Cloud Solutions for State NGS Bioinformatics and Storage

Joel R. Sevinsky, Ph.D.

InFORM 2017

Thursday, Nov 9th, 9:15-10:15 AM
The Cloud’s Big Secret

There is no cloud
it’s just someone else’s computer
What is cloud computing?

Cloud Computing = Someone else’s computer
What is cloud computing?

Cloud Computing = Someone else’s computer

Virtualization

Internet for connectivity
What is a virtual machine?

Standard Computer
1 box = 1 computer
What is a virtual machine?

Standard Computer
1 box = 1 computer

VM Server
1 box = many computers
Large Scale Cloud Resources

Amazon/Google/Microsoft Cloud

Hypervisor
Cloud Computing is HUGE!

Amazon/Google/Microsoft Cloud
Cloud in Current Workflows

Amazon Web Services

Hypervisor

OS
BaseSpace

OS
App

Printer
Cloud in Current Workflows

Amazon Web Services

CDC CE

Hypervisor

OS
BaseSpace

OS
App

Hypervisor

OS
CE

OS
App

---

CDPHE
Laboratory Services Division
Department of Public Health & Environment
Cloud in Current Workflows

Amazon Web Services

CDC CE

NCBI SRA (AWS)
Cloud in Current Workflows

Google Cloud

Amazon Web Services

CDC CE

NCBI SRA (AWS)
Cloud Bioinformatics at CDPHE

Amazon Web Services

Google Cloud

NCBI SRA (AWS)

Hypervisor

OS

BaseSpace

OS

App

Hypervisor

Linux

StaPH-B

OS

App

Hypervisor

OS

App

Hypervisor

No PHI from CDPHE is uploaded to the cloud! All data is publically available through NCBI.
Cloud Bioinformatics at CDPHE

VM instances

- Name: qc-instance, Zone: us-west1-a, Recommendation: None, Internal IP: 10.138.0.2, External IP: None, Connect: SSH
- Name: staphb-0-3-arizona, Zone: us-central1-f, Recommendation: None, Internal IP: 10.128.0.4, External IP: None, Connect: SSH
- Name: staphb-0-3-kraken, Zone: us-central1-f, Recommendation: None, Internal IP: 10.128.0.7, External IP: None, Connect: SSH
- Name: staphb-0-3-logan, Zone: us-central1-f, Recommendation: None, Internal IP: 10.128.0.2, External IP: None, Connect: SSH
- Name: staphb-0-3-new-mexico, Zone: us-central1-b, Recommendation: None, Internal IP: 10.128.0.6, External IP: None, Connect: SSH
- Name: staphb-0-3-wyoming, Zone: us-central1-f, Recommendation: None, Internal IP: 10.128.0.5, External IP: None, Connect: SSH
- Name: staphb-0-5, Zone: us-central1-c, Recommendation: None, Internal IP: 10.128.0.3, External IP: None, Connect: SSH
Disclaimer: VMs Use the Command Line

Don’t fear the command line. Fear is the path to the dark side. Fear leads to anger. Anger leads to hate. Hate leads to suffering.
VM images contain pre-installed pipelines for SOP bioinformatics

Ctrl-c

Ctrl-v
VM images contain pre-installed pipelines for SOP bioinformatics

Then (APHL Bioinformatics Primer)

Steps for performed analysis - all copy and paste

1. basemount Basespace  #you will be required to paste auth url in your browser
2. mkdir SEQ056  #whatever you want to call your analysis
3. cp /home/staphb/Basespace/Projects/SEQ056_QC_Listeria/Samples/PN*/Files/* .gz ./SEQ056/
4. cd SEQ056
5. mkdir clean
6. (for i in *R1_001.fastq.gz; do b=`basename $i _R1_001.fastq.gz`; run_assembly_shuffleReads.pl $b"_R1_001.fastq.gz" $b"_R2_001.fastq.gz" > clean/$b.fastq; echo $b; run_assembly_trimClean.pl -i clean/$b.fastq -o clean/$b.cleaned.fastq.gz --nosingletons; rm clean/$b.fastq; done;)
    >& cleaning.log &
7. cd clean
8. run_assembly_readMetrics.pl *.fastq.gz --fast --numcpus 8 --e 3000000 | sort -k3,3n > readMetrics.tsv
9. more readMetrics.tsv

<table>
<thead>
<tr>
<th>File</th>
<th>avgReadLength</th>
<th>totalBases</th>
<th>minReadLength</th>
<th>maxReadLength</th>
<th>avgQuality</th>
<th>numReads</th>
<th>PE?</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNUSAL003218_S1_L001.cleaned.fastq.gz</td>
<td>201.14</td>
<td>59452601</td>
<td>62</td>
<td>251</td>
<td>35.72</td>
<td>295576</td>
<td>yes</td>
<td>19.81</td>
</tr>
</tbody>
</table>
VM images contain pre-installed pipelines for SOP bioinformatics

Steps for performed analysis - all copy and paste

1. basemount Basespace  #you will be required to paste auth url in your browser
2. mkdir SEQ056        #whatever you want to call your analysis
3. cp
   /home/staphb/Basespace/Projects/SEQ056_QC_Listeria/Samples/PN*/Files/* .gz ./SEQ056/
4. cd SEQ056
5. mkdir clean
6. (for i in *R1_001.fastq.gz; do b=`basename $i _R1_001.fastq.gz`; run_assembly_shuffleReads.pl $b"_R1_001.fastq.gz" $b"_R2_001.fastq.gz" >cLean/$b.fastq; echo $b; run_assembly_trimClean.pl -i clean/$b.fastq -o clean/$b.cleaned.fastq.gz --nosingletons; rm clean/$b.fastq; done;)
   >& cleaning.log &
7. cd clean
8. run_assembly_readMetrics.pl *.fastq.gz --fast --numcpus 8 -e 3000000 |sort -k3,3n > readMetrics.tsv
9. more readMetrics.tsv

Now:

Copy and paste one command

<table>
<thead>
<tr>
<th>File</th>
<th>avgReadLength</th>
<th>totalBases</th>
<th>minReadLength</th>
<th>maxReadLength</th>
<th>avgQuality</th>
<th>numReads</th>
<th>PE?</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNUSAL003218_S1_L001</td>
<td>201.14</td>
<td>59452601</td>
<td>62</td>
<td>251</td>
<td>35.72</td>
<td>295576</td>
<td>yes</td>
<td>19.81</td>
</tr>
</tbody>
</table>
Cluster Detection Pipeline
E. coli O26 Outbreak

hqSNP Tree

Pairwise hqSNP Matrix
Bioinformatics Resource Support

Google Cloud

VM instances

<table>
<thead>
<tr>
<th>Name</th>
<th>Zone</th>
<th>Recommendation</th>
<th>Internal IP</th>
<th>External IP</th>
<th>Connect</th>
</tr>
</thead>
<tbody>
<tr>
<td>qc-instance</td>
<td>us-west1-a</td>
<td></td>
<td>10.138.0.2</td>
<td>None</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-3-arizona</td>
<td>us-central1-f</td>
<td></td>
<td>10.128.0.4</td>
<td>None</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-3-kraken</td>
<td>us-central1-f</td>
<td></td>
<td>10.128.0.7</td>
<td>None</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-3-logan</td>
<td>us-central1-f</td>
<td></td>
<td>10.128.0.2</td>
<td>35.193.123.27</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-3-new-mexico</td>
<td>us-central1-b</td>
<td></td>
<td>10.128.0.6</td>
<td>35.184.193.39</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-3-wyoming</td>
<td>us-central1-f</td>
<td></td>
<td>10.128.0.5</td>
<td>35.193.77.40</td>
<td>SSH</td>
</tr>
<tr>
<td>staphb-0-5</td>
<td>us-central1-c</td>
<td></td>
<td>10.128.0.3</td>
<td>None</td>
<td>SSH</td>
</tr>
</tbody>
</table>

CO  NM  WY
Reasons to Use Cloud Computing

• All resources outside of Bionumerics are cloud based on the Google Cloud Platform, both processing and storage capacity.

• Advantages:
  • On-demand, expandable, flexible, inexpensive solution with no capital costs. Currently spend about $0.28/hour while processing on 8 CPU 30 GB RAM Linux VM.
  • Work outside the state IT network allowing full admin rights and open source software utilization.
  • All data transfer of large data sets occurs outside state IT network, relieving the burden on limited network bandwidth.
  • When using Basespace, data is transferred directly to the cloud VM, processed, and data/analysis are stored in a Google storage bucket. Only results are downloaded to the state network.
  • All images created within the CDPHE program and instantly shareable and can be replicated at other state PHLs for consistency of bioinformatics analysis.
  • Yet to be tested but cloud computing should be an asset for CLIA validation due to the ability to lock down images for stability and validation.

• Disadvantages:
  • Not every state PHL has access to the Google Cloud and Basespace.
Current Status

• CG Pipeline from CDC PulseNet for MiSeq data QC
• Microbial Identification (MASH, SeqSero, SerotypeFinder, and more)
• Kraken (for identification of contamination)
• Reference free clustering (UPHL pipeline)
• Reference based clustering (lyve-SET used by PulseNet)
• Connectivity to BaseSpace, NCBI, Google Storage Buckets, and more
Future Plans and Improvements

• More pipelines, hopefully through StaPH-B collaborations

• Replicate virtual machines in AWS and MS Azure
  • Volunteers???

• Distribute, distribute, distribute
  • We can talk with your IT departments

• Promote usage on organisms where established CDC/PHL infrastructure is not in place
Thanks

• CDPHE
  • Logan Fink – APHL/CDC Bioinformatics Fellow extraordinaire

• UPHL
  • Kelly Oakeson – Reference free clustering

• CDC PulseNet
  • Lee Katz – creator of lyve-SET, assisted with implementation on VM

• CDC OAMD
  • Funding and ideas
  • Bioinformatics Resource Support Sites (CO, MN, MA)
EXPERTS IN BIOINFORMATICS USE HIGH-PERFORMANCE COMPUTING TO INTEGRATE DATA FROM EPIDEMIOLOGIC INVESTIGATIONS AND GENOMIC SEQUENCING. THEY CONNECT THE DOTS, EXPLORING HOW PATHOGENS SPREAD, MUTATE, AND DEVELOP ANTIMICROBIAL RESISTANCE TO HELP SOLVE OUTBREAKS.

BIOINFORMATICS

Joel Sevinsky (CO)
joel.sevinsky@state.co.us

Logan Fink (CO)
logan.fink@state.co.us

StaPH-B Members
Kelly Oakeson (UT)
Kevin Libuit (VA)
Sean Wang (MN)
And many others