Infrastructure Solutions for Bioinformatics: Virginia’s Approach

Lauren Turner, PhD
Lead Scientist
VA’s Bioinformatics Infrastructure Development: A Story in Three Phases

• Pre-2016: Reliance on external partners for analysis
• 2016 – 2018: Cloud resource through academic collaboration
• 2017 – Present: On-premises resources
• 2019: Agency contract for cloud resources
Considerations

• What are our compute requirements?
• Is the environment scalable? Dynamic?
• How will the environment be managed?
• Who will administer the environment?
• What is the process for accessing and developing the environment?
• Do policies exist for access or use of open source software, Linux OS, cloud resources?
• What is the longevity of the environment or infrastructure?
• Can the environment be accessed remotely?
Phase I: Back in 2016...

- Partnership with University of Virginia School of Medicine Bioinformatics Core Director
  - Virginia Governor’s Data Internship Program
  - Project-based work requiring compute environment
  - Amazon Web Services via university account
    - Analysis of reads stored publically on NCBI
  - Goal: Evaluate bioinformatics for on-demand state PHL use
Logistics of the Academic Partnership

Academic Advisor:
- Infrastructure provision
- AWS account administrator
- Intern AWS access
- Intern AWS environment orientation
- Bioinformatics/scripting SME

Intern:
- Pipeline development
- Pipeline execution in AWS
- Project dataset analysis

Agency Sponsor:
- Identification of problem
- Identification of needs
- Providing datasets
- Evaluation of results
Logistics of the Academic Partnership

**Considerations**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are our compute needs?</td>
<td>Not sure, exploratory phase</td>
</tr>
<tr>
<td>Is the environment scalable?</td>
<td>✓</td>
</tr>
<tr>
<td>Is the environment dynamic?</td>
<td>✓</td>
</tr>
<tr>
<td>How will the environment be managed?</td>
<td>Overseen by UVA partner</td>
</tr>
<tr>
<td>Who will administer the environment?</td>
<td>Overseen by UVA partner</td>
</tr>
<tr>
<td>Do policies exist for access/use?</td>
<td>Not at the agency-level</td>
</tr>
<tr>
<td>What is the longevity of the infrastructure?</td>
<td>Depends on longevity of the partnership!</td>
</tr>
<tr>
<td>Remote access?</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Major benefit:** Access to a dynamic, scalable environment with intern defined processing and storage specifications.

**Major drawback:** Access to environment dependent on university partner affiliation. Without further oversight and approval, use limited to development needs.
Phase II: (2017) On Premises Infrastructure Goals:

- Capacity independent from university collaboration
- State procurement and management/administration of hardware
- Production environment for validated bioinformatics pipeline execution by DCLS Scientists
- Open-source software access for pipeline development in “the sandbox”
Logistics of Local Infrastructure

- Agency approval required for:
  - Procurement of Linux OS workstation
  - Installation and use of open source software
- Oversight of development and production environments
- Review and approval of software for production use

- Caveats:
  - Education on open source software applications
  - Sequestration of Linux OS and open source software
  - Only de-identified pathogen sequences permitted
  - No PHI/PII
  - No human gDNA sequencing
## Local Infrastructure Logistics

### Considerations

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are our compute needs?</td>
<td>High-Range Workstation</td>
</tr>
<tr>
<td>Is the environment scalable?</td>
<td>X</td>
</tr>
<tr>
<td>Is the environment dynamic?</td>
<td>X</td>
</tr>
<tr>
<td>How will the environment be managed?</td>
<td>DCLS IT</td>
</tr>
<tr>
<td>Who will administer the environment?</td>
<td>DCLS IT</td>
</tr>
<tr>
<td>Do policies exist for access/use?</td>
<td>Developed during project</td>
</tr>
<tr>
<td>What is the longevity of the infrastructure?</td>
<td>Limited – 5 years?</td>
</tr>
<tr>
<td>Remote access?</td>
<td>X</td>
</tr>
</tbody>
</table>

**Major benefit:** Agency/DCLS managed development and production environment.

**Major drawbacks:** Policies and procedures developed *in situ*, IT administrative support critical, maintenance of environment balanced with all other IT needs, limited scalability and no transportability.
Phase III: (2019) Agency-Managed Cloud Resources

Goals:

• Access to dynamic and scalable compute environment
• Reduction of long-term infrastructure costs
• Increased accessibility for remote computing
• More cost efficient long-term data storage
• Production and development environments
• Increased user capacity
Advancing Toward Cloud Computing

- Virginia Information Technology Agency (VITA) engagement since 2017
- VITA’s enterprise cloud solution policies developed alongside DCLS’ Needs
- DCLS’ requests for dynamic, scalable resources supported and reinforced by agency and PHL community:
  - Technical requirements meetings
  - Interagency conference calls and agreements
  - PHL Bioinformatician consultation on costs, applications, scalability and administration of cloud resources
Agency-Managed Cloud Resources (2019)

- Access via a VITA-managed contract with second party provider
- DCLS serving as a test case, first user PaaS
## Logistics for Agency Managed Cloud Resources

<table>
<thead>
<tr>
<th>Considerations</th>
<th>Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are our compute needs?</td>
<td>Dynamic</td>
</tr>
<tr>
<td>Is the environment scalable?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is the environment dynamic?</td>
<td>Yes</td>
</tr>
<tr>
<td>How will the environment be managed?</td>
<td>Unysis (VITA Contractor)</td>
</tr>
<tr>
<td>Who will administer the environment?</td>
<td>Unysis (VITA Contractor)</td>
</tr>
<tr>
<td>Do policies exist for access/use?</td>
<td>VITA compliant</td>
</tr>
<tr>
<td>What is the longevity of the infrastructure?</td>
<td>Not constraint</td>
</tr>
<tr>
<td>Remote access?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Major benefit:** Access to a dynamic and scalable compute environment that complies with Commonwealth IT policies.

**Major drawbacks:** To be determined as the project progresses.
Takeaways

• Academic partnerships and student internships laid a foundation to explore bioinformatics applications and PHL use
• Bioinformatics needs may outpace the progress of agency resources and policies for use
• Development of justification documents and examples of other state’s approaches provided useful point of reference for engagement of state agencies and service providers
• Cloud environments will best meet the future bioinformatics needs of DCLS, accessing this system in a VITA compliant manner will allow DCLS to build capacity for analysis of more sensitive data in the future
Acknowledgements

DCLS/DGS
• Dr. Denise Toney
• Kevin Libuit
• Ed Scherer
• Troy Westerman
• Erik Clod-Svensson
• Emily Hopkins
• Willie Andrews
• Linoise Edassery
• Pam Stephenson
• Barbara Garnett
• Johnathan Barbett
• Sandra Gill

VITA
• Demetrias Rogers

Dr. Stephen Turner

APHL

State PHL
• Joel Sevinskey (CO)
• Pascal Lapierre (NY)
• Mike Palumbo (NY)
• Sean Wang (MN)
• Kelsey Florek (WI)