Building an Event Analytics Pipeline with Kafka, ksqlDB, and Druid

Hellmar Becker, Senior Sales Engineer

About Me

Hellmar Becker
Sr. Sales Engineer at Imply
Lives near Munich

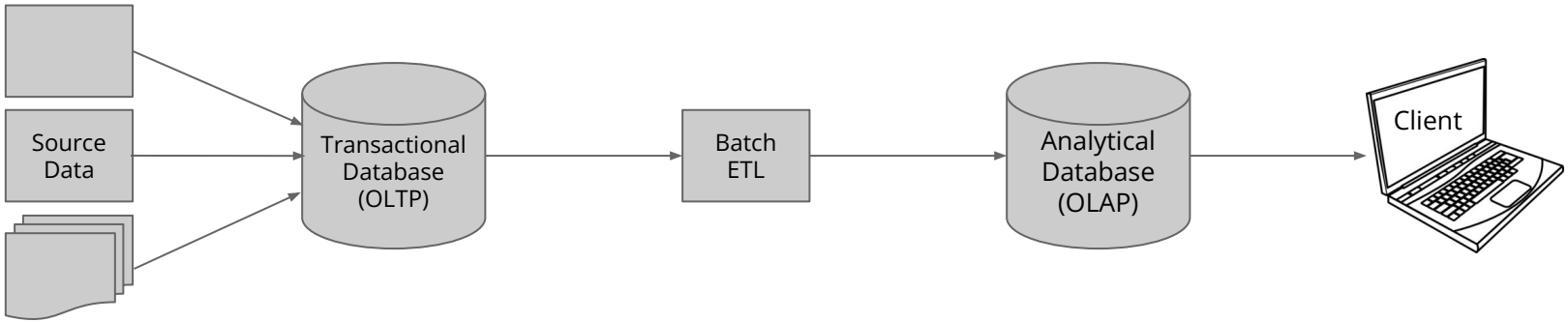
hellmar.becker@imply.io
https://www.linkedin.com/in/hellmarbecker/
https://blog.hellmar-becker.de/

Agenda

- The Case for Streaming Analytics
- How to Prepare Your Data: Streaming ETL
- How to Analyze Your Data: Streaming Analytics
- Apache Druid - A Streaming Analytics Database
- K2D - A Streaming Analytics Architecture
- Live Demo
- Q&A

The Case for Streaming Analytics

- Analytics - "the process of discovering, interpreting, and communicating significant patterns in data."
- OLAP = Online Analytical Processing
- Classical:



But that's not enough anymore!

The Case for Streaming Analytics (contd.)

- The Big Data Hype gave us the *Lambda Architecture*
- Separate paths for batch and realtime
- One common serving layer
- Complex, hard to reconcile

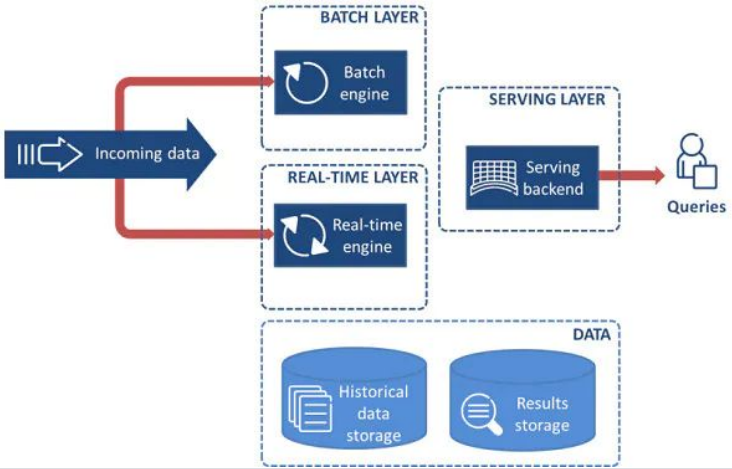


Image source: https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa

The Case for Streaming Analytics (contd.)

- 2014 Jay Krepps: Kappa Architecture
- Avoids having separate code paths for batch and streaming

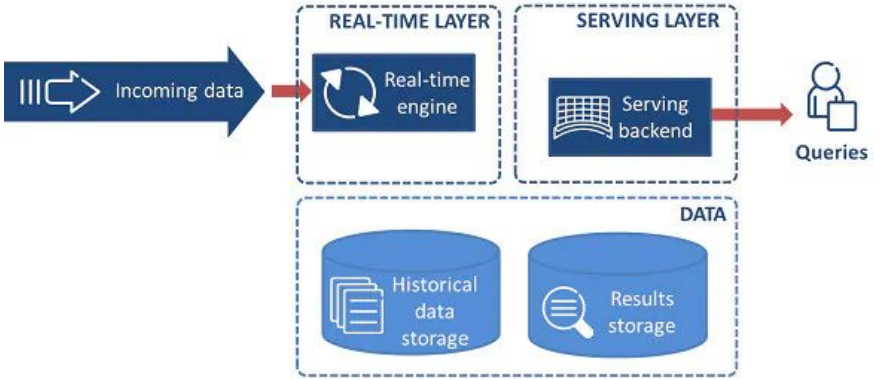


Image source: https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa

How to prepare your Data: Streaming ETL

ETL = Extract, Transform, Load

Let's focus on the Transform part

Simple Event Processing = 1 event at a time

- Filter
- Transform
- Cleanse

Complex Event Processing = Relate events to each other

- Windowing
- Aggregations
- Joins
- Enrichment

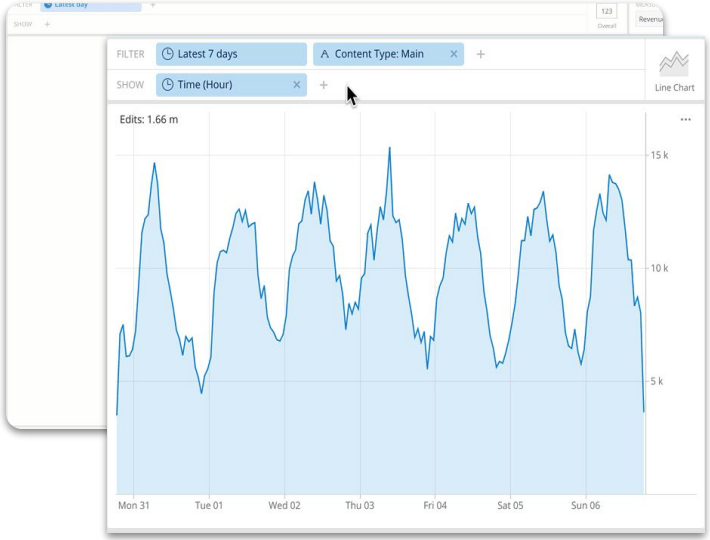
ksqlDB is a tool by Confluent that does streaming ETL using *streaming SQL*

How to analyze your data: Streaming Analytics with Druid

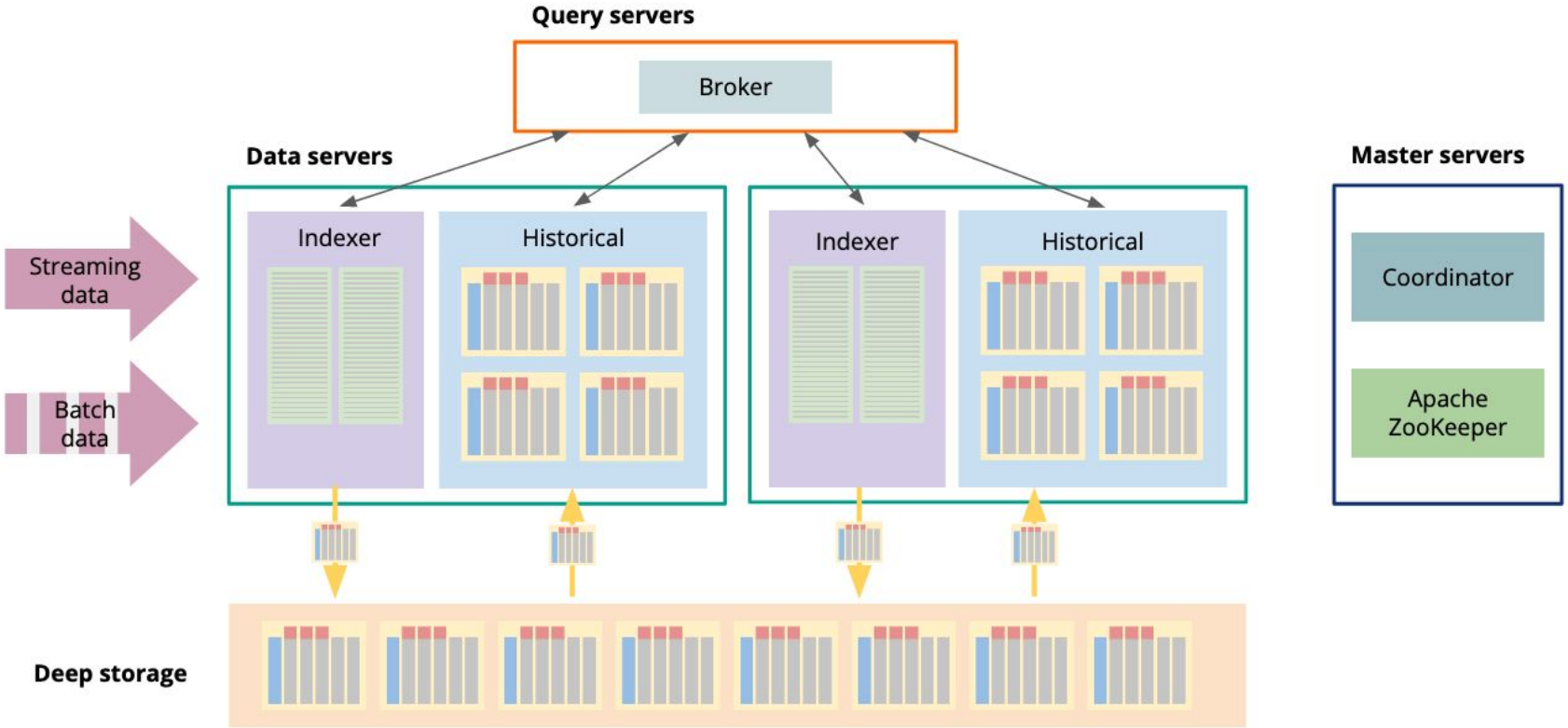
For analytics applications that require:

1. **Sub-second queries at any scale**
   Interactive analytics on TB-PBs of data
2. **High concurrency at the lowest cost**
   100s to 1000s QPS via a highly efficient engine
3. **Real-time and historical insights**
   True stream ingestion for Kafka and Kinesis

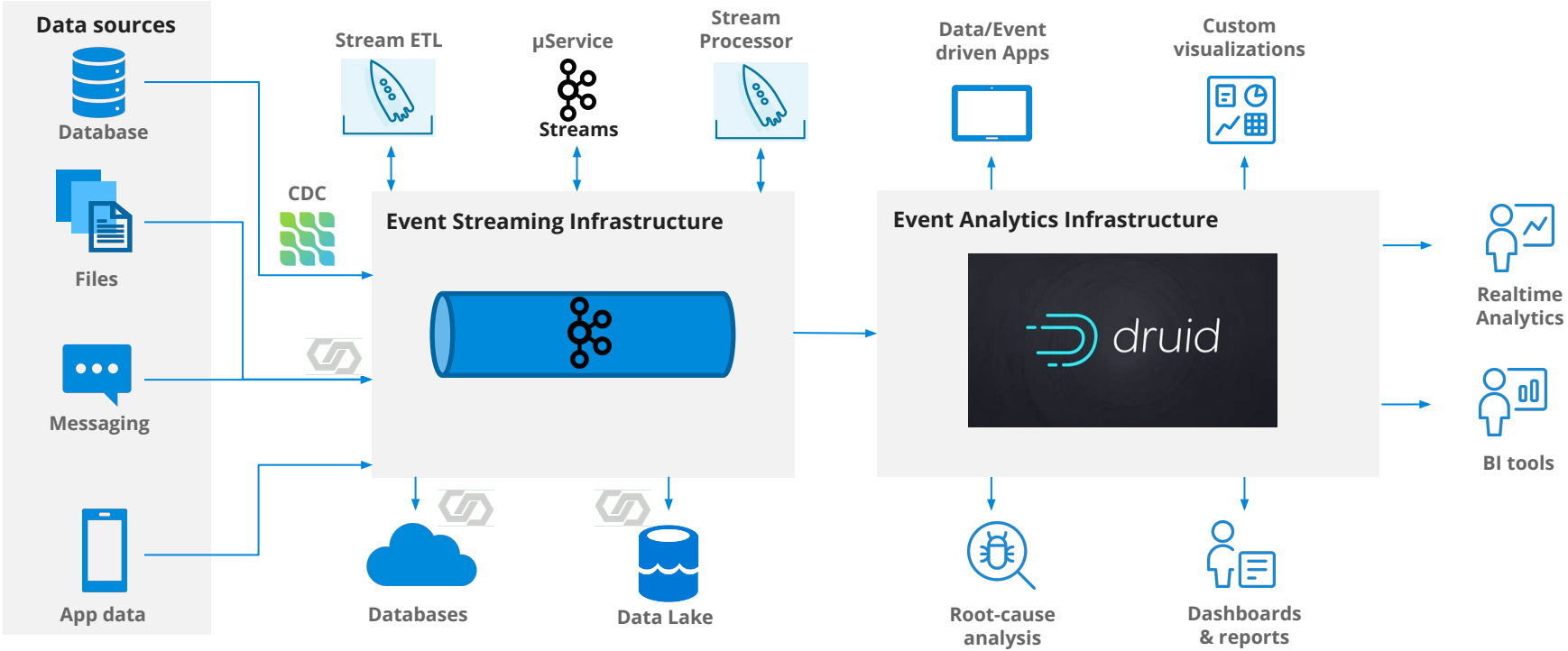
★ Plus, **non-stop reliability** with automated fault tolerance and continuous backup



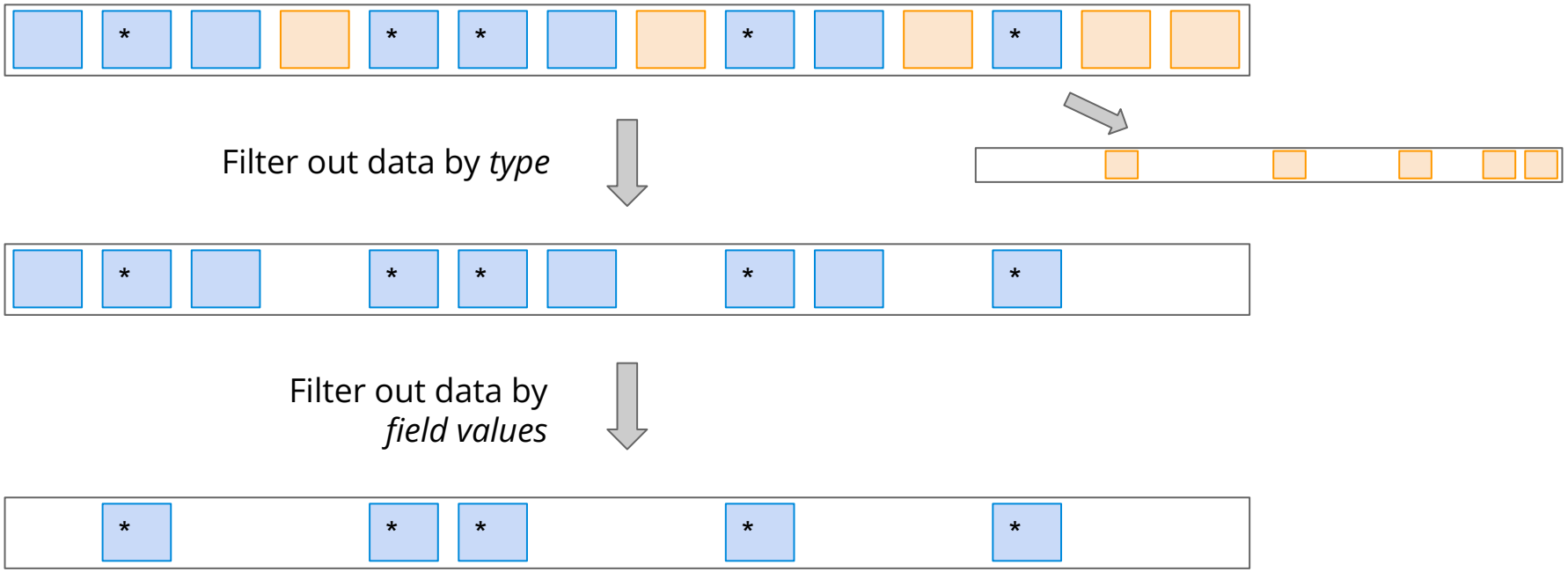
Why do you need a Streaming Analytics Database?



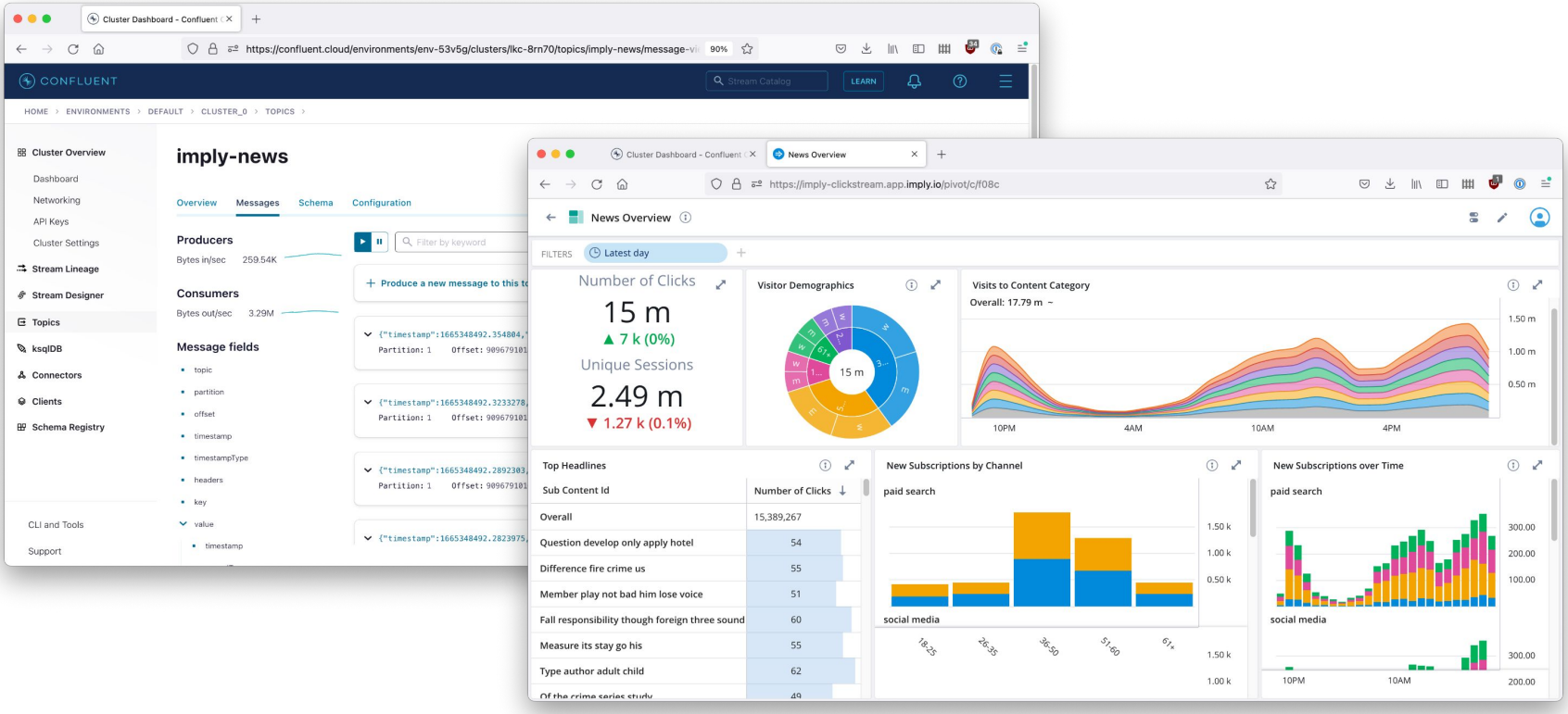
K2D Architecture - Kafka to Druid



Preprocessing - What we are going to do today



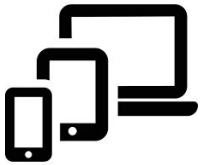
Use Case: Publisher Clickstream Data

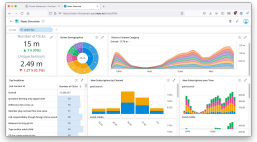


Demo Architecture

Data Production	Delivery	Processing	Storage	Query	Visualisation

Analytics Pipeline





Tracking, Transactions	**Kafka** as an event streaming platform	Preprocessing with **ksqlDB**: - Filter - Enrich - Transform	**Apache Druid** - Elastic storage model (in production backed by cloud storage) Highly scalable, built in DR	**Apache Druid** Analytical queries against realtime, detail data	**Imply Pivot** - a data exploration and adhoc analytics GUI for Druid

imply

Live Demo

Learnings

- **Kafka** and **Druid** complement each other
- Use **ksqlDB** for
  - Preprocessing
  - Enrichment
  - Materialized views
- Use **Druid** for
  - Scalable analytical applications
  - Adhoc data exploration
  - OLAP style analysis
- Integration is easy with native integration APIs

Questions

hellmar.becker@imply.io
https://www.linkedin.com/in/hellmarbecker/
https://blog.hellmar-becker.de/