Data Mapping at a Billion Dollar Self-Driving Startup

PEPR 2022
Marc-Antoine Paré
Tech Lead, Privacy Infrastructure

Previously: differential privacy for the Department of Energy’s “Secure Energy Algorithm Testbed”
Contents

Introduction to the privacy challenge of AVs

Background on Data Mapping at Cruise

What We Built

Use Cases and Findings
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 The two Slack channels related to this project have been archived.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
Tens of millions of full-text searchable fields
<table>
<thead>
<tr>
<th>Container</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml-prod.datasets.dataset_trial_collection</td>
<td>g2_high_res_front_image_raw</td>
</tr>
<tr>
<td>ml-prod.datasets.dataset_trial_collection</td>
<td>g2_med_res_front_image_raw</td>
</tr>
<tr>
<td>ml-stg.datasets.dataset_trial_collection</td>
<td>g2_high_res_front_image_raw</td>
</tr>
<tr>
<td>ml-stg.datasets.dataset_trial_collection</td>
<td>g2_med_res_front_image_raw</td>
</tr>
<tr>
<td>ml-dev.datasets.dataset_trial_collection</td>
<td>g2_high_res_front_image_raw</td>
</tr>
<tr>
<td>ml-dev.datasets.dataset_trial_collection</td>
<td>g2_med_res_front_image_raw</td>
</tr>
</tbody>
</table>

Ad-hoc metadata searches
### Regex search support

**Filters**
- Created After
- Verified Labels

**Showing 1 to 50 of ~8,985 entries**

<table>
<thead>
<tr>
<th>Container</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>logs-prd.container_logs.av_bot_orchestrator_20220420</td>
<td>jsonPayload.waypoints.point.latitude</td>
</tr>
<tr>
<td>Container</td>
<td>Field</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>dw-ingest.raw_road.waypoints_v1</td>
<td>waypoints.passenger.phone_number</td>
</tr>
<tr>
<td>dw-ingest-stg.raw_road.waypoints_v1</td>
<td>waypoints.passenger.phone_number</td>
</tr>
<tr>
<td>fleet-dev.commercial.ux</td>
<td>phone_no</td>
</tr>
<tr>
<td>fleet-dev.commercial.ux</td>
<td>operator_phone_no</td>
</tr>
</tbody>
</table>
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%3Aon-cruise%2Cperson-name%2Cphone-number%2Cprecise-location%3Ageocoord%2Cprecise-location%3Astreet-address
<table>
<thead>
<tr>
<th>Container</th>
<th>Field</th>
<th>Verified Labels</th>
<th>Table Created At</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>fleet-dev.ops.inspections_data</td>
<td></td>
<td>-</td>
<td>2021-11-04T00:00:00Z</td>
<td>Sample</td>
</tr>
<tr>
<td>fleet-data.ops.test</td>
<td></td>
<td>-</td>
<td>2021-10-30T00:00:00Z</td>
<td>Sample</td>
</tr>
<tr>
<td>data-infra.analytics_projects.prediction_model_testing_times</td>
<td></td>
<td>-</td>
<td>2021-10-22T23:31:25Z</td>
<td>Sample</td>
</tr>
<tr>
<td>ground-truth.user_views.rec_vs_assessment</td>
<td></td>
<td>-</td>
<td>2020-09-14T21:57:03Z</td>
<td>Sample</td>
</tr>
<tr>
<td>data-infra.data_dev.simulation_ds4_test_per_bag</td>
<td>_0_3g_perc</td>
<td></td>
<td>2020-04-12T07:04:54</td>
<td>Sample</td>
</tr>
<tr>
<td>data-infra.data_dev.simulation_ds4_test_per_bag</td>
<td>_0_47g_perc</td>
<td></td>
<td>2020-04-12T07:04:54</td>
<td>Sample</td>
</tr>
</tbody>
</table>
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?
Metadata Scans

Schema ingestion with Google Cloud Asset Inventory

Implemented as a BigQuery query
## 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).

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Metadata Scan Coverage</th>
<th>Metadata Scan Precision</th>
<th>Content Scan</th>
<th>Content Scan Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cruise Email Address</td>
<td>98%</td>
<td>LOW</td>
<td>100%</td>
<td>HIGH</td>
</tr>
<tr>
<td>Street Address</td>
<td>50%</td>
<td>HIGH</td>
<td>100%</td>
<td>MED</td>
</tr>
<tr>
<td>Geolocation</td>
<td>50%</td>
<td>HIGH</td>
<td>100%</td>
<td>MED</td>
</tr>
<tr>
<td>Person Name</td>
<td>10%</td>
<td>HIGH</td>
<td>100%</td>
<td>MED</td>
</tr>
<tr>
<td>Phone Number</td>
<td>100%</td>
<td>HIGH</td>
<td>100%</td>
<td>LOW</td>
</tr>
</tbody>
</table>
Scanning Locked Down Data

Least Complex

Blanket READ access for Privacy Engineering

Most Complex

In-situ scanners
Scanning Locked Down Data

Least Complex

Blanket READ access for Privacy Engineering

Most Complex

In-situ scanners

Secure data sampling service
SELECT *
FROM `{table_id}`
TABLESAMPLE SYSTEM (1 PERCENT)
LIMIT 1000

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

Returns 1,000 randomly sampled rows
Results are cached for 24 hours
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
<table>
<thead>
<tr>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
<th>37.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
</tr>
<tr>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
</tr>
<tr>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
</tr>
<tr>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
</tr>
<tr>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
<td>37.7</td>
</tr>
</tbody>
</table>

[1] precise-location:geocoord
[3] skip
[4] expected empty or redirected

data-infra.dbt_test_pipeline.marcus_brody_2022

Enter a label number: [4]

(indiana) 1:indiana-cli* 2:up- 3:npm 4:~/.src/indiana
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
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.
### Table Schema

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type</th>
<th>Mode</th>
<th>Policy Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>_vin</td>
<td>STRING</td>
<td>NULLABLE</td>
<td>privacy: archived</td>
</tr>
<tr>
<td>sp_header</td>
<td>RECORD</td>
<td>NULLABLE</td>
<td>privacy: archived</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>seq</td>
<td>INTEGER</td>
<td>privacy: archived</td>
</tr>
<tr>
<td>stamp</td>
<td>RECORD</td>
<td>NULLABLE</td>
<td>privacy: archived</td>
</tr>
<tr>
<td></td>
<td>secs</td>
<td>INTEGER</td>
<td>privacy: archived</td>
</tr>
<tr>
<td></td>
<td>nsecs</td>
<td>INTEGER</td>
<td>privacy: archived</td>
</tr>
<tr>
<td></td>
<td>frame_id</td>
<td>STRING</td>
<td>privacy: archived</td>
</tr>
</tbody>
</table>
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

---

**Ellen Nadeau**  Today at 10:29 AM

Yall. someone asked for sec approval of contractor access to a variety of BQ tables. [@marc.pare](mailto:marc.pare) immediately could identify which datasets in question have sensitive info. SO COOL to see Indiana in action! That would have been a whole different review process without that ability to search the data within the datasets requested.

 влад

---

**Dr. Marcus B.**  10:33 AM

Hey Marc, nice work on Indiana. I was wondering if there were any plans to scan logging output like in Stackdriver or Humio.

---

**Jock Lindsey**  5:16 PM

Nice!! 🌞🌈

I didn't realise we could use the `google-beta` provider like this:

```yaml
resource "google_data_catalog_taxonomy" "privacy_taxonomy" {
  provider = google-beta
}
```

When I have a moment I'm going to try this out in our TFE workspaces

---

**René B**

Hey Marc - Thanks again for your time yesterday, I have been diving into the documentation and has been extremely helpful! I wanted see if you had any concerns if I referenced your
We’re hiring!

getcruise.com/careers