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.

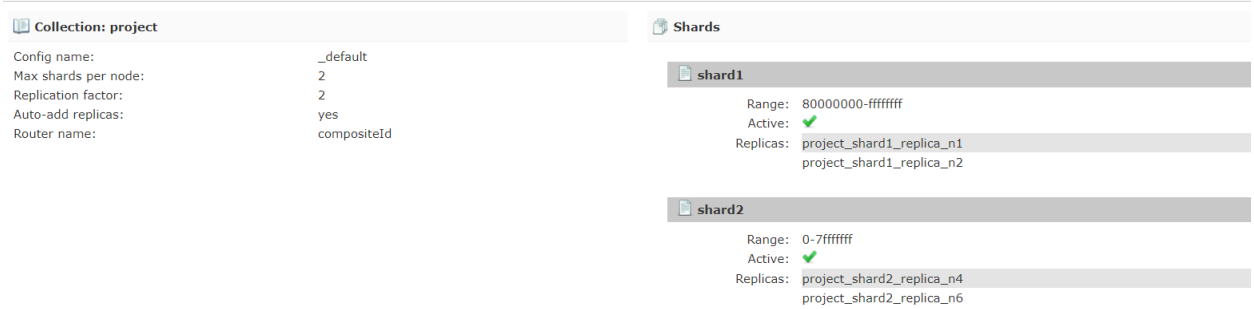
```
"responseHeader":{
  "zkConnected":true,
  "status":0,
  "QTime":17,
  "params":{
    "q":"*:*",
    "_":"1623226560164"}},
```


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)



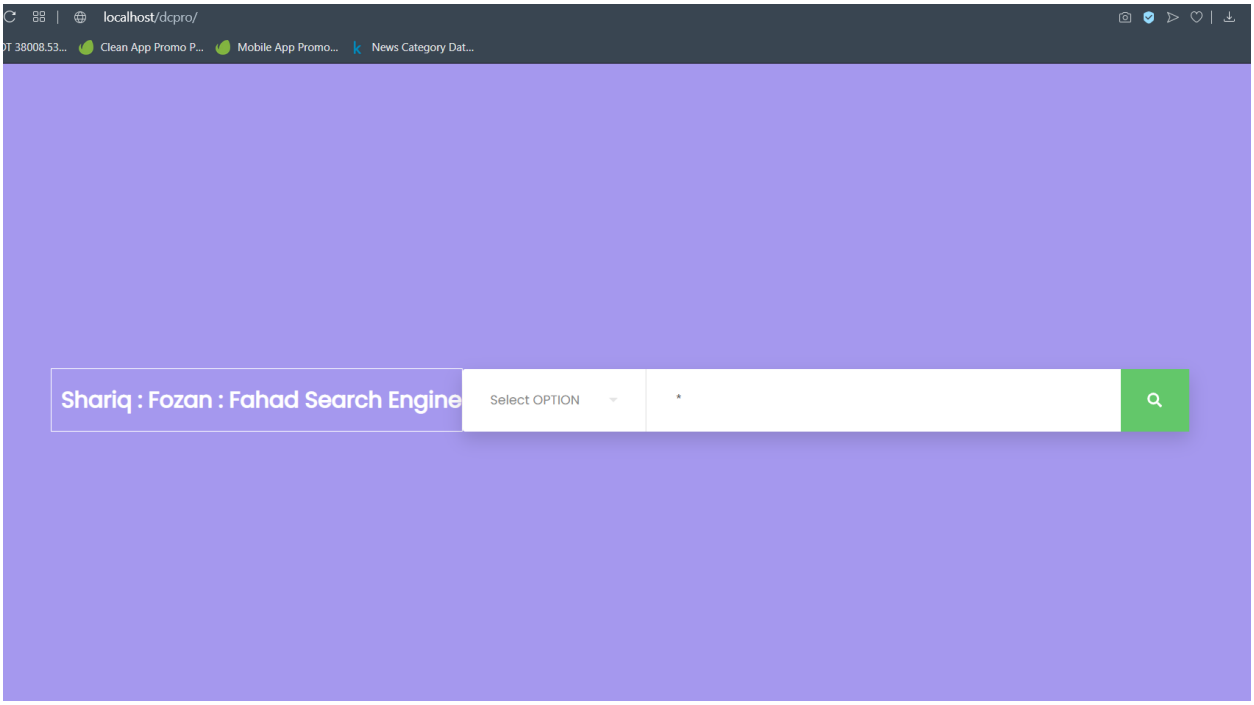
Application :

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="Selct 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>
```

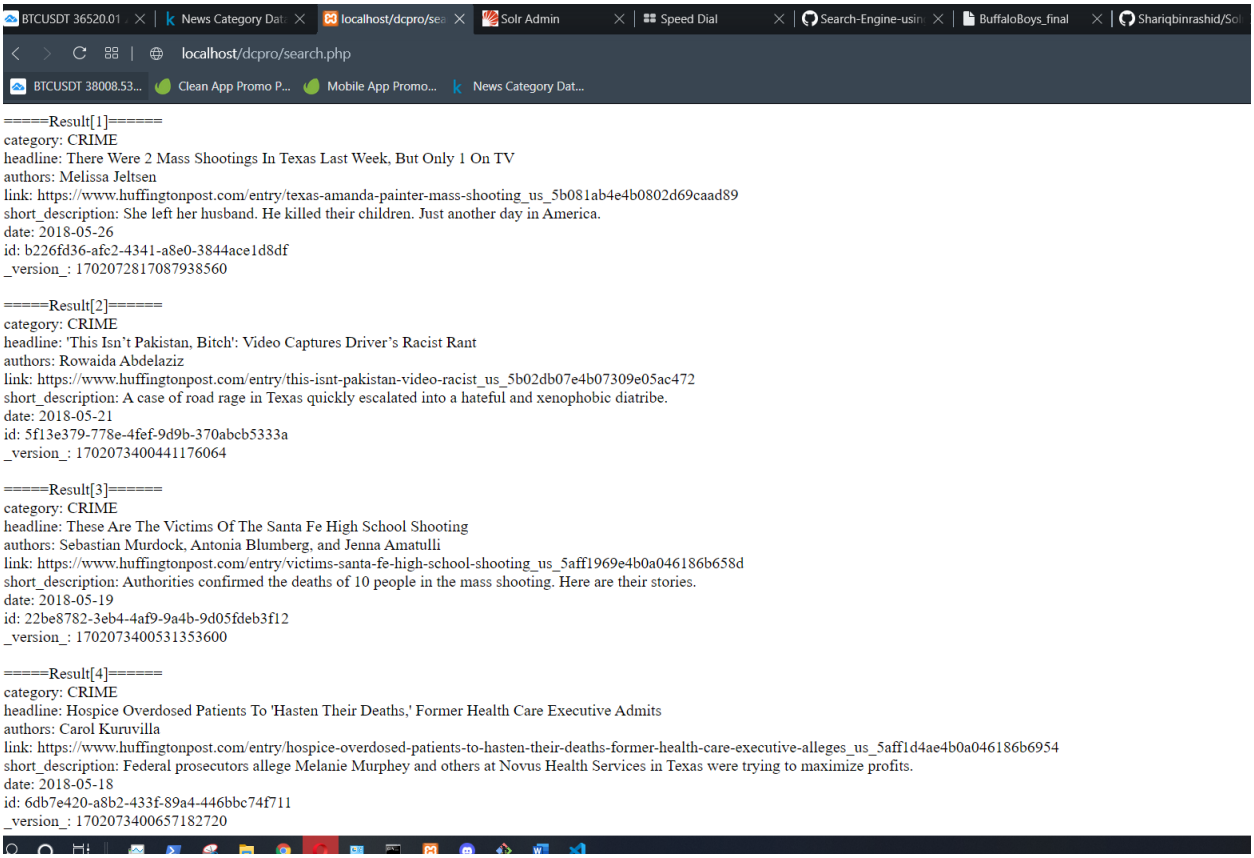
And its value on search field than that query will be sent to ***search.php*** using **Get** method.

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**.



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.