Analytics platform

Rob Gardner
University of Chicago

Ilija Vukotic
University of Chicago

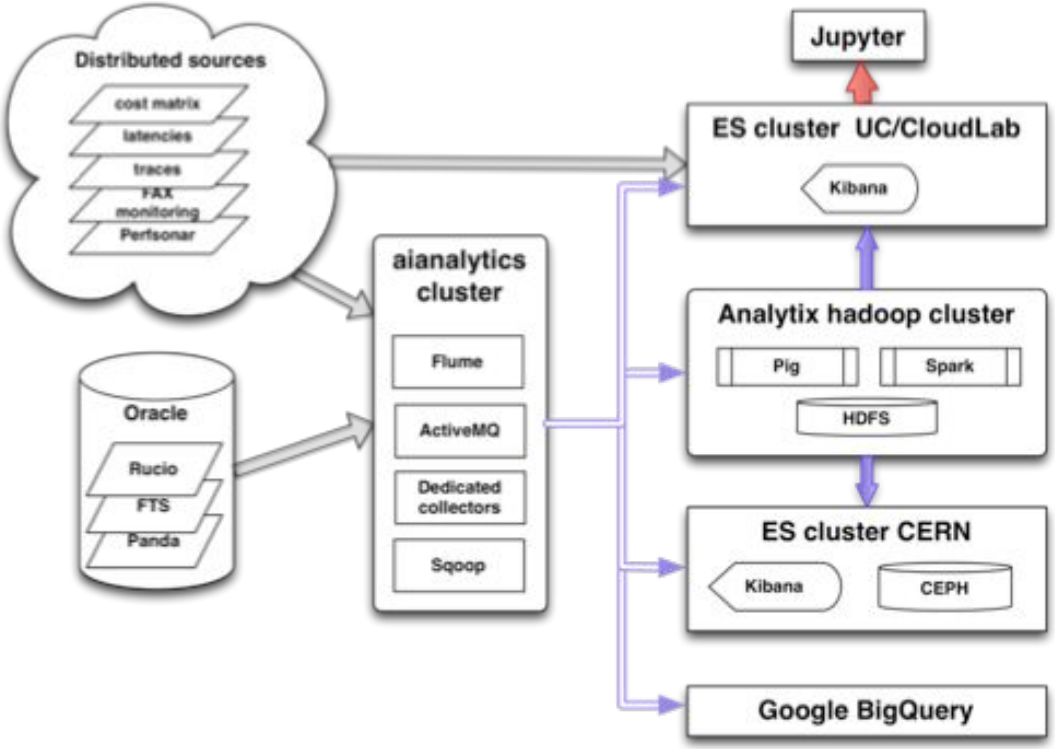
Current system

Ingress

- Cron pig scripts
- Sqoop scripts
- Flume services
- Python collectors

Storage

- 3 ES clusters
- Two Hadoop clusters
- Google BigQuery



Current situation - HDFS data

Bulk of data is also stored in HDFS

Analytix cluster is much more performant now, still people don't use it for interactive analysis. There are two ways to use it (let's forget direct java M/R):

- Pig - steep learning curve, often need user defined functions in python or java, but code easily readable
- Spark - much faster, much nicer for data preprocessing for ML methods.

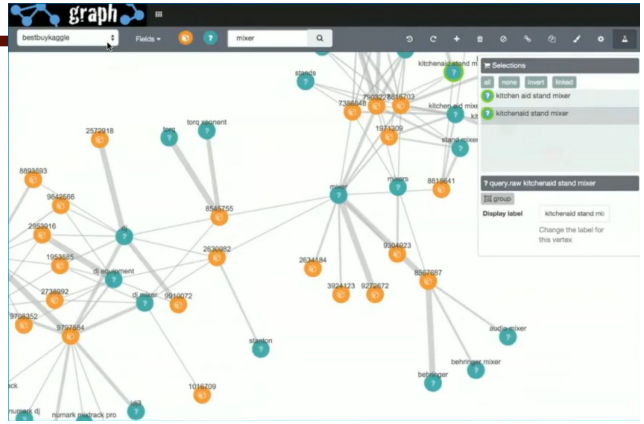
Current situation - Elasticsearch

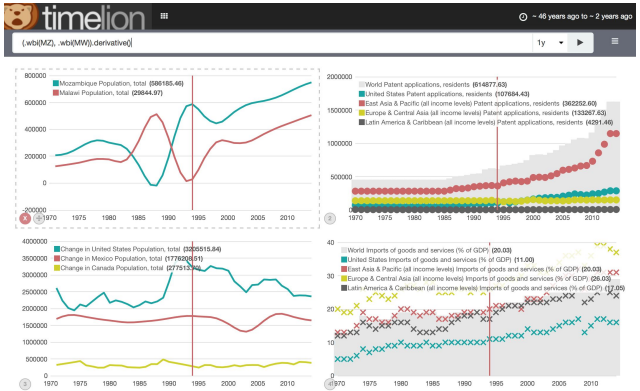
All the data in ES at UC/CloudLab. Jobs and tasks data in ES @CERN too.

Kibana is widely used, people are slowly gaining experience. New tools are being added - some of it very powerful/high level.

ES is very convenient to get the data to your code (simple or ML) due to simple but powerful API. ~15 people are using it that way.

Even more convenient is if one has resources to crunch data locally.





Google BigQuery data

Panda jobs and tasks data are exported to BigQuery - starting from May 1st - 240M records/300GB.

It is very easy and fast to make sql-like queries. Very nice REST API. Question is where should we use it and how much would it cost.

New UC analytics cluster

Brand new nodes:

- 5 data nodes (R630, 64GB RAM, 4x800 GB SSD, 10Gbps NIC)
- 3 head nodes (same config sans storage)

Will be used for WLCG and OSG networking data, FAX and StashCache monitoring, OSG job accounting, Tier2 services monitoring, tests of event and geometry indexing, etc.

In operation this Fall

Pending changes

Services are all well written so system is stable and does not require much work.

Still it would be better to make it leaner:

Move aianalytics ES to central ES @ CERN. Currently only used by Shaojun Sun for Panda Logs. Could be done in a week.

Move most of the ATLAS specific data to ES@CERN. Currently CERN cluster is not performant enough (with only jobs and tasks info going there). We communicate with Ulrich S. and Pablo S. We should not expect improvements before end of the year.

Simplify UC cluster. Currently the cluster is geographically distributed. That's both positive and negative. System is more complex but more resilient to network outages. We will drop resources from CloudLab when it becomes possible.

Services - Alarms, Alerts

One of the first services we should develop and deploy of top on the Analytics Platform.

There is a (very expensive) service (Watcher).

We need a web page showing alarms, a way for people to un/subscribe to alerts, create their own. Should quickly decide on technology to do it.

Will need some more resources to run all the alarms/alerts codes, we don't want people doing ML interfering with its operation.

Jupyter

UC - one beefed up machine with 2xGPUs, code in github. ~10 regular users.

CERN - SWAN (beta)

backend is CERNbox

Other hardware sources around: Lyon, SLAC, ...

Discussion with David R. and Michael K. on best ways to provide ML software. http://bit.ly/2aCJqUa

Even if we don't lead this effort we can gain a lot but helping establish and then use the standard.

Installed

Root numpy
Numpy
Scipy
Matplotlib
Pandas
XGboost
Scikit Learn
Scikit images
Theano
TensorFlow
Keras
H5py
BLAS / ATLAS / LAPACK libraries
Cuda and CuDNN
ZLib

Indexing physics event data

A service to make some very simple visual cuts just now went public (http://opendata.cern.ch/collection/ATLAS).

There is a CERN project trying to provide Jupyter resources to open access with EOS data as a backend.

We want to try a serious full physics event data indexing in Elasticsearch.

We need to know how does it scale in terms of data size, search performance, delivery performance.

Surely it will work good enough for outreach (through opendata, ATLASrift projects).

Will it be performant enough for ML investigations?

Last steps of physics analysis?

To be partly done for CHEP 2016

If yes could it serve as an EventService?

Persistification and delivery of the GeoModel data

GeoDB is not convenient for event viewers.

We want it :

persistified in a new way,

indexed,

searchable,

deliverable as JSON via REST interface,

maybe even tessellated.

To be partly done for CHEP 2016

This will provide something similar to google maps tiling with different Level Of Details.

Summary

Now we have:

- Experience in getting different kinds of data imported, indexed.
- Quite a bit of data collected.
- Experience in simple data analysis.

It's time to move further:

- Move support to CERN as much as possible
- Provide a platform to do advanced data analysis
- Close the feedback loop
- Add non-ADC data
- Be a data backend to different outreach platforms