EBU750U COVID-19 Alternative Assessment

Joint Programme Assessments 2019/20

EBU750U Cloud Computing

Answering this paper requires 2 hours; Answers to be submitted within the allocated 48 hours window.

Answer ALL questions

INSTRUCTIONS

1. **You must NOT share any content from this document during the assessment period.**
2. Your answers must be typed or written clearly and legibly with black or blue colour **and in English**.
3. You need to submit your answers BEFORE the allocated deadline.
4. **Read the instructions on the inside cover of the questions sheet.**

Examiners

Dr Gokop Goteng, Dr Atm Alam

Filename: 1920_EBU750U_ALT2020_1

Instructions

This is an open-book exam, which should be completed within 2 hours. You MUST submit your answers within 48 hours from the exam being released.

You MUST complete the exam on your own, without consulting any other person. You MAY NOT check your answers with any other person.

You can refer to textbooks, notes and online materials to facilitate your working, if you provide a direct quote, or copy a diagram or chart, you must cite the source.

Before you start the assessment

1) Read the questions thoroughly and understand them.

2) Ensure you have all the resources you require to complete and upload the final assessment.

3) If you require any assistance, **raise the issue via the messaging section of this assessment on QMPlus**, immediately.

During the assessment session

1) Use the supplied answer sheet document to enter your answers. Start on a new page for each question. Make sure it is clear which question number you are answering.

2) Type your answers in the supplied answer sheet; hand-written answers (including sketches or equations) can be scanned and incorporated into the answer sheet. Please save your work at least every 15 minutes so that you do not risk losing it.

3) When completed answering all questions, save the file as pdf before uploading, **only pdf will be accepted**, any other file format will not be accepted.

4) Your submission must be your own work, and you must ensure that you do not break any of the rules in the Academic Misconduct Policy.

Submitting the Assessment

1) You will have 48 hours from the start of the scheduled assessment time – do not leave submissions too close to the deadline. NO late submission will be accepted, no exceptions.

2) Make sure you upload and submit the final version before the deadline.

5) Please be aware that submissions will be subject to review, including but not limited to plagiarism detection software.

If you have any problems relating to access or submitting during the exam period, please contact the email (it-issues@qmbupt.org), state the module code in the subject, and clearly state your name and student ID and any issues you are experiencing. You must use either @qmul.ac.uk or @bupt.edu.cn email address. Requests from external email addresses will not be processed.

Question 1

a) Alibaba Cloud has a compute cluster which consists of 100 Virtual Central Processing Units (VCPUs) and 40% of the VCPUs are dedicated for serial or sequential processing activities.

[15 marks]

i) Calculate the TWO system efficiencies of this Alibaba Cloud cluster using "fixed workload" and then using "scaled workload" in TWO separate calculations.

(10 marks)

ii) If the Alibaba Cloud cluster has a total Mean Time To Failure (MTTF) of 600 days and an average Mean Time To Repair (MTTR) of 2 days, calculate the High Availability (HA) of the cluster, showing all steps of your calculations.

(5 marks)

b) Amazon Web Services (AWS) provides cloud services that are highly available, reliable and have disaster recovery and failover redundant systems. Describe the services and infrastructures that AWS uses to provide highly available and disaster recovery systems for its databases, elastic compute cloud (EC2) and data storage.

[4 marks]

c) An Information Technology (IT) company maintains a physical computer cluster and wants to use VMWare virtualisation technology to create virtual servers for increased efficiency and performance. The company has sent its Systems Administrators to be trained on how to manage the VMWare and virtual cluster. Can you explain in at least FOUR ways the Systems Administrators will use VMWare technologies to manage both the physical and virtual machines in the newly created virtual cluster?

[4 marks]

d) A cloud user has sent an un-encrypted data to another user via a public cloud network. Describe how to ensure that the data is not tampered with while in transit by another malicious user.

[2 marks]

Question 2

a) This question is about Virtual Private Cloud (VPC):

[15 marks]

i) Describe a CIDR and a Subnet as used in AWS VPC and why CIDR has changed the way networking is done compared to 20 years ago.

(4 marks)

ii) An e-commerce company has its web server and database hosted in AWS VPC 10.0.0.0/16. The VPC consists of a public subnet 10.0.1.0/24 and private subnet 10.0.2.0/24. Describe the VPC IP address range "10.0.0.0/16" in terms of which part of the IP address will change or remain unchanged and explain the meaning and implications of the CIDR "/16". Also describe where the web server and database should be placed in terms of the two subnets created.

(11 marks)

b) Write a Graphics Processing Units (GPU) program using the Compute Unified Device Architecture (CUDA) that computes a Fibonacci sequence of numbers. Discus the drawbacks/disadvantages in using GPU to compute Fibonacci sequence and suggest the best types of problems or applications that suit GPU CUDA programming model.

[10 marks]

Question 3

a) Consider a dataset for the current COVID-19 pandemic from the data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The timeseries dataset contains the global daily confirmed cases, recovered cases and death tolls from 22 January 2020 for all affected countries. Each record represents a country's COVID-19 status where various details are captured [*types in brackets*]:

`date [String], countryName [String], confirmedCases[String], recoveryCases[String], deathTolls [String].`

This data can be fed as input to a MapReduce job as a set of key/value pairs (`String date, CountryData data`). The keys are Strings with the unique id (date), and the values are **CountryData** objects with the full details (and methods to access each of the fields). For example,

- **`CountryData.getCountryName()`** will return the **`countryName`** field of the input row.
- **`CountryData.getCategory()`** will return either of the three categories: **`confirmedCases`**, **`recoveryCases`**, **and `deathTolls`** field of the input row.
- **`CountryData.getCases(date, category)`** will return the number of cases of the mentioned **category** on the mentioned **date** of the input data.

Write a MAP function and a Reduce function for a given input that computes the most global cases (**confirmed and recovery cases**) and deaths (**death tolls**) for each day from the CSSE's COVID-19 dataset. You can use pseudocode to write the specification, and use a diagram to illustrate the data flow between input, map, reduce, and output blocks. Marks will be given for precision and conciseness.

Note: you can assume that there is a method called **`computeMax(List<Pair> cases)`**, which returns the country that has the daily global maximum cases in a list. This should be used in the **`reduce`** method.

[6 marks]

b) Hadoop job execution involves many computation tasks. Indicate with **arrows** (→), which daemons (right part of **Figure 1**) are responsible for the Hadoop computation tasks (left part of **Figure 1**).

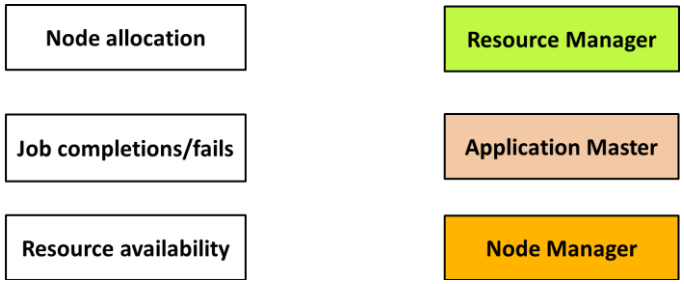

Figure 1

Note that you can copy this figure to your answer script and connect the corresponding blocks.

[3 marks]

c) The **Combiner** has the same structure as the reducer (i.e., the same method signature), but it must comply with certain rules. In this context, state TRUE or FALSE "***If a Combiner function is called more than 2^n times (where n is the number of Mappers in your program), the output from the Reducer will not be the same***."

A. TRUE

B. FALSE

[2 marks]

d) When processing large datasets, the need for joining data by a common key can be very useful. Explain three ways of **joining datasets** in Hadoop.

[6 marks]

e) This question is about **performance** in Map/Reduce.

[8 marks]

i) Define the concept of speedup in parallel computing.

(1 mark)

ii) Assume that in a current design, 40% of a computation job in an application can be performed using up to 2 processors (i.e., in parallel). Now, the application can be redesigned with either of the following two choices so that the performance in terms of speedup is improved.

a. Increase the number of processors to 4.
b. Use 2 processors but add features that will allow the applications to use them for 80% of execution.

Which will you choose, and why? Show the working out.

(7 marks)

Question 4

a) How does a Spark Execution Architecture work? Explain with a suitable diagram and appropriate labeling.

[6 marks]

b) The following questions relate to **Content Delivery Networks** (CDNs):

[7 marks]

i) What is a Content Delivery Network? Name one example of a Content Delivery Network other than Akamai, ChinaCache, and Limelight.

(2 marks)

ii) Why does a Content Delivery Network place servers all around the world?

(3 marks)

iii) Explain what load balancing is in the context of a Content Delivery Network.

(2 marks)

c) The following questions relate to **Cloud Databases**:

[7 marks]

i) What is **Brewer's CAP Theorem**? Explain the three properties of CAP.

(4 marks)

ii) Apache **Casandra** is a NoSQL distributed database that provides high data availability across a number of machines. The following diagram (**Figure 2**) shows a Cassandra ring of **5** machines. All information in the Cassandra ring is saved with a Replication Factor of **2**.

The Client makes a data read request to **Node #1** (the coordinator). The grey heptagon indicates the hash value of the key the client requested. Which node will serve the data to the client? Where will the replica of the data be stored? Explain your answer.

(3 marks)

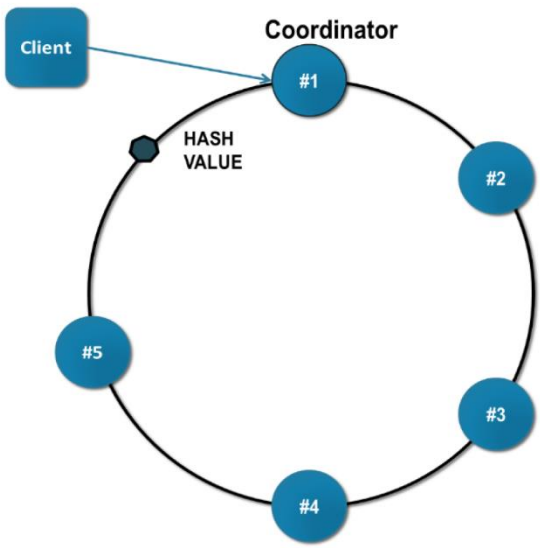

Figure 2: Cassandra Ring

d) This question is about **distributed graph processing**

[5 marks]

i) What is **graph partitioning**? Why is it necessary? Discuss the role of graph partitioning in distributed graph processing systems.

(3 marks)

ii) Explain the relationships between **graph partitioning** and **performance**. Would a *bad* partitioning decision results in worse performance? If so, **why**?

(2 marks)

END OF PAPER

DO NOT WRITE ON THIS PAGE.