

Introduction

Next-generation Sequencing (NGS) data imposes major challenges for processing due to its volume and complexity¹. Analysis of NGS data is generally based on many third party software, which are sometimes complex to install, configure and usually have dependencies that may arise in portability and reproducibility issues. Thus, it is necessary to develop infrastructures to store, manage and analyse massive genomic data in an efficient, scalable and reproducible way.

Materials and Methods

We have developed a Docker² container-based infrastructure for NGS data analysis comprising Bioinformatics and Big Data tools (Hadoop³ and Spark⁴). This setup was deployed in a classroom of 20 commodity hardware nodes at the Computing Center in Escuela Superior de Ingeniería y Tecnología (Universidad de La Laguna, Tenerife).

We implemented Docker images for the Hadoop and Spark components and a set of bioinformatics tools (**Figure 1**), including QC applications (FastQC, MultiQC, Qualimap2), aligners (BWA), variant callers (GATK4⁵, Platypus), and other supplementary software (JupyterLab).

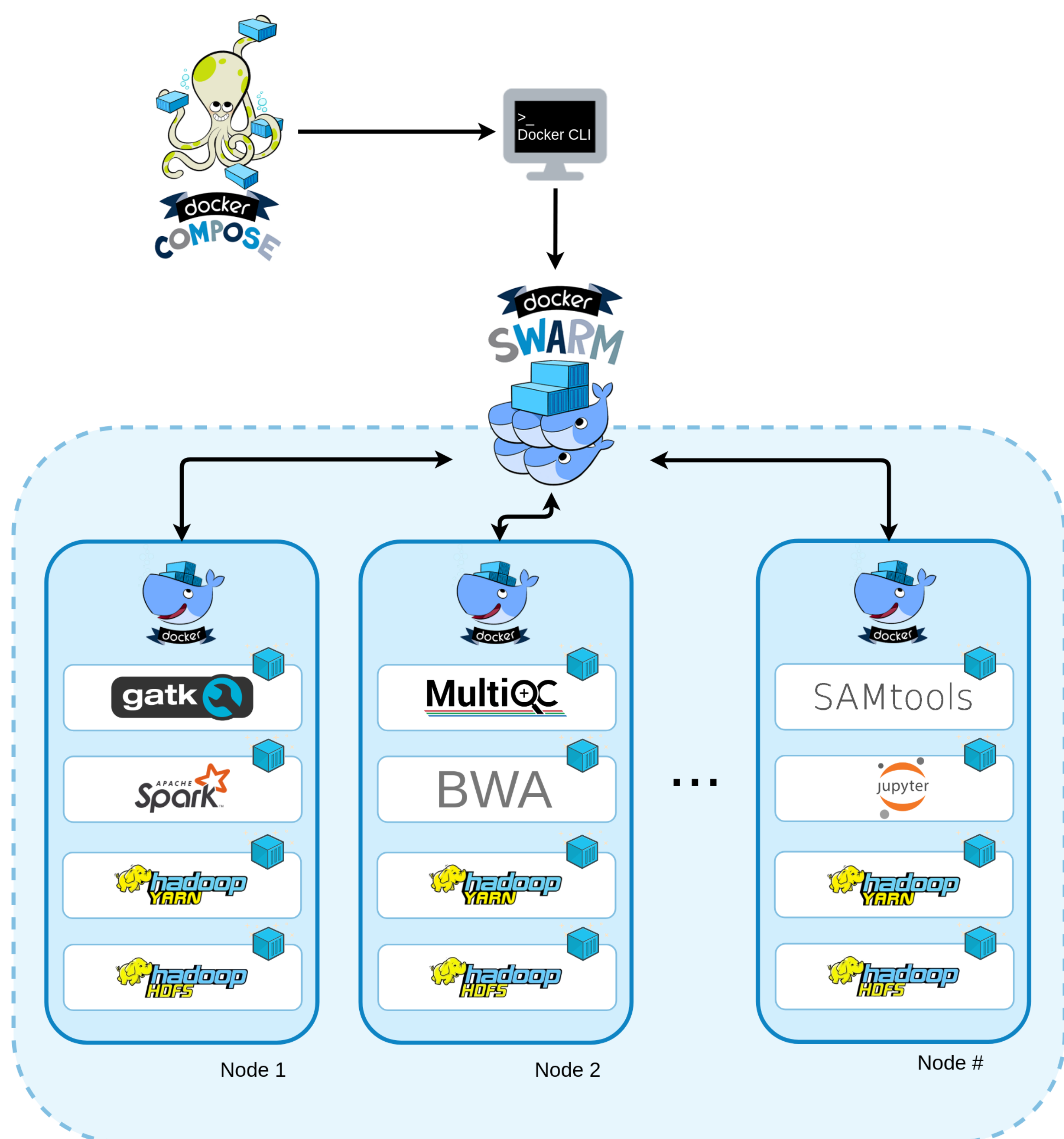


Figure 1. Overview of the architecture

Results

Docker Compose is used for multi-container definition and Docker Swarm for container orchestration and cluster management.

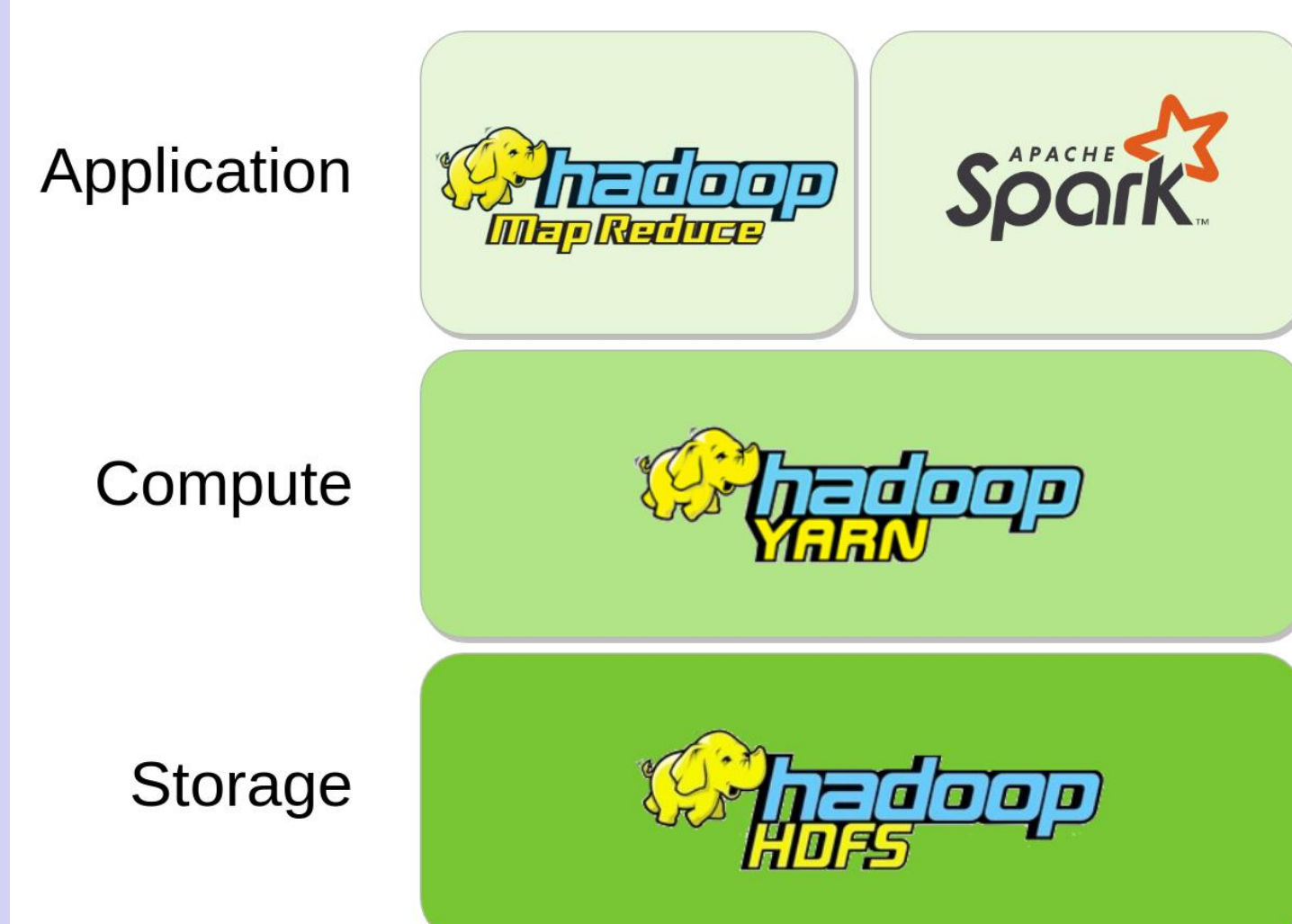


Figure 2. Spark on YARN overview

GATK4 container is configured to take advantage of the Spark cluster so Spark-enabled GATK tools can be run and parallelized throughout the cluster⁶.



Similarly, a Docker image containing JupyterLab is available for research and educational purposes. This image is also prepared to submit tasks to the Spark cluster.

The images provide many other tools ready to use right out of the box (**Figure 3**), and more applications can be easily integrated as needed.

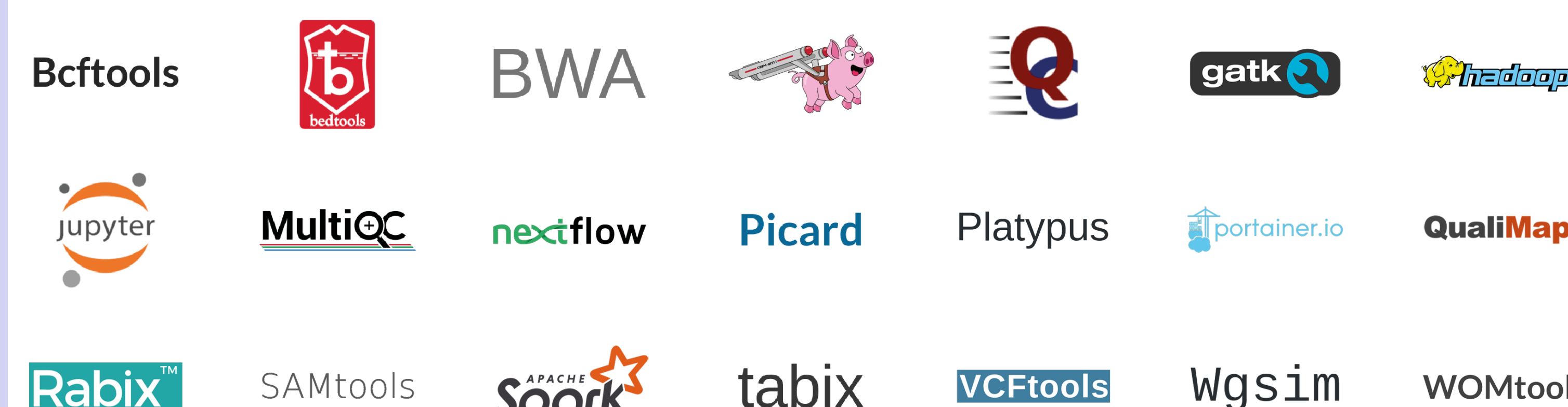


Figure 3. Dockerized software

There is an interface also available where the different services provided by the dockerized tools are exposed and can be accessed through a web browser.

Conclusions

- We have established a Docker-based bioinformatics platform using existing equipment.
- This solution allows a non-exclusive usage; faculty students can keep using the computers inside the classroom during data processing.
- This setup can be easily used as execution environment for genomics pipelines written in languages that support Docker containers such as WDL, CWL or Nextflow (**Figure 4**).

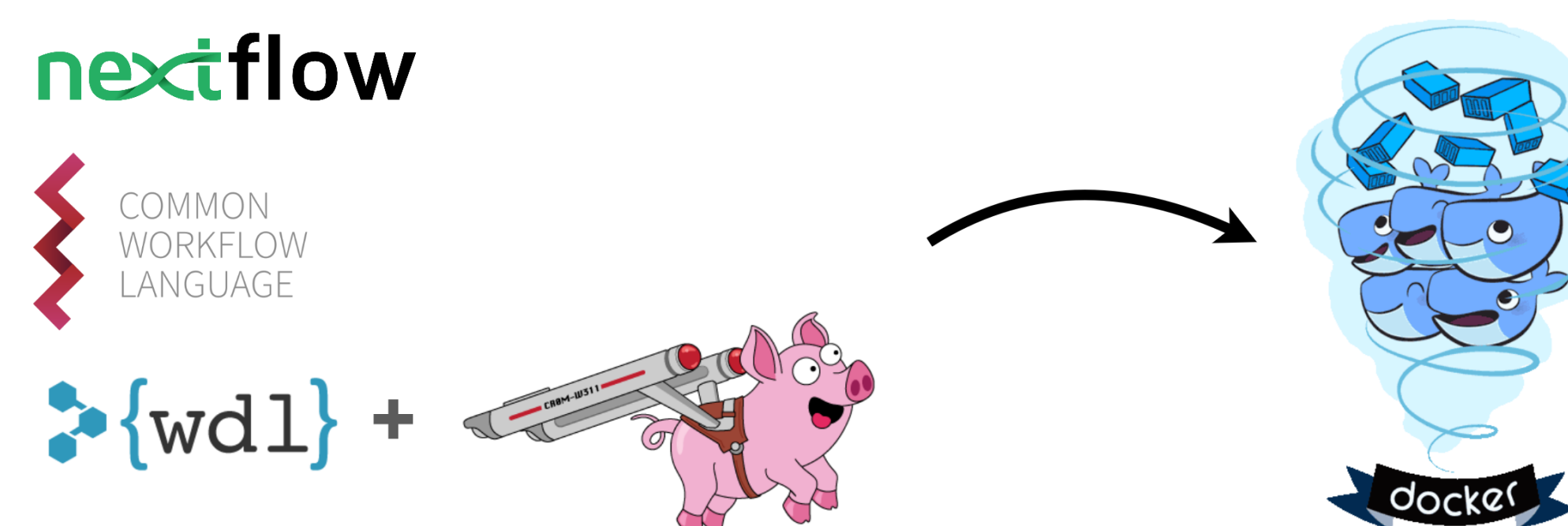
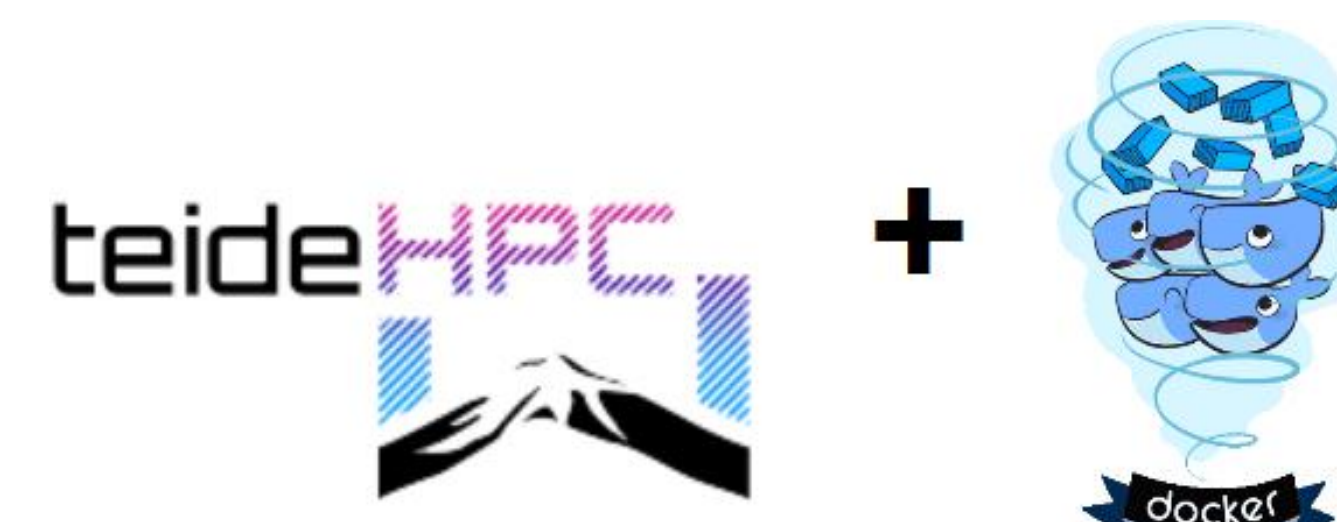


Figure 4. This Docker-based infrastructure makes it easy to build out complex deployment pipelines. See poster 19 to learn more about WDL-based pipelines development

Future work

- Implement this approach in the ITER's Teide-HPC supercomputer.



- Retrieve metrics from containers execution.
- Perform benchmarkings.

Availability

Source code and documentation



<https://github.com/lubertorubior/docker-esit>

Contact data

Contact data and poster download



Acknowledgements

Funded by Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE) and also supported by the CEDel program (Centro de Excelencia de Desarrollo e Innovación, Cabildo de Tenerife). Special thanks to the TARO research group at Universidad de La Laguna for their guidance and support.



References

1. Z.D. Stephens et al., *PLoS Biology*, vol. 13, pp. e1002195, 2015.
2. D. Merkel, *Linux J.*, vol. 2014, mar. 2014.
3. K. Shvachko et al., *Proc. Mass Storage Syst. Technol. (MSSST)*, 2010, pp. 1–10
4. M. Zaharia et al., *Commun ACM*, vol. 59, pp. 56–65, oct. 2016.
5. G.A. Auwera et al., *Current Protocols in Bioinformatics*, vol. 43, pp. 11.10.33, 2013.
6. N. Tucci et al., *arXiv preprint arXiv:1806.00788*, 2018.