

# Single view of ... *Everything*

**Streaming 4.4 billion events  
with MongoDB & Apache Kafka**



**Simon Aubury**

Principal Data  
Engineer Architect

# Where

Are these guys from?





Australia

New Zealand

Asia

# *Why*

Build a single customer view?

iag



# We've a lot of data

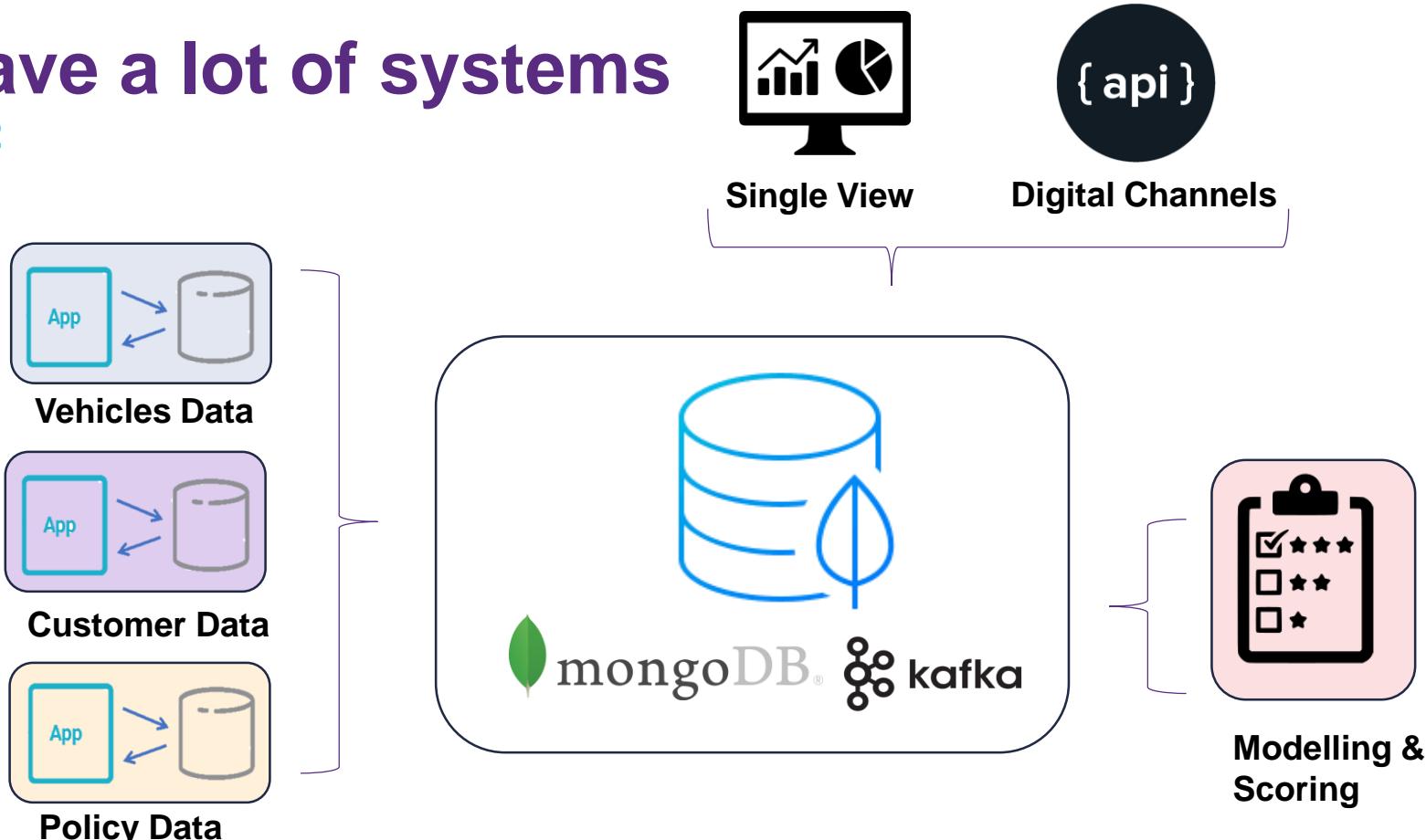
## Context #1

mvyear	mvmake	mvmodel	mvbody
1886	RUDGE	PENNY FARTHING	MBIKE
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1896	FORD	QUADRICYCLE	CONVT
1900	MERCEDES		CONVT



# We have a lot of systems

Context #2



# We want to tie it together

## Context #3

CGU

For Advisers About Us Claims

Car & Vehicle ▾ Home & Contents ▾ Travel Lifestyle ▾ Business

## BACKING AMBITIOUS AUSSIES FOR OVER 165 YEARS.

It's our ambition to help you live yours.

Learn more

Get a quote →

NRMA INSURANCE

Insurance ▾ Claims ▾ Renewals & payments ▾ Contact us

Have you been affected by the recent storm? During severe weather and natural disasters, our phones can be quite busy. If you have damage to your home or car, visit our emergency page to claim online.

Get a quote Retrieve a quote

Loyalty Discount We reward you for your entire relationship with us. Calculate your discount to see how much you can save.

Making a claim is easy Find out what you'll need and how to get your claim started.

Existing customer Learn more about how to access and manage your policy online.

Help is not something we do. Help is who we are. Watch the film

2014 TOYOTA AURION MC146N  
Comprehensive Car Insurance  
Policy expires 13 Mar 2018  
This year's premium \$751.09

368 GEORGE STREET, SYDNEY NSW 2000  
Home Buildings Policy automatically renewes in 3 days  
Policy expires 13 Jun 2017  
Next year's premium \$751.98

# Focused Insights

Profile

Basic Profile  
Title: MR SIMON  
First Name: B  
Middle Names: AUBURY  
Last Name: not collected  
Date of Birth: [REDACTED]

Address: WEST RYDE NSW 2114

Map: Shows locations in Sydney, including West Ryde, Chatswood, and Balmain.

Profile

Best Contact  
Home Phone: not collected  
Work Phone: not collected  
Mobile Phone: [REDACTED]  
Email: [REDACTED]

Products

Vehicles

Customer Health

Marketing

Communication History

Relationships  
Same Address: [REDACTED] IN  
Same Address: [REDACTED] BURY

Activity timeline

- 2017 Nov Took out CTP insurance on a 2009 HOLDEN BARINA
- 2016 Nov Took out comprehensive insurance on a 2006 HONDA JAZZ
- Invalid date Invalid date Lodged a Collision claim on a 2003 FORD FOCUS
- 2019 Jan Took out CTP insurance on a 2009 HOLDEN BARINA

Property Factors

Property Attributes

Land size: 590m<sup>2</sup>

Slope: 1.80

Elevation: 36m AHD

Aspect: 247° W/S

Nearest pub/bar/club: 1.50 mins

Environmental Factors

Drive time to points of interest:  
Nearest education: 1.50 mins  
Nearest emergency: 1.50 mins  
Nearest shop: 3.30 mins  
Train station: 1.50 mins  
Nearest pub/bar/club: 1.50 mins

Profile  
Address: LOT 308 60 D AVENUE, WEST RYDE NSW 2114

Property Factors  
Flood Risk: N

Activity timeline

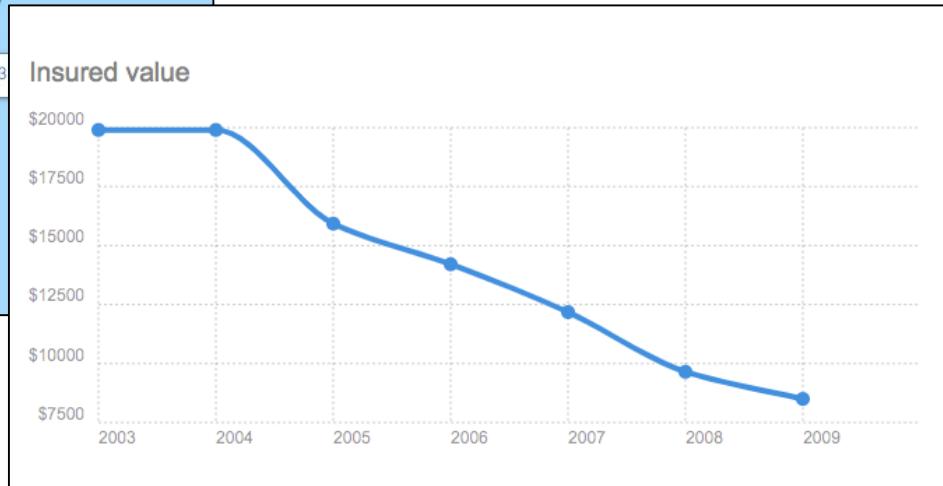
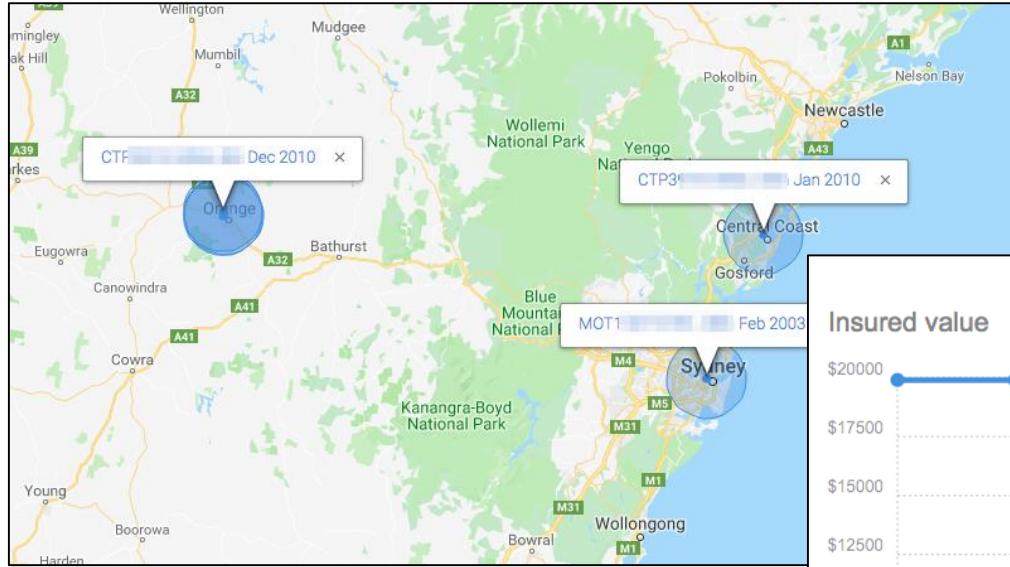
- Apr Took out a Collision claim on a 2016 HONDA CR-V
- 2006 Nov Took out CTP insurance on a 2006 HONDA JAZZ
- 2003 Feb Took out comprehensive insurance on a 2003 FORD FOCUS

2016 HONDA CR-V

NSW  
Nov 2017

18-11-2016: [REDACTED] 3525  
BILLING CLIENT, DRIVER, INSURED, MAIL ADDRESS

# Refocused Data



# How

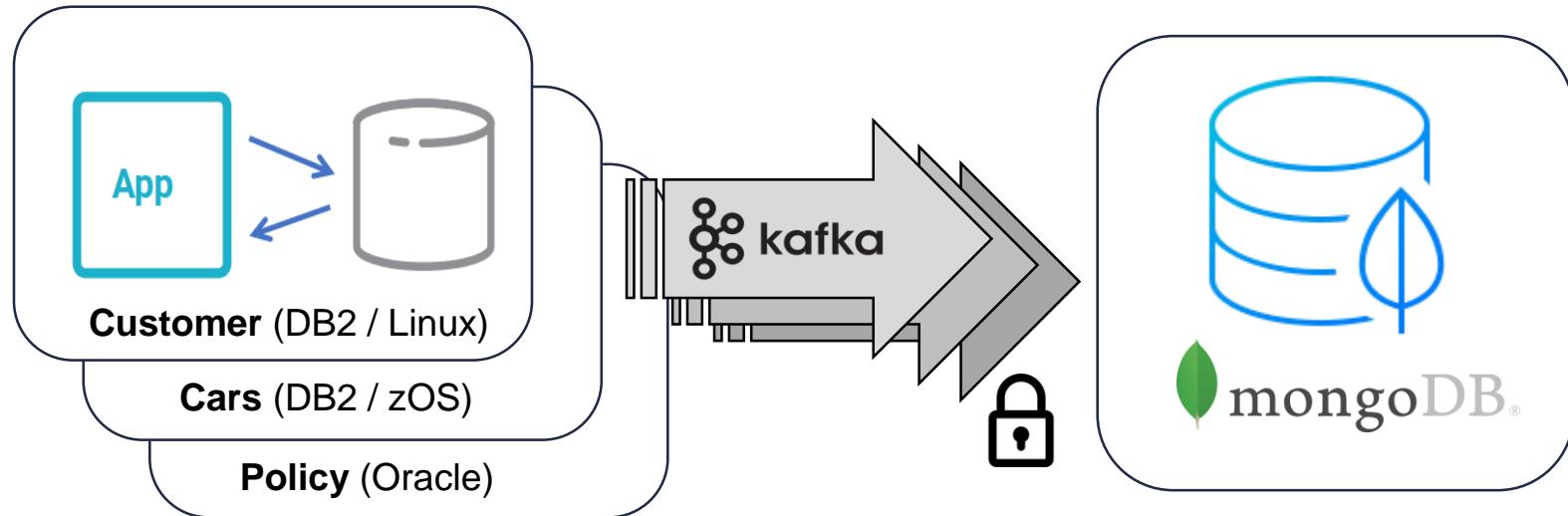
Did we build this?

iag



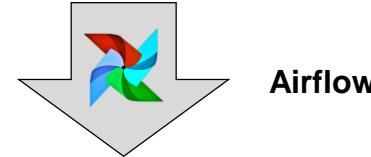
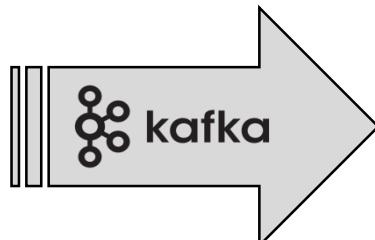
# System of record acquisition

Architecture of capture



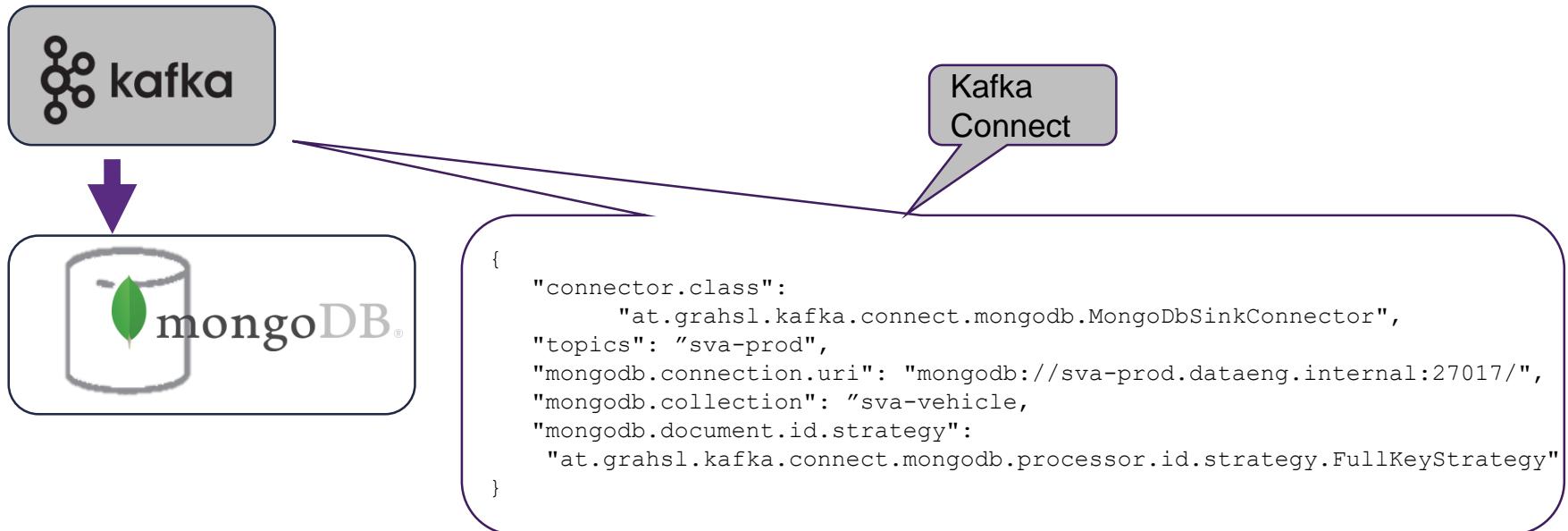
# Insight Generation

## Architecture of insights



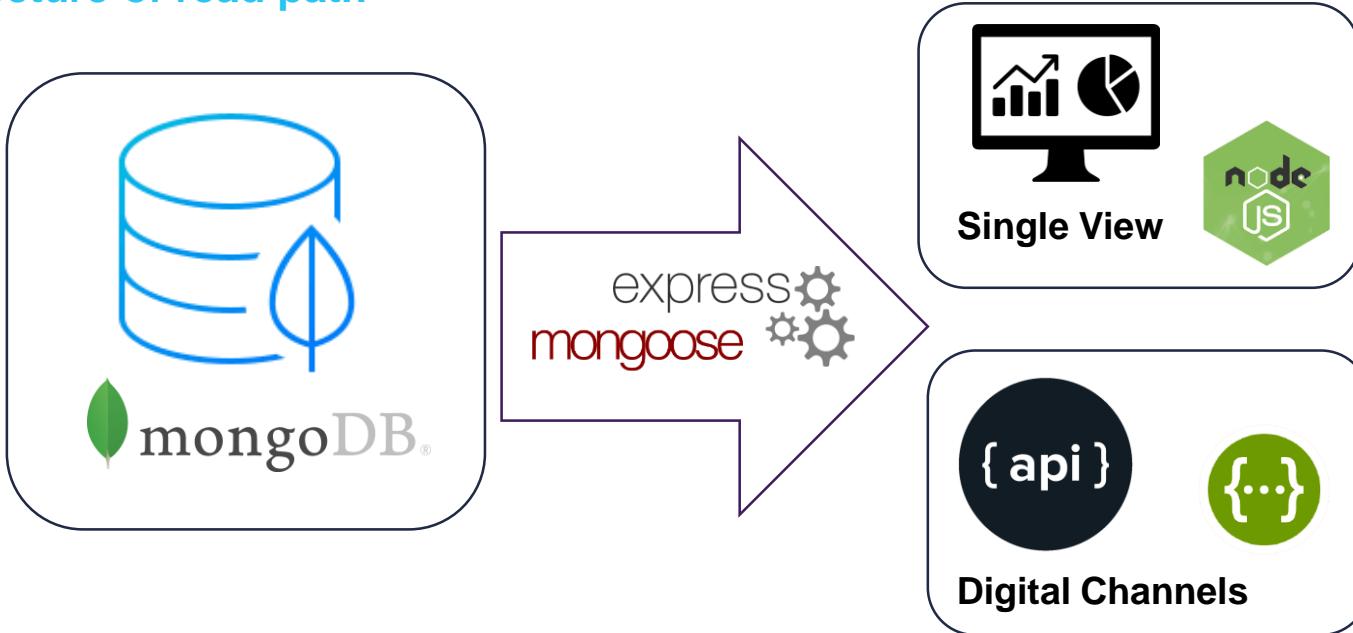
# Kafka Connect Sink

## Architecture of write path



# Serving Tier

Architecture of read path



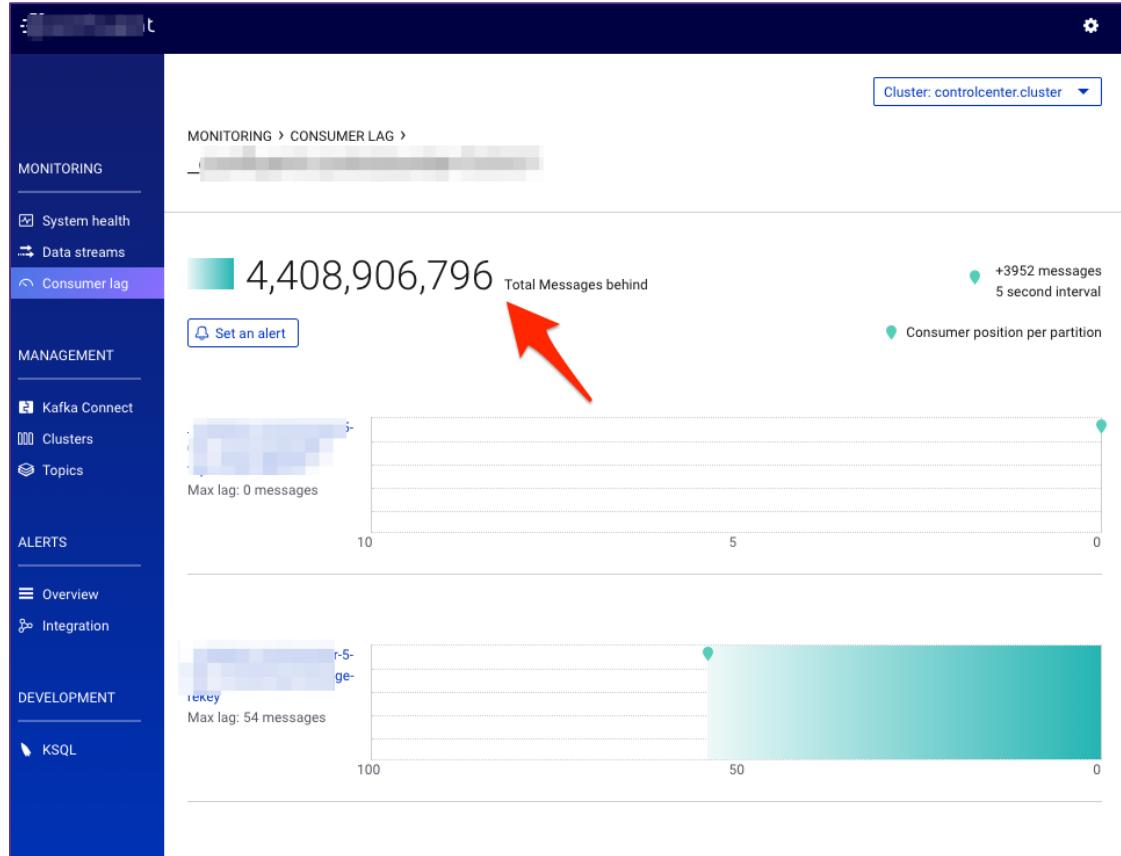
# What did we discover?

Slow to fast ... to *really* fast!



# Challenge

Lots of data



# Horizontal scaling?

## Theory

The diagram illustrates horizontal scaling through three cards:

- MongoDB Card:** Describes horizontal scaling as dividing the system dataset and load over multiple servers. It notes that while the overall speed or capacity of a single machine may not be high, each machine handles a subset of the overall workload, potentially providing better efficiency than a single high-speed high-capacity server. It also mentions the trade-off of increased complexity in infrastructure and maintenance.
- Kubernetes Card:** Explains that scaling out a Deployment ensures new Pods are created and scheduled to Nodes with available resources. It states that scaling will increase the number of Pods to the new desired state. Kubernetes supports [autoscaling](#) of Pods, but it is outside the scope of this tutorial. Scaling to zero is also possible, and it will delete all Pods of the specified Deployment.
- Kafka Card:** States that to scale out, you simply start another instance of your stream processing application, e.g. on another machine. The instances of your application will become aware of each other and automatically begin to share the processing work.

1. <https://docs.mongodb.com/manual/sharding/>
2. <https://kubernetes.io/docs/tutorials/kubernetes-basics/scale/scale-intro/>
3. <https://www.confluent.io/blog/elastic-scaling-in-kafka-streams/>

# Reality: Healthy Tech Competition



9:00 AM

Underwhelming ... it's time to use more docker apps.  
600/sec with 35% cpu on stream server



9:17 AM  
Scaled



9:24 AM

Ingestion stats on mongo: 2M records. Pretty impressive  
considering the number of queries and index hits ...



10:01 AM  
I can bump to 32



9:47 AM  
Looking good on the 16 CPUs - Mongo is keeping  
up with Kafka on the ingest now



10:03 AM  
Nice - starting to hammer. 40% across all CPU's



10:30 AM  
Hah - mongo's winning ! Topic drained



10:46 AM  
New record - we're at 5,070 records per second  
18 million / hour

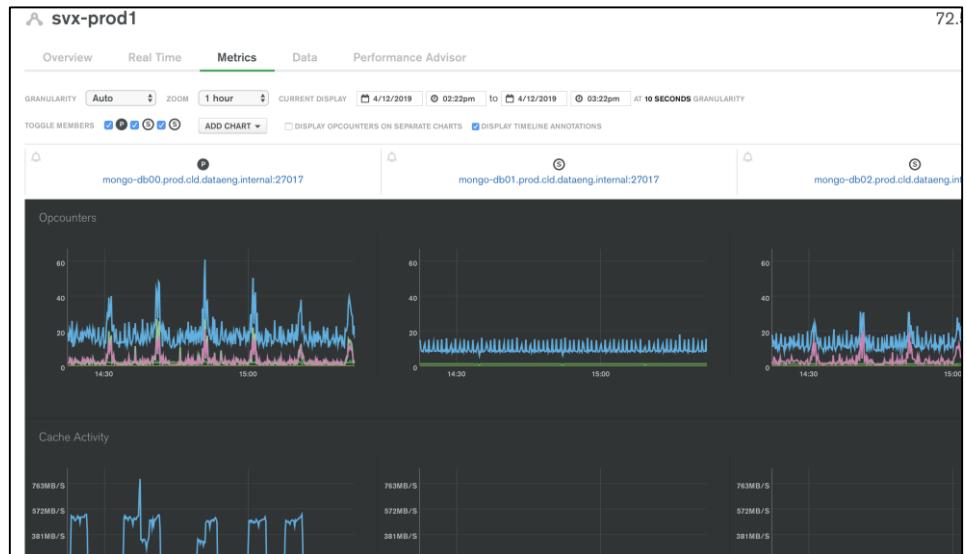
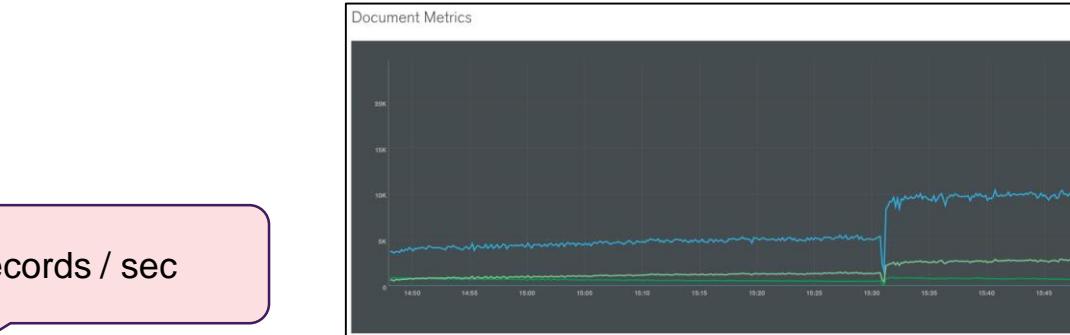
# Horizontal scaling ... meet efficient code

In numbers

30 records / sec



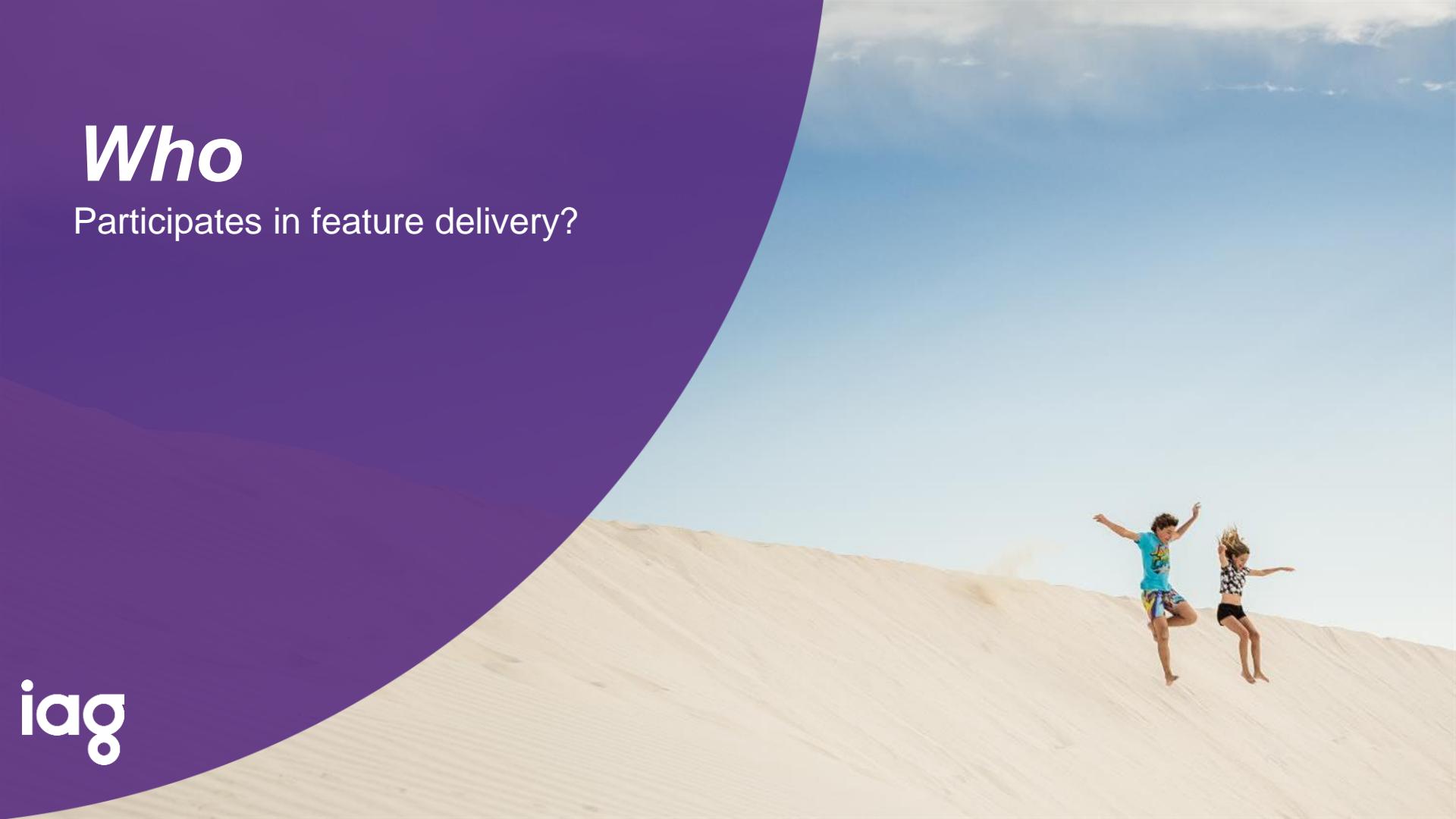
33,500 records / sec



# Who

Participates in feature delivery?

iag



# Feature Sprint

JIRA interface showing the 'All sprints' board. The board has sections for '12 To Do' and '4 In Progress'. Issues include TIS-37, TIS-10, TIS-68, TIS-17, TIS-20, and TIS-26. Filters at the top include 'Product', 'UI', 'Server', 'Only My Issues', and 'Recently Updated'.

svx.policy\_insights document view. The document details a policy entry with a red box highlighting specific fields:

```
_id: Object
  svx_policy_key: "HUON-604"
  cap_product: "POLICY AND PRODUCT"
  cap_model: "RETENTION OFFER"
  cap_feature: "PERCENTILE"
  kafka_id: 1651
  svx_policy_key: "HUON-CTP-04"
  cap_product: "POLICY AND PRODUCT"
  cap_model: "RETENTION OFFER"
  cap_feature: "PERCENTILE"
  feature_value_numeric: 30
  feature_value_character: null
```

A UI interface showing search results for 'addresses' and 'customers'. The 'addresses' section includes a POST button for '/addresses/{rapid\_address\_end\_point}'. The 'customers' section includes several GET methods for customer-related endpoints like '/customers/{\_id}' and '/customers/{\_id}/activities'.

A dashboard titled 'Products' showing CAP Features. It includes a table comparing current products across different categories (e.g., NOVA, MOTOR) based on CTP values. It also displays 'LOW Expected profit' and 'LOW Chance of renewal' metrics.

Category	Product	Current	All	CTP
NOVA	NOVA	60	65	45
	NOVA	60	65	56
MOTOR	MOTOR	60	65	25
	MOTOR	60	65	25

# *What*

Does this mean to our customers  
& stakeholders?

iag



# Unlocking Geospatial Analytics

## What we couldn't do with our RDBMS

Analytics team

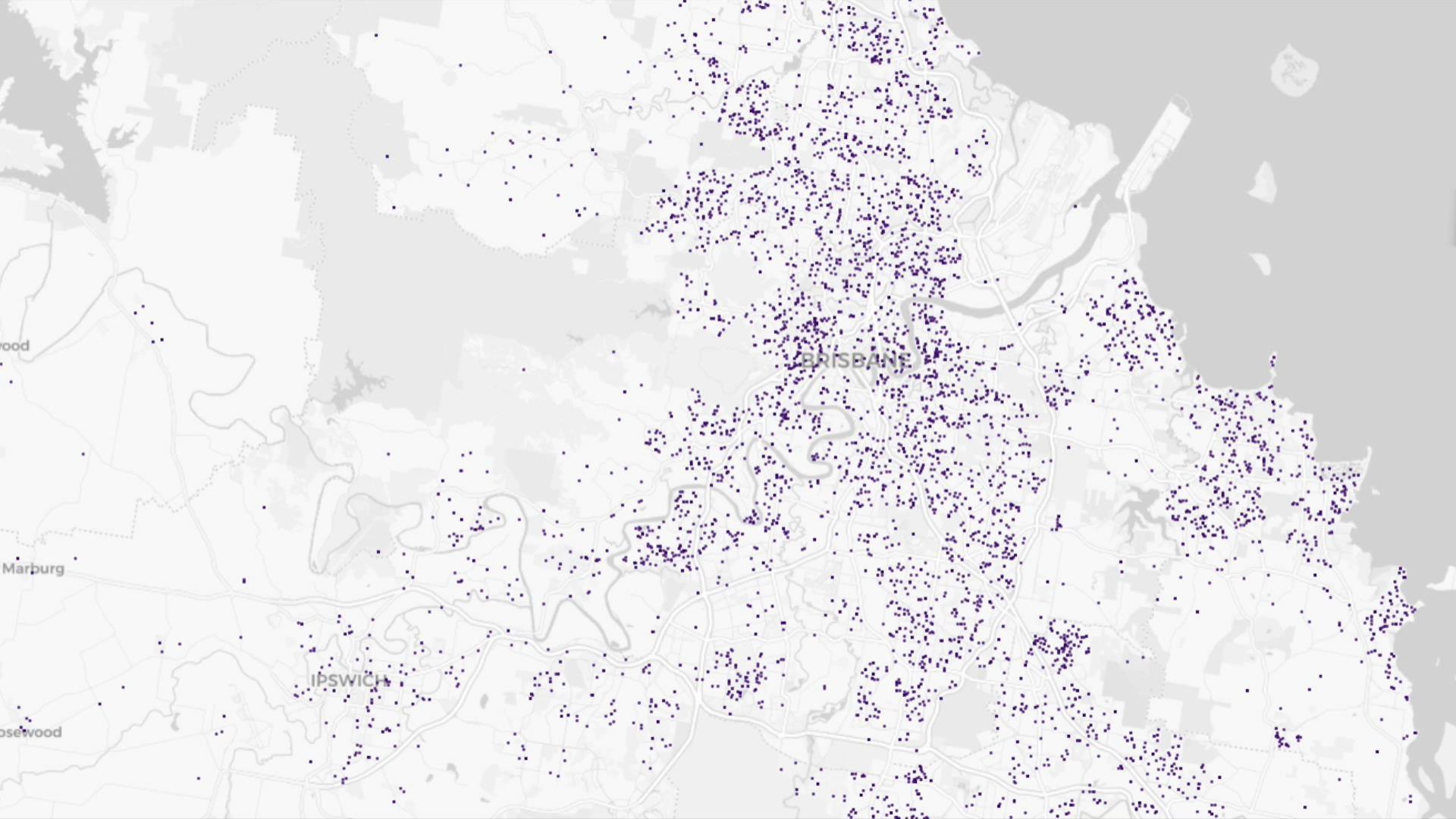
- 15 years of geospatial policy history
- Goal: understand correlations between population growth and policy purchases
- This business problem had been “on the table” and unsolved for over 3 years
- The following analysis generated from Mongo was done fast ... very, very fast



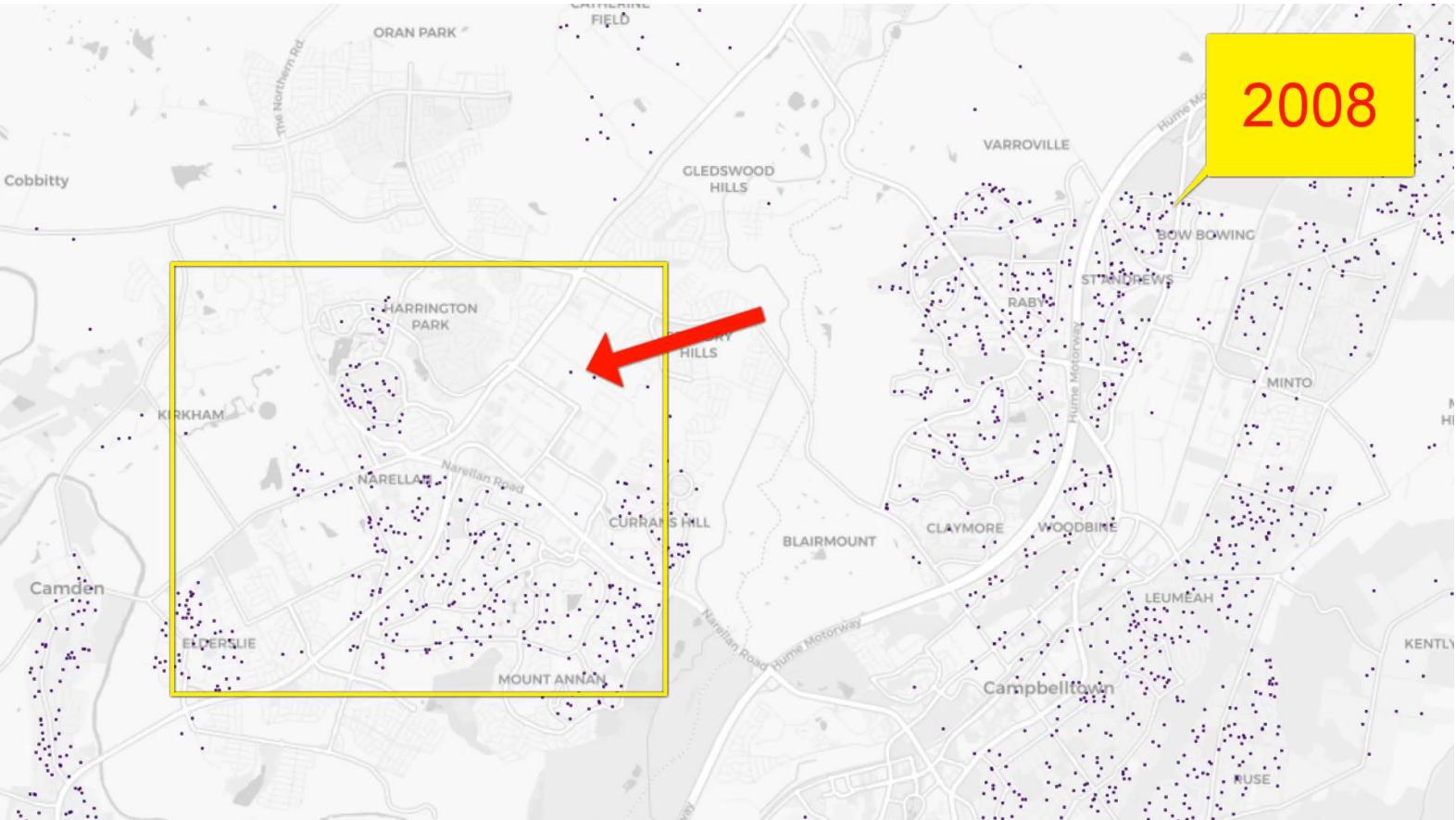
# Geospatial Over Time

“Creating a duplicated policy map from traditional IAG data systems not only involved complex queries but was error prone due to data quality issues.

SVx and the mongo platform brings the data to a single place and allows easy extraction for multiple different use cases”



# Geospatial Over Time



# Conclusions

## In summary

- Why build a single customer view?
- How did we build this?
- Who participates in feature delivery?
- What does this mean to our customers & stakeholders?



# *Thank-you!*

Any questions???

iag

