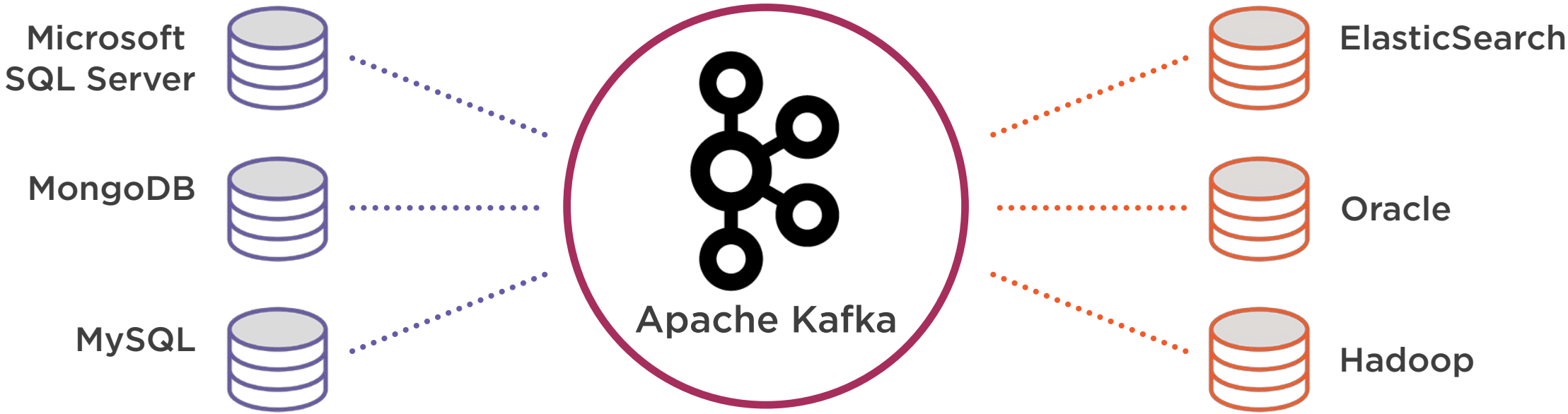
Getting Started with Apache Kafka

WHY APACHE KAFKA

Ryan Plant
COURSE AUTHOR

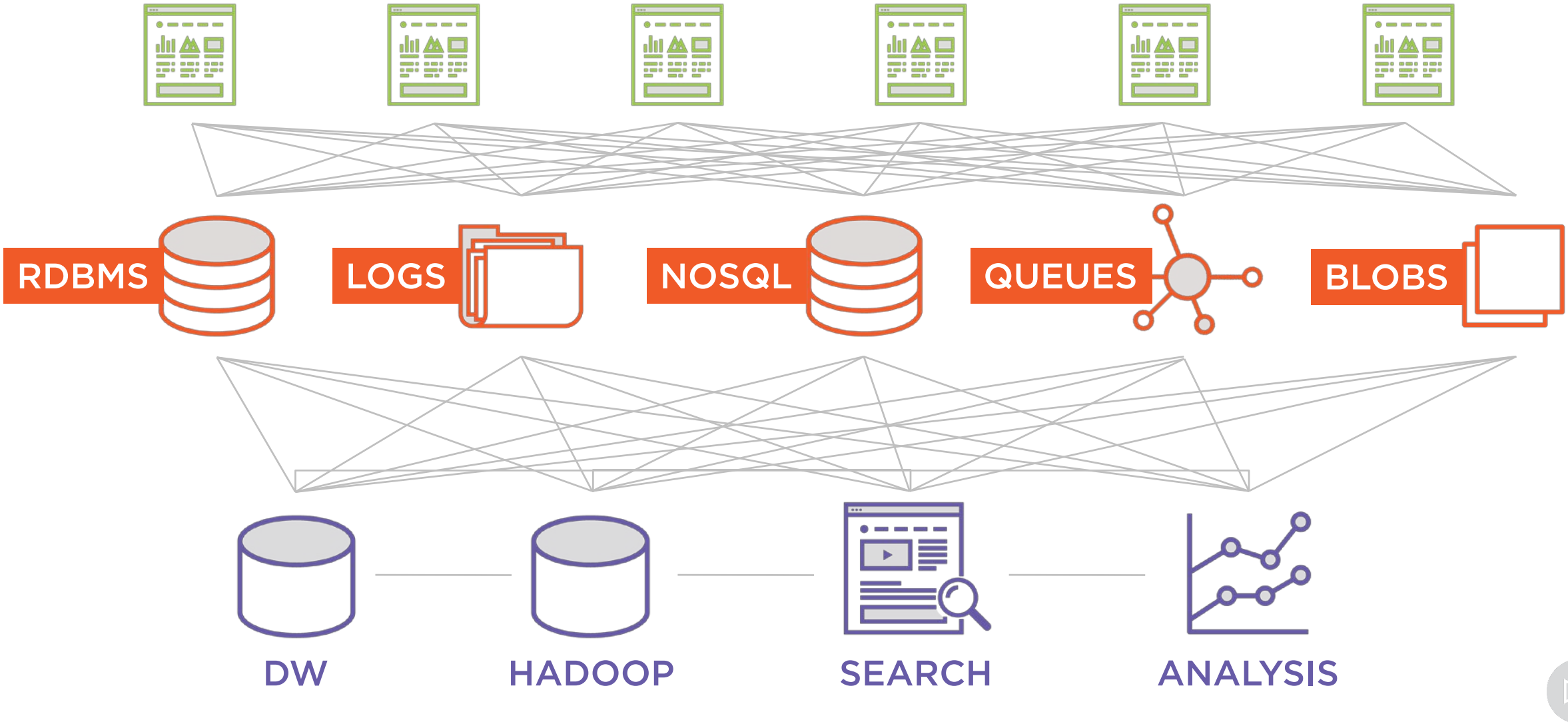
@ryan_plant blog.ryanplant.com

What Is Apache Kafka?



"A high-throughput distributed messaging system."

What a Typical Enterprise Looks Like



Database replication

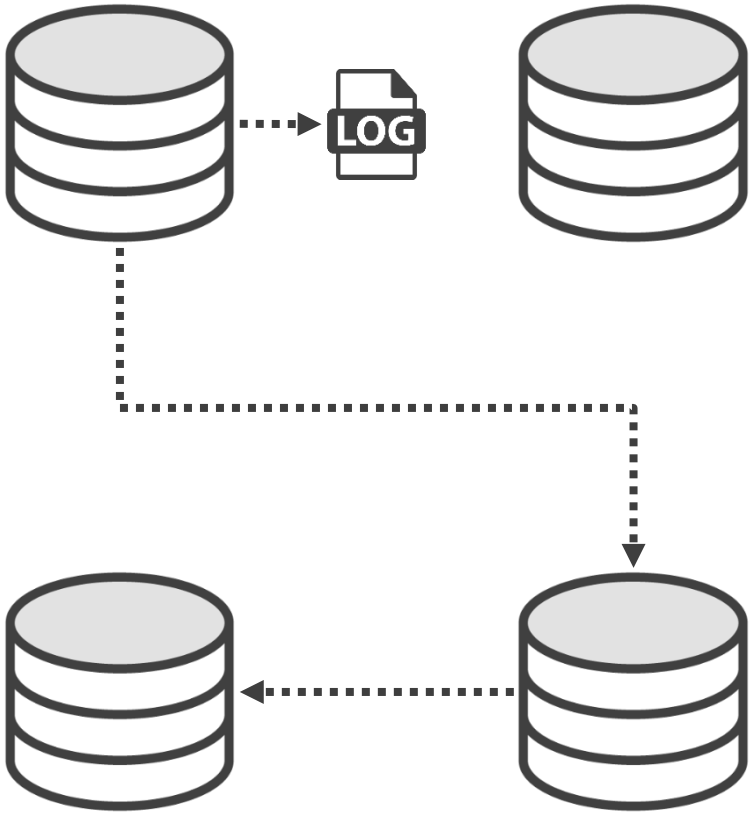
Log shipping

Extract, Transform, and Load (ETL)

Messaging

Custom middleware magic

Database Replication and Log Shipping



RDBMS to RDBMS only

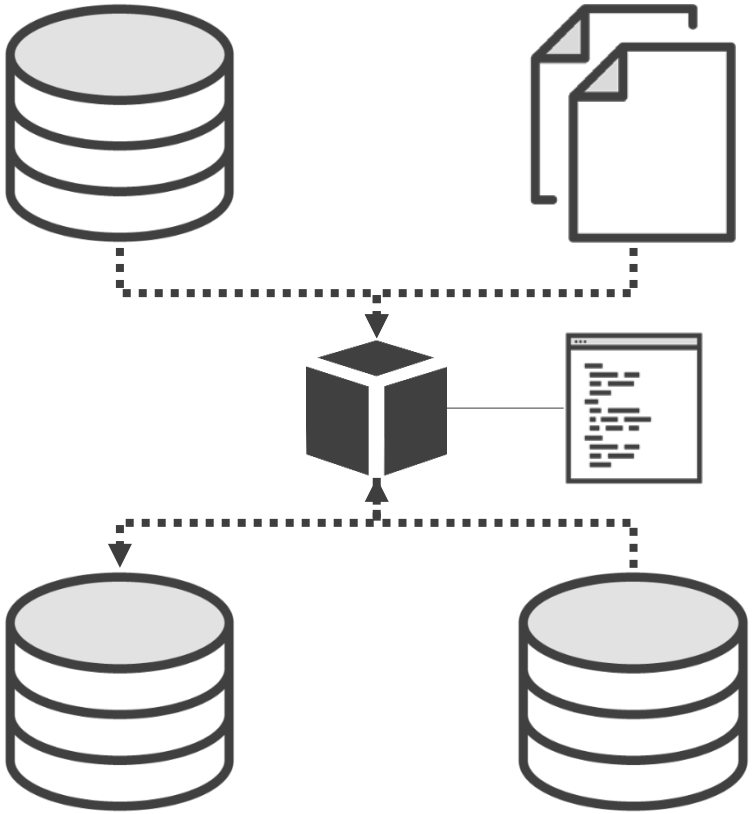
Database-specific

Tight coupling (schema)

Performance challenges (log shipping)

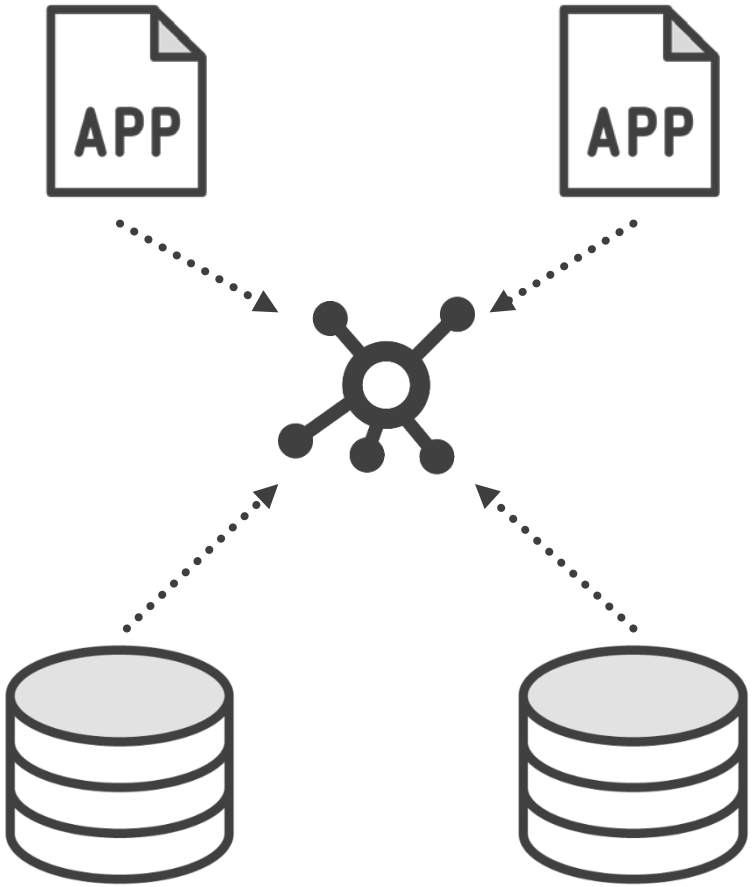
Cumbersome (subscriptions)

Extract, Transform, and Load (ETL)



- Typically proprietary and costly
- Lots of custom development
- Scalability challenged
- Performance challenged
- Often times requires multiple instances

Messaging



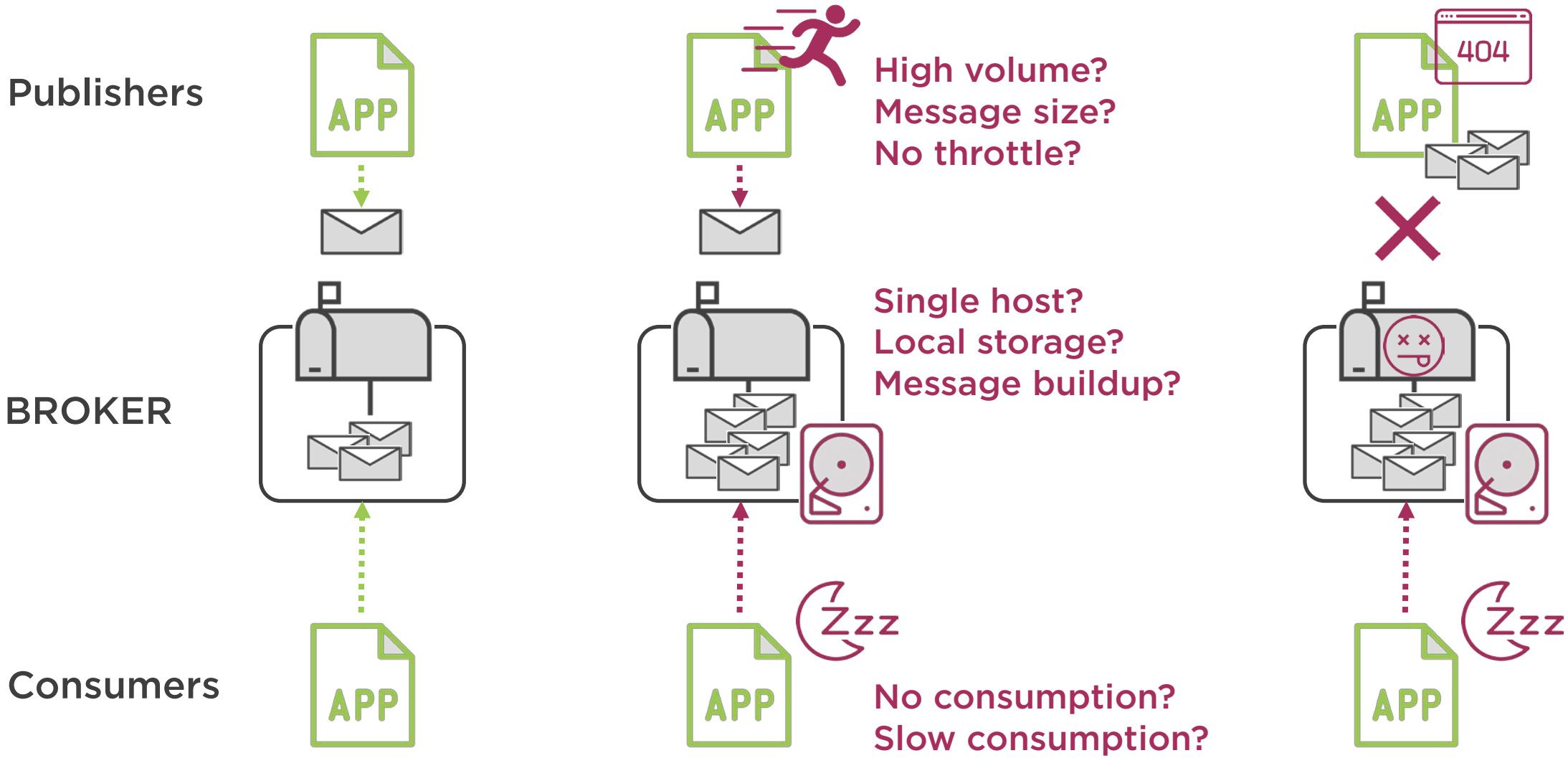
Limited scalability

Smaller messages

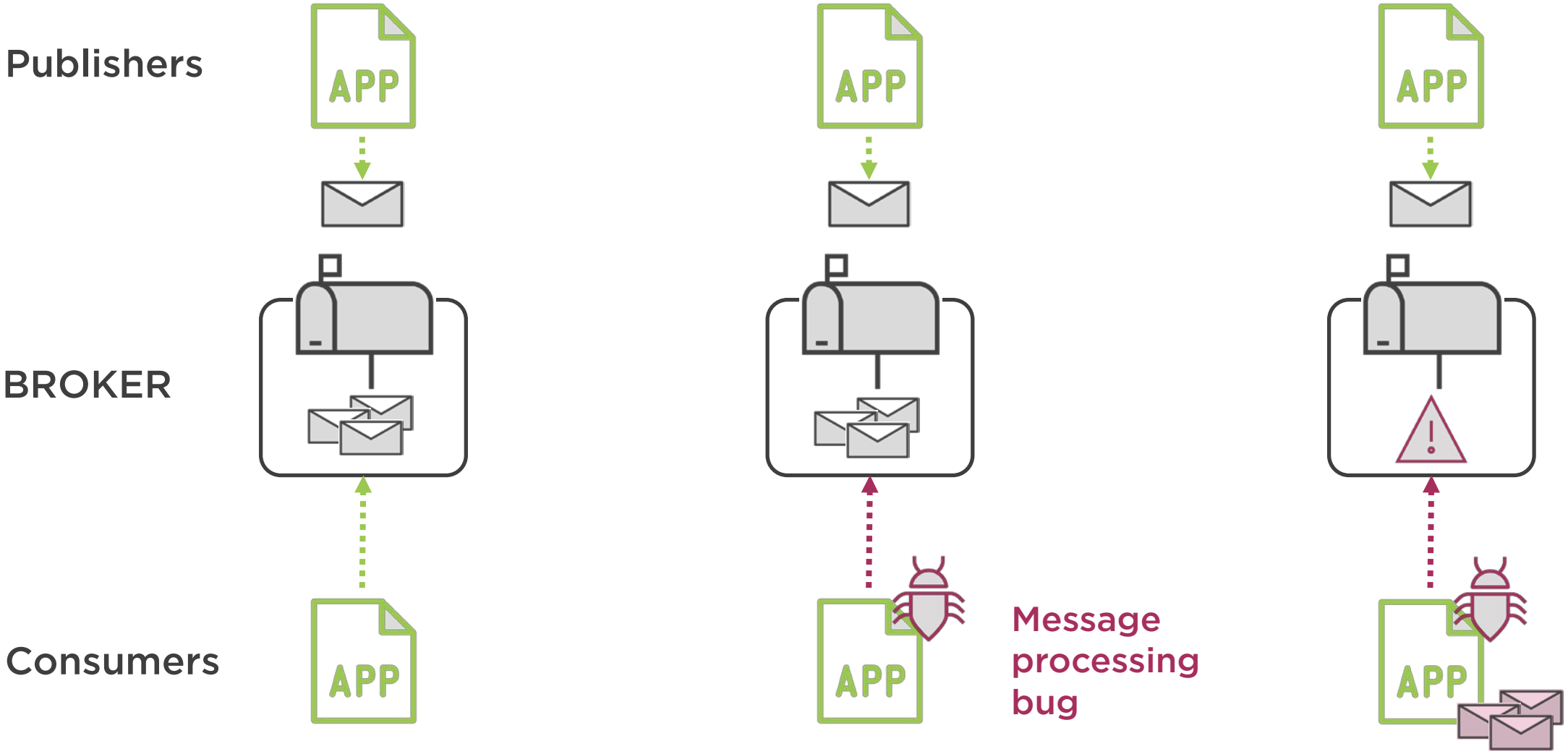
Requires rapid consumption

Not fault-tolerant (application)

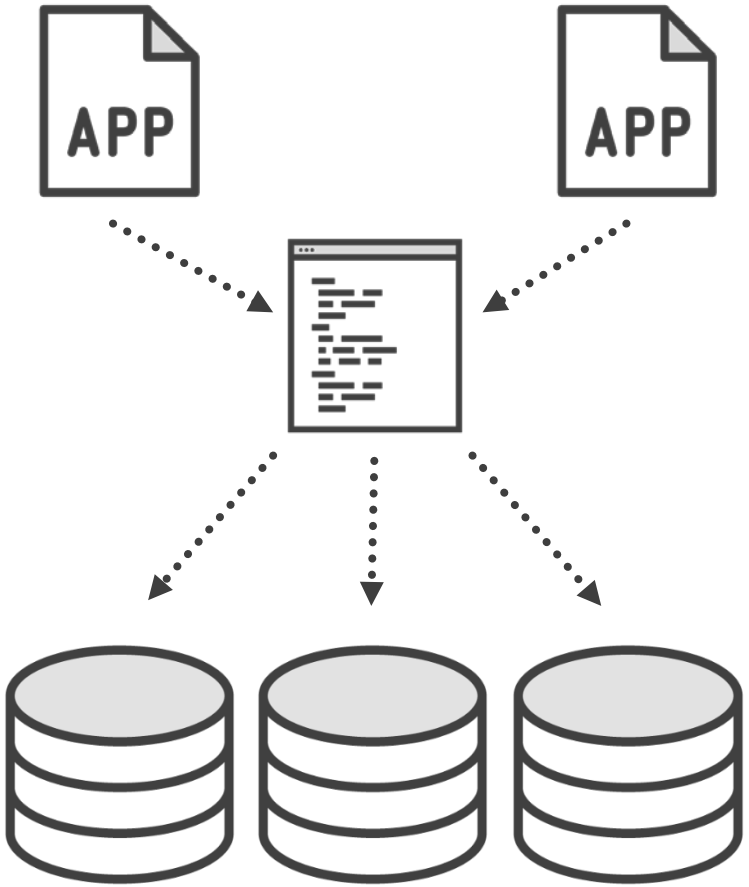
Perils of Messaging Under High Volume



Perils of Messaging With Application Faults



Middleware Magic



Increasingly complex

Deceiving

Consistency concerns

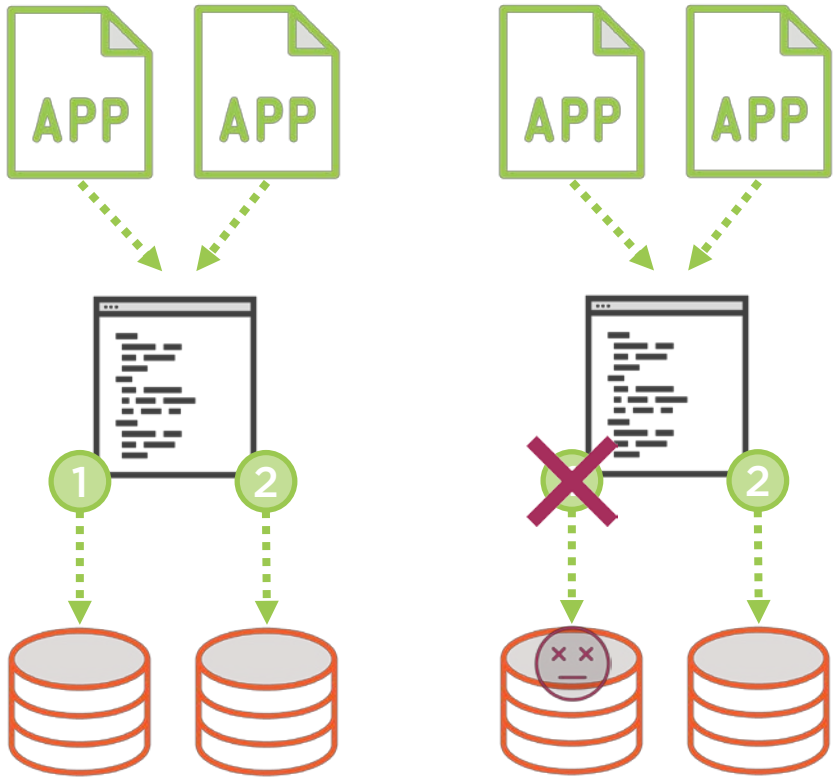
Potentially expensive

Middleware Challenges

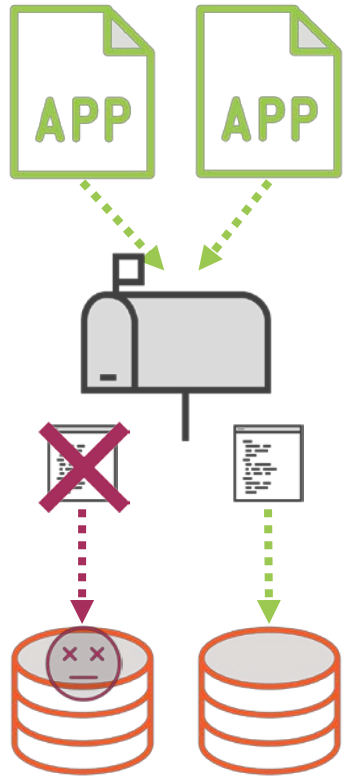
Multi-write pattern

Atomic transaction

Coordination logic



Message broker pattern



Competing consumers

Non-consuming consumer

Isn’t There a Better Way?

To move data around:

- Cleanly
- Reliably
- Quickly
- Autonomously

That's What LinkedIn Asked in 2010...

High Volume:

- Over 1.4 trillion messages per day
- 175 terabytes per day
- 650 terabytes of messages consumed per day
- Over 433 million users

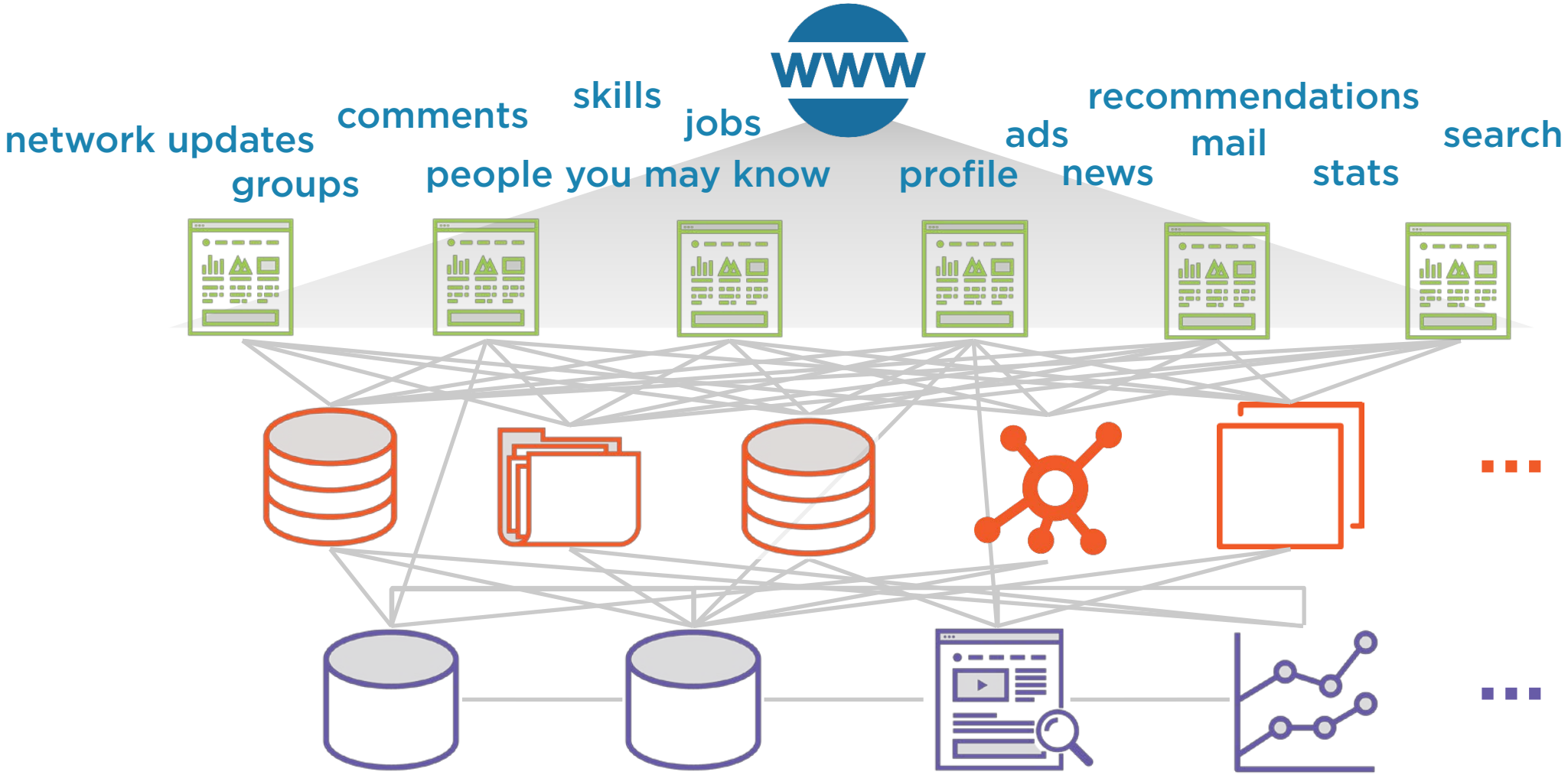
High Velocity:

- Peak 13 million messages per second
- 2.75 gigabytes per second

High Variety:

- Multiple RDBMS (Oracle, MySQL, etc.)
- Multiple NoSQL (Espresso, Voldemort)
- Hadoop, Spark, etc.

Pre-2010 LinkedIn Data Architecture



Franz Kafka

kaf•ka•esque /’káf, kə, ɛsk/ | adjective

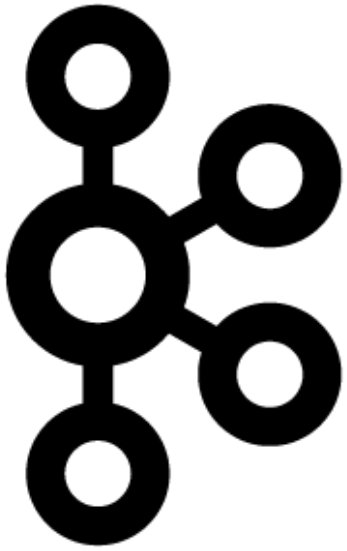
Basically it describes a nightmarish situation which most people can somehow relate to, although strongly surreal.

synonyms: surreal, lucid, spoilsbury toast boy

Usage: “Whoa! This flick is way kafkaesque…”

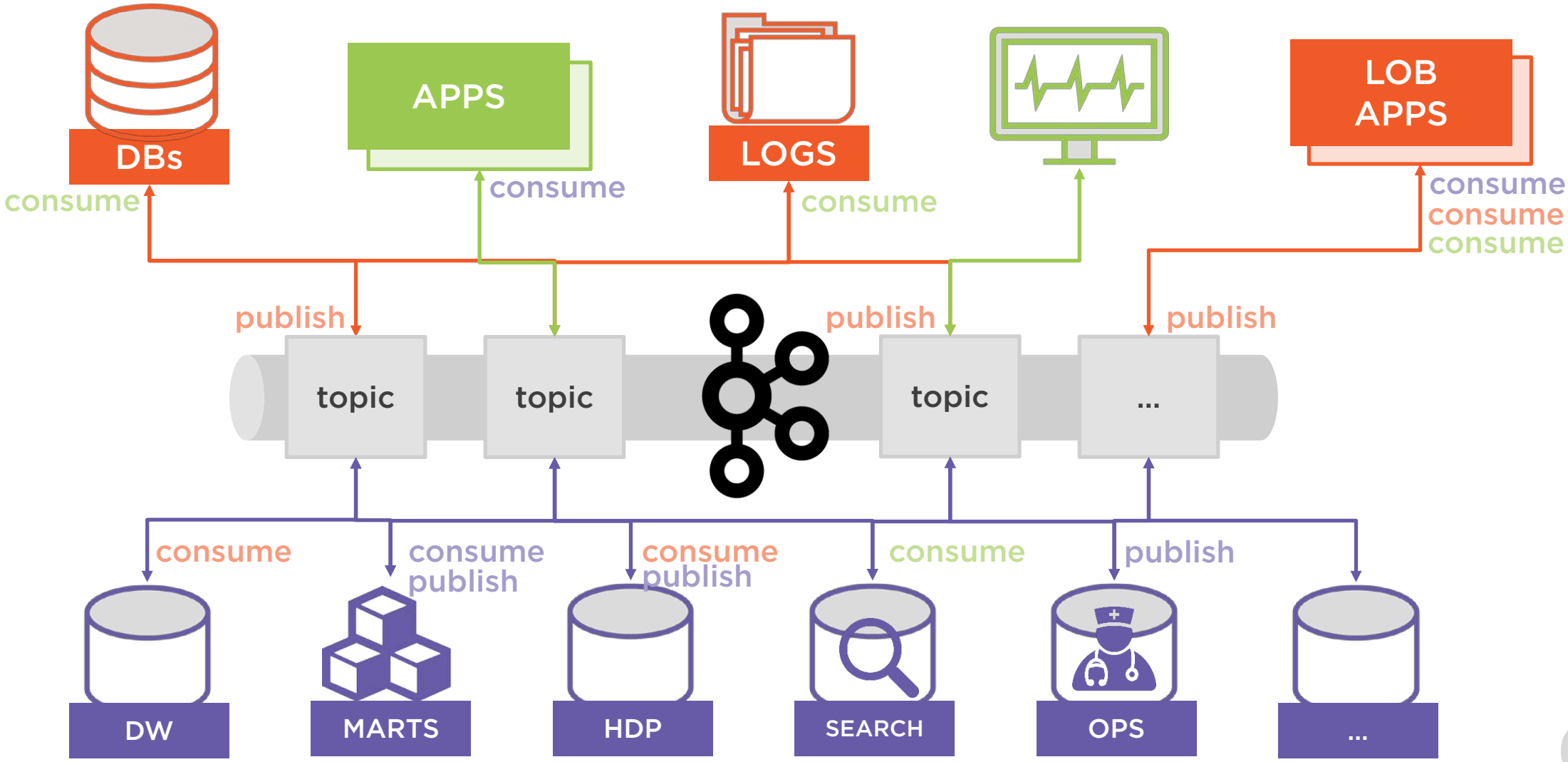
Source: Urban Dictionary

Next-generation Messaging Goals



- High throughput
- Horizontally scalable
- Reliable and durable
- Loosely coupled Producers and Consumers
- Flexible publish-subscribe semantics

Post-2010 LinkedIn Data Architecture



Timeline of Events

2003
LinkedIn Launch

2009
Kafka Inception
Development begins

2010
Initial Kafka Deployment @ LinkedIn

2011
Kafka Open Sourced
Apache Software Foundation

Today
1.1 Trillion messages per day @ LinkedIn

Apache Kafka Adoption

7X since 2015

Yahoo
Etsy
Microsoft
Bing
Mailchimp

Uber
Oracle
Goldman Sachs
Netflix
PayPal

Square
Coursera
IBM
Pintrest
Twitter

Airbnb
Spotify
Ancestry
LinkedIn
Hotels.com

Summary

Kafka is a distributed messaging system

Designed to move data at high volumes

Addresses shortcomings of traditional data movement tools and approaches

Invented by LinkedIn to address data growth issues common to many enterprises

Open-sourced under Apache Software Foundation in 2012

First-choice adoption for data movement for hundreds of enterprise and internet-scale companies