



Building a Real-Time Data Pipeline: Apache Kafka at LinkedIn

Hadoop Summit 2013

Joel Koshy

June 2013

Network update stream



Amit Kumar via Twitter 3h
akumar On the hunt for a stellar product manager for @lily (rare, such fellas). Recommendations? There's a \$10k referral bonus you know.
Favorite 13 Retweet 4 Reply · 1 day ago

News currently trending on LinkedIn Today

Not on LinkedIn? Here's How to Start. [Infographic] [wspnews.com](#)
It may sound trite, but if you want to be hired in 2012, you are going to need a LinkedIn profile. The website tops all other social networks when it comes to finding jobs, and adoption of the site is still growing. The site...

Like (2) · Comment (2) · Share · 1 day ago

Nathan Pinno, rakesh gawani and 30 others like this

Show 7 more comments

Rebecca Gettys I got to the point where I considered switching my actual resume for my linkedin profile
22 hours ago

sunder singh Get closer to Africa every 60 days [technogreedy.com](#) Enjoy a whopping 30% off on 18 issues by simply renewing your subscription every year. If you subscribe more than 10 copies of an issue, an additional discount of 10% shall be given *Taxi Free to contact us at subscription@thetimesofafrica.com
18 hours ago

Add a comment...

Paul Tarjan via Twitter 3h
ptarjan I posted 4 photos on Facebook in the album "The beard comes off"
<http://it.com/82c9r7p>
The beard comes off | Facebook [facebook.com](#)
Facebook is a social utility that connects people with friends and others who work, study and live around them. People use Facebook to keep up with friends, upload an unlimited number of photos, post links and videos, and learn...
Favorite 13 Retweet 4 Reply · 1 day ago

LinkedIn "If you want to be hired in 2012, you are going to need a LinkedIn profile."

Not on LinkedIn? Here's How to Start. [Infographic] [wspnews.com](#)
It may sound trite, but if you want to be hired in 2012, you are going to need a LinkedIn profile. The website tops all other social networks when it comes to finding jobs, and adoption of the site is still growing. The site...

Like (2) · Comment (2) · Share · 1 day ago

rakesh gawani, Travis Howard and 88 others like this

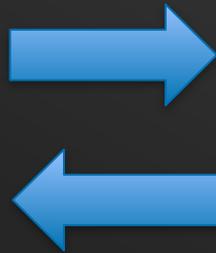
Show 10 more comments

Turning You to Trust
15 hours ago

We have a lot of data.

We want to leverage this data to build products.

Data pipeline



People you may know

People You May Know ^{beta}

See people from different parts of your professional life

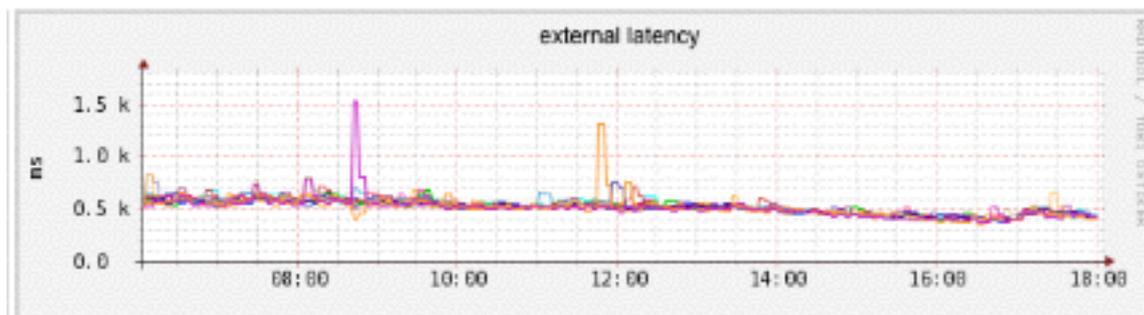
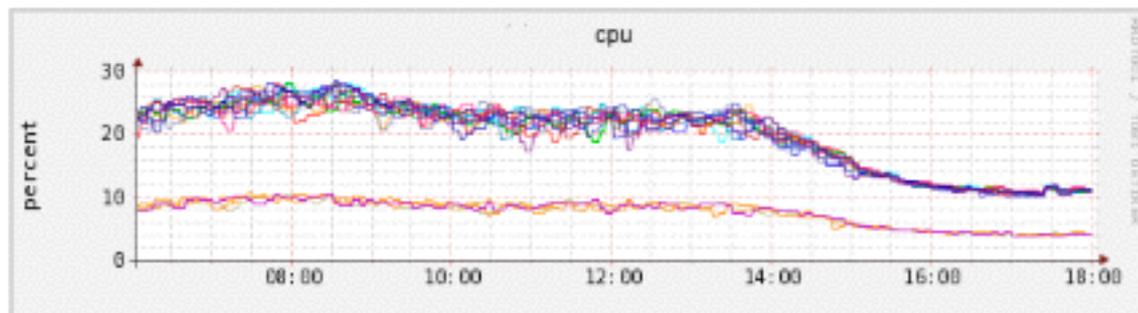
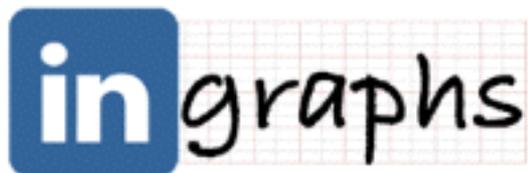


All Suggestions / LinkedIn (12)

[Connect All](#)

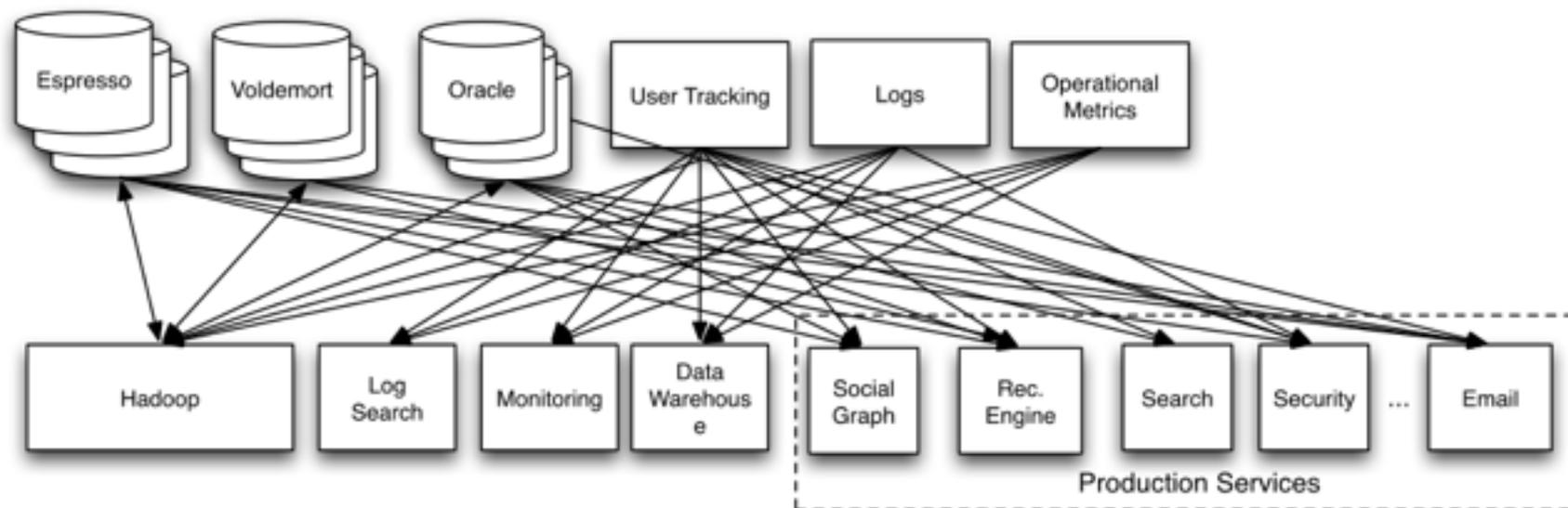
 <p>Brad Mauney ^{2nd} Senior Product Manager, Search & Social Graph at LinkedIn Mountain View, California</p> <p>Connect  5 shared connections</p>	 <p>Albert Wang ^{2nd} Senior User Experience Designer at LinkedIn Mountain View, California</p> <p>Connect  127 shared connections</p>
 <p>Sam Shah ^{2nd} Principal Engineer at LinkedIn Mountain View, California</p> <p>Connect  22 shared connections</p>	 <p>Tan Nhu ^{2nd} Senior Web Developer at LinkedIn Mountain View, California</p> <p>Connect  16 shared connections</p>
 <p>Vinodh Jayaram ^{2nd} Software Engineering Manager at LinkedIn Mountain View, California</p> <p>Connect  10 shared connections</p>	 <p>Andy Chen ^{2nd} Software Engineer at LinkedIn Mountain View, California</p> <p>Connect  78 shared connections</p>

System and application metrics/logging

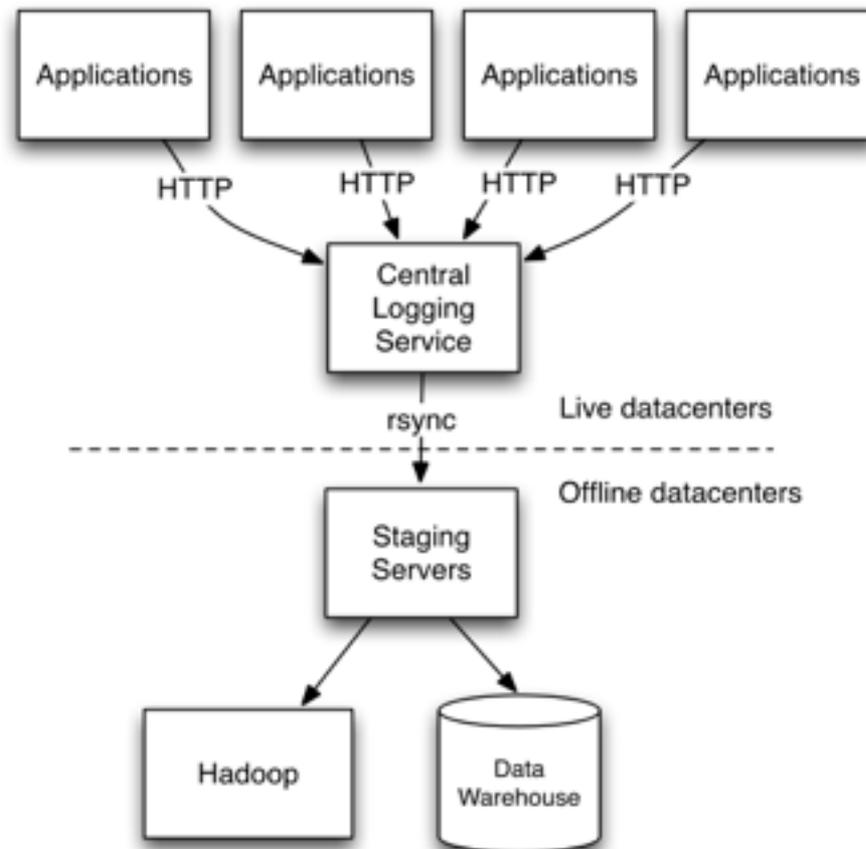


How do we integrate this variety of data
and make it available to all these systems?

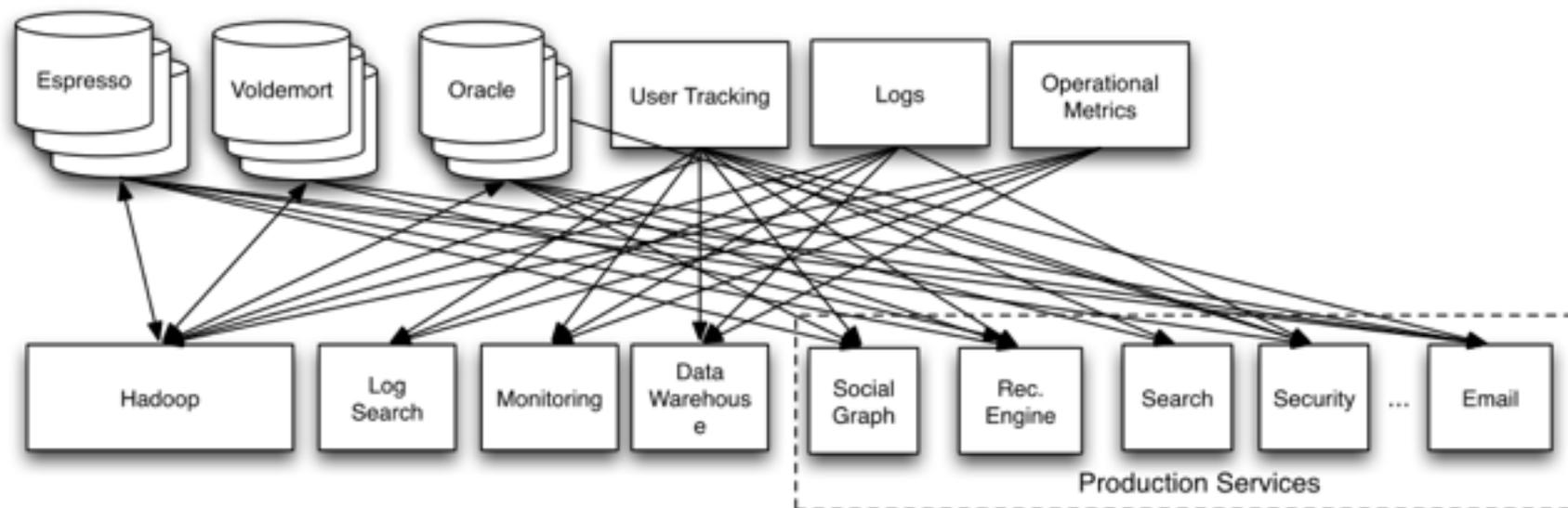
Point-to-point pipelines



LinkedIn's user activity data pipeline (circa 2010)



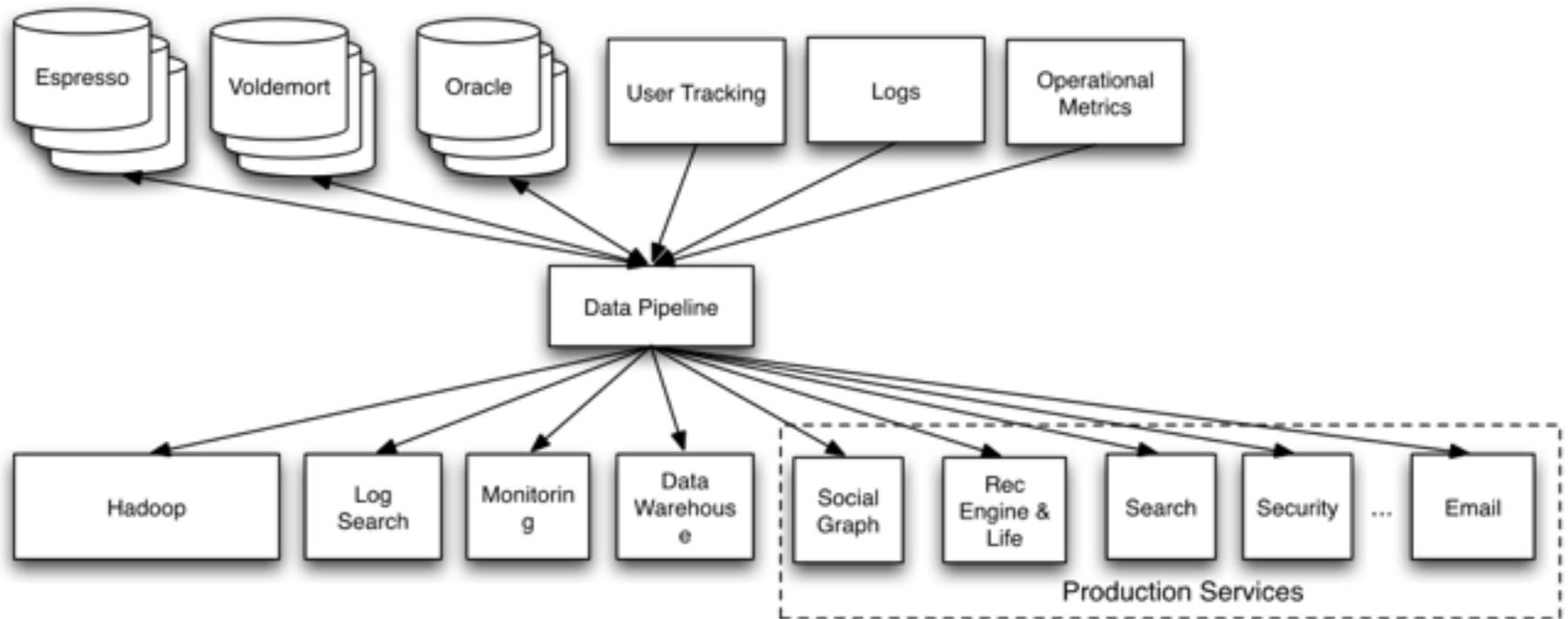
Point-to-point pipelines



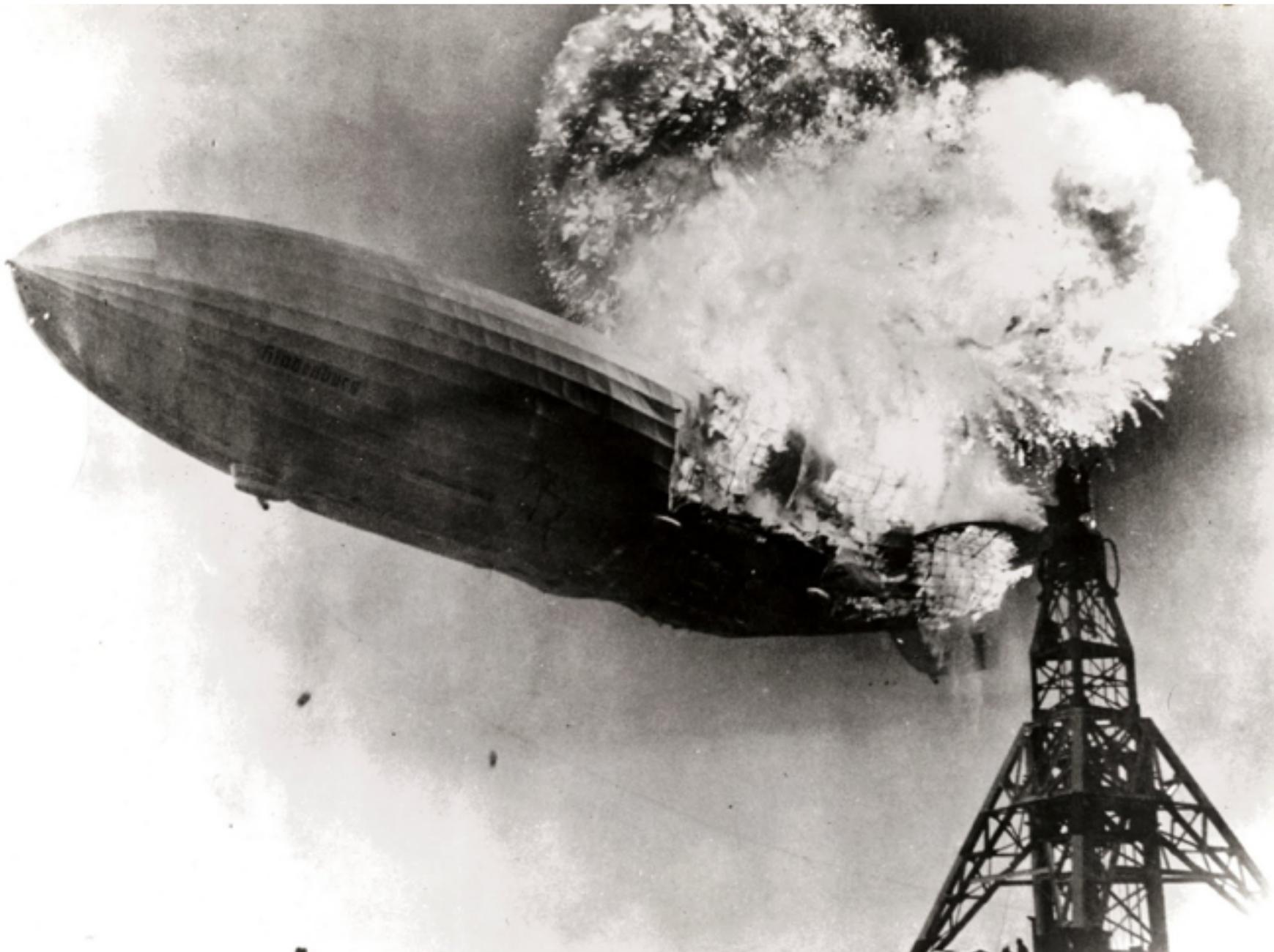
Four key ideas

1. Central data pipeline
2. Push data cleanliness upstream
3. O(1) ETL
4. Evidence-based correctness

Central data pipeline



First attempt: don't re-invent the wheel



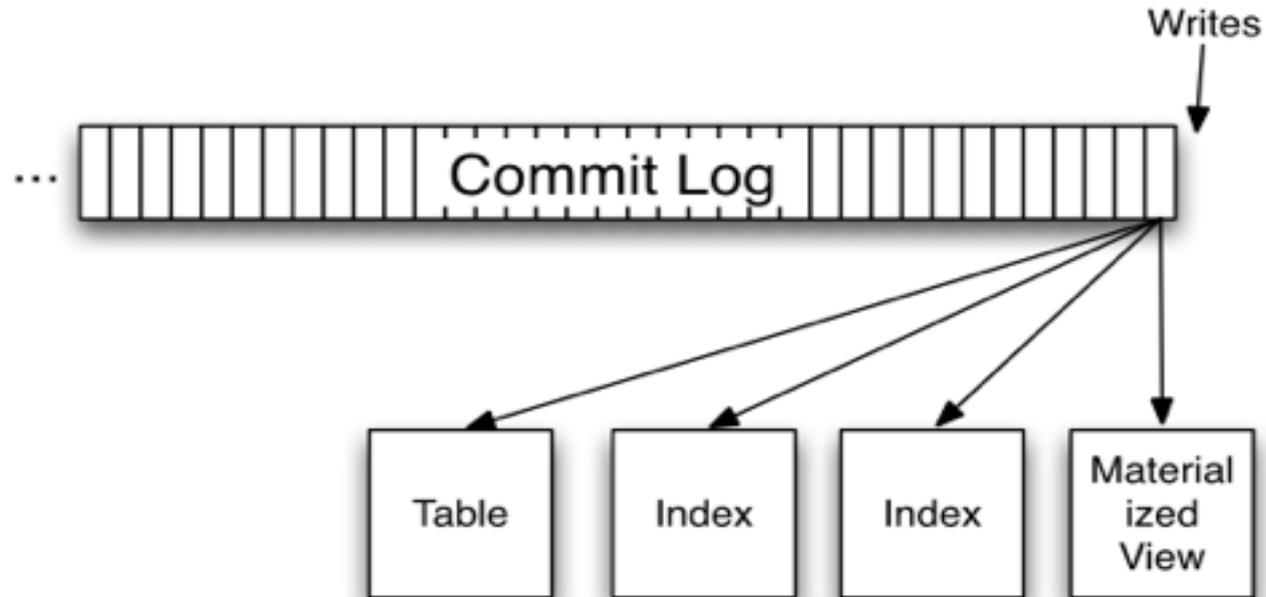
HADOOP SUMMIT 2013



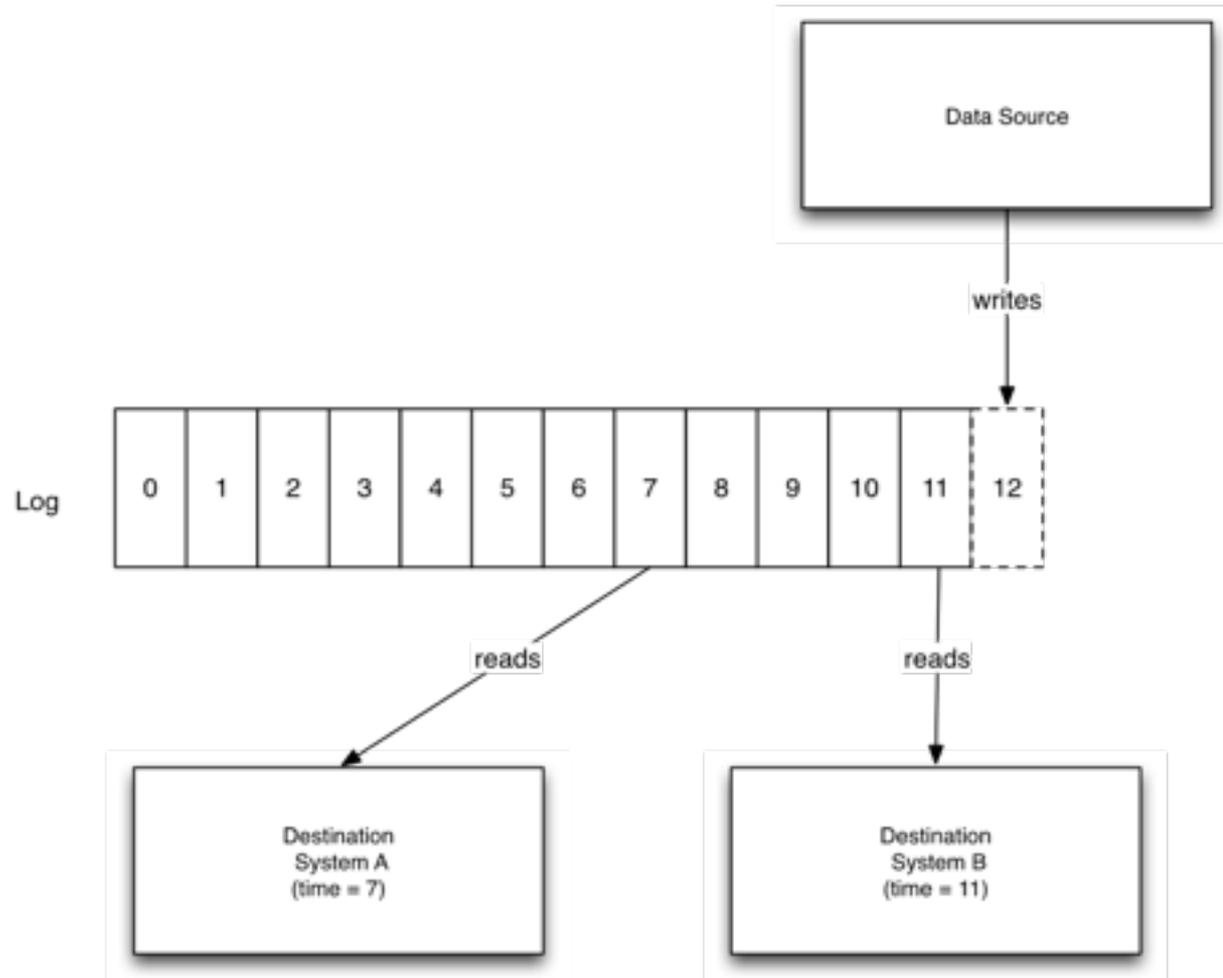
Second attempt: re-invent the wheel!

Use a central commit log

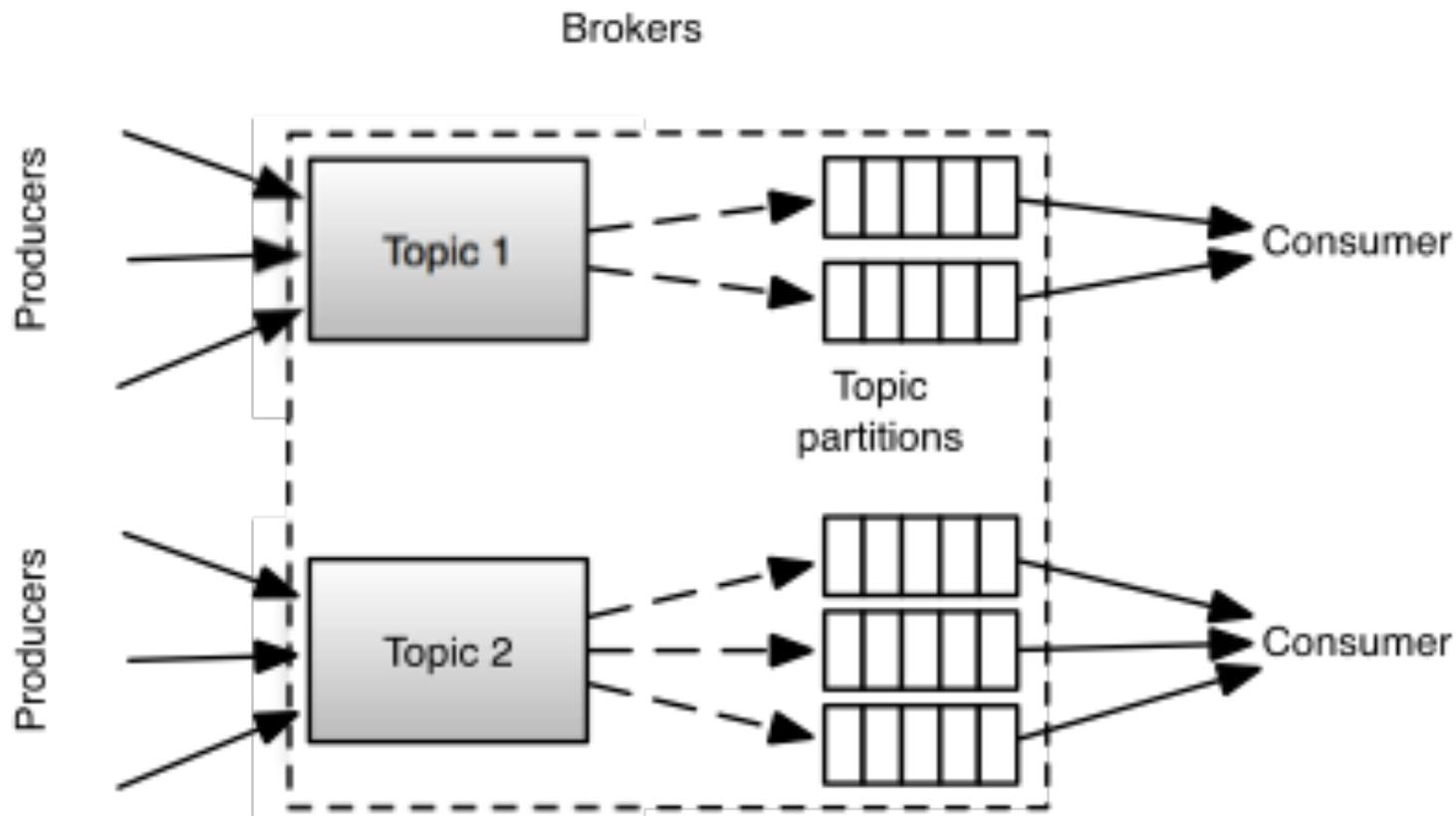
What is a commit log?



The log as a messaging system



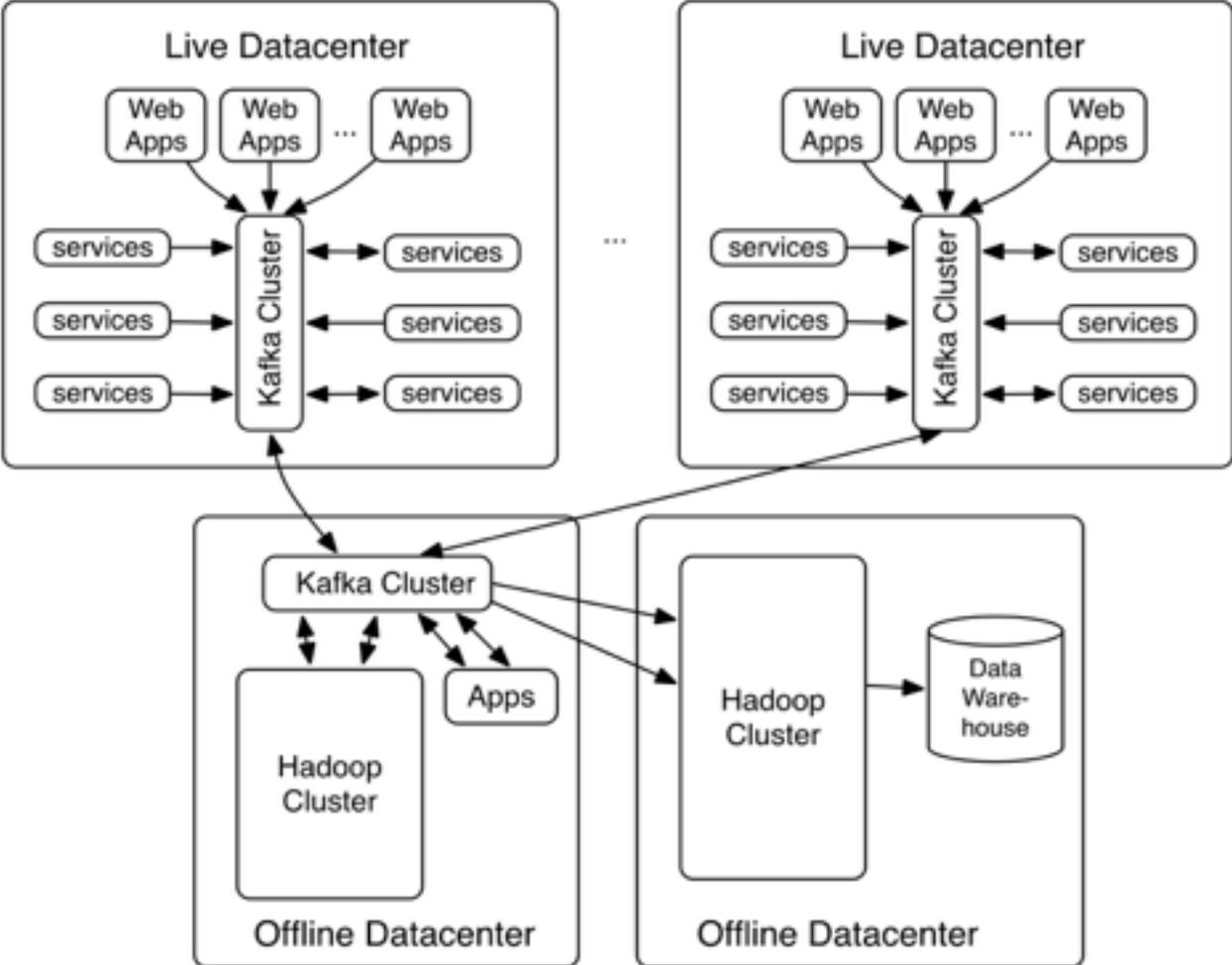
Apache Kafka



Usage at LinkedIn

- 16 brokers in each cluster
- 28 billion messages/day
- Peak rates
 - Writes: 460,000 messages/second
 - Reads: 2,300,000 messages/second
- ~ 700 topics
- 40-50 live services consuming user-activity data
- Many ad hoc consumers
- Every production service is a producer (for metrics)
- 10k connections/colo

Usage at LinkedIn



Four key ideas

1. Central data pipeline
2. Push data cleanliness upstream
3. O(1) ETL
4. Evidence-based correctness

Standardize on Avro in data pipeline

```
{
  "type": "record",
  "name": "URValidationRequestEvent",
  "namespace": "com.linkedin.event.usv",
  "fields": [
    {
      "name": "header",
      "type": {
        "type": "record",
        "name": "TrackingEventHeader",
        "namespace": "com.linkedin.event",
        "fields": [
          {
            "name": "memberId",
            "type": "int",
            "doc": "The member id of the user initiating the action"
          },
          {
            "name": "timeMs",
            "type": "long",
            "doc": "The time of the event"
          },
          {
            "name": "host",
            "type": "string",
            ...
          }
        ]
      }
    }
  ]
}
```

Four key ideas

1. Central data pipeline
2. Push data cleanliness upstream
3. O(1) ETL
4. Evidence-based correctness

Hadoop data load (Camus)

- Open sourced:
 - <https://github.com/linkedin/camus>
- One job loads all events
- ~10 minute ETA on average from producer to HDFS
- Hive registration done automatically
- Schema evolution handled transparently

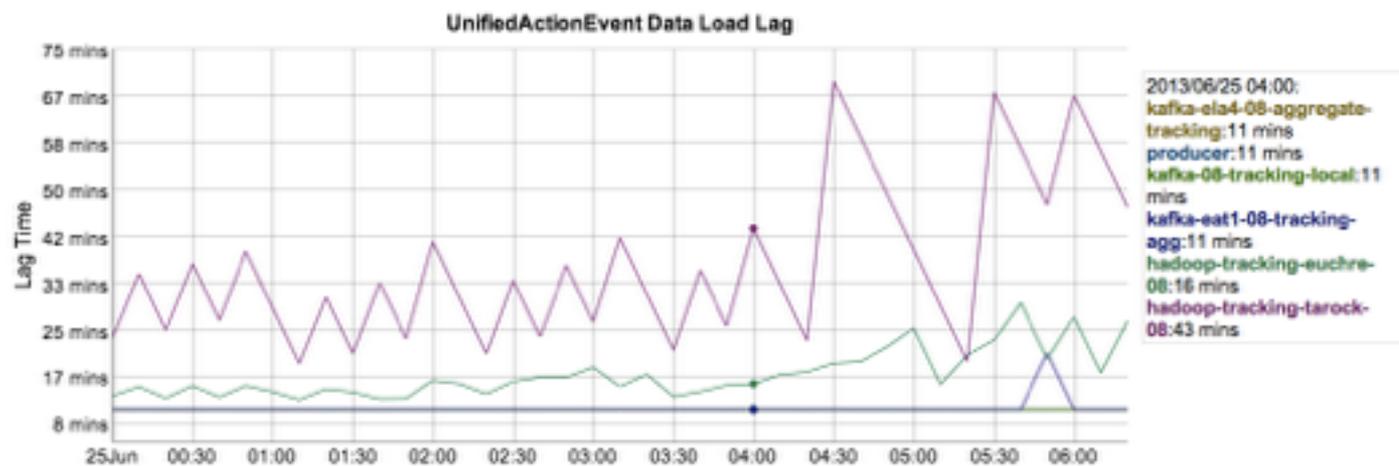
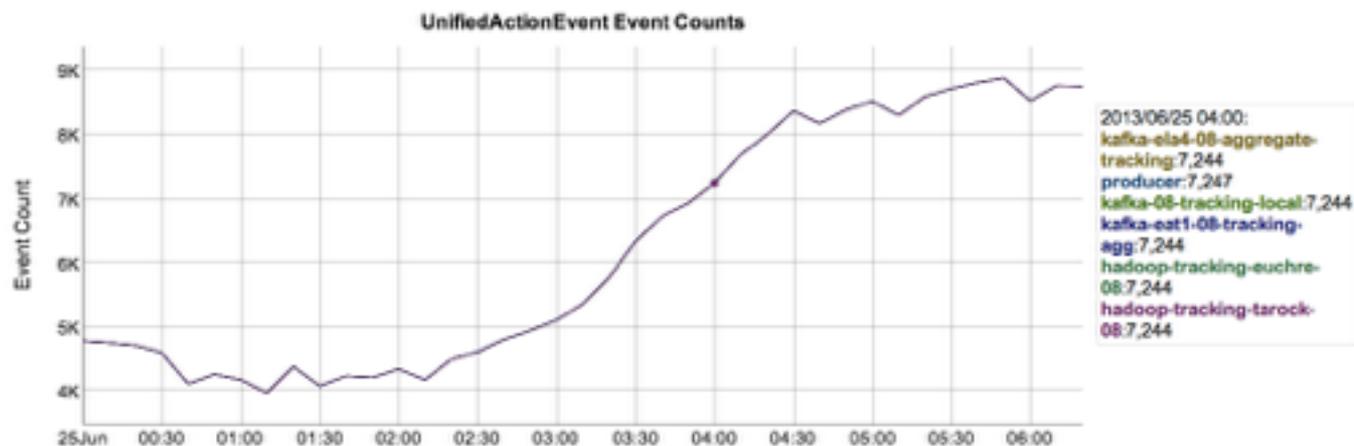
Four key ideas

1. Central data pipeline
2. Push data cleanliness upstream
3. O(1) ETL
4. Evidence-based correctness

Does it work?

“All published messages must be delivered to all consumers (quickly)”

Audit Trail



Kafka replication (0.8)

- Intra-cluster replication feature
 - Facilitates high availability and durability
- Beta release available
 - <https://dist.apache.org/repos/dist/release/kafka/>
- Rolled out in production at LinkedIn last week

Join us at our user-group meeting tonight @ LinkedIn!

- Thursday, June 27, 7.30pm to 9.30pm
- 2025 Stierlin Ct., Mountain View, CA
- <http://www.meetup.com/http-kafka-apache-org/events/125887332/>
- Presentations (replication overview and use-case studies) from:
 - RichRelevance
 - Netflix
 - Square
 - LinkedIn

