Modeling the Costs of Cloud Computing Tools via Galaxy for Anvil

Peiyuan Xu [1][2], Enis Afgan [3], Michael Schatz [2][3]

[1] Department of Applied Mathematics and Statistics, [2] Department of Computer Science, [3] Department of Biology

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Abstract

Cost is one of the largest barriers towards migrating biomedical related analysis in the cloud. Researchers have limited information about the expected costs of running analysis tools, which in turn makes adopting cloud solutions very difficult. The open-source Galaxy for Anvil project (galaxyproject.org) works to better understand the resource requirements and estimate job cloud costs. The project leverages real-world data from usegalaxy.org that contains information about historic job information for hundreds of thousands of jobs and historical tool data for thousands of Galaxy tools. For this project, we specifically mined the data for the resource requirements and usage data of the ~10 most popular tools and will study how changing parameters influence those requirements. This project will in turn yield impactful insights for a large number of researchers interested in adopting cloud solutions.

Background of My Work

During Summer 2021, I worked two months in the lab of Bloomberg Distinguished Professor Michael Schatz on the Galaxy for Anvil research project. I collaborated with Professor Michael Schatz, Dr. Enis Afgan, as well as Dannon Baker, Keith Suderman, Dave Clements, Victor Wen, and Bridget Carr. My work is presented at the July 2021 ISMB workshop led by Professor Michael Schatz. It is also accepted towards the November 2021 Genomic Informatics Conference in Cold Spring Harbors, Maryland as well as towards February 2022 Advanced Genomic Biotechnology (AGBT) conference. In addition, my work has been presented at the August 2021 Galaxy for Anvil roundtable discussion, April 2022 Mini-Science Gateway, as well as during JHU weekly lab meetings. I am currently still actively involved in this project and intends to make further contributions in the months ahead.

Methods and Materials

This Galaxy for Anvil project consists of three phases:
1) Mining existing tool data
2) Systematic benchmarking of selected tools
3) Dissemination of results via a lookup table or API

I mainly worked on Phase I in which I mine existing tool data from large-scale datasets on usegalaxy.org that contain historical information about the job costs as well as resource consumption and usage data of the 1000+ Galaxy tools over the past couple of years.

For Phase I, I wrote several Python scripts and SQL queries for information retrieval (see here). These queries extract tool properties which includes:

1) Total and average CPU time
2) Total and average memory
3) Number of jobs
4) Number of users

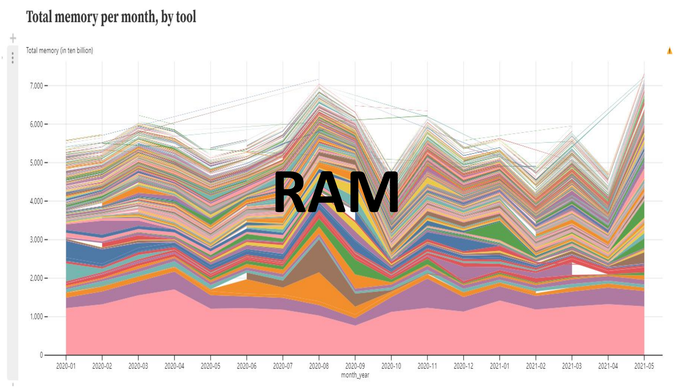
I analyzed these metrics from both the May 2021 timespan as well as the entire January 2020 to May 2021 timespan for the 1000+ Galaxy tools.

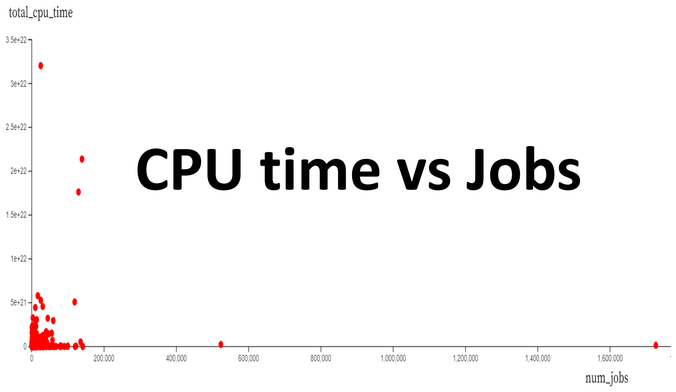
From there I identified the 10 most popular tools in terms of the average number of users across all months. For these ten tools, I created plots showing the resource requirements of these tools (see example plot in Results)

Results

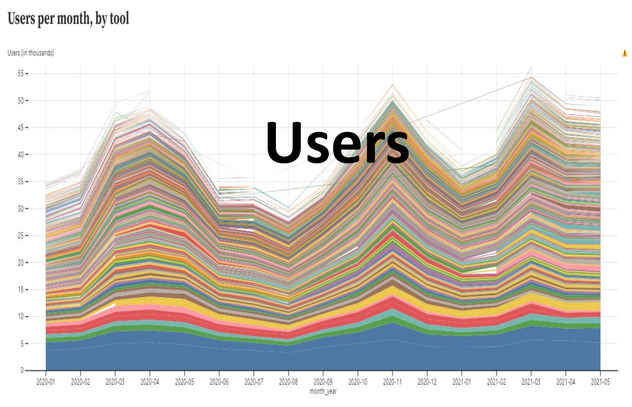
The results about the resource requirements and usage data about the Galaxy tools are mostly published in the Cloud Costs interactive Observable HQ notebook located here and co-authored with Dannon Baker. Below are some examples of visualizations of the research findings and benchmarking results in the form of stacked area charts and zoomable scatterplots.

1) Resource Consumption





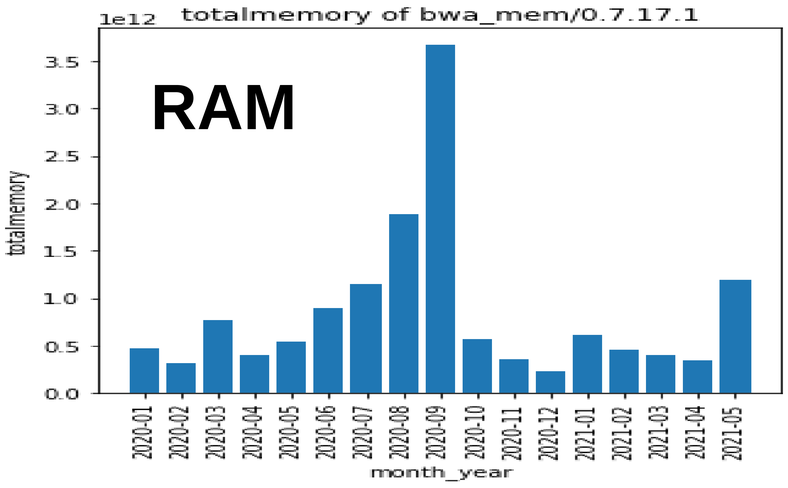
2) Usage



I then extracted the ten most popular tools in terms of the average number of users across all months and created several visualization of the resource consumptions (e.g., CPU time and memory) and usage data (e.g., num jobs) of each of the ten tools separately.

The interactive HQ Observable dashboard of the resource requirements of each of the ten most popular tools from usegalaxy.org in terms of the job data is still in progress.

Below, we have one of the ~10 most popular tools is bwa-mem and the plots of the total memory of that tool from January 2020 to May 2021 is shown below. However, this is a static plot, and I am currently in the process of making interactive Observable plot of this and of many others.



Conclusion

So far, I have mined the data of the Galaxy tools and jobs from January 2020 to May 2021 from usegalaxy.org to extract useful tool properties such as CPU time and memory, as well as usage information like jobs and users. I used the average number of users across each of the 1000+ Galaxy tools to extract the ~10 most popular tools. My work and the interactive visualizations published in Observable notebooks will yield impactful insights for the scientific community and software developers interested in adopting cloud solutions. In the future, I plan to contribute to the benchmarking efforts of Phase II as well as finish creating the interactive HQ dashboard of the ~10 most popular tools.

Acknowledgement

This project will not have been possible without the help of my mentors: Dr. Enis Afgan and Professor Michael Schatz. They have provided me with valuable advice and guidance throughout the project.

Special thanks to the research scientists and to my peers in this project for helping out and supporting me. Thanks to my collaborators Dannon Baker, Keith Suderman, Dave Clements, Victor Wen, and Bridget Carr.