



Real-time User Analysis with the Pitt-Google Alert Broker

Troy Raen

University of Pittsburgh | Pitt-Google Collaboration

Pitt-Google Alert Broker Collaboration



Troy Raen

Lead Developer
Grad Student
University of Pittsburgh

on behalf of...



Michael Wood-Vasey (PI)

Professor
University of Pittsburgh



Ross Thomson

Solutions Architect
Google



Christine Mazzola Daher

Grad Student
University of Pittsburgh



Daniel Perrefort

Research Assistant Professor
University of Pittsburgh

Pitt-Google Alert Broker



- Selected as a **full-stream alert broker for LSST**
 - Currently processing and distributing the **ZTF** alert stream
 - In early stage dev. for **LIGO / Virgo** stream
- Broker operates 100% on Google Cloud
 - Explicitly cloud-based model leverages Google-managed compute and storage services
- **Users can easily leverage Google Cloud services for their own analysis and storage** and/or move data out of the Cloud

Founding Motivations

1. Provide broad public access to the LSST alert stream
 - full stream, curated streams, database catalogs
2. Facilitate scientific collaboration with a low barrier to entry

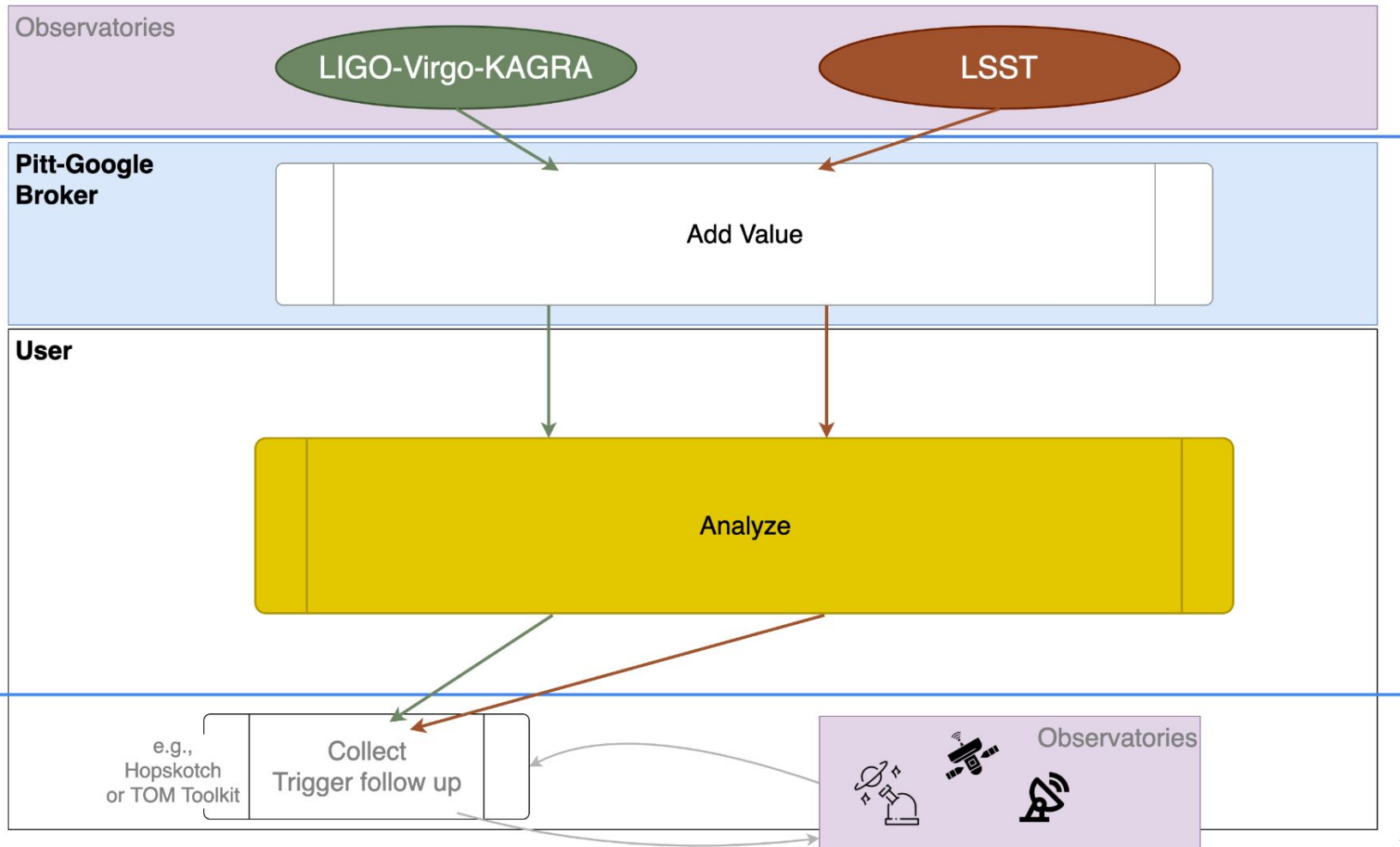
An example of:

Real-time User Analysis

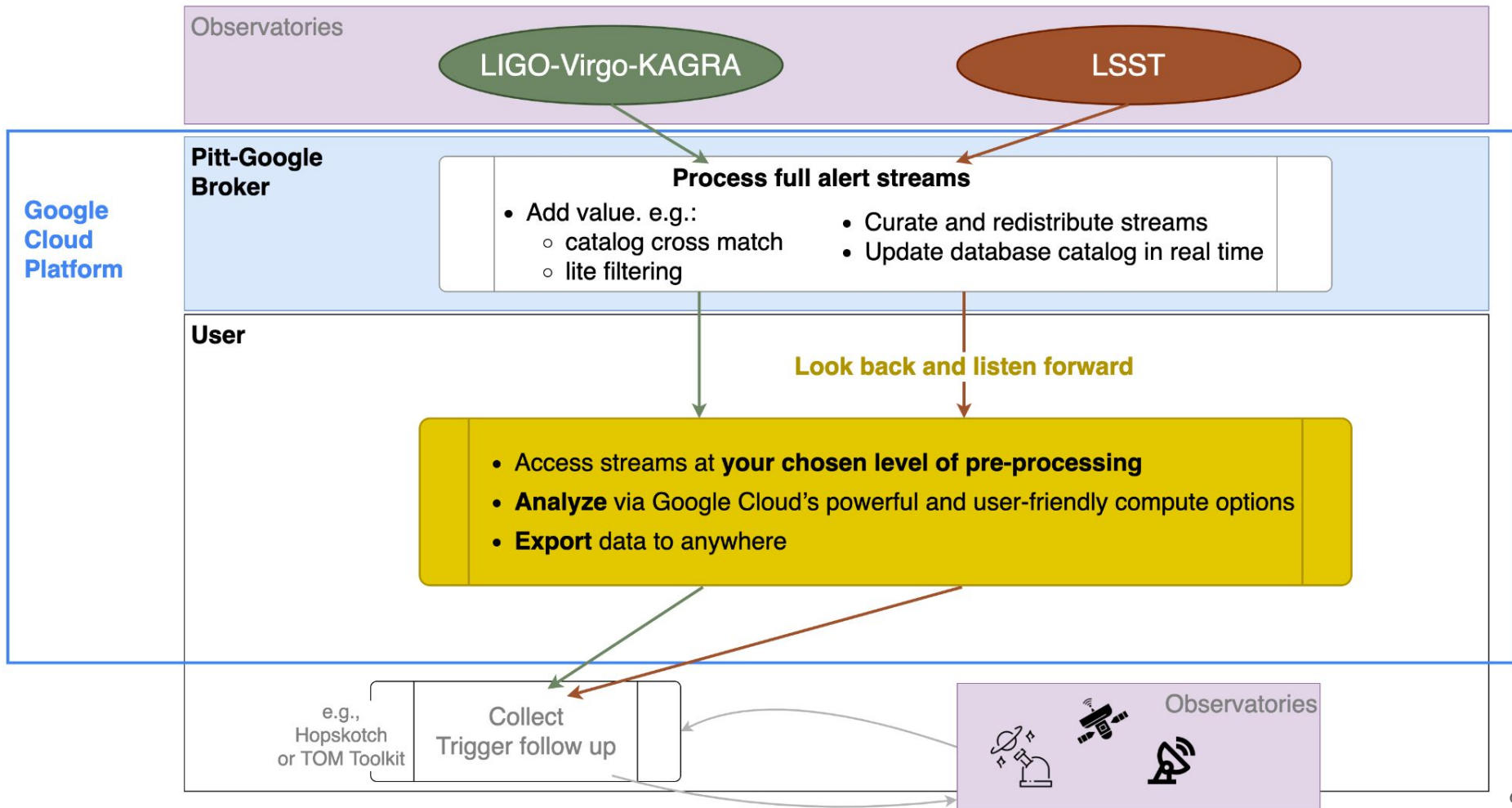
of MMA streams via

Pitt-Google broker and Google Cloud Platform

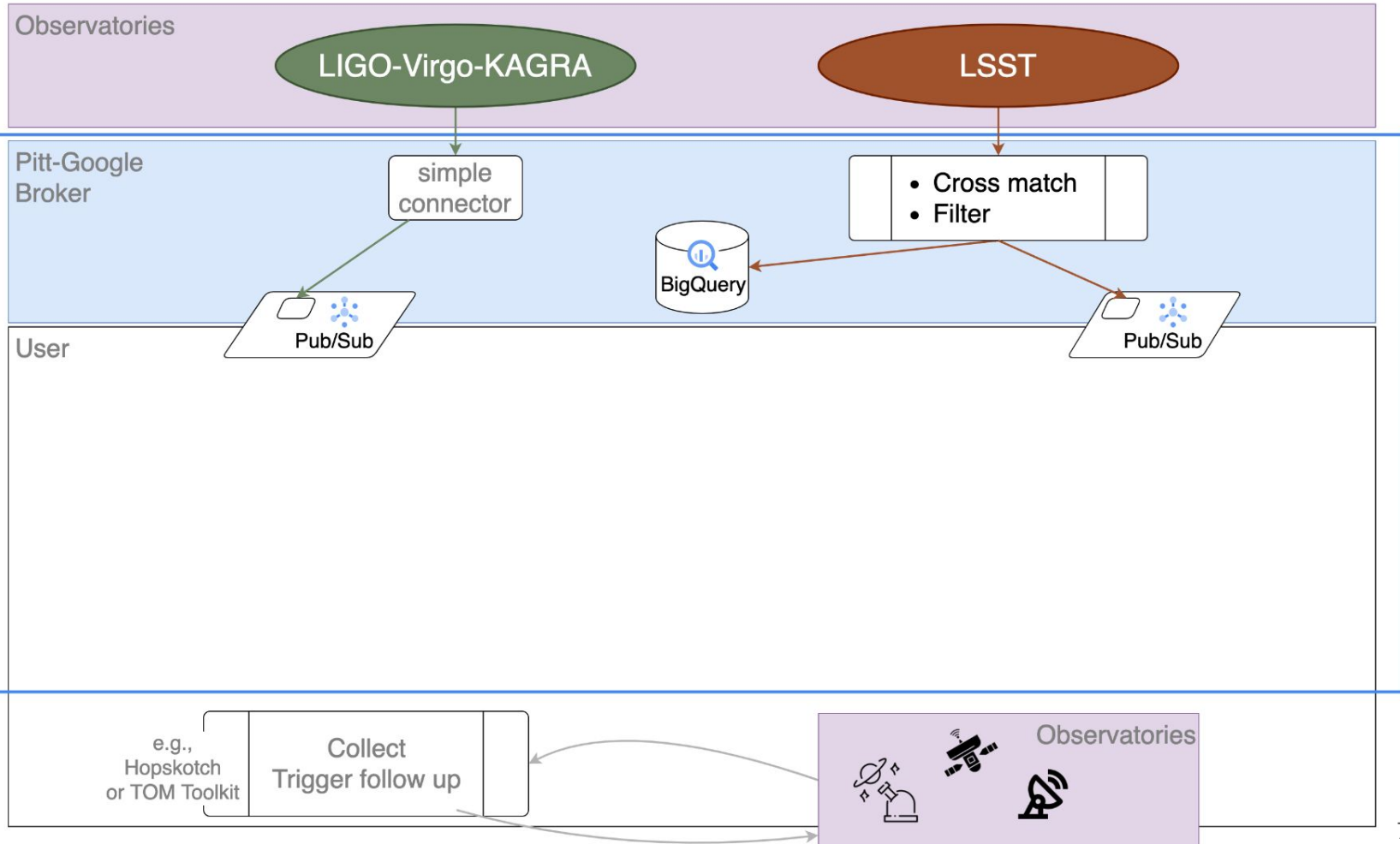
User analysis via Pitt-Google broker and Google Cloud Platform



User analysis via Pitt-Google broker and Google Cloud Platform



User analysis via Pitt-Google broker and Google Cloud Platform



Data Resources

Message Streams via Pub/Sub

- *Full stream*
- Semantically compressed (every alert, but with less data)
- Pure (real/bogus filter)
- Value added:
 - [Cross match](#) (e.g., Gaia, AllWISE, Census of the Local Universe)
 - Classification (e.g., *SuperNNova* (Möller & de Boissière, 2020), variable stars)
- Filtered (e.g., *likely extragalactic*, likely supernovae)
Completeness prioritized over purity.

Pub/Sub

- Streaming message service
 - functionally similar to Apache Kafka, but easier to use
- **Push to any HTTP endpoint for event-driven processing**
 - e.g., Google Cloud Run, AWS Lambda
- Pull from anywhere via APIs (Python, REST, gRPC, CLI, ...)

Catalogs in BigQuery

- *Alerts*
- [DIA Source](#)
- DIA Object
- *Value added*

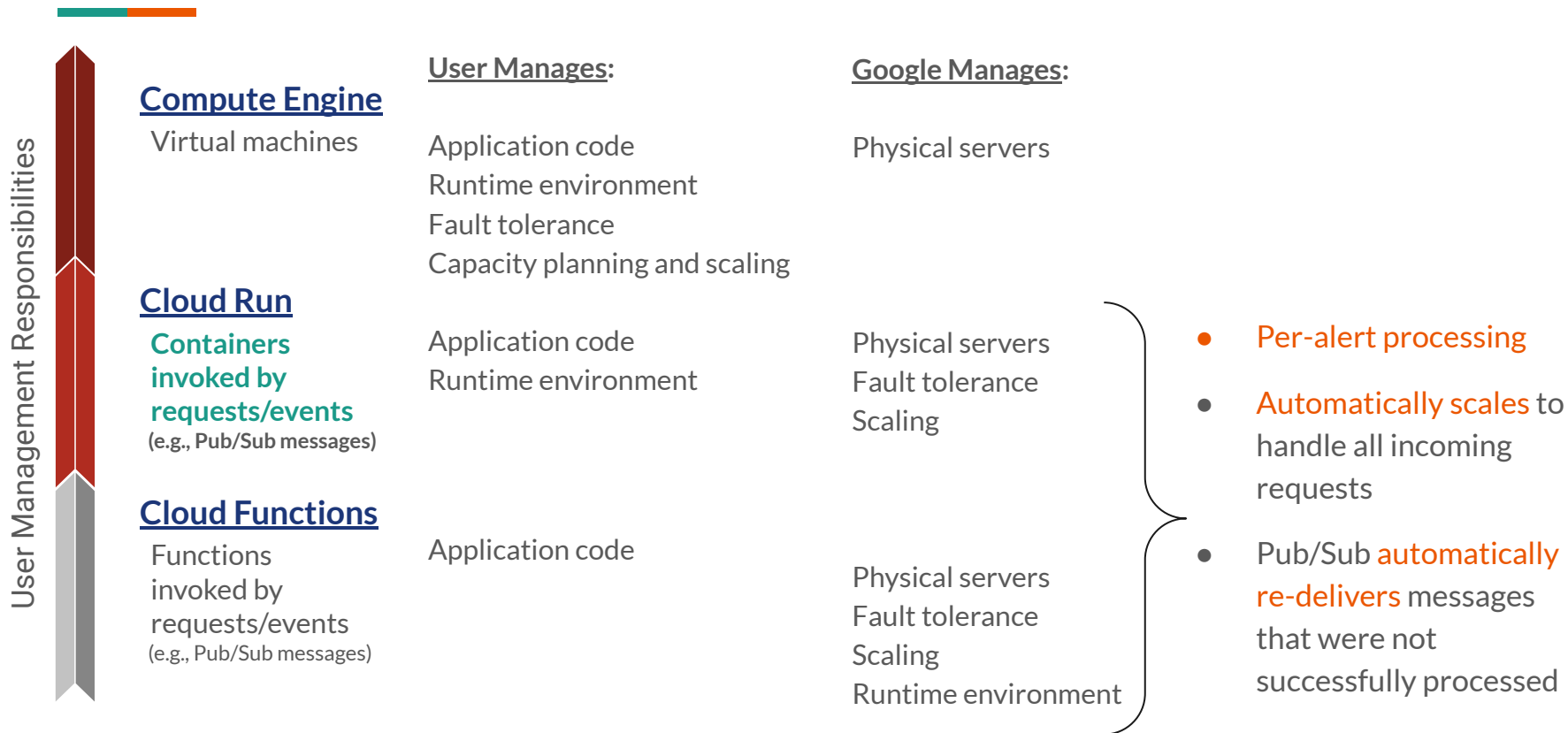
BigQuery

- Data warehouse. Relational. SQL access.
- **Optimized for complex analytical queries on massive datasets**
- **Real-time streaming inserts, immediately queryable**
- Built-in geographic information system (GIS), useful for cross matching
- TAP/ADQL available

File access from [Cloud Storage](#) - *complete, original alert packets*

italics = currently available for ZTF

Compute on Google Cloud (some options)



Low-Latency Processing



Time delta between:

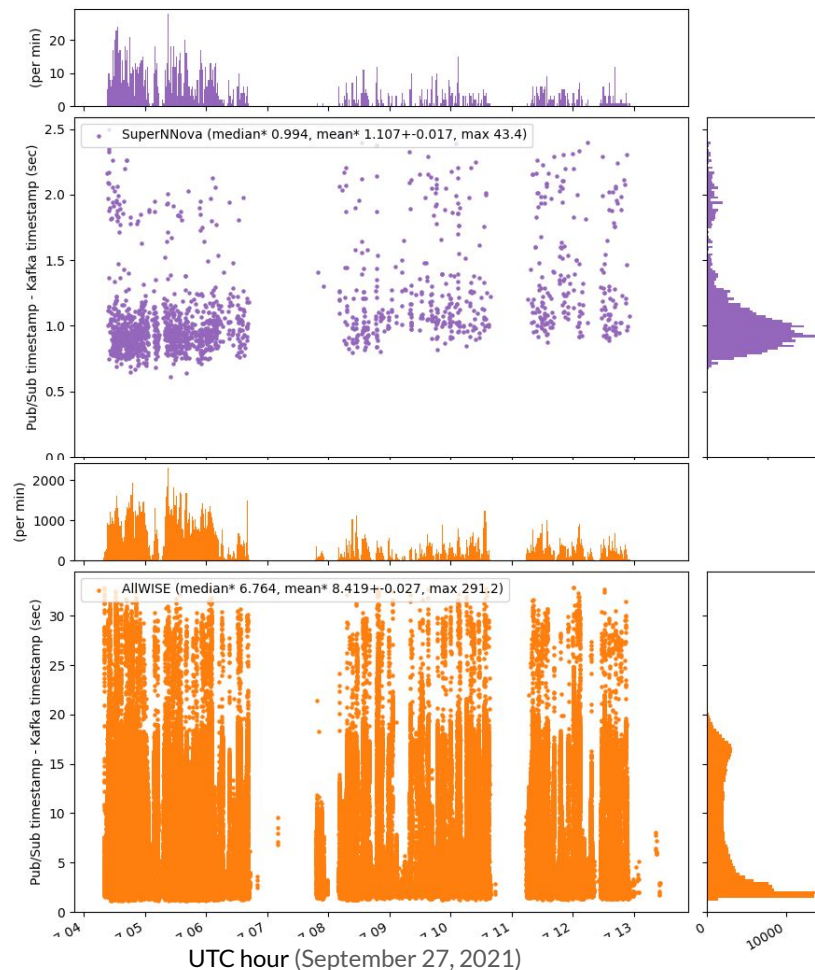
- **ZTF alert production** at IPAC (Kafka timestamp)
- **Pitt-Google message production** after processing thru the indicated module (Pub/Sub timestamp)

SuperNNova classification

- **~1 sec median latency**
- Includes transfer from ZTF, ingestion to Pitt-Google, and 2 previous pipeline modules (running on Cloud Functions)

Cross match with AllWISE catalog in BigQuery

- **~7 sec median latency**
- This is worst-case scenario! Cone search for 3 nearest neighbors in AllWISE catalog, stored in an **unoptimized** BigQuery table.



User Access

User Accounts

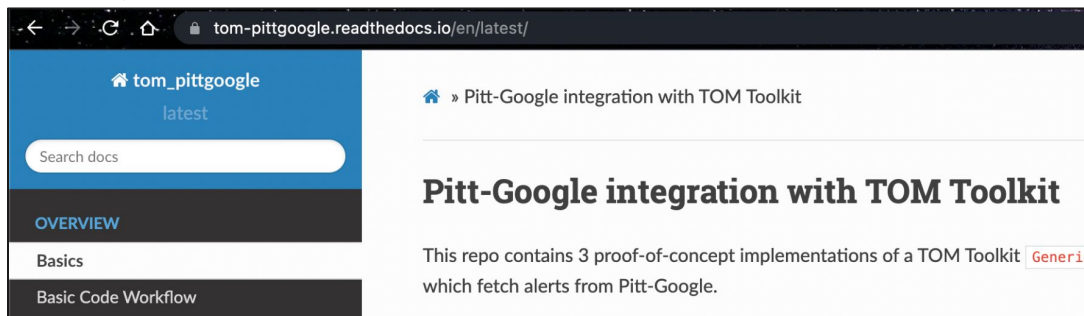
- Pitt-Google broker ingests data into Google Cloud services (e.g., Pub/Sub, BigQuery).
- **Users access data through Google Cloud Platform directly**, not through Pitt-Google.
 - Need a Google account (e.g., Gmail address) and a free Google Cloud Platform project (~1 min.). [See our docs](#).

Data Access Methods

- **API access to everything** (e.g., **Python**, Java, REST, CLI, etc.), in and out of the Cloud.

Integration example: [Pitt-Google's TOM Toolkit plugin](#)

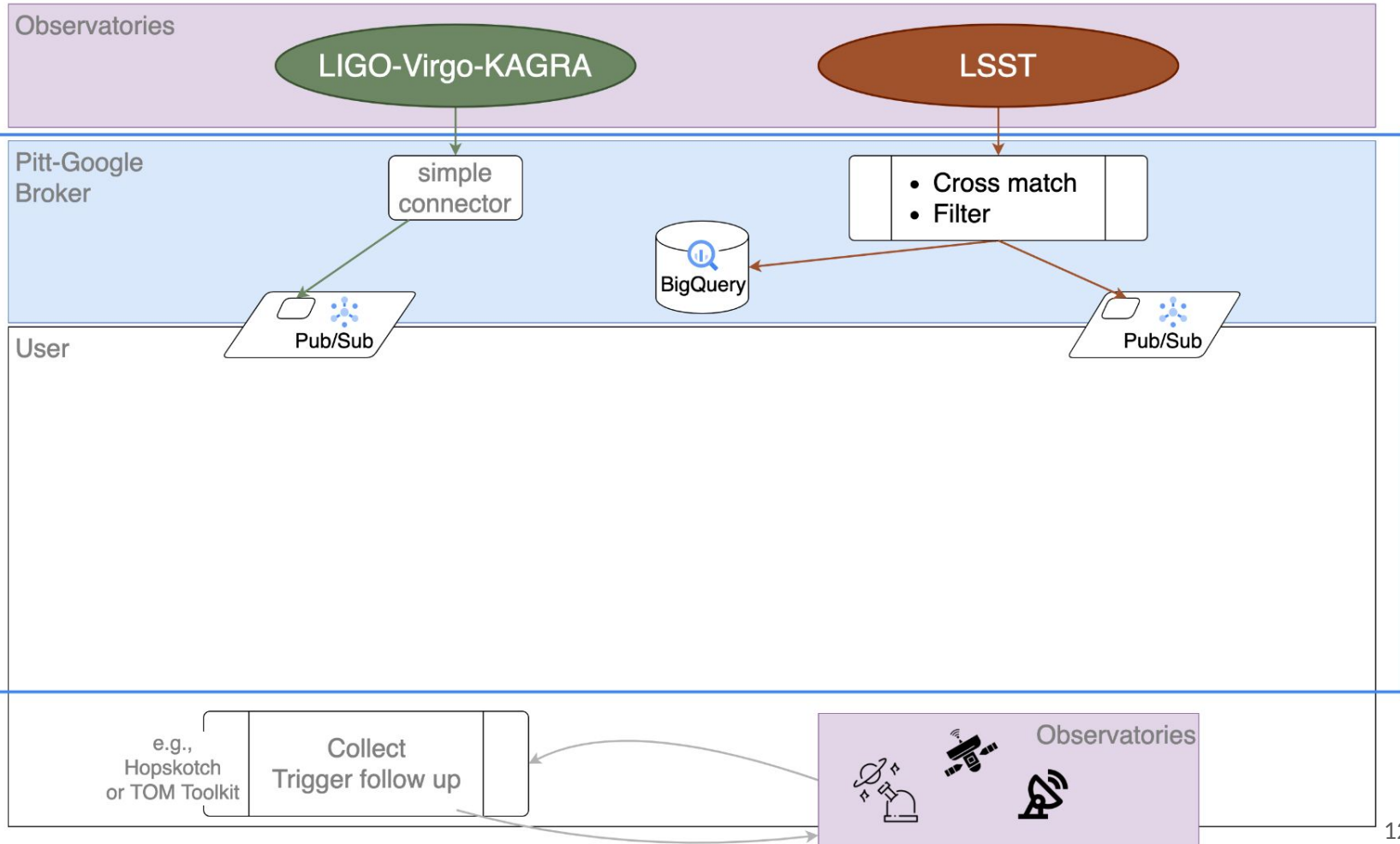
- Web console
- Easy to move data around within the Google Cloud via Pub/Sub.
 - e.g., Simple to develop and maintain data event-driven pipelines by processing on managed compute services.



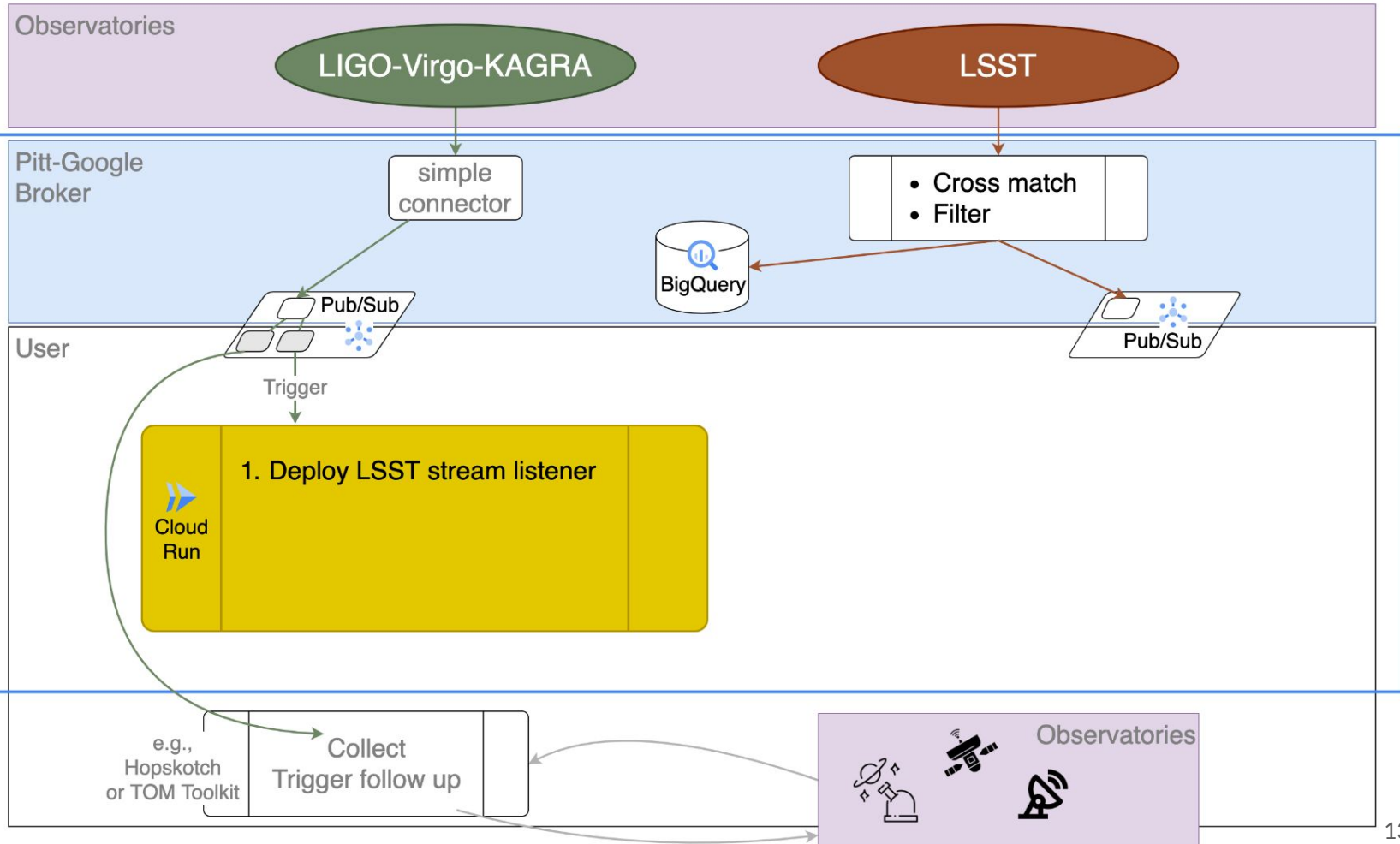
Pricing Structure

- There is a **Free-Tier** for most services
- After free-tier, **pay-as-you-go** and **only pay for what you actually use**.

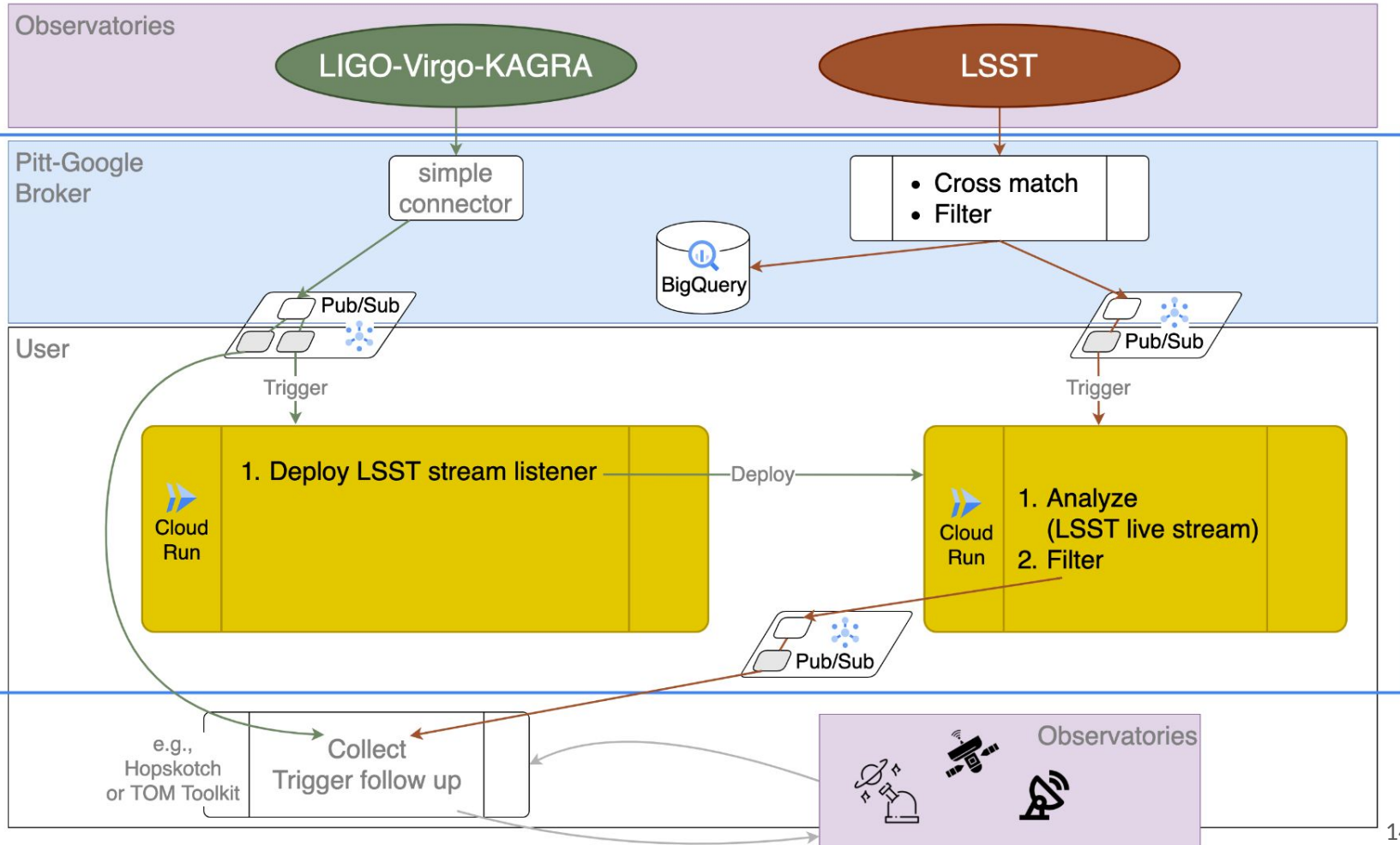
User analysis via Pitt-Google broker and Google Cloud Platform



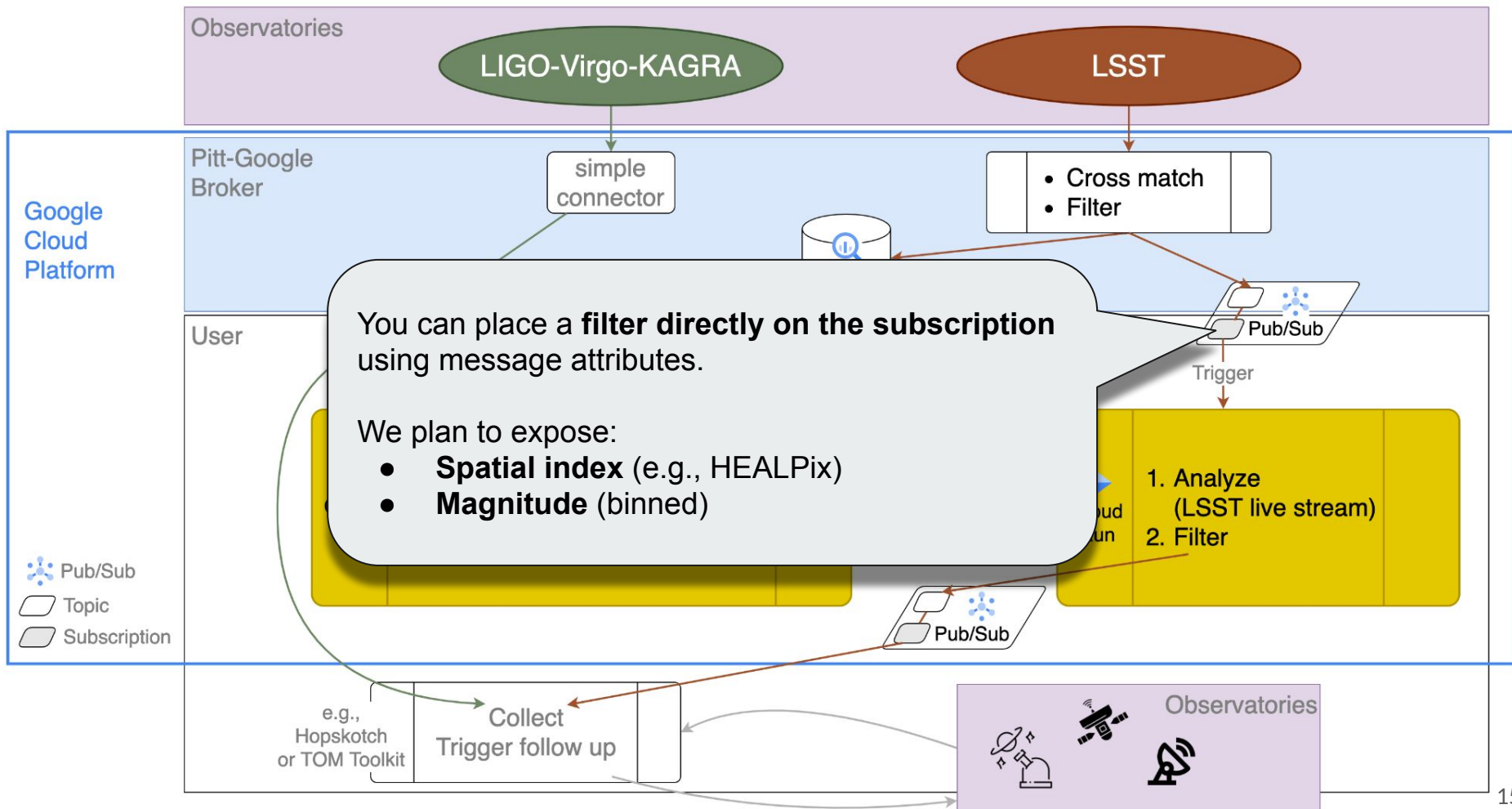
User analysis via Pitt-Google broker and Google Cloud Platform



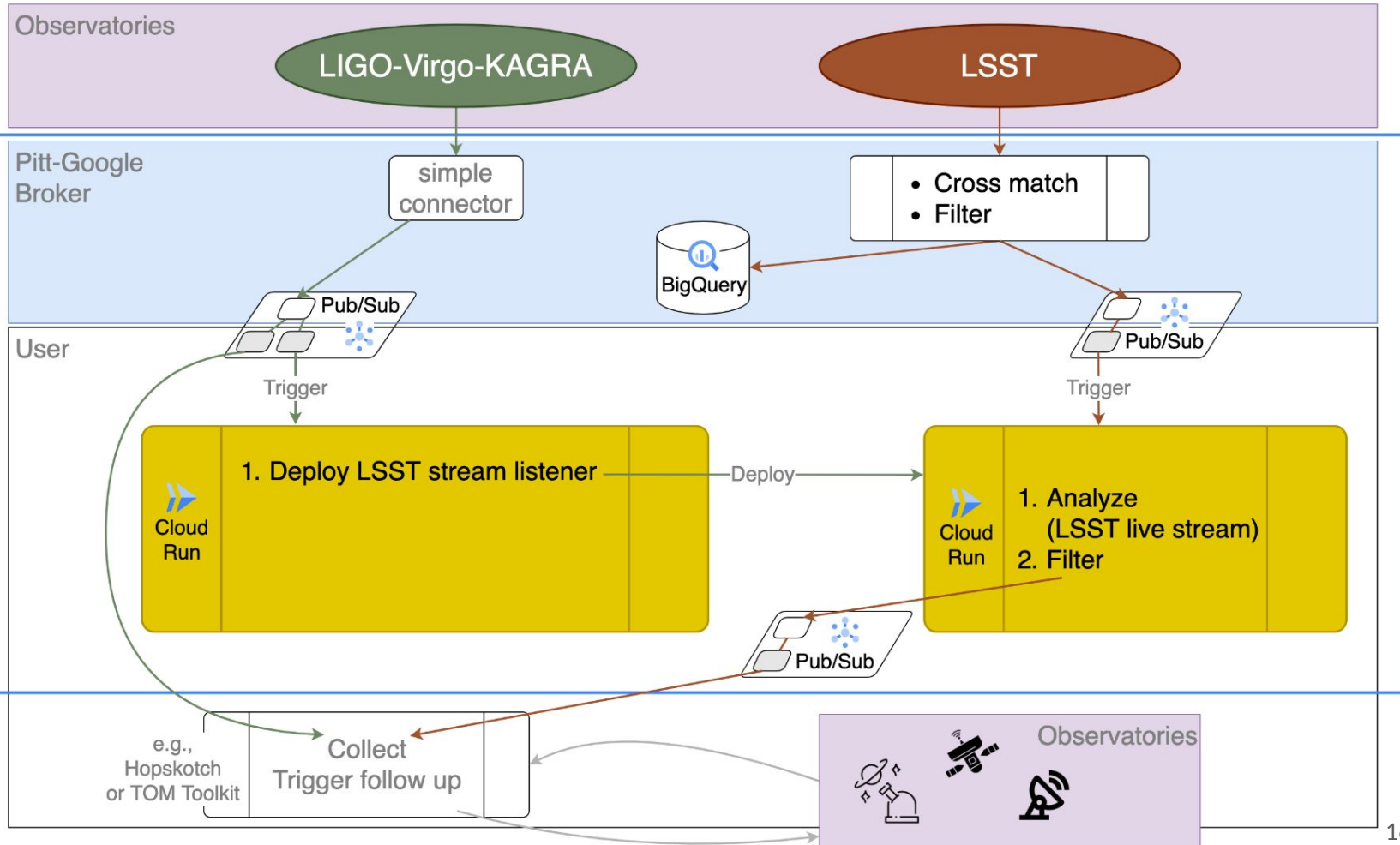
User analysis via Pitt-Google broker and Google Cloud Platform



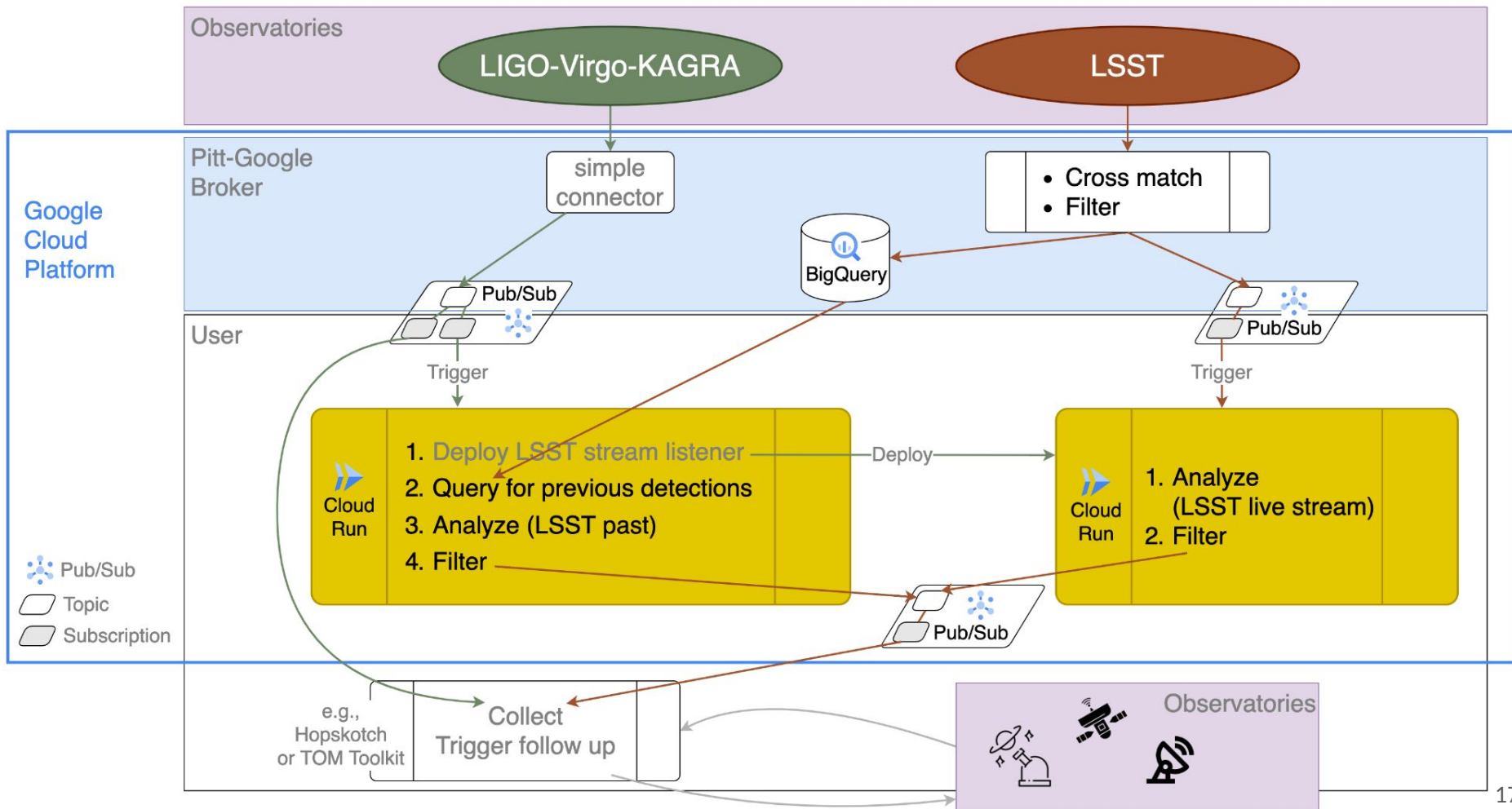
User analysis via Pitt-Google broker and Google Cloud Platform



User analysis via Pitt-Google broker and Google Cloud Platform



User analysis via Pitt-Google broker and Google Cloud Platform



User Analysis Cost Estimates

There are MANY factors at play that will affect the cost.

This is a **generous estimate that assumes you process a lot of data each night.**

Yearly cost estimate

\$810.88

Cloud Run	Resource estimates	Usage estimates	Nightly cost estimate
	2 CPUs	0.5 sec / event	
	512 MiB memory	10^5 events / night	
	\$0.67		

BigQuery	Operation	Free Tier	Standard Price
	query	1 TB / month	\$5.00 / TB
	Nightly data estimate		Nightly cost estimate at standard price
	0.25 TB		\$1.25

Pub/Sub	Operation	Free Tier	Standard Price
	data delivery	10 GiB / month	\$40 / TiB
	Nightly data estimate		Nightly cost estimate at standard price
	10^5 alerts * 80 kb / alert		
	= 0.0073 TiB		\$0.29
	Operation		Standard Price
	egress to Europe		\$0.12 / GB
	Nightly data estimate		Nightly cost estimate at standard price
	10^3 alerts * 80 kb / alert		
	= 0.08 GB		\$0.01

[Cloud Run pricing](#) | [BigQuery Pricing](#) | [Pub/Sub pricing](#) | [Google Cloud Pricing Calculator](#)

Takeaways



Pitt-Google Alert Broker

Ingesting alert streams into Google Cloud Platform
and adding value (e.g., cross matches) such that
users can easily take advantage of Google's advanced, big-data technology solutions for themselves.

Will be a **full-stream LSST broker**.

Currently running **ZTF**. In early dev. stage to run **gravitational wave alert streams**.

Users can

Compute next to the data without managing servers, VMs, scaling or fault tolerance.

Export the data of interest (API access to everything).

Troy Raen | Slack @troyraen on LSSTC | Email troy.raen@pitt.edu