

Efficient Web Archive Searching

CS 4624: Multimedia, Hypertext, and Information Access
Dr. Edward Fox

Ming Cheng, Xiaolin Zhou, Jinyang Li, Yijing Wu, Lin Zhang
Virginia Tech, Blacksburg, VA 24061
May 7, 2020

Outline



1 *Project Overview*

2 *Project Design*

3 *Implementation*

4 *Benchmark &
Result*

5 *Future Plans*

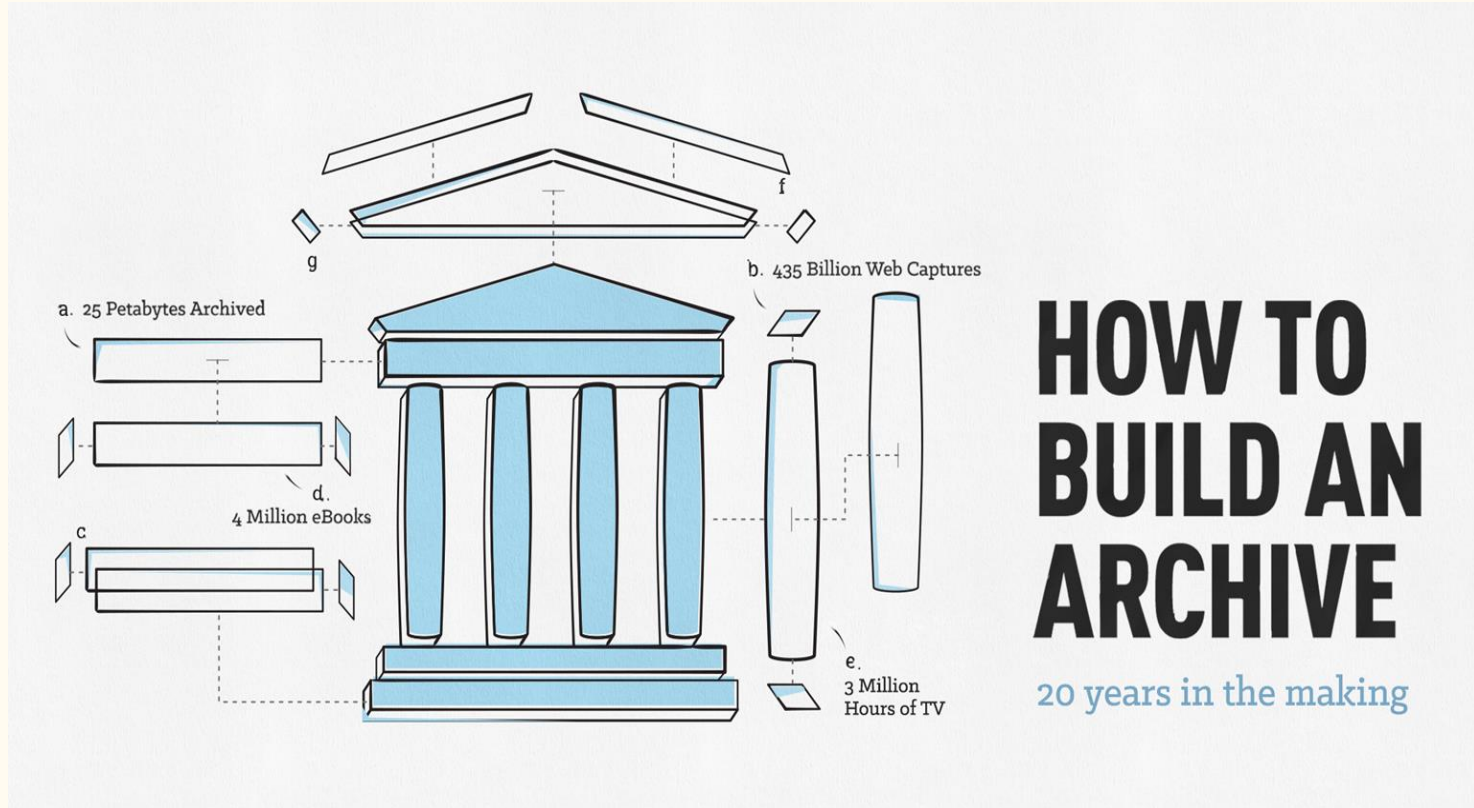
6 *Project Timeline*

7 *Acknowledge-
ments*

8 *Reference*



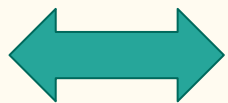
Project Overview



Project Design



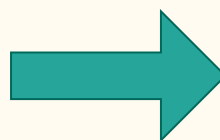
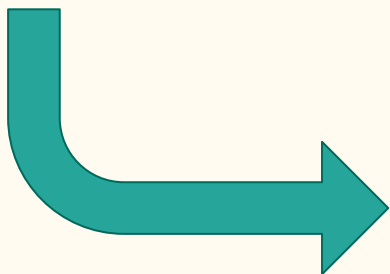
Parquet



APACHE
ARROW



Hashing



BenchMark
Result

Implementation

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



Simple Hash Algorithm Breakdown

1. Protocols + Domain Only

`https://canvas.vt.edu/courses/104585`

2. Remove Special Characters

`https://canvas.vt.edu/`

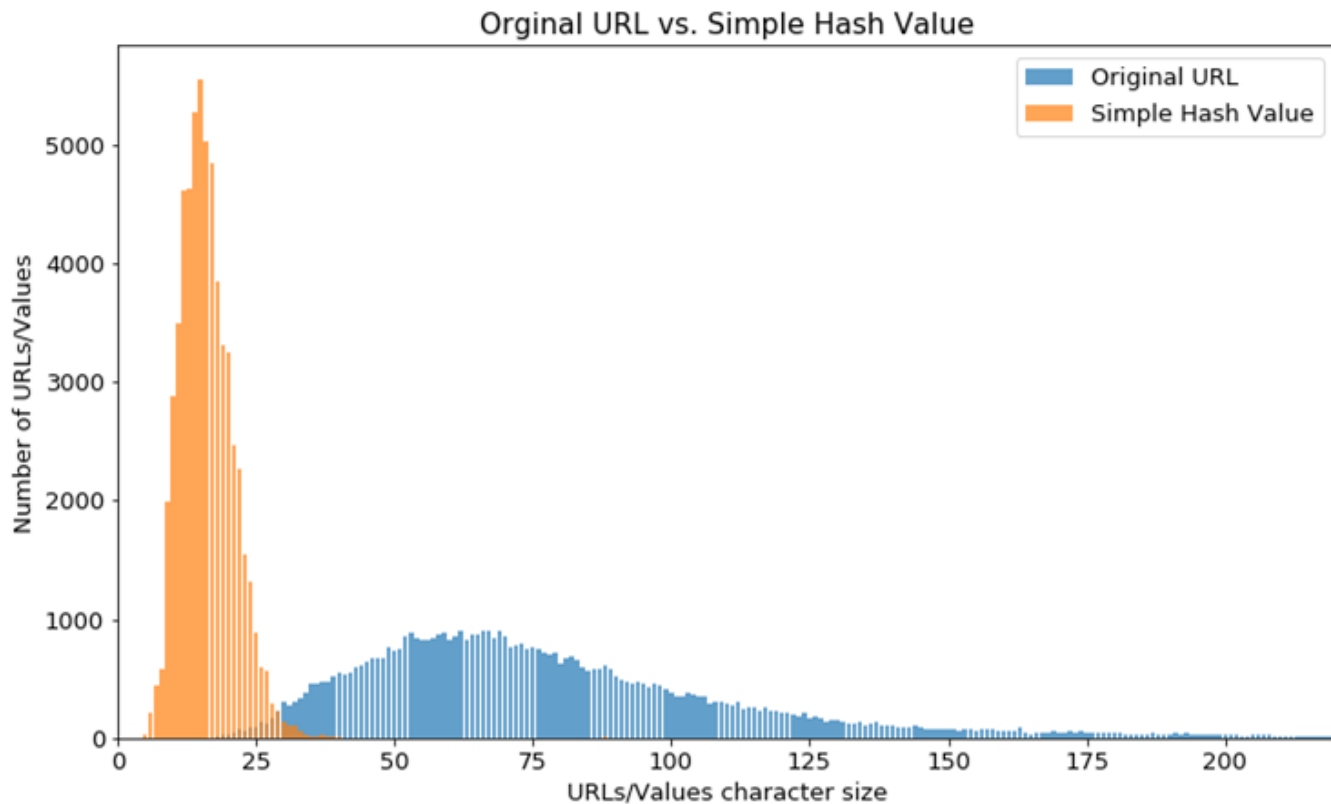
3. Simple Hash

<code>https</code>	<code>-> 0</code>		first-level domain		<code>canvasvt</code>
<code>http</code>	<code>-> 1</code>		2 bytes 0-9, a-z, A-Z		
			$2 \times 62 = 3844 > \#FLD$		

4. Result

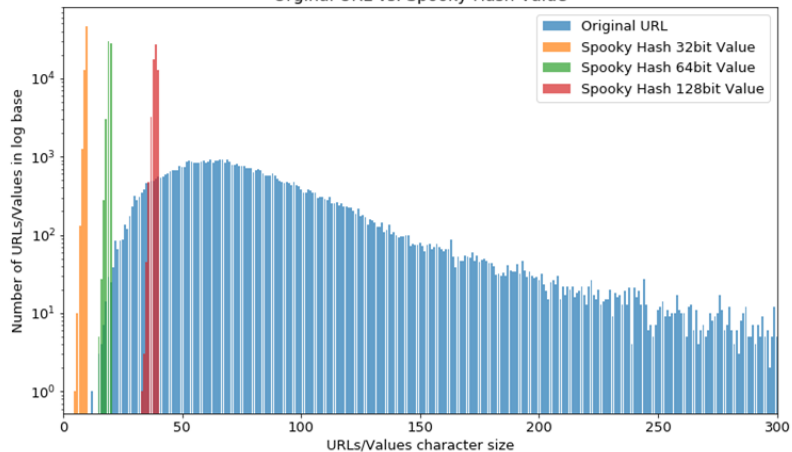
`0Q2canvasvt`

Simple Hash

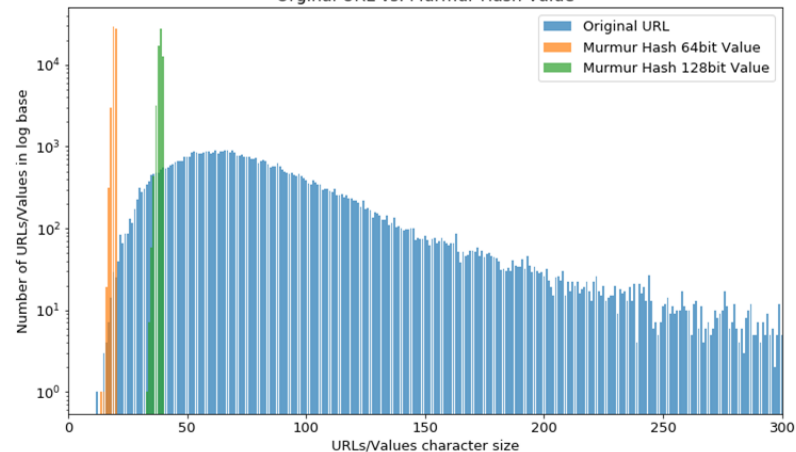


Spooky Hash, Murmur Hash, City Hash, XX Hash

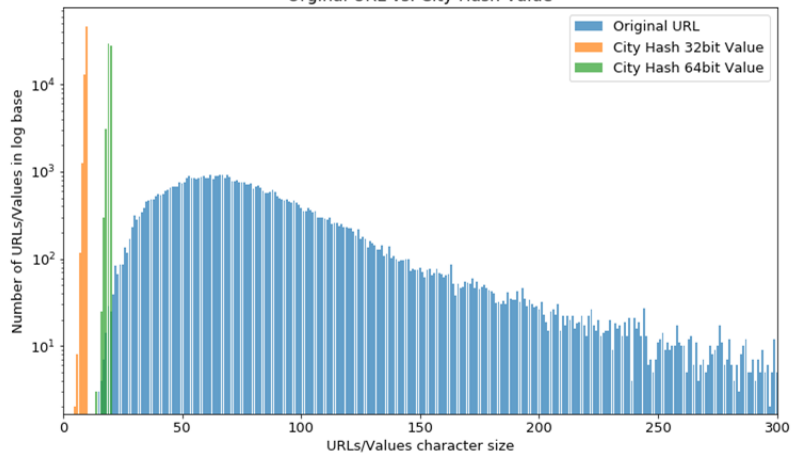
Original URL vs. Spooky Hash Value



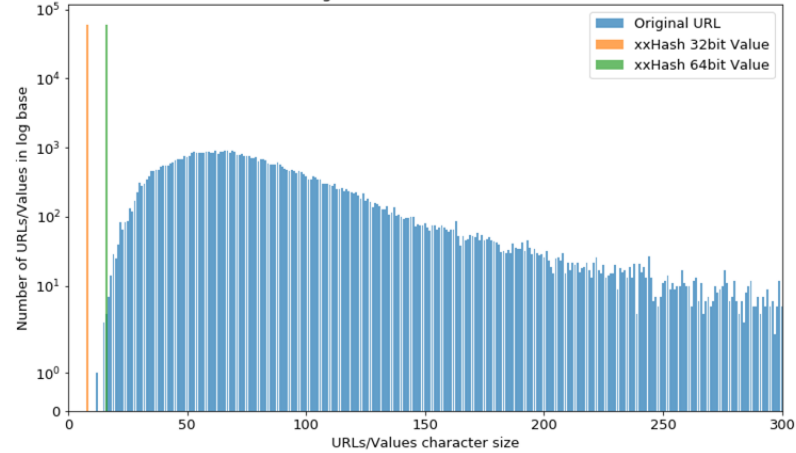
Original URL vs. Murmur Hash Value



Original URL vs. City Hash Value

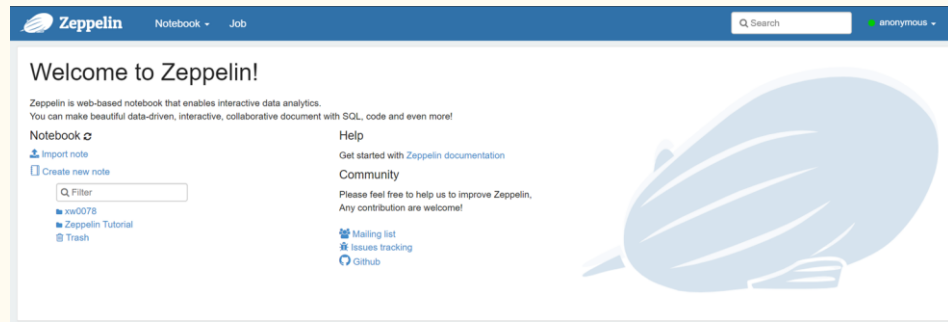


Original URL vs. xxHash Value



Benchmark

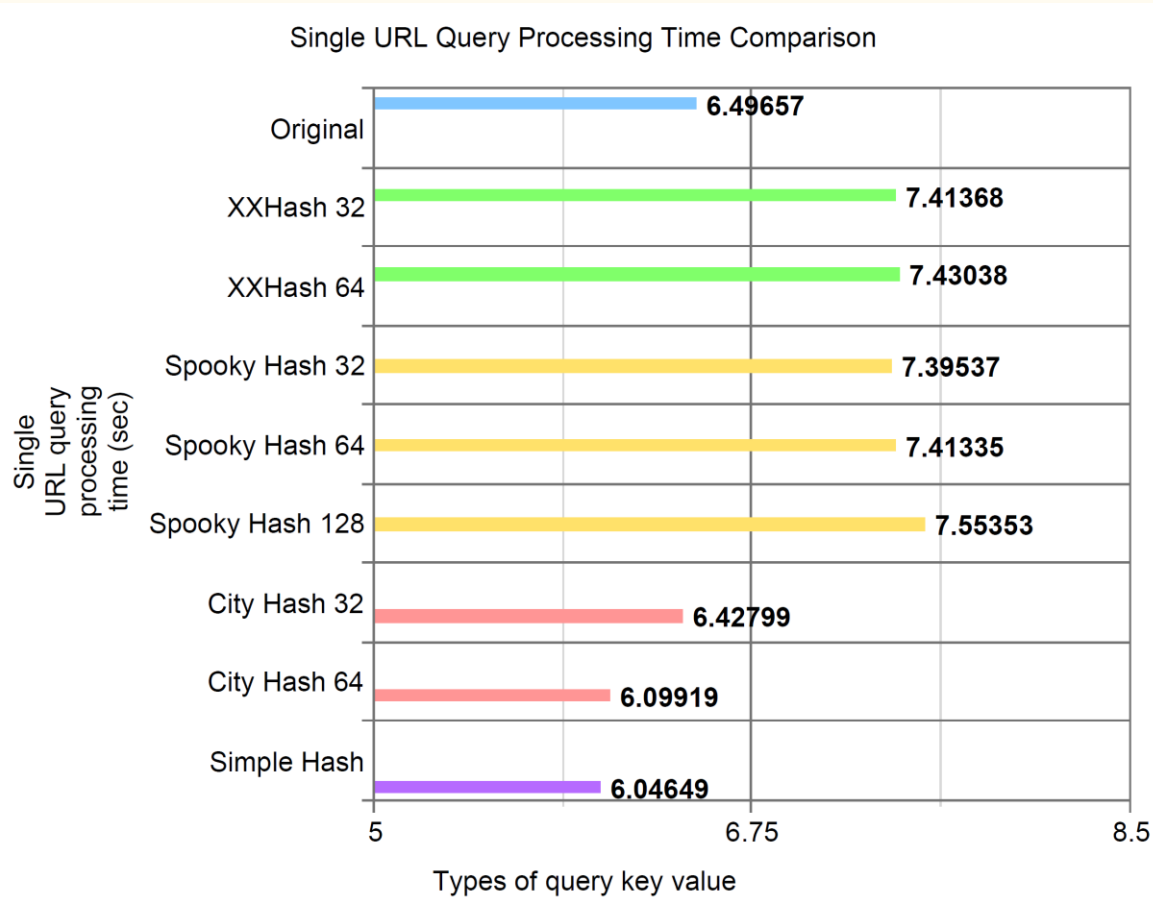
Tool: Zeppelin



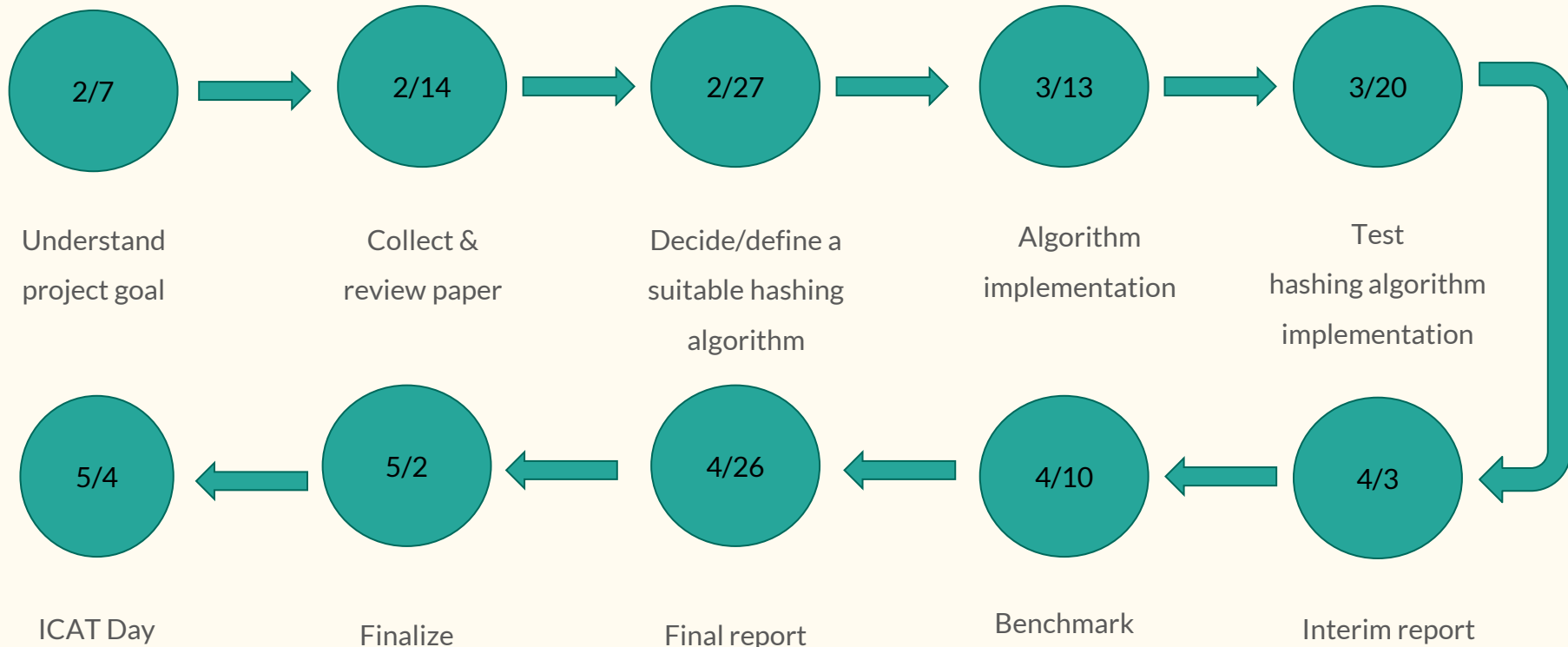
1. All hash results + original database \longrightarrow Combined.parquet
2. Run Zeppelin Web Server, and go to: localhost:8080
3. Load Parquet file, random pick 1000 samples
4. Clean cache, start query by originalURL and hash results
5. Calculate average single URL query time

Result

Based on 250 randomly selected samples.



Project Timeline



Acknowledgements

Client

Xinyue Wang



Supervisor

Dr. Edward A. Fox



Consultant

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>

