*Building a **Flexible Data Infrastructure** for Streaming Video*

Michael Skariah, CTO
Datazoom - The Video Data Infrastructure Platform
www.datazoom.io

mhv/2019
July 2019, Denver, CO

About Michael Skariah

Michael Skariah is the CTO of Datazoom which he joined in November 2017. As a 15+ year veteran of Software Technology Management, with a focus on Software Architecture, Software Development, and Integration. He is an experienced technology leader and was the first employee at two previous start-ups.

His experience includes managing globally distributed Big Data teams to deliver multiple product offerings including Analytics and Personalization products in all phases of the development. He excels at structuring complex tasks across large teams to deliver solutions on time and on schedule.

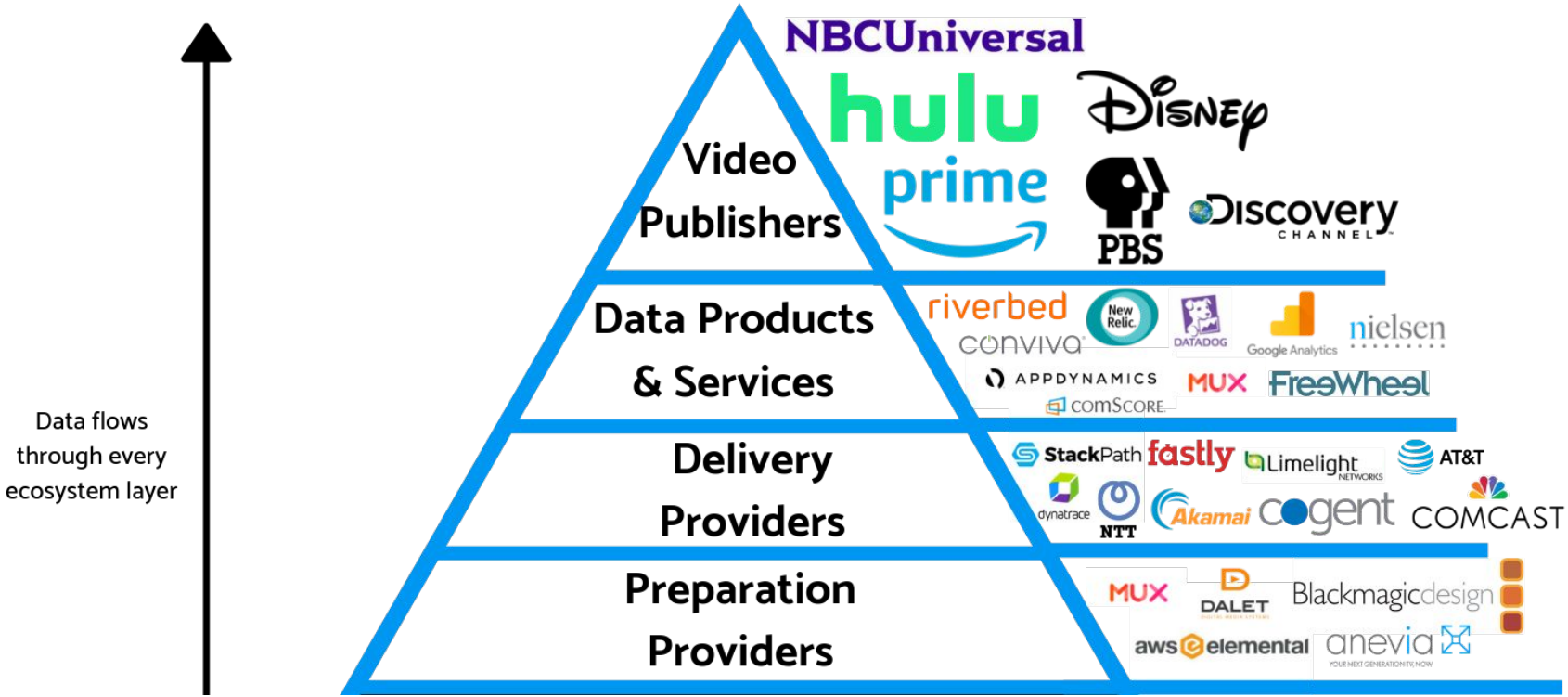
He was previously the Director of Engineering at Ooyala.

"Organisations struggle to corral data into useable and actionable intelligence. **Without clean**, **relevant**, and **labeled data**, organisations are **stymied in their efforts to move aggressively on AI**, which CEOs overwhelmingly 'agree' will have a significant impact on their business **within the next five years**."

PwC's 22nd Annual CEO Survey (1,378 CEOs in +90 territories), 2019

What ***data*** do video technicians have today? How good is it? What else do they need?

Who suffers from video data problems?



Review of Current Data Challenges & Questions

Embedded Subjectivity

Can we consistently reproduce our results? Subjectivity is carried into all downstream analysis & functions.

Data Silos & Restricted Access

Do we have all relevant data in one place?

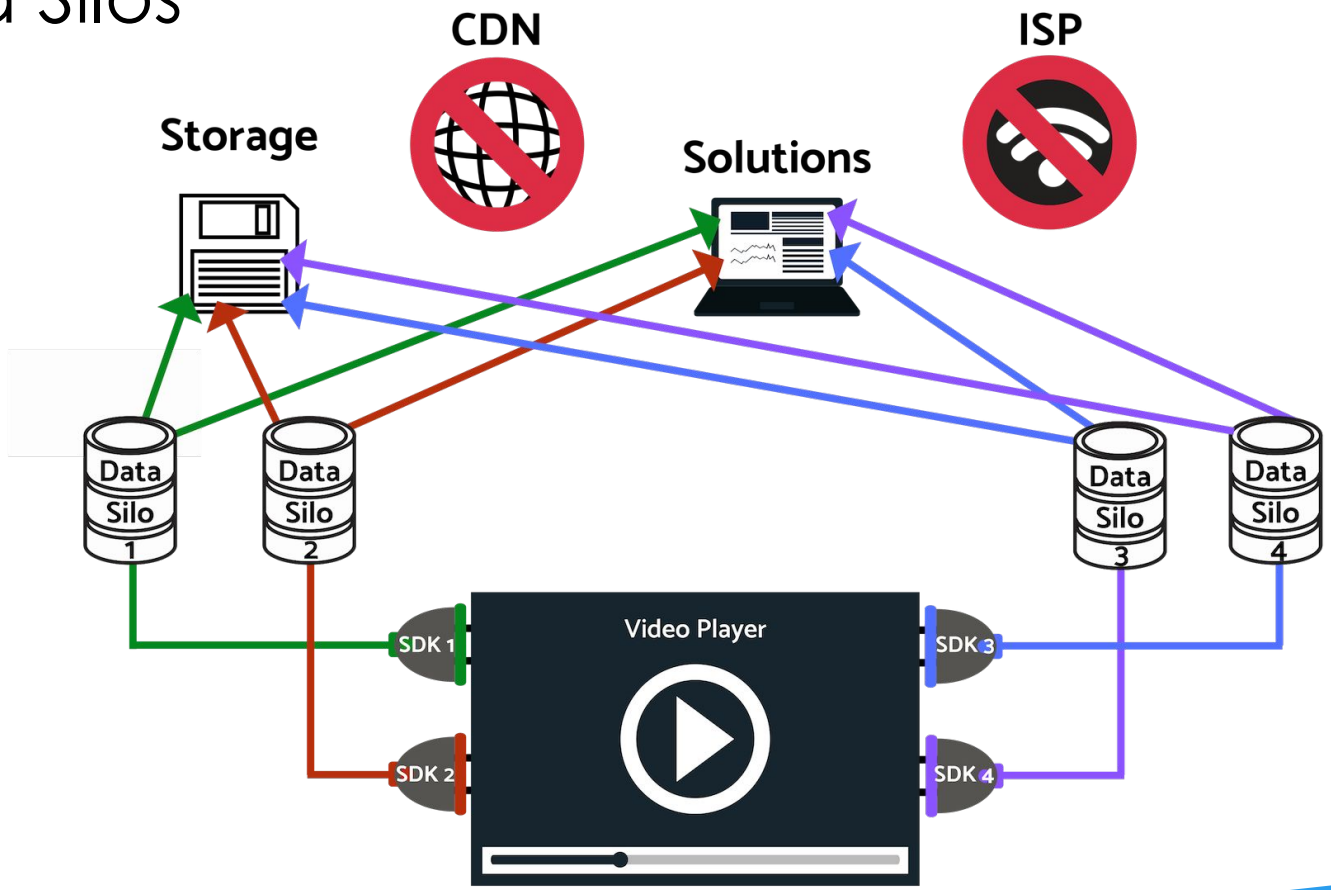
Lack of Data Standardization

Is data prepared to be used AI and ML systems to use it?

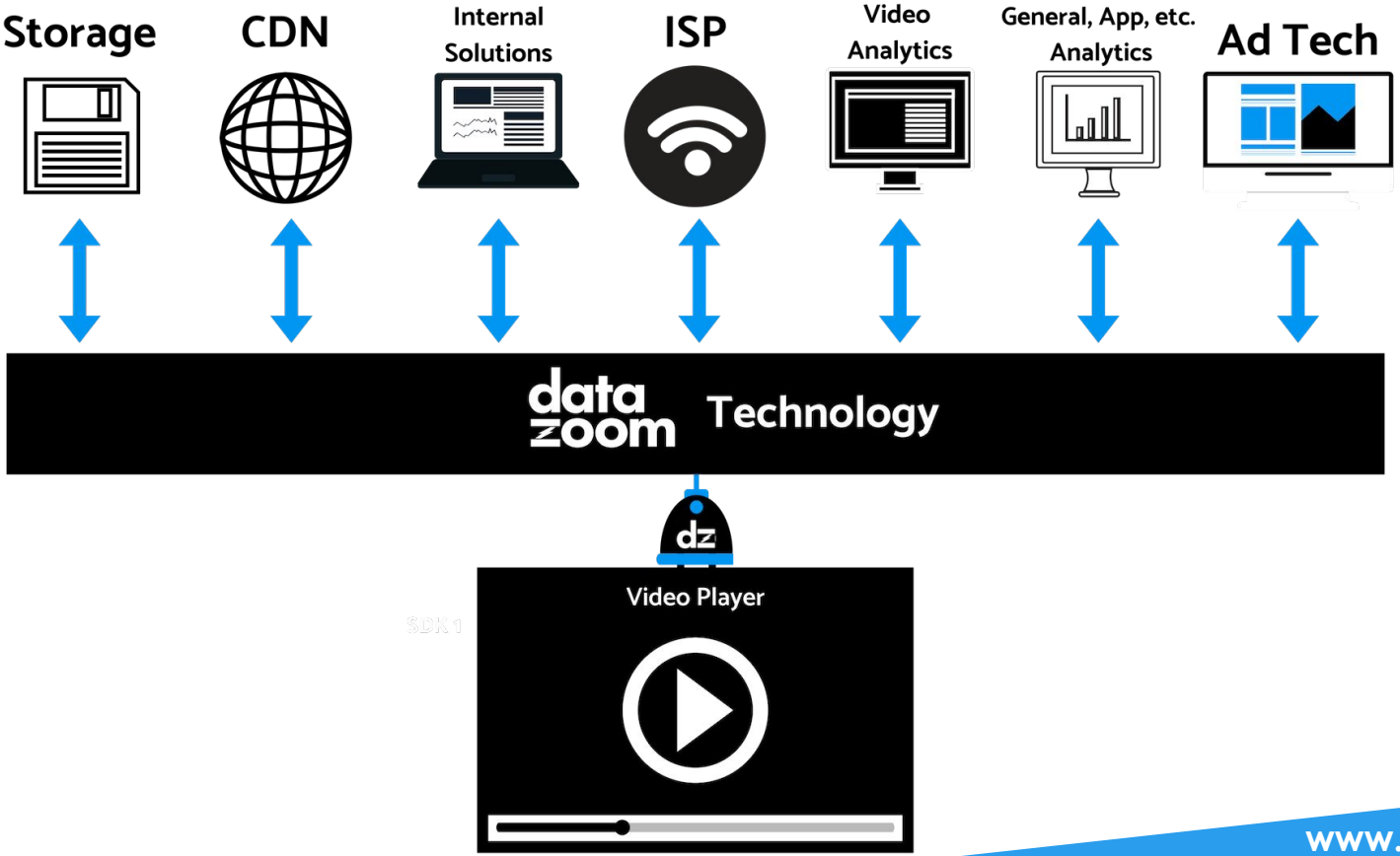
Slow Data

Do we have access to data in a relevant time-frame so that real-time Automation can have meaningful results?

Data Silos

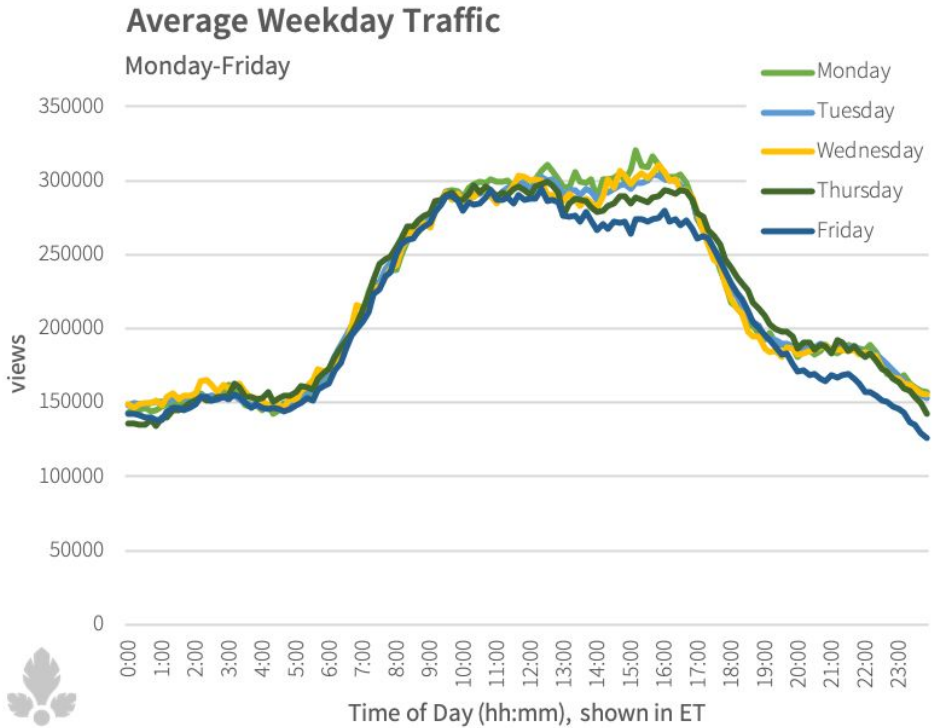

CDN
ISP
Storage
Solutions
Data Silo 1
Data Silo 2
Data Silo 3
Data Silo 4
SDK 1
SDK 2
SDK 3
SDK 4
Video Player

How Datazoom simplifies the "data supply chain"



Ensuring High Availability to Guarantee Autoscaling

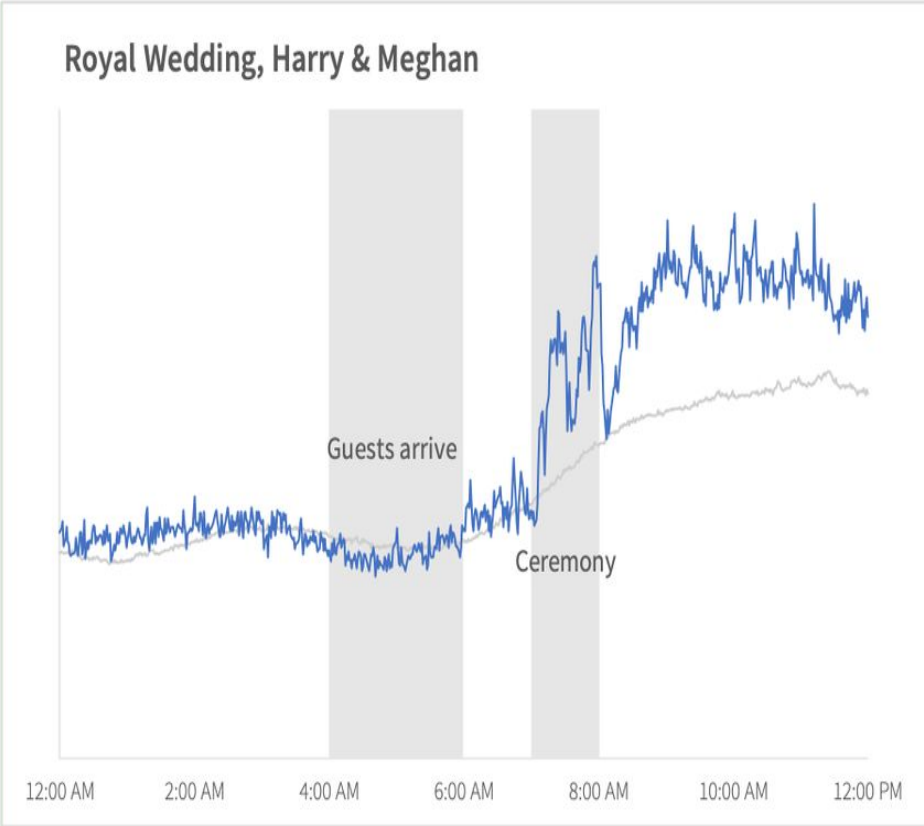
The Stability of (Live)Streaming...



Generally, we have systems designed to respond to general and expected user behavior.

Source: Pars.ly

...is unpredictable

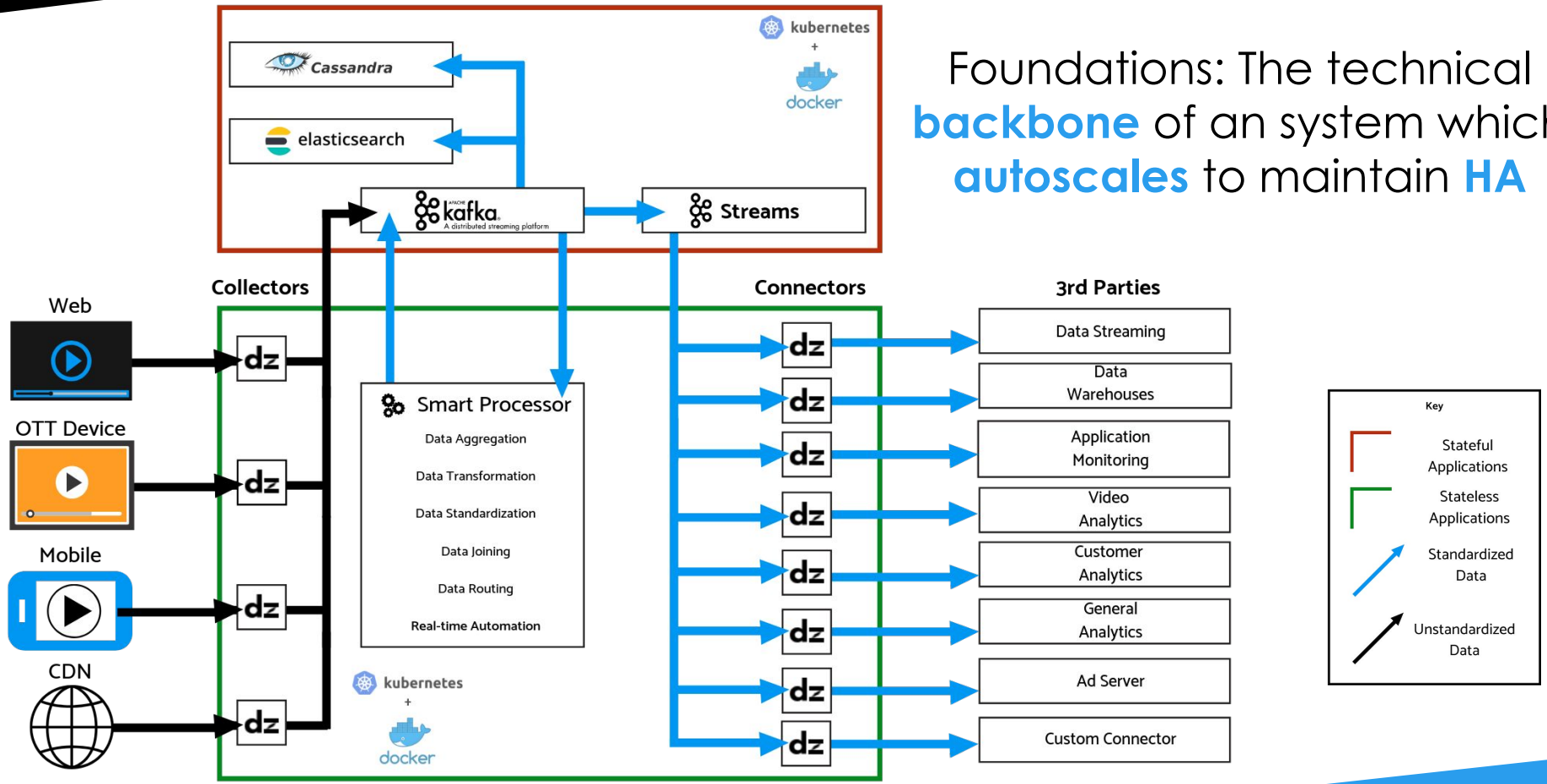


Source: Pars.ly

But we must meet the user demand, it doesn't work in reverse. When we consider live events like the wedding of Harry Windsor and Meghan Markle, our "generally expected patterns" fail us.

As the scale of users continues to grow, and the importance of streaming surpasses traditional television for live events, maintaining a highly available data infrastructure to respond to these changes becomes increasingly important.

Foundations: The technical **backbone** of an system which **autoscales** to maintain **HA**



Three Requirements for the Platform

1. Robust - stable even when exposed to different unforeseen events
2. Resilient - quickly recover from disorders and become stable
3. Reliable - ability to work without failure and meeting SLA's

Factors influencing Reliability

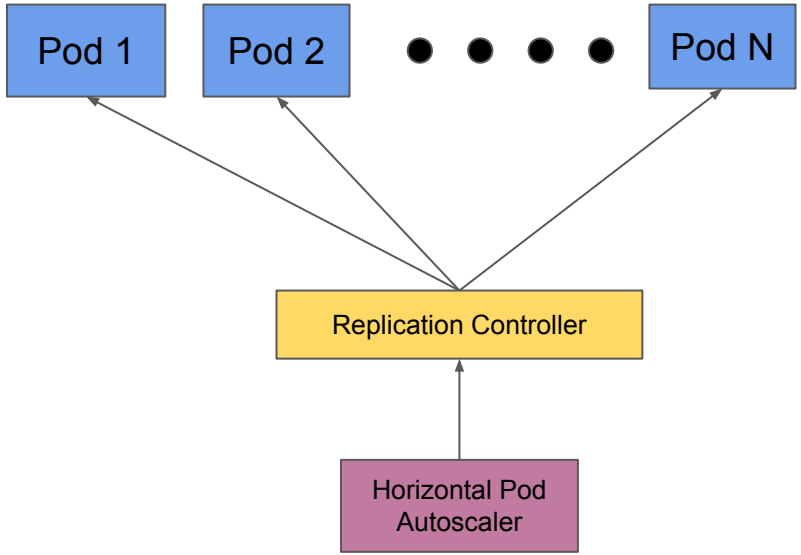
Availability: Deployed in multiple availability zones across the globe

Replication: Multiple pods for a service to ensure HA

Autoscaling: Cluster level and Pod level (**Horizontal Pod Scaling; HPA)** technology using following signals

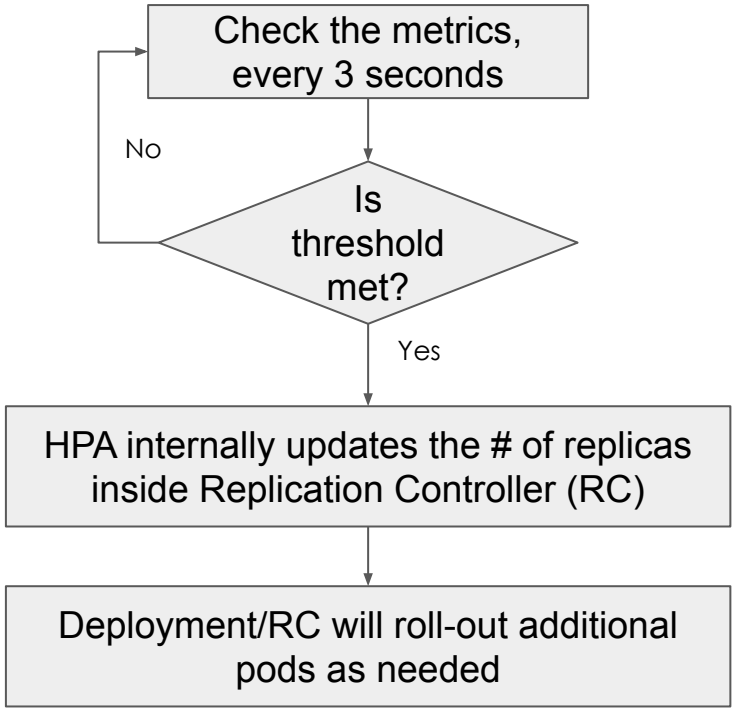
- CPU
- Memory
- Custom
- Rate

Horizontal Pod Scaling

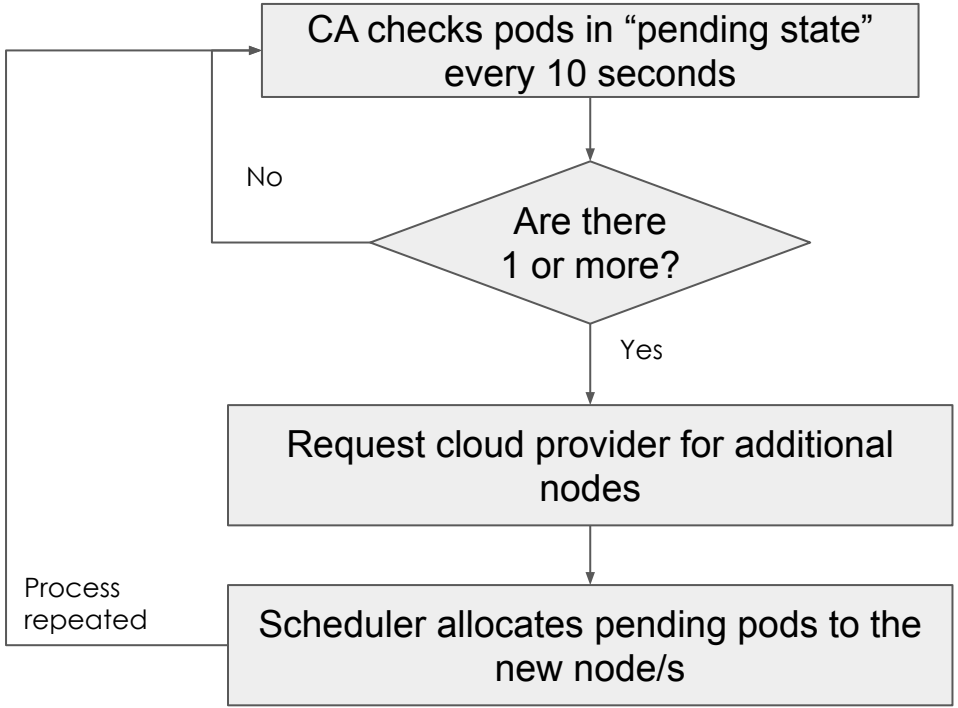


Type / Entity	Nodes	Pods
Horizontal	# of nodes	# of pods
Vertical	resources for a node	resources for a pod

Pod scaling flow



Cluster scaling flow



Timings on scale up and scale down

Pods:

Scale up: within 3 seconds

Scale down: 45 min inactivity

Node:

Scale Up: 5 min

Scale down: dependent on pod allocation

Example with custom metrics

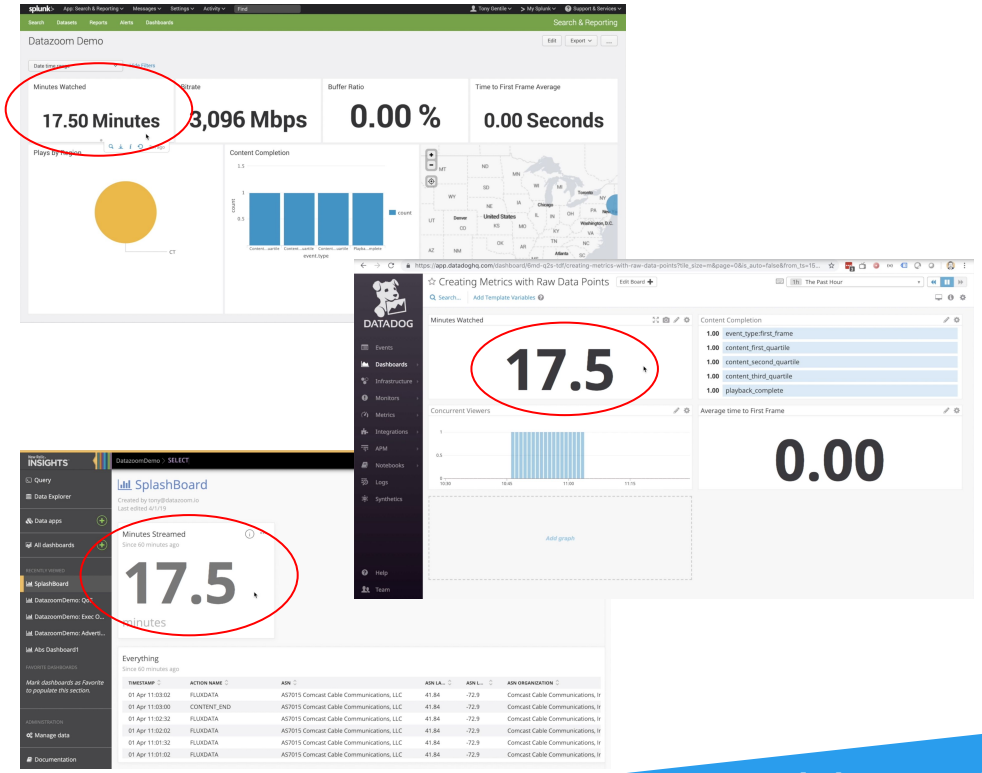
Configuration for event receiver service

```
kind: HorizontalPodAutoscaler
apiVersion: autoscaling/v2beta1
metadata:
  name: dz-message-broker-hpa
spec:
  scaleTargetRef:
    kind: Deployment
    name: dz-event_receiver
  minReplicas: 3
  maxReplicas: 20
  metrics:
  - type: Pods
    pods:
      metricName: concurrent_events
      targetAverageValue: 50000
```

Conclusions and Applications

Joining and Visualizing CDN Logs (Fastly) with player data, visualized in Splunk

Enriching and Aligning Systems for Video Analytics (Splunk, Datadog, New Relic Insights)



*These are just samples. Email michael@datazoom.io for more use-cases like this.

Questions ?

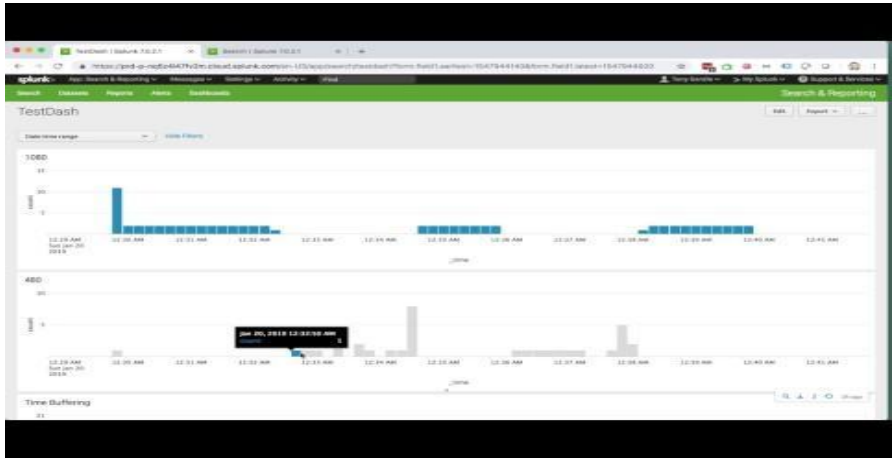
Want a copy? email: **michael@datazoom.io**

*Building a **Flexible Data Infrastructure** for Streaming Video*

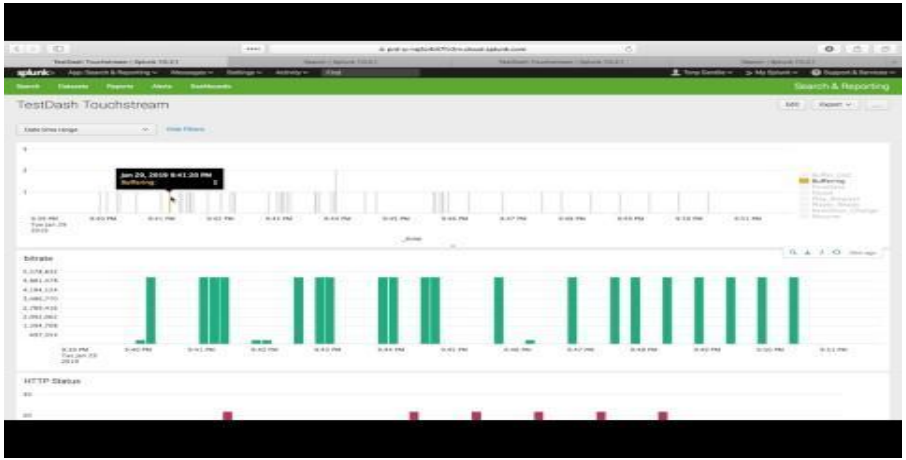
Michael Skariah, CTO
michael@datazoom.io
Datazoom - The Video Data Infrastructure Platform
www.datazoom.io

Examples and Additional Materials

Videos Demonstrating Use Cases



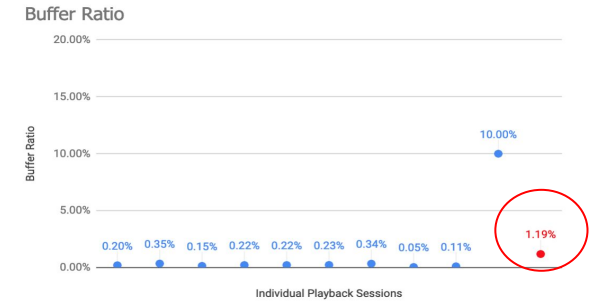
Joining and Visualizing CDN Logs (Fastly) with player data, visualized in Splunk

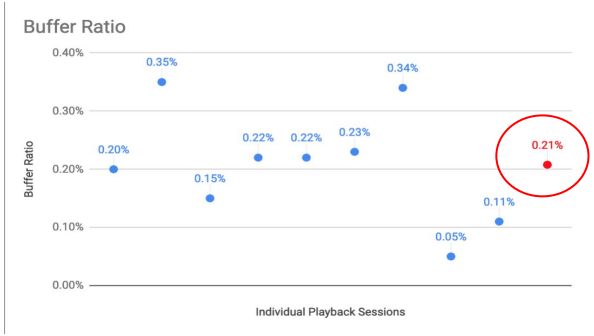


Validating Synthetic QoE tests (Touchstream) with real-time playback data, viewed in Splunk

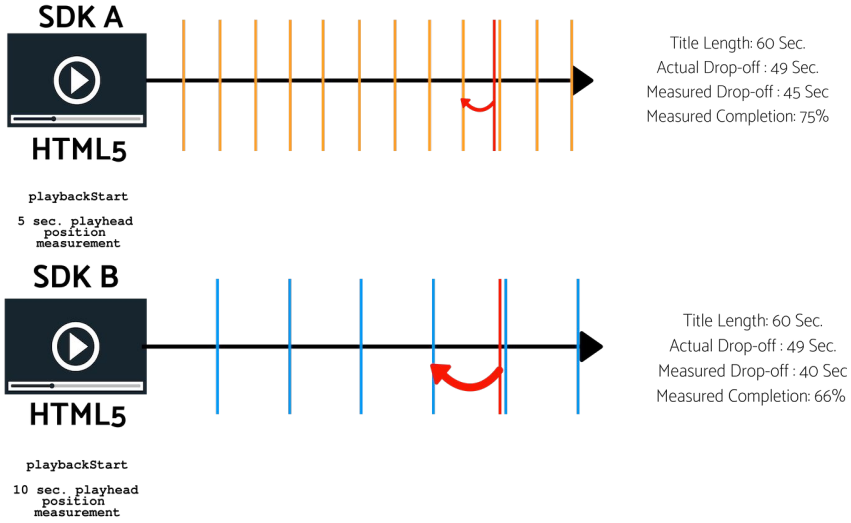
Metrics Limit Accuracy & Precision.

Different data **calculations** are used.
Is buffer ratio 1.19% or .21%?

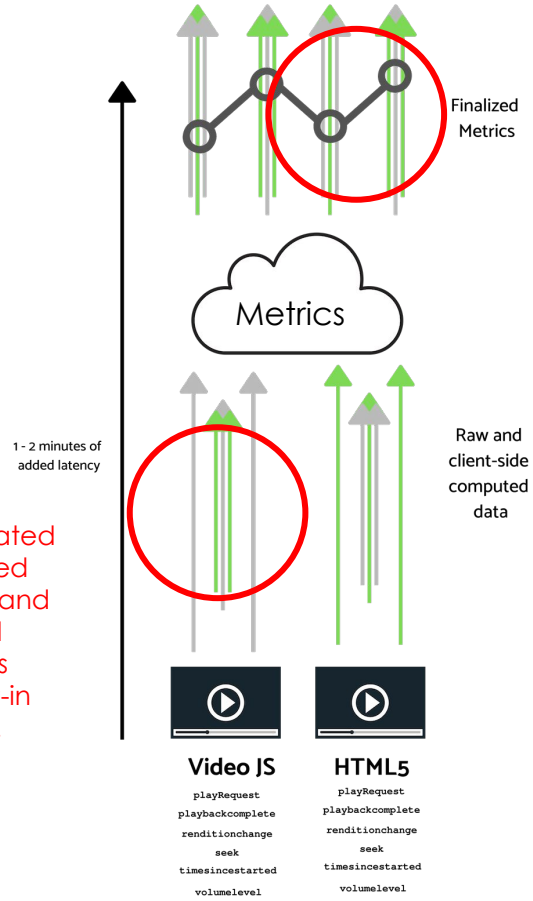




Different data **measurement** is taken.
Is completion rate 75% or 66%?

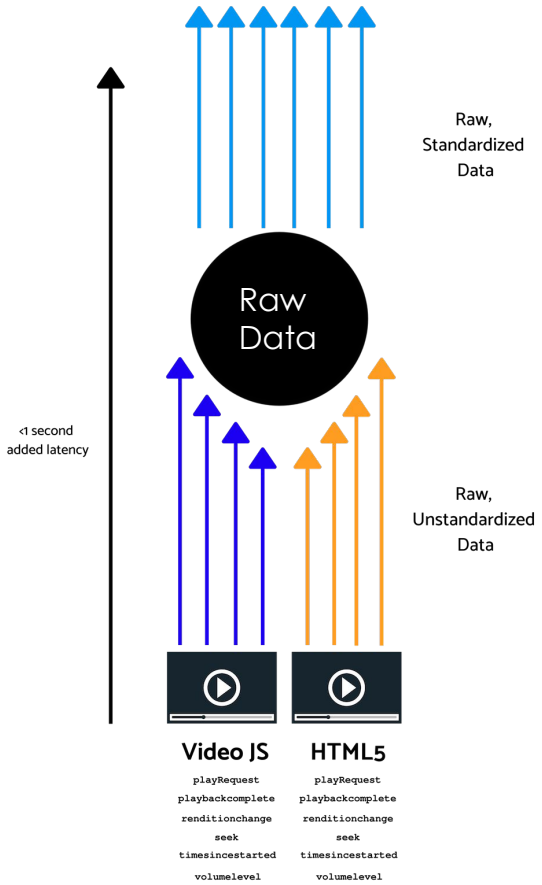


Granular raw data opens new opportunities.



*Pre-calculated data created client-side, and obfuscated calculations create built-in subjectivity.

Moving from Subjective to **Objective** Measurement.



Datazoom: The holistic 'datatecture' for streaming video