letgo

Data governance in streaming at scale

Sebastián Ortega @ letgo

About us

letgo

Carol H.
Hey! I love the bike and I'd like to buy it. How does $25 sound?

Me
Tell you what... Let's leave it at $30 and I'll throw in a bike lock

Founded in 2015

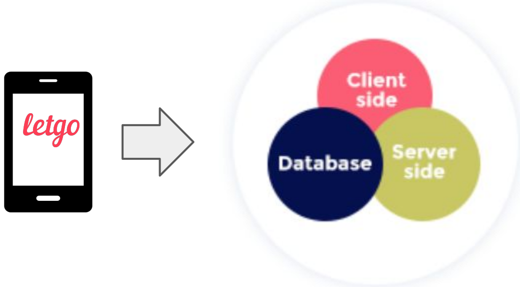
Phenomenal growth

Focused on Turkey

Offices in Barcelona and Istanbul

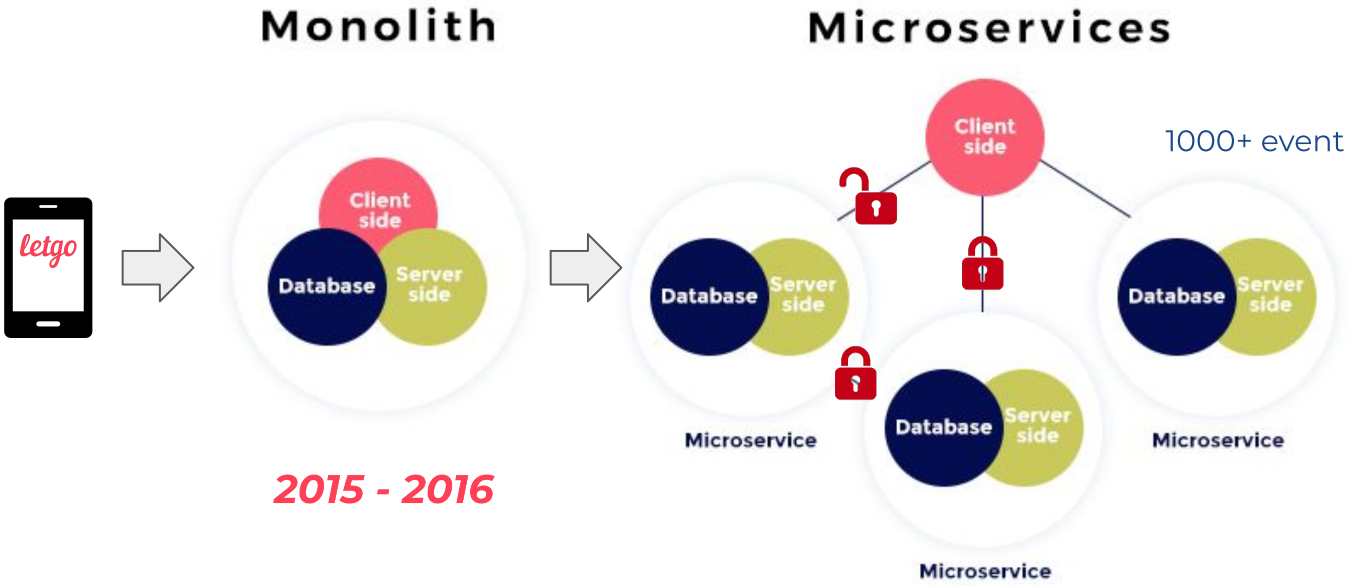
From Monolith to Microservices

Monolith



2015 - 2016

From Monolith to Microservices



1000+ event types!

2015 - 2016

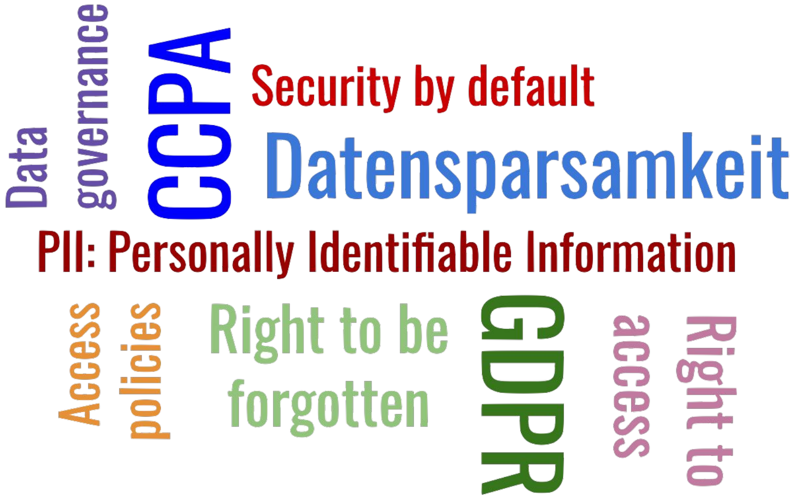
2017 - now

Data Governance in streaming at scale?

letgo

Data Governance in streaming at scale

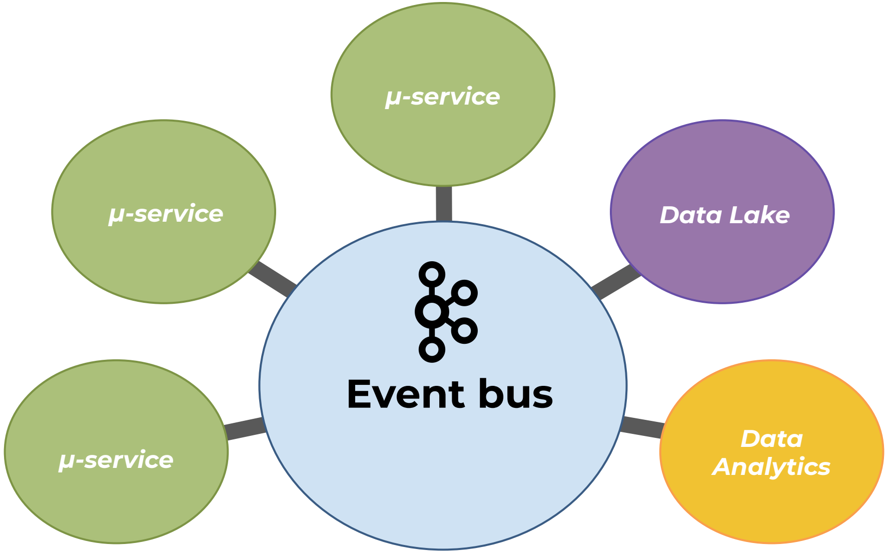
- Data governance means
  - a minimum of data quality and integrity
  - catalogue of data and which parts are sensitive
  - security controls
  - data minimization



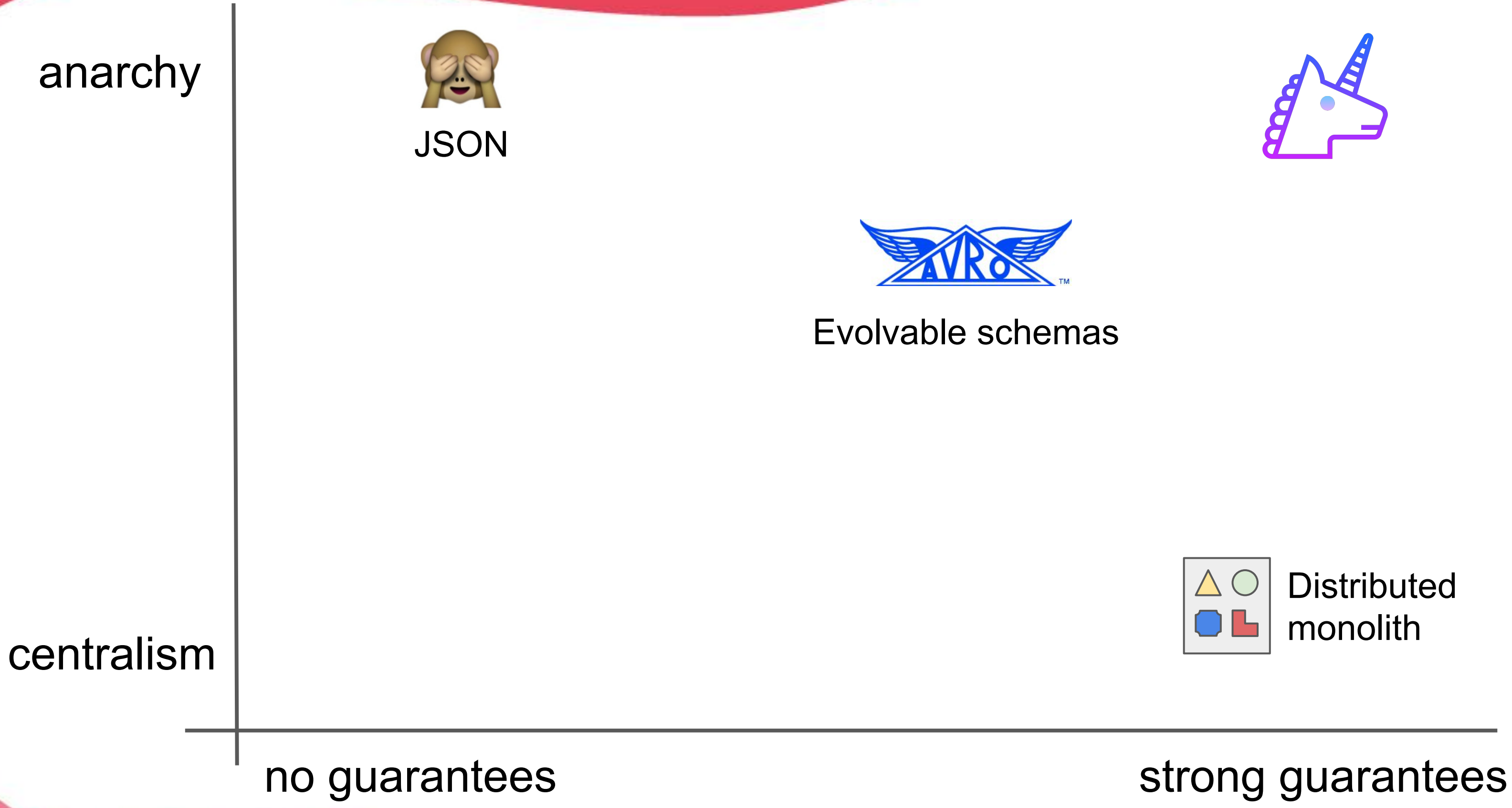
Data Governance in streaming at scale

Data Governance in streaming at scale

- Dimensions in which you need to scale
    - Sheer data: Kafka-powered event bus
    - Organizational: data and metadata-driven self-service architecture
    - Data team owns the platform, teams own their own data



Schemas to enable governance without killing freedom



anarchy

JSON

Evolvable schemas

Distributed monolith

centralism

no guarantees

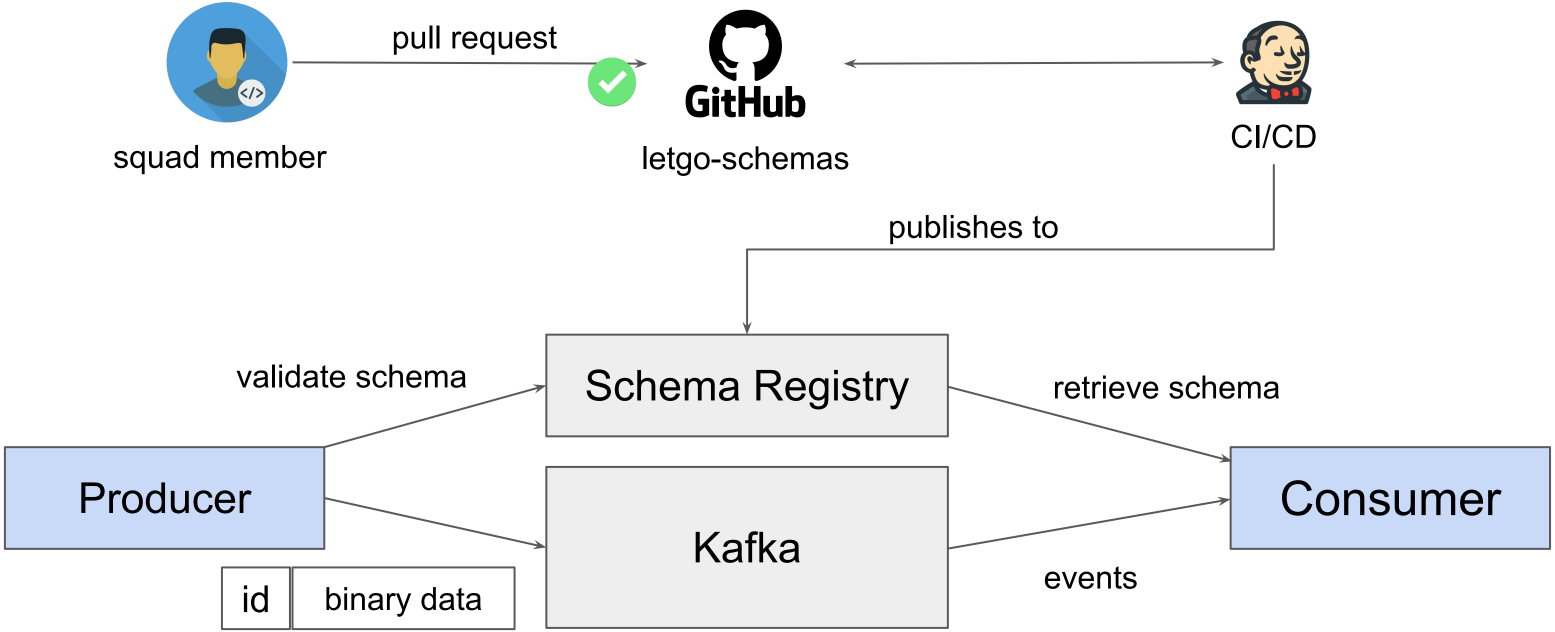
strong guarantees

Avro schemas

- Schema-driven serialization protocol
- Support for JSON and binary representations
  - Compact, fast to parse
  - Need tools to inspect
- Polyglot
- Evolvable schemas

```
{
 "type": "record",
 "namespace": "com.letgo.squad1",
 "name": "ProductVisit",
 "fields": [
   {
     "name": "product_id",
     "type": "string",
     "doc": "Identifier of the
visited product"
   },
   ...
```

Schema management



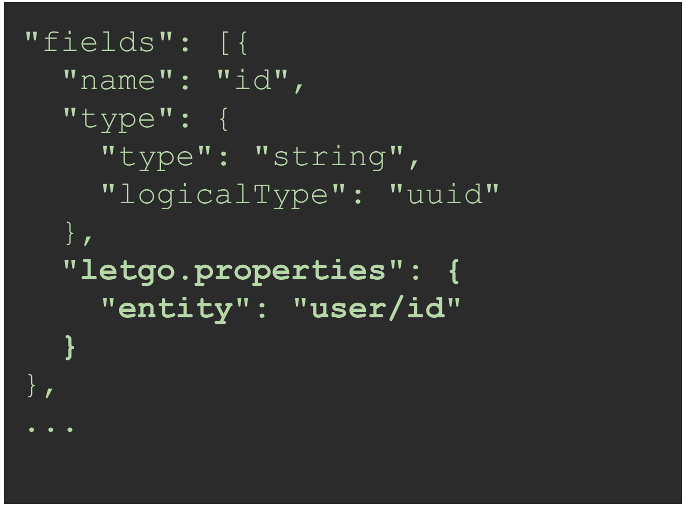
Schemas to manage PII

Where is ~~Waldo~~ what sensitive information?

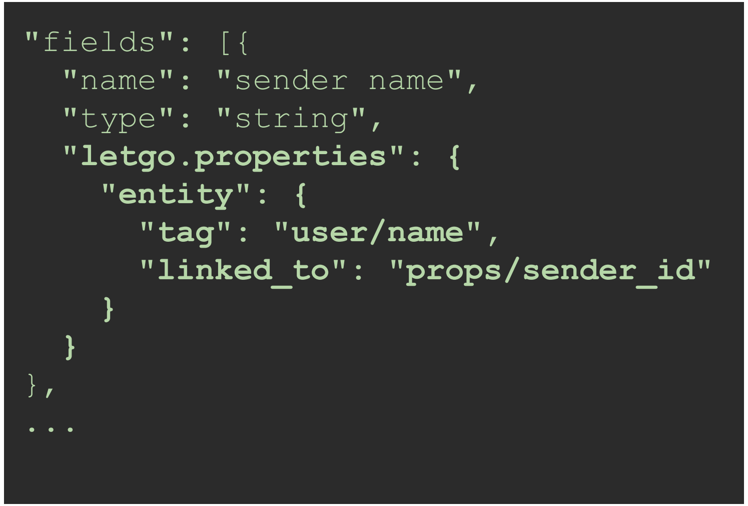
- We have many hundreds of domain events
- An email or user_id might appear under any field name (maybe different names in the same event)
- Avoid writing a pile of ifs
- Tagging
  - Source team has the knowledge
  - Schema is already being written by source team
  - Avro can be extended with more metadata

PII tagging

- Added `letgo.properties` to extend any Avro type definition
- More complex versions of the entity exists to avoid ambiguity



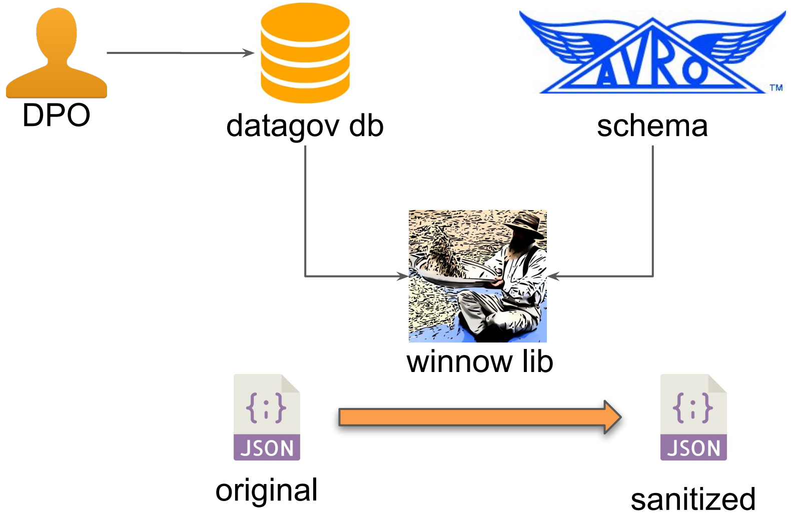
```
"fields": [{
  "name": "id",
  "type": {
    "type": "string",
    "logicalType": "uuid"
  },
  "letgo.properties": {
    "entity": "user/id"
  }
},
...
```



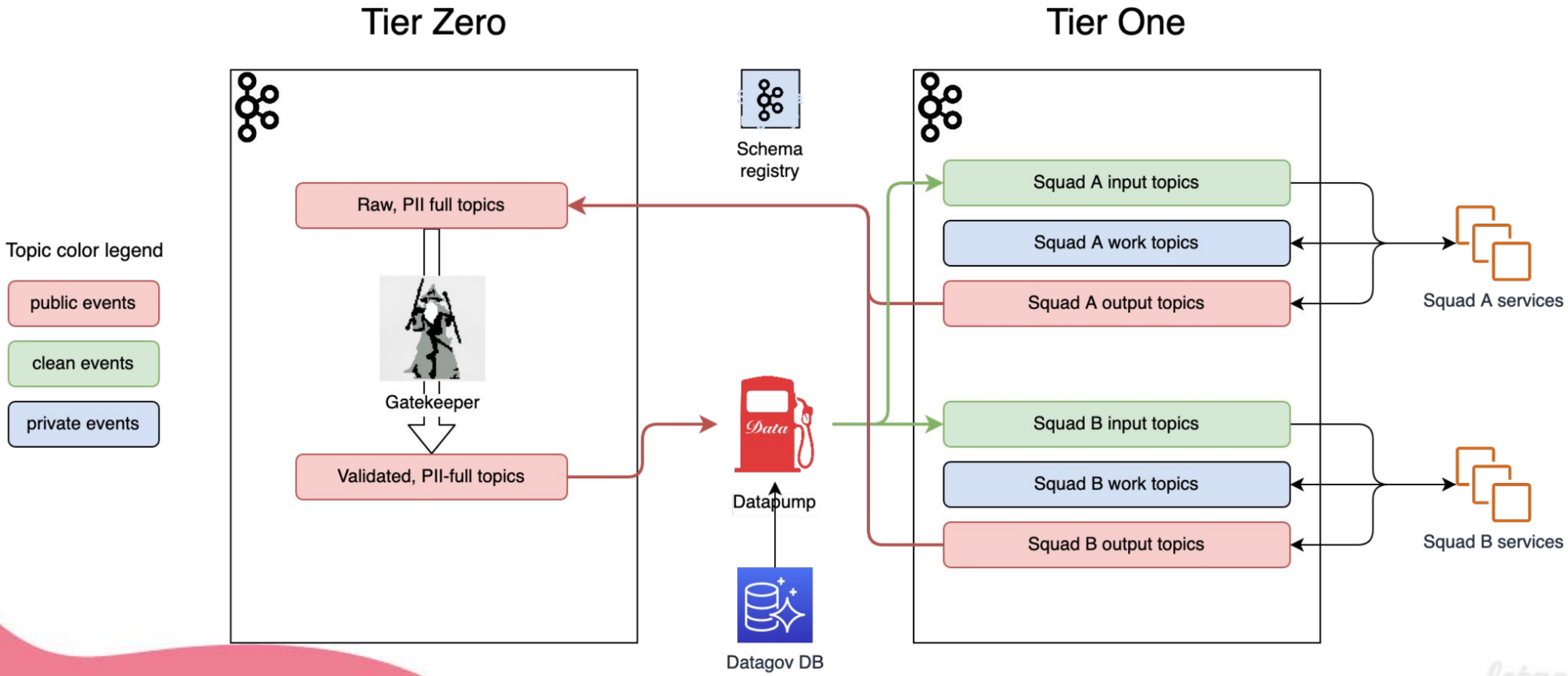
```
"fields": [{
  "name": "sender name",
  "type": "string",
  "letgo.properties": {
    "entity": {
      "tag": "user/name",
      "linked_to": "props/sender_id"
    }
  }
},
...
```

PII tagging

- Internal library `winnow`
- Parse letgo metadata
- Combine it with business unit PII ACLs
- Governance actions
    - Remove fields
    - Hash fields



Metadata & data life cycle



letgo

Data platform key pillars

- **Self-service** data platform, teams are the data owners
- **Metadata-driven**: schemas and ACLs drive the behavior of the platform
- **Evolvability**: schemas can evolve without stop-the-world coordination
- **Privacy-safe defaults**: no PII flows between business units by default

Q&A

letgo

We are hiring!

Data Engineering

Team:

3 x Data Engineers

1 x Site Reliability Engineer

Open Positions:

2 Data Engineer

Data Science

Team:

5 x Data Scientists

1 x ML Engineer

Business Intelligence

Team:

2 x TnS Analysts

5 x Data Analysts

Open Positions:

1x BI Analyst

https://www.linkedin.com/company/letgo/jobs/

https://careers.olxgroup.com

Thank you