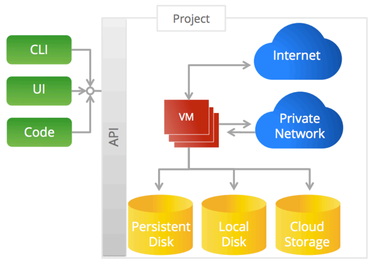
Deploying Big Data Applications to the Cloud

Anchal Agrawal and Professor Robert J. Brunner
Laboratory for Cosmological Data Mining, University of Illinois at Urbana-Champaign

Cloud vs. Local Machines



Google Compute Engine. Google and the Google logo are registered trademarks of Google Inc., used with permission.

- With terabytes of data at hand, downloading and processing data on local machines can be very challenging.
- Moving data-intensive applications to the cloud eliminates stress on local computing resources and simplifies the deployment process.
- Using the power of cloud computing, large jobs can be run in parallel across a computing cluster.

Linux Containers or Virtual Machines?

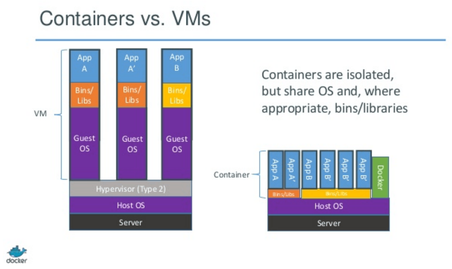
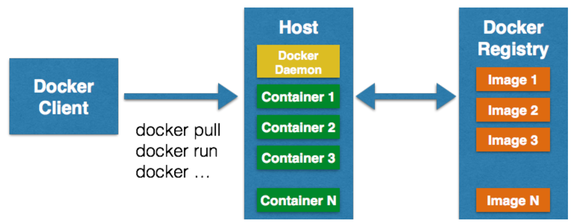


Image source: Docker Inc.

- A virtual machine consists of an entire operating system with the overhead of device drivers and memory management.
- Linux containers provide fast and lightweight process virtualization.
- By sharing the host OS, containers incur a small memory overhead. They start up instantly and have better performance.

An Introduction to Docker

- Docker is an open-source container platform that enables developers to run and manage distributed applications.
- Docker consists of the Docker engine (a runtime daemon), Docker Hub (an online registry for sharing Docker containers) and the Docker command line client which is used to manage containers.
- Apps built with Docker run on multiple platforms such as OS X or Windows computers and Linux servers.
- A Docker image consists of a base operating system such as Ubuntu along with a custom software stack and is used to create containers.
- Users can build custom containers on top of a base OS by installing libraries. The resulting container can be shared on Docker Hub.



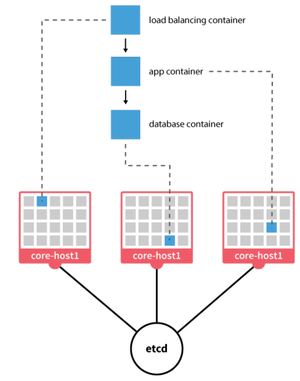
Docker's architecture. Image source: devopscube.com

Implemented Use Cases

- Used containers to run Hadoop Streaming jobs with Map and Reduce scripts written in Python.
- Used Pachyderm to run a MapReduce job. Pachyderm uses Docker containers instead of the JVM, which is used by the Hadoop ecosystem.
- Launched IPython notebook servers inside Docker containers.

Docker Containers as a Cluster

- Swarm is Docker's native clustering system that runs multiple Docker hosts as a single virtual host. Swarm can be configured with the Docker command line client.
- There are several third party Docker cluster management tools such as Deis and Shipyard.
- CoreOS is the preferred Linux distribution for Docker clusters. It comes with etcd (a distributed key-value store for service discovery) and fleet, a native command-line tool for cluster management.



A webapp running on a CoreOS cluster. Image source: coreos.com/using-coreos/

Future Work

Building a Docker-based cloud computing cluster that allows users to run Python data analysis jobs in parallel over multiple hosts.

References

- docs.docker.com/introduction/understanding-docker/
- rightscale.com/blog/cloud-management-best-practices/docker-vs-vms-combining-both-cloud-portability-nirvana/
- docs.docker.com/swarm/
- coreos.com/using-coreos/
- pachyderm.io/
- hadoop.apache.org/docs/r1.2.1/streaming.html
- devopscube.com/what-is-docker/
- deis.io/
- shipyard-project.com/

aagrawa4@illinois.edu | bigdog@illinois.edu