290N Winter 2015 HW1

Team:

Sachin Rathod - rathod@umail.ucsb.edu
Sahaj Biyani - sahajbiyani@umail.ucsb.edu

Performance:

- **Indexing Time and index size:**

Number of docs	Indexing Time (secs)	Index Size
49723	124.834	221 MB
30886	77.115	87MB
51758	123.245	180MB

Estimation for larger number of documents:

Number of docs	Indexing Time (secs)	Index Size
1 million	2500 s ~ 41 mins	3.37 GB
10 million	25000 s ~ 7 hours	33.7GB
100 million	250000 s ~ 70 hours	337 GB

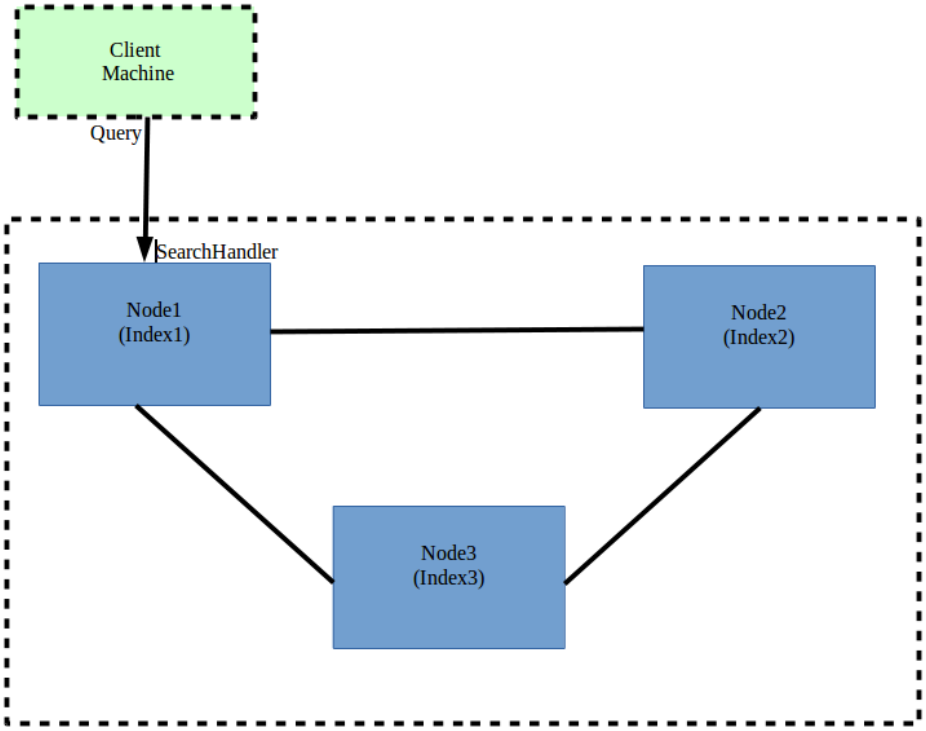
- **Response Times:**

Example Query	Response Time (ms) (On single machine)	Response Time (ms) (distributed search)
Formation	9 (1688 results)	44 (6333 results)
Dark Matter	9 (944 results)	31 (3330 results)
Program AND Systems	20 (1603 results)	68 (5792 results)
Skeletal OR Analysis	87 (9275 results)	214 (24082 results)

Architecture:

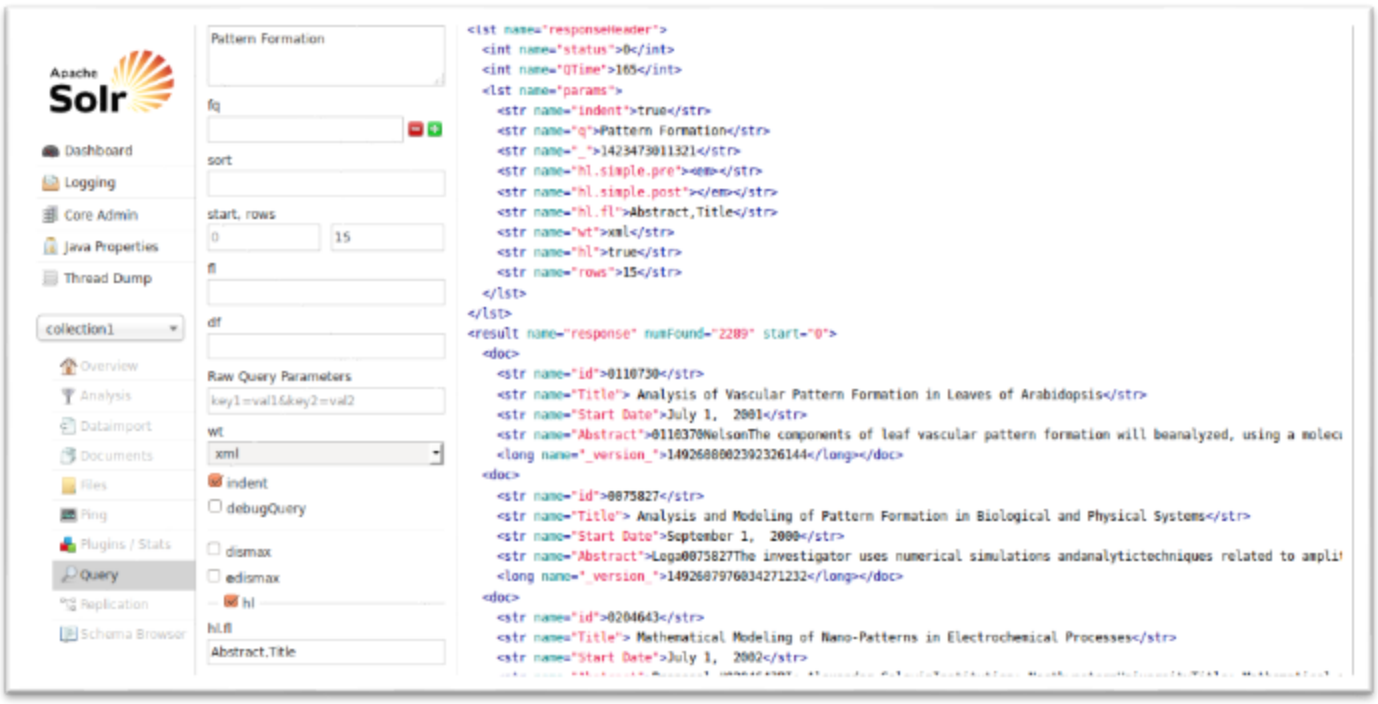
NSF Dataset was parsed to generate xml files with various fields such as Abstract, Title, Start Date etc. Each of the three parts of the dataset are indexed at three different machines running solr instances (on Amazon EC-2 servers).
At each machine, *schema.xml* is modified to include fields from NSF dataset and solrconfig.xml is modified to includes solr instance addresses for the *shards* of the dataset.

Following is the high level architecture:



Example Query: "Pattern Formation"

First 7 results have the exact phrase match in Title ->Good Results!



Lucene Scoring Function:

$score(q,d) = coord(q,d) \cdot queryNorm(q) \cdot$	$\sum_{t \text{ in } q}$	$(tf(t \text{ in } d) \cdot idf(t)^2 \cdot t.getBoost() \cdot norm(t,d))$

tf(t in d) correlates to the term's *frequency*

idf(t) stands for Inverse Document Frequency.

coord(q,d) is a score factor based on how many of the query terms are found in the specified document.

queryNorm(q) is a normalizing factor

t.getBoost() is a search time boost of term *t* in the query *q* as specified in the query text Ex. "Pattern^4 Matching"

norm(t,d) encapsulates a few (indexing time) boost and length factors: *Document boost, Field boost, lengthNorm*