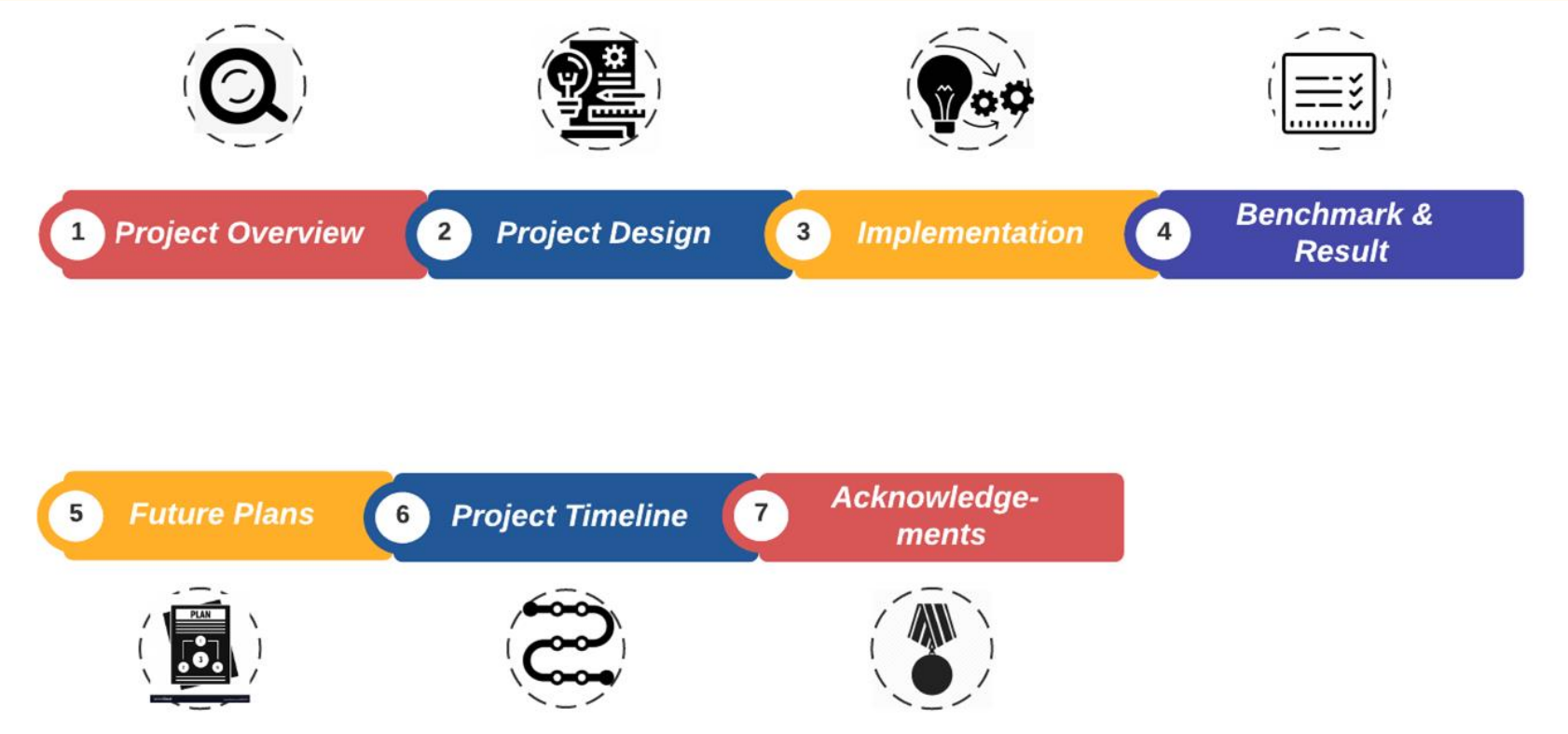
Efficient Web Archive Searching

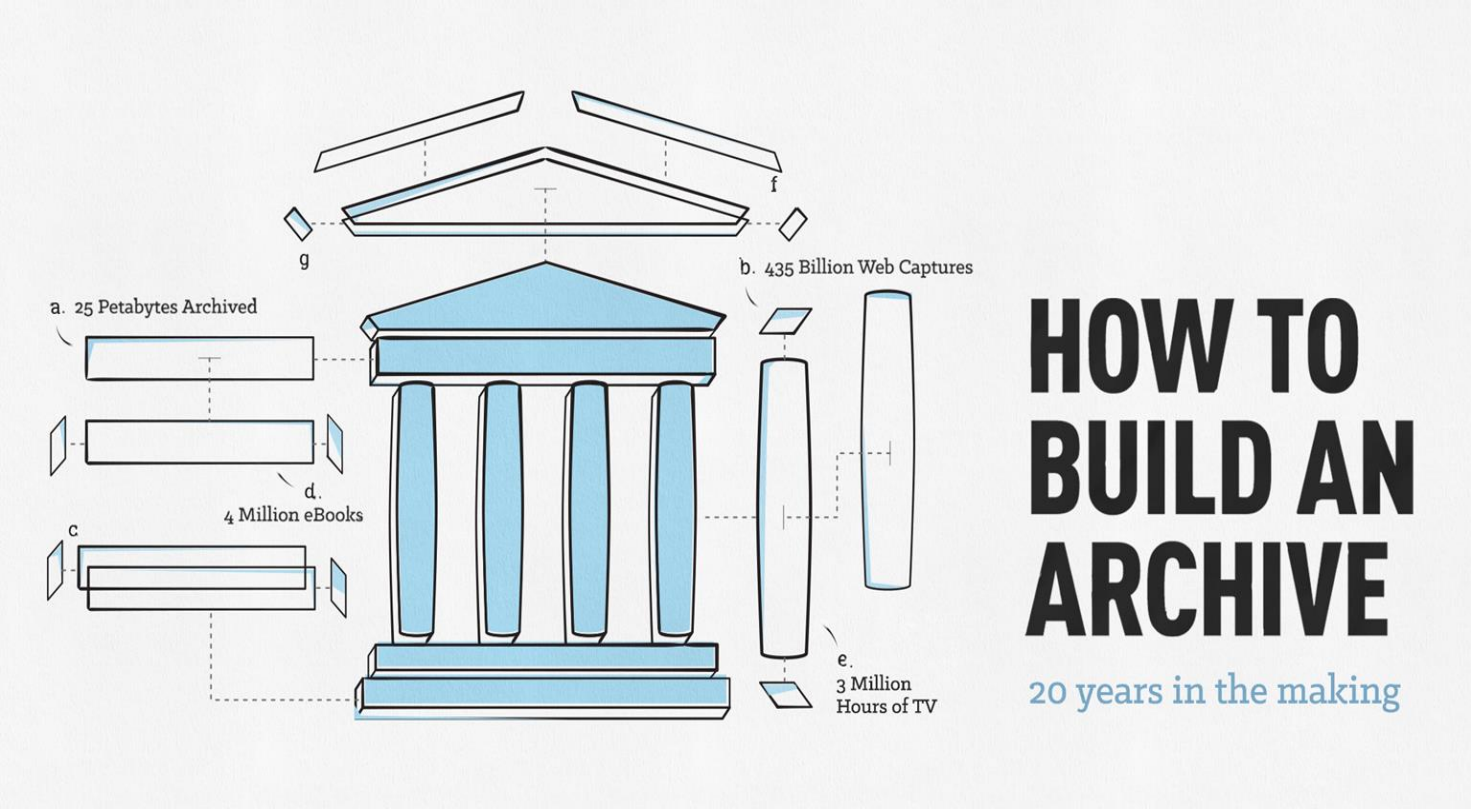
CS4624 - Multimedia/Hypertext
Dr. Edward Fox

Ming Cheng, Xiaolin Zhou, Jinyang Li, Yijing Wu, Lin Zhang
Virginia Tech, Blacksburg VA 24061
April 28, 2020

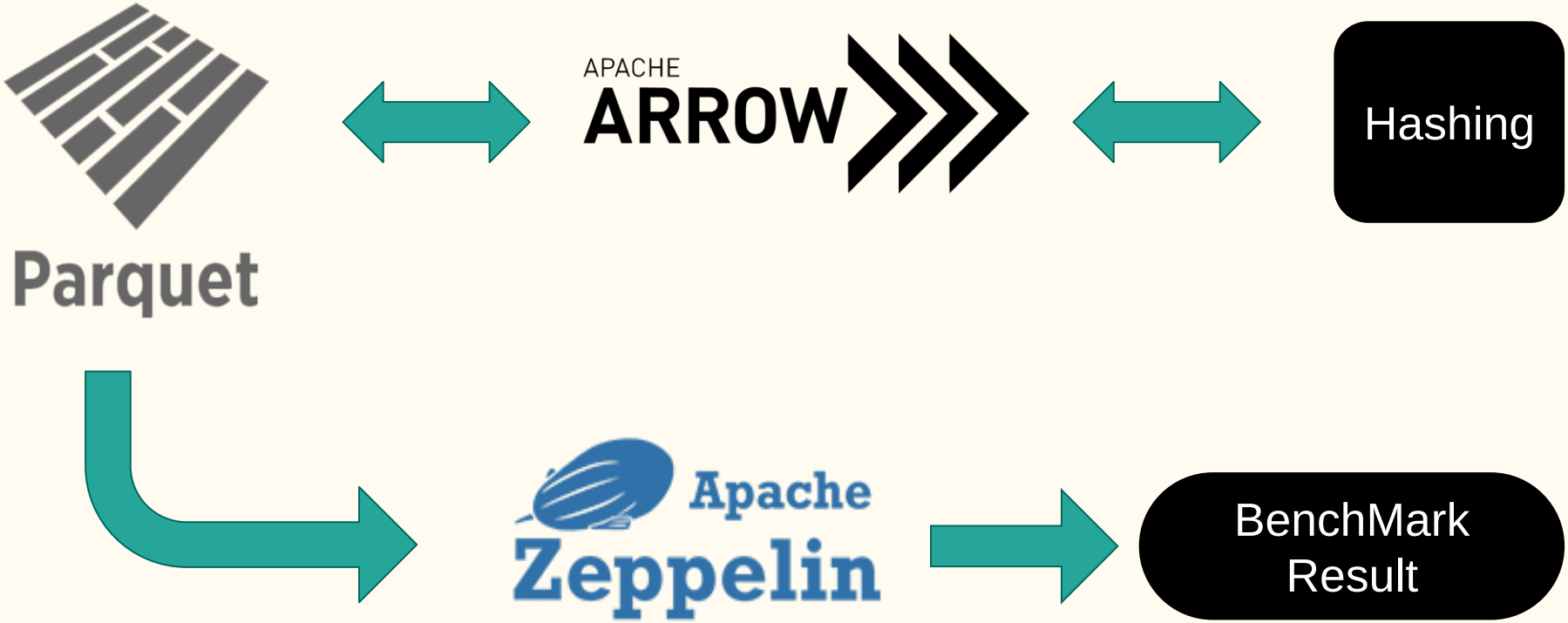
Outline


1 Project Overview
2 Project Design
3 Implementation
4 Benchmark & Result
5 Future Plans
6 Project Timeline
7 Acknowledge-ments
PLAN

Project Overview



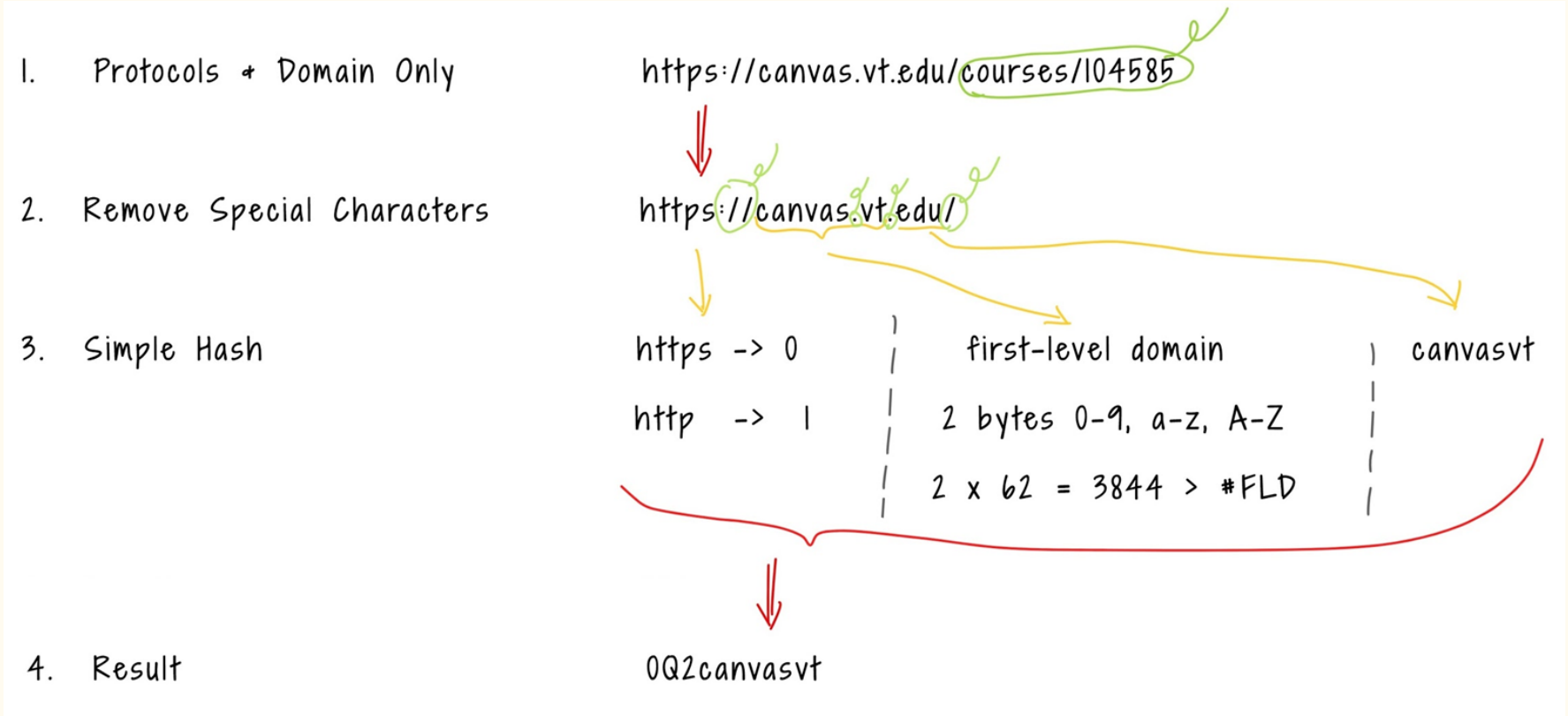
Project Design



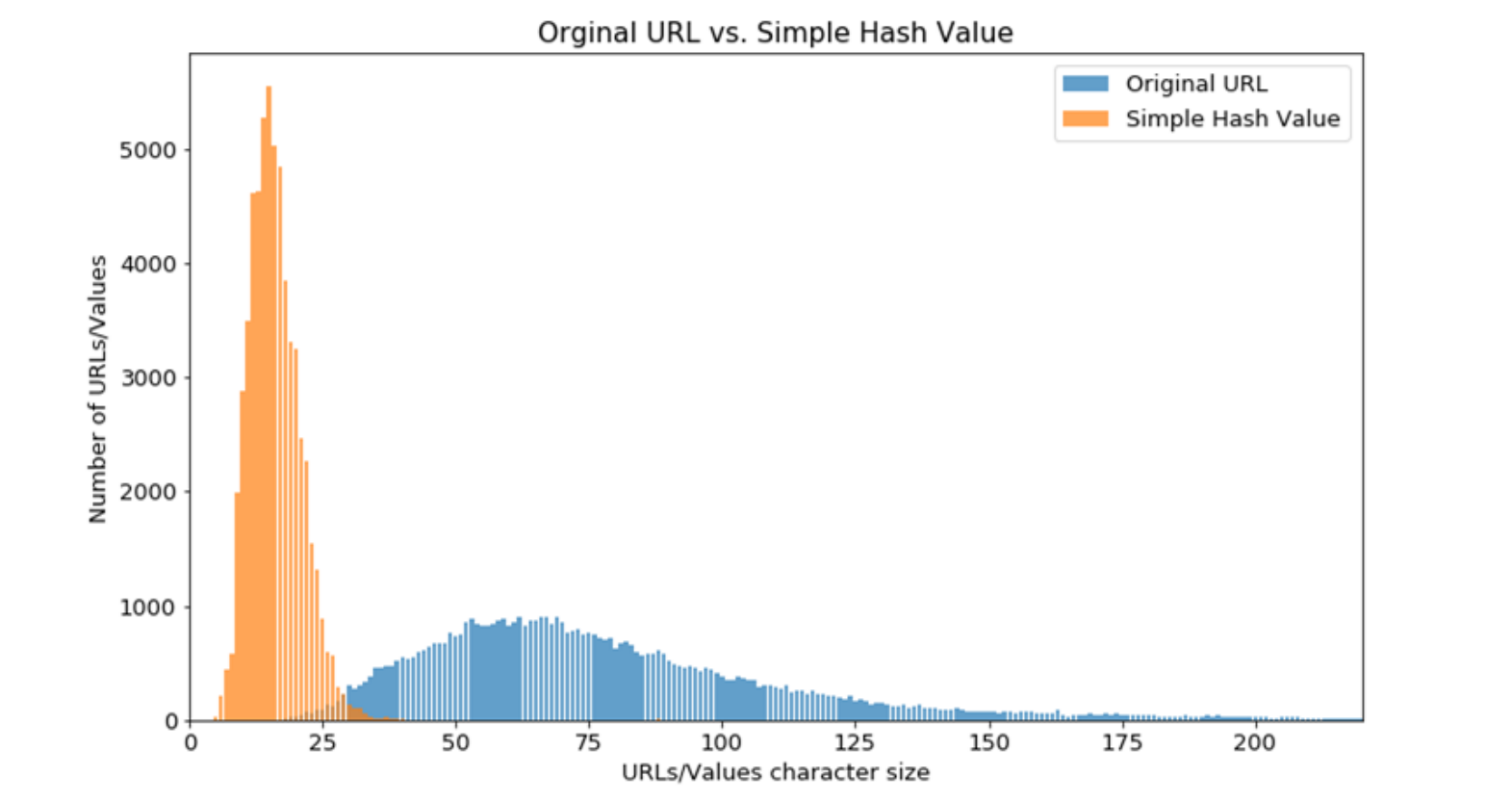
Implementations

- Our Implementation:
  - Simple Hash
- Existing algorithms:
  - Spooky Hash
  - Murmur Hash
  - City Hash
  - XXHash

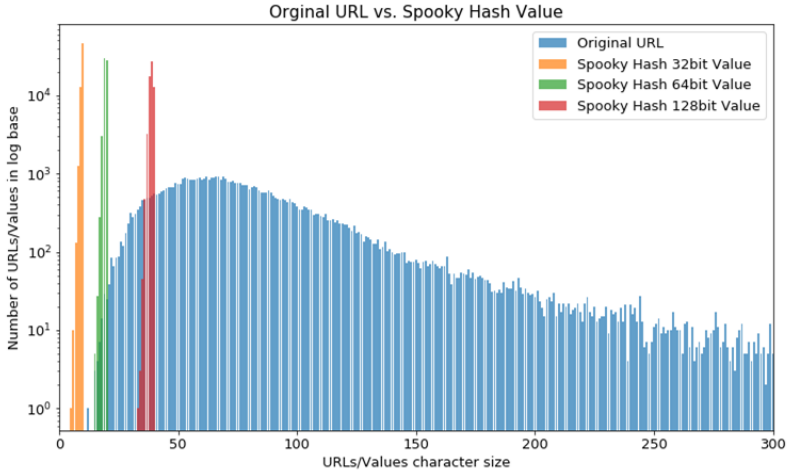
Simple Hash Algorithm Breakdown

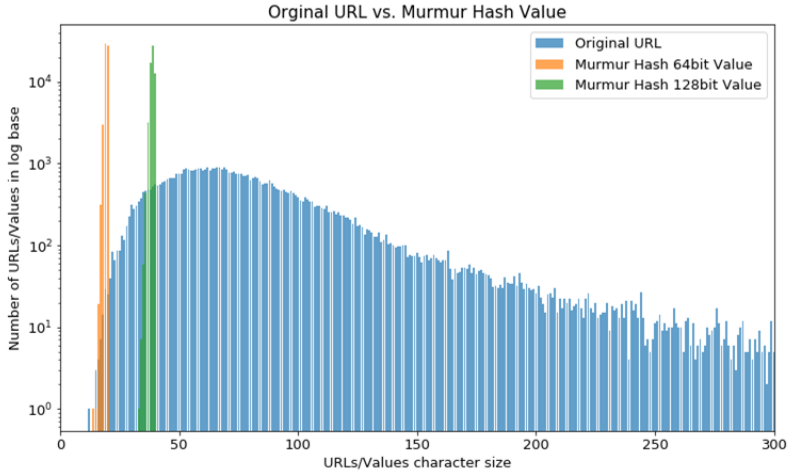


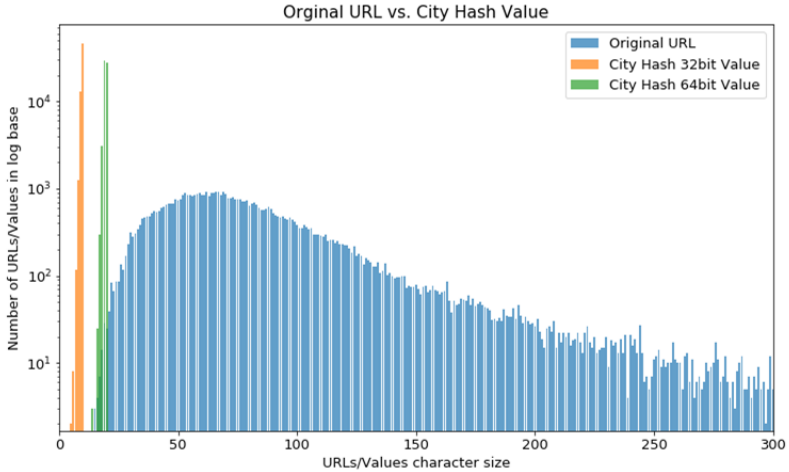
Simple Hash

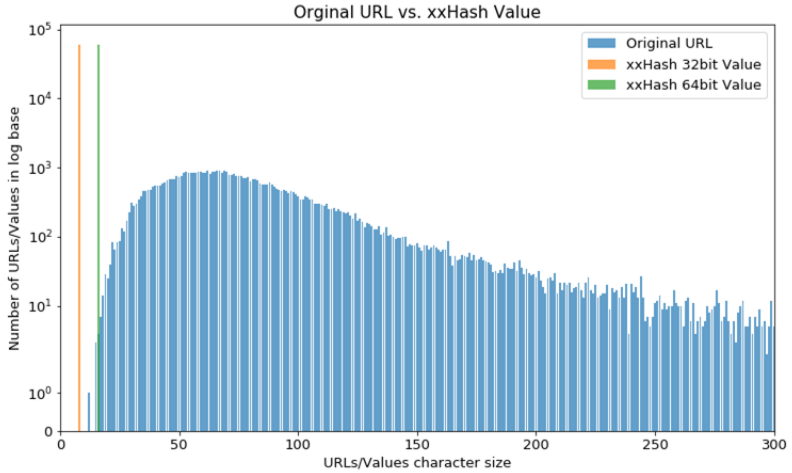

Orginal URL vs. Simple Hash Value
Original URL
Simple Hash Value
Number of URLs/Values
0
1000
2000
3000
4000
5000
0
25
50
75
100
125
150
175
200
URLs/Values character size

Spooky Hash, Murmur Hash, City Hash, XX Hash

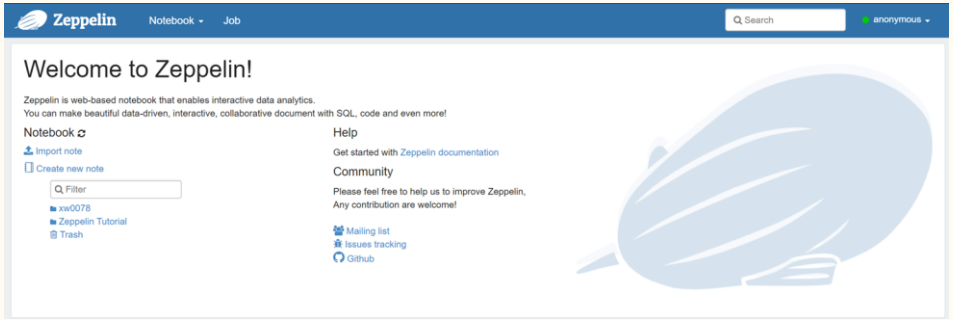








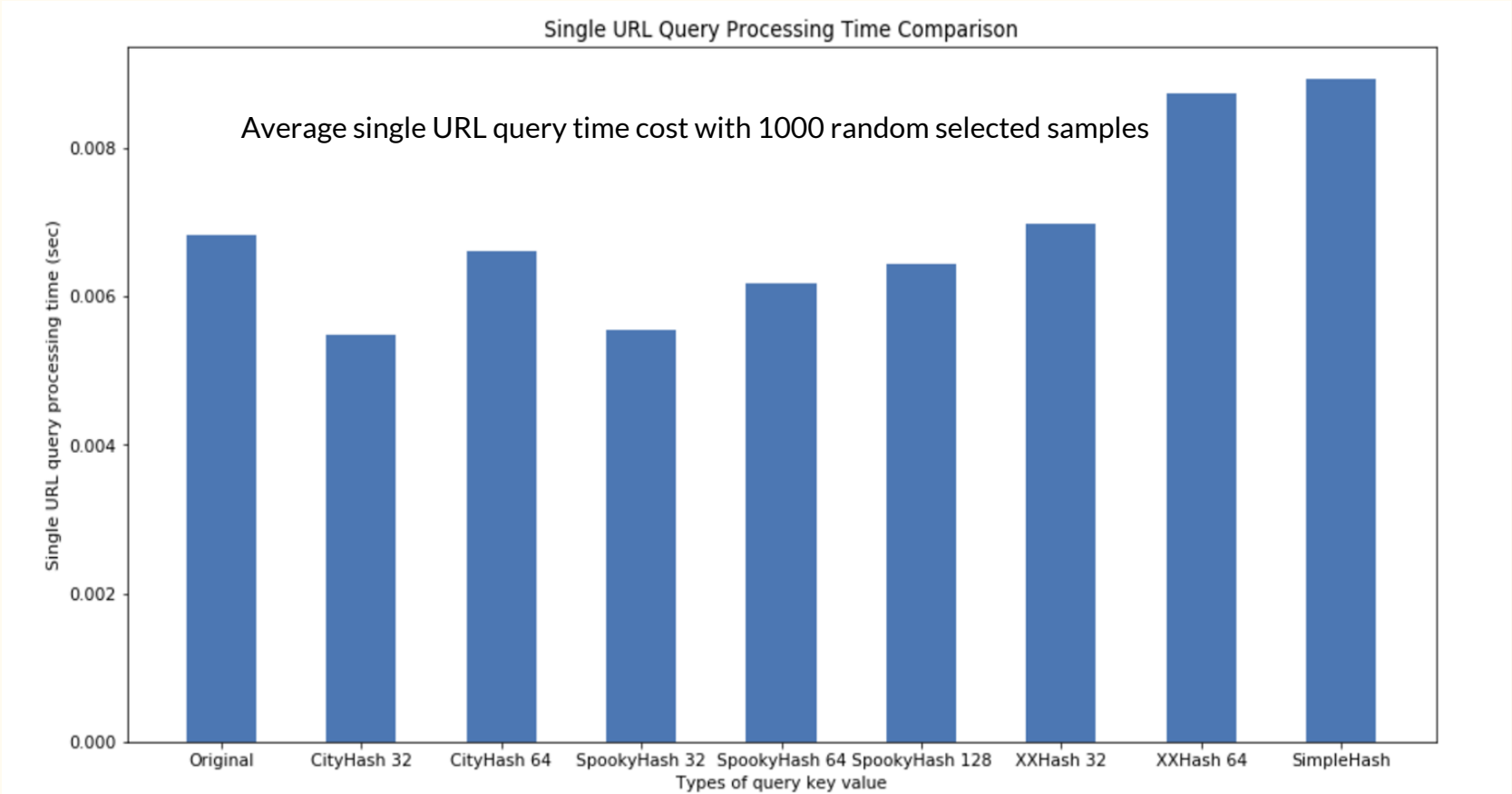
Benchmark



Tool: Zeppelin

1. All hash results + original database → Combined.parquet
2. Run Zeppelin Web Server, and go to: localhost:8080
3. Load parquet file, random pick 1000 samples
4. Clean cache, start query by originalURL and hash results
5. Calculate average single URL query time

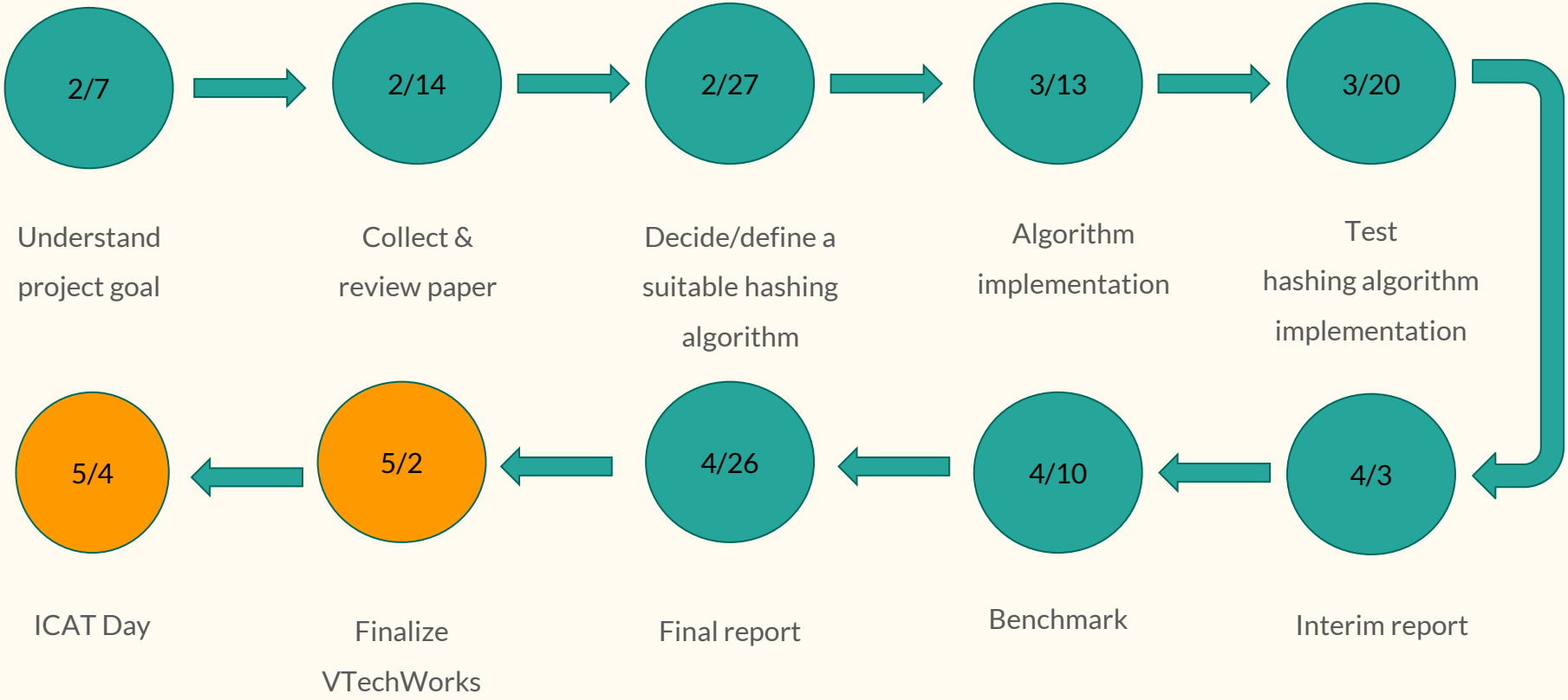
Result


Single URL Query Processing Time Comparison
Average single URL query time cost with 1000 random selected samples
Single URL query processing time (sec)
0.000
0.002
0.004
0.006
0.008
Original
CityHash 32
CityHash 64
SpookyHash 32
SpookyHash 64
SpookyHash 128
XXHash 32
XXHash 64
SimpleHash
Types of query key value

Future Plans

1. Reduce collisions.
2. Larger benchmark database.
3. Run benchmark multiple times and take the average.

Project Timeline



Acknowledgements

Xinyue Wang

Dr. Edward A. Fox

Dr. Lenwood S. Heath

Reference

Miguel C, Daniel G, Francisco C, and Mário S. (2013). A survey of web archive search architectures. Retrieved From :https://doi.org/10.1145/2487788.2488116

Helge H, Vinay G, and Avishek A. (2016). ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. Retrieved From: https://doi.org/10.1145/2910896.2910902

Ben-David, A., & Huurdeman, H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. Retrieved From: https://doi.org/10.7227/ALX.0022

Elliot J and Scott K. (2009). Architecture of the internet archive. Retrieved From: https://doi.org/10.1145/1534530.1534545


Q
A