Elastic cloud computing for microbial genomics analytics Part I: Concepts, back-end using Nectar and ownCloud

Hawkesbury Institute for the Environment

Kaitao Lai[1] ‡, Alexie Papanicolaou[1] and Thomas Jeffries[1]
[1]Soil Biology and Genomics Theme, Hawkesbury Institute for the Environment, Western Sydney University, Sydney, Australia‡Kaitao.Lai@westernsydney.edu.au

Abstract

The ever-increasing demand for high-throughput metagenomic data analysis has led to an unsustainable demand for computational training and resources. Elastic computing (EC) is a service oriented computing mode that computing resources can be scaled up and down easily by the cloud service provider, such as Nectar, and Amazon Web Services (AWS). We provide a visual web application platform with a Web 2.0 user interface and an automatic scaling analysing service to support wet-lab staffs for metagenomic data analysis. This platform can request and startup virtual machine resource after that researchers submit metagenomic data for analysing by using automatic scaling analysing service.

This platform has been designed to be an easy-to-use web application framework for the organisation and analysis of high-throughput metagenomic data, with a focus on bacterial 16S rRNA sequencing datasets and fungal ITS sequence datasets. It integrates QIIME (Quantitative Insights Into Microbial Ecology), a set of bioinformatics software packages and scripts, and an automatic scaling analysing service, with interactive visualisations (ownCloud web interface) to enable researchers to receive results from demultiplexing and quality filtering, OUT picking, taxonomic assignment, and diversity analysis with submitted sequencing data. These results can be used to do metagnomic analysis for microbial communities.

System architecture and workflows

This system consists of personal cloud storage ownCloud with ownCloud application CRATEIT, a plugin for data publishing, as front-end components, cloud storage and VM instance, computing resource, in Nectar as back-end components. This diagram below illustrates the workflows after the user submit job and request analysis services. Raw sequences data from HIE biologist users are uploaded through CRATEIT application in ownCloud to cloud storage in Nectar using cloud storage software SWIFT. The work service then starts up VM instance by computing resource software nova command, and then VM instance runs QIIME commands from automatic workflow to process and analysis the data. Once the analysis work completed, the compressed output file is stored in object storage. Finally, front-end side downloads the output file and available for user to download to local machine from ownCloud.

Project benefits and perspective

The ability to produce large size of results using QIIME with automatic scaling anlysing function provides a cost effective method for wet-lab staffs for metagenomic data analysis without the need of Linux training. Web-lab staffs could spend less time on metagenomic data analysis. Future directions are centered around visual analytics and interaction that will facilitate semi-automated analysis.

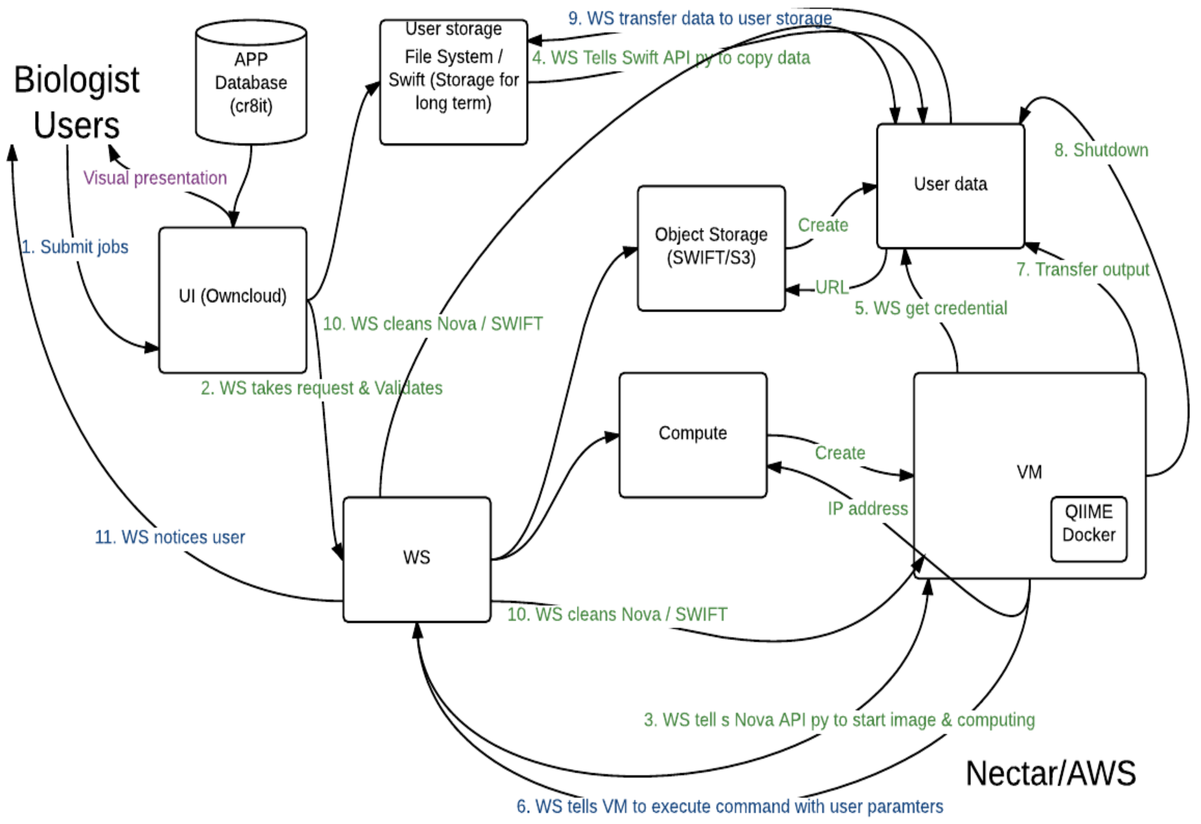

Figure 1 – Infrastructure for Biowebapp

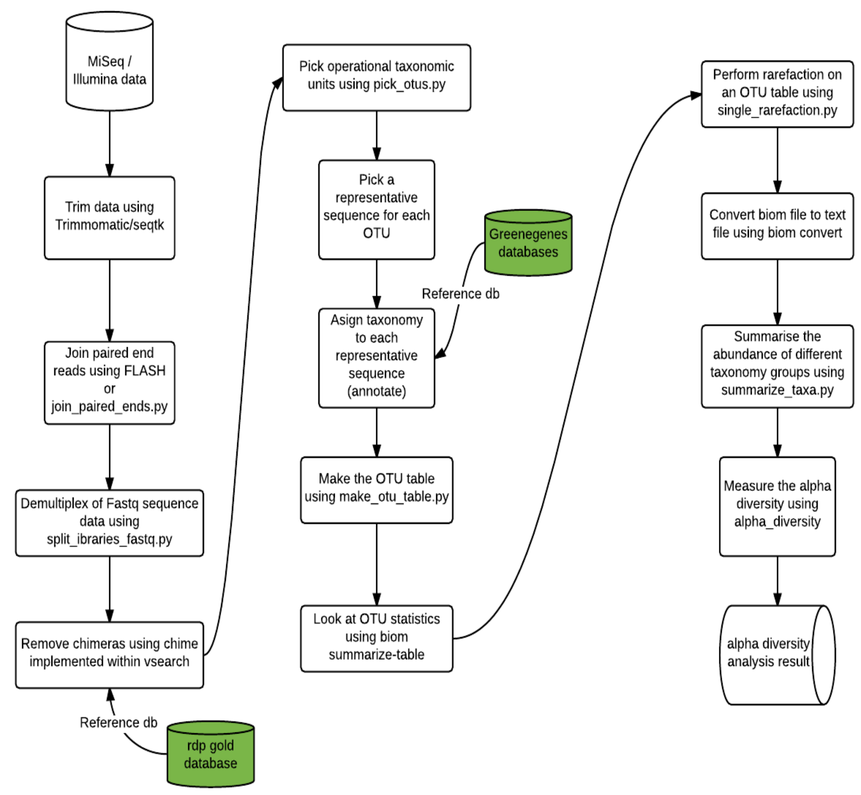

Figure 2 – Data analysis workflow for bacterial 16s rRNA sequencing data analysis

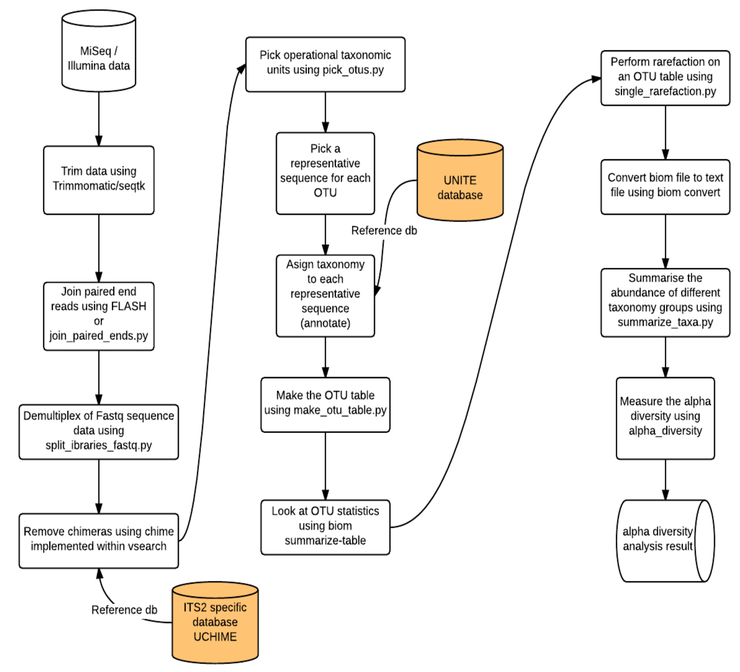

Figure 3 – Data analysis workflow for fungal ITS2 sequence data analysis