TOWARDS Using Spark, PRAGMA Cloud, Deep Learning, and Virtualized GPUs to Find Museum Specimens Contaminated with Mercury Salts

Matthew Collins, Nadya Williams, Gaurav Yeole
PRAGMA34 - Tokyo, Japan

iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.

Deep learning problems in biodiversity

A Few Uses:

- Taxonomic identification (1)
- Morphometric analysis (2)
- Characterization (3)

Challenge: Few biologists have access and skills in deep learning and AI

(1) https://link.springer.com/article/10.1186/s12862-017-1014-z
(2) http://www.bioone.org/doi/abs/10.1666/08068.1
(3) https://blogs.nvidia.com/blog/2018/02/22/ai-for-biodiversity-informatics/

Mercury staining of herbarium sheets

An old method of preserving plants pressed on paper sheet in museums was mercuric chloride

Stained (right) and unstained herbarium sheets

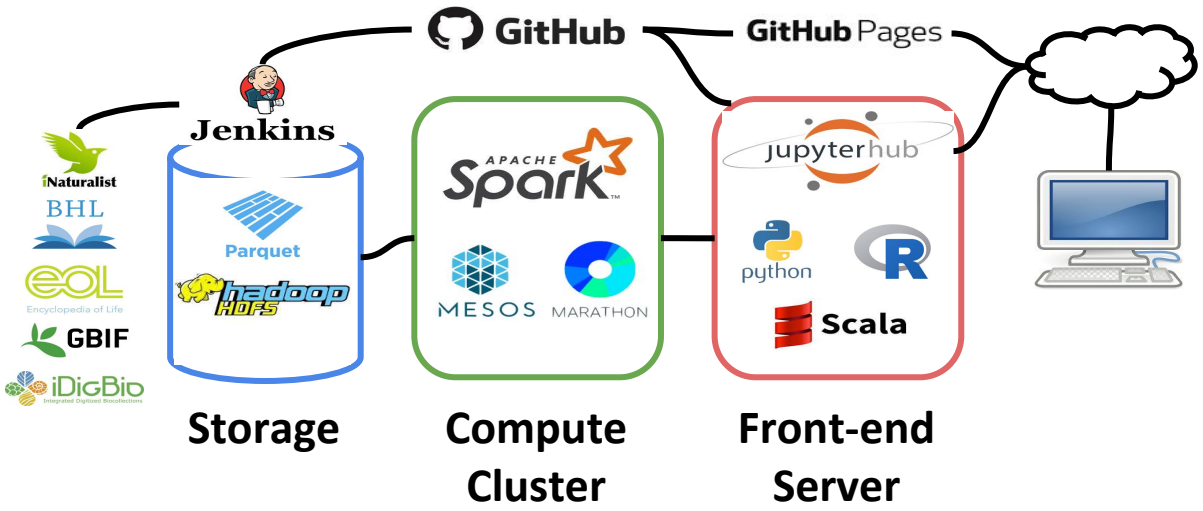
The Smithsonian Institution's paper

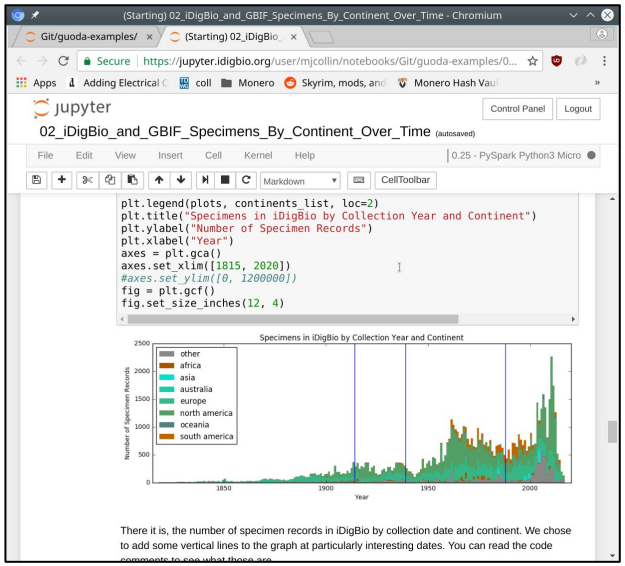
Data scientists and curators at the Smithsonian Institution found they could detect mercury staining with a deep learning model:

Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, Dorr L (2017) Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal 5: e21139. https://doi.org/10.3897/BDJ.5.e21139

Jupyter notebooks make resources available to researchers

Web-based programming environment that can execute code on **remote infrastructure**

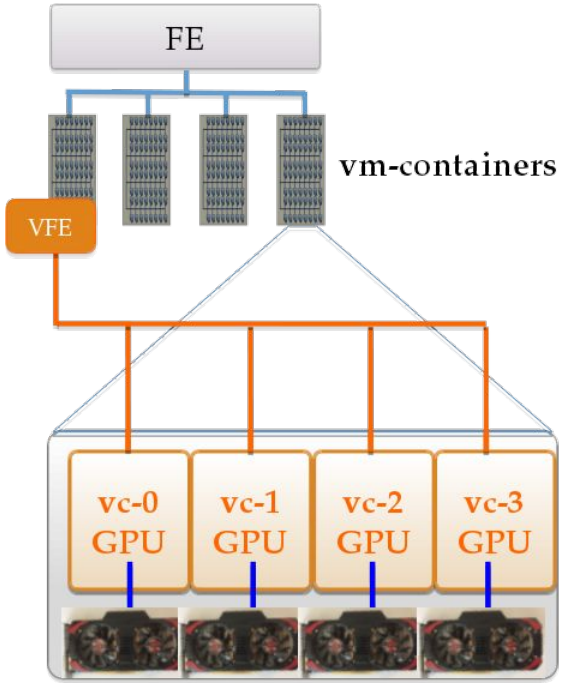




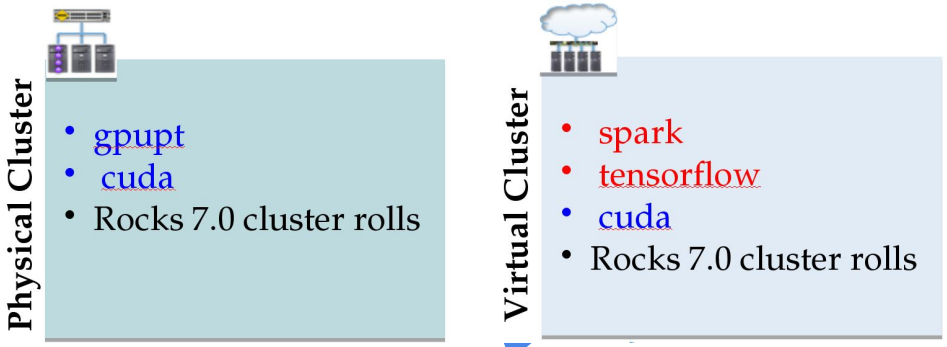
http://guoda.bio/jupyter

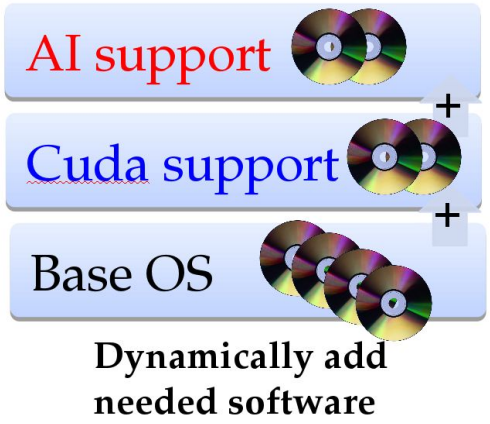
GPU-based virtual cluster setup

- On a physical host build rolls: cuda, gpupt, tensorflow, spark
- Prepare physical host for GPU pass-through
  - Add GPU cards: GeForce GTX 1060 (consumer grade), Tesla C2075, Tesla K20x
  - Enable VT-D extensions in BIOS
  - Activate Vt-d extensions in the kernel
  - “Attach” GPUs to the vm-container (gpupt roll)
  - Verify GPU cards work with (cuda roll)
  - “Detach” GPUs from the VM-container (gpupt roll)
- Create a virtual cluster
  - Run virtual FE anywhere (non-GPU node)
  - Run virtual compute nodes on GPU-enabled vm-container
  - Connect GPU to a virtual machine through the hypervisor and allocate a full GPU capability



Software architecture

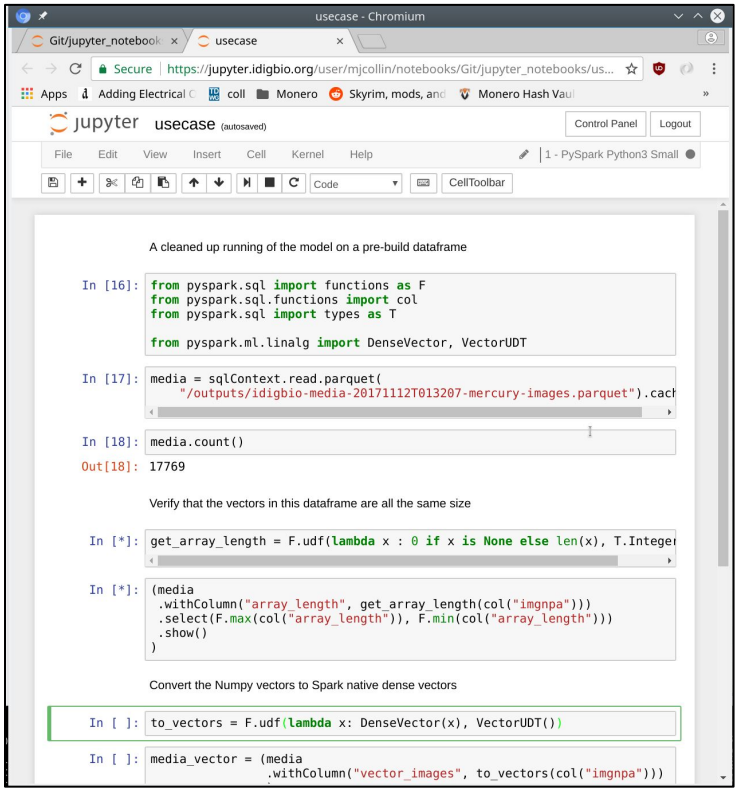




- cuda https://github.com/nbcrrolls/cuda
  - NVIDIA drivers 384.90
  - cuda toolkit 8.0.61
- gpupt https://github.com/pragmagrid/gpupt
  - Commands to manipulate GPU assignment (attach, detach, list info, status):
  - Provide correct parameters from PCI bus to the hypervisor (address, name)
  - Enables creation of a portion of VM xml description for the GPU
  - Recompile qemu-kvm that understands –enable-kvm flag
- spark https://github.com/pragmagrid/spark
  - Apache Spark analytics engine
- tensorflow https://github.com/pragmagrid/tensorflow
  - cuDNN
  - Tensorflow …

Development environment

- **Python 3.6**
- **Virtualenv** - library management
- **Spark** - distributed computing engine
- **Tensorflow** - deep learning library
- **Keras** - higher-level deep learning on top of Tensorflow
- **Elephas** - distributed Keras on Spark
- **Parquet** - columnar data storage format



Demo

https://github.com/acislab/pragma-cloud-spark

www.idigbio.org

SDSC SAN DIEGO SUPERCOMPUTER CENTER

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All images used with permission or are free from copyright.