CSC443 Assignment2 Report

Weijun Zeng(zengwei)
Xiaoyi Zhao(zhaoxi30)

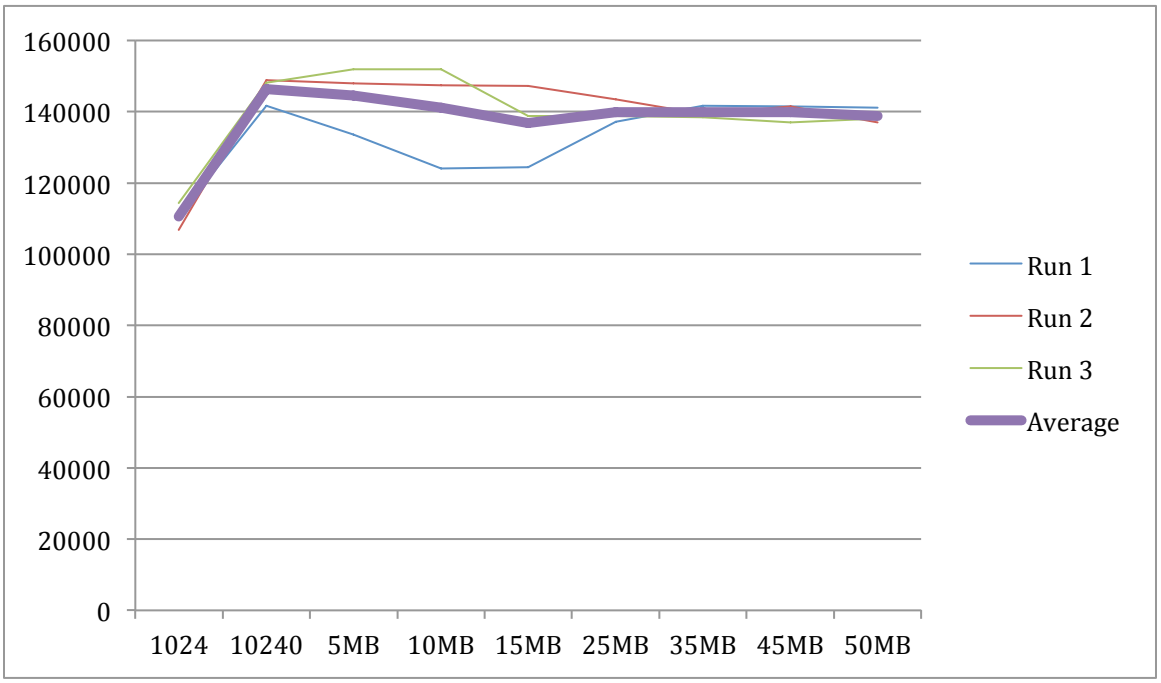
2.3 Record Serialization

According to PIAZZA, we are not required to deliver an answer for this part.

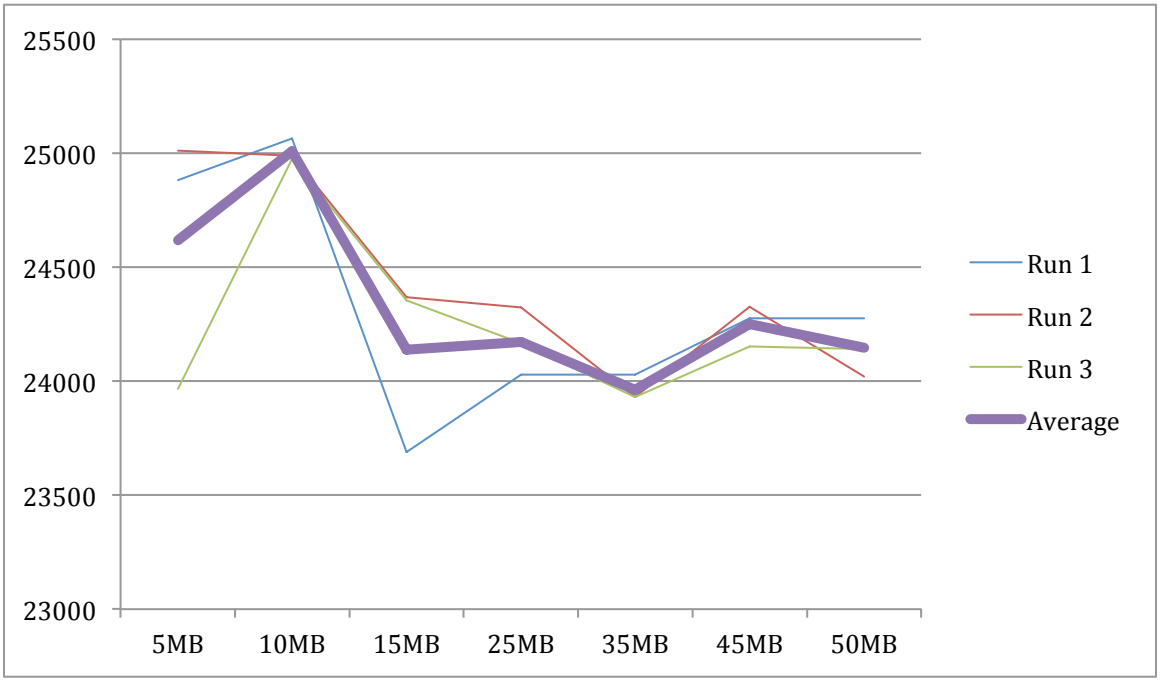
3.2 Page Layout Experiment

- **Plot the performance (records / second) versus page size for write and read.**

Write (records/s vs page size)



Read (records/s vs page size)



- **Discuss why page based format is superior to storing records using a CSV file.**

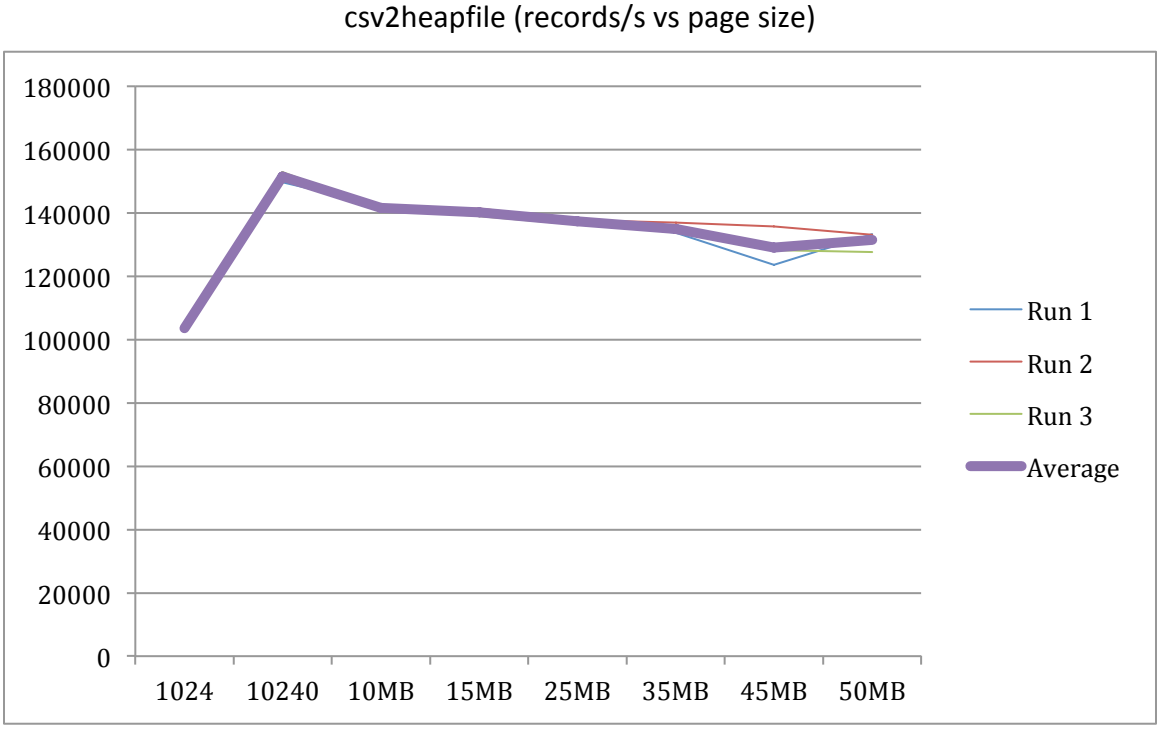
Page based format takes advantage at data management. It has directory pages, which can help perform all of the operations (insert, delete, search, update) quicker. Also, generating 50,000 tuples of records of CSV file takes more than 5 minutes while write 50,000 records to page file only takes couple of seconds.

- Discuss the shortcomings of the way we organize pages.

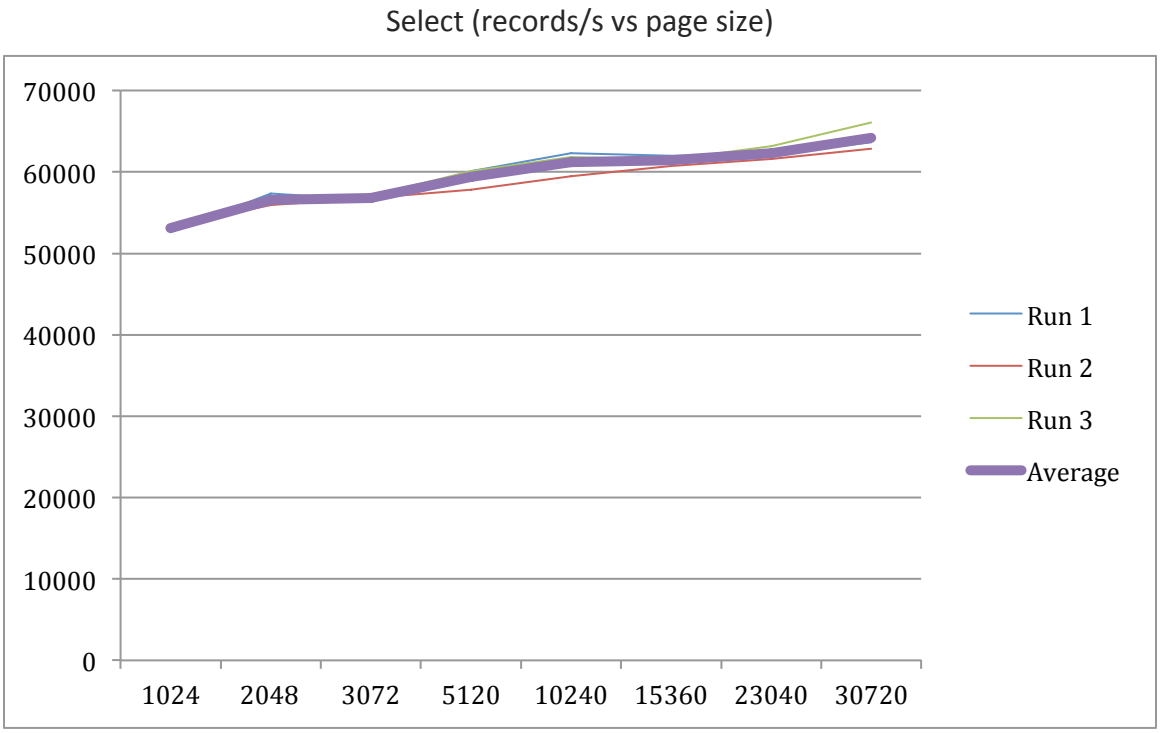
It may be over simplified since the structure is very simple. Also everything is fixed such as attribute length, attribute counts. It is not flexible. It cannot combine pages after deletion. Slot is left empty after a deletion is performed. So the size of the file is fixed even if there aren't that much of data in the file.

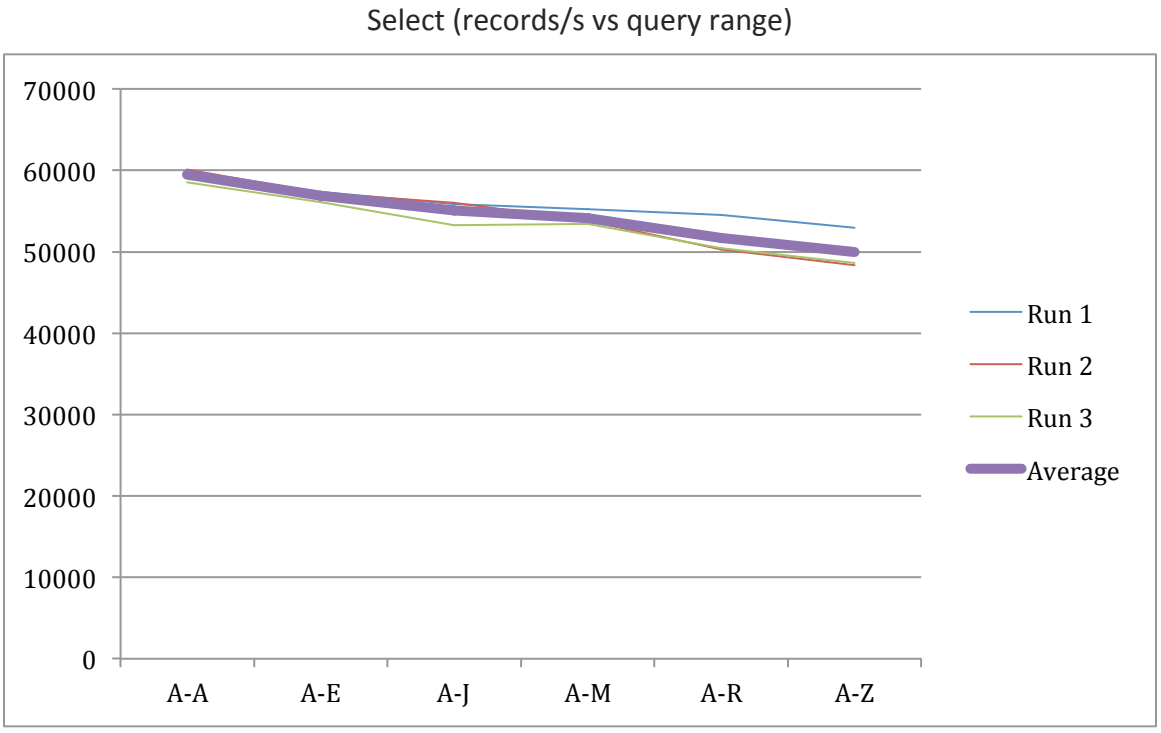
4.3 Heap File Experiment

- Measure the performance of csv2heapfile, comment on how the page size affects the performance of load.



- Measure the performance of the query versus page size.





- Comment on the choice of page size and the effects of the range from *start* and *end* on the performance of the query.

First graph plots the performance for SELECT operation versus different page sizes. As you can see, the performance is the lowest at the smallest page size – 1024bytes. And it is steadily increasing. The performance is the highest at the largest page size – 30720bytes. From the plot, we can conclude that the larger the page size is, the better the performance this select query is.

Second graph plots the performance of SELECT operation versus different range of start and end. As you can see, the performance is the highest at the smallest range – AAAAAAAAAA to AAAAAAAAAA. And it is steadily decreasing. The performance is the lowest at the largest range – AAAAAAAAAA to ZZZZZZZZZZ. From the plot, we can conclude that the smaller the range is, the better the performance it is.