Google Cloud

Next '22

Discussion: How to optimize your data ingestion

by migrating from Pulsar to Pub/Sub

Oct 11–13

Andres Margalef

Software Expert
Mercado Libre

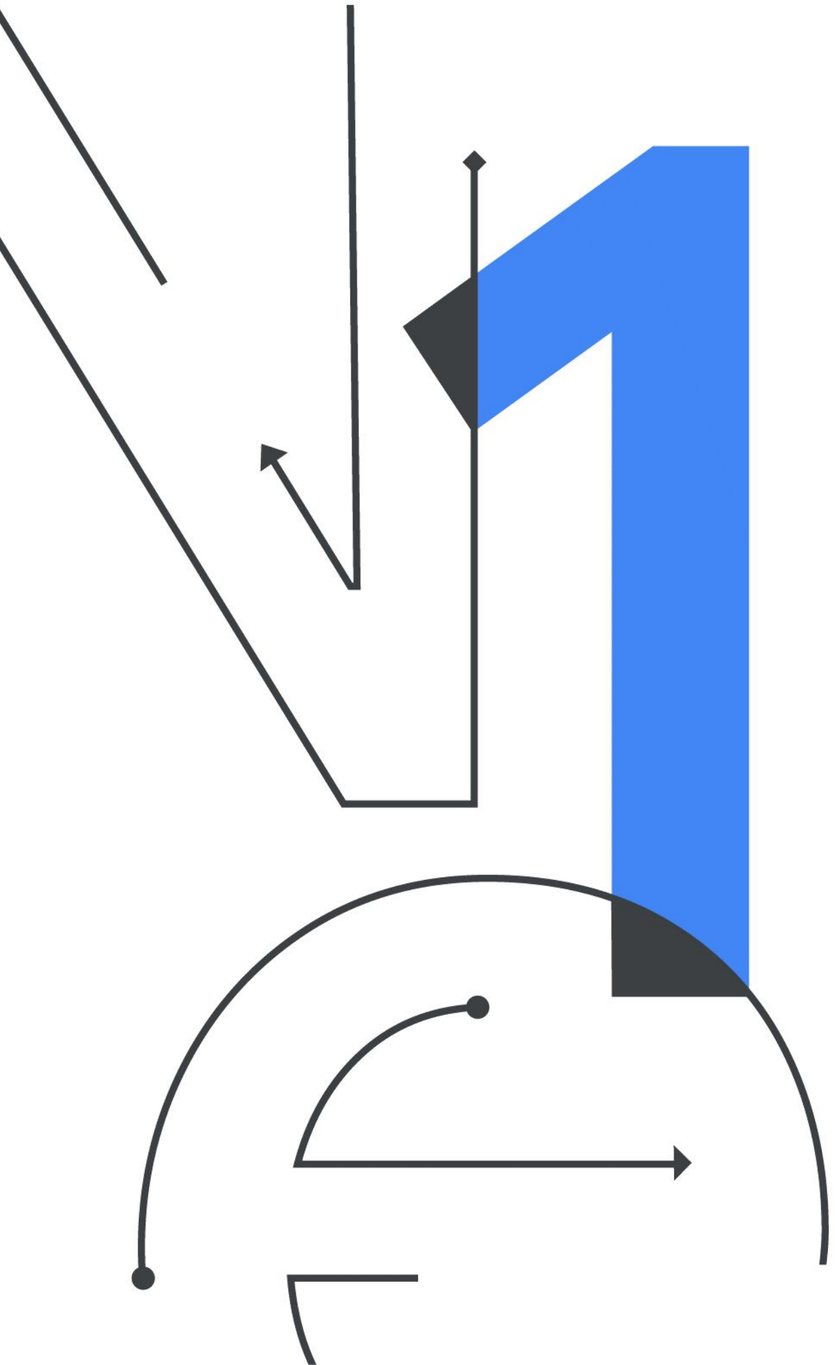
Shane Glass

Developer Advocate,
Data Integration
Google Cloud

Wei Hsia

Developer Advocate
Google Cloud

Contents



Data Ingestion

The evolution of a “high value” question

What has happened?

Batch Analytics

What is happening?

Real-time Analytics

What should happen?

Continuous Intelligence

Mercado Libre

Largest e-commerce, fintech & logistics company in LATAM

2x Grow YoY

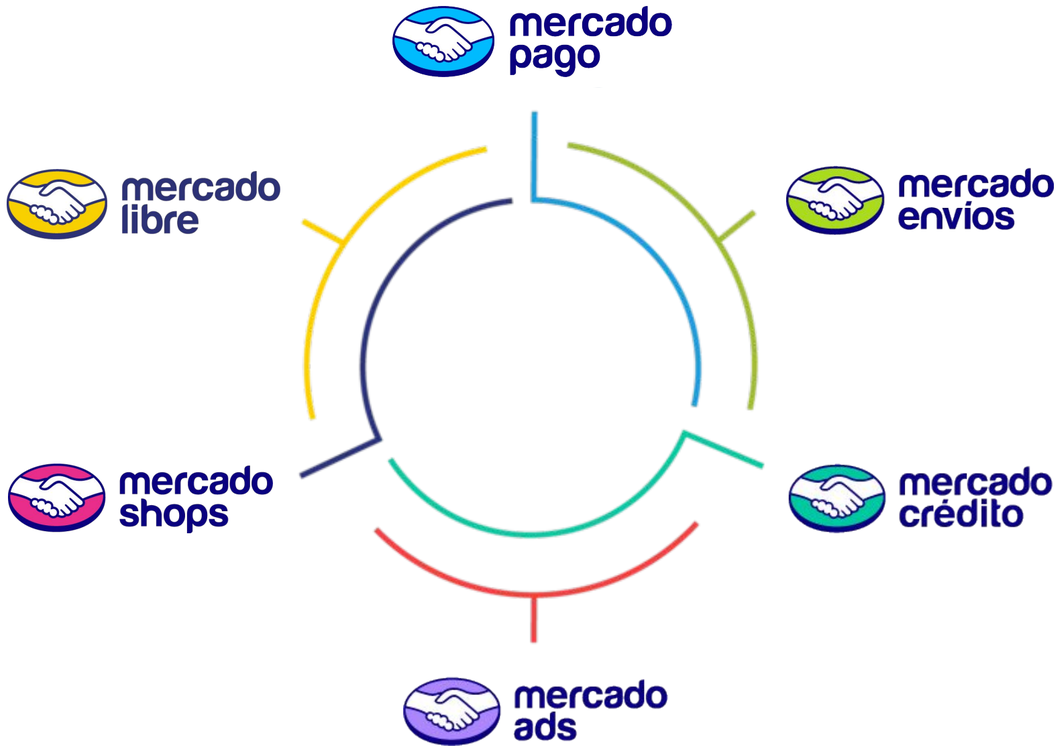
75M+ Active Users

18 Countries

250M+ items sold per quarter

34k+ Employees

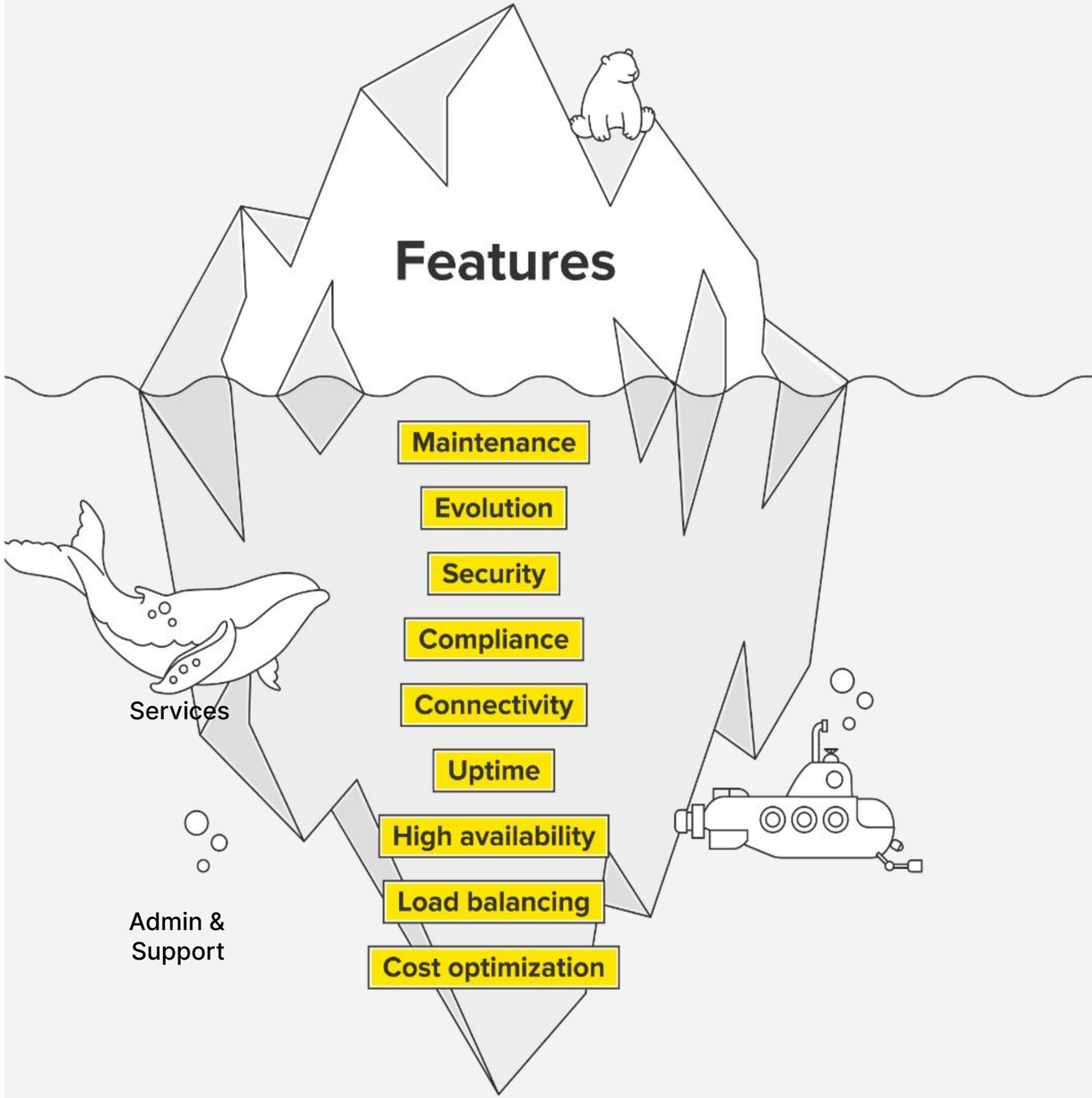
12k Engineers



Mercado Libre’s Fury

Fury

- PAAS solution on top of the cloud providers
- SAAS as abstraction layers on top of backend services
- Technology migration without impact
- Fine-grained governance by each SAAS mantainers
- Multi-* (multi zone / multi region / multi cloud)



Mercado Libre's Fury

Hot links

Related tools

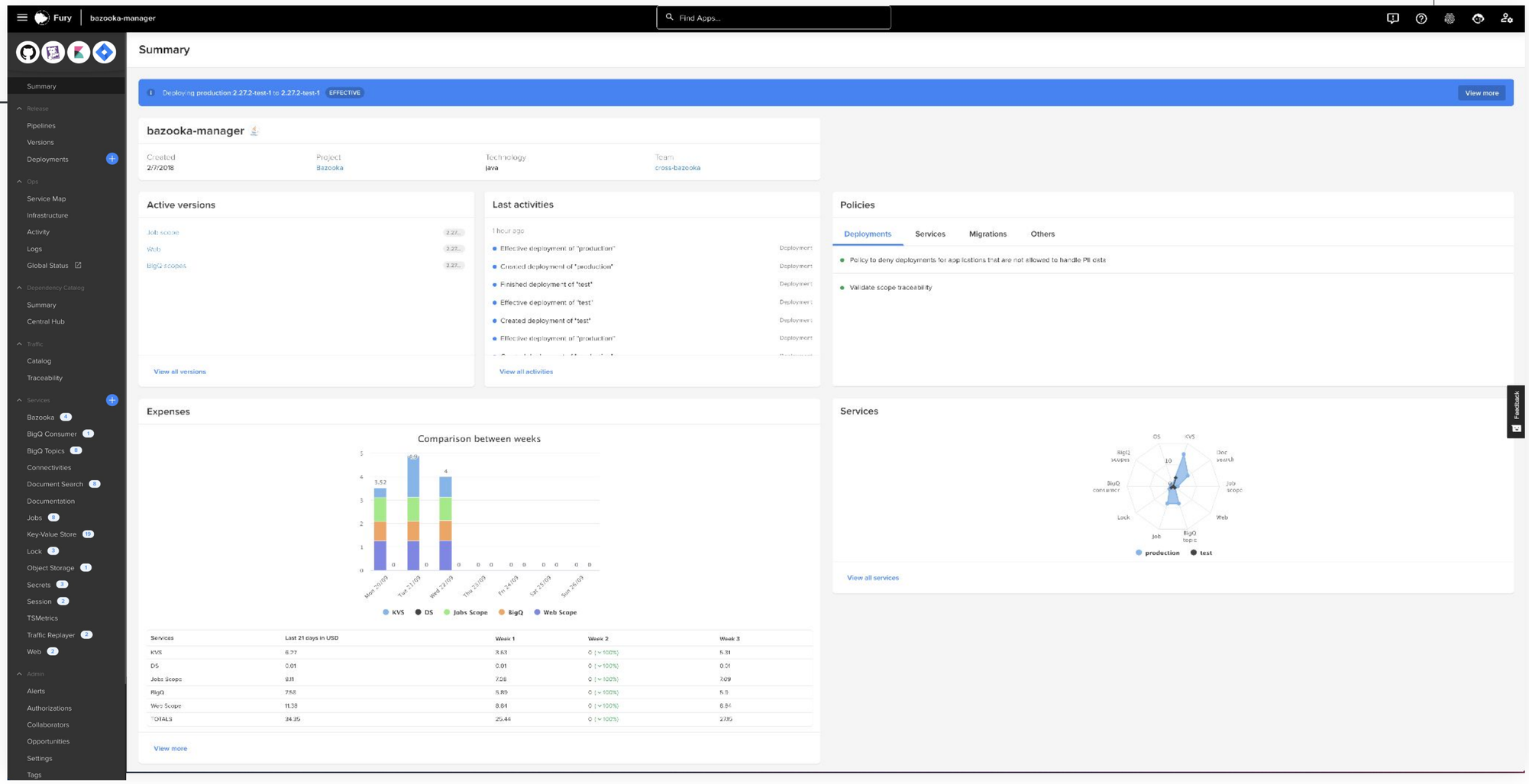
Ops

Dependencies

Traffic

Services

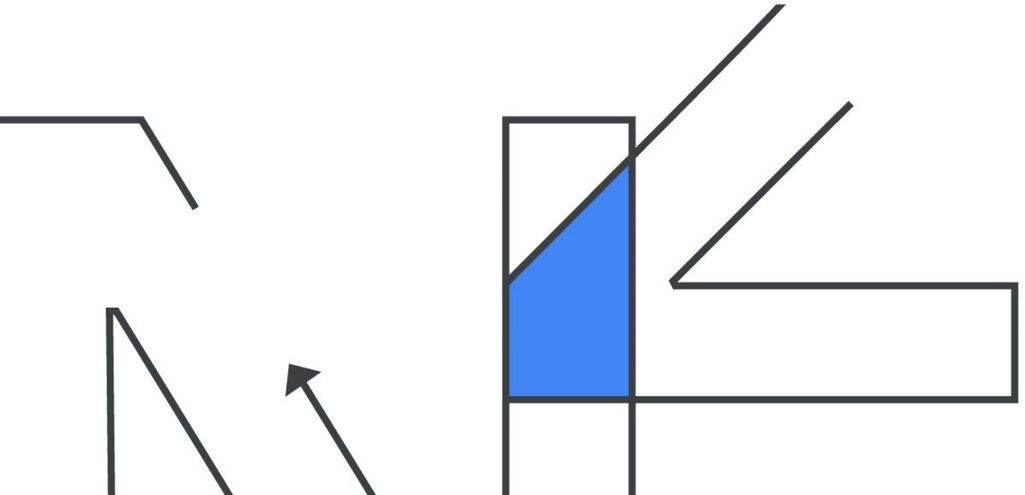
Admin & Support



Mercado Libre’s BigQueue

What is BigQueue

- Publish-Subscribe abstraction layer
- Apache Pulsar as backend service



Why is this important?

- Common standard for async communication between microservices
- Technology evolution without user impact.
- Governance and control on the service usage

Some numbers...

- **~1.5 GB/s** Publisher
- **~4.0 GB/s** Subscriber
- **35M** Produced MPM
- **85M** Delivered MPM
- **80M** Filtered MPM

Pulsar to Pub/Sub

What wasn’t working?

Manageability

- Too many moving parts
- Needed to scale small team
- Upgrades and updates

Scalability

- Grew, and grew...
- Constantly evaluating infrastructure
- Manually scaling, spikes of traffic

Cost

- Peaks and valleys
- Overprovisioned infrastructure
- On calls

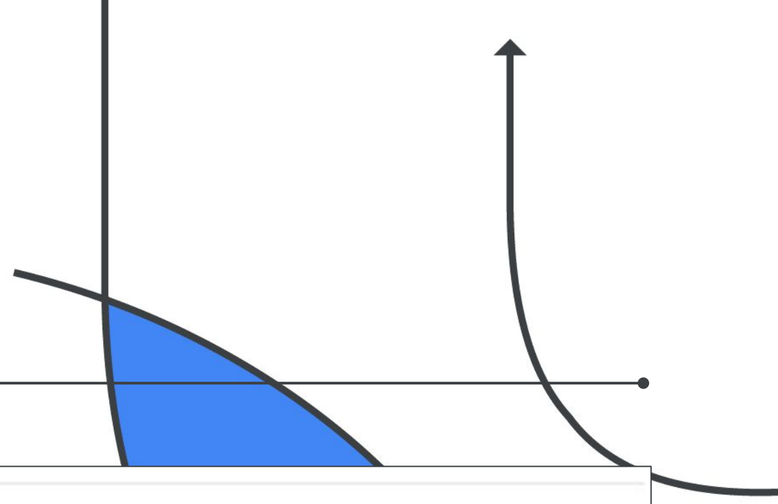
Pulsar infrastructure

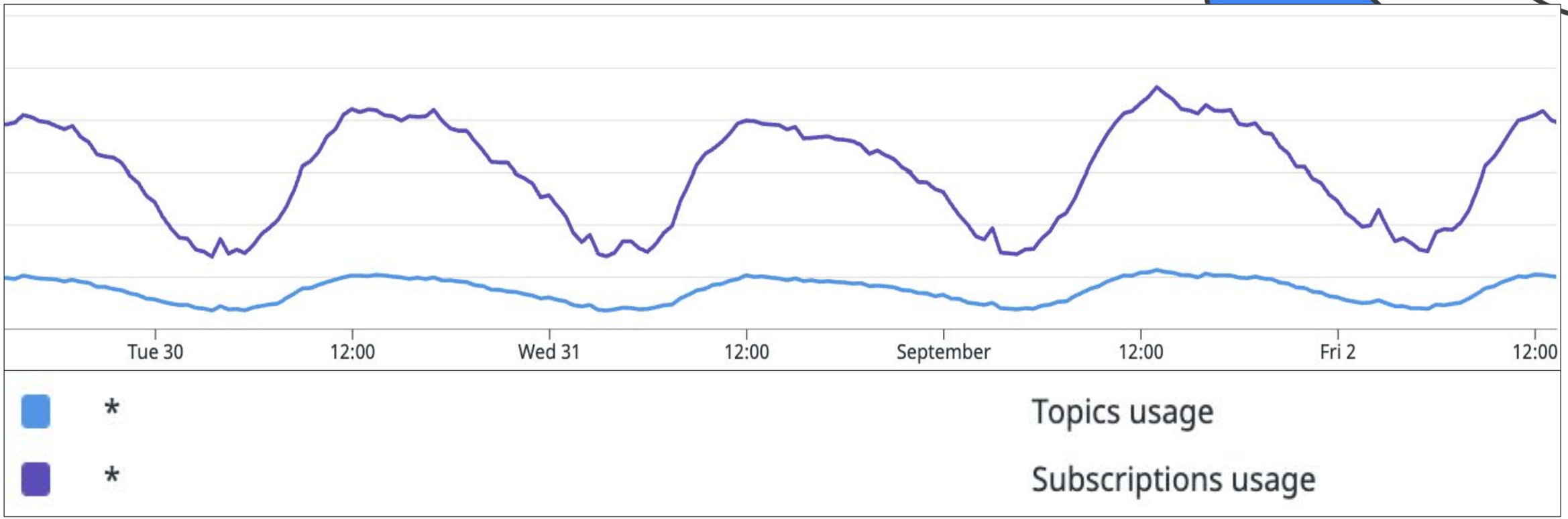
Our Pulsar setup needed to have high performance infrastructure to sustain Mercado Libre's growing needs.

Plus the team had to manage it all!

Hardware	Size
Multi Zone Clusters	+30
Instances	~1500
Provisioned Storage	+2 PB
% Storage usage	**~ 50%**

Consumers sleep

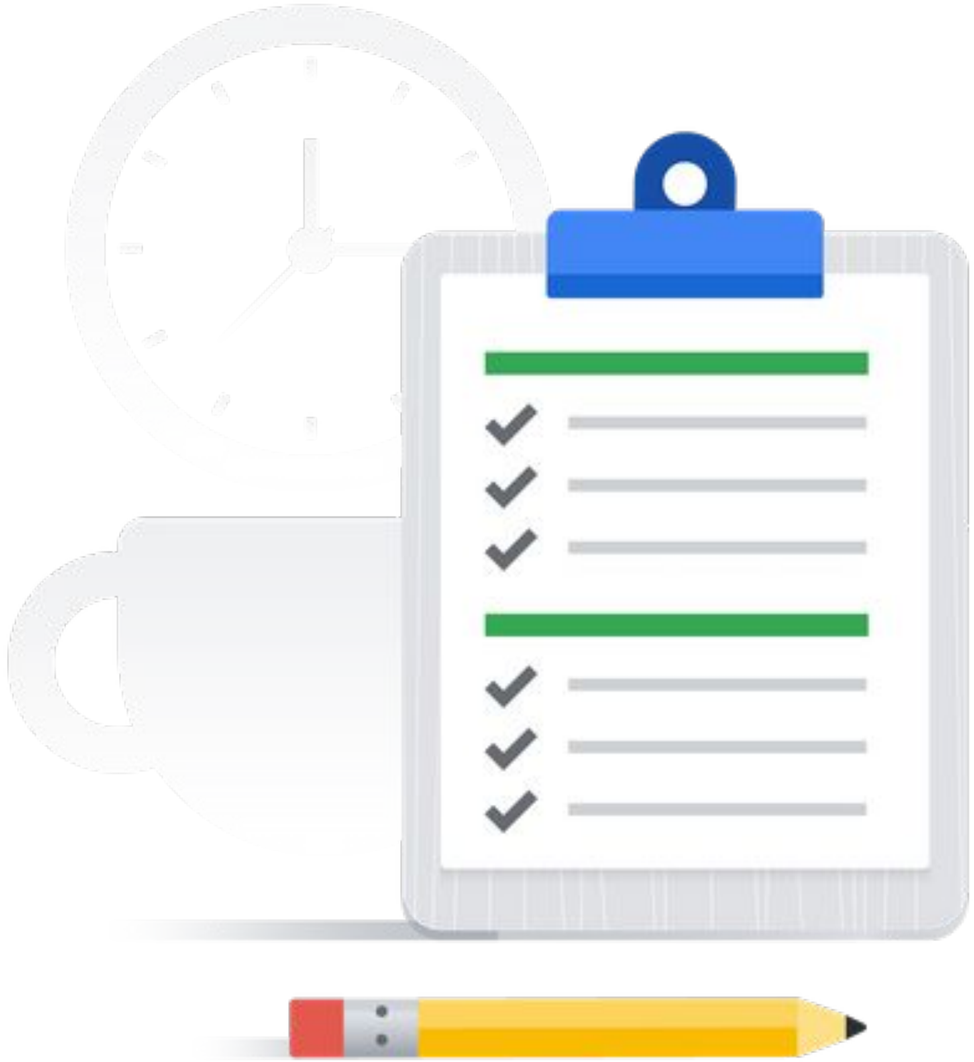




What did Mercado Libre want to solve?

Solution needed to:

- Be scalable
- Reduce strain on team
- Reduce TCO
- Be flexible, stable



Evaluated Pub/Sub

Serverless experience

No capacity planning with Cloud Pub/Sub

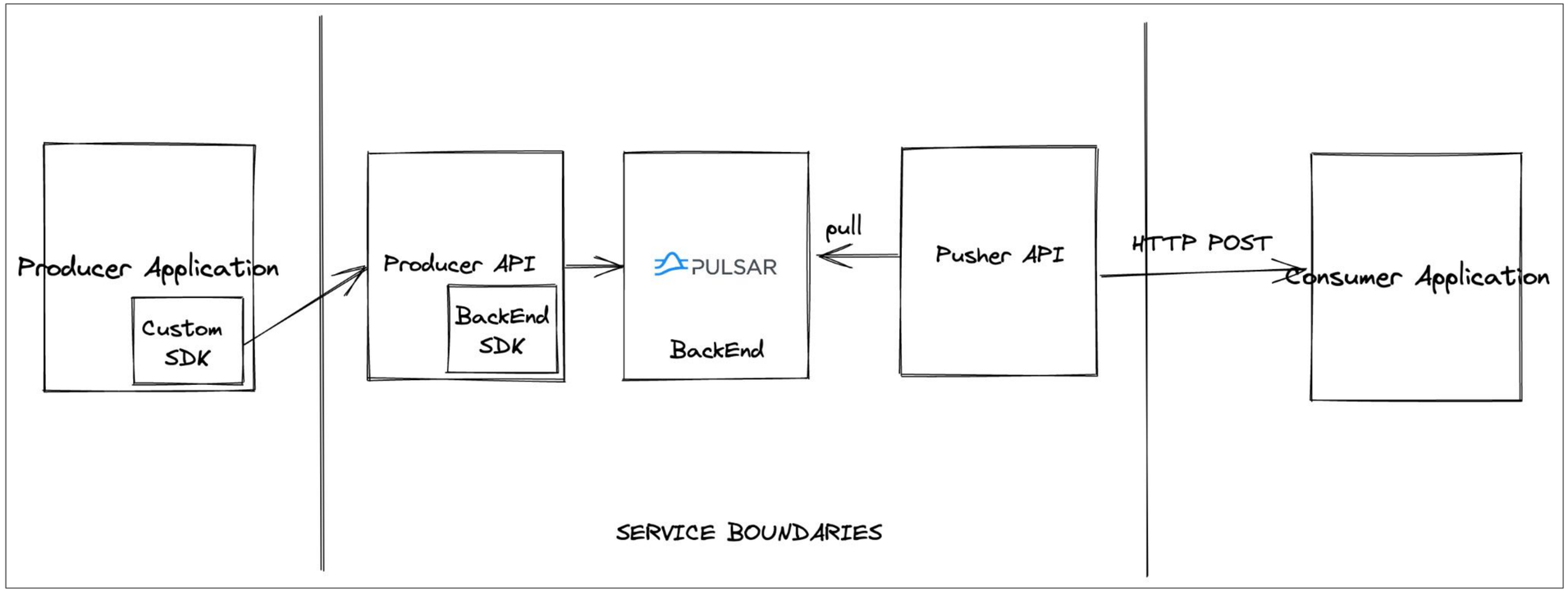
Auth, Authz, Encryption

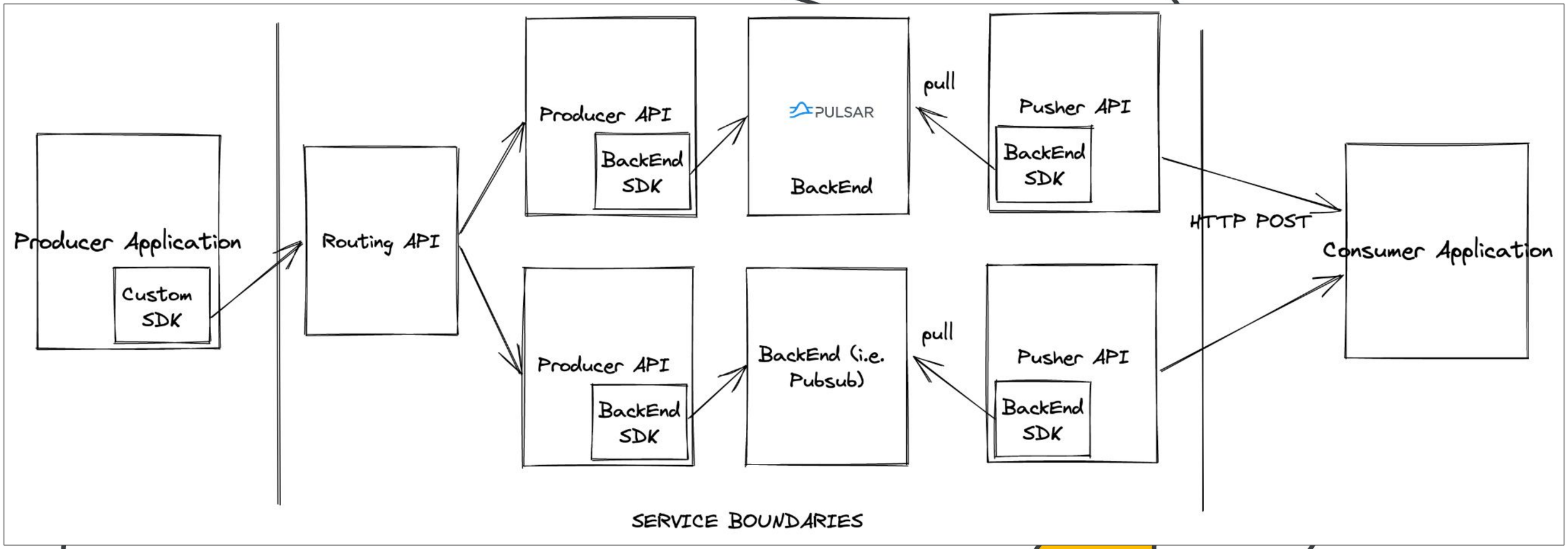
Built-In policy, security certifications

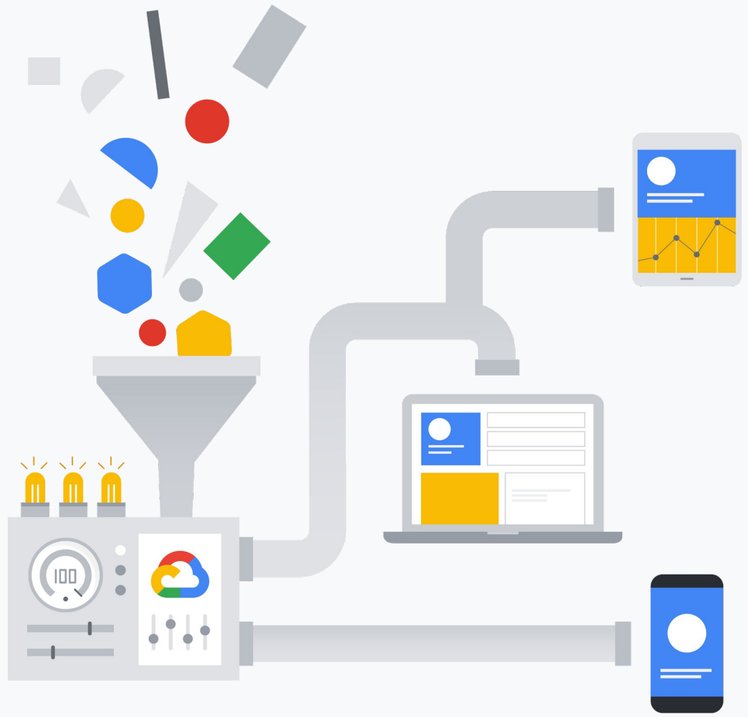
High reliability with 3 zone replication

Pay per byte

Developers didn’t feel a thing




Producer Application
Custom SDK
Routing API
Producer API
BackEnd SDK
Producer API
BackEnd SDK
PULSAR
BackEnd
BackEnd (i.e. Pubsub)
pull
pull
Pusher API
BackEnd SDK
Pusher API
BackEnd SDK
HTTP POST
Consumer Application
SERVICE BOUNDARIES



Lessons learned

Proper CLI or frontend would have saved a lot of time and reduced complexity.

- Relying only on configurations in documents instead of frontends to control the system does not scale
- Maintaining two backend in simultaneous is complex
- Lack of dashboard with current setup is painful
- Using CLI instead of HTTP endpoints would have prevented many human errors

Considerations

Cost of moving to managed service

Cheaper!

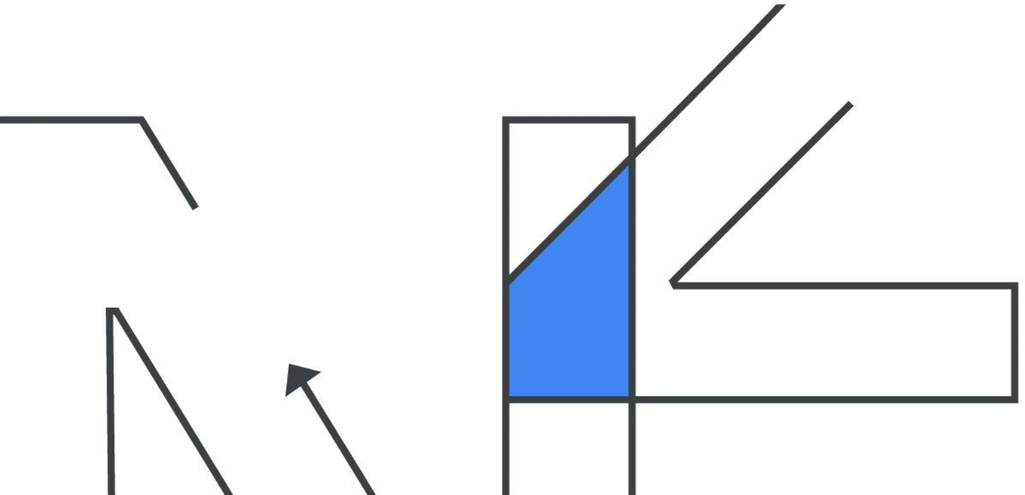
- Pay-as-you-use
- Expertise freed up to enhance business
- Less infrastructure maintenance windows at night

Latency

Introduced network latency between systems.

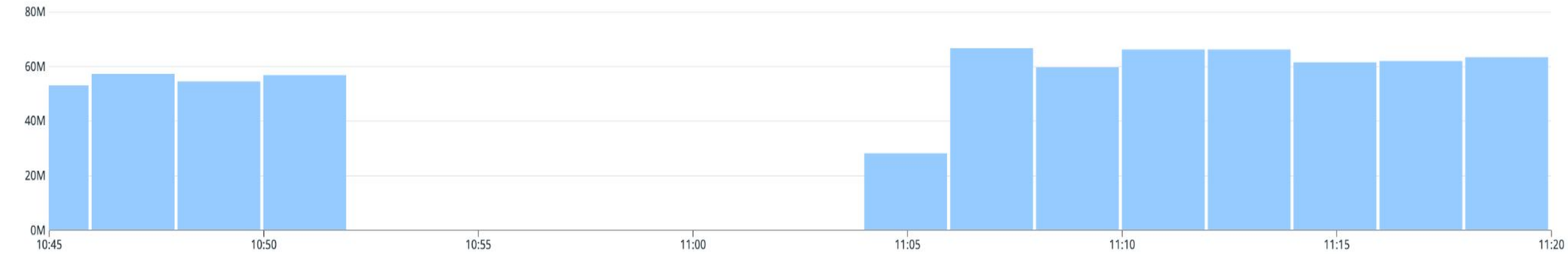
Met migration requirements

- Went from 2ms-3ms range to 9ms-10ms but was acceptable.
- Low latency for time between message produced and message consumed in healthy subscribers < 150 ms

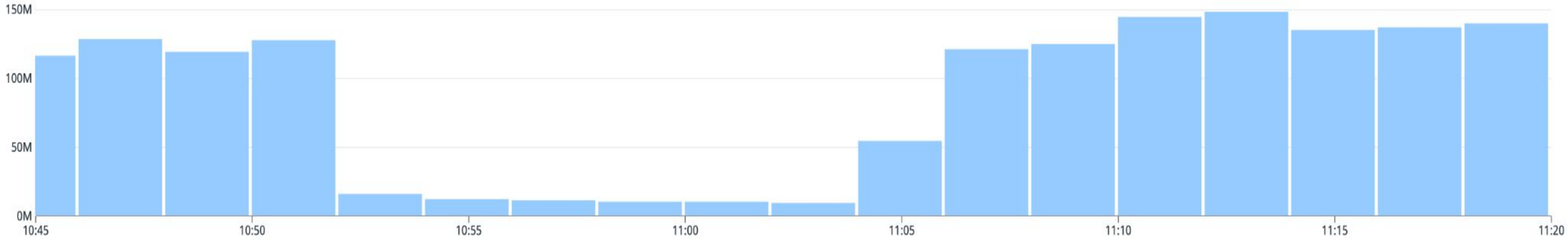


Elastic Scaling

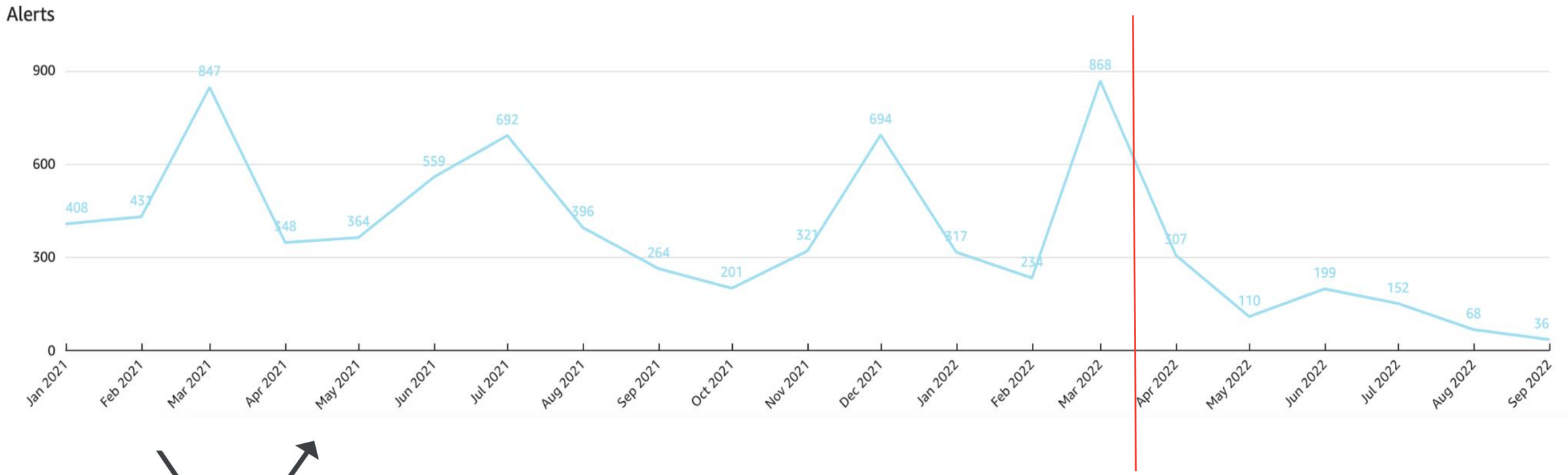
High spike of traffic producing messages to topics



High spike of traffic consuming messages from subscriptions



On call alerts



100% in Pub/Sub

Ecosystem

The modern approach to continuous intelligence

Early attempts: Pinpoint Streaming data engines

Server-based

- Inflexible
- Lacks scalability
- Bottlenecked

Stand-alone

- Disconnected
- Incomplete
- Insufficient

Engineering-focused

- Complex
- Resource intensive
- Business adverse

Modern approach: Unified intelligence platform that brings together streaming and batch data in one tool

Serverless

- Elastic
- Performance at any scale
- Boundless possibilities

Fully integrated

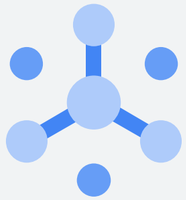
- Connected in all dimensions
- Complete business capabilities

Customer focused

- Simplified
- Automated
- Business friendly

Pub/Sub

Hyperscale, subscription messaging when people, applications and machines need to connect insight to everything



Pub/Sub

Messaging and event ingestion for real-time analytics

Scalable, durable event ingestion and delivery
Serverless, auto-scaling & auto-provisioning; 1KB/s to 100GB/s with consistent performance; 3x3 data replication; storage for up to 31 days

Scaled Publish/Subscribe Pattern
Up to 10,000 subscriber apps per topic; independent capacity & billing for publishers and subscribers; push & pull delivery modes

Global Routing
Publish and subscribe to events regardless of geography

Deeply Integrated
Scalable analytics with Dataflow, serverless actions with Functions, built in monitoring, audit logging, and compliance.

Dataflow

The backbone of data analytics on Google Cloud, with data pipeline optimization and automation that removes the complexity of continuous intelligence

Dataflow

Highly automated, self-healing data pipeline optimization for streaming and batch

Ultra simplicity
Serverless, auto-provisioning, and self healing; automated performance, work balancing, unified batch and streaming under one lane

Accessible ML
Out of the box support for MLOps (TFX, Kubeflow), ML Inferencing (PyTorch, TensorFlow, and scikit-learn), GPU and large scale data processing

Best of OSS and optimized platform
Innovate with Apache Beam SDK; create pipelines in your language of choice; optimize workloads with built in monitoring and observability

Built to scale
Designed and built for horizontal and vertical scaling (first in industry). Optimize pipeline performance at any scale; maximize utilization to save costs

Read more: https://cloud.google.com/blog/topics/developers-practitioners/dataflow-backbone-data-analytics

BigQuery

Cloud data warehouse, combining and analyzing all data, all analytics, for all users at all times

BigQuery

Next generation cloud data warehouse

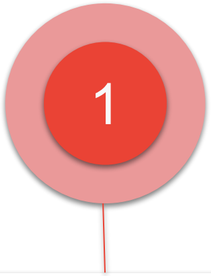
Built for real-time and historical, highly complex analytics
Serverless, auto-scaling & auto-provisioning, plus a streaming buffer for streaming analytics at any scale (billions of events per second)

Designed for universal analytics use cases
Suitable for all data and all analytics; perfect for real time and just in time data and analytics

Architected with intelligence for simplified operations
Automated performance, maintenance, and management to eliminate the need for DBAs

Embedded with machine learning and location analytics
Operates ML use cases without an MLOps platform; make complex geographic analytics as easy as SQL

HackFest to Accelerate Time to Value



What's the HackFest?

- 2 hrs content
- Demo & deep dive
- Customer use case specific demo

What do you get from a HackFest?

- Demonstrate your use cases on GCP
- Access and training to new offering
- Accelerate to Pilot
- It's ***FREE*** - Contact your rep to sign up

Customer Quotes:

- "Good value of our time. Helped us to understand GCP capabilities better.."
- "Accelerated bulk of the team to have better understanding of GCP streaming capabilities. Built more trust and team bonding to execute the project."

Thank you

Google Cloud

Next '22