



Genomics in the Cloud

Book Club - Week 5

December 28, 2020

Agenda

- Chapter 4: First Steps in the Cloud
- Additional resources
- Working session / open discussion





Our guest
speaker

Dr. Geraldine
Van der
Auwera

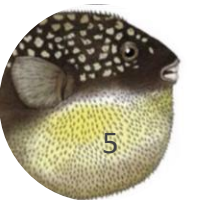


Chapter 4: First Steps in the Cloud

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.

Google Cloud Platform (GCP)

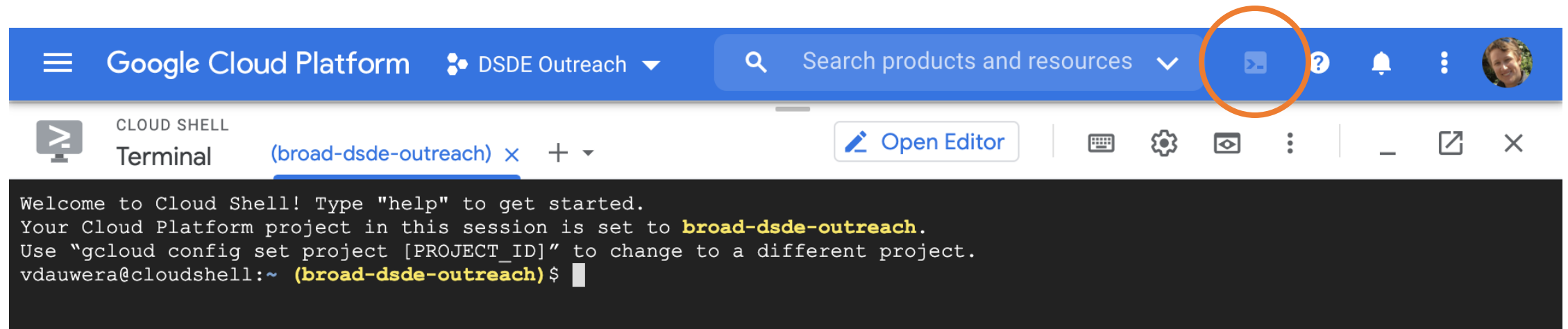
- Why Google Cloud / what applies to other clouds
- Account setup / get your \$300 free credits!
- Services used in the book
 - **Cloud shell (free)** -- basic virtual machines (VM), good for quick tasks and practice
 - **Google Cloud Storage (GCS)** via gsutil (file management utility)
 - **Google Cloud Engine (GCE)** -- custom virtual machines (VMs), for real work
 - Life Sciences API (née Pipelines API) – batch execution on GCE (via Cromwell, Terra)
 - Dataproc (Spark) option under the hood in Terra Jupyter Notebooks (Ch 12)
- Related / of interest:
 - BigQuery (datastore) – for tabular data e.g. cohort variants



Cloud Shell basics

- Free service – basic VM (not enough for genomics)

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

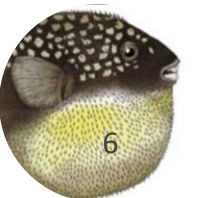


The screenshot displays the Google Cloud Platform (GCP) console interface. At the top, a blue header bar contains the 'Google Cloud Platform' logo, a dropdown menu for 'DSDE Outreach', a search bar with the text 'Search products and resources', and several utility icons. An orange circle highlights the 'Cloud Shell' icon (a blue terminal window) in the top right corner of the header. Below the header, the 'Cloud Shell' terminal window is open, showing a 'Terminal' tab for the project '(broad-dsde-outreach)'. The terminal content includes a welcome message, instructions on how to change the project, and the current shell prompt 'vdauwera@cloudshell:~ (broad-dsde-outreach)\$'.

Google Cloud Platform DSDE Outreach Search products and resources

CLOUD SHELL Terminal (broad-dsde-outreach) x + Open Editor

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to broad-dsde-outreach.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
vdauwera@cloudshell:~ (broad-dsde-outreach)$
```



Managing files & buckets with `gsutil`

- List bucket contents

```
$ gsutil ls gs://genomics-in-the-cloud
```

- Copy/move contents to/from local disk & between buckets

```
$ gsutil cp gs://bucket1/readme.txt .
```

```
$ gsutil cp readme.txt gs://bucket2/stuff/
```

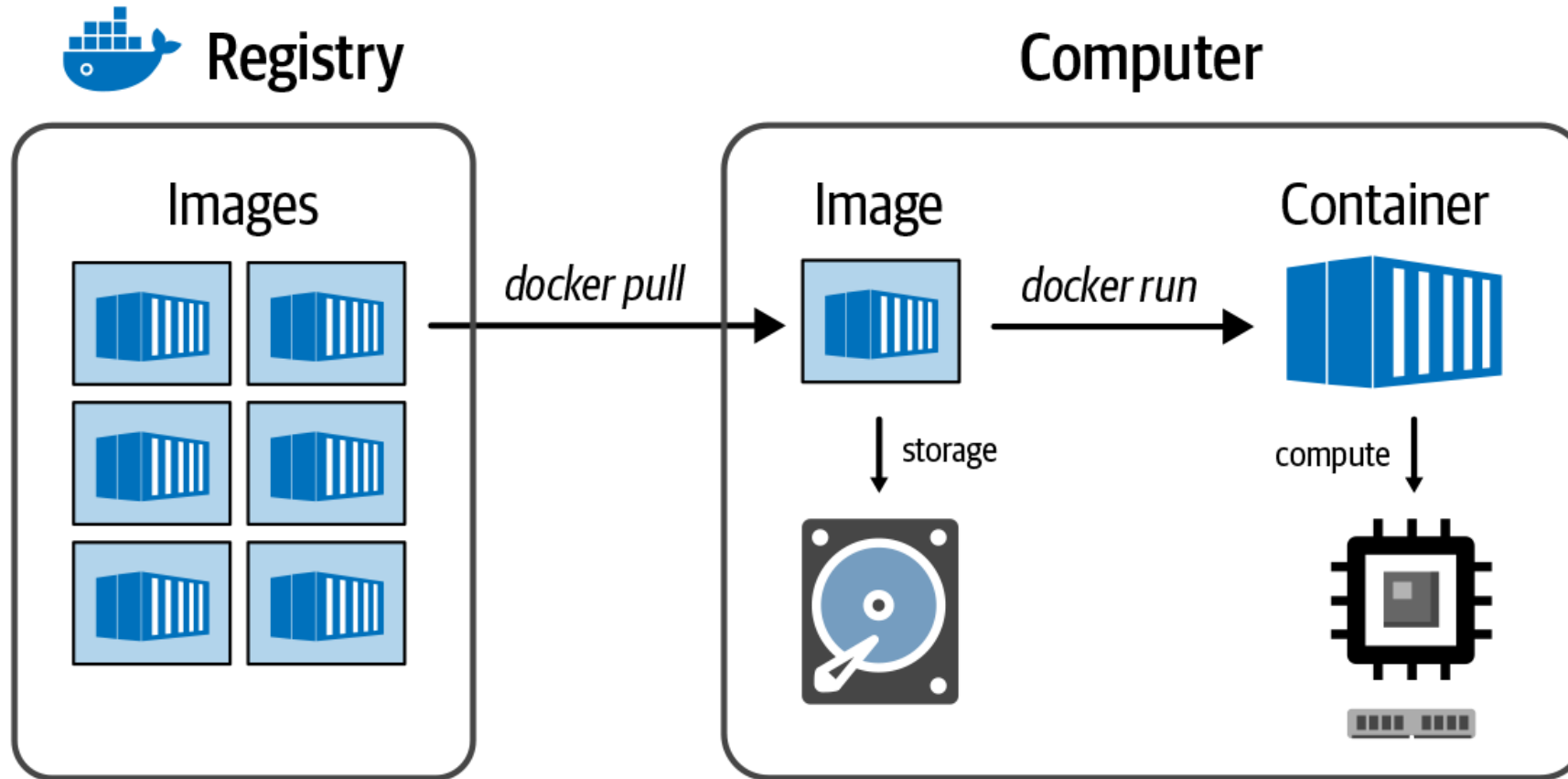
```
$ gsutil cp gs://bucket1/readme.txt gs://bucket2/stuff/
```

- Create own bucket

```
$ gsutil mb gs://my-totally-unique-bucket
```



Key Docker commands: **pull** and **run**



Pulling and running in practice

- Retrieve a container image

```
$ docker pull library/ubuntu
```

- Spin up the container interactively

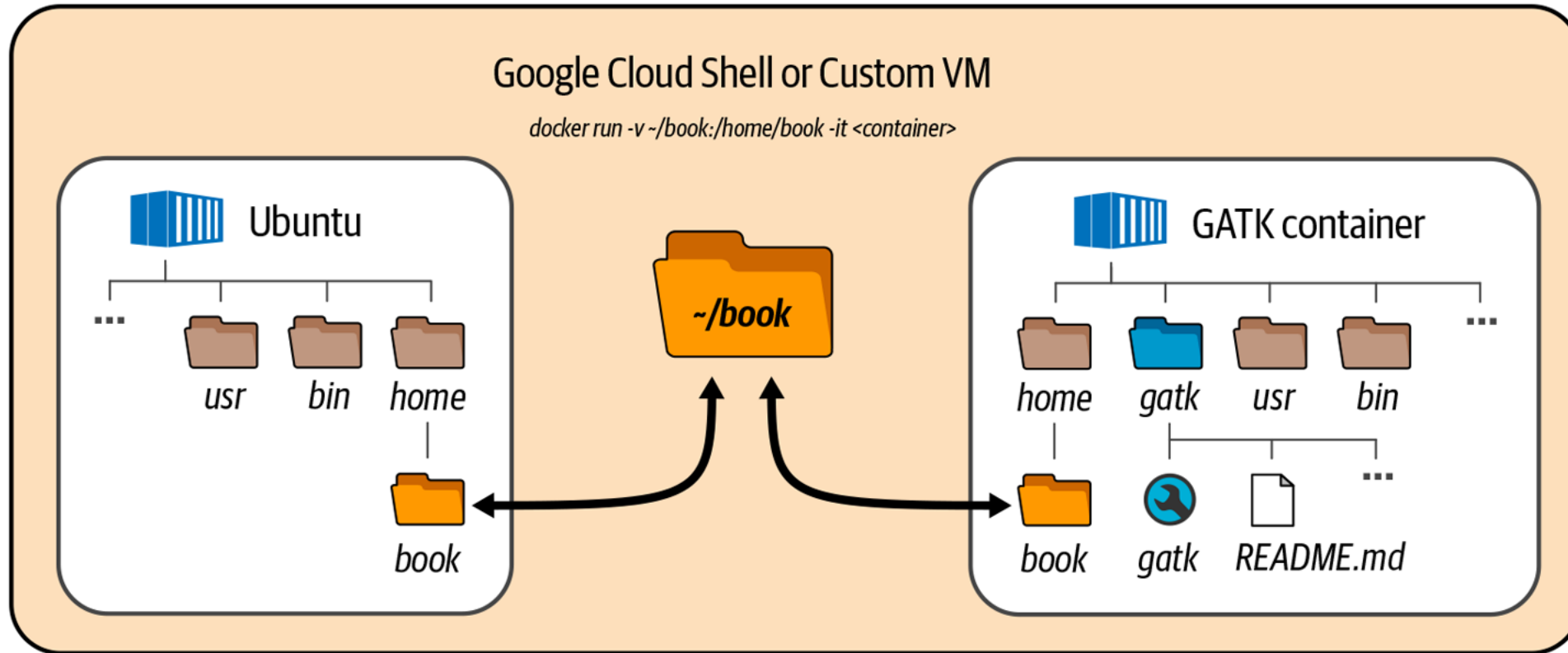
```
$ docker run -it library/ubuntu /bin/bash  
root@0b87869c76f7:/#
```

- Spin it up with a mounted volume

```
$ docker run -v ~/dir1:/home/dir2 -it library/ubuntu /bin/bash
```




Using a mounted volume





Custom VMs on GCE


[←](#) Create an instance


To create a VM instance, select one of the options:


 **New VM instance**
Create a single VM instance from scratch


 **New VM instance from template**
Create a single VM instance from an existing template

 **Marketplace**
Deploy a ready-to-go solution onto a VM instance

Name 
Name is permanent

Region 
Region is permanent

Zone 
Zone is permanent


Machine configuration 

Machine family

Machine types for common workloads, optimized for cost and flexibility

Series
N1
Powered by Intel Skylake CPU platform or one of its predecessors

Machine type




vCPU


2


Memory

7.5 GB

[CPU platform and GPU](#)

Container 
☐ Deploy a container image to this VM instance. [Learn more](#)

Boot disk 




New 100 GB standard persistent disk

Image

Ubuntu 18.04 LTS

\$59.08 monthly estimate
That's about \$0.081 hourly
Pay for what you use: No upfront costs and per second billing
[Details](#)



11

Important VM config items

- Region/Zone – physical location of datacenters -> compartmentalized services
- Machine Type – processor(s) & memory
- Container – to preinstall; convenient option but not used here
- Boot Disk – OS distribution (from predefined list or custom)



Log in over SSH & authenticate

- Secure connection to VM
 - Once logged in, can use it like a 'normal' machine
- Authenticate with `gcloud init`
 - Ensures credentials are set up to access resources
- DO NOT SHARE ACCESS TO VM IF USING PERSONAL CREDENTIALS
 - **Service accounts** exist for sharing resources & automating things



Set up data / GATK container

- Copy book data from bucket to VM*

```
$ mkdir ~/book
```

```
$ gsutil -m cp -r gs://genomics-in-the-cloud/v1/* ~/book/
```

- Install Docker on VM

```
$ curl -sSL https://get.docker.com/ | sh
```

- Pull official GATK container from Google Cloud Registry (also available in Dockerhub)

```
$ docker pull us.gcr.io/broad-gatk/gatk:4.1.3.0
```

- Spin up container in interactive mode with mounted volume

```
$ docker run -v ~/book:/home/book -it us.gcr.io/broad-gatk/gatk:4.1.3.0 /bin/bash  
root@0c83569e76b8:/#
```

* GATK can read/write directly from/to GCS – this is left for later in the book for simplicity + demonstrate generally applicable usage



Test that GATK runs

```
# gatk
Usage template for all tools (uses --spark-runner LOCAL when used with a Spark tool)
  gatk AnyTool toolArgs
Usage template for Spark tools (will NOT work on non-Spark tools)
  gatk SparkTool toolArgs [ -- --spark-runner <LOCAL | SPARK | GCS> sparkArgs ]
Getting help
  gatk --list          Print the list of available tools
  gatk Tool --help     Print help on a particular tool
Configuration File Specification
  --gatk-config-file   PATH/TO/GATK/PROPERTIES/FILE
gatk forwards commands to GATK and adds some sugar for submitting spark jobs
  --spark-runner <target> controls how spark tools are run
    valid targets are:
      LOCAL:      run using the in-memory spark runner
      SPARK:      run using spark-submit on an existing cluster
                  --spark-master must be specified
                  --spark-submit-command may be specified to control the Spark submit command
                  arguments to spark-submit may optionally be specified after --
      GCS:        run using Google cloud dataproc
                  commands after the -- will be passed to dataproc
                  --cluster <your-cluster> must be specified after the --
                  spark properties and some common spark-submit parameters will be translated
                  to dataproc equivalents
  --dry-run          may be specified to output the generated command line without running it
  --java-options 'OPTION1[ OPTION2=Y ... ]' optional - pass the given string of options to
                  the java JVM at runtime.
                  Java options MUST be passed inside a single string with space-separated values
```



Connect IGV genome browser to GCS

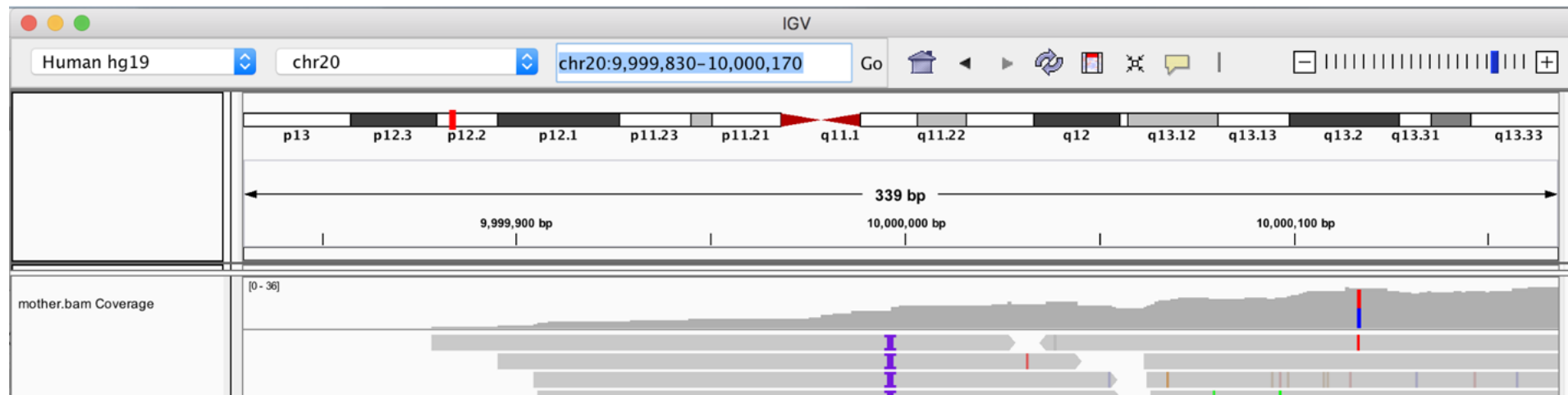
Load from URL

File URL:

*Specify url to an index file. **Required for BAM and indexed files***

Index URL:

OK Cancel



Remember to stop your VM!

genomics


Filter VM instances

×

?

Columns

▼

<input type="checkbox"/> Name ^	Zone	Recommendation	Internal IP	External IP	Connect	
<input type="checkbox"/>  genomics-book	us-east4-a		10.150.0.2 (nic0)	35.194.83.111	SSH ▼	<div><div>⋮</div><div><div>Start</div><div>Stop</div><div>Reset</div><div>Delete</div></div><div><div>View network details</div><div>View logs</div></div></div>



Additional resources

- Google Cloud console
 - <https://console.cloud.google.com/>
- Terra “New to Cloud” orientation page
 - <https://terra.bio/resources/new-to-cloud/>
- Re-sizing persistent disk space
 - <https://cloud.netapp.com/blog/google-cloud-persistent-disk-how-to-resize-and-use>
- Changing machine types
 - <https://cloud.google.com/compute/docs/instances/changing-machine-type-of-stopped-instance>





Thank you for joining us today!

Next week: Chapter 5

Next meeting: January 4, 2021