NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 2 EXAMINATION 2022-2023

CE4123/CZ4123/SC4023 – Big Data Management

Apr/May 2023 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 5 questions and comprises 6 pages.
2. Answer **ALL** questions.
3. This is a closed-book examination.
4. All questions do NOT carry equal marks.

---

1. For each the following questions, select one answer choice and write your answer in the answer sheet.

   (a) Which of the following statement is correct?
   A. Variance is one of big data 5 V's.
   B. In a distributed system, if one data replica loses, then the computed results will not be accurate.
   C. In a distributed system, data must be partitioned equally at each machine.
   D. In a distributed system, strict (replica) consistency usually means low availability.

   (2 marks)

   (b) Which of the following statements about data models is NOT correct?
   A. Both key-value model and graph model have been used in NoSQL databases.
   B. Column-store always uses schema-less data model.
   C. In key-value data model, the key uniquely decides the value.
   D. Edges in graph model have labels and directions.

   (2 marks)

Note: Question No. Q1 continues on Page 2

(c) Which of the following statements about LSM-tree is NOT correct?
   A. In LSM-tree, Bloom filters can reduce page reads when calling *Put*(*k*, *v*).
   B. LSM-tree employs an out-of-place update scheme.
   C. In an LSM-tree of size ratio 4, the number of key-value pairs currently stored in Level-2 can be 2 times of the number of key-value pairs stored in Level-1.
   D. Fence pointers can reduce the I/O cost.

(2 marks)

(d) Which of the following statements is NOT correct?
   A. Wide-Column database is one type of NoSQL databases.
   B. If compression technique is applied in column store, decompression may not be done before querying the compressed column.
   C. When updating a whole tuple, column store faces bigger challenges than row store.
   D. In column store, if tuple size is smaller than disk page size, then reading a tuple only requires reading one disk page.

(2 marks)

(e) We have a 10-integer array H[0-9] in main memory: 9, 8, 7, 6, 5, 4, 3, 2, 1, 0. Cache size is 4 (integers), and cache line size is 2 (integers). Assume that the cache replacement mechanism is first-cached-first-evicted. Which of the following access patterns gives the best cache performance?
   A. H[1], H[3], H[5], H[7], H[9], H[8], H[0], H[2], H[4], H[6]
   B. H[0], H[2], H[4], H[6], H[8], H[9], H[1], H[3], H[5], H[7]
   C. H[0], H[2], H[4], H[6], H[8], H[7], H[9], H[5], H[3], H[1]
   D. H[4], H[5], H[6], H[7], H[0], H[1], H[2], H[3], H[8], H[9]

(2 marks)

2. Consider the task of analyzing monthly salaries from a dataset. There are N salaries sequentially stored on the disk and each salary is a positive integer. Suppose a disk page can contain at most B integers. The main memory can contain at most M integers, and $M>100B$. Both N and M are multiples of B. Design algorithms for the following cases and explain the detailed procedures.

   (a) If $M<N<50M$, design an algorithm to compute the median salary (i.e., the median of the N integers) using $O(N/B)$ I/Os.

   (10 marks)

Note: Question No. Q2 continues on Page 3

(b) If B=4096 and $N>M$, design an algorithm to compute the highest 8000 salaries (the largest 8000 integers among the N integers) using O(N/B) I/Os.

(10 marks)

3. Figure Q3 shows a table R of four columns Ra, Rb, Rc, Rd, and a table S of four columns Sa, Sb, Sc, Sd. Both R and S are stored in a column store and all values in R and S are integers. Consider the query "**SELECT** SUM(Ra) **FROM** R, S **WHERE** Rc=Sb and Ra<6 and Rb>2 and Sa>2 and Sc<8 and Sd>3" and answer the following questions.

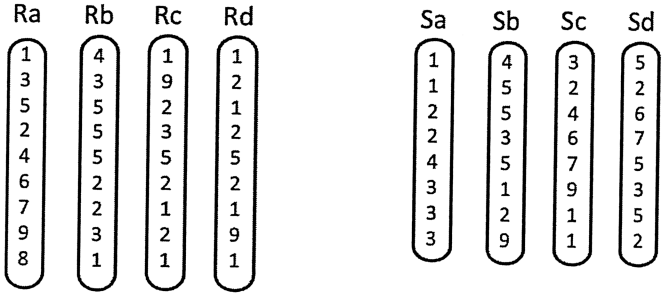

Figure Q3

(a) Draw the flow chart for the query.

(5 marks)

(b) Show how sorting can be respectively applied in R and S to make the scan more efficient. Draw the flow chart after sorting.

(5 marks)

(c) If an integer occupies 4 bytes, please explain how to apply the dictionary-based compression on column Ra and column Sa, and show one possible encoding of each value in Ra and Rs. Also compute the respective compression rates for Ra and Rs.

(5 marks)

4. Consider a *Student* table containing three columns (*studentID*, *courseID*, *semester*). Each tuple in the *Student* table records that a student registered for the course in the corresponding semester. Also consider a *Professor* table containing three columns (*professorID*, *courseID*, *semester*). Each tuple in the *Professor* table records that a professor teaches a course in the corresponding semester. A course may open in multiple semesters. If a student fails a course, he may retake the course in later semesters. **Example tuples** for the two tables are as follows.

studentID	courseID	semester
S001	C001	2021S1
S002	C001	2021S1
S002	C002	2021S2
S003	C001	2021S2

Table Q4.1: Example Student Table

professorID	courseID	semester
P001	C001	2021S1
P002	C002	2021S1
P001	C001	2021S2

Table Q4.2: Example Professor Table

The *Professor* table and the *Student* table are stored together in a file named *input_file*, with each tuple per line. There is an additional attribute to indicate whether this tuple is from the *Student* Table or the *Professor* Table. Based on the above example tuples, the file content is as follows.

Student-Table S001;C001;2021S1
Student-Table S002;C001;2021S1
Student-Table S002;C002;2021S2
Student-Table S003;C001;2021S2
Professor-Table P001;C001;2021S1
Professor-Table P002;C002;2021S1
Professor-Table P001;C001;2021S2

Note: Question No. Q4 continues on Page 5

Please use MapReduce for the following scenarios and write down the pseudocode. Your pseudocode should start with a *Map* function that takes each line of the *input_file* as the input. The key in the *Map* function is the additional attribute (e.g., "*Student-Table*") in the line, and the value in the *Map* function is the remaining of the line (e.g., *S001;C001;2021S1*). You also need to design *Reduce* function if necessary. You can use multiple MapReduce jobs.

(a) Use MapReduce to collect for each student (represented by *studentID*) the number of distinct courses (represented by *courseID*s) he has registered.

(5 marks)

(b) Use MapReduce to collect the courses (represented by *courseID*s) that have more than 50 registered students for at least two semesters. For example, a course will be output if there are 55 students for 2021S1 and 60 students for 2021S2.

(8 marks)

(c) Use MapReduce to output every pair of (student, professor) (represented by *studentID* and *professorID*) that the student has attended at least one courses taught by the professor.

(12 marks)

5. Consider an LSM-tree with a size ratio T. The memory buffer (Level 0) can store at most M key-value pairs.

(a) Consider T=10 and M=1000, and the LSM-tree is a leveling LSM-tree which is initially empty. If we call 1,000,000 times of function *put*(), finally the LSM-tree has L levels (i.e., Level-0 to Level-(L-1)). Please list all possible values of L and explain your answer.

(6 marks)

(b) Consider T=3 and M=2, and the LSM-tree is a leveling LSM-tree which is initially empty. If after inserting 11 distinct key-value pairs, the LSM-tree has the following layout shown in Figure Q5b (e.g., (1, v_1) is a key-value pair). Please compute how many possible insertion orders will lead to the LSM-tree layout and explain your answer.

(7 marks)

Note: Question No. Q5(b) continues on Page 6

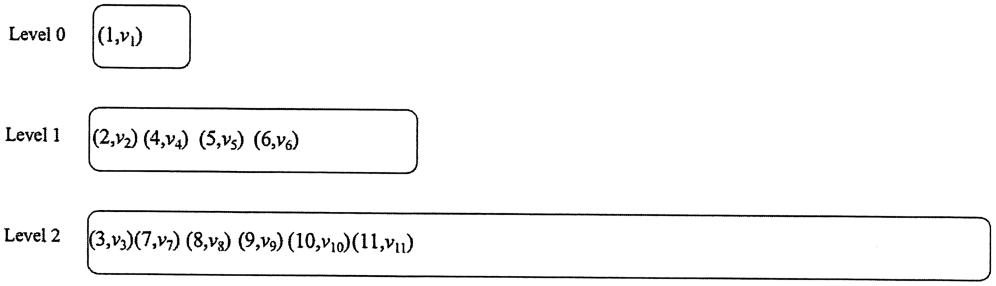

Figure Q5b

(c) Consider T=3 and M=2, and all the possible insertion orders in Q5(b). If we apply the same insertion order to a **tiering** LSM-tree which is initially empty, what are the possible tiering LSM-tree layouts after the insertions? Please draw all possible layouts and explain your answer. (Note: you need to describe each tier in Level-1 and Level-2 in your drawing.)

(10 marks)

(d) Consider T=4 and M=2, and insert the same sequence of key-value pairs into an empty leveling LSM-tree and an empty tiering LSM-tree. Design an insertion sequence such that after the insertions, there will be 6 key-value pairs in Level-1 of the tiering LSM-tree while only 2 key-value pairs in Level-1 of the leveling LSM-tree. Also, after the insertions both LSM-trees have 8 key-value pairs in their Level-2. Please explain your answer.

(7 marks)

END OF PAPER

CE4123 BIG DATA MANAGEMENT
CZ4123 BIG DATA MANAGEMENT
SC4023 BIG DATA MANAGEMENT

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.