DOCKERPEDIA: A KNOWLEDGE GRAPH OF DOCKER IMAGES

Maximiliano Osorio
Universidad Técnica Federico Santa María, Chile
mosorio@inf.utfsm.cl

Carlos Buil-Aranda
Universidad Técnica Federico Santa María, Chile
cbuil@inf.utfsm.cl

Hernán Vargas
Universidad Técnica Federico Santa María, Chile
hvargas@inf.utfsm.cl

Docker

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another

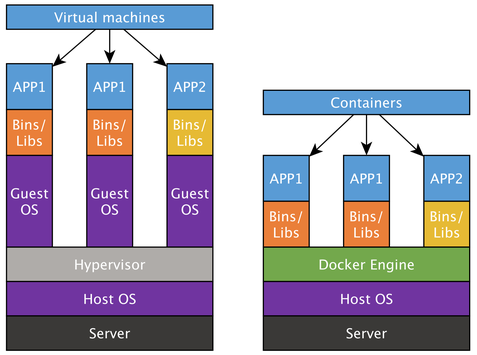


Figure 1: Containers vs. Virtual Machines (VMs): What's the Difference?

Containers share the machine's OS system kernel and therefore do not require an OS per application

Docker images are black boxes

Docker does not control what packages are in the images, whether the image will deploy correctly or the images might have any security problem.

> **Daniel Walsh, Consulting Engineer at Red Hat said**
>
> Docker is about running **random crap** from the internet as root on your host

DockerPedia is an RDF linked dataset that stores the information about Docker images hosted in Docker Hub including 4.5 million of images, its layers and packages. We also provide vulnerability analysis of these packages obtained with Clair.

Use cases: Security analysis

Our proposal allows creating a visualization tool, which uses the data available at DockerPedia, to help Docker users search and compare between different Docker images, allowing them to find software distributions which fit their needs and monitor the state of Docker repositories over time.

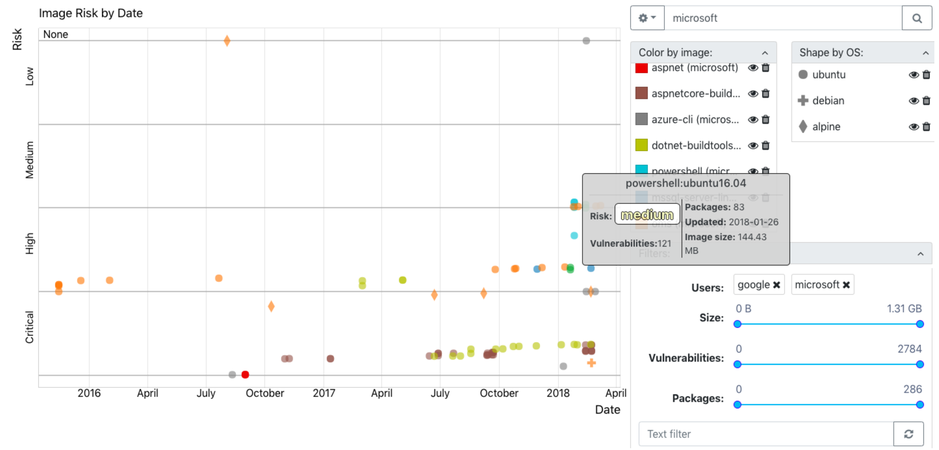


Figure 2: Risk of user images

jena-fuseki

Apache Jena Fuseki 2: SPARQL 1.1 server with web UI, backed by Apache Jena's TDB triple store

D

Label:
Last updated: 2017-10-09
Installed packages: 35
Image size: 68 MB

Vulnerabilities:

High (4)

libxml2	2.9.4-r3	CVE-2017-16931
curl	7.55.0-r0	CVE-2017-8817
curl	7.55.0-r0	CVE-2017-8816
curl	7.55.0-r0	CVE-2017-8818

Medium (9)

Figure 3: Image vulnerabilities: jena-fuseki

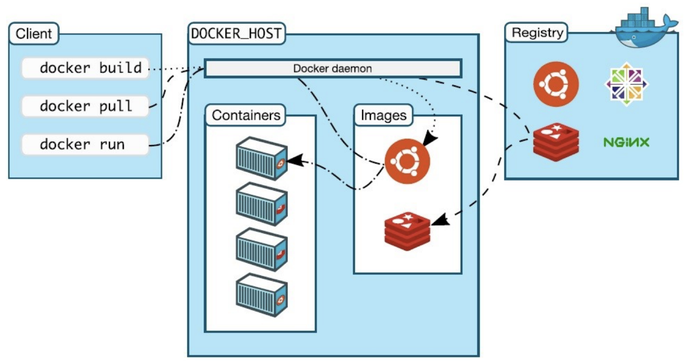
DockerHub

Anyone has the chance to create and store images into the Docker Hub registry by first creating a descriptor file called Dockerfile.
This descriptor describes what software packages will be within the image, builds the image and finally uploads it to Docker Hub.

Dockerfile	Layers
FROM node:8-slim	71150592d86e
ENV dir /var/www/ldf-server	0c3fc3ea9a44
ADD . ${dir}	446b26d60eba
RUN apt-get update && \ apt-get install -y g++ make python && \ cd ${dir} && npm install && \ apt-get remove -y g++ make python && apt-get autoremove -y && \ rm -rf /var/cache/apt/archives	ffc8d437eea6
EXPOSE 3000	9a9f8e67172c
WORKDIR ${dir}	6d413323fa27
ENTRYPOINT ["node", "bin/ldf-server"]	e71ebff76d8a
CMD ["--help"]	ef9807c35b51

Figure 4: Layers of a image

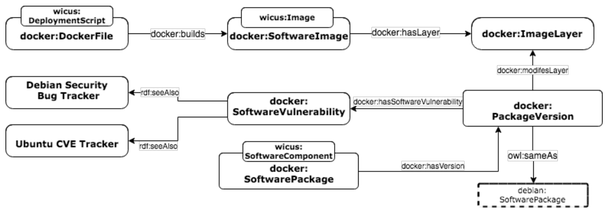
Docker Hub is an online registry for Docker image repositories. Currently stores more than 4.5 million images in two types of repositories: official and community. Official repositories contain verified images such as *Nginx, Red Hat* and *Docker*. Community repositories are not verified but can be created by any user or organization.



Our work

To extract the information from the Docker images hosted in Docker Hub, we performed a search over its free text box to obtain all the Docker images.
In February 2018, this search returned 1,363,510 Docker repositories and 4,608,443 images composed of 4,593,602 community images and 14,841 official images. The total size of these images is 53.47 PB.
After that, we use the tool from the Clair project to detect vulnerabilities within each downloaded image.

The ontology first imports abstract classes and relations from the Docker Ontology [1] for some of the Docker concepts and the WICUS ontology [2] for the software experiment reproducibility concepts.



Details about the dataset

Docker images	4,608,443	Images on DockerHub
Docker images	157,632	Images analyzed at DockerPedia
Number of packages	44.194	Packages analyzed at DockerPedia
Links to Debian	13.136	Links to the Debian dataset
Triples	>100,000,000	Triples in the dataset

Fig. 1: Details about the dataset: number of triples

Acknowledgements

This work was supported by Fondecyt Project 11170714, *Instituto Milenio de Investigación sobre los Fundamentos de los Datos* and with the support of the *Dirección de Postgrado y Programas*, UTFSM, Chile.

References

[1] Da Huo, Jaroslaw Nabrzyski, and Charles Vardeman. "Smart Container: an Ontology Towards Conceptualizing Docker." In: *International Semantic Web Conference (Posters & Demos)*. 2015.

[2] Idafen Santana-Perez et al. "Reproducibility of execution environments in computational science using Semantics and Clouds". In: *Future Generation Computer Systems* 67 (2017), pp. 354–367.