SolrEcosystem

This page documents that relate to the broader Solr Ecosystem of Search and Information Retrieval and may be of interest to Solr users. Some of these projects may have features that can be used directly in conjunction with Solr. You may also be interested in the IntegratingSolr page which documents client libraries and projects for integrating Solr with other applications not specific to Search/IR.

- Solr distributions / forks
- Data acquisition
    - Crawlers And Connectors
    - Pipelines / Document Processing
    - Indexing
- Monitoring

Solr distributions / forks

- Rivulet Enterprise Search
- Constellio Enterprise
- Lucidworks Fusion - commercially supported platform extending and enhancing Solr with various enterprise demanded capabilities, including security, connectors, indexing and querying pipelines, signal handling, analytics, etc.
- Solandra - A tight integration of Solr and Cassandra. The result is Solr with the awesome scalability properties of Cassandra.

Data acquisition

There are numerous ways to bring data into Solr. Many people roll their own solution or use the DataImportHandler

Crawlers And Connectors

Web, email, and file crawlers (alphabetically).

- Aperture (web, email, file)
- Crawl-Anywhere (web) Solr Info
- DataImportHandler (email, file)
- Droids (none) Solr Info
    - Presently, more of a framework for a crawler.
- Heritrix (web)
- ManifoldCF (web, file) Solr Info
- Norconex Collectors (web, file) Solr Info
- Nutch (web) Solr Info (included as part of the Nutch Tutorial)

Pipelines / Document Processing

Frameworks for flexible document processing. See DocumentProcessing for more background and criteria for a proposal. Some crawlers/connectors have their own pipeline capability and they are not repeated here.

- Aspire (by Search Technologies) - integrates with Solr. Not open-source but free.
- Hydra (by Findwise) - integrates with Solr.
- Behemouth - outputs to Solr
- OpenPipe (alt) Solr Info
- OpenPipeline
- ETL (Extract Transform Load) – many are applicable; these are a couple notable ones:
    - Talend Open Studio (TOS)
        - Custom Talend components for SOLR can be found on Talend Forge Exchange with associated doc and tutorial
    - Kettle (Pentaho)
    - CloverETL

A common problem amongst the ETLs is that each step in the pipeline accepts and emits records in a fixed flat schema, they don't support dynamic name-value pairs. And these are not document oriented; if you want to pass a DOM of some kind then you serialize it into a field. However, the ETLs are all far more mature than nascent document or XML oriented pipelines.

- ESBs (Enterprise Service Buses) – not listed; various
- One of the XProc implementations (an XML pipeline spec) such as Calabash

Indexing

Generating the Lucene/Solr Index

Hadoop:

- Cascading - Solr "Tap"
- Katta - KattaIntegration

Talend:

- High speed Insert Plugin
- High speed Update Plugin

Monitoring

Tools or services for monitoring Solr-specific performance metrics.

- SPM for Solr
- SolrMeter - stress test tool
- SolrJMeter - stress test tool using JMeter