EasyMapReduce: Leverage Docker and Spark to Scale Any Data Processing Tool in MapReduce Fashion

UPPSALA UNIVERSITET

Marco Capuccini[a,b,*], Ola Spjuth[a]

Challenges

- **Increasing size of datasets** is challenging for existing data processing tools
- **Distributed computing is hard**, and the effort of reimplementing each tool in a workflow is not sustainable

➔ **Need for a generic way to scale existing software**

Methods

The next-generation MapReduce framework: a cluster computing engine for the processing of large-scale datasets [1].

The leading containerization platform: it allows to wrap software stacks, avoiding virtualization. It assures that the analysis will always run the same [2].

Results

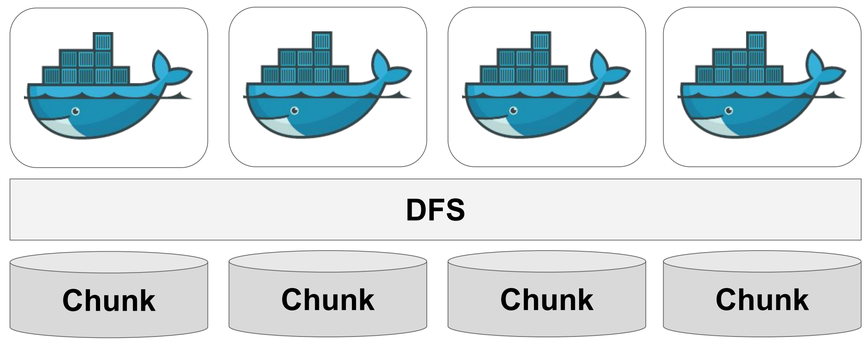
We implemented **EasyMapReduce** to provide the means to run existing serial tools in MapReduce fashion.

Fork me on GitHub

How it works

Wrap your tool with Docker

Submit to a Spark cluster



Locality-aware scheduling and fault tolerance are for free

Example: DNA GC count

```
spark-submit --class se.uu.farmbio.easymr.EasyMap \
  --master spark://cluster-endpoint \
  easymr-0.0.1.jar \
  --imageName ubuntu:14.04 \
  --command \
    'cat /input | fold -1 | grep [gc] | wc -l > /output' \
  /input/path/dna.txt /output/path/count_by_line.txt \
  --trimCommandOutput
```

```
spark-submit --class se.uu.farmbio.easymr.EasyReduce \
  --master spark://cluster-endpoint \
  easymr-0.0.1.jar \
  --imageName ubuntu:14.04 \
  --command \
    'expr $(cat /input1) + $(cat /input2) > /output' \
  /input/path/count_by_line.txt /output/path/sum.txt \
  --trimCommandOutput
```

[1] Spark: Cluster Computing with Working Sets. Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. HotCloud 2010.
[2] Docker and the Three Ways of DevOps. John Willis, Docker. Retrieved from: https://goo.gl/8WBsXk (on date 2016-10-10).

a. Department of Pharmaceutical Biosciences, Uppsala University, Sweden
b. Department of Information Technology, Uppsala University, Sweden
* marco.capuccini@it.uu.se

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654241