Wolfram Wingerath

Real-Time Processing Explained

A Survey of Storm, Samza, Spark & Flink

Architecture

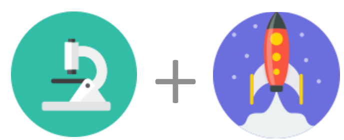
Baqend | code.talks

About me
Wolfram Wingerath

PhD Thesis & Research

Distributed Systems Engineer

Research:
- Real-Time Databases
- Stream Processing
- NoSQL & Cloud Databases
- …

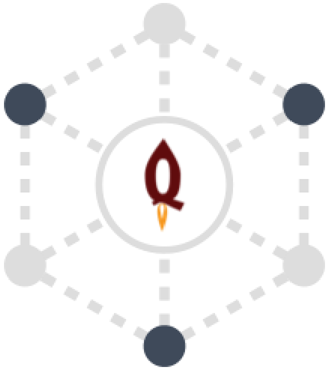


Practice:
- Backend-as-a-Service
- Web Caching
- Real-Time Database
- …

Universität Hamburg

BaQend
www.baqend.com

Who We Are



Our Product

Speed Kit:
- Accelerates *Any* Website
- Pluggable
- Easy Setup

test.speed-kit.com

Our Services

- Web & Data Management Workshops
- Performance Auditing
- Implementation Services

consulting@baqend.com

Baqend

Wolfram Wingerath

ww@baqend.com

Distributed Systems Engineer

Outline

Introduction
Big Data in Motion

System Survey
Big Data + Low Latency

Wrap-Up
Summary & Discussion

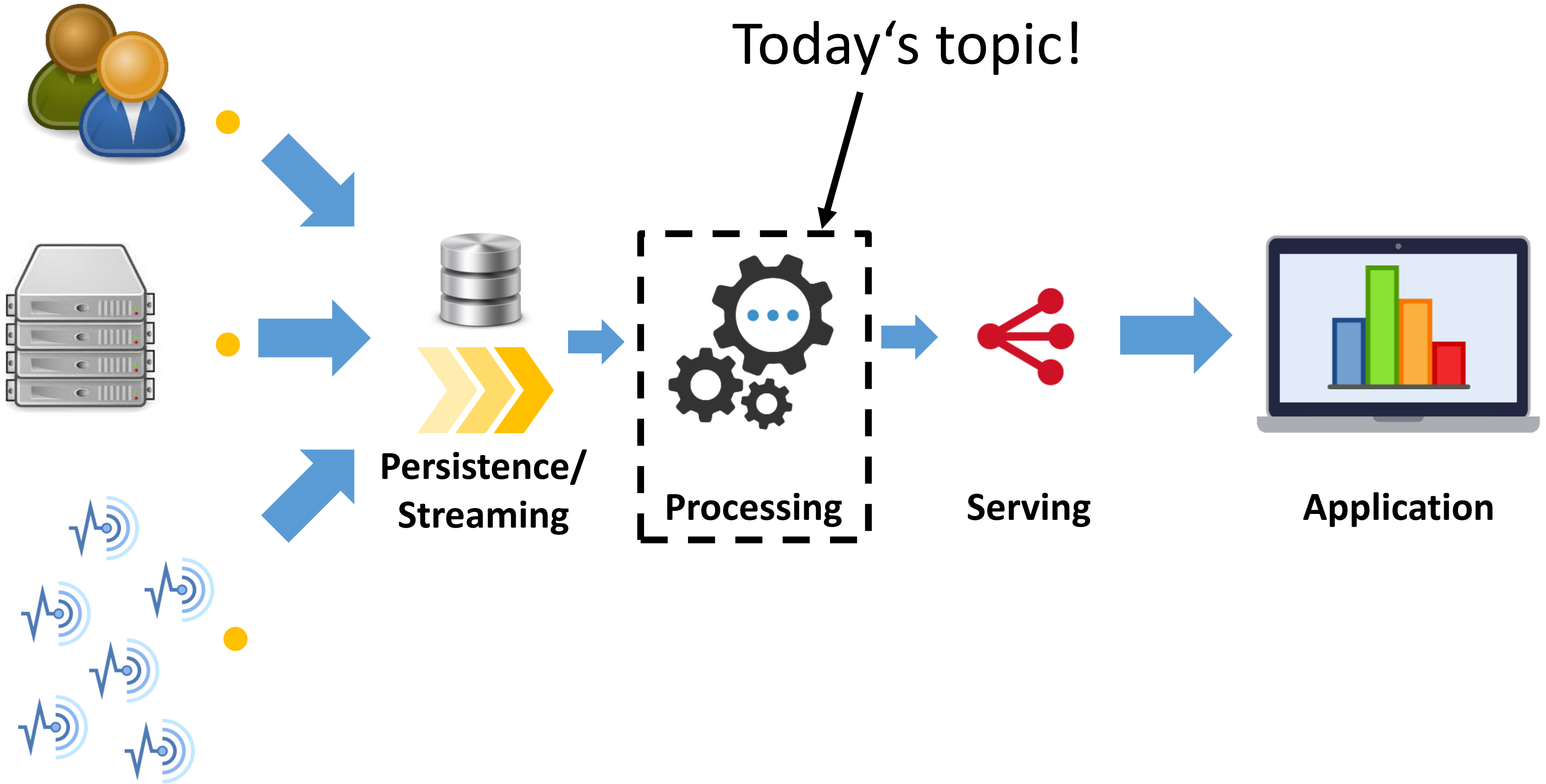
Future Directions
Real-Time Databases

- **Big Picture:**
  - A Typical Data Pipeline
  - Processing Frameworks
- **Processing Models:**
  - Batch Processing
  - Stream Processing
- **Streaming Architectures:**
  - Lambda Architecture
  - Kappa Architecture
  - Typical Operators
  - Exemplary Use Case

IN PRACTICE

Scalable Data Processing

A Data Processing Pipeline



Data Processing Frameworks

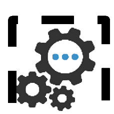
Scale-Out Made Feasible

Data processing frameworks **hide complexities of scaling**, e.g.:

- **Deployment -** code distribution, starting/stopping work
- **Monitoring -** health checks, application stats
- **Scheduling -** assigning work, rebalancing
- **Fault-tolerance** - restarting workers, rescheduling failed work

Running in cluster

Running on single node



Scaling out

INTRODUCTION

Batch vs Stream Processing

Big Data Processing Frameworks

What are your options?

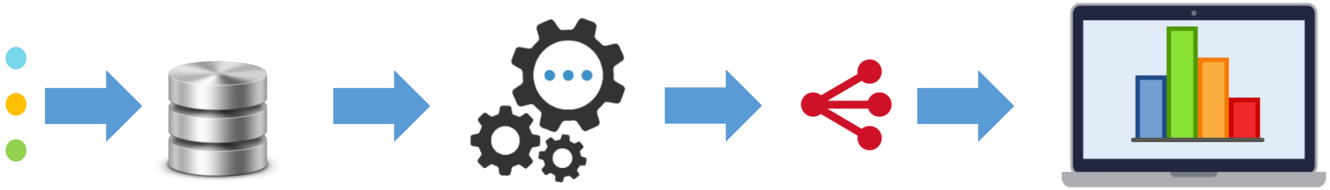
Batch Processing

„Volume“

- **Cost-effective** & Efficient
- **Easy to reason about**: operating on complete data

But:

- **High latency**: jobs periodic (e.g. during night times)

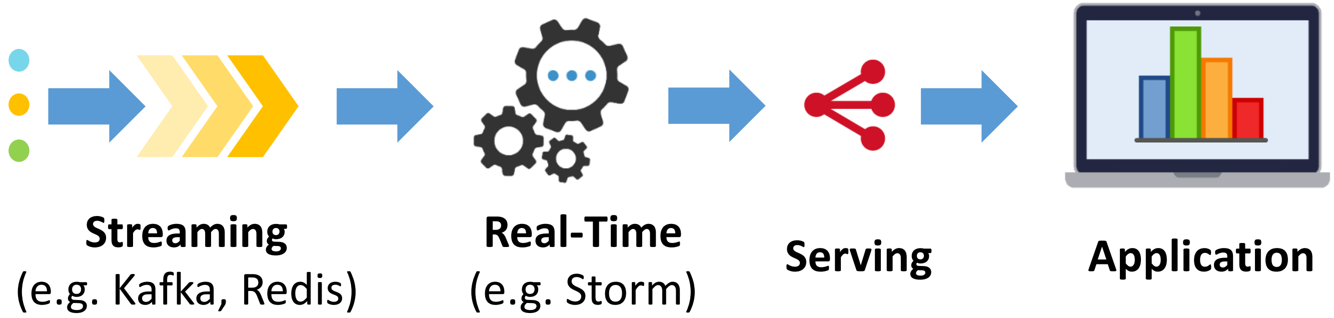


Persistence (e.g. HDFS) | **Batch** (e.g. MapReduce) | **Serving** (e.g. HBase) | **Application**

Stream Processing

„Velocity“

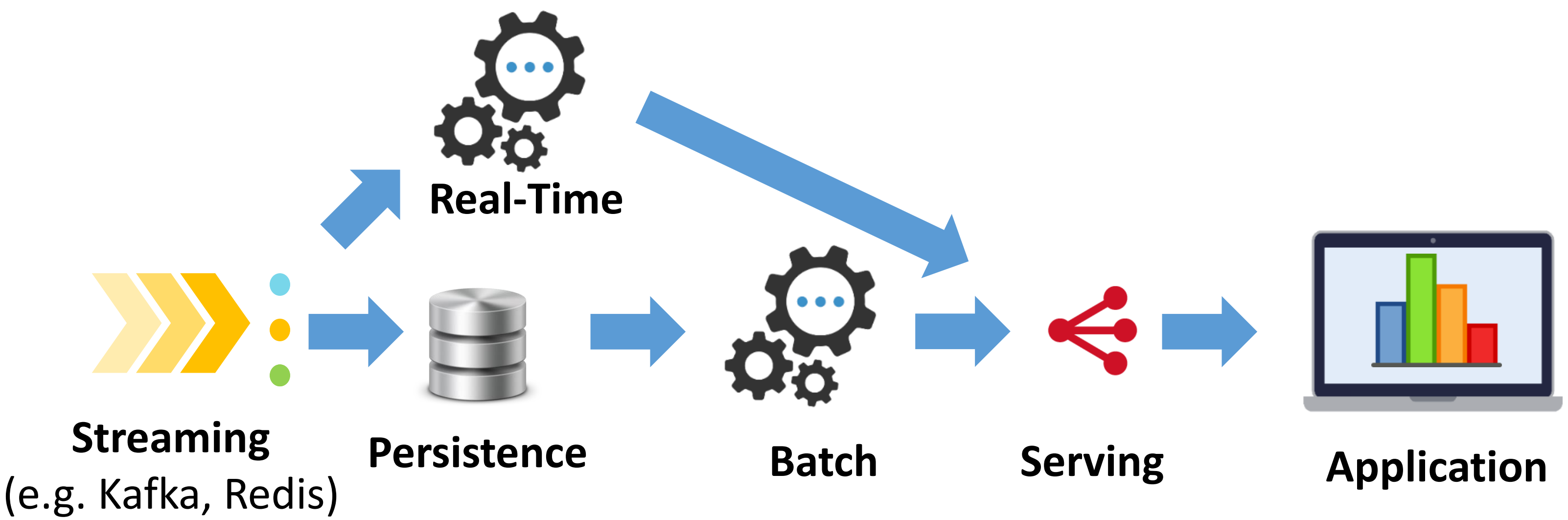
- Low end-to-end latency
- Challenges:
  - **Long-running jobs** - no downtime allowed
  - **Asynchronism** - data may arrive delayed or out-of-order
  - **Incomplete input** - algorithms operate on partial data
  - More: fault-tolerance, state management, guarantees, ...



Lambda Architecture

$\text{Batch}(D_{\text{old}}) + \text{Stream}(D_{\Delta\text{now}}) \approx \text{Batch}(D_{\text{all}})$

- **Fast** output (real-time)
- **Data retention + reprocessing** (batch)
  → **„eventually accurate“** merged views of real-time & batch
  Typical setups: Hadoop + Storm (→ Summingbird), Spark, Flink
- High complexity 2 code bases & 2 deployments

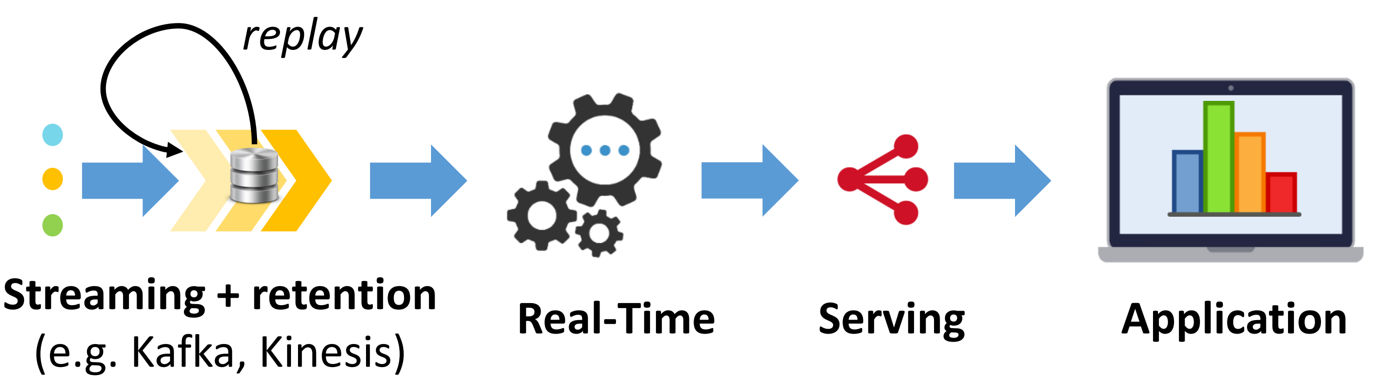


Nathan Marz, *How to beat the CAP theorem* (2011)
http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html

Kappa Architecture

Stream(D_{all}) = Batch(D_{all})

- **Simpler** than Lambda Architecture
- **Data retention** for history
- Reasons against Kappa:
    - Existing **legacy batch system**
    - **Special tools** only for a particular batch processor
    - Only **incremental** algorithms

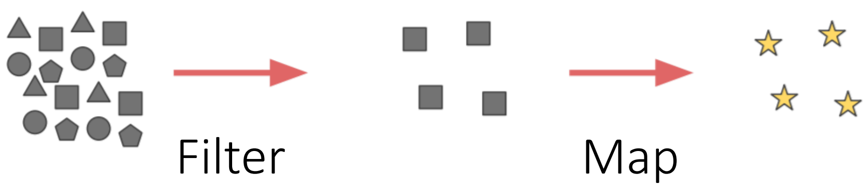


Jay Kreps, *Questioning the Lambda Architecture* (2014)
https://www.oreilly.com/ideas/questioning-the-lambda-architecture

Typical Stream Operators

Examples

Filter & Transform

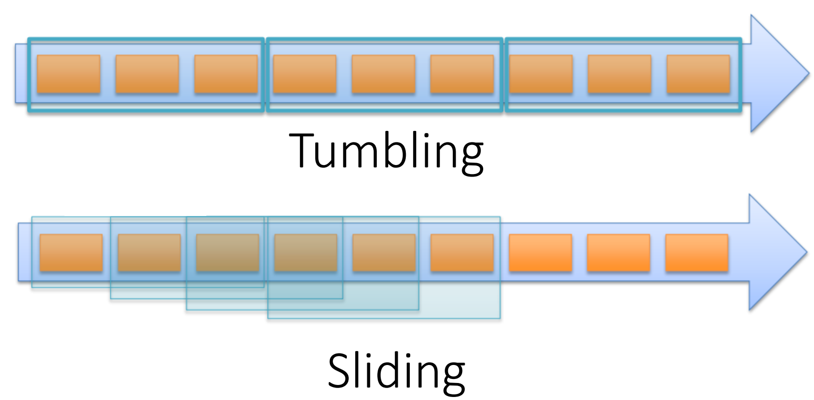


Group

Aggregates

SUM()
COUNT()

Windows



https://www.infoq.com/presentations/stream-processors-databases

https://www.infoq.com/presentations/stream-processing-apache-flink

Typical Use Case

Example from Yahoo!

Input

- **Read** Ad tracking data **from Kafka**

Filter

- **Discard** useless data

Project

- **Extract** relevant fields

Group

- By Ad **campaign**

Window

- Ad views per **10-min-window**

https://yahooeng.tumblr.com/post/135321837876/benchmarking-streaming-computation-engines-at

Wrap-up

Data Processing

- Processing frameworks abstract from **scaling issues**

Batch processing

- easy to reason about
- extremely efficient
- huge input-output latency

Stream processing

- quick results
- purely incremental
- potentially complex to handle

- **Lambda Architecture**: batch + stream processing
- **Kappa Architecture**: stream-only processing

Outline

Introduction
Big Data in Motion

System Survey
Big Data + Low Latency

Wrap-Up
Summary & Discussion

Future Directions
Real-Time Databases

- Processing Models:
  Stream ⟷ Batch
- Stream Processing Frameworks:
  - Storm
  - Trident
  - Samza
  - Flink
  - Other Systems

SURVEY

Popular Stream Processing Systems

Processing Models

Batch vs. Micro-Batch vs. Stream

stream | **micro-batch** | **batch**

Flink

STORM Trident

samza

hadoop

Spark Streaming

Amazon Elastic MapReduce

low latency ↔ high throughput

Storm

„Hadoop of real-time“

Overview

- **First** production-ready, well-adopted stream processor
- **Compatible**: native Java API, Thrift, distributed RPC
- **Low-level**: no primitives for joins or aggregations
- **Native stream processor**: latency < 50 ms feasible
- **Big users**: Twitter, Yahoo!, Spotify, Baidu, Alibaba, ...

History

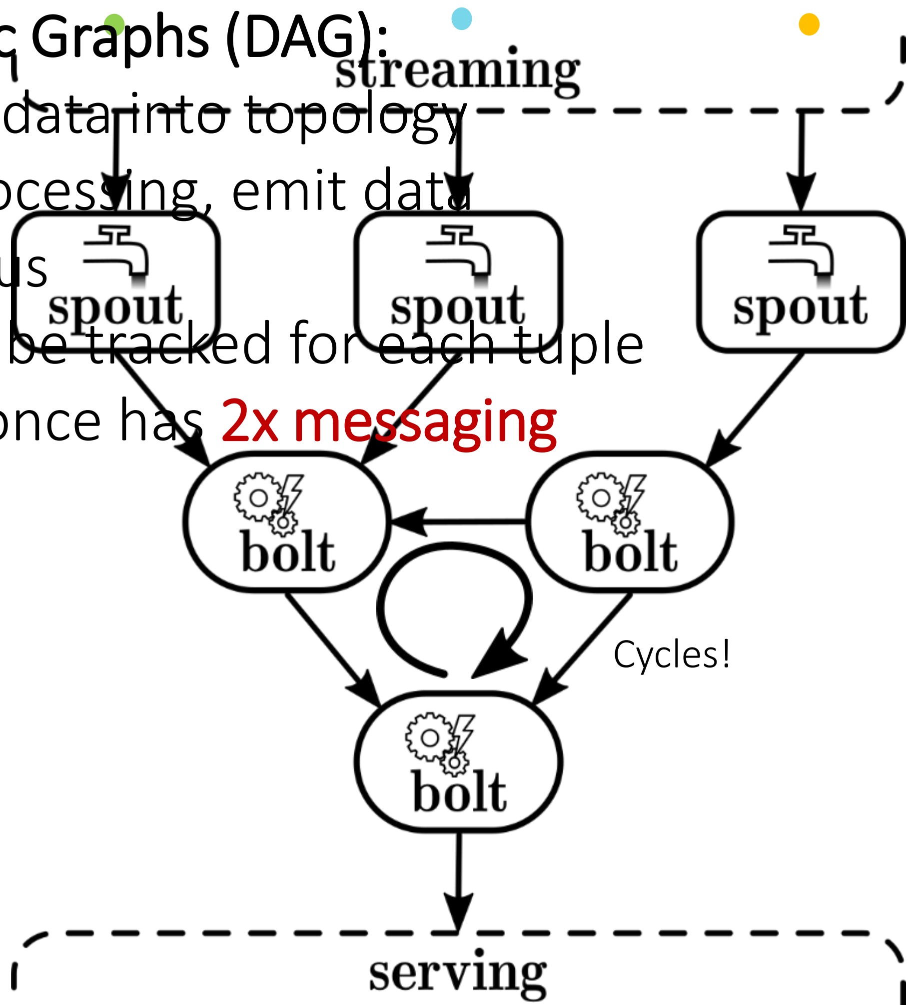
- **2010**: developed at BackType (acquired by Twitter)
- **2011**: open-sourced
- **2014**: Apache top-level project

Dataflow

Directed Acyclic Graphs (DAG):

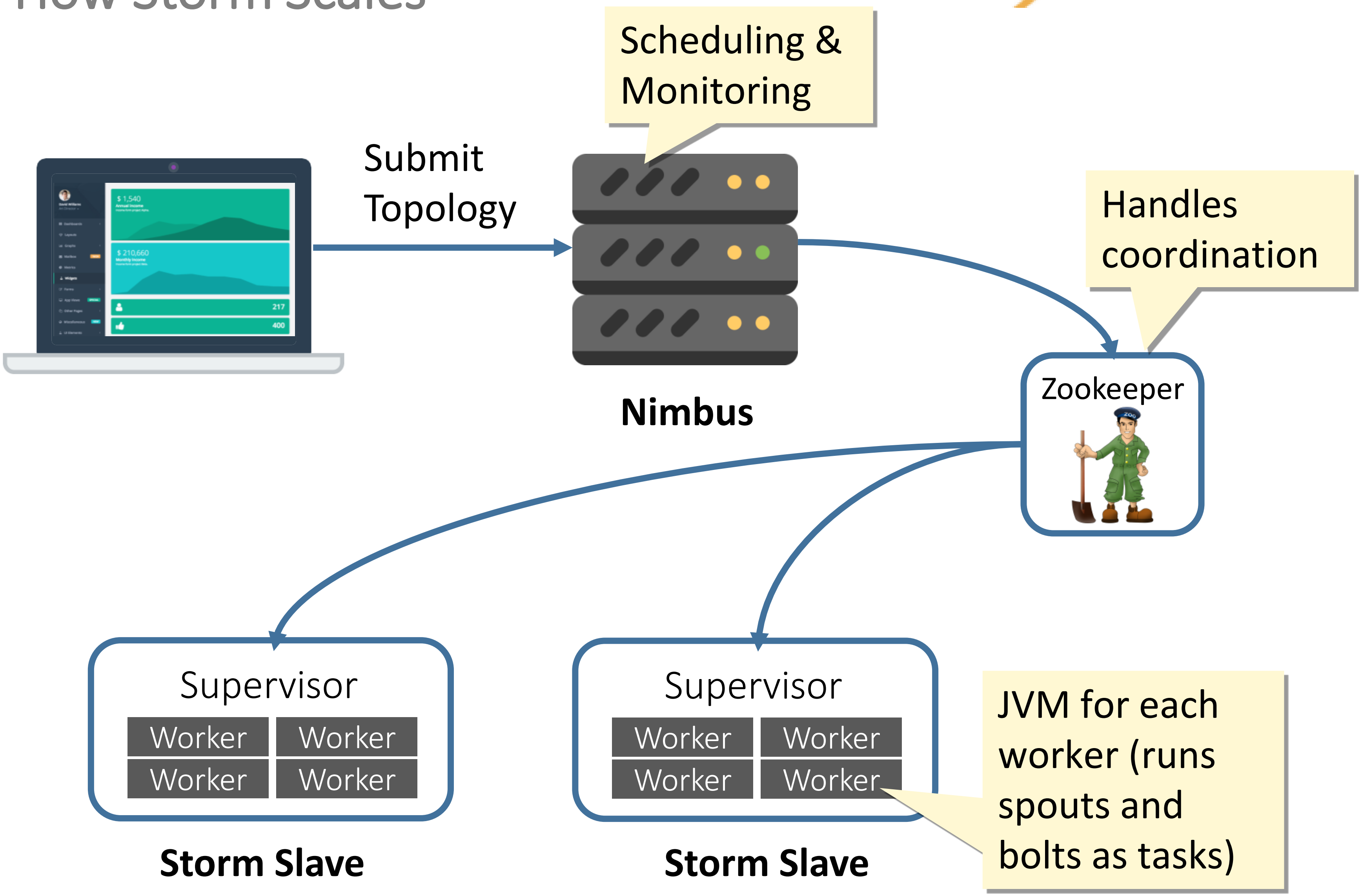
- **Spouts**: pull data into topology
- **Bolts**: do processing, emit data
- Asynchronous
- Lineage can be tracked for each tuple

  → At-least-once has 2x messaging overhead



Cluster Architecture

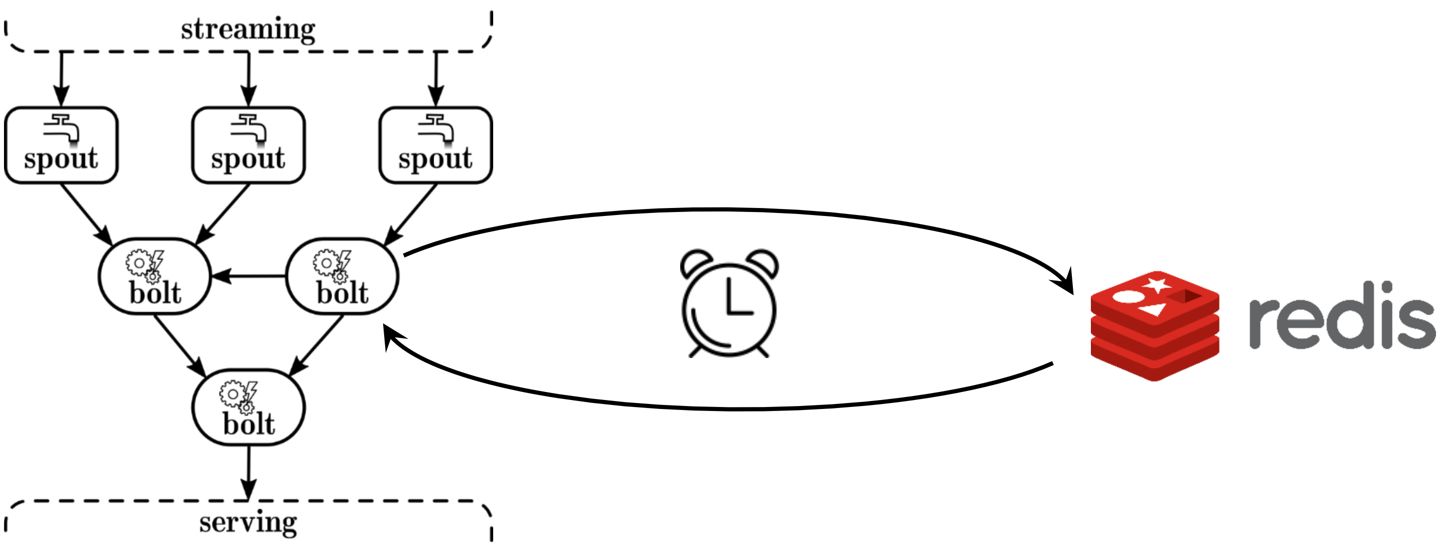
How Storm Scales



State Management

Recover State on Failure

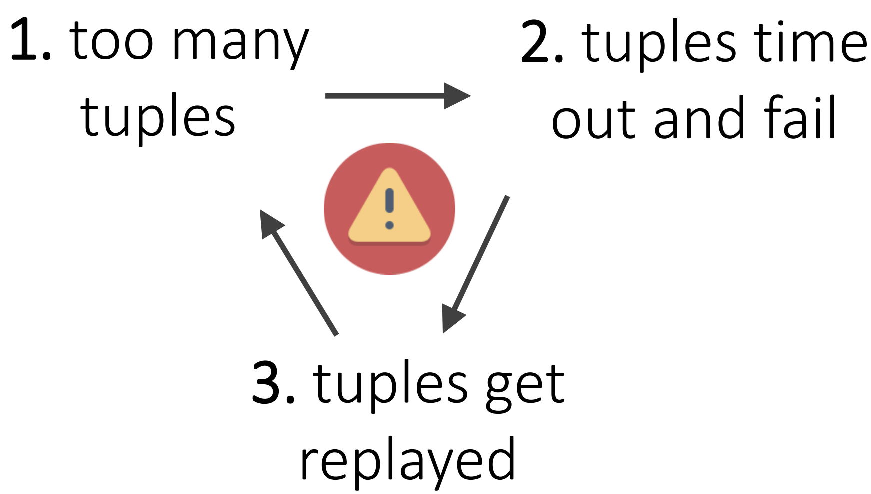
- **In-memory** or **Redis**-backed reliable state
- *Synchronous state communication* on the critical path
  → infeasible for large state



Back Pressure

Throttling Ingestion on Overload

STORM



Approach: monitoring bolts‘ inbound buffer

1. Exceeding **high watermark** → throttle!
2. Falling below **low watermark** → full power!

Trident

Stateful Stream Joining on Storm

STORM Trident

Overview:

- Abstraction layer on top of Storm
- Released in 2012 (Storm 0.8.0)
- **Micro-batching**
- **New features**:
  - High-level API: aggregations & joins
  - Strong ordering
  - Stateful exactly-once processing

    → Performance penalty

Trident

Partitioned Micro-Batching

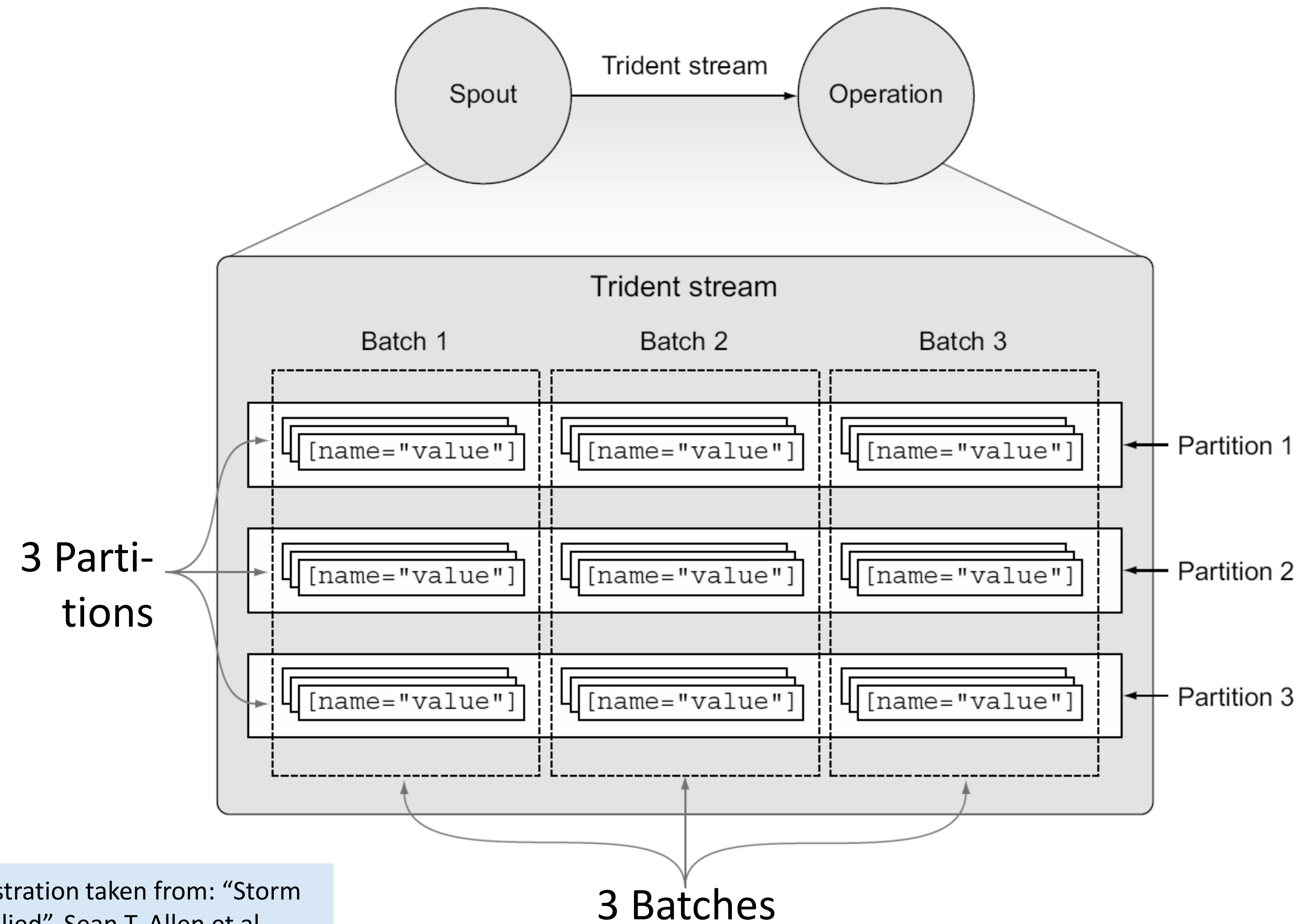


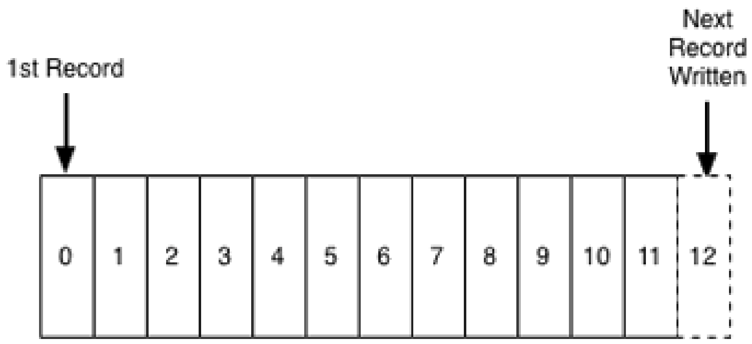
Illustration taken from: "Storm applied", Sean T. Allen et al.

Samza

Real-Time on Top of Kafka

Overview

- Co-developed with **Kafka**
  → **Kappa Architecture**
- **Simple**: only single-step jobs
- Local state
- Native stream processor: low latency
- **Users**: LinkedIn, Uber, Netflix, TripAdvisor, Optimizely, ...



History

- Developed at **LinkedIn**
- **2013**: open-source (Apache Incubator)
- **2015**: Apache top-level project

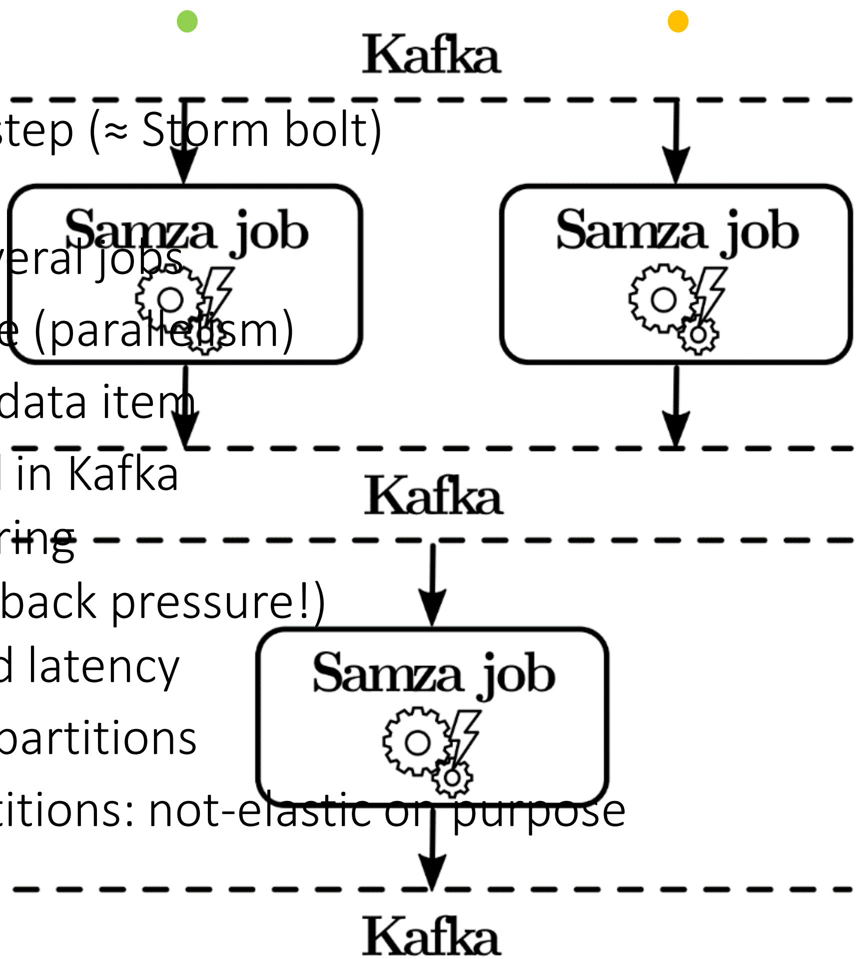
Illustration taken from: Jay Kreps, *Questioning the Lambda Architecture* (2014)
https://www.oreilly.com/ideas/questioning-the-lambda-architecture (2017-03-02)

Dataflow

Simple By Design

samza

- **Job**: processing step (≈ Storm bolt)
  → Robust
  → But: often several jobs
- **Task**: job instance (parallelism)
- **Message**: single data item
- **Output persisted** in Kafka
  → Easy data sharing
  → Buffering (no back pressure!)
  → But: Increased latency
- **Ordering** within partitions
- Task = Kafka partitions: not-elastic on purpose



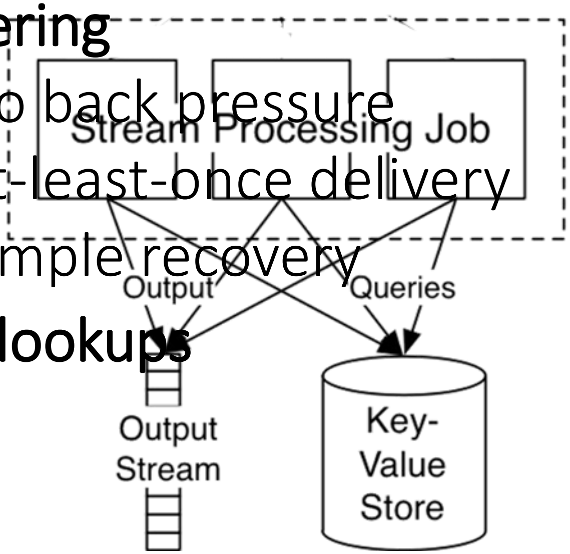
Martin Kleppmann, *Turning the database inside-out with Apache Samza* (2015)
https://www.confluent.io/blog/turning-the-database-inside-out-with-apache-samza/ (2017-02-23)

Samza

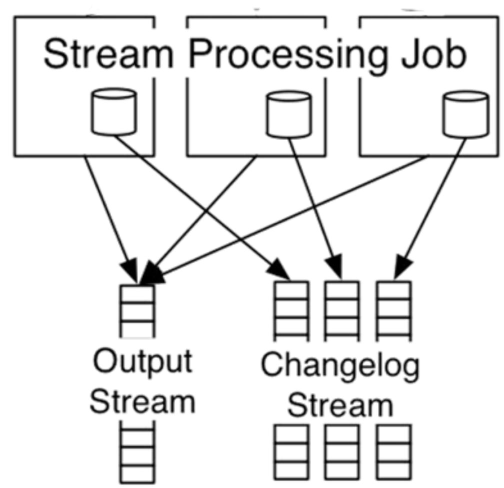
Local State

Advantages of local state:

- **Buffering**
  → No back pressure
  → At-least-once delivery
  → Simple recovery
- **Fast lookups**



Remote State



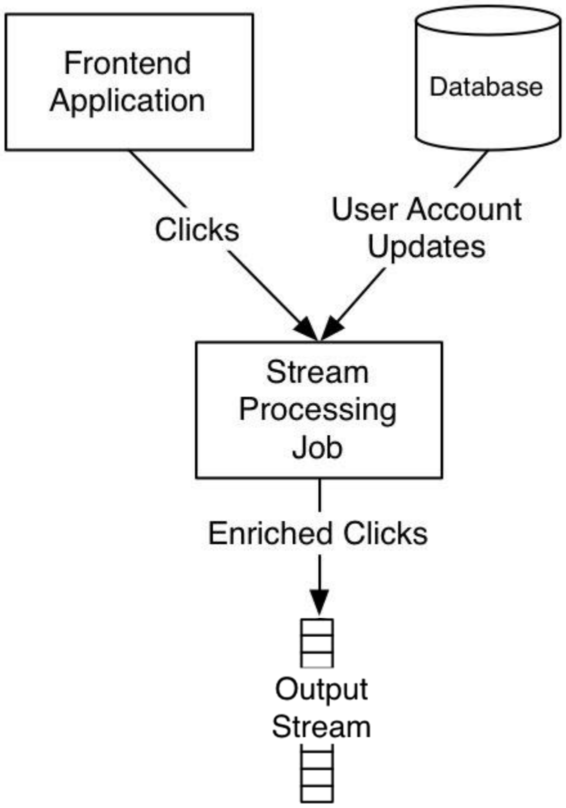
Local State

Illustrations taken from: Jay Kreps, *Why local state is a fundamental primitive in stream processing* (2014)
https://www.oreilly.com/ideas/why-local-state-is-a-fundamental-primitive-in-stream-processing (2017-02-26)

Dataflow

Example: Enriching a Clickstream

samza



Example: the *enriched clickstream* is available to every team within the organization

Illustration taken from: Jay Kreps, *Why local state is a fundamental primitive in stream processing* (2014)
https://www.oreilly.com/ideas/why-local-state-is-a-fundamental-primitive-in-stream-processing (2017-02-26)

State Management

Straightforward Recovery

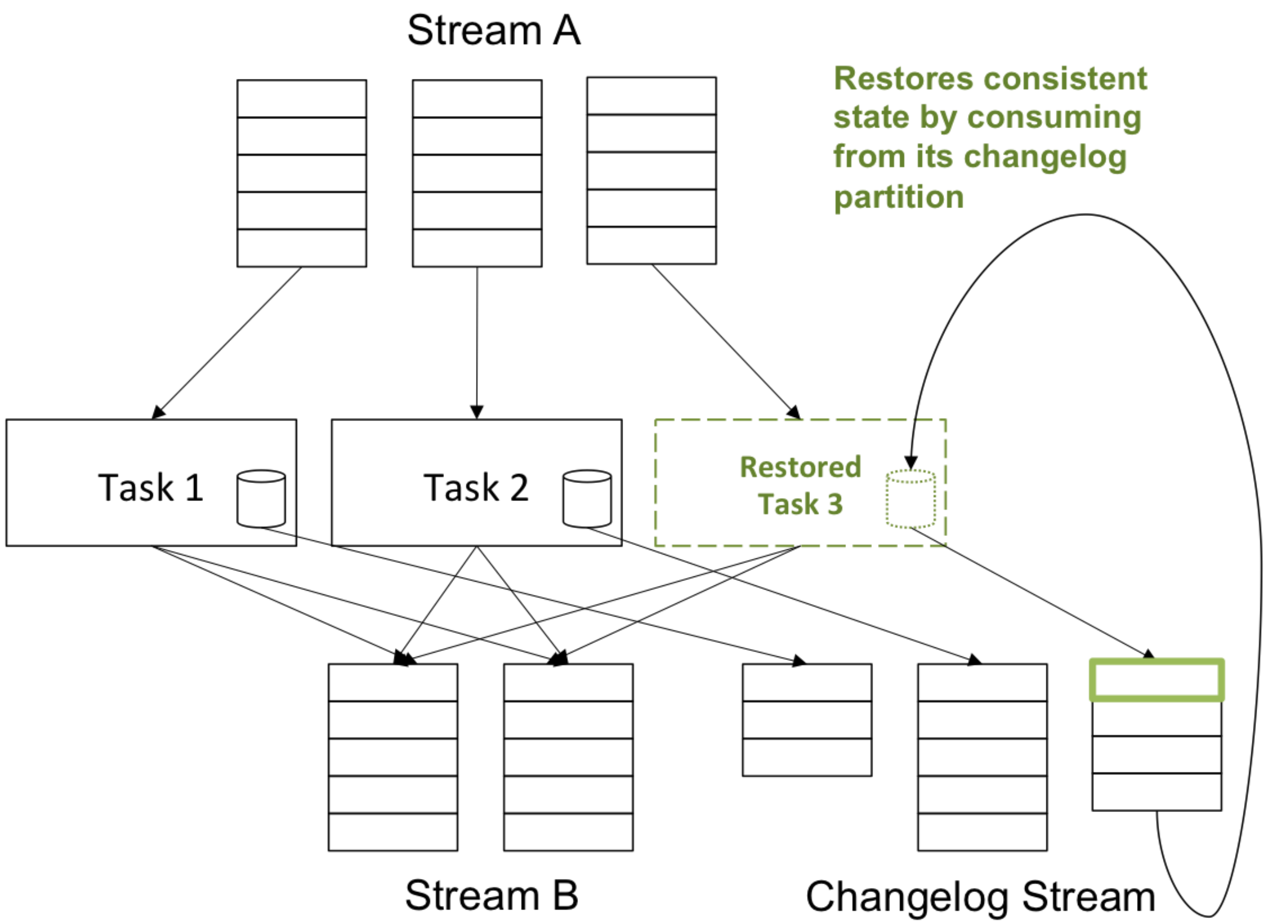


Illustration taken from: Navina Ramesh, *Apache Samza, LinkedIn's Framework for Stream Processing* (2015)
https://thenewstack.io/apache-samza-linkedins-framework-for-stream-processing (2017-02-26)

Spark

„MapReduce successor“

Overview

- **High-level API**: immutable collections (RDDs)

Core	SQL	MLlib	GraphX	Spark Streaming

- **Community**: 1000+ contributors in 2015
- **Big users**: Amazon, eBay, Yahoo!, IBM, Baidu, ...

History

- **2009**: developed at UC Berkeley
- **2010**: open-sourced
- **2014**: Apache top-level project

Spark Streaming

Overview

- **High-level API**: DStreams (~Java 8 Streams)
- **Micro-Batching**: seconds of latency
- **Rich features**: stateful, exactly-once, elastic

History

- **2011**: start of development
- **2013**: Spark Streaming becomes part of Spark Core

Spark Streaming

Core Abstraction: DStream

Resilient Distributed Data set (RDD)

- **Immutable** collection & **deterministic** operations
- **Lineage** tracking:
  → state can be reproduced
  → periodic checkpoints reduce recovery time

DStream: Discretized RDD

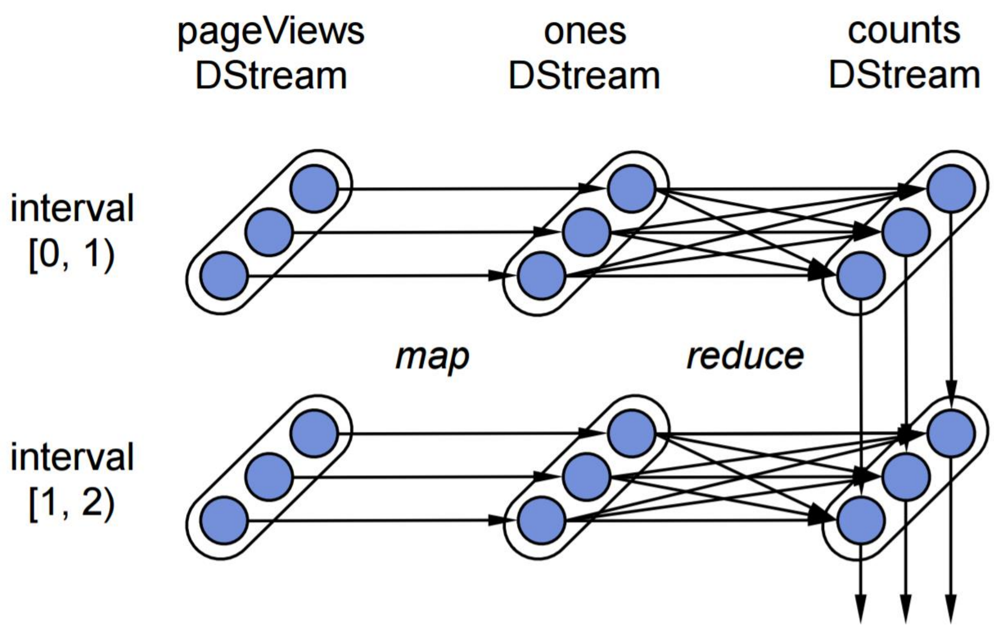
- **RDDs are processed in order**: no ordering within RDD
- RDD scheduling ~50 ms → latency >100ms

Illustration taken from:
http://spark.apache.org/docs/latest/streaming-programming-guide.html#overview (2017-02-26)

Example

Counting Page Views

```
pageViews = readStream("http://...", "1s")
ones = pageViews.map(event => (event.url, 1))
counts = ones.runningReduce((a, b) => a + b)
```



Zaharia, Matei, et al. "Discretized streams: Fault-tolerant streaming computation at scale." *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013.

Flink

Overview

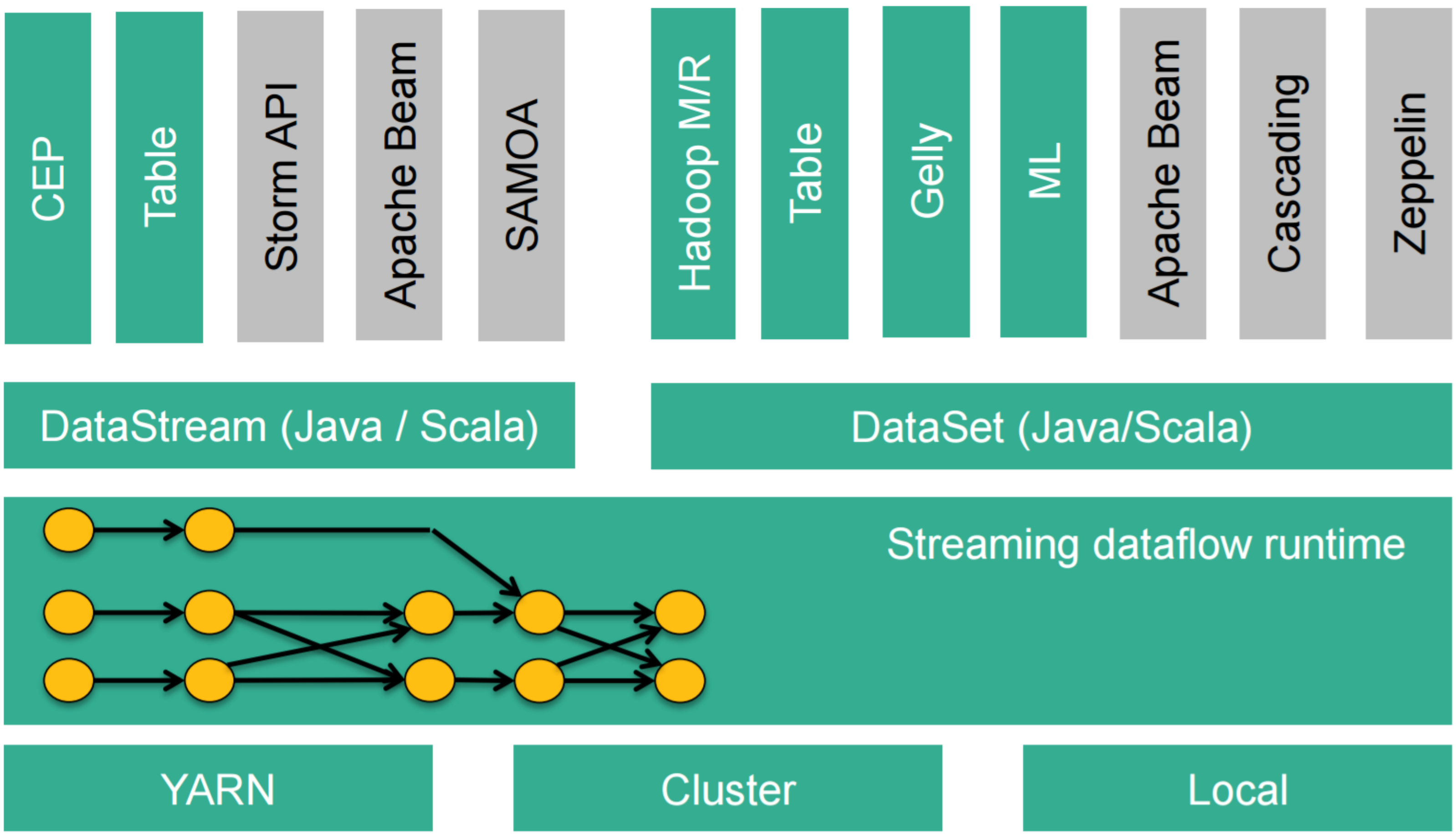
- **Native stream processor:** Latency <100ms feasible
- **Abstract API** for stream and batch processing, stateful, exactly-once delivery
- **Many libraries**: Table and SQL, CEP, Machine Learning , Gelly…
- **Users**: Alibaba, Ericsson, Otto Group, ResearchGate, Zalando…

History

- **2010**: start as **Stratosphere** at TU Berlin, HU Berlin, and HPI Potsdam
- **2014**: Apache Incubator, project renamed to Flink
- **2015**: Apache top-level project

Architecture

Streaming + Batch

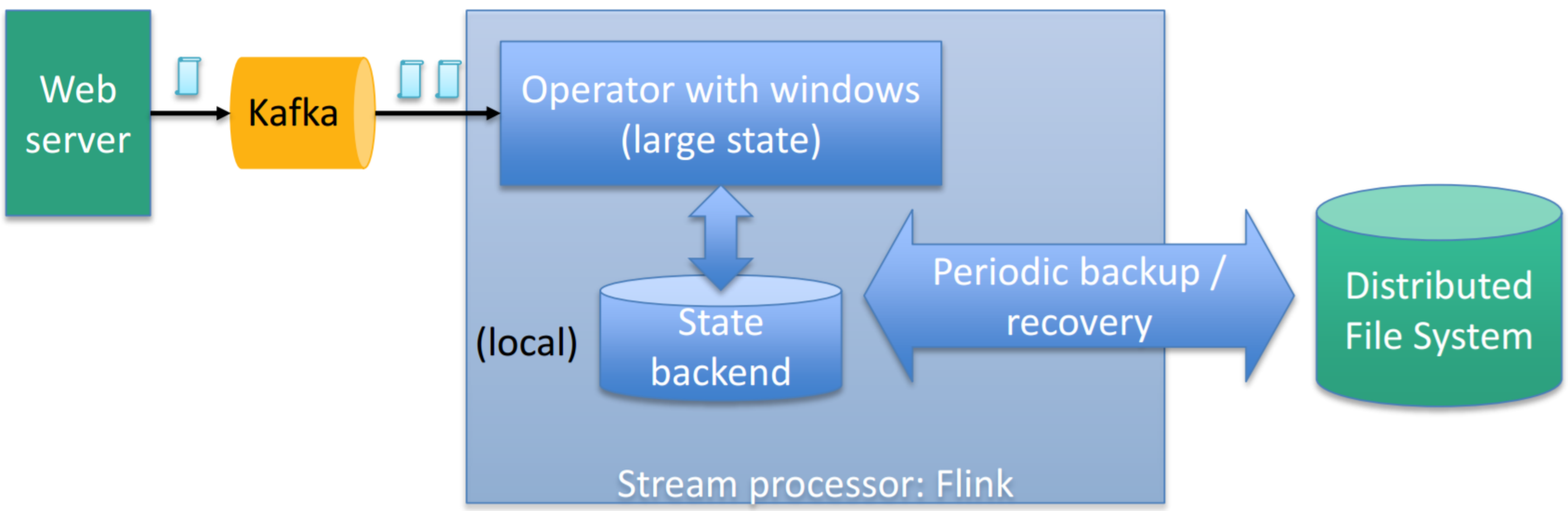


https://www.infoq.com/presentations/stream-processing-apache-flink

Managed State

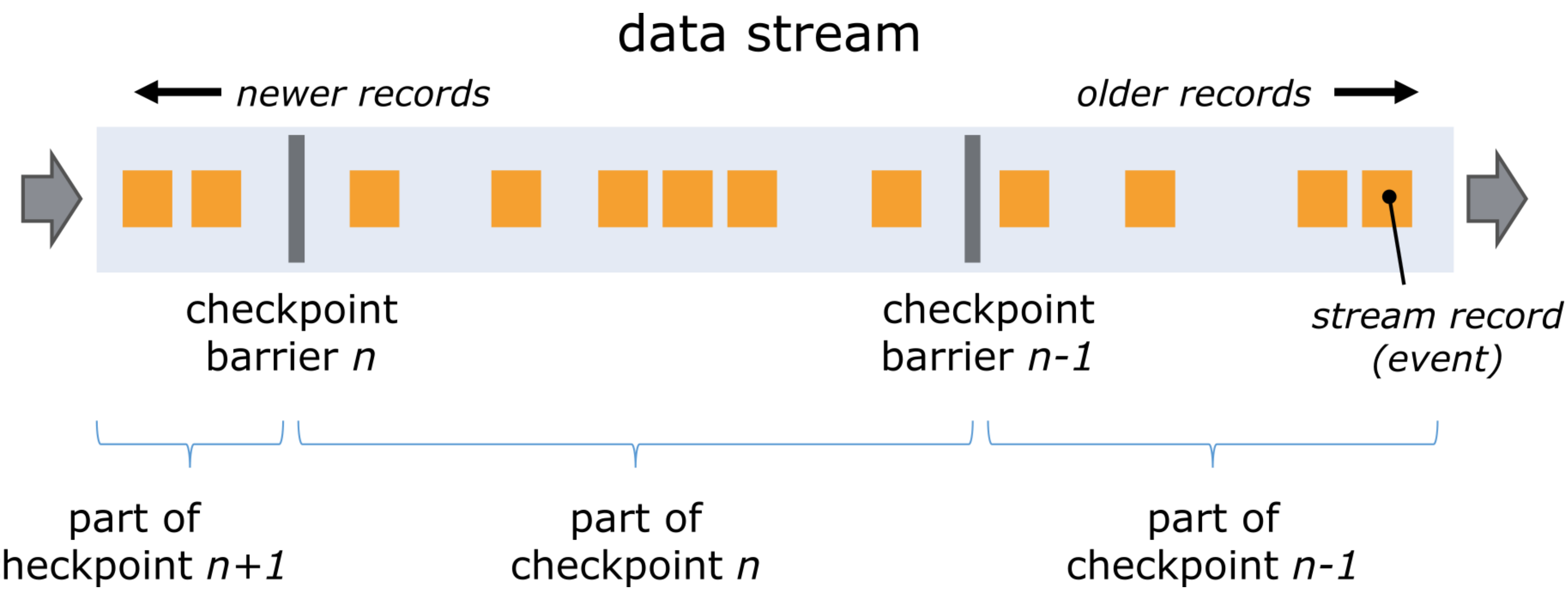
Streaming + Batch

- Automatic **Backups** of local state
- Stored in **RocksDB,** Savepoints written to **HDFS**



https://www.infoq.com/presentations/stream-processing-apache-flink

Highlight: Fault Tolerance
Distributed Snapshots



- **Ordering** within stream partitions
- Periodic **checkpoints**
- **Recovery**:
  1. *reset state* to checkpoint
  2. *replay data* from there

→ **Exactly-once**

Illustration taken from: https://ci.apache.org/projects/flink/flink-docs-release-1.2/internals/stream_checkpointing.html (2017-02-26)

Outline

Introduction
Big Data in Motion

System Survey
Big Data + Low Latency

Wrap-Up
Summary & Discussion

- Comparison Matrix
- Other Systems
- One-Line Takeaway

Future Directions
Real-Time Databases

WRAP UP

Side-by-side comparison

Comparison

	Storm	Trident	Samza	Spark Streaming	Flink (streaming)
Strictest Guarantee	at-least-once	exactly-once	at-least-once	exactly-once	exactly-once
Achievable Latency	≪100 ms	<100 ms	<100 ms	<1 second	<100 ms
State Management	○ (small state)	○ (small state)	✓	✓	✓
Processing Model	one-at-a-time	micro-batch	one-at-a-time	micro-batch	one-at-a-time
Backpressure	✓	✓	no (buffering)	✓	✓
Ordering	✗	between batches	within partitions	between batches	within partitions
Elasticity	✓	✓	✗	✓	✓

Performance
Yahoo! Benchmark

- Based on **real use case**:
  - Filter and count ad impressions
  - 10 minute windows

“**Storm** […] and **Flink** […] show **sub-second latencies** at relatively high throughputs with **Storm** having the **lowest 99th percentile** latency. **Spark** streaming […] supports high **throughputs**, but at a relatively **higher latency**.”

From https://yahooeng.tumblr.com/post/135321837876/benchmarking-streaming-computation-engines-at

Other Systems

Heron

Apex

Dataflow

Beam

Kafka Streams

IBM InfoSphere Streams

And even more: Kinesis, Gearpump, MillWheel, Muppet, S4, Photon, …

Outline

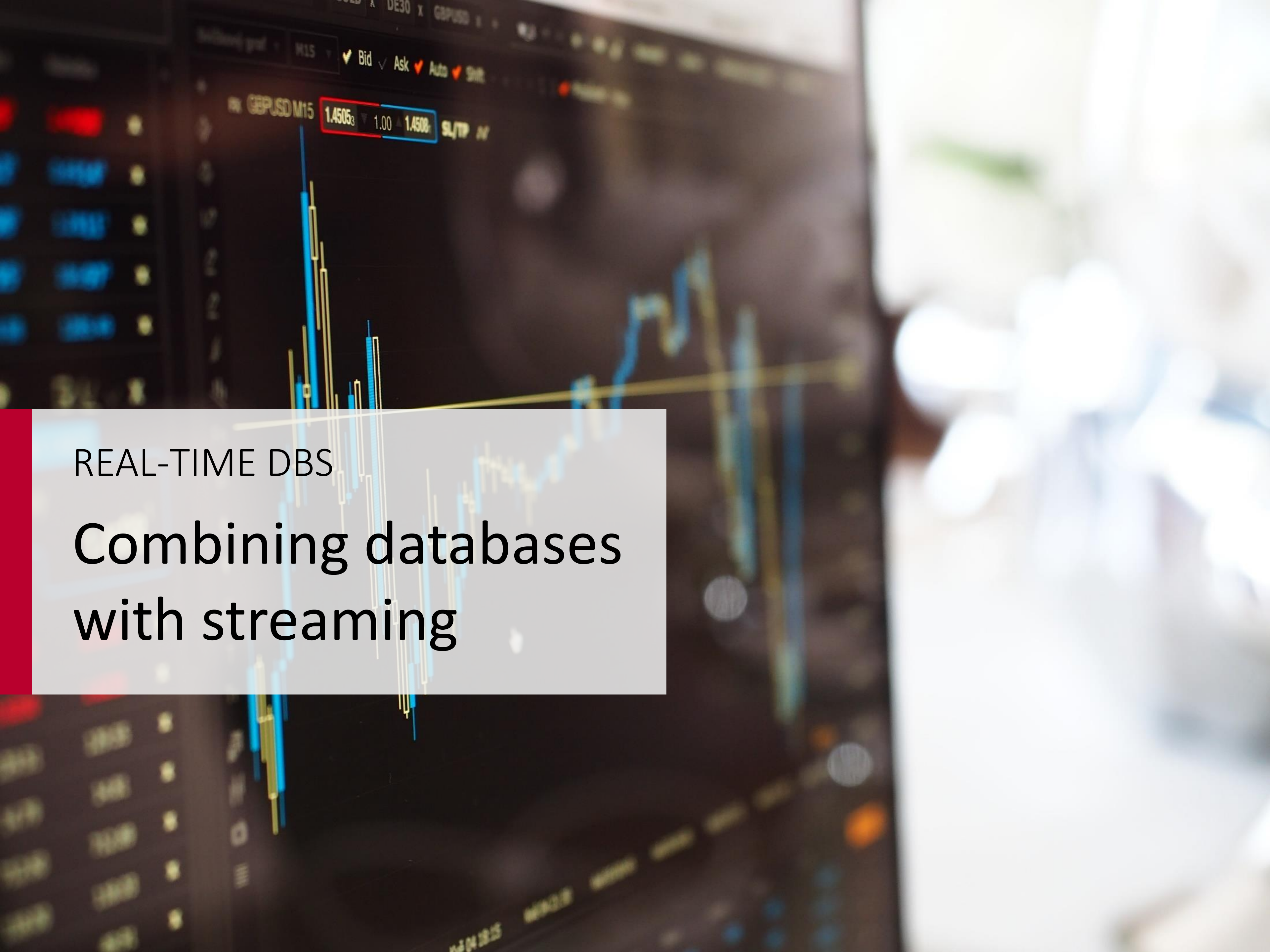
Introduction
Big Data in Motion

System Survey
Big Data + Low Latency

Wrap-Up
Summary & Discussion

Future Directions
Real-Time Databases

- **Real-Time Databases:**
  - **Why** Push-Based Database Queries?
  - **Where** Do Real-Time Databases Fit in?
- **Comparison Matrix:**
  - Meteor
  - RethinkDB
  - Parse
  - Firebase
  - Baqend
- **Use Cases at Baqend:**
  - Query Caching
  - Real-Time Queries

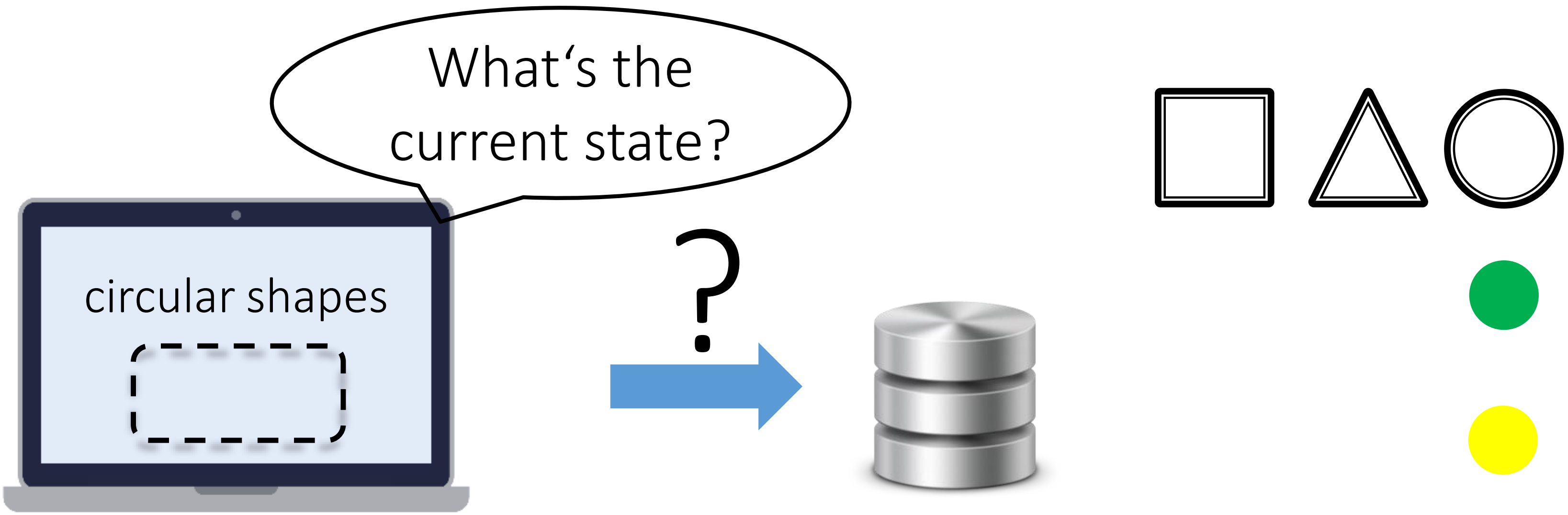


REAL-TIME DBS

Combining databases with streaming

Traditional Databases

No Request? No Data!



Query maintenance: periodic polling

→ Inefficient

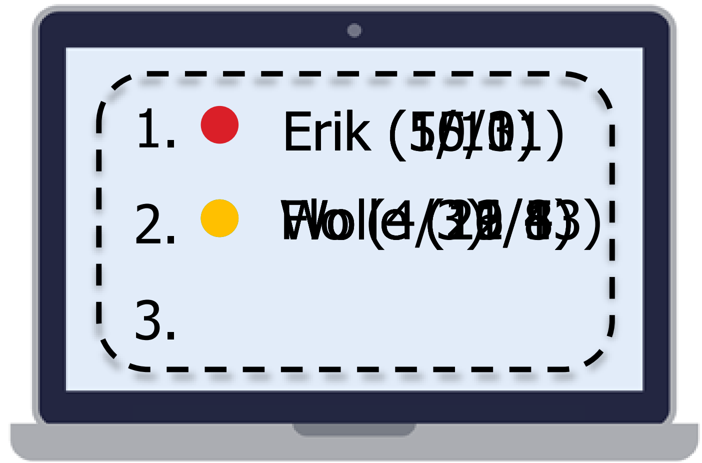
→ Slow

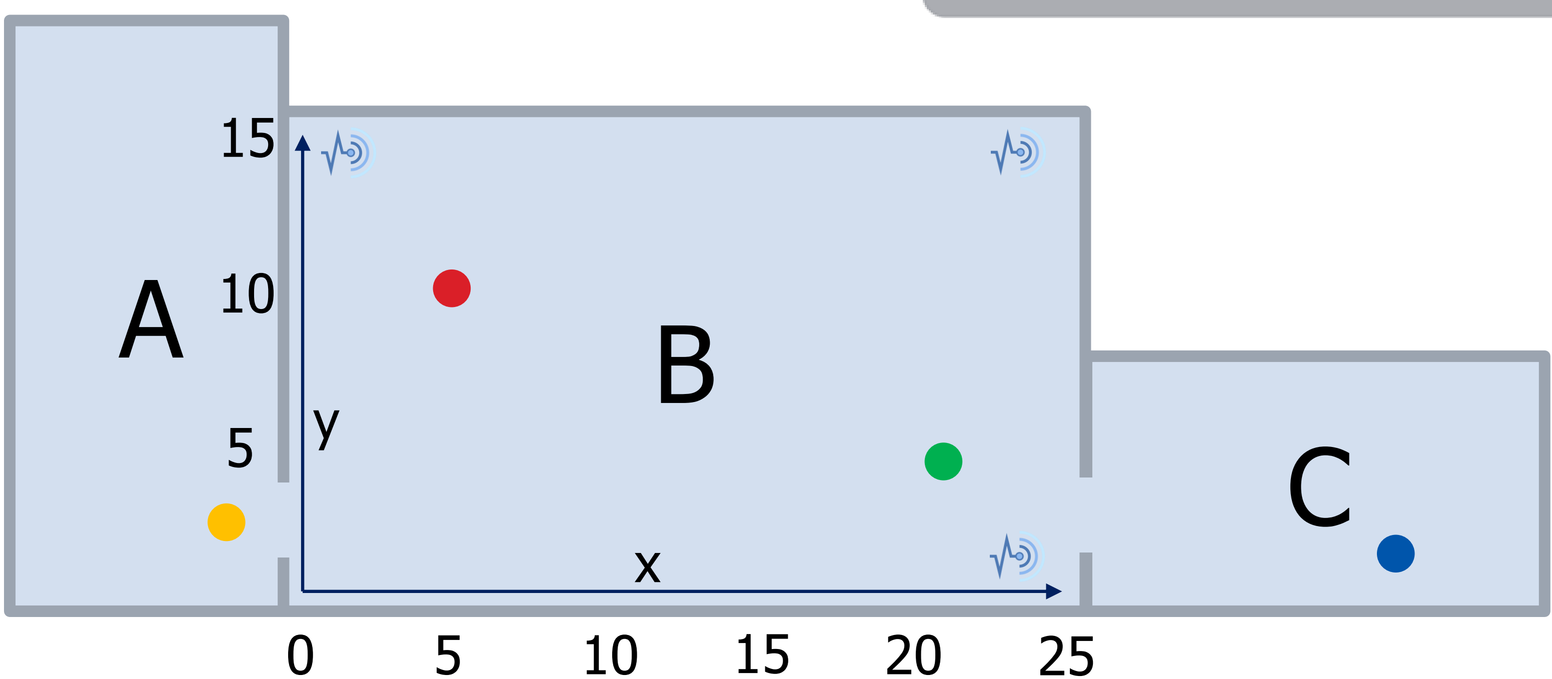
Push-Based Access For Evolving Domains

Self-Maintaining Results

Find people in Room B:

```
SELECT name, x, y
  FROM People
  WHERE x BETWEEN 0 AND 25
    AND y BETWEEN 0 AND 15
  ORDER BY name ASC
```





Data Management Overview

DBMS vs. Real-Time DB vs. Stream Processing

Database Management	Real-Time Databases	Stream Processing
static collections	evolving collections	ephemeral streams

pull-based ⟵ ⟶ push-based

Real-Time Databases

In a Nutshell

	Meteor Poll-and-Diff	Meteor Oplog Tailing	RethinkDB	Parse	Firebase
Scales with write TP	✓	✗	✗	✗	✗
Scales with no. of queries	✗	✓	✓	✓	? (100k connections)
Composite queries (AND/OR)	✓	✓	✓	✓	○ (AND In Firestore)
Sorted queries	✓	✓	✓	✗	○ (single attribute)
Limit	✓	✓	✓	✗	✓
Offset	✓	✓	✗	✗	○ (value-based)

USE CASE

How This Makes the Web Faster

Why latency matters

100ms 400ms 500ms

Average: 9,3s

Loading…

% Traffic

% Visitors

% Revenue

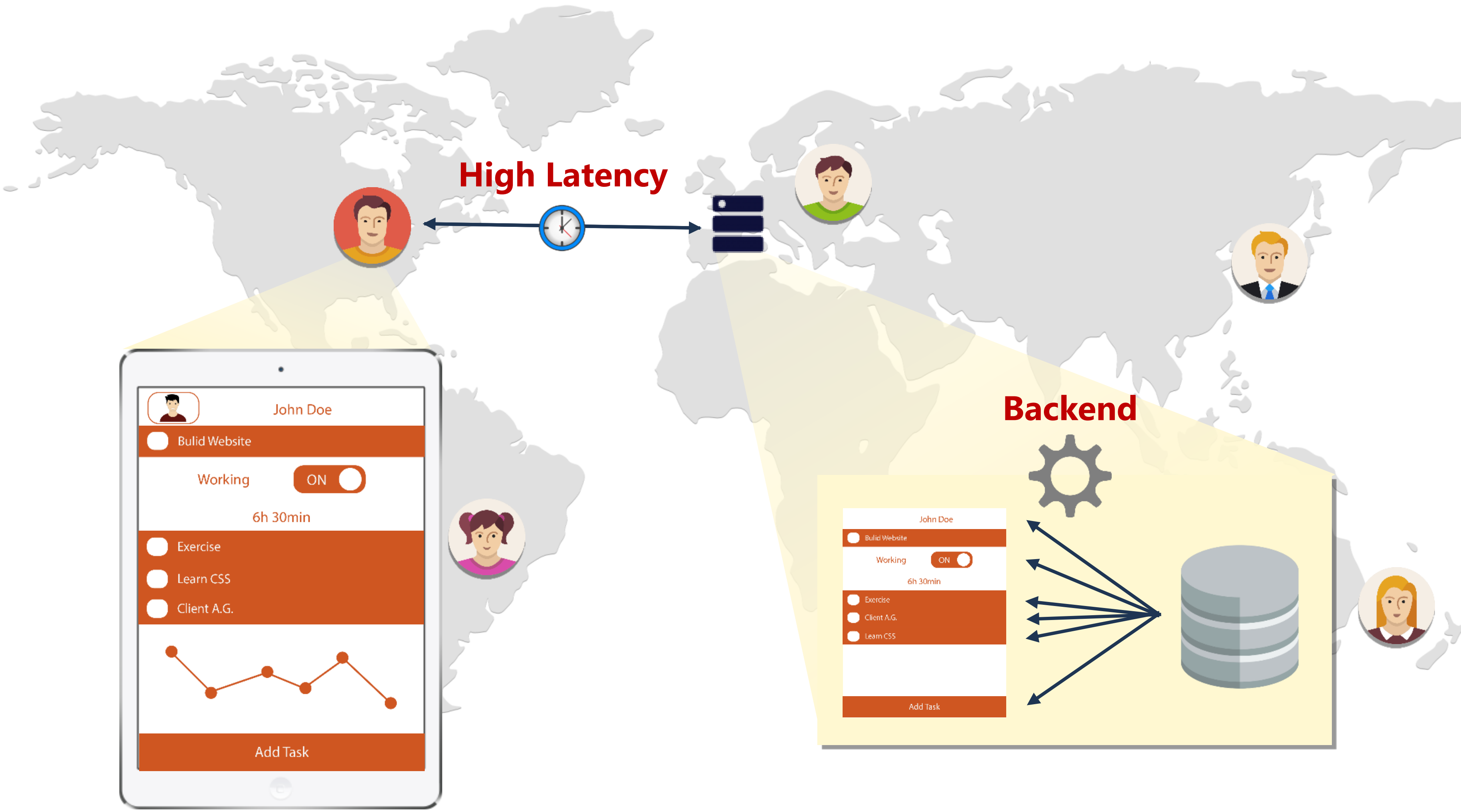
Google

YAHOO!

amazon.com

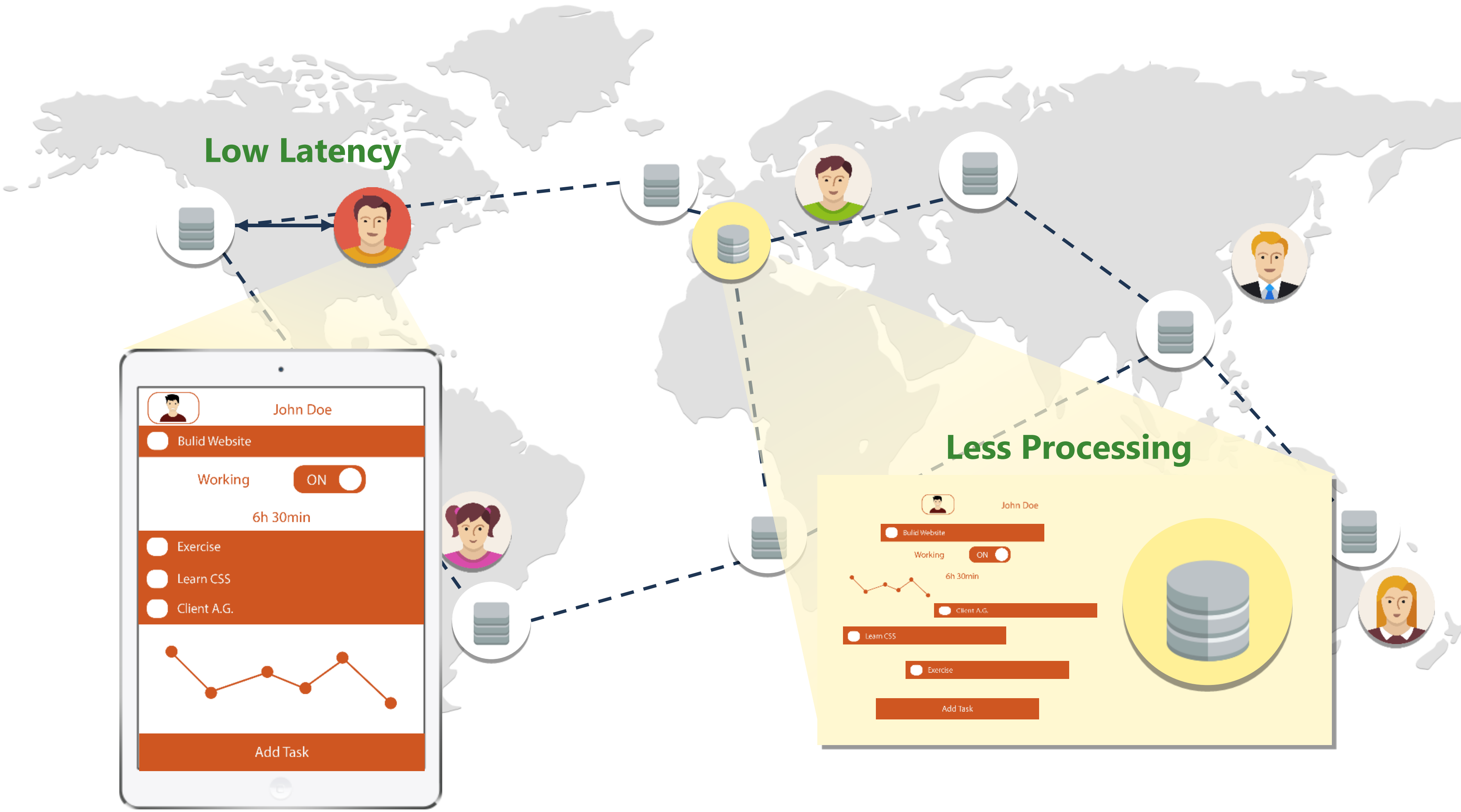
The Problem

Two Bottlenecks: Backend & Latency



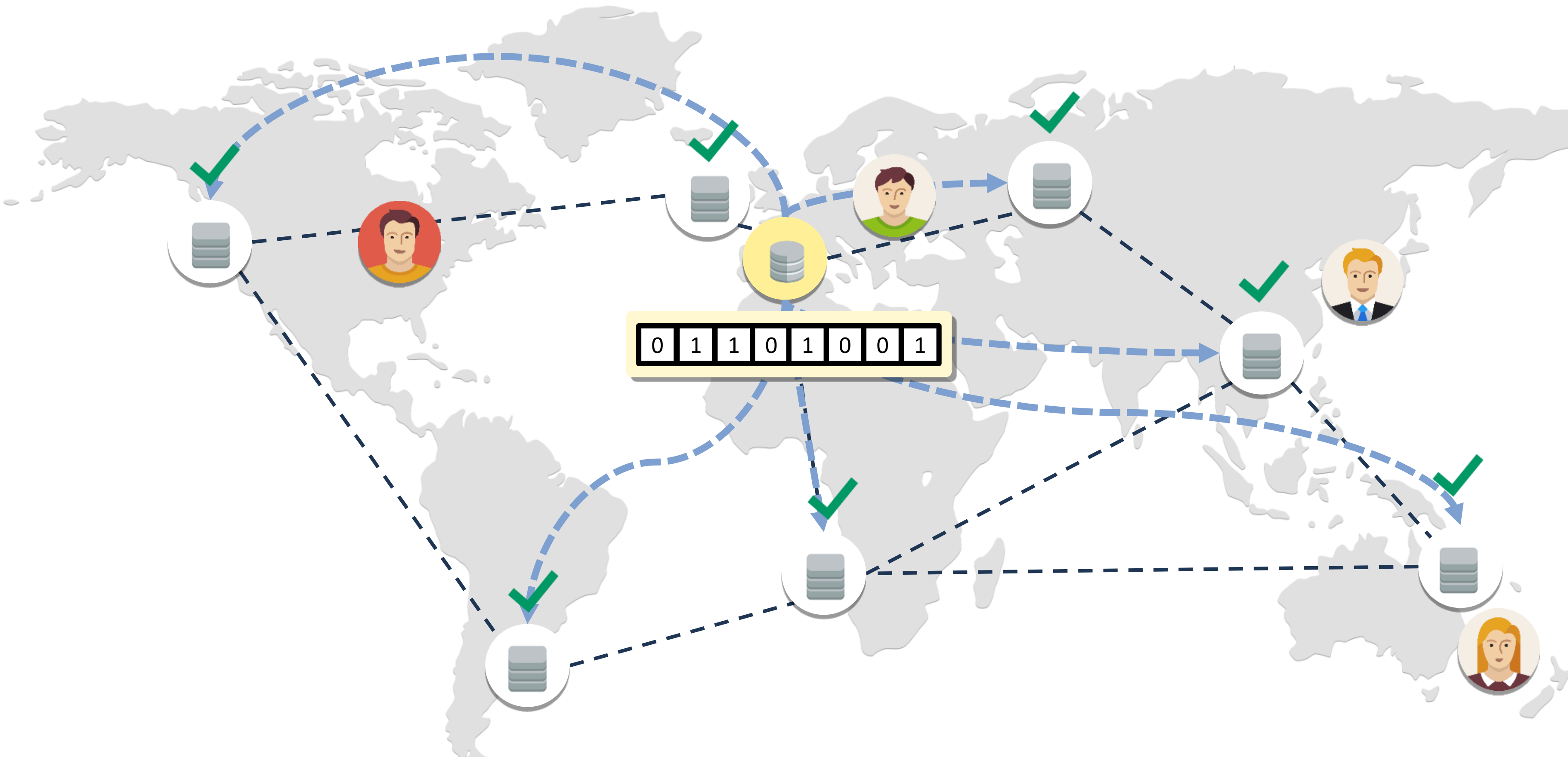
Solution: Global Caching

Fresh Data from Ubiquitous Web Caches



New Caching Algorithms

Solve Consistency Problem



New Caching Algorithms

Solve Consistency Problem

F. Gessert, F. Bücklers, und N. Ritter, „ORESTES: a Scalable Database-as-a-Service Architecture for Low Latency“, in *CloudDB 2014*, 2014.

F. Gessert, S. Friedrich, W. Wingerath, M. Schaarschmidt, und N. Ritter, „Towards a Scalable and Unified REST API for Cloud Data Stores“, in *44. Jahrestagung der GI*, Bd. 232, S. 723–734.

F. Gessert und F. Bücklers, „ORESTES:
auf Cloud-Datenbanken“, in Informatik

ath, S. Friedrich, und N. Ritter, „The
d Caching in the Age of Cloud Data

F. Gessert und F. Bücklers, *Performanz*
vermittels der Web-Caching-Hierarchie

eb-Caching von Datenbankobjekten im

M. Schaarschmidt, F. Gessert, und N. R
Persistence“, in BTW 2015.

t, „Who Watches the Watchmen? On
rking“, in BTW 2015.

S. Friedrich, W. Wingerath, F. Gessert,
Survey“, in *44. Jahrestagung der Gesel*
704.

d-Datenbanken in Forschung und

W. Wingerath, F. Gessert, S. Friedrich, N. Ritter „Real-time stream processing for Big Data“, *Big Data Analytics it - Information Technology*, 2016

F. Gessert, N. Ritter „Scalable Data Management: NoSQL Data Stores in Research and Practice“, *32nd IEEE International Conference on Data Engineering, ICDE*, 2016

F. Gessert, W. Wingerath, S. Friedrich, N. Ritter “NoSQL Database Systems: A Survey and Decision Guidance”, *Computer Science - Research and Development*, 2016

F. Gessert, N. Ritter „Polyglot Persistence“, *Datenbank Spektrum*, 2016.

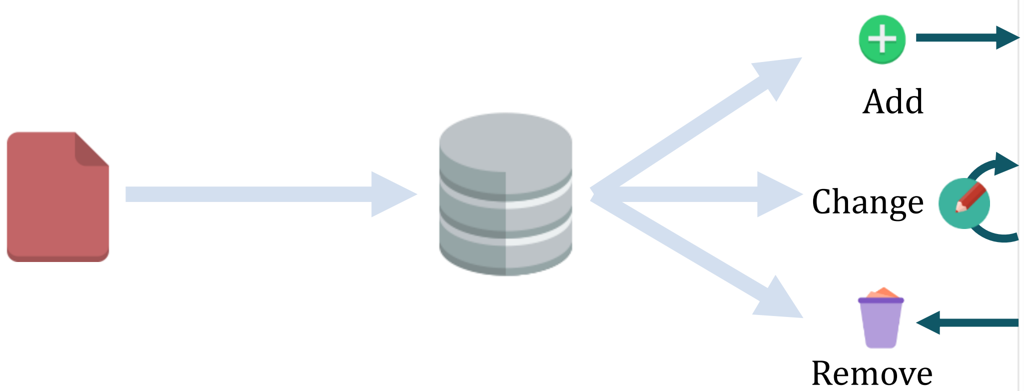
8 Years Research & Development

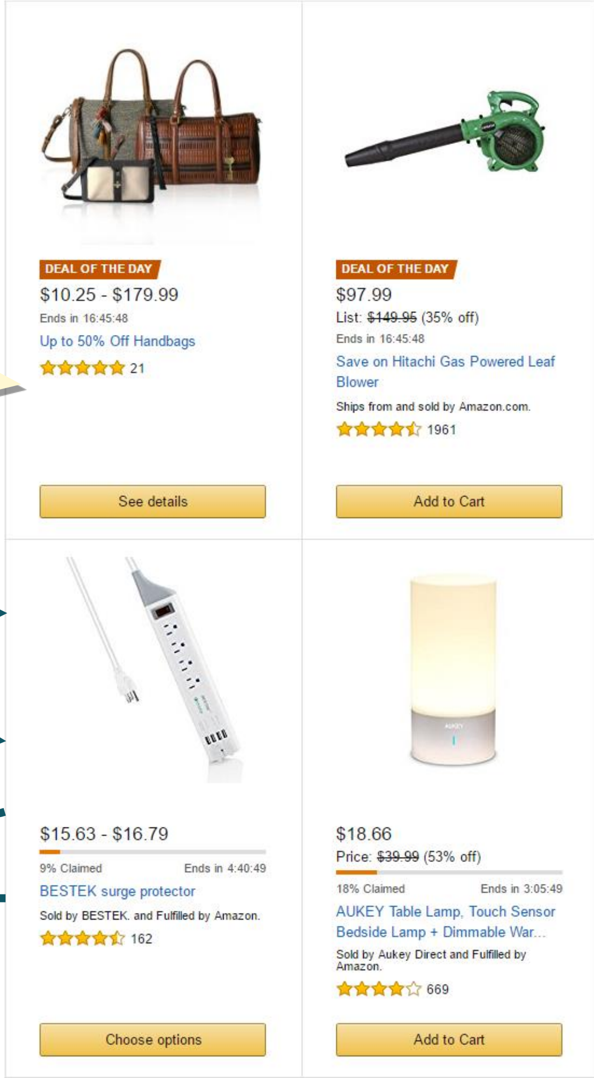
Universität Hamburg

InvaliDB

Invalidating DB Queries

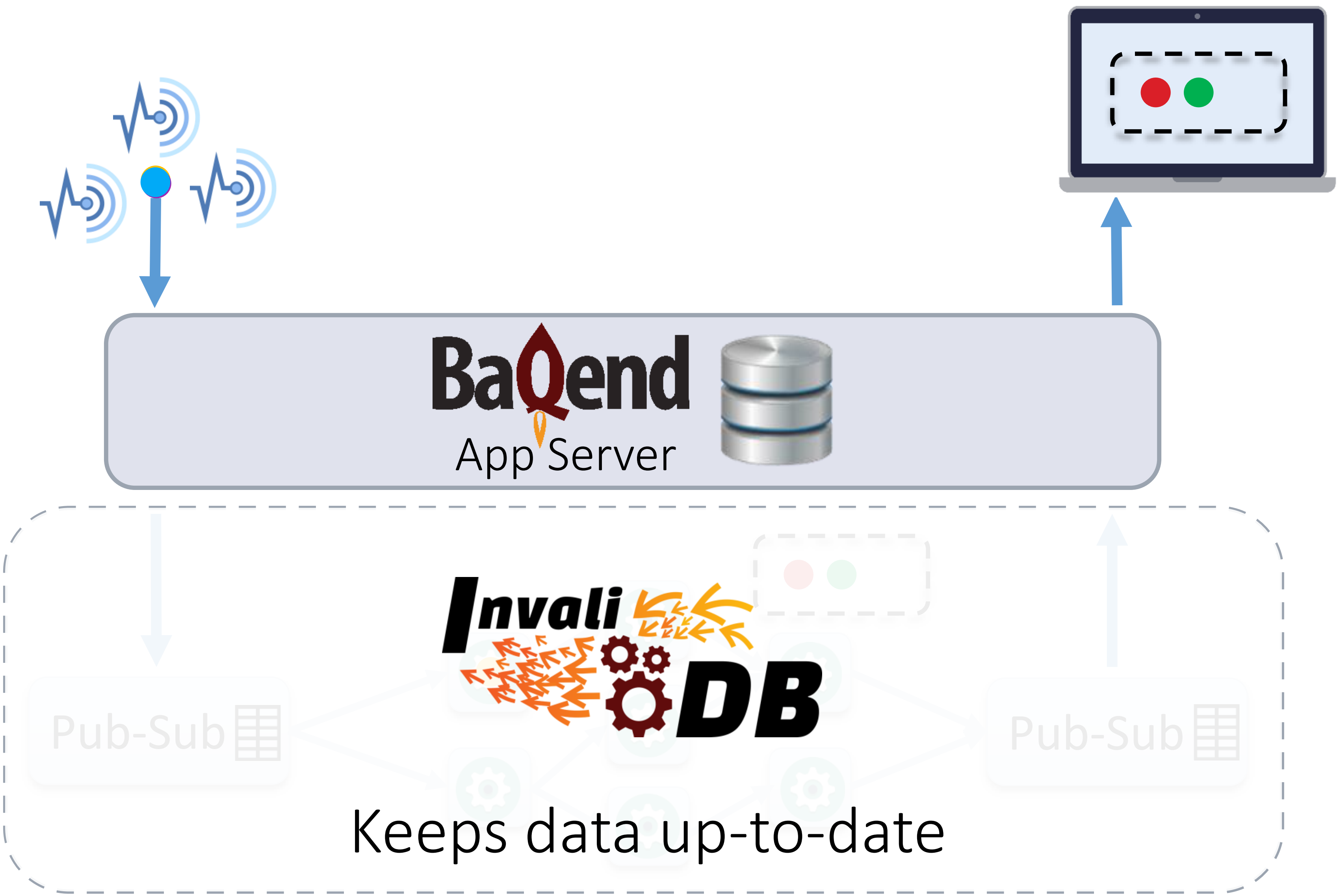
How to **detect changes to query results**:
„Give me the most popular products that are in stock.“





Going Real-Time

Query Caching & Subscribing



InvaliDB

Filter Queries: Distributed Query Matching

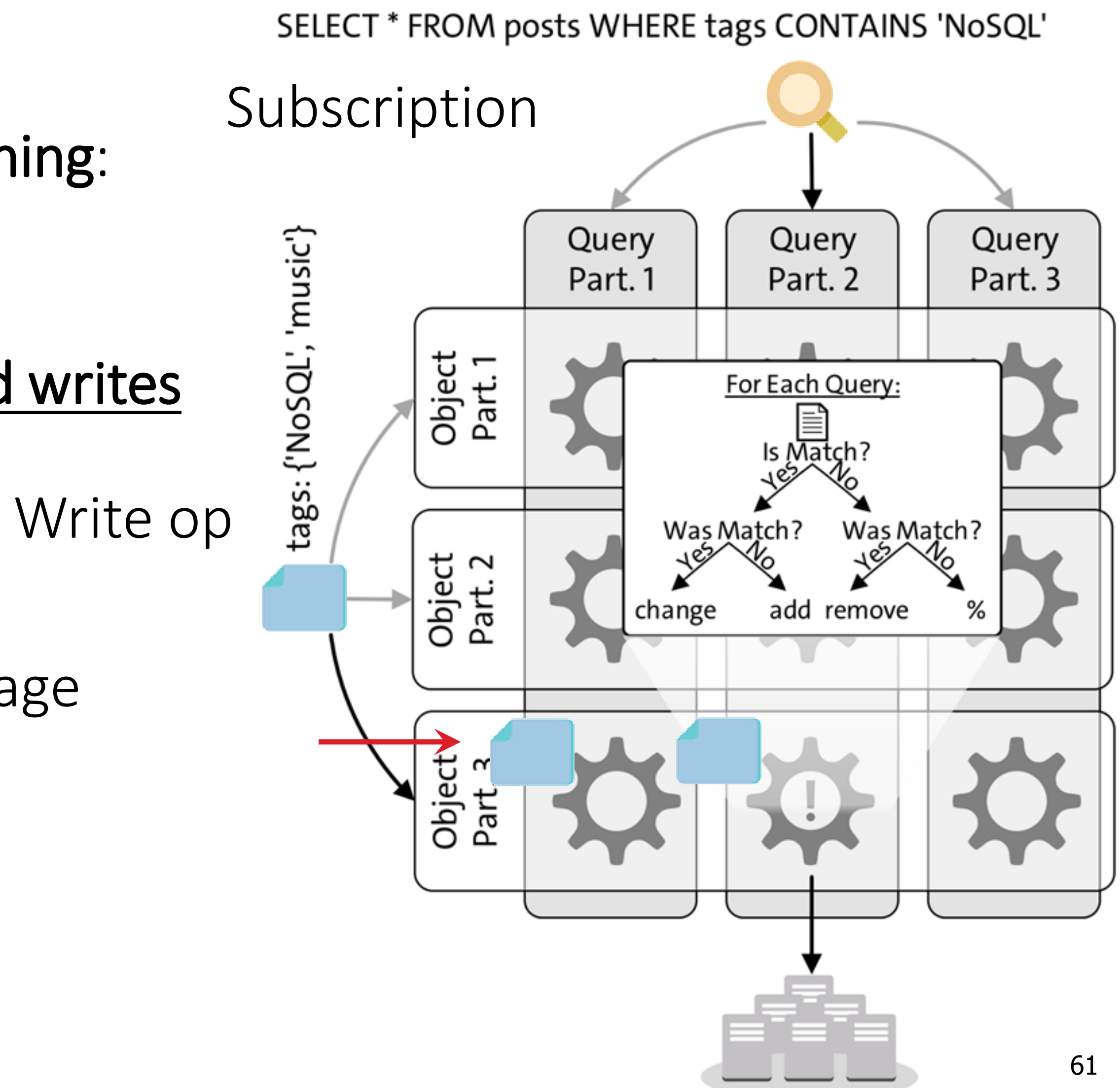
Two-dimensional partitioning:

- *by Query*
- *by Object*

→ **scales with queries and writes**

Implementation:

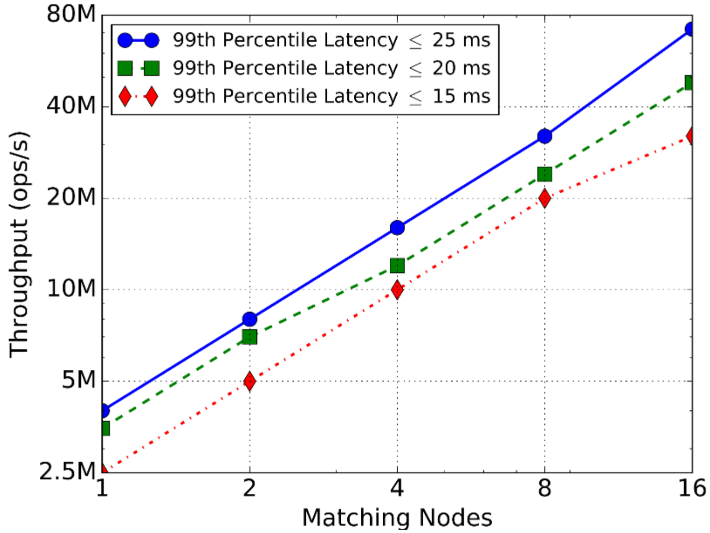
- Apache Storm & Java
- MongoDB query language
- Pluggable engine



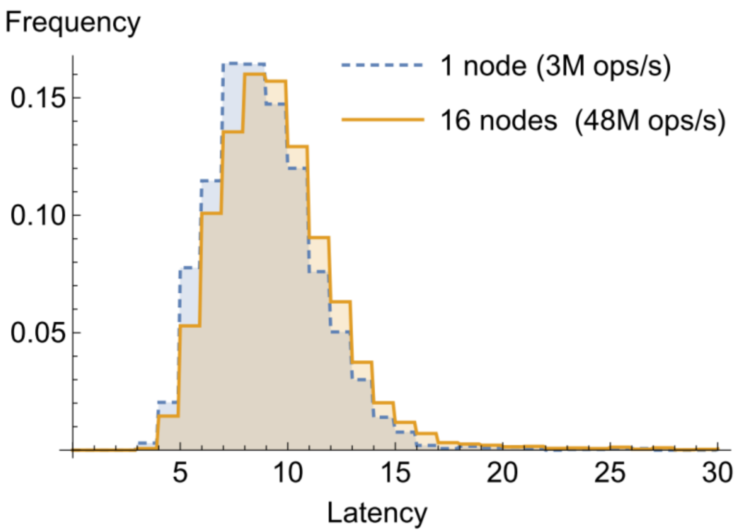
Baqend Real-Time Queries

Low Latency + Linear Scalability

Linear Scalability



Stable Latency Distribution



Quaestor: Query Web Caching for Database-as-a-Service Providers
VLDB '17

Programming Real-Time Queries

JavaScript API

```
var query = DB.Tweet.find()
              .matches('text', /my filter/)
              .descending('createdAt')
              .offset(20)
              .limit(10);
```

Static Query

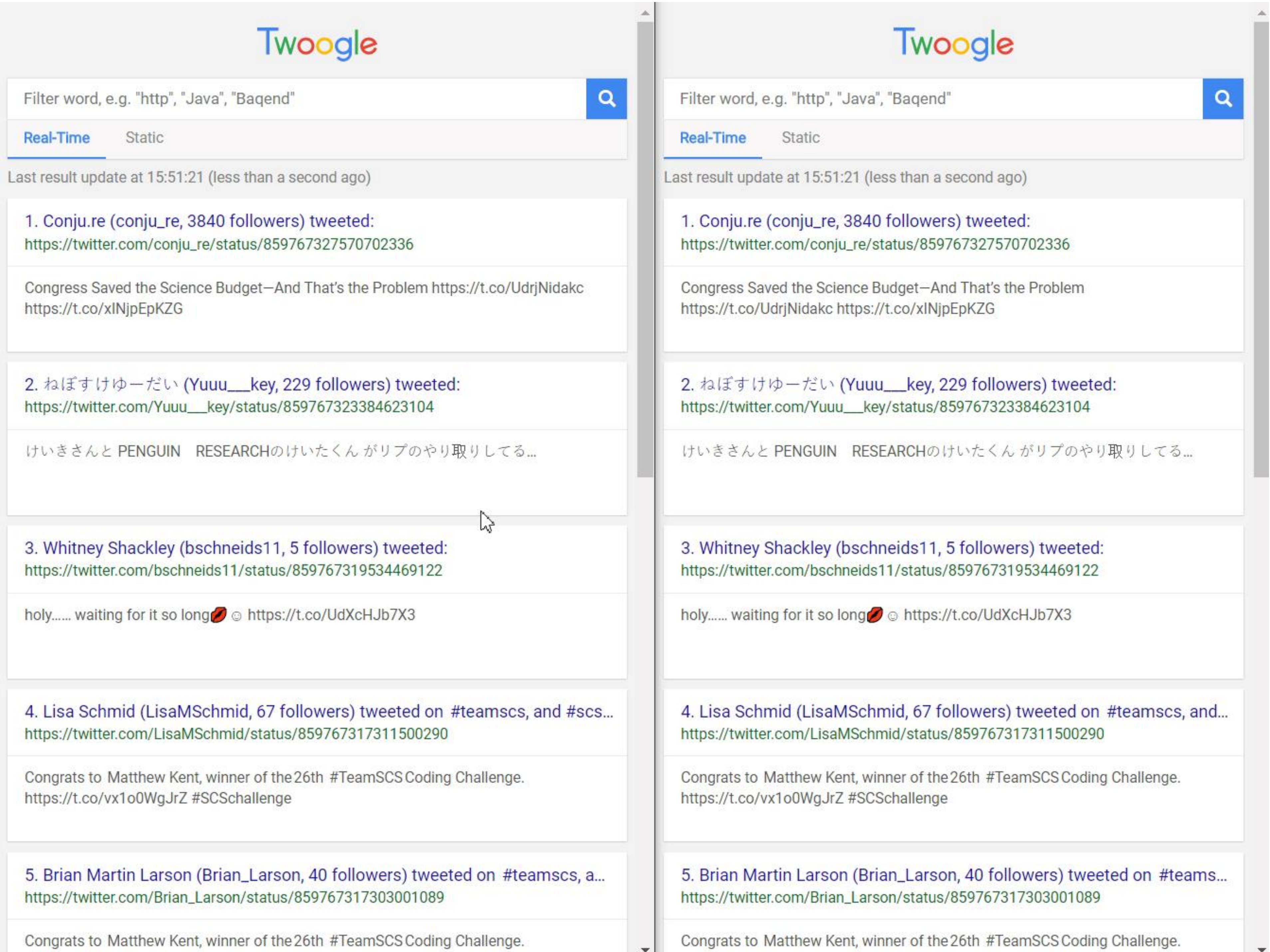
```
query.resultList(result => ...);
```

Google

Real-Time Query

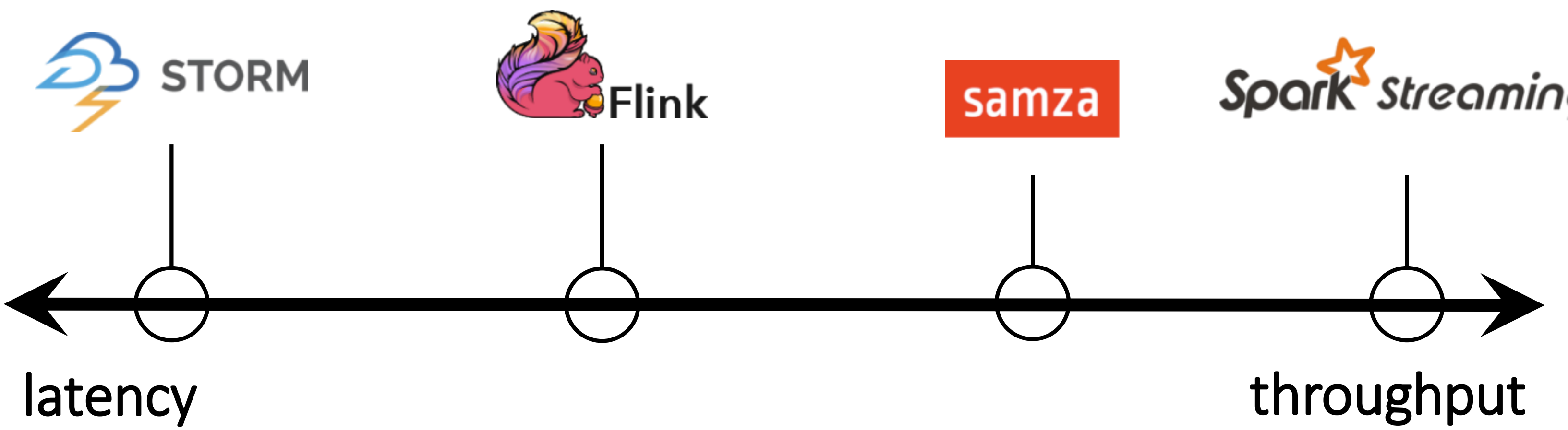
```
query.resultStream(result => ...);
```

Twoogle


Twoogle
Filter word, e.g. "http", "Java", "Baqend"
Real-Time
Static
Last result update at 15:51:21 (less than a second ago)
1. Conju.re (conju_re, 3840 followers) tweeted:
https://twitter.com/conju_re/status/859767327570702336
Congress Saved the Science Budget—And That's the Problem https://t.co/UdrjNidakc https://t.co/xlNjpEpKZG
2. ねぼすけゆーだい (Yuuu__key, 229 followers) tweeted:
https://twitter.com/Yuuu__key/status/859767323384623104
けいきさんと PENGUIN　RESEARCHのけいたくん がリプのやり取りしてる...
3. Whitney Shackley (bschneids11, 5 followers) tweeted:
https://twitter.com/bschneids11/status/859767319534469122
holy...... waiting for it so long💋☺ https://t.co/UdXcHJb7X3
4. Lisa Schmid (LisaMSchmid, 67 followers) tweeted on #teamscs, and #scs...
https://twitter.com/LisaMSchmid/status/859767317311500290
Congrats to Matthew Kent, winner of the 26th #TeamSCS Coding Challenge. https://t.co/vx1o0WgJrZ #SCSchallenge
5. Brian Martin Larson (Brian_Larson, 40 followers) tweeted on #teamscs, a...
https://twitter.com/Brian_Larson/status/859767317303001089
Congrats to Matthew Kent, winner of the 26th #TeamSCS Coding Challenge.
Twoogle
Filter word, e.g. "http", "Java", "Baqend"
Real-Time
Static
Last result update at 15:51:21 (less than a second ago)
1. Conju.re (conju_re, 3840 followers) tweeted:
https://twitter.com/conju_re/status/859767327570702336
Congress Saved the Science Budget—And That's the Problem https://t.co/UdrjNidakc https://t.co/xlNjpEpKZG
2. ねぼすけゆーだい (Yuuu__key, 229 followers) tweeted:
https://twitter.com/Yuuu__key/status/859767323384623104
けいきさんと PENGUIN　RESEARCHのけいたくん がリプのやり取りしてる...
3. Whitney Shackley (bschneids11, 5 followers) tweeted:
https://twitter.com/bschneids11/status/859767319534469122
holy...... waiting for it so long💋☺ https://t.co/UdXcHJb7X3
4. Lisa Schmid (LisaMSchmid, 67 followers) tweeted on #teamscs, and...
https://twitter.com/LisaMSchmid/status/859767317311500290
Congrats to Matthew Kent, winner of the 26th #TeamSCS Coding Challenge. https://t.co/vx1o0WgJrZ #SCSchallenge
5. Brian Martin Larson (Brian_Larson, 40 followers) tweeted on #teams...
https://twitter.com/Brian_Larson/status/859767317303001089
Congrats to Matthew Kent, winner of the 26th #TeamSCS Coding Challenge.

Summary

- **Stream Processors:**



- **Real-Time Databases** integerate Storage & Streaming
- **Learn more**: slides.baqend.com

BaQend
ww@baqend.com
@baqendcom