstep S to the correct recipients in superstep $S + 1$.
- Giraph employs ZooKeeper to enforce barrier waits.
- Giraph disallows nodes from sending messages to other nodes they do not share an edge with.

14. (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 key/value pairs.
- Bag: an ordered set of fields.
- Map: an ordered set of fields.
- Map: a collection of tuples.

Free-form questions.

15. **[HDFS]** Consider a small cluster with 20 machines: 19 DataNodes and 1 NameNode. Each node in the cluster has a total of 2 Terabyte hard disk space and 2 Gigabyte of main memory available. The cluster uses a block-size of 64 Megabytes (MB) and a replication factor of 3. The master maintains 100 bytes of metadata for each 64MB block.

(a) (1 point) Lets upload the file `wiki_dump.xml` (with a size of 600 Megabytes) to HDFS. Explain what effect this upload has on the number of occupied HDFS blocks.

(b) (2 points) Figure 1 shows an excerpt of `wiki_dump.xml`'s structure. Explain the relationship between an HDFS block, an InputSplit and a record based on this excerpt.

```
<dump time="1483027930">
  <page id="EN3234">
    ...
    ...
    ...
  </page>
  <page id="DE5434">
    ...
    ...
    ...
  </page>
  ...
</dump>
```

} 80.2 MB

} 0.6 MB

Figure 1: Excerpt of `wiki_dump.xml`. Each Wikipedia page is stored within a `<page>` element. The element with id EN3234 contains 80.2 Megabytes of textual content.

- (c) (2 points) You are the only user of the cluster and write a Hadoop job to extract information from `wiki_dump.xml`. You want to speed up the job by testing different block size configuration: besides the existing 64 MB configuration, you also consider 32 MB and 128 MB block sizes. Which configuration do you think will lead to the fastest job execution? Explain why.

- (d) (2 points) Let us assume no files are currently stored on HDFS. You are given 100 million files, each one with a size of 100 Kilobytes. How many of those can you upload successfully to the cluster, considering the storage restrictions (memory/disk) on the NameNode and the DataNodes? Explain your answer.

16. (2 points) **[Hadoop]** What is the purpose of *speculative execution*? Why is it usually restricted to the map phase and not used in the reduce phase?

20. **[MapReduce Design Patterns]** You are given a file `bib.sci` with a large scientific bibliography, each entry containing the authors' name, the title of the work and the publication venue. You are asked to produce for each pair of authors the number of papers they have written together, that is, both author names appear in the author list in arbitrary order. For any two authors that have co-authored one or more papers, exactly one output record should be produced. Author pairs that have not written a publication together should not appear in the output. As an example, consider the following toy bibliography with 3 entries (you can assume each entry to be one line in `bib.sci`):

Felix Hill, KyungHyun Cho, Anna Korhonen, Yoshua Bengio. Learning to Understand Phrases by Embedding the Dictionary. TACL 4: 17-30 (2016)

Felix Hill, KyungHyun Cho, Sebastien Jean, Coline Devin, Yoshua Bengio. Not All Neural Embeddings are Born Equal. CoRR abs/1410.0718 (2014)

Felix Hill, Anna Korhonen. Concreteness and Subjectivity as Dimensions of Lexical Meaning. ACL (2) 2014: 725-731

You can further assume the availability of a method `extractAuthors()` to retrieve the list of authors from each entry. Your program should output the following pairs (note that their order is arbitrary):

(Coline Devin,Yoshua Bengio) 1
(Coline Devin,Felix Hill) 1
(Felix Hill,KyungHyun Cho) 2
(Felix Hill,Sebastien Jean) 1
(KyungHyun Cho,Sebastien Jean) 1
(Coline Devin,Sebastien Jean) 1
(Coline Devin,KyungHyun Cho) 1
(Sebastien Jean,Yoshua Bengio) 1
(KyungHyun Cho,Yoshua Bengio) 2
(Felix Hill,Yoshua Bengio) 2
(Anna Korhonen,Felix Hill) 2
(Anna Korhonen,Yoshua Bengio) 1
(Anna Korhonen,KyungHyun Cho) 1

- (a) (2 points) In class, we covered two design patterns (*Pairs* and *Stripes*) that are able to solve this task. Pick one of the two design patterns and apply it to the task described here. Name the design pattern you picked and write **pseudo-code** to solve the task.

- (b) (1 point) In **pseudo-code** write a useful combiner for this task that works in combination with the mapper/reducer of answer (a).

- (c) (1 point) Briefly explain why the combiner you wrote in (b) is likely to speed up the execution time of the job. Explain why it is not guaranteed to speed up the execution time.

- (d) (2 points) Given the task, which of the two design patterns (*Pairs* or *Stripes*) will benefit more from the use of a combiner? Explain your answer.

- (e) (2 points) Given the task, which of the two design patterns (*Pairs* or *Stripes*) will scale seamlessly, i.e. perform without fail no matter how large the corpus becomes? Explain your answer.

21. **[Pig]** You have been given a file that contains the meta-data of all movies that were produced in the last 50 years: `movies.dat`. The first three lines of this file look as follows:

```
(Open Season,9/29/2006) [budget#$85M,gross#$197M]      {(RT,48%),(MC,49)}  
(Surf's Up,6/8/2007)   [budget#$100M,gross#$149M]    {(RT,78%),(MC,64),(IMDB,4.5)}  
(Arthur Christmas,11/23/2011) [budget#$100M,gross#$147M]   {(MC,69),(RT,92%)}
```

As evident, the schema definition contains complex data types. If translated to a relational schema, the first line of this file would look as shown in Figure 3. The primary key is the composite of the first two columns in Figure 3.

Movie title	Release date	Budget	Gross income	Rotten Tomatoes (RT) score	Metacritic (MC) score
Open Season	9/29/2006	\$85M	\$197M	48%	49

Figure 3: Relational schema representation of the first line of `movies.dat`.

- (a) (2 points) Write a Pig script that outputs for each release year, the three movies with the highest budget.

- (b) (1 point) List all Pig operators you have used in (a) that force a reduce phase.

- (c) (2 points) You have received a second dataset `movies2.dat`, which contains information about some of the movies' availability (either 'yes' or 'no') on popular streaming services. The first three lines of this file look as follows:

```
(Open Season,9/29/2006) [netflix#yes,hulu#yes]
(Arthur Christmas,11/23/2011) [netflix#no]
(The Smurfs 2,7/31/2013) [netflix#yes,pathethuis#yes,hulu#no]
```

[continuation of (c)] Write a Pig script, that outputs for each streaming service the average gross income of the movies streamed on it.

- (d) (2 points) The script you wrote in (c) requires a join between the two datasets `movies.dat` and `movies2.dat`. Describe for the two of the following join implementations under which circumstances they will deliver accurate results and a faster execution time than the default join.

fragment-replicate join:

skewed join:

- (e) (2 points) Describe how Pig's skewed join is implemented through Hadoop jobs.

22. **[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 *continuous distributed sampling*? Explain your answer.

THE END.