Apache Kafka at LinkedIn

Guozhang Wang

BDTC 2014

About Me

Guozhang Wang

Senior Software Engineer at LinkedIn

San Francisco Bay Area | Computer Software

Current	The Apache Software Foundation, LinkedIn
Previous	Cornell University, LinkedIn, Microsoft
Education	Cornell University

View profile as ▼

445 connections

www.linkedin.com/in/guozhangwang/

Contact Info

Agenda

- **_Overview of Kafka_**

- _Kafka Design_

- _Kafka Usage at LinkedIn_

- _Roadmap_

- _Q & A_

Why We Build Kafka?

We Have a lot of Data

- *User activity tracking*
  - *Page views, ad impressions, etc*

- *Server logs and metrics*
  - *Syslogs, request-rates, etc*

- Messaging
  - *Emails, news feeds, etc*

- Computation derived
  - *Results of Hadoop / data warehousing, etc*

.. and We Build Products on Data

Newsfeed

Recommendation

Recommendation

Search

4,002 results for **software engineer linkedin**   Sort by **Relevance** ▾

Senior **Software** Quality Assurance **Engineer**
Epicor Software
Newburgh, NY · Dec 9, 2013
▸ 2 people in your network · Similar   [Save Job] ▾

Software Engineer in Test
LinkedIn
New York · Dec 7, 2013
▸ 1,995 people in your network · Similar   [Save Job] ▾

Principal **Software Engineer**
Symantec
USA - California, Mountain View - HQ · Nov 25, 2013
▸ 2 connections to the poster · Similar   [Save Job] ▾

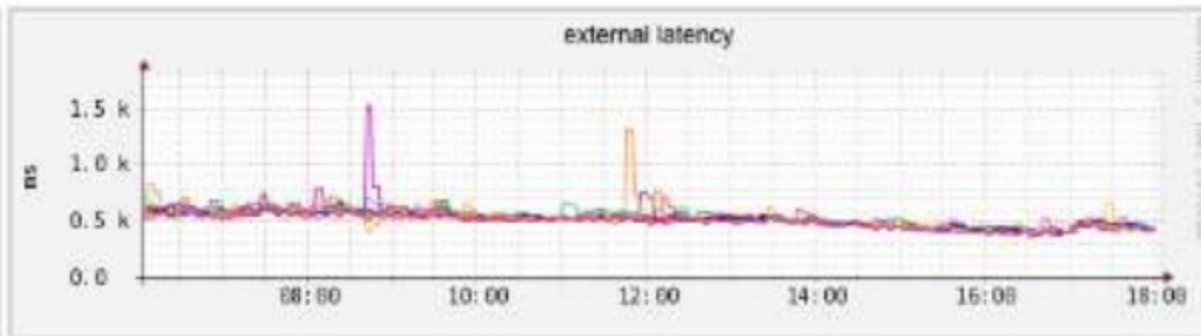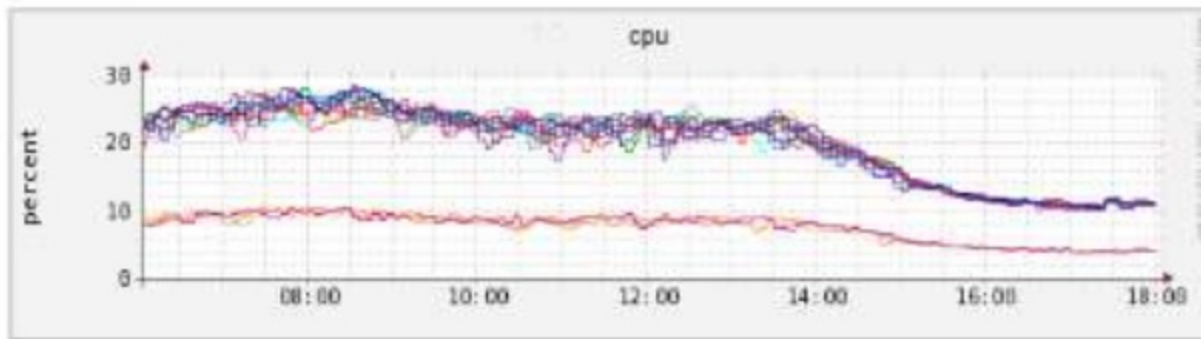Test **Engineer** OR **Software** Development
Engineer in Test (SDET)
Andiamo Partners
Newark, NJ · Dec 10, 2013
▸ 3 people in your network · Similar   [Save Job] ▾
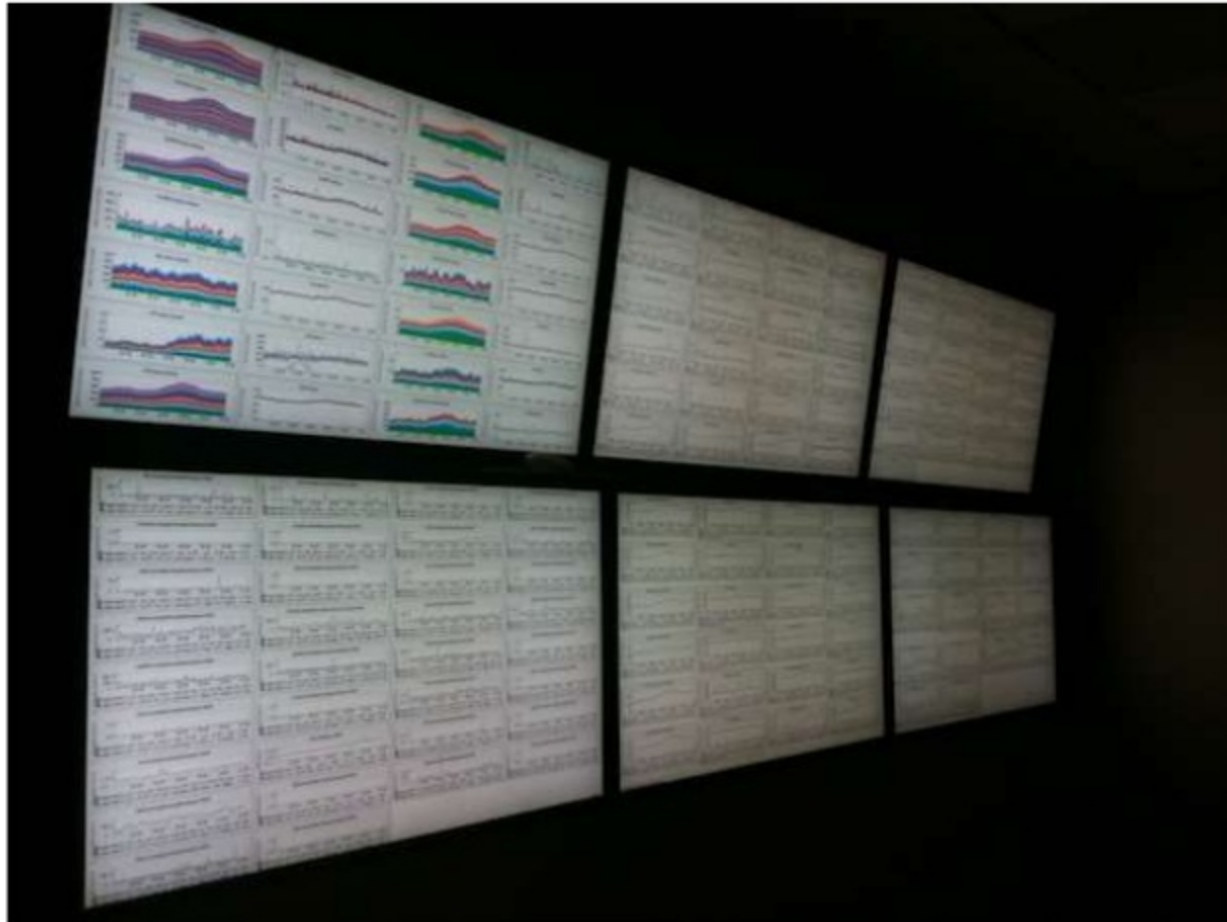
Senior **Software** Developer - C#
Epicor Software
Newburgh, NY · Dec 9, 2013
▸ 2 people in your network · Similar   [Save Job] ▾
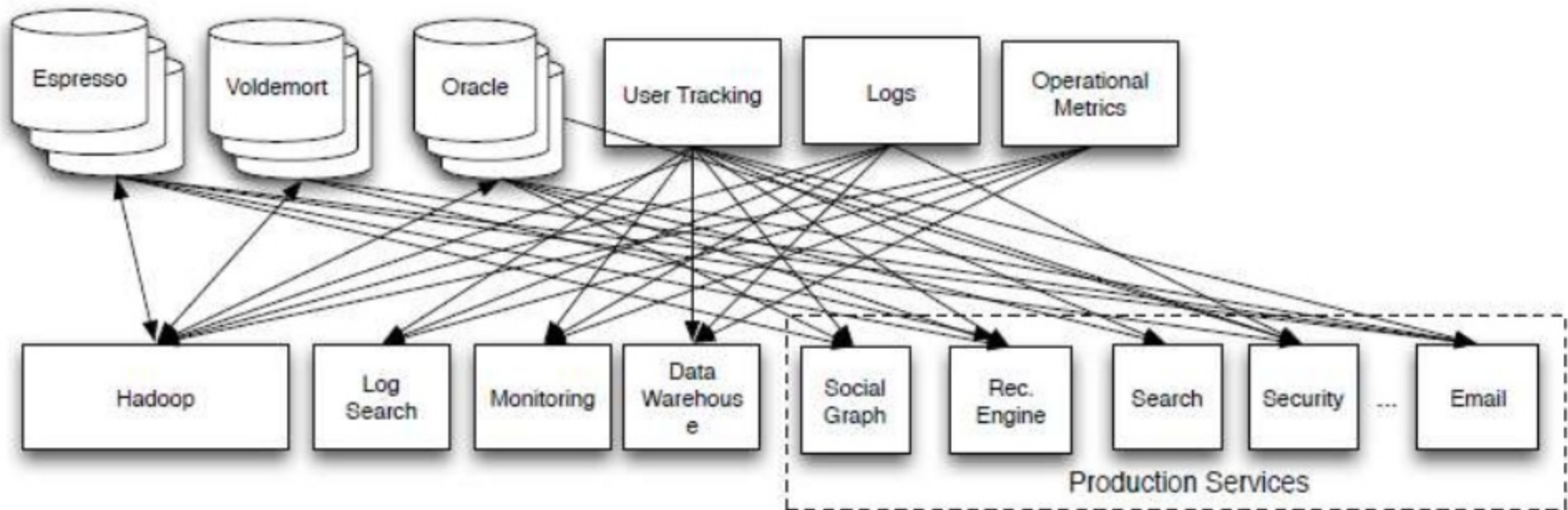
Metrics and Monitoring

.. and a LOT of Monitoring

The Problem:

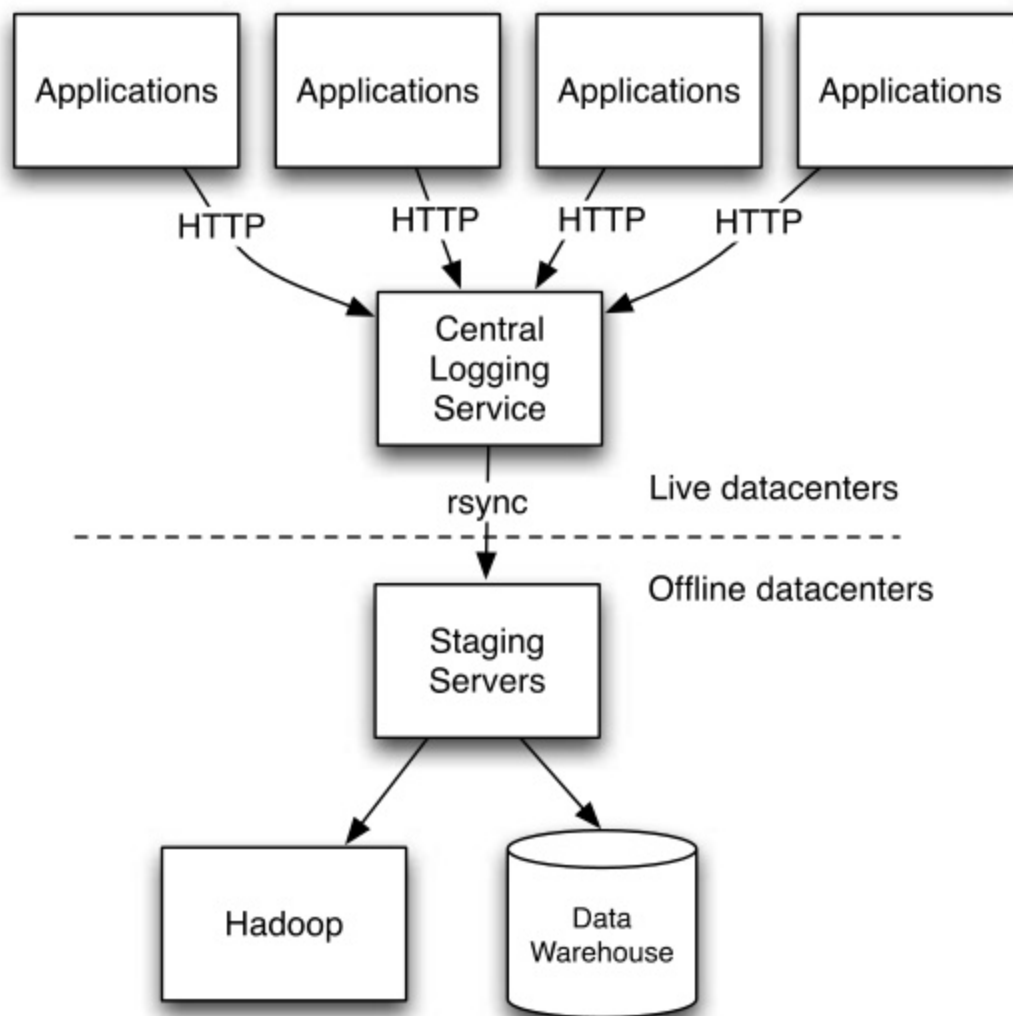*How to integrate this variety of data and make it available to all products?*
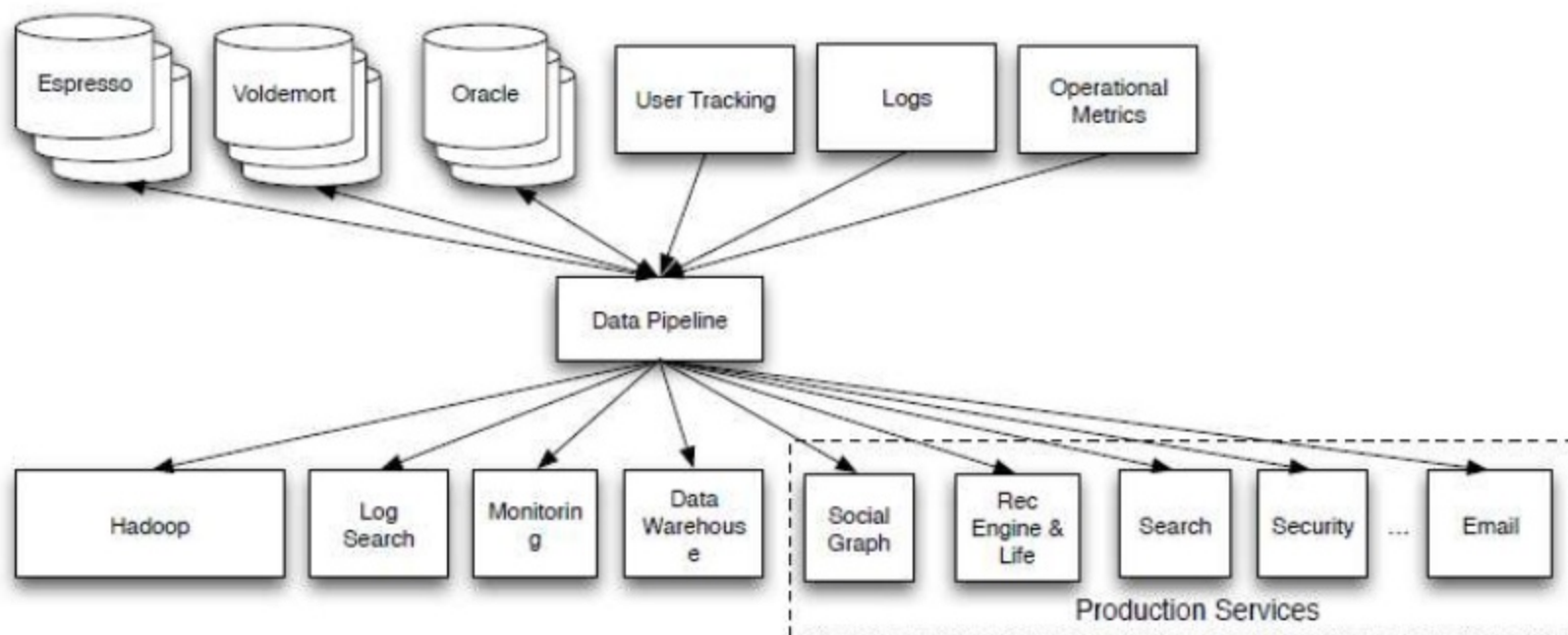
Life back in 2010:
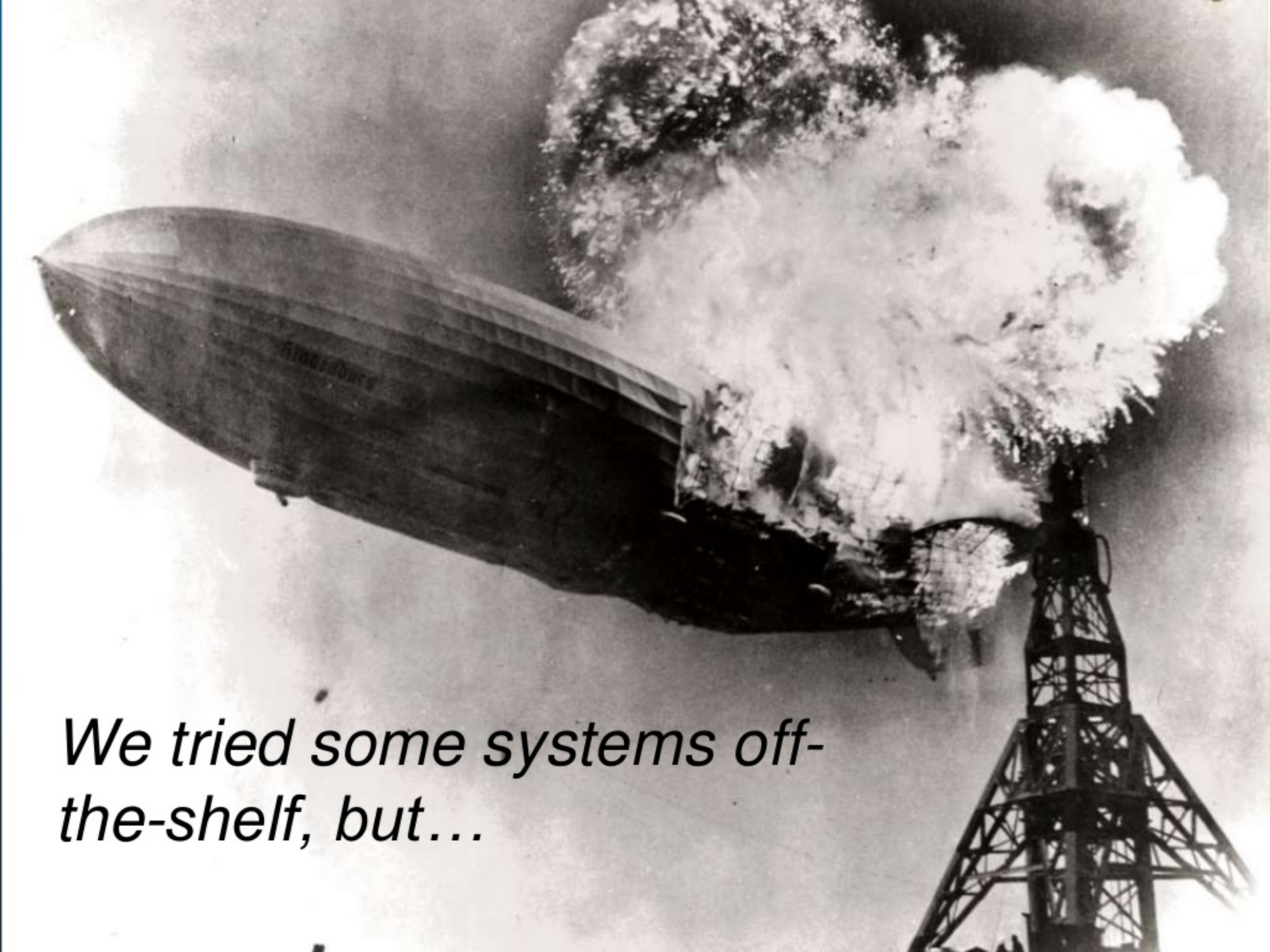# Point-to-Point Pipelines

Example: User Activity Data Flow

What We Want

- A *centralized* data pipeline

We tried some systems off-the-shelf, but…

What We *REALLY* Want

- A **centralized** data pipeline that is

  - Elastically scalable

  - Durable

  - High-throughput

  - Easy to use
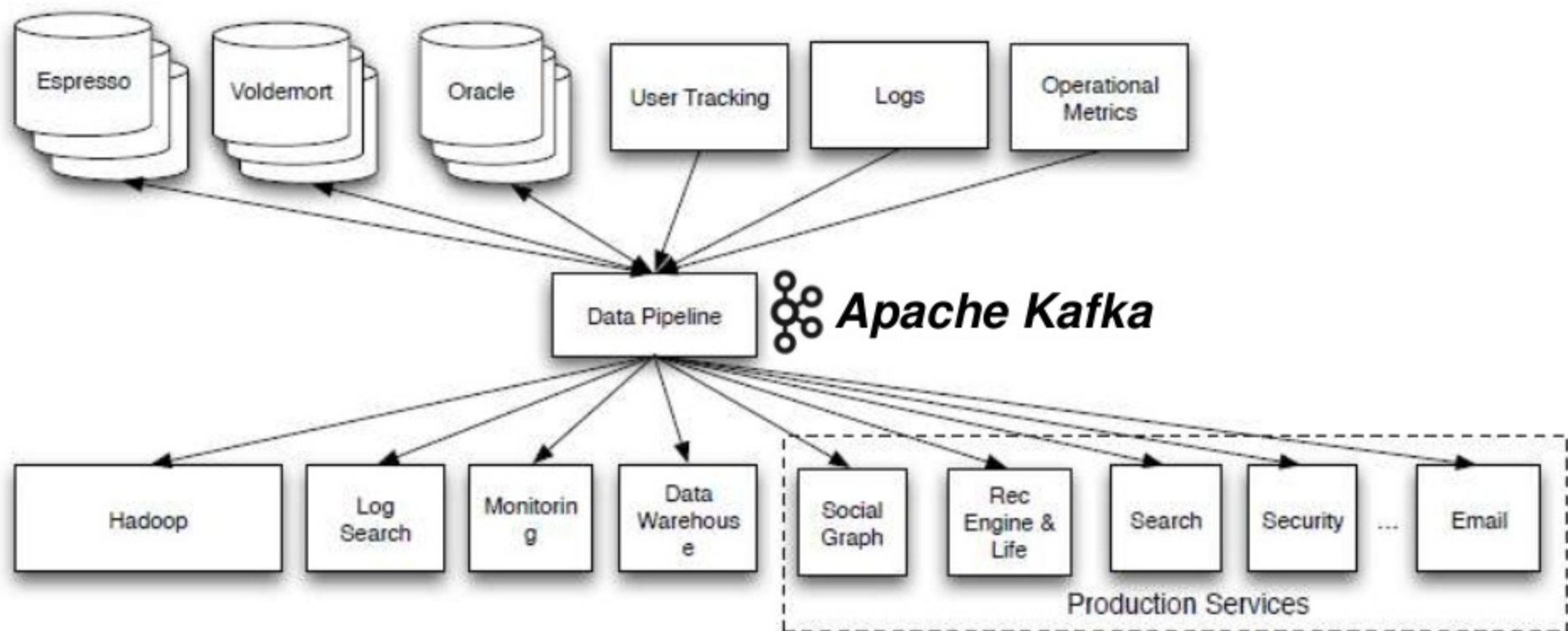
Apache Kafka

- A distributed pub-sub messaging system

  - Scale-out from groundup

  - Persistent to disks

  - High-Throughput (10s MB/sec per server)

Life Since Kafka in Production

- Developed and maintained by 5 Devs + 2 SRE