Techniques for Large-Scale Data: Exam

University of Gothenburg | Chalmers University of Technology
Department of Computer Science and Engineering
Period 4, 2019 (DIT872 (and re-exam in DIT871) / DAT345)
Instructors: Dr. Alexander Schliep (Examiner) with Dr. Graham Kemp

Information:

- The exam takes place from 8:30–12:30 on Wednesday, June 5, 2019.
- The instructors or a representative will visit the exam room at 9:30 and 11:00 (Tel. 076-608 69 63).
- You can earn are a total of 60 points in 8 questions in the exam. Additionally, the clicker percentage points will be added to the exam points as bonus points; the maximum will be six bonus points (10% of the total exam points) added for a 100% iClicker result.
- Grades for GU students (DIT872) are normally determined as follows: $\geq$ 70 % for grade VG; $\geq$ 40 % for grade G. Grades for Chalmers students (DAT345) are normally determined as follows: $\geq$ 80 % for grade 5; $\geq$ 60 % for grade 4; $\geq$ 40 % for grade 3.

Instructions:

- You may use one A4-sized sheet (back and front) of prepared formulas and notes, but all work must be your own. No photocopies or print-outs of slides, books, material off the web. It must be handed in with your exam questions. Please, write the exam code on it. *Please, do not write your name on it.*
- Begin the answer to each question on a new page. Write page number and question number on **every** page.
- Write clearly; unreadable = wrong!
- Fewer points are given for unnecessarily complicated solutions.
- Indicate clearly if you make any assumptions that are not given explicitly in the question.
- Show **ALL** your work. You will get little or no credit for an unexplained answer. Please indicate why a specific computation or transformation is appropriate. The points of each question appear in parentheses; use this for guiding your time.
- There is no need to compute numerical answers; you may leave Binomials, factorials and fractions, should they arise, as is.
- **No electronic devices of ANY kind!** Please store **all** your devices in your bag and not on your person. Any device found at your seat, **even if it is turned off**, will be considered cheating and reported!
- Printed English language dictionaries—including dictionaries translating to and from another language to English—are allowed. Electronic dictionaries are not.

Question 1 [*5 points total*]

A media article discusses data gathered about public transport usage in London:

> Transport for London (TfL) now has a bird's eye view of the estimated 4.1 million journeys taken on its network each day.
> It knows where people get on and off and it can start to see patterns in the data — for instance, someone who uses the system during the day but not at peak times is likely to be a student or a retired person, someone who has one day a week when they don't use the network may work from home that day, someone who takes a brief diversion along their usual route may be dropping off a child at nursery.
> "The data can be used to inform future expansion, whether we need to add a bus route or increase the frequency of trains, to alleviate capacity issues by informing people about the most crowded times and places and generally helps us to understand customers better," said ... a data scientist at TfL.

Discuss whether the data gathered by Transport for London raises potential ethical issues.

Question 2 [*15 points total*]

The following CYPHER statement was generated using the "Arrow" tool.

```
CREATE
  (‘0‘ :Course {code:"DAT345",name:"Techniques for large-scale data"}) ,
  (‘1‘ :Lecture {day:"Wednesday",time:"10:00"}) ,
  (‘2‘ :Room {room_name:"HA3"}) ,
  (‘3‘ :Teacher {forename:"Graham"}) ,
  (‘4‘ :Lecture {day:"Monday",time:"10:00}) ,
  (‘5‘ :Teacher {forename:"Alexander"}) ,
  (‘0‘)-[:‘HAS_LECTURE‘ ]->(‘1‘),
  (‘1‘)-[:‘IN_ROOM‘ ]->(‘2‘),
  (‘1‘)-[:‘GIVEN_BY‘ ]->(‘3‘),
  (‘0‘)-[:‘HAS_LECTURE‘ ]->(‘4‘),
  (‘4‘)-[:‘IN_ROOM‘ ]->(‘2‘),
  (‘4‘)-[:‘GIVEN_BY‘ ]->(‘5‘)
```

(a) [*3 pts*] Sketch how this information would appear in the "Arrow" tool.

(b) [*4 pts*] Suppose we want to store the same information as triples in a Semantic Web database. What triples would we have?

(c) [*3 pts*] Give RDFS or OWL 2 statements that model the metadata describing the *relationships* in this example.

(d) [*3 pts*] Based on the triples in your answer to part (b), and assuming many other similar triples, write a SPARQL query that finds the forenames of teachers who give a lecture in room HA3 on Wednesday.

(e) [*2 pts*] RDF triples can be stored in a relational database in different ways. One simple approach is to have a single table with three columns storing subject, predicate and object values. Another approach is to build a *hexastore*. What is a *hexastore*?

Question 3 [*3 points total*]

NoSQL systems are sometimes described as being *schemaless*. What does this mean? Give an advantage and a disadvantage of a database being *schemaless*.

Question 4 [*8 points total*]

(a) [*6 pts*] Suppose we have relations $S(a,b)$ and $T(c,d)$ and query:

```
SELECT d
FROM   S, T
WHERE  b=50 and a=c
```

Draw *three* alternative logical query plans for this query.
Discuss the efficiency of the logical query plans.

(b) [*2 pts*] Describe two decisions that are made during *physical plan generation*.

Question 5 [*9 points total*]

(a) [*4 pts*] A transaction processing system might experience a *system error* or a *system crash*. Describe these kinds of failures, and how they differ.

(b) [*5 pts*] Suppose a relational database contains relation *Employees(empId, name, salary, dept, city)*, that this relation is stored in 30 disc blocks, and that each city has employee records stored (on average) in 5 disc blocks.

Suppose that two kinds of task are performed on this relation:

- task 1: inserting a new row;
- task 2: finding the employee identifiers for a given $city$.

For each of these tasks, state how many disc block transfers will be needed if:

i) there are no indexes;

ii) there is an index on $city$ (assume that this index fits into a single disc block).

Suppose that 90% of operations performed on this relation are task 1 (inserting new rows) and 10% are task 2 (finding the employee identifiers for given cities).

iii) Discuss whether it would be better to have an index on $city$, or to have no indexes.

Question 6 [*4 points total*]

Recall the definition of the memory hierarchy in computer architectures. At the top are the CPU registers.

(a) [*1 pts*] What are the (most important) different levels of the memory hierarchy?

(b) [*1 pts*] Give approximate latency (order of magnitude suffices) for access to data in lower levels compared to access to CPU registers.

(c) [*1 pts*] What is the main technical factor causing addition of the intermediate levels in the memory hierarchy?

(d) [*1 pts*] Which consequences result from latency across the memory hierarchy both for the design of serial and parallel programs?

Question 7 [*8 points total*]

During the course we learned about four main paradigms for writing parallel code for computations: multi-threaded programming (MT), message passing (MP), map-reduce (MR), and Spark (SP) as an example of cluster computing.

Consider the task of computing histograms and summary statistics (mean, standard deviation, min, max; not median) for a given array of numbers read from a file.

Sketch (either graphically, with bullet points, or commented pseudo-code) and explain how the communication between parallel tasks is handled in two different out of the four frameworks (you may choose which two) using the task of computing histograms and summary statistics for univariate observations as an example.

You do not have to explain File I/O. Do explain the data flow and how the execution progresses from an array of numbers (MT, MP) respectively tuples (MR, SP) to the final result of a histogram and summary statistics.

Question 8 [*8 points total*]

As a junior data scientist recently hired to Reynholm Industries you are tasked with overhauling and redesigning the IT infrastructure and accelerating existing computational workloads. The COO's team runs daily analysis (DA) on orders, shopping baskets, online advertising, and website visits. The total amount of data is about 200GB per day collected in Reynholm Industries' data center. The output are summary statistics (histograms, averages, etc.) and nice plots enabling monitoring the state of affairs. It is desirable to collect historical data, but the size of the data makes it impossible to keep all the data in one database on one machine.

Before your arrival the IT people split up the database over fifty machines and demand keeps growing. The way the data is split up, it is necessary to query all the databases if one wants to summarize data, say compute the total of sales, for one specific customer over Reynholm Industries' lifetime. The querying is done in parallel on the fifty machines, so it is reasonably time-efficient for one query. However, as all the servers are busy for each query, this approach is neither energy-efficient nor time-efficient if many queries are run concurrently. Obviously, for customers who rarely shop at Reynholm Industries queries for sales to that customer would return no result on most of the fifty servers which hold the historical sales data in a distributed manner.

(a) [*4 pts*] How do you impress your new colleagues by reducing the total computational effort needed for computing the total of sales for one specific customer?

(b) [*2 pts*] To assure Reynholm Industries' competitiveness, management suggests to investigate deployment of machine learning models trained with the large amount of customer data. Propose an appropriate hard- and parallel software environment supporting such advanced computations and argue which specific aspects influence your decision.

(c) [*2 pts*] Your boss wants to accelerate computations for one specific analysis task. The two options are (N) to buy more of the existing type of cluster nodes (e.g., Intel-based SMP general purpose computers with 128GB RAM and disks) or (H) buy tensor-unit hardware accelerators, which can execute specific types of operations very rapidly. Describe your decision process including the variables you need to determine.