Docker and containers for Data Science

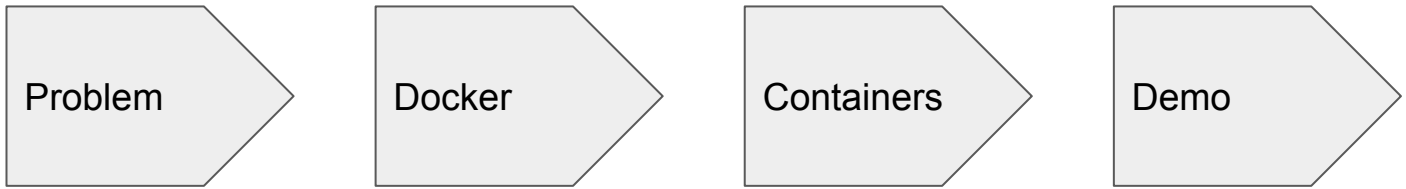
SDS 27/07/2020

Who am I

- Computer scientist working with Data
- EngD in computer science student at the University of St Andrews. MSc in Data Science. BSc in Computer Science.
- DKUK Volunteer
- I’ve worked on BI, DWH, PM, AI/ML and data analytics
- Playing with Docker since 2016
- https://darenasc.github.io
- @darenasc
- https://github.com/darenasc

Problem

- Configurations and installation of software
- Updates, failures, restores, of applications can be tedious
- Working on multiple projects in the same machine competing for resources, problems with library versions
- Installations using local machines resources



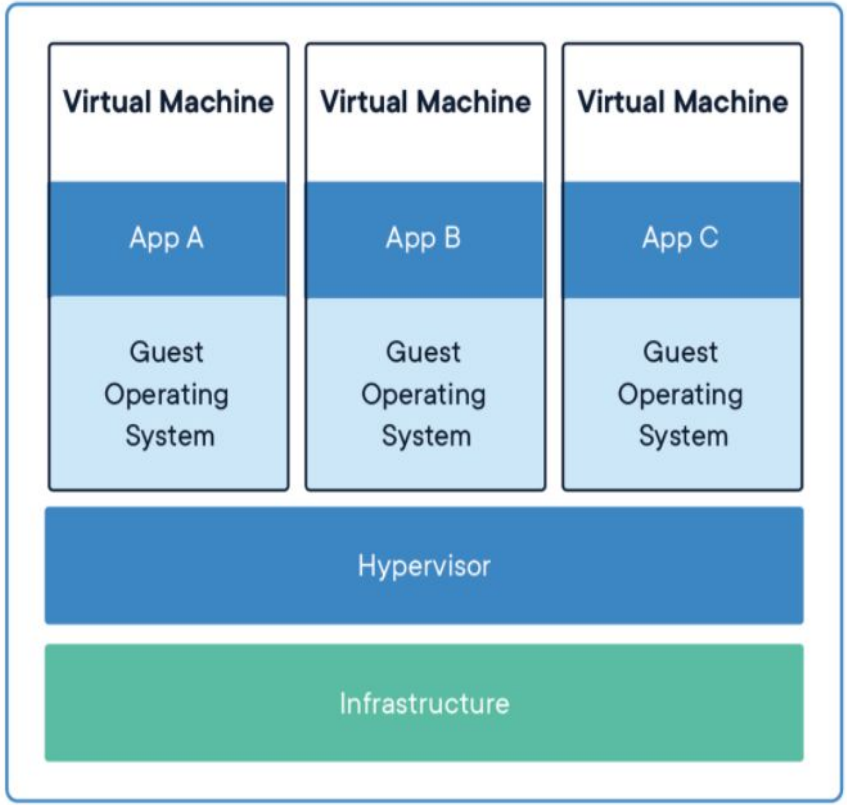
What is Docker

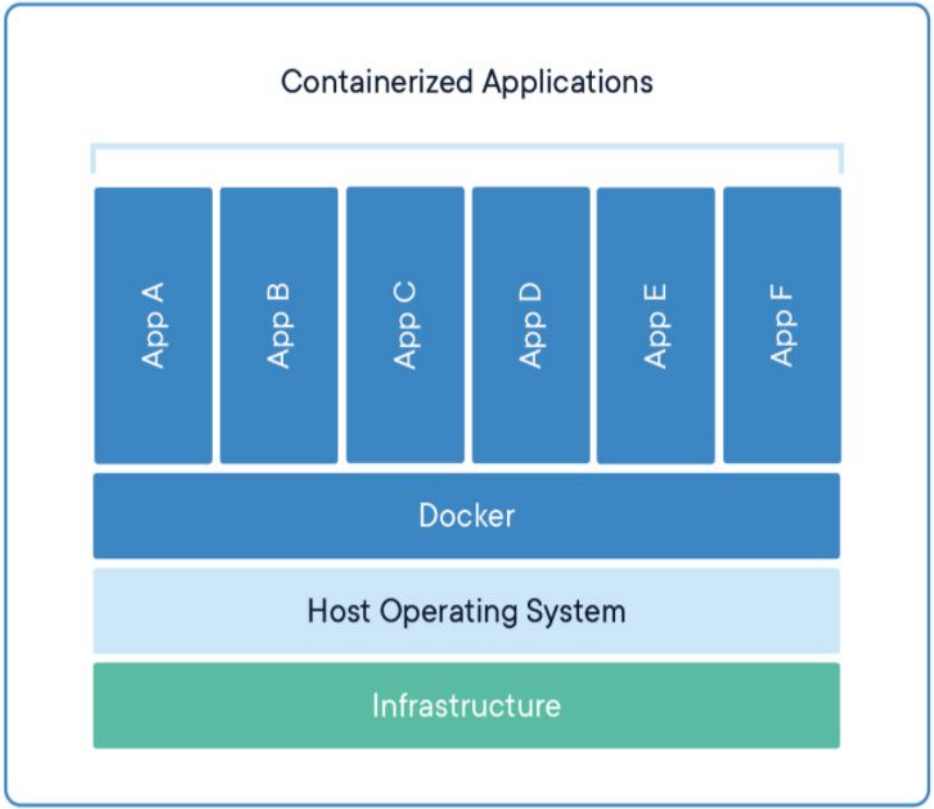
- Founded in 2010
- Open Sourced in 2013
- Platform that offers an API to use the local resources of the host machine
- Containers are not virtual machines
- Docker runs on Windows, Mac, and Linux
- It has good documentation!

Containers

- Development can happen locally and the software will run in a laptop, cluster or in the cloud with no modifications.
- Containers share resources with the host OS
- Portability. The “but it works on my machine!” excuse is no longer valid.
- Lightweight. You can run dozens of containers in the same machine
- Containers are an old concept. Docker wrapped and extended the existing Linux container technology

Virtual Machines and Containers





Images and Containers

- Docker uses **images** that work as templates

- Images run in **containers**

Dockerfiles

FROM

COPY

WORKDIR

RUN

Docker commands

```
docker ps

docker pull

docker run

docker restart [stop, start]

docker inspect

docker exec
```

What can I do with a container?

- Assign CPU, RAM memory, disk space, attach it to a network.
- Attach a persistent storage like for a database or file system
- One of the most common use case is use it for microservices
- Components of a data science architecture (an architecture that will allow you to **store**, **explore**, **analize**, **process** and **present** ALL the data in your organization)
- Divide the problem in its components and implement them in containers

DEMO

- Getting images
- Run containers
- Expose ports
- Persistent storage

https://github.com/darenasc/docker-demo

Questions?

References

- Top questions for getting started with Docker
- Manage data in Docker