
AWS
re:Invent

ARC310

Managing the serverless challenge: Running big data on AWS Lambda

Opher Dubrovsky
Big Data Dev Lead
Nielsen

Boaz Ziniman
Principal Developer Advocate
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How you can build data pipelines using serverless



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.


DataOut
250 billion events per day
Running on serverless!

This is a true story.

The events described in this talk took place at Nielsen Marketing Cloud in 2019ć2020.



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

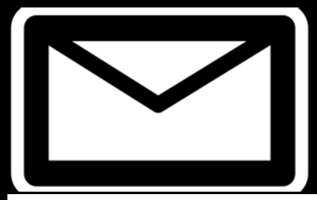
September 5, 2019
10:45 AM



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

September 5, 2019 10:45 AM



[redacted] com> Thu, Sep 5, 2019, 10:45 AM

to [redacted] me, [redacted]

Hi Exelate/Nielsen,

Two weeks ago an engineer flagged the incoming QPS from your side to our API, this QPS is still so intense our infrastructure is at a strain coping with it.

- We would like to understand why there are so many requests being made?

DDoS attack

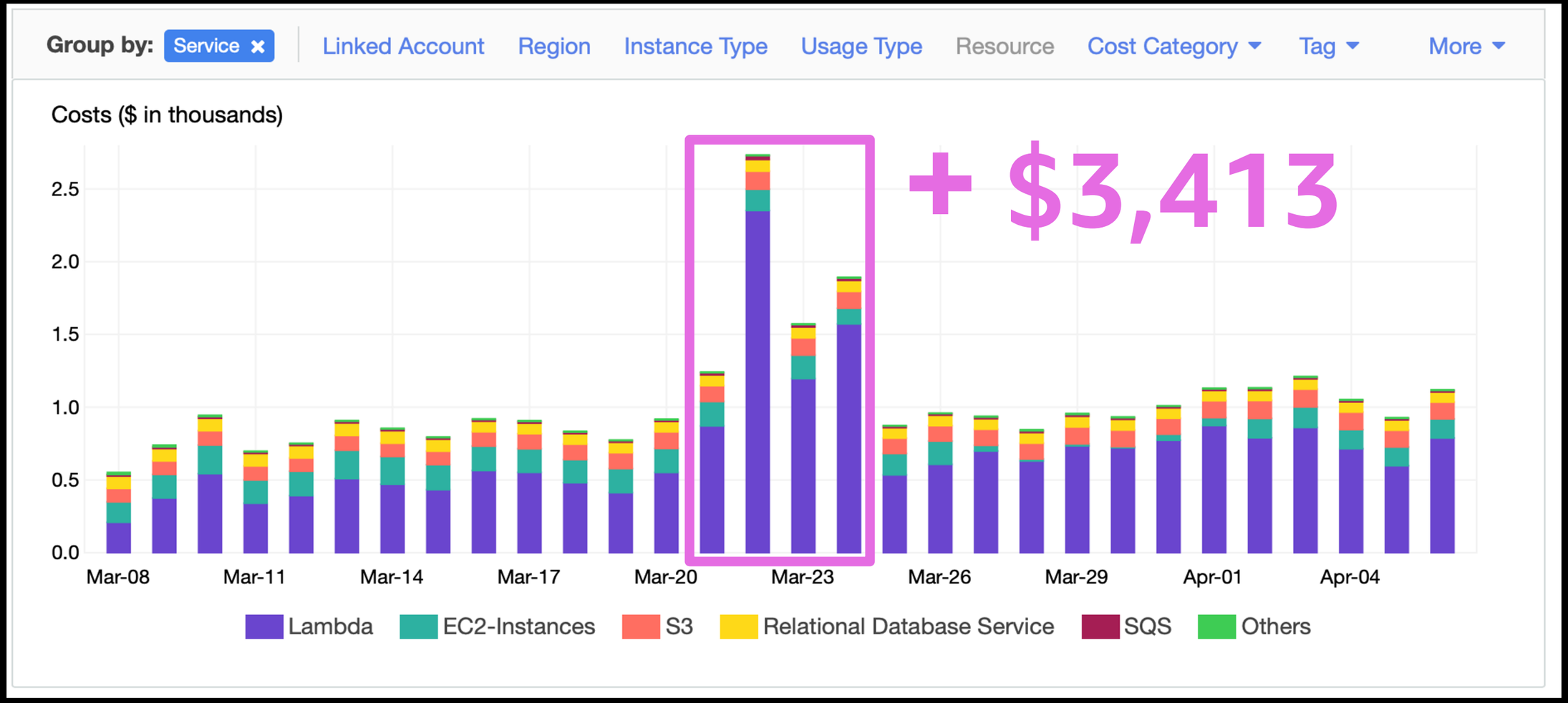
March 21, 2020
02:30 PM



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Mar 21, 2020



Reprocessing bug

Powerful tools = large-scale problems



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Building a massive serverless data pipeline
- Data and stream serverless patterns
- Avoiding pitfalls
- Takeaways



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

nielsen

Nielsen Marketing Cloud



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Nielsen Marketing Cloud (NMC)



- Acquired by Nielsen in 2015 (eXelate)
- Builds marketing segments

- **Data used for:**
  - Targeting (campaigns)
  - Business decisions

NMC in a Nutshell

nielsen

aws

Cloud Native

~60TB/day
5 PetaByte

Spark

~6000 nodes/day

~6 million/day

nielsen

The DMP problem



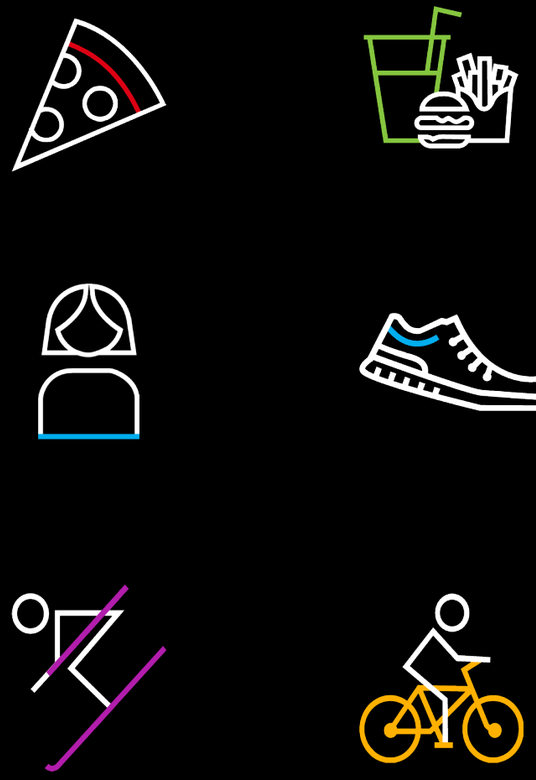
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Segments → Platforms → Campaigns

nielsen

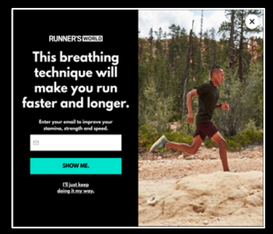
Segment



Upload

amazonadvertising

Run Campaigns





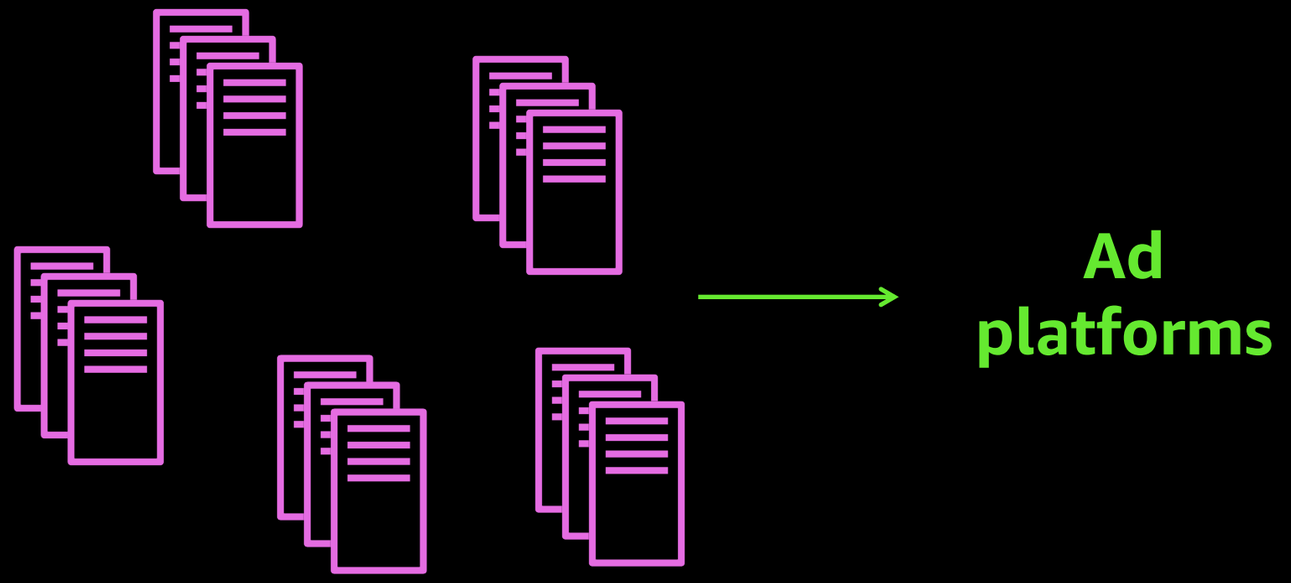






Our old system – lots of scripts

- Scaling was hard
- Constant support
- No visibility
- Development for each platform



Constant problems + slow fix time !

nielsen

Looking for a better way



AWS re:Invent

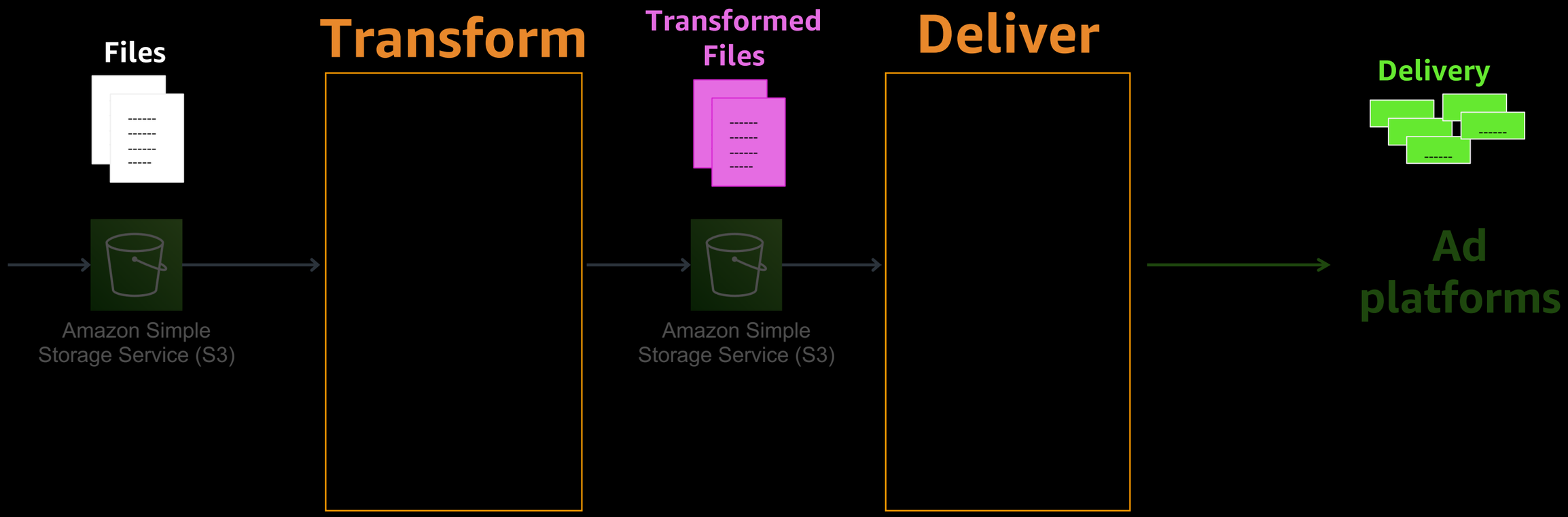
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

System goals

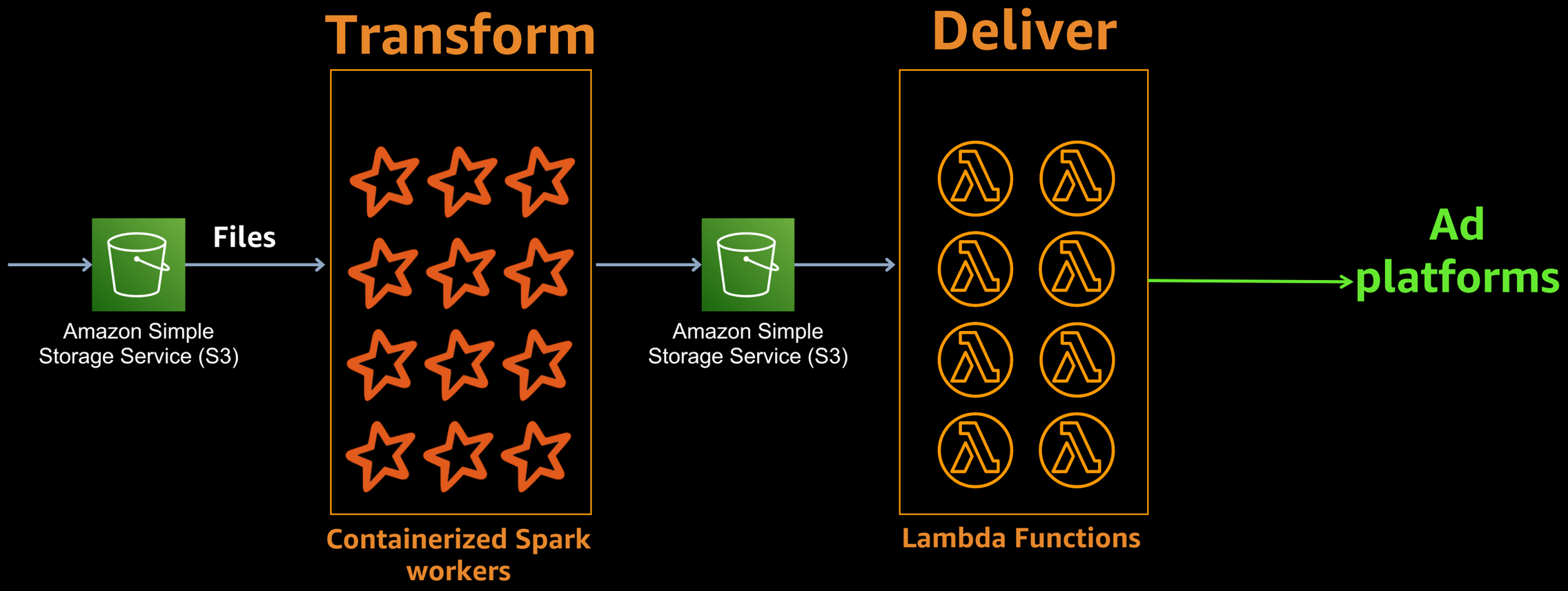
nielsen

- Scales well + handle bursts
- Cost effective
- Easy transformations
- Centralized metrics

DataOut – Basic architecture

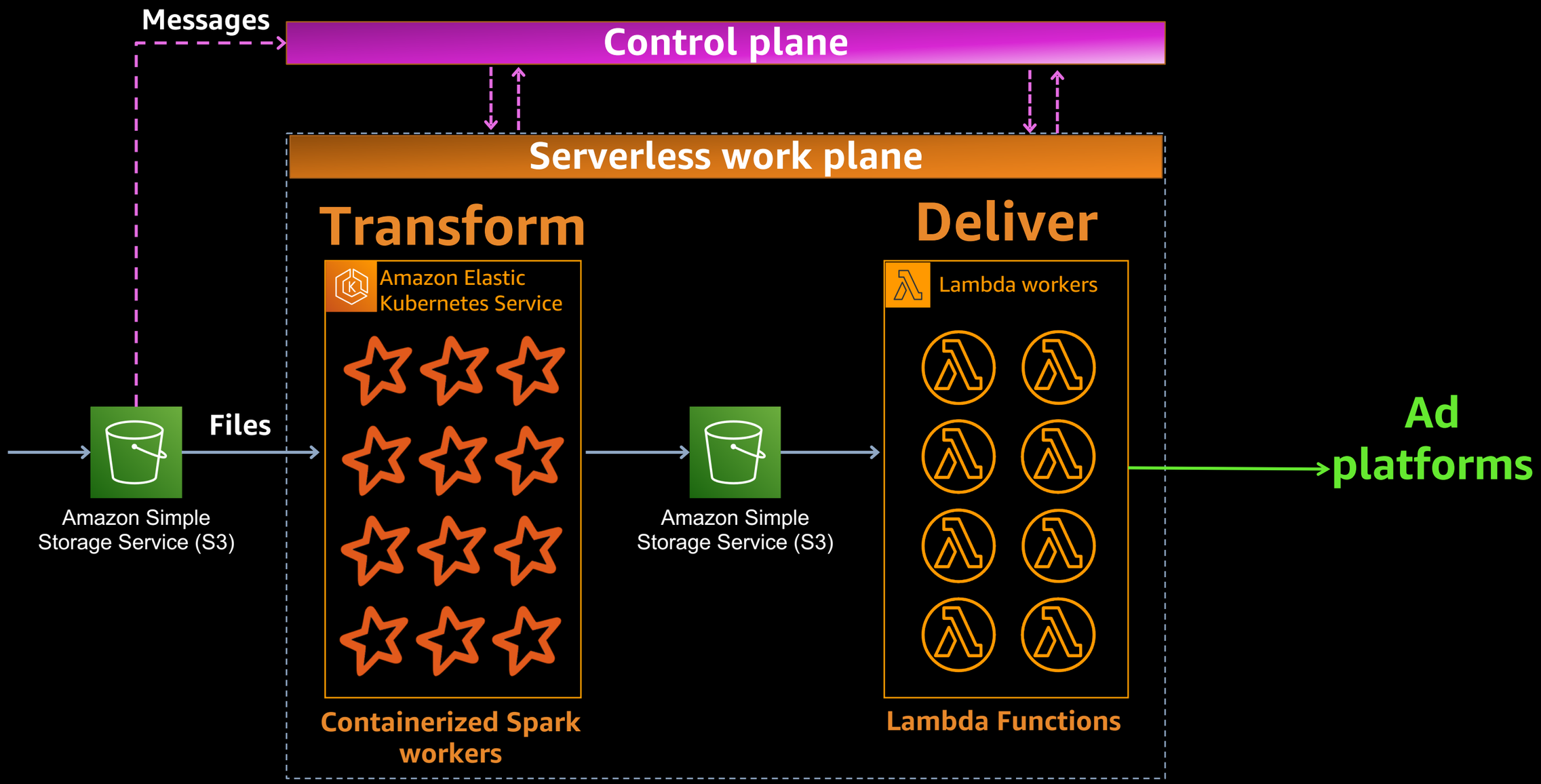


DataOut – Basic architecture

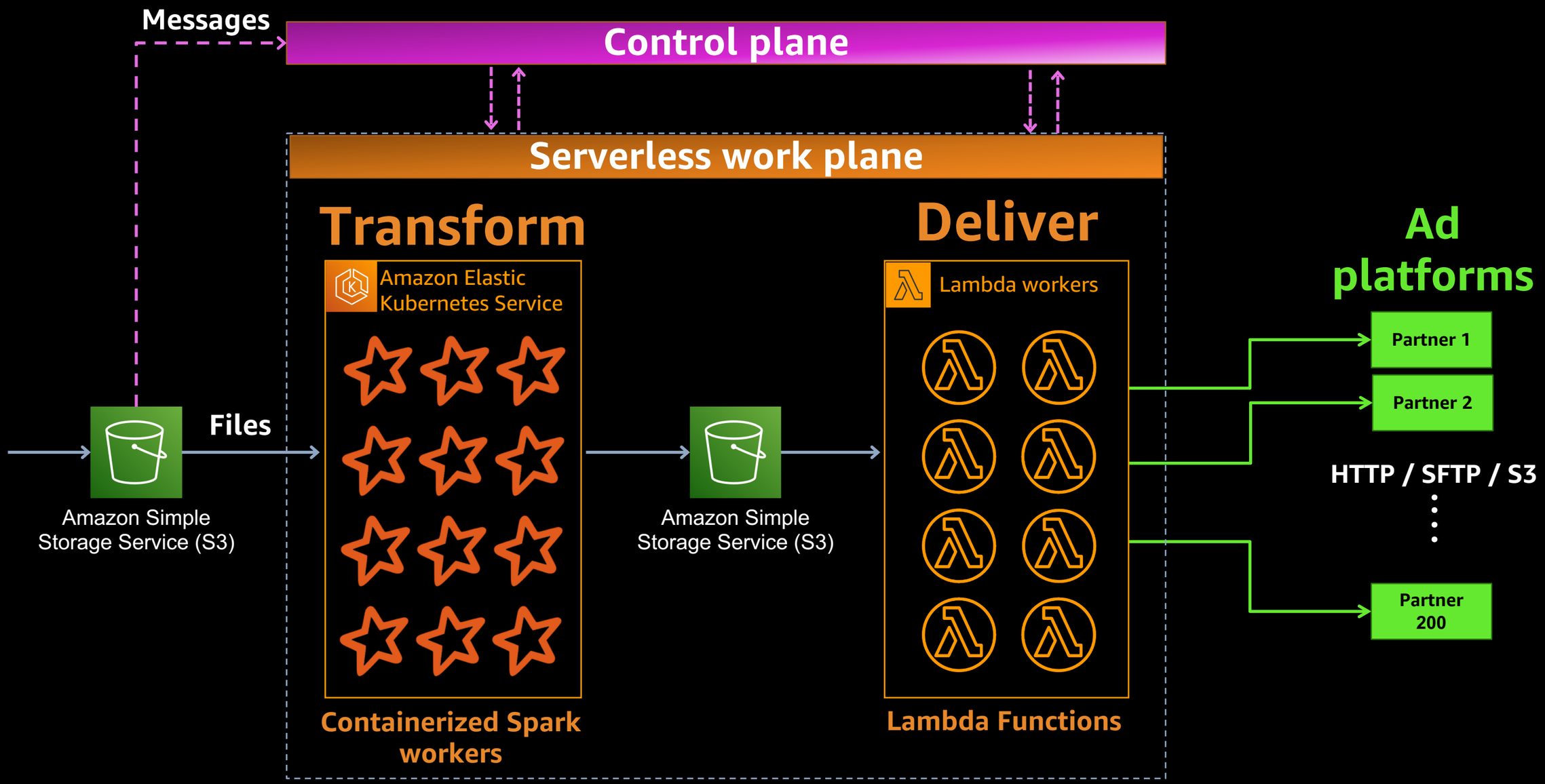


DataOut – Basic architecture

nielsen



DataOut – Basic architecture



nielsen

Transform

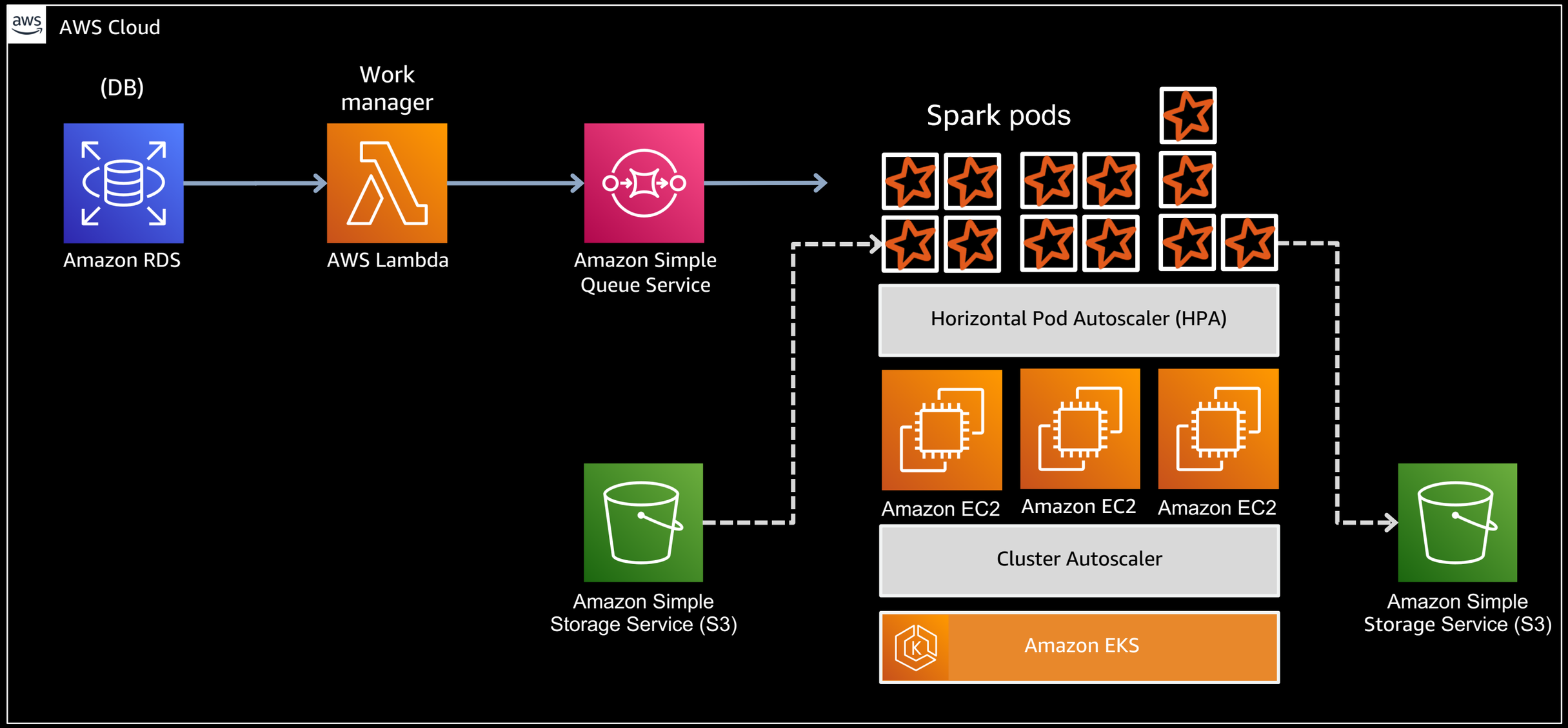


AWS re:Invent

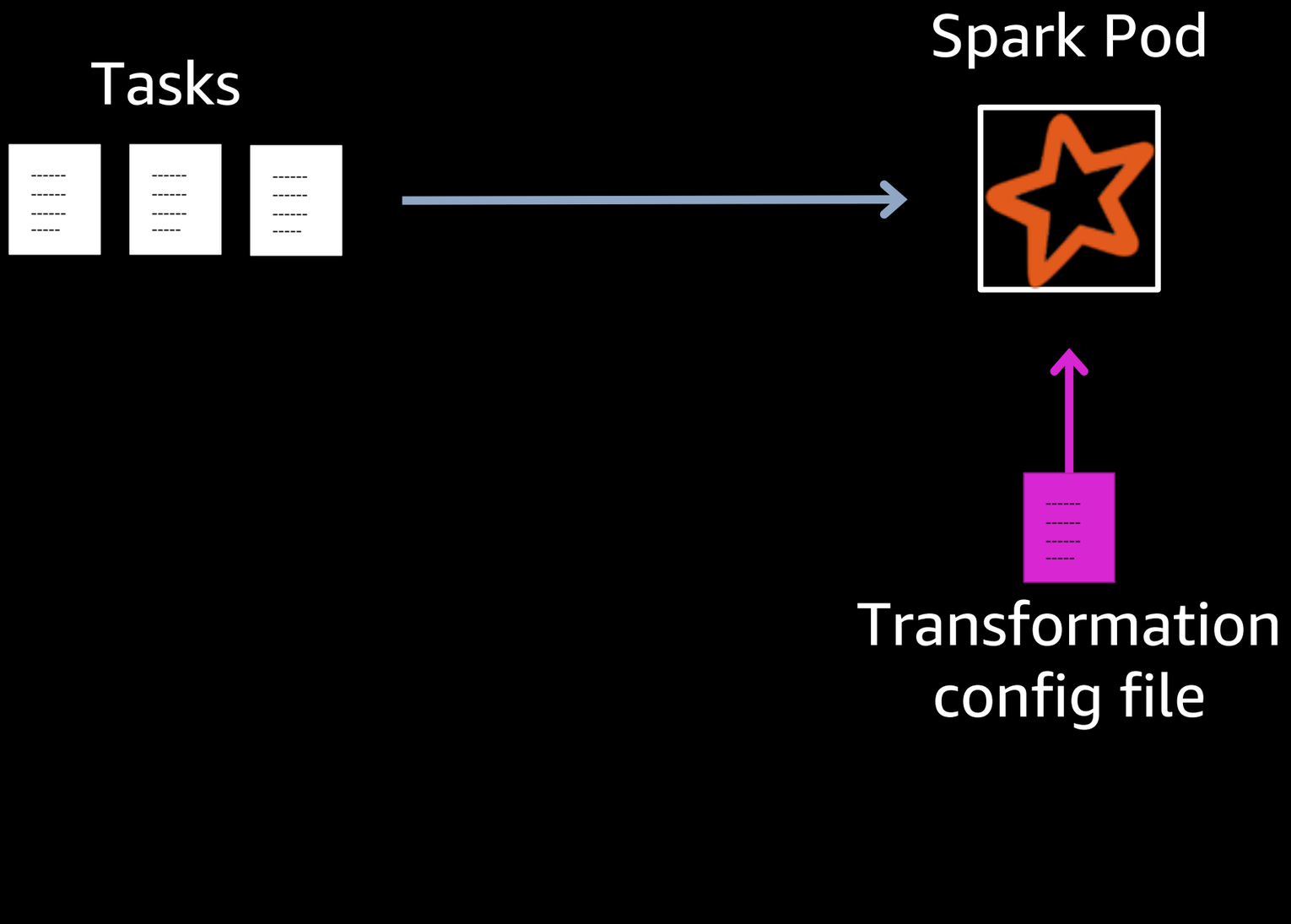
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Transform – Spark Pods / EKS

nielsen



Transformations are config files



EKS scaling

nielsen

Tasks

Spark pods

EC2



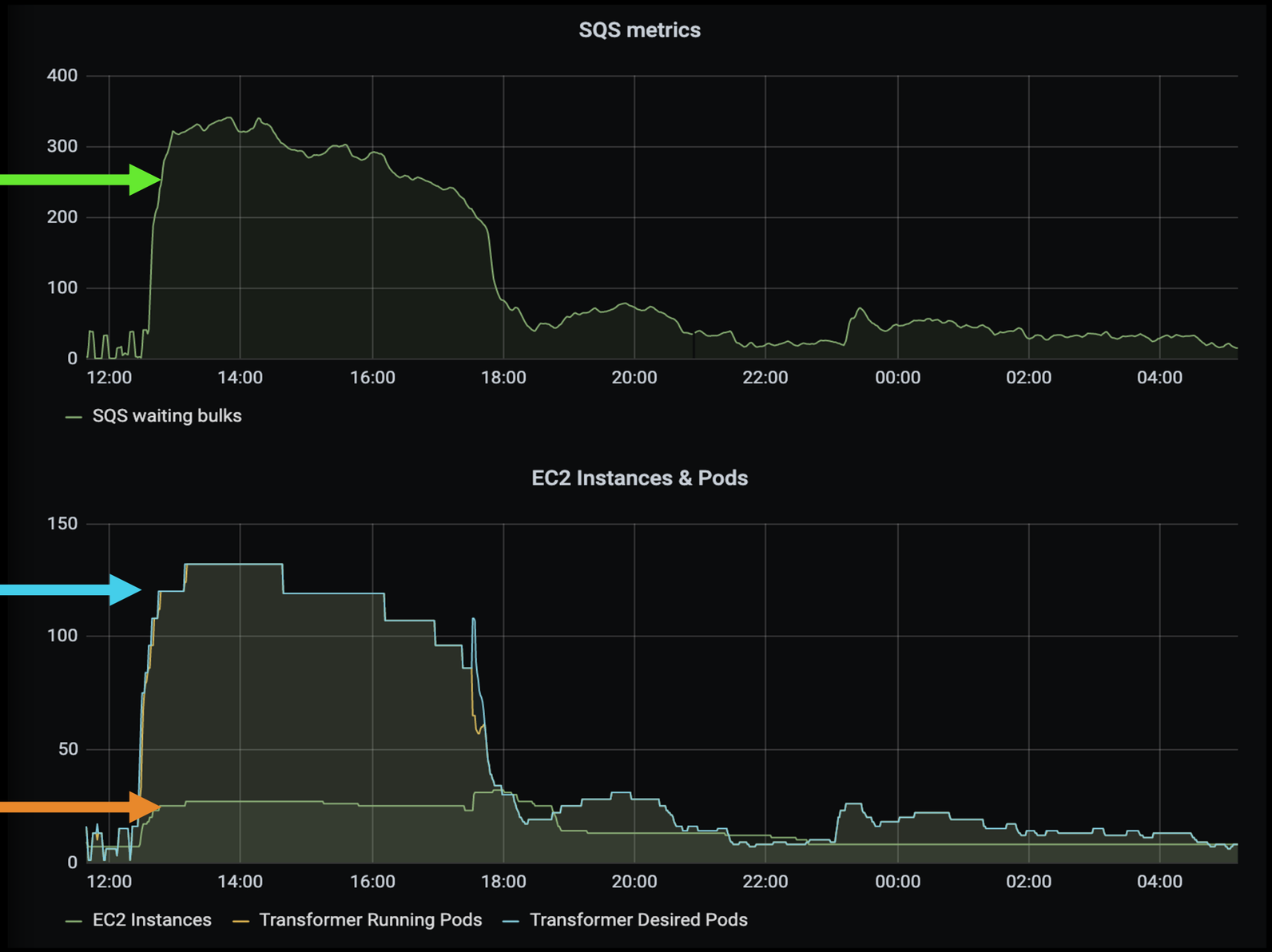
Handling bursts

nielsen

Tasks

Spark Pods

EC2



nielsen

Deliver



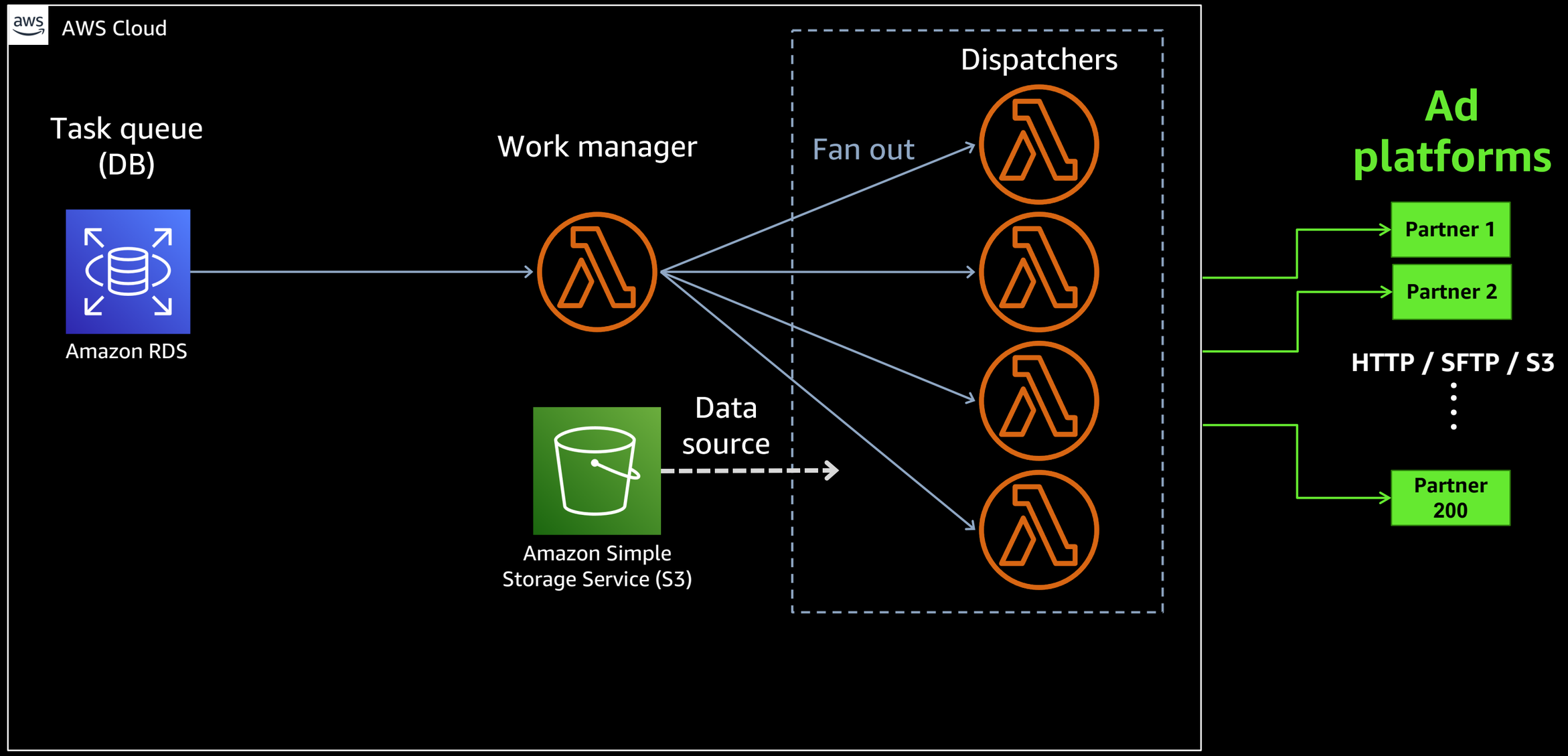
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

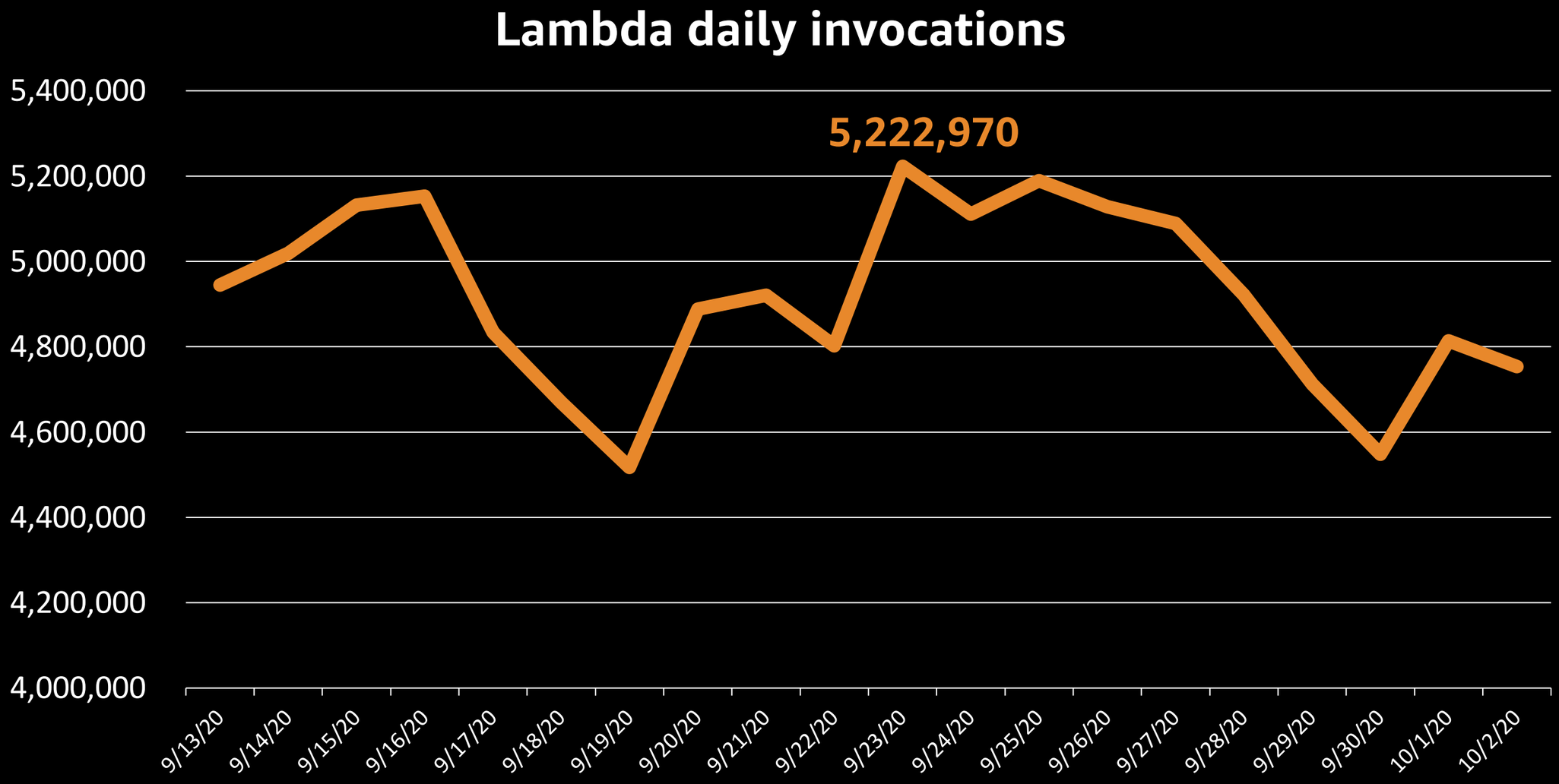
Deliver – Lambda

nielsen

SCALING IS AUTOMATIC



Lambda invocations



DataOut – WE DID IT !



250 billion events per day

Running on serverless!

Serverless pattern
Data processing pipeline

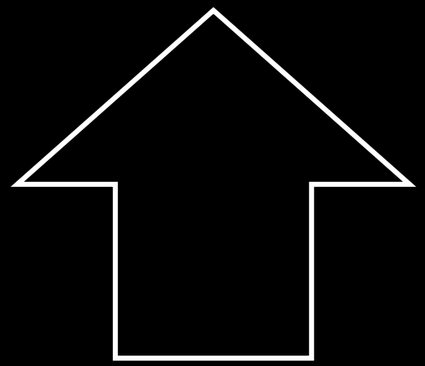


AWS re:Invent

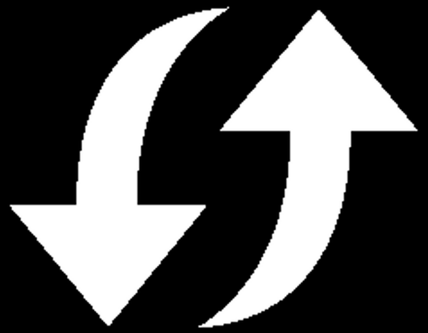
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern data

TYPICAL CHARACTERISTICS



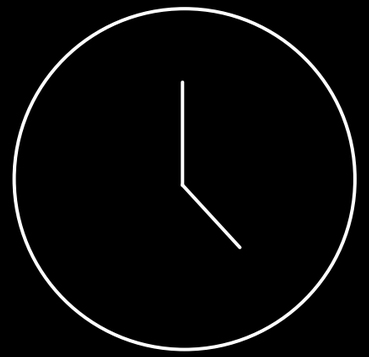
High volume



Continuous

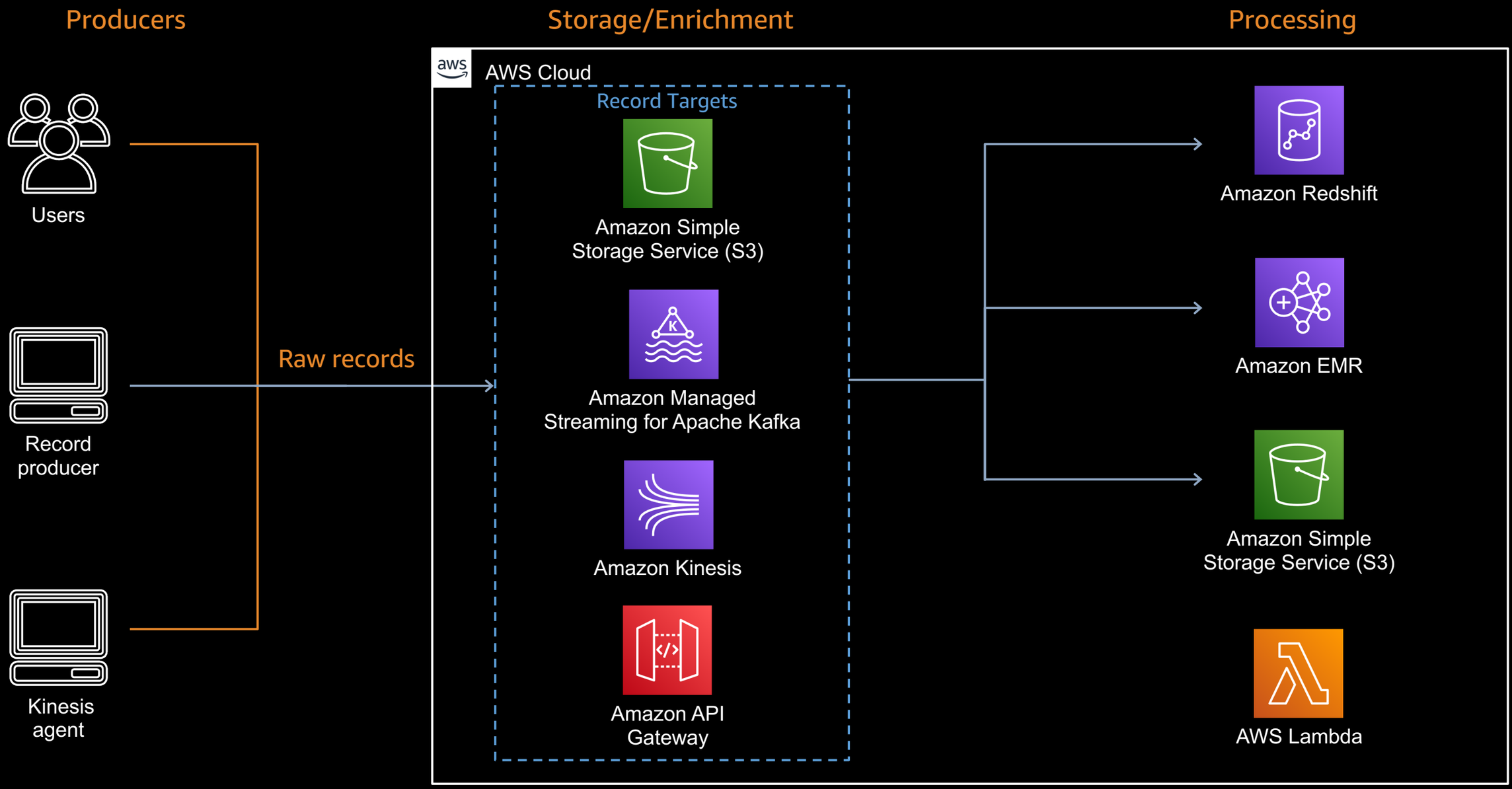


Spiky traffic

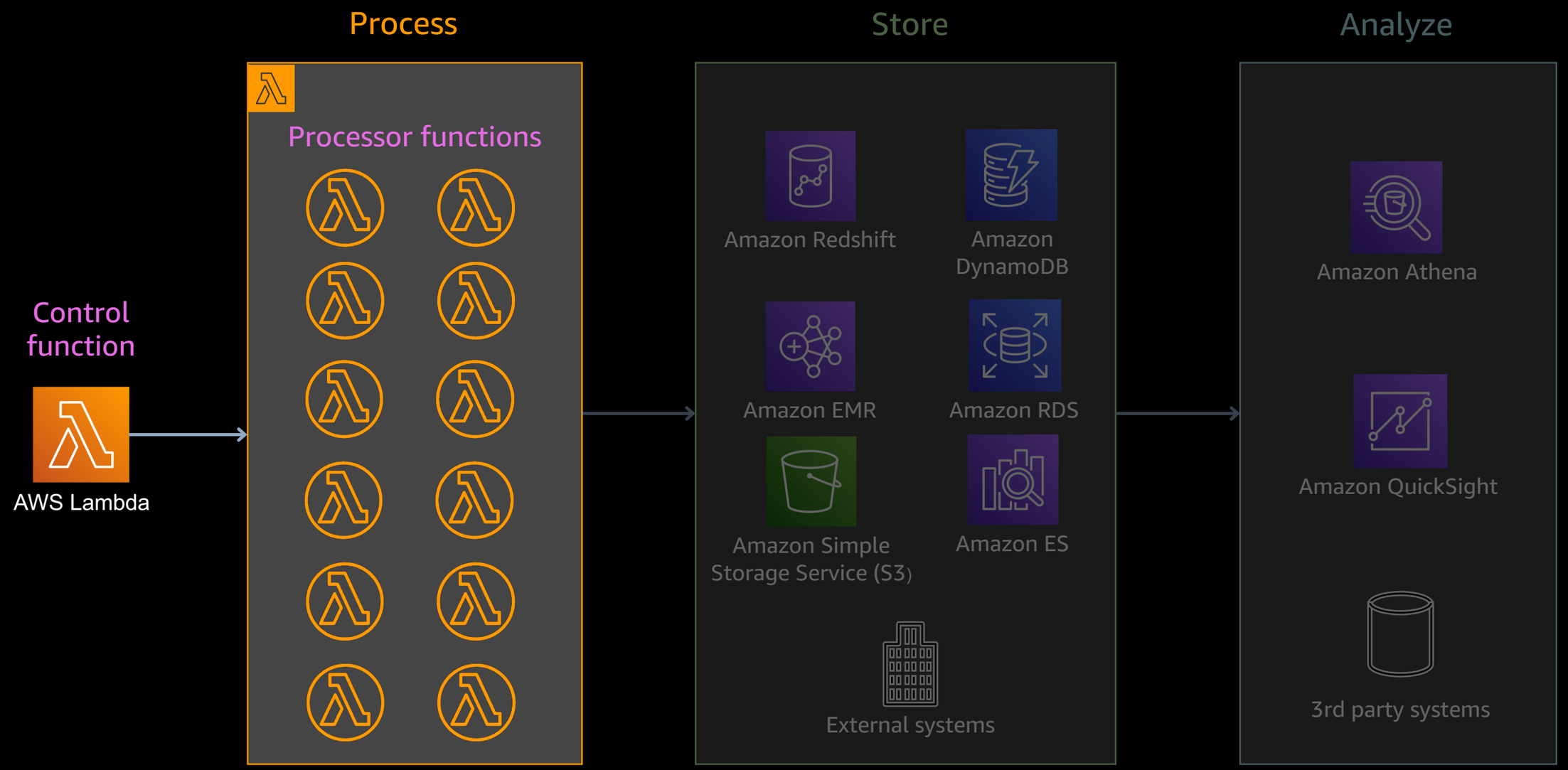


Low latency

Data ingestion



Data processing



nielsen

Avoiding pitfalls



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

This is a true story.

The events described in this talk took place at Nielsen Marketing Cloud in 2019–2020.



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

nielsen



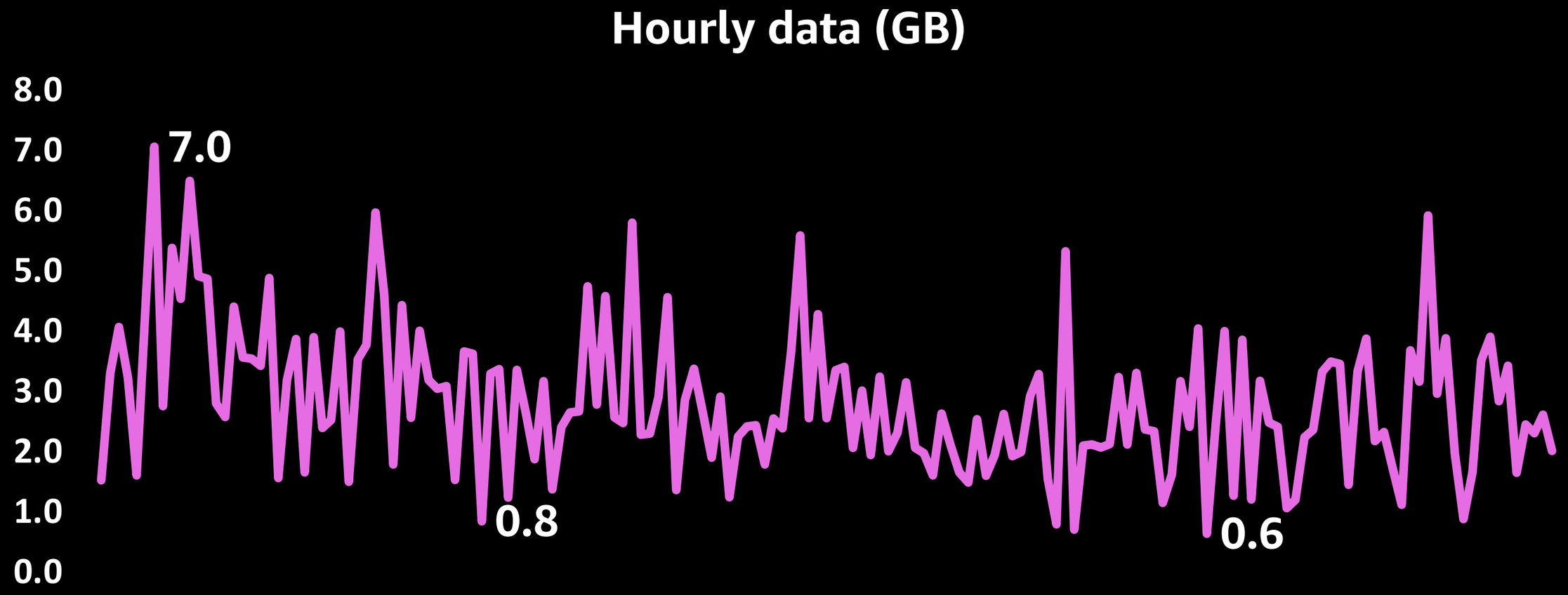
Preventing DDoS
Being good neighbors

(Rate limiting)

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Accounts have variations

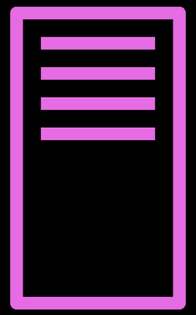


12x scaling each day

3rd party scaling might be hard

nielsen

- No Scaling



- Limited Scaling



Solution – Rate limiting

nielsen

Goals

- Flexible time periods
- Support multiple limit types

Strategy per account

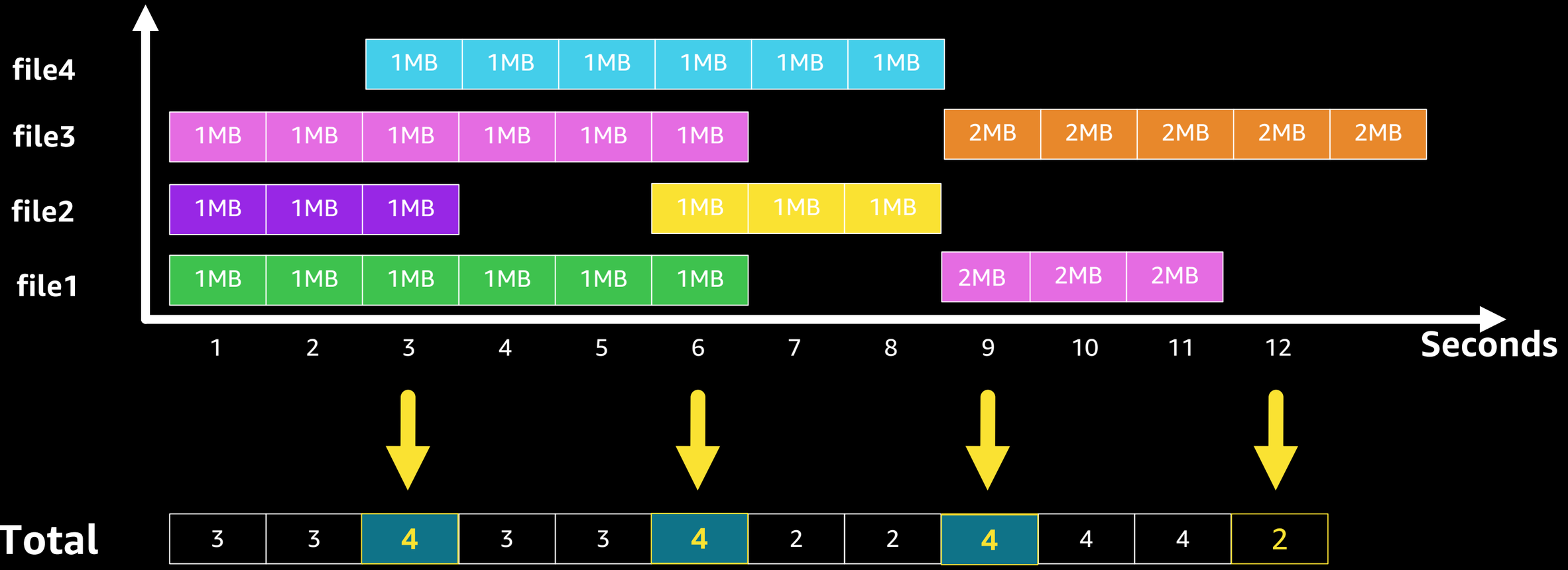
nielsen

- Set period + limitation
- Available strategies:

Strategy	Settings
Size	Period (sec) + MB
Tasks	Period (sec) + tasks
HTTP requests	Period (sec) + # of requests

Example limit by size for account

- Limit: 240MB/min = 4MB/sec
- We calculate every ~3 sec



Problem: How to determine rate/sec ?

nielsen

- We know the file size (say, 40MB)
- We don't know the time to send it

Solution

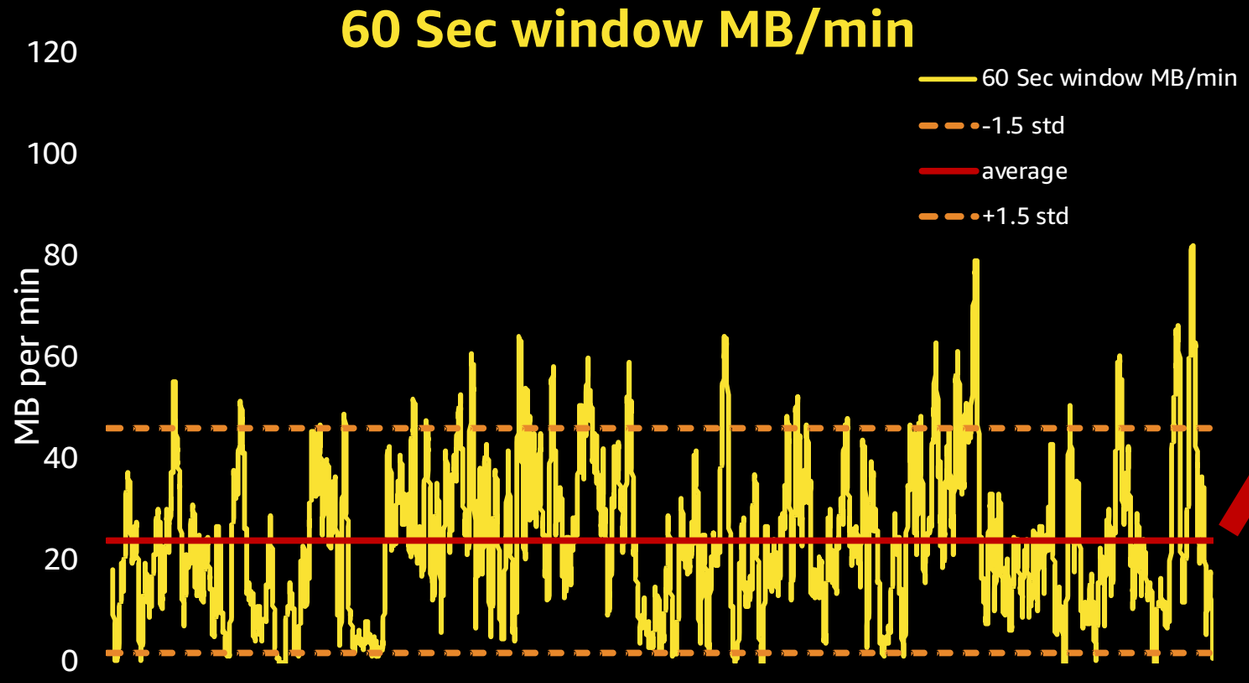
- Calculate hourly stats per account: **t = a*(file_size)+b**
- For each file, calculate expected rate/sec

- Example: 40MB, t = 2min → $^{1}/_{3}$ MB/sec

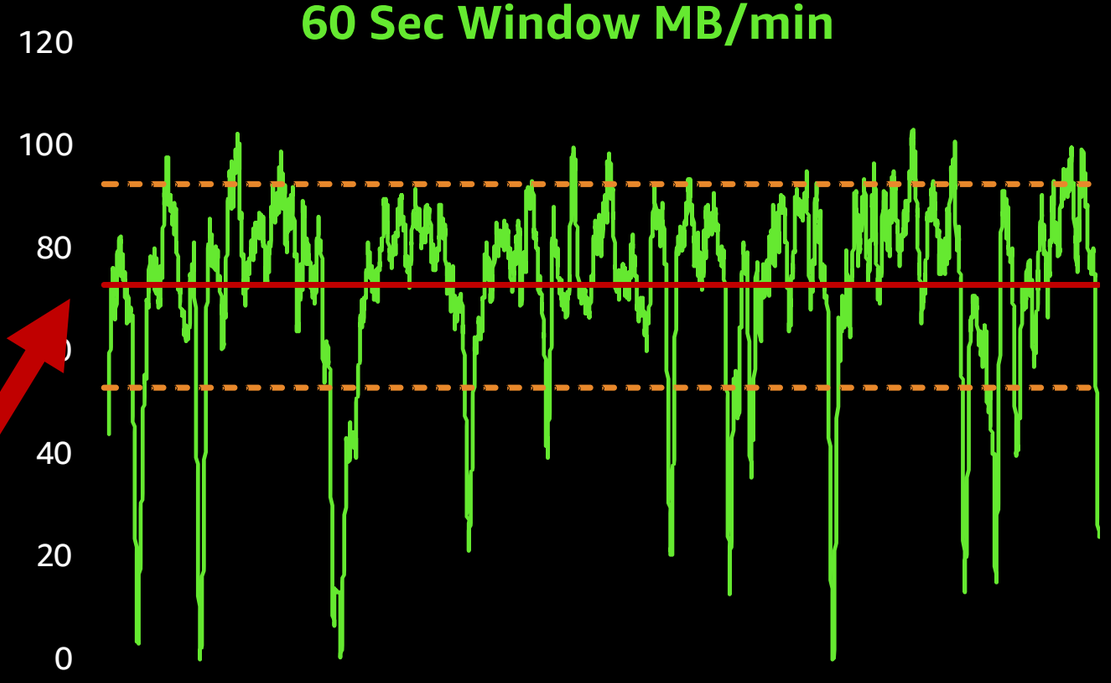
Side benefit – Send more data

nielsen

BEFORE: 28% utilization



AFTER: 69% utilization



2.5x more data

nielsen

Protecting internal systems
Being good roommates



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

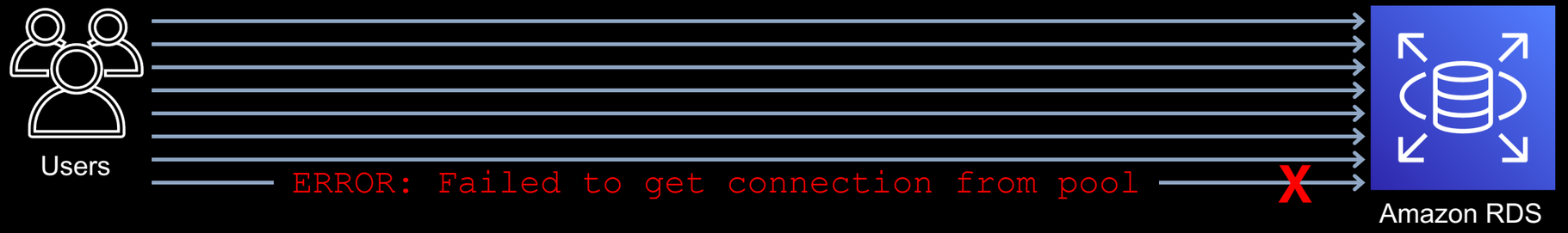
The problem (non-serverless systems)

nielsen

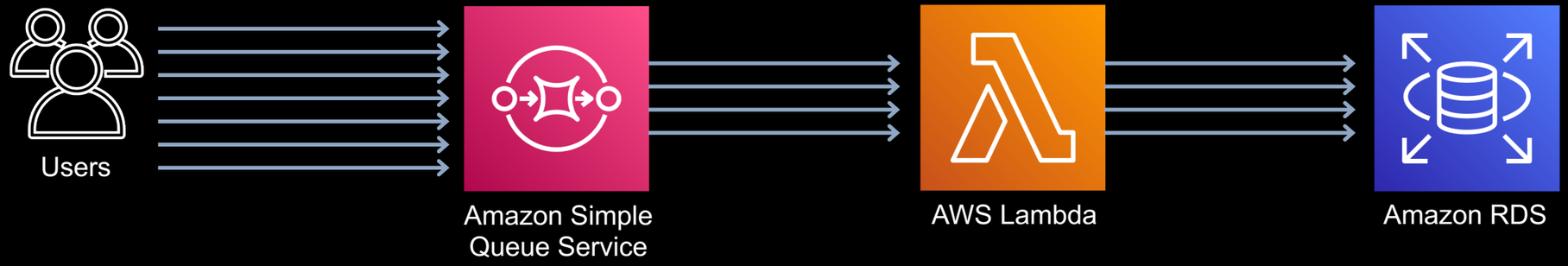
- Limited scaling
- May have a hard time responding to flooding
- Unexpected resources could limit the system

3x–12x scaling each day

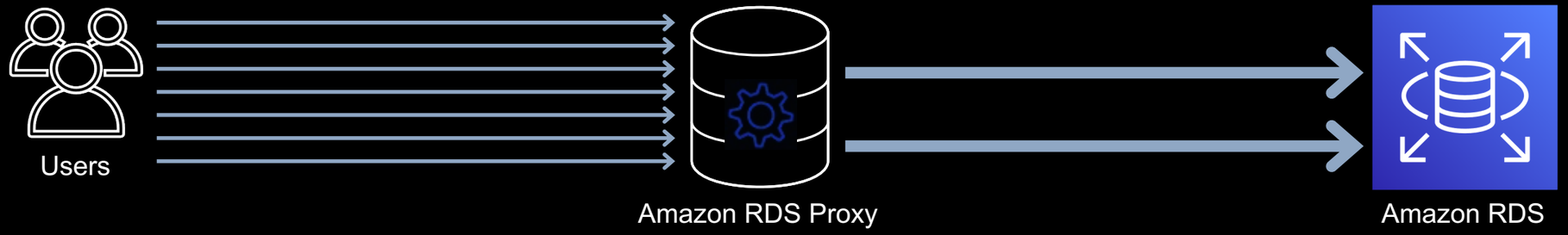
Handling RDS connections



Add buffers (queues)

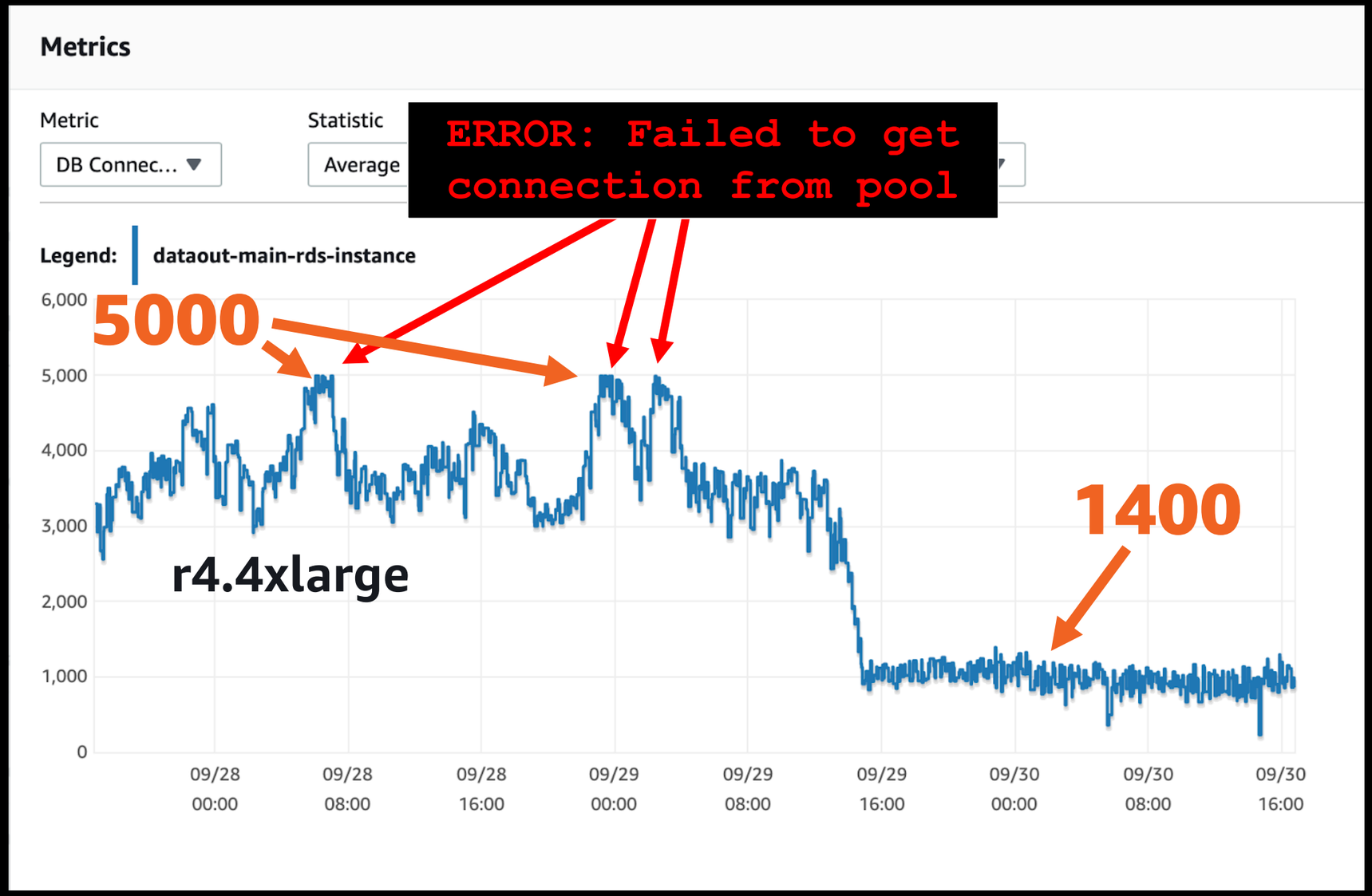


Use aggregators (**RDS Proxy**)



Amazon RDS Proxy

nielsen



nielsen

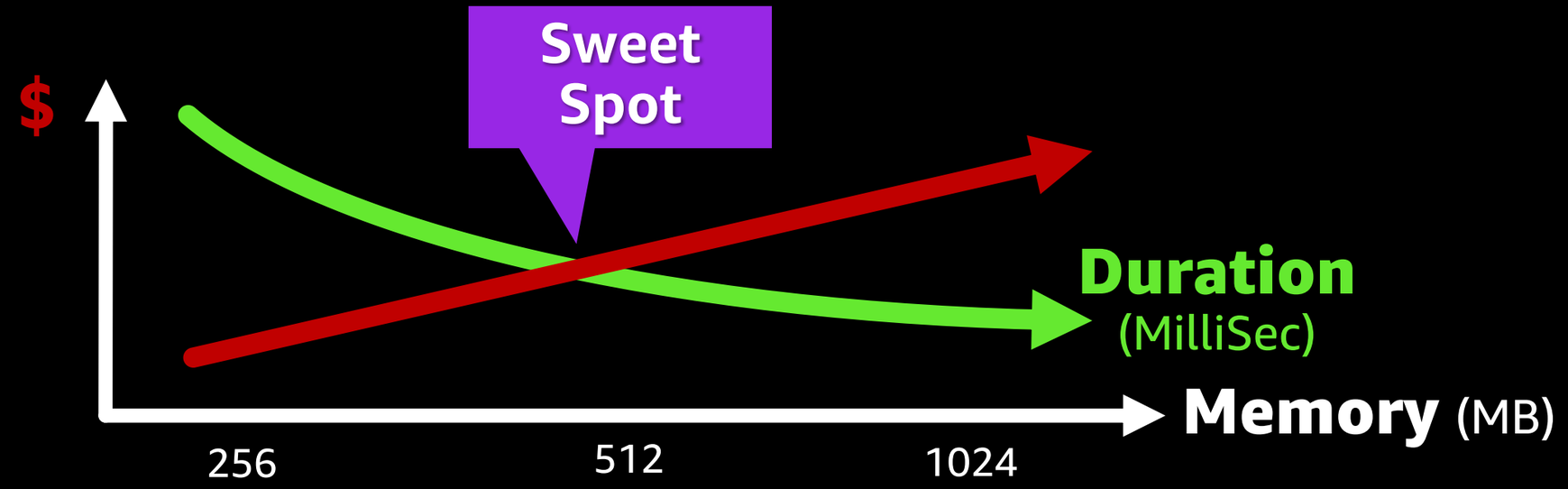
Driving down costs
+ avoiding mishaps



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How does Lambda pricing work ?



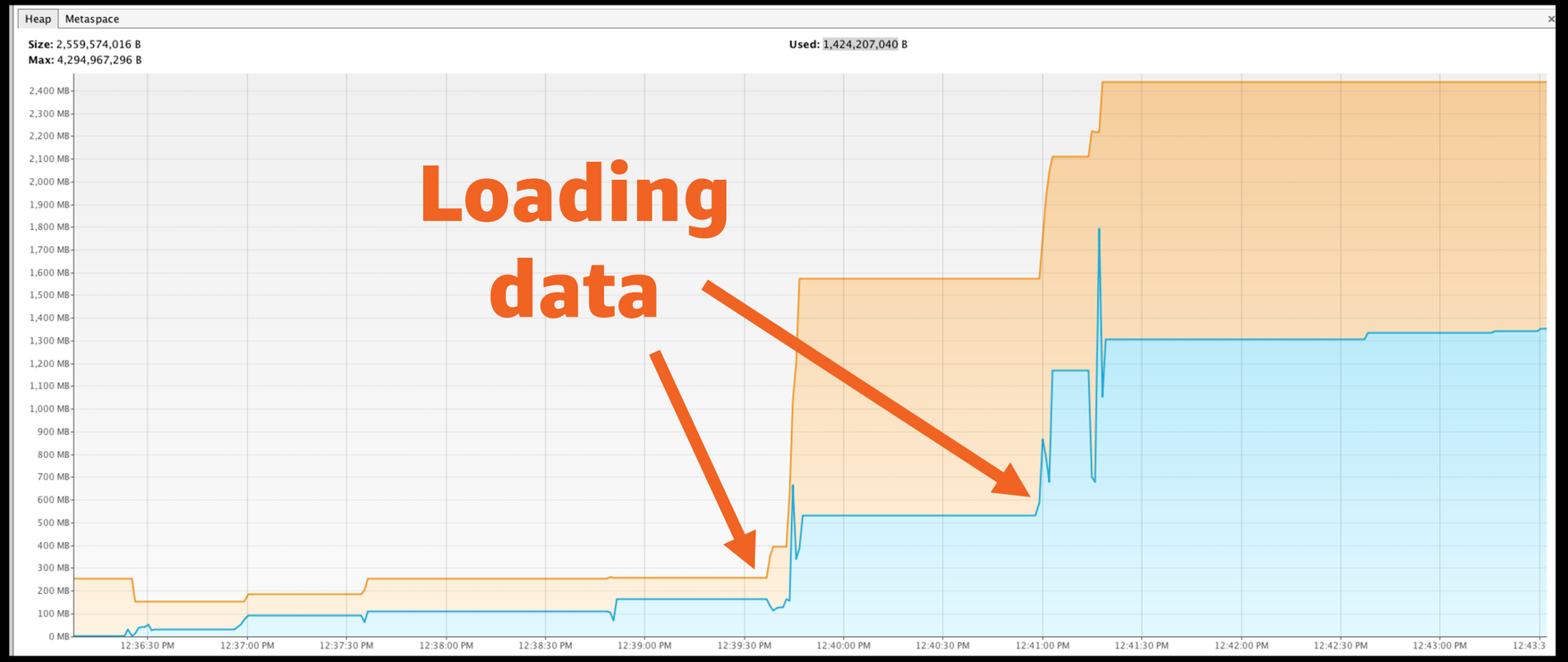
$$$$$ = memory * duration

But . . .

- Memory = code footprint + working memory
- Memory allocation affects CPU and bandwidth

Profiling

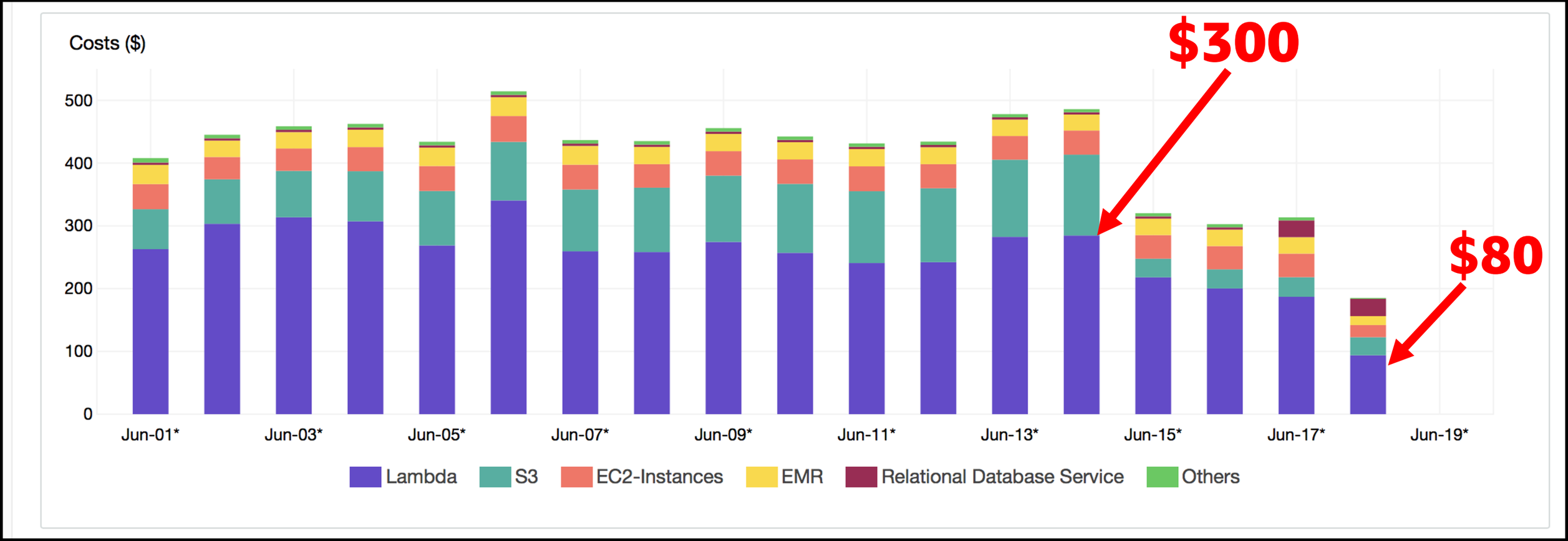
nielsen



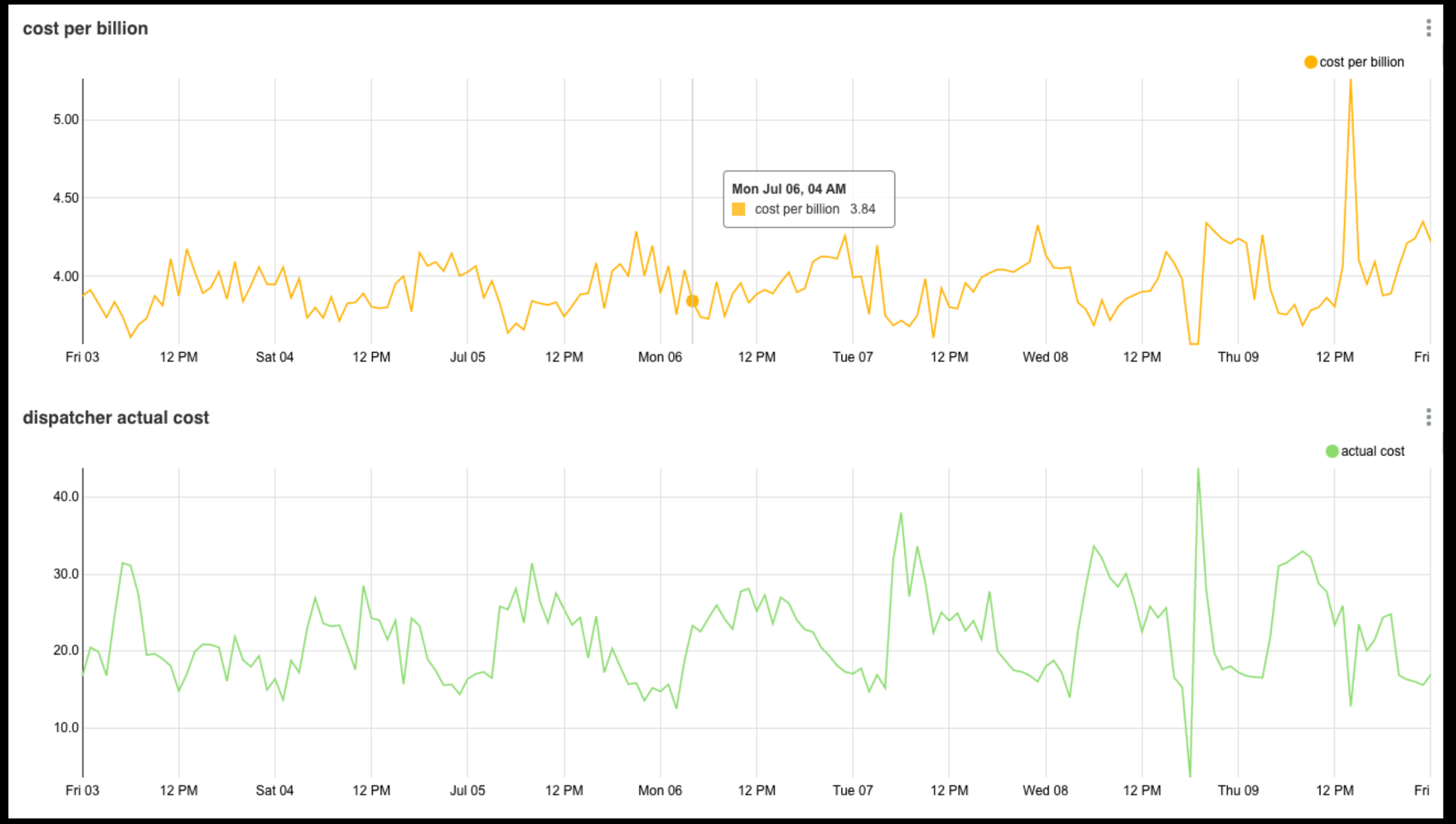
The memory hogs:

Data, HTTP connections

Total cost improvement



Monitoring costs



nielsen

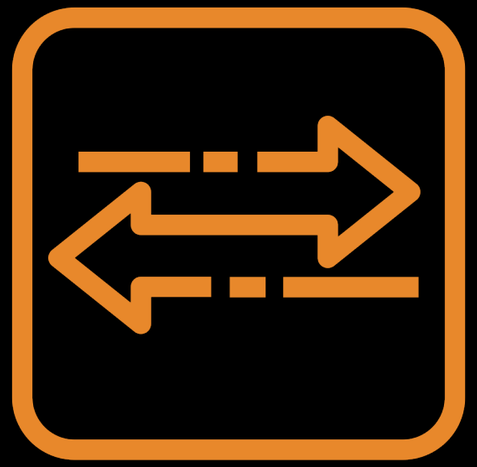
Wrap up



AWS re:Invent

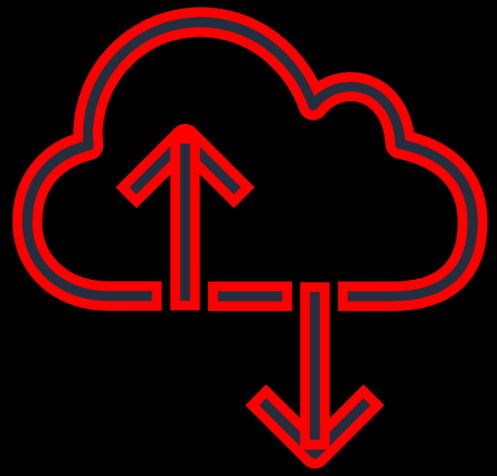
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Wrapping up



Massive

serverless data pipeline



Scales

according to traffic



Costs

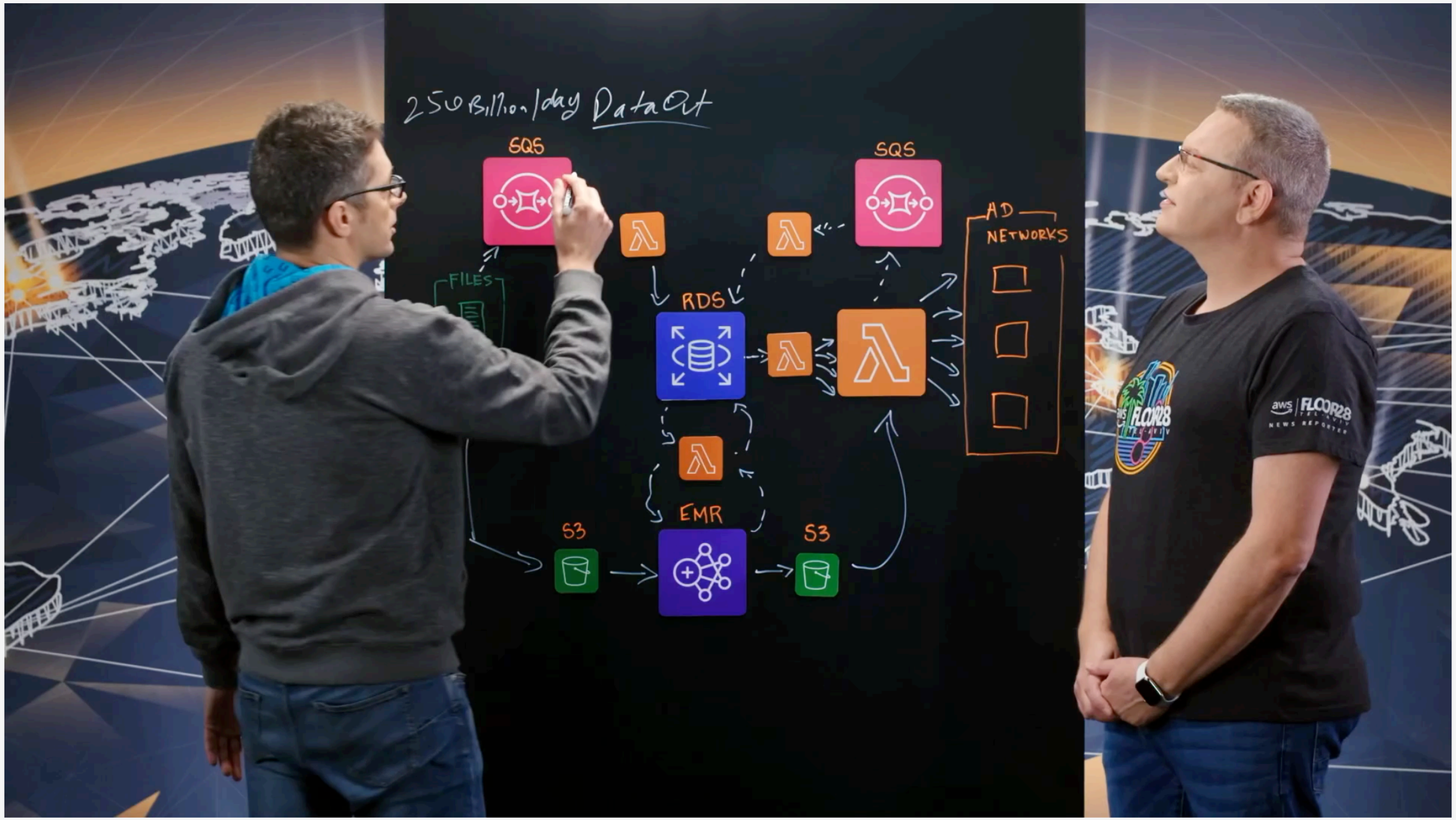
$4.25/Billion events

Takeaways - serverless is ready for prime time

- Massively and seamlessly scale on demand
- Rethink your architecture – the serverless way
- Serverless is the fastest way to build a modern application
- Watch the video – https://bit.ly/TMA-Serverless

(32) Nielsen: Processing 55TB

https://bit.ly/TMA-Serverless



#AWS

Nielsen: Processing 55TB of Data Per Day with AWS Lambda

2.3K 23 SHARE SAVE

91,607 views • Sep 23, 2020

Architecture resources

AWS Well-Architected Framework

The official best practices for Architecting in the AWS Cloud

https://aws.amazon.com/architecture/well-architected

AWS Well-Architected Labs

Hands-on labs to help you learn, measure, and build using architectural best practices

https://wellarchitectedlabs.com/

AWS Architecture Center

Official AWS repository for all Architecture resources

https://aws.amazon.com/architecture

AWS Solutions Library

Vetted reference implementations and Well-Architected patterns

https://aws.amazon.com/solutions/



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Opher Dubrovsky

Big Data Dev Lead
Nielsen
@jumpja

Boaz Ziniman

Principal Developer Advocate
AWS
@ziniman

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.