

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

Hellmar Becker, Senior Sales Engineer



About Me



Hellmar Becker Sr. Sales Engineer at Imply Lives near Munich



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

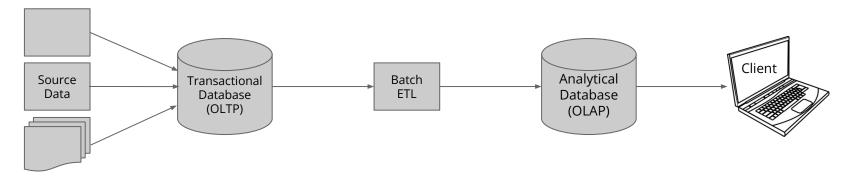


Agenda

- 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:

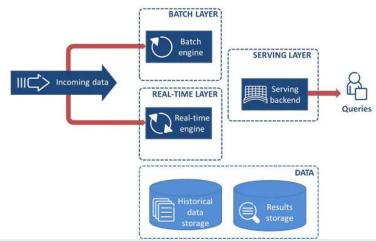


But that's not enough anymore!



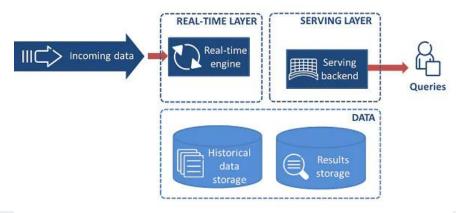
The Case for Streaming Analytics (contd.)

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



The Case for Streaming Analytics (contd.)

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



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

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

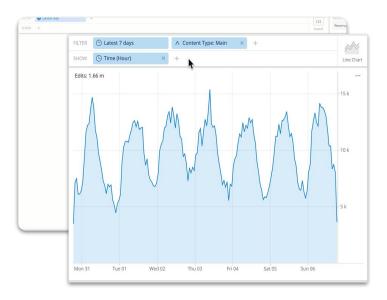
How to analyze your data: Streaming Analytics with Druid

For analytics applications that require:

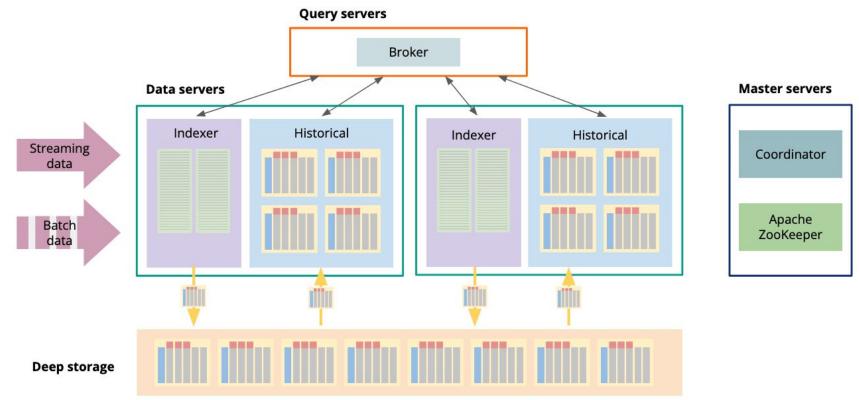
- **1 Sub-second queries at any scale** Interactive analytics on TB-PBs of data
- 2 High concurrency at the lowest cost 100s to 1000s QPS via a highly efficient engine

3 Real-time and historical insights True stream ingestion for Kafka and Kinesis

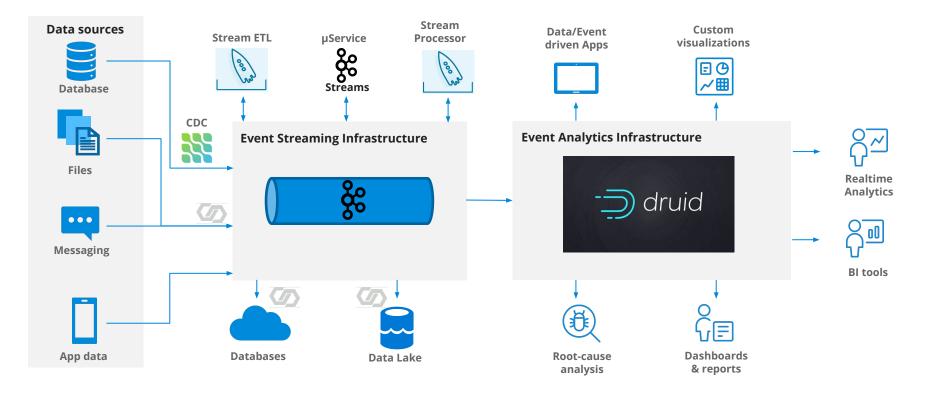




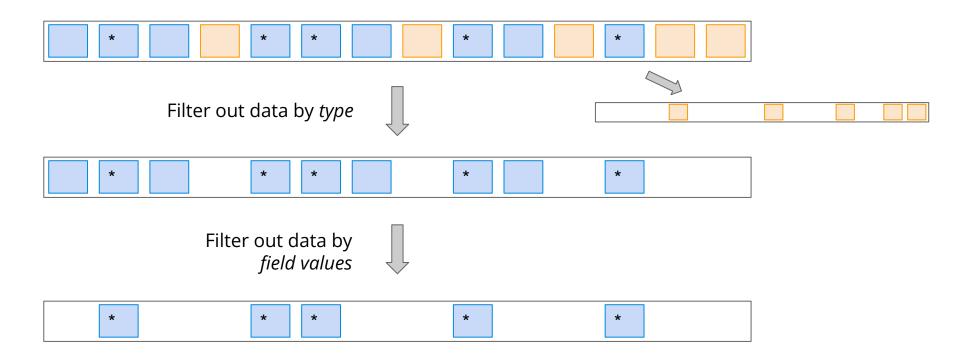
Why do you need a Streaming Analytics Database?



K2D Architecture - Kafka to Druid



Preprocessing - What we are going to do today

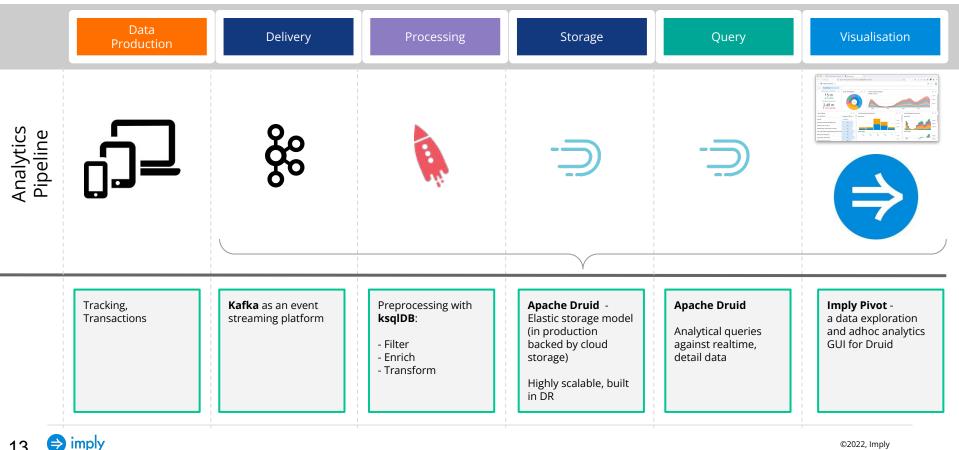




Use Case: Publisher Clickstream Data

	shboard - Confluent 🛛 +				2 III\ 🗉 🗰	M					
$\leftarrow \rightarrow \mathbb{C} \ \textcircled{a}$	O A and https://confluent.cl	oud/environments/env-53v5g/clusters/lkc-	8rn70/topics/imply-news/message-vic 90% 9	am Catalog							
	DEFAULT > CLUSTER_0 > TOPICS >			,							
B Cluster Overview	imply-news		Cluster Dashboard - Confluent : X News Overview X +								
Dashboard Networking	Overview Messages Schema Configuration		← → O A ≈ [±] https://imply-clickstream.app.imply.io/plvot/c/f08c ← ■ News Overview ①					☆ ♡ Ł M III ## ♥ ◎ ≅			
API Keys Cluster Settings	Producers Bytes in/sec 259.54K	► II Q. Filter by keyword	FILTERS (S Latest day +							a	5 / (2)
⇒ Stream Lineage	Consumers	+ Produce a new message to this to	-	Visitor Demographics	(i) 🖍	Visits to Content Catego Overall: 17.79 m ~	ory				() 🖍
Topics ksqlDB	Bytes out/sec 3.29M Message fields • topic • partition • offset • timestamo	 ("timestamp":1665348492.354804, Partition: 1 0ffset:99979101 ("timestamp":1665348492.323278, Partition: 1 0ffset:990679101 	15 m ▲ 7 k (0%)	3 3 2	*	-					1.50 m
& Connectors			Unique Sessions								
 Glients 			2.49 m ▼ 1.27 k (0.1%)	er of	10PM		4AM		10AM 4PM		
	 timestampType 	 {"timestamp":1665348492.2892303, 	Top Headlines	1	New Subscriptio	ns by Channel		() 🖍	New Subscriptio	ns over Time	1 2
	 headers key 	Partition: 1 Offset: 909679101	Sub Content Id Overall	Number of Clicks ↓	paid search				paid search		.
CLI and Tools Support	valuetimestamp	{"timestamp":1665348492.2823975,	Question develop only apply hotel	54				1.50 k	h	at the	200.00
			Difference fire crime us Member play not bad him lose voice	55				0.50 k	.	.,/////////////////////////////////////	100.00
			Fall responsibility though foreign three sound	60 55	social media	÷ *	51 ₁₀ 67.		social media		
			Measure its stay go his Type author adult child	62	18-13	Construction of the second	57, 67 _×	1.50 k 1.00 k	10PM	10AM	200.00
			Of the crime series study	49							

Demo Architecture



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/