aws efs kubernetes tensorflow pipeline spark

Sandor Magyari
Mon, Jan 15, 2018

Amazon Elastic File System on Kubernetes

At Banzai Cloud we provision different frameworks and tools like Spark, Zeppelin and most recently Tensorflow, all running on our Pipeline PaaS (built on Kubernetes).

One of Pipeline's early adopter is running a **Tensorflow** `Training Controller` using GPUs on AWS EC2 wired into our CI/CD pipeline and needed significant parallelization for reading training data. We have introduced support for Amazon Elastic File System and will make it publicly available in the forthcoming release of Pipeline. Beside **Tensorflow** they also use EFS for **Spark Streaming checkpointing** - instead of S3 (note that we don't use HDFS at all).

This post would like to walk you through the gotchas of **EFS** on Kubernetes and give you a clear idea about the benefits, before we dig into the Tensorflow and Spark Streaming examples in the next related posts, hence:

- At the end of the blog you will understand how EFS works on k8s
- You will be able to provision and use EFS with or without Pipeline
- Appreciate the simplicity and automation of what Pipeline does