CloudyCluster

# Case Study: CloudyCluster on AWS for Topic Modeling

## TOPIC MODELING IN THE DICE LAB

With the rapid increase of data analysis complexity and the availability of data libraries, more researchers are finding it difficult to get the time and resources required to perform their research in traditional data centers. CloudyCluster and the Amazon Web Services (AWS) Marketplace offer a simple solution for those who need compute and storage on demand. Topic Modeling is one area of research that lends itself well to the CloudyCluster model, confirmed by Dr. Amy Apon, Professor and Chair of the Division of Computer Science in the School of Computing at Clemson University.

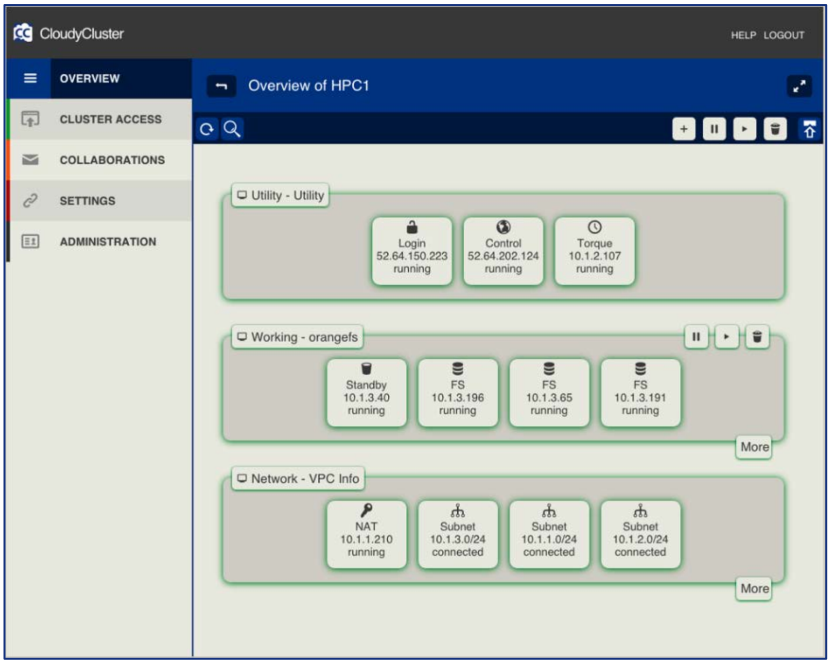
The Data Intensive Computing Ecosystems (DICE) lab at Clemson University is utilizing CloudyCluster for scientific computations in the area of scalable machine learning, and in particular, in topic modeling. Topic modeling is a method for understanding the content of a huge set (corpus) of text documents, to find overall themes and to make sense of the information in the set of documents. For example, topic modeling can be used to find the themes in a large set of news articles or social media posts. More than just keywords, topic models consider the distribution of topics and themes across sets of documents, and can also be used to model how documents change over time.

Topic modeling, which is extremely compute-intensive, is based on Latent Dirichlet Allocation (LDA), published by Blei, Ng, and Jordan in 2003. This method has been implemented many times. One recent implementation by Google, PLDA+, uses message passing in a distributed cluster environment to speed up the calculation of topics in a very large corpus.

## READY WHEN YOU ARE

Although the university has a large data center for academic research, the process of topic modeling requires large amounts of storage space and long-running models. Apon's team found it difficult to wait for required storage and compute resources, and CloudyCluster provided a fast viable alternative for their research needs.

Among its many advantages, CloudyCluster provided a simple and fast means to transfer their workflow from the university data center to the AWS cloud. With its easy UI and walk-through process, CloudyCluster allowed Apon's team to port the project workflow from the university data center to the cloud in under an hour. No longer competing with other researchers on campus for data center resources, the research team eliminated wait time and were able to run jobs whenever they needed. Additionally, CloudyCluster allows research assistants to focus on the research instead of losing time to system administration.



**USER-FRIENDLY CLOUDYCLUSTER INTERFACE**

## JOB SCALABILITY

Due to the scope of resources required for topic modeling, CloudyCluster offers another advantage with its scalability. CCQ allows users to let CloudyCluster determine and provision compute resources as needed, providing faster time to results with ease. CCQ also integrates tightly with the AWS Spot Market allowing users to save up to 90% off of the AWS On-Demand price. This allows users to perform more computation for the same price with minimal effort.

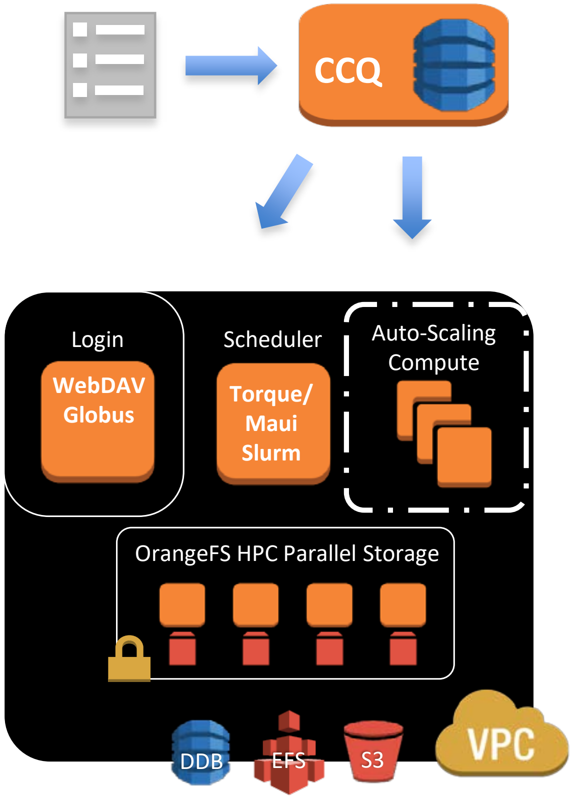
"We are using CloudyCluster to execute PLDA+ in the Amazon Web Services Cloud. With CloudyCluster and AWS we have access to a massive amount of resources that allow us to explore topics in a wide range of documents simultaneously" says Dr. Apon. She continues, “We expect that our calculations will lead to a better understanding of how topic models work and how they can be applied to the study and understanding of the themes and change of themes in many kinds of documents such as scientific journals, patents, historical documents, and more.”

## FASTER ACCESS

Because Amazon hosts many public datasets, including the Common Crawl dataset, CloudyCluster and AWS offer another advantage. For researchers running analysis on these public datasets already stored on Amazon, access to this data is faster. Storage space required for the same job in a data center is eliminated, as well as download time. Faster access means more efficient computation and a faster time to results.

As Apon’s team discovered, CloudyCluster offers several advantages for compute-intensive workflows on large data sets. Ease of porting to the cloud, job scalability, accessibility and elimination of wait times, and reduced time to results make CloudyCluster a valuable means to accomplishing research goals.

## ABOUT CLOUDYCLUSTER



CloudyCluster Technical Architecture

With virtually limitless resources available, HPC in the Amazon Web Services (AWS) cloud offers both faster computation and time to results. Cloudy Cluster makes it easy for anyone to set up an HPC cluster in the AWS cloud, including Compute, Storage, and Data Transfer.

Anyone can quickly and easily use CloudyCluster to run HPC and BigData jobs on AWS and CloudyCluster and AWS are available 24/7/365, with no waiting for hardware and installation time. Best of all, CloudyCluster lets you pause your HPC environment whenever you are not running a job, so you pay only for the processing time and storage you need.

For additional information about how CloudyCluster can support your work, visit *http://www.cloudycluster.com* and sign up for a free $100 AWS credit and schedule a free web meeting to help you get started.