

# Transforming the house hunting experience

Alex Burmester, Alexander Kanarsky, Trulia



San Francisco, CA 94114



HOMES ▾

Buy

Rent

Advice

Mortgage

Local Info

Find a Pro

Trulia Mobile



Alex Burmester ▾

## 94114 Real Estate & Homes For Sale — 128 homes found

For Sale

Recently Sold

Sort by

Featured



### \$765,000 49 Seward St #1, San Francisco CA

FEATURED



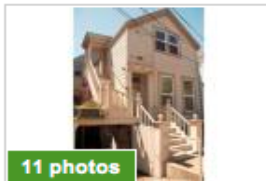
14 photos

2 br / 1 ba  
Condo  
Castro (94114)  
Paragon Real Estate Group



### \$899,000 408 Vicksburg St, San Francisco CA

FEATURED



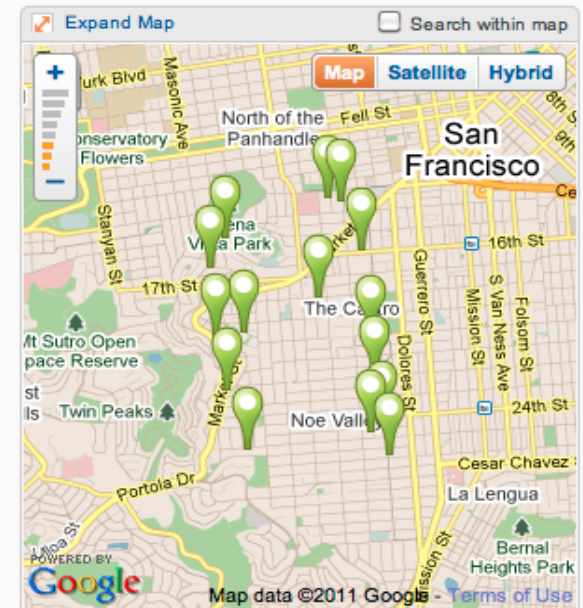
11 photos

1,643 sqft / \$547/sqft  
Multi-Family Home  
Noe Valley (94114)  
Coldwell Banker Residential...



List view

Large map



- New listings email alerts
- Save this search
- My searches
- Contact local agents

**LOCATION**  
City, neighborhood, ZIP

Add nearby locations

94114

Add city, zip or n'hood

**REFINE SEARCH**  
Set price, beds, baths, etc.

Reset and start over

Search Within

Location

Price

Min

Max

Custom price range

Presented by

**Lucid**  
IMAGINATION

**LUCENE**  
REVOLUTION

**trulia**  
real estate search

# About Us

- Trulia: Founded in 2005
- Technology company in downtown San Francisco transforming real estate search.
- 14M unique visitors per month
- Large and active user community
- Strong industry relationships: agents, brokers, over 500k registered professionals
- Helping consumers make informed decisions
- About the speakers

# Trulia's Marketplace

## Consumers

- \$1Tr Real Estate Economy
- ~5M Home sales/yr
- ~16M Leases/yr
- >70M Homeowners



### Most Compelling Product:

Comprehensive data, unique insights, great UX



## Professionals

- \$20Bn Real Estate Mktg
- \$7Bn Online Today
- >1M Realtors
- >1M Landlords/Managers
- >1M Local Contractors



### Essential Marketing Solution:

Huge exposure, qualified customers, powerful tools

We create value by connecting consumers and professionals

# Problem overview

- Real Estate search specifics
- Traditional Search: Agent based/MLS
- No single MLS country wide
- Largest financial transaction for most people with potential for huge life tradeoffs
- Lots of data but silo'd off with data quality issues and differing disclosure rules by area.

# The House Hunting Process

- Buying, investing, renting
- How online search changes this
- Trulia was first site to mash-up lots of data nationwide, maps, heatmaps, tax info, rentals, street-view
- Focus on speed
- Relevancy of data
- Amount of information
- Targeted *local* RE search experience

# Challenges

- Integrating large diverse datasets
- Data quality and freshness
- Delivering the right data to the right people
- Scaling for growth
- Meaning of Location, location, location

# Legacy Approach

- MLS-like database search
- MySQL 4.x
- Limitations on data processing
- Limitations on scalability
- Replication problems
- Update speed problems

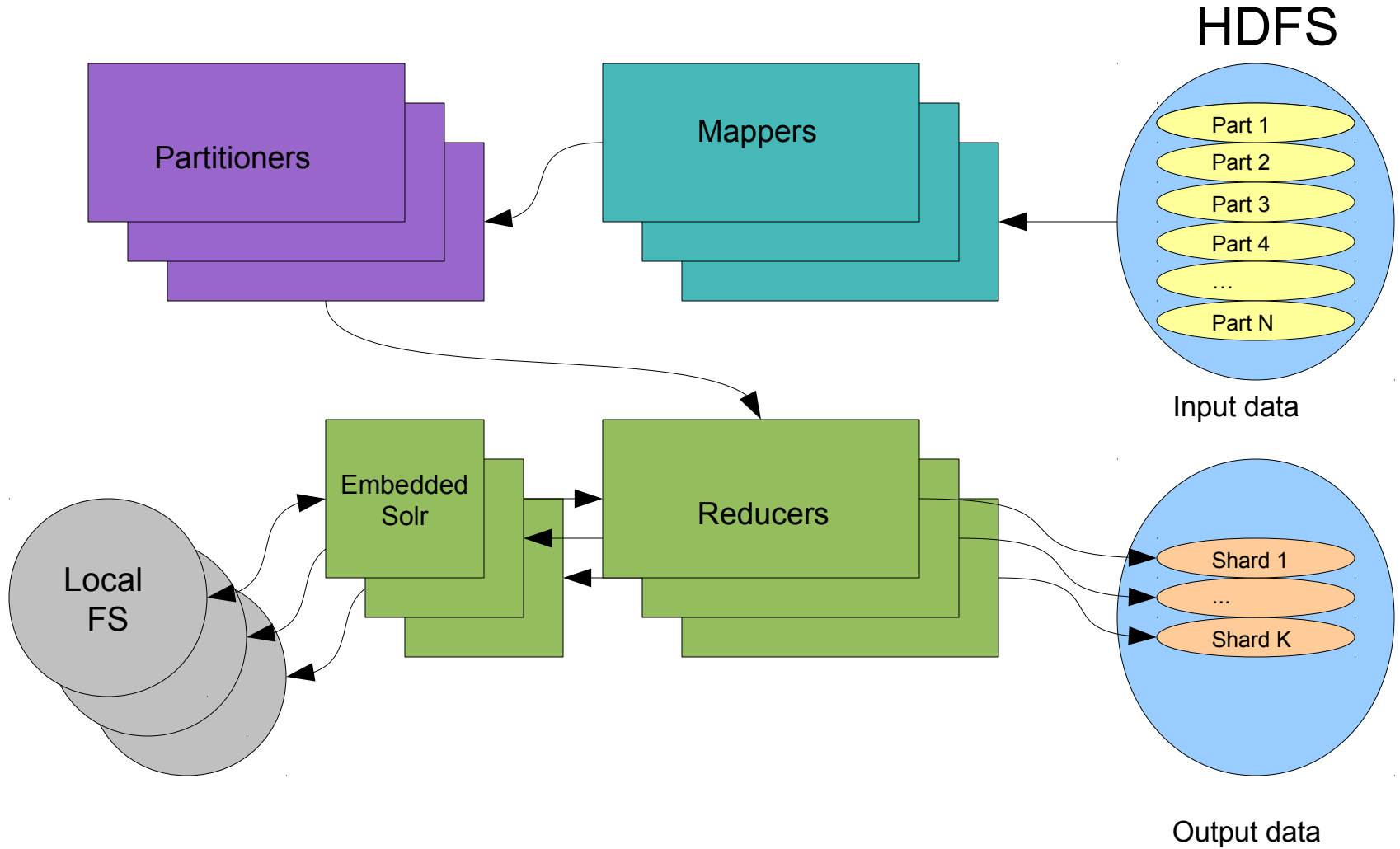
# Why Solr?

- Fast, flexible attribute search, faceting
- Non-uniform data handling, dynamic schema
- Easy, uniform API for indexing, query
- Excellent Full-Text Search
- Indexing, Search scale great (Hadoop indexing, distributed queries, replication)
- Stored content: can be used as a data store
- Add-Ons: Geospatial search, Field Collapsing, Automatic Data Import via DIH
- Strong Developers/Users Community

# Hadoop Indexing

- Based on SOLR-1301 patch
- Reduce-side indexing
- Map-side partitioning (sharding)
- Input Data in HDFS
- Output Indexes in HDFS
- Local FS for temporary data processing
- 120+ mil. documents index in less than 1 hour
- Paired with Index Manager

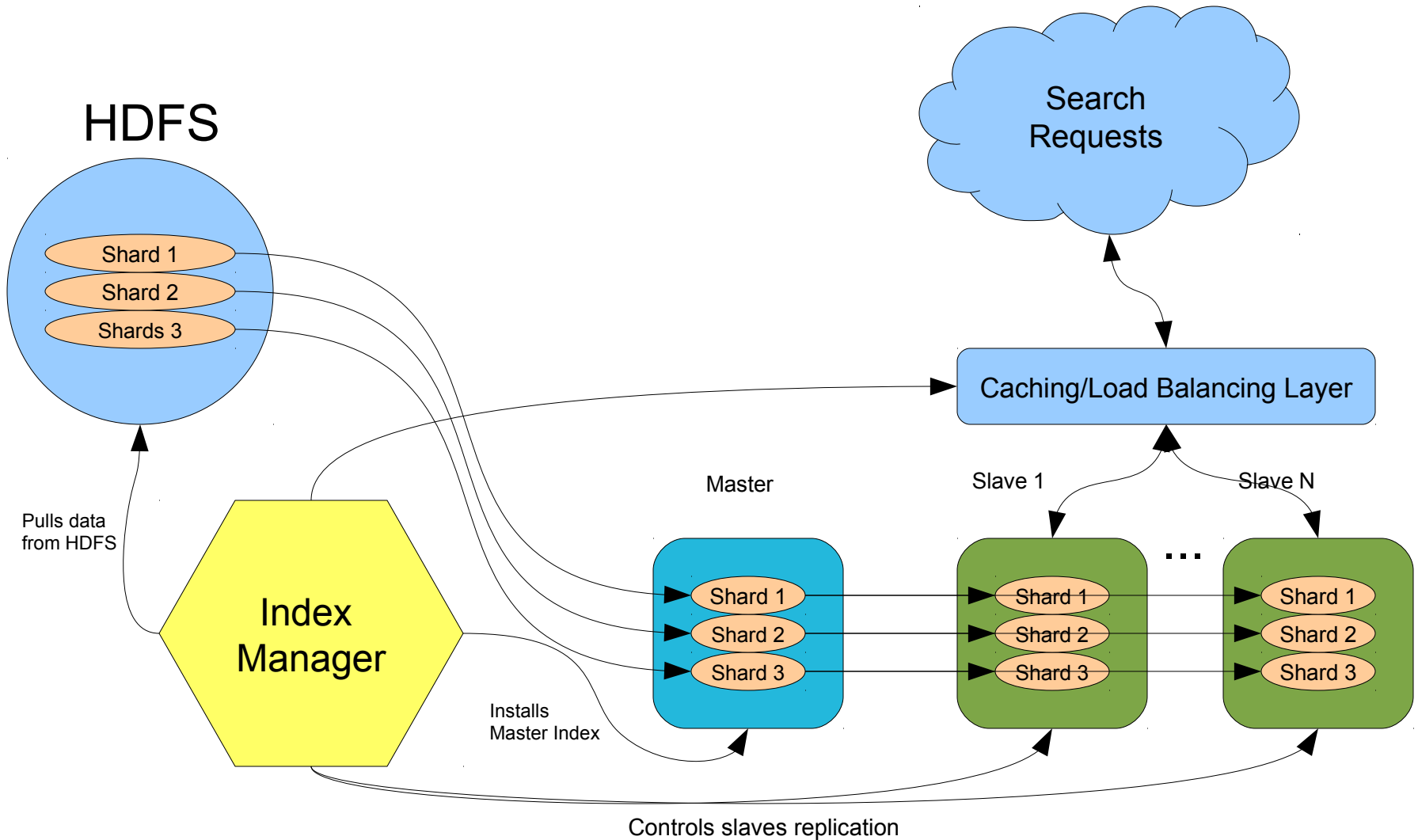
# Hadoop Indexing



# Index management

- Index Manager: custom controller
- Manages index deployment from HDFS
- Controls load balancing
- Manages forced replication on slaves
- Interacts with caching layer
- Handles Direct indexing
- Visualizes the state of the indexes
- Alternatives: Katta, Solr Cloud

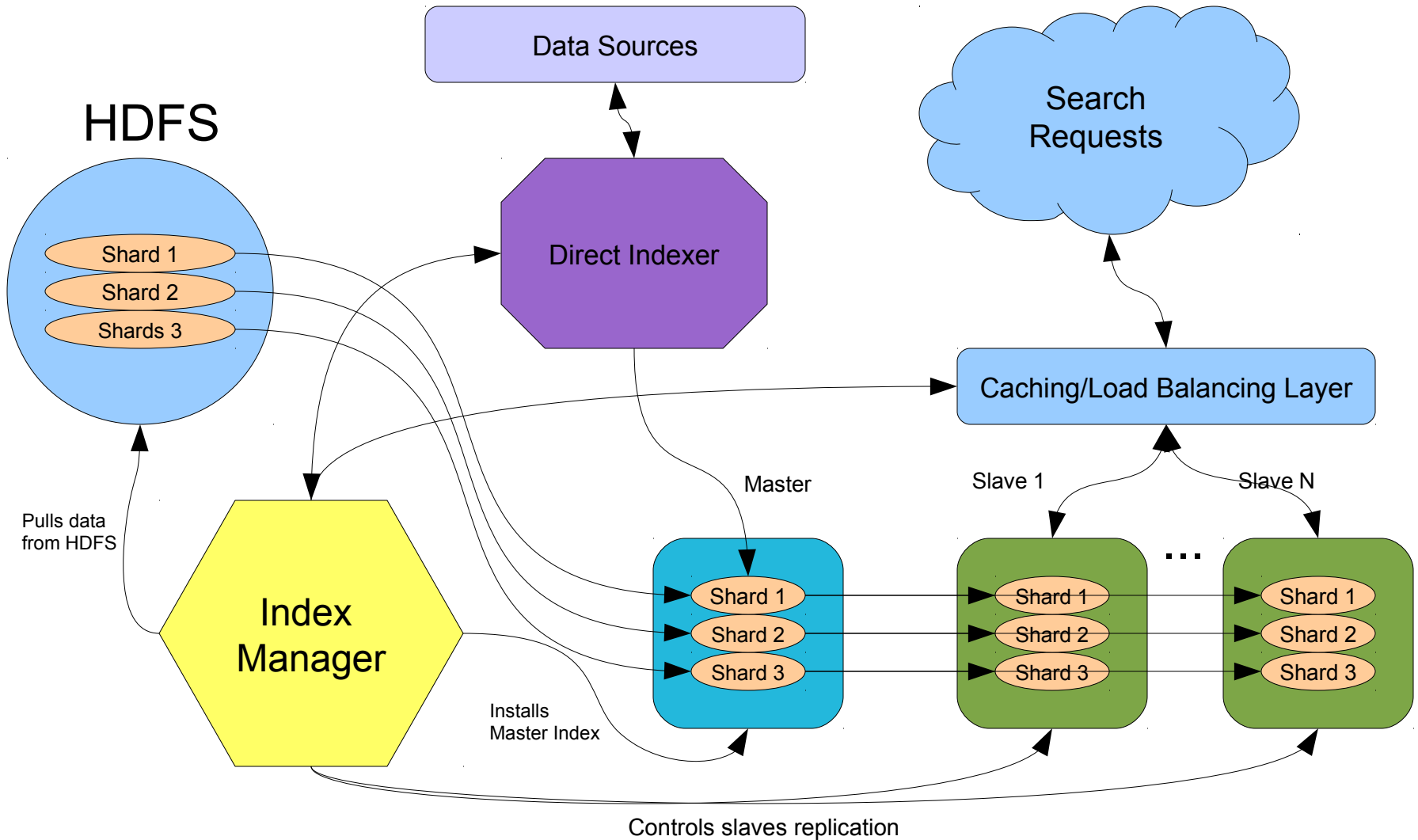
# Index Management



# Direct Indexer

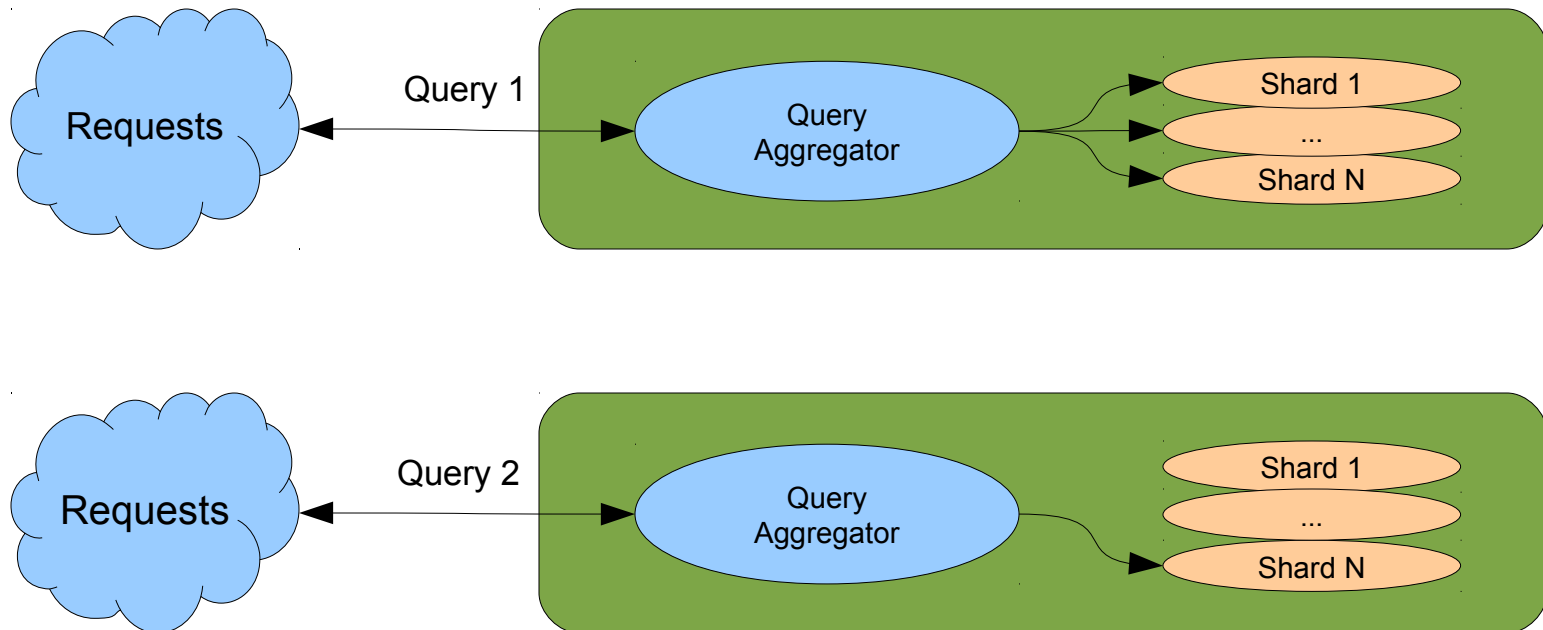
- Handles various updates between the batches
- Patches, pictures updates etc.
- Only does certain type of documents
- Directly updates master index shards
- Interacts with Index Manager (scope of changes)
- Runs on a regular basis

# Direct Indexer



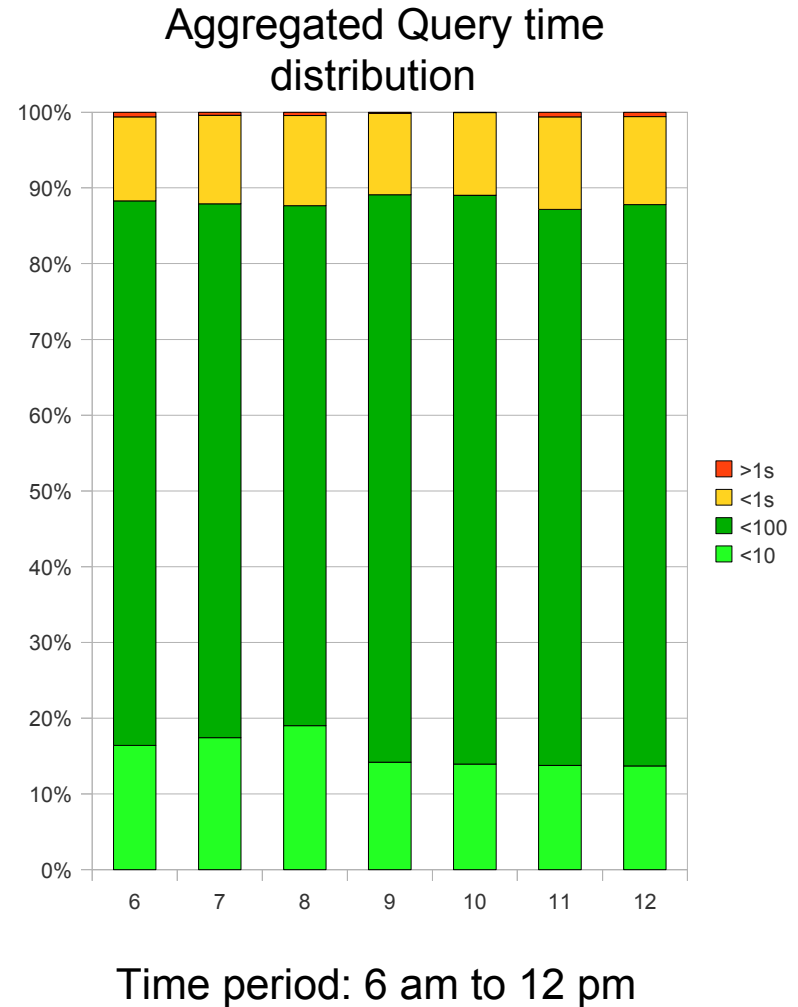
# Distributed Search

Query Aggregator uses the same partitioning schema as Hadoop Indexer, so only the relevant shards are queried. Query extra parameters define the scope.



# Distributed Search

- Many servers, many shards
- Request is sent only to relevant shards
- For search I/O wait matters
- So SSDs rule :-)
- 30-50 qps per shard
- 90% of queries < 100 ms,
- 99% within 1sec
- Separate indexes for some documents
- 50% of queries < 10 ms
- Caching/load balancing matters



# Geospatial search

- Local search is super important
- Mobile search is naturally spatial-oriented
- Communities, neighborhoods, enclaves
- POI search
- Comparable Properties
- Personalized Search: Factoring census data in
- How to enable it with Solr?
- No standard solution yet

# Geospatial Solr/Lucene

- SOLR-733 patch
- Local Lucene / Local Solr (Patrick O'Leary)
- Spatial Lucene in Lucene 3.1
- JTeam's SSP (Chris Male)
- SOLR-2155 patch (David Smiley)
- Lucene-Spatial-Playground (David Smiley, Chris Male, Ryan McKinley)
- Spatial functions in Solr 3.1

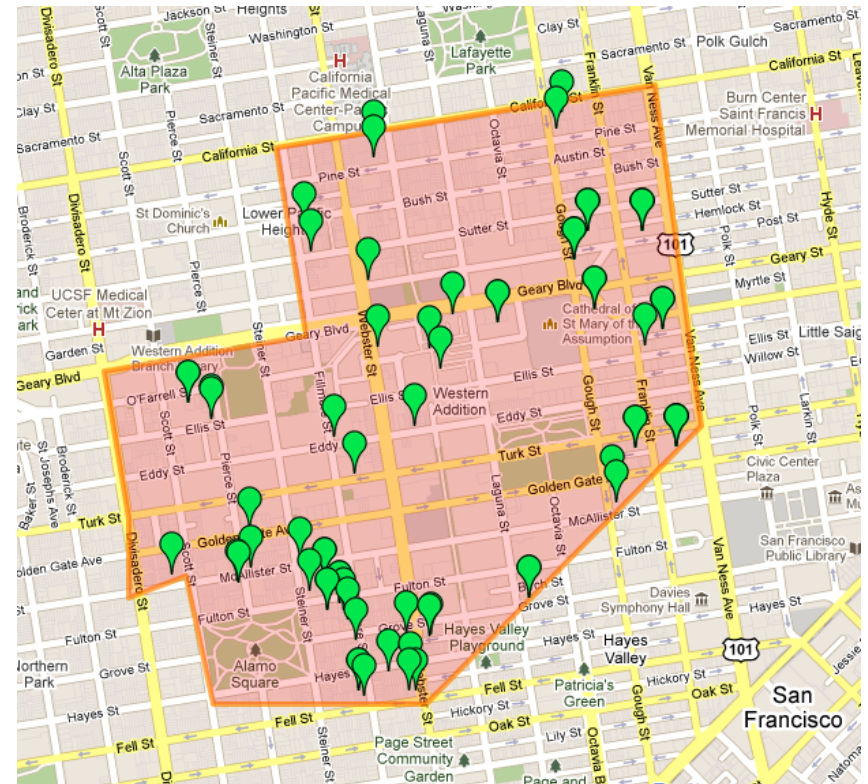
# Geospatial search

- What kind of spatial search do we need?
- Radius search
- Polygon search
- Sort by distance is needed for radius search
- Implementation chosen: SSP-based, fast geometry

- INFO: [active] webapp= path=/select/ params={fl=streetAddress\_s,lat,lng&wt=kml&q={!spatial+polygons%3D37.77960208641734,-122.43867874145508;37.783265262376574,-122.43945121765137;37.78414711095678,-122.4330997467041;37.788827515747435,-122.43404388427734;37.79038758480464,-122.4221134185791;37.781908551707666,-122.420654296875;37.774988939930005,-122.42932319641113;37.774988939930005,-122.4360179901123;37.77817746896081,-122.43687629699707;37.777634750327046,-122.43867874145508;37.77960208641734,-122.43867874145508;}\*:\*&rows=1000} hits=75 status=0 QTime=6

# Geospatial Search

- Circle search: given (lat, lng, radius) return list of docs with lat/lng indexed
- Sorting: distance from the center
- Polygon search: given a polygon (set of lat/lng pairs) return list of docs with lat/lng indexed
- Bounding boxes: lat,lng range queries
- Postponed distance calculation (query response write time)



# Geospatial Search

- Spatial query parser: `{!spatial <params>}`
- Can be wrapped as a filter query
- Facet by spatial queries
- Planar geometry: fast, but has limitations
- Good for our use cases, YMMV
- Polygon compression: Google's delta-encoding
- TBD: shapes indexing

# Wrap Up

- We are reshaping real estate search - with the right tools!
- Hadoop + Solr way is great for frequent batch indexing
- Solr Search *is* fast - and highly scalable through sharding and replication
- Geospatial Solr search is still not standardized, lots of progress in development
- Open Source: Trulia contributes

# Links

- SOLR-1301: Hadoop Indexing Contrib
  - [issues.apache.org/jira/browse/SOLR-1301](https://issues.apache.org/jira/browse/SOLR-1301)
- Spatial Solr Plugin (SSP)
  - [www.jteam.nl/products/spatialsolrplugin.html](http://www.jteam.nl/products/spatialsolrplugin.html)
- SSP Polygons/Polylines Extension
  - [sourceforge.net/projects/ssplex/files/](https://sourceforge.net/projects/ssplex/files/)

# Contact Information

- Trulia
  - [www.trulia.com](http://www.trulia.com)
  - [www.linkedin.com/company/trulia](http://www.linkedin.com/company/trulia)
- Alex Burmester
  - [aburmester@trulia.com](mailto:aburmester@trulia.com)
- Alexander Kanarsky
  - [akanarsky@trulia.com](mailto:akanarsky@trulia.com)

# The End

- Q & A
- Trulia is hiring!
  - [www.trulia.com/jobs](http://www.trulia.com/jobs)
  - Work in downtown San Francisco with great people
  - Looking for engineers specializing in search, distributed data processing, visualization/data science, mobile platforms, front end and more.
- Thank you for coming!