



Genomics in the Cloud

Book Club - Week 11

February 8, 2021

Agenda

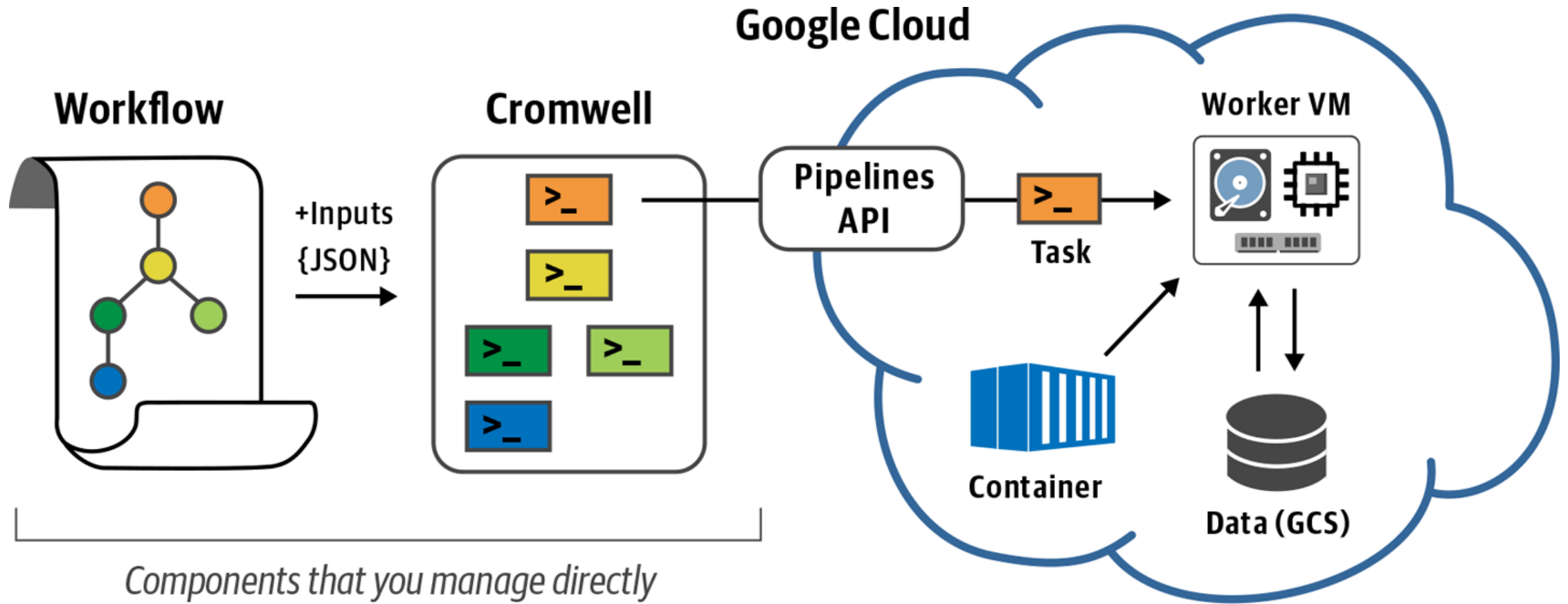
- Chapter 10: Running Single Workflows at Scale with Pipelines API
- Additional resources
- Open discussion



Chapter 10: Running Single Workflows at Scale with Pipelines API

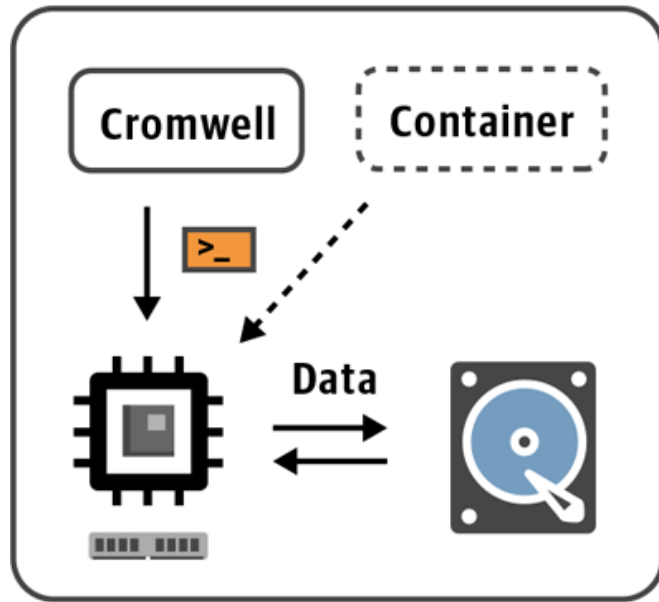
Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

Cromwell + PAPI overview



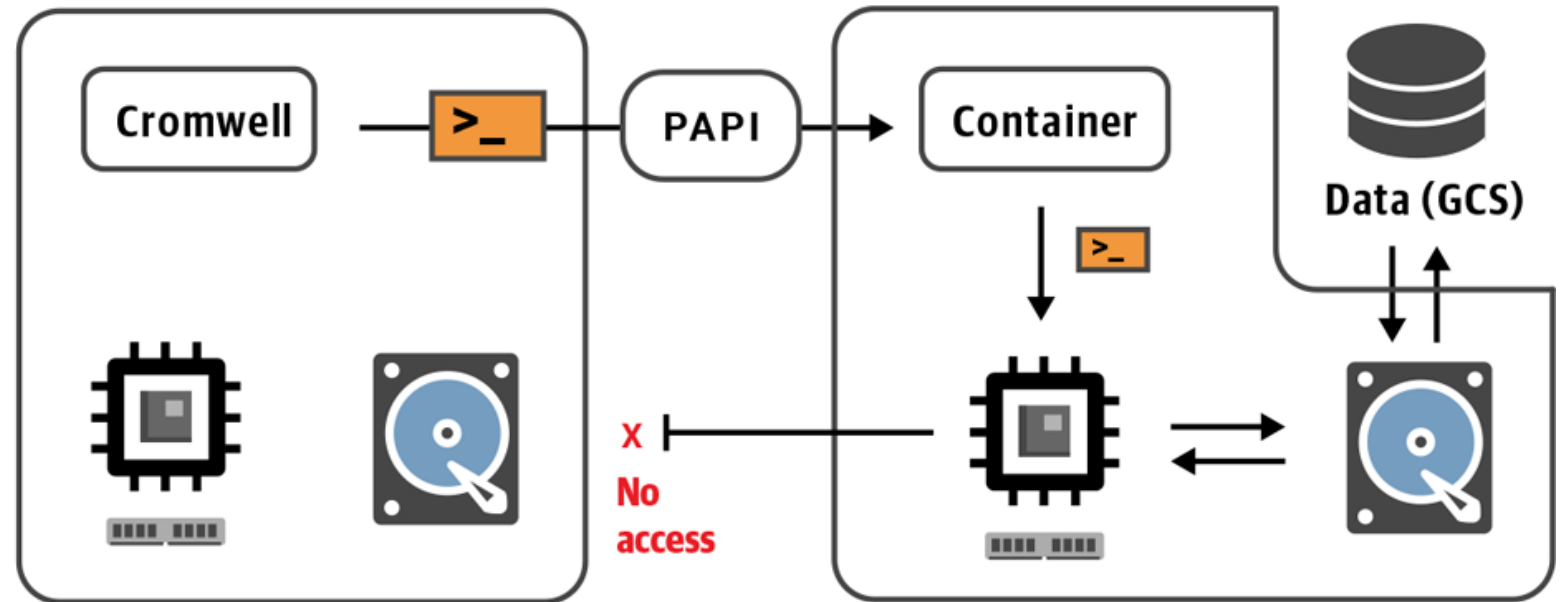
Local vs PAPI execution

Work inside single VM



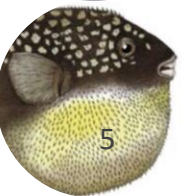
VM

Dispatching work to other VMs



Master VM

Worker VM



Enabling genomics APIs in GCP



Compute Engine API

Google

Compute Engine API



Genomics API (deprecated)

Google

Uploads, processes, queries, and searches Genomics data in the cloud.



Google Cloud Storage JSON API

Google

Lets you store and retrieve potentially-large, immutable data objects.

<https://console.cloud.google.com/apis>



1. Getting started

```
$ cat ~/book/code/config/google.conf
```

```
$ mkdir ~/sandbox-10
```

```
$ cp ~/book/code/config/google.conf ~/sandbox-10/my-google.conf
```

```
$ export CONF=~/sandbox-10
```

```
$ export BIN=~/book/bin
```

```
$ export WF=~/book/code/workflows
```

```
$ export BUCKET="gs://my-bucket"
```

```
$ nano ~/sandbox-10/my-google.conf
```

In **my-google.conf**, change <google-project> to your project name



2. Running scattered HaplotypeCaller via PAPI

```
$ gcloud auth application-default login
```

```
$ cat $WF/scatter-hc/scatter-haplotypecaller.gcs.inputs.json
```

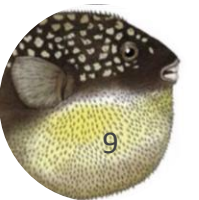
```
$ java -Dconfig.file=$CONF/my-google.conf -jar $BIN/cromwell-48.jar \  
  run $WF/scatter-hc/scatter-haplotypecaller.wdl \  
  -i $WF/scatter-hc/scatter-haplotypecaller.gcs.inputs.json
```

```
# Check status  
https://console.cloud.google.com/home/dashboard
```

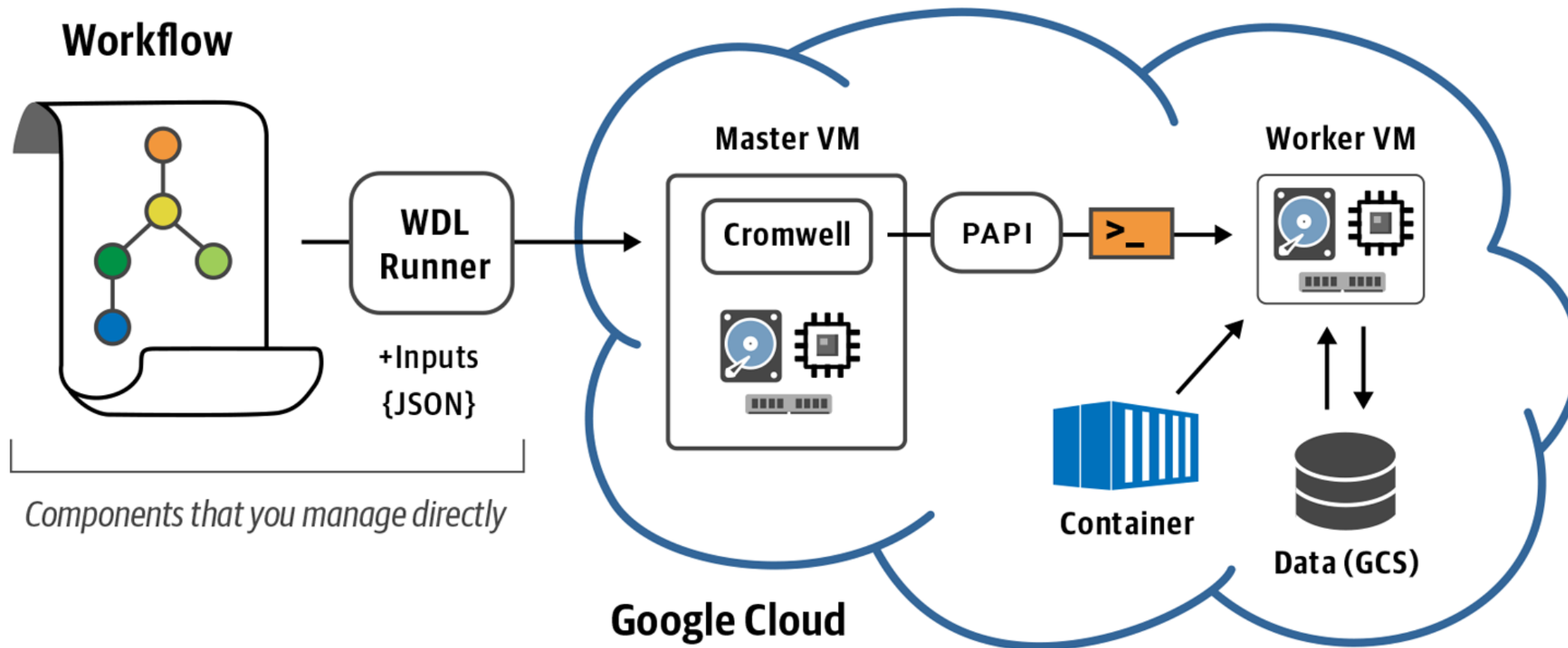


Suggested cost-saving optimizations

1. Dynamic sizing for resource allocation
 - Evaluate file sizes at runtime before requesting VMs
2. File streaming to GATK4 tools
 - Use `localization_optional: true`
 - GCP-only, not available for Picard tools bundled in GATK
3. Preemptible VM instances
 - 20% of list price
 - Saving your progress (checkpointing) now available



WDL Runner



3. WDL Runner

```
$ export WR_CONF=~/.book/code/config  
$ export WR_PIPE=~/.book/wdl_runner/wdl_runner
```

Unset the default compute zone







```
$ gcloud config set compute/zone ""
```

```
$ gcloud alpha genomics pipelines run \  
  --pipeline-file $WR_PIPE/wdl_pipeline.yaml \  
  --regions us-west1 \  
  --inputs-from-file WDL=$WF/scatter-hc/scatter-haplotypecaller.wdl, \  
  WORKFLOW_INPUTS=$WF/scatter-hc/scatter-haplotypecaller.gcs.inputs.json, \  
  WORKFLOW_OPTIONS=$WR_CONF/empty.options.json \  
  --env-vars WORKSPACE=$BUCKET/wdl_runner/test/work, \  
  OUTPUTS=$BUCKET/wdl_runner/test/output \  
  --logging $BUCKET/wdl_runner/test/logging
```



4. Monitoring WDL Runner

```
$ cd ~/book/wdl_runner  
$ bash monitoring_tools/monitor_wdl_pipeline.sh <operation_ID>
```




<input type="checkbox"/> Name ^	Zone
<input type="checkbox"/>  genomics-book	us-east4-a
<input type="checkbox"/>  google-pipelines-worker-49df01d13f4e9a8a425fc9c3d7da91b7	us-central1-b
<input type="checkbox"/>  google-pipelines-worker-4dfc38f4ed8642c2e39d3cbd013410fd	us-central1-b
<input type="checkbox"/>  google-pipelines-worker-50bf05a598c0bfbb64e7c6761b01b030	us-central1-b
<input type="checkbox"/>  google-pipelines-worker-f4628e21ce5d31017f0ef3cac27f829c	us-central1-b
<input type="checkbox"/>  google-pipelines-worker-f4b02a3582e27f2c215da8d20a7a0371	us-east4-a

The WDL Runner Server VM!



WDL Runner output

[Buckets](#) / [genomics-book](#) / [wdl_runner](#) / test

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified
<input type="checkbox"/>	 logging	110.63 KB	application/octet-stream	Standard	12/15/19, 4:42:05 AM UTC-5
<input type="checkbox"/>	 output/	—	Folder	—	—
<input type="checkbox"/>	 work/	—	Folder	—	—



Additional resources

- Cromwell documentation
 - <https://cromwell.readthedocs.io/en/stable/tutorials/PipelinesApi101/>
 - <https://cromwell.readthedocs.io/en/stable/backends/Google/>
 - <https://cromwell.readthedocs.io/en/stable/RuntimeAttributes/>
- GCP Life Sciences API (beta)
 - <https://cloud.google.com/life-sciences/docs/reference/gcloud-examples>
- Preemptible VMs
 - <https://cloud.google.com/preemptible-vms/>
- WDL Runner
 - <https://github.com/broadinstitute/wdl-runner>





Thank you for joining us today!

Next week: Chapter 11

Next meeting: February 15, 2021