

# Data analysis on the cloud

Roumeliotis Fotis

Giannakopoulou Teodora

Supervisor: Spyridon Trigazis



# Who are we?

We are High-school students that came at CERN through the greek HSSIP.

Roumeliotis Fotis



Trikala

Giannakopoulou Theodora



Komotini

# Description of the project

With our arrival, we started working on a project in the IT-department by the name “Analyzing massive datasets in the cloud”. During this project we learnt about:

- Cloud compute resources (Virtual Machines with diff OS, Containers, Volumes)
- Storage Systems/solutions
- Clusters (Kubernetes, HTcondor)

# Cloud computing

- What are clouds made of?
- Linux servers mostly.

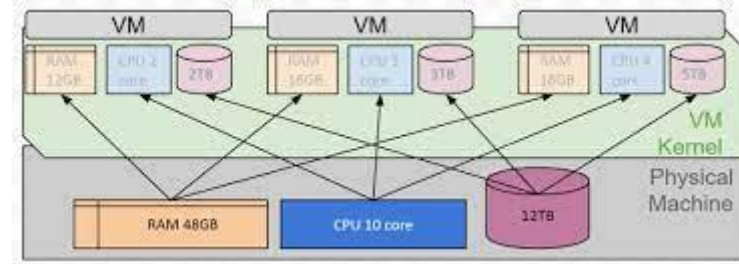


# Motivation of Cloud Computing

- Can be used by anyone in the world to access compute resources and large datasets
- It's very important for scientists in order to store, share and gain access to scientific data
- Is used in experiments (like LHCb, ALICE, ATLAS, CMS at CERN)
- It helps analyse massive amounts of data and plays an important role in the evolution of science.

# Cloud compute resources (I)

## Virtual machines



Virtual servers simulate physical servers and they come with advantages compared to physical hardware:

- Security: If one is compromised the others are safe
- Flexibility: Access by different people, hardware partitioning

VMs give us the chance to create a new one in just a few minutes.

# Cloud compute resources (II)

## Containers

However server virtualisation is presenting some problems like CPU and memory overhead.

- Reproducible environments with container images
- Isolation → Security
- Resource limits → Resource management

# Storage Systems

- AFS: a user's personal catalogue that can be found through different servers
- EOS: that's where all scientific data is stored, with much larger capacity than AFS
- cernbox: upload pictures/archives and save them on the cloud (backed by EOS)
- openstack block storage volumes: gives you the chance to create volumes and attach them to different servers.



# OpenStack Infrastructure

Production since 2013

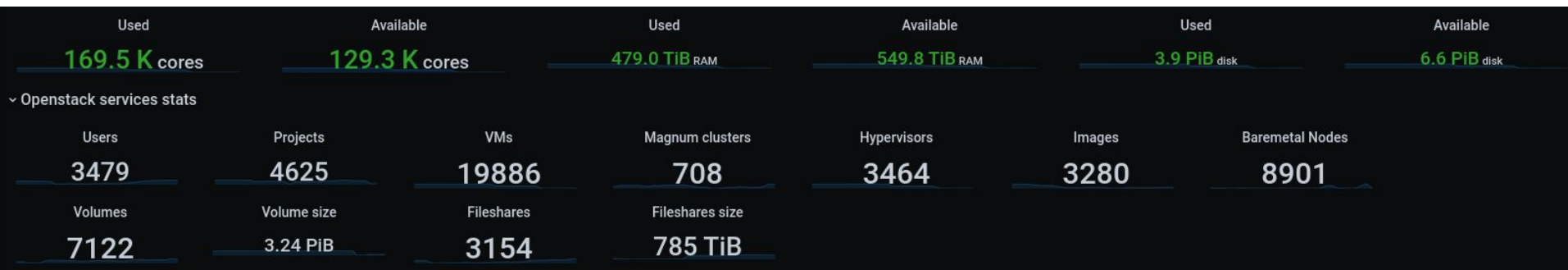
~ 170,000 cores

~20,000 vm running

~700 kubernetes clusters running

~ 3600 cluster nodes

~8,900 Physical servers (430,000 cores)



# Clusters

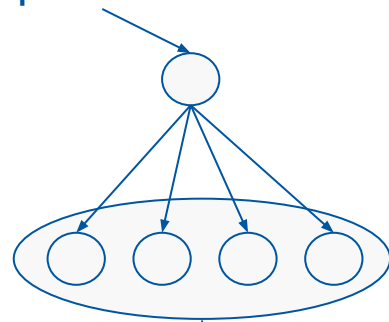
When we talk about clusters we mean objects of the same kind, which are synchronized and work as a team.

Control plane: Scheduling and configuration of workloads.

Data plane: Runs workloads and access user data.

Examples: Kubernetes, HTCondor (Batch), Slurm (HPC)

Control plane



Data plane

# Kubernetes

“Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.” (<https://kubernetes.io/>)

- Run in any cloud (on Premise, AWS, GCP, Azure, etc)
- Application/Experiment lifecycle
- Scalable workloads

# Examples

- Creation of personal VMs with different OS and clusters  
[https://clouddocs.web.cern.ch/tutorial/openstack\\_command\\_line.html](https://clouddocs.web.cern.ch/tutorial/openstack_command_line.html)
- Creation and attachment of volumes to different servers
- Aviator <https://gitlab.cern.ch/cloud-infrastructure/aviator>
- Event reconstruction /analysis of the CMS experiment (S'Cool Lab)  
<https://github.com/cms-opendata-education/scool-lab-sc18-opendata/>
- Word count: a programme on Python that counts words on a given text.
- Math examples

# Accessing computing resources

- Access LXPLUS by signing in (You can use putty and choose different hosts like lxplus8.cern.ch with CentOS 8)
  - “LXPLUS (Linux Public Login User Service) is the interactive logon service to Linux for all CERN users. The cluster LXPLUS consists of public machines provided by the IT Department for interactive work.”  
(lxplusdoc.web.cern.ch)
- Gain access and authentication by switching to your project using *export OS\_PROJECT\_NAME="<name of the project>"*

# Interacting with OpenStack

```
[thgianna@lxplus800 ~]$ openstack server list --name thgianna-c8 -c Name -c Status -c Image -c Flavor
```

Name	Status	Image	Flavor
thgianna-c8	ACTIVE	C8 - x86_64 [2021-09-01]	m2.medium

```
[thgianna@lxplus800 ~]$ openstack server show froumeli-personal
```

Field	Value
OS-EXT-AZ:availability_zone	cern-geneva-c
OS-EXT-STS:power_state	Running
OS-EXT-STS:vm_state	active
OS-SRV-USG:launched_at	2021-09-14T12:14:26.000000
addresses	CERN_NETWORK=188.184.102.246, 2001:1458:d00:3b::100:3ed
created	2021-09-14T12:13:06Z
flavor	m2.small (17895)
id	7b5a5b74-9116-4f83-9f5f-542dc40ab6cc
image	CC7 - x86_64 [2021-09-01] (69ba9cec-17e4-4082-a3d8-a680db0a1421)
key_name	froumeli-personal-laptop
name	froumeli-personal
project_id	9b461b5c-df46-42f3-9464-f7ef19dbc69a
status	ACTIVE
user_id	froumeli

# Interacting with kubernetes

```
[thgianna@lxplus800 ~]$ kubectl get nodes
```

NAME	STATUS	ROLES	AGE	VERSION
thgianna-cluster-01-gtct2vrsnukb-master-0	Ready	master	3d3h	v1.21.1
thgianna-cluster-01-gtct2vrsnukb-node-0	Ready	<none>	3d3h	v1.21.1

```
[thgianna@lxplus800 ~]$ kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
aviator-6ccd95446c-8gnlm	1/1	Running	0	3d1h

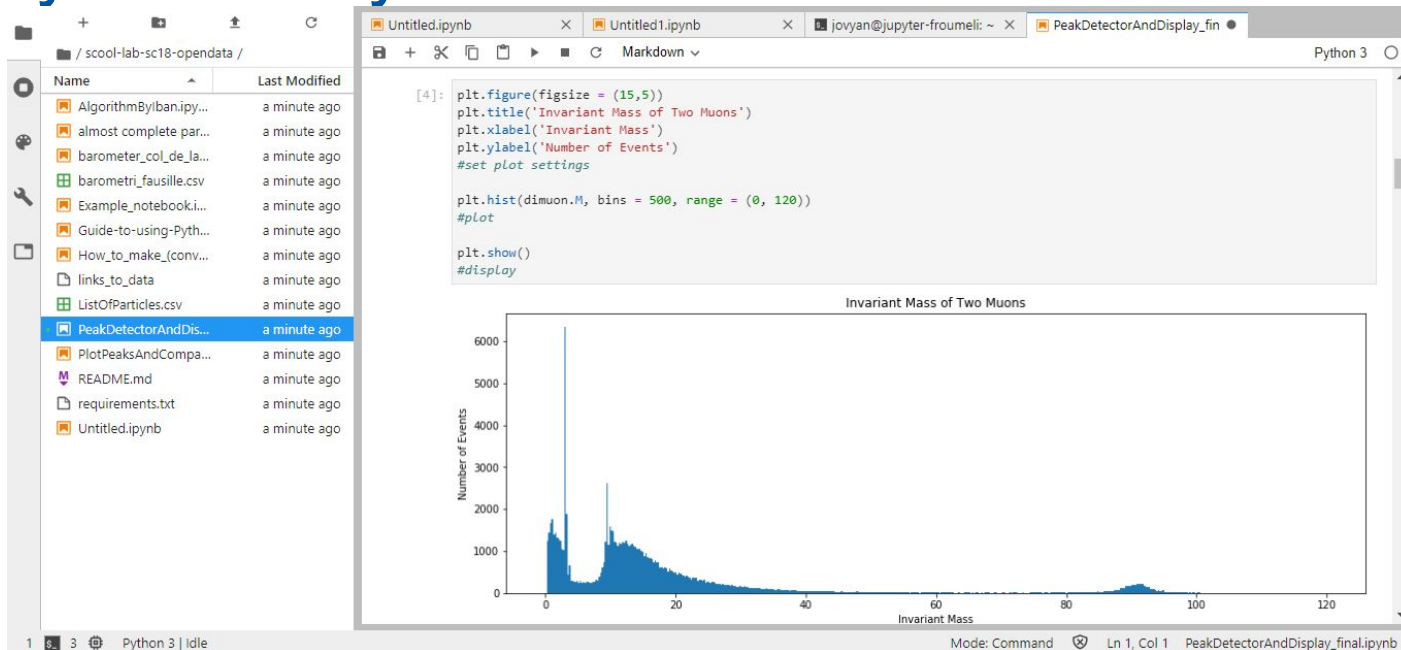
```
[thgianna@lxplus800 ~]$ kubectl get svc
```

NAME	TYPE	CLUSTER-IP	EXTERNAL-IP	PORT(S)	AGE
aviator	NodePort	10.254.58.141	<none>	80:32592/TCP	3d1h (188.184.72.56:32592)



# Interactive Physics Analysis

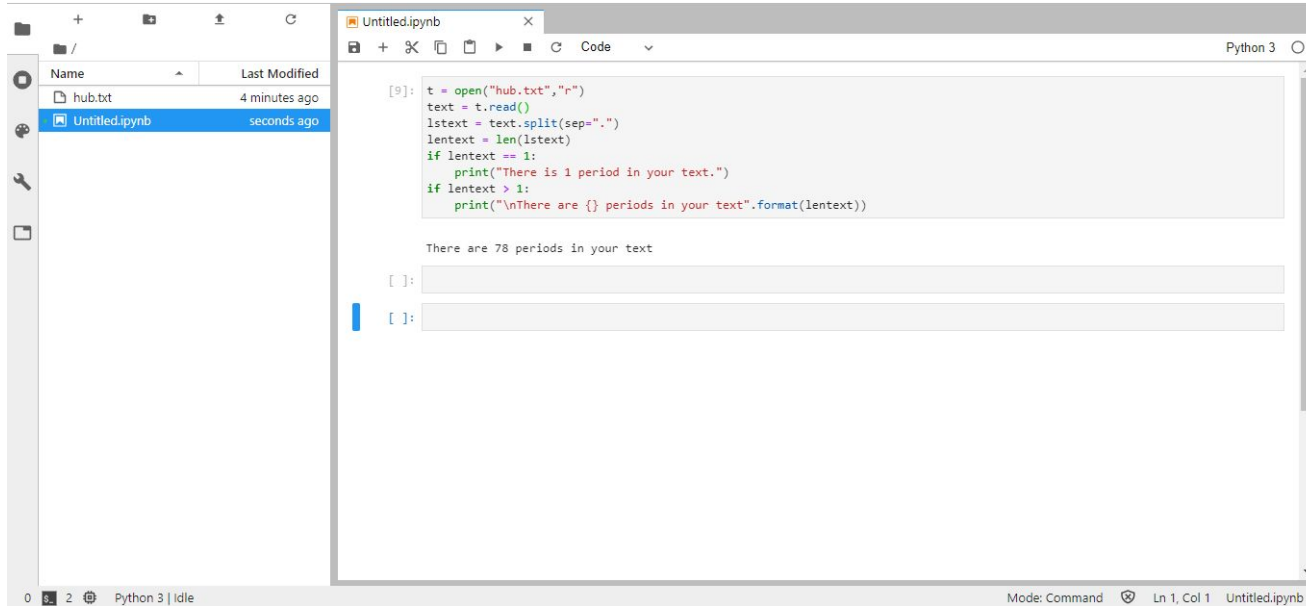
- [hub.cern.ch](http://hub.cern.ch)
- [ml.cern.ch](http://ml.cern.ch)



<https://github.com/cms-opendata-education/scool-lab-sc18-opendata/>



# Example word count program on jupyter



The screenshot displays a Jupyter Notebook environment. On the left, a file browser shows a directory with files 'hub.txt' (modified 4 minutes ago) and 'Untitled.ipynb' (modified seconds ago). The main area contains a code cell with the following Python code:

```
[9]: t = open("hub.txt", "r")
text = t.read()
ltext = text.split(sep=".")
lentext = len(ltext)
if lentext == 1:
    print("There is 1 period in your text.")
if lentext > 1:
    print("\nThere are {} periods in your text".format(lentext))
```

Below the code, the output of the cell is displayed: "There are 78 periods in your text". Below the output, there are two empty code cells, each starting with "[ ]:". The status bar at the bottom indicates "Python 3 | Idle", "Mode: Command", and "Ln 1, Col 1 Untitled.ipynb".

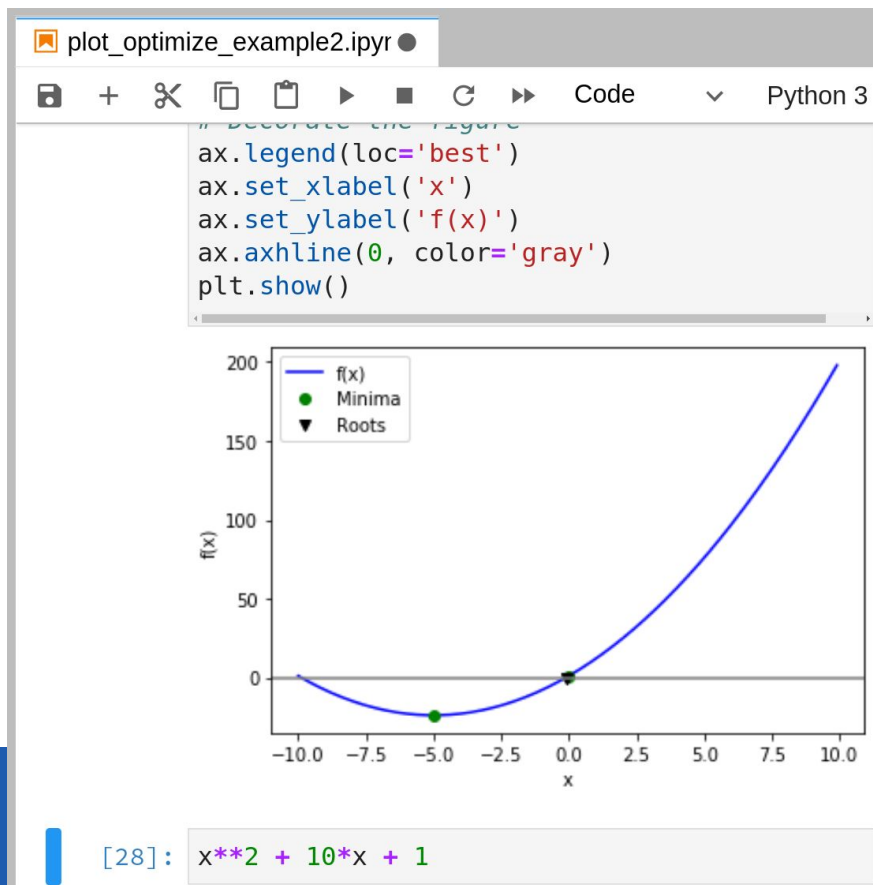
# Examples we can use at home

## Compute function root and minima

[https://scipy-lectures.org/intro/scipy/auto\\_examples/plot\\_optimize\\_example2.html](https://scipy-lectures.org/intro/scipy/auto_examples/plot_optimize_example2.html)

<https://matplotlib.org/stable/gallery/showcase/integral.html>

<https://scipy-lectures.org/>



# Links

- <https://jupyter-tutorial.readthedocs.io/en/latest/first-steps/install.html>
- <https://jupyter-docker-stacks.readthedocs.io/en/latest/>
- <https://colab.research.google.com/>
- <https://www.kaggle.com/>
- <https://docs.docker.com/get-docker/>
- <https://getfedora.org/> <https://opensource.com/article/18/5/dual-boot-linux>
- <https://docs.microsoft.com/en-us/windows/wsl/install>
- <https://www.virtualbox.org/wiki/Downloads>

# QUESTIONS

Thank you for your time!

