Project Report - Implementing a Sorted File

Arpan Banerjee

UFID: 9359-9083

arpanbanerjee@ufl.edu

Krutantak Patil

UFID: 5615-6343

Krutantakb.patil@ufl.edu

1) Steps to compile and run - code, tests and gtests

_(_Assuming .bin files would be provided to us in <u>dbfiles</u> folder)

  i. **make main** – Command to build the main program.
  ii. **./main** – Command to run the main executable.
  iii. **make test.out** – Command to build the tests.
  iv. **./test.out** – Command to run the tests.
  v. **make bigQTest.out** – Command to build the gtests.
  vi. **./gtestScript.sh** – Command to run the gtests. (<u>gtest</u> folder contains data required for running 4 test cases written in BigQTest.cc file)

2) BigQ file implementation

Data members

  i. **int runlength** – Used to store length of each run.
  ii. **File runFile** – Used to store records and pages in the file.
  iii. **OrderMaker sortOrder** – Pointer to page holding recent record additions to the file.
  iv. **Pipe *in** – Pointer to produce records and add to pages from .bin files.
  v. **Pipe *out** – Pointer to consume sorted and merged records from .bin files.

Member functions

  i. **BigQ** ()

    a. _Default constructor used in BigQTest.cc to initailize BigQ object for testing_.

  ii. **BigQ** (Pipe &in, Pipe &out, OrderMaker &sortorder, int runLength)

    a. Paramaterized constructor called to read data from input pipe, sort it using two-pass multiway merge algorithm and then store using output pipe.
    b. A worker thread is created using pthread in this constructed.

  iii. **worker** (void *workerThread)

    a. It is a worker routine which first reads data from input pipe and computes sorted runs. Then, using priority queue, it performs k-way merging and stores back using the output pipe.
    b. runCount is initialized to 0.
    c. overflow map is used to store overflow papes.

iv.   **initFile** (File &runfile)

    a. This function is used to create and open an in memory file used to stored sorted runs.

v.   **tpmmsPhase1** (Pipe *in, OrderMaker *sortOrder, int &runCount, int runLength, File &runFile, map<int,Page*> &overflow)

    a. Its the phase 1 of Two-Pass Multiway Merge Sort Algorithm.
    b. Records are retrieved from the inpute pipe and added to pages until the page is full. Once the page is full, the next record is added to a new page and page counter is incremented by 1.
    c. If, either there are no more records in the inpute pipe or page count equals runLength, then pages sorted and stores in in-memory file as runs.

vi.   **generateRuns** (vector<Record *> &records, int &runCount, int runLength, File &runFile, map<int,Page*> &overflow)

    a. Records are sorted using Phase1Compare operator and stored as runs.
    b. After sorting and generating runs, if, pageCount is equal to or more than runlength, then pages are stored in the overflowmap.

vii.   **tpmmsPhase2** (Pipe *out, OrderMaker *sortOrder, int runCount, int runLength, File &runFile, map<int,Page*> overflow)

    a. Its the phase 2 of Two-Pass Multiway Merge Sort Algorithm.
    b. Priority queue is used to merge k-runs.
    c. Runs are merged using Phase2Compare operator and then written to output pipe.
    d. K-way merging algorithm is used to read runs and merge them

viii.   **Phase1Compare**(OrderMaker *inSortOrder)

a. It is used to sort records using CNF order and ComparisonEngine class.

   ix.   **Phase2Compare**(OrderMaker *inSortOrder)

a. It is used to merge records using CNF order and ComparisonEngine class.

3) ./test.out results for 1GB data

  i.   **./runTestCases.sh**

```
C BigQ.h      G BigQ.cc      G BigQTest.cc      ≡ output1.txt ×

a2test › ≡ output1.txt
  1   n_nationkey: [6], n_name: [FRANCE], n_regionkey: [3], n_comment: [refully final requests. regular, ironi]
  2   n_nationkey: [19], n_name: [ROMANIA], n_regionkey: [3], n_comment: [ular asymptotes are about the furious multipliers. express dependencies nag above the ironically ironic account]
  3   n_nationkey: [22], n_name: [RUSSIA], n_regionkey: [3], n_comment: [ requests against the platelets use never according to the quickly regular pint]
  4   n_nationkey: [23], n_name: [UNITED KINGDOM], n_regionkey: [3], n_comment: [eans boost carefully special requests. accounts are. carefull]
  5   n_nationkey: [13], n_name: [JORDAN], n_regionkey: [4], n_comment: [ic deposits are blithely about the carefully regular pa]
  6   n_nationkey: [4], n_name: [EGYPT], n_regionkey: [4], n_comment: [y above the carefully unusual theodolites. final dugouts are quickly across the furiously regular d]
  7   n_nationkey: [11], n_name: [IRAQ], n_regionkey: [4], n_comment: [nic deposits boost atop the quickly final requests? quickly regula]
  8   n_nationkey: [20], n_name: [SAUDI ARABIA], n_regionkey: [4], n_comment: [ts. silent requests haggle. closely express packages sleep across the blithely]
  9   n_nationkey: [10], n_name: [IRAN], n_regionkey: [4], n_comment: [efully alongside of the slyly final dependencies. ]
 10   consumer: removed 25 recs from the pipe
 11   ********************************************************************************************************************************
 12   producer: opened DBFile dbfiles/customer.bin
 13   producer: inserted 150000 recs into the pipe
 14   consumer: removed 150000 recs from the pipe
 15   ********************************************************************************************************************************
 16   o_orderkey: [4515876], o_custkey: [100685], o_orderstatus: [F], o_totalprice: [510062], o_orderdate: [1993-11-02], o_orderpriority: [4-NOT SPECIFIED], o_clerk: [Clerk#000000105], o_shippriority: [0], o_comment: [carefully accounts: slyly i
 17   o_orderkey: [2185667], o_custkey: [51796], o_orderstatus: [F], o_totalprice: [511360], o_orderdate: [1992-10-08], o_orderpriority: [1-URGENT], o_clerk: [Clerk#000000574], o_shippriority: [0], o_comment: [. deposits wake quickly unusual dep
 18   o_orderkey: [2199712], o_custkey: [66790], o_orderstatus: [O], o_totalprice: [515532], o_orderdate: [1996-09-30], o_orderpriority: [2-HIGH], o_clerk: [Clerk#000000650], o_shippriority: [0], o_comment: [ the final, ironic deposits inte]
 19   o_orderkey: [3586919], o_custkey: [24049], o_orderstatus: [F], o_totalprice: [522644], o_orderdate: [1992-11-07], o_orderpriority: [1-URGENT], o_clerk: [Clerk#000000924], o_shippriority: [0], o_comment: [are alongside of the pending deposi
 20   o_orderkey: [2232932], o_custkey: [13940], o_orderstatus: [O], o_totalprice: [522721], o_orderdate: [1997-04-13], o_orderpriority: [2-HIGH], o_clerk: [Clerk#000000245], o_shippriority: [0], o_comment: [lithely stealthy accounts are slyly a
 21   o_orderkey: [4576548], o_custkey: [108931], o_orderstatus: [O], o_totalprice: [525591], o_orderdate: [1997-12-26], o_orderpriority: [1-URGENT], o_clerk: [Clerk#000000336], o_shippriority: [0], o_comment: [he slyly ironic requests. regular,
 22   o_orderkey: [3043270], o_custkey: [144617], o_orderstatus: [O], o_totalprice: [530604], o_orderdate: [1997-02-12], o_orderpriority: [5-LOW], o_clerk: [Clerk#000000699], o_shippriority: [0], o_comment: [riously final deposits? ]
 23   o_orderkey: [4722021], o_custkey: [128120], o_orderstatus: [F], o_totalprice: [544089], o_orderdate: [1994-04-07], o_orderpriority: [1-URGENT], o_clerk: [Clerk#000000230], o_shippriority: [0], o_comment: [al, express pinto beans are after
 24   o_orderkey: [1750466], o_custkey: [21433], o_orderstatus: [F], o_totalprice: [555285], o_orderdate: [1992-11-30], o_orderpriority: [4-NOT SPECIFIED], o_clerk: [Clerk#000000040], o_shippriority: [0], o_comment: [ ironic packages. even notor
 25   consumer: removed 1500000 recs from the pipe
 26   ********************************************************************************************************************************
 27   n_nationkey: [7], n_name: [GERMANY], n_regionkey: [3], n_comment: [l platelets. regular accounts x-ray: unusual, regular acco]
 28   n_nationkey: [19], n_name: [ROMANIA], n_regionkey: [3], n_comment: [ular asymptotes are about the furious multipliers. express dependencies nag above the ironically ironic account]
 29   n_nationkey: [22], n_name: [RUSSIA], n_regionkey: [3], n_comment: [ requests against the platelets use never according to the quickly regular pint]
 30   n_nationkey: [23], n_name: [UNITED KINGDOM], n_regionkey: [3], n_comment: [eans boost carefully special requests. accounts are. carefull]
 31   n_nationkey: [4], n_name: [EGYPT], n_regionkey: [4], n_comment: [y above the carefully unusual theodolites. final dugouts are quickly across the furiously regular d]
 32   n_nationkey: [10], n_name: [IRAN], n_regionkey: [4], n_comment: [efully alongside of the slyly final dependencies. ]
 33   n_nationkey: [11], n_name: [IRAQ], n_regionkey: [4], n_comment: [nic deposits boost atop the quickly final requests? quickly regula]
 34   n_nationkey: [13], n_name: [JORDAN], n_regionkey: [4], n_comment: [ic deposits are blithely about the carefully regular pa]
 35   n_nationkey: [20], n_name: [SAUDI ARABIA], n_regionkey: [4], n_comment: [ts. silent requests haggle. closely express packages sleep across the blithely]
 36   consumer: removed 25 recs from the pipe
 37
```

5) gTest results:

i. **TEST(BigQTest, GenerateRuns_OrderTbl) –**

a. This function produces records from gtests/orders.bin file, and test whether runs are created and written to in-memory file. Here we check it for records for multiple and overflow pages.

ii. **TEST(BigQTest, GenerateRuns_NationTbl) –**
   a. This function produces records from gtests/nation.bin file, and test whether runs are created and written to in-memory file. Here we check it for records for a single page.

iii. **TEST(BigQTest, Phase1Compare) –**
   a. This function assert sorting of records using Phase1Compare operator. Records are produced from gtests/actual.bin file and then sorted. Here we match the sorted record vector with the expected record vector produced using gtests/expected.bin file.

iv. **TEST(BigQTest, Phase2Compare) –**
   a. This function assert sorting of records using Phase2Compare operator. Records are produced from gtests/actual.bin file and then sorted. Here we match the sorted record vector with the expected record vector produced using gtests/expected.bin file. Since the sorting techinque is not consistent, we ASSERT_FALSE here.

(ensure that gtest folder contains actual.bin, expected.bin, nation.bin, & orders.bin files required for testing BigQ.cc file)

```
arpan@arpan-laptop:~/Desktop/Krutantak/Git/database-system-implementation/a2test$ ./gtestScript.sh

** IMPORTANT: MAKE SURE THE INFORMATION BELOW IS CORRECT **
 catalog location:      catalog
 tpch files dir:        tpch-dbgen/
 heap files dir:        dbfiles/

specify sort ordering (when done press ctrl-D):

[==========] Running 4 tests from 1 test case.
[----------] Global test environment set-up.
[----------] 4 tests from BigQTest
[ RUN      ] BigQTest.GenerateRuns_CustomerTbl
[       OK ] BigQTest.GenerateRuns_CustomerTbl (11 ms)
[ RUN      ] BigQTest.GenerateRuns_NationTbl
[       OK ] BigQTest.GenerateRuns_NationTbl (0 ms)
[ RUN      ] BigQTest.Phase1Compare
[       OK ] BigQTest.Phase1Compare (0 ms)
[ RUN      ] BigQTest.Phase2Compare
[       OK ] BigQTest.Phase2Compare (0 ms)
[----------] 4 tests from BigQTest (11 ms total)

[----------] Global test environment tear-down
[==========] 4 tests from 1 test case ran. (11 ms total)
[  PASSED  ] 4 tests.
arpan@arpan-laptop:~/Desktop/Krutantak/Git/database-system-implementation/a2test$
```