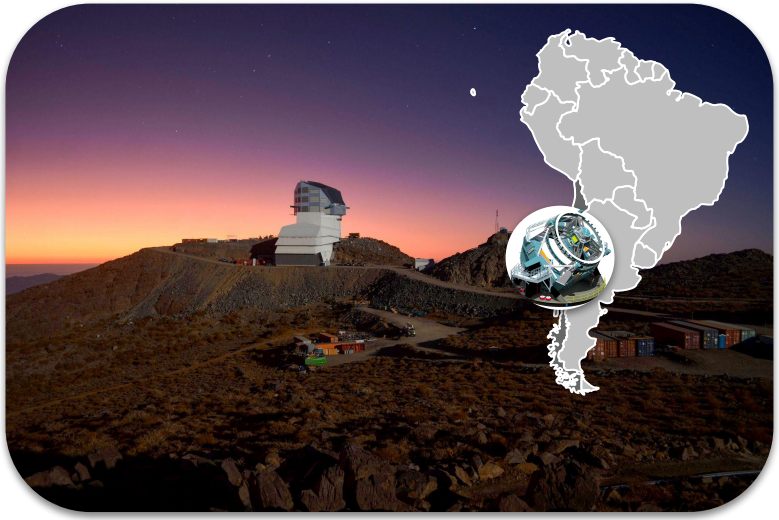
Vera Rubin Observatory

See also ET-PP/ET-EIB workshop @ Geneva

Julien Peloton (IJCLab)

15/11/2023

Rubin data products



Rubin Observatory (2025+)

- **20TB of images / night**
- **1TB of alerts / night**: x100-x1000 above current streams
- *Everything matters a priori*

Now

Raw Data
Sequential 30s image, 20TB/night

60s

Prompt Data Product
Difference Image Analysis
Alerts: up to 10 million per night

24h

Prompt Products DataBase
Images, Object and Source catalogs from DIA
Orbit catalog for ~6 million Solar System bodies

Year

Annual Data Release
Accessible via the LSST Science Platform & LSST Data Access Centers.

End

Final 10yr Data Release
Images: 5.5 million x 3.2 Gpx
Catalog: 15PB, 37 billion objects

Rubin alert system

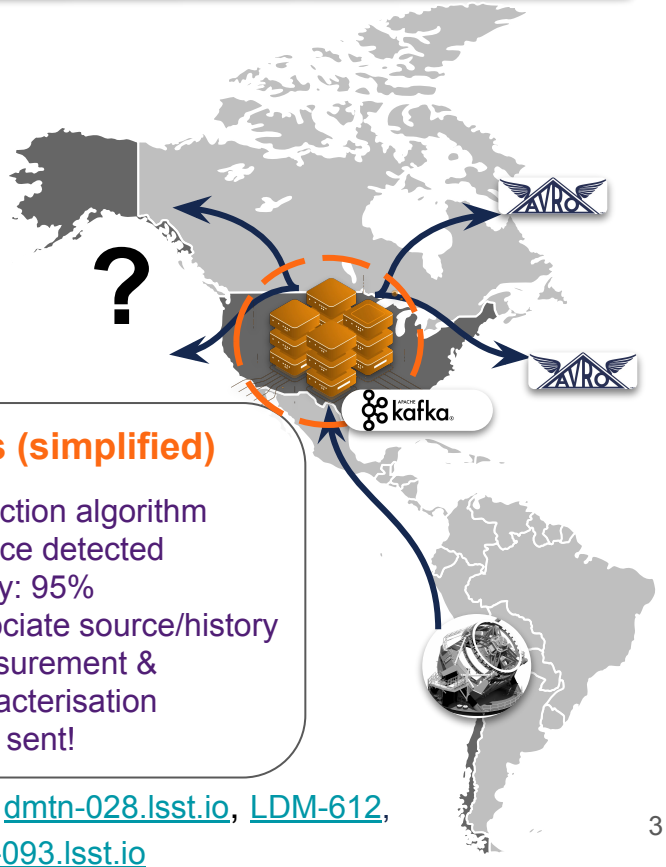
Image data sent from Chile to the USA. Alert system will identify sources that move or vary within 60 seconds.

- Sources packaged with contextual information into world-public alert packets for distribution.

Suite of open source technologies considered for distributing alerts

- Binary serialization format: Apache Avro
- Alert distribution: Apache Kafka

Prototyping on ZTF (Palomar)



Steps (simplified)

1. Detection algorithm
2. Source detected
   Purity: 95%
3. Associate source/history
4. Measurement & characterisation
5. Alert sent!

Technical notes: dmtn-028.lsst.io, LDM-612, ls.st/dpdd, dmtn-093.lsst.io

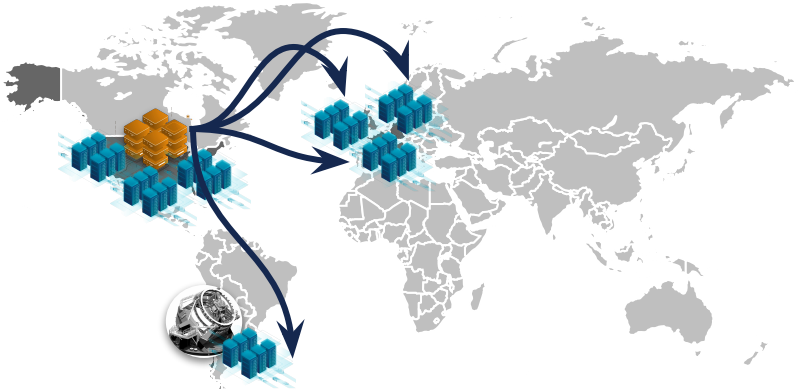
Rubin brokers

Rubin will send the full alert stream to seven brokers; others and individuals will operate downstream.

- ALERCE, AMPEL, ANTARES, Babamul, Fink, Lasair, Pitt-Google

Serve a large scientific community by ingesting, **classifying**, filtering, and redistributing alerts. Classification is a community-driven effort.

All prototyping on ZTF (300k alerts/night), and test deployment of the Rubin Alert Distribution system in the Google Cloud.



Fink: cloud-based broker

60+ members, 15+ scientific topics covered

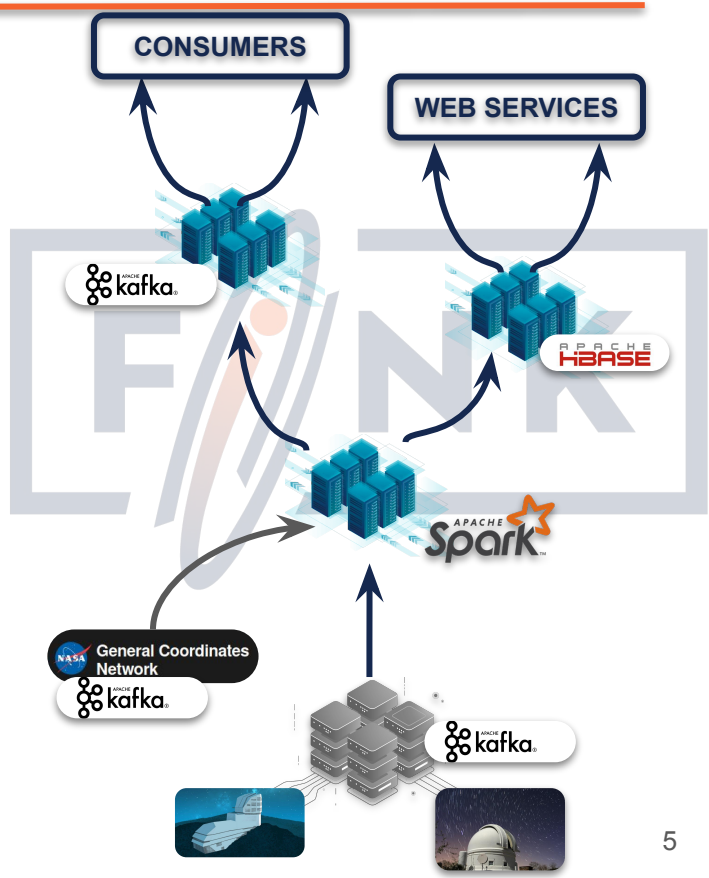
- Community-driven scientific roadmap

Services deployed on large OpenStack clouds (UPSaclay & CC-IN2P3)

- Computing (Spark), database (HBase), streaming (Kafka), storage (Ceph & HDFS)
- Autoscaling based on the load

Operating 24/7 since 2019, serving 100+ unique users per day (scientists & follow-up facilities).

Tested up to 50M alerts/night. Science database of 7TB (200M events).



Main computing challenge

As data sets become bigger and more complex, most state-of-the-art computer science tools do not benefit the scientific community at large

Domain experts are the crucial agent for scientific discoveries

- Huge legacy of codes…
- … but they rarely meet computing requirements (sorry, true story)

Stronger interplay between the computing model & user software

- Software engineering role is increasing
  - Tailored service to integrate codes developed by the community
  - Infrastructure should be created to adapt to specific user needs
- Regularly training is a key for long term sustainability

Conclusion

Low-latency challenge for Rubin alerts is dominated by the computing

- Fink scientific roadmap is defined by the community of users
- Model of computing: survey → brokers ↔ scientific community

Fink: processing is centralised, science is decentralised

- Cloud computing, with the elastic provisioning, allows to scale out resources to match the demand from the community
- Brokers provide data, computing & engineering services for the community
- Set of open source components chosen to be the backbone of the structure

Various challenges remain: user-driven & evolving analysis, open & big data, interoperability for multi-messenger & multi-wavelength analyses…

To go further

Why 60 seconds latency on the Rubin side? *see e.g. DMS-REQ-0004, LSE-61. Fiber networks Chile → USA (20TB images each 30 seconds) + DIA processing.*

Why 7 brokers? *No single team can cover all topics from variable & transient astronomy. Money-wise, perhaps easier to also integrate worldwide.*

How big is one alert? At which rate they are sent? *An alert is about 100KB. Alerts are sent to brokers by bursts of ~10,000 every 30 seconds.*

How long it takes to process one alert? *The processing time depends how many treatments we want to perform, which depends on the user needs.*

How does the users being served relate to the N events/night? *Eventually not all users want all alerts. The role of the broker is to reduce the N events/night to M (N>>M) events per night, and per science case of interest.*