Daniele Venzano
<venza@brownhat.org>

Docker Meetup Sophia Antipolis
16/03/2017

Intro

- A bit about me
- Data analytics in the real world
- Zoe: data analytics as a service
  - Architecture
  - ZApps
- Demo (Jupyter notebook with Apache Spark)

A bit about Daniele Venzano

- Software engineering (2006)
  - Linux embedded systems, kernel drivers, graphical interfaces

- Research (2011)
  - Code analysis, OpenFlow, automatic bug detection

- Research engineer (2013)
  - Research: virtualization, networking, distributed systems performance
  - Zoe: development and project management

What is “big” data ?

- The four Vs of Big Data:
  - Variety
    - Connect many different kinds of data: twitter feeds, audio files, images, web logs
  - Veracity
    - Is the data representative/of good quality?
  - Velocity
    - streaming, latency
  - Value
    - What is the objective of the analysis?

Data science

- Three phases
  - Data exploration
    - Data scientists “play” with data to extract value
    - On a subset of data on their laptop
  - Integration
    - DevOps transform semi-prototypes into production jobs
    - Try to replicate in production ad-hoc environments
  - Production jobs
    - Example: update e-commerce prices once per hour according to customer’s Twitter mood
    - Find bugs when running on the full set of data

How Zoe helps ?

- Three phases
  - Data exploration
    - On a subset of data on their laptop
    - Can work on full set of data from the start
  - Integration
    - Try to replicate in production ad-hoc environments
    - Reproducible environments thanks to Docker images
  - Production jobs
    - Find bugs when running on the full set of data
    - Fast iteration between development and production

Zoe

- Zoe is…
  - A software you can install
  - Specialized for data analytics
  - Generic: can run any data-driven framework
    - Apache Spark, Tensorflow, Flink, …
  - A user-friendly layer on top of:
    - A simple Docker engine (laptop install)
    - Docker Swarm
    - Kubernetes (end of March)
  - Efficient: smart scheduling for high resource utilization

Zoe top to bottom

User (data scientists, DevOps engineers, researchers)

ZApp

Pre-made by an administrator
Or
Created by users (needs to write Dockerfiles)

Zoe

Zoe itself with its UI, APIs and scheduling component

Backend

Docker, Kubernetes, etc.

Demo time!

- Zoe from the command-line
- Eurecom production deployment
- List and inspect executions
- Check containers in Swarm
- Take a look at the Algorithmic Machine Learning ZApp

Users and contributors

- Zoe is used by:
  - Eurecom for EU projects, research and teaching
  - Air France/KLM
  - KPMG
- Zoe development is funded by:
  - KPMG (dedicated development team)
  - The EU commission through the IOStack project

Thank you!

Daniele Venzano

venza@brownhat.org

http://zoe-analytics.eu

The future

- Three-months development cycles
  - Releases at end of March, June and September
- Focus on:
  - User interaction: ZApp store, workflow oriented web interface
  - Resource utilization: advanced scheduling techniques, dynamic resource allocation
  - Industrialization: CI, automated testing, user management
  - Storage: interaction with existing data lakes

Zoe architecture