

# Execution of Bioinformatic Pipeline Using Galaxy Platform on Cloud

Swapnaja More Department of Computer Science, Y.M. College, Bharati Vidyapeeth Pune. Maharashtra,India <sup>1</sup>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

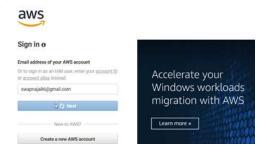
The coming storm of data forces scientists to find easy and genuine storage and computational methods. According to sequencing instruments growth, configuring and using specialized software to interpret the data necessitates computer hardware. professional technical support, and bioinformaticians(Lathan, 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 provided assessment is metagenomics for 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 email account, finally decide an AWS user name and a password. This picks, later on, to log into your Galaxy server. (Fig. 1a &b)



# © IJPMN, Volume 8, Issue 3, December-2021



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

Login				
Please Sign In				
Before you can launch any appliances via launch and monitor appliances you have la convenience.				
Login via one of the sites listed below:				
flage with Friedbook	Ging	in with Google	<b>V</b> Lope	with Twitter

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



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

# International Journal Of Public Mental Health And Neurosciences ISSN: 2394-4668

# (An Official publication of Sarvasumana Association)

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, cloudaware 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).

# © IJPMN, Volume 8, Issue 3, December-2021



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)* 

g GVL Dashboard	i			<ul> <li>Home</li> </ul>	O Admin	0 Abo
Genomics Hoome to the GVI. Deal		al Lab Disabboard is a portal through which you can access all services on your CVI. Instance.				
nstance S or swapnaja-g	5011100	2S virtual-lab-19-09-10t04-18				
Application service		Description	Status	Access link		
Ξ.	Galaxy	Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research	0	http://3.220.19 Username: ma Password: <cu< td=""><td>qu ngie laun</td><td>d&gt;</td></cu<>	qu ngie laun	d>
45	CloudMan	CloudMan is a cloud manager that orchestrates the steps required to provision and manage compute clusters on cloud infrastructure. Use CloudMan to start and manage your Galaxy service and to add additional nodes to your compute cluster.	0	http://3.220.19 Username: ub/ Pasaword: 4ch	ntu	5
	Luburdu Desktop	Luburitu is a lightweight desktop environment through which you can run desktop applications on your virtual machine. You can also access the GVL commandine utilities through the desktop.	0	http://3.220.19 Username: ub/ Password: -ick	intu	6
	SSH	You can login to your virtual machine remotely through an SSH slient.	0	ssh obuntu@). Username: ub: Password: </td <td>ntu</td> <td>5</td>	ntu	5

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

# International Journal Of Public Mental Health And Neurosciences ISSN: 2394-4668

# (An Official publication of Sarvasumana Association)

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).

✓CloudLaunch		<b>Q</b> Catalog	<public appliances<="" th=""><th>🖋 Wy Applances 🔸</th><th>≜liser (swapnaja) +</th></public>	🖋 Wy Applances 🔸	≜liser (swapnaja) +
My Appliances					
This page is a list of active appliances you have la	aunched. 🖸				
2 ( ) 이렇게 많다. 것이 같은 것은 이가 없을 것 같은 것을 가지 않는 것이 있다.					
Name	Created	Status	Access ad	dress	Actions

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.

You are u	sing the following Amazon EC2 resources in the US	East (N. V	irginia) regior	1:
4	Running Instances		3	Elastic IPs
0	Dedicated Hosts		0	Snapshots
4	Volumes		0	Load Balancers
1	Key Pairs		3	Security Groups
0	Placement Groups			
Lean	n more about the latest in AWS Compute from AWS r	e:Invent b	y viewing the	EC2 Videos 🕼 .
To start u Launch	Instance sing Amazon EC2 you will want to launch a virtual so Instance vill launch in the US East (N. Virginia) region	erver, know	n as an Ama	zon EC2 instance.
To start u Launci Note: Your	sing Amazon EC2 you will want to launch a virtual se Instance	erver, know		zon EC2 instance. ed Events
To start u Launch Note: Your Service	sing Amazon EC2 you will want to launch a virtual so Instance virtual so instances will launch in the US East (N. Virginia) region be Health			ed Events
To start u Laund Note Your Service S	sing Amazon EC2 you will want to launch a virtual so Instance virtual so instances will launch in the US East (N. Virginia) region be Health		Schedul	ed Events . Virginia):
To start u Launch Note: Your Service Service S US El	sing Amazon EC2 you will want to launch a virtual so Instance virtual so instances will launch in the US East (N. Virginia) region b Health Status:		Schedul US East (N	ed Events . Virginia):
To start u Launch Note: Your Service S O US Ei Availabil	sing Amazon EC2 you will want to launch a virtual se t Instance v instances will kunch in the US East (N. Virginia) region b Health istatus: sst (N. Virginia): ity Zone Status:		Schedul US East (N	ed Events . Virginia):

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

# © IJPMN, Volume 8, Issue 3, December-2021



#### CloudMan Console

This is CloudMan - an application that allows you to manage this cloud cluster and the services pr configuration has been completed (indicated by a notification popup), you can start using the cluster an For more information on the system features, see the <u>wis</u>).

Shut do	wn	Add worker nodes *	Remove worker nodes *
luster status			
Cluster name:	swapnaja-ge	nomics-virtual-lab-19-09-10t04-	18
Disk status:	33G / 54G (6	196)	
Worker status:	Requested:	2 Available: 2 Idle: 2	
	Applications	Data	

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).

### **CloudMan Admin Console**

This admin panel is a convenient way to gain insight into the status of individual CloudMan services as well as to control those servic. Services should not be manipulated unless absolutely necessary. Please keep in mind that the actions performed by the service-control 'buttone' are basic in that they assume things will operate as expected. In other words, minimal special or handling for recovering services exists. Also note that clicking on a service action button will initiate the action; there is additional confirmation required.

#### Galaxy controls

Use these controls to administer functionality of Galaxy.

- Access Galaxy
- Current Galaxy admins: swapnaja86@gmail.com

Status

- Set Galaxy admin users what will the on?
- swapnaja86@gmail.com Set admin users
  Gataxy is at revision: 4a345099c3 (release 18.05 branch) from 01 Jul 2018
- Services controls

Use these controls to administer individual application services managed by CloudMan. Currently running a Galaxy type of cluster w transient' storage type.

#### Service name

PSS	Completed	Log	Stop	Start	Restart
ClouderaManager (beta)	Unstarted	Log	Stop	Start	Restart
Cloudgene (beta)	Unstarted	Log	Stop	Start	Restart
Galaxy	Running	Log	Stop	Start	Restart
GalaxyReports	Shut down	Log	Stop	Start	Restart
Nginx	Running	Log	Stop	Start	Restart
NodeJSPraxy	Running	Log	Stop	Start	Restart
Postgres	Running	Log	Stop	Start	Restart
ProFTPd	Running	Log	Stop	Start	Restart
Pulsar	Unstarted	Log	Stop	Start	Restart
Slurmetid	Running	Log	Stop	Start	Restart
Slurmd	Running	Log	Stop	Start	Restart
Supervisor	Running	Log	Step	Start	Restart
File systems					
Name	Status	Usage		Controls	
galaxyIndices	Running	12.2.447	VOL Z GB (1%)	- 22	Details
transient_nfs	Running	_	153.3-G8 (61%)	- 26	Details
galaxy	Running		53.3 GB (61%)	- 22	Details
			100000		

+ Add new

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). <u>Galaxy CloudMan: delivering</u> <u>cloud compute clusters</u>. BMC Bioinformatics, BioMed Central.

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

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

[4] Afgan, E., K. Krampis, et al. (2015). <u>Building and provisioning bioinformatics environments on public and private Clouds</u>. 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." <u>PloS one 10(10)</u>: e0140829.

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

## © IJPMN, Volume 8, Issue 3, December-2021



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

[8] Harjinder Kaur , M. S. G. (2019). "Role of Big Data in Cloud Computing: A Review." <u>IJERT</u> 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." <u>GigaScience</u> **8**(2).

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

[11] Krampis, K., T. Booth, et al. (2012). "Cloud BioLinux: preconfigured and on-demand bioinformatics computing for the genomics community." <u>BMC Bioinformatics. 2012 Mar 19;13:42.</u> 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." <u>Handbook of virtual environments: Design, implementation, and applications</u>: 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. <u>Encyclopedia of Bioinformatics and Computational</u> <u>Biology</u>. Oxford, Academic Press: 144-151.

# © IJPMN, Volume 8, Issue 3, December-2021