Search Discover Analyze

Large Scale Search, Discovery and Analytics with Solr, Mahout and Hadoop

Grant Ingersoll
Chief Scientist
Lucid Imagination

Search is Dead, Long Live Search

Good keyword search is a commodity and easy to get up and running

The Bar is Raised

Relevance is (always will be?) hard

Holistic view of the data AND the users is critical

User Interaction

User Interaction

Access

Access

Content Relationships

Content Relationships

Topics

Quick Background and needs

Architecture

Abstract

Practical

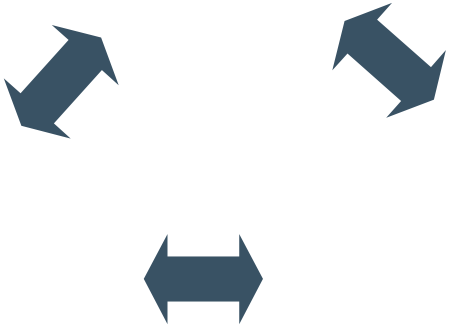
SDA In Practice

Components

Challenges and Lessons Learned

Wrap Up

Why Search, Discovery and Analytics (SDA)?



User Needs

Real-time, ad hoc access to content

Aggressive Prioritization based on Importance

Serendipity

Feedback/Learning from past

Business Needs

Deeper insight into users

Leverage existing internal knowledge

Cost effective

What Do Developers Need for SDA?

Fast, efficient, scalable search

Bulk and Near Real Time Indexing

Handle billions of records w/ sub-second search and faceting

Large scale, cost effective storage and processing capabilities

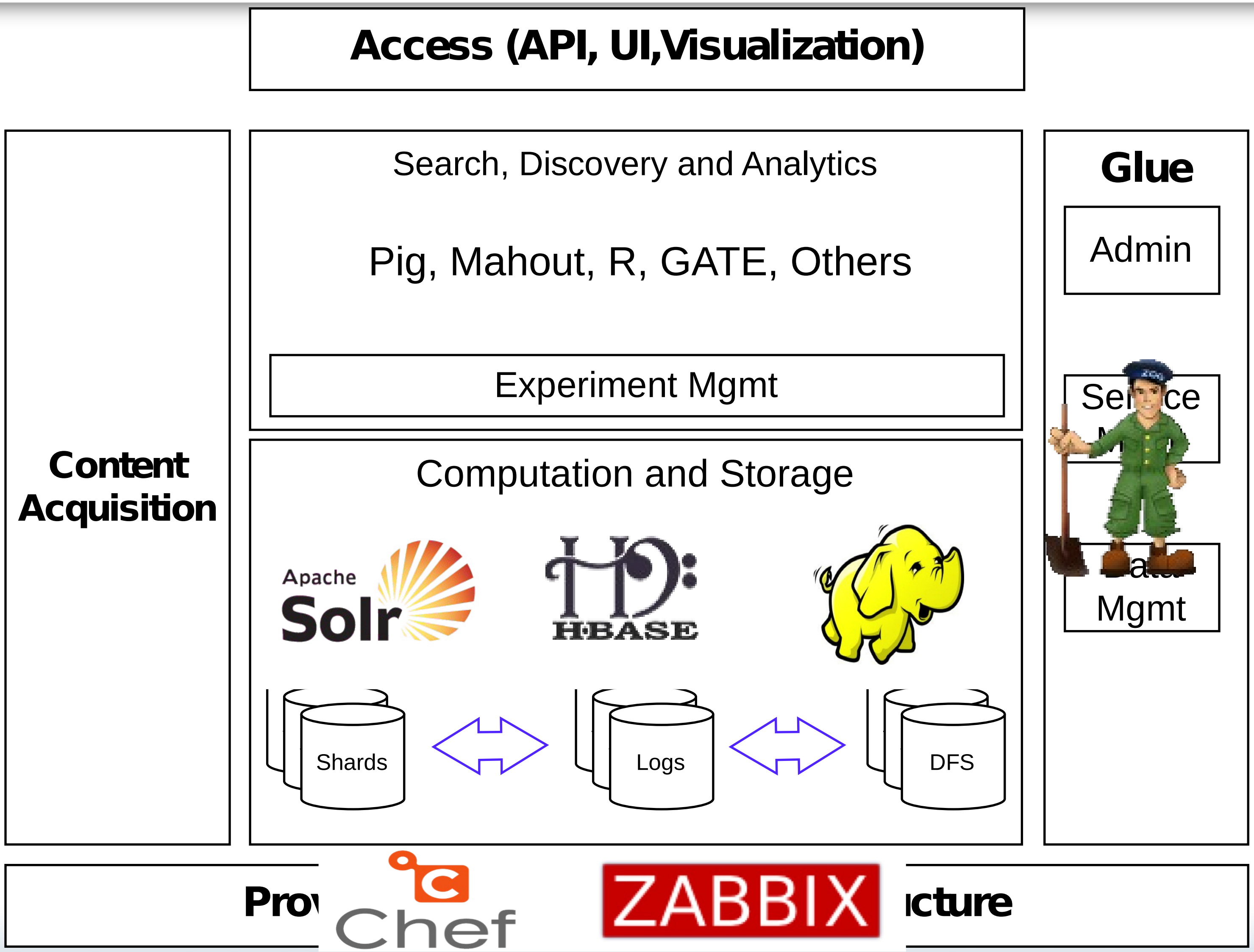
Need whole data consumption and analysis

Experimentation/Sampling tools

Distributed In Memory where appropriate

NLP and machine learning tools that scale to enhance discovery and analysis

Abstract -> Practical SDA Architecture



Access (API, UI,Visualization)

Content Acquisition

Search, Discovery and Analytics

Pig, Mahout, R, GATE, Others

Experiment Mgmt

Computation and Storage

Shards

Logs

DFS

Glue

Admin

Ser ce

a
Mgmt

Prov cture

Computation and Storage

Solr

- SolrCloud
- Document Storage?
- Document Index

Hadoop

- WebHDFS
- Small file are an unnatural act
- Stores Logs, Raw files, intermediate files, etc.

HBase

- User Histories
- Document Storage?
- Metric Storage

Challenges

- Who is the authoritative store? Solr or HBase?
- Real time vs. Batch
- Where should analysis be done?

Search In Practice

Three primary concerns

Performance/Scaling

Relevance

Operations: monitoring, failover, etc.

Business typically cares more about relevance

Devs more about performance (and then ops)

Search with Solr: Scaling and NRT

SolrCloud takes care of distributed indexing and search needs

Transaction logs for recovery

Automatic leader election, so no more master/worker

Have to declare number of shards now, but splitting coming soon

Use CloudSolrServer in SolrJ

NRT Config tips:

1 second soft commits for NRT updates

1 minute hard commits (no searcher reopen)

Search: Relevance

ABT – Always Be Testing

Experiment management is critical

Top X + Random Sampling of Long Tail

Click logs

Track Everything!

Queries

Clicks

Displayed Documents

Mouse/Scroll tracking???

Phrases are your friend

Discovery Components

Serendipity

- Related Items
- Topics
- Recommendations
- Did you mean?
- More Like This
- Trends
- Stat. Interesting Phrases

Organization

- Clustering
  - Named Entities
- Importance
- Time Factors
- Faceting
- Classification

Data Quality

- Duplicates
  - Boosts
  - Length
- Document factor Distributions

Challenges

- Many of these are intense calculations or iterative
- Many are subjective and require a lot of experimentation

Discovery with Mahout

Mahout’s 3 “C”s provide tools for helping across many aspects of discovery

Collaborative Filtering

Classification

Clustering

Also:

Collocations (Statistically Interesting Phrases)

SVD

Others

Challenges:

High cost to iterative machine learning algorithms

Mahout is very command line oriented

Some areas less mature

Aside: Experiment Management

Plan for running experiments from the beginning across Search and Discovery components

Your analytics engine should help!

Types of Experiments to consider

Indexing/Analysis

Query parsing

Scoring formulas

Machine Learning Models

Recommendations, many more

Make it easy to do A/B testing across all experiments and compare and contrast the results

Analytics Components

Commonly used components

Solr

R Stats

Hive

Pig

Commercial

- Starting with Search and Discovery metrics and analysis gives context into where to make investments for broader analytics

Analytics in Practice

Simple Counts:

Facets

Term and Document frequencies

Clicks

Search and Discovery example metrics

Relevance measures like Mean Reciprocal Rank

Histograms/Drilldowns around Number of Results

Log and navigation analysis

Data cleanliness analysis is helpful for finding potential issues in content

Wrap

Search, Discovery and Analytics, when combined into a single, coherent system provides powerful insight into both your content and your users

Solr + Hadoop + Mahout

Design for the big picture when building search-based applications

Find me

- http://www.lucidimagination.com

grant@lucidimagination.com
@gsingers