



# Data Infrastructure at LinkedIn

Shirshanka Das

XLDB 2011



# Me



- UCLA Ph.D. 2005 (Distributed protocols in content delivery networks)
- PayPal (Web frameworks and Session Stores)
- Yahoo! (Serving Infrastructure, Graph Indexing, Real-time Bidding in Display Ad Exchanges)
- @ LinkedIn (Distributed Data Systems team): Distributed data transport and storage technology (Kafka, Databus, Espresso, ...)

# Outline

- **LinkedIn Products**
- Data Ecosystem
- LinkedIn Data Infrastructure Solutions
- Next Play

# LinkedIn By The Numbers

- 120,000,000+ users in August 2011
- 2 new user registrations per second
- 4 billion People Searches expected in 2011
- 2+ million companies with LinkedIn Company Pages
- 81+ million unique visitors monthly\*
- 150K domains feature the LinkedIn Share Button
- 7.1 billion page views in Q2 2011
- 1M LinkedIn Groups


\* Based on comScore, Q2 2011


# Member Profiles

LinkedIn Account Type: Pro Tom Quiggle Add Connections

Home Profile Contacts Groups Jobs Inbox 92 Companies News More People Search... Advanced

**Earn Your M.S. in Finance - 15 month program designed for busy working professionals. Click for info!** From: Saint Mary's College of California




**Jeff Weiner** 1st 

CEO at LinkedIn  
Mountain View, California | Internet


**Suggest Connections**

[Send Jeff a message](#)  
[Suggest a profile update for Jeff](#)  
[Save Jeff's Profile](#) ?

LinkedIn Premium 

**Get Hired Faster**  
with Job Seeker Premium


- Get noticed with a Job Seeker Badge
- Move to the top as a Featured Applicant
- Contact recruiters directly with InMail

[Learn More](#) 

**Jeff's Activity**

**Jeff Weiner** likes this update:

**Mike Gamson** Another step in the right direction for the Khan Academy. . .




**Khan Academy Integrates With Digital Textbooks** mashable.com

The 12-minute lectures that Bill Gates has called "the start of a revolution" will now be linked with the material in some digital textbooks.

4 hours ago • Like (4) • Comment • Share


**Jeff Weiner @dtunkelang** adds great examples to recent talk re: the importance of understanding mutual fit when recruiting.



**Dream. Fit. Passion.** thenoisychannel.com · via Daniel Tunkelang

A few days ago, our CEO Jeff Weiner led a


**Jeff Weiner @dtunkelang** adds great examples to recent talk re: the importance of understanding mutual fit when recruiting.




**Dream. Fit. Passion.** thenoisychannel.com · via Daniel Tunkelang


A few days ago, our CEO Jeff Weiner led a session at LinkedIn on how to "close" candidates — that is, how to persuade candidates to join your team once you have found and interviewed them. Since not everyone has the opportunity...


5 hours ago • Like (1) • Comment • Send a message • Share • See all activity


**Current** **CEO at LinkedIn** 

**Member, Board of Directors at DonorsChoose** 

**Member, Board of Directors at Malaria No More**

**Past** Executive in Residence at Accel Partners 

Executive in Residence at Greylock 

Executive Vice President Network Division at Yahoo! 

[see all](#) ▾

**Education** University of Pennsylvania - The Wharton School

**Recommendations** 8 people have recommended Jeff

**Connections** 500+ connections

**Websites** [Company Website](#)

**Twitter** [Follow](#) @jeffweiner

**Public Profile** <http://www.linkedin.com/in/jeffweiner08>

[Share](#) [PDF](#) [Print](#) [vCard](#) [Flag](#)

**Summary**

Internet executive with over 16 years of experience, including general management of mid to large size organizations, corporate development, product development, business operations, and strategy.

LinkedIn

5

# Signal - faceted stream search

Search within checked filters

**Try these searches**  
Computer Software, San Francisco Bay Area  
Computer Software  
 Search for your company, competitors or favorite product.

**Network**   
 By Me (0)  
 1st Connections (6)  
 2nd Connections (5)  
 3rd + Everyone (0)

**Company**   
 LinkedIn (11)  
 EMC (1)

**Location**   
 San Francisco Bay Area (11)

**Industry**

**Time**

**School**

**11 Updates**

**Harold Lee** Talk by John Ousterhout on RAMCloud at LinkedIn today at 4pm, open to the public, RSVP via the link if interested

**RAMCloud: Scalable High-Performance Storage Entirely in DRAM - A...** events.linkedin.com 2nd  
In recent years DRAM has played a larger and larger role in storage systems, driven by the demands of large-scale Web applications. However, DRAM is still used primarily in limited or special-purpose ways, such as a cache for...  
Like · Comment · Share · 5 days ago

**Kapil Surlaker** Tech talk on RAMCloud by John Ousterhout Today!

**RAMCloud: Scalable High-Performance Storage Entirely in DRAM - A...** events.linkedin.com 1st  
In recent years DRAM has played a larger and larger role in storage systems, driven by the demands of large-scale Web applications. However, DRAM is still used primarily in limited or special-purpose ways, such as a cache for...  
Like (1) · Comment · Share · 5 days ago  
 Ganesh Narasimhan likes this

**Jingwei Wu**

**RAMCloud: Scalable High-Performance Storage Entirely in DRAM** engineering.linkedin.com 2nd  
Come by LinkedIn Headquarters on Wednesday, October 12 for a public tech talk "RAMCloud: Scalable High-Performance Storage Entirely in DRAM". John Ousterhout, Professor of Computer Science at Stanford University, will be...  
Like · Comment · Share · 6 days ago

# People You May Know

Add Connections Colleagues Classmates **People You May Know**

## Filter By...

### Current Company

- All Companies
- LinkedIn (36)
- PayPal (27)
- Yahoo! (14)
- Google (8)
- Cisco Systems (3)

### Past Company

- All Companies
- PayPal (20)
- Yahoo! (18)
- LinkedIn (15)
- Oracle (13)
- IBM (11)

### School

- All Schools
- University of California, Los Angeles (24)
- Indian Institute of Technology, Delhi (15)
- Stanford University (11)
- San Jose State University (8)
- University of California, Santa Cruz (6)

## Import contacts >

It's easy to search your email contacts and quickly grow your network



**Sudip Nag** (2nd)

Senior Director, Software Development at Xilinx Inc  
8 connections in common

[+ Connect](#) | [x](#)



**Ross Bundy** (2nd)

Senior QA Engineer at Riverbed Technology  
6 connections in common

[+ Connect](#) | [x](#)



**Jun-Hong (June) Cui** (2nd)

Associate Professor & Assistant Dean, School of Engineering, University of Connecticut  
7 connections in common

[+ Connect](#) | [x](#)



**Amer Marji** (3rd)

Technology Consultant at Accenture

[+ Connect](#) | [x](#)

**Preeth Eldhose** (3rd)

QA Director at BroadVantage Inc & Stocks Day Trading Investor (at Preeth Inc)

[+ Connect](#) | [x](#)



**Sudipto Mukhopadhyay** (3rd)

Sr. Member of Technical Staff at VMware

[+ Connect](#) | [x](#)

**Ashidhara Das** (3rd)

Independent Higher Education Professional

[+ Connect](#) | [x](#)

**Michael Durand** (2nd)

-

3 connections in common

[+ Connect](#) | [x](#)



**Alexandro Sentinelli** (2nd)

Research System Engineer at STMicroelectronics  
5 connections in common

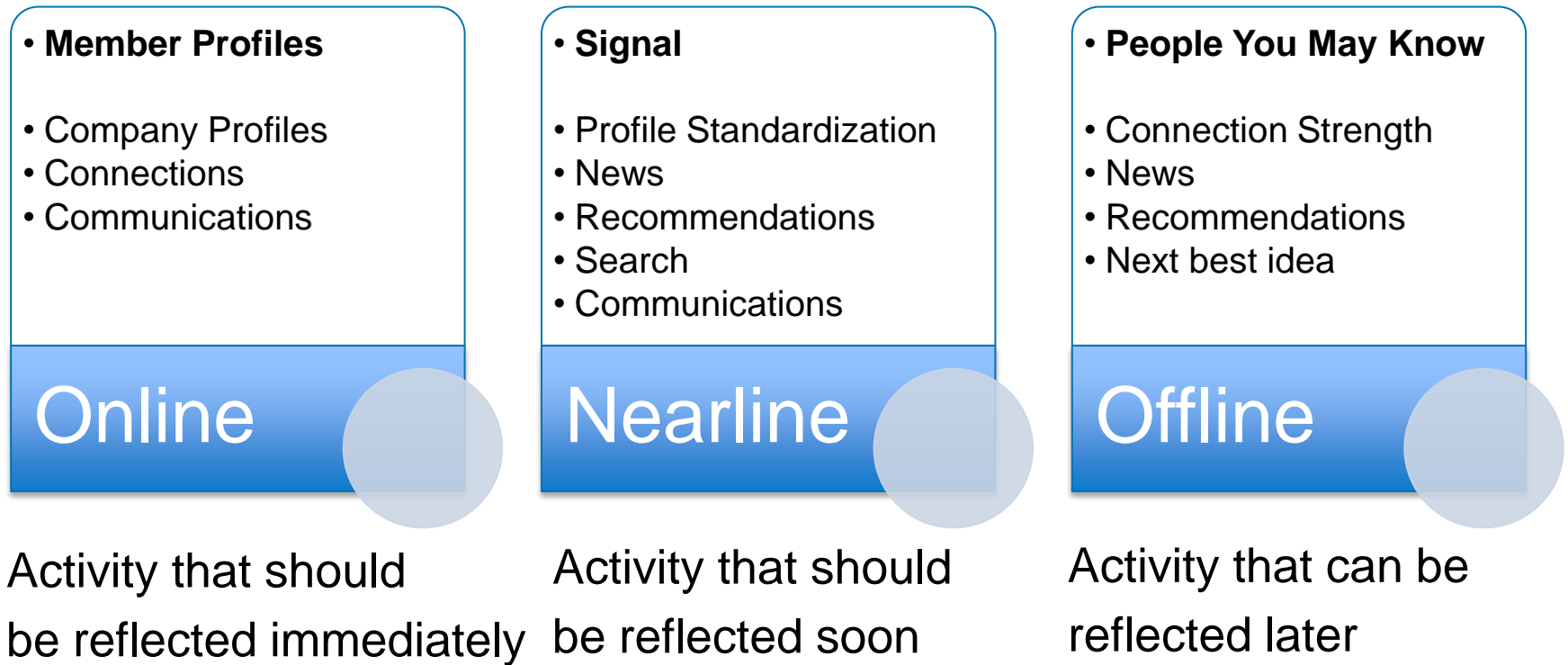
[+ Connect](#) | [x](#)

# Outline

- LinkedIn Products
- **Data Ecosystem**
- LinkedIn Data Infrastructure Solutions
- Next Play



# Three Paradigms : Simplifying the Data Continuum



# Data Infrastructure Toolbox (Online)

Capabilities	Systems
Key-value access	Voldemort
Rich structures (e.g. indexes)	Espresso Oracle
Change capture capability	
Search platform	Zoie, Bobo, Sensei
Graph engine	D-Graph

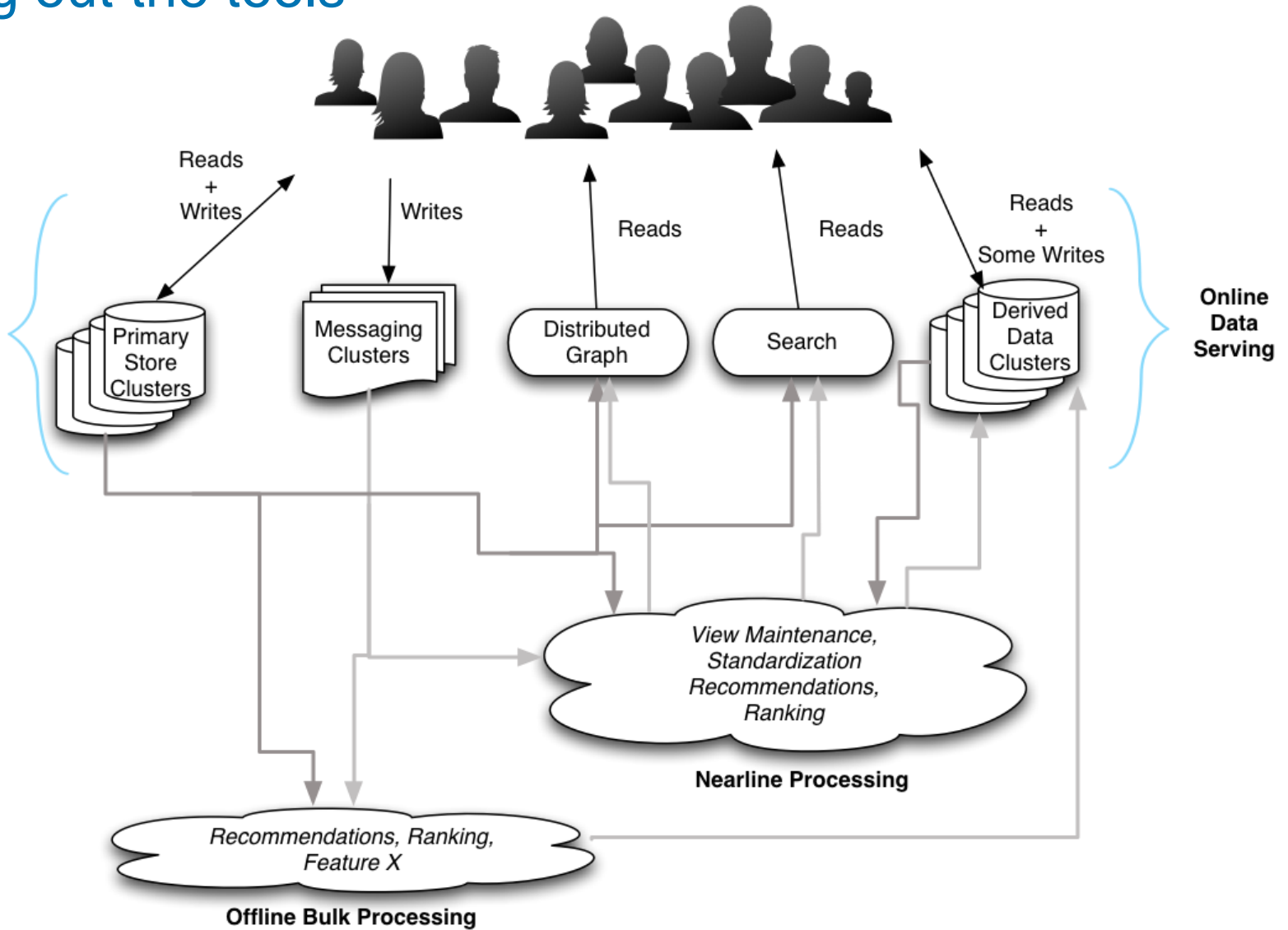
# Data Infrastructure Toolbox (Nearline)

Capabilities	Systems
Change capture streams	Databus
Messaging for site events, monitoring	Kafka
Nearline processing	<i>Coming Soon!</i>

# Data Infrastructure Toolbox (Offline)

Capabilities	Systems
Machine learning, ranking, relevance	Hadoop, Hive, Pig Azkaban, RDBMS
Analytics on Social gestures	<i>Coming Soon!</i>

# Laying out the tools



# Outline

- LinkedIn Products
- Data Ecosystem
- **LinkedIn Data Infrastructure Solutions**
- Next Play

# Focus on four systems in Online and Nearline

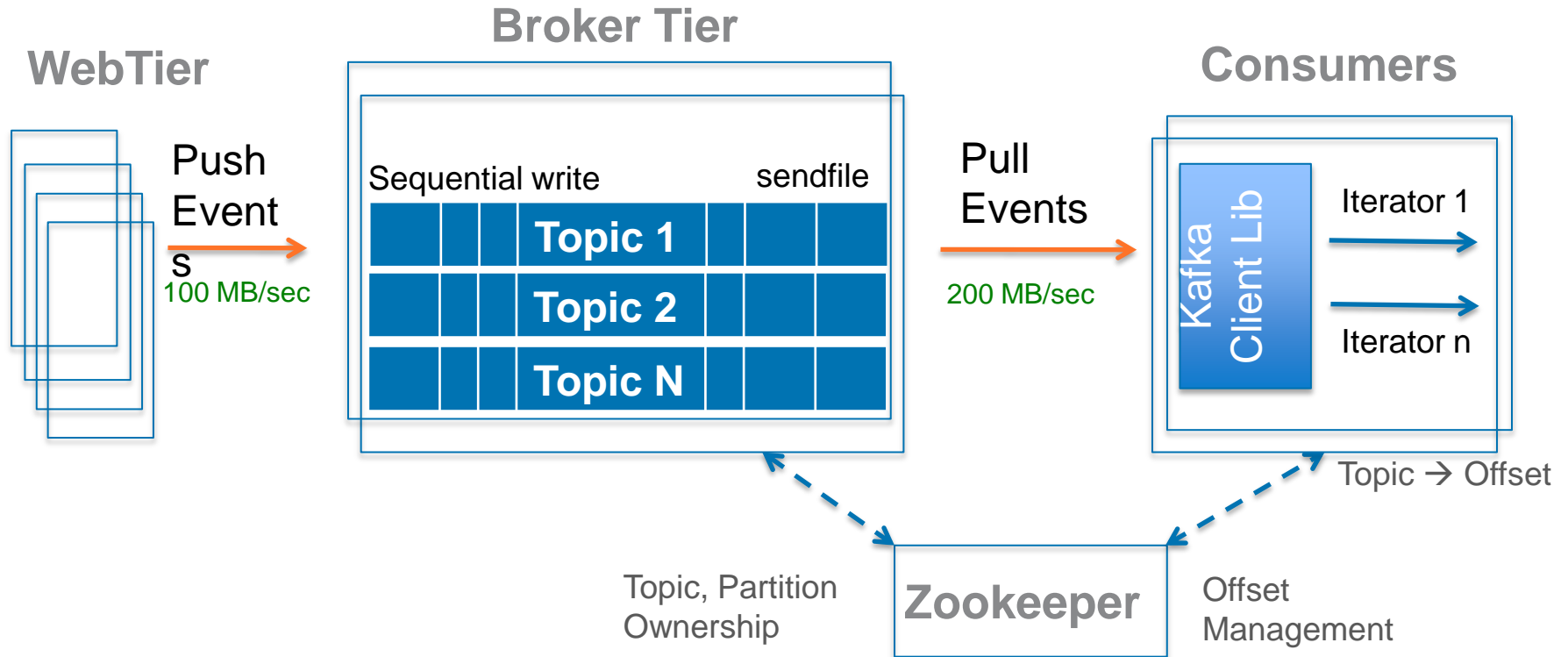
- Data Transport
  - Kafka
  - Databus
- Online Data Stores
  - Voldemort
  - Espresso

LinkedIn Data Infrastructure Solutions

## Kafka: High-Volume Low-Latency Messaging System



# Kafka: Architecture



## Scale

- Billions of Events
- TBs per day
- Inter-colo: few seconds
- Typical retention: weeks

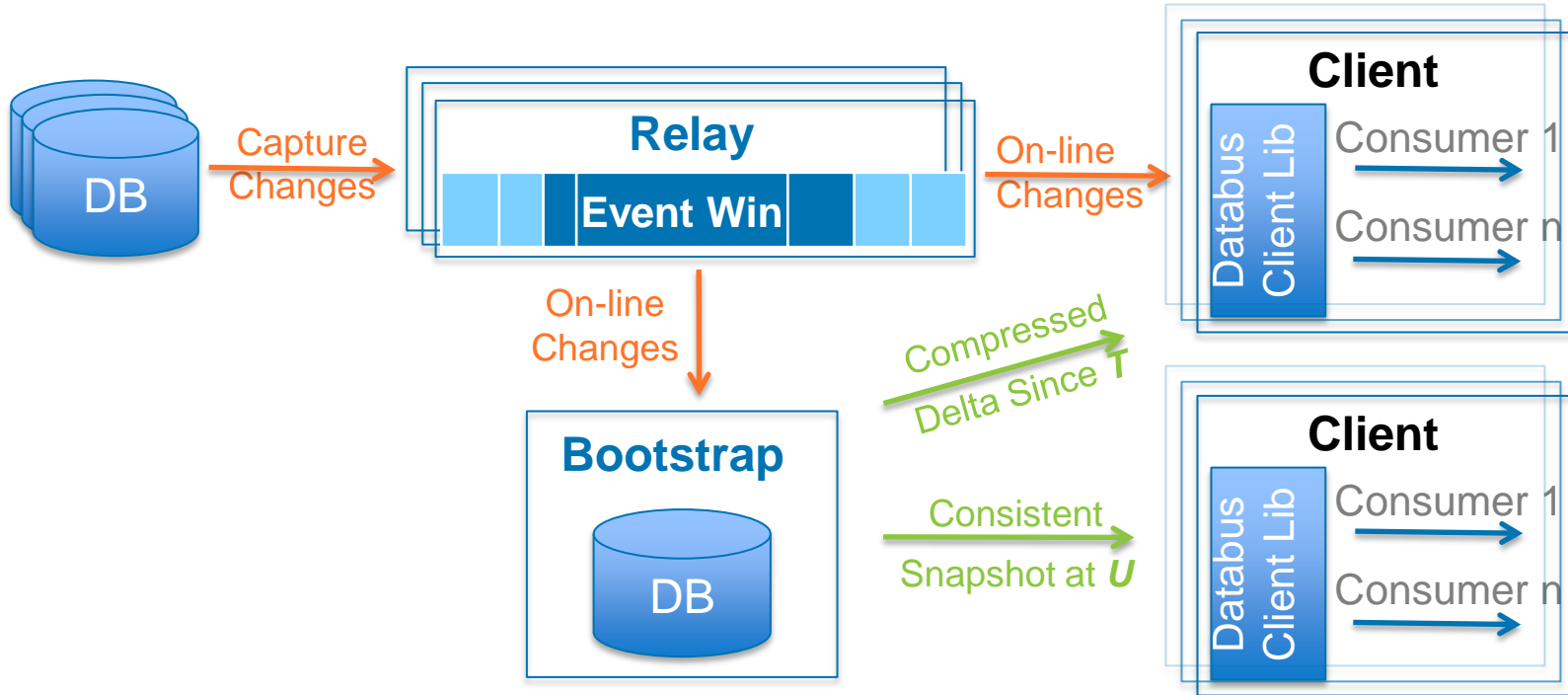
## Guarantees

- At least once delivery
- Very high throughput
- Low latency
- Durability

LinkedIn Data Infrastructure Solutions

## Databus : Timeline-Consistent Change Data Capture

# Databus at LinkedIn



## Features

- Transport independent of data source: Oracle, MySQL, ...
- Portable change event serialization and versioning
- Start consumption from arbitrary point

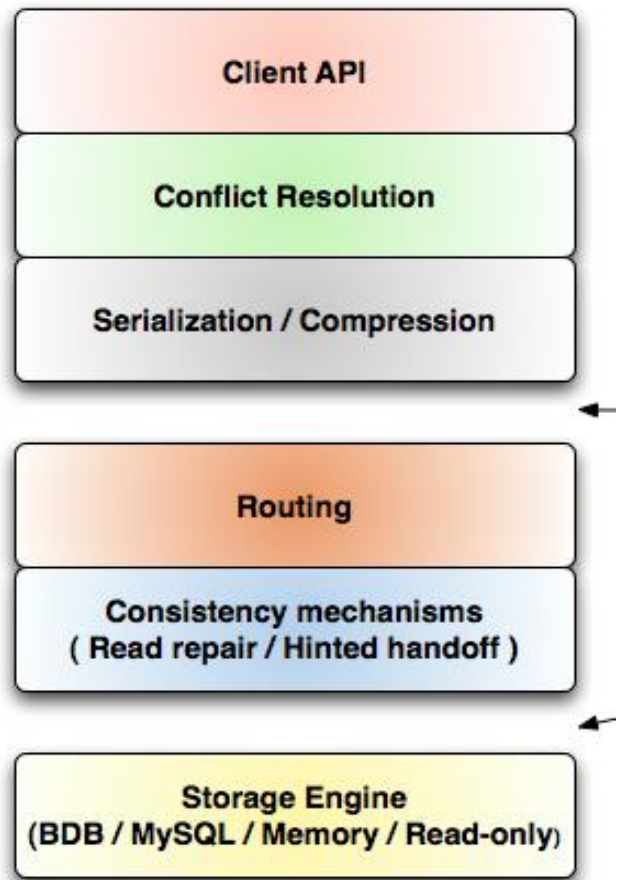
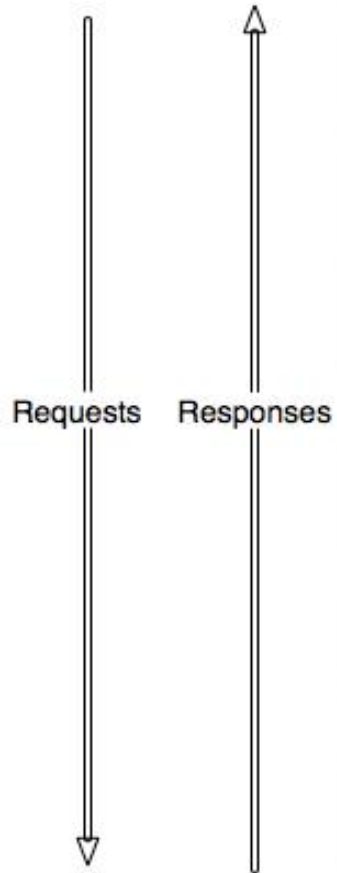
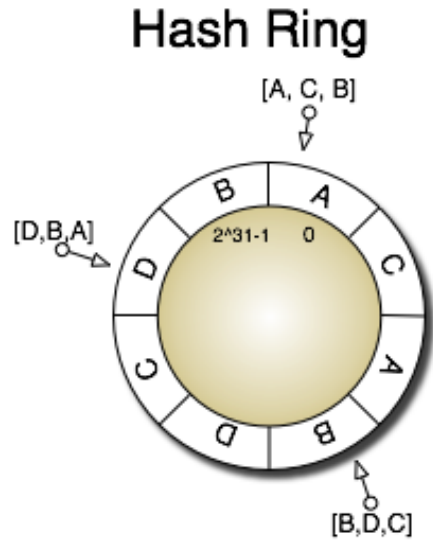
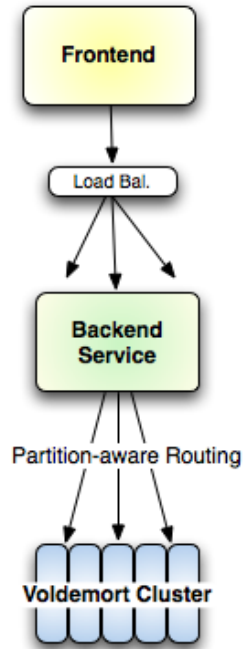
## Guarantees

- Transactional semantics
- Timeline consistency with the data source
- Durability (by data source)
- At-least-once delivery
- Availability
- Low latency

LinkedIn Data Infrastructure Solutions

**Voldemort: Highly-Available Distributed Data Store**

# Voldemort: Architecture



## Highlights

- Open source
- Pluggable components
- Tunable consistency / availability
- Key/value model, server side "views"

## In production

- Data products
- Network updates, sharing, page view tracking, rate-limiting, more...
- Future: SSDs, multi-tenancy

LinkedIn Data Infrastructure Solutions

## Espresso: Indexed Timeline-Consistent Distributed Data Store



# Espresso: Key Design Points

- Hierarchical data model
  - InMail, Forums, Groups, Companies
- Native Change Data Capture Stream
  - Timeline consistency
  - Read after Write
- Rich functionality within a hierarchy
  - Local Secondary Indexes
  - Transactions
  - Full-text search
- Modular and Pluggable
  - Off-the-shelf: MySQL, Lucene, Avro

# Application View

## Mailbox Database

Message Metadata Table

MemberId	MsgId	Value Blob
bob	1	Invitation to join LinkedIn
bob	2	Job opportunity
bob	3	Request for referral
tom	1	Invitation to join LinkedIn
tom	2	Job opportunity

Message Details Table

MemberId	MsgId	Value Blob
bob	1	"Dear Bob,...."
bob	2	"Hello there,...."
bob	3	"Good morning, "
tom	1	"Hi Tom,..."
tom	2	"Interesting opportunity"

Mailbox Aggregates Table

MemberId	Value Blob
bob	unread:20, total:100
tom	unread: 2, total: 25



# Partitioning

Mailbox Database

Message Metadata Table		
MemberId	MsgId	Value Blob
bob	1	Invitation to join LinkedIn
bob	2	Job opportunity
bob	3	Request for referral
tom	1	Invitation to join LinkedIn
tom	2	Job opportunity

Mailbox Aggregates Table	
MemberId	Value Blob
bob	unread:20, total:100
tom	unread: 2, total: 25

Message Details Table		
MemberId	MsgId	Value Blob
bob	1	"Dear Bob,...."
bob	2	"Hello there,...."
bob	3	"Good morning, "
tom	1	"Hi Tom,...."
tom	2	"Interesting opportunity"

Mailbox Database - Partition 1

Message Metadata Table		
MemberId	MsgId	Value Blob
bob	1	Invitation to join LinkedIn
bob	2	Job opportunity
bob	3	Request for referral

Mailbox Aggregates Table	
MemberId	Value Blob
bob	unread:20, total:100

Message Details Table		
MemberId	MsgId	Value Blob
bob	1	"Dear Bob,...."
bob	2	"Hello there,...."
bob	3	"Good morning, "

Mailbox Database - Partition 2

Message Metadata Table		
MemberId	MsgId	Value Blob
tom	1	Invitation to join LinkedIn
tom	2	Job opportunity

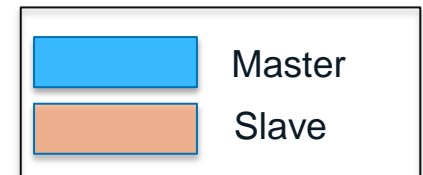
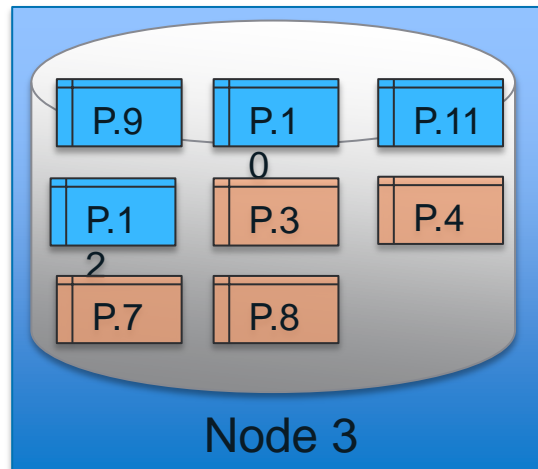
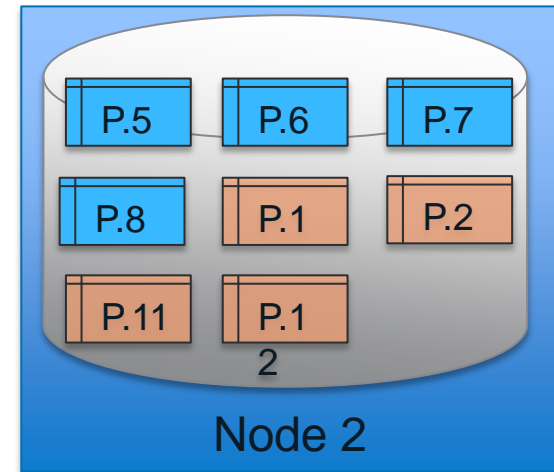
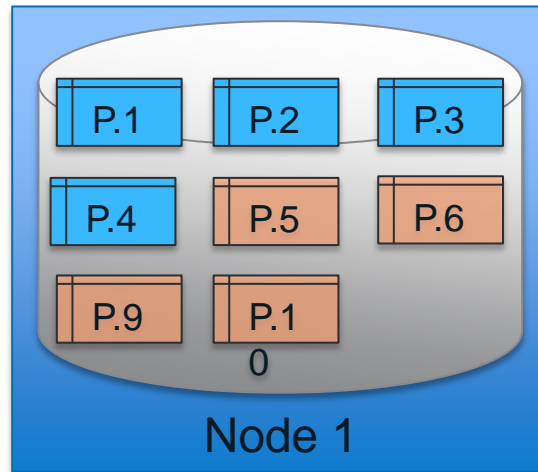
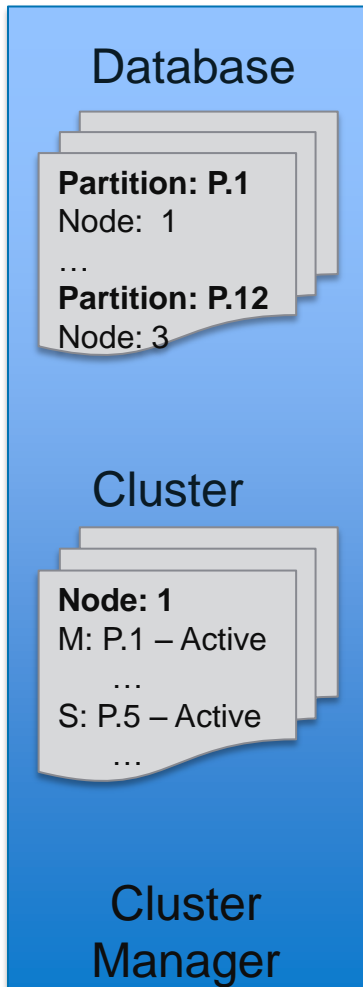
Mailbox Aggregates Table	
MemberId	Value Blob
tom	unread: 2, total: 25

Message Details Table		
MemberId	MsgId	Value Blob
tom	1	"Hi Tom,...."
tom	2	"Interesting opportunity"

# Partition Layout: Master, Slave

3 Storage Engine nodes, 2 way replication



# Espresso: API

- REST over HTTP
- *Get Messages for bob*
  - GET /MailboxDB/MessageMeta/bob
- *Get MsgId 3 for bob*
  - GET /MailboxDB/MessageMeta/bob/3
- *Get first page of Messages for bob that are unread and in the inbox*
  - GET /MailboxDB/MessageMeta/bob/?query="+isUnread:true  
+isInbox:true"&start=0&count=15

# Espresso: API Transactions

- Add a message to bob's mailbox
  - transactionally update mailbox aggregates, insert into metadata and details.

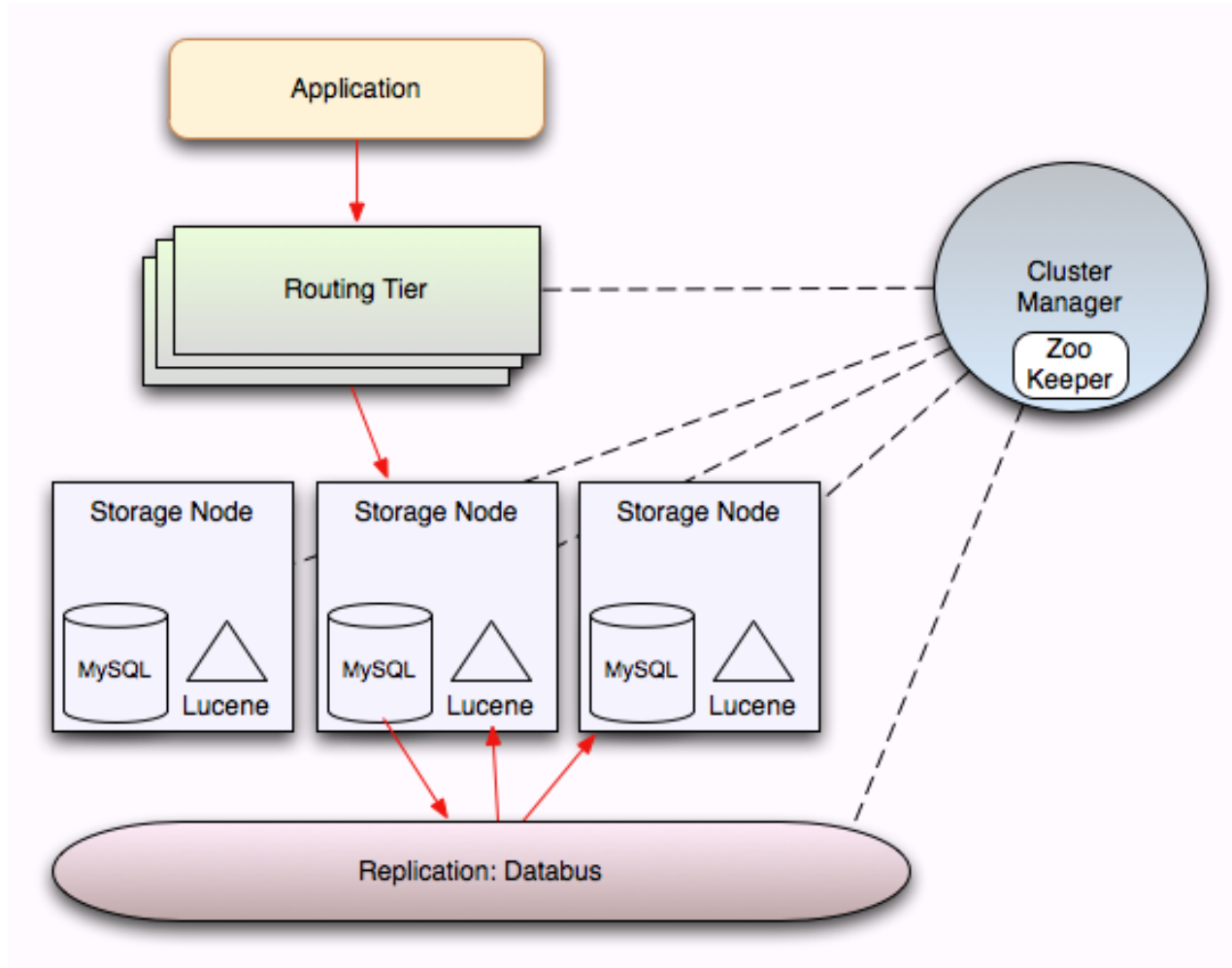
```
POST /MailboxDB/*/bob HTTP/1.1
Content-Type: multipart/binary; boundary=1299799120
Accept: application/json
--1299799120
Content-Type: application/json
Content-Location: /MailboxDB/MessageStats/bob
Content-Length: 50
{"total":"+1", "unread":"+1"}

--1299799120
Content-Type: application/json
Content-Location: /MailboxDB/MessageMeta/bob
Content-Length: 332
{"from":"...", "subject":"...", ...}

--1299799120
Content-Type: application/json
Content-Location: /MailboxDB/MessageDetails/bob
Content-Length: 542
{"body":"..."}

--1299799120-
```

# Espresso: System Components



# Espresso @ LinkedIn

- First applications
  - Company Profiles
  - InMail
- Next
  - Unified Social Content PI
  - Member Profiles
  - Many more...

The screenshot shows the LinkedIn Messages inbox for a user. At the top, there are navigation tabs for 'Overview', 'Careers', 'Products & Services', and 'Analytics'. Below these, there are buttons for 'Messages' (with a count of 4) and 'Invitations' (with a count of 35). A search bar labeled 'Search Inbox' and a 'Compose Message' button are also visible. The main area displays a list of messages, each with a checkbox, a profile picture, the sender's name, the subject of the message, and the date. The messages listed are:

Select	Sender	Subject	Date
<input type="checkbox"/>	Bill Umoff	Tech lead opportunity with Quantcast (RTB platform)	Oct 5
<input type="checkbox"/>	Musaab At-Taras	Innovate Invitation	Oct 5
<input type="checkbox"/>	Jackie Moore	Seeking Senior Level Engineer/Architect	Sep 26
<input type="checkbox"/>	Niloy Mukherjee (Responded)	RE: Hi !	Sep 19
<input type="checkbox"/>	Pocket Gems	Sequoia's Hottest Start-up Seeks Top Engineers	Sep 15
<input type="checkbox"/>	Alexander Lawrence	Developing Software for International Markets - Alexander Lawrence invites you to join!	Sep 6
<input type="checkbox"/>	M.P. Singh (Replied)	Hi	Aug 31
<input type="checkbox"/>	Kevin Lane	Software Design at WhatsApp, Inc.	Aug 2
<input type="checkbox"/>	Linas Mikalcius	RE: Follow Up From Riviera Partners	Jun 30

## Espresso: Next steps

- Launched first application Oct 2011
- Open source 2012
- Multi-Datacenter support
- Log-structured storage
- Time-partitioned data

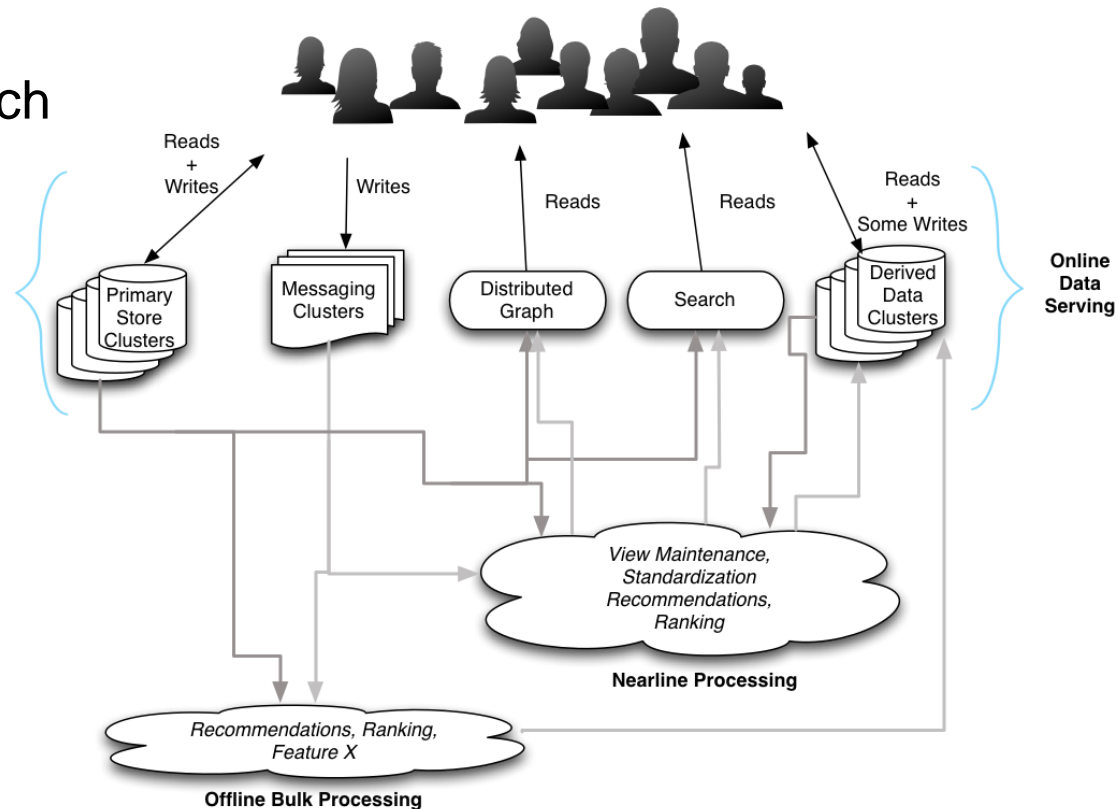
# Outline

- LinkedIn Products
- Data Ecosystem
- LinkedIn Data Infrastructure Solutions
- **Next Play**



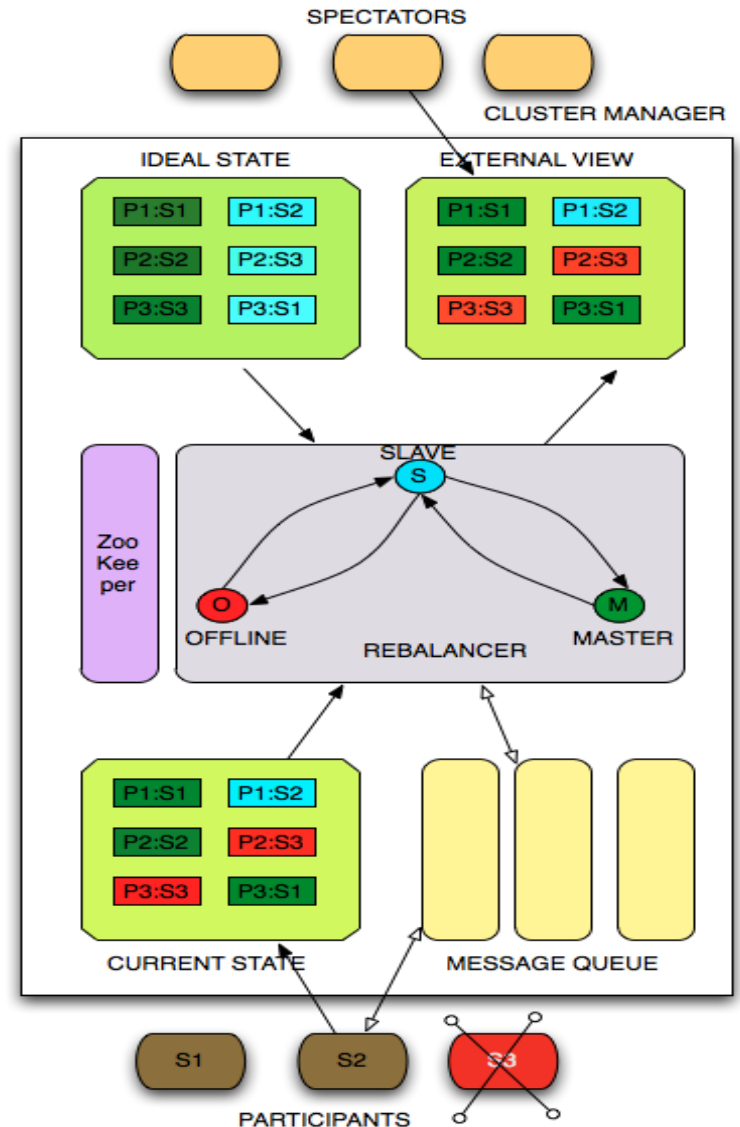
# The Specialization Paradox in Distributed Systems

- Good: Build specialized systems so you can do each thing really well
- Bad: Rebuild distributed routing, failover, cluster management, monitoring, tooling



# Generic Cluster Manager: Helix

- Generic Distributed State Model
- Centralized Config Management
- Automatic Load Balancing
- Fault tolerance
- Health monitoring
- Cluster expansion and rebalancing
- Open Source 2012
- Espresso, Databus and Search



## Stay tuned for

- Innovation

- Nearline processing
- Espresso eco-system
- Storage / indexing
- Analytics engine
- Search

- Convergence

- Building blocks for distributed data management systems

Thanks!

# Appendix

# Espresso: Routing

- Router is a high-performance HTTP proxy
- Examines URL, extracts partition key
- Per-db routing strategy
  - Hash Based
  - Route To Any (for schema access)
  - Range (future)
- Routing function maps partition key to partition
- Cluster Manager maintains mapping of partition to hosts:
  - Single Master
  - Multiple Slaves

# Espresso: Storage Node

- Data Store (MySQL)
  - Stores document as Avro serialized blob
  - Blob indexed by (partition key {, sub-key})
  - Row also contains limited metadata
    - Etag, Last modified time, Avro schema version
- Document Schema specifies per-field index constraints
- Lucene index per partition key / resource

# Espresso: Replication

- MySQL replication of mastered partitions
- MySQL “Slave” is MySQL instance with custom storage engine
  - custom storage engine just publishes to databus
- Per-database commit sequence number
- Replication is Databus
  - Supports existing downstream consumers
- Storage node consumes from Databus to update secondary indexes and slave partitions