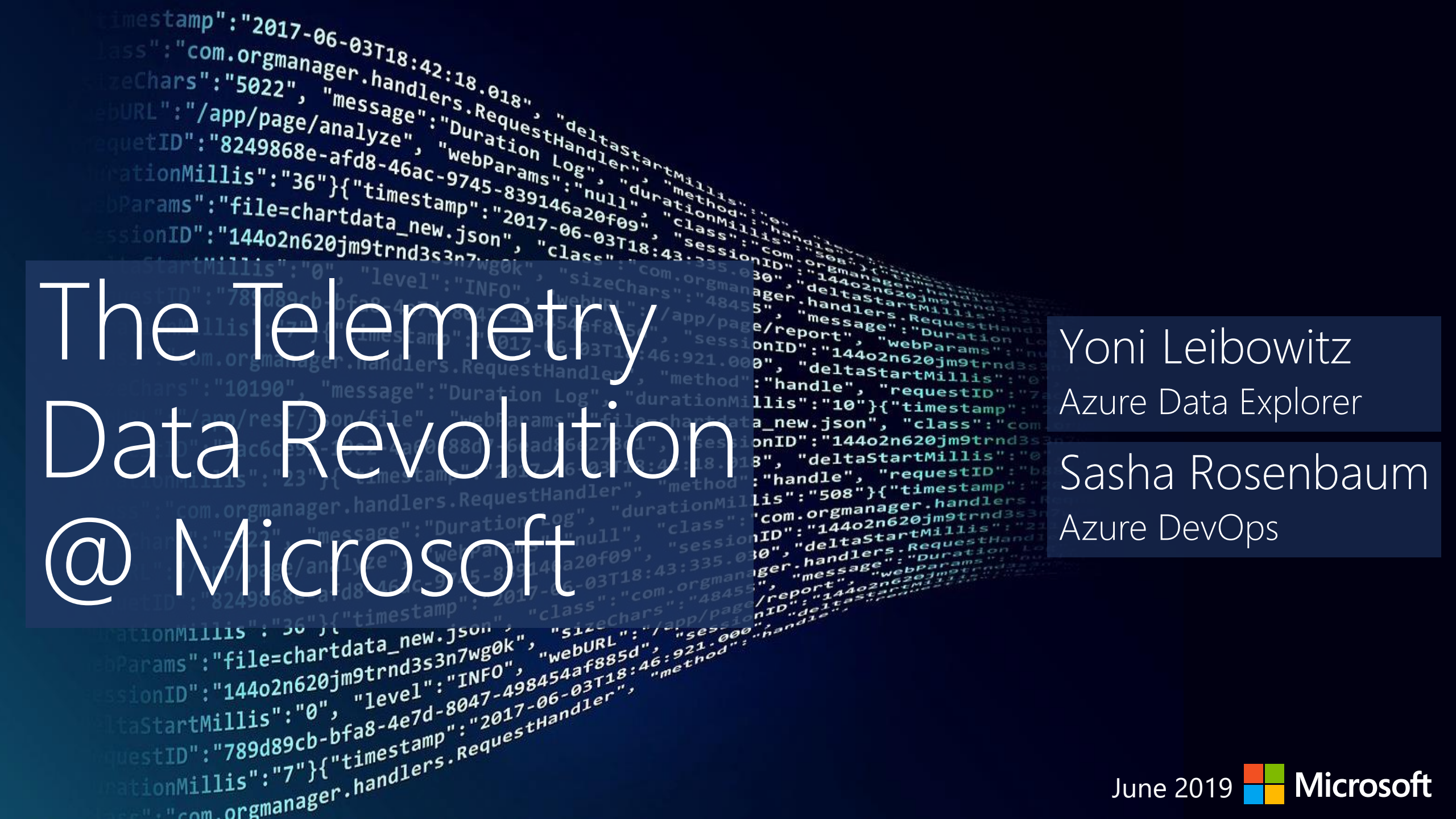


# The Telemetry Data Revolution @ Microsoft


Yoni Leibowitz
Azure Data Explorer

Sasha Rosenbaum
Azure DevOps


June 2019

Microsoft

hi |

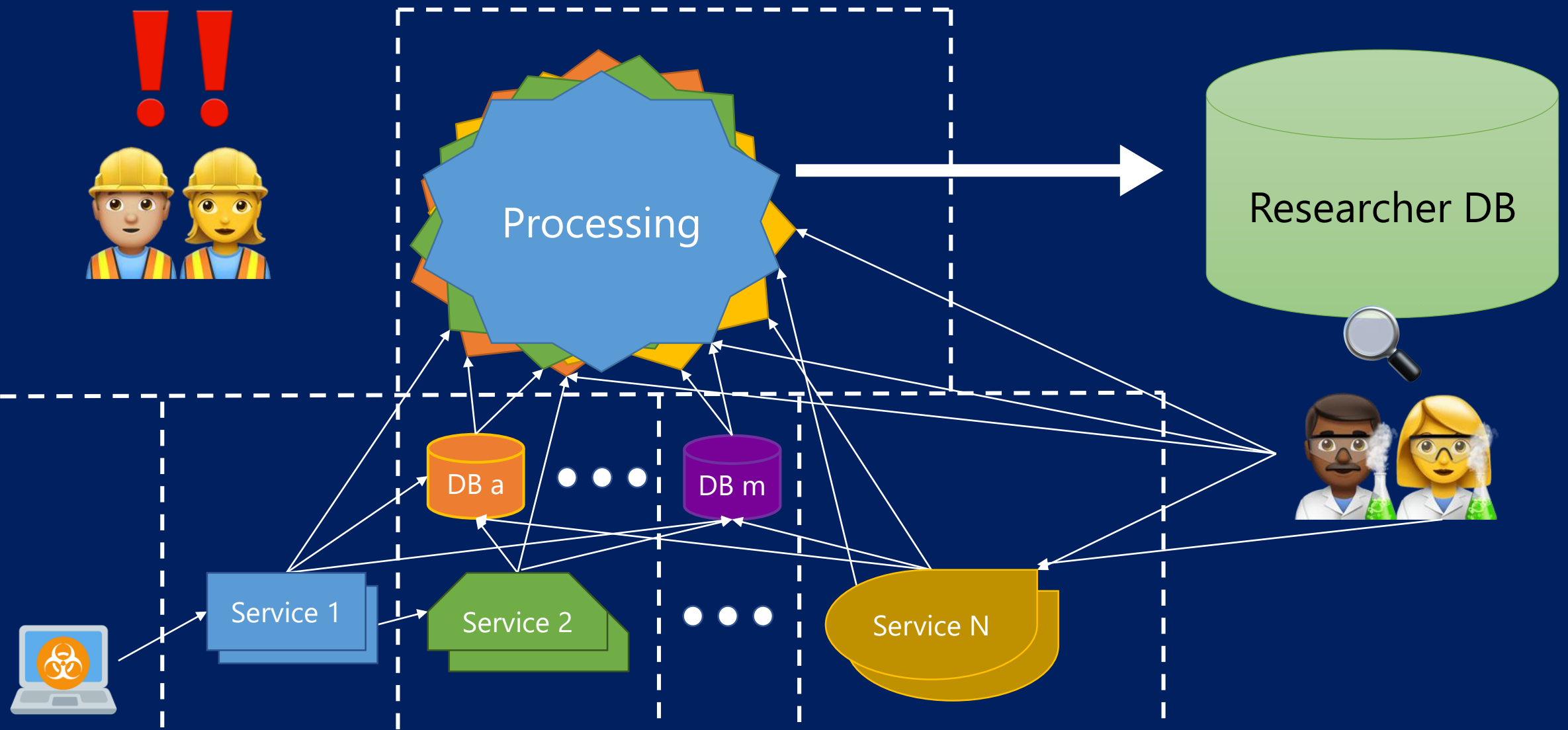
yonileibo

yonileibowitz

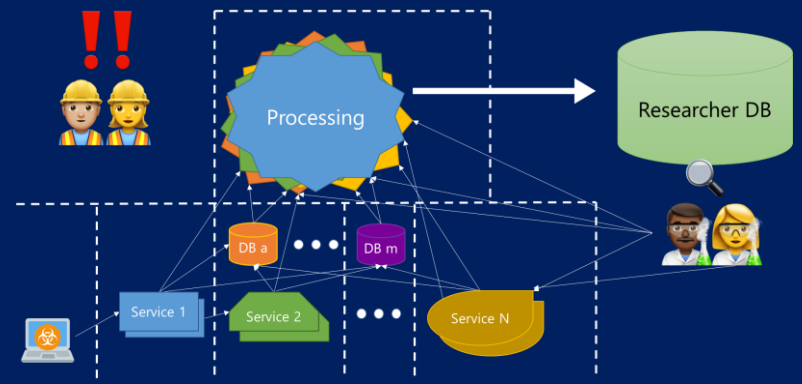
yonil@microsoft.com

https://aka.ms/kusto.blog

# back in 2014 …


Processing
Researcher DB
DB a
DB m
Service 1
Service 2
Service N

# telemetry | challenges




- growing **Variety, Volume, Velocity**
  - multiple origins – apps | services | devices | …
  - structured | semi-structured | free-text
  - MB → GB → TB → PB → EB
  - continuously generated

- driving *smart* **automation** for
  detection | investigation | remediation | validation

- **democratizing** the data
  - enabling everyone to *quickly* reach *meaningful* insights
  - while meeting security & compliance requirements

# telemetry | solutions ?

> open source

> 3rd party

> in-house

> mix-and-match

meanwhile,
one quick elevator ride away ... |

# **A**zure **D**ata e**X**plorer | "Kusto"

High **V**olume
High **V**elocity
High **V**ariety
structured
semi-structured
free-text

powerful & simple analytics query language

fully-managed

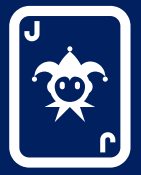
a **big data** **analytics** **cloud platform**

**optimized** **for interactive queries**

extreme performance over large data sets

rapid iterations to explore the data, low latency

# let's start with a demo | 

"what's the trick ?" | 

# a database for "log" data

- **optimized for append-only data streams ("logs")**
  - no in-place updates or single-record delete
  - deletion is done in bulk, e.g. when data ages out
    - exception: re-writing data when purging (GDPR)

- **entities**
  - cluster – database(s) – table(s) – column(s)
  - a table is a collection of data shards ("extents")

# storage

- ingested data is stored in a **proprietary format**
  - unique ingestion and query performance rely on this
  - for best performance, data should be ingested prior to query
    - [preview] query data directly from the data lake, without ingestion

- **tiered storage model**
  - compressed data is stored in durable Azure blob storage
  - ingested data is cached
    - in SSD (local VM or managed disk)
    - in RAM

# a unique combination of well-known technologies

> **sharding** allows scaling-out and distributing queries over N nodes

> separation to **columns** allows ignoring entire columns which are not relevant for a query

> **compression** allows loading data from storage quickly

> **indexing** allows ruling-out data quickly

> **low-granularity indexing** means fast ingestion and low overhead
  > data scanning may still be required, but is efficient due to **column store**

fast ingestion, fast queries | 

# Kusto Query Language | KQL

```
DimClusters
| project Cluster, Account, Division = tolower(Division)
| join (
    KustoLogs
    | where Timestamp > ago(1d)
    | where Level == "Error"
    | summarize Count = count() by Source
) on $left.Cluster == $right.Source
| where isnotempty(Division)
| top-nested 3 of Division by sum(Count) desc,
  top-nested 5 of Account by sum(Count) desc
```

- functional
- follows a data-flow model
- allows intermixing of operators
  - to be done at any order
  - common relational operators:
    - filters | joins | aggregation | sorting
  - special data operators: multi-value expand | parsing | top-nested | …
  - dedicated support for time-series, user analytics, ML (python | R)
  - make this all fluent & super easy: intellisense, syntax highlighting
- **SQL** is supported too

# JAKH | just another Kusto hacker | 

# (R)evolution

Ideation

Foundation

Resource commitment
Incubation
"friends & family" exposure

1st major internal customer

Growing the Internal customer base

Establishment
Chosen as the Telemetry solution for multiple divisions

Data platform for Microsoft services

Heading for GA

013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019

IDEATION & CONCEPT

VALIDATION

SCALE

Telemetry Analytics - Internal

1st party validation
Developing modern analytics

SCALE

Analytics Data platform for products

PUBLIC PREVIEW

Interactive analytics
Big data platform

GA

# #poweredByADX

Microsoft services using ADX as their data platform + exposing KQL

## Azure Monitor

Log Analytics

Application Insights

## Security Products

Windows Defender ATP

Azure Security Center

Azure Sentinel

## IoT

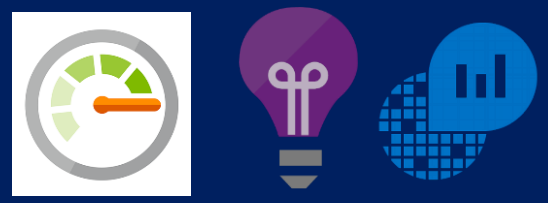
Time Series Insights

Azure IoT Central

Microsoft Connected Vehicle Platform

## Gaming

Azure PlayFab



# Azure DevOps  ADX | 

Sasha Rosenbaum

Azure DevOps PM

@DivineOps

Sasha Rosenbaum

Azure DevOps PM

@DivineOps

# Introducing Azure DevOps

## Azure Boards

Deliver value to your users faster using proven agile tools to plan, track, and discuss work across your teams.

## Azure Pipelines

Build, test, and deploy with CI/CD that works with any language, platform, and cloud. Connect to GitHub or any other Git provider and deploy continuously.

## Azure Repos

Get unlimited, cloud-hosted private Git repos and collaborate to build better code with pull requests and advanced file management.

## Azure Test Plans

Test and ship with confidence using manual and exploratory testing tools.

## Azure Artifacts

Create, host, and share packages with your team, and add artifacts to your CI/CD pipelines with a single click.

**https://azure.com/devops**

# DevOps at Microsoft

Azure DevOps is the toolchain of choice for Microsoft engineering with over 100,000 internal users

https://aka.ms/DevOpsAtMicrosoft

**442k**
Pull Requests per month

**155**
Petabytes of build artifacts managed

**28k**
Work items created per day

**2.4m**
Private Git commits per month

**3.5k**
Open Source repos

**12k**
Employees contributing to open source

**82,000**
Deployments per day

Data: Internal Microsoft engineering system activity, March 2019

# Oh no, we have to learn another query language!..

Searching large amounts of unstructured data is easy!!!

# Usage: then and now

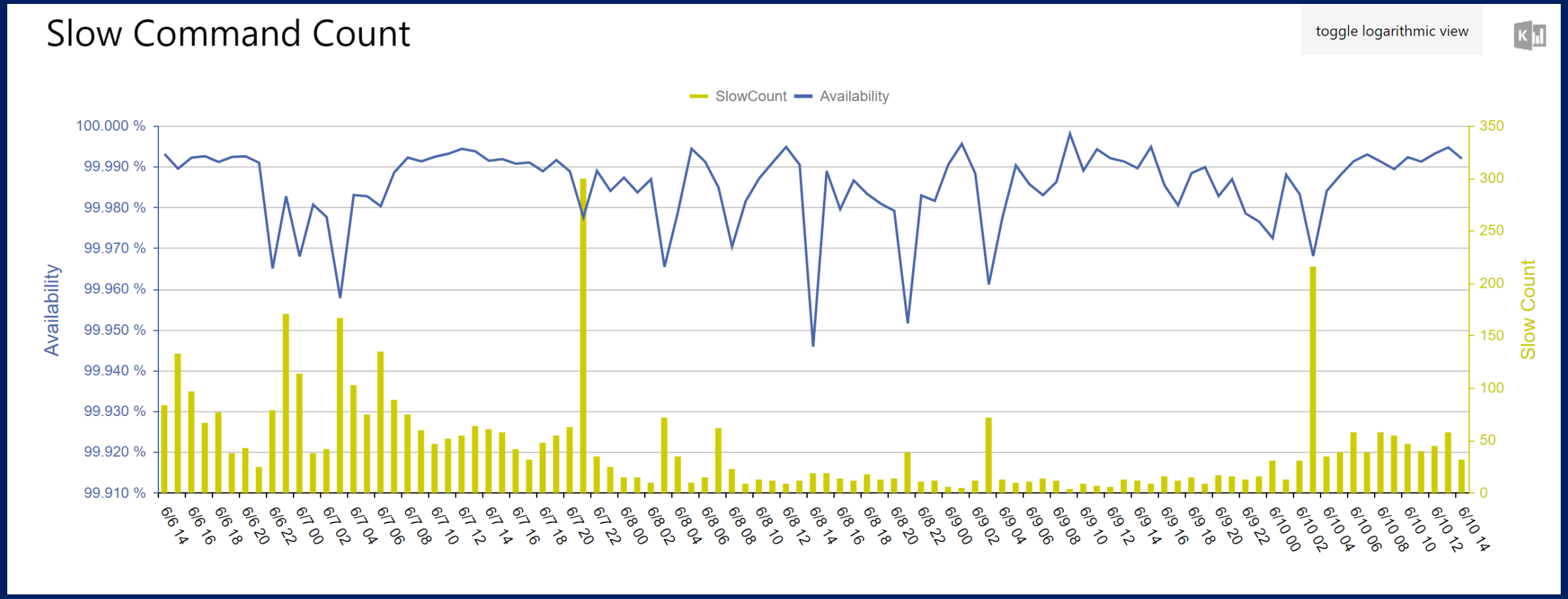
- 3 days of recent data
- 1 ADX cluster
- Diagnostics only

- 28 days of recent data
- Additional ADX cluster for historical data
- 3 different use cases

# Use Cases

1. Diagnostics
2. Analytics
3. Alerts

# Analytics Dashboards



## Alerting Tool

# Queries

Path

/Personal Storage/v-angold/PeakHoursStatsFor28daysV2.csl

```
55  let setPercentIssuesThreshhold = 2.0;              // Specify the percentage floor filter value of the sample Issues Percent.  Se
56  let setFailCountThreshold = 1;                     // Filter out commands with low lookback fail counts during the lookback count
57  let setSlowCountThreshold = 1;                     // Filter out commands with low lookback slow counts during the lookback count
58  //
59  let setTimeOffSet = 1h;
60  //
61  // Method for finding the command counts during a time span.
62  // Parameters:  daysAgo                    The number of days from today.  0d = today,  1d = yesterday, 7d = 1 week ago.  Note th
63  //              setActivityStatusValue     Specified the command counts that pass (0), fail (1) or are slow (2).
64  let CommandTelemetry = (daysAgo:timespan, setActivityStatusValue:int) {
65      ActivityLog
66      | where Feature != "Unknown" // Rid empty GUIDs
67      | where PreciseTimeStamp between ((currentDateTime - daysAgo - LookbackTimeSpan) .. (currentDateTime - daysAgo))
68      | where ActivityStatus == setActivityStatusValue
69      | where (Service == setService or setService  == "All")
70          and (Feature == setFeature or setFeature == "All")
71          and (ScaleUnit == setScaleUnit or setScaleUnit == "All")
72          and (IsExceptionExpected == toint(expectedException) or expectedException == "All")
73      | summarize count() by TimeFrame = bin(PreciseTimeStamp, binSize), IsExceptionExpected,  Command, Application, Feature, Servic
74      | summarize CommandCounts = count() by Command, Application, Feature, Service
75  };
```

Cancel  Save

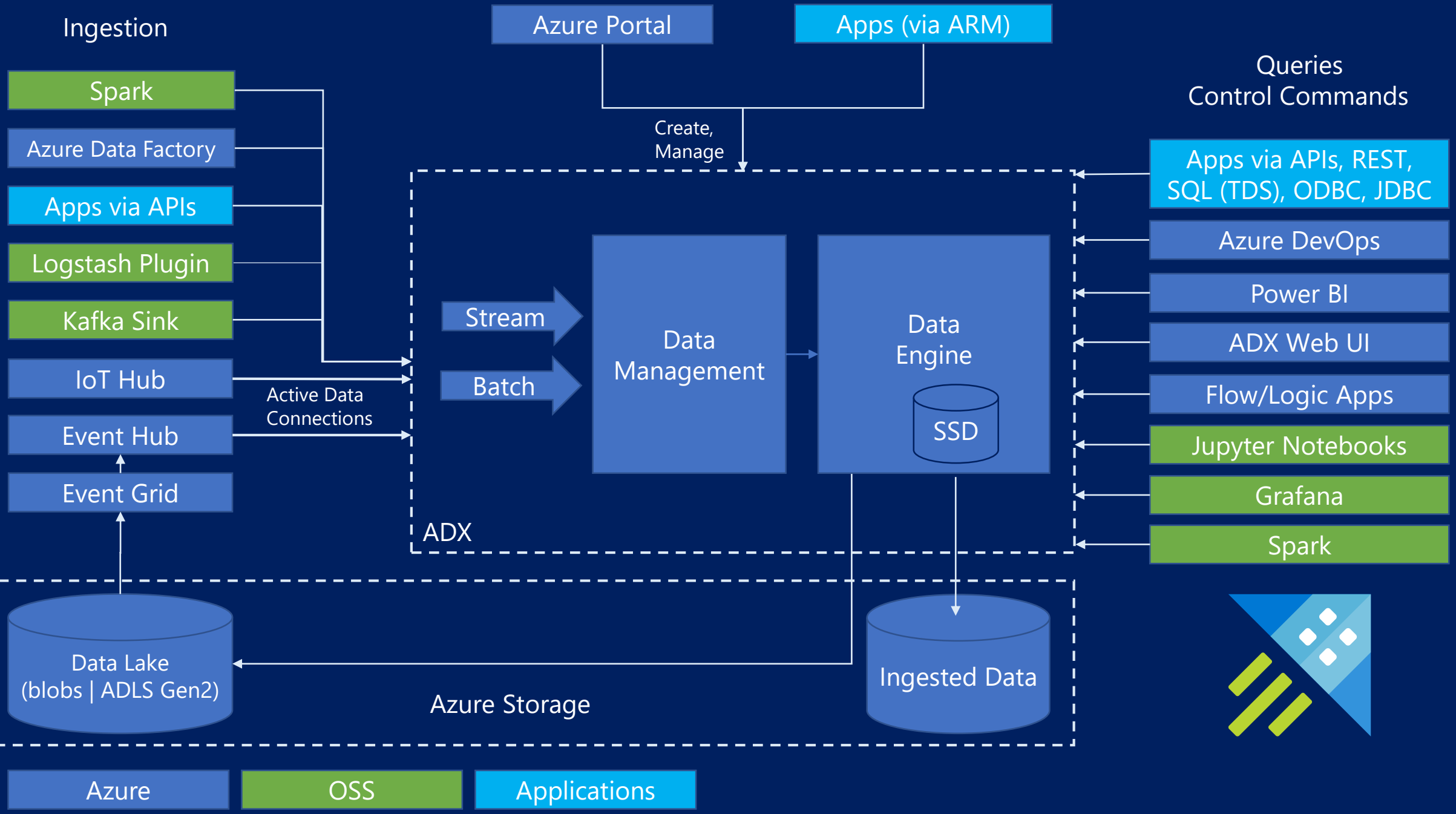
# scale

**675K**
**CORES**

**0.55 EXABYTES**
**TOTAL DATA SIZE**

**13 PETABYTES**
**PER DAY**

**> 30K**
**MONTHLY UNIQUE USERS @ MICROSOFT**


Ingestion
Azure Portal
Apps (via ARM)
Queries
Control Commands
Spark
Azure Data Factory
Apps via APIs
Logstash Plugin
Kafka Sink
IoT Hub
Event Hub
Event Grid
Create,
Manage
Active Data
Connections
Stream
Batch
Data
Management
Data
Engine
SSD
ADX
Apps via APIs, REST,
SQL (TDS), ODBC, JDBC
Azure DevOps
Power BI
ADX Web UI
Flow/Logic Apps
Jupyter Notebooks
Grafana
Spark
Data Lake
(blobs | ADLS Gen2)
Azure Storage
Ingested Data
Azure
OSS
Applications

# what are people saying ? |

"

Kusto is AMAZING.
It's one of those rare tools that once you start using, you can't do without.

*It went viral in our team in no time.*

*For those using it for deep investigations, it's as addictive as Candy Crush!*


Adi | Engineering Manager

"

The capabilities are phenomenal, users love it, and anyone who sees it is impressed.
This was built by engineers for engineers, and it shows.

*If I were to mention taking it away, I would face a massive revolt from the ranks*

Ashwin | Engineering Director

"

When I show Azure users Kusto queries *their jaws drop*.

*People love it* ♡ *it's awesome*

Scott | Executive VP

# The data revolution outside of Microsoft has already begun

*This seems impossible. What's the trick? Is this software or magic?*

For us this was *game-changing*.

We were in a world where queries were taking minutes to hours.

And here - results are coming back to us in 1, 2 or 3 seconds

DocuSign®

"

Kusto is one of the supreme offerings in Azure to date.

Often, while showing results to customers, questions come up – Many times I can just write a quick query and answer it.

*It's addictive* – It's sometimes hard to communicate how exactly, it just is"

BASF

"

We love every inch of Kusto.

Everyone is loving speed and the ability of Kusto to chew our sh** schema.

Every possible query flies *faster than an SR-71 Blackbird*

DODO PIZZA

join the data revolution | 

# Thank you | 👏