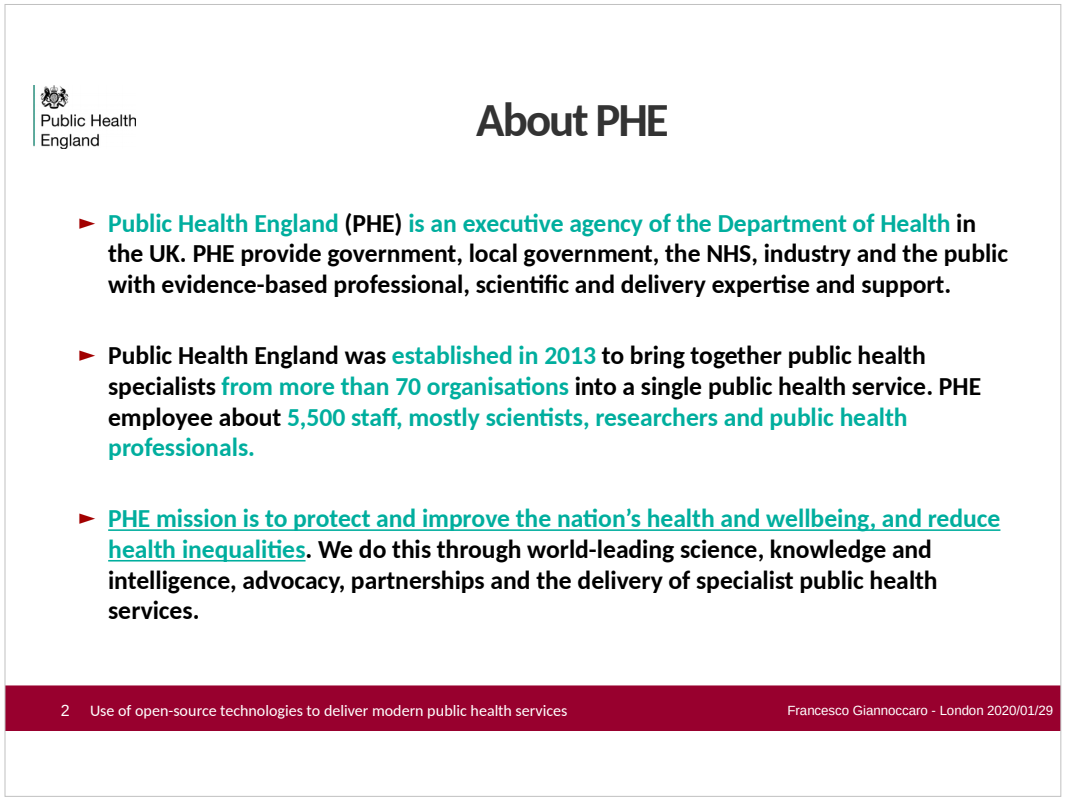
Use of open source cloud technologies to deliver modern public health services

Francesco Giannoccaro
29th January 2020

Public Health England

Protecting and improving the nation's health



HIGHLIGHT THE AMBITIOUS AND INSPIRING MISSION PUBLIC HEALTH ENGLAND HAS.

HOW THIS MISSION ALIGN TO THE OPEN SOURCE VALUES, IN THAT PHE AIMS TO DELIVER INNOVATIVE PUBLIC HEALTH SERVICES TO EVERYONE INDEPENDENTLY HOW RICH THEY ARE, REDUCING HEALTH INEQUALITY.

Public Health England

Wide range of public health services

PHE deliver a wide range of public health services including

- **research and scientific publications based on mathematical models such as Spatial Metapopulation Model for transmissible disease (eg Flu/Smallpox), predictive models applied to the Anthrax, inference problem to be able to infer: likely size of outbreak, location of source, spatial extent, etc**

- **pathogen genomics service, based on whole genome sequencing, for pathogen typing, surveillance and outbreak investigation. More than 100,000 bacterial and viral genomes have been sequenced since the service launch in 2014** (updated to Q1/2018)
- **campaigns such as: Be Clear on Cancer, Act FAST, Stoptober, Change4Life**

TALK ABOUT THE WORK PHE DOES IN GENERAL.

HIGHLIGHT THE SCIENTIFIC WORK DONE BY PHE IN THE SPACE OF ANTIMICROBIAL RESISTANCE (MONITORING BACTERIA AND VIRUS THAT ARE DEVELOPING RESISTANCE TO ANTIBIOTICS)

AS WELL AS THE SCREENING (AND IMMEDIATE DIAGNOSIS) OF PATIENTS THAT MAY HAVE BEEN EXPOSED TO AGGRESSIVE PATHOGENS

PREDICTIVE MODELS TO SIMULATE OUTBREAK OF TRANSMISSIBLE DISEASES

IN CASE OF OUTBREAK: SCREENING OF PEOPLE TRAVELLING FROM COUNTRIES CONSIDERED AT RISK

CAMPAIGNS ON CANCER, AND HOW SOCIAL BEHAVIOURS AND LIFE STYLE AFFECT OUR HEALTH AND WELL BEING

Public Health England

Challenges of structuring a central ICT department merging more than 70 different organisations

We focused on re-shaping the ICT infrastructure to better support PHE scientific community and designing a technology innovation roadmap to shift from a restricted stack of proprietary technologies into a more Open orchestrated ecosystem

initial technology stack managed by our central ICT department: the main focus was on supporting business as usual requirements

some of the new open source technologies that we have introduced during the redesign and innovation of our infrastructure to support the PHE scientific community

MERGING MORE THAN 70 ORGANISATION IN A SINGLE WELL INTEGRATED IS A BIG CHALLENGE

INITIALLY THE CENTRAL IT DEPARTMENT WAS FOCUSED ON **SUPPORTING B.A.U. requirements**

SCIENTIFIC COMPUTING SYSTEM WERE not supported, FRAGMENTED DEPLOYED AND MAINTAINED BY SMALL TEAM IN DIFFERENT DEPARTMENTS AND SCIENTIFIC UNITS

Public Health England

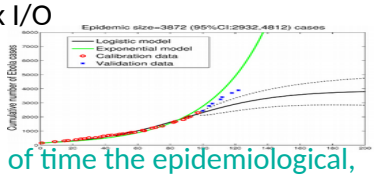
Use of HPC in Public Health England - Background

The **Bioinformatics** Unit

- Processing and analysing DNA for diagnostics and surveillance of infectious diseases. Samples received from patients with unidentified and potentially aggressive pathogens (bacteria and virus) that need urgent identification.
- Mainly **high throughput computing**: many small jobs, lots of CPU and disk I/O

The **Statistic Modeling and Economics** Department

- Running multiple real time models and simulations to predict expected pandemic disease dynamics, supporting national vaccination policy and control of antimicrobial resistance;
- **Traditional MPI** HPC: larger jobs, lots of CPU and small amount of disk I/O



Emergency Response Department

- Running multiple models and simulations to better understand ahead of time the epidemiological, social, behavioral drivers that exacerbate the risks posed by infectious disease threats, including bioterrorism;
- Mainly **traditional MPI** HPC and some GPU: larger jobs, lots of CPU and moderate disk I/O.

SLIGHTLY DIFFERENT TYPE OF WORKLOAD

TWO OF THEM run CPU BOUND applications

THE BIOINFORMATICS HPC SYSTEM has been designed to support HIGH THROUGHPUT COMPUTING , to run I/O BOUND workloads

Public Health England

Technology innovation program to develop scientific computing

ICT in 2015/2016 designed and implemented a **technology innovation program** underpinned by the ICT strategy, focused on **supporting the scientific community** within PHE. This work included:

1) HPC Cloud to support the needs for high performance and high throughput computing, based on the open source technology **OpenStack**

2) PHE scientific data catalogue: as a prerequisite to enable **Big Data analytics** capability in 2016 we started deploying **iRODS** an open-source technology to help our scientists to organise, catalogue, aggregate and share scientific data

3) Cross-platform orchestration systems to centrally mange heterogeneous hybrid and multi clouds and virtualization ecosystems, including Vmware, Openstack, AWS, Azure, GCP (based on **ManageIQ/CloudForms**)

4) Platform as a Service solution to support containerised applications: to support application life-cycle management, scalability and portability (solution based on **OKD/OpenShift**)

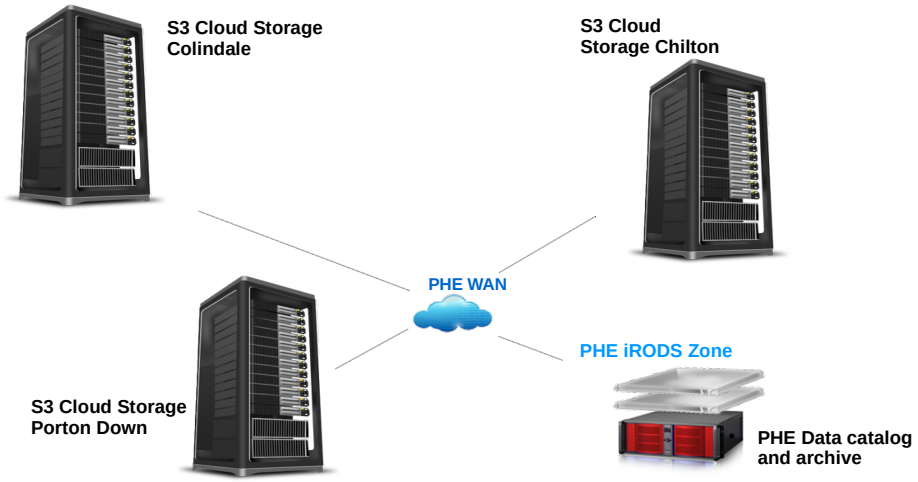
IN 2015 WE SHAPED THE NEW ICT STRATEGY WITH PRIMARY FOCUS FOCUS ON SUPPORTING THE SCIENTIFIC COMPUTING: **OPEN SOURCE FIRST** → WHY: openess, open science, accessibility and collaboration

HOW DIFFERENT OUR EVERY DAY LIFE WOULD BE IF THE **WORLD WIDE WEB HAD BEEN PATENTED**?

WHAT WOULD THE WORLD LOOK LIKE IF THE **HUMAN GENOME** WAS THE **INTELLECTUAL PROPERTY OF A COMPANY**? IT ALMOST HAPPENED

Public Health England

geo-distributed storage system (10 Petabyte) using on-premise cloud storage for scientific data cataloguing and archiving



Genomics is set to become the biggest source of data on the planet, overtaking the leading heavyweights – astronomy, YouTube and Twitter. Genome sequencing currently produces a staggering 25 petabytes of digital information per year. The amount of DNA sequencing data produced around the world is doubling approximately every seven months.
https://sangerinstitute.blog/2019/03/04/getting-smart-about-artificial-intelligence/

THERE IS AN IMPORTANT PREREQUISITE TO ENABLE BIG DATA ANALYTICS, DATA MASH UP, MACHINE LEARNING ETC:
MAKE DATA EASILY BROWSABLE, SEARCHABLE DISCOVERABLE, SHAREABLE

THE INCREDIBLE LARGE AMOUNT OF DATA THAT PHE IS PRODUCING AND MANAGING IS MOSTLY UNSTRUCTURED, WITH INCONSISTENCY IN NAMING POLICY FOR FILES
OFTEN NOT EASILY ACCESSIBLE, IN SILOS OR IN STORAGE SYSTEMS NOT EASILY DISCOVERABLES

NEEDED A TECHNOLOGY TO ALLOW THE CREATION OF SCIENTIFIC DATA CATALOGUE
WITH THE CAPABILITY OF ADDING AS MANY METADATA AS SCIENTIST WANTS/NEEDS
NO NEED TO MOVE DATA FROM WHERE THEY STORED
BUT STILL HAVE A WAY TO BROWSE AND SEARCH

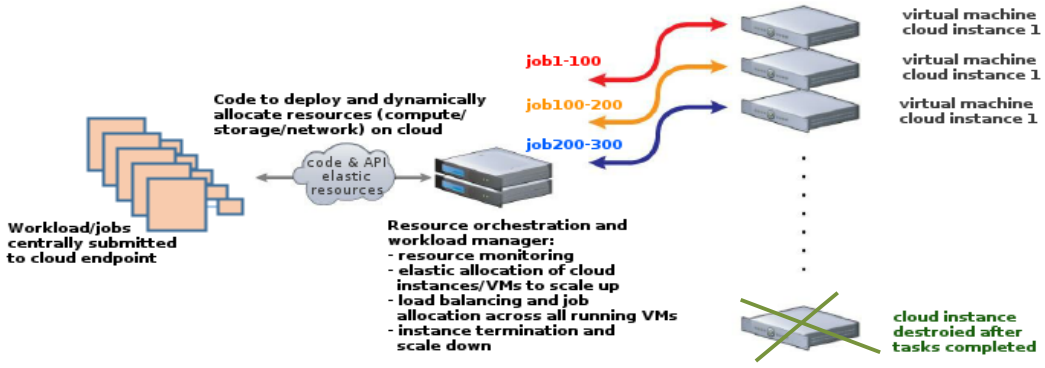
NEED FOR HYBRID CLOUD – ON PREMISE DATA/IO INTENSIVE WORKLOAD
AND ELASTIC BURSTING ON PUBLIC CLOUD FOR CPU INTENSIVE WORKLOADS

Public Health England

Benefits of using cloud technologies for different type of workloads
HPC workloads vs BAU and web application

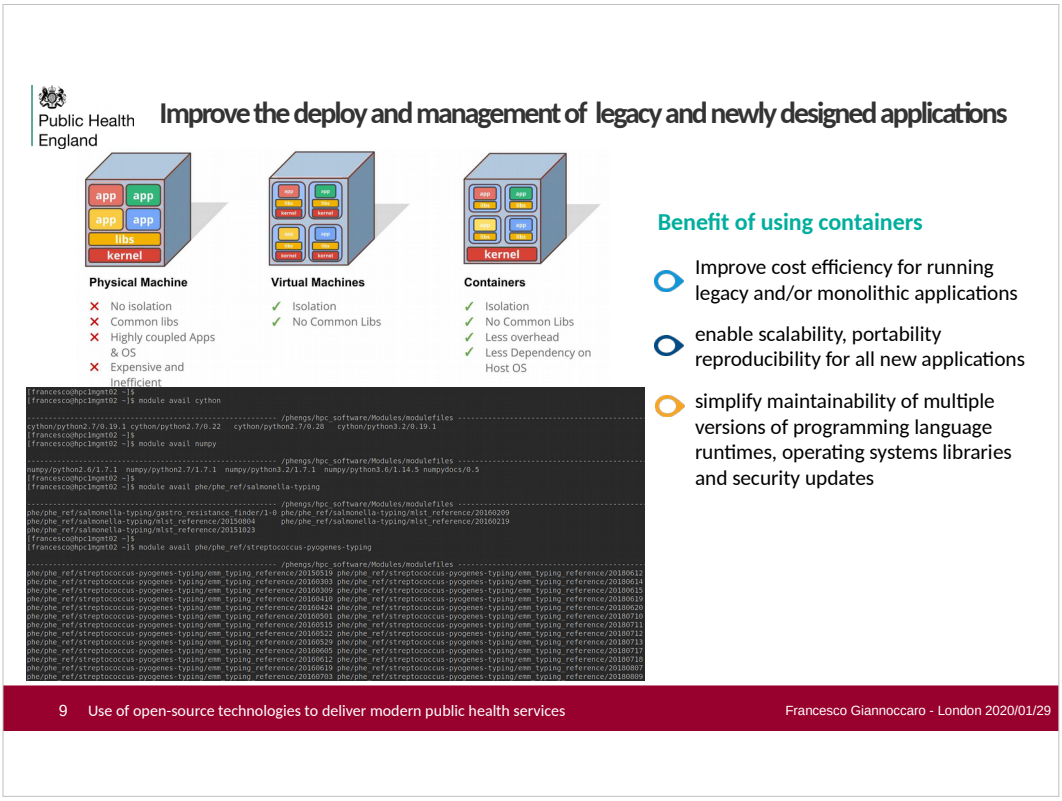
The on premise OpenStack platform delivered in PHE gives the organization the ability to use innovative cloud technologies to provide additional compute capacity (scale up and down on demand) to run data intensive workloads, without the risk of incurring in ever growing revenue costs generated by using public/commercial clouds.

To get the benefits of using IaaS cloud technologies it is required to use programming languages and APIs capable of managing code defined infrastructure components (compute, storage, networks) to make elastic use of resources, spawning cloud instances that only run for the time required to complete specific computational tasks and are then destroyed, releasing the allocated resources therefore minimizing the cost.



The benefits of using cloud technology ultimately reside in implementing pipelines capable of elastic use of resources

COST DIFFERENCE OF ON-PREMISE VS PUBLIC in IaaS for data intensive pipelines are order of magnitude different



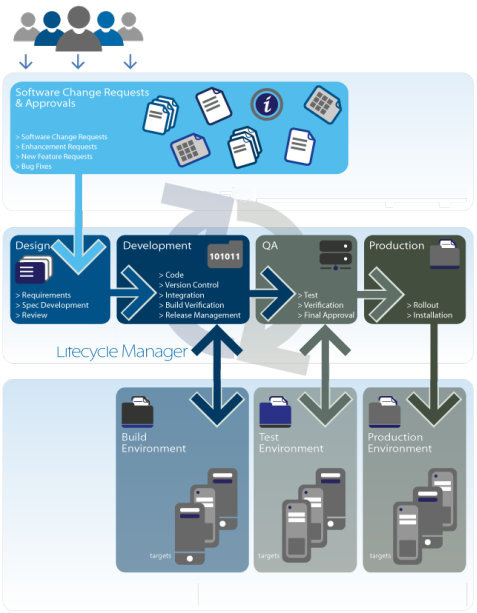
Improve cost efficiency and maintainability of legacy/monolithic application &
provide scalability, portability/reproducibility to new cloud-native application

Benefit of containers (similarity between web applications and HPC clusters modules):

challenges of maintainability of multiple versions of programming language runtimes, operating systems libraries and security updates
similarity between web applications and HPC clusters modules. In scientific computing containers (such as Singularity) are making the difference in improving portability and reproducability

Public Health England

Managing PHE Infrastructure as Code, improving automation and centralized configuration practices – phase 1 : GitLab, Ansible / Tower



The broad ecosystem of technologies that we chose to implement will accelerate service delivery and reduce operational costs through a **centralized deployment, configuration and orchestration management systems** that will provide:

- **automatic provisioning**
- centralised configuration, auditing and change tracking
- ensure **compliance** and governance
- resource quota enforcement

This will ensure that systems will be configured through configuration definition files, as scripts. This will also mean that **applying configuration** updates with code/scripts will be **fast and consistent**, allowing engineers to provision many servers faster and with less risks of error than any human could do through "type and click" procedures.

HIGHLIGHTS THE BENEFIT OF BUILDING AND ECOSYSTEM OF OPEN SOURCE TECHNOLOGY TO SUPPORT AND FOSTER AUTOMATION AND DEVOPS CULTURE/PRACTICES

GITLAB

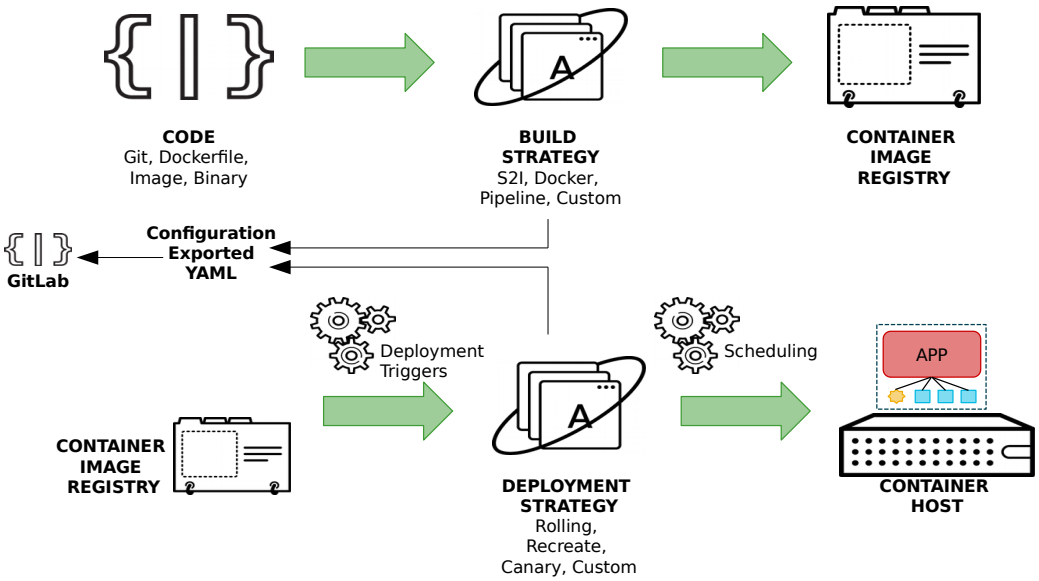
ANSIBLE

ANSIBLE AWX / TOWER

JENKINS (slightly anticipate content of next 2 slides, saying the benefit of having all these tools in a single well integrated PaaS: OKD/OpenShift)

Public Health England

OpenShift to leverage the benefits of using a well Integrated devops tools ecosystem – phase 2



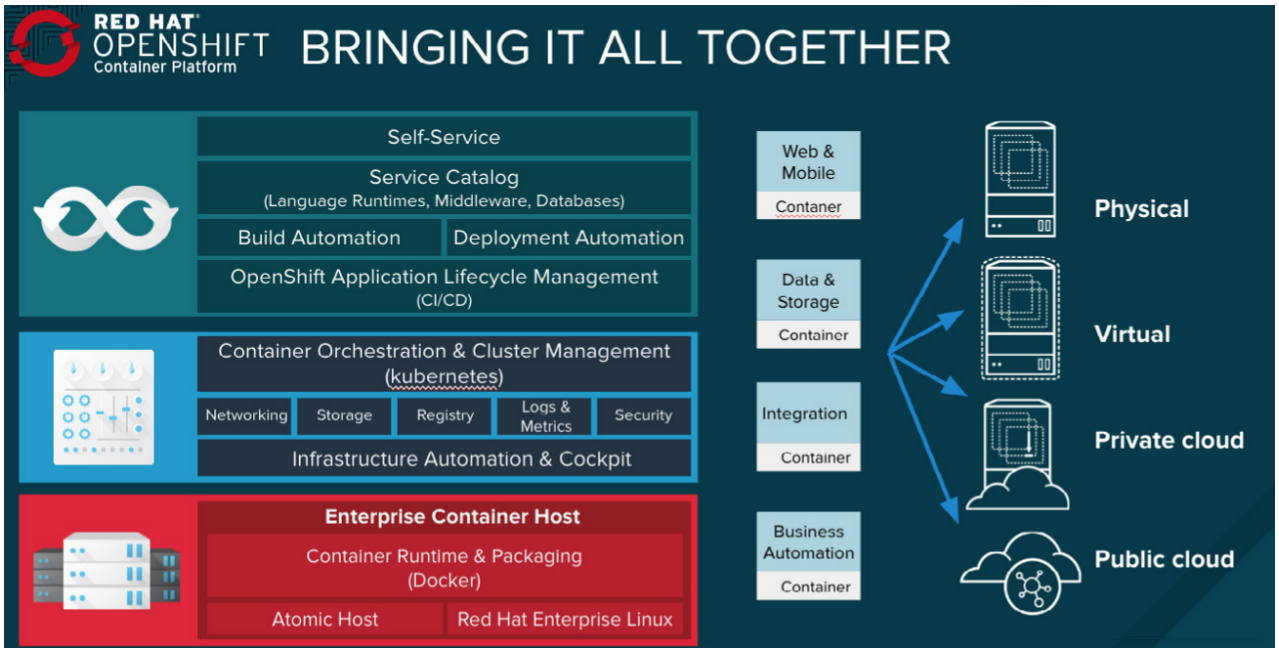
DEVOPS – FULL APPLICATION LIFECYCLE
SHARED STORAGE
NETWORKING ISOLATION
DNS
SECURE CONTAINER REGISTRY (without this PHE would never be able to use insecure resource such as Docker Hub)

LIBRAY OF PRE BUILT COMPONENTS
GITLAB
JENKINS

ENABLING AND FACILITATING DEVOPS CULTURE/ COLLABORATION

Public Health England

OpenShift simplify the management of cloud native apps, enabling autoscaling and the use of a well integrated devops tools ecosystem



PHE initial user case (in addition to HPC/Singularity)

New Intranet beta - used Jenkins pipelines for CI and image promotion

NTBSS - exploring .NET Core as deployment platform

Needs to use Kerberos to authenticate to MS SQL Cluster

'Sidecar' container keeps Kerberos credentials fresh

Maintained by IT department, app developers don't need to concern themselves with the details

UK National Screening Committe Recommendations beta - replacing a legacy non-GDS compliant web site.

Greenfield development

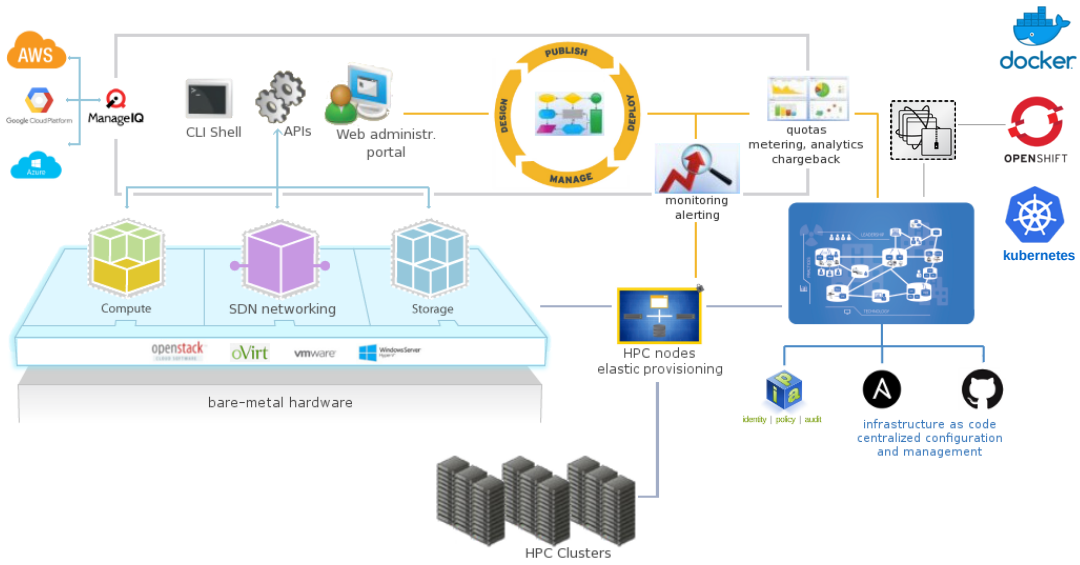
Will use Jenkins pipelines for CI/CD

Current focus in facilitate use of: JupyterHub, Tensorflow and GPU resources

Public Health England

Overview of the PHE hybrid multi cloud infrastructure

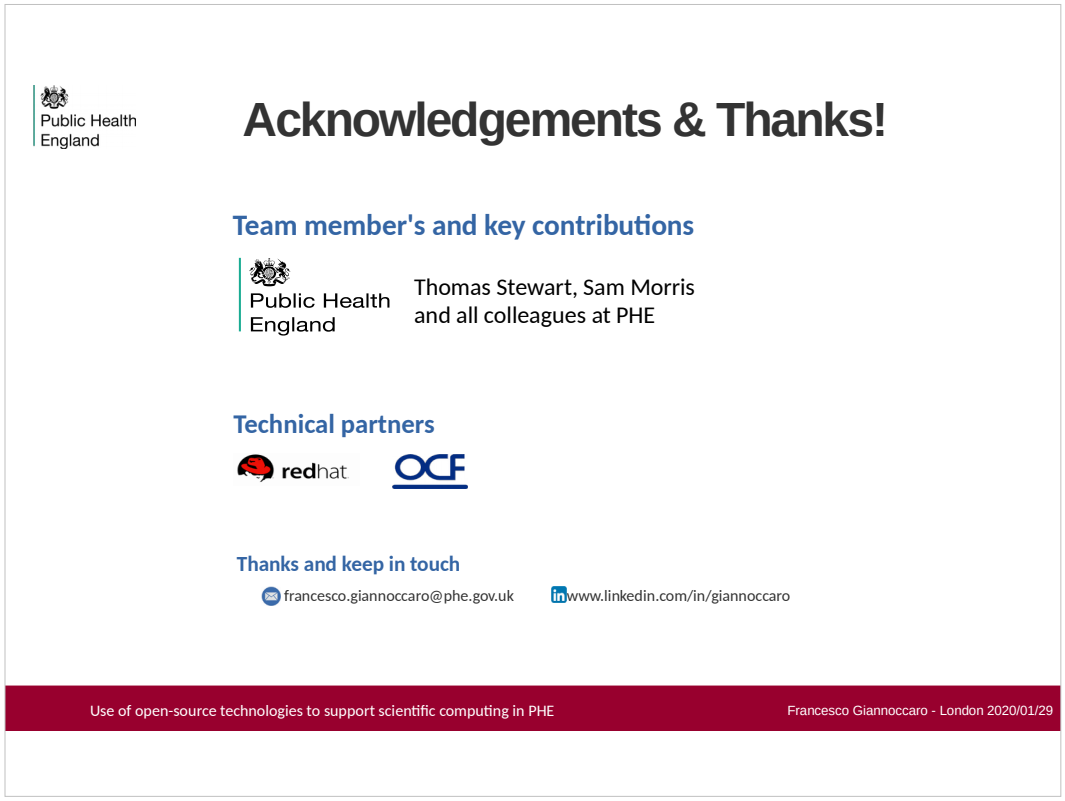
Multi-site/geo distributed infrastructure based on hybrid/multi cloud and containers platforms



THE CURRENT INFRASTRUCTURE

USING A WIDE ECOSYSTEM OF OPEN SOURCE TECHNOLOGIES

OVIRT
OPENSTACK
KUBERENETES
ANSIBLE
FREE IPA
GITLAB
MANAGE IQ
AND OTHER



MAKE SURE TO HAVE ENOUGH TIME IN THE ASSIGNED SLOT TO GIVE THANKS TO THE OPEN SOURCE COMMUNITY FOR MAKING OPEN SCIENCE POSSIBLE