A Serverless Framework for Distributed Bulk Metadata Extraction

By Tyler J. Skluzacek
Data generated at each phase of the lifecycle

- Acquire
- Clean
- Use/Reuse
- Publish
- Preserve/Destroy
Many scientists push data into a **data lake**

**Data Lake**: an unstructured collection of files such that schema applied on read, but not write *(i.e., put anything in → figure out schema when you take it out)*  

[Khine, ‘18]
Without active curation, a data lake will become a data **swamp**.

**Data Swamp:** a data lake that is **difficult to navigate** or is **missing critical informational elements** such that files cannot be accessed, discovered, or reused. [Hai, ‘16]
To avoid ‘swamping’, we need an index of rich searchable metadata

```json
"object_type": "image"
"image_type": "photograph"
"entities": ["dog", "tree", "leaves"]
"file_size_mb": 2.0
"created_on": "06-05-2021T00:00"
"owner": ...
```
We built a system to extract these metadata...

... but we found that data are huge and distributed across heterogeneous computing machinery.
Xtract remotely orchestrates extraction plans across distributed data

Leverages funcX to enable scalable and remote execution of lightweight extraction functions

Workflow:
Crawl repository
Group files by applying grouping function
Make and execute processing location decisions for each file
Execute extraction plan
Validate metadata documents
Scalable to at least 2,048 concurrent HPC workers

ImageSort too lightweight to scale well with batch size of 8

MatIO sees reasonable (but imperfect) scaling up to 4,096 workers

Data:
- 200,000 Materials Data tasks (1.1 TB) from the Materials Data Facility
- 80,000 tasks (14 GB) from the Common Objects in Context Training Set
Optimizations facilitate higher task throughputs and decreased execution time.

Max throughput achieved at 8x8 (64) family batch size.

**Batching:** Total task throughput for executing materials science extraction functions on Midway2 with client-side (Xtract) batches and internal funcX batches.
When transfer is necessary, Xtract processes files nearly as quickly as they arrive. Crawling never a bottleneck.

Transfers packaged as ‘blocks’ with maximum 10GB or 20,000 files.

Bulk metadata extraction times for an MDF subset (50,000 files) processed on 4—32 Midway2 nodes.
We can process 60TB (2.2 million groups) using the Theta supercomputer in just over 6 hours.

We notice a large throughput spike early on...

Throughput spike caused by long-running ASE extractions.

Checkpoint by writing metadata to file system until full batch completed.

Longest Extractor
- ase
- yaml
- csv
- xml
- json
- dft
Extracting a Google Drive repository on a Kubernetes cluster showcases the compute flexibility.

### Invocations and Extraction Times

<table>
<thead>
<tr>
<th>Extractor</th>
<th>Total Invocations</th>
<th>Avg. Extract Time (s)</th>
<th>Avg. Transfer Time (s)</th>
<th>Avg. File Size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>3539</td>
<td>2.76</td>
<td>1.38</td>
<td>0.559</td>
</tr>
<tr>
<td>Tabular</td>
<td>333</td>
<td>0.21</td>
<td>0.31</td>
<td>0.024</td>
</tr>
<tr>
<td>Null-Value</td>
<td>333</td>
<td>0.84</td>
<td>0.30</td>
<td>0.024</td>
</tr>
<tr>
<td>Images</td>
<td>774</td>
<td>1.06</td>
<td>0.80</td>
<td>4.0</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>1</td>
<td>2.2</td>
<td>5.9</td>
<td>14.0</td>
</tr>
</tbody>
</table>

Invocations and Extraction Times for 5 extractors run on a Graduate Student’s Google Drive repository.
Conclusion

Xtract enables metadata extraction on:
• big data
• distributed data
• data on heterogeneous cyberinfrastructure

Xtract is made possible by...

Contact information:
Tyler J. Skluzacek
skluzacek@uchicago.edu