Hadoop on Containers

What are containers?

Containers help to pack lots of applications onto a single bare metal host. They are different from a virtual machine (VM) in that they don't maintain a copy of the full operating system (OS) and share the kernel with the host and other containers. This also means that they are less isolated than VMs (read less secure). Docker Engine is to dockers as a hypervisor is to VMs.

Popular container orchestration tools

Docker Swarm, Kubernetes, Apache Mesos, Amazon ECS, CoreOS, Red Hat OpenShift. All of these help with container deployment and configuration, scaling, upgrades, security, etc.

Major challenges of running Hadoop on containers

1. Lack of support for truly stateful applications (Hadoop, Spark, Kafka need this but containers are stateless in nature). Hadoop was never meant to be run as a microservice.

2. Hadoop needs HDFS (read persistent storage). Though, Kubernetes has started providing PersistentVolumes, this still doesn't address the requirements for persistent storage (storage that exists when container dies).

3. Kubernetes has a project to support HDFS but it is still under development. It may take some time before it is production ready.

4. Lack of LDAP/AD support, only basic networking support available.

Some companies that offer containerized solutions using one of the orchestration tools

1. BlueData EPIC- has its own container orchestration layer but is also experimenting with Kubernetes and a solution based on Kubernetes exits. See this video for a demo.

2. Robin Systems – this company exists because container orchestration tools like Kubernestes, OpenShift etc. were only well suited for stateless applications (collection of microservices) although both of them now have a kind-of stateful solution. See this paper for details.

3. Portworx – acts like a plugin and can use Kubernetes, Mesos, etc. Like BlueData, it aims at reducing the complexity involved at managing containers via orchestration tools.

4. Altiscale (now acquired by SAP) – not much information available on their website other than the fact that Altiscale used to offer container based big data solutions.

References

1. Anant Chintamaneni 2017. “Big Data and Container Orchestration with Kubernetes (K8s)”. https://www.bluedata.com/blog/2017/12/big-data-container-orchestration-kubernetes-k8s/ . [Online; accessed 18-July-2018].

2. Sam Charrington 2015. “Running Hadoop on Docker, in Production and at Scale”. https://thenewstack.io/running-hadoop-docker-production-scale/. [Online; accessed 18-July-2018].

3. Susan Hall 2017. "Six Gotchas with Running Docker Containers on Hadoop". https://thenewstack.io/docker-hadoop-theres-good-bad-ugly/. [Online; accessed 18-July-2018].

4. Tom Phelan 2017. "Hadoop and Spark on Docker: Ten Things You Need to Know". https://www.bluedata.com/blog/2017/08/hadoop-spark-docker-ten-things-to-know/. [Online; accessed 17-July-2018].

5. Thomas Phelan 2016. "Lessons Learned Running Hadoop and Spark in Docker Containers". https://www.slideshare.net/BlueDataInc/lessons-learned-running-hadoop-and-spark-in-docker-containers. [Online; accessed 19-July-2018].

6. Portworx 2017. "A Production Ops Guide to Deploying Hadoop in Docker Containers". https://docs.portworx.com/applications/hadoop-docker.html. [Online; accessed 19-July-2018].