The Journal of Open Source Software

hotsub: A batch job engine for cloud services with ETL framework

Hiromu Ochiai[1]

1 National Cancer Center Japan

DOI: 10.21105/joss.01069

Software

- Review
- Repository
- Archive

Submitted: 04 November 2018
Published: 12 November 2018

Summary

Because of the rapid accumulation of biomedical data all over the world, developing a platform for analzing them using high-performance computational infrastructure has become increasingly important in many biological and medical fields. Nowadays, cloud computing is getting a lot of attention since they can promote the sharing of data and reproducible analytical workflows across institutions. On the other hand, there has yet been no decisive practice on how to set up analytical workflows in cloud computing resources.

One possible approach is what we call on-demand Extraction Transformation Load (ETL) framework. The overview of this framework is as follows:

1. Each input file (e.g., FASTQ file) is first extracted from a storage area (e.g., Amazon Simple Storage Service) to each virtual machine within a computing area (e.g. Amazon Elastic Compute Cloud (Amazon EC2))
2. Each input file is transformed into an output file (e.g., FASTQ to BAM conversion).
3. Each generated output file is loaded to the storage area stopping and deleting each VMs.

There are several frameworks which realize on-demand ETL framework in the cloud computing environment, which is often provided by cloud computing vendors (AWS Batch by Amazon Web Service or Azure Batch by Microsoft) or third parties (dsub by Google Genomics). However, with these current ETL implementations, since commonly used data across VMs (e.g., reference genomes) is downloaded to individual VMs, we need to pay particular attention to the excessive load of network and storage, and deployment and transferring of data according to cost charging policy of each provider.

Here we propose a novel framework, on-demand Extended Extraction Transform Load (ExETL), in which commonly necessary data is first loaded to a pre-built shared data instance, mounted by each computing VM and used across VMs. We demonstrate that this framework reduces the payment cost a lot in several cases. Besides, we would like to argue that sharing of analytical workflows will be enhanced since users can safely try the workflows without special caution of the location of the associated database.

We have developed a software implementing the proposed ExETL framework, hotsub (https://github.com/otiai10/hotsub).

Since 'hotsub' uses Docker and Docker Machine, users of 'hotsub' don't have to care about acquiring VMs on cloud services nor setting up environment for computing. Handling infrastructures and runtimes are automated by 'hotsub'.

Statement of need

Even just for basic ETL framework provided by cloud services, it's necessary to configure the managed services on web-console of each cloud service. By using `hotsub`, on the other hand, users don't have to configure VMs on web-console.

In addition, `hotsub` suggests and implements **ExTL** framework, which solves potential problems simple ETL frameworks by AWS Batch, ECS, and dsub have. By using simple ETL framework for bio-informatics, downloading huge reference genome on **each computing instance** could be inefficiency of network traffic and instance time.

If your resources are located on Google Cloud Storage, you can just use `--provider` option to change which platform your computing resources will be launched on, with the same command line interface of `hotsub`. It helps the ecosystem of sharing workflows with someone using different cloud services.

References

- AWS Batch — Easy and Efficient Batch Computing Capabilities - AWS
- Batch - Compute job scheduling service | Microsoft Azure
- DataBiosphere/dsub - Open-source command-line tool to run batch computing tasks and workflows on backend services such as Google Cloud.
- Enterprise Container Platform | Docker
- Docker Machine | Docker Documentation