

Building a large scale SaaS app

Open Source, Storage and Scalability

Dan Hanley, CTO

<http://www.magus.co.uk>

14 March, 2008

Agenda

Who are Magus?

- What do we do?
- Who do we do it for?

How do we do it?

- SOA
- Scalability
- Storage
- F/OSS

The Magus proposition

- Leading provider of innovative web-content engineering solutions to global corporations
- Specialise in **managed** applications that help clients build value from their online assets and from the wider web
- Three main applications:
 - **ActiveStandards**
 - **RemoteSearch**
 - **CrucialInformation**
- Delivering solutions since 1995

Our managed applications

Delivering Software-as-a-service (ASP model)

- **ActiveStandards** designed to help companies stay on-brand, on-line by tracking and managing corporate web standards compliance, worldwide
- **RemoteSearch** a multi-site search engine, providing integrated search frameworks for enterprise websites
- **CrucialInformation** a premium current awareness service delivering high-quality, strategic intelligence from the web and syndicated services

ActiveStandards

MAGUS Active Standards
Demonstration

What is this product?
Magus Active Standards enables you to successfully roll-out and monitor brand and other web standards across your entire web presence.

How do I use it?
Click on a country name to see a detailed report on that country website's status. Buttons and tabs give you extra information.

What's the next step?
To find out more, or to request a custom demonstration please contact sales@magus.co.uk or call +44 (207) 019 4700

Position	Country	Region	Pages checked	Errors	Checkpoints failed	Appeals
=1	Japan	Asia	342	0		0
=1	UK	Europe	788	0		0
=1	South Africa	Asia	304	0		0
4	France	Europe	507			
5	Germany	Europe	388			
6	Corporate.com	None	2113			
7	Argentina	Americas	290			
8	USA	Americas	545			
9	Czech Republic	Europe	278			

MAGUS To find out

Information Home
Summary for corporate.com
Select an archived report ▾

CHECKPOINTS REPORTS DIAGNOSTICS

6 November 2006 [Hide checkpoints with errors] RESULTS


01 Brand

ID	Checkpoint	Errors	Appeals
1.1	Only the corporate fonts are used	0	0
1.2	Only colours contained in the brand palette are used	0	0
1.3	Brand/product name presentation should follow guidelines	5	2
1.4	Ensure images used relate to brand themes	0	0
1.5	All images should conform to specified dimensions	0	0
1.6	All images should follow rules for file size	9	3
1.7	Links to other websites should not open in a new window	0	0
1.8	Feature boxes should not exceed 6 lines in length	12	0
1.9	External links should open in a new window	0	0
1.10	All website pages should contain links to the corporate homepage	0	0
1.11	No more than 1 Flash animation should appear in the right hand column	0	0
1.12	All thumbnail images should contain a page link	18	0
1.13	All pages should contain correct 'topbar' links	0	0
1.14	No more than 5 feature boxes should appear on a page	0	0


Click on an item in the list on the left, to show the error/report data.

MAGUS To find out more about Magus Active Standards, contact sales@magus.co.uk or call +44 (207) 019 4700
© Magus 2006

RemoteSearch



[Shell.com](#) | [Shell Directory](#)
[accessibility](#) | [help](#) | [contact](#) | [sitemap](#)

search 

[www.shell.com](#)

[Home](#)

[Shell for Motorists](#)

[Shell for Businesses](#)

[Shell for the Home](#)

[Shell Directory](#)

Search - www.shell.com

Enter Your Search Criteria

Search for using

include PDFs

Results: 37 search results found
 Page 1 of 4 pages of search results - view: 1 | 2 | 3 | 4


Environment and Society - In The Shell Sustainability Report 2005
 Shell in alternative energy: \$1 billion invested since 2000. Broadest alternative energy use of biofuels. Hydrogen filling stations in five countries. One of the largest wind power projects. Leading developer of next generation solar technology.

Investor Centre - Volkswagen, Shell and Iogen to study feasibility of producing ethanol in Germany
 Letter of intent signed at Detroit Auto Show, Detroit, January 8, 2006 - Volkswagen, Shell and Iogen Corporation announced a joint study to assess the economic feasibility of producing cellulose ethanol in Germany. This advanced biofuel production can cut CO2 emissions by 90% compared with conventional fuels.

Media Centre - Volkswagen, Shell and Iogen to study Feasibility of Producing Ethanol in Germany
 Detroit, January 8, 2006 - Volkswagen, Shell and Iogen Corporation announced a joint study to assess the economic feasibility of producing cellulose ethanol in Germany. This advanced biofuel production can cut CO2 emissions by 90% compared with conventional fuels.

Shell in Australia - Shell's Position on Ethanol in Fuels
 Although Shell in Australia does not currently use ethanol in its fuels, we believe we should also believe that a lot of further development, discussion and consultation needs to be done.

Shell Nigeria - Shell: sustainability at the heart of our business
 Royal Dutch Shell plc today released its ninth annual report on its environmental and social commitment to help meet the world's current and future energy needs in an environmentally sound and sustainable way.



feel good, look good
and get more out of life

Our brands
Our values
Our company

[Home](#) > Search results > information

Search


259 results matching **information**

Search again:


Optionally restrict your search to:

-
-

1. [FAQs](#)
 This page lists a selection of frequently asked questions and answers that should help you in your search.
www.unilever.com/resources/faq.asp
 Language: English
2. [Nutrition](#)
 We can make a difference to the diets of millions of people. The challenge is to make our products the healthy choice for consumers without compromising taste, convenience and affordability.
www.unilever.com/ourvalues/environment-society/sus-dev-report/nutrition/
 Language: English
3. [Information management technology](#)
 Information Management and Technology (IM & T) gives you the opportunity to improve the way we do business.
www.unilever.com/ourcompany/careers/our_people/information_management_and_technology/
 Language: English
4. [Unilever plans faster growth](#)
 22/02/2000: Unilever today detailed its plans to accelerate top line growth and step up its investment in research and development.




Brand information




[Can't find it?](#)

Unilever websites



Latest publications



[Download library](#)

Resources

- [Contact us](#)
- [Download library](#)
- [FAQs](#)
- [Site map](#)
- [RSS](#)

Fast track on this site for

- [investors](#)
- [journalists](#)
- [job seekers](#)
- [CSR analysts](#)
- [food service professionals](#)
- [suppliers](#)

CrucialInformation

This screenshot shows the 'Cardiovascular' section of the Crucial Information website. The left sidebar contains a navigation menu with categories like 'Companies' (BAE Systems, Citigroup, GSK, Sainsbury, Shell) and 'Topics' (Cardiovascular, Knowledge Management, Power Generation, Global Brand Management, TMT). The main content area lists several articles, including 'Drug Interaction Inhibits Tamiflu', 'Ablynx Enters into a Research an...', 'ARCA Discovery Names Richard B...', 'Psychological, Behavioral Therap...', 'EPIX Announces Canadian Approv...', 'Coenzyme Q10: Should You Take...', 'Abbott to Expand Presence in Lip...', 'Pharmacopeia Announces the Ap...', and 'Boston Scientific Announces Japa...'. At the bottom, a search bar shows '388 items found' and 'Page 1 of 39'.

This screenshot shows the 'Global Brand Management' section of the Crucial Information website. The top navigation bar includes 'Home', 'About this service', and 'Contact us'. The left sidebar lists 'Companies' (BAE Systems, Citigroup, GSK, Sainsbury, Shell) and 'Topics' (Cardiovascular, Knowledge Management, Power Generation, Global Brand Management, TMT). The main content area features a search bar with 'Global Brand Management' selected as the filter. Below the search bar, a list of articles is displayed, including 'Branding is a way of business', 'Tesco to go for China branding', 'Bloggers score low on trust scale in region', 'Malaysians look at reputation, brand when buying online', 'Keep close control to brand successfully', 'Names in the Age of Branding Gone Wild', 'Online Reputation Management: The New PR', 'Mark Ritson on branding: Don't be scared of...', 'Vneshtorgbank to become VTB under rebranding', and 'Use transparent marketing for brand success'. At the bottom, a search bar shows '41 items found' and 'Page 1 of 5'.

Social Networking

The screenshot shows the IET Discover website interface. At the top, the logo 'IET discover' is displayed with 'BETA' and the tagline 'Collaborative intelligence in science, engineering and technology'. Navigation links 'find', 'share', and 'connect' are visible. Below the header, there are tabs for 'All Discover' (selected) and 'My Discover'. A secondary navigation bar includes 'Bookmarks', 'Groups', 'People', and 'Industry news'. The main content area is titled 'Discover bookmarks' and contains a search bar with 'nanotech' entered. Below the search bar, it shows '1 - 10 of 40 bookmarks matching "nanotech" in IET Discover'. Two search filters are shown: 'Full text' (selected) and 'Tags only'. The results list two items:

- Convergence of Nanotechnology and Green Building is Creating Fresh Economic Opportunities and Environmental Benefits**
http://www.azonano.com/News.asp?NewsID=4989
In this edition of AZoNano News, as well as bringing you the latest news from the international Nanotechnology industry, we'll be talking with Stephan Stucklin from Nanosurf about the compact size of the ... [More...]
First added: 24 September 2007, by: Dan Hanley
Rate this bookmark: ★★☆☆☆ 1 rating [Report as spam]
Tags: fabrication, nanorods, nanotech, nanotechnology
- Nanotechnology - Foresight Nanotech Institute**
http://www.foresight.org/
Advancing Beneficial Nanotechnology Foresight is the leading think tank and public interest institute on nanotechnology. Founded in 1986, Foresight was the first organization to educate society about ... [More...]
First added: 17 September 2007, by: David Wiblin
Rate this bookmark: ★★☆☆☆ 2 ratings [Report as spam]
Tags: nanotech, nanotechnology, advance, beneficial, foresight [More...]

On the right side of the page, there is a 'Results analysis' section with 'Help' and 'Get browser toolbar' buttons. Below it is a 'Related' section with tabs for 'tags', 'groups', and 'people'. The 'tags' tab is active, showing a list of related terms such as 'biotech', 'nanotechnology', 'environmental protection agency', and 'nanotechnology'.

Our clients

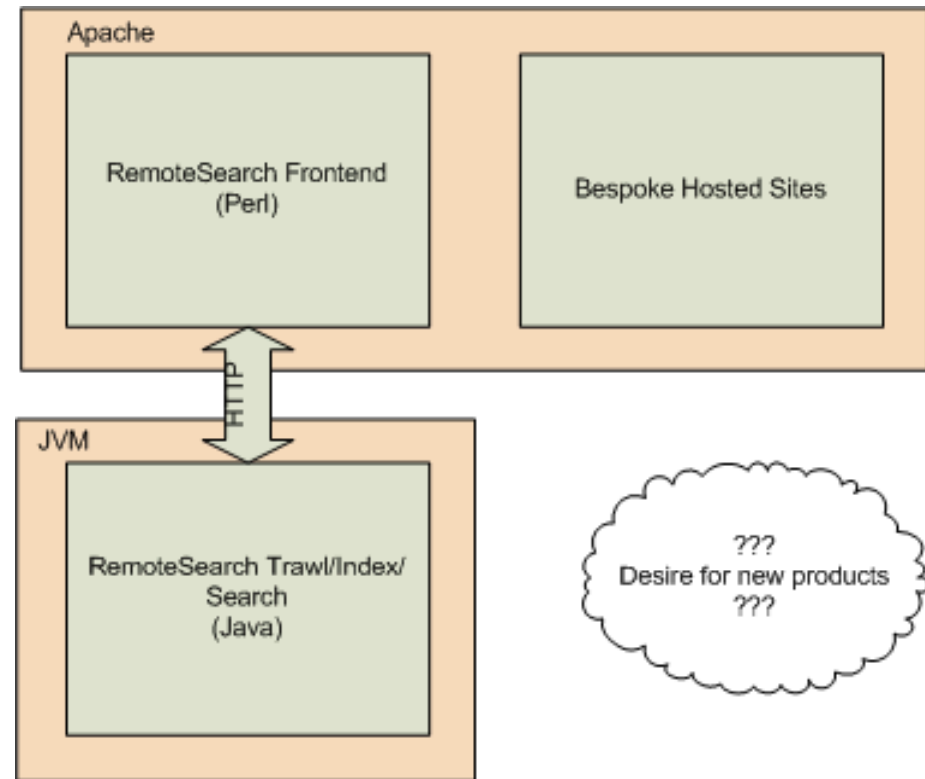


MAGUS

Technically - where we were

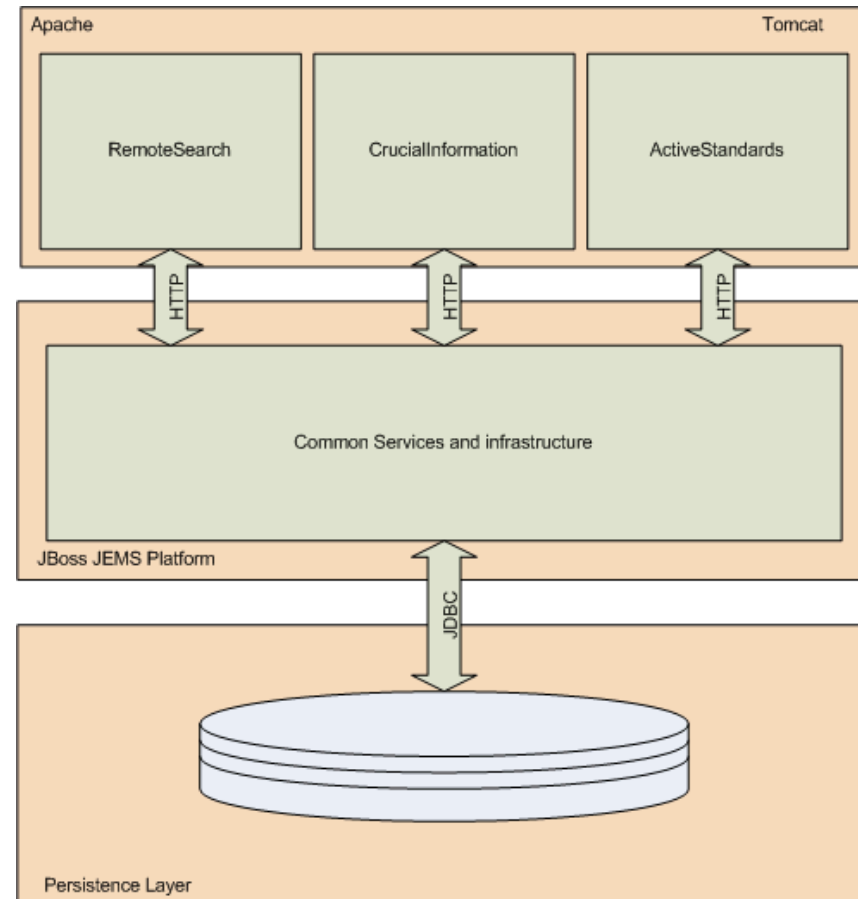
1 product

- Web design business
- All home grown
- No appservers
- No failover
- No common infrastructure
- Scalability worries
- No version control
- Unclear methodology



Technically – where we are now

- 3 main applications
- Bespoke capability
- Common infrastructure
- Platform of services
- Fault tolerant
- Scalable
- Defined process & methodology



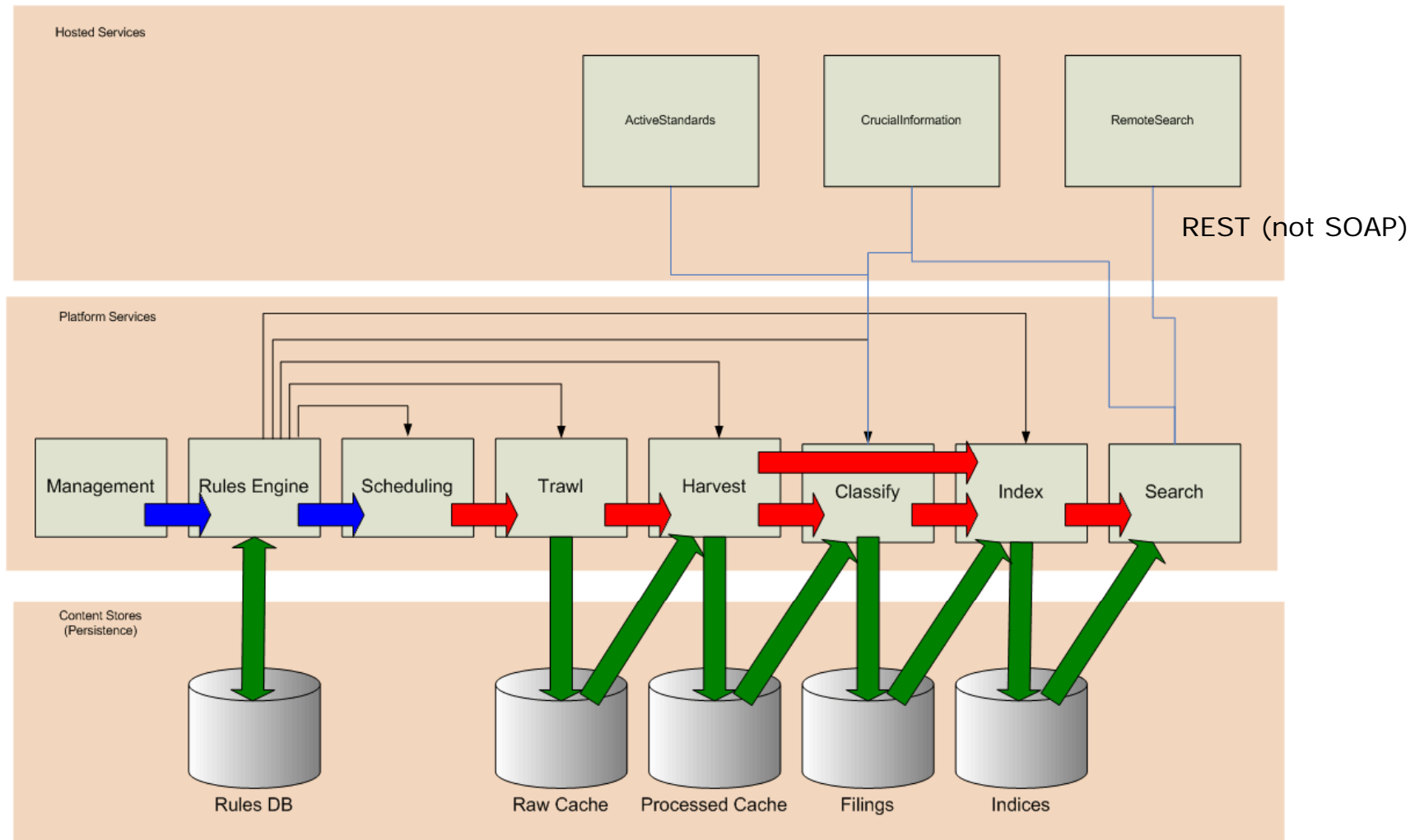
Approach

- Do a lot with a little – 35 people, punching above our weight
- Don't reinvent the wheel
- Extract commonality – keep it DRY

The components of the stack

- Trawl
- Harvest
- Index
- Search
- Analysis
- Monitor
- Routing
- Store
- Quartz
- ClientEngine
- Profile
- LinkChecker

Logical architecture

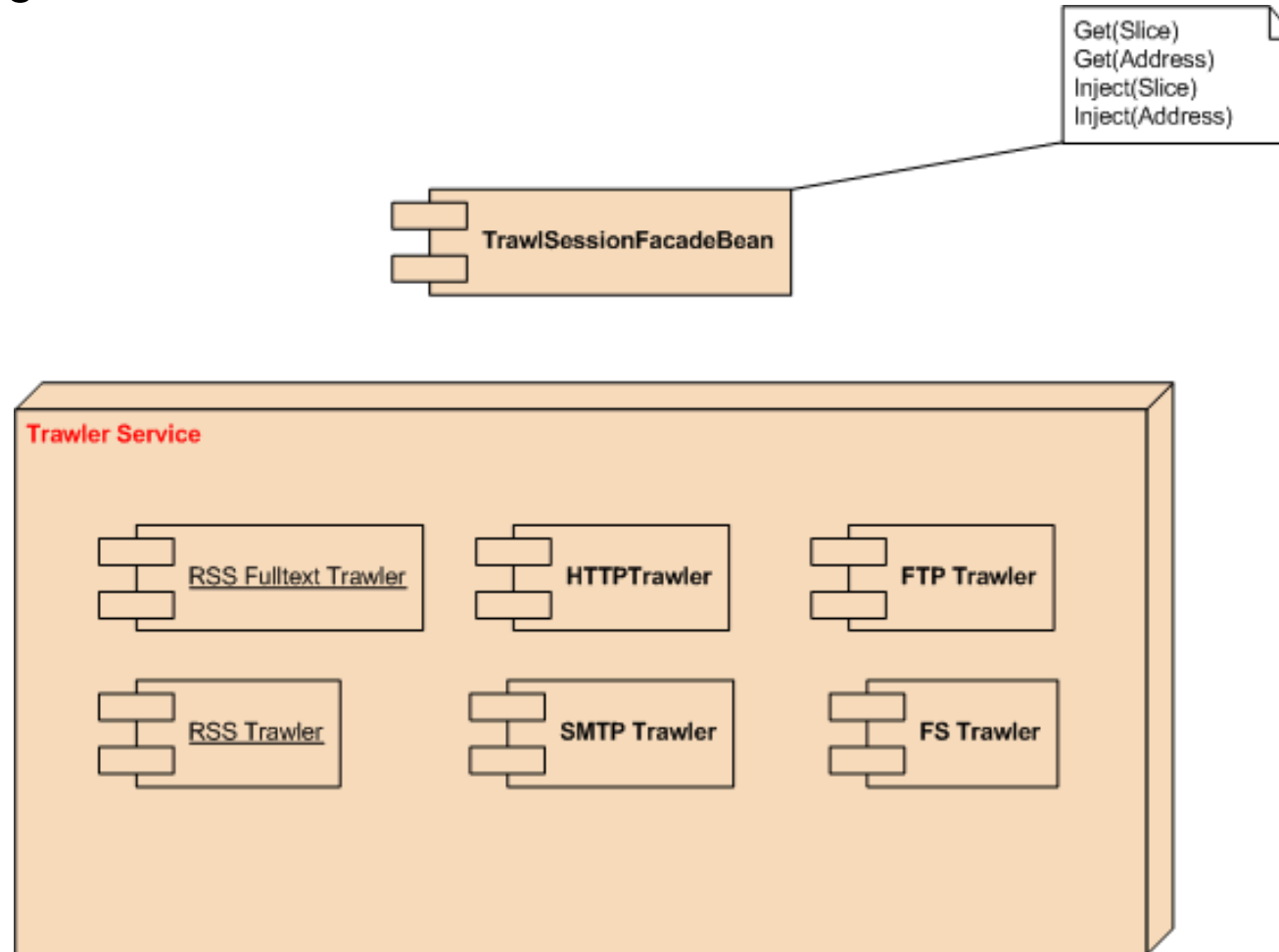


Trawl

- Responsible for managing the gathering of data in its raw form into the Store.
- Currently have Trawlers for:
 - HTTP
 - FTP (several flavors)
 - RSS, Atom etc
 - SMTP
 - Google
 - Technorati
 - Moreover
 - FT (several flavors)

Trawler service

Pluggable architecture based on JMX Mbean service

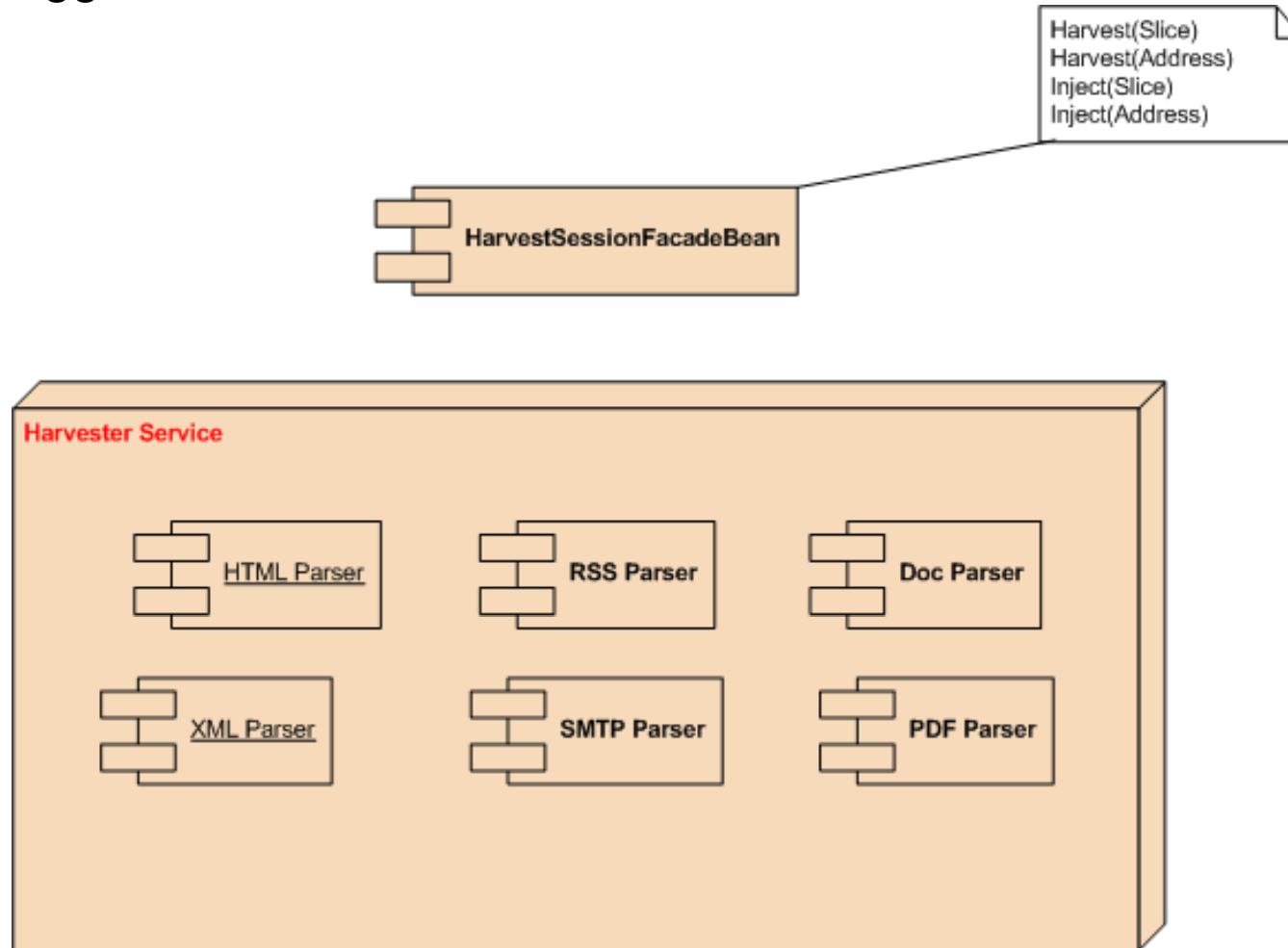


Harvest

- Responsible for extracting explicit data from Links and storing the fielded data in the database, and the non fielded data in the Store.

Harvest service

Pluggable architecture based on JMX Mbean service



Index

- Responsible for building, purging, maintaining indices.

Search

- Responsible for searching indices and delivering results.

Analysis

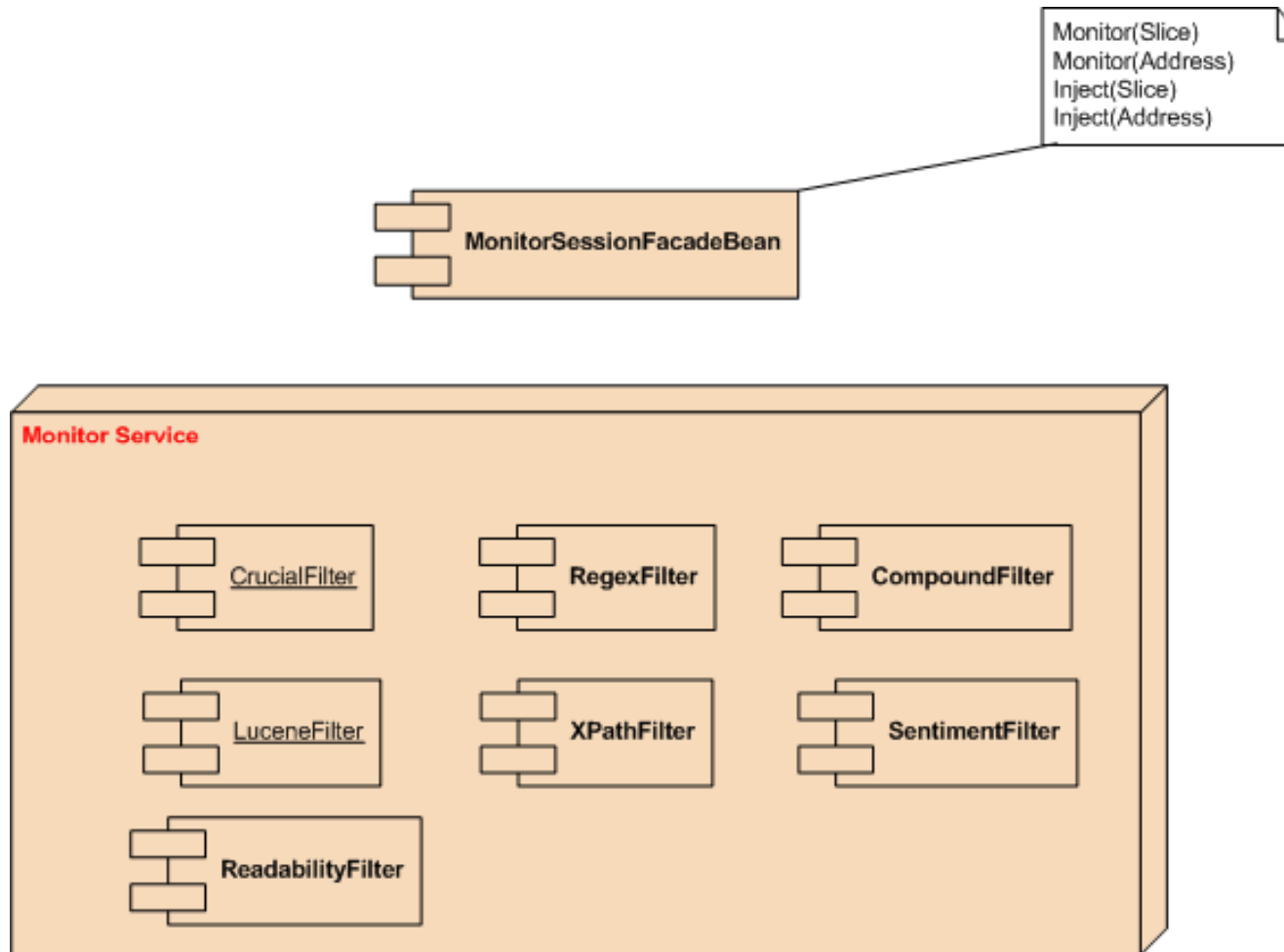
- Responsible for deriving scores for information implicit in the page
 - Sentiment
 - Readability
 - Language detection etc

Monitor

- Badly named, should be called “Classifier”
- Responsible for creating filings between Links and Categories.
- A Link can be a bookmark, news item, blog article etc.
- A Category can be Users Bookmarks, News Topic, an AST Guideline etc.

Classifier (monitor) service

Pluggable architecture based on JMX Mbean service



LinkChecker

- Responsible for checking the life of links and removing them correctly from the system when they have expired.

Routing

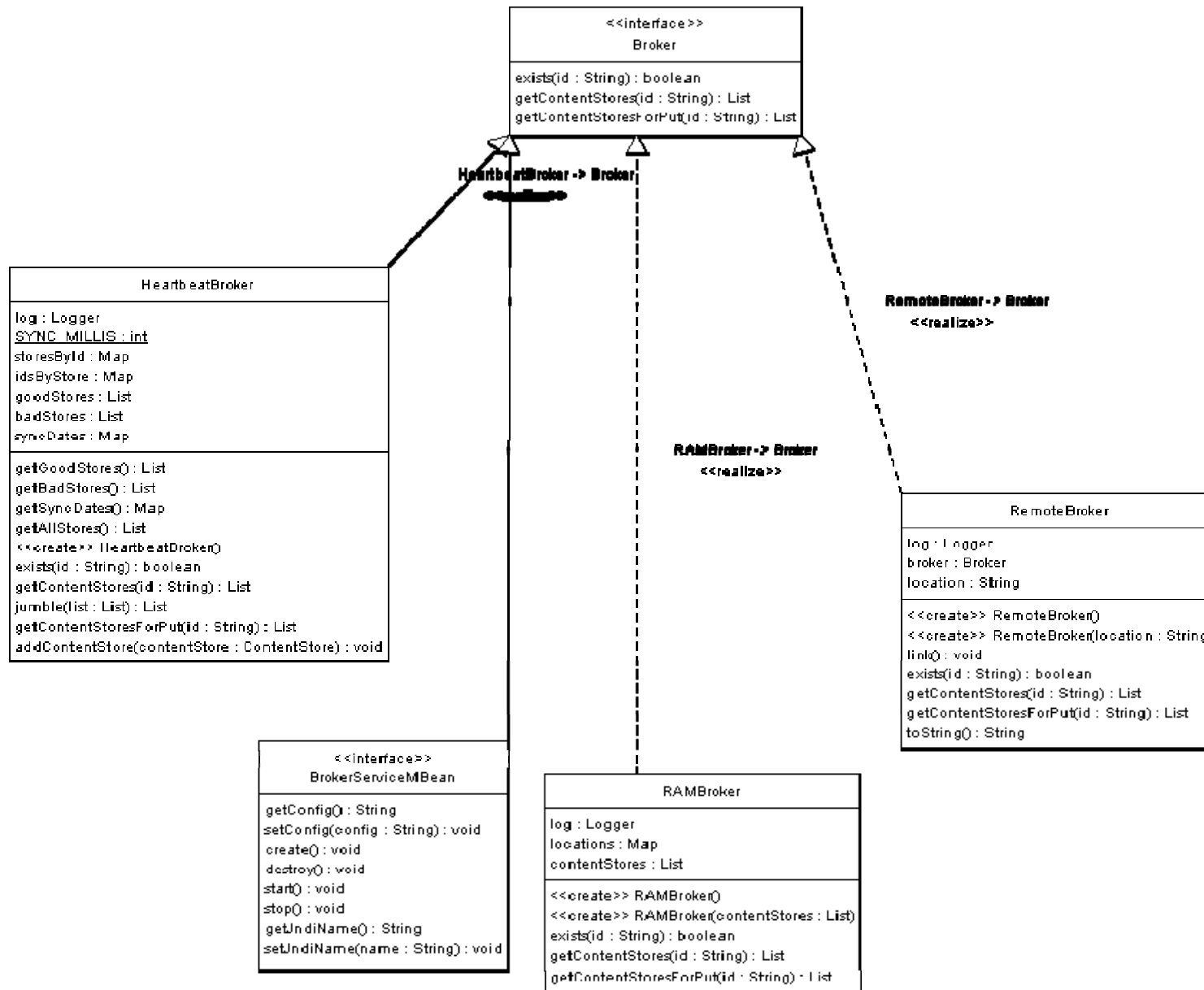
- Manages the workflow of jobs through the stack
- Has the capability to dynamically loadbalance workloads.

Content stores

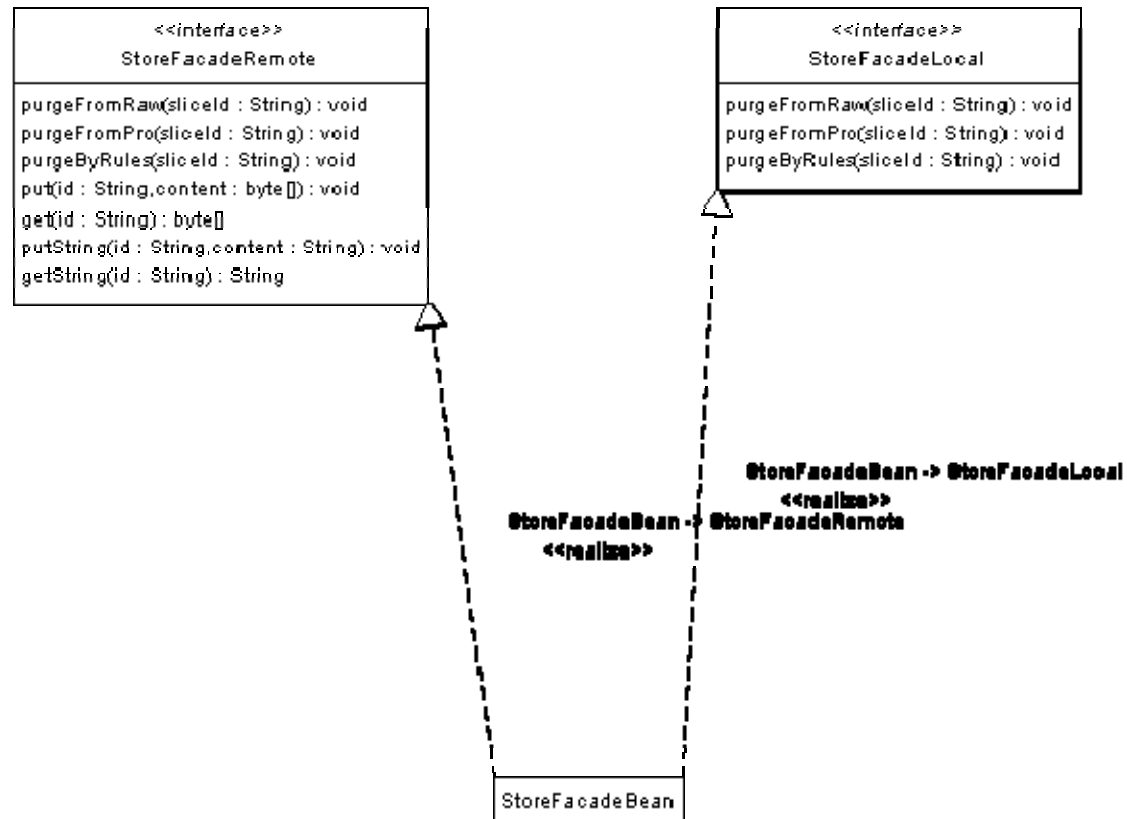
- We needed a multiple terabyte (currently 24 TB) distributed, fail safe, filesystem
- NFS was crumbling under load
- ZFS was vapourware
- Lustre was too complex
- We built our own!
- Magus Contentstores, responsible for holding both the raw and processed non fielded content of links which have been trawled and harvested

Content stores - configuration

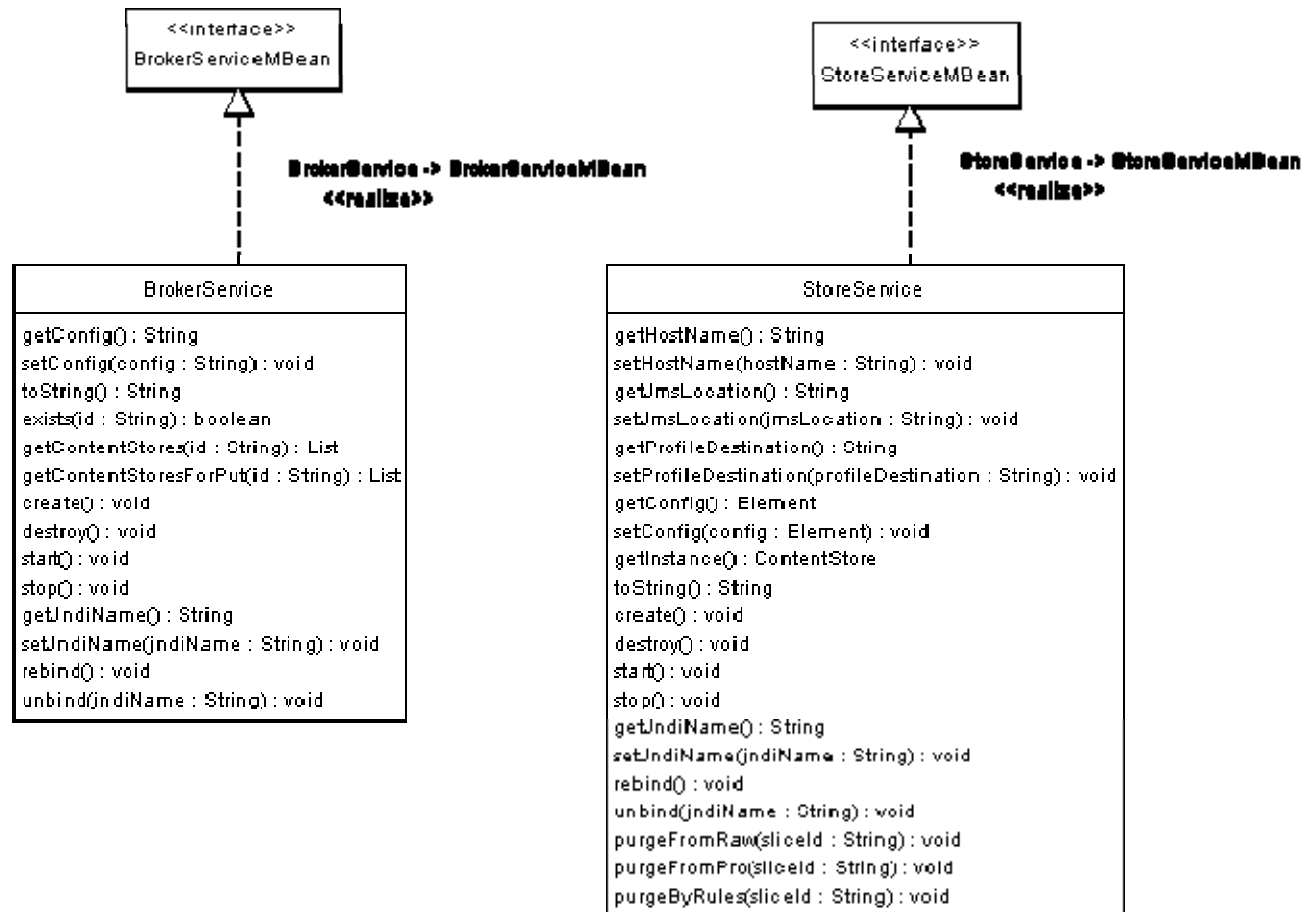
```
<mbean code="uk.co.magus.store.service.StoreService" name="magus.service.store:service=StoreServiceLocalCalls">
  <attribute name="JndiName">magus/services/StoreServiceLocalCalls</attribute>
  <attribute name="Config">
    <TryEachStripeStore>
      <List>
        <MirrorStore>
          <List>
            <RemoteStore>nas:1299;StoreServiceRemoteCallsInvokeTarget</RemoteStore>
            <RemoteStore>m4:1099;StoreServiceRemoteCallsInvokeTarget</RemoteStore>
          </List>
        </MirrorStore>
        <MirrorStore>
          <List>
            <RemoteStore>nas:1199;StoreServiceRemoteCallsInvokeTarget</RemoteStore>
            <RemoteStore>m5:1099;StoreServiceRemoteCallsInvokeTarget</RemoteStore>
          </List>
        </MirrorStore>
      </List>
    </TryEachStripeStore>
  </attribute>
  <depends>jboss:service=Naming</depends>
</mbean>
```



Store Interfaces



Store JMX Beans



Contentstore - engines

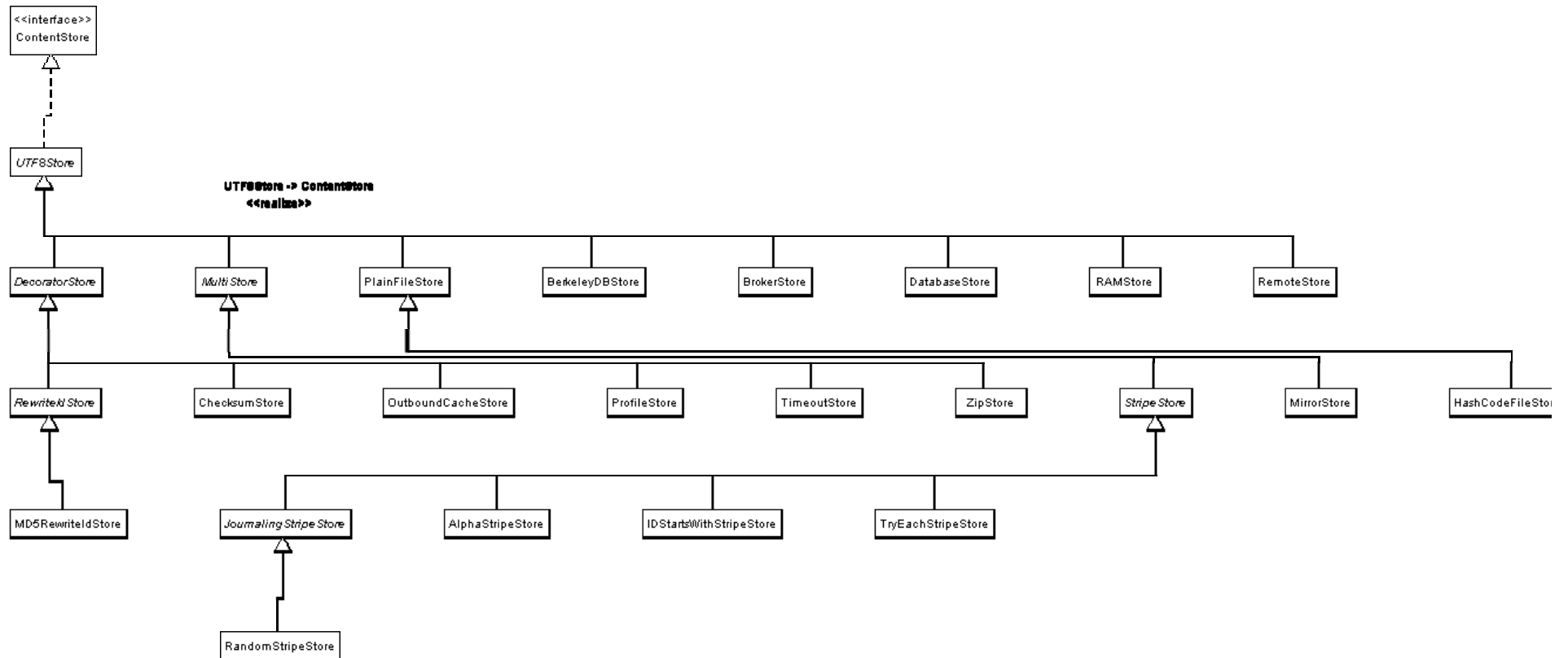
Can use many types of engine on a node

Currently supports:

- Mysql
- SleepyCat
- Filesystem

These can be decorated to enhance functionality

Content Store Classes



Quartz

- Responsible for firing messages on time.
- The “heartbeat” of the stack.

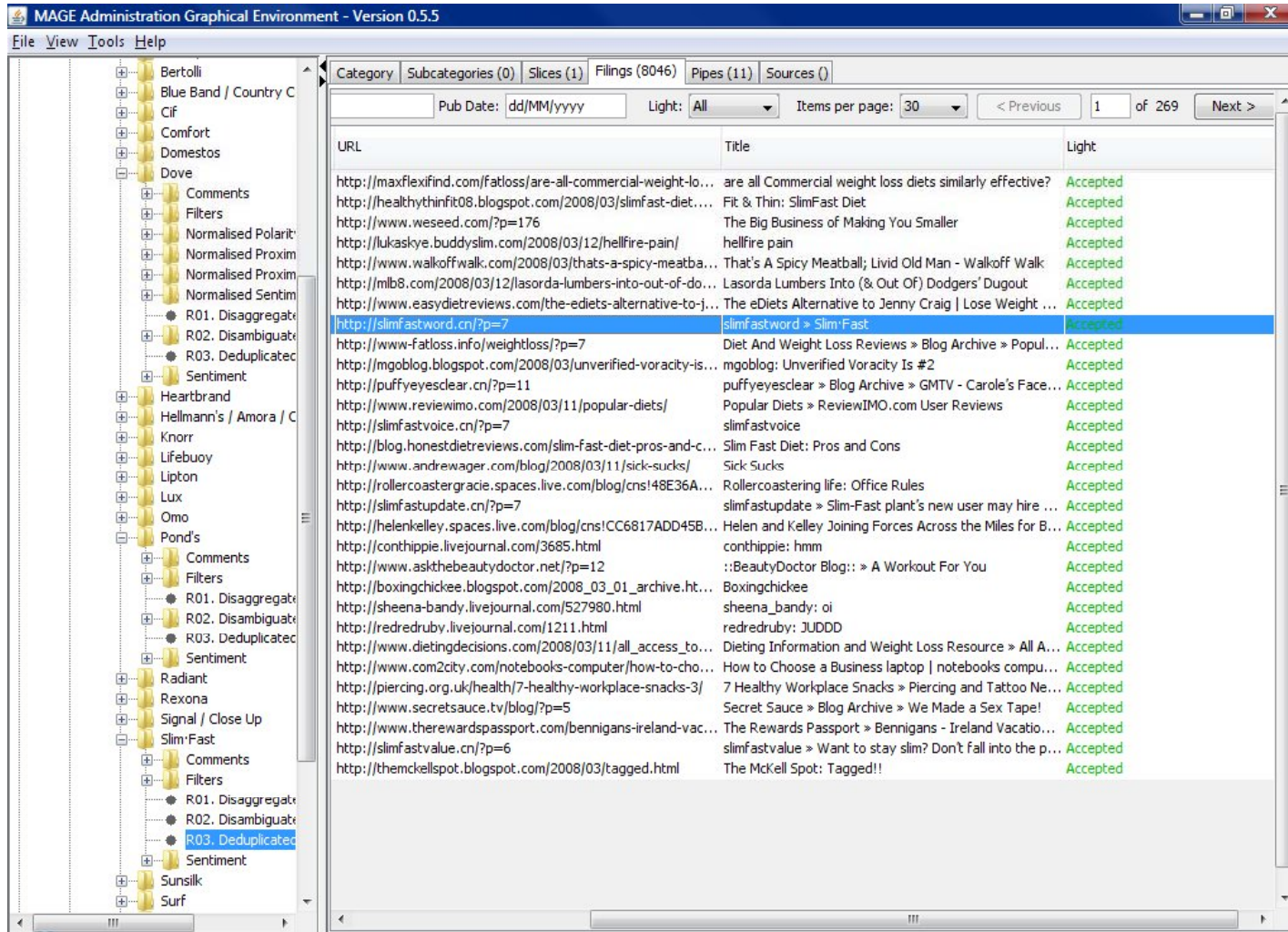
Client Engine

- Responsible for stack based processing for Client Applications.
- Keeps “heavy lifting” out of the Web Tier.
- Coordinates Client Applications requests across multiple stack services.

Management Application

- Manage taxonomy
- Manage rules
- Manage scheduling
- Focus on managing the business
- Leave service management to JMX or web consoles
- Swing

Management App



Management App

The screenshot shows a web browser window titled "MAGE Administration Graphical Environment - Version 0.5.5". The browser address bar displays the URL: http://www.shell.com/home/content/in-en/shell_for_motorists/site_locator/Ahm_khokhra_road.html. The page content includes the Shell logo, navigation links (Accessibility, Help, Sitemap, Search), and a "Shell Station Locator" section for "AHMEDABAD - Khokhra Road". A map is displayed with a red border, showing the location of the Shell Retail Outlet relative to the Kataria Maruti Show Room, Anupama Cinema, and Apparel Park. The map also indicates directions to the Baroda Express Highway, CG Road, and Rakhhyal Industrial Area. Below the map, the address is listed: Shell India Marketing Pvt. Ltd, Retail Outlet Premises, TP No. 7, Plot No. 6, Ambika Mills, Near Anupama Cinema. The browser status bar at the bottom shows "Done" and "AST Highlighted Content".

Management App

The screenshot shows the 'Pipe Edit' application window with the following configuration:

- Source:** R03. Deduplicated
- Destination:** Proximity (positive)
- Pipe ID:** d2c5abc112b932da0112b9d25f6800e3
- Default Light:** Green (selected)
- Filter Details:**
 - Filter Tree:** New filter
 - Profile:** Filter name: New filter, Notes: (empty)
 - Filter type:** LuceneFilter
 - Category:** General Filter Type
- Info:** Filter ID: d2c5abc112b932da0112b9d24e9500e2, Filter last updated: 31 May 2007 - 03:19:58 PM
- Terms:**

```
"slimfast Ace"~3
"slimfast Adore"~3
"slimfast advantage"~3
"slimfast Amazing"~3
"slimfast Award"~3
"slimfast Beautiful"~3
"slimfast best"~3
"slimfast Better"~3
"slimfast Boost"~3
"slimfast Breakthrough"~3
```

At the bottom, there are buttons for 'Save & Close', 'Cancel', 'Apply Changes', and 'Test this pipe', along with a dropdown menu set to 'on most recent 100 items'.

Management App

Slice Edit: Unilever Blog Monitoring slice for Most-commented links (FULL INDEX)

Name: **Unilever Blog Monitoring slice for Most-commented links (FULL INDEX)** Slice ID: 55d632f41367759f011371d377840b7e

Overview Configuration Categories Starting URLs / Rules Control Search

Display Name: Unilever Blog Monitoring slice for Most-commented links (FULL INDEX) Home page: http://www.magus.co.uk

Internal depth: 0 Collect internal links on page Use body digest for trawl limit Trawler Type: Blog Link Update Trawler

Max downloads: 100 Collect external links on page Big job slice: Yes Title Field Strategy: use regex

Content threshold: 0 Download images in html pages Download stylesheets in pages

Content Length

User Agent: Encoding: auto Search Term:

Duplicate match: Link Match Remote Search Slice Blog Feed Id:

Use body digest for deduping Obey Robots.txt Client name:

Dedup on content digest Is remote search slice

Lock links on slice

Cron: 29 40 6,20 * * ? Edit Delete (is saved independently of other changes to this slice)

Regexpert: Priority: Normal Safe Name: unileverblogmonitoringsliceformostcommentedlinksfullindex

Navigator: Account:

Username: Page expiry: 0

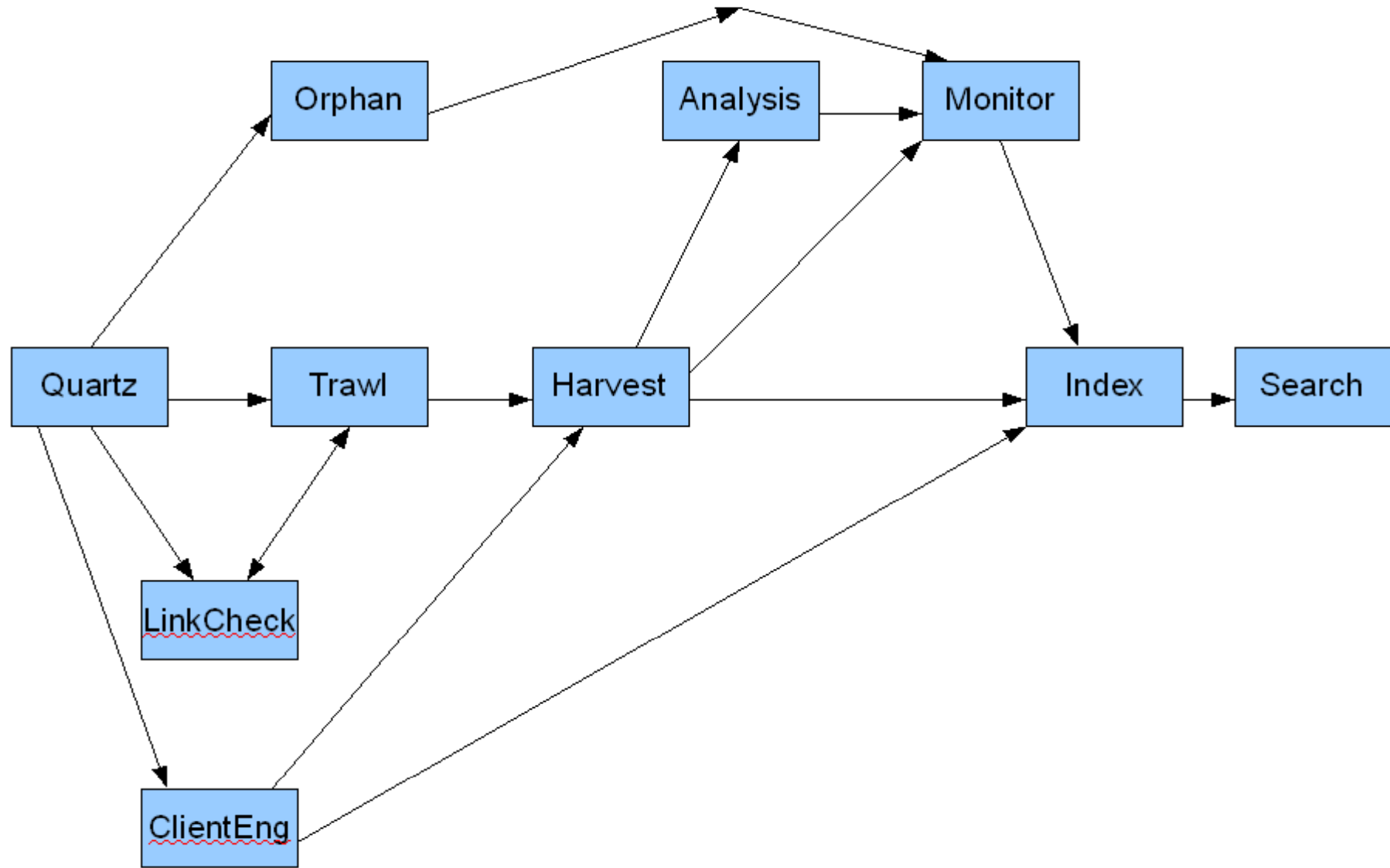
Password: Locale:

Save & Close Cancel Apply Changes

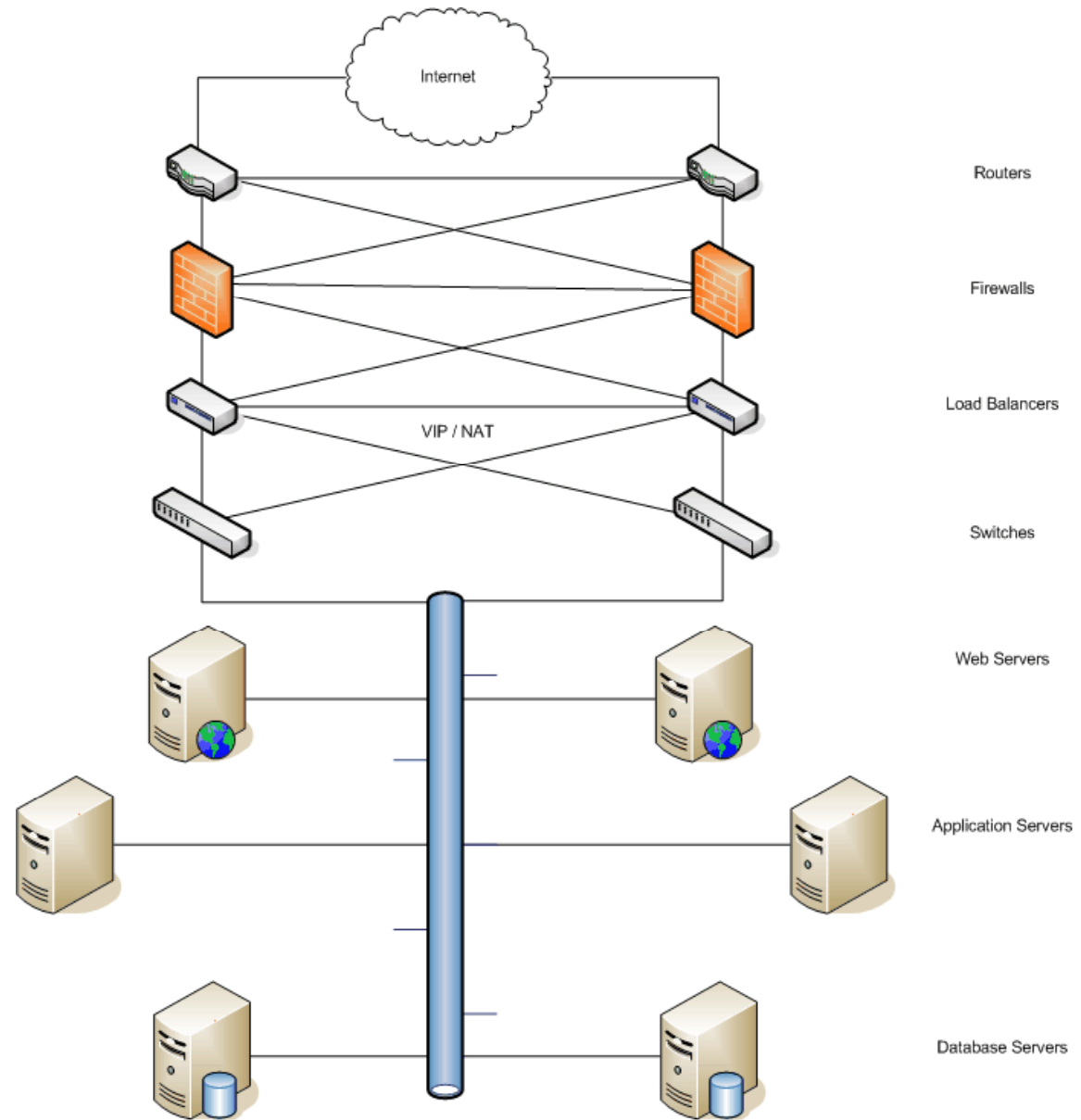
MAGUS

Profile

- An internal service used to collect metrics on system wide performance



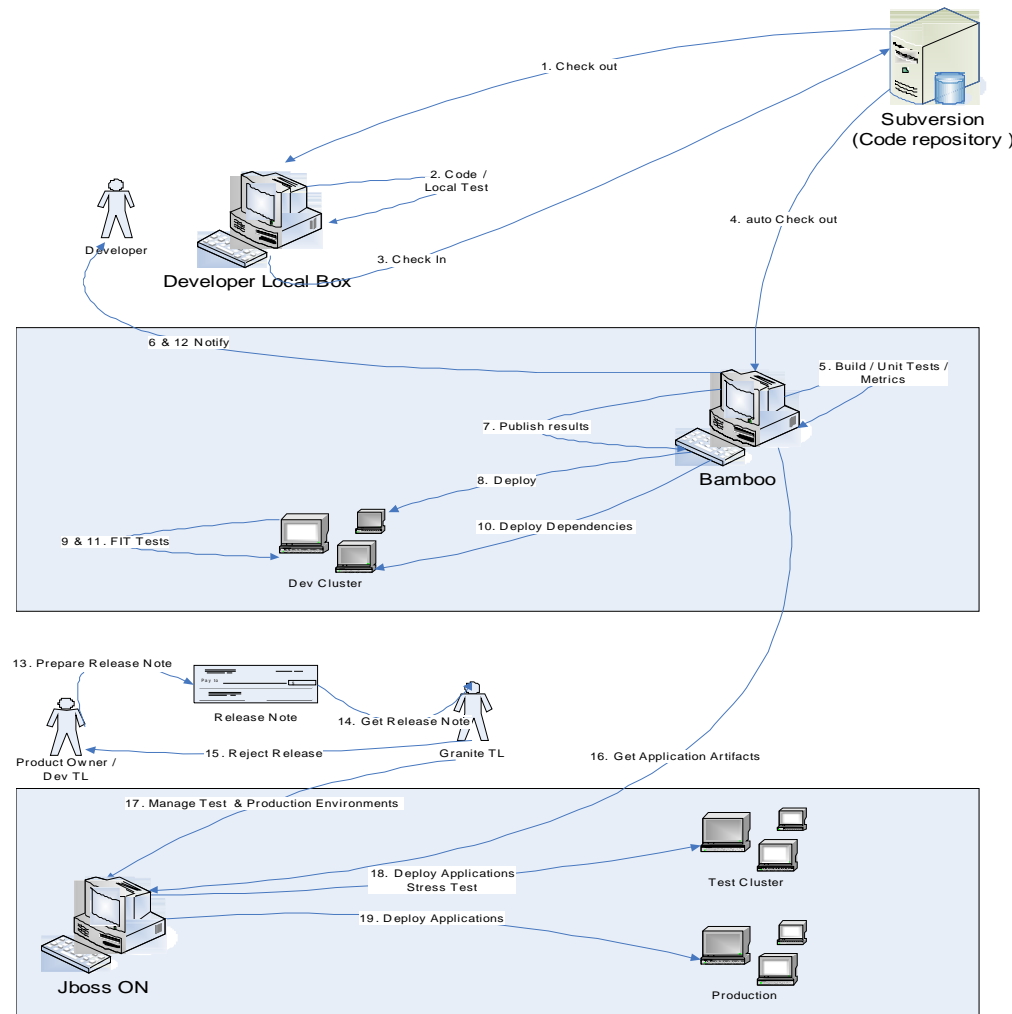
Infrastructure architecture



Methodology

- Agility – sprints
- Issue tracking – Jira
- Regular, scheduled, deployments
- Consolidated build & version control

Deployment



Throughput

- 11,000 sources in system
- ~16,000,000 pages rolling store
- ~200,000 new pages per day
- Average < 2 minutes from page detection to fully classified and indexed.

Cost comparisons

- Apples and oranges?

Proprietary			
Product	Per CPU	CPUs	Total
Oracle	20,000.00	10	\$200,000
Weblogic AS	10,000.00	38	\$380,000
MS Windows Server	3,919.00	48	\$188,112
Visual Team Studio	1,000.00	12	\$12,000
ClearCase	4,125.00	1	\$4,125
Jira	2,000.00	1	\$2,000
Autonomy IDOL bundl	75,000.00	2	\$150,000
IBM Intelligent Datami	132,000.00	1	\$132,000
Verity K2	50,000.00	2	\$100,000
			\$1,068,237
			£580,531.26
			€349,629.77

Licence Free			
Product	Per CPU	CPUs	Total
MySql	\$0.00	10	\$0.00
Jboss AS	\$0.00	38	\$0.00
Redhat/Ap:	\$0.00	48	\$0.00
Eclipse	\$0.00	12	\$0.00
Subversion	\$0.00	1	\$0.00
Trac	\$0.00	1	\$0.00
Carrot2	\$0.00	12	\$0.00
LingPipe	\$0.00	12	\$0.00
Lucene	\$0.00	8	\$0.00
UIMA	\$0.00	12	\$0.00
			\$0.00

Thank you

Questions?