Lucene
Revolution Conference
October 7-8, 2010

TNR Global
Scalable Web and Search Solutions™

Migrating from Fast ESP to Lucene Solr Search Platform

Presented by Michael McIntosh

TNR Global
Scalable Web and Search Solutions™

Introduction

- Michael McIntosh
  - » Search Architect & VP of Enterprise Search at TNR Global
  - » 10+ Years in Search, 15+ Years in Software Development
  - » Core Member of Lycos Search Engine Team (1997-2001)
- TNR Global, LLC
  - » Web Development and Search Integration Services
  - » LAMP Stack (Linux, Apache, MySQL, PHP, Python, Perl)
  - » Search Integrators using Java, Python, Ruby and C#
  - » Search Engine (Fast ESP, Solr, OmniFind and More)

Agenda

- Define Our Challenges
- Outline Potential Solution
- Identify Core Components
- Explore Specific Use Cases
- Highlights What Was Learned

The Problem

- Largest Clients using Fast ESP for Linux
- No Future in Fast ESP for Linux Platforms

Um, ESP? I think our future together may be in serious jeopardy…

The Problems (cont.)

- Largest Clients using Fast ESP for Linux
- No Future in Fast ESP for Linux Platforms
- Lacking Dynamic Fields & Robust Facet Support
- Limited Ability Modify Result Ranking Algorithm
- Proprietary Code & Limited Community Support

The Problems (cont.)

- Search Migration Path for ESP Clients
- Both Structured & Unstructured Content
- Scalable, Fault-Tolerant, Production Quality
- Content Taxonomy and Drill-Down Navigation
- Web Crawling, HTML & Multi-Page Documents

The Solution

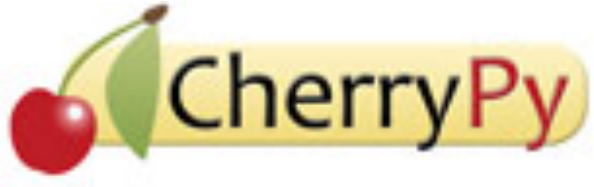
- Apache Solr Search Platform
  - » Robust and Powerful Search Feature Set
  - » Active and Passionate Development Community
  - » Good Lucene and Solr Development Documentation
  - » Community Experts and Commercial Support Options

The Solution (cont.)

- Open-Source Tools for Missing Functionality
  - » Pypes - Document-Centric Processing Pipeline
  - » Heritrix - Highly Configurable Web Crawler
  - » Supervisor - Cluster Node Services Controller

The Solution (cont.)

- ESP Specific Code Migration
  - » Refactor to Decouple Tightly Integrated ESP Code
  - » Utilize RESTful Service Oriented Architecture Solutions
  - » Using CherryPy for Python Based Services
  - » Using Jetty for Java Based Services



The Solution (cont.)

- Platform Agnostic Code Transition
  - » Content Connectors - Database and XML Data Feeds
  - » Content Transformers - ESP FastXML Readers Available
  - » Content Feeding - Trivial to Import Structured Documents

```
<?xml version="1.0" encoding="UTF-8"?>
<documents>
  <document id="http://www.aperturescience.com/item?id=14602&amp;tsId=1931068">
    <element name="catalog_id"><value>1931068</value></element>
    <element name="catalog_name"><value>Aperture Science Catalog</value></element>
    <element name="item_id"><value>4096</value></element>
    <element name="item_category"><value>/Storage/Containers</value></element>
    <element name="item_name"><value>Companion Cube</value></element>
     ...
  </document>
  ...
</documents>
```

APERTURE LABORATORIES

Key Migration Concerns

- What are deal-breakers for our clients?
  - » Solution MUST support highly structured content
  - » Solution MUST support unstructured web content
  - » Solution MUST support parametric search features
  - » Solution MUST support hierarchal taxonomy faceting
  - » Solution MUST support faceting on dynamic fields
  - » Solution MUST support scalable search/indexing architecture
  - » Solution MUST support fault-tolerance & partial fail-over

Key Migration Challenges

- Crawling Unstructured Web Content
  - » Millions of documents from 3rd party websites
  - » Mixture of dynamic and static website content
  - » Mixture of very high and very low quality content
  - » Need to Support HTML and PDF at a minimum
- Feeding Highly Structured XML Content
  - » Millions of products with domain-specific attributes
  - » Mixture of manually and automatically classified content
  - » Taxonomy and structure in nearly constant state of flux

Crawling Web Content with Solr

- Heritrix Web Crawler
  - » Internet Archive's Open-Source Web Crawler
  - » Very Powerful and Highly Configurable Features
  - » Can be configured to mimic ESP crawler behaviors
  - » Can cache documents for later content feeding
  - » Already had experience working with this tool

Feeding Web Content with Solr

- YouSeer API
  - » Open-source search engine framework
  - » Built on top of other open source components
  - » Part of SeerSuite framework.
  - » Utilizes Heritrix for crawling and Solr for indexing
  - » Simple and convenient to use

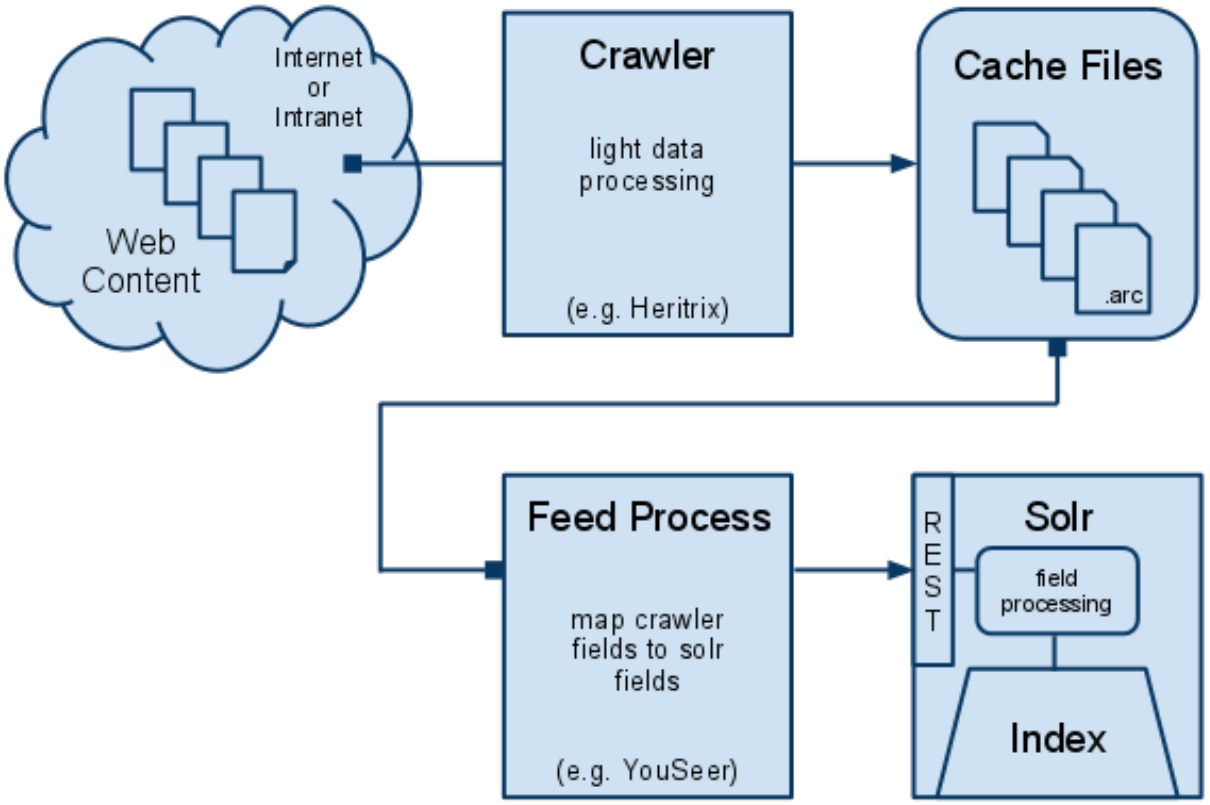
YouSeer

Crawling & Feeding Web Content

HERITRIX

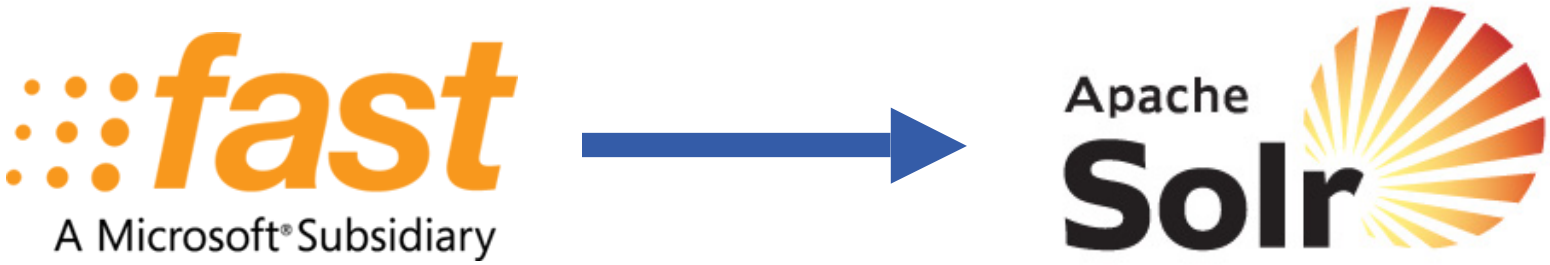
YouSeer

Apache Solr



Feeding Product Content with Solr

- Solr supports XML, JSON, CSV out-of-the-box
- We already transform content to ESP FastXML
- Many options for data import, easily scriptable
- ESP prefers denormalized content, Solr does too



ESP FastXML Content Example

```
<?xml version="1.0" encoding="UTF-8"?>
<documents>
  <document id="http://www.aperturescience.com/item?id=14602&amp;tsId=1931068">
    <element name="catalog_id"><value>1931068</value></element>
    <element name="catalog_name"><value>Aperture Science Catalog</value></element>
    <element name="item_id"><value>4096</value></element>
    <element name="item_category"><value>/Storage/Containers</value></element>
    <element name="item_name"><value>Companion Cube</value></element>
     ...
  </document>
 <document id="http://www.aperturescience.com/item?id=14647&amp;tsId=193764">
    <element name="catalog_id"><value>193764</value></element>
    <element name="catalog_name"><value>Aperture Science Catalog</value></element>
    <element name="item_id"><value>2048</value></element>
    <element name="item_category"><value>/Supplies/Baking</value></element>
    <element name="item_name"><value>Cake Ingredient #42</value></element>
     ...
  </document>
  ...
</documents>
```

Solr Taxonomy Faceting Approach

- At initial pass, Solr does not appear to currently support taxonomy faceting
  - » There are several ways around this including patches
  - » It is relatively easy to resolve if taxonomy is shallow
  - » Taxonomy Faceting Support is around the corner

Electronics » Camera & Photo » Digital Cameras

Our Taxonomy Faceting Approach

- We used fields in schema for top-level and second-level taxonomy categories
  - » Top Level Field Named "Family"
  - » Second Level Field Named "Category"
  - » The facet field are selected based upon user-selection
  - » If no family value selected, faceting occurs on family
  - » If family is selected, faceting occurs on category
  - » If family/category selected, no need to taxonomy facet

Product Attribute Faceting Approach

- We used dynamic fields to store attributes
  - » Attribute name is family_category_attribute=value
  - » We do not facet on attributes until at least Family Selected
  - » During feeding we capture family/category/attribute maps
  - » The front-end leverages f/c/a map to know what to facet
  - » Using this approach, can have preferred attribute field
  - » Only most relevant fields faceted on for each Fam/Cat

Solr Migration: Pros / Cons

- ESP Features That We Miss…
  - » We miss the really nice administration interface
  - » We miss the really nice monitoring interfaces
  - » We miss the numerous content data connectors
  - » We miss the processing pipeline & doc processors
- Solr Features That We Love…
  - » Open-Source, Completely Customizable
  - » Dynamic Fields and Runtime Faceting Support
  - » Active and Passionate Development Community

What We Have Learned about Solr…

- If you have mostly structured data…
  - » With denormalization, it should be trivial to import
  - » You have many ways to get content into Solr
  - » You overall development time could short
  - » There are a lot of people using Solr in this way
- If you have mostly unstructured data…
  - » You need to find a good crawling solution
  - » You will not have all that you need out-of-the-box
  - » Crawling 3rd party content can be a daunting task

Questions?

- Contact Us!
  - Website: http://www.tnrglobal.com
  - E-Mail: info@tnrglobal.com
  - Phone: 413.425.1499

Thank you for your time!