The Future of Data Pipelines

John Hammink

Developer Advocate, Evangelist @ Aiven

john@aiven.io

All Things Open 2019, Raleigh NC

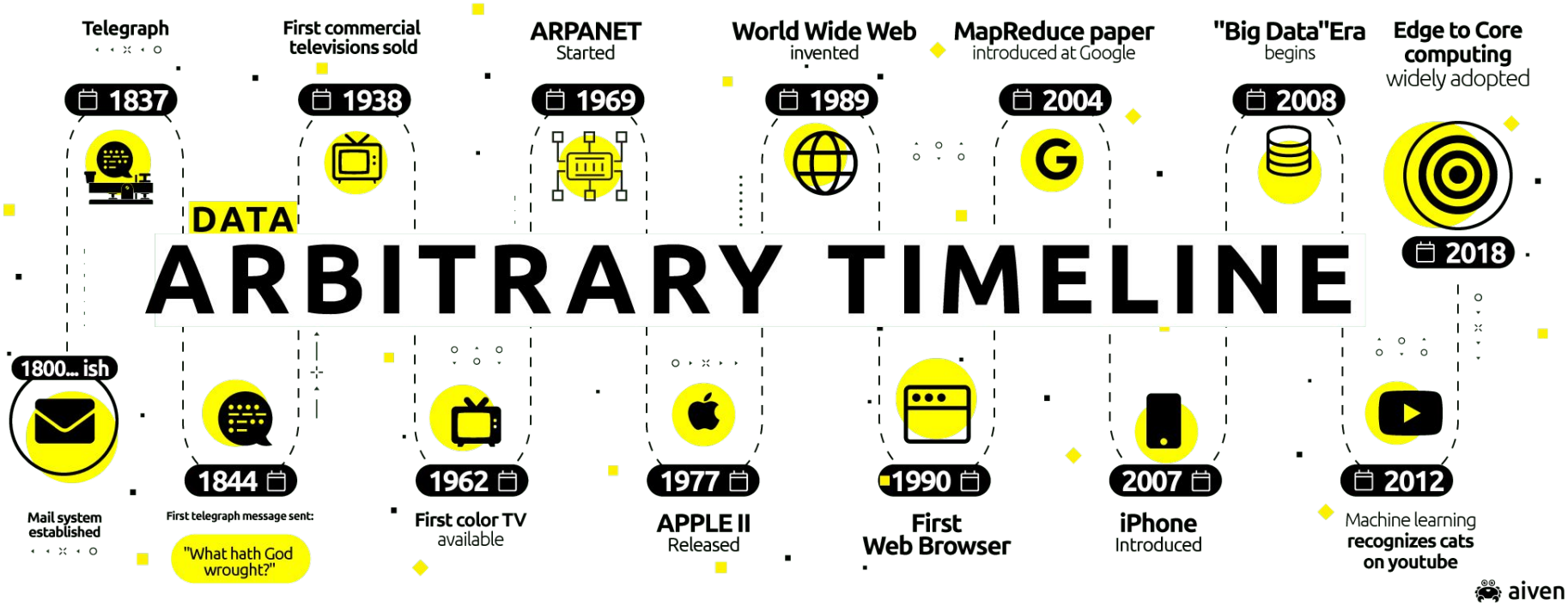
The Future of Data Pipelines

- Let's start with where the **future looked** from **about a year ago**
- Let's assume that **pub-sub** and particularly technologies like **Kafka** have created a **pivot point**
- Let's follow up with a **discussion of trends** we're all seeing

The Future of Data Pipelines

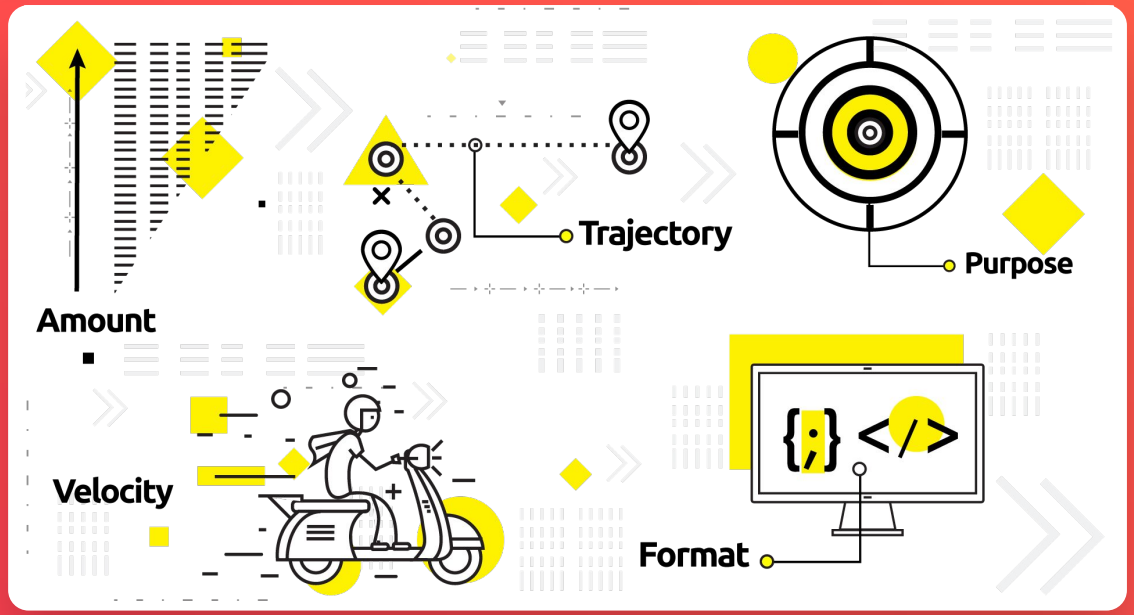
- Background
- Learnings
- Some numbers on data
- Functionality
- Design
- Compliance
- Reliability/Performance/Usability/Other
- Conclusions
- Q & A

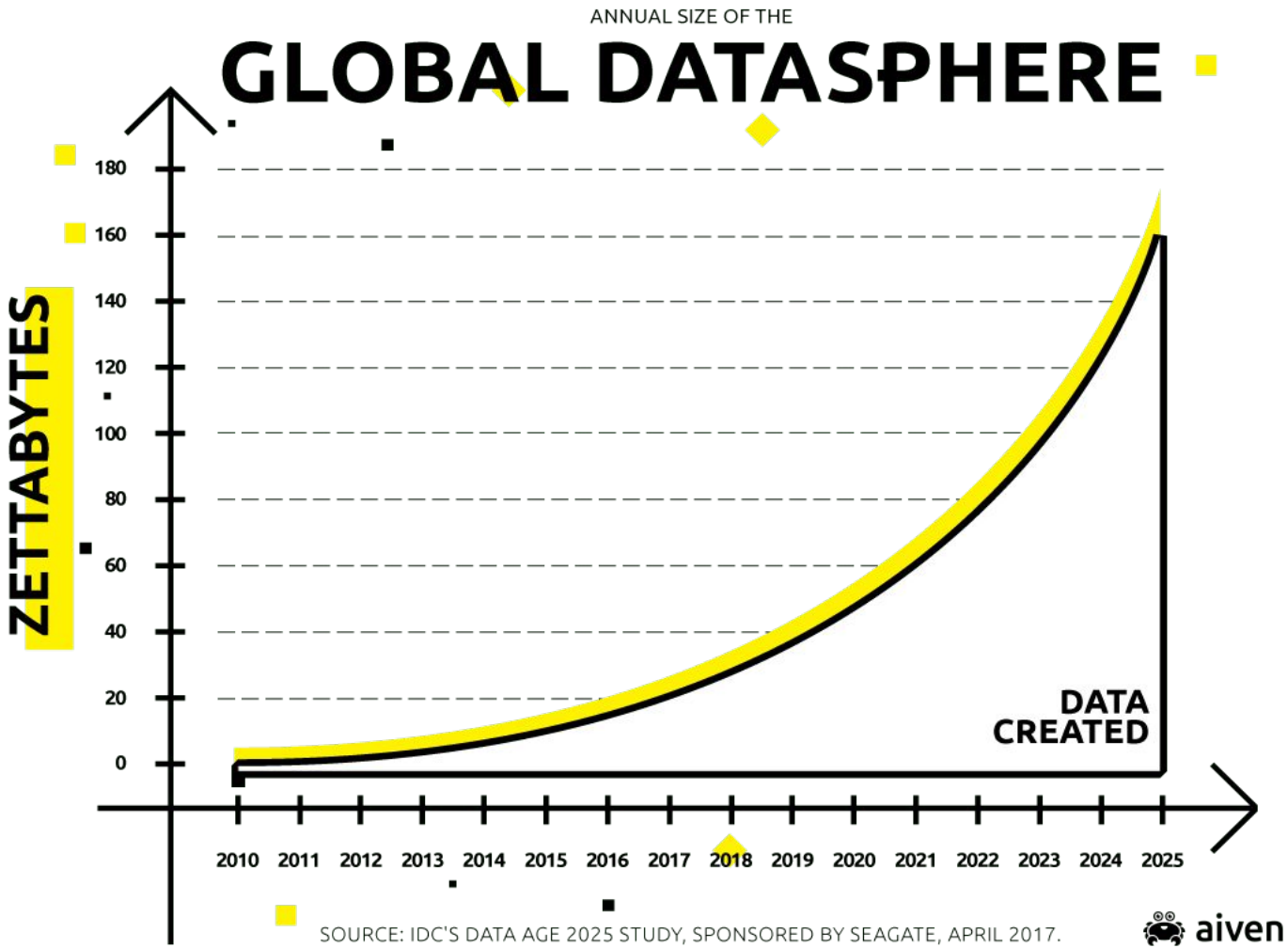
An arbitrary timeline of data



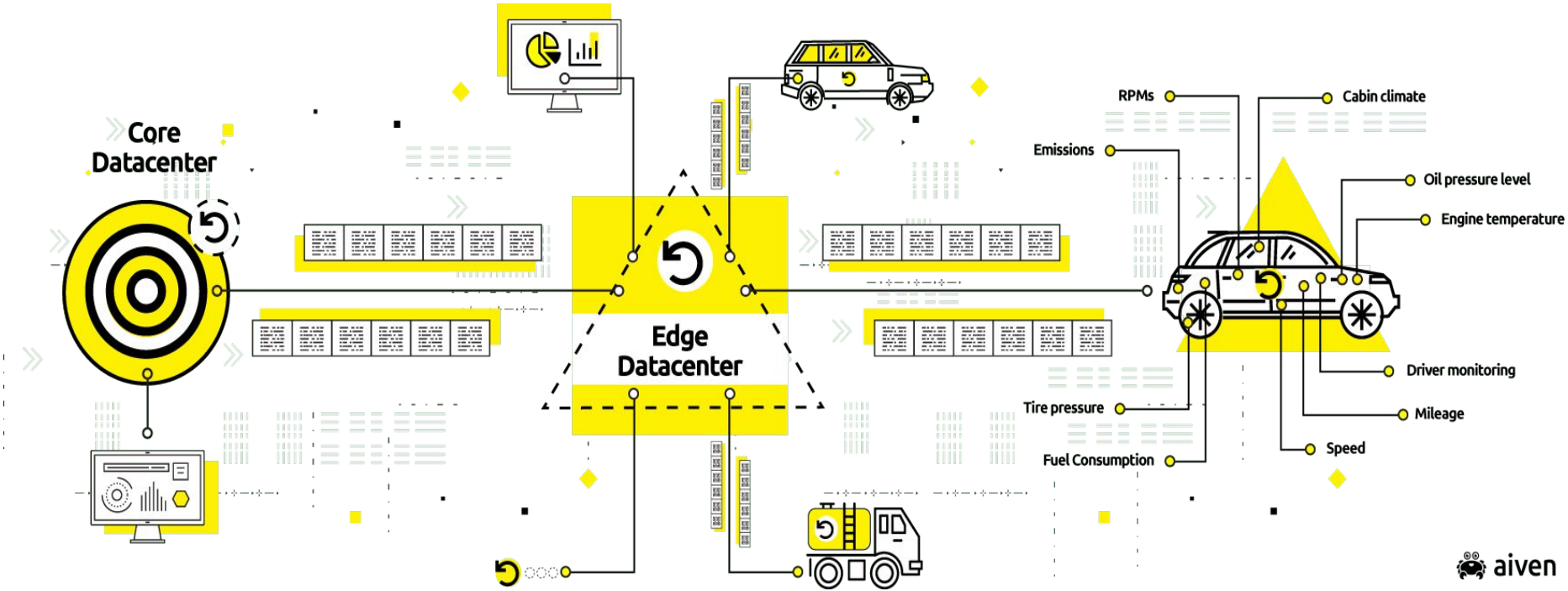
What have we learned?

Pipeline data has grown/transformed in terms of:



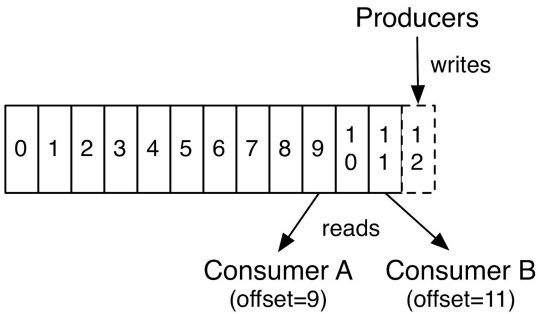

ANNUAL SIZE OF THE
GLOBAL DATASPHERE
ZETTABYTES
180
160
140
120
100
80
60
40
20
0
DATA CREATED
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
SOURCE: IDC'S DATA AGE 2025 STUDY, SPONSORED BY SEAGATE, APRIL 2017.
aiven

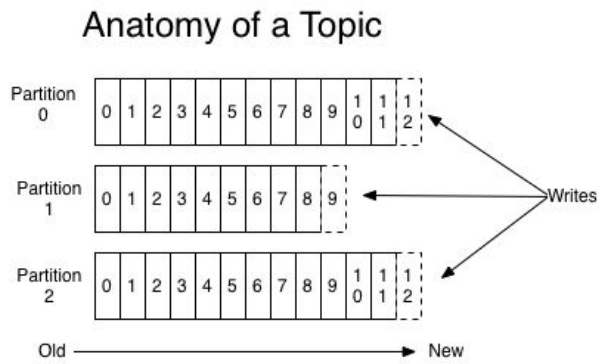
Core-to-edge computing



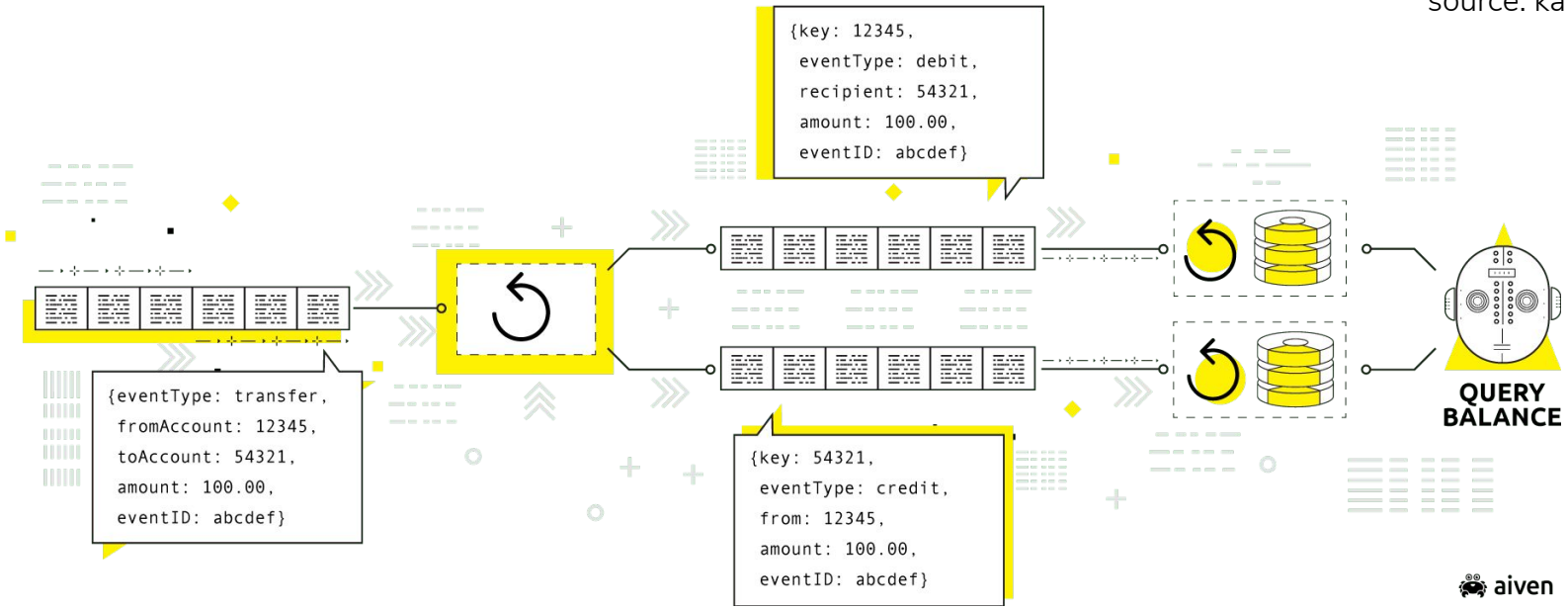
Functionality

• can autoscale, shard, tolerate partitions	• can capture, fix and requeue errored events
• can be troubleshot and configured on fly	• data can be used for AI/ML
• agnostic: accomodates all possible formats	• can self-perpetuate and automate continuous improvements





source: kafka.apache.org



Source: Martin Kleppmann's Kafka Summit 2018 Presentation, Is Kafka a Database?

Other design considerations

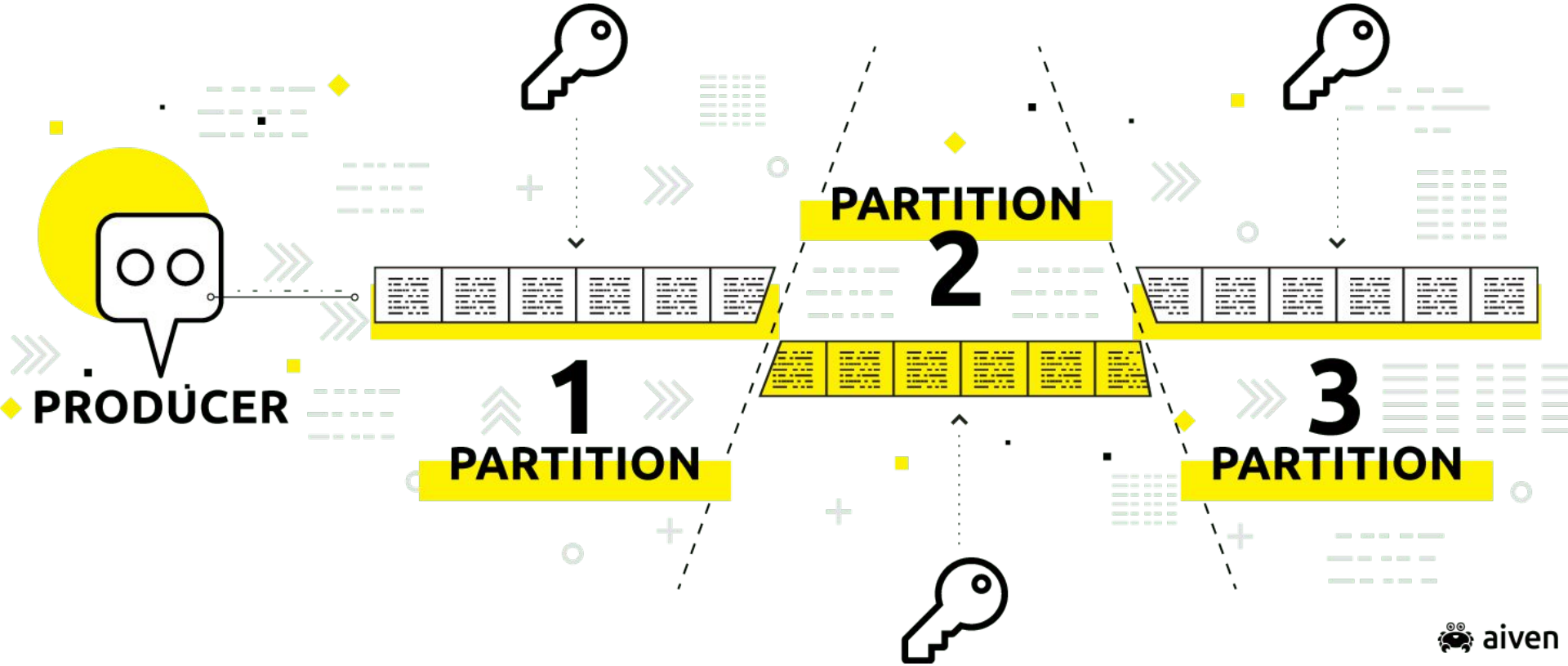
Kill switch - *when things go wrong; a way to stem the data flow.*

Query on the streams - like KSQL, Apache Flink, SQL on Amazon Kinesis, etc. Ultimately: one query interface for all data.

Need to accomodate - core-to-edge, growing data volume, near real-time velocity, bi-directionality

Distributed ledgers - pipelines can support these as tunably consistent distributed datastores

Compliance



Reliability/Performance/Usability/Other

- $MTTF \longrightarrow \infty$

Components: from ***interpreted*** to ***native*****?**

Metadata: will continue to transform, to offer even more space savings.

Pipelines: from ***hard-coded*** to ***drag and drop*** designs?

Which **data** is left ***volatile*** and which is ***stored***?

Wrapping up

Check out **The Future of Data Pipelines** at

https://aiven.io/blog/the-future-of-data-pipelines/

Discussion