Sensitive data active catalog. How to control sensitive data in real time

Limassol, Cyprus 2024

PERCONA UNIVERSITY

Speaker

Aleksandr Sungurov

Information Security Architect

alexander.sungurov@exness.com

@Banzay021

Awesome company

- Delivery company
- A large number of customers
- Clients PII
- Card date
- A lot of data that needs to be processed quickly

- Annoying “bugs”

WHY

Hard to have full control over sensitive data

01

We don't know exactly where sensitive data are

02

Typically, there is no single approach to working with sensitive data

03

Difficult to track the movement of data

04

Data quality issues

05

It is long/impossible to search for all sensitive data storage locations

06

Difficult to find and localize a data leak

07

Lack of a culture of working with data

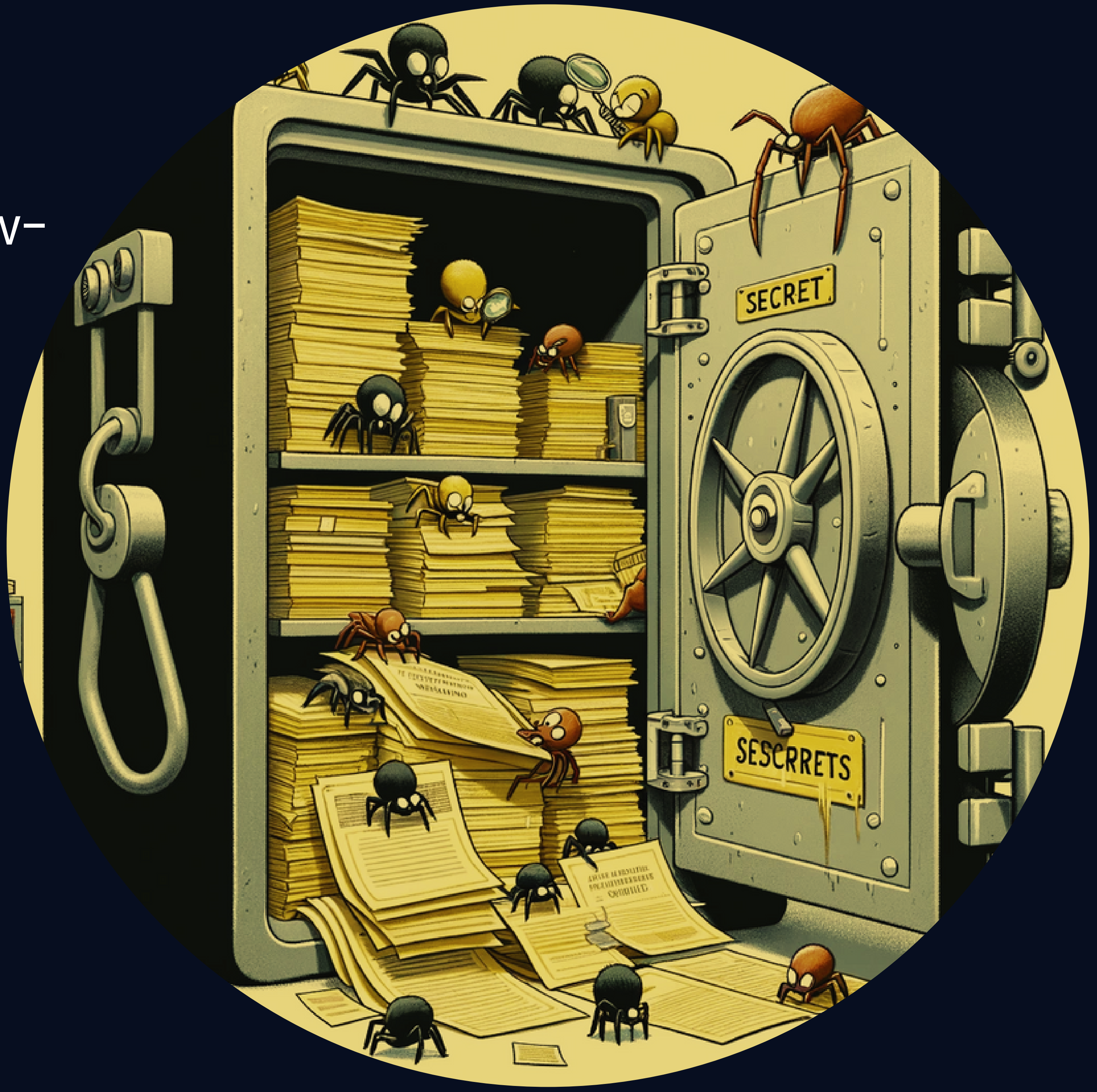
08

Classification of critical data

Critical

Payment card data
Personal data (clients)
Personal data (employees)
Business sensitive data: Company strategy and know-how
Business sensitive data: Company financial data
Business sensitive data: Trading Information
Financial data
Medical data
Video surveillance footage
User authentication data
Keys, passwords, secrets for financial operations
Masked / anonymized confidential data
System data
Crypto addresses
Internal corporate data
Public data

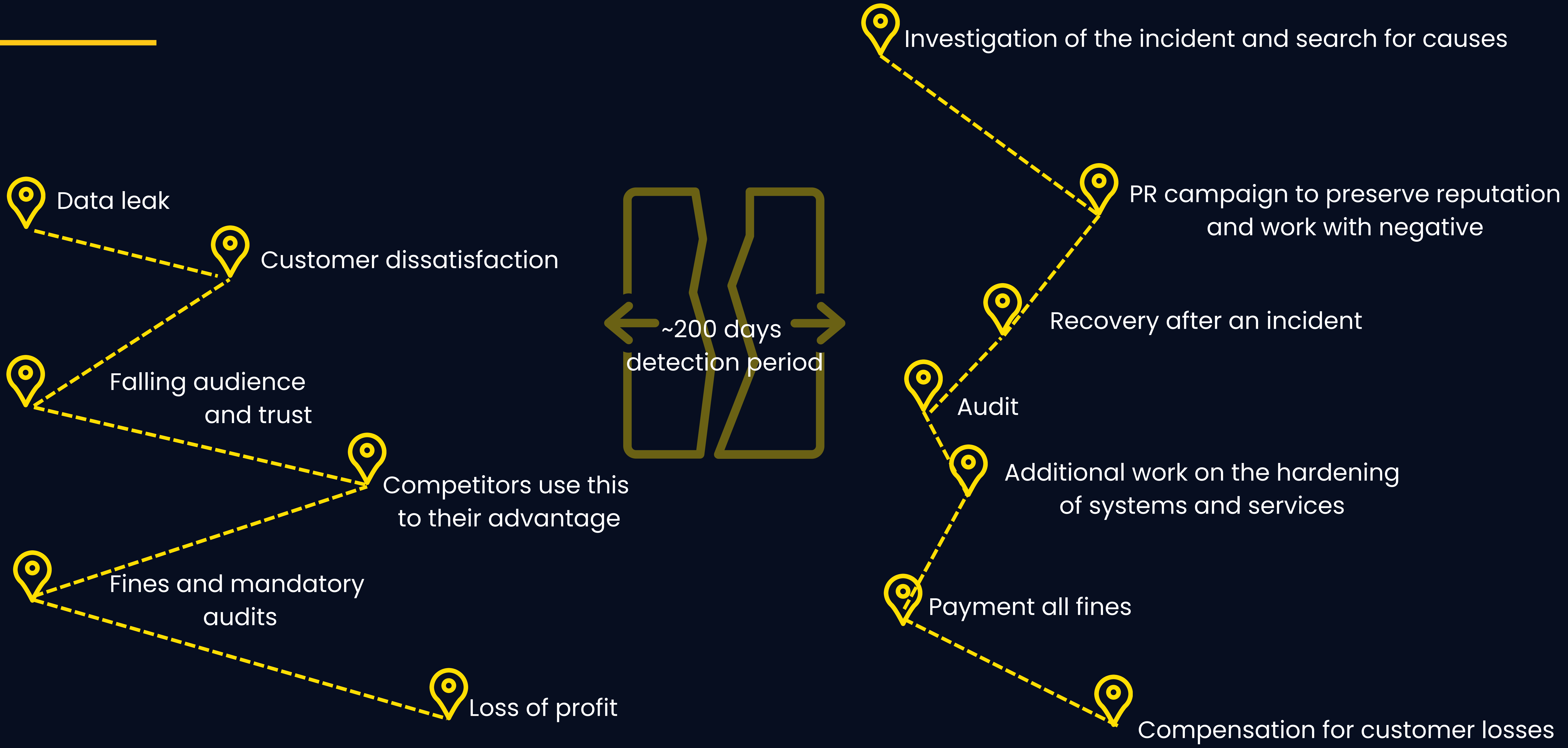
non-Critical



Classification of critical data

- Passwords
- API keys (both external and internal)
- One-time Confirmation/Signature Codes (OTP)
- Encryption Keys
- Authentication Tokens
- Sessions (client, applications)
- Client keys/Secrets
- Private keys (SSH, PGP, RSA, ECD, etc.)
- Verification codes
- Private cert key
- Signature Keys (PEP, CAP)
- Other secrets

What can this lead to



Data leak

Customer dissatisfaction

Falling audience and trust

Competitors use this to their advantage

Fines and mandatory audits

Loss of profit

~200 days detection period

Investigation of the incident and search for causes

PR campaign to preserve reputation and work with negative

Recovery after an incident

Audit

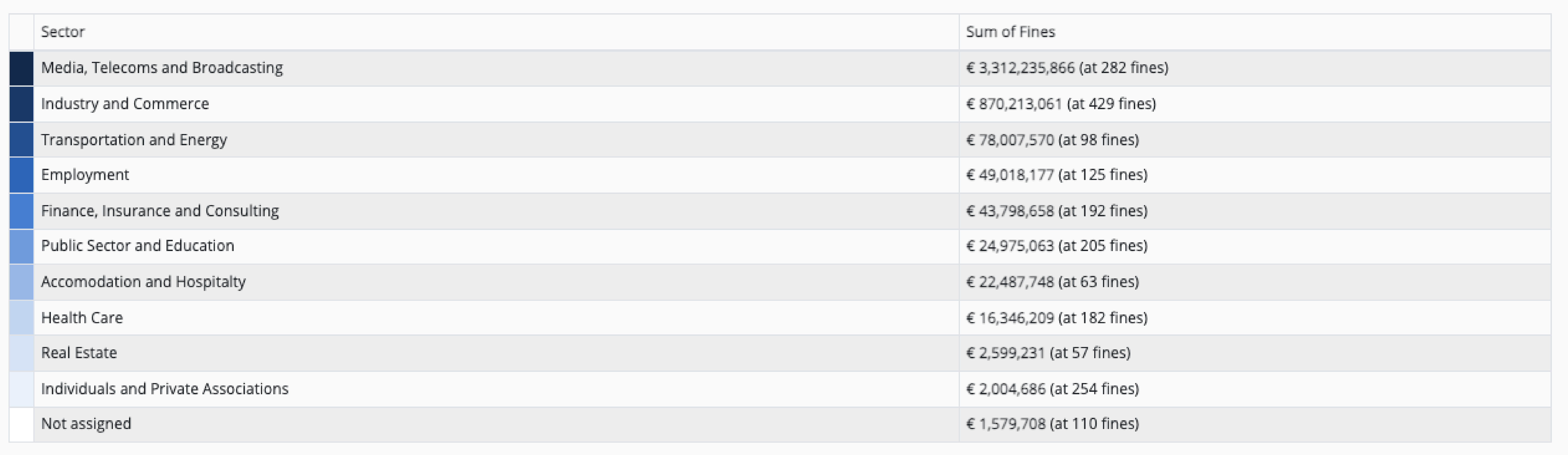
Additional work on the hardening of systems and services

Payment all fines

Compensation for customer losses

GDPR

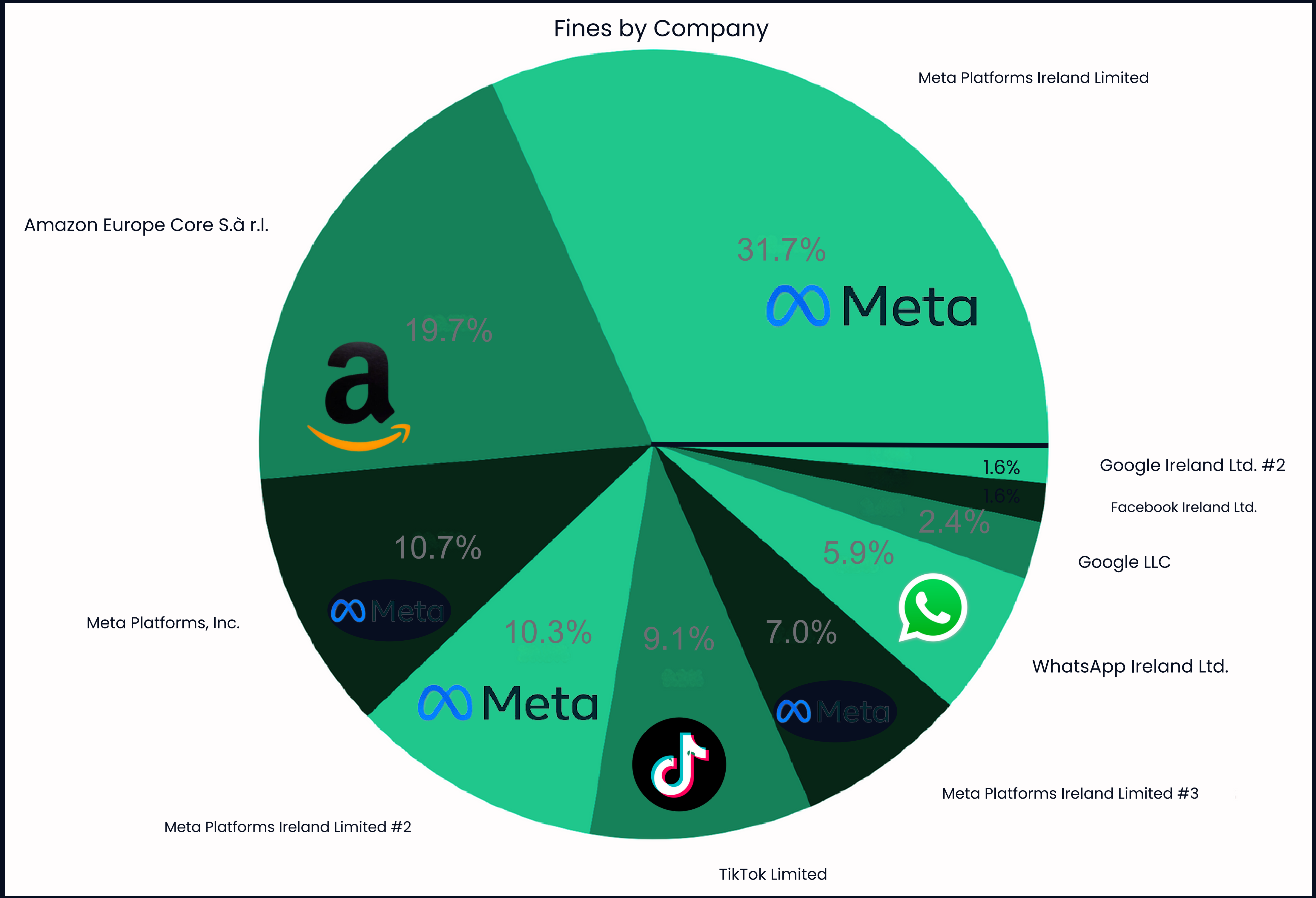
- Up to 4% of the company's annual turnover
- Determined by the results of an independent audit

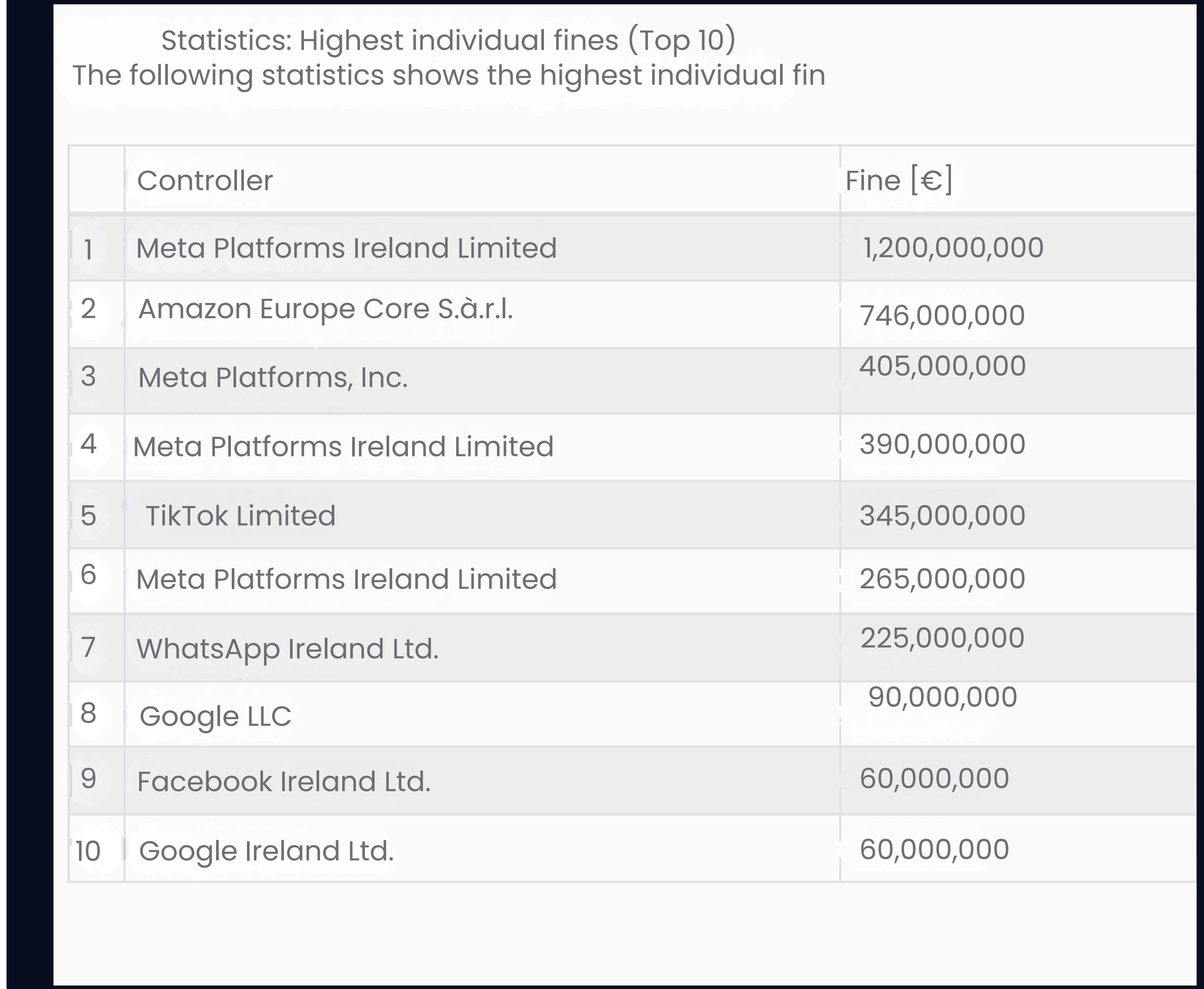


	Sector	Sum of Fines
■	Media, Telecoms and Broadcasting	€ 3,312,235,866 (at 282 fines)
■	Industry and Commerce	€ 870,213,061 (at 429 fines)
■	Transportation and Energy	€ 78,007,570 (at 98 fines)
■	Employment	€ 49,018,177 (at 125 fines)
■	Finance, Insurance and Consulting	€ 43,798,658 (at 192 fines)
■	Public Sector and Education	€ 24,975,063 (at 205 fines)
■	Accomodation and Hospitalty	€ 22,487,748 (at 63 fines)
■	Health Care	€ 16,346,209 (at 182 fines)
■	Real Estate	€ 2,599,231 (at 57 fines)
■	Individuals and Private Associations	€ 2,004,686 (at 254 fines)
□	Not assigned	€ 1,579,708 (at 110 fines)

GDPR Fines Statistics

GDPR





Statistics: Highest individual fines (Top 10)
The following statistics shows the highest individual fin

	Controller	Fine [€]
1	Meta Platforms Ireland Limited	1,200,000,000
2	Amazon Europe Core S.à.r.l.	746,000,000
3	Meta Platforms, Inc.	405,000,000
4	Meta Platforms Ireland Limited	390,000,000
5	TikTok Limited	345,000,000
6	Meta Platforms Ireland Limited	265,000,000
7	WhatsApp Ireland Ltd.	225,000,000
8	Google LLC	90,000,000
9	Facebook Ireland Ltd.	60,000,000
10	Google Ireland Ltd.	60,000,000

GDPR Fines Statistics

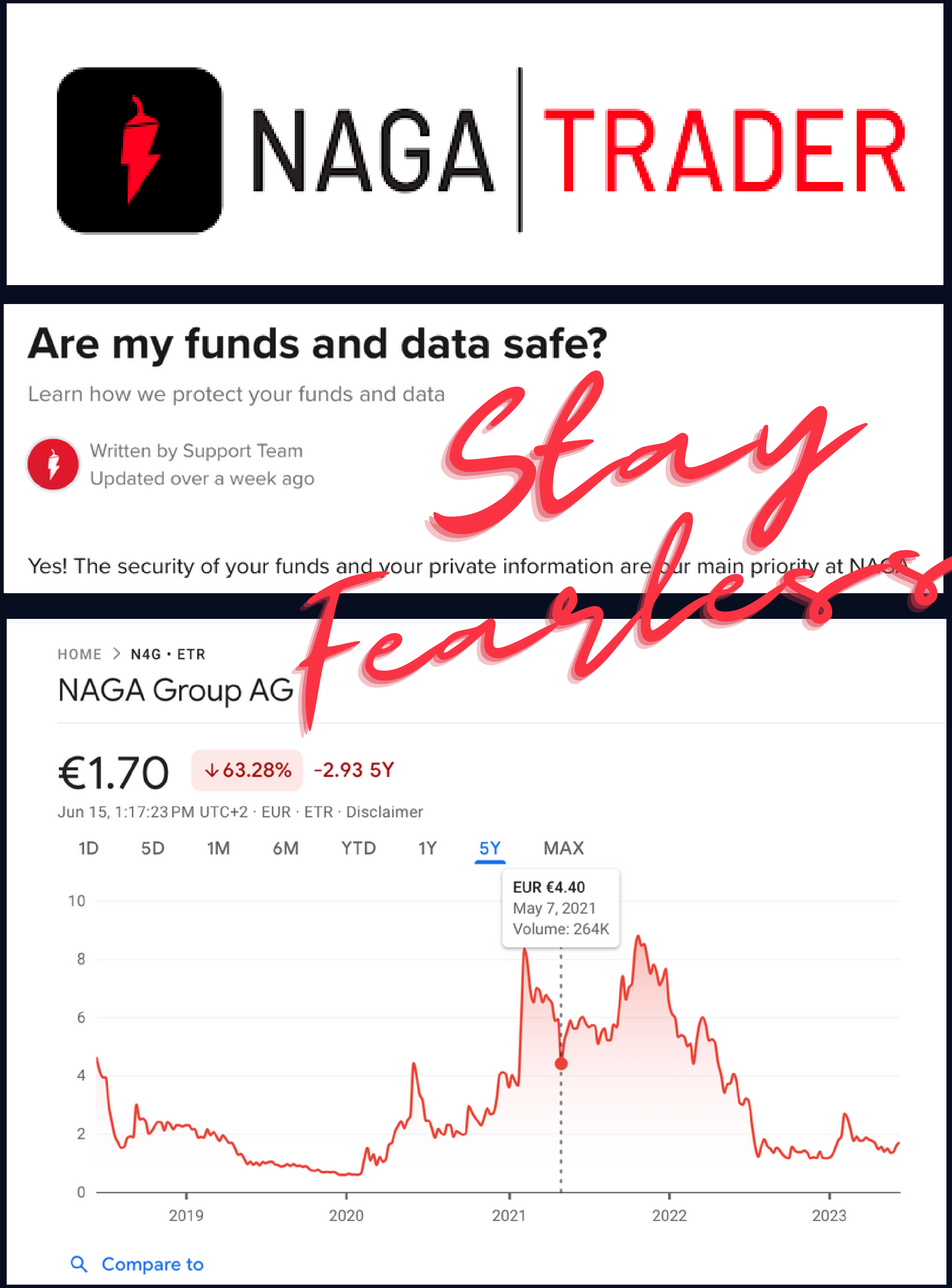
Example

Notification of a personal data breach by NAGA Markets Europe Ltd

NAGA Markets Europe Ltd reported a data breach in May 2021, where an unknown individual accessed their database.

This breach compromised the personal information of about 342,000 customers including:

- names
- postal addresses
- email addresses
- phone numbers.



Are my funds and data safe?

Learn how we protect your funds and data

Written by Support Team
Updated over a week ago

Yes! The security of your funds and your private information are our main priority at NAGA

HOME > N4G · ETR

NAGA Group AG

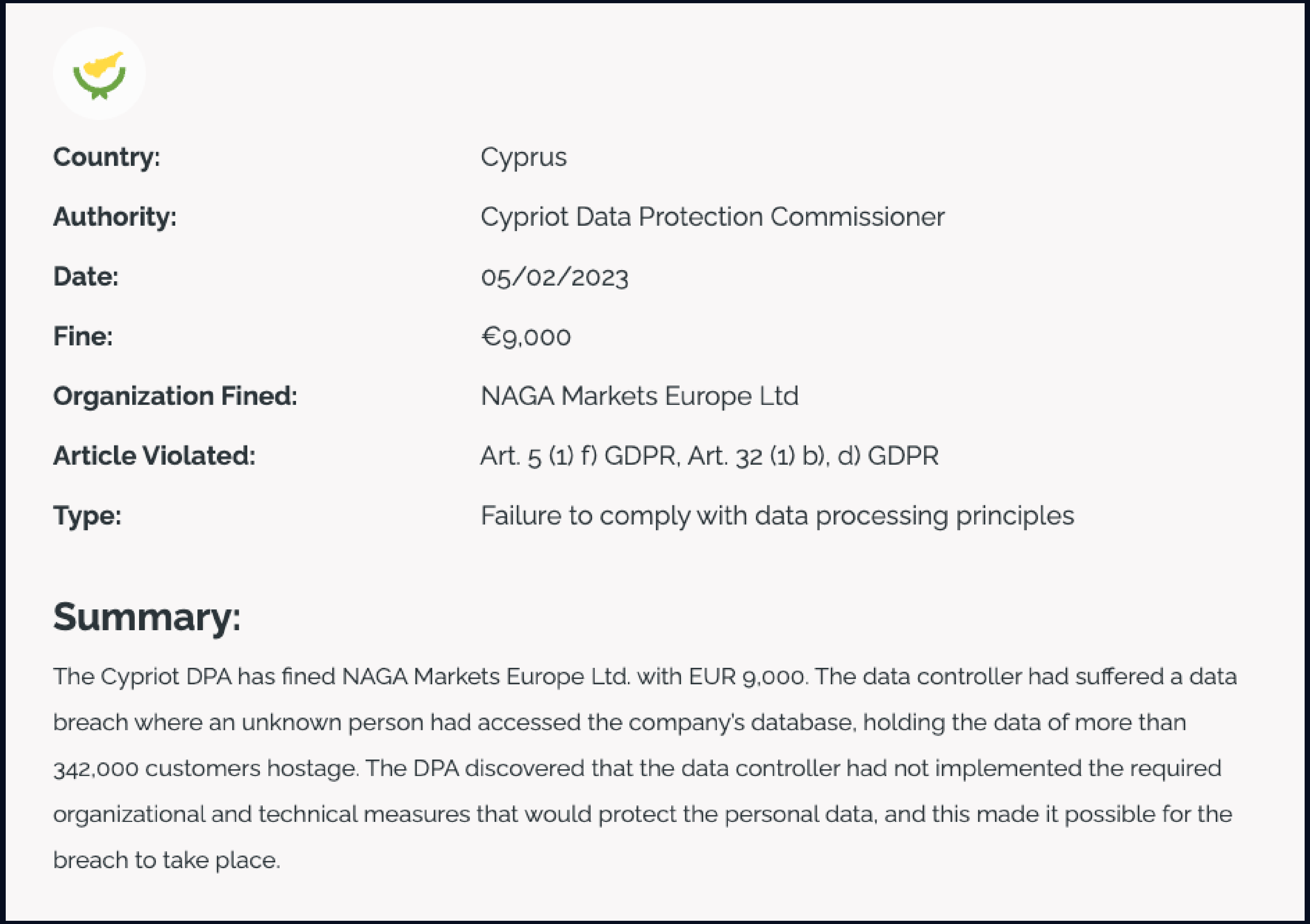
€1.70 ↓63.28% -2.93 5Y

Jun 15, 1:17:23 PM UTC+2 · EUR · ETR · Disclaimer

Example

The amount of the penalty for one account

9000 / 342 000 = 0,0263 euro per user



Country:	Cyprus
Authority:	Cypriot Data Protection Commissioner
Date:	05/02/2023
Fine:	€9,000
Organization Fined:	NAGA Markets Europe Ltd
Article Violated:	Art. 5 (1) f) GDPR, Art. 32 (1) b), d) GDPR
Type:	Failure to comply with data processing principles

Summary:

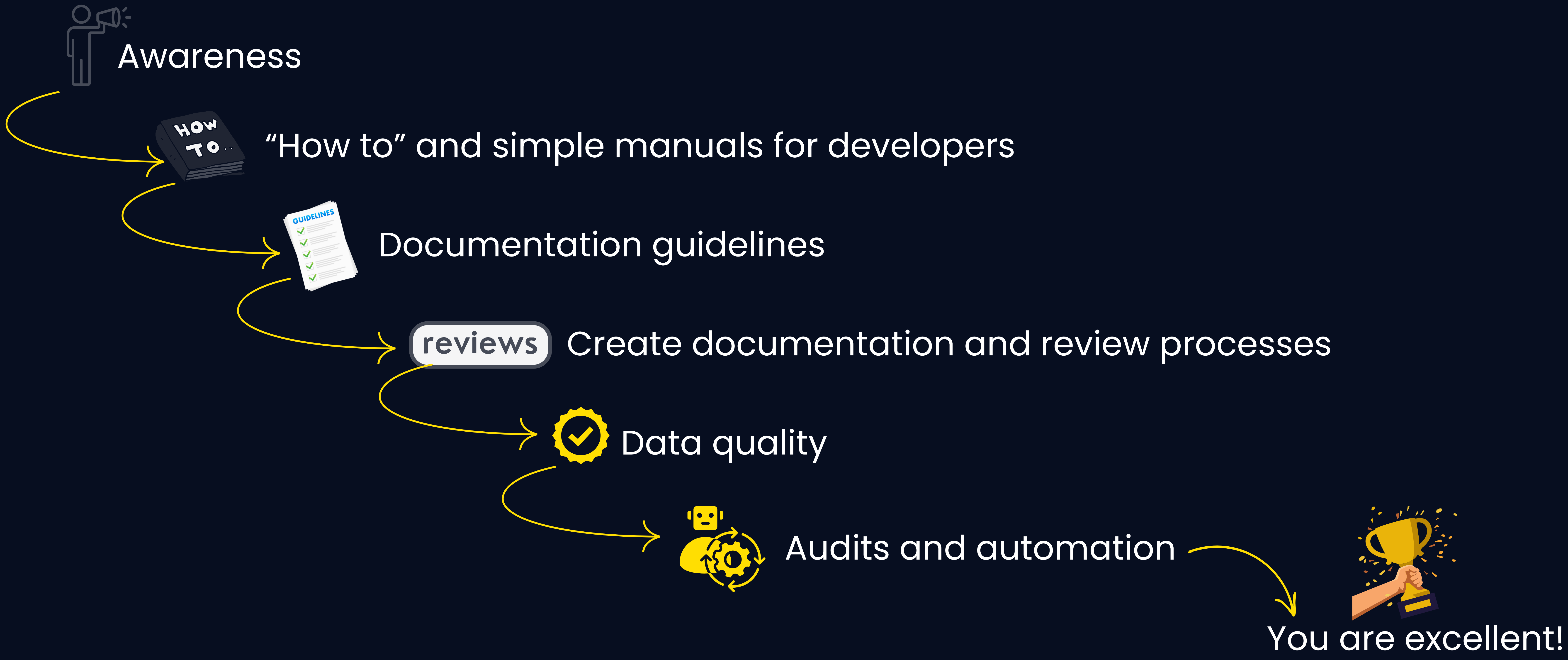
The Cypriot DPA has fined NAGA Markets Europe Ltd. with EUR 9,000. The data controller had suffered a data breach where an unknown person had accessed the company's database, holding the data of more than 342,000 customers hostage. The DPA discovered that the data controller had not implemented the required organizational and technical measures that would protect the personal data, and this made it possible for the breach to take place.

Path

What can lead us to Sensitive data active catalog ?

It takes time to assemble a spaceship...

Solutions



Awareness

"How to" and simple manuals for developers

Documentation guidelines

Create documentation and review processes

Data quality

Audits and automation

You are excellent!

What we need?

- Data owners
- Data quality metrics
- Data artifact inventory
- Data Usage Controls
- Event-Driven Approach
- Data Lifecycle

Processes

- Education (Data)
- Up to date “How to”
- Documentation process
- Data quality checks
- Sensitive data searching

Data active catalog

- Centralised store of metadata (producers, data schemas)

- Unified data pipelines and infra
- Message sampling services
- Policy as code for documentation

Tools and services

- Store metadata
- Search for data
- Message sampling services
- Policy as code tools

Data Catalogs

A data catalog is a tool designed to manage an organization's data assets. It provides a centralized inventory of available data, making it easier for users to find and understand data within an organization.

- Enhanced Data Discovery
- Improved Data Governance
- Better Collaboration
- Informed Decision Making

Open-source Data Catalogs

ODD

Backend: Postgres
Data Ingestion:
Postgres ✅
Vertica
ClickHouse ✅
DBT ✅
Kafka ✅
Argo
Tableau

Disadvantages:
No Custom sources
No Data Domains

link

Amundsen

Backend: Neo4j
+ Elastic
Data Ingestion:
Postgres ✅
Vertica ✅
ClickHouse
DBT
Kafka
Argo
Tableau

Disadvantages:
No Catalog of sources
No community support

link

Open Metadata

Backend: MySQL
+Elastic
Data Ingestion:
Postgres ✅
Vertica✅
ClickHouse✅
DBT✅
Kafka✅
Argo
Tableau✅

Disadvantages:
Requires the latest versions of all supporting products.

link

DataHub

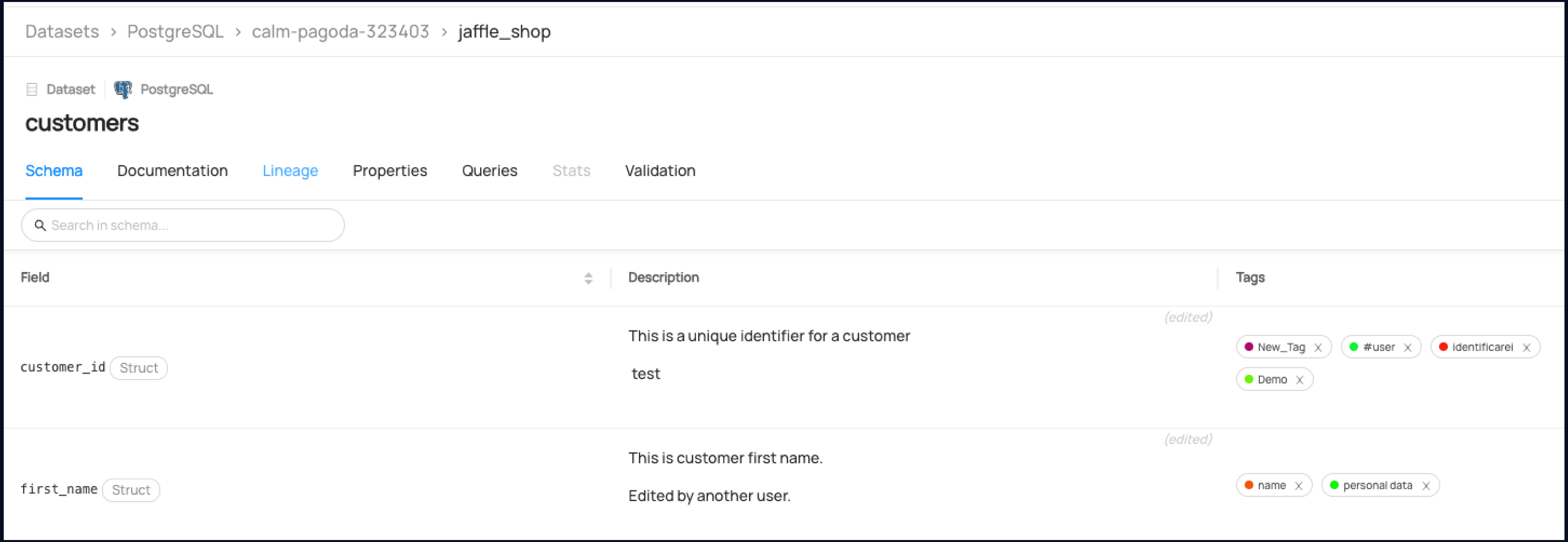
Backend: Neo4j
+ Elastic
Data Ingestion:
Postgres ✅
Vertica ✅
(can use via sqlalchemy)
ClickHouse ✅
DBT✅
Kafka✅
Argo
Tableau✅

Disadvantages:
No* (for us)

link

DataHub

DataHub is an open-source metadata platform for the modern data stack.



Datasets › PostgreSQL › calm-pagoda-323403 › jaffle_shop

Dataset PostgreSQL

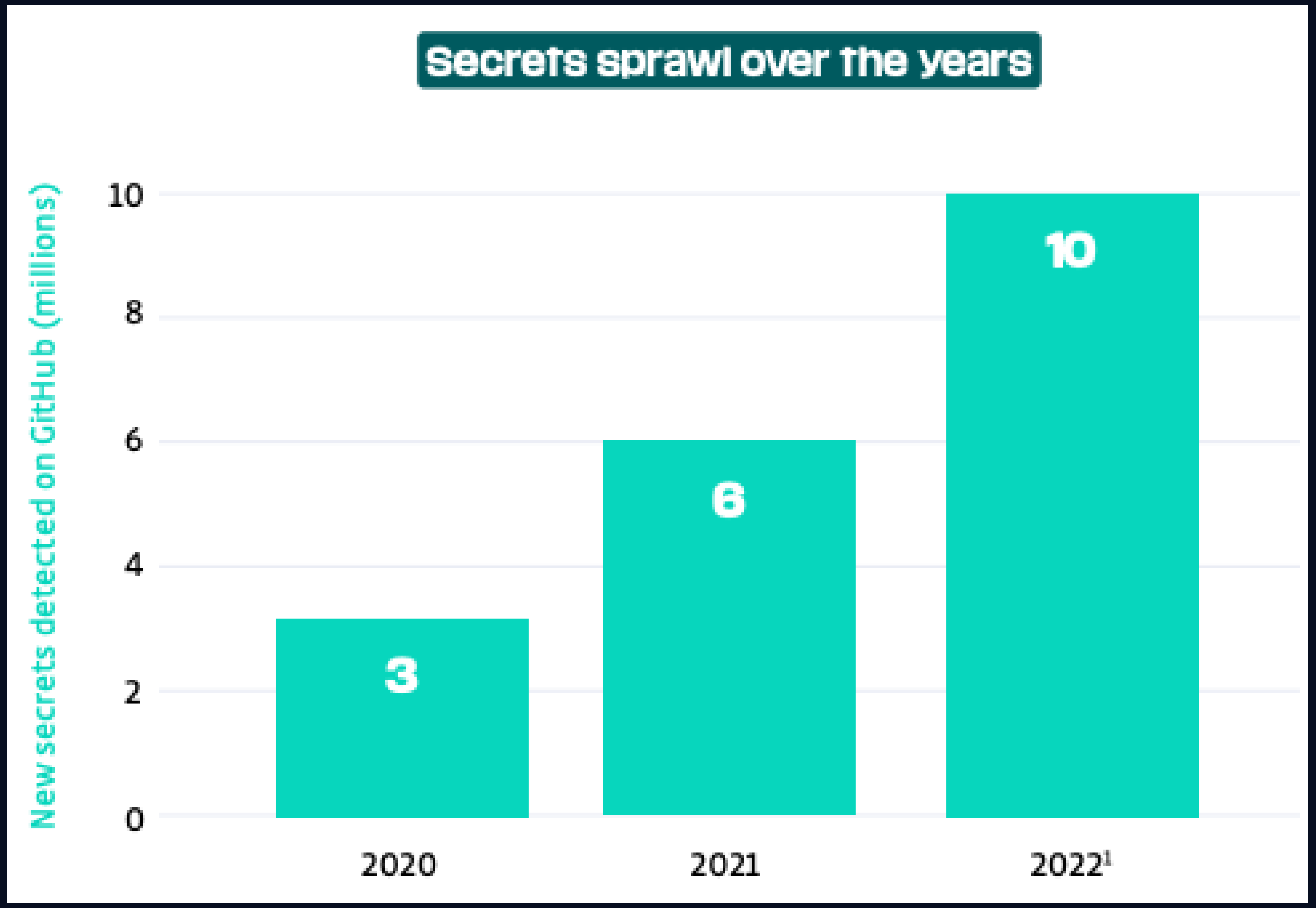
customers

Field	Description	Tags
customer_id Struct	*(edited)* This is a unique identifier for a customer test	New_Tag X #user X identificarel X Demo X
first_name Struct	*(edited)* This is customer first name. Edited by another user.	name X personal data X

Search for critical data tools

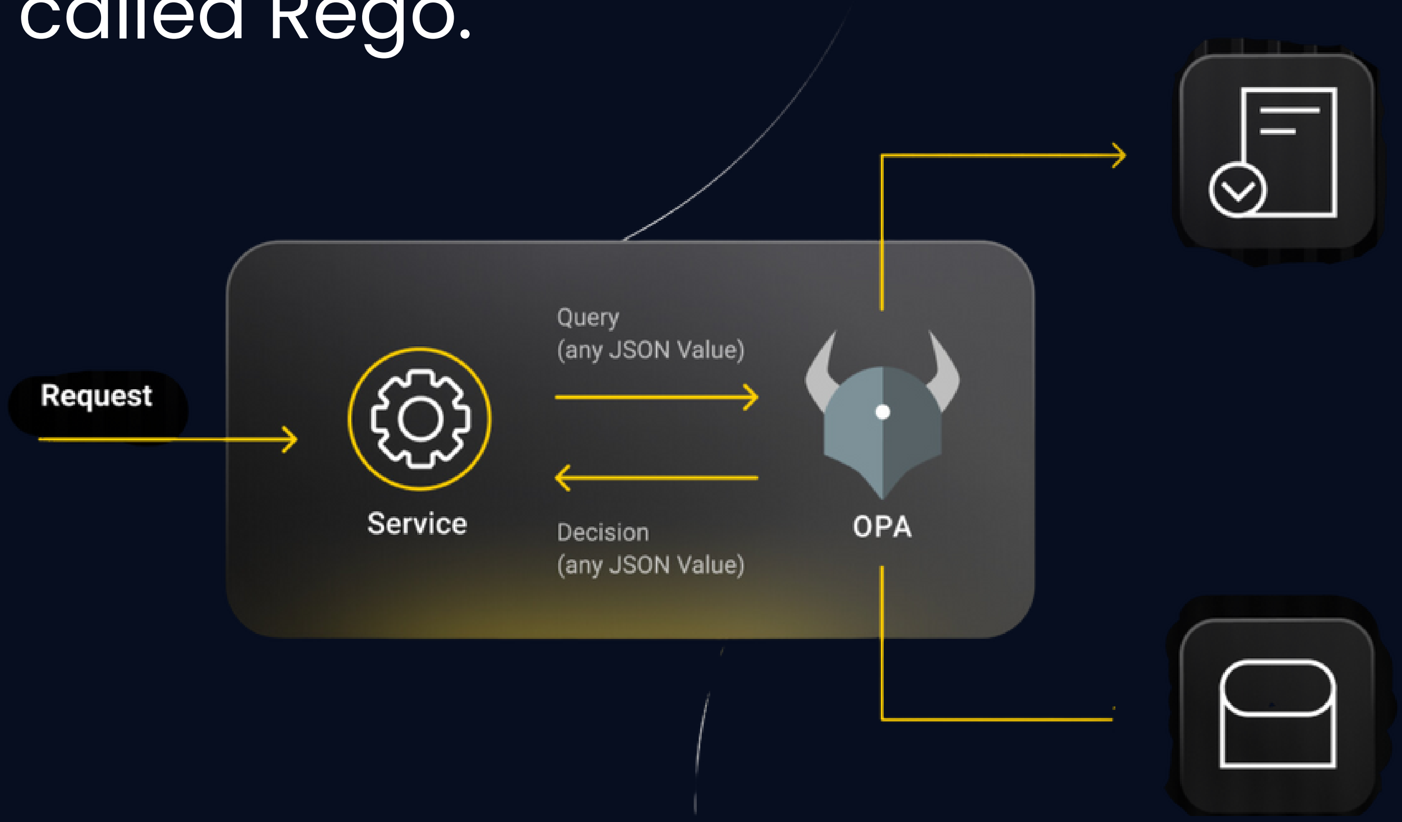
Git

- Gitleaks link
- Talisman link
- Trufflehog link



Policy as code tools

Open Policy Agent is an open source, general purpose policy engine created by the Cloud Native Computing Foundation. It provides a framework for policy as code in any domain, based on a high-level declarative language called Rego.



link

Schema Registry for Kafka

Schema Registry provides a centralized repository for managing and validating schemas for topic message data

link

Search for critical data tools

K8S

Intercepts and samples traffic

soveren

Data Bases

Need Agents

imperva

Data Security Fabric

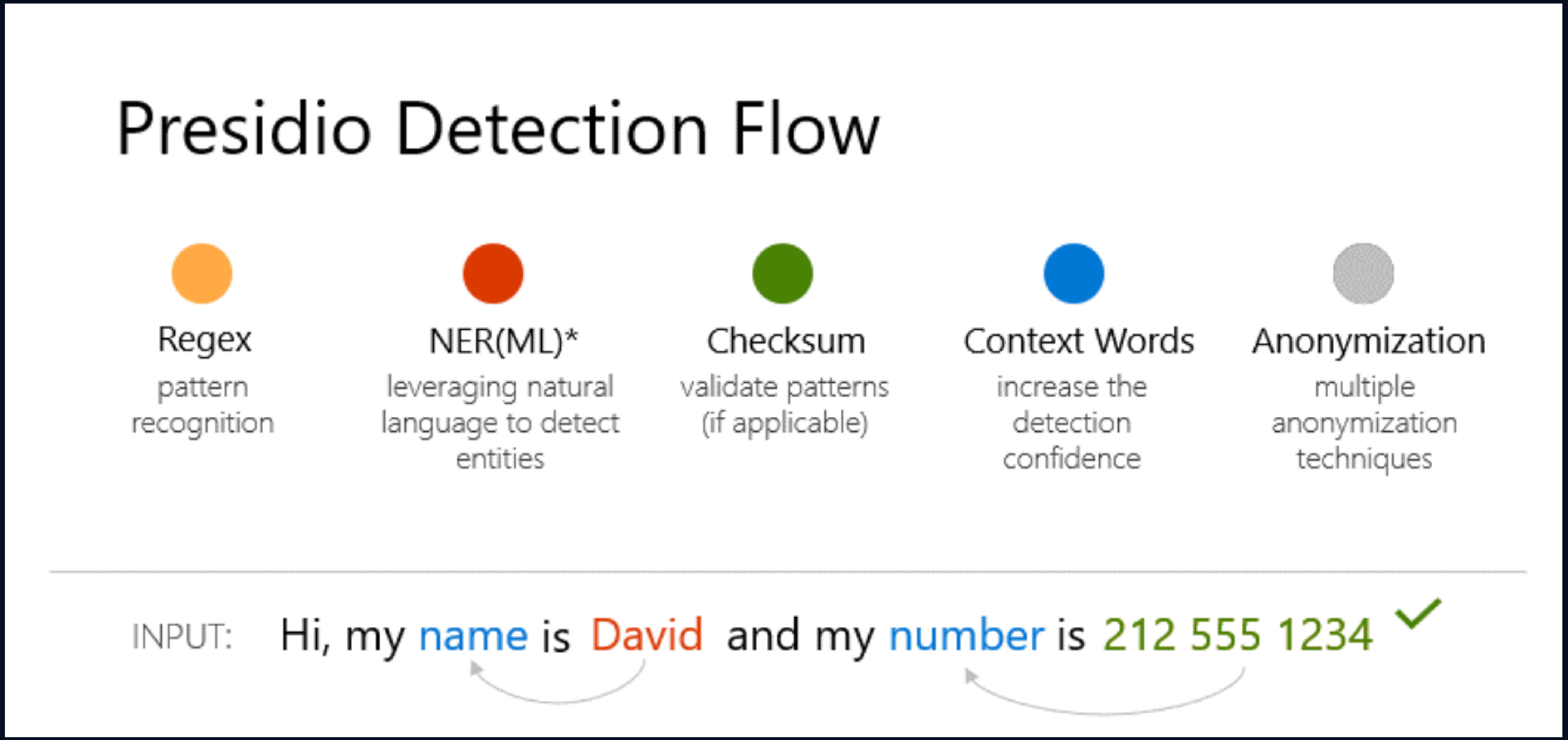
SPIRION™

TRICENT!

BigID

AI tool: Microsoft Presidio

Presidio (Origin from Latin praesidium ‘protection, garrison’) helps to ensure sensitive data is properly managed and governed.



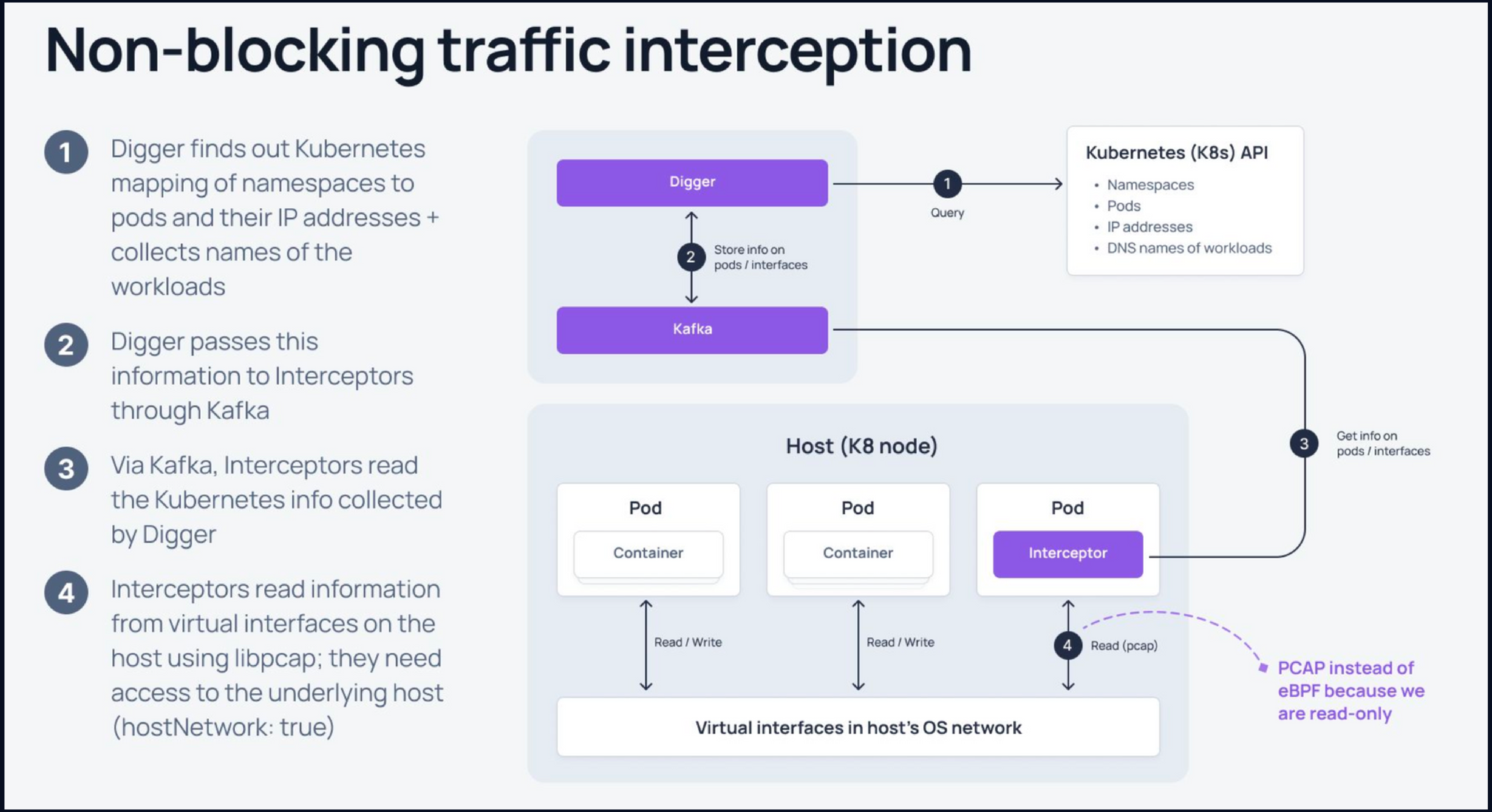
Presidio Detection Flow

- Regex: pattern recognition
- NER(ML)*: leveraging natural language to detect entities
- Checksum: validate patterns (if applicable)
- Context Words: increase the detection confidence
- Anonymization: multiple anonymization techniques

INPUT: Hi, my name is David and my number is 212 555 1234 ✓

link

Soveren



Non-blocking traffic interception

1. Digger finds out Kubernetes mapping of namespaces to pods and their IP addresses + collects names of the workloads
2. Digger passes this information to Interceptors through Kafka
3. Via Kafka, Interceptors read the Kubernetes info collected by Digger
4. Interceptors read information from virtual interfaces on the host using libpcap; they need access to the underlying host (hostNetwork: true)

PCAP instead of eBPF because we are read-only

Soveren data types

Right now Soveren works with the following data types:

- Person
- Passport
- Phone number
- IP address
- Birth date
- Tax number
- Email address
- Gender
- Pension number
- Location
- US Driver license
- Credit or debit Card
- IBAN

Solution Architecture

What we have

- Event driven approach
- Kafka as a service for teams
- K8S
- Data bases under load
- DataHub
- The desire to know about the quality of data and its movement

Approach

- Do not load databases with crawlers
- Sample messages from kafka
- Validate data schemas
- Minimize false positive detects
- Identify all critical data in the company
- Update information in Data Hub and CMDB

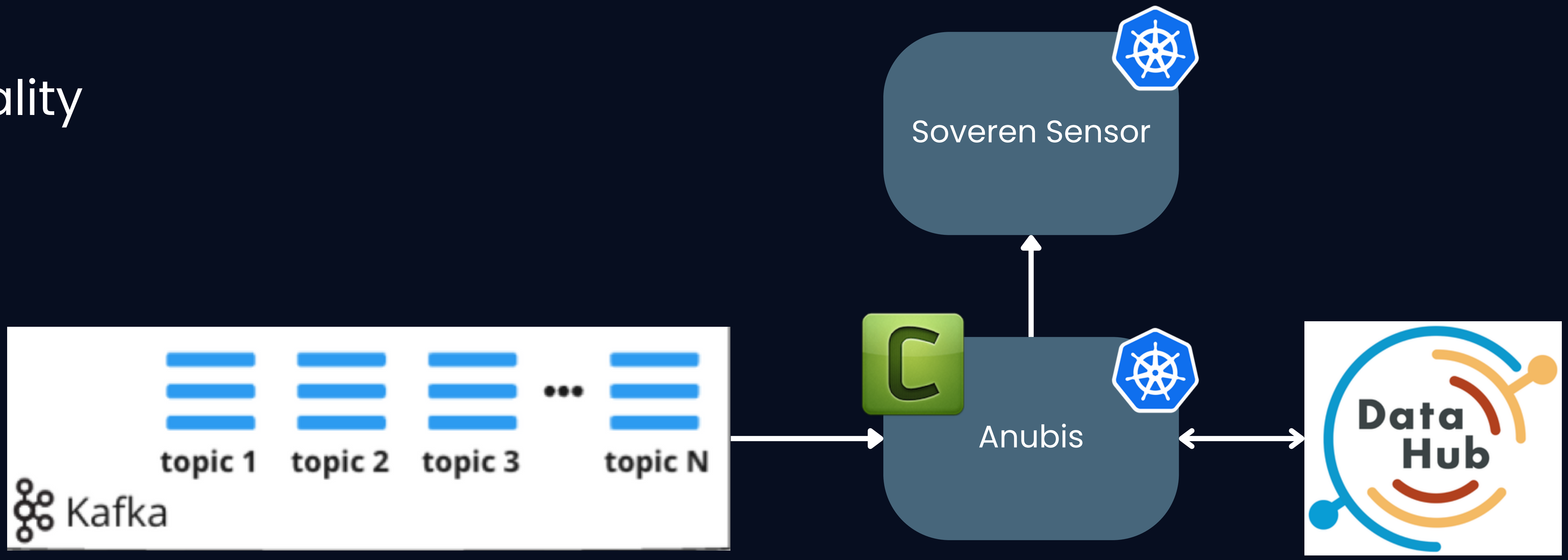
Anubis

If developing is not fun, then why do it?

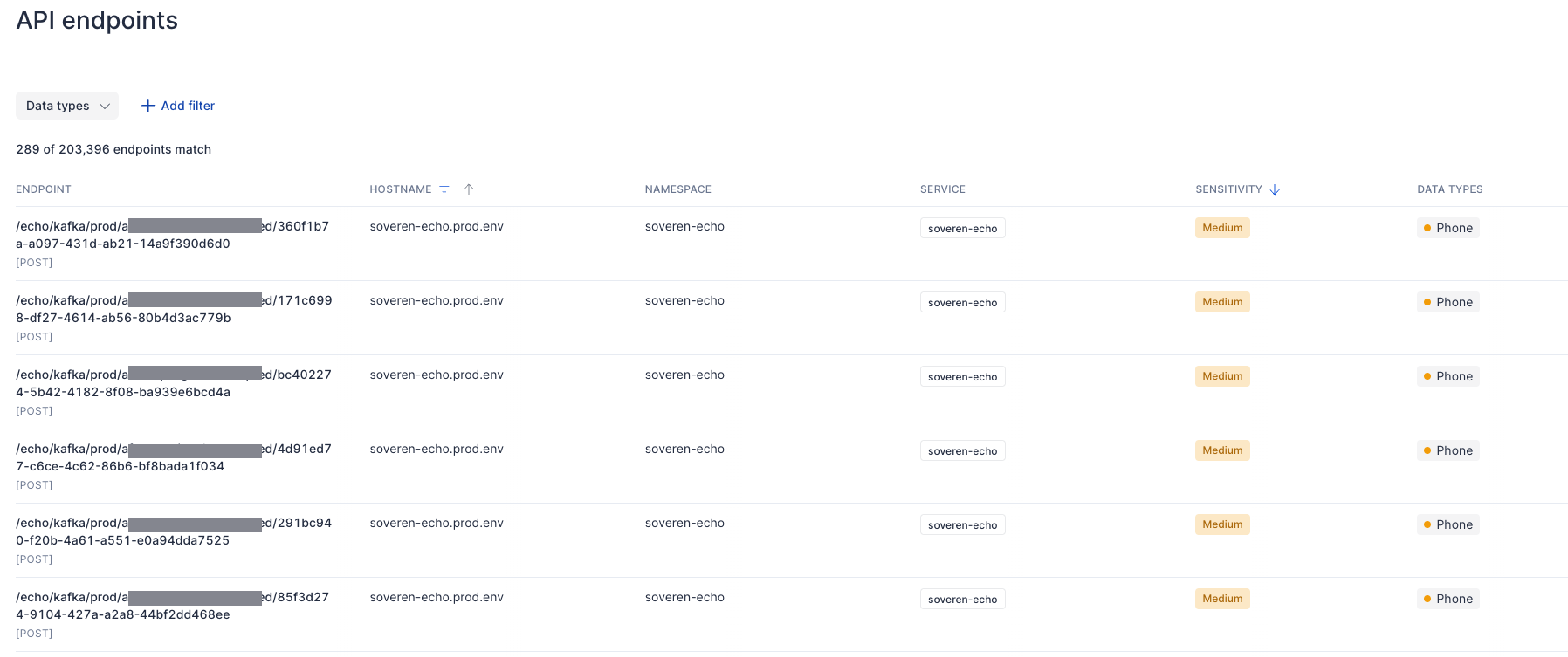
Solution Architecture now

Service for sampling messages from kafka – “Anubis”

- Search for critical data
- Assessment of data quality
- Enrich findings
- Report to DataHub
- Report to CMDB



Anubis at work

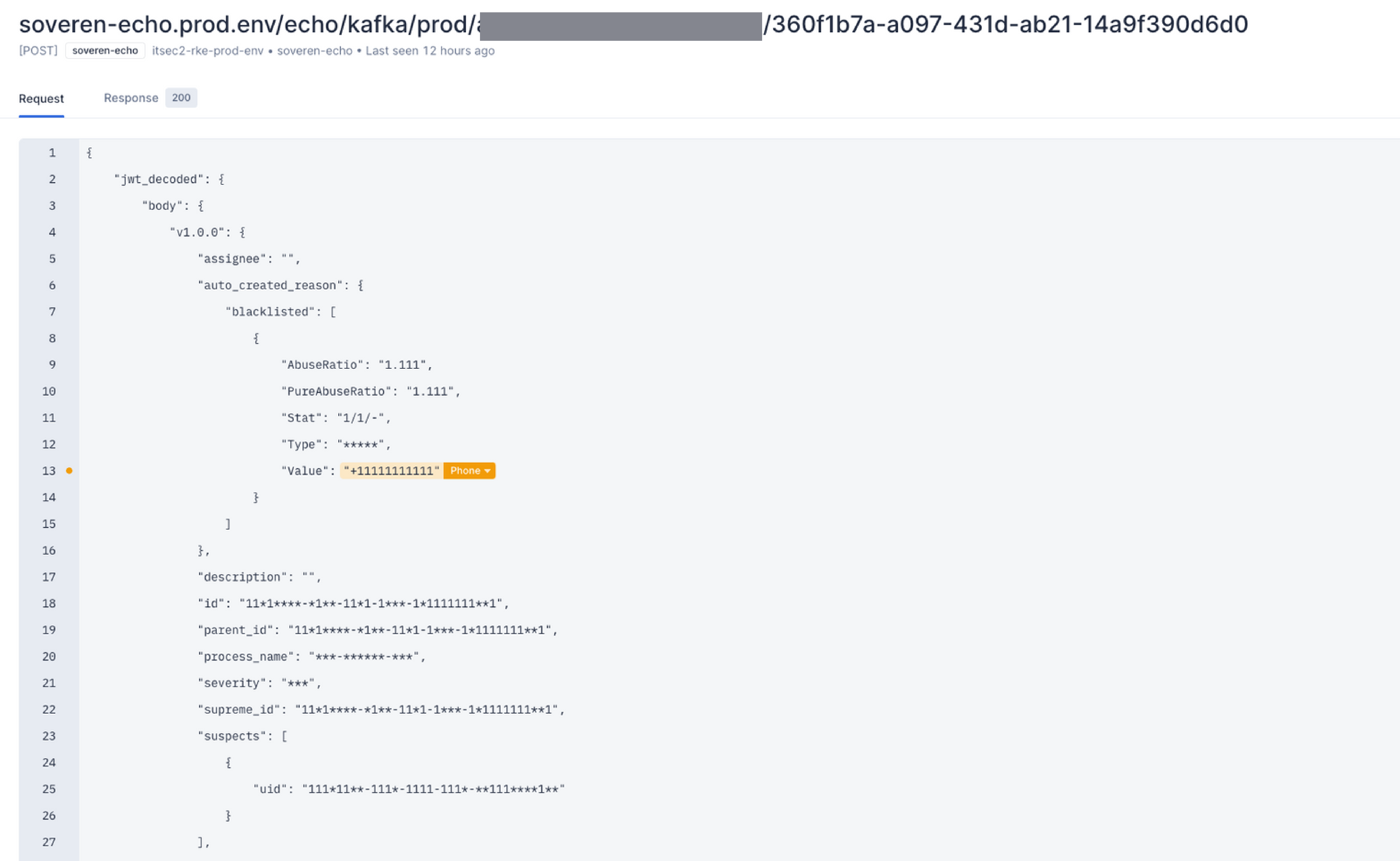


API endpoints

289 of 203,396 endpoints match

ENDPOINT	HOSTNAME	NAMESPACE	SERVICE	SENSITIVITY ↓	DATA TYPES
/echo/kafka/prod/a[redacted]d/360f1b7a-a097-431d-ab21-14a9f390d6d0 [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone
/echo/kafka/prod/a[redacted]d/171c6998-df27-4614-ab56-80b4d3ac779b [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone
/echo/kafka/prod/a[redacted]d/bc402274-5b42-4182-8f08-ba939e6bcd4a [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone
/echo/kafka/prod/a[redacted]d/4d91ed77-c6ce-4c62-86b6-bf8bada1f034 [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone
/echo/kafka/prod/a[redacted]d/291bc940-f20b-4a61-a551-e0a94dda7525 [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone
/echo/kafka/prod/a[redacted]d/85f3d274-9104-427a-a2a8-44bf2dd468ee [POST]	soveren-echo.prod.env	soveren-echo	soveren-echo	Medium	● Phone

Anubis at work

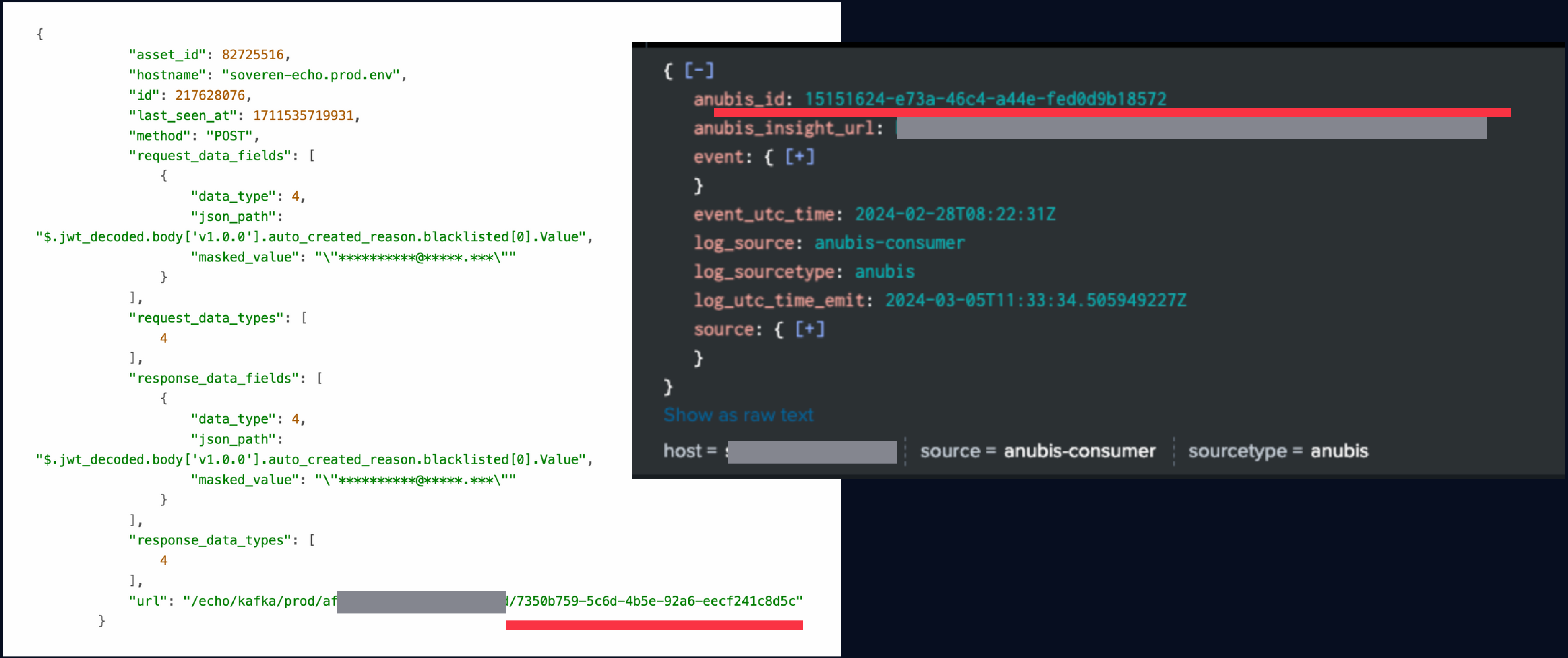


soveren-echo.prod.env/echo/kafka/prod/[redacted]/360f1b7a-a097-431d-ab21-14a9f390d6d0

[POST] soveren-echo itsec2-rke-prod-env • soveren-echo • Last seen 12 hours ago

```
{
    "jwt_decoded": {
        "body": {
            "v1.0.0": {
                "assignee": "",
                "auto_created_reason": {
                    "blacklisted": [
                        {
                            "AbuseRatio": "1.111",
                            "PureAbuseRatio": "1.111",
                            "Stat": "1/1/-",
                            "Type": "*****",
                            "Value": "+11111111111"
                        }
                    ]
                },
                "description": "",
                "id": "11*1****-*1**-11*1-1***-1*1111111**1",
                "parent_id": "11*1****-*1**-11*1-1***-1*1111111**1",
                "process_name": "***-******-***",
                "severity": "***",
                "supreme_id": "11*1****-*1**-11*1-1***-1*1111111**1",
                "suspects": [
                    {
                        "uid": "111*11**-111*-1111-111*-**111****1**"
                    }
                ],
```

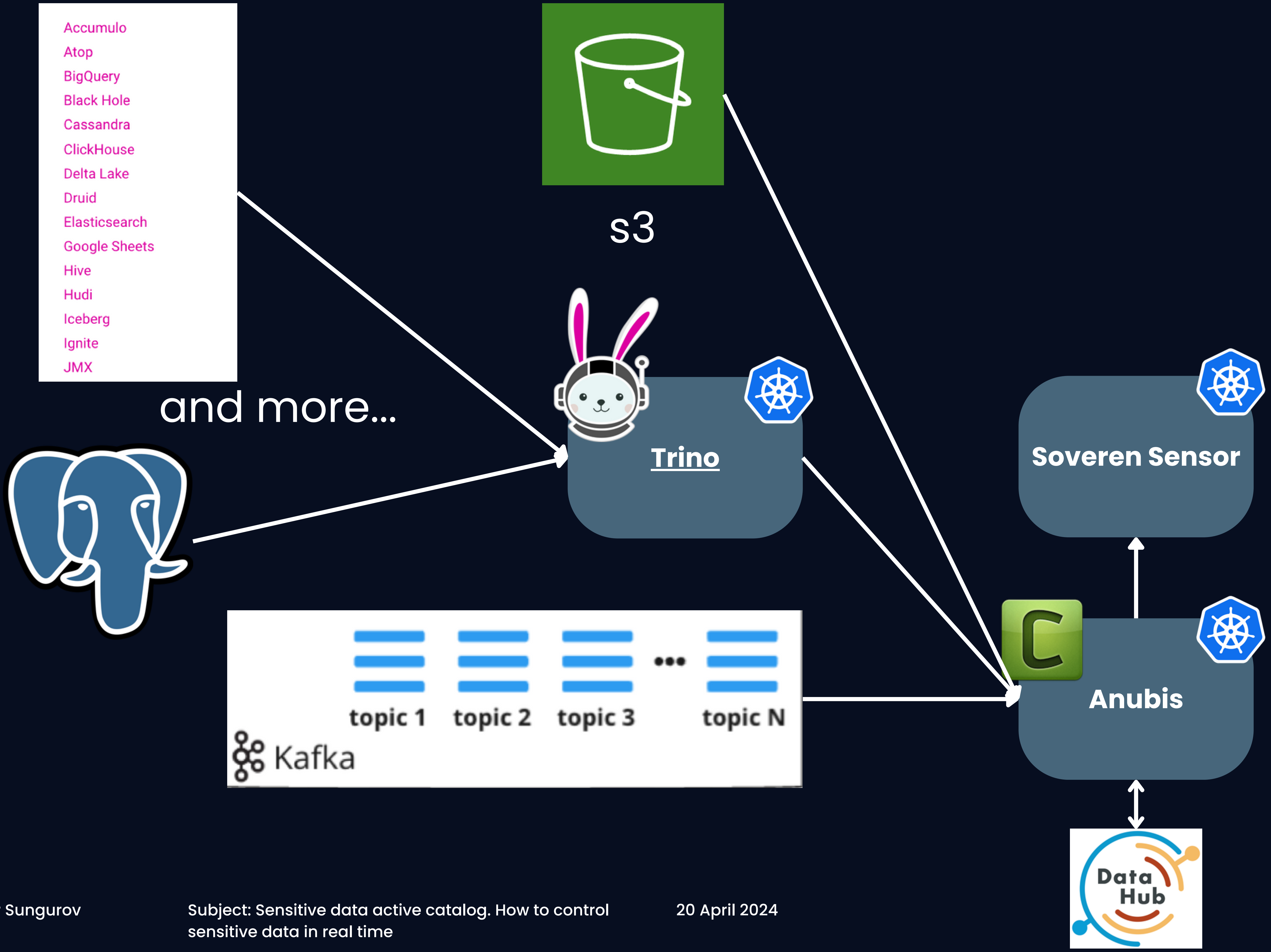
Anubis at work



```
{
            "asset_id": 82725516,
            "hostname": "soveren-echo.prod.env",
            "id": 217628076,
            "last_seen_at": 1711535719931,
            "method": "POST",
            "request_data_fields": [
                {
                    "data_type": 4,
                    "json_path":
"$.jwt_decoded.body['v1.0.0'].auto_created_reason.blacklisted[0].Value",
                    "masked_value": "\"**********@*****.***\""
                }
            ],
            "request_data_types": [
                4
            ],
            "response_data_fields": [
                {
                    "data_type": 4,
                    "json_path":
"$.jwt_decoded.body['v1.0.0'].auto_created_reason.blacklisted[0].Value",
                    "masked_value": "\"**********@*****.***\""
                }
            ],
            "response_data_types": [
                4
            ],
            "url": "/echo/kafka/prod/af[redacted]/7350b759-5c6d-4b5e-92a6-eecf241c8d5c"
        }
```

```
{ [-]
   anubis_id: 15151624-e73a-46c4-a44e-fed0d9b18572
   anubis_insight_url: [redacted]
   event: { [+]
   }
   event_utc_time: 2024-02-28T08:22:31Z
   log_source: anubis-consumer
   log_sourcetype: anubis
   log_utc_time_emit: 2024-03-05T11:33:34.505949227Z
   source: { [+]
   }
}
Show as raw text
host = [redacted] | source = anubis-consumer | sourcetype = anubis
```

Solution Architecture to be



- Accumulo
- Atop
- BigQuery
- Black Hole
- Cassandra
- ClickHouse
- Delta Lake
- Druid
- Elasticsearch
- Google Sheets
- Hive
- Hudi
- Iceberg
- Ignite
- JMX

and more...

s3

Trino

Soveren Sensor

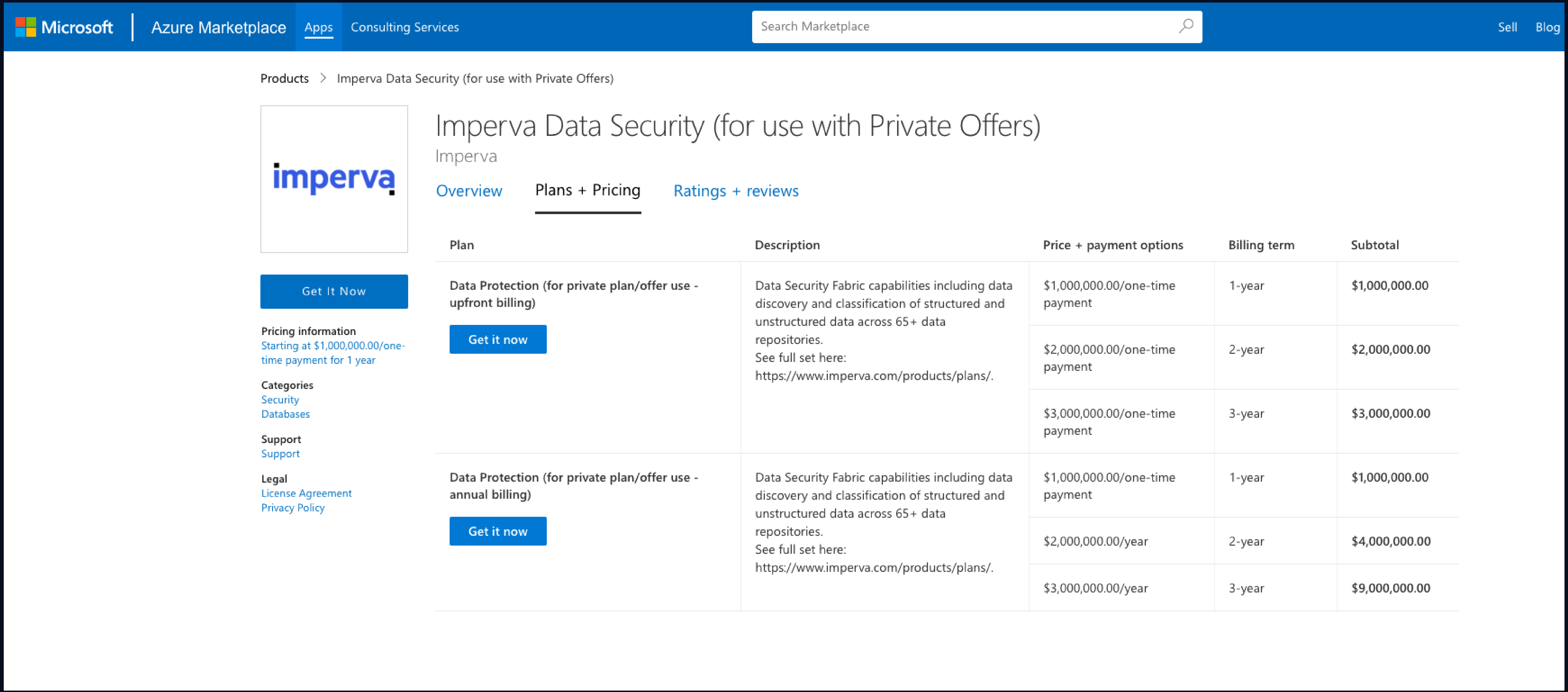
Anubis

topic 1 topic 2 topic 3 ••• topic N

Kafka

Data Hub

There's something strange...



Products > Imperva Data Security (for use with Private Offers)

Imperva Data Security (for use with Private Offers)

Imperva

Pricing information
Starting at $1,000,000.00/one-time payment for 1 year

Categories
Security
Databases

Support
Support

Legal
License Agreement
Privacy Policy

Plan	Description	Price + payment options	Billing term	Subtotal
Data Protection (for private plan/offer use - upfront billing) Get it now	Data Security Fabric capabilities including data discovery and classification of structured and unstructured data across 65+ data repositories. See full set here: https://www.imperva.com/products/plans/.	$1,000,000.00/one-time payment	1-year	$1,000,000.00
		$2,000,000.00/one-time payment	2-year	$2,000,000.00
		$3,000,000.00/one-time payment	3-year	$3,000,000.00
Data Protection (for private plan/offer use - annual billing) Get it now	Data Security Fabric capabilities including data discovery and classification of structured and unstructured data across 65+ data repositories. See full set here: https://www.imperva.com/products/plans/.	$1,000,000.00/one-time payment	1-year	$1,000,000.00
		$2,000,000.00/year	2-year	$4,000,000.00
		$3,000,000.00/year	3-year	$9,000,000.00

exness

Author: Aleksandr Sungurov

Subject: Sensitive data active catalog. How to control sensitive data in real time

20 April 2024

exness

Thank you

slides → bit.ly/PER2024

alexander.sungurov@exness.com
@Banzay021
www.linkedin.com/in/alexander-sungurov/

Limassol, Cyprus 2024
PERCONA UNIVERSITY