CS511 SP24 Final Project Update 2

Author: Yangchen Ye, Xixiang Liu, Zihan Shan

This week's vector

This week's vector is: does column sketch work well under micro-benchmark settings on dataset that most fit its strength. This will be a starting point and baseline to conduct extended benchmarks in end-to-end settings and different types of dataset.

This week's plan

To answer the question, we implemented the evaluation infrastructure of the project, which is a general scanner framework with predicate evaluation capability. Specifically we implemented the sacnner algorithm described in the column sketch paper based on manual SIMD.

In addition, to prepare for the microbenchmark and end-to-end experiment, we started building a custom tool datamimic, which is able to generate parquet dataset with different statistical property (distribution, NDV, sortedness, etc).

This week's result

We run micro benchmark experiments comparing performance gain from using column sketch to evaluate predicate on a 500MB floating point dataset and got promising results:

```
Benchmark f64, selectivity=0.25/Baseline/0
                        time:   [171.69 ms 172.10 ms 172.59 ms]
Benchmark f64, selectivity=0.25/Column sketch(No SIMD)/0
                        time:   [70.211 ms 70.321 ms 70.439 ms]
Benchmark f64, selectivity=0.25/Column sketch (SIMD)/0
                        time:   [28.495 ms 28.567 ms 28.634 ms]
```

The infrastructure and evaluation can be found at project repository

In addition, we worked out the first prototype of datamimic, which could be used to generate zipf distributed dataset.

Next week's vector

Now that we've got a minimal viable product of column sketch, next week's vector would be to come up with reasonable first-step integration with apache parquet. Specifically we will try out several serialization/deserialization design of the column sketch data structure and compressed code. Next week's focus would be to come up with a reasonable serialization plan and study the storage overhead.

Next week's plan

We will start with a 'hello-world' style prototype. We will implement a binary which accepts a parquet table, apply column sketch on the column and output a new parquet table with extra columns which are compressed code. We will also emit a separate binary file containing the serialized compression map for the columns. We will study the effectiveness of this scheme, and if it works well, we could implement better integration by merging the data structure file into custom parquet pages.

One sentence per person

Xixiang Liu: I reviewed the background materials again, and I scrutinized the code written by my teammates to get ready for my work in the future.

Zihan Shan: I build the basic version of testing data generator, which is able to generate parquet file filling with columns of data following given distribution type like normal distribution. I will work on combining this part into our testing framenwork.

Yangchen Ye: I built the evaluation infrastructure and reproduced the SIMD scanner, and conducted micro benchmark about the performance of column sketch.