AGENDA

- ArchiveSpark Overview/Recap
- Benchmarking
- Demo
ARCHIVESPARK

- Framework - efficient data access, extraction and derivation on Web archive data
  - ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation
    - Helge Holzmann, Vinay Goel, Avishek Anand
    - Published in JCDL 2016
    - Nominated for the Best Paper Award
- OS project - https://github.com/helgeho/ArchiveSpark
TWO TECHNIQUES

1. **Pre-generated CDX metadata index**
   - Smaller dataset
   - Reduction based on Web archive metadata

2. **Incremental filtering workflow**
   - “Extract only what you need”
   - Augment -> Filter -> Repeat

- **Concept – Enrichments**
  - ArchiveSpark Record extension
  - Featured – StringContent.scala, Html.scala, Json.scala, Entities.scala, Prefix.scala, ...
  - Custom – mapEnrich[Source, Target](sourceField, targetField) (f: Source => Target)
WORKFLOW
FLEXIBLE DEPLOYMENT

- Ultimately a Scala/Spark library
- Environments
  - Standalone solitary Spark instance
  - Local HDFS-backed Spark cluster
  - Large-scale YARN/Mesos-orchestrated cluster running Cloudera/Hortonworks
- Quickstart – Docker (latest version)

- ArchiveSpark version – 2.1.0
  - Spark 2.0.2
  - Scala 2.11.7 -> Java 8
WARC FILES

- **Standard Web archiving format** - ISO 28500
- **Single capture of web resource at a particular time**
  - Header section - Metadata (URL, timestamp, content length...)
  - Payload section - Actual response body (HTML, JSON, binary data)
    - HTTP Response - HTTP headers (origin, status code)

```
WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:9e6f625c-74f6-4f9b-bec0-cebec1102f4f>
WARC-Date: 2015-06-13T18:07:56Z
WARC-IP-Address: 54.231.65.41
Content-Type: application/octet-stream
WARC-Payload-Digest: sha1:1B672GHV56GLGGWVWVSVSLF30F578B
Content-Length: 18196
WARC-Block-Digest: sha1:LI22BQHM9F2H00909099808H2EC3004

HTTP/1.1 200 OK
x-amz-id-2: Bw5C6R6XsX+Cr35QXhSbhaZts79kqG3oxjyajGz5/4+I8yFaz6K1VjLQQGr0t
x-amz-request-id: D774C57D374B03F3
Date: Sat, 13 Jun 2015 18:08:35 GMT
Content-Type: image/png
Content-Length: 17850
Server: AmazonS3
```
CDX INDEX

- “Reduced” WARC file
  - WARC metadata
  - Pointers to WARC records – offsets in WARC file

- CDX
  - Header – specifies metadata fields contained in the index
  - Body – typically 9 – 11 fields
    - Original URL, SURT, date, filename, MIME type, response code, checksum, redirect, meta tags, compressed offset
TOOLS

- **CDX Writer**
  - Python script for CDX extraction

- **Jupyter Notebook**
  - Web application for code sharing/results visualization

- **Warcbase** (only benchmarking)
  - “State-of-the-art” platform for managing and analyzing Web archives
  - Hadoop/HBase ecosystem - CDH
  - Archive-specific Scala/Java objects for Apache Spark and HBase
  - HBase command-line utilities - IngestFiles
BENCHMARKING

- Evaluation of 3 approaches
  1. ArchiveSpark
  2. Pure Spark using Warcbase library
  3. HBase using Warcbase library

- Preprocessing
  - ArchiveSpark – CDX index files extraction
  - HBase – WARC ingestion

- ArchiveSpark Benchmark subproject
  - Requirements:
    - Built and included Warcbase
    - sbt assemblyPackageDependency -> sbt assembly
ENVIROMENT

- Development
  - Cloudera Quickstart VM - CDH 5.8.2

- Benchmarking
  - Cloudera CDH 5.8.2 cluster hosted on AWS (courtesy of Dr. Zhiwu Xie)
  - 5-node cluster consisting of m4.xlarge AWS EC2 instances
    - 4 vCPUs
    - 16 GiB RAM
    - 30 GB EBS storage
    - 750 Mbps network
BENCHMARK 1 - SMALL SCALE

- Filtering & corpus extraction

- 4 scenarios
  - Filtering of the dataset for a specific URL (one URL benchmark)
  - Filtering of the dataset for a specific domain (one domain benchmark)
  - Filtering of the dataset for a date range of records (one month benchmark)
  - Filtering of the dataset for a specific active (200 OK) domain (one active domain benchmark)

- Dataset - example.warc.gz
  - One capture of archive.it domain
  - 261 records
  - 2.49 MB
BENCHMARK 1 - RESULTS

One Url

<table>
<thead>
<tr>
<th></th>
<th>ArchiveSpark</th>
<th>Spark</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>2.477421343</td>
<td>1.359588305</td>
<td>1.127581107</td>
</tr>
<tr>
<td>Max</td>
<td>11.06495323</td>
<td>2.070165325</td>
<td>2.556766825</td>
</tr>
<tr>
<td>Min</td>
<td>0.620396273</td>
<td>0.787934256</td>
<td>0.583918894</td>
</tr>
<tr>
<td>Average/No outliers</td>
<td>0.850671088</td>
<td>1.316165381</td>
<td>0.74553106</td>
</tr>
</tbody>
</table>

One domain (text/html)

<table>
<thead>
<tr>
<th></th>
<th>ArchiveSpark</th>
<th>Spark</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1.362194425</td>
<td>1.347103491</td>
<td>1.371052337</td>
</tr>
<tr>
<td>Max</td>
<td>1.959518501</td>
<td>1.873476908</td>
<td>1.722339119</td>
</tr>
<tr>
<td>Min</td>
<td>1.003479605</td>
<td>0.745562134</td>
<td>1.026370613</td>
</tr>
<tr>
<td>Average/No outliers</td>
<td>1.362194425</td>
<td>1.347103491</td>
<td>1.371052337</td>
</tr>
</tbody>
</table>

One month online

<table>
<thead>
<tr>
<th></th>
<th>ArchiveSpark</th>
<th>Spark</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1.661291172</td>
<td>4.280218437</td>
<td>4.173602424</td>
</tr>
<tr>
<td>Max</td>
<td>2.124806716</td>
<td>5.588506367</td>
<td>5.602307053</td>
</tr>
<tr>
<td>Min</td>
<td>1.336782951</td>
<td>3.251106002</td>
<td>3.271654471</td>
</tr>
<tr>
<td>Average/No outliers</td>
<td>1.609789445</td>
<td>4.269318795</td>
<td>4.09674533</td>
</tr>
</tbody>
</table>

One domain (text/html) online

<table>
<thead>
<tr>
<th></th>
<th>ArchiveSpark</th>
<th>Spark</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.60295292</td>
<td>1.497542806</td>
<td>1.468800563</td>
</tr>
<tr>
<td>Max</td>
<td>0.712397826</td>
<td>2.122264735</td>
<td>1.710818983</td>
</tr>
<tr>
<td>Min</td>
<td>0.547894969</td>
<td>1.073288654</td>
<td>1.11752257</td>
</tr>
<tr>
<td>Average/No outliers</td>
<td>0.590792374</td>
<td>1.347723263</td>
<td>1.441909627</td>
</tr>
</tbody>
</table>
Filtering & corpus extraction

4 scenarios

- Filtering of the dataset for a specific URL (one URL benchmark)
- Filtering of the dataset for a specific domain (one domain benchmark)
- Filtering of the dataset for a specific active (200 OK) domain (one active domain benchmark)
- Filtering of the dataset for pages containing scripts (pages with scripts benchmark)

Dataset - WIDE collection

- Internet Archive crawl data from Webwide Crawl (02/25/2011)
- 214470 records
- 9064 MB
- 9 files - approx. 1 GB
BENCHMARK 2 – PREPROCESSING

- CDX Extraction
  - 4 minutes 41 seconds

- HDFS Upload
  - 2 minutes 46 seconds

- HBase Ingestion x9
  - 1 file – (1 minute 10 seconds <-> 1 minute 32 seconds)
  - Sequential ingestion – approx. 13 minutes 54 seconds
**BENCHMARK 2 - RESULTS**

![Graph showing duration in seconds for different benchmarks]

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Duration (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArchiveSpark</td>
<td>1.57127569, 2.2226522, 1.86293519, 2.73961916, 2.64634819, 3.03771583, 3.3022695, 3.09726886</td>
</tr>
<tr>
<td>Spark</td>
<td>57.5994247, 67.0127334, 66.3912017, 122.647549, 123.213224, 125.03373, 181.210311, 183.119922</td>
</tr>
<tr>
<td>Hbase</td>
<td>1.77862944, 1.7610249, 1.73715926, 2.06726184, 1.78162828, 2.01815477, 1.83698445, 2.03299462, 1.8755871</td>
</tr>
</tbody>
</table>
BENCHMARK 2 - RESULTS

One Domain (text/html)

<table>
<thead>
<tr>
<th></th>
<th>ArchiveSpark</th>
<th>Spark</th>
<th>HBase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6305868291</td>
<td>98.095503</td>
<td>0.796309112</td>
</tr>
<tr>
<td>2</td>
<td>1.912719245</td>
<td>119.37970841</td>
<td>1.05643289</td>
</tr>
<tr>
<td>3</td>
<td>2.439430593</td>
<td>115.2829777</td>
<td>1.151400879</td>
</tr>
<tr>
<td>4</td>
<td>3.730593796</td>
<td>214.1183215</td>
<td>1.643759031</td>
</tr>
<tr>
<td>5</td>
<td>3.691433171</td>
<td>219.90471652</td>
<td>1.4184737981</td>
</tr>
<tr>
<td>6</td>
<td>5.204209412</td>
<td>225.0313651</td>
<td>2.438475912</td>
</tr>
<tr>
<td>7</td>
<td>6.182943747</td>
<td>323.5355116</td>
<td>2.3122200972</td>
</tr>
<tr>
<td>8</td>
<td>5.897939707</td>
<td>327.4338609</td>
<td>2.639451323</td>
</tr>
<tr>
<td>9</td>
<td>7.406937993</td>
<td>327.153154</td>
<td></td>
</tr>
</tbody>
</table>

DURATION IN SECONDS

One Domain (text/html)
**BENCHMARK 2 - RESULTS**

One Domain (text/html) Online (Status Code - 200)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Duration in Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArchiveSpark</td>
<td>1.08825498</td>
</tr>
<tr>
<td>Spark</td>
<td>97.8362223</td>
</tr>
<tr>
<td>HBase</td>
<td>0.7369705</td>
</tr>
</tbody>
</table>

One Domain (text/html) Online

- **ArchiveSpark**
  - 1.08825498
  - 1.26664041
  - 1.48358385
  - 2.62414807
  - 3.25672284
  - 4.24379717
  - 4.53168404
  - 7.29249372
  - 6.6079556

- **Spark**
  - 97.8362223
  - 117.852994
  - 115.479023
  - 225.376197
  - 223.129106
  - 224.254291
  - 432.276701
  - 397.342915
  - 504.426315

- **HBase**
  - 0.7369705
  - 1.11761995
  - 1.26732586
  - 1.53378846
  - 1.69809579
  - 1.90836823
  - 3.42893737
  - 3.54753302
  - 3.47453209
BENCHMARK 2 - RESULTS

Web Pages (text/html) with Scripts

<table>
<thead>
<tr>
<th>ArchivSpark (HTML)</th>
<th>ArchivSpark (StringContent)</th>
<th>Spark</th>
<th>Hbase</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.33707137</td>
<td>68.59872419</td>
<td>113.8777769</td>
<td>42.70015677</td>
</tr>
<tr>
<td>187.167541</td>
<td>167.034233</td>
<td>138.2625233</td>
<td>64.82519567</td>
</tr>
<tr>
<td>208.1246111</td>
<td>185.7237064</td>
<td>165.723834</td>
<td>85.1789321</td>
</tr>
<tr>
<td>362.0677655</td>
<td>319.3848116</td>
<td>254.2081722</td>
<td>115.8876752</td>
</tr>
<tr>
<td>396.5672534</td>
<td>344.311085</td>
<td>268.7816292</td>
<td>164.9606675</td>
</tr>
<tr>
<td>380.5400389</td>
<td>325.1620758</td>
<td>260.1414318</td>
<td>192.2859835</td>
</tr>
<tr>
<td>523.3000353</td>
<td>456.6902816</td>
<td>300.6188545</td>
<td>213.2043155</td>
</tr>
<tr>
<td>538.7206097</td>
<td>482.2359012</td>
<td>351.0332913</td>
<td>223.7166384</td>
</tr>
<tr>
<td>552.6327346</td>
<td>497.7889328</td>
<td>378.4675011</td>
<td>238.895708</td>
</tr>
</tbody>
</table>

DURATION IN SECONDS
ACKNOWLEDGEMENT & DISCLAIMER

- This material is based upon work supported by following grants:
  - IMLS LG-71-16-0037-16: Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse
  - NSF IIS-1619028, III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)

- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.
THANK YOU

http://tinyurl.com/zejgc9f