International Journal Of Public Mental Health And Neurosciences
ISSN: 2394-4668
(An Official publication of Sarvasumana Association)

Execution of Bioinformatic Pipeline Using Galaxy Platform on Cloud

Swapnaja More
Department of Computer Science, Y.M. College,
Bharati Vidyapeeth
Pune. Maharashtra,India
[1]swapnaja86@gmail.com

Ajit D. More
Department of Computer Applications and system Studies,
IMED, Bharati Vidyapeeth
Pune. Maharashtra, India
ajit.more@bharatividyapeeth.ed

Abstract - The emergence of NGS (next-generation sequencing) correlated with the possibility of a tsunami of genome data which would flood storage systems and crush computing clusters in different genome informatics ecosystems. The deployment of compute cluster in Amazon EC2 cloud needs knowledge associated efforts. The solution presented in this paper making it possible for researchers how to deploy bioinformatic pipeline with exactly required computing power along with existing analysis software, to handle the ongoing overflow data.

Index Terms—*Next Generation Sequencing, galaxy, cloudMan, Amazon web services. Elastic Compute cloud (EC2).*

I. Introduction

The coming storm of data forces scientists to find easy and genuine storage and computational methods. According to sequencing instruments growth, configuring and using specialized software to interpret the data necessitates computer hardware, professional technical support, and bioinformaticians(Lathan, Tracey et al. 2002). The researcher's requirements of computing resources fluctuate widely over time. Buying and maintaining a fixed amount of computing resources are very costly for institutional clusters(Harjinder Kaur 2019). Hence, the structured and cost-beneficial option is provided by the cloud for this situation. Since the cloud architecture is elastic and scalable, we can assign resources precisely when needed and dynamically scale them up or down as our needs change over time. Outside providers maintain these compute clouds(Afgan, Krampis et al. 2015). At usegalaxy.org, the accessible public server is provided, known as Use galaxy, which anyone can use as primary. To control the galaxy, such as installing tools, managing users, creating groups, etc(Goecks, Nekrutenko et al. 2010)., Through the user interface user must become an administrator. Admin privileges available to only registered users. Local galaxy can be installed on our local instance of the galaxy from scratch and become an admin(Afgan, Baker et al. 2016). CloudMan (usecloudman.org) allows researchers to handily deploy, customize, and share their entire cloud analysis environment, including data, tools, and configure//////////tions(Afgan, Baker et al. 2010). This platform improves accessibility of cloud resources, tools, and contributes toward reproducibility and transparency of research solutions.

II. Galaxy Installation

The Cloud-based bioinformatics framework integrates all the previously listed methods. It provides an overall solution in which the Galaxy framework can be deployed and configerd on clouds, auto-scale cloud resources, user-specific tools can be tailored, high-performance data transfer capacities, and a semantic verification mechanism is available. The platform eliminates the significant previously existing usage barriers. For resource consumption, the Amazon EC2 pay-as-you-go billing model offers a scalable and elastic sequence analysis environment. In order to validate the efficiency, a performance assessment is provided for metagenomics in bioinformatics(Goecks, Nekrutenko et al. 2010).

III. Launching Genomic Virtual Laboratory Platform On AWS Cloud

In order to sign up to AWS, you need a credit card, an e-mail account, finally decide an AWS user name and a password. This picks, later on, to log into your Galaxy server. (Fig. 1a &b)

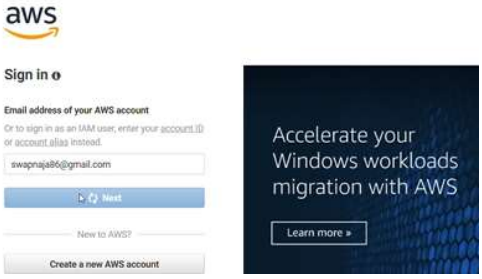
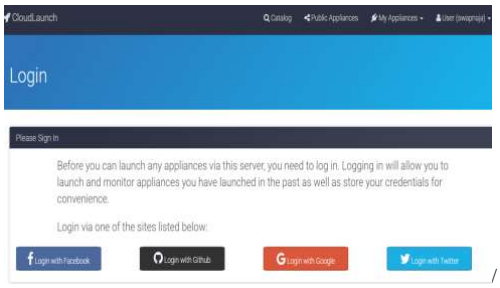


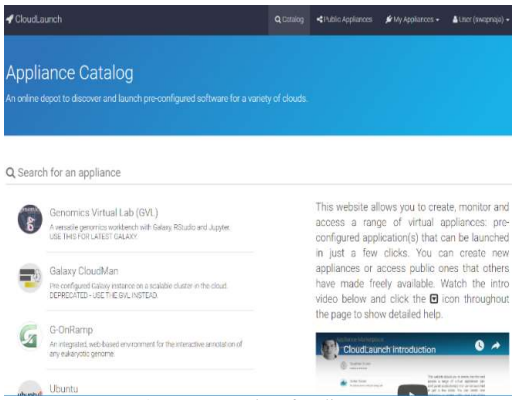
Fig.1. a) Signing page for Amazon Web Services.



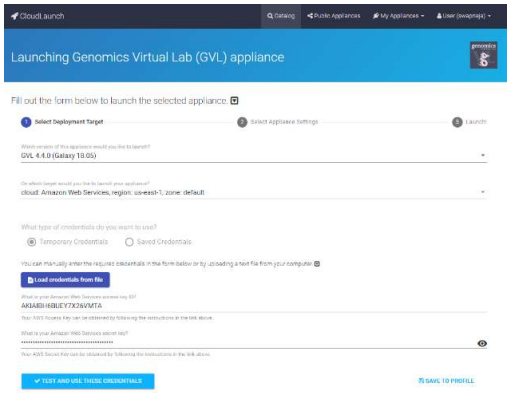
b) Cloud Launch login to monitor appliances.

Download the Security Credentials and Access Credentials section, which lists an active account and an Access Key ID, and use Secret Access Key. (Fig.2 c &d)

A. Starting with Cloudman



c) Browse a catalog of appliances



d) Launch of chosen application with access credential.

Fig. 2. Functionality available through cloudlaunch

The selected unit can be started by selecting and launching properties from a variety of cloud providers. The user interfaces to launch the GVL framework give the GVL dashboard appliance-specific start-up parameters for primary virtual machines and instance services.

The Genomic Virtual Laboratory (GVL) is a complete data analysis platform that comes as a complete package with various pre-installed and configured software (Galaxy, RStudio, Jupyter) with cloud compute infrastructure(Afgan, Krampis et al. 2015). Every instance of the GVL can be dynamically scaled by CloudMan and customized with additional tools. In the GVL Galaxy instance, inbuilt overall 200 tools are present. GVL will enable researchers to automatically access tools and platforms for genomic analyses as a working workbench to improve the solution's replicability and usability(Afgan, Sloggett et al. 2015).

Genome study is concerned with analyzing and interpreting enormous experimental data available nowadays through public genomic in different complex workflows. New algorithm and tool development processes are rapidly applied to keep up with new 'omic' technologies(Berger, Peng et al.). Many visualization options are available for searching experimental data and public genomic catalogs such as GBrowse, IGV, UCSC Genome Browser. Galaxy, Genpattern, Yabi, Mobyle, and Chipster are different analysis platforms provide biologists with skill in programming to design analysis workflows and release on High Throughput Computing (HTC) clusters(Le, Tran et al. 2018).

However, the fact is that the required tools, plan of actions, data services and required platforms for genomics studies are complicated to install and also customization requires considerable computational and storage resources. It usually involves a regular maintenance to keep the software, data and hardware up to date. It is also the case that a single workflow platform is not often sufficient for all the steps of a real-world analysis(Afgan, Krampis et al. 2015). This is because analyses often involve visualization and evaluation of processing steps, requiring a combination of various analysis, data-wrangling and visualization tools to carry out an end-to-end analysis.

B. Executing metagenomic data analysis workflow with galaxy cloudman

The most common Cloud tools in public clouds are Amazon Web Services (AWS), HP Cloud, and Google Compute Engine, where users only pay for the resources they use. As a result, several cloud management software applications have been created to serve a functional purpose in bioinformatics by customizing cloud resources. Besides, cloud-aware frameworks, platforms, and virtual laboratories have been built to combine several applications' functionality. Galaxy on the Cloud is a cloud-based Galaxy framework with pre-configured settings. The emergence of research-oriented cloud computing has opened the door to developing bioinformatics analysis support on these widely available national infrastructure facilities and public clouds(Hu, Weng et al. 2018; Hiltemann, Boers et al. 2019).

Galaxy on cloud, such as Amazon EC2 (Elastic Compute) Cloud, offers a range of advantages, including a better user experience, usage-based pricing, on-demand resource configuration, increased resource efficiency, and faster processing(Afgan, Baker et al. 2021). The GVL machine image is used to create cloud-launchable instances, quickly launched and set up using a launcher web application (Fig. 3).

C. Launch GVL4.4.0(Galaxy18.05)

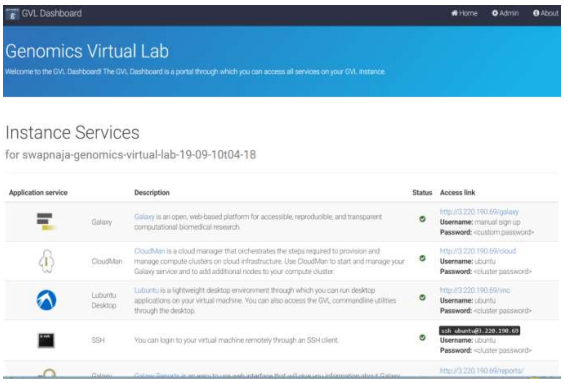


Fig. 3. Launch GVL4.4.0(Galaxy18.05) appliance among various cloud providers and launch properties.

For every launched instance following services run one by one

- The GVL Dashboard is an all-in-one place to access all resources and current status. Sometimes it has been the default page for certain self-launched GVL cases(Fig. 4.3).
- Turning each of the GVL in the cloud to a virtual cluster and managing the cloud resources, CloudMan scale it by incorporating worker node.
- Galaxy is a web-enabled standard web analysis tool that can deploy jobs around the cluster on the cloud. Researchers can adjust the galaxy through Galaxy Toolshed and Galaxy Data Managers (Bedoya-Reina, Ratan et al. 2013).
- RStudio Server and IPython Notebookare web-based platforms for performing statistical analysis using the R programming language and algorithmic analysis on Python code.
- Remote Desktop is a web-based remote access application that is part of the Linux OS(Krampis, Booth et al. 2012).

Linux provides an environment with command-line access with pre-installed reference data to the device and all necessary tools in the Galaxy environment with complete administrative and sash control. The GVL has been introduced and is currently accessible on the Australian Research Cloud and Amazon Web Services (Sydney region). The field of genomics has emerged as one of the challenging areas that require an intricate ecosystem. It has various technologies including instruments, computers, and data to support multi-step pipelines technologies, computers, and data—all sharpened to assist multi-step pipelines utilizing numerous tools and handling data in gigabytes(Krampis, Booth et al. 2012). The current status of the deployed applications is shown on a dashboard of launched appliances (Fig. 4).

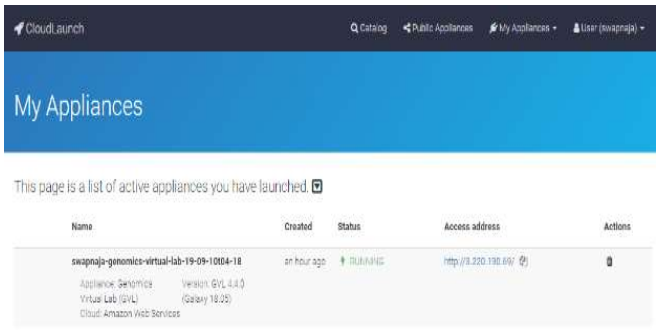


Fig. 4. Active appliances for GVL cloud Amazon Web Services

The fact of processing large amounts of genomic data necessitates a robust data analysis workbench. As a result, the data analysis process involves a high degree of maintenance expenses and technological skills, limiting the entry researchers' focus on biology.

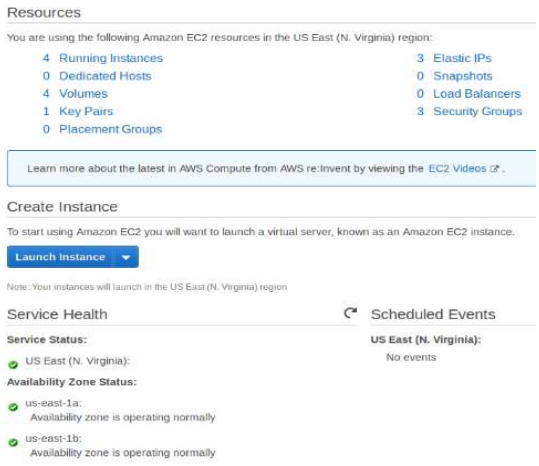


Fig.5. Amazon EC2 Resources in the US East (N.Verginia)region.

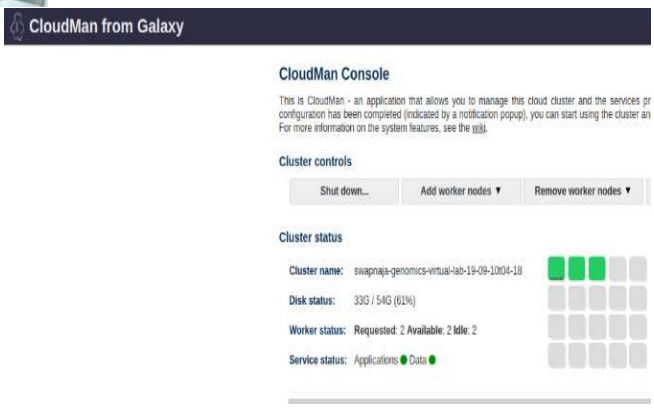


Fig. 6. Cloud cluster and services provided by CloudMan from Galaxy

Through the CloudMan web interface, one can scale the size of the cloud cluster at runtime by adding or removing worker instances comprising the cluster. As the use of a given cluster expands, users may consume the space associated with the given cluster (Fig 4.5 and 4.6).

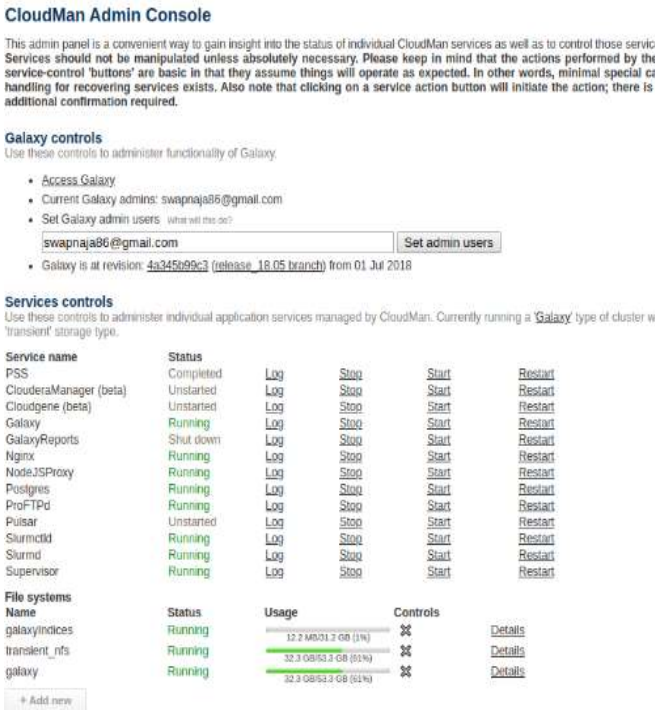


Fig. 4.7 CloudMan Admin Console and Service controls to administer individual application services.

The changes supported at this level of instance customization include modifications to the file systems managed by CloudMan. The available file systems are listed on the CloudMan Admin console. Modifying contents of these file systems allows you to customize your instance of Galaxy, install or modify tools, as well as modify reference genomes used by Galaxy tools(YeeLow 2019). A change made holds the ownership of the directories (Fig 4.7)

IV. Conclusion

As NGS becomes an essential tool to provide solution for analysis of sequence, it is an important key factor which implemented by biomedical researchers. Galaxy Cloud marks this by integrating the available Galaxy interface with automated management of cloud computing resources. Galaxy allows users to implement existing workflows which are tested and proven with test data and also allows constructing their own analyses for novel tasks. Galaxy Cloud instance are granted and managed wholly by the user who created them, and can be used capably in secure private clouds but it is cost effective. Thus Galaxy Cloud provides a solution that conserves user control and privacy, makes complex analysis accessible, and allows the use of practically unlimited as required compute resources.

Acknowledgment

This work has been supported by Department of Computer Science, Y.M. College, Pune and Department of Computer Applications and system Studies, IMED, Bharati Vidyapeeth, Pune for strong intramural support for experimental, computational work and for the critical reading of this manuscript.

References

[1] Afgan, E., D. Baker, et al. (2010). Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics, BioMed Central.

[2] Afgan, E., D. Baker, et al. (2021). "Galaxy CloudMan: d/elivering cloud compute clusters." BMC Bioinformatics **11 Suppl 12**: S4.

[3] Afgan, E., D. Baker, et al. (2016). "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update." Nucleic Acids Research **44**(W1): W3-W10.

[4] Afgan, E., K. Krampis, et al. (2015). Building and provisioning bioinformatics environments on public and private Clouds. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

[5] Afgan, E., C. Sloggett, et al. (2015). "Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud." PloS one **10**(10): e0140829.

[6] Berger, B., J. Peng, et al. "Computational solutions for omics data." Nat Rev Genet. 2013 May;14(5):333-46. doi: 10.1038/nrg3433.

[7] Goecks, J., A. Nekrutenko, et al. (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol. 2010;11(8):R86. doi: 10.1186/gb-2010-11-8-r86. Epub 2010 Aug 25.

[8] Harjinder Kaur , M. S. G. (2019). "Role of Big Data in Cloud Computing: A Review." IJERT Volume 08, Issue 07**.**

[9] Hiltemann, S. D., S. A. Boers, et al. (2019). "Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur." GigaScience **8**(2).

[10] Hu, Z., X. Weng, et al. (2018). "Metagenomic next-generation sequencing as a diagnostic tool for toxoplasmic encephalitis." Annals of Clinical Microbiology and Antimicrobials **17**(1): 45.

[11] Krampis, K., T. Booth, et al. (2012). "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community." BMC Bioinformatics. 2012 Mar 19;13:42. doi: 10.1186/1471-2105-13-42.

[12] Lathan, C. E., M. R. Tracey, et al. (2002). "Using virtual environments as training simulators: Measuring transfer." Handbook of virtual environments: Design, implementation, and applications: 403-414.

[13] Le, V.-V., H. Tran, et al. (2018). "Taxonomic assignment for large-scale metagenomic data on high-perfomance systems." Journal of Computer Science and Cybernetics **33**: 119.

[14] YeeLow, C. B. L. Y. (2019). Pipeline of High Throughput Sequencing. Encyclopedia of Bioinformatics and Computational Biology. Oxford, Academic Press**:** 144-151.