Brown Dog: An Elastic Data Cyberinfrastrure for Autocuration and Digital Preservation

Jay Alameda
National Center for Supercomputing Applications
University of Illinois at Urbana-Channpaign

Acknowledgements

The Brown Dog team & coauthors:

Smruti Padhy, Jay Alameda, Rui Liu, Edgar Black, Liana Diesendruck, Mike Dietze, Greg Jansen, Praveen Kumar, Rob Kooper, Jong Lee, Richard Marciano, Luigi Marini, Dave Mattson, Barbara Minsker, Chris Navarro, Marcus Slavenas, William Sullivan, Jason Votava, Inna Zharnitsky, Kenton McHenry

This material is based upon work supported by the National Science Foundation under Grant Number NSF ACI-1261582: “CIF21 DIBBs: Brown Dog”

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.

The Problem

- Large collections of **unstructured and/or un-curated** data
- Many data types and file formats
- Variety of existing software
- Short life span of digital data and software
- Hinders reproducibility of scientific results

An Example: Ecosystems and Climate Change

M. Dietze, K. McHenry, A. Desai, "Model-data Synthesis and Forecasting Across the Upper Midwest: Partitioning Uncertainty and Environmental Heterogeneity in Ecosystem Carbon," NSF DBI-1062547, 2011-2014

M. Dietze, K. McHenry, A. Desai, "ABI Development: The *PEcAn* Project - A Community Platform for Ecological Forecasting," NSF DBI-1457890, 2015-2019

- Towards regional-scale high resolution estimates of plant life and carbon storage
- Scientific workflow and data assimilation system connecting a variety of models within the Ecology community to a variety of data sources
- Grown to 52 developers over the past 3 years
  - NCSA / U. Illinois, BU, Brookhaven National Lab, University of Wisconsin, University of Notre Dame, Utah State, Columbia University, Pacific Northwest National Laboratory, DuPont Pioneer, Exeter College, UK, U. Arizona, Dartmouth College

Ecosystems and Climate Change

- Models:
  - Ecosystem Demography (ED)
  - SIPNET
  - DALEC
  - …
- Data:
  - Biofuel Ecophysiological Trait and Yield Database (BETY)
  - Forest Inventory and Analysis (FIA)
  - North American Regional Reanalysis (NARR)
  - North American Carbon Program (NACP)
  - Food and Agriculture Organization (FAO)
  - …

Ecosystems and Climate Change

- Data with Unstructured Aspects:
  - MODIS (Multi-spectral)
  - Lidar
  - Palsar (Radar)
  - Aviris (Airborne Infrared Spectrometer)
  - Landsat (Images)

  - Published results (e.g. tables, figures, plots)
    - Manually done to ingest into BETY

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data
  - Document

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image
- Spatial

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image
- Spatial
- Tabular

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image
- Spatial
- Tabular
- Weather

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image
- Spatial
- Tabular
- Weather
- 3D

Ecosystems and Climate Change

- Settlement Vegetation data
- Born Physical
  - Paper, Microfiche, Alphanumeric/Color coded on vellum sheets
- Born Digital
  - PDF, JPEG, GIF, TIFF, XLS, XLSX, CSV, SHP, netCDF, HDF5, XML, GRIB, GRIB2, geoTIFF, DBF, BIL, BIP, ARC, SDTS, SRTM, IMG, UA, LGW, SXW, ODS
  - Ad hoc formats:
    - Spreadsheets
    - Databases
    - Services
    - R Data
    - Matlab Data

- Document
- Image
- Spatial
- Tabular
- Weather
- 3D
- Archive, Database, Filesystem, …

What we need

A system/framework that

- Enables access to data contents irrespective of file formats
- Extracts metadata from data content and does automatic curation
- Uses existing conversion/extraction/data analysis tools
- Is extensible – easily add new tools
- Is dynamically scalable
- Is easy to use

CIF21 DIBBs: Brown Dog

- PI: Kenton McHenry, Ph.D.
- Co-PI: Jong Lee, Ph.D.
- Co-PI: Barbara Minsker, Ph.D.
- Co-PI: Praveen Kumar, Ph.D.
- Co-PI: Michael Dietze, Ph.D.

Brown Dog – A framework for autocuration

- Data Access Proxy (DAP)
  - File format conversions
  - Example – png to pdf
- Data Tilling Service (DTS)
  - Extraction of metadata, signatures or derived products from a file's content
  - Example – Face extraction, text extraction using OCR, table from pdf, previews
- Tools Catalog (TC)
  - Allows to add new conversion/extraction tools to the DAP/DTS
- Elasticity Module (EM)
  - Scales Up/Down DAP/DTS

DAP DTS TC EM

Data Access Proxy (Data format conversion)

- REST API
- Largely Reversible
- Software Servers
  - 3rd party software, library, external service
- Wrapper Scripts (Converters)

```
#Application name (Version)
#File types supported (e.g. document, depth, image, …)
#Comma separated list of supported input formats
#Comma separated list of supported output formats
```

Describe

```
#Call external application and/or carry out conversion
…
```

Convert File

Adding Converters to Software Server within DAP

Example

```
#!/bin/sh
#ImageMagick (v6.5.2)
#image
#bmp, dib, eps, fig, gif, ico, jpg, jpeg, jp2, pcd, pdf, pgm,
pict, pix, png, pnm, ppm, ps, rgb, rgba, sgi, sun, svg, tga,
tif, tiff, ttf, x, xbm, xcf, xpm, xwd, yuv
#bmp, dib, eps, gif, jpg, jpeg, jp2, pcd, pdf, pgm, pict,
png, pnm, ppm, ps, rgb, rgba, sgi, sun, svg, tga, tif, tiff,
ttf, x, xbm, xpm, xwd, yuv

output_filename=$(basename "$2")
output_format="${output_filename##*.}"

#Output PGM files as ASCII
if [ "$output_format" = "pgm" ]; then
        convert "$1" -compress none "$2"
else
        convert "$1" "$2"
fi
```

Data Tilling Service (Metadata Extraction)

- REST API
- Extractors
- Use any existing tool
- Python library - pyClowder

```
extractors.connect_message_bus(extractorName=extractorName,
                               messageType=messageType,
                               rabbitmqURL=rabbitmqURL,
                               rabbitmqExchange=rabbitmqExchange,
                               processFileFunction=process_file,
                               checkMessageFunction=check_message)
```

Connect

```
def process_file(parameters):
    global extractorName
    inputfile=parameters['inputfile']

    # call actual program
    result = subprocess.check_output(['wc', inputfile], stderr=subprocess.STDOUT)
    (lines, words, characters, filename) = result.split()
```

Work on File

```
extractors.upload_file_metadata(mdata=metadata,
parameters=parameters)
```

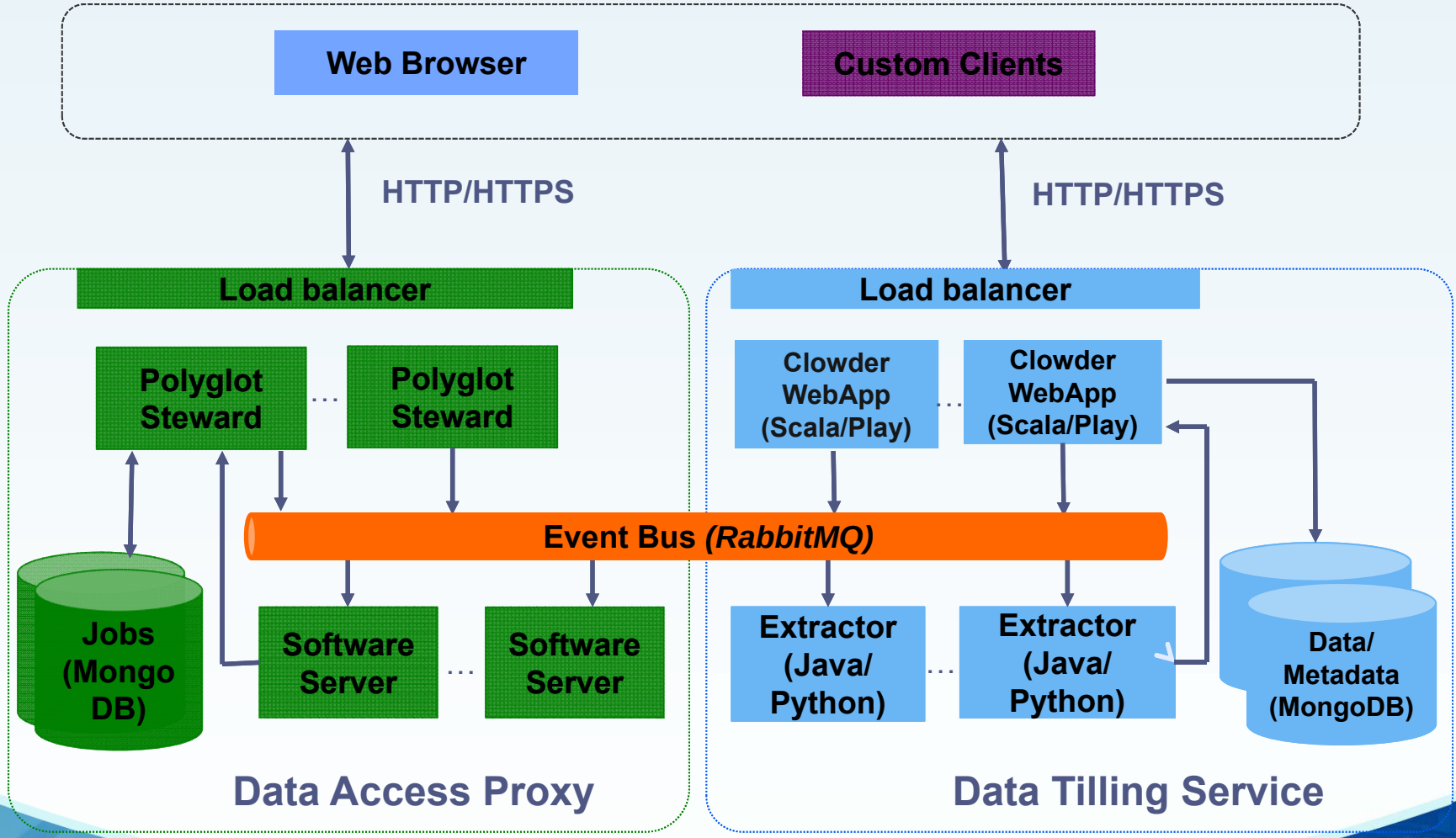
Return Metadata

Creating a Python extractor using pyClowder for DTS

Example

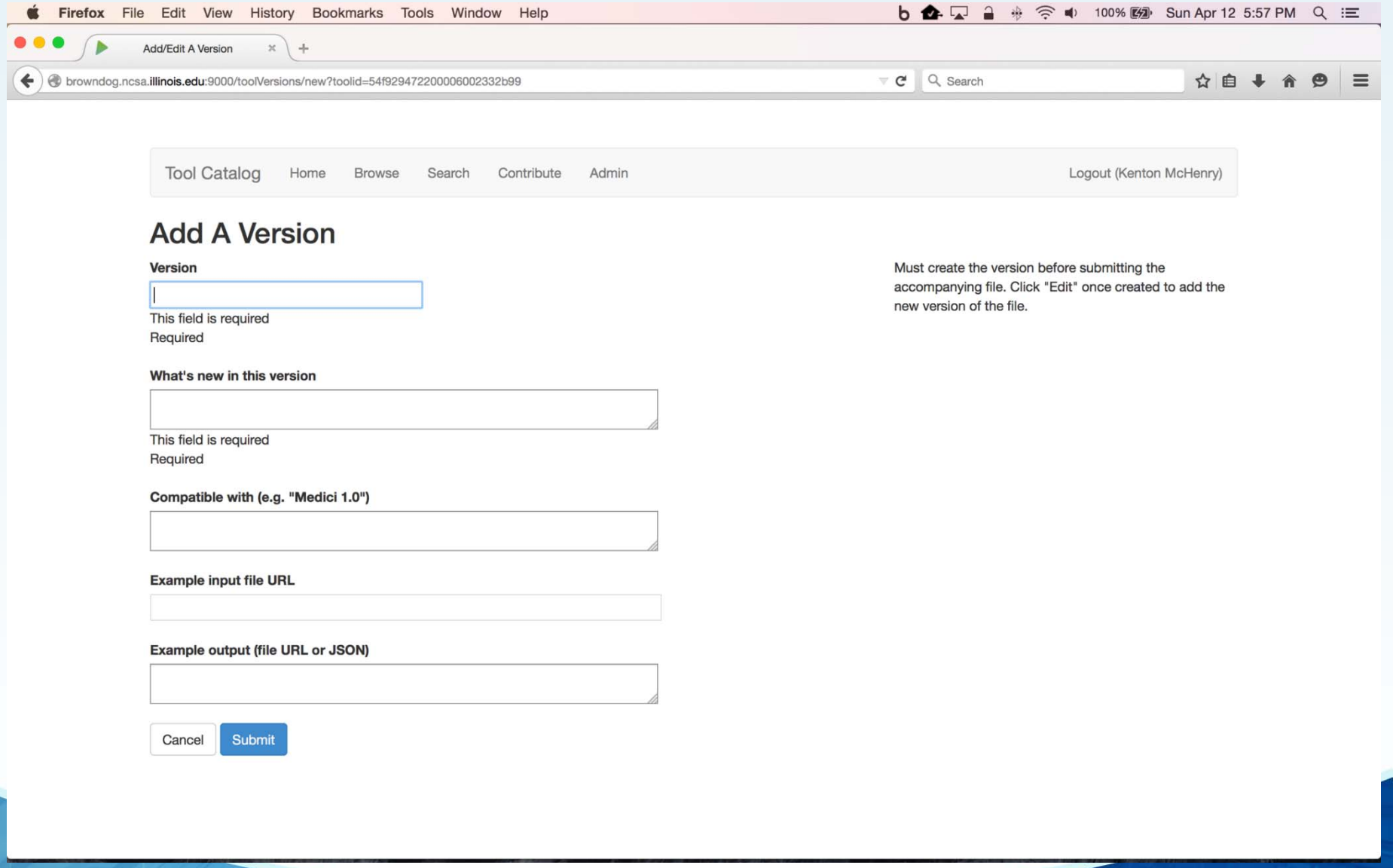
```
#!/usr/bin/env python
import subprocess
import logging
from config import *
import pymedici.extractors as extractors

def main():
    global extractorName, messageType, rabbitmqExchange, rabbitmqURL

    #set logging
    logging.basicConfig(format='%(levelname)-7s : %(name)s -  %(message)s', level=logging.WARN)
    logging.getLogger('pymedici.extractors').setLevel(logging.INFO)

    #connect to rabbitmq
    extractors.connect_message_bus(extractorName=extractorName, messageType=messageType, processFileFunction=process_file,
        rabbitmqExchange=rabbitmqExchange, rabbitmqURL=rabbitmqURL)

# ----------------------------------------------------------------------
# Process the file and upload the results
def process_file(parameters):
    global extractorName

    inputfile=parameters['inputfile']

    # call actual program
    result = subprocess.check_output(['wc', inputfile], stderr=subprocess.STDOUT)
    (lines, words, characters, filename) = result.split()

    # store results as metadata
    metadata={}
    metadata["extractor_id"]=extractorName
    metadata['lines']=lines
    metadata['words']=words
    metadata['characters']=characters

    # upload metadata
    extractors.upload_file_metadata(mdata=metadata, parameters=parameters)

if __name__ == "__main__":
    main()
```

The Brown Dog Services Architecture



DAP DTS TC EM

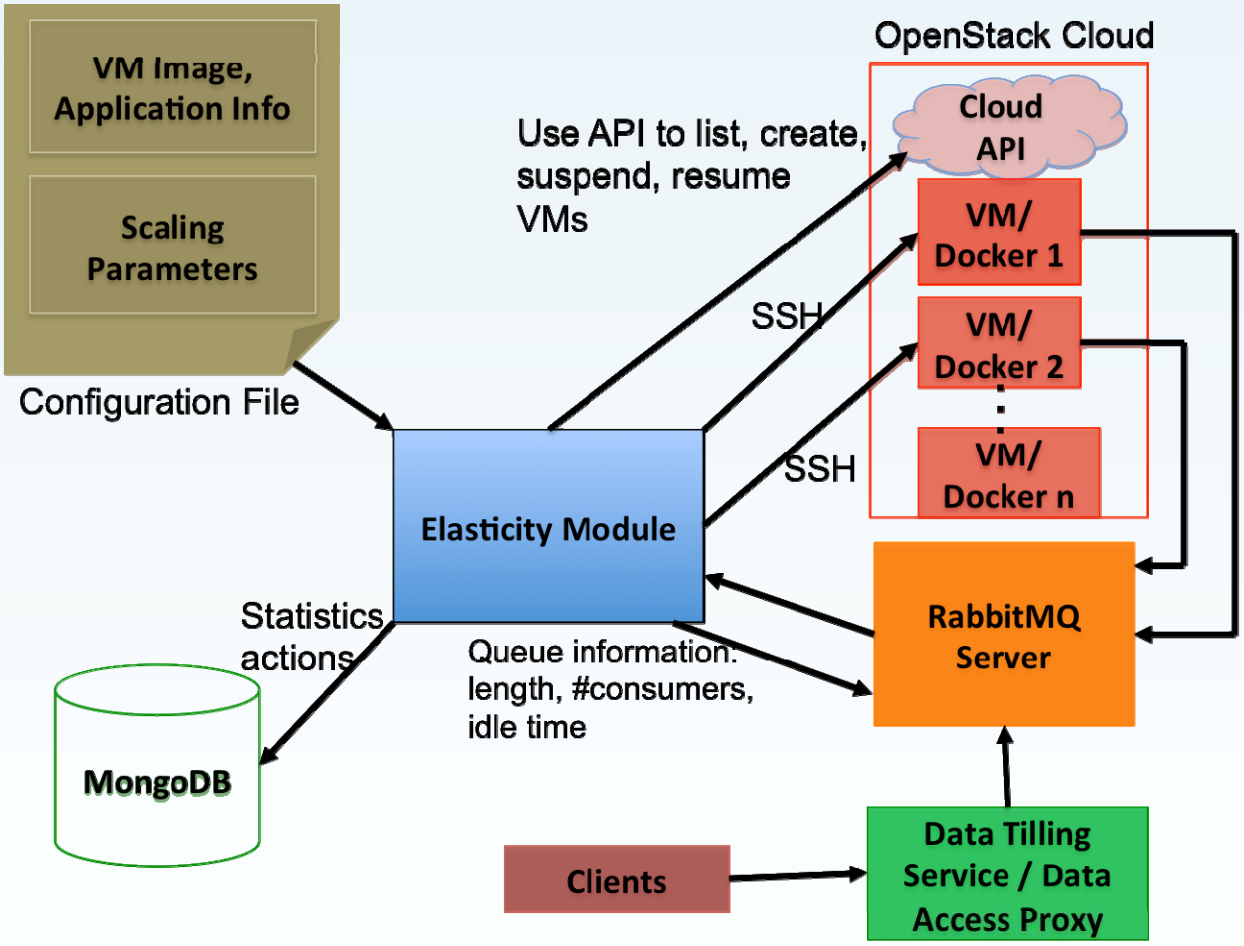
Tools Catalog



Elasticity

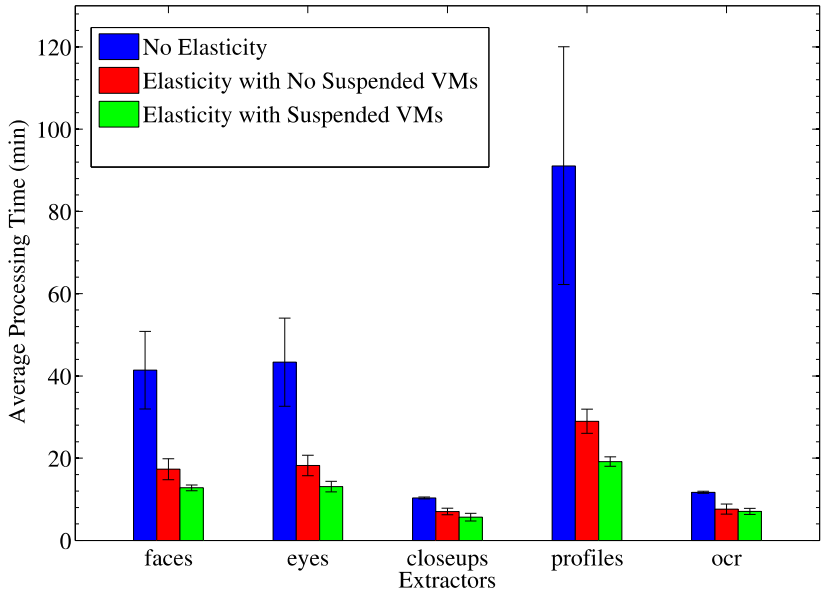
- Automatically scales up/down DAP/DTS based on the user demands
- Leverages cloud computing IaaS
- Supports a variety virtual machine/container frameworks
- Leverages HPC resources to batch execute jobs in long queues
- Focuses on DTS extractors and DAP Software Servers

Elasticity Module Architecture



Elasticity – Performance Evaluation

- Tested with Open CV (Computer Vision) extractors
  - faces, eyes, profiles and closeups
- Tested with OCR extractor with around 1200 test images



Processing time is reduced by 70% and 80% if started with suspended VMs

Summary

- Huge diversity in data and analysis
- Programmable Interface – various client applications
- Automatically scales up/down
- Place to preserve/reuse software/tools
- Integrable with scientific workflow system
- Resuable modules

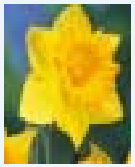
Brown Dog Services- Software Components, Cloud/HPC Resources

Clowder

Polyglot

RabbitMQ

Versus



Daffodil

mongoDB

openstack™ CLOUD SOFTWARE

NCSA Nebula
An OpenStack Cloud Resource

docker

XSEDE
Extreme Science and Engineering Discovery Environment

Project website:
http://browndog.ncsa.illinois.edu/

NSF NCSA

Thank You

Questions ?

@NCSABrownDog