Statistical reports and data analytics with distributed computing

Gabriella Azzopardi
Openlab Summer Student 2015

Supervisors: Axel Voitier
Filippo Maria Tilaro

› **19/08/2015**

EN Engineering Department

ICE Industrial Controls and Electronics Group

SIEMENS

CERN openlab

Background image: Shutterstock

Introduction

What is Data Analytics?

- Helps organisations better understand their data
- Used to discover patterns and useful information

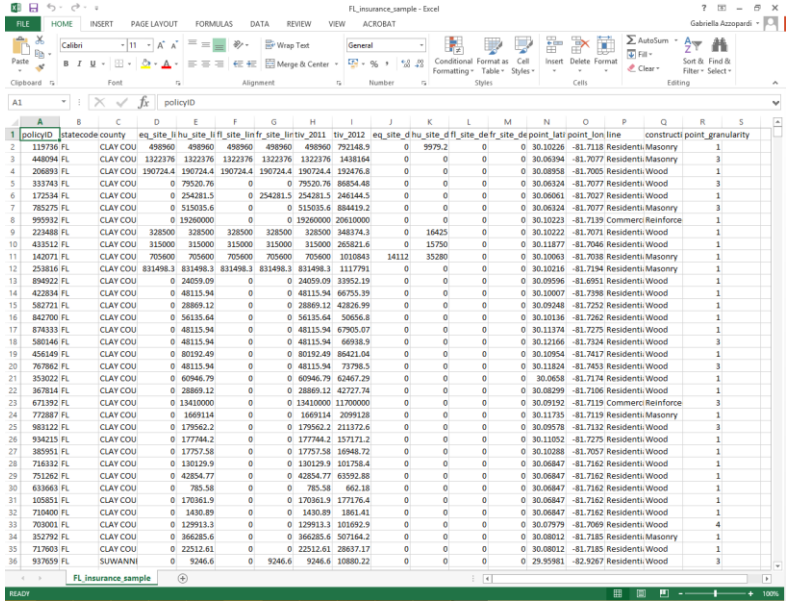
What is Data Analytics as a Service (DAaaS)?

- Enables users to perform Data Analytics themselves
- From data acquisition to end-user visualisation, reporting and interaction

Background image: Shutterstock

Data Analytics as a Service (Front-End)

Before



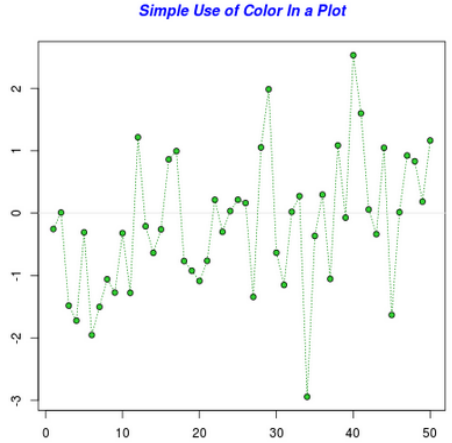
policyID	statecode	county	eq_site_li	hu_site_li	fl_site_lin	fr_site_lin	tiv_2011	tiv_2012	eq_site_d	hu_site_d	fl_site_de	fr_site_de	point_lati	point_lon	line	constructi	point_granularity
119736	FL	CLAY COU	498960	498960	498960	498960	498960	792148.9	0	9979.2	0	0	30.10226	-81.7118	Residenti	Masonry	1
448094	FL	CLAY COU	1322376	1322376	1322376	1322376	1322376	1438164	0	0	0	0	30.06394	-81.7077	Residenti	Masonry	3
206893	FL	CLAY COU	190724.4	190724.4	190724.4	190724.4	190724.4	192476.8	0	0	0	0	30.08958	-81.7005	Residenti	Wood	1
333743	FL	CLAY COU	0	79520.76	0	0	79520.76	86854.48	0	0	0	0	30.06324	-81.7077	Residenti	Wood	3
172534	FL	CLAY COU	0	254281.5	0	254281.5	254281.5	246144.5	0	0	0	0	30.06061	-81.7027	Residenti	Wood	1
785275	FL	CLAY COU	0	515035.6	0	0	515035.6	884419.2	0	0	0	0	30.06324	-81.7077	Residenti	Masonry	3
995932	FL	CLAY COU	0	19260000	0	0	19260000	20610000	0	0	0	0	30.10223	-81.7139	Commerci	Reinforce	1
223488	FL	CLAY COU	328500	328500	328500	328500	328500	348374.3	0	16425	0	0	30.10222	-81.7071	Residenti	Wood	1
433512	FL	CLAY COU	315000	315000	315000	315000	315000	265821.6	0	15750	0	0	30.11877	-81.7046	Residenti	Wood	1
142071	FL	CLAY COU	705600	705600	705600	705600	705600	1010843	14112	35280	0	0	30.10063	-81.7038	Residenti	Masonry	1
253816	FL	CLAY COU	831498.3	831498.3	831498.3	831498.3	831498.3	1117791	0	0	0	0	30.10216	-81.7194	Residenti	Masonry	1
894922	FL	CLAY COU	0	24059.09	0	0	24059.09	33952.19	0	0	0	0	30.09596	-81.6951	Residenti	Wood	1
422834	FL	CLAY COU	0	48115.94	0	0	48115.94	66755.39	0	0	0	0	30.10007	-81.7398	Residenti	Wood	1
582721	FL	CLAY COU	0	28869.12	0	0	28869.12	42826.99	0	0	0	0	30.09248	-81.7252	Residenti	Wood	1
842700	FL	CLAY COU	0	56135.64	0	0	56135.64	50656.8	0	0	0	0	30.10136	-81.7282	Residenti	Wood	1
874333	FL	CLAY COU	0	48115.94	0	0	48115.94	67905.07	0	0	0	0	30.11374	-81.7275	Residenti	Wood	1
580146	FL	CLAY COU	0	48115.94	0	0	48115.94	66938.9	0	0	0	0	30.12166	-81.7324	Residenti	Wood	3
456149	FL	CLAY COU	0	80192.49	0	0	80192.49	86421.04	0	0	0	0	30.10954	-81.7417	Residenti	Wood	1
767862	FL	CLAY COU	0	48115.94	0	0	48115.94	73798.5	0	0	0	0	30.11824	-81.7453	Residenti	Wood	3
353022	FL	CLAY COU	0	60946.79	0	0	60946.79	62467.29	0	0	0	0	30.0658	-81.7174	Residenti	Wood	1
367814	FL	CLAY COU	0	28869.12	0	0	28869.12	42727.74	0	0	0	0	30.08299	-81.7106	Residenti	Wood	1
671392	FL	CLAY COU	0	13410000	0	0	13410000	11700000	0	0	0	0	30.09192	-81.7119	Commerci	Reinforce	3
772887	FL	CLAY COU	0	1669114	0	0	1669114	2099128	0	0	0	0	30.11735	-81.7139	Residenti	Masonry	1
983122	FL	CLAY COU	0	179562.2	0	0	179562.2	211372.6	0	0	0	0	30.09578	-81.7132	Residenti	Wood	3
934215	FL	CLAY COU	0	177744.2	0	0	177744.2	157171.2	0	0	0	0	30.11052	-81.7275	Residenti	Wood	1
385951	FL	CLAY COU	0	17757.58	0	0	17757.58	16948.72	0	0	0	0	30.10288	-81.7057	Residenti	Wood	1
716332	FL	CLAY COU	0	130129.9	0	0	130129.9	101758.4	0	0	0	0	30.06847	-81.7162	Residenti	Wood	1
751262	FL	CLAY COU	0	42854.77	0	0	42854.77	63592.88	0	0	0	0	30.06847	-81.7162	Residenti	Wood	1
633663	FL	CLAY COU	0	785.58	0	0	785.58	662.18	0	0	0	0	30.06847	-81.7162	Residenti	Wood	1
105851	FL	CLAY COU	0	170361.9	0	0	170361.9	177176.4	0	0	0	0	30.06847	-81.7162	Residenti	Wood	1
710400	FL	CLAY COU	0	1430.89	0	0	1430.89	1861.41	0	0	0	0	30.06847	-81.7162	Residenti	Wood	1
703001	FL	CLAY COU	0	129913.3	0	0	129913.3	101692.9	0	0	0	0	30.07979	-81.7069	Residenti	Wood	4
352792	FL	CLAY COU	0	366285.6	0	0	366285.6	507164.2	0	0	0	0	30.08012	-81.7185	Residenti	Masonry	1
717603	FL	CLAY COU	0	22512.61	0	0	22512.61	28637.17	0	0	0	0	30.08012	-81.7185	Residenti	Wood	1
937659	FL	SUWANNI	0	9246.6	0	9246.6	9246.6	10880.22	0	0	0	0	29.95981	-82.9267	Residenti	Wood	3

Background image: Shutterstock

Data Analytics as a Service (Front-End)

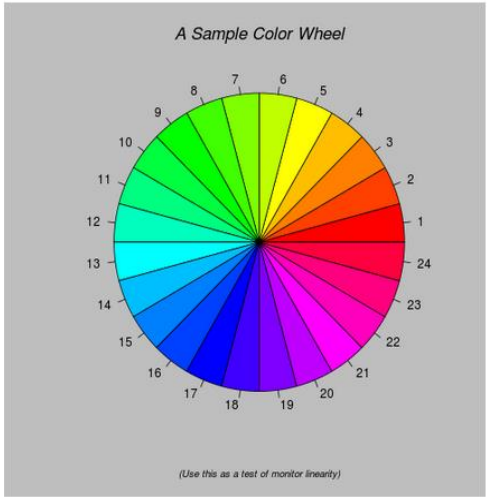
After

```
x <- stats::rnorm(50)
opar <- par(bg = "white")
plot(x, ann = FALSE, type = "n") +
abline(h = 0, col = gray(.90)) +
lines(x, col = "green4", lty = "dotted") +
points(x, bg = "limegreen", pch = 21) +
title(main = "Simple Use of Color In a Plot",
      xlab = "Just a Whisper of a Label",
      col.main = "blue", col.lab = gray(.8),
      cex.main = 1.2, cex.lab = 1.0, font.main = 4, font.lab = 3)
```




```
par(bg = "gray")

pie(rep(1,24), col = rainbow(24), radius = 0.9) +
title(main = "A Sample Color Wheel", cex.main = 1.4, font.main = 3) +
title(xlab = "(Use this as a test of monitor linearity)",
      cex.lab = 0.8, font.lab = 3)
```



Background image: Shutterstock

Performing Data Analytics

The Problem

- 3000+ signals analysed each day
- Number of signals is expected to increase

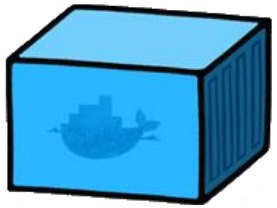
Proposed Solution

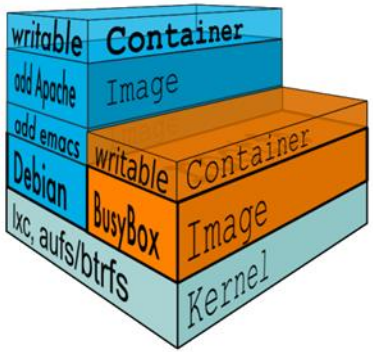
- Parallelize execution across a cluster of nodes
- Execute each component using Docker

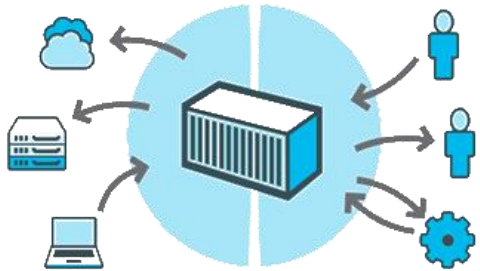
Background image: Shutterstock

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







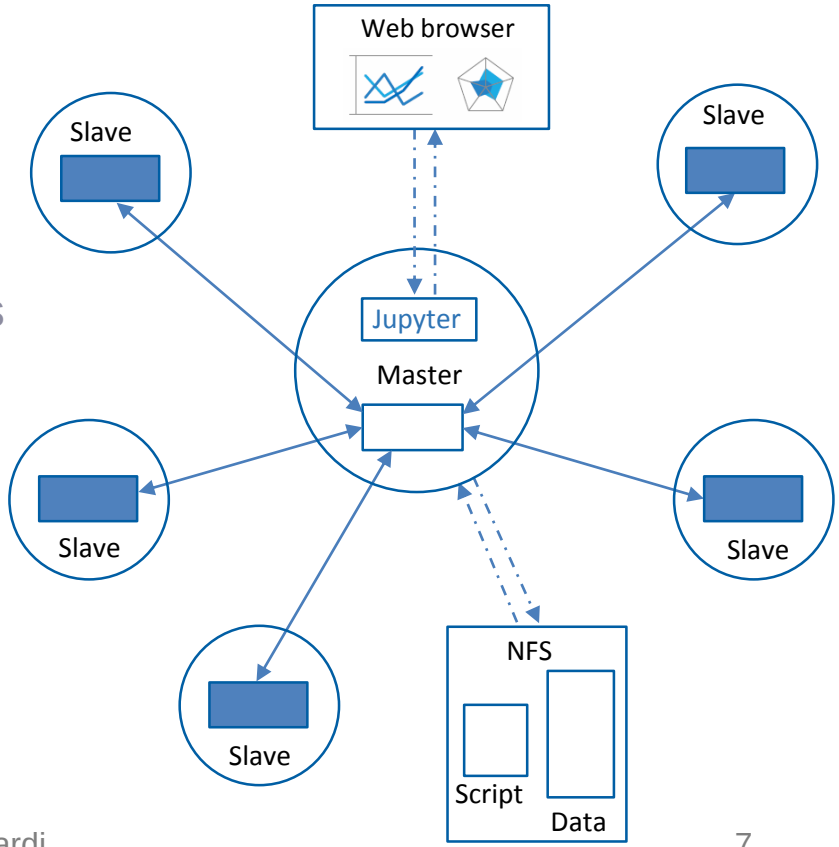
Why?

- Simple installation
- Easily scalable
- Straightforward to maintain

Background image: Shutterstock

Data Analytics as a Service (Back-end)

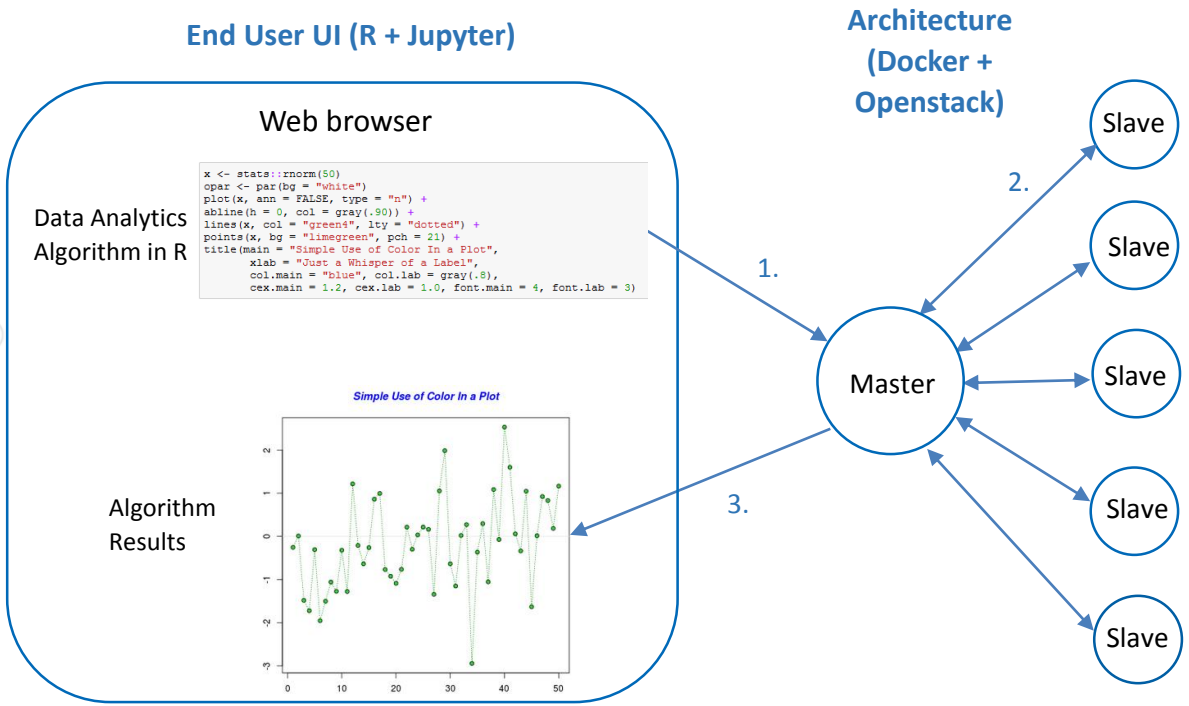
- Network File system (NFS):
  - Data Analytics algorithm to be run
  - Data to be used
- Master Container: Divide work across slaves
- Slave Container: Run allocated task in R
- Jupyter Container: Hosts Jupyter in browser
- Jupyter through Web browser:
  - Write, compile, share code
  - View and share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



Background image: Shutterstock

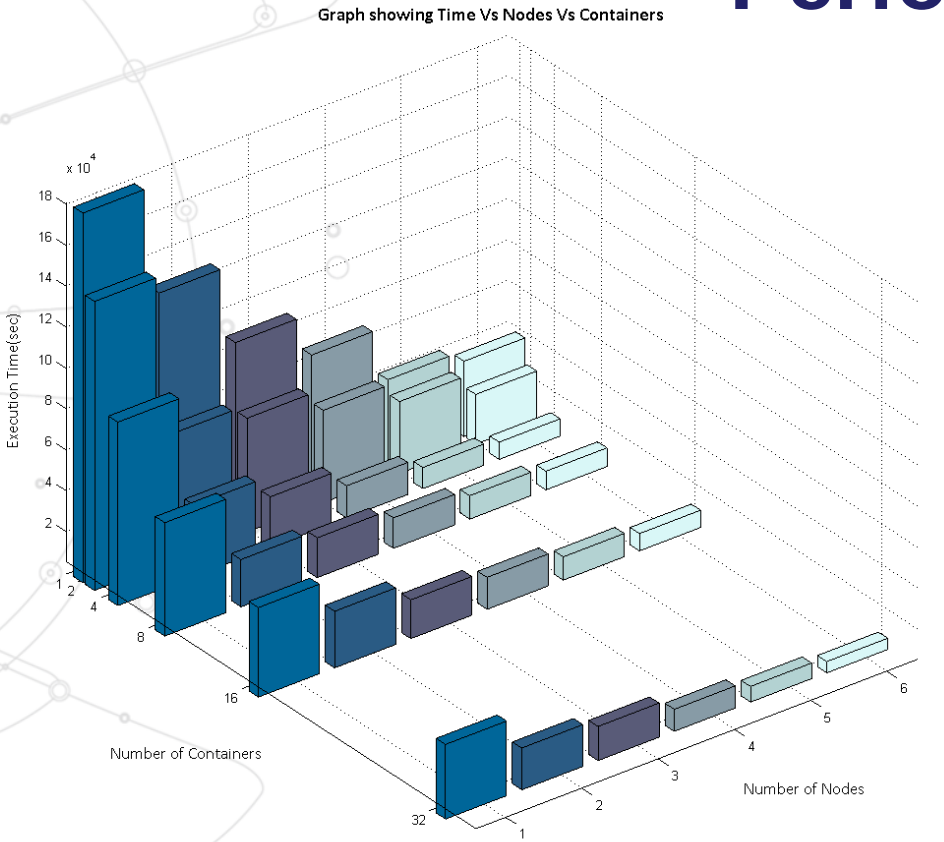
DAaaS with Distributed Computing

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



Background image: Shutterstock

Performance Gains



Future Work:

- Find alternative to NFS, such as DFS

Background image: Shutterstock

Thank you!

Background image: Shutterstock