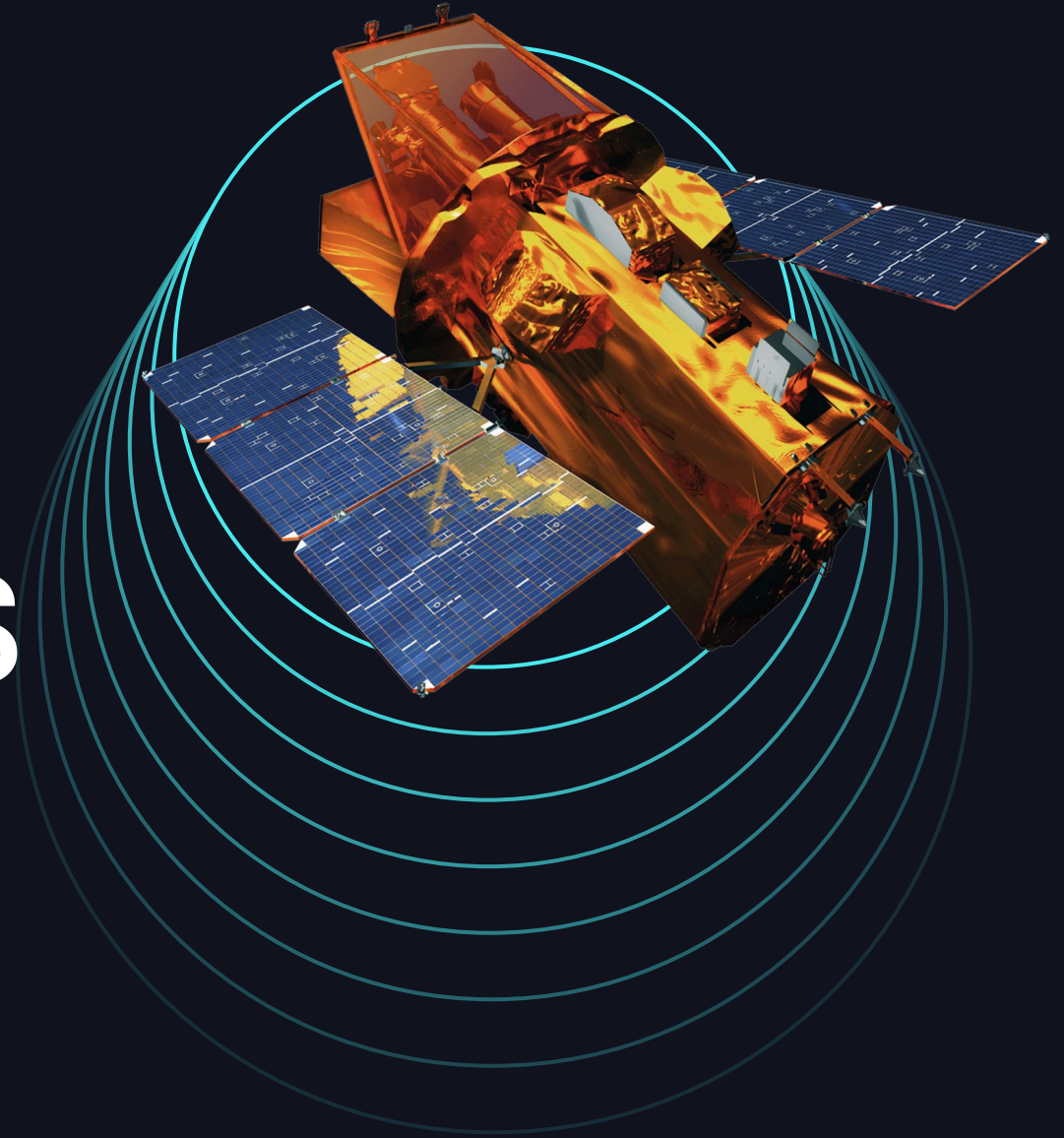


STREAMING DATA PIPELINES FROM SUPERNOVAS TO LLMS



Frank Munz, Databricks
June 2024

Product safe harbor statement

This information is provided to outline Databricks' general product direction and is for **informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all



Supernovas, Black Holes and GRBs

1 Gamma Ray Burst (GRB) ~ energy of the sun over its lifetime

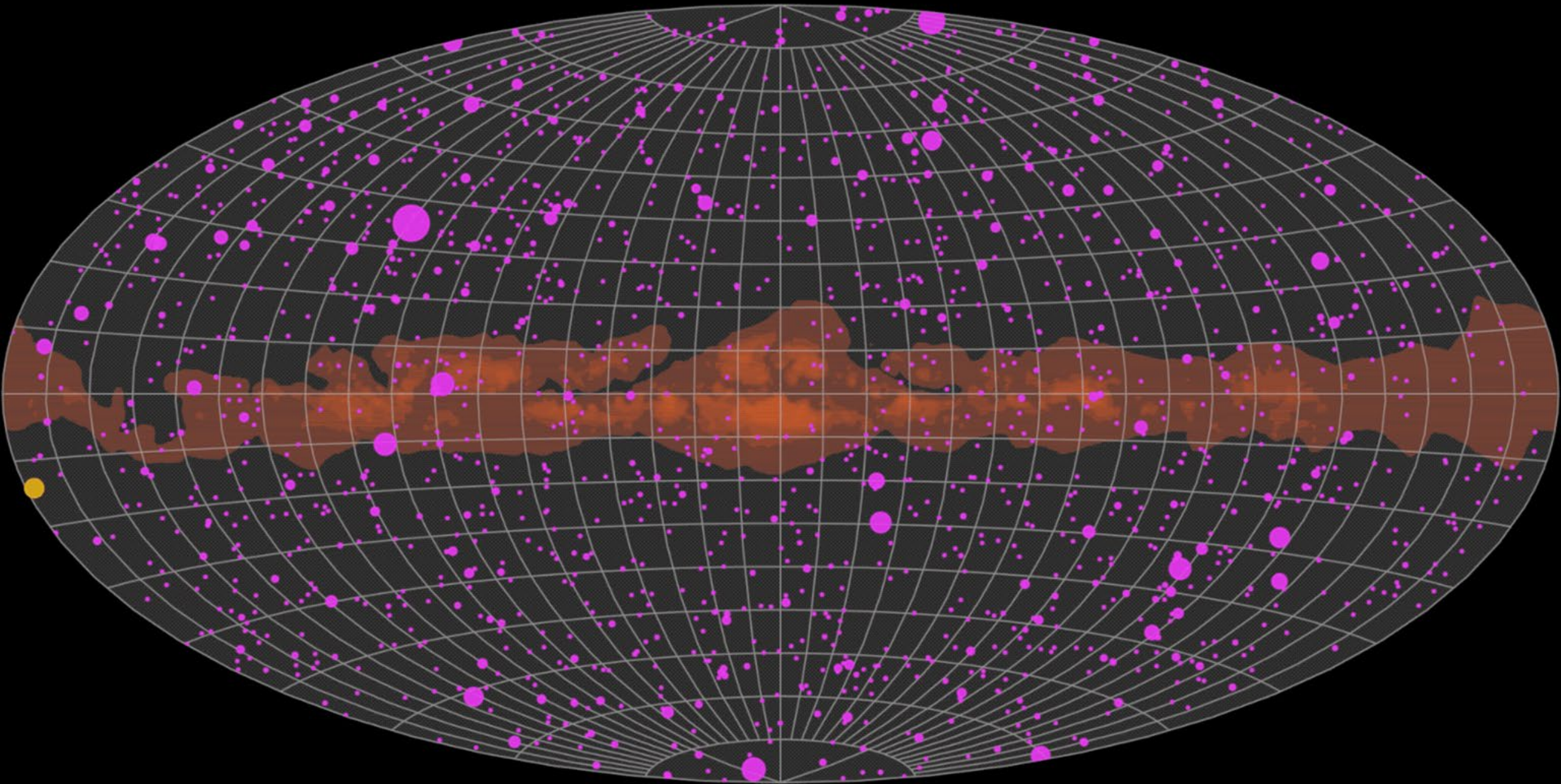
- < 2 seconds: **merger** of neutron stars or a **neutron star and a black hole**
- > 2 seconds: **collapse** of a **massive star** (> 30 solar masses)

Supernova

- Massive stellar explosions at the end of a star's life
- Can leave behind a black hole or neutron star

Black Hole

- Can form from the merger of 2 neutron stars or 2 black holes
- Extremely dense regions of space with immense gravitational pull



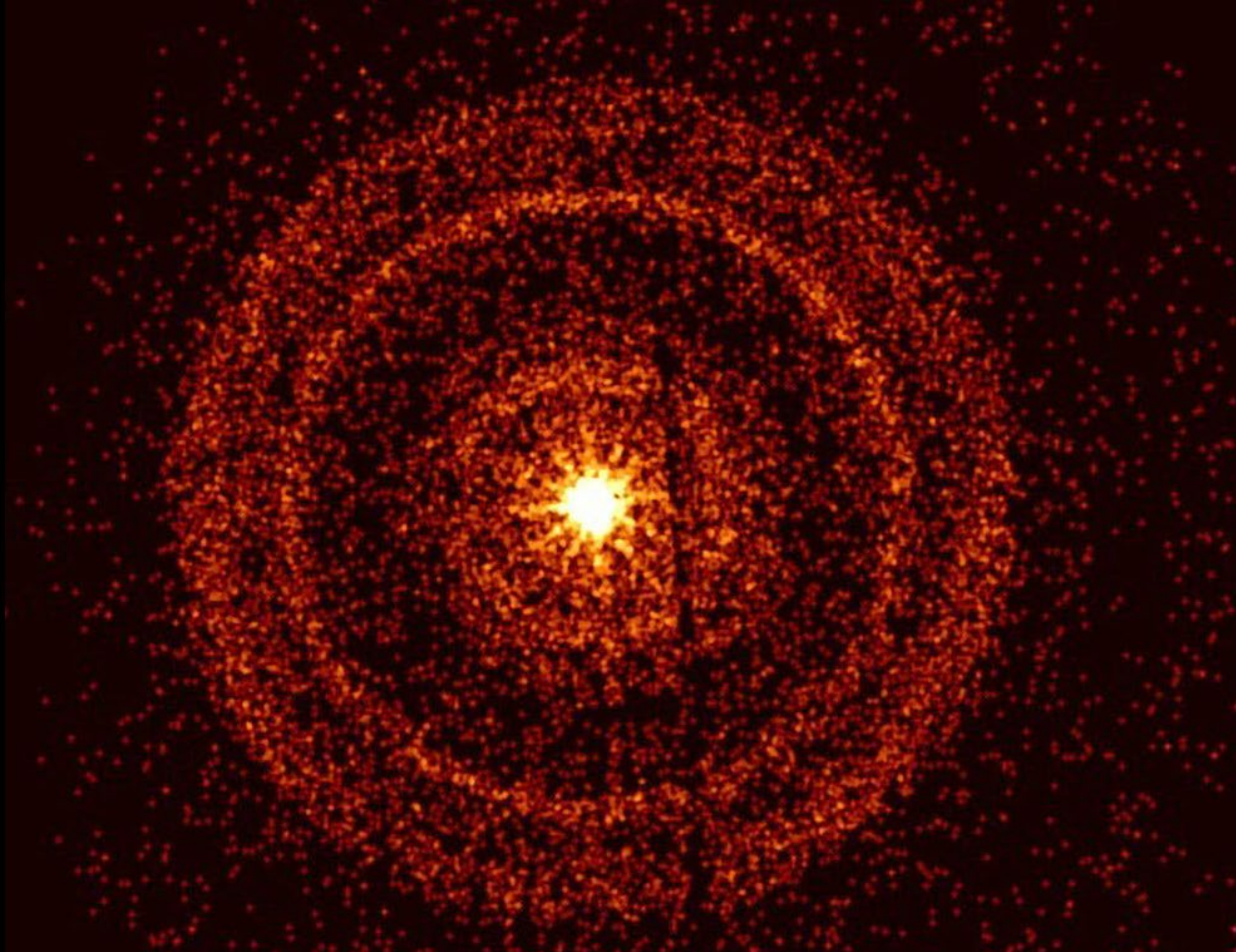
Neil Gehrels Swift Observatory

Launched 2004, Data transmitted via
Gamma-ray Coordinates Network (GCN)

Key Instruments:

- **Burst Alert Telescope (BAT):**
Locates GRBs across a wide field of view.
- **X-ray Telescope (XRT):**
Observes afterglow of GRBs in X-ray wavelengths
- **Ultraviolet/Optical Telescope (UVOT):**
Captures optical and ultraviolet emissions





BOAT – GRB 221009A

(Closest and) Brightest Of All Times Gamma Ray Burst

- Detected on Oct 9, 2022 simultaneously by **Swift** and **Fermi** telescopes
- Originated 2.4 billion light-years away (**1.9 bn ago**) in Sagitta
- Lasted over **10 hours**, with 10 minute initial burst
- 5,000 VHE photons detected, previous record was ~100
- Brightest GRB afterglow ever recorded
- Thought to occur only **once every ~10,000 years**

Event time
vs ingestion
time?



IceCube Neutrino Observatory


Located in Antarctica

- Detects high-energy neutrinos from extreme cosmic environments
- 5,160 digital optical modules (DOMs)
- Embedded in **a cubic km of ice**
- Ice serves as detector medium and background radiation shield
- Neutrinos produce **Cherenkov radiation, detected by DOMs**

- **Data transmitted via Gamma-ray Coordinates Network (GCN)**
-> alerts astronomers for quick follow-up observations


NASA uses Apache Kafka: GCN Project

Judith Rascusin's (NASA) Talk @ Current.io 2023




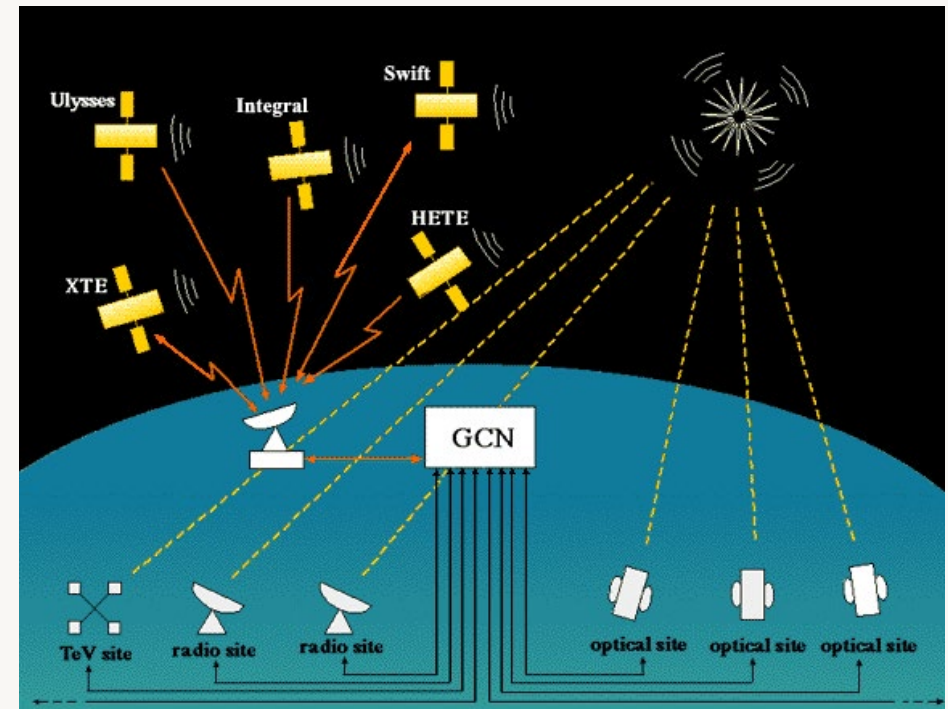
Two Types of GCN Data Products

GCN Notices	GCN Circulares
<pre> TITLE: GCN/PERMI NOTICE NOTICE DATE: Wed 26 Aug 22 21:10:07 UT NOTICE TYPE: Permi-GM Flight Position SOURCE_NUM: 43 STRUCID_NUM: 42912247 294.2526 (+190.456 12s) (20000), 294.4166 (+190.456 08s) (1950), 294.2126 (+191.067 1000s), +71.8488 (+718 52' 52" (constant), +71.8294 (+718 47' 26" (1970)), GEM_RASID: 5.30 (5m radius, statistical plus systematic) GEM_SYSTEM: 1016 (constant) DATA_TRIGGER: 22.60 (alpha) DATA_TIME: 1.424 (sec) GEM_DATE: 1997 J200: 238 000: 20789724 GEM_TIME: 79782.72 000 (22:09:42.72) 00 GEM_YEAR: 20 GEM_YEARID: 150-00 (04s) DATA_TIME_OFFSET: 1-0000 (04s) MWD_RATIO: 0.34 SOL_ANGLEFROM: 3 (version number of) MWD_SYSTEM: 934 GM JOB_NAME_PREFIX: 44. Generic Transient OBJECTNAME: G.C.G. 1-1, G.C.G. G.C.G. G.C.G. GEM_POSITION: 154.082 (+126.246 12s) +10.208 (csga 55' 51") GEM_POSITION: 94.000 (long) -0.2 (lat) (axis of beam) MWD_POSITION: 258.344 (+126.126 14s) -22.274 (-22d 13' 54") MWD_OFFSET: 92.44 (04s) MWD_OFFSET: 43 (1) GEM_COMMENT: 101.02 11.53 (lat) latitude (in lat of the burst (in transients)) </pre>	<pre> TITLE: GCN CIRCULAR NUMBER: 2428 SUBJECT: GM 200820: Permi GM detection DATE: 20/08/27 21:10:00 GMT FROM: C. Malacaria (NASA-MPO/PSDA) and C. Heesen (IAS) report on behalf of the Permi GM team. *At 22:09:42.72 UT on 28 August 2020, the Permi Gamma-ray Burst Monitor (GM) triggered and located GM 200820 (trigger 420172087 / 200820821). The on-ground calculated location, using the GM trigger data, was reported in GCN 20392. The GM light curve shows an exceptionally bright long GM with a duration (90%) of about 7.4 s (10-100 keV). The time-averaged spectrum from 20-0.003 s to 70+ 12.544 s in best fit by a Band function with slope = -1.0(1.3) +/- 3.5. kT = alpha = -0.41 +/- 0.21, and beta = -2.52 +/- 0.04. The event (21ms (10-100 keV) in this time interval is (1.414 +/- 0.0618-04 seg/cm^2). The 1-0.003-sec peak photon flux measured starting from 70% 1 s in the 10-1000 keV band is 110(1) +/- 0.1 ph/s/cm^2. The spectral analysis results presented above are preliminary. Final results will be published in the GM GM Catalogue. https://heasarc.gsfc.nasa.gov/W3Browse/permi/fmnight.html The Permi GM team and GM team members are the following: </pre>
<ul style="list-style-type: none"> • By and for machines • Fixed, predefined format • Schema specific to each notice type 	<ul style="list-style-type: none"> • By and for humans (some automated) • Freeform text (with established style) • Citable (but not peer-reviewed)



2023 | THE NEXT GENERATION OF KAFKA SUMMIT

ORGANIZED BY:  CONFLUENT



GCN Notices: machine generated
 + Circulares: human generated


[Link to Judith's talk](#)



Get Your OIDC Credentials

<https://gcn.nasa.gov/quickstart>

An official website of the United States government [Here's how you know](#) ▾

 **General Coordinates Network** Missions Notices Circulars Documentation frank.munz@databricks.com ▾

New Announcement Feature, Code of Conduct, Circular Revisions. See [news and announcements](#)

Start Streaming GCN Notices

1 Sign in / Sign up 2 **Select Credentials** 3 Customize Alerts 4 Get Sample Code

2 of 4 Select Credentials

Client credentials allow your scripts to interact with GCN on your behalf. Select one of your existing client credentials, or create a new one.

<code>f3</code> (created 4 months ago) scope: <code>gcn.nasa.gov/kafka-public-consumer</code> client ID: <code>12uku1alv4grcn2qv27o8hf587</code>	Delete	Select →
--	------------------------	--------------------------

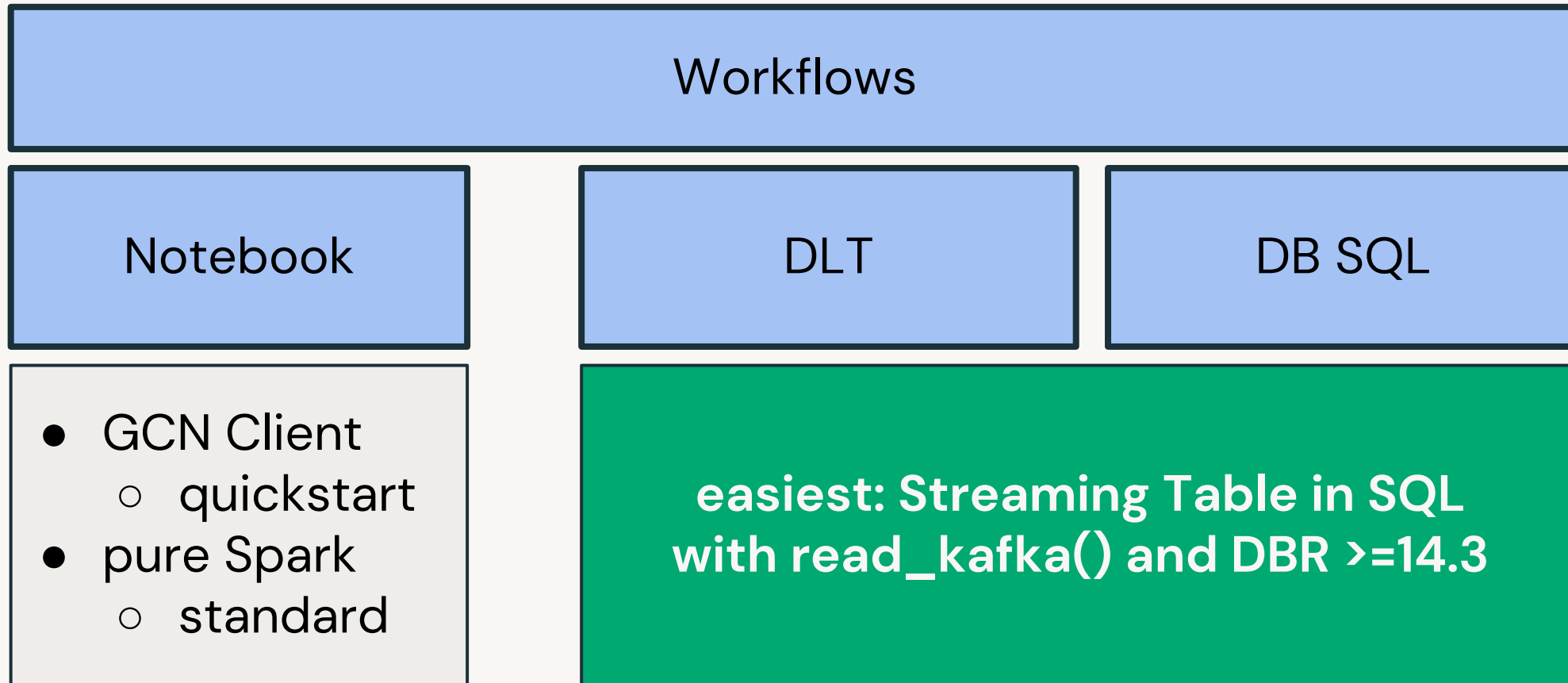


The Data Intelligence Platform
supports streaming data from the
ground up

The main actors (Ingest)

Ingest Streaming Data from Apache Kafka

(we cover the human written circulars later...)



Now, show me the code

Notebook with GCN Kafka Wrapper

Wraps Confluent Kafka Client

```
from gcn_kafka import Consumer

topics = ['gcn.classic.text.SWIFT_POINTDIR']
config = {'auto.offset.reset': 'earliest'}

consumer = Consumer(config,
                    client_id='abc...',
                    client_secret='xyz...',
                    domain='gcn.nasa.gov')
consumer.subscribe(topics)

while True:
    for message in consumer.consume(timeout=1):
```

KAFKA message

msg = TITLE: GCN/SWIFT NOTICE
NOTICE_DATE: Fri 03 May 24 04:16:31 UT
NOTICE_TYPE: SWIFT Pointing Direction
NEXT_POINT_RA: 213.407d {+14h 13m 38s} (J2000)
NEXT_POINT_DEC: +70.472d {+70d 28' 20"} (J2000)
NEXT_POINT_ROLL: 2.885d
SLEW_TIME: 15420.00 SOD {04:17:00.00} UT
SLEW_DATE: 20433 TJD; 124 DOY; 24/05/03
OBS_TIME: 900.00 [sec] (=15.0 [min])
TGT_NAME: RX J1413.6+7029
TGT_NUM: 3111759, Seg_Num: 10
MERIT: 60.00
INST_MODES: BAT=0=0x0 XRT=7=0x7 UVOT=12525=0x30ED
SUN_POSTN: 40.78d {+02h 43m 07s} +15.81d {+15d 48' 31"}
SUN_DIST: 93.68 [deg] Sun_angle= -11.5 [hr] (East of Sun)
MOON_POSTN: 338.61d {+22h 34m 27s} -12.48d {-12d 28' 49"}
MOON_DIST: 113.09 [deg]
MOON_ILLUM: 31 [%]
GAL_COORDS: 113.36, 45.10 [deg] galactic lon,lat of the pointing direction
ECL_COORDS: 143.56, 69.70 [deg] ecliptic lon,lat of the pointing direction
COMMENTS: SWIFT Slew Notice to a preplanned target.
COMMENTS: Note that **preplanned targets** are overridden by any new BAT Automated Target.
COMMENTS: Note that preplanned targets are overridden by any T00 Target if the T00 has a higher Merit Value.
COMMENTS: The **spacecraft longitude,latitude at Notice_time is 247.70,10.86 [deg]**.
COMMENTS: This Notice was ground-generated -- not flight-generated.

What that SWIFT notice means:

Swift Alert: Pointing towards RX J1413.6+7029

On Friday, May 3rd, 2024, at 04:16:31 UT, the Swift telescope is scheduled to point towards a preplanned target, RX J1413.6+7029. This celestial object is located at a Right Ascension of 213.407 degrees (or 14 hours, 13 minutes, and 38 seconds) and a Declination of +70.472 degrees (or +70 degrees, 28 minutes, and 20 seconds).

The telescope will begin its slew to this target location at 04:17:00.00 UT, which will take approximately 15 minutes to complete. Once in position, Swift will observe RX J1413.6+7029 for 900 seconds, or 15 minutes, using its Burst Alert Telescope (BAT), X-ray Telescope (XRT), and Ultraviolet/Optical Telescope (UVOT).

At the time of observation, the Sun will be at a position of 40.78 degrees (or 2 hours, 43 minutes, and 7 seconds) and +15.81 degrees (or +15 degrees, 48 minutes, and 31 seconds), with a Sun angle of -11.5 hours (or East of the Sun). The Moon will be at a position of 338.61 degrees (or 22 hours, 34 minutes, and 27 seconds) and -12.48 degrees (or -12 degrees, 28 minutes, and 49 seconds), with a Moon illumination of 31%.

It's worth noting that this observation is part of a preplanned target list, but it may be overridden by a new BAT Automated Target or a Target of Opportunity (TOO) with a higher merit value. Additionally, the spacecraft's longitude and latitude at the time of observation will be 247.70 degrees and 10.86 degrees, respectively.

Ingest and Transform Easily with Delta Live Tables Pipelines

The best way to do ETL on the Databricks Data Intelligence Platform

```
-- incrementally ingest
CREATE STREAMING TABLE raw_data
AS
SELECT *
FROM cloud_files ("/raw_data",
"json")

-- incrementally transform
CREATE MATERIALIZED VIEW clean_data
AS
SELECT timestamp, id, target
FROM LIVE.raw_data
```



Accelerate ETL development

Declare **SQL or Python** and DLT automatically orchestrates the DAG, handles retries, changing data



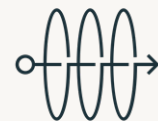
Automatically manage your infrastructure

Automates complex tedious activities like **recovery, auto-scaling, and performance optimization**



Ensure high data quality

Deliver reliable data with built-in **quality controls, testing, monitoring, and enforcement**

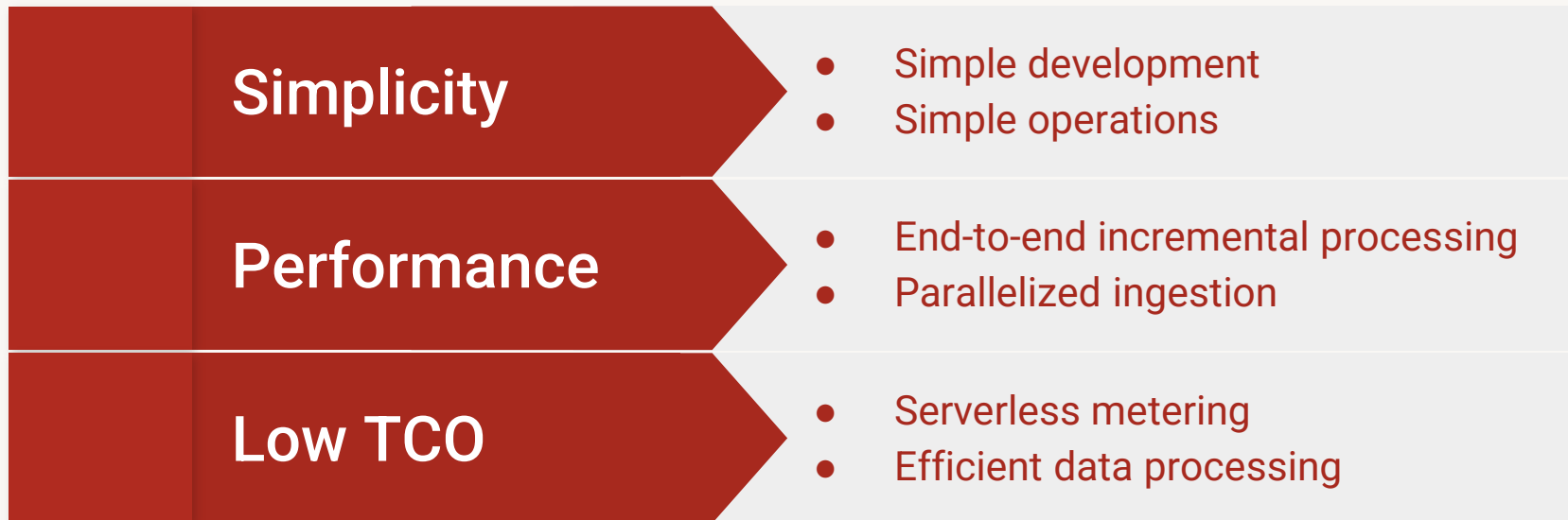


Unify batch and streaming

Get the simplicity of SQL with freshness of streaming with one **unified API**

Delta Live Tables with serverless compute

The simplest way to build data pipelines



Delta Live Tables

Ingest: Streaming Table in SQL with read_kafka()

```
1 CREATE OR REPLACE STREAMING TABLE raw_space_events AS
2 SELECT offset, timestamp, value::string as msg
3 FROM STREAM read_kafka(
4   bootstrapServers => 'kafka.gcn.nasa.gov:9092',
5   subscribe => 'gcn.classic.text.SWIFT_POINTDIR',
6   startingOffsets => 'earliest',
7
8   -- params kafka.sasl.oauthbearer.client.id
9   `kafka.sasl.mechanism` => 'OAUTHBEARER',
10  `kafka.security.protocol` => 'SASL_SSL',
11  `kafka.sasl.oauthbearer.token.endpoint.url` => 'https://auth.gcn.nasa.gov/oauth2/token',
12  `kafka.sasl.login.callback.handler.class` => 'kafkashaded.org.apache.kafka.common.security.oauthbearer.secured.
13  OAuthBearerLoginCallbackHandler',
14
15  `kafka.sasl.jaas.config` =>
16  |
17  |   kafkashaded.org.apache.kafka.common.security.oauthbearer.OAuthBearerLoginModule required
18  |   clientId="7u2rpivvxxxxxxxxxxxxxxxxxxxx"
19  |   clientSecret="1errfm2jdgl0uolkb78kjnf8v94eyyyyyyyyyyyyyyy" ;
20  |
21 );
```

Delta Live Tables

Transformation: Materialized View using PIVOT and type casts

```
1 CREATE OR REPLACE MATERIALIZED VIEW split_events
2 COMMENT "Split Swift event message into individual rows"
3 AS
4 WITH extracted_key_values AS (
5     SELECT
6         timestamp,
7         split_part(line, ':', 1) AS key,
8         TRIM(SUBSTRING(line, INSTR(line, ':') + 1)) AS value
9     FROM (
10        SELECT
11            timestamp,
12            explode(split(msg, '\n')) AS line
13        FROM (LIVE.raw_space_events)
14    )
15    WHERE line != ''
16 ),
17 pivot_table AS (
18     SELECT *
19     FROM (
20         SELECT key, value, timestamp
21         FROM extracted_key_values
22     )
23     PIVOT (
24         MAX(value) FOR key IN ('TITLE', 'NOTICE_DATE', 'NOTICE_TYPE', 'NEXT_POINT_RA', 'NEXT_POINT_DEC',
25             'NEXT_POINT_ROLL', 'SLEW_TIME', 'SLEW_DATE', 'OBS_TIME', 'TGT_NAME', 'TGT_NUM', 'MERIT',
26             'INST_MODES', 'SUN_POSTN', 'SUN_DIST', 'MOON_POSTN', 'MOON_DIST', 'MOON_ILLUM', 'GAL_COORDS',
27             'ECL_COORDS', 'COMMENTS')
28     )
29 )
30 SELECT timestamp, TITLE, CAST(NOTICE_DATE AS TIMESTAMP) AS NOTICE_DATE, NOTICE_TYPE, NEXT_POINT_RA,
31     NEXT_POINT_DEC, NEXT_POINT_ROLL, SLEW_TIME, SLEW_DATE, OBS_TIME, TGT_NAME, TGT_NUM, CAST(MERIT AS
32     DECIMAL) AS MERIT, INST_MODES, SUN_POSTN, SUN_DIST, MOON_POSTN, MOON_DIST, MOON_ILLUM, GAL_COORDS,
33     ECL_COORDS, COMMENTS
34 FROM pivot_table
```

Demo Swift DLT Pipeline with Data Intelligence

Data Rooms

AI/BI Genie


Enable business users to interact with data with LLM-powered Q&A

Natural language -> answers in text and visualizations

Curate dataset-specific experiences with custom instructions

Powered by Databricks SQL & DatabricksIQ

Works with DLT Streaming Tables and Materialized Views



Title
GCN Messages

Description
Describe what data is available in this space and what type of questions users can ask.

Default warehouse
[Stopped] FrankDWH-Serverless Preview

Tables
Choose tables to use for answering questions in the space. It is best to keep the scope for each space as small as possible. Data access is governed by Unity Catalog permissions.

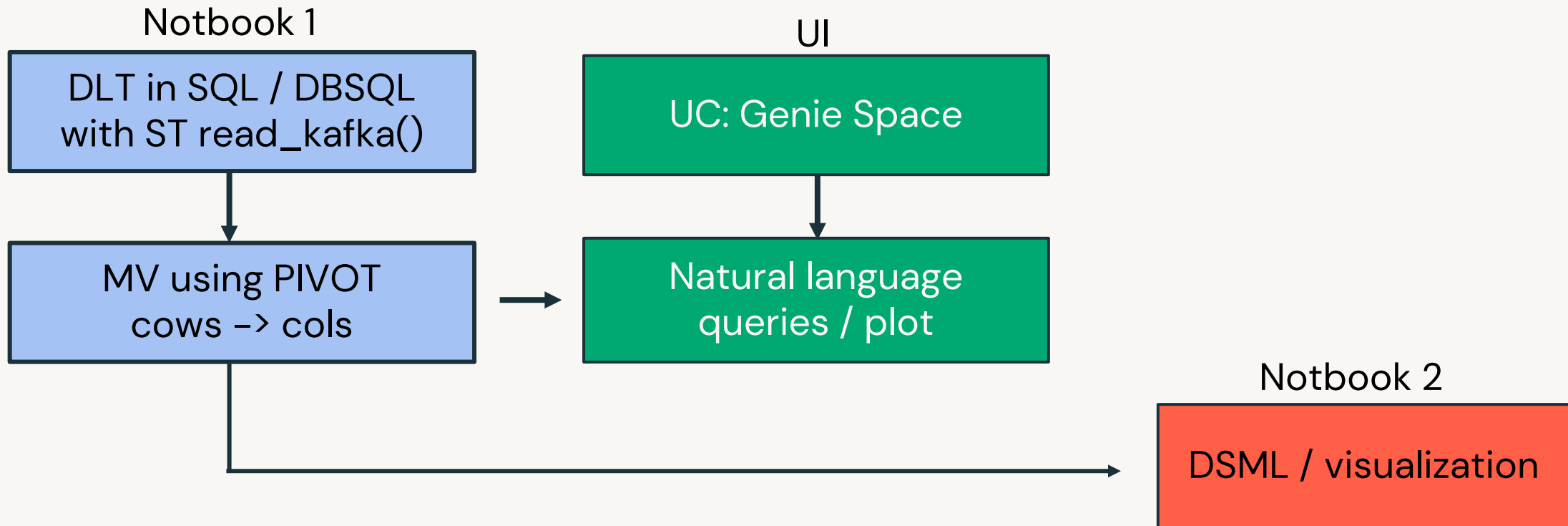
Catalog	Schema	Table
demo_frank	nasa	raw_events
demo_frank.nasa.raw_events		

Sample questions



Demo Genie

SWIFT Analytics – Back of an envelope architecture



Genie or Databricks Assistant?

Databricks Assistant

Technical User

Developer with SQL / Python

Tabular data

Technical or data tasks

- Fix this Python code
- document this table
- write me a SQL query

Genie

Business User

No programming

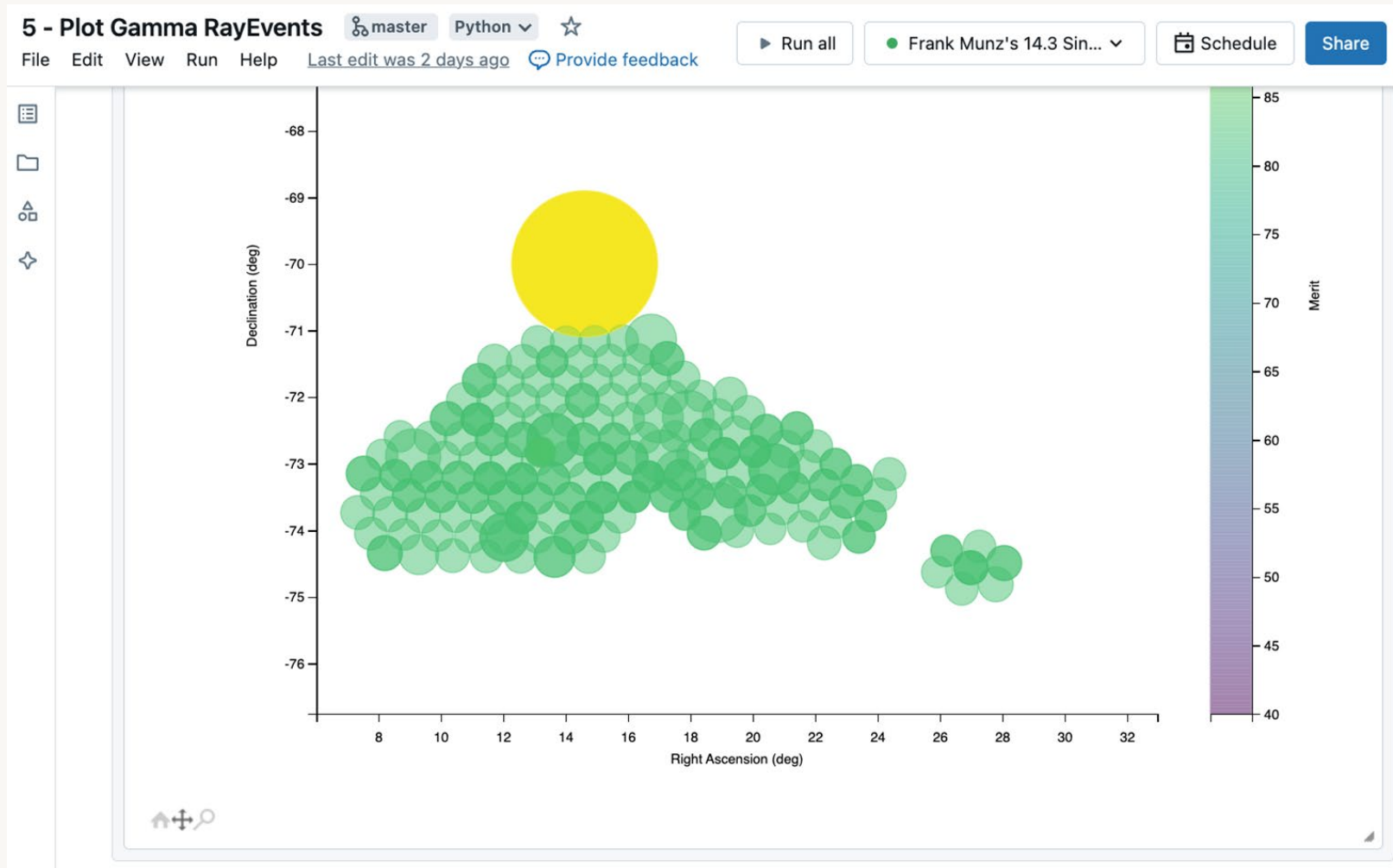
Tabular data

Answer business questions such as

- Who were my fastest growing customers last quarter?
- Explain me this data set



Scientific Notebook Visualization



Genie Visualization

General Instructions

Add general instructions on how you want Genie to behave.

```
* if asked for coordinates take the first part of
columns NEXT_POINT_RA
as RA and next_point_dec as DEC
* from RA and DEC values such as "261.952d {+17h
27m 48s} (J2000)" just take the first part "261.
952" before the "d" as a degree value and drop
the rest.
```

Discard

Save

Example SQL Queries

Add example queries that Genie can learn from.

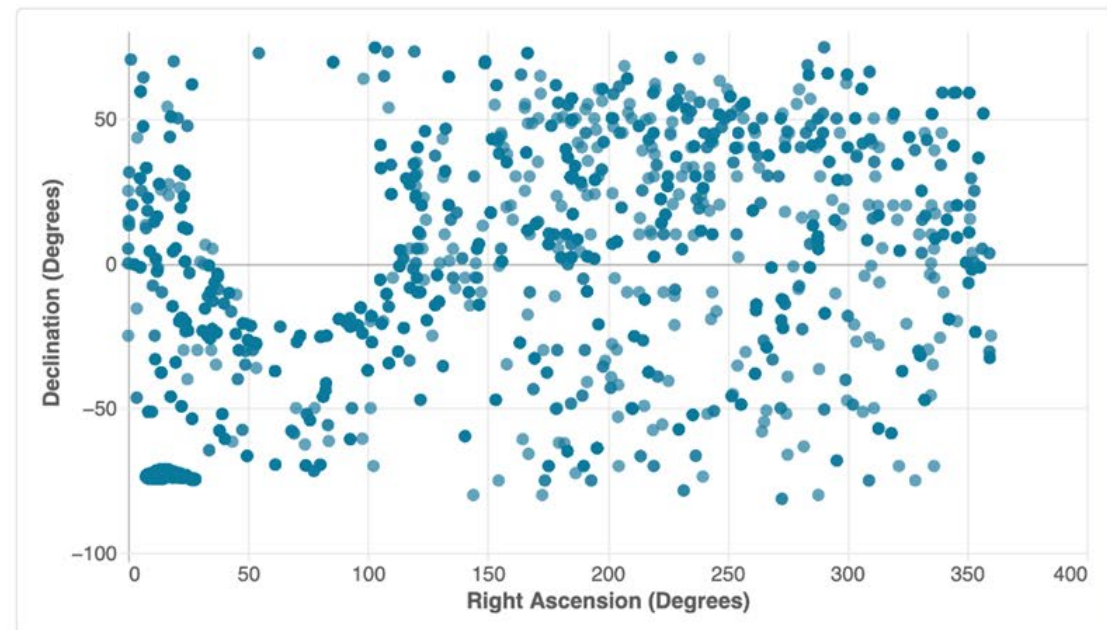
+ Add example query

F Frank Munz



Visualize

Genie



> Show generated code



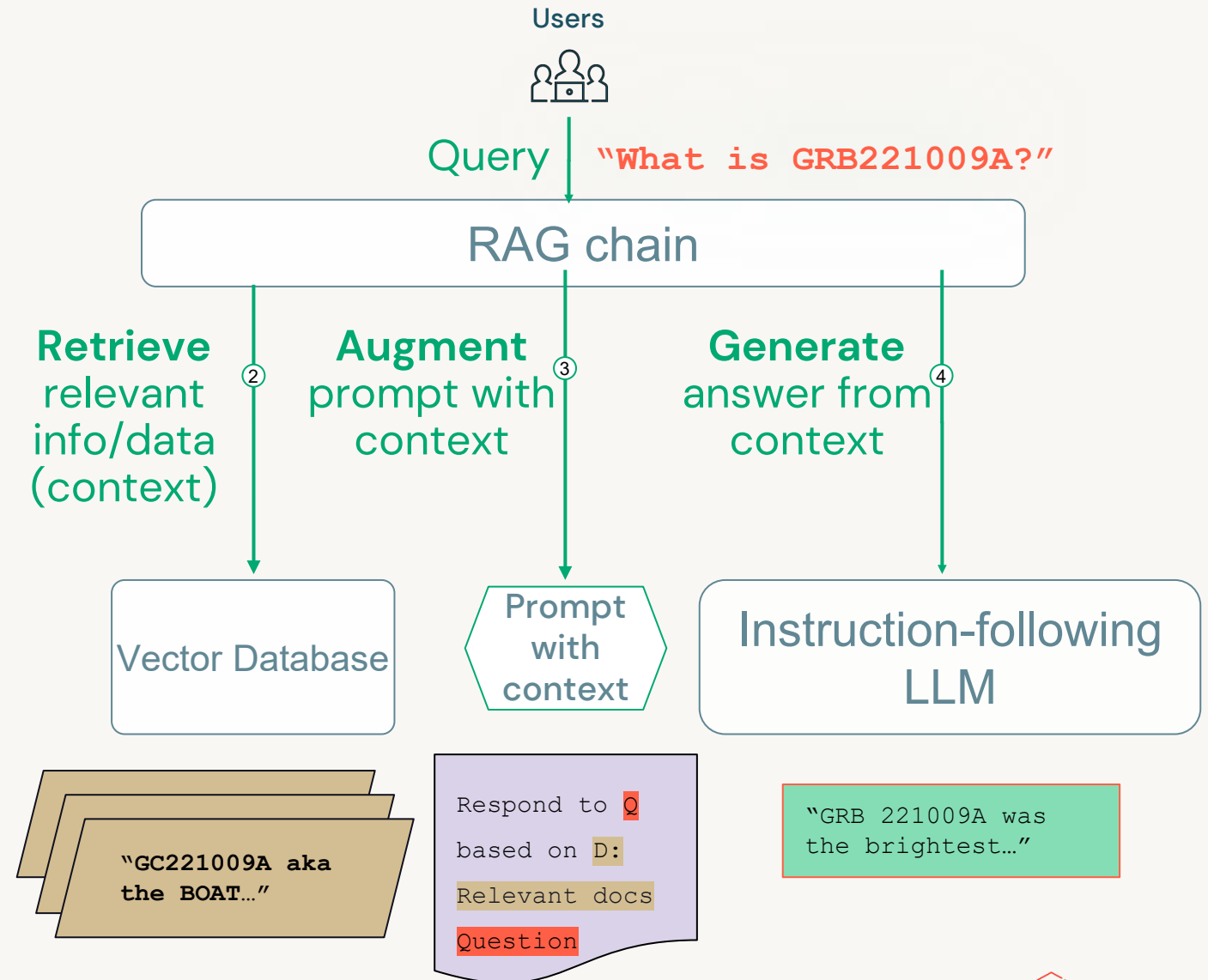
RAG with DBRX / LLama3

Compound AI chat bot based on
36,000 NASA Circulars

Retrieval Augmented Generation (RAG)

RAG uses LLMs as *reasoning engines*, rather than as static models.

Your data
+
an LLM “brain”



Unity Catalog Lineage

Data Lineage for demo_frank.circulars.circulars_chunked

Last 3 months

Volume

demo_frank.nasa.unpack
frank.munz@databricks.com

/Volumes/demo_frank/nasa/unpack/archiv...

Streaming table

demo_frank.circular_pipeline.raw_circulars
frank.munz@databricks.com

bibcode	string
body	string
circularId	bigint
createdOn	bigint
editedBy	string
editedOn	bigint
email	string
subject	string
submitter	string
version	bigint
eventId	string
submittedHow	string
format	string
_rescued_data	string

Hide columns

Materialized view

demo_frank.circular_pipeline.proc_circulars
frank.munz@databricks.com

id	bigint
created	timestamp
body	string
submitter	string

Table

demo_frank.circulars.circulars_chunked
frank.munz@databricks.com

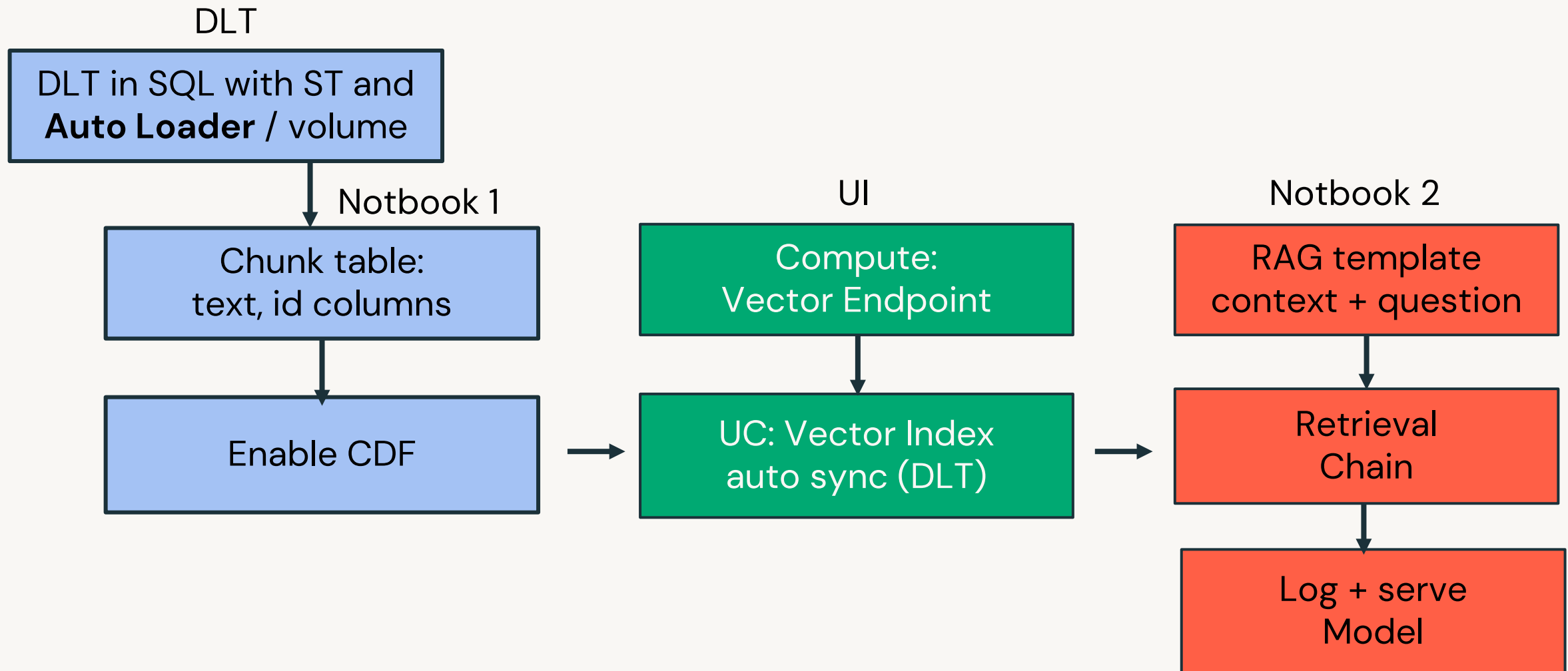
id	bigint
created	timestamp
body	string
submitter	string
chunked_text	string
chunk_id	string

Hide columns



Demo RAG + LLM

Circulars RAG – Back of envelope architecture



**There is no good model
without good data**

Summary and Conclusion

Conclusion

- You are just one copy and paste of a SQL command away from exploring streaming data from a NASA satellite.
- Simply enable Genie on any UC table,
E.g. DLT Streaming Tables or Materialized Views
- Ask Genie natural language questions and create plots
 - Genie writes SQL for you
 - Add your own instructions (2 instructions made notebook obsolete)
 - Instructions work with functions

Conclusion

- RAG adds (context based text) data context to an LLM query
 - The template matters a lot → prompt engineering
 - Fresher data
 - Less hallucinations
- Use Data Intelligence:
Assistant & DBRX and other LLMs for coding support!
- Explore the new RAG Framework and tooling
- TLDR: It's all about the platform

THANK YOU!

Judith Rascusin (NASA)

Alex, Nicolas, Raghu, Praveen, Neil, Eric (Databricks)

& all of YOU!

