Data Infrastructure at LinkedIn

Lei Gao

http://www.linkedin.com/in/gaolei

Outline

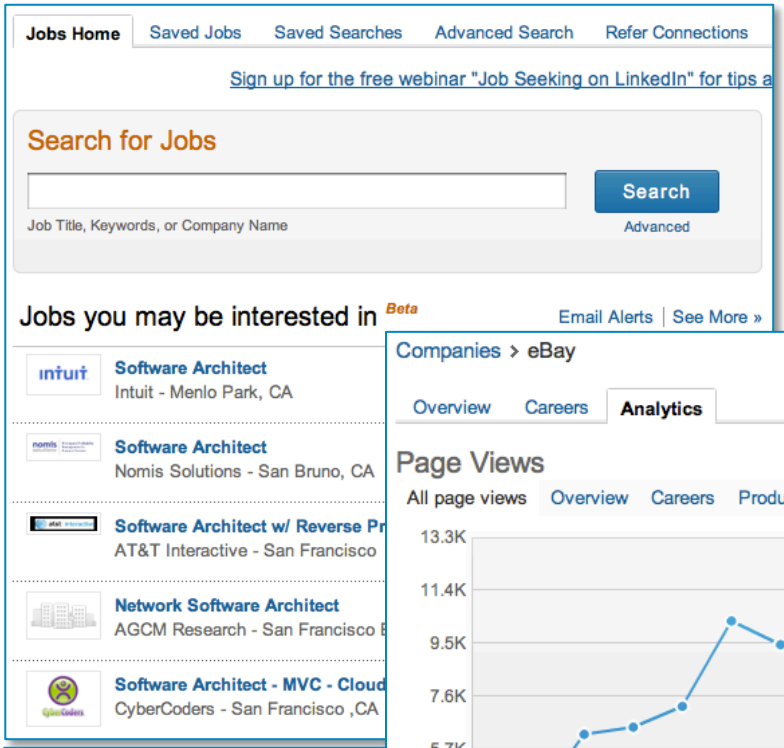
- LinkedIn Products
- Data Ecosystem
- LinkedIn Data Infrastructure Solutions
- Next Play

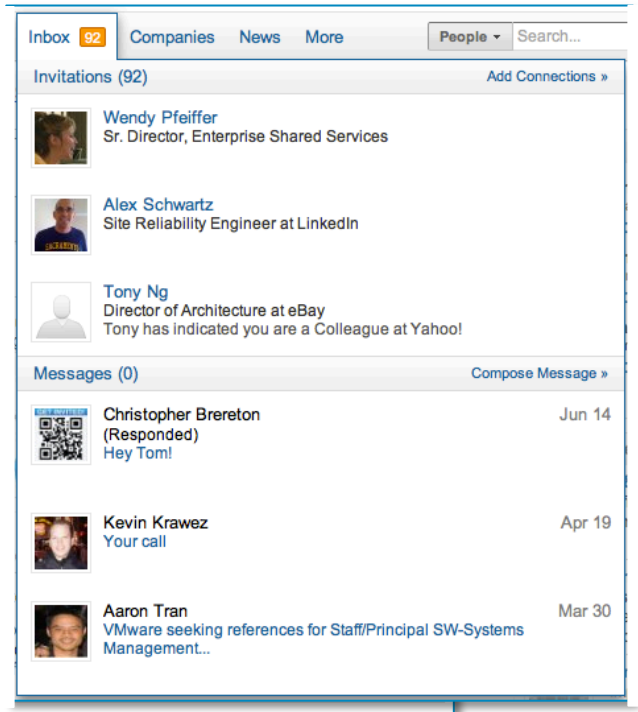
LinkedIn By The Numbers

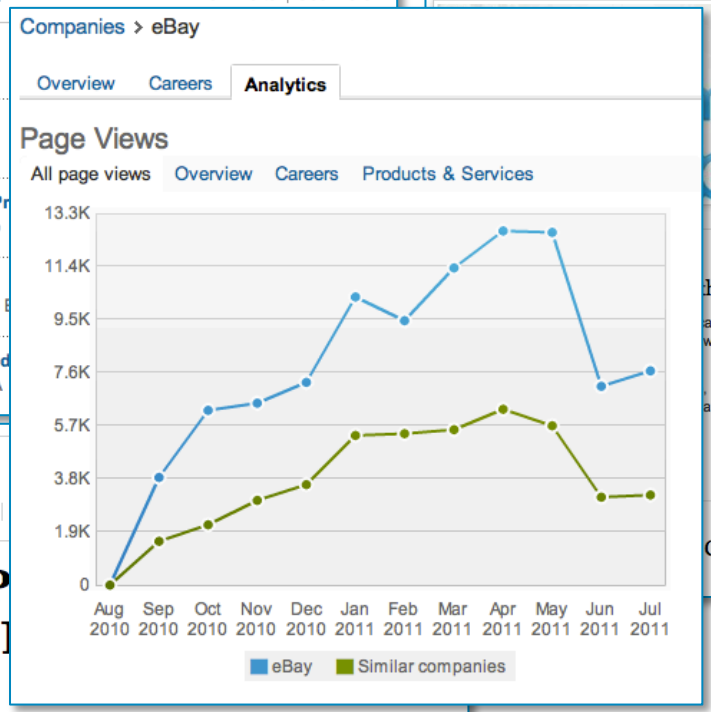
- 150M + users*
- ~ 4.2B People Searches in 2011**
- >2M companies with LinkedIn Company Pages**
- 16 languages
- 75% of Fortune 100 Companies use LinkedIn to hire***

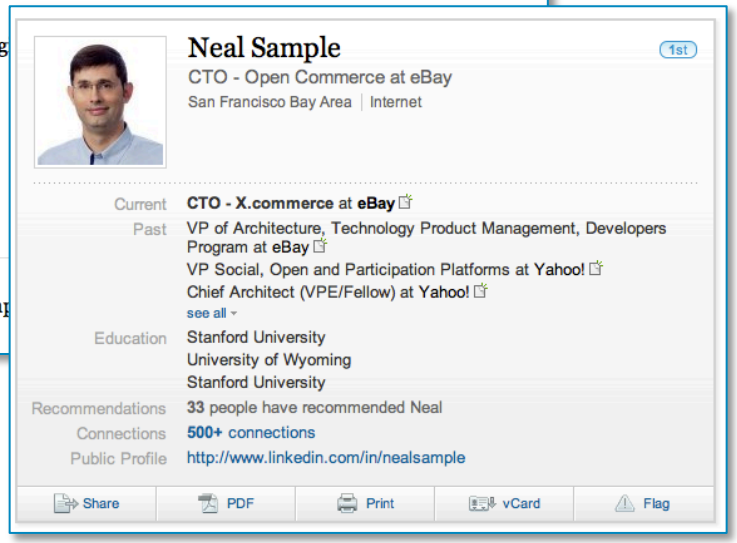
* As of February 9th 2012
** As of December 31st 2011
*** As of September 30th 2011

Broad Range of Products & Services

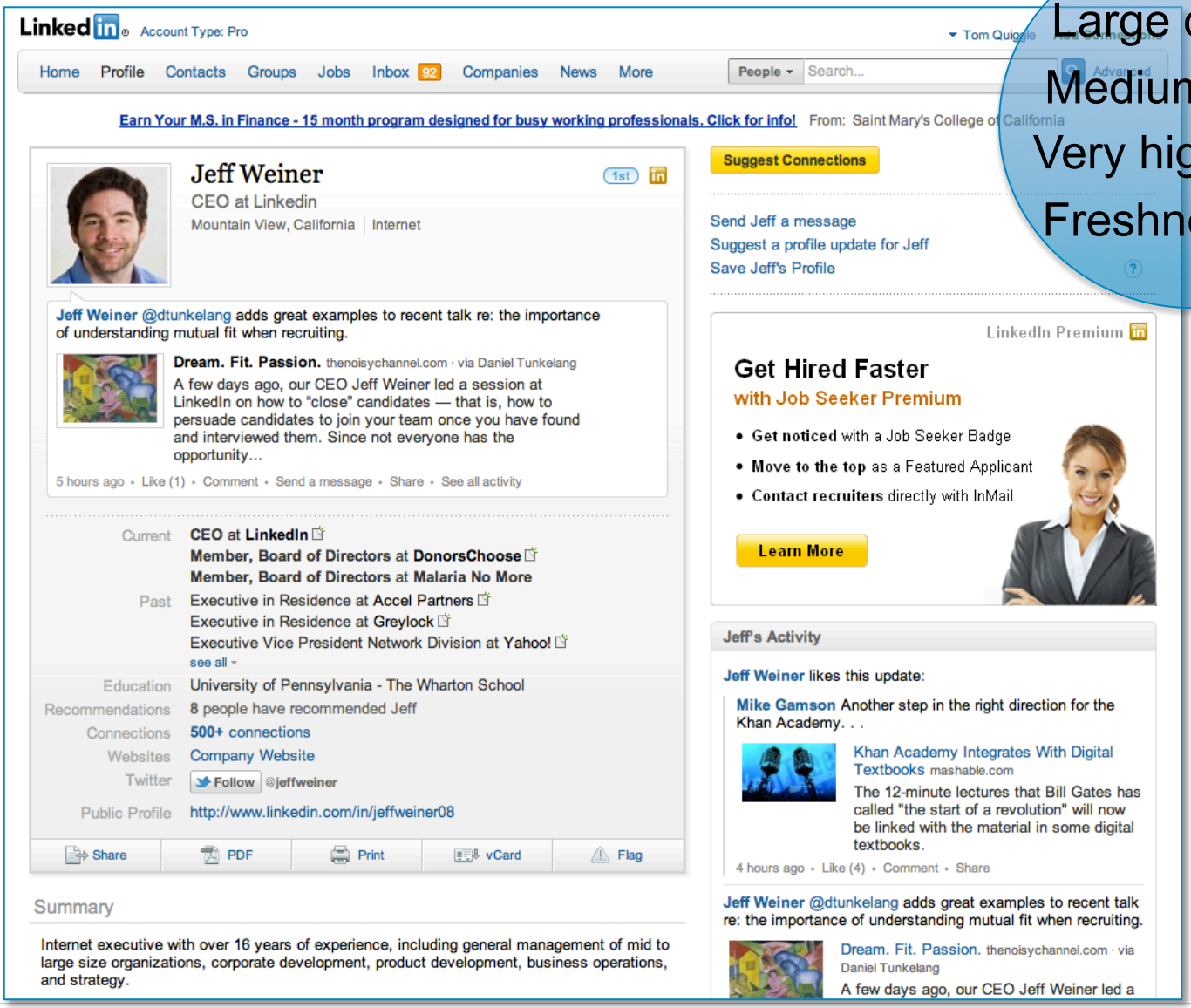








User Profiles



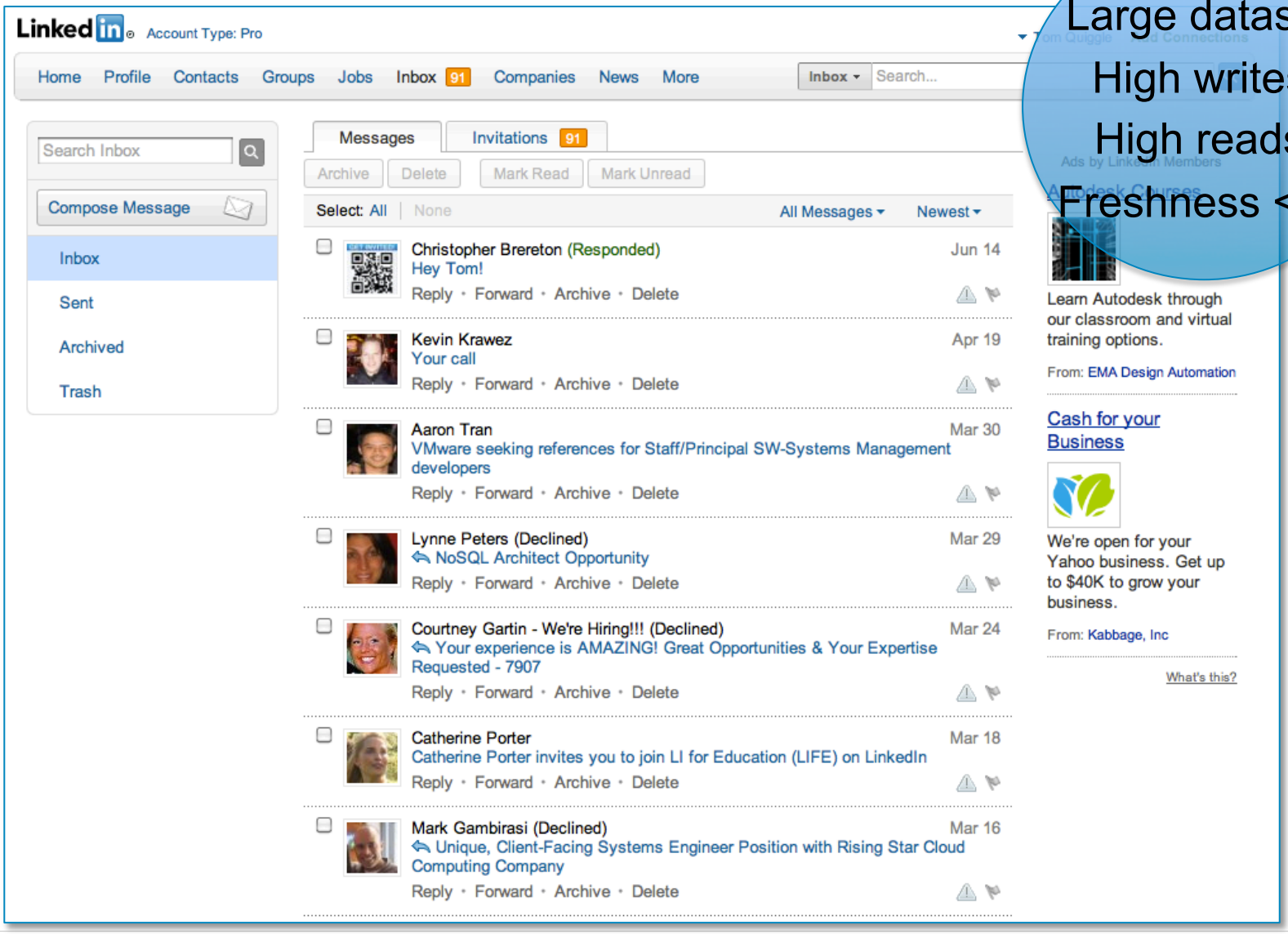
Large dataset

Medium writes

Very high reads

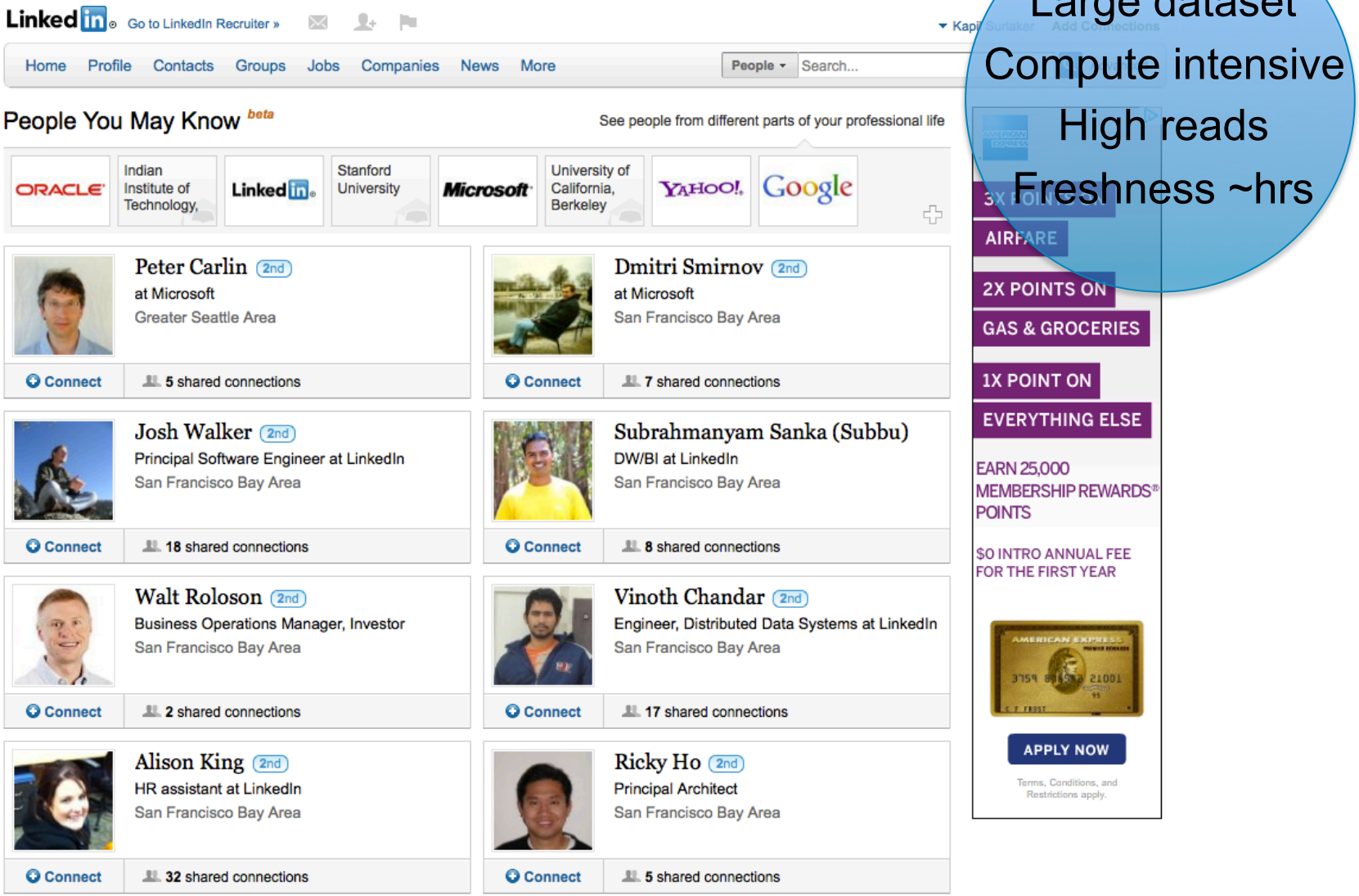
Freshness <1s

Communications



Large dataset
High writes
High reads
Freshness <1s

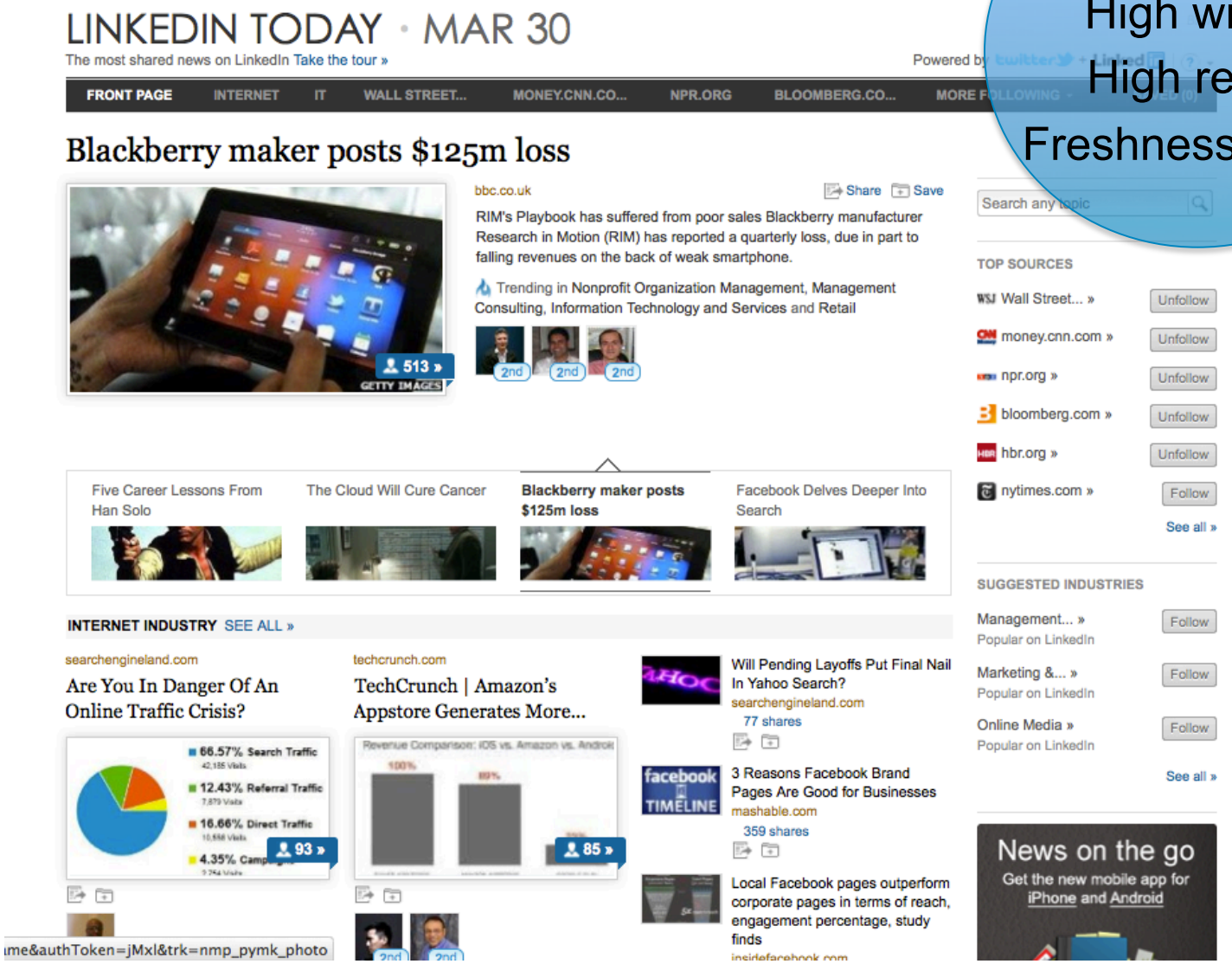
People You May Know



Large dataset
Compute intensive
High reads
Freshness ~hrs

LinkedIn Today

Moving dataset
High writes
High reads
Freshness ~mins



Outline

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

Three Paradigms : Simplifying the Data Continuum

- **Member Profiles**
- Company Profiles
- Connections
- Communications

Online

Activity that should be reflected immediately

- **Linkedin Today**
- Profile Standardization
- News
- Recommendations
- Search
- Communications

Nearline

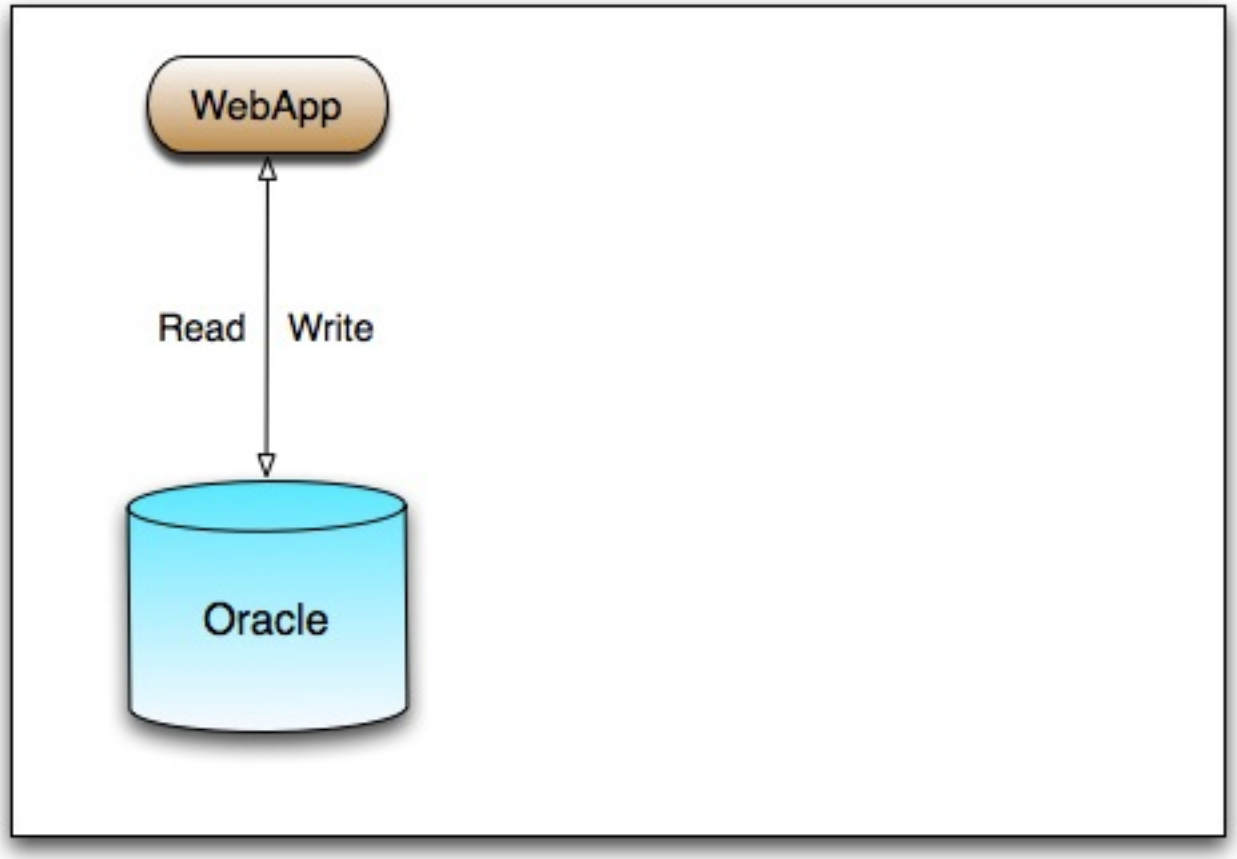
Activity that should be reflected soon

- **People You May Know**
- Connection Strength
- News
- Recommendations
- Next best idea

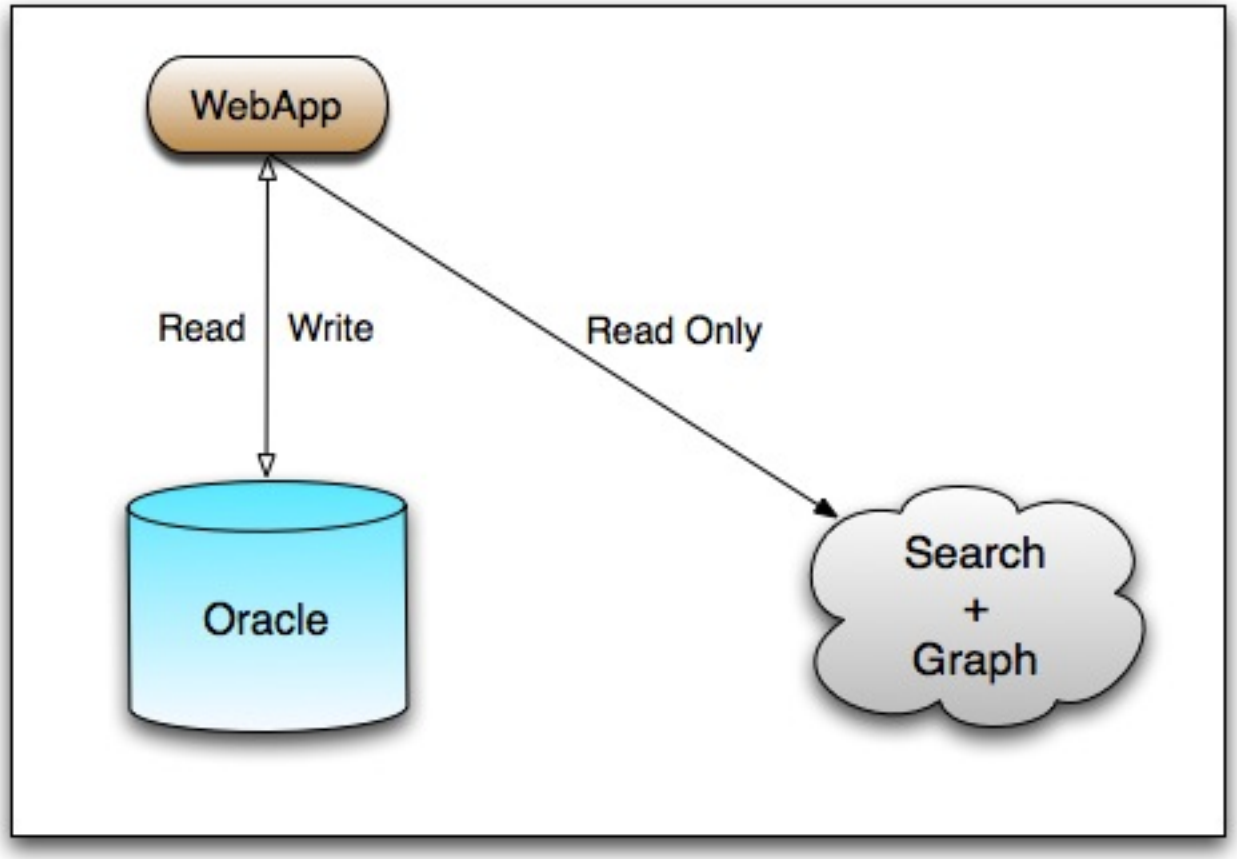
Offline

Activity that can be reflected later

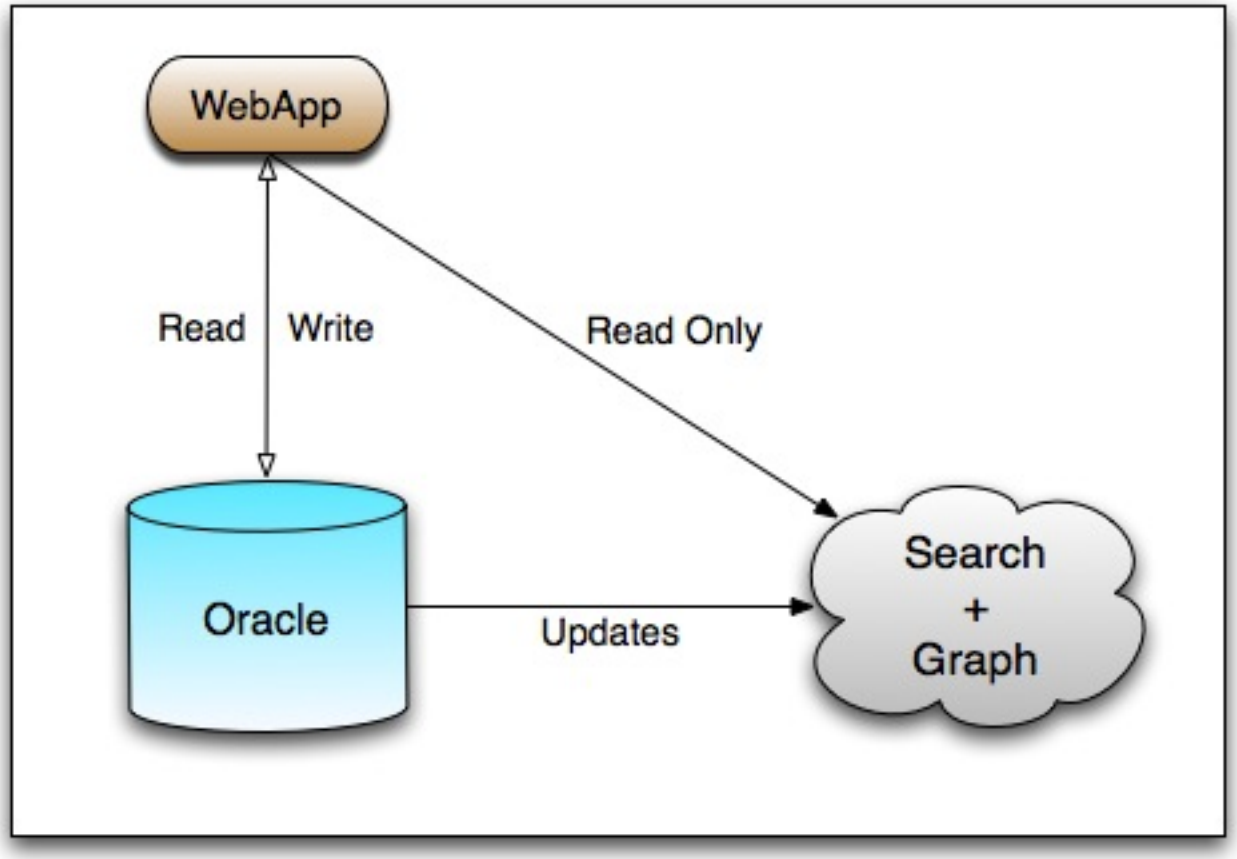
LinkedIn Product Architecture



LinkedIn Product Architecture



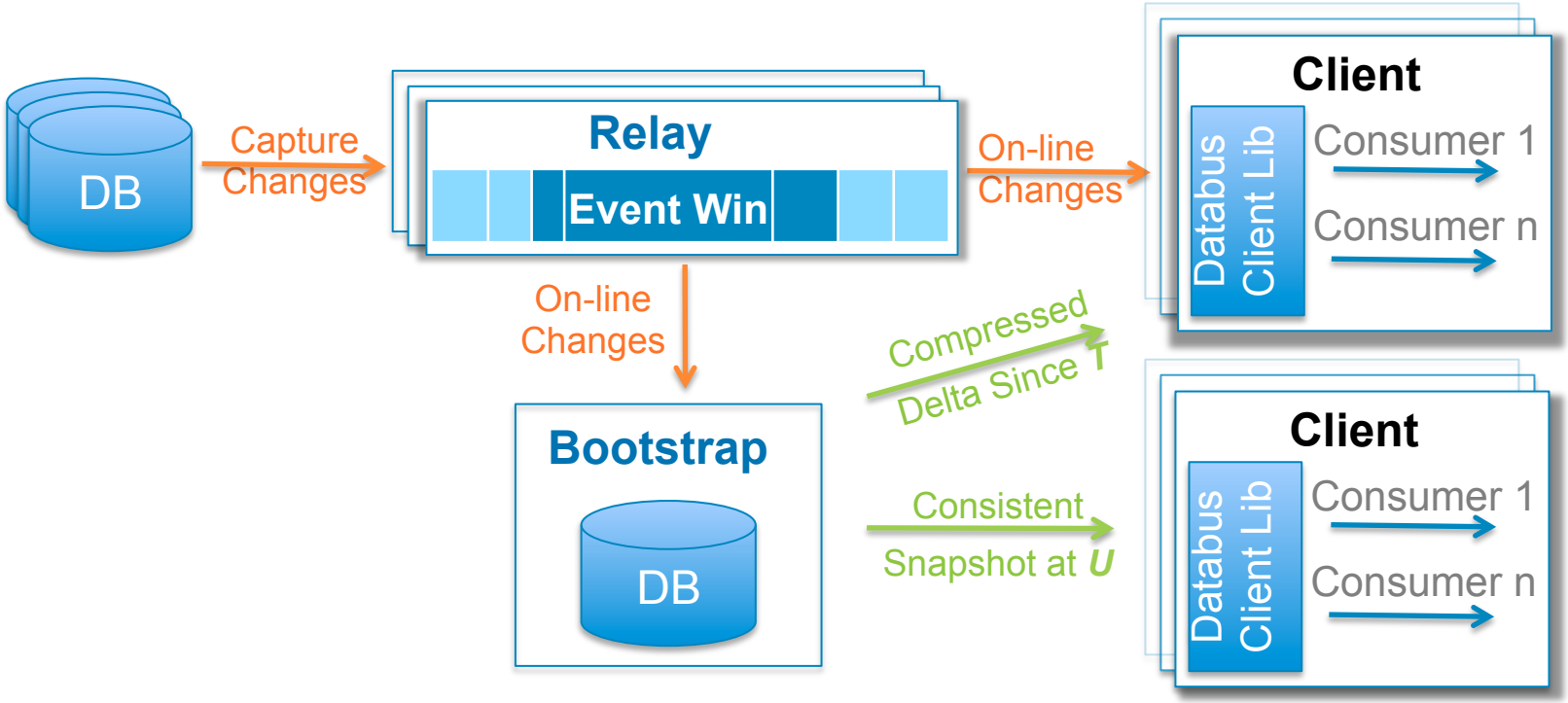
LinkedIn Product Architecture



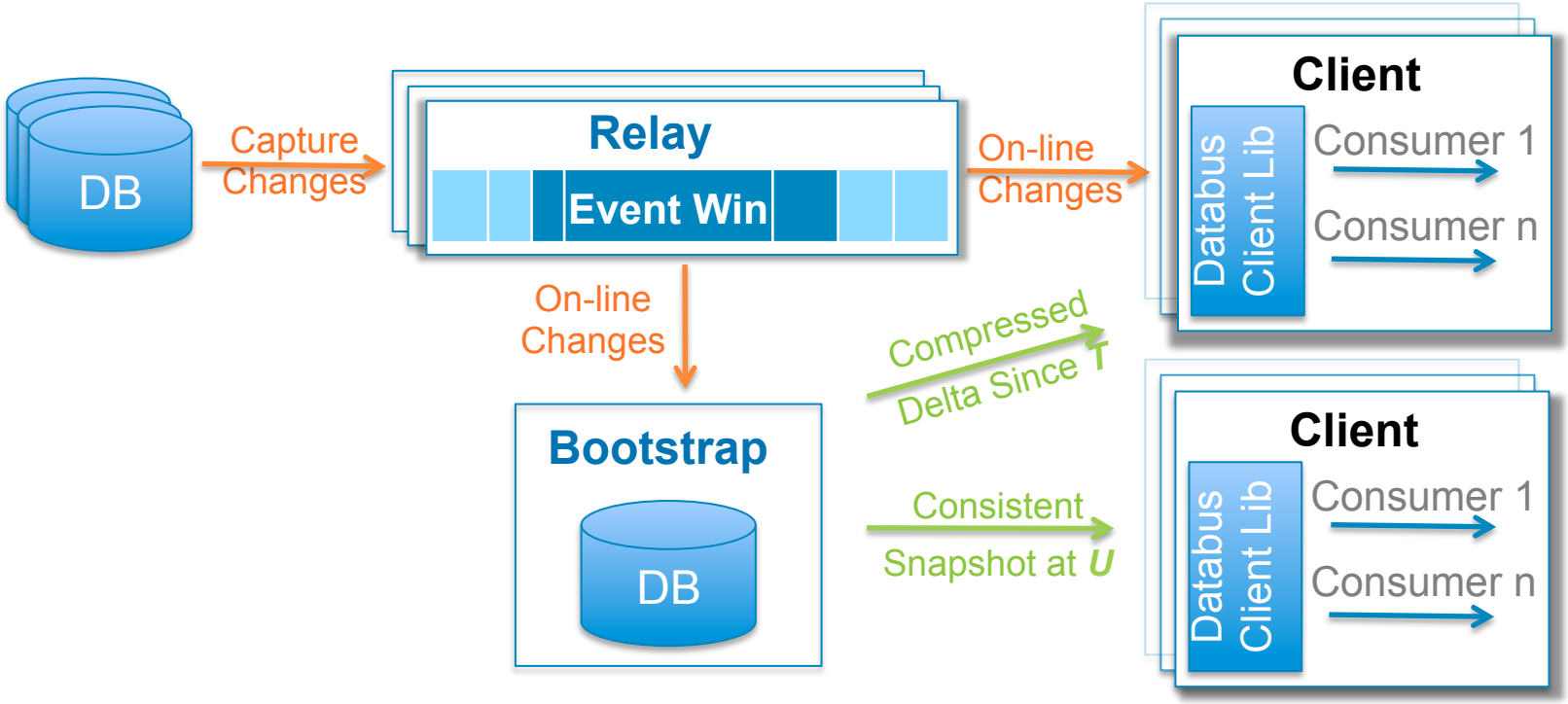
LinkedIn Data Infrastructure Solutions

Databus : Timeline-Consistent Change Data Capture

Databus at LinkedIn



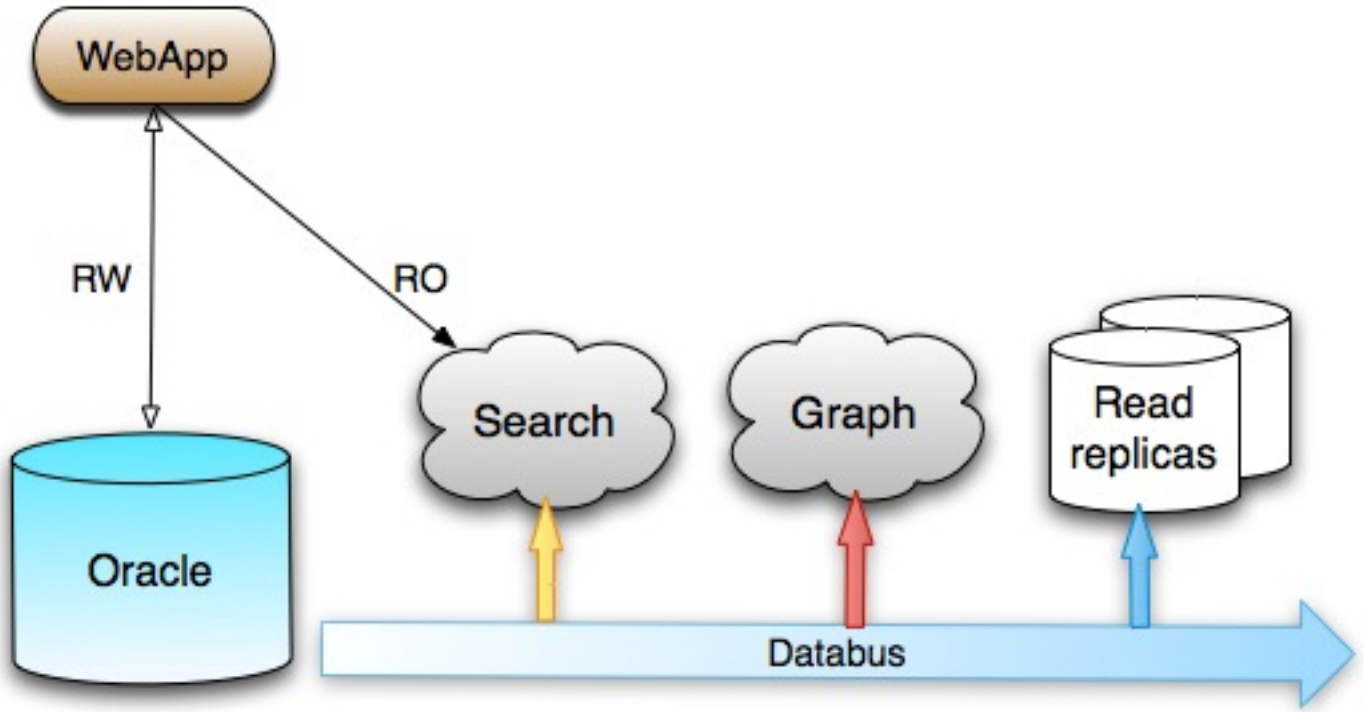
Databus at LinkedIn



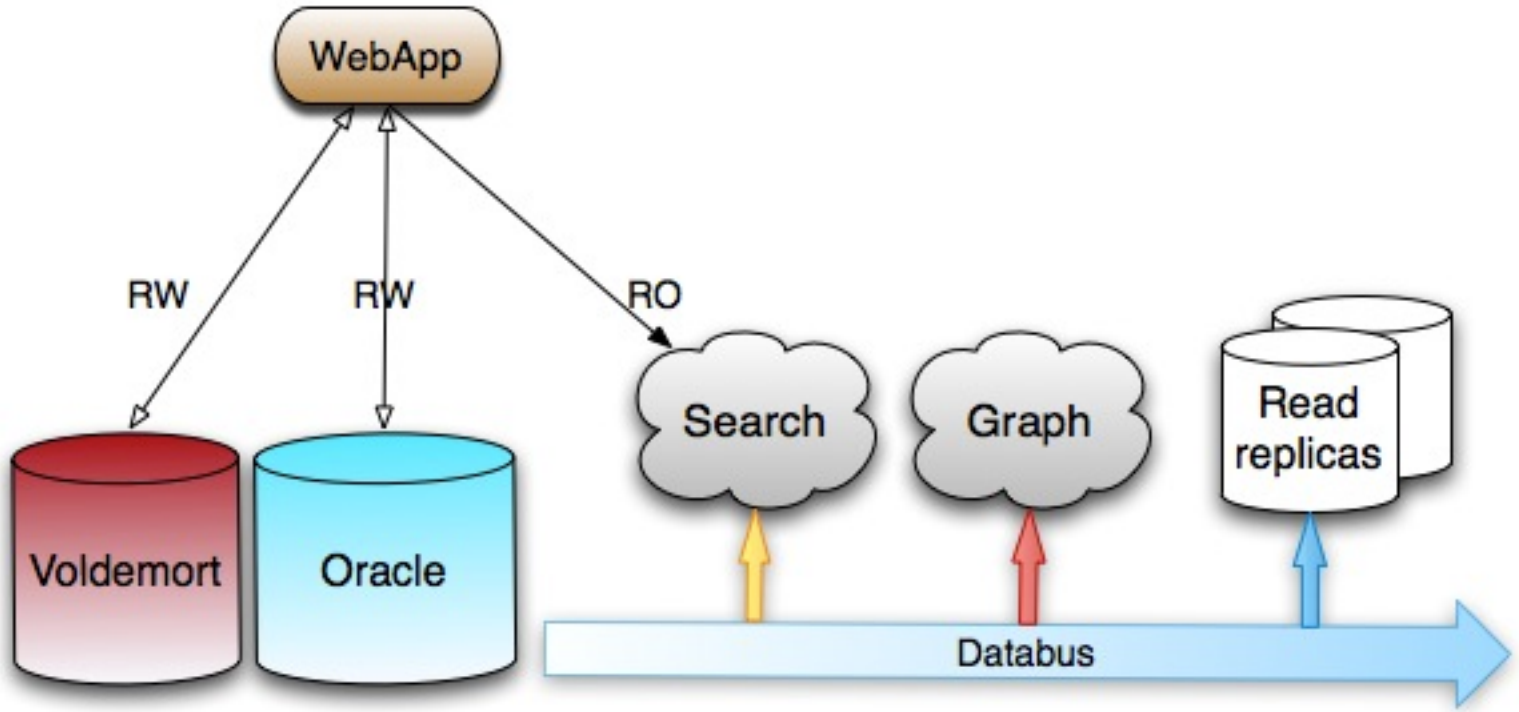
- Transport independent of data source: Oracle, MySQL, …
- Transactional semantics
- In order, at least once delivery

- Tens of relays
- Hundreds of sources
- Low latency - milliseconds

LinkedIn Product Architecture



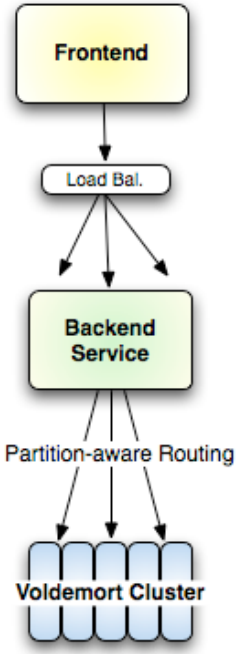
LinkedIn Product Architecture

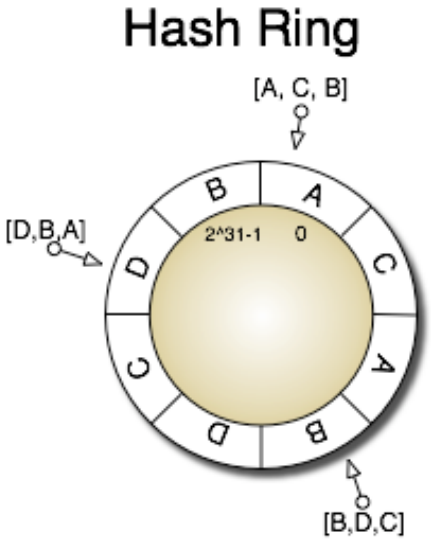


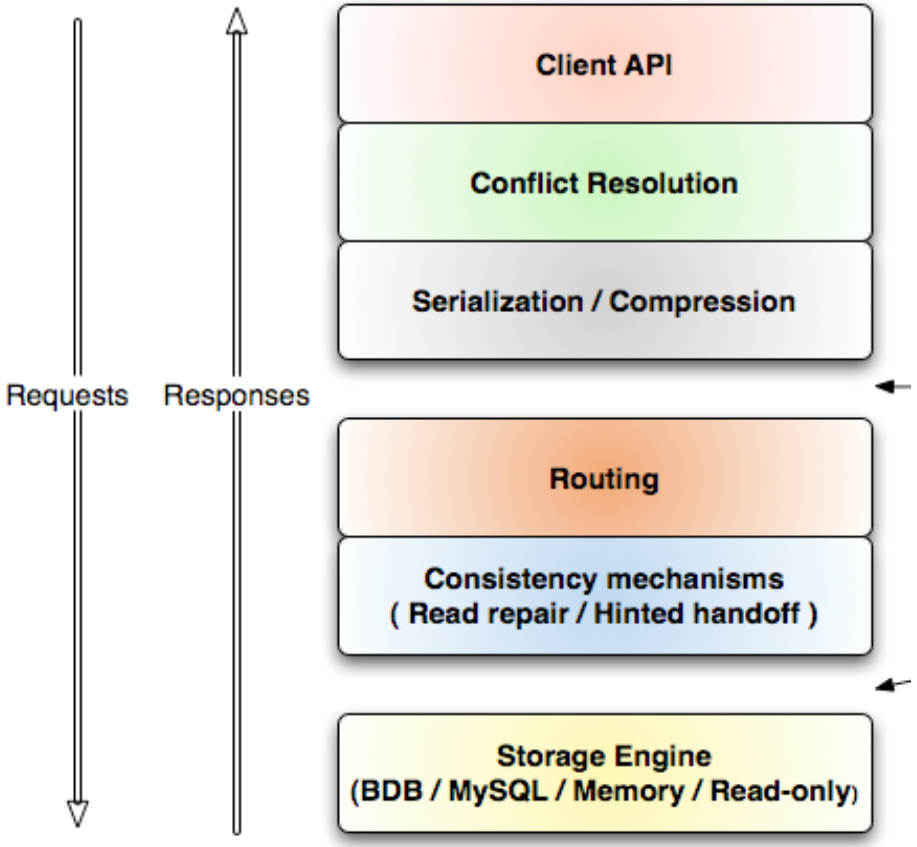
LinkedIn Data Infrastructure Solutions

Voldemort: Highly-Available Distributed KV Store

Voldemort: Architecture



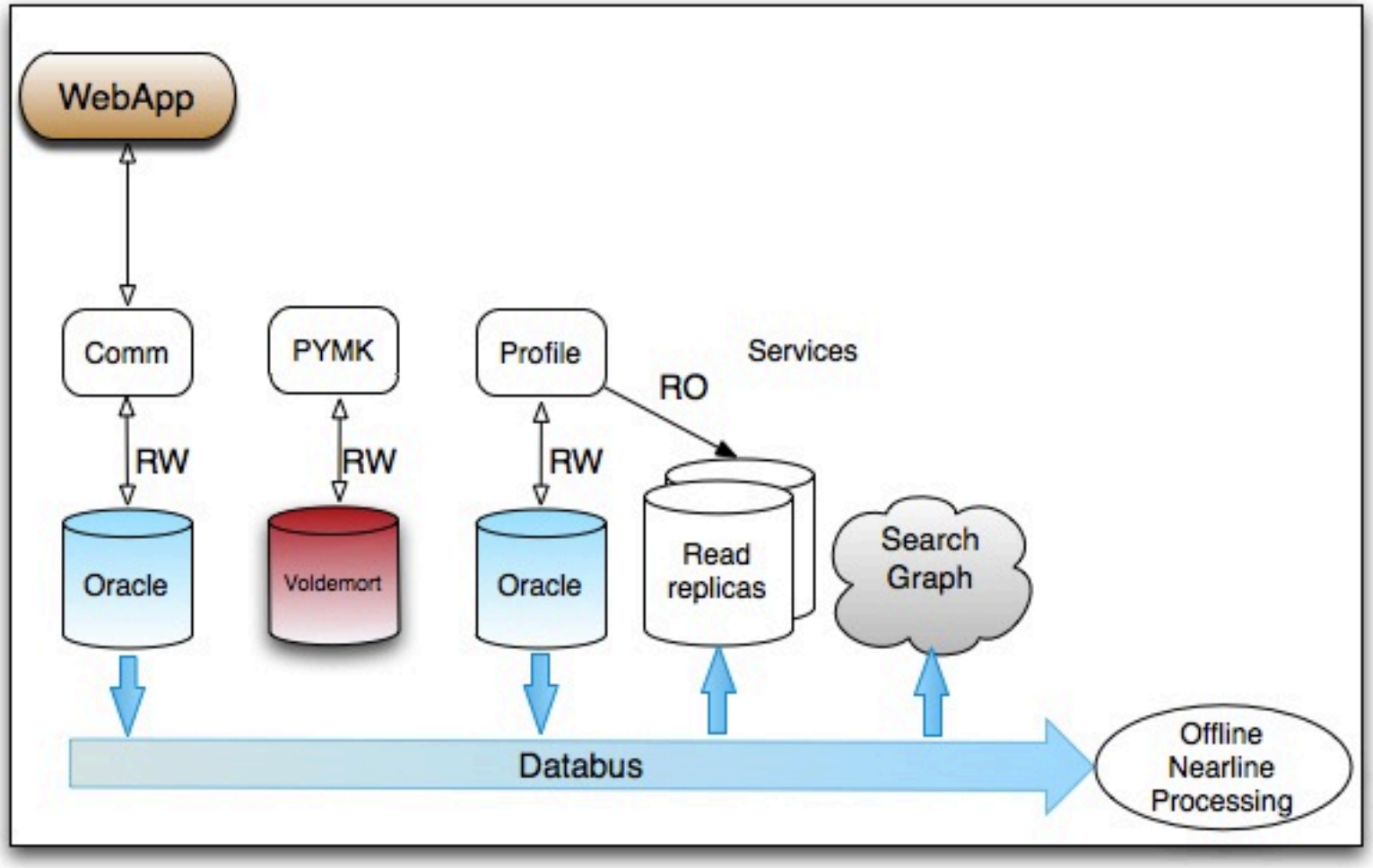




- Pluggable components
- Tunable consistency / availability
- Key/value model, server side “views”

- 10 clusters, 100+ nodes
- Largest cluster – 10K+ qps
- Avg latency: 3ms
- Hundreds of Stores
- Largest store – 2.8TB+

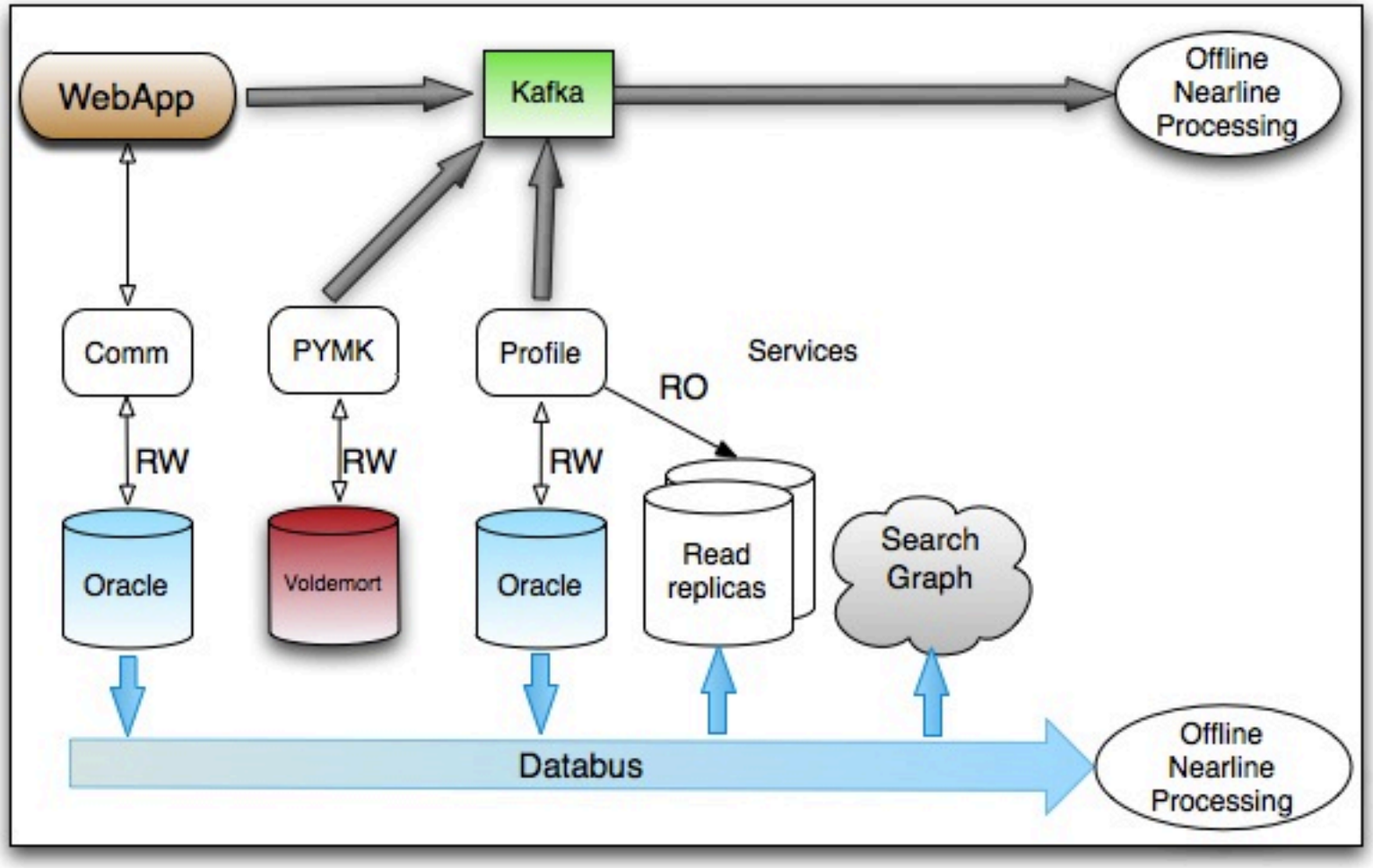
LinkedIn Product Architecture



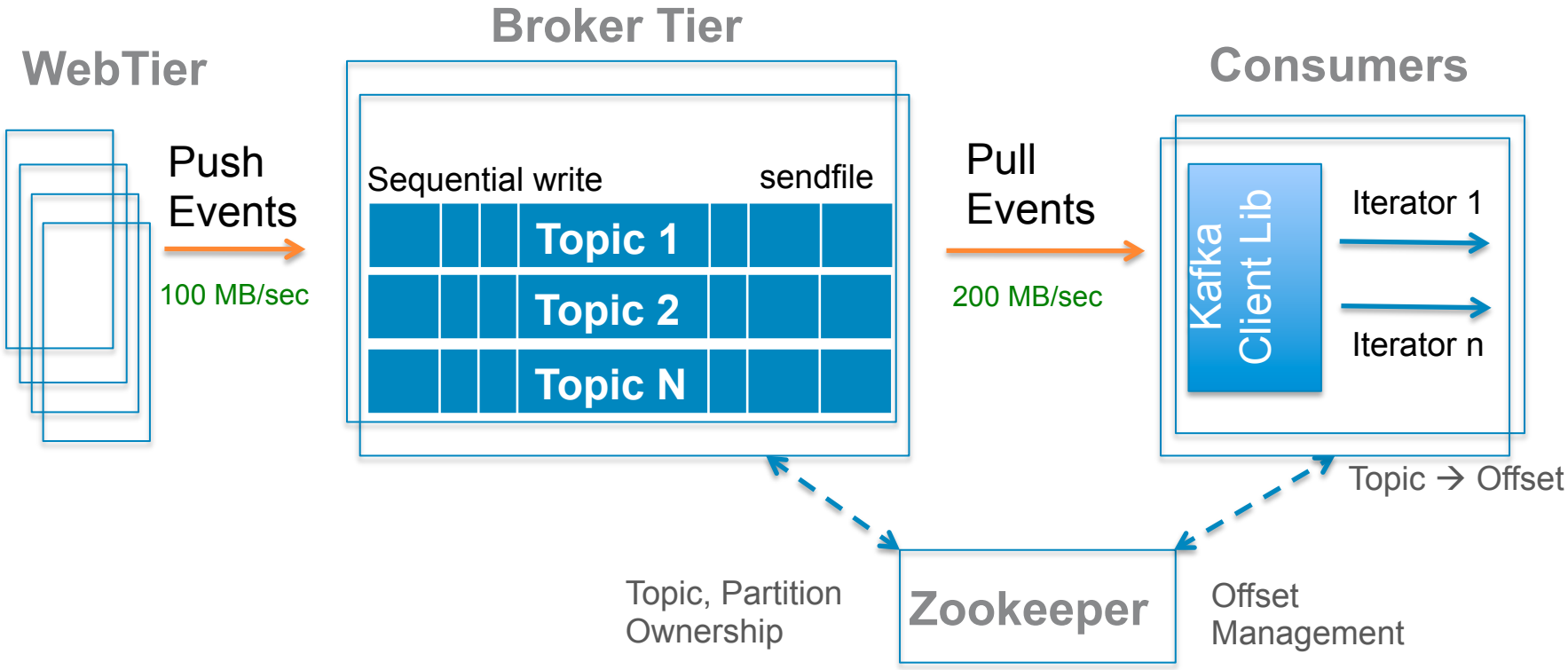
LinkedIn Data Infrastructure Solutions

Kafka: High-Volume Low-Latency Messaging System

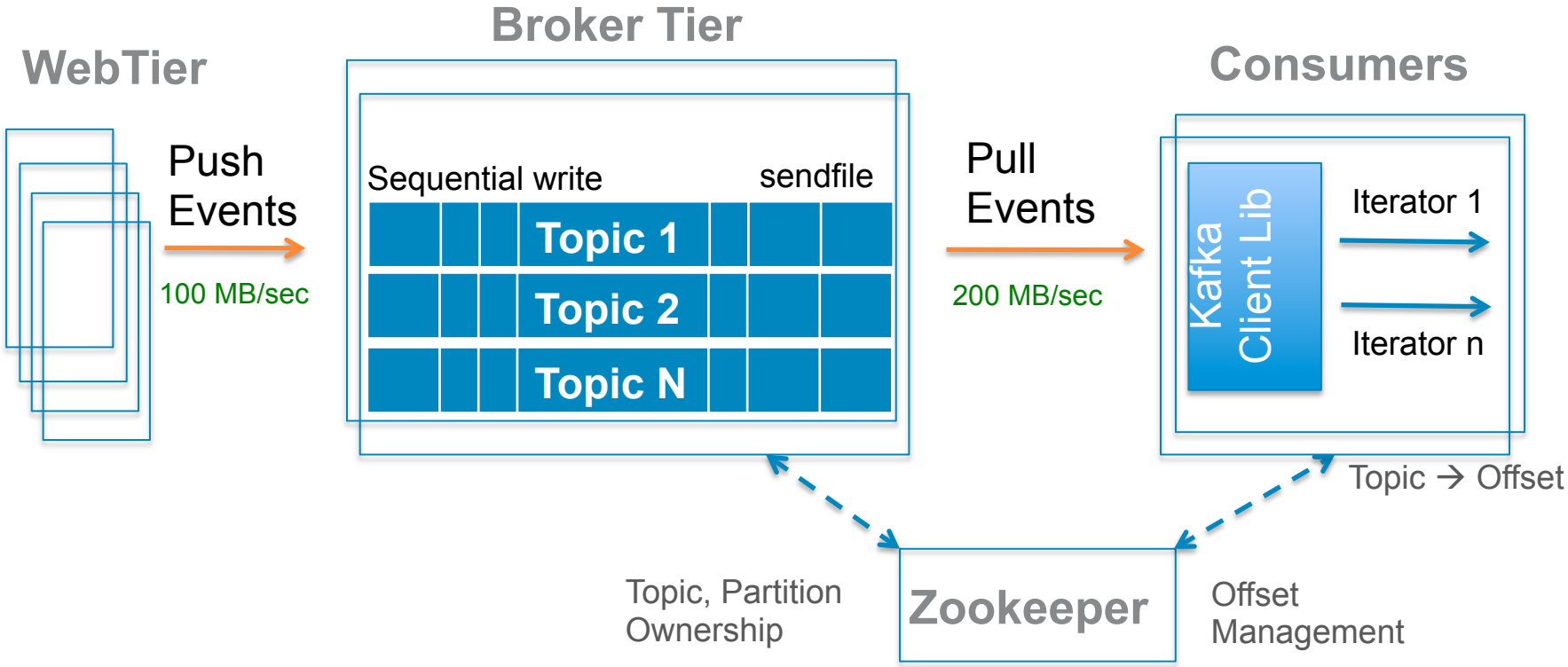
LinkedIn Product Architecture



Kafka: Architecture



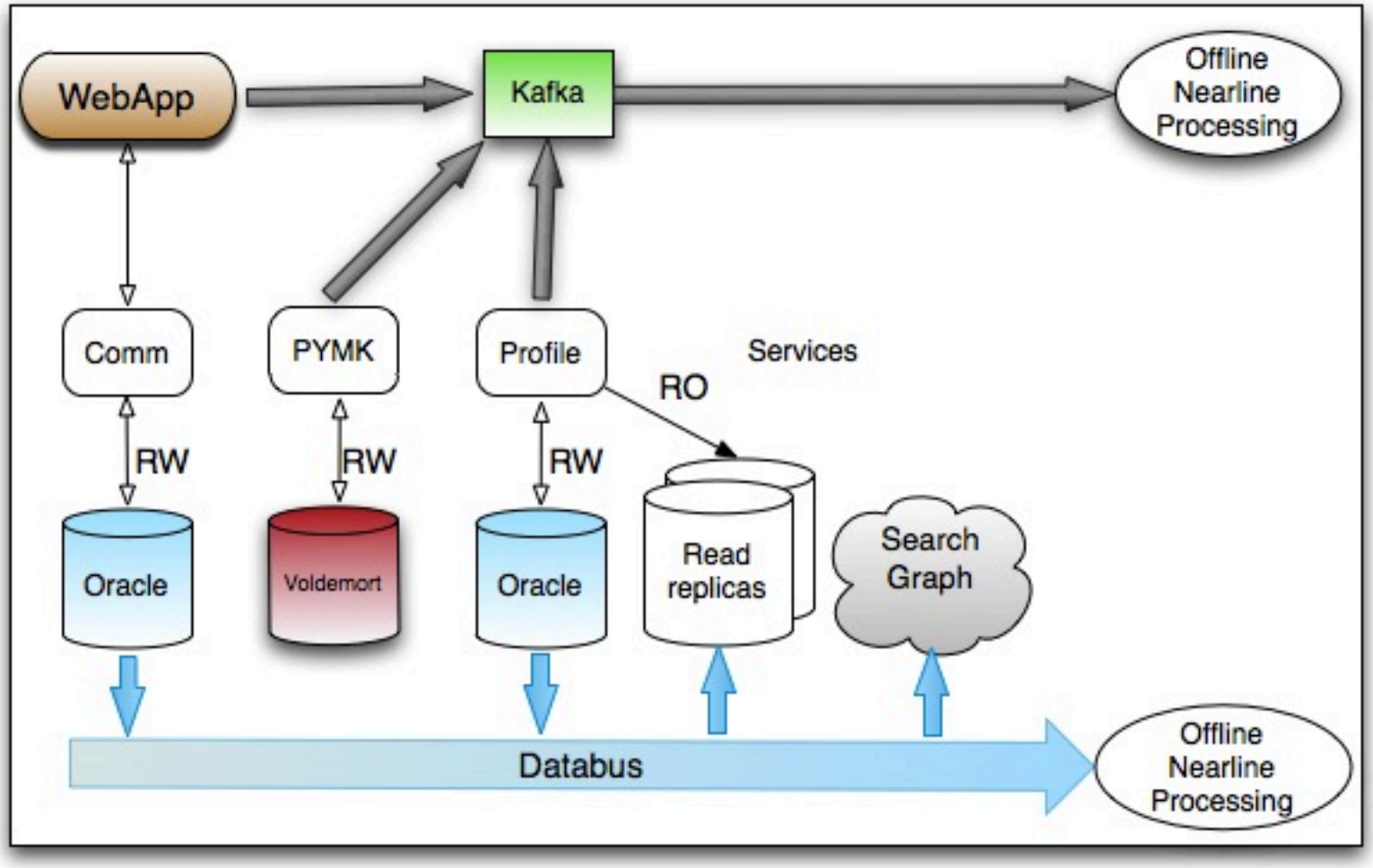
Kafka: Architecture



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

- Billions of Events, TBs per day
- 50K+ per sec at peak
- Inter and Intra-cluster replication
- End-to-end latency: few seconds

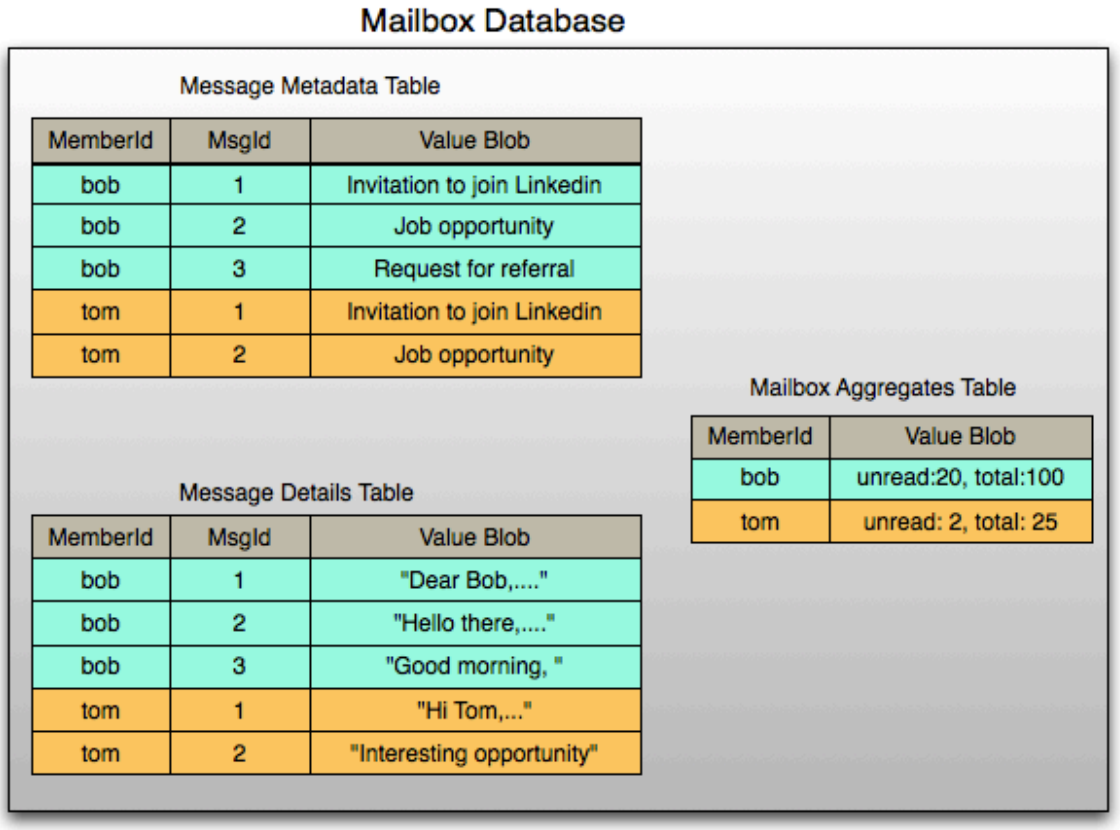
LinkedIn Product Architecture



LinkedIn Data Infrastructure Solutions

Espresso: Indexed Timeline-Consistent Distributed Data Store

Application View



Hierarchical data model

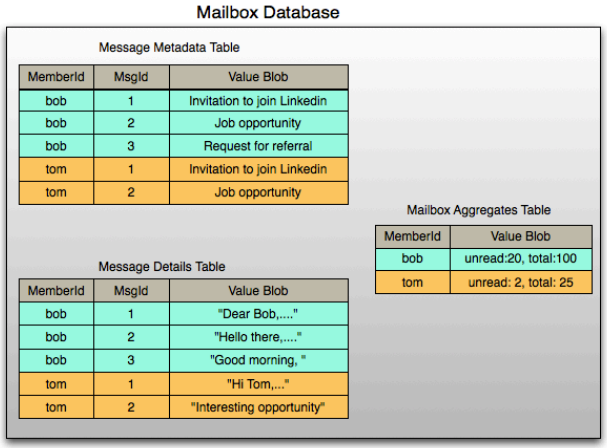
Rich functionality on resources

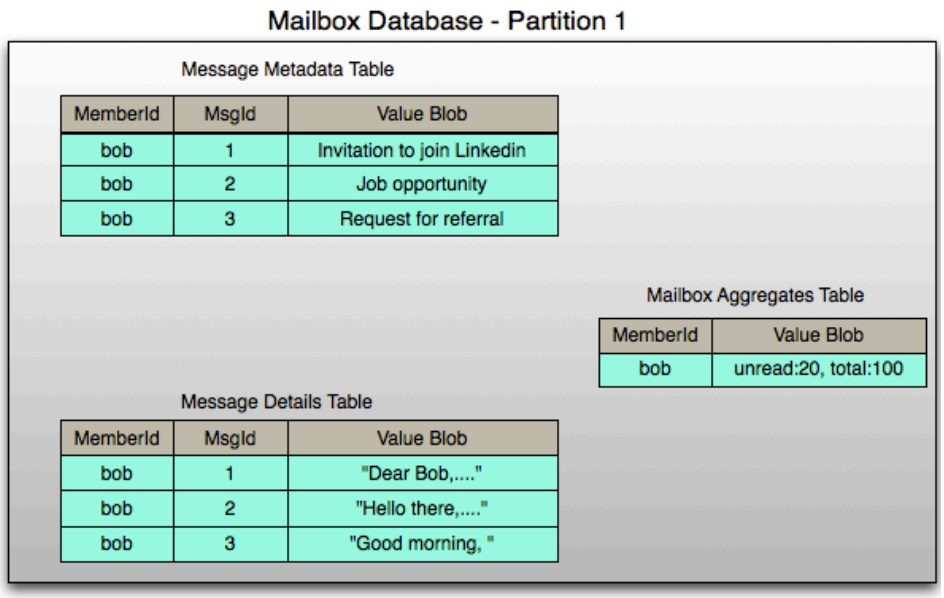
- ✓ Conditional updates
- ✓ Partial updates
- ✓ Atomic counters

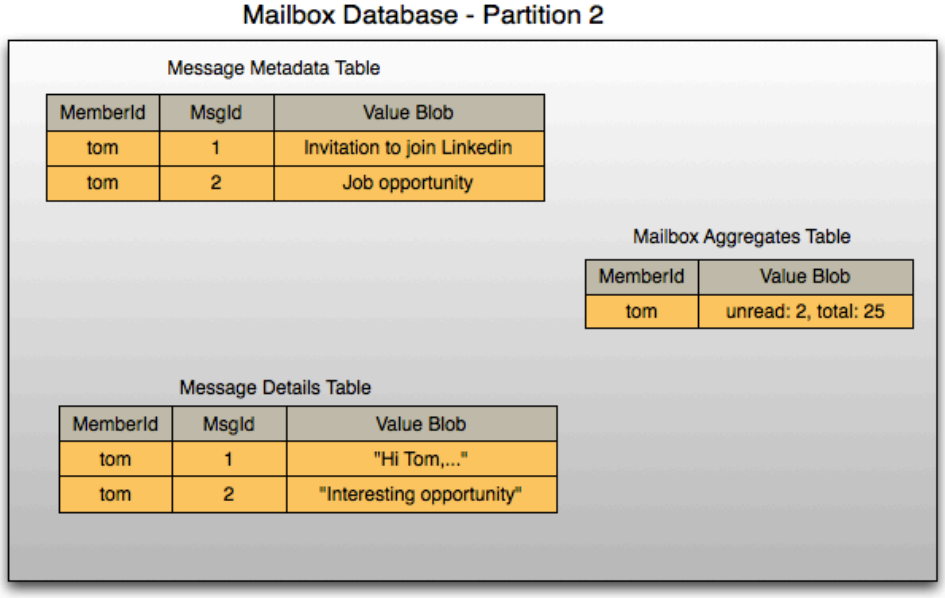
Rich functionality within resource groups

- ✓ Transactions
- ✓ Secondary index
- ✓ Text search

Partitioning

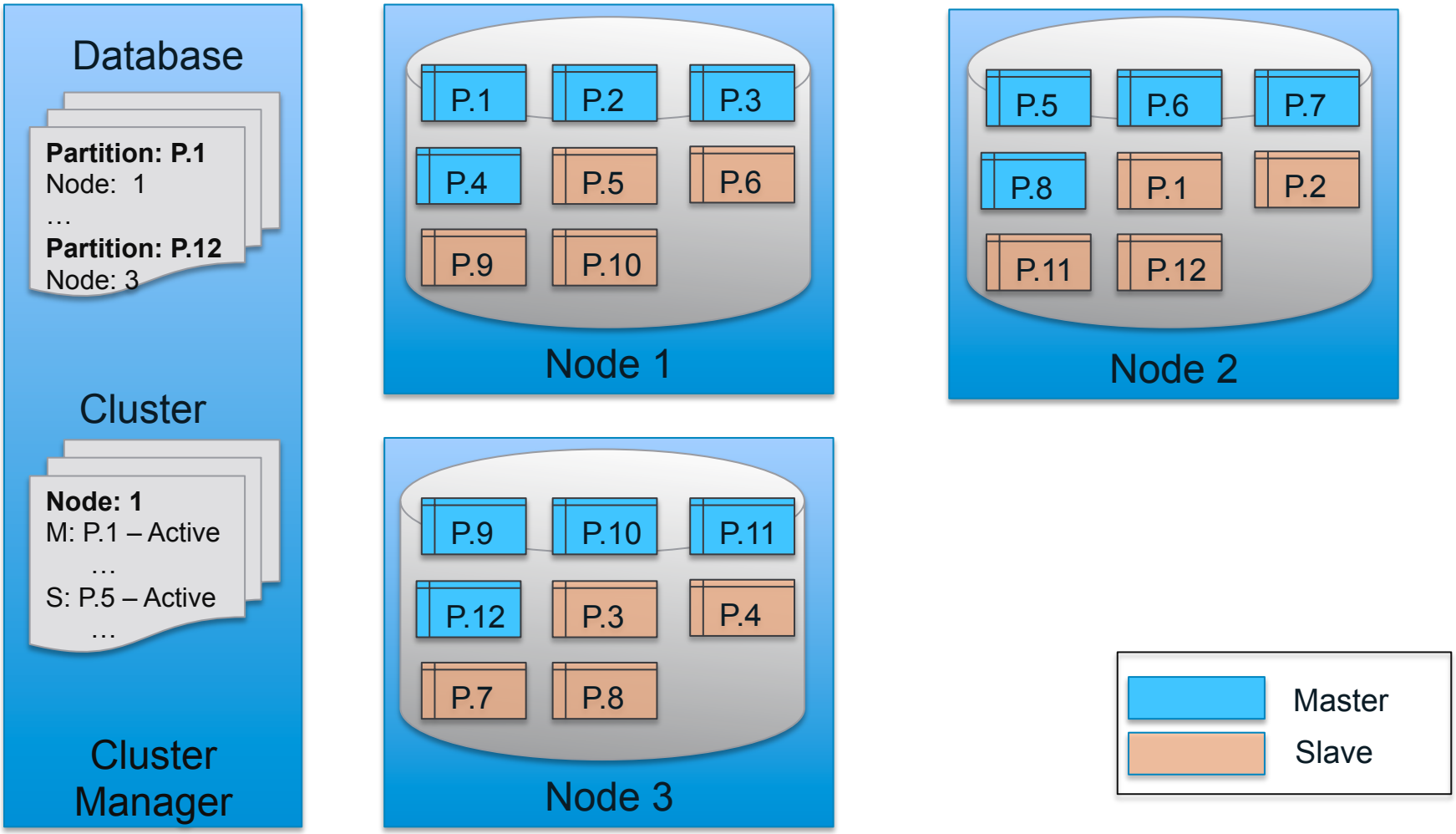




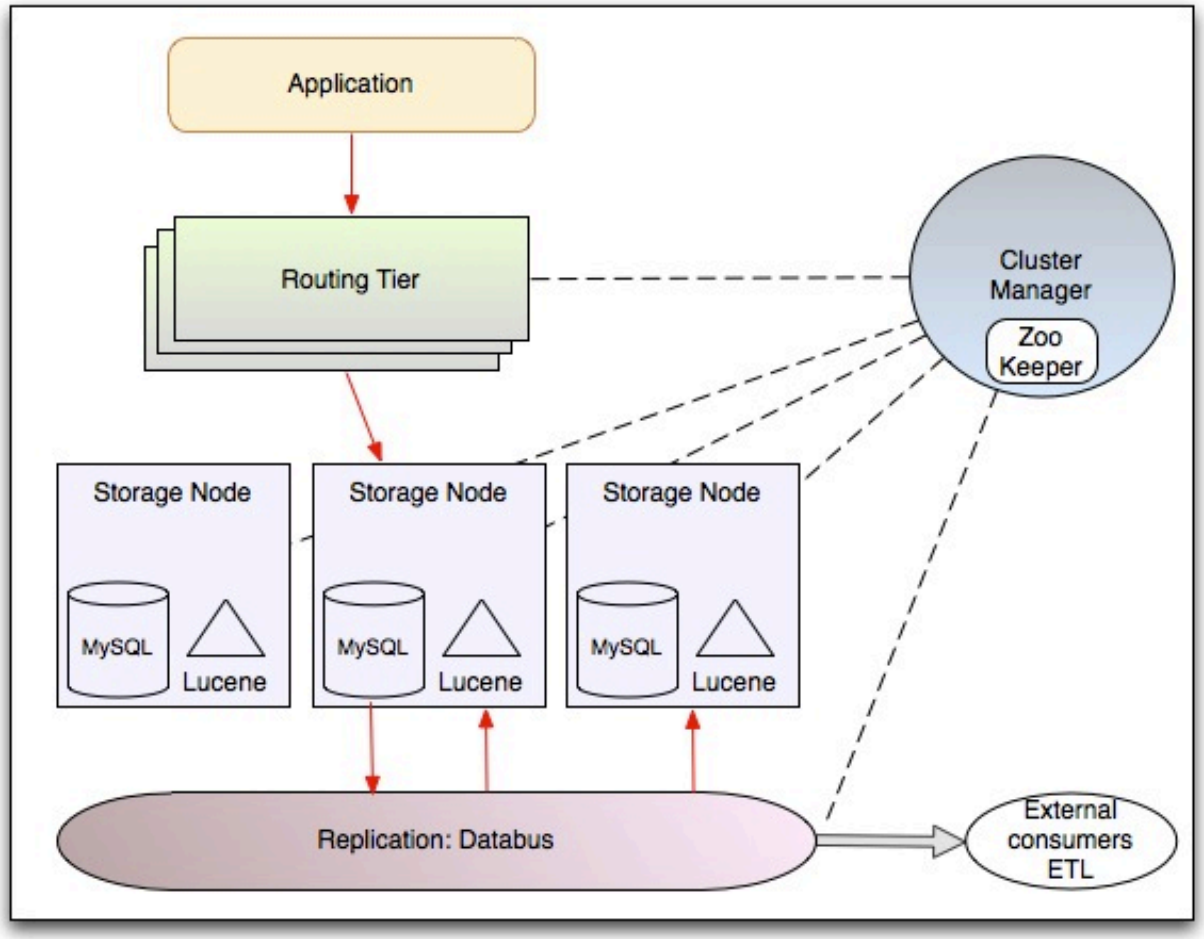


Espresso Partition Layout: Master, Slave

3 Storage Engine nodes, 2 way replication

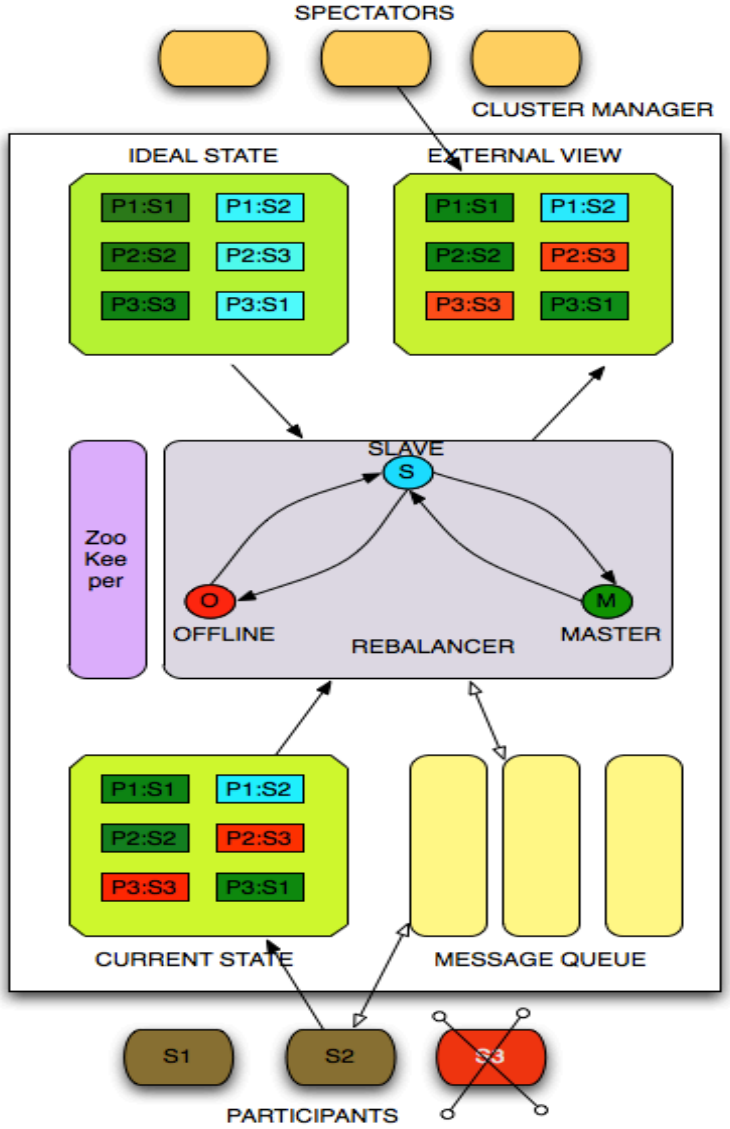


Espresso: System Components



Generic Cluster Manager: Helix

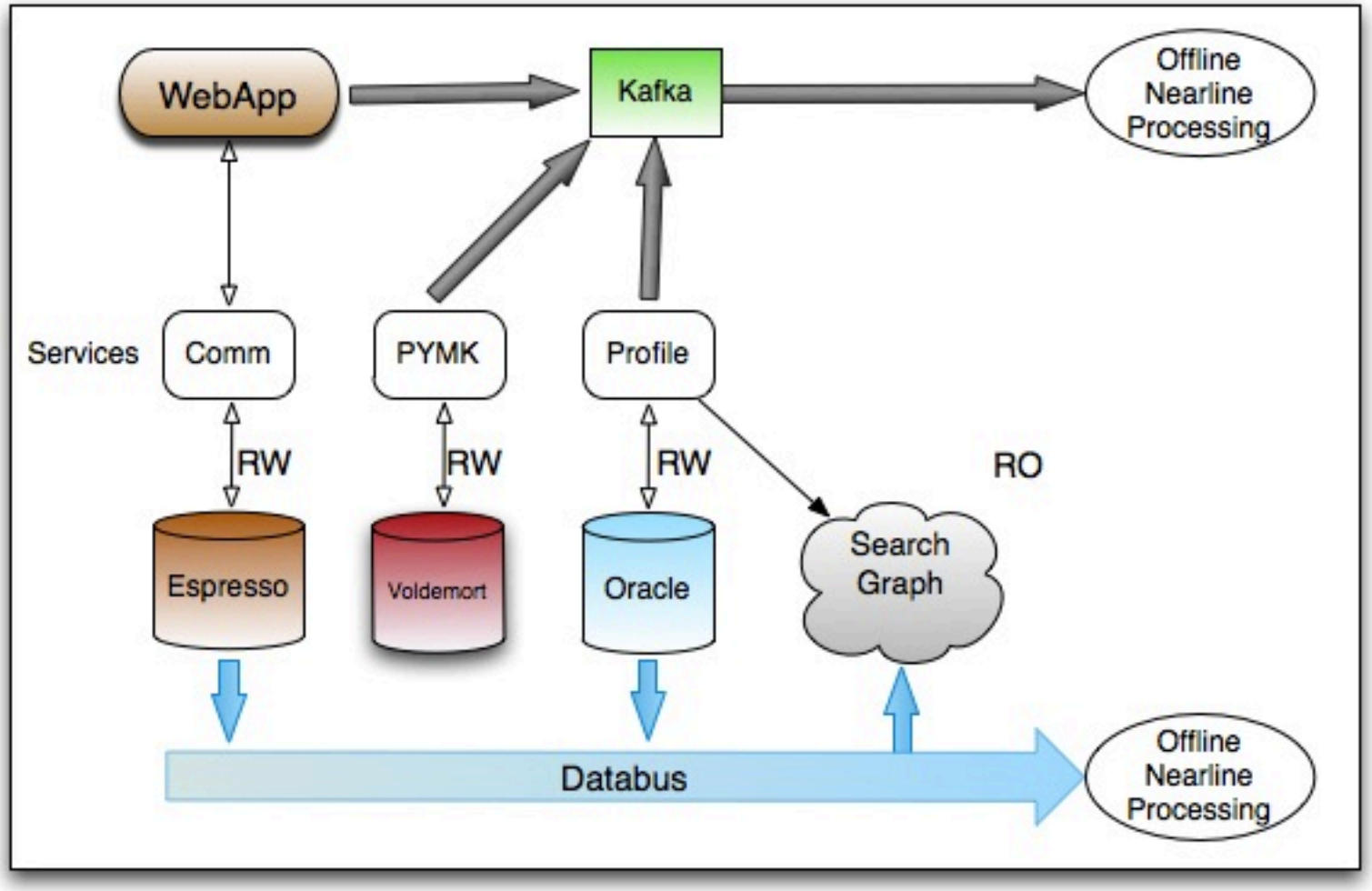
- Generic Distributed State Model
- Centralized Config Management
- Automatic Load Balancing
- Fault tolerance
- Health monitoring
- Cluster expansion and rebalancing
- Espresso, Databus and Search
- Open Source Apr 2012
- https://github.com/linkedin/helix



Espresso@Linkedin

- Launched first application Oct 2011
- Open source 2012
- Future
  - Multi-Datacenter support
  - Global secondary indexes
  - Time-partitioned data

LinkedIn Product Architecture



Acknowledgments

Siddharth Anand, Aditya Auradkar, Chavdar Botev, Vinoth Chandar, Shirshanka Das, Dave DeMaagd, Alex Feinberg, John Fung, Phanindra Ganti, Mihir Gandhi, Lei Gao, Bhaskar Ghosh, Kishore Gopalakrishna, Brendan Harris, Rajappa Iyer, Swaroop Jagadish, Joel Koshy, Kevin Krawez, Jay Kreps, Shi Lu, Sunil Nagaraj, Neha Narkhede, Sasha Pachev, Igor Perisic, Lin Qiao, Tom Quiggle, Jun Rao, Bob Schulman, Abraham Sebastian, Oliver Seeliger, Adam Silberstein, Boris Shkolnik, Chinmay Soman, Subbu Subramaniam, Roshan Sumbaly, Kapil Surlaker, Sajid Topiwala, Cuong Tran, Balaji Varadarajan, Jemiah Westerman, Zach White, Victor Ye, David Zhang, and Jason Zhang

Questions?