

LinkedIn Search & Lucene



John Wang

<http://www.linkedin.com/in/javasoze>

Lucene Revolution
10/8/2010

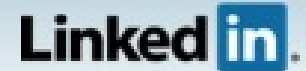
Agenda

- About LinkedIn
- Search @ LinkedIn
- People Search @ LinkedIn
- Lucene @ LinkedIn
- Challenges
- Open Source Projects @ LinkedIn
- Demo?!

About LinkedIn

- "Startup"
 - HQ in Mountain View
 - Founded in 2003
- As of Oct. 2010
 - Largest Professional Social-Network
 - 80+ million members
 - 170 countries
 - US ~ 50%
 - Fastest Growing Country is India
 - 26th Site on the web by traffic (Alexa)
 - 16th Site in the U.S. by traffic (Alexa)
 - #1 in NL by traffic (Quantcast)
 - Growth
 - ~ 1M new members every 9 days (>1 sec)

Search Properties @ LinkedIn



- **People Search**
- Jobs Search
- News Search
- Forum Search
- Group Search
- Company Search
- Reference Search
- Addressbook Search
- Answers Search
- **LIAR - (LinkedIn Advanced Recommendation)**
 - Jobs You May Like
 - Talent Match
 - and more ...
- ... and more

People Search @ LinkedIn



- Largest search by request
- Richest functionality
 - realtime, live updates
 - faceted navigation integration
 - personalized, networking integration
 - multi-lingual
 - structured and unstructured information
 - contextual query segmentation & classification
 - name spellcheck support
- Fully distributed, horizontally & hierarchically scalable
- Elastic, online expansion
- no index/cache warming
- no downtime for index optimization

People Search @ LinkedIn

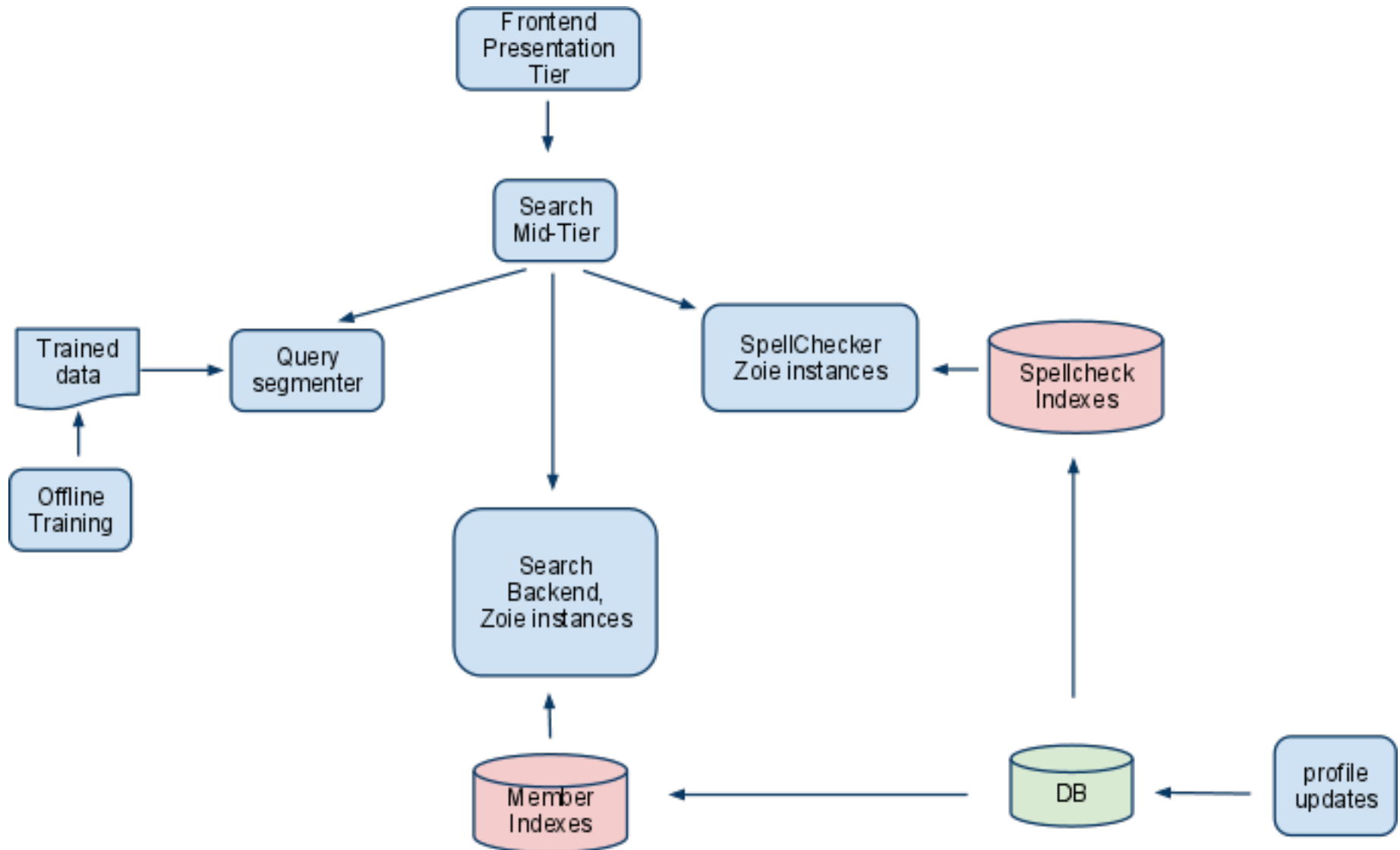
- details

- distribution
 - 5 million members/docs per partition
 - 2 partitions per node, partially balanced
 - 17+ partitions
 - 11 replications
 - 11 brokers
 - new partition/node every N days, no need to restart cluster
- load
 - Peak to Trough ratio: 5
 - QPS can peak at around 200 on a node
 - each node 99% latency < 100ms
 - max query length > 30 with > 10 boolean clauses

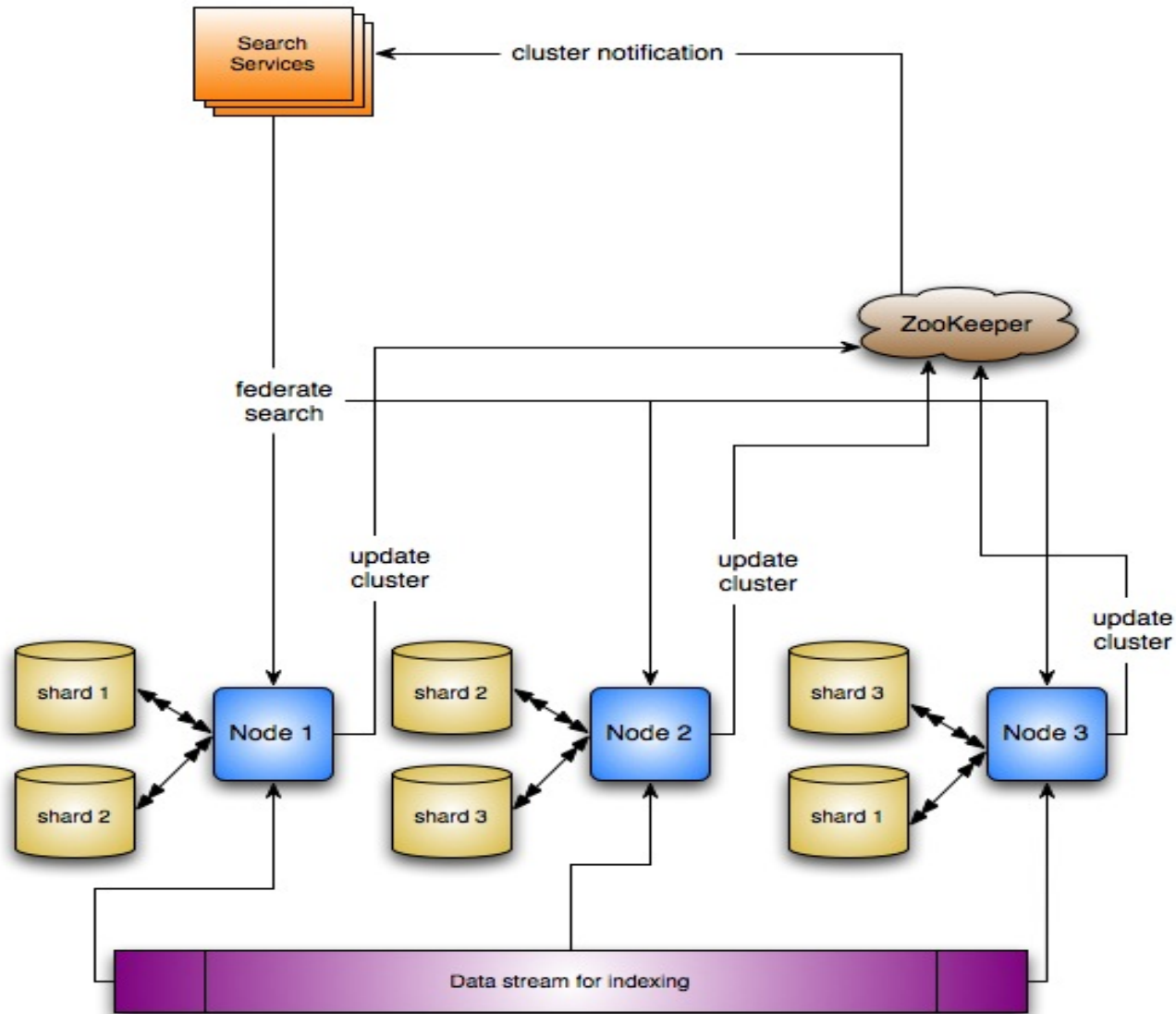
People Search @ LinkedIn - Requirements

- Real-time indexing/searching
 - Time between when user updates a profile and being able to find him/herself by that update need to be near-instantaneous
- Faceted Search
 - Usability-lab demonstrated the need for a paradigm shift
- Distributed/Partitioned Index
 - Horizontally scalable architecture to accommodate a fast growing member base
- Performance, Performance, Performance

Architecture - People Search

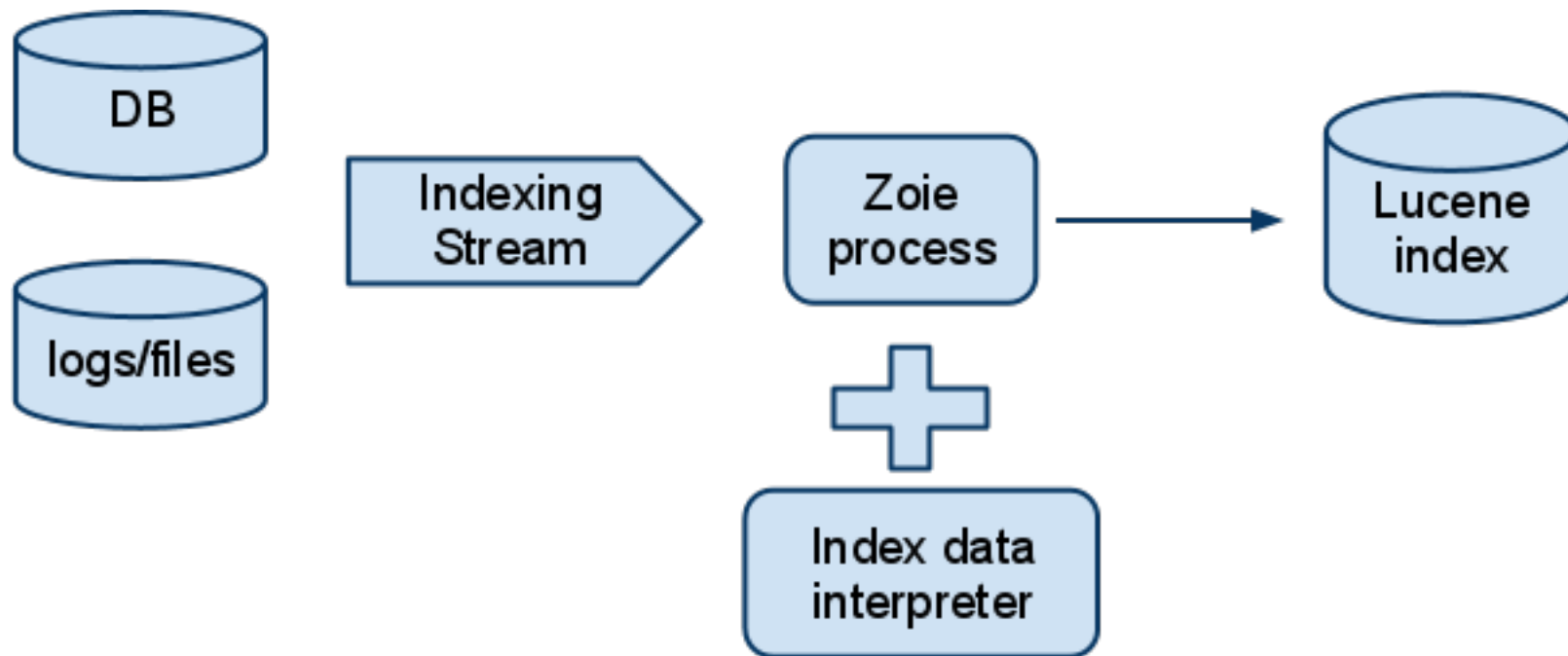


Architecture - search cluster



Stream indexing model

- Eventual consistency (important!)
- Low indexing latency (what does that mean?)



Why Lucene?

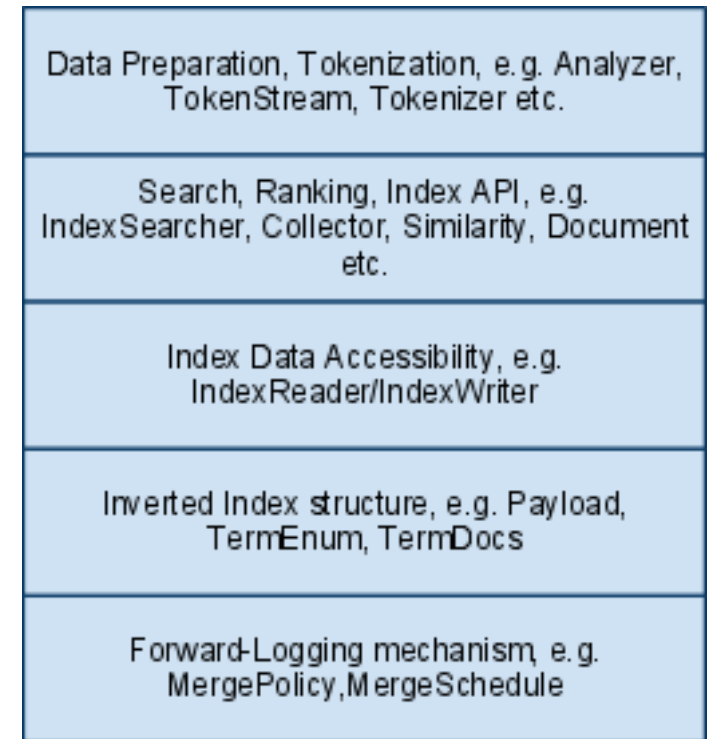
- Open-Source, Free (APL 2.0)
- Apache project
- Java-based
- Active community (support, QA etc.)
- Fast
- Incremental indexing support - important!
- Customizable - access to term postings etc.
- Strong committer team - very performance focused

Lucene @ LinkedIn

- Highly leveraged
- 2.9.1 - Huge migration from 2.4.2
- Pure Lucene
 - Currently, no need for a LI-Lucene variant
- Open source projects built on top of it:
 - Zoie/Hourglass
 - Bobo
 - Kamikaze
 - ...

Layers of Lucene

- Many ways of customizing Lucene
- Modularized/Isolated ways of improvements
 - Realtime branch
 - Flexible Indexing branch
- Tons of Analyzer contributions
- Can be used at any level
- Our customizations:
 - Sorting framework
 - FieldCache framework
 - Segment merge framework



Challenges

- Distributed
 - partitioning
 - dynamic cluster management
- Realtime
 - stream updates
 - low indexing latency
 - low maintenance
 - segment merges (GC-like effect)
- Faceted Search
 - different types of facets
 - join with external data
 - faceting on runtime information

Next Steps

- Facet value scoring
- Push model for indexing
 - preserving eventual consistency
- Enhanced indexing api
- Leverage new Lucene features, esp:
 - flex indexing
 - realtime support
- More plugins:
 - IndexInterpreters
 - FacetHandlers
 - etc.
- performance,performance,performance

Take-aways

- Latency - important, even when scaling horizontally
- Search application, not just search backend
- Understanding your ACID-needs
- Assumptions can bite you
- Social search:
 - speed of information flow
 - security
- Leverage:
 - open source
 - community
 - contributions

Open source projects @ LinkedIn

- <http://sna-projects.com>
- To name a few:
 - Voldemort (distributed K-V Store)
 - Bobo, Zoie, Sensei (Distributed realtime facet search stack)
 - Kamikaze (sorted int-set compression library)
 - Krati (super-fast K-V store)
 - Norbert (cluster management and RPC system based on Zookeeper)
 - ...
- Find them interesting?

We are hiring!

and we want you!

Q/A

???