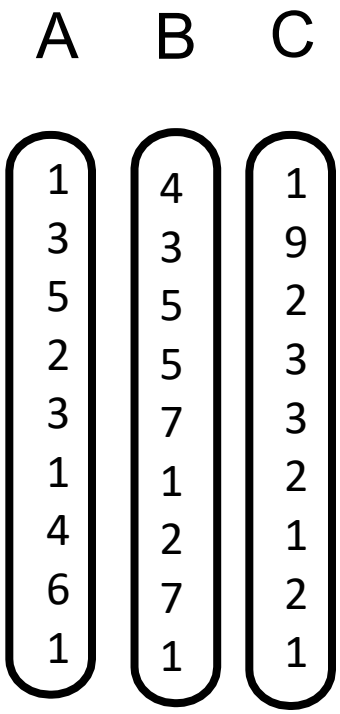
CE/CZ 4123 Big Data Management Tutorial 5
Column Stores

Question 1

Given column store table T as follow.

(1) Give the flow chart using "column at a time" for the query "***SELECT*** *min(B)* ***FROM*** *T* ***WHERE*** *A>3 and C>5*"

(2) Give the procedure using "column at a time" for the query "***SELECT*** *sum(C)* ***FROM*** *T* ***WHERE*** *A<10 and B>3*"



Question 2

Redo Question 1 using "vector at a time". Assume that the vector size is 3.

Question 3

Suppose we are querying a **Student's** information table with three columns **Name**, **Email**, **Age**. Given a query of the following form:
"***SELECT*** *Name* ***FROM*** *Student* ***WHERE*** *predicate (Email) and predicate (Age)*".
Here, a predicate applied on a column is a filtering function (e.g., **Email** ending with *ntu.edu.sg,* **Age**>19). We define the selectivity of a predicate by the percentage of

the qualified results in the corresponding column. Assume that the selectivity of predicate(**Email**) is p and the selectivity of predicate(**Age**) is q, where $0<p<1$, and $0<q<1$. Let page size be P. We assume column width $w < P$ and each value in a column is contained in a page. Consider two options in scanning columns: scanning **Email** first and scanning **Age** first.

(1) If the column widths are the same (denoted by w), please analyze which is better.

(2) If the width of **Email** is $2w$, and the widths for **Name** and **Age** are w, then which option is better?