Galaxy Cluster to Cloud - Genomics at Scale

Enis Afgan[1,2], Dannon Baker[1], John Chilton[3], Nate Coraor[3], The Galaxy Team, James Taylor[1]
[1]Department of Biology
Johns Hopkins University
Baltimore, MD, USA
{enis.afgan, dannon, jxtx}@jhu.edu
[2]Centre for Informatics and Computing
Rudjer Boskovic Institute (RBI)
Zagreb, Croatia
[3]Department of Biochemistry and Molecular Biology
Penn State University
University Park, PA, USA
{jxc755, ndg1}@psu.edu

ABSTRACT
Fueled by the radically increased capacity to generate data over the past decade, the field of biomedical research has been constrained by the ability to analyze data. Galaxy, an open genomics and biomedical research platform, has been democratizing access to data analysis tools with its effective and accessible web interface. However, the scale of data and the scope of tools required have proven to be a significant challenge for any monolithic deployment of the Galaxy application. We have found that a distributed and federated approach to utilizing compute and storage resources is necessary. This paper describes the ongoing efforts in creating a ubiquitous platform capable of simultaneously utilizing dedicated as well as on-demand cloud resources.

Categories and Subject Descriptors
H.2.4 [**Information Storage and Retrieval**]: Systems and Software – *distributed systems*

General Terms
Performance, Design, Experimentation.

Keywords
Cloud computing, data analysis, genomics, accessibility, federation

1. INTRODUCTION: GALAXY AS A RESOURCE
Modern day genomics has been stereotyped as a big data science - and rightfully so: the cost of DNA sequencing has plummeted in the past decade [1] and the ability to generate data has spread from large research centers and institutions to individual departments and small labs. This ability to rapidly generate data has cultivated significant activity and advancements in medical diagnosis and treatment [2]. The reality, however, is that data alone is not sufficient to obtain desired results because the data must first be analyzed, which is far from a simple task [3].

These circumstances have generated many new potential users for science gateways by targeting the data analysis process - users that are primarily interested in plugging problem-specific parameters into a computer program and observing the output. Yet, the majority of biomedical data analysis software available are command line tools that require a significant level of technical expertise before they can be used (e.g., command line access, tool installation, configuration). Additionally, with the growing size of input datasets, these tools need to be executed on advanced computer infrastructure, ideally, in parallel [4]. This situation presents significant obstacles for fast-paced data analysis that has become a requirement for achieving improvements in the medical diagnosis and treatment.

In response to this changing landscape, we have developed the Galaxy application [5], [6] - a web-based platform for easily interacting with both tools and data. Dubbed a platform for "data intensive biomedical research", Galaxy has been under continuous development for nearly 10 years and focuses on enabling accessible, reproducible and transparent computational science. On the front end it is a web interface that allows users to interact with tools, chain those tools into workflows, share individual data or entire analyses, as well as visualize results in a variety of portable formats. On the back end, it makes it easy for any tool developer to take a standard UNIX command line tool and automatically present a web interface for it. Galaxy can run jobs on a local machine or submit those jobs to a large range of cluster and grid job schedulers. Galaxy very effectively manages both data and tools, providing a simple user interface for experimentalists while still exposing an extensible API to enable downstream developers and afford portability.

As both a public service and a demonstration of the capabilities of the Galaxy platform we provide a free and publicly accessible instance of the application at *usegalaxy.org*, known as Galaxy Main. This instance allows anyone to upload data and utilize a wide array of tools for their data analysis needs. The impact of this service is evident from the number of jobs run and the number of registered users over the past several years (see Figure 1).

While great efforts are made to make Galaxy Main an exceptionally accessible and useful resource, it has some limitations, such as usage quotas and a predetermined toolset. One alternative to using the Galaxy Main portal is to install a local instance of Galaxy. With the basic dependencies being automatically managed, it is a straightforward process to install the application on a local machine[1] or a cluster[2]. Still, there are formidable challenges that must be overcome to provide a complete Galaxy platform due to the reliance on tools and reference datasets, which need to be installed independently on the underlying system. Without these associated tools and data,

9th SC Gateway Computing Environments Workshop '14, Nov 21, 2014, New Orleans, LA, USA.

[1] https://wiki.galaxyproject.org/Admin/GetGalaxy
[2] https://wiki.galaxyproject.org/Admin/Config/Performance/ProductionServer

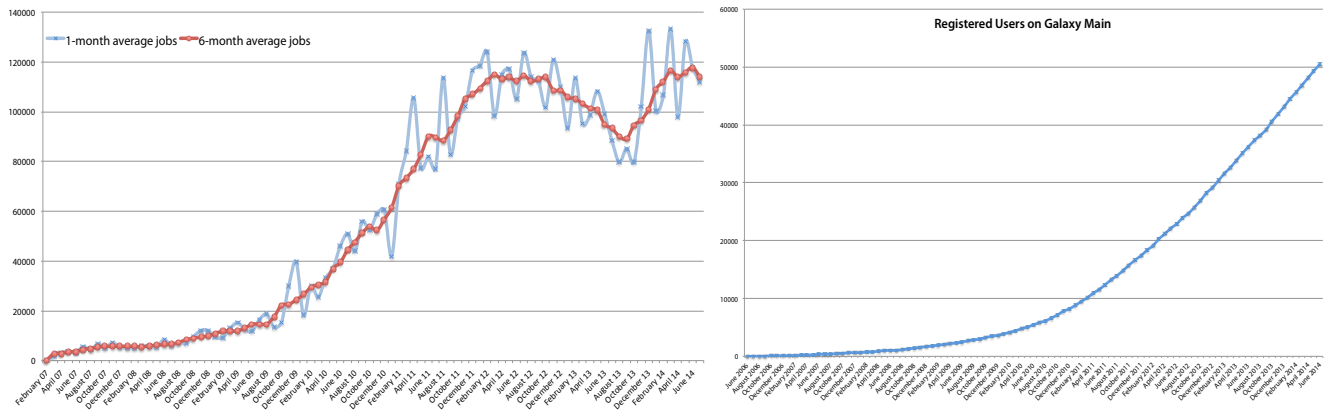

Figure 1. The number of jobs run on the Galaxy Main instance (left) and the number of registered users (right).

the portal is not particularly useful in a research context. Progress is being made to alleviate those obstacles by providing an 'app store' like functionality for installing bioinformatics tools (see the Galaxy ToolShed [7]) and easily populating the platform with desired reference datasets (see Galaxy Data Managers [8]).

With a relatively low computational burden imposed by the Galaxy application itself, the primary requirements for a local Galaxy are the availability of compute and storage hardware requisite to the tasks imposed by the jobs users choose to run, and of course, the resultant datasets. This can impose a not insignificant infrastructural burden and, with many labs not having the upfront funds to purchase computational infrastructure or perhaps having only the occasional need for data analysis, this becomes an undesirable option. In response, we have made the entire Galaxy platform available on various cloud computing resources [9]. This solution allows someone to create a virtual dynamically scalable compute cluster that comes preconfigured with the Galaxy application, a suite of the most common domain-specific tools, and gigabytes of reference genomic data. Each cluster is further customizable and can be shared as a unit [10]. In the theme of Galaxy accessibility we provide a simple full stack management interface, CloudMan [11], which requires only a web browser to use. This solution is our best attempt at a compromise between the accessibility of Galaxy Main and the flexibility of a local install while not requiring any dedicated local computer hardware. Technically, the solution replicates a traditional cluster environment using flexible cloud infrastructure, which simplifies the transition for typical domain tools so they do not require any modification to be able to utilize the flexibility offered by the Cloud.

2. EVOLUTION: THE RESOURCES ARE CONVERGING

The three outlined solutions for utilizing the Galaxy platform for performing data analysis offer a well-rounded set of choices: Galaxy Main offers unprecedented accessibility while local installs offer maximum flexibility and Galaxy on the Cloud fills the void between the two extremes. In addition to the Galaxy Main instance maintained by the Galaxy project there are over 50 additional public Galaxy instances maintained by various groups around the globe. Each instance is configured differently and primarily caters to the requirements of the group maintaining it. One outcome of this is that the domain scientist has the choice of which instance they use and they might be able to find suitable tools for their analysis on any of those public instances rather than having to deploy their own Galaxy. Simultaneously, dedicated Galaxy on the Cloud instances offer exceptional training opportunities, sharing capabilities, and are very good for targeted analyses, small groups, and individual researchers in need of a dedicated and customizable analysis platform. Taking all of these deployment options together, the Galaxy user is presented with a wide range of choices to accommodate their data analysis requirements.

This growing number of choices, however, yields confusion among users and poses new challenges. With each type of analysis or job submission, questions arise regarding the choice of an instance to use. Instead, it would be more favorable for a user to simply work on their data analysis, ideally on any chosen instance, while all the technical details of leveraging all the available choices are automatically addressed.

Technically, a majority of the challenges associated with enabling this flexibility in Galaxy revolve around availability: whether it be resource, data, artifact (e.g., tools, workflows, provenance), or bandwidth. Historically, the Galaxy application has worked around this problem by keeping a locally accessible copy of each of these components (see Figure 2 for a Galaxy platform architecture diagram). The result of this approach was that one works on a particular Galaxy instance while all other Galaxy instances are totally unrelated and disconnected. In order to work on another instance, one has to explicitly export the data from the source instance, import the data on the destination instance and ensure compatible tools and reference data are available, including correct versions, which is a significant administrative endeavor.

In order to bridge this divide, we have been implementing and leveraging a number of features of the Galaxy platform and will describe here an *architectural blueprint* for assembling them into a cohesive unit to deliver a ubiquitous Galaxy capable of efficiently running in a federated fashion (Figure 2):

Data availability is possibly the most difficult challenge due to the size of data being operated on; continuously moving data between instances is simply not a technically feasible solution. Instead, data needs to be decoupled from the application that manages it. To achieve this, the Galaxy application implements the notion of an Object Store - a pluggable file management interface that acts as a layer between Galaxy and any user datasets. Implementing the Object Store interface for various storage mediums (an abstract hierarchical store, Amazon S3, iRODS, and various local disk object stores are currently implemented while other cloud storage options such as Dropbox

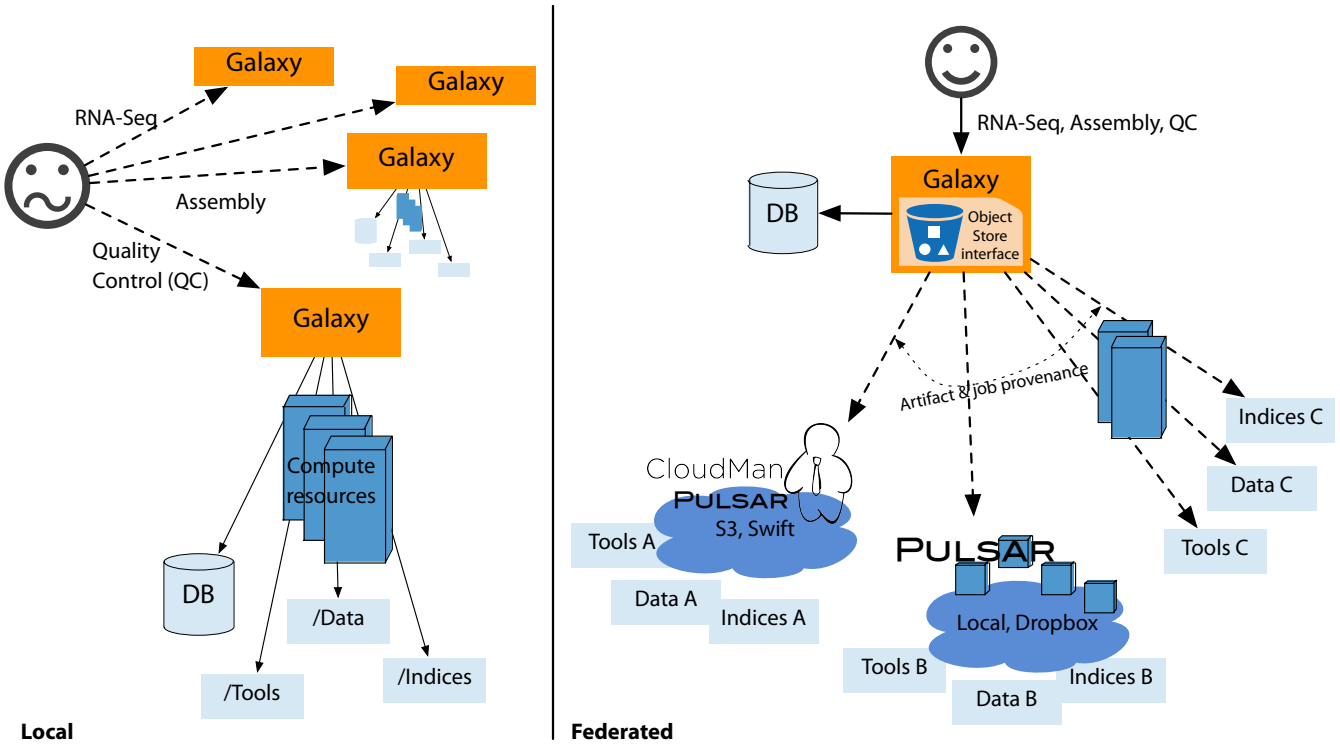

Figure 2. A high-level architecture of the components required to run a Galaxy platform (left: local; right: federated).

are planned) allows datasets to be 'physically' disconnected from a particular instance of Galaxy while the application can still access and interact with them. A user will thus eventually be able to associate self-provisioned external storage resources (S3, etc.) with a Galaxy account and move beyond the imposed quota or limitations on any given instance.

Resource availability is the ability to leverage any and all available computational resources. The Galaxy platform provides the notion of pluggable and dynamically configurable job runners where each job can potentially be submitted to any number of different resources. In combination with CloudMan and the Object Store, this has the potential to leverage external resource availability by allowing users to dynamically acquire additional resources on demand or to leverage data locality to schedule jobs. The ability to execute jobs on a variety of resources alone is not sufficient though - it needs to be possible to automatically interact with those resources and exchange job and provenance information. Galaxy Pulsar[3] is a standalone light-weight server that can be deployed on dedicated or transient cloud resources which also currently provides a method for handling this remote job execution, ensuring the requisite provenance information is exchanged. Long-term, a more integrated solution is very desirable where a Galaxy instance can simply 'advertise' its availability and capacity to the full community of Galaxy instances. This would allow a user to register resources available for use in the Galaxy ecosystem as well as ensure any desired level of data privacy by keeping the data at a specific location and performing the computation on resources local to the data.

Artifact availability is being addressed from multiple directions. Most notably, the Galaxy ToolShed represents a scriptable repository of Galaxy artifacts and enables easy installation of the required ones, including specific tool versions and tool dependencies. Installing tools from the ToolShed requires administrative privileges on the destination Galaxy instance and hence can be seamlessly utilized only on a limited subset of instances one has access to. Utilizing the cloud solution offers (a) a potentially suitable set of pre-installed tools and (b) grants the user the required privileges to install the desired artifacts. Alternatively, allowing users to register and couple their instances to a globally available one, makes it possible to leverage that particular instance (and its toolset) as part of an otherwise more comprehensive resource set. Finally, integration with Docker is currently being explored as a lightweight mechanism for provisioning required artifacts on demand across the described ecosystem.

3. IMPLEMENTATION: A STEPWISE APPROACH

Once combined, the features described above fulfill the basic requirements of the ecosystem required for performing genomics research at scale, yet much work remains to truly integrate these features into a federated job execution platform for Galaxy. As a step in this direction, we are transitioning CloudMan from solely being a dedicated, standalone cloud manager into a cloud management container capable of executing arbitrary Galaxy tasks. CloudMan is already well-versed at managing cloud resources and offers a service-oriented architecture. Configured with flexible Galaxy job running as one of the services, it can be used as an on-demand resource-provisioning engine behind Galaxy's job execution mechanism. As an example, one could, using the API (e.g., via BioBlend [12]), spawn new cloud instances, which would be configured with a Pulsar service. These instances could be linked as job execution destinations to the base

[3] https://pulsar.readthedocs.org/

instance, which would include these new cloud resources into its job resource pool and start utilizing the expanded resource capacity. Finally, integration with the S3 Object Store and advanced scheduling techniques will be implemented that leverage data locality and opportunistic resource utilization.

To achieve this, we are currently abstracting CloudMan's core cloud management features from the functional service features. This will allow for the core framework to be easily and quickly deployed on arbitrary cloud resources (thus facilitating rapid resource provisioning) and provide a layer of cross-cloud interoperability for all implemented functional services. To further reduce the time required to procure an operational system deployed services will leverage pre-built cloud components as required by the implemented services (e.g., file system snapshots; downloadable, hence globally available, file system archives; and hosted file systems for reference data that are leveraged in a read only mode by multiple instances simultaneously). Finally, more complete API coverage for interacting with deployed systems (e.g., service management) will allow for the necessary level of interaction between the initiating Galaxy instance and freshly deployed systems.

4. CONCLUSIONS

Overall, we are evolving an accessible, long-standing, free public gateway from running on a set of dedicated compute resources towards a ubiquitous compute platform capable of accommodating all the requirements and corner cases of a research data analysis. This evolution is being realized in stages driven and enabled by the available technology and innovation. Thus far, we have successfully maintained a dedicated gateway and evolved it to scale to many thousands of users. Simultaneously, we have replicated the available gateway functionality in the form of atomic cloud deployments. Going forward, we are blurring the distinction between the two. We believe that providing a single nexus of Galaxies (i.e., Galaxy Main) that is globally available and linking it to the plethora of storage and compute resource, including other Galaxy instances, that are seamlessly utilized is the simplest way possible to utilize the provided functionality and help solve today's data analysis research challenges.

5. ACKNOWLEDGMENTS

The efforts of the Galaxy Team (Dan Blankenberg, Dave Bouvier, Martin Cech, Dave Clements, Carl Eberhard, Jeremy Goecks, Sam Guerler, Jennifer Jackson and Anton Nekrutenko) were instrumental for making this work happen. This project was supported through grant number HG005542 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG006620 and NSF grant DBI 0543285. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

6. REFERENCES

[1] E. C. Hayden, "Technology: The $1,000 genome," *Nature*, vol. 507, no. 7492, pp. 294–5, Mar. 2014.

[2] A. Katsnelson, "Momentum grows to make 'personalized' medicine more 'precise,'" *Nat. Med.*, vol. 19, no. 3, p. 249, Mar. 2013.

[3] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *J. Big Data*, vol. 1, no. 1, p. 2, Jun. 2014.

[4] M. Baker, "Next-generation sequencing: adjusting to data overload," *Nat. Methods*, vol. 7, no. 7, pp. 495–499, Jul. 2010.

[5] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol.*, vol. 11, no. 8, p. R86, Jan. 2010.

[6] E. Afgan, J. Goecks, D. Baker, N. Coraor, A. Nekrutenko, and J. Taylor, "Galaxy - a Gateway to Tools in e-Science," in *Guide to e-Science*, X. Yang, L. Wang, and W. Jie, Eds. Springer, 2011, pp. 145–177.

[7] D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor, and A. Nekrutenko, "Dissemination of scientific software with Galaxy ToolShed," *Genome Biol.*, vol. 15, no. 2, p. 403, Jan. 2014.

[8] D. Blankenberg, J. E. Johnson, J. Taylor, and A. Nekrutenko, "Wrangling Galaxy's reference data," *Bioinformatics*, vol. 30, no. 13, pp. 1917–9, Jul. 2014.

[9] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor, "Harnessing cloud computing with Galaxy Cloud," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 972–974, Nov. 2011.

[10] E. Afgan, B. Chapman, and J. Taylor, "CloudMan as a platform for tool, data, and analysis distribution," *BMC Bioinformatics*, vol. 13, p. 315, 2012.

[11] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy CloudMan: delivering cloud compute clusters," *BMC Bioinformatics*, vol. 11 Suppl 1, p. S4, 2010.

[12] C. Sloggett, N. Goonasekera, and E. Afgan, "BioBlend: automating pipeline analyses within Galaxy and CloudMan," *Bioinformatics*, vol. 29, no. 13, pp. 1685–6, Jul. 2013.