



**jazzed** ABOUT SOLR

PEOPLE AS A SEARCH PROBLEM

# ABOUT ME

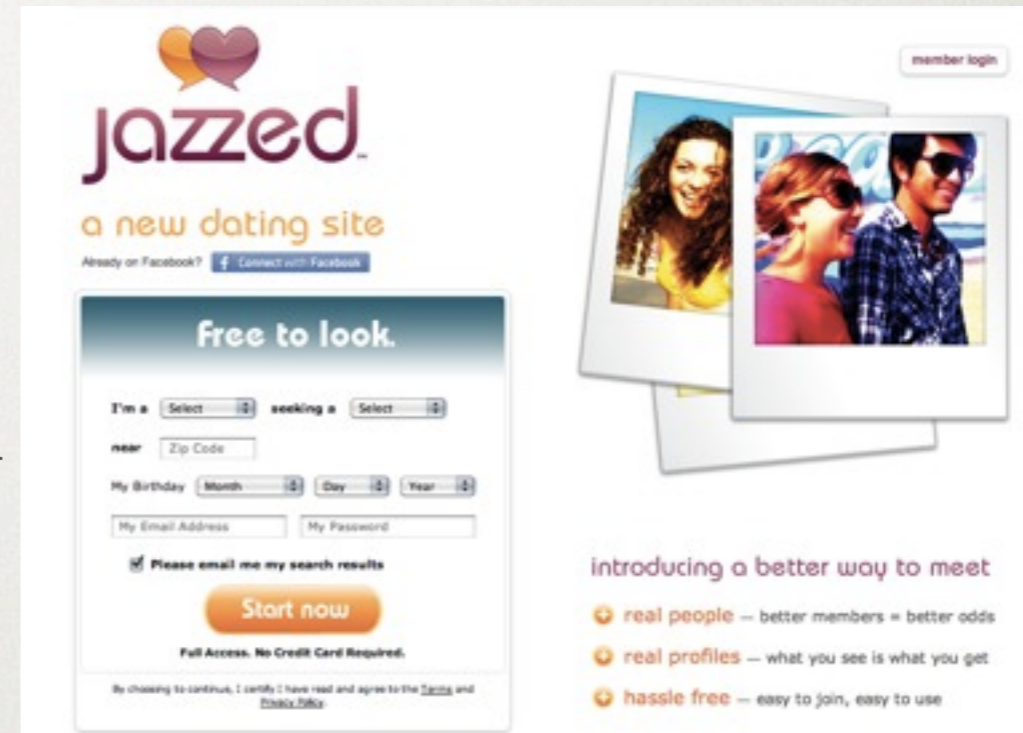
---

- Building websites since 1996, Java since 1997
- Prior web search experience
- Building and scaling eHarmony products since 2002

# WHAT IS JAZZED

---

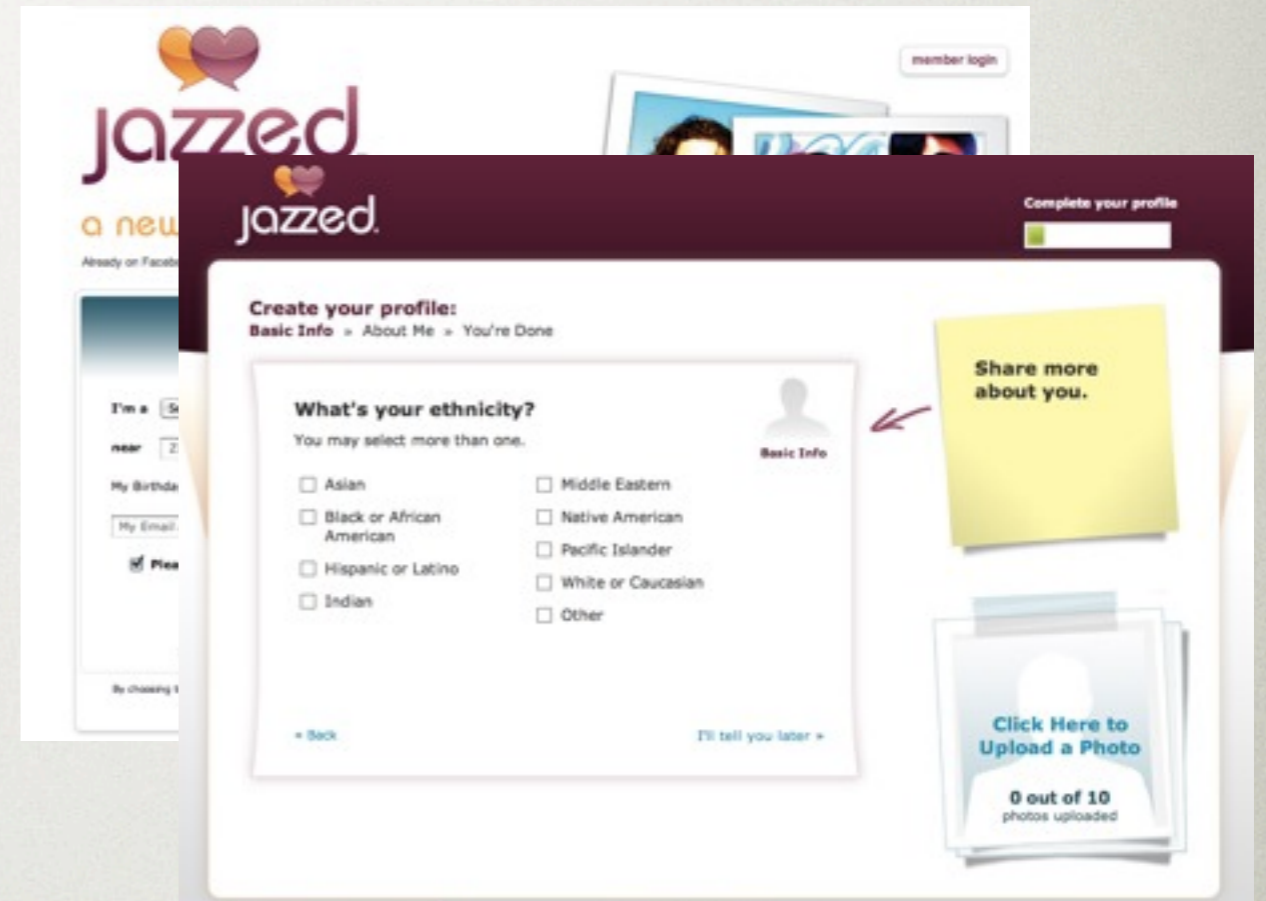
- Subscription Based Dating Site
- Incubated by eHarmony



# WHAT IS JAZZED

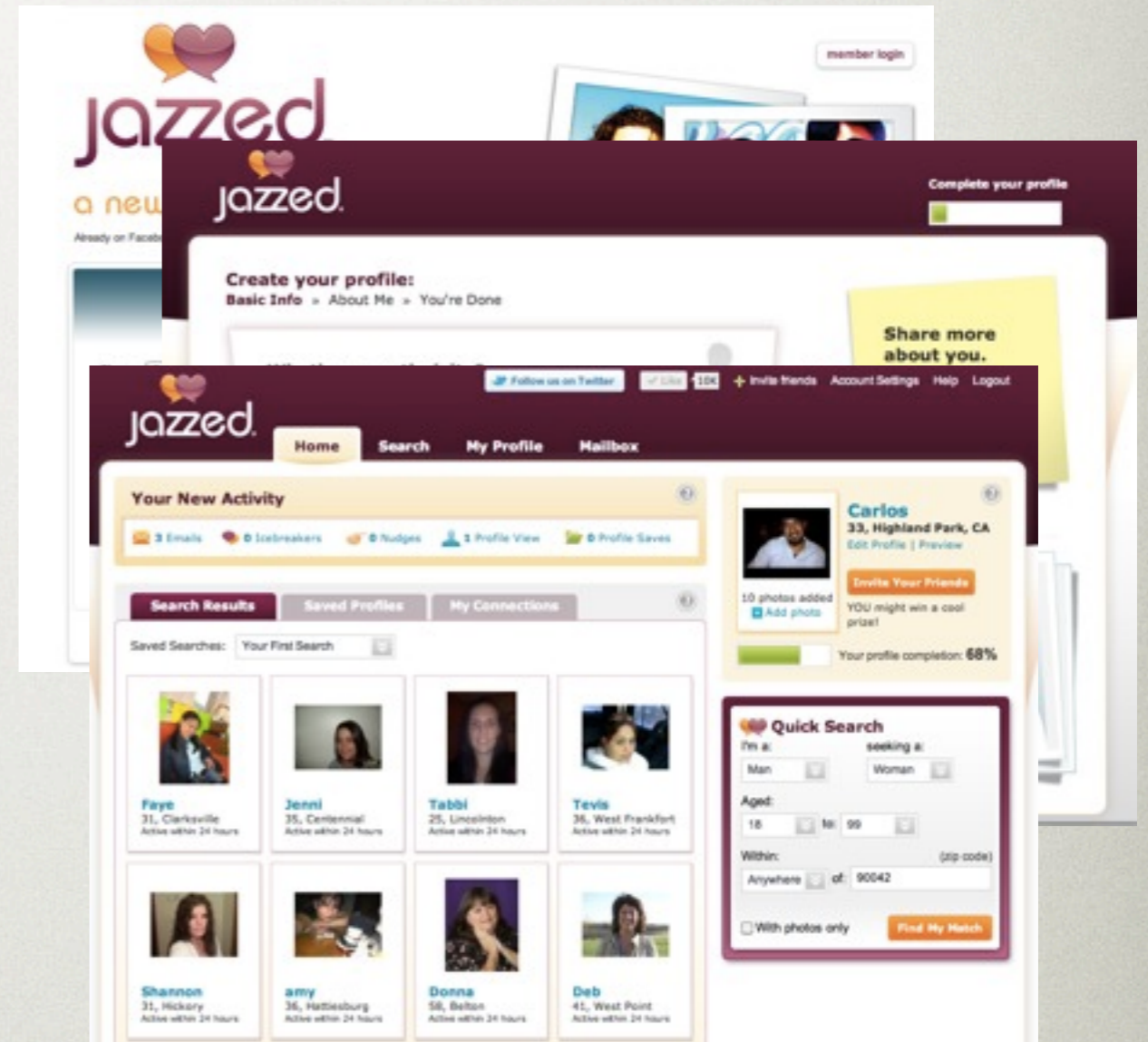
---

- Create a profile
- Search for others
- View their photos
- Privately Communicate



# WHAT IS JAZZED

- Create a profile
- Search for others
- View their photos
- Privately Communicate



# WHAT IS JAZZED

- Create a profile
- Search for others
- View their photos
- Privately Communicate



# WHAT IS JAZZED

- Create a profile
- Search for others
- View their photos
- Privately Communicate



# HOW IS IT DIFFERENT?

---

- Covers broader range of relationships
- Easy to get started
- Real profiles screened by machine and humans
- Fast, effective search oriented tools

# JAZZED STATS

---

- Started Fall 2009
- Beta Summer 2010
- Launched October 2010
- 100,000s of Profiles
- 1,000s of Searches Daily

# JAZZED ARCHITECTURE

---

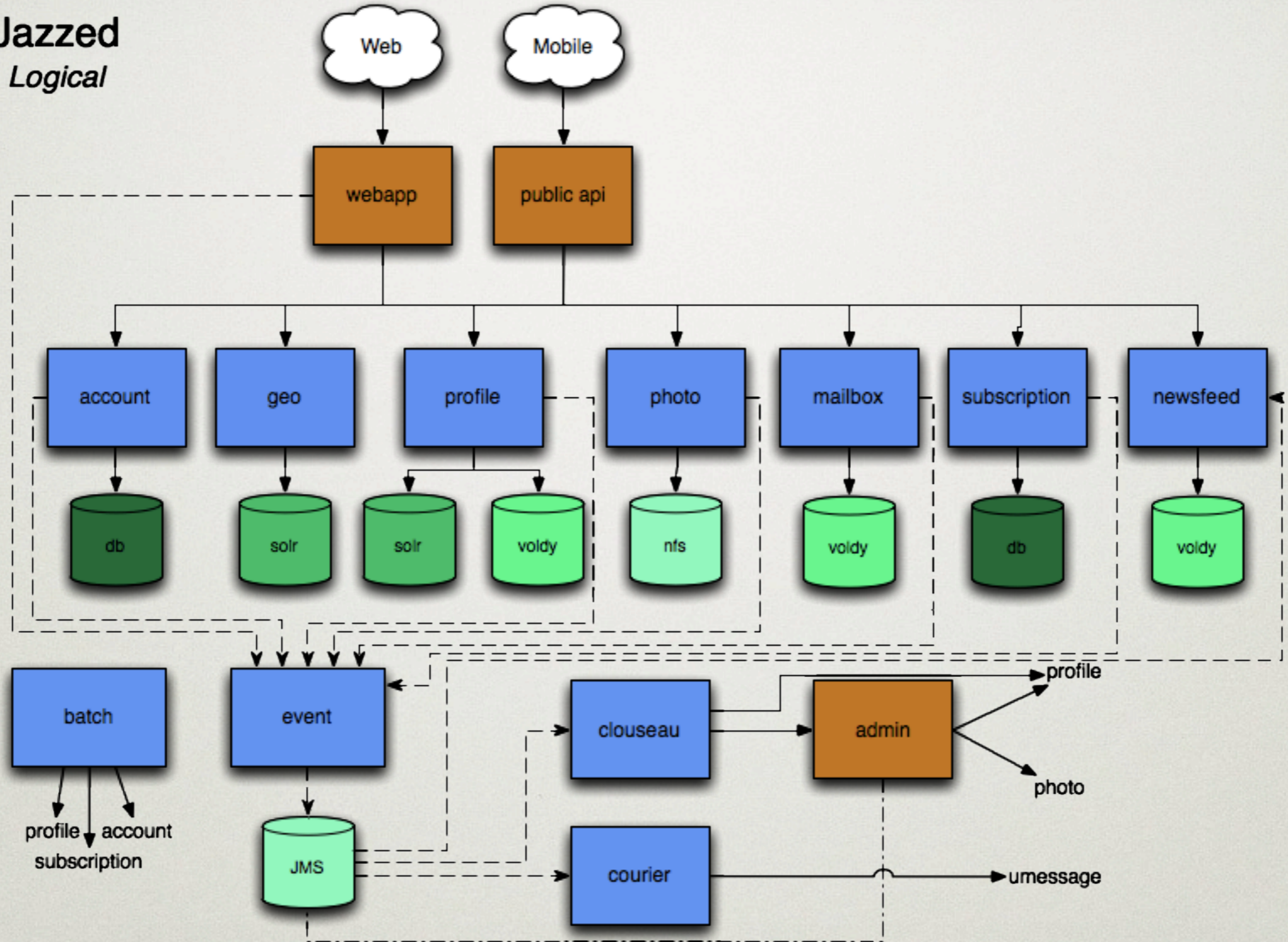
- Event-driven SOA
- REST, JSON, EIP, Not-only-SQL
- Technology incubation

# TECH STACK

---

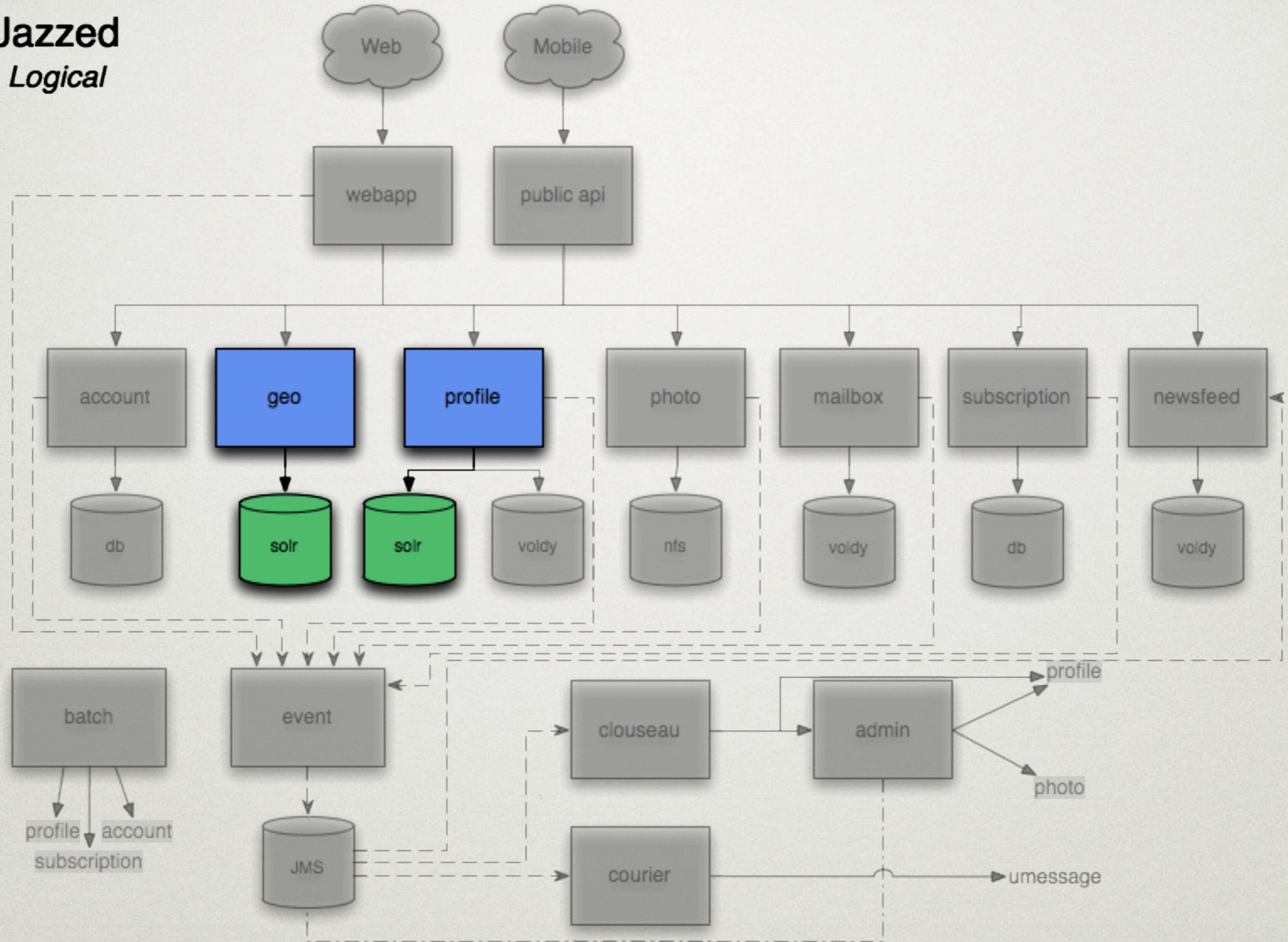
- Java 6, Spring 3, Jersey 1.1, JMS (AQMP)
- RHEL 4, Oracle 11g, Voldemort 0.81, Solr 1.4.1, NFS

# Jazzed Logical



# Jazzed

Logical



# NOT COVERED

---

- Distributed Search
- Caching Strategies
- Data Import
- Analyzers / Tokenizers

# WHY LUCENE?

---

- Proven Solid IR library
- Prefer Open Source Solutions
- Not Only SQL
- Flexible Ranking
- Pluggable

# WHY SOLR

---

- Performant, Extensible, RESTful Service
- Configuration, Schema, Multicores
- Admin Interface
- Replication, Backups, Monitoring

# OPEN SOURCE

---

- Strengthens Engineering Team
- Be apart of great community
- Not Brochure-ware

# NOT ONLY SQL

---

- One solution does not fit all
- Prefer availability over consistency
- Horizontal Scaling over Vertical

# FLEXIBLE RANKING

---

- Query Strategies
  - Boolean Algebra
  - Vector Space Analysis
  - Hybrids
- Extensive Function Support
- Index and Query Boosting

# ...OH MY!

---

- Standard Plugins - Geospatial\*,  
Faceting, Spelling, MoreLikeThis
- Full Text with Highlighted Results
- Client agnostic

# INEVITABLE QUESTION

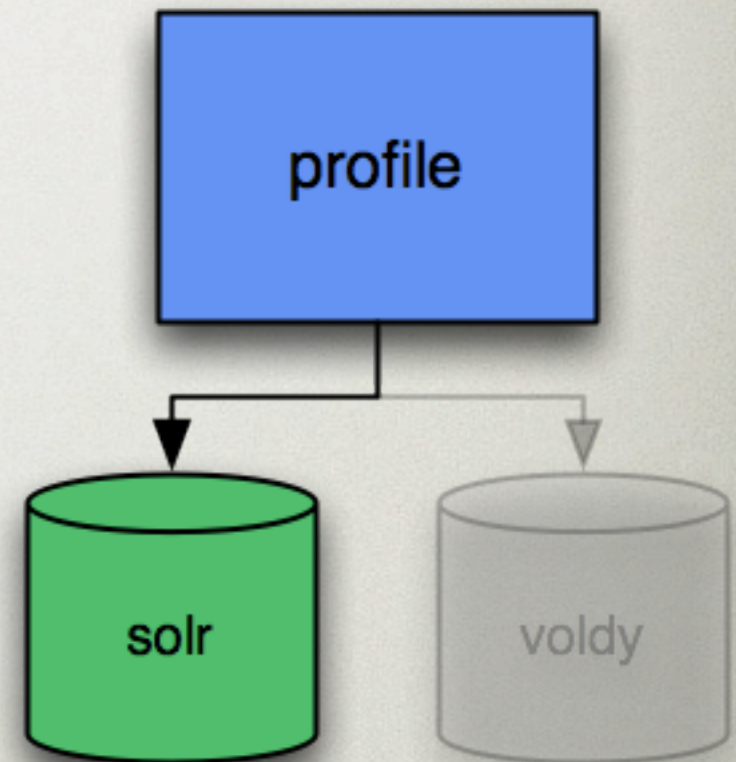
---

- “Does it scale?”
- Solr POC Benchmark
  - 10 Million profiles
  - >200 queries / sec under 100ms 90th
  - Default tuning until 5 million profiles

# PROFILE SERVICE

---

- RESTful Hybrid Data Service
- Public, Private, Attributes
- Event Producer



# PROFILES

---

- Mostly structured
- Categories - Eye Color, Desired Ethnicity
- Dates - Birthdate
- Numbers - Coordinates, Age Range
- Text -Name, Headline

# INVERTING PEOPLE

---

- Stored as an inverted index
- Index random accessed by term

Term	Document
MALE	1, 3, 5, 7, 9
FEMALE	2, 4, 6, 8, 10
HAIR_RED	8
HAIR_BLOND	1, 2, 5, 6
EYE_BLUE	1, 2, 3, 10
EYE_BROWN	4, 5, 6, 7, 8, 9
fun	1, 3, 7, 9
funny	2, 4, 6, 10
beach	1, 2, 3, 4, 5, 6, 7, 8

# SCHEMA DESIGN

---

- Single “Table”
- One-to-many = multi-value fields
- Individual vs Composite Fields
  - copyTo and have both!

# FIELD CONSIDERATIONS

---

- Stored or not
- Indexed or not
- Multivalued - desires fields
- Type

# SOLR TYPES USED

---

*The 't' is for Trie*

- tdate, tint, tfloat\* - birthdate, loginAt
- text - all text
- string - id, non indexed text
- random - good for random sorts
- *enum* - for all enumerations

# DATA DUPLICATION

---

- By function - numberPhotos & hasPhotos
- By relationship - hiddenBy & hidden
- By analysis - name & text

# SAVING PROFILES

---

- Updating is in memory operation
- No partial updates
- Commit means flush index changes
- Autocommit on maxDocs, maxTime or both

# WHY ALSO VOLDEMORT

---

- Private profiles can not be stale
- Many fields not searchable or viewable by others
- Isolate queries from fetch by id

# QUERYING

---

- Superset of Lucene
- Efficient Range Queries
- Multiple Query Handlers
  - Dismax, Boost, Geo

# RECALL VS PRECISION

---

- Focus on recall when corpus is small
- Precision once it is at critical mass

# BOOLEAN QUERIES

---

- Default operator set to AND
- +gender:FEMALE +seeking:MALE  
+eyeColor:EYE\_BLUE +hairColor:  
(HAIR\_RED, HAIR\_BLONDE)
- Sort order is important

# HYBRID QUERIES

---

- Default operator set to OR
- +gender:FEMALE +seeking:MALE  
eyeColor:EYE\_BLUE hairColor:  
(HAIR\_RED, HAIR\_BLONDE)

# WHY YOU'RE LUCKY IF YOU LIKE REDHEADS

---

- Inverse Document Frequency (IDF)
  1. Blue eyed, redheads
  2. Blue eyed, blonds
  3. Redheads
  4. Blonds
- Rarer is favored over more common
- More fields matched = higher ranking

# BOOSTING

---

- Query time by importance
- `eyeColor:EYE_BLUE^2`  
`hairColor:HAIR_BLOND`

# FILTER FIELDS

---

- Useful for roles and other lists
- `-hidden:(2 4 6)`

id	hidden
1	2, 4, 6
2	1

# FILTER FIELDS

---

- Useful for roles and other lists
- `-hidden:(2 4 6)`
- `-hiddenBy:1`

id	hidden
1	2, 4, 6
2	1

id	hiddenBy
1	2
2	1
4	1
6	1

# DATE MATH

---

- Simplifies query preprocessing
- +birthDate:[NOW / DAY+1DAY-36YEAR  
TO NOW / DAY-25YEAR]

# DATE MATH

---

- Simplifies query preprocessing
- +birthDate:[NOW / DAY+1DAY-36YEAR  
TO NOW / DAY-25YEAR]

Between 25 and 35 years old

## DISTANCE SEARCHING

---

- lat, lon, distance
- SolrLocal by Patrick O'Leary
- Additional overhead ~90ms per query
- Superseded in Solr 3.1

# TESTING QUERIES

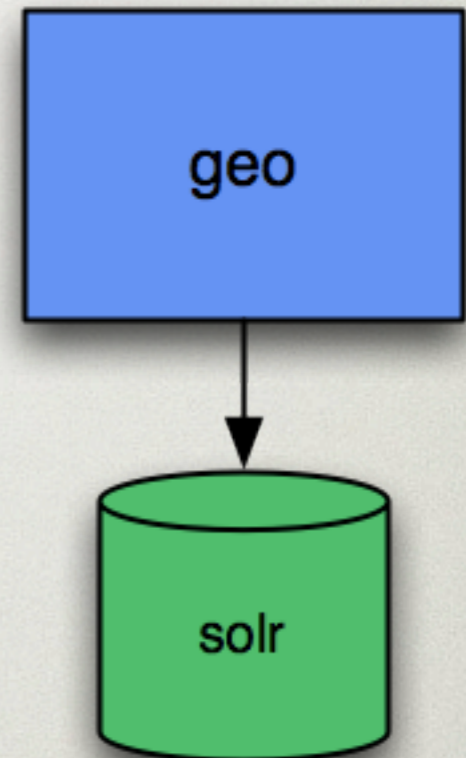
---

- Log queries and ids returned
- Version your search strategies
- Improve one thing at a time

# GEO SERVICE

---

- Read-mostly service
- Fields - Postal Code, Country, State, Cities, Lat, Lon
- Usage - Registration  
Validation, City Selection



# OPERATIONS

---

- Servlet container and filesystem
- Jetty 6, 64 Java 6 JVM
- 8G Heap `-XX:+UseCompressedOops`

# OPERATIONS

---

- Active / Passive
- Layer 7 Load balancing
- Nightly snapshots
- Eventually SolrCloud

# MULTICORE

---

- Run multiple schemas on the same
- Hot swappable for backwards compatible changes
- private / public profiles

# SECURITY

---

- No security provided
- At minimum secure your UpdateHandler
- Separate Cores

```
<delete>  
  <query>*:*</query>  
</delete>
```

# FUTURE

---

- Solr 3.1
- Mutual Matching
- Faceting / Guided Search
- Incorporating spelling
- Hierarchies, categories, better ranking models

# FACETING

- Returns counts with query results
- Efficient
- Guides the user toward precision

**SEARCH:** We found **4357** items!

NARROW YOUR CHOICES

YOUR SELECTIONS:

SHOES

STYLES

Bootie (486)  
Platform (438)  
Rain Boot (268)  
Slouch (175)  
Wedges (114)  
Peep Toe (57)  
Moccasin (56)  
Chukka (35)  
Athletic (20)  
High Tops (13)  
Pumps (2)

OCCASION

Casual (3098)  
Dress (1203)  
Outdoor (450)  
Office & Career (248)  
Work & Duty (35)  
Athletic (26)

NEWEST MOST POPULAR NAME L

NEW



**Harley-Davidson**  
Danica  
**\$130.00**

NEW

THANK YOU

[JTUBERVILLE@EHARMONY.COM](mailto:JTUBERVILLE@EHARMONY.COM)

TWITTER: @JTUBERVILLE