Statistical reports and data analytics with distributed computing

Gabriella Azzopardi
Openlab Summer Student 2015

Supervisors: Axel Voitier
Filippo Maria Tilaro

19/08/2015
What is Data Analytics?
› Helps organisations better understand their data
› Used to discover patterns and useful information

What is Data Analytics as a Service (DAaaS)?
› Enables users to perform Data Analytics themselves
› From data acquisition to end-user visualisation, reporting and interaction
Data Analytics as a Service (Front-End)

Before
Data Analytics as a Service (Front-End)

After

Simple Use of Color in a Plot

A Sample Color Wheel

(Use this as a feel of monitor linearity)
Performing Data Analytics

The Problem
› 3000+ signals analysed each day
› Number of signals is expected to increase

Proposed Solution
› Parallelize execution across a cluster of nodes
› Execute each component using Docker
What is Docker?

- Docker: build, ship and run distributed applications
- Container: lightweight and isolated
- Image: layers are directly on the Kernel
- Hub: build, update, download images

Why?

- Simple installation
- Easily scalable
- Straightforward to maintain
Data Analytics as a Service (Back-end)

- Network File system (NFS):
  - Data Analytics algorithm to be run
  - Data to be used
- Master Container: Divide work across slaves
- Slave Container: Run allocated task in R
- Jupyter Container: Hosts Jupyter in browser
- Jupyter through Web browser:
  - Write, compile, share code
  - View and share data and results
DAaaS with Distributed Computing

1. User writes R script and it is run on Master.
2. Master divides task across slaves
3. Results are displayed in browser
Future Work:

- Find alternative to NFS, such as DFS
Thank you!