Funnel

Task Execution Service
Who am I?

- Alex Buchanan
- Computational Biology @ Oregon Health Science University
- Software Engineer / Data Wrangler

- We build tools to make data science better, faster, easier.
Funnel: Overview
Life of a Task

Create Task

BWA  GATK  CWL  etc.
Life of a Task

Create Task → Assign to worker

BWA
Life of a Task

Create Task → Assign to worker → Download inputs → Run commands → Upload outputs

BWA
Life of a Task

- Create Task
- Assign to worker
- Download inputs
- Run commands
- Upload outputs

Status

BWA
Life of a Task

- Slurm
- Dynamo
- Postgres
- Mongo
- etc.

- Slurm
- AWS
- Google Cloud
- SGE
- etc.

- Filesystem
- S3
- GS
- HTTP
- etc.

- Local
- Docker
- Singularity
- Conda
- etc.

- Filesystem
- HTTP/FTP
- Google Cloud
- S3
- etc.
Life of a Task

GA4GH Task Execution Schemas (TES)

Slurm
Dynamo
Postgres
Mongo
etc.

Slurm
AWS
Google Cloud
Slurm
etc.

Filesystem
S3
GS
HTTP
etc.

Local
Docker
Singularity
Conda
etc.

Filesystem
HTTP/FTP
Google Cloud
S3
etc.

BWA

Status
Dashboard

tinyurl.com/funnel-slides
CLI

```bash
> funnel run 'echo hello $NAME' --env NAME=Alex --cpu 2
bc mk887ntf6l9671q0ng

> funnel task get --view minimal bcmk887ntf6l9671q0ng
{
    "id": "bcmk887ntf6l9671q0ng",
    "state": "COMPLETE"
}

> python generate-tasks.py | funnel task create
bcmkbhfntf6l9671q0o0
bcmkbhfntf6l9671q0og
bcmkbhfntf6l9671q0p0
bcmkbhfntf6l9671q0pg
bcmkbhfntf6l9671q0q0

> funnel task cancel bcmkbhfntf6l9671q0pg

> funnel task list --state RUNNING | jq '.logs[0].startTime'
```
<table>
<thead>
<tr>
<th>Database</th>
<th>File Storage</th>
<th>Compute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embedded</td>
<td>S3</td>
<td>Embedded</td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>Google Cloud Storage</td>
<td>autoscaling</td>
</tr>
<tr>
<td>MongoDB</td>
<td>OpenStack Swift</td>
<td>AWS Batch</td>
</tr>
<tr>
<td>DynamoDB</td>
<td>HTTP</td>
<td>Traditional HPC</td>
</tr>
<tr>
<td>Google Cloud Datastore</td>
<td>Local files</td>
<td>○ SGE</td>
</tr>
<tr>
<td>Postgres</td>
<td>FTP</td>
<td>○ HTCondor</td>
</tr>
<tr>
<td></td>
<td>Azure</td>
<td>○ Slurm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>○ PBS/Torque</td>
</tr>
</tbody>
</table>

**Deployment**
- Systemd
- AWS Batch
- App Engine

**Events**
- Kafka
- Google Cloud Pub/Sub
- RabbitMQ

**Metrics**
- Prometheus
- Grafana

*Wishlist items in blue*
Task API

**CreateTask(Task) ID**

POST  /v1/tasks

**GetTask(ID) Task**

GET   /v1/tasks/123

**ListTasks() List of Tasks**

GET   /v1/tasks

**CancelTask(ID)**

POST  /v1/tasks/123:cancel

GA4GH Task Execution Schemas (TES)

[github.com/ga4gh/task-execution-schemas](github.com/ga4gh/task-execution-schemas)
{  
  "name": "md5sum example",
  "description": "A simple md5sum command.",
  "tags": {
    "custom-tag-1": "tag-value-1",
    "custom-tag-2": "tag-value-2",
  },

  "resources": {
    "cpuCores": 1,
    "ramGb": 1.0,
    "diskGb": 100.0,
  },

  "inputs": [{
    "url": "s3://my-bucket/path/input.txt",
    "path": "/inputs/in"
  }],

  "outputs": [{
    "url": "s3://my-bucket/path/md5sum_output",
    "path": "/outputs/stdout"
  }],

  "executors": [{
    "image": "alpine",
    "command": ["md5sum", "/inputs/in"],
    "stdout": "/outputs/stdout",
    "env": {
      "MY_VAR": "ONE",
    }
  }],
continued...
...continued from previous

```
"id": "bcm2kgnntf6l9671q0m0",
"state": "COMPLETE",
"creationTime": "2017-11-14T11:49:04.427163701-08:00",

"logs": [{
  "startTime": "2017-11-14T11:49:04.433593468-08:00",
  "endTime": "2017-11-14T11:49:08.487707039-08:00",
  "metadata": {
    "hostname": "worker-1",
  },
  "systemLogs": [
    "task was assigned to worker 1",
    "docker command: docker run -v /vol:/data alpine md5sum",
  ],
  "outputs": [{
    "url":  "s3://my-bucket/output-data/results.txt",
    "path":  "/outputs/results.txt",
    "sizeBytes": 123
  }],
}

"logs": [{
  "startTime": "2017-11-14T11:49:05.127885125-08:00",
  "endTime": "2017-11-14T11:49:08.484461502-08:00",
  "stdout":  "Hello, Funnel!",
  "stderr":  "",
  "exitCode": 0,
},
]
```

Overall task

Task attempt (automated for retries)

Flexible system logs

Uploaded files

Executor details
Funnel: Goals

Goals
- Easy to deploy, maintain
- Adaptable, cross-platform
- Simple API
- Easy to access, debug tasks
- Enable ecosystem of tools and plugins
Ecosystem

JSON + REST are great for building tools.
Ecosystem: Bunny

CWL workflow engine built by Seven Bridges Genomics.

github.com/rabix/bunny
Ecosystem: Cromwell

WDL + CWL workflow engine build by The Broad Institute.

github.com/broadinstitute/cromwell
Ecosystem: Galaxy

Experimental TaskRunner plugin for Galaxy.

Hackathon this weekend.

Galaxy TaskRunner prototype
Ecosystem: Tesseract

Send python function calls to remote machines.

github.com/ohsu-comp-bio/tesseract

```python
from tesseract import Tesseract, FileStore

def say_hello(name):
    return "hello, " + name

fs = FileStore("./test_store/")
r = Tesseract(fs, "http://funnel.compbio")
r.with_resources(
    cpu_cores=1,
    ram_gb=4,
    disk_gb=None,
    docker="python:2.7",
    libraries=["cloudpickle"]
)

future = r.run(say_hello, "Alex")
result = future.result()
print(result)
```
Ecosystem: Gaia

Workflow engine with simple REST API.

github.com/bmeg/gaia
Ecosystem: TESK

GA4GH Task API implementation. Focused on Kubernetes and federation.

github.com/EMBL-EBI-TSI/TESK
Ecosystem: cwl-tes

CWL engine that submits tasks via TES.
Based on cwltool.

[github.com/common-workflow-language/cwl-tes](https://github.com/common-workflow-language/cwl-tes)
Future Work

- **Compute**
  - Google Pipelines API
  - Kubernetes
  - Mesos

- **Storage**
  - FTP
  - SSH (scp, rsync)
  - Azure

- **Deployment / Admin**
  - App Engine, Beanstalk, Lambda, etc.

- **Extensions**
  - Call caching ([issues/545](https://tinyurl.com/funnel-slides))
  - Download caching ([issues/5](https://tinyurl.com/funnel-slides))
  - Task Dependencies ([issues/39](https://tinyurl.com/funnel-slides))
Future Work: Auth

- API access
- Storage
- Docker
- Access compute as user (e.g. user account on Google Cloud VM)
Thanks!

Alex Buchanan   Adam Struck   Ryan Spangler   Allison Creason   Brian Walsh   Kyle Ellrott

Web: ohsu-comp-bio.github.io/funnel/
Slides: tinyurl.com/funnel-slides
Fin
Life of a Task

Create Task ➔ Assign to worker ➔ Download inputs ➔ Run commands ➔ Upload outputs

BWA

Status