Data Mapping at a Billion Dollar Self-Driving Startup

PEPR 2022

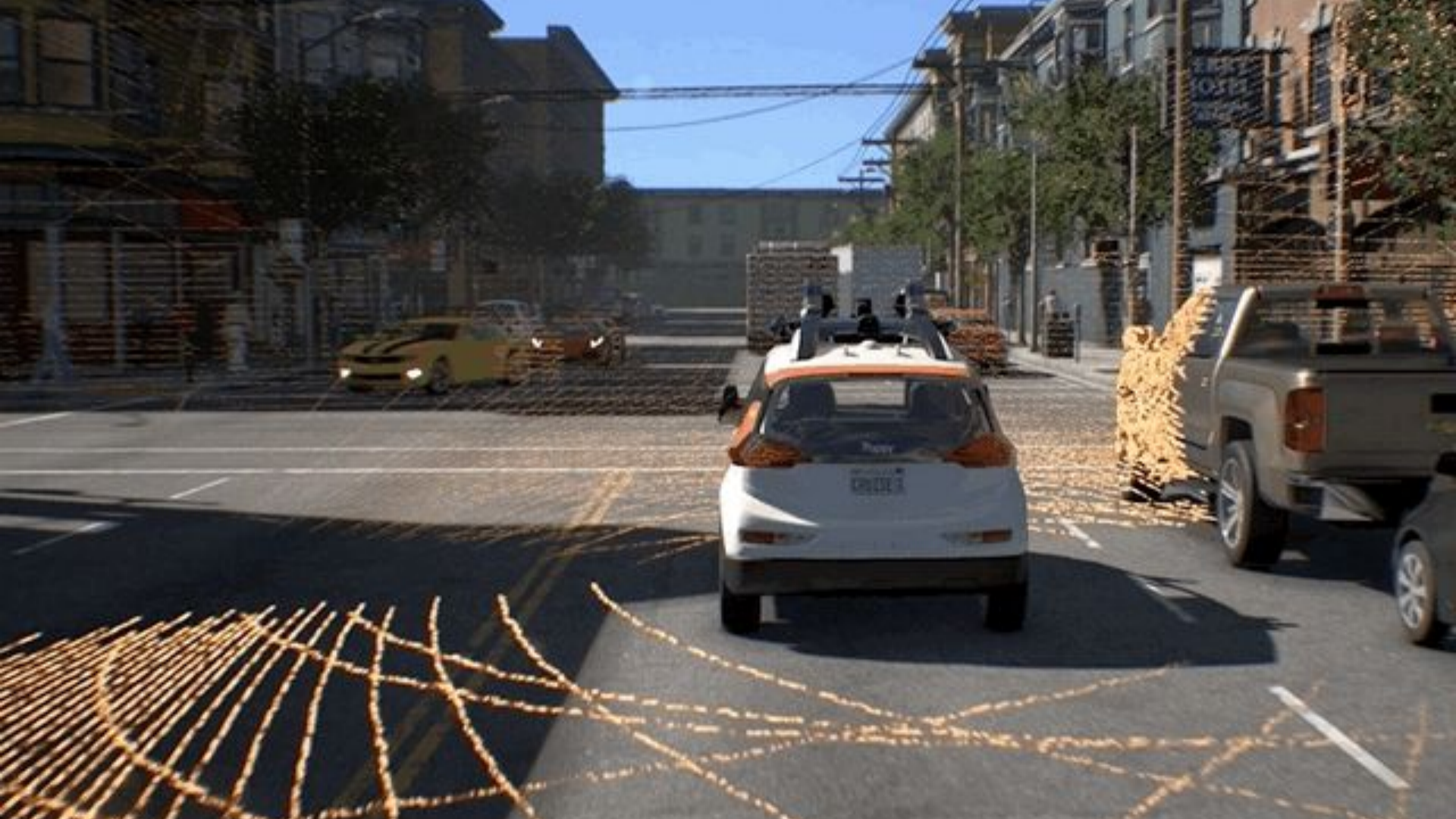
Who

Marc-Antoine Paré

Tech Lead, Privacy Infrastructure

Previously: differential privacy for the Department of Energy's "Secure Energy Algorithm Testbed"

Contents



cruise

Meta-Mapping

Document Past Data Mapping Efforts

An application for manual tagging of data across Drives, Segments, Events, and Scenarios for machine learning workflows

A manually maintained mapping of applications that handle data with data sensitivity classification

A spreadsheet that classifies each stream of data collected by Cruise AVs (almost 2,000 streams!) as sensitive/non-sensitive data.

A dataset that lists data types for Cruise with additional metadata about desired retention periods.

A data lineage system for one data platform (Google Cloud Dataflow)

A data catalog for Machine Learning datasets.

An attempted “automated service inventory” application that was never launched.

A manual maintained “access and deletion” spreadsheet enumerates almost 2,000 data fields across Cruise datasets (all SaaS applications)

An instance of the GCP Data Catalog to surface AV data for analysts, but metadata has not been updated since December 2020. The two Slack channels related to this project have been archived.

A pilot installation of OneTrust by the Legal team.

Lessons **Learned** from Past Data Mapping Efforts

Lesson 1

Manual labelling isn't enough

Lesson 2

Mapping is context specific (road map vs topographic maps)

Lesson 3

Aggressively trim scope

Lesson 4

Take use cases to the finish line (you've labeled data, so what?)

Lessons **Applied** to Our MVP

Lesson 1

Manual labelling isn't enough

Build automated sensitive data detectors

Lesson 2

Mapping is context specific (road map vs topographic maps)

Map for Privacy Engineering only

Lesson 3

Aggressively trim scope

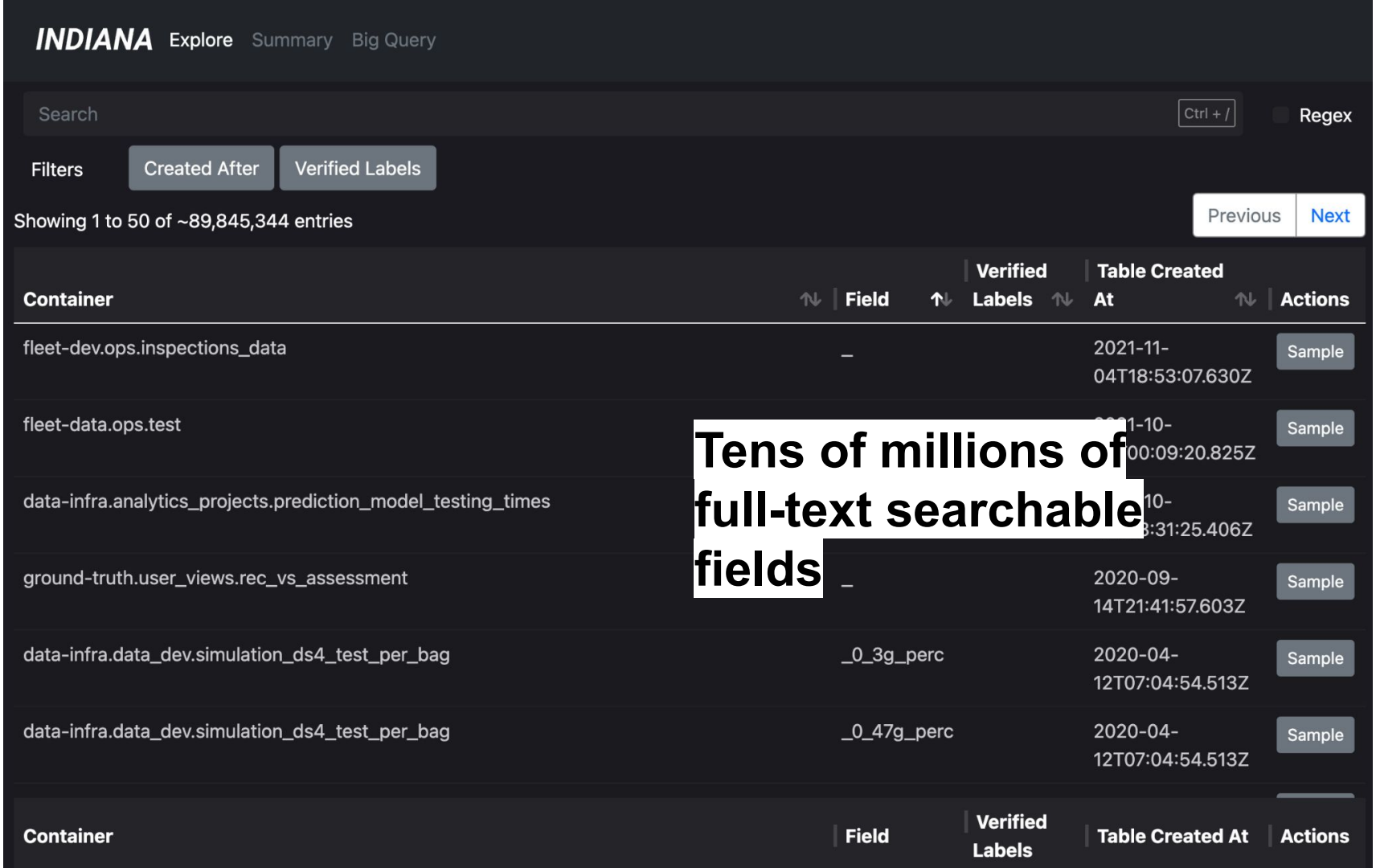
Focus (at first) on two high value BigQuery projects

Lesson 4

Take use cases to the finish line (you've labeled data, so what?

Archive abandoned sensitive data

What We Built


INDIANA Explore Summary Big Query
Search
Ctrl + /
Regex
Filters
Created After
Verified Labels
Showing 1 to 50 of ~89,845,344 entries
Previous
Next
Container
Field
Verified Labels
Table Created At
Actions
fleet-dev.ops.inspections_data
_
2021-11-04T18:53:07.630Z
Sample
fleet-data.ops.test
Sample
data-infra.analytics_projects.prediction_model_testing_times
Sample
ground-truth.user_views.rec_vs_assessment
_
2020-09-14T21:41:57.603Z
Sample
data-infra.data_dev.simulation_ds4_test_per_bag
_0_3g_perc
2020-04-12T07:04:54.513Z
Sample
data-infra.data_dev.simulation_ds4_test_per_bag
_0_47g_perc
2020-04-12T07:04:54.513Z
Sample
Tens of millions of full-text searchable fields

INDIANA Explore Summary Big Query

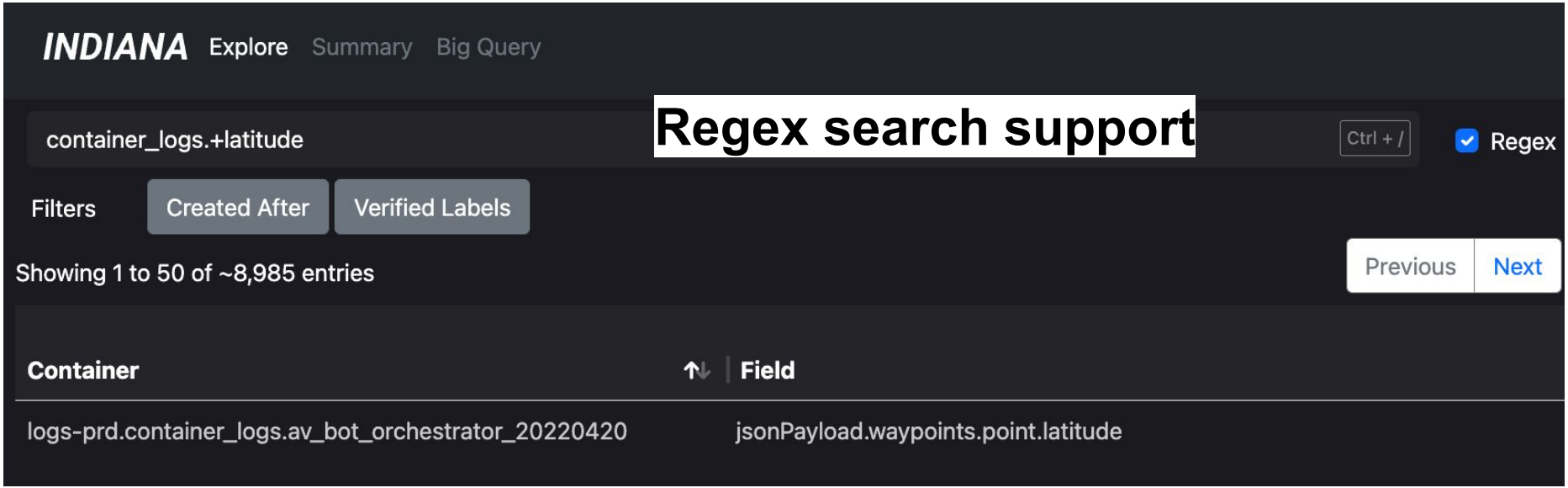
Ad-hoc metadata searches

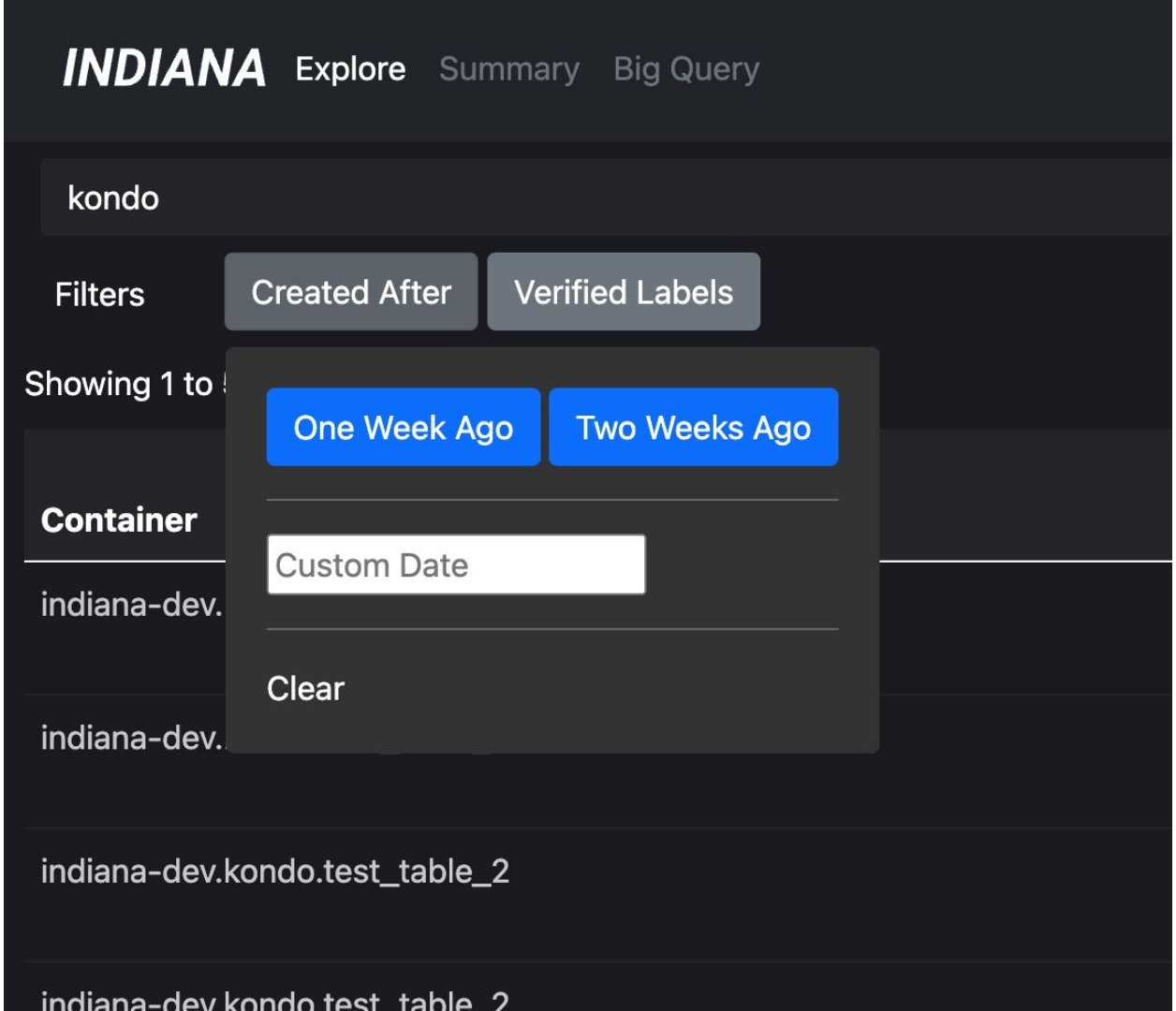
image_raw

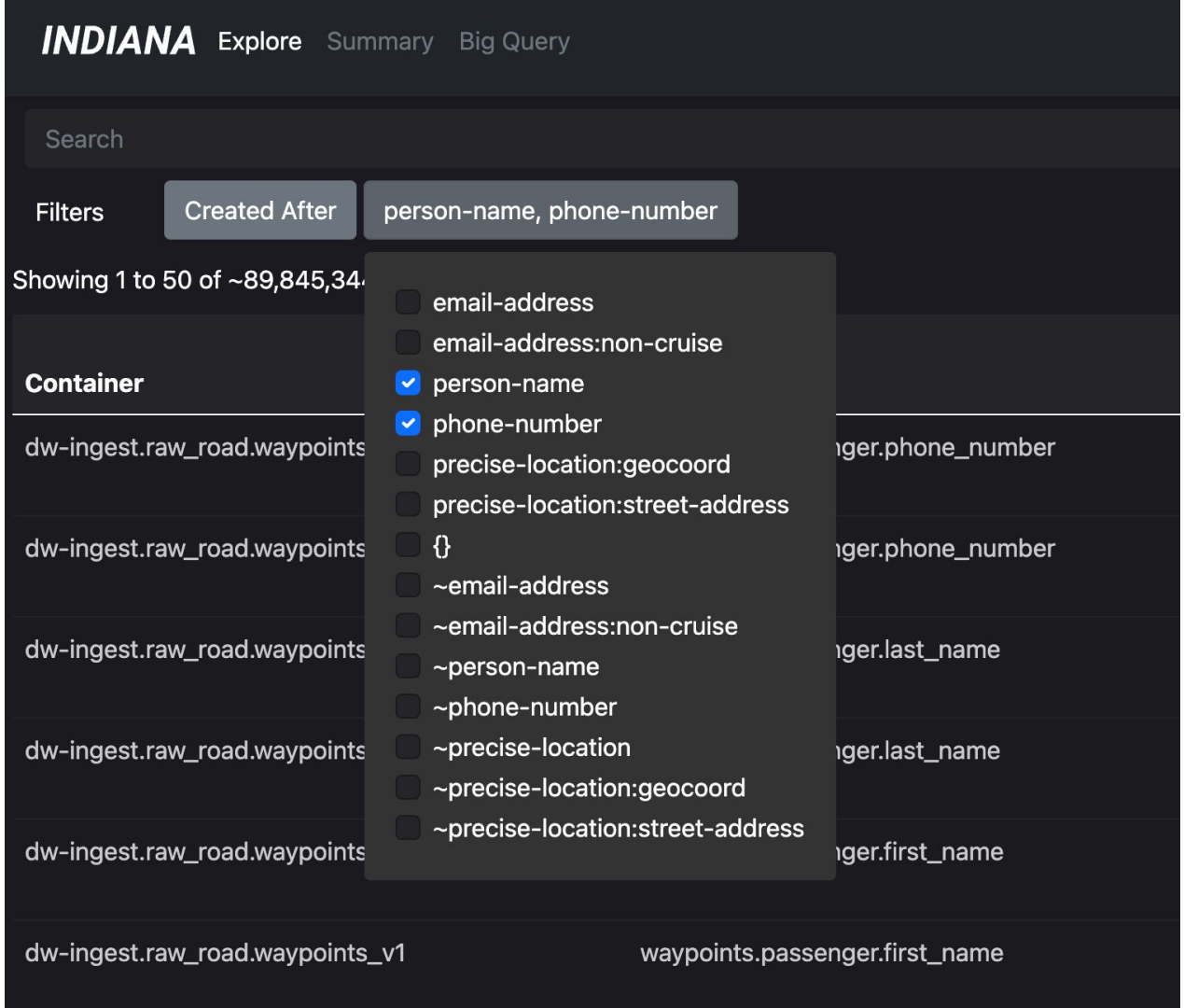
Filters | Created After | Verified Labels

Showing 1 to 50 of ~9,146 entries

Container	Field	V La
ml-prod.datasets.dataset_trial_collection	g2_high_res_front_image_raw	
ml-prod.datasets.dataset_trial_collection	g2_med_res_front_image_raw	
ml-stg.datasets.dataset_trial_collection	g2_high_res_front_image_raw	
ml-stg.datasets.dataset_trial_collection	g2_med_res_front_image_raw	
ml-dev.datasets.dataset_trial_collection	g2_high_res_front_image_raw	
ml-dev.datasets.dataset_trial_collection	g2_med_res_front_image_raw	


INDIANA
Explore
Summary
Big Query
container_logs.+latitude
Regex search support
Ctrl + /
Regex
Filters
Created After
Verified Labels
Showing 1 to 50 of ~8,985 entries
Previous
Next
Container
Field
logs-prd.container_logs.av_bot_orchestrator_20220420
jsonPayload.waypoints.point.latitude


INDIANA
Explore
Summary
Big Query
kondo
Filters
Created After
Verified Labels
Showing 1 to
One Week Ago
Two Weeks Ago
Custom Date
Clear
Container
indiana-dev.
indiana-dev.
indiana-dev.kondo.test_table_2
indiana-dev.kondo.test_table_2


INDIANA
Explore
Summary
Big Query
Search
Filters
Created After
person-name, phone-number
Showing 1 to 50 of ~89,845,34
email-address
email-address:non-cruise
person-name
phone-number
precise-location:geocoord
precise-location:street-address
{}
~email-address
~email-address:non-cruise
~person-name
~phone-number
~precise-location
~precise-location:geocoord
~precise-location:street-address
Container
dw-ingest.raw_road.waypoints
ger.phone_number
dw-ingest.raw_road.waypoints
ger.phone_number
dw-ingest.raw_road.waypoints
ger.last_name
dw-ingest.raw_road.waypoints
ger.last_name
dw-ingest.raw_road.waypoints
ger.first_name
dw-ingest.raw_road.waypoints_v1
waypoints.passenger.first_name

INDIANA Explore Summary Big Query

Search Ctrl + / Regex

Filters Created After phone-number

Showing 1 to 50 of ~89,845,344 entries

Previous Next

Container	Field	Verified Labels	Table Created At
dw-ingest.raw_road.waypoints_v1	waypoints.passenger.phone_number	phone-number	2020-10-02T01:05:41.452Z
dw-ingest-stg.raw_road.waypoints_v1	waypoints.passenger.phone_number	phone-number	2020-09-16T01:30:59.165Z
fleet-dev.commercial.tmp_ux	phone_number	phone-number	2022-01-28T19:13:24.075Z
fleet-dev.commercial.ux	phone_no	phone-number	2021-05-18T03:42:05.028Z
fleet-dev.commercial.ux	operator_phone_no	phone-number	2021-09-03T14:12:09.731Z

Activity

All | Comments | Work Log | History | Activity | Transitions | Time In Status

Links Hierarchy

added a comment - 13/Aug/21 9:52 AM

Marc Paré do any of these tables have sensitive data as flagged by indiana?

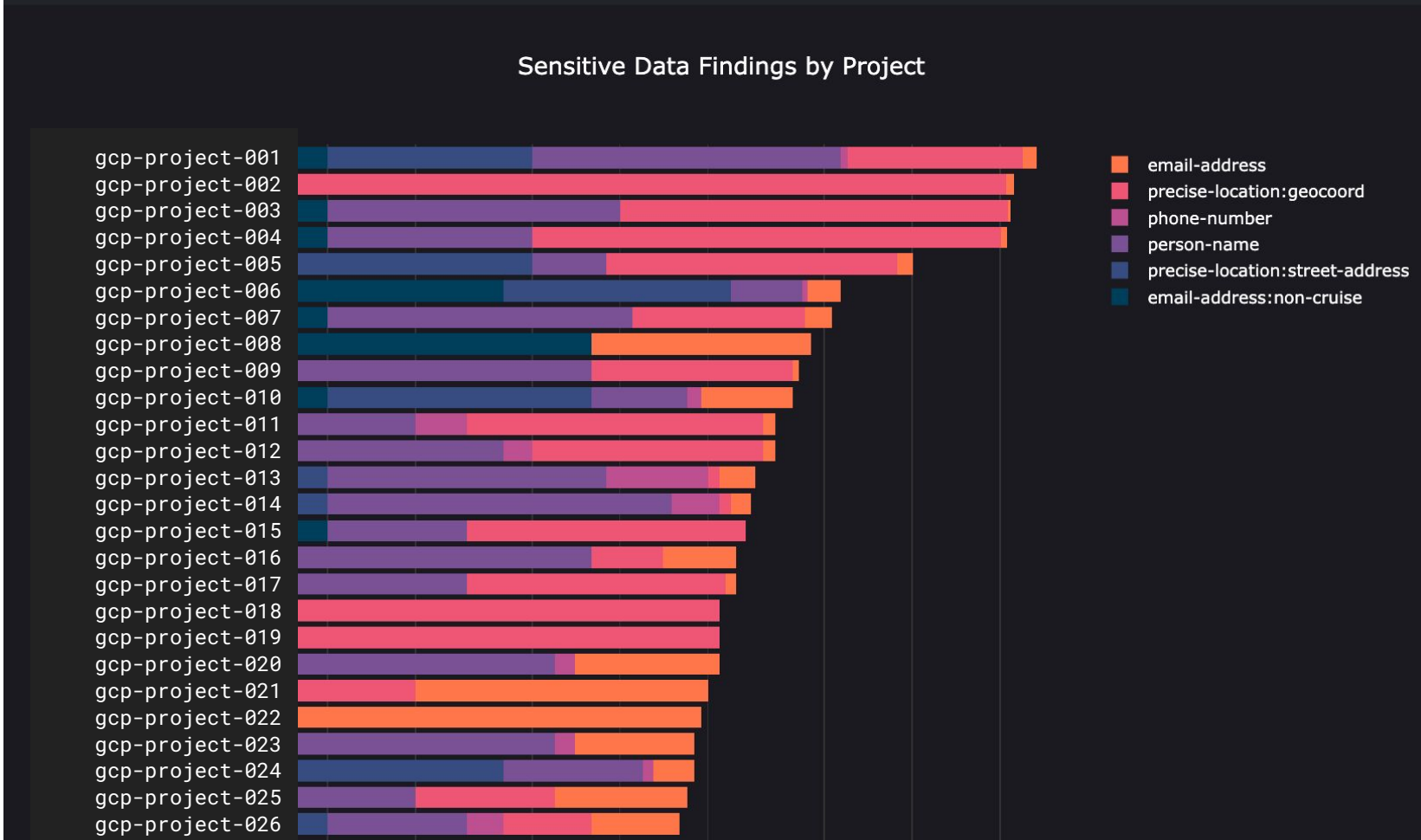
 Marc Paré added a comment - 13/Aug/21 9:56 AM

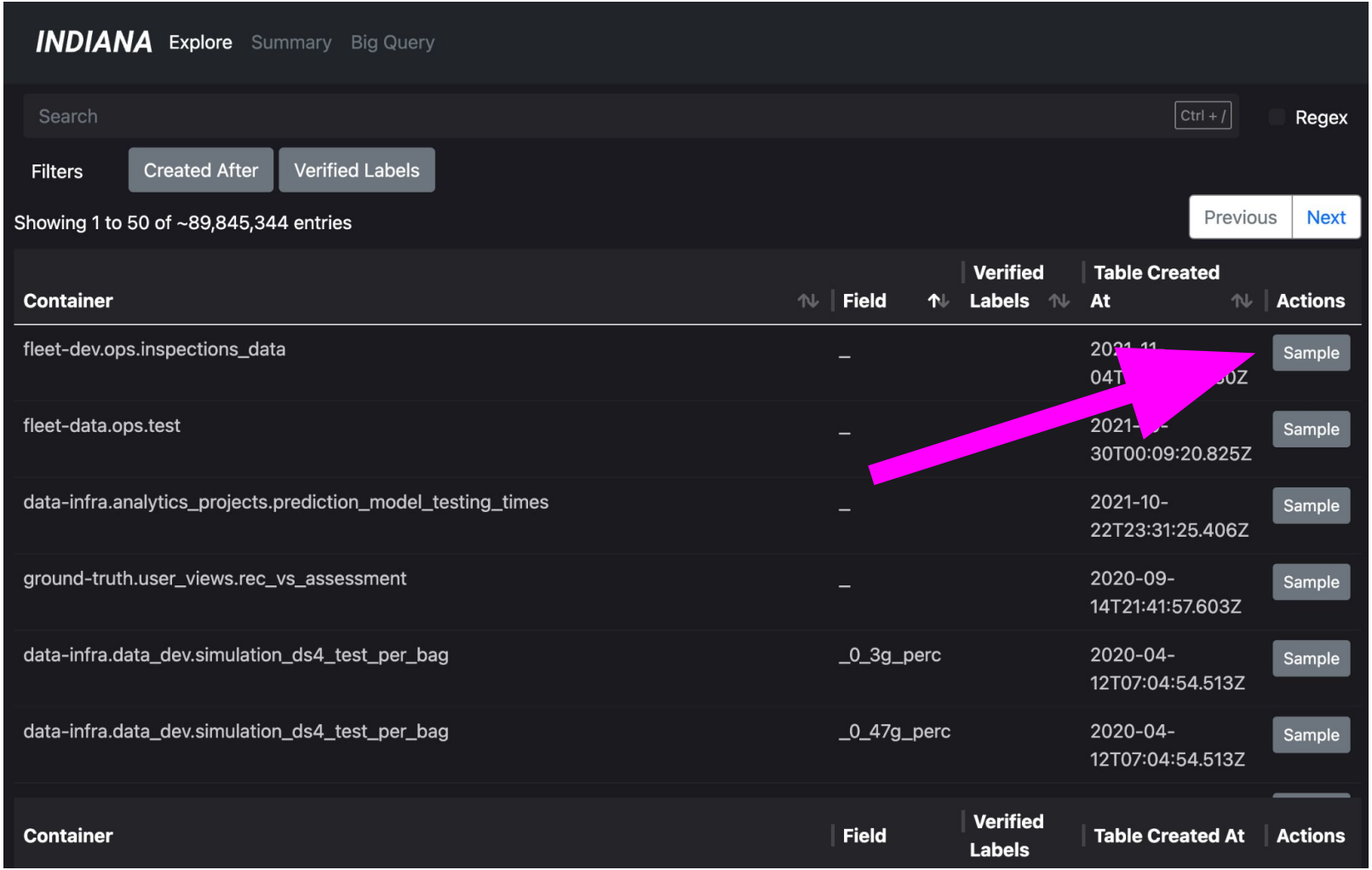
Quite a bit (197 tagged fields)

https://indiana.dev.paasapps.robot.car/?query=analytics%7Cbppe%7Ccalibration%7Ccamera_cloud_calibration%7Ccar_logs%7Ccar_metr address%3Anon-cruise%2Cperson-name%2Cphone-number%2Cprecise-location%3Ageocoord%2Cprecise-location%3Astreet-address

---

INDIANA Explore Summary Big Query




INDIANA
Explore
Summary
Big Query
Search
Ctrl + /
Regex
Filters
Created After
Verified Labels
Showing 1 to 50 of ~89,845,344 entries
Previous
Next
Container
Field
Verified Labels
Table Created At
Actions
fleet-dev.ops.inspections_data
_
Sample
fleet-data.ops.test
_
2021-
30T00:09:20.825Z
Sample
data-infra.analytics_projects.prediction_model_testing_times
_
2021-10-22T23:31:25.406Z
Sample
ground-truth.user_views.rec_vs_assessment
_
2020-09-14T21:41:57.603Z
Sample
data-infra.data_dev.simulation_ds4_test_per_bag
_0_3g_perc
2020-04-12T07:04:54.513Z
Sample
data-infra.data_dev.simulation_ds4_test_per_bag
_0_47g_perc
2020-04-12T07:04:54.513Z
Sample
Container
Field
Verified Labels
Table Created At
Actions

Web Application Sample Queries

- **Ad-hoc data mapping**
  - Are there any columns with the text “aws_secret”?
  - Does a field with a suspicious name actually store sensitive information?
- **Handling data access requests**
  - Are the tables that a contractor is requesting access to sensitive?
    - Do any of the fields have verified labels?
    - Can I spot check all the fields in the schema?
- **Data flow visibility**
  - Have any sensitive fields been created in the last two weeks?
- **Summary Visualization for risk assessment**
  - Which BigQuery projects have the most sensitive data?
  - Which BigQuery projects store customer phone numbers?

cruise

Metadata Scans

Schema ingestion with Google Cloud Asset Inventory

Implemented as a BigQuery query

```
SCANS = [
{
    "tag": "precise-location:street-address",
    "scanner": RegexScan("(?<!ip)address"),
    "weight": 0.5
},
{
    "tag": "precise-location:geocoord",
    "scanner": RegexScan("latitude|longitude|dropoff_loc|pickup_loc|vehicle_i
    "weight": 0.5
},
{
    "tag": "precise-location:geocoord",
    "scanner": RegexScan("arrived_waypoint|next_destination_(lat|long)|approx
    "weight": 0.7
},
{
    "tag": "person-name",
    "scanner": RegexScan("first_name|firstname|lastname|last_name"),
    "weight": 1.0
},
{
    "tag": "phone-number",
    "scanner": RegexScan("phone_number|phone"),
    "weight": 1.0
},
{
    "tag": "email-address",
    "scanner": RegexScan("email"),
    "weight": 1.0
},
```

Content Scans

Full content scanned using the Google Cloud DLP “Inspect” API using a combination of ML and rules-based matchers

Scanners were highly customized for accuracy on Cruise data (none worked out of the box)

Data Type	Metadata Scan Coverage	Metadata Scan Precision	Content Scan	Content Scan Precision
Non-Cruise Email Address	98%	LOW	100%	HIGH
Street Address	50%	HIGH	100%	MED
Geolocation	50%	HIGH	100%	MED
Person Name	10%	HIGH	100%	MED
Phone Number	100%	HIGH	100%	LOW

Scanning Locked Down Data

Least Complex

Most Complex

Blanket READ access
for Privacy Engineering

In-situ scanners

Scanning Locked Down Data

Least Complex

Most Complex

Blanket READ access
for Privacy Engineering

In-situ scanners

Secure data sampling service

Scanning Locked Down Data

```
https://trowel/v1/
 sample/google/bigquery/<project>/<dataset>/<table>
```

=>

Returns 1,000 randomly sampled rows
Results are cached for 24 hours

```
SELECT *
  FROM `{table_id}`
  TABLESAMPLE SYSTEM (1 PERCENT)
LIMIT 1000
```

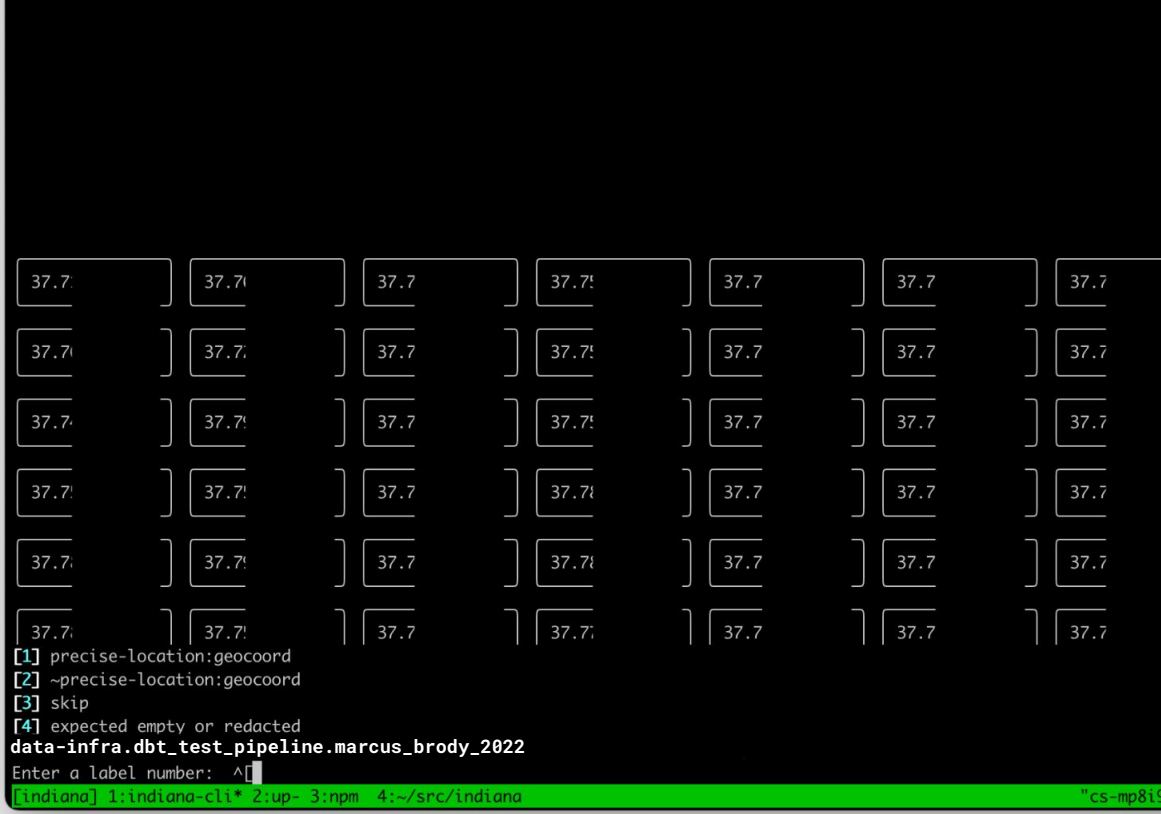
Scanning Locked Down Data

Trowel Threat Mitigation

- Greatly simplified abuse monitoring
- Technical controls for large scale data exfiltration
- Technical controls for targeted data searches
- Prevent over-provision of Privacy Engineering user accounts

cruise

Human in the loop label verification



cruise

Automated Scanning Summary

Covered **100%** of BigQuery projects

14PB of underlying data covered

Millions tables scanned

Tens of millions of fields scanned

Thousands of sensitive fields identified

Use Case: Archiving Abandoned Data

Surprisingly, these valuable data are mostly abandoned. Sensitive data mapping by the Indiana initiative uncovered high rates of abandonment for sensitive data. **55% of tables flagged with sensitive tables have no queries in the last 3 months.** Looking back a year, the number doesn't change much. 43% of tables have no queries in the last twelve months.

Explorer

+ ADD DATA

EDITOR

DTIEN_C...

COMPOSE NEW

_archived

Found 42 results. Broaden search to all projects.

- cruise-data-infra-pro...
  - analytics_projects
    - VRU_sce_tko_rdi_ba...
    - dtien_cruise_path_tt...
    - dtien_sce_proxy_CIR...
    - dtien_sce_proxy_cro...
    - dtien_sce_proxy_eva...
    - dtien_sce_proxy_eva...
    - dtien_sce_proxy_eva...
    - dtien_sce_proxy_eva...
    - dtien_sce_proxy_gho...
    - dtien_sce_proxy_psc...
    - dtien_sce_proxy_psc...

dtien_cruise_p...

QUERY SHARE COPY DELETE EX

SCHEMA DETAILS PREVIEW

Table schema

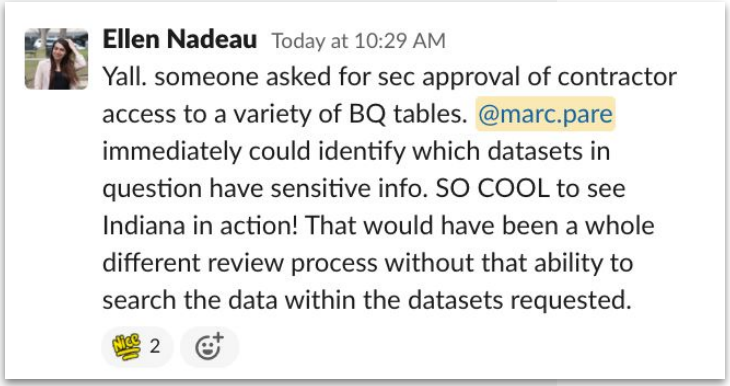
Filter Enter property name or value

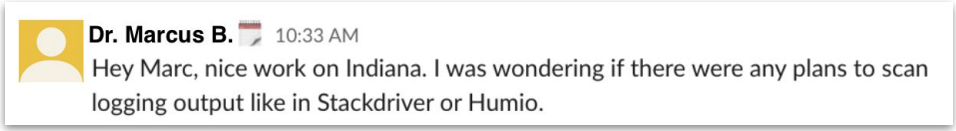
Field name	Type	Mode	Policy Tags
_vin	STRING	NULLABLE	privacy : archived
sp_header	RECORD	NULLABLE	
seq	INTEGER	NULLABLE	privacy : archived
stamp	RECORD	NULLABLE	
secs	INTEGER	NULLABLE	privacy : archived
nsecs	INTEGER	NULLABLE	privacy : archived
frame_id	STRING	NULLABLE	privacy :

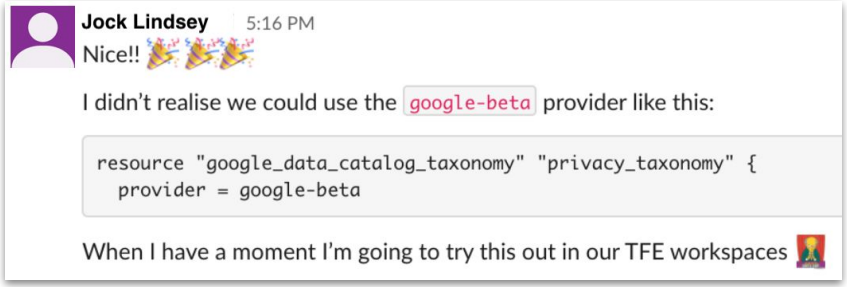
You don't have access to this field.

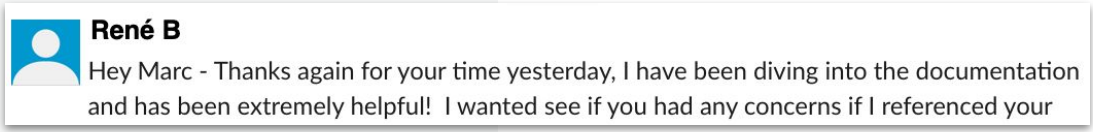
Many Teams Impacted Along the Way

- Enabled **Business Continuity and Disaster Response**
- Found gaps in a **Detection and Response** data stream
- Found gaps in a **Data Infra** data access logs
- **Legal** used Indiana and its background research to develop data buckets for law enforcement requests
- Experience with BigQuery Column-Level permissions informed the design of improved **People** team access controls









We’re hiring!

getcruise.com/careers