Smoke kills storage space

Michael Sklar Chess

Christophe Joseph Rimann

Nigel Schuster

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Project Idea

- Smoke is an efficient lineage engine
  - Lineage is the process of identifying the source of an aggregate result
  - Fotis built this awesome engine
- Lineage requires significant storage cost
  - Reverse List compression
  - Bitmap compression

Hypothesis

- Zipf Distribution:
  - Primary research:
    - Inverted List:
      - Size reduced by 8x
        - Wang 2017
      - Time remains similar
  - Stretch goal:
    - Bitmap:
      - Consistent performance

Work to be done

- Build a test harness to benchmark functionality with Smoke test data (Zipf Distribution, TPC-C)
- Use this harness to test compression algorithm
    - SIMDPforDelta
    - SIMD BP128
    - ...
    - We will each test a selection of the compression algorithms outlined by the Wang 2017 paper
- Integrate into Smoke

Necessary Resources

- Smoke source code
- Computing Resources for testing purposes:
  - Server-class machine (Ubuntu 14.04, 64GiB 2133MHz DDR4, 3.1GHz Intel Xeon E5-1607 v4)
    - We will likely need google cloud credit for this
  - MacBook Pro (macOS Sierra 10.12.3, 8GiB 1600MHz DDR3, 2.9GHz Intel Core i7)
    - We will attempt to approximate this hardware using our own machines
  - These machines are the same used in the original Smoke paper

Works Cited

http://db.ucsd.edu/wp-content/uploads/2017/03/sidm338-wangA.pdf

http://www.cs.columbia.edu/~fotis/pubs/techr/smoke_extended.pdf

https://github.com/lemire/SIMDCompressionAndIntersection