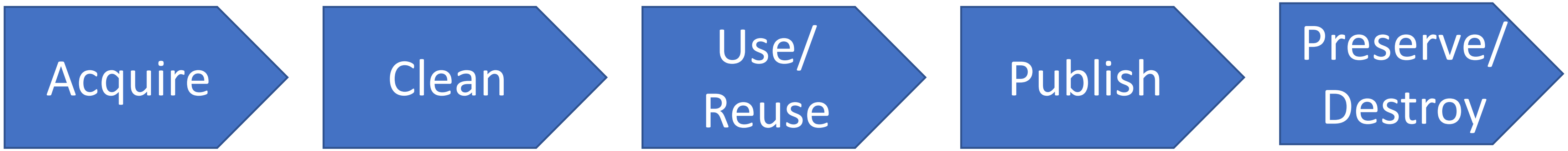
A Serverless Framework for Distributed Bulk Metadata Extraction

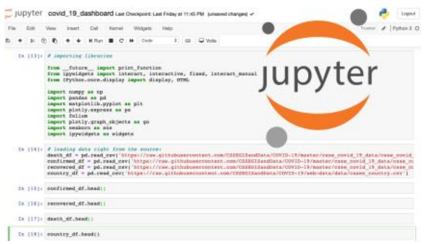
By Tyler J. Skluzacek

globus labs

THE UNIVERSITY OF CHICAGO

Data generated at each phase of the lifecycle

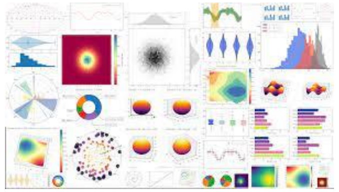




jupyter covid_19_dashboard Last Checkpoint: Last Friday at 11:45 PM (unsaved changes)

```
In [13]: # importing libraries

         from __future__ import print_function
         from ipywidgets import interact, interactive, fixed, interact_manual
         from IPython.core.display import display, HTML

         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import plotly.express as px
         import folium
         import plotly.graph_objects as go
         import seaborn as sns
         import ipywidgets as widgets

In [14]: # loading data right from the source:
         death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
         confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
         recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
         country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')

In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```

pandas



Keras

The Materials Data Facility (MDF)

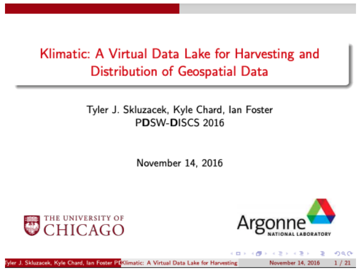
A simple way to publish, discover, and access materials datasets

DLHub

Data and Learning Hub for Science

https://www.dlhub.org

globus



Klimatic: A Virtual Data Lake for Harvesting and Distribution of Geospatial Data

Tyler J. Skluzacek, Kyle Chard, Ian Foster
PDSW-DISCS 2016

November 14, 2016

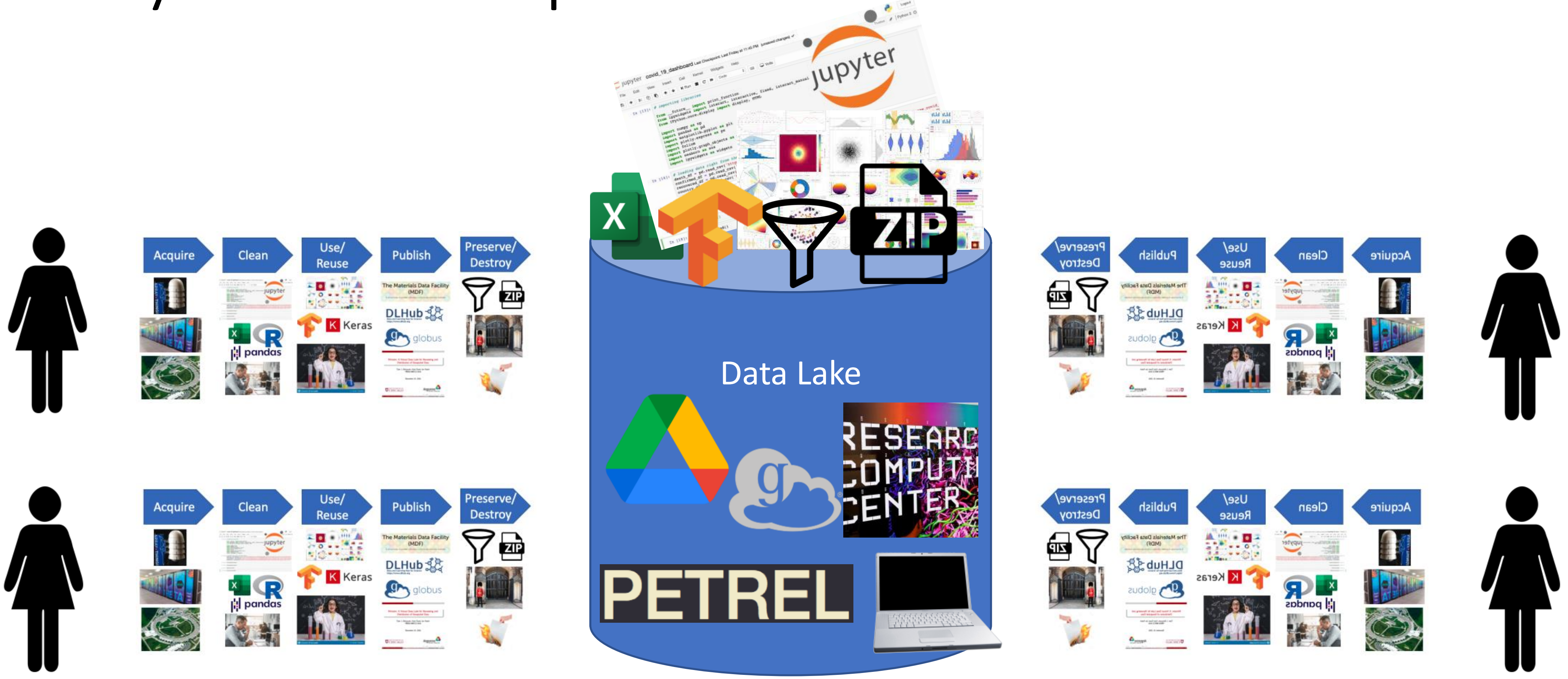
THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Tyler J. Skluzacek, Kyle Chard, Ian Foster PDKlimatic: A Virtual Data Lake for Harvesting November 14, 2016 1 / 21

ZIP

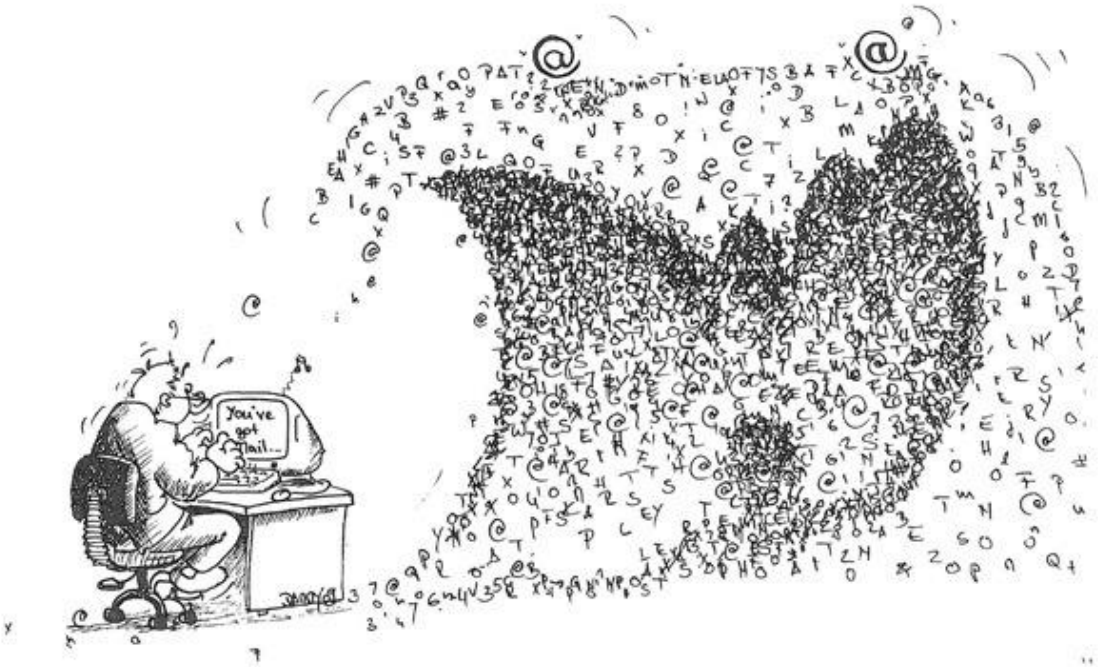
Many scientists push data into a **data lake**



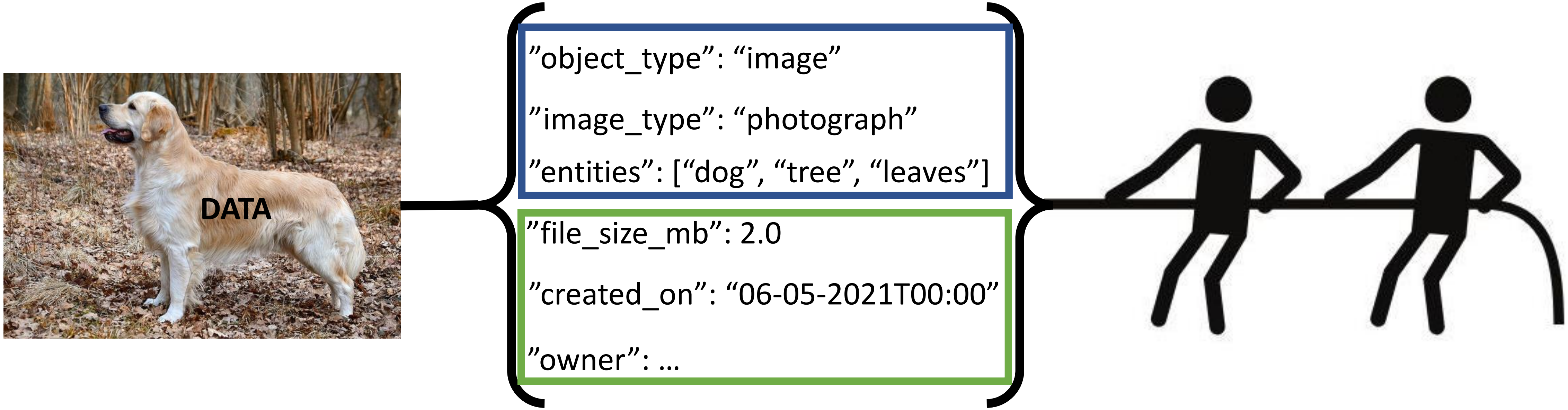
Data Lake: an unstructured collection of files such that schema applied on read, but not write *(i.e., put anything in → figure out schema when you take it out)* [Khine, '18]

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



”object_type”: “image”

”image_type”: “photograph”

”entities”: [“dog”, “tree”, “leaves”]

”file_size_mb”: 2.0

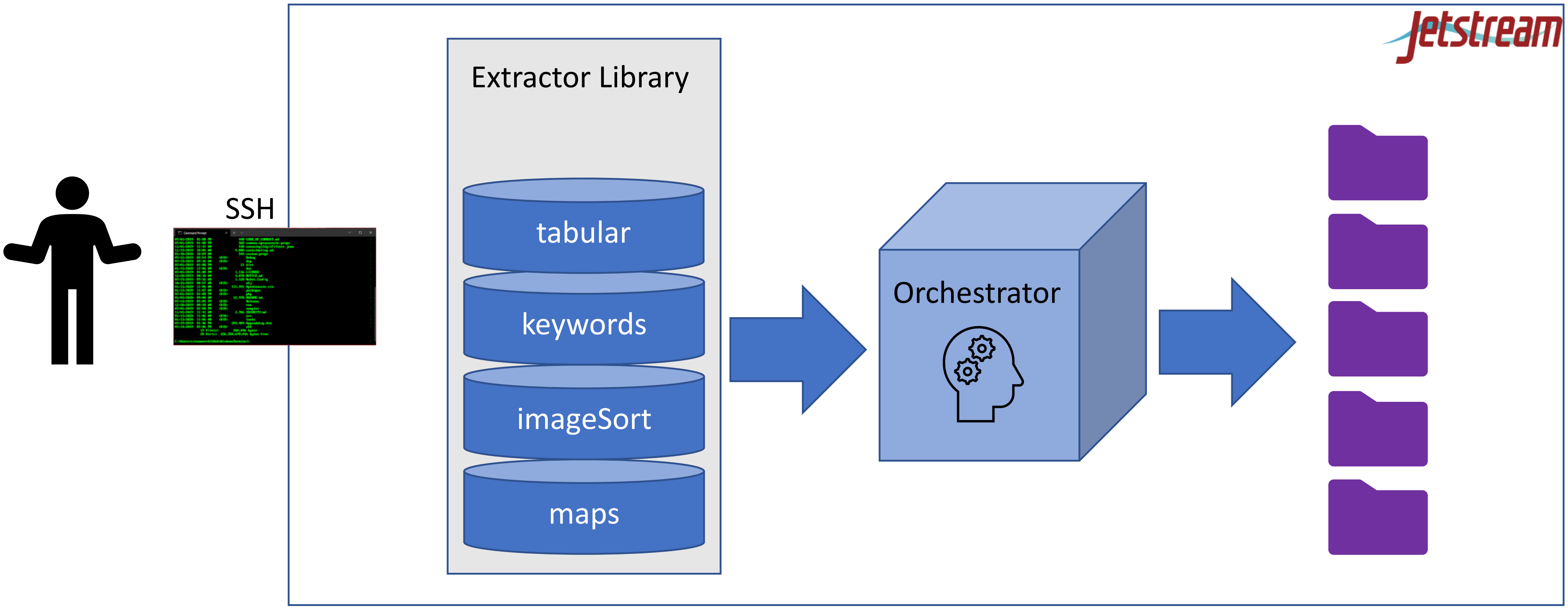
”created_on”: “06-05-2021T00:00”

”owner”: …

🟦 = content

🟩 = context

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:

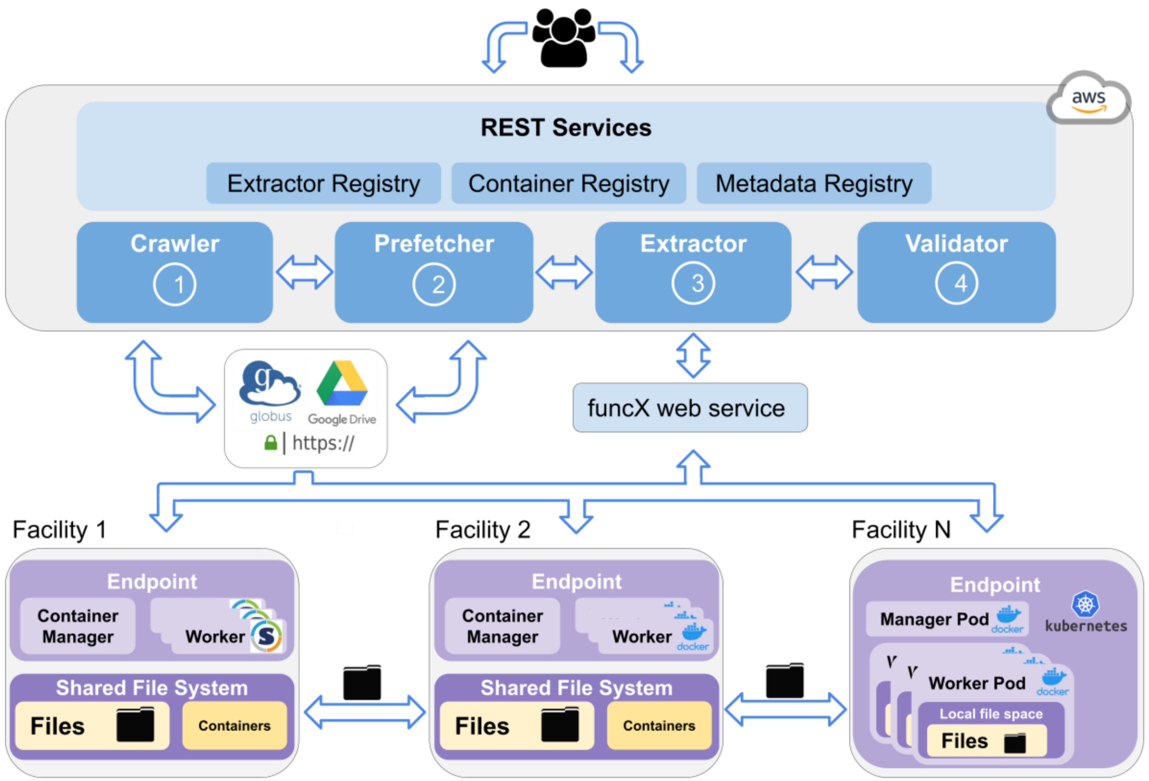
Crawl repository

Group files by applying grouping function

Make and execute processing location decisions for each file

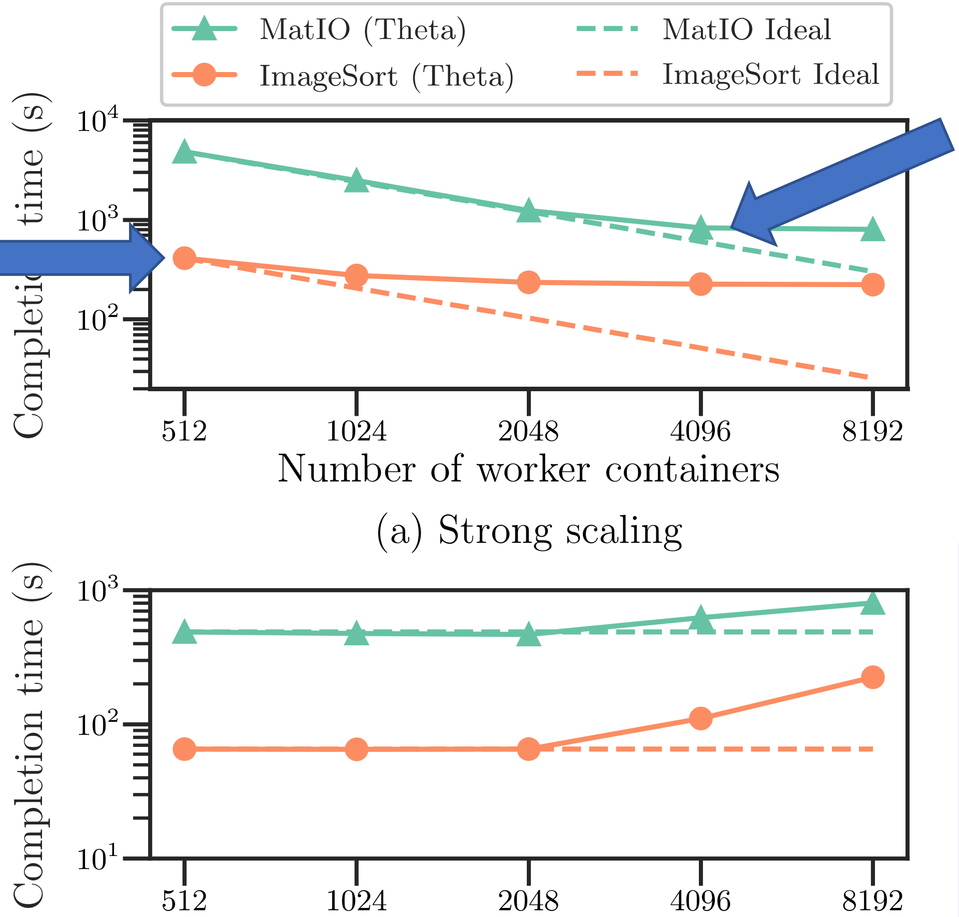
Execute extraction plan

Validate metadata documents



Scalable to *at least* 2,048 concurrent HPC workers

ImageSort too lightweight to scale well with batch size of 8



(a) Strong scaling

(b) Weak scaling

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

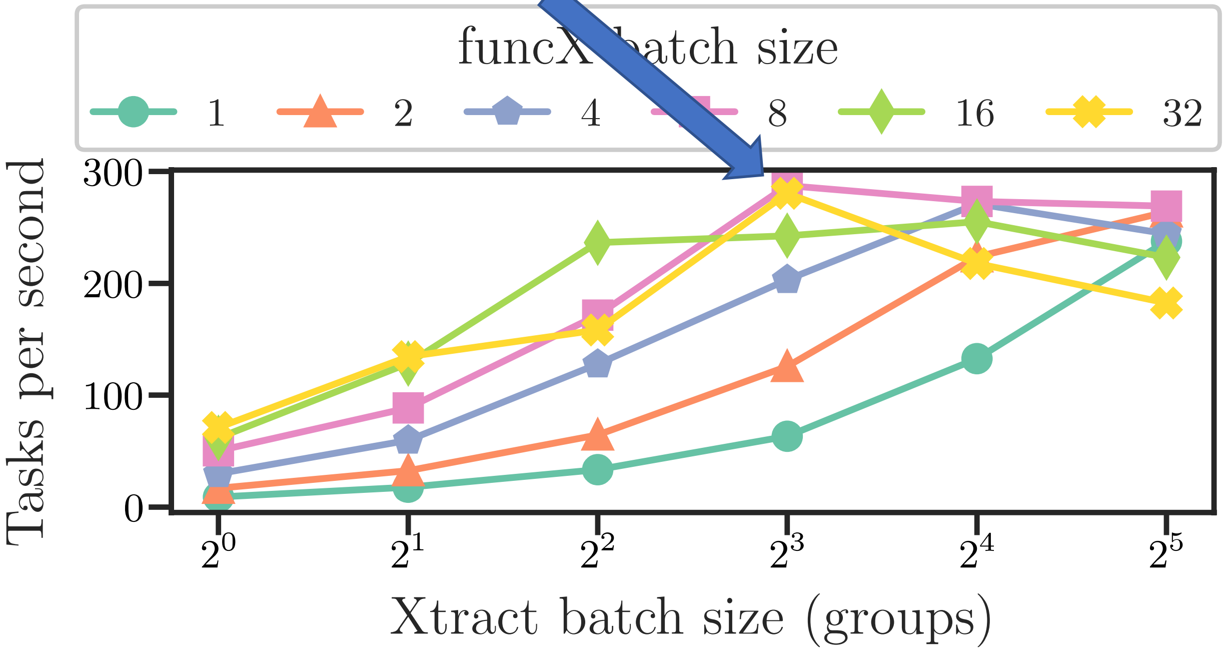
Data:

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

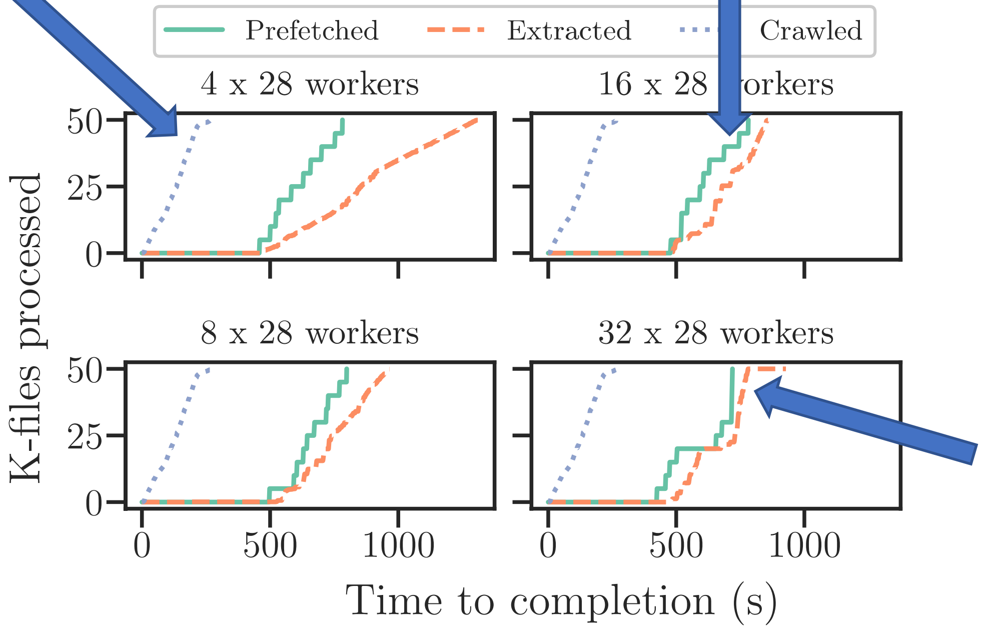


Batching: Total task throughput for executing materials science extraction functions on Midway2 with client-side (Xtract) batches and internal funcX batches.

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

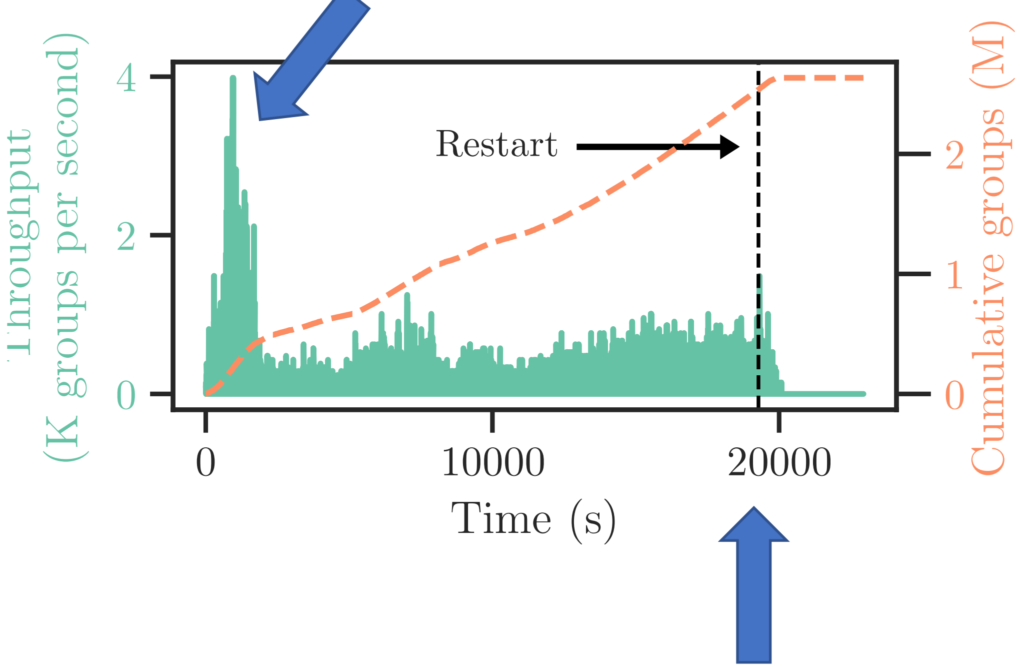


32 nodes x 28 workers: extraction line “hugs” the prefetch line

Bulk metadata extraction times for an MDF subset (50,000 files) processed on 4—32 Midway2 nodes.

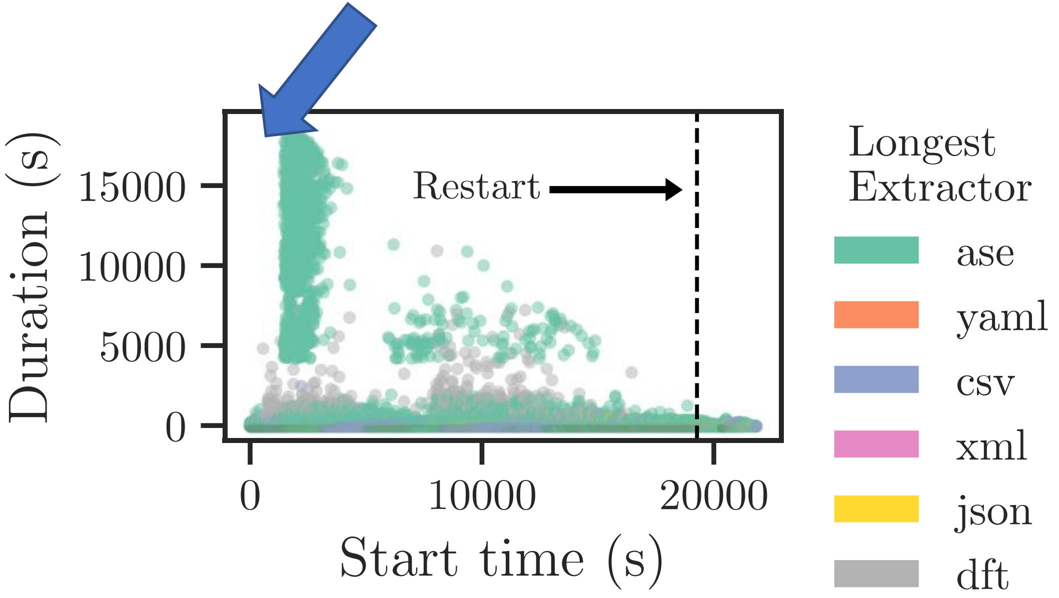
We can process 60TB (2.2 million groups) using the Theta supercomputer in **just over 6 hours**

We notice a large throughput spike early on…

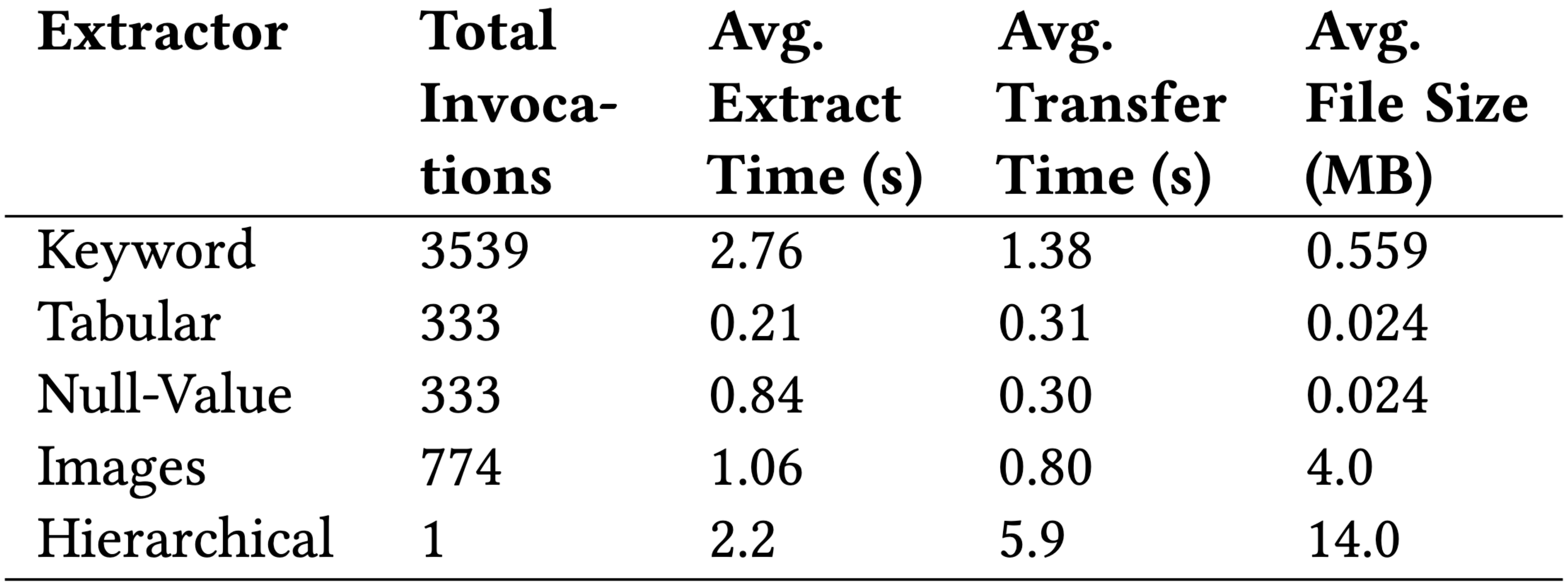


Checkpoint by writing metadata to file system until full batch completed

Throughput spike caused by long-running ASE extractions



Extracting a Google Drive repository on a Kubernetes cluster showcases the compute flexibility



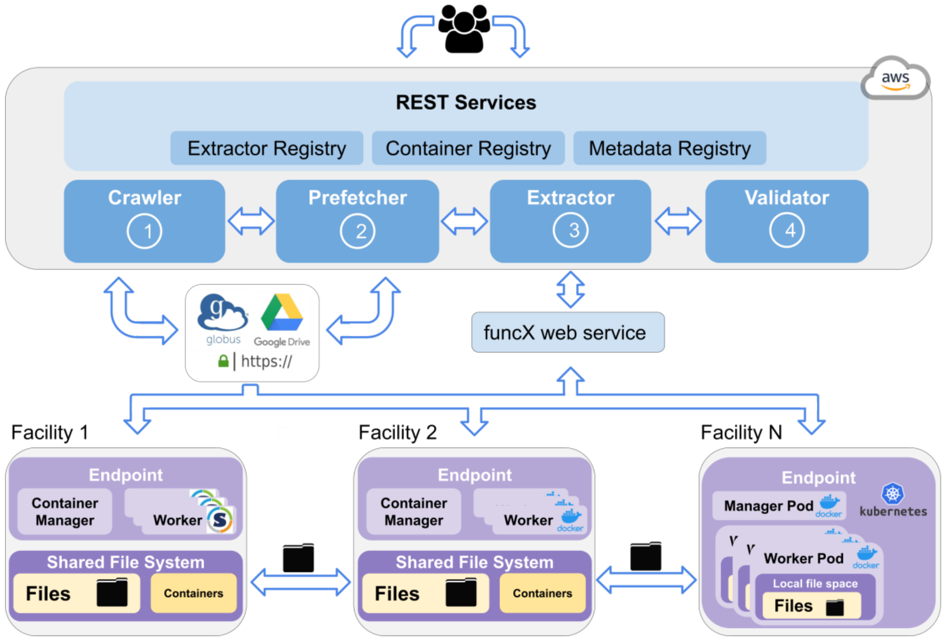
Extractor	Total Invocations	Avg. Extract Time (s)	Avg. Transfer Time (s)	Avg. File Size (MB)
Keyword	3539	2.76	1.38	0.559
Tabular	333	0.21	0.31	0.024
Null-Value	333	0.84	0.30	0.024
Images	774	1.06	0.80	4.0
Hierarchical	1	2.2	5.9	14.0

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…

funcX

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