UNIVERSITÉ Concordia UNIVERSITY

LAB ASSIGNMENT REPORT - 1

HARISH JAYASANKAR - 40105791

SAI CHARAN DUDUKA - 40103928

An Nguyen - 40087621

WINTER 2020

COURSE NAME: Advanced Database Technology and Applications

COURSE NUMBER: COMP 6521

INSTRUCTOR: N. Shiri

CONCORDIA UNIIVERSITY

PROJECT DESCRIPTION:

We have been given two input files which are of large dataset. The dataset consists of relational data in form of rows and columns and we need to merge two files without any duplicates based on Employee ID with main memory available as 5MB or 10MB.

AIM:

1. To efficiently merge two files T1 and T2 consists of data (500,000 tuples) each without any duplicates which is to run with 5 MB of main memory
2. To efficiently merge two files T1 and T2 consists of data (1000,000 tuples) each without any duplicates which is to run with 10 MB of main memory

DESIGN PHASE FOR ALGORITHM:

Phase 1:

In Phase 1 we divide the two input files into sub list which in our file is called sub lists which are sorted based on Employee ID. The sub lists contain number of records which are present in dataset. The buffer size is calculated based on available memory. When the buffer size limit is exceeded we will write sorted data to the disk.

To sort the data, we have used the comparator concept and each sub list consists of same amount of data.

Steps Followed:

- Fill buffer with data.
- Sort using best and efficient sorting algorithm.
- Write sorted into sub lists when the limit is reached.
- Repeat until all records put into sub list.

Phase 2:

In Phase 2 we divide Memory in (M-1) Buffers and 1 output buffer where M stands for Main Memory. And then use M-1 to read K sorted Chunks where K stands for sorted subsists. Then we merge the values one by one removing duplicates based on Employee ID. We do this merging operation until we are getting a single output file.

ANALYSIS:

Analysis for 5MB of main memory with 50,000tuples in each T1 and T2 Files (100,000 tuples):

Processing time for phase 1: **0.685s**
Processing time for phase 2: **1.073s**
Total Time: **1.759s**
No of output tuples: 5553
No of Sub lists: 9
No of I/O’s for phase 1: 18
No of I/O’s for phase 2: 26
Total no of I/O’s = 54

Analysis for 10MB of main memory with 50,000tuples in each T1 and T2 Files (100,000 tuples):

Processing time for phase 1:**0.441s**
Processing time for phase 2: **0.572s**
Total Time: **1.013s**
No of output tuples: 5553
No of Sub lists: 6
No of I/O’s for phase 1: 12
No of I/O’s for phase 2: 30
Total no of I/O’s = 42

Analysis for 5MB of main memory with 100,000tuples in each T1 and T2 Files (200,000 tuples):

Processing time for phase 1: **1.125s**
Processing time for phase 2: **2.615s**
Total Time: **3.823s**
No of output tuples: 5553
No of Sub lists: 15
No of I/O's for phase 1: 30
No of I/O's for phase 2: 40
Total no of I/O's = 75

Analysis for 10MB of main memory with 100,000tuples in each T1 and T2 Files (200,000 tuples):

Processing time for phase 1: **0.783s**
Processing time for phase 2: **2.12s**
Total Time: **2.903s**
No of output tuples: 5553
No of Sub lists: 9
No of I/O's for phase 1: 18
No of I/O's for phase 2: 18
Total no of I/O's = 36

Analysis for 5MB of main memory with 500,000tuples in each T1 and T2 Files (1,000,000 tuples):

Processing time for phase 1: **6.813s**
Processing time for phase 2: **11.957s**
Total Time: **18.78s**
No of output tuples:5553
No of Sub lists: 75
No of I/O's for phase 1: 150
No of I/O's for phase 2: 196
Total no of I/O's = 346

Analysis for 10MB of main memory with 500,000tuples in each T1 and T2 Files (1,000,000 tuples):

Processing time for phase 1: **9.282**
Processing time for phase 2: **5.081**
Total Time: **15.092**
No of output tuples:5553
No of Sub lists: 43
No of I/O's for phase 1: 86
No of I/O's for phase 2: 71
Total no of I/O's = 157

Analysis for 20MB of main memory with 500,000 tuples in each T1 and T2 Files (1,000,000 tuples):

Processing time for phase 1: **7.876**
Processing time for phase 2: **4.785**
Total Time: **12.681**
No of Sub lists: 21
No of output tuples: 5553
No of I/O's for phase 1: 42
No of I/O's for phase 2: 35
Total no of I/O's = 77

Analysis for 5MB of main memory with 1,000,000 tuples in each T1 and T2 Files (2,000,000 tuples):

Processing time for phase 1: **13.356s**
Processing time for phase 2: **23.59s**
Total Time: **36.964s**
No of Sub lists:150
No of output tuples: 5553
No of I/O's for phase 1: 300
No of I/O's for phase 2: 391
Total no of I/O's = 691

Analysis for 10MB of main memory with 1,000,000 tuples in each T1 and T2 Files (2,000,000 tuples):

Processing time for phase 1: **18.315s**
Processing time for phase 2: **11.233s**
Total Time: **29.543s**
No of Sub lists: 85
No of output tuples: 5553
No of I/O's for phase 1: 170
No of I/O's for phase 2: 143
Total no of I/O's = 313

Analysis for 20MB of main memory with 1,000,000 tuples in each T1 and T2 Files(2,000,000 tuples):

Processing time for phase 1: **15.439s**
Processing time for phase 2: **16.497s**
Total Time: **31.754s**
No of Sub lists: 41
No of output tuples: 5553
No of I/O's for phase 1: 82
No of I/O's for phase 2: 58
Total no of I/O's = 140

Junit:

We have implemented test cases for testing memory calculation method and final merged output file. All the test cases are **passed**.

Team Members Contribution:

We implemented the pair programming concept, we all three worked together as a team on coding and report.