

**BIG DATA
PROCESSING
IN THE CLOUD:
A HYDRA/SUFIA
EXPERIENCE**

**Collin Brittle
Zhiwu Xie**

**Helsinki
June 2014**

WHO?

VirginiaTech *Invent the Future* | Center for Digital Research and Scholarship
University Libraries

Library web • Summon
Enter keyword search
A to Z index | Library staff

QUICKLINKS

Center for Digital Research and Scholarship

Off Campus Sign in

Collections
Consulting
Events
Research

Center for Digital Research and Scholarship

Virginia Tech Libraries' Center for Digital Research and Scholarship strives to facilitate excellence in digital research. The center is built on three core themes: partnerships, technology, and services. We strive to partner with individual faculty, research labs, centers, and institutes to solve academic-based problems through applied research in

Contact us
General questions can be sent to our email address.
More detailed questions can be directed to the CDRS staffer responsible for the service or resource. See our people page for a list of staffers.

CDRS news
Faculty Week: Publishing support from the University



VirginiaTech

Smart Infrastructure Laboratory



.....

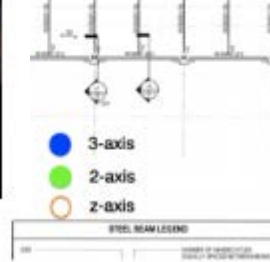
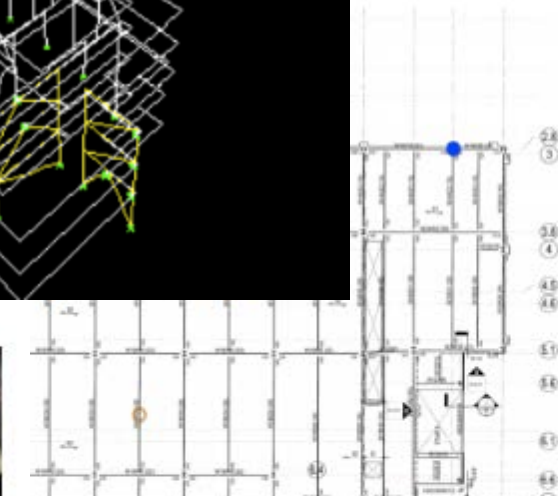
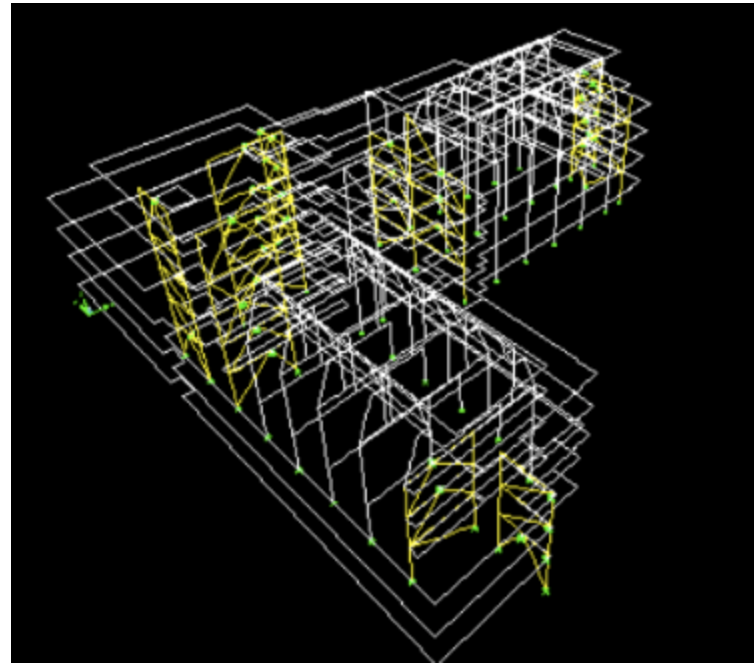
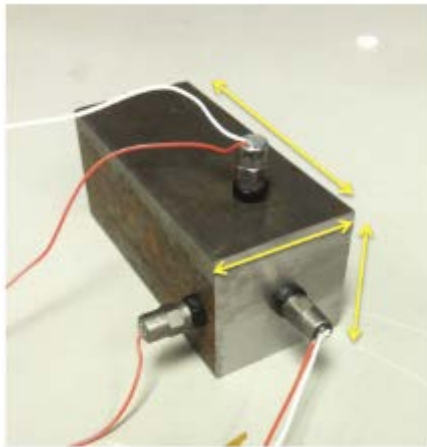
WHAT?



WHY?



SENSORS



SMART INFRASTRUCTURE



Acceleration

acoustics

Temperature

strain

wind and flow

load



DATA SHARING

- **Encourage exploratory and multidisciplinary research**
- **Foster open and inclusive communities around**
 - modeling of dynamic systems
 - structural health monitoring and damage detection
 - occupancy studies
 - sensor evaluation
 - data fusion
 - energy reduction
 - evacuation management
 - ...



CHARACTERIZATION

- **Compute intensive**
- **Storage intensive**
- **Communication intensive**
- **On-demand**
- **Scalability challenge**



COMPUTE INTENSIVE

- **About 6GB raw data per hour**
- **Must be continuously processed, ingested, and further processed**
- **User-generated computations**
- **Must not interfere with data retrieval**



STORAGE INTENSIVE

- **SEB will accumulate about 60TB of raw data per year**
- **To facilitate researchers, we must keep raw data for an extended period of time, e.g., ≥ 5 years**
- **VT currently does not have an affordable storage facility to hold this much data**
- **Within XSEDE, only TACC's Ranch can allocate this much storage**



COMMUNICATION INTENSIVE

- **What if hundreds of researchers around the world each tried to download hundreds of TB of our data?**



ON DEMAND

- **Explorative and multidisciplinary research cannot predict the data usage beforehand**



SCALABILITY

- **How to deal with these challenges in a scalable manner?**



BIG DATA + CLOUD

- **Affordable**
- **Elastic**
- **Scalable**



FRAMEWORK REQUIREMENTS

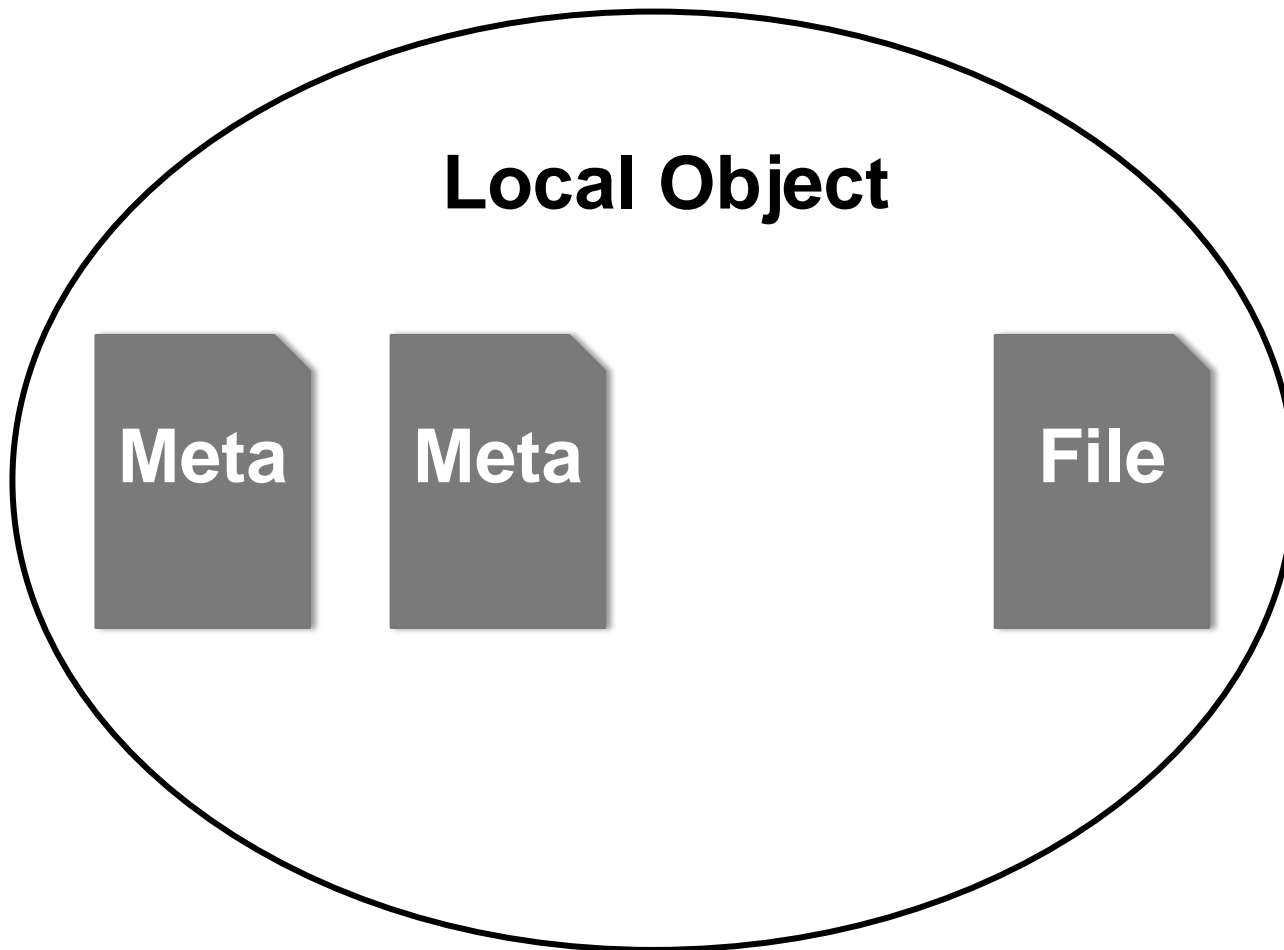
- **Mix local and remote content**
- **Support background processing**
- **Be distributable**



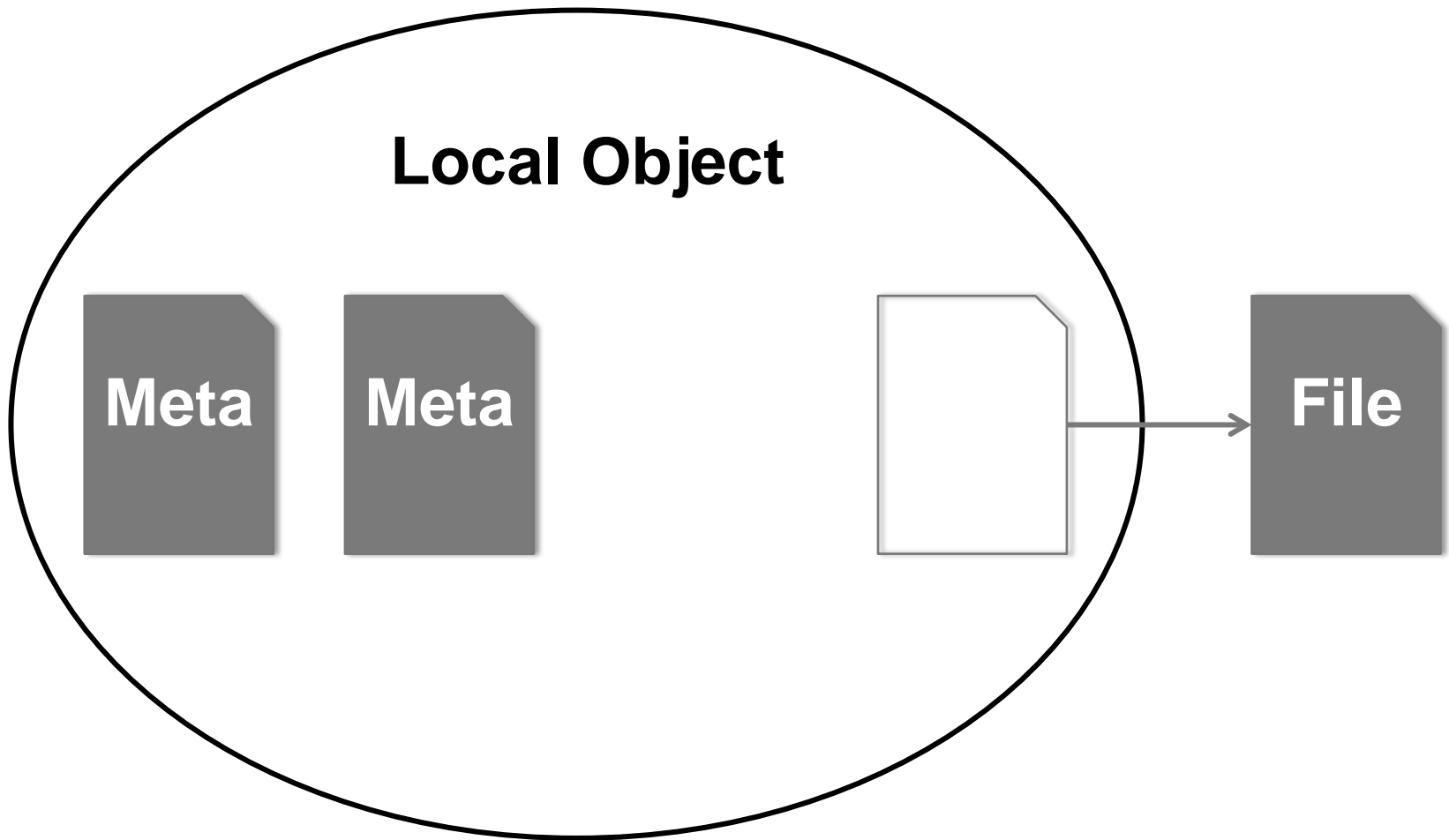
FRAMEWORK REQUIREMENTS

- **Mix local and remote content**
- Support background processing
- Be distributable

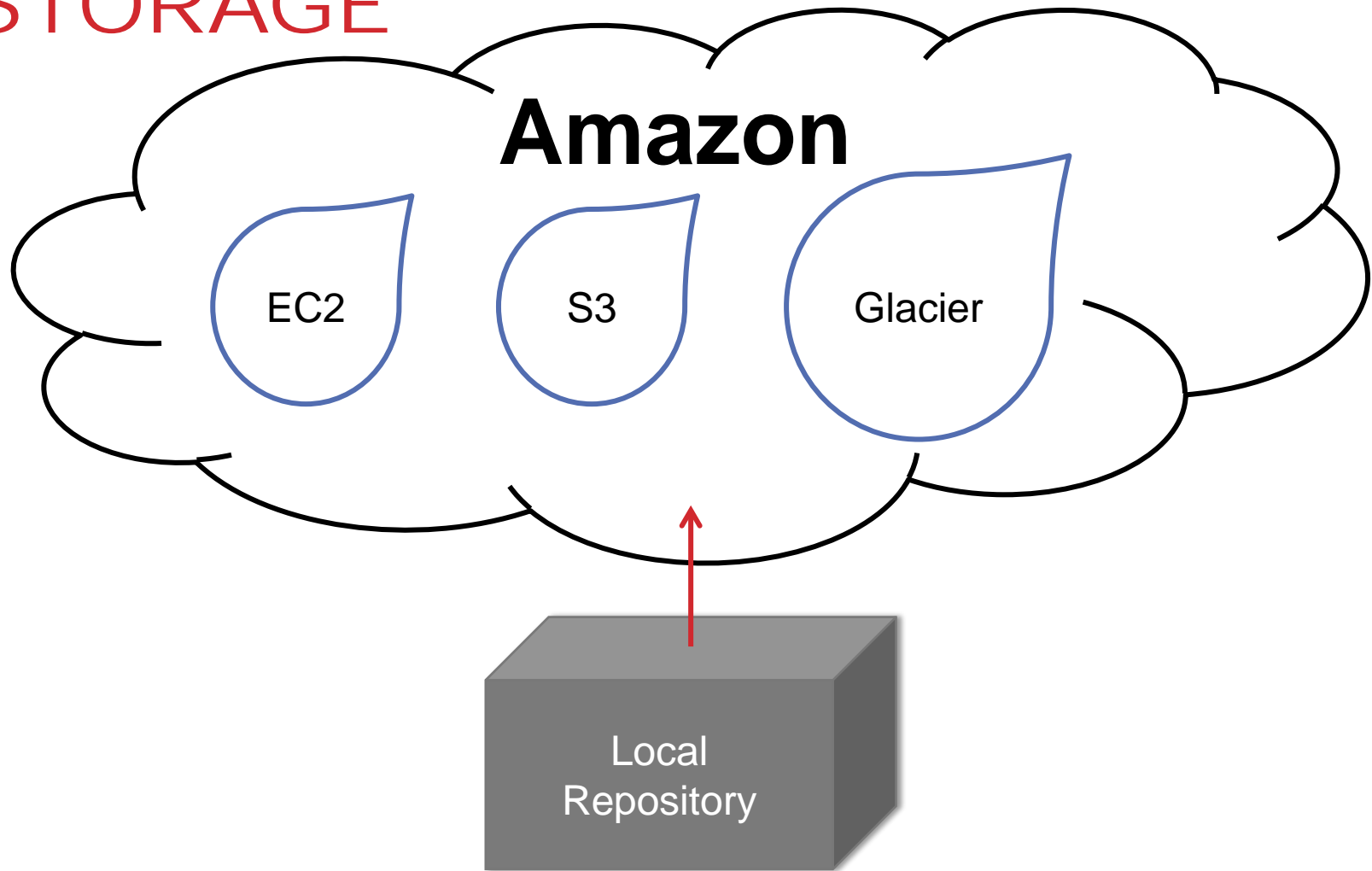
OBJECTS AND DATASTREAMS



OBJECTS AND DATASTREAMS



REMOTE STORAGE

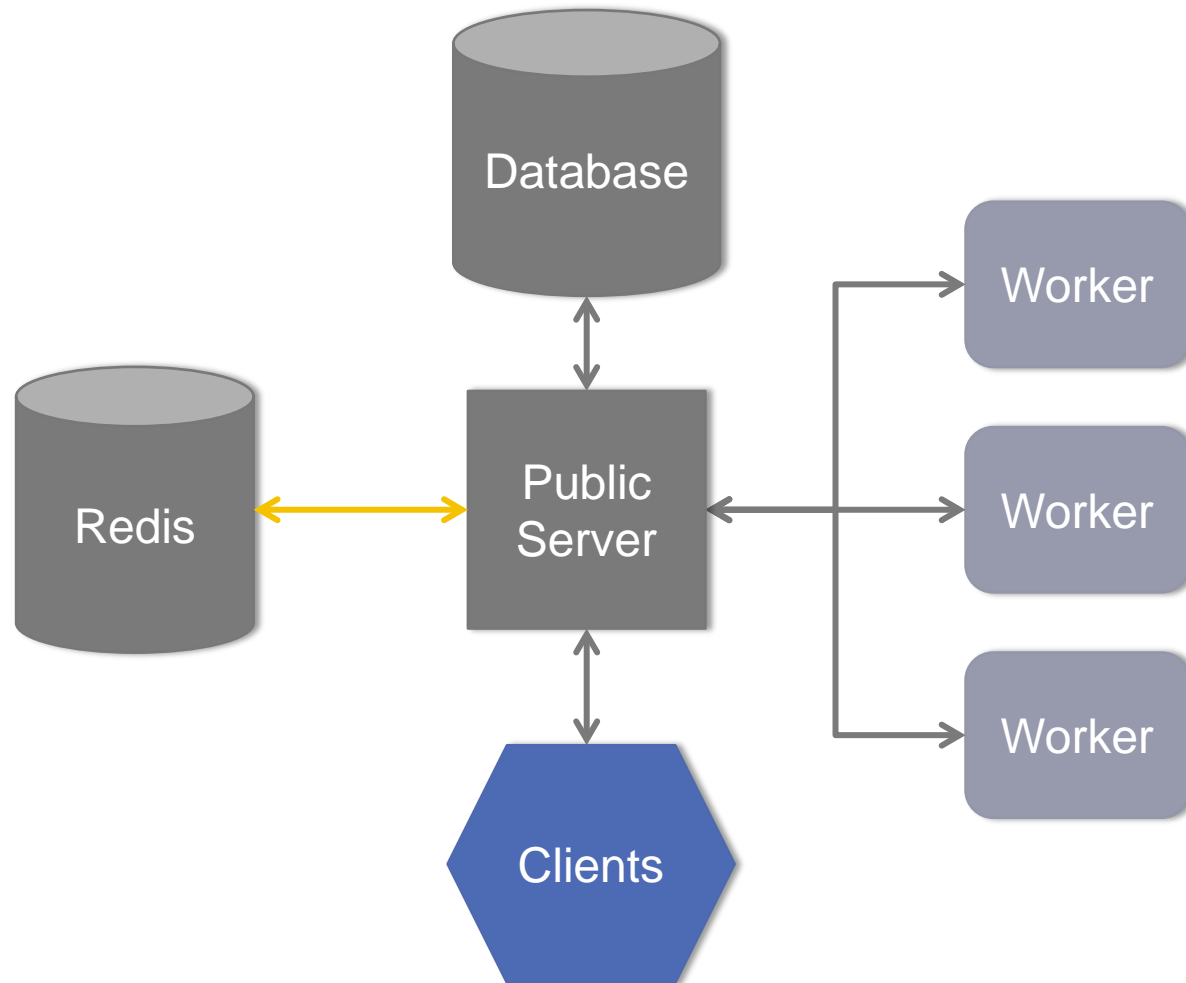




FRAMEWORK REQUIREMENTS

- Mix local and remote content
- **Support background processing**
- Be distributable

BACKGROUND PROCESSING





FROM QUEUES TO THE CLOUD

0100
0010

1100
0011

1010
0101

0101
0101

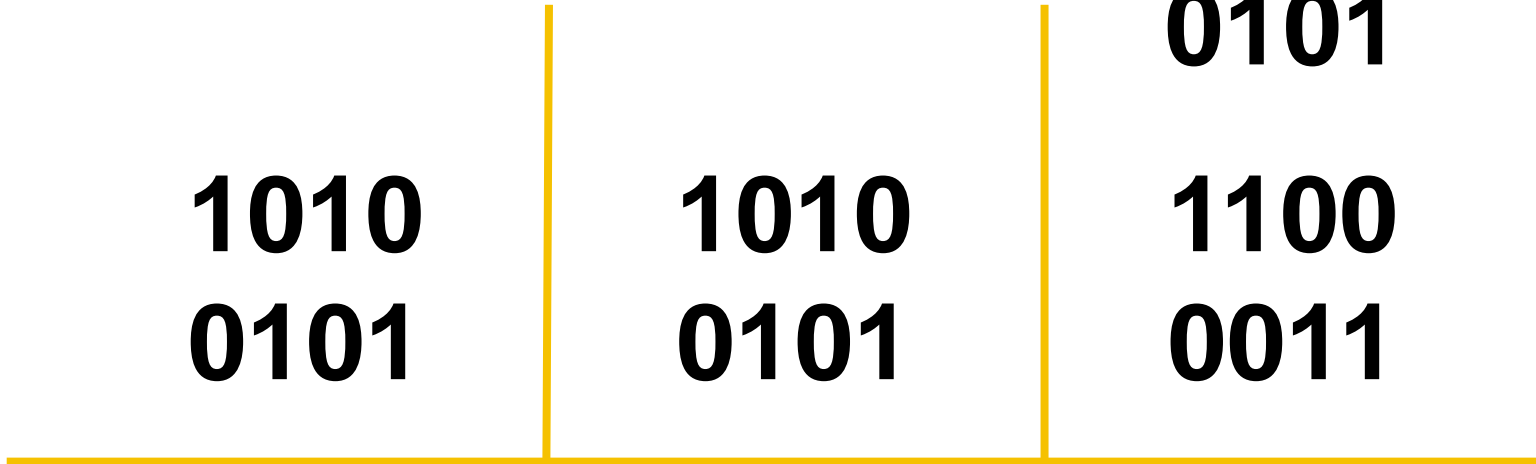


FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101
1100
0011



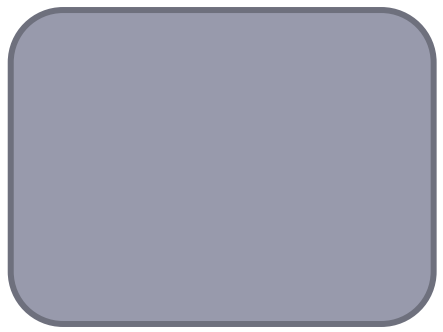


FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101
1100
0011





FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101

1100
0011



FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101

1100
0011



FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101

1100
0011

0011
1100



FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101





FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101



FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101



FROM QUEUES TO THE CLOUD

1010
0101

1010
0101

1010
0101

1111
0000



FROM QUEUES TO THE CLOUD

	1010	1010
	0101	0101



QUEUEING

The screenshot shows the Resque overview page in a browser. The page title is "Resque" and the URL is "localhost:3000/resque/overview". The navigation menu includes "Overview", "Working", "Failed", "Queues", "Workers", and "Stats". The "Queues" section is active, displaying a table of registered queues. Below the table, it shows "0 of 0 Workers Working".

Name	Jobs
<u>audit</u>	0
<u>average</u>	0
<u>batch_update</u>	0
<u>characterize</u>	0
<u>data</u>	0
<u>event</u>	0
<u>high</u>	0
<u>low</u>	0
<u>failed</u>	0

0 of 0 Workers Working

Where	Queue	Processing
Nothing is happening right now...		

Name	Jobs
<u>audit</u>	0
<u>average</u>	0
<u>batch_update</u>	0
<u>characterize</u>	0
<u>data</u>	0
<u>event</u>	0
<u>high</u>	0
<u>low</u>	0
<u>failed</u>	0

0 of 0 Workers Working

The list below contains all workers which are currently running a job.

Where	Queue	Processing
Nothing is happening right now...		

QUEUEING

Queues
The list below contains all the registered queues with the number of jobs currently in the queue. Select a queue from above to view all jobs currently pending on the queue.

Name	Jobs
audit	0
average	0
batch_update	0
characterize	0
data	4
event	10
high	1
low	1
failed	0

1 of 1 Workers Working
The list below contains all workers which are currently running a job.

Where	Queue	Processing
Collins-MacBook-Air.local:18839	AVERAGE	Sufia::Resque::MarshaledJob just now

Powered by Statpad v1.25.2
Connected to Redis namespace sufia:development on redis://localhost:6379/0

Name	Jobs
audit	0
average	1
batch_update	1
characterize	0
data	4
event	5
high	1
low	1
failed	0

1 of 1 Workers Working

The list below contains all workers which are currently running a job.

	Where	Queue	Processing
		AVERAGE	Sufia::Resque::MarshaledJob just now



FRAMEWORK REQUIREMENTS

- Mix local and remote content
- Support background processing
- **Be distributable**



FROM QUEUES TO THE CLOUD

0101
0101

0101
0101





1100

0011

1010

0101

0010

0100

1100

0011

1010

0101

0010

0100

1100

0011

1010

0101

0010

0100





FROM QUEUES TO THE CLOUD

1100
0011

1100
0011

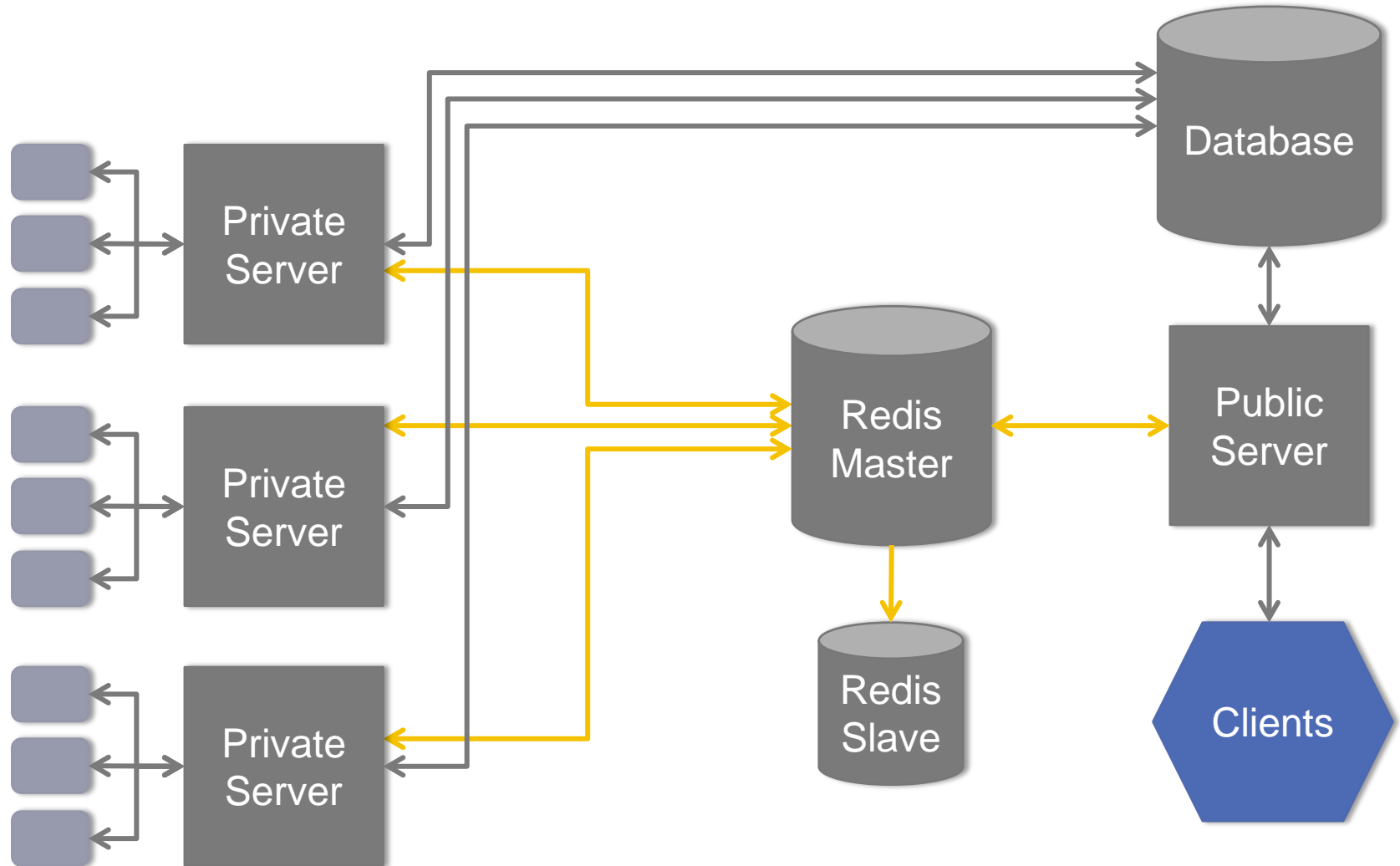
1010
0101



0010
0100

0000
0010

DISTRIBUTED PROCESSING



SCALE UP

The screenshot shows the Redis Queue Manager interface. At the top, there are tabs for Overview, Working, Failed, Queue, Workers, and Stack. The Overview tab is active. Below the tabs, there is a section titled "QUEUES" with a table listing various queues and their job counts. The queues listed are null, average, batch_update, character, bit, count, high, low, and total. The 'average' queue has 4 jobs. Below the queues section, there is a section titled "1 of 1 Workers Working" with a table showing a worker named 'Cofax-McBook-Air.local:1923' processing a job in the 'average' queue. The job being processed is 'Sufia::Resque::MarshaledJob'.

Name	Jobs
null	0
average	0
batch_update	0
character	0
bit	4
count	10
high	1
low	1
total	0

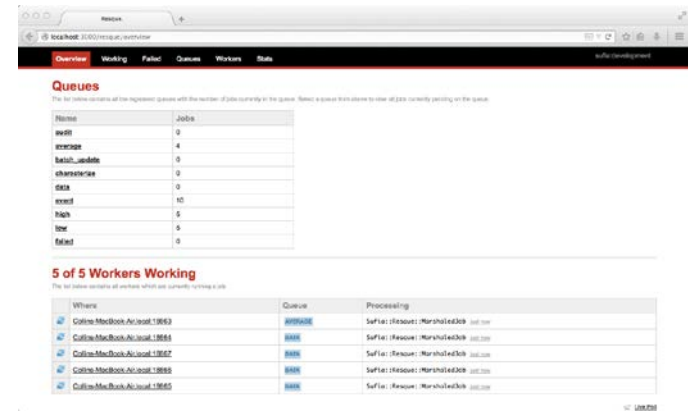
Where	Queue	Processing
Cofax-McBook-Air.local:1923	average	Sufia::Resque::MarshaledJob [1923]

1 of 1 Workers Working

The list below contains all workers which are currently running a job.






	Where	Queue	Processing
		AVERAGE	Sufia::Resque::MarshaledJob <u>just</u> <u>now</u>

SCALE UP



5 of 5 Workers Working

The list below contains all workers which are currently running a job.

	Where	Queue	Processing
		AVERAGE	Sufia::Resque::MarshaledJob <u>just now</u>
		DATA	Sufia::Resque::MarshaledJob <u>just now</u>
		DATA	Sufia::Resque::MarshaledJob <u>just now</u>
		BATCH_UPDATE	Sufia::Resque::MarshaledJob <u>just now</u>
		AVERAGE	Sufia::Resque::MarshaledJob <u>just now</u>

WE CHOSE SUFIA

WHAT IS SUFIA?

- **Ruby on Rails framework...**
- **Based on Hydra...**
- **Using Fedora Commons...**
- **And Resque**



FRAMEWORK REQUIREMENTS

- **Mix local and remote content**
- **Support background processing**
- **Be distributable**

QUESTIONS?

rotated8 (who works at) vt.edu