

Semester Project Report

Distributing Computing

Team Members

Shariq Bin Rashid (242431)

Fozan Shahid (256852)

Fahad Fahim (241542)

Project description in detail :

This search engine, fetch and display data of News articles and headlines using Solr for distributive computing concept.

Dump Processing & Parsing:

• News Articles:-

For News Articles we have parsed news data and headline. We have parsed news using SAX parser and extracted title, date, news body out of it.

Below is the sample data.json file.

```
{  
  "category": "ENTERTAINMENT",  
  "headline": "Amy Poehler: A 'Parks And Recreation' Revival Would Be 'Amazing'",  
  "authors": "Curtis M. Wong",  
  "link": "https://www.huffingtonpost.com/entry/amy-poehler-parks-and-recreation-revival_us_5b06e39de4b07c4ea1061910",  
  "short_description": "Co-star Nick Offerman also seemed on board... under one (incredible) condition.",  
  "date": "2018-05-24"  
}
```

There are multiple Pages in the file separated by comma.

Indexing:-

We have configured Schema.xml according to the requirements and data.json and Wiki.xml is then placed in the SOLR_HOME/example/exampledocs folder to proceed with indexing.

We start jetty by giving java -jar start.jar command and post the file to Solr for indexing with command java -jar post.jar *.json.

```

"responseHeader":{
  "zkConnected":true,
  "status":0,
  "QTime":17,
  "params":{
    "q":"*:*",
    "_":"1623226560164"}}},
" " " " " " 1000 " .

```

Solr :

We have indexed around 200853 lines in a two shards which includes 80mb of data.

We also tried setting up two external servers using zookeeper but it requires port forwarding from ISP, so we have used shards on same IP, but it shows same setup environment as shown below which is functionality of Distributive Computing.

Host	Node	CPU	Heap	Disk usage	Requests	Collections	Replicas
192.168.56.1 Windows 10 15.8Gb Java 15 Load: -1 show details...	8983_solr Uptime: 10h 21m show details...	1%	96%	1.5Mb	RPM: 0.06 p95: 41ms	project	project_s1r1 (978 docs) project_s2r4 (858 docs)
	8984_solr Uptime: 10h 17m show details...	1%	22%	1.5Mb	RPM: 0 p95: 54ms	project	project_s1r2 (978 docs) project_s2r6 (858 docs)

Collection: project	Shards
Config name: _default Max shards per node: 2 Replication factor: 2 Auto-add replicas: yes Router name: compositeId	<div> shard1 Range: 80000000-ffffffff Active: Replicas: project_shard1_replica_n1 project_shard1_replica_n2 </div> <div> shard2 Range: 0-7fffffff Active: Replicas: project_shard2_replica_n4 project_shard2_replica_n6 </div>

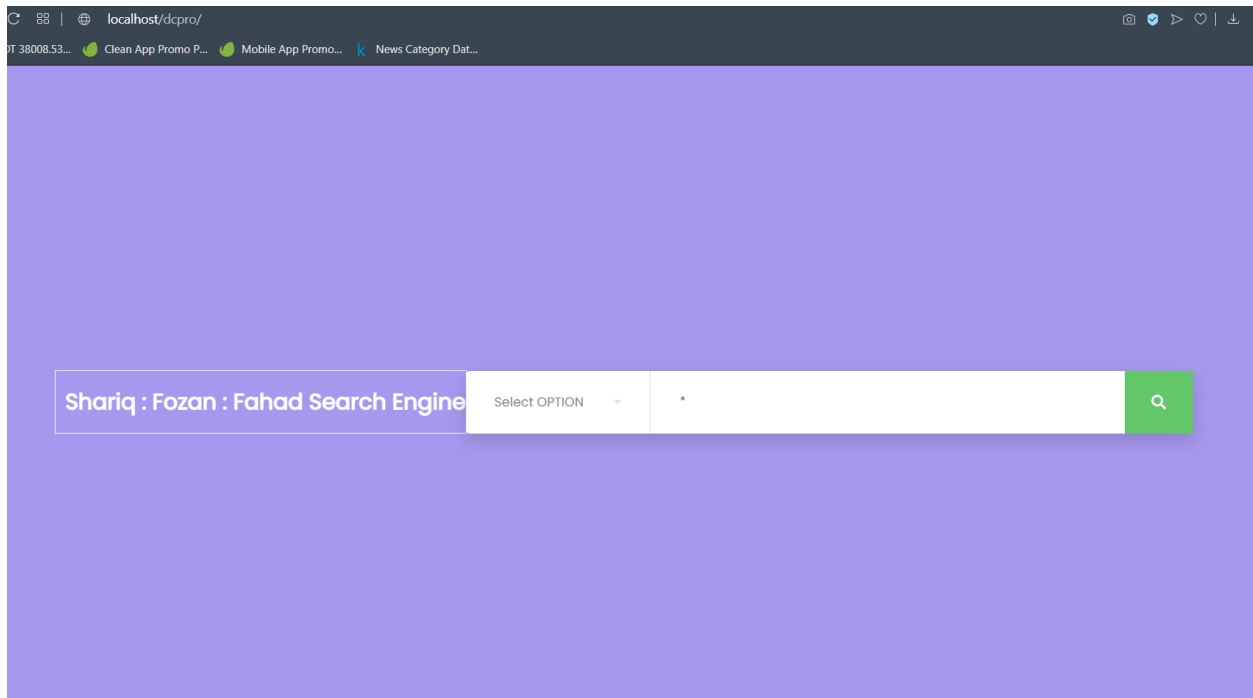
Application :

We have use **XAMP** server to host website on localhost.

Application is built with:

- Html/css
- JavaScript
- PHP

The interface between Solr and main application, navigation from one page to another, fetching Solr data and displaying it is written in javascript and php.



Features:

User can search related to following key:

```
<select data-trigger="" name="option" >
  <option placeholder="Select CATEGORY">Select OPTION</option>
  <option>*</option>
  <option>category</option>
  <option>headline</option>
  <option>authors</option>
  <option>link</option>
  <option>short_description</option>
  <option>date</option>
</select>
</div>
div>
```

And its value on search field than that query will be sent to *search.php* using **Get** method.

Query Processing :

When a user enters a query, the documents are fetched from Solr in json format using following

url :

"http://localhost:8983/solr/search?query

json data returned from Solr is formatted using php function **display**.

```
====Result[1]====
category: CRIME
headline: There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV
authors: Melissa Jeltsen
link: https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89
short_description: She left her husband. He killed their children. Just another day in America.
date: 2018-05-26
id: b226fd36-afc2-4341-a8e0-3844ace1d8df
_version_: 1702072817087938560

====Result[2]====
category: CRIME
headline: 'This Isn't Pakistan, Bitch': Video Captures Driver's Racist Rant
authors: Rowaida Abdelaziz
link: https://www.huffingtonpost.com/entry/this-isnt-pakistan-video-racist_us_5b02db07e4b07309e05ac472
short_description: A case of road rage in Texas quickly escalated into a hateful and xenophobic diatribe.
date: 2018-05-21
id: 5f13e379-778e-4fef-9d9b-370abcb5333a
_version_: 1702073400441176064

====Result[3]====
category: CRIME
headline: These Are The Victims Of The Santa Fe High School Shooting
authors: Sebastian Murdock, Antonia Blumberg, and Jenna Amatulli
link: https://www.huffingtonpost.com/entry/victims-santa-fe-high-school-shooting_us_5aff1969e4b0a046186b658d
short_description: Authorities confirmed the deaths of 10 people in the mass shooting. Here are their stories.
date: 2018-05-19
id: 22be8782-3eb4-4af9-9a4b-9d05fdeb3f12
_version_: 1702073400531353600

====Result[4]====
category: CRIME
headline: Hospice Overdosed Patients To 'Hasten Their Deaths,' Former Health Care Executive Admits
authors: Carol Kuruvilla
link: https://www.huffingtonpost.com/entry/hospice-overdosed-patients-to-hasten-their-deaths-former-health-care-executive-alleges_us_5aff1d4ae4b0a046186b6954
short_description: Federal prosecutors allege Melanie Murphey and others at Novus Health Services in Texas were trying to maximize profits.
date: 2018-05-18
id: 6db7e420-a8b2-433f-89a4-446bbc74f711
_version_: 1702073400657182720
```

GitHub repository of Project:

<https://github.com/Shariqbinrashid/SolrSerachEngine>

Contribution:

Shariq bin rashid: Development

Fozan Shahid: Server setup and testing

Fahad Fahim: Research , optimization and documentation

Conclusion:

This search engine work with concepts of **Distributive** Computing. Indexing data to multiple shards on **SOLR** shows realtime server working environment and then fetching it through json format on search Engine.