

# WEB CRAWLER using Apache Nutch with Elasticsearch and Kibana :-

**BY: Akhil Sourav ( S20180010007)**

## **POINTS TO NOTE:**

1. Instead of attaching all the un-necessary files from **Apache Nutch, Elastic search and Kibana** . I have only attached the **IMPORTANT** files in the **zip** which I have **configured or changed** .Except these files everything else remains the same. And also whole folder size is too big for uploading.  
..
2. I have performed all the tasks and codes on my Windows Subsystem (UBUNTU) and for each command I have include the Screenshot alongwith it.

Lets begin ,

## **1. Setting up java Environment :**

(1.1) Java runtime(JRE) and development Environment(JDK)

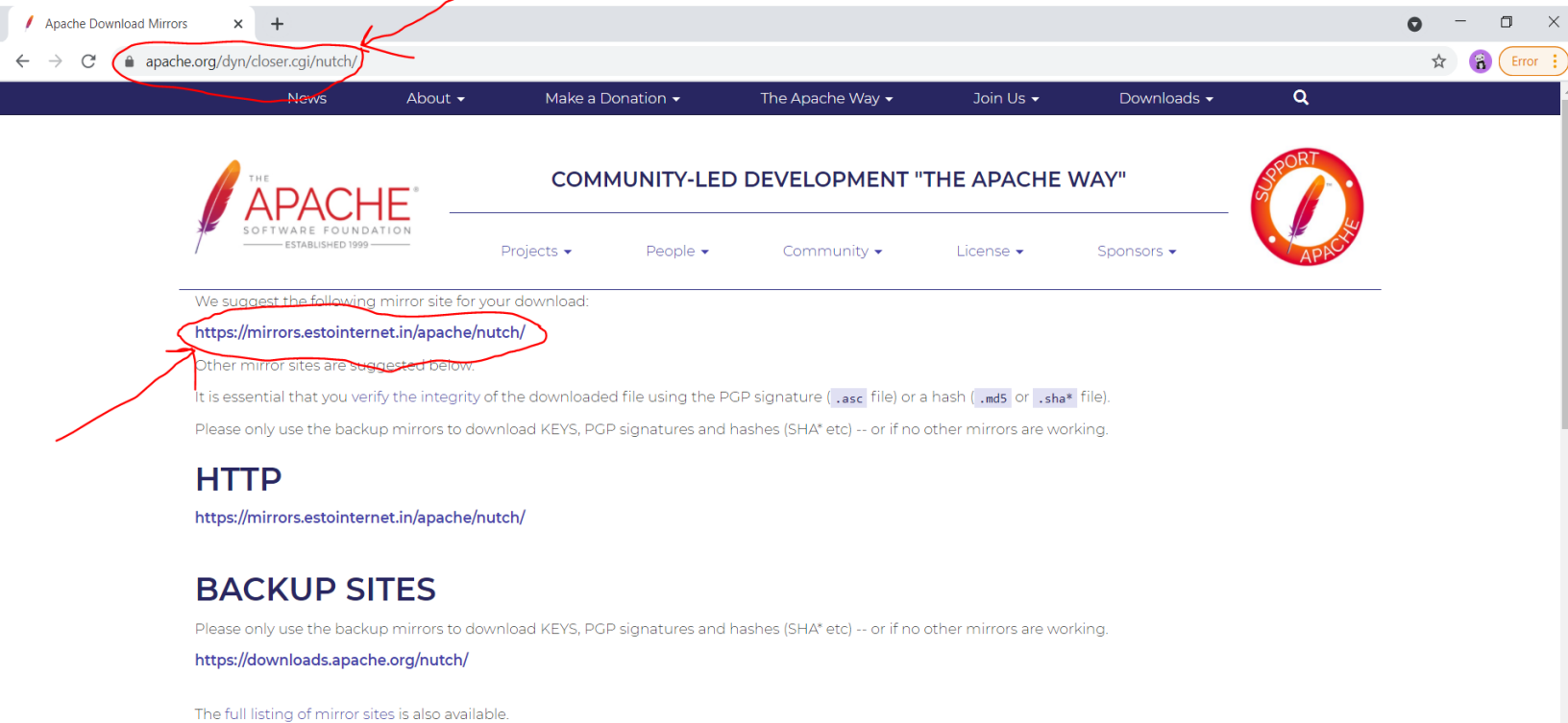
```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ java --version
openjdk 11.0.10 2021-01-19
OpenJDK Runtime Environment (build 11.0.10+9-Ubuntu-0ubuntu1.20.04)
OpenJDK 64-Bit Server VM (build 11.0.10+9-Ubuntu-0ubuntu1.20.04, mixed mode, sharing)
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$
```

(1.2) Setting JAVA\_HOME in the bash file.

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ export JAVA_HOME=$(readlink -f /usr/bin/java | sed "s:bin/java::")
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-amd64/
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$
```

## 2. Setting Up APACHE Nutch :

### 2.1 Go to this Website and download Apache Nutch



The screenshot shows a web browser window with the address bar displaying `apache.org/dyn/closer.cgi/nutch/`. The page header includes navigation links: News, About, Make a Donation, The Apache Way, Join Us, and Downloads. The main content area features the Apache Software Foundation logo and the text "COMMUNITY-LED DEVELOPMENT 'THE APACHE WAY'". Below this, a list of links (Projects, People, Community, License, Sponsors) is visible. A message states: "We suggest the following mirror site for your download:" followed by the URL `https://mirrors.estointernet.in/apache/nutch/`, which is circled in red. Other mirror sites are suggested below, and a note emphasizes verifying the integrity of the downloaded file using PGP signatures or hashes. The page also includes sections for HTTP and BACKUP SITES, with the backup site URL `https://downloads.apache.org/nutch/` highlighted.

We suggest the following mirror site for your download:

<https://mirrors.estointernet.in/apache/nutch/>

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha\* file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA\* etc) -- or if no other mirrors are working.

### HTTP

<https://mirrors.estointernet.in/apache/nutch/>

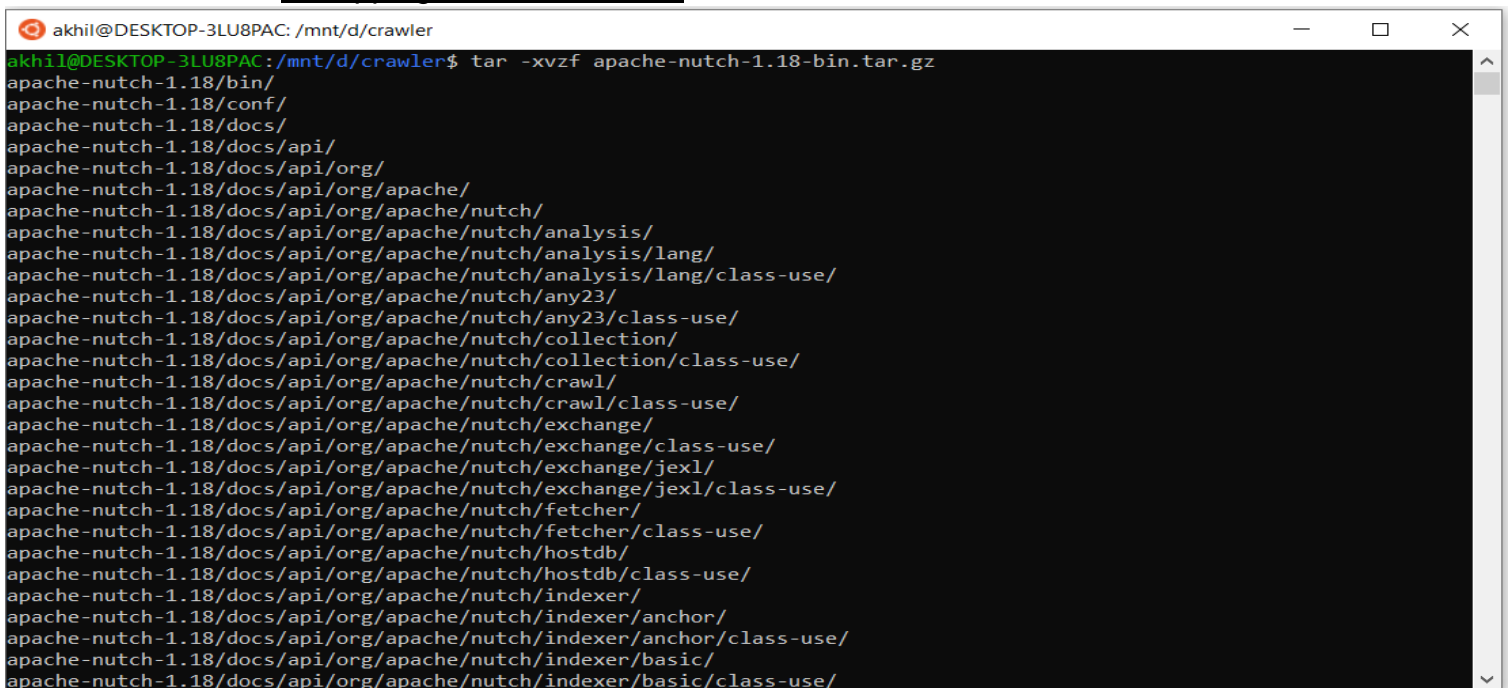
### BACKUP SITES

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA\* etc) -- or if no other mirrors are working.

<https://downloads.apache.org/nutch/>

The full listing of mirror sites is also available.

### 2.2 Unzipping APACHE Nutch :



```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ tar -xvzf apache-nutch-1.18-bin.tar.gz
apache-nutch-1.18/bin/
apache-nutch-1.18/conf/
apache-nutch-1.18/docs/
apache-nutch-1.18/docs/api/
apache-nutch-1.18/docs/api/org/
apache-nutch-1.18/docs/api/org/apache/
apache-nutch-1.18/docs/api/org/apache/nutch/
apache-nutch-1.18/docs/api/org/apache/nutch/analysis/
apache-nutch-1.18/docs/api/org/apache/nutch/analysis/lang/
apache-nutch-1.18/docs/api/org/apache/nutch/analysis/lang/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/any23/
apache-nutch-1.18/docs/api/org/apache/nutch/any23/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/collection/
apache-nutch-1.18/docs/api/org/apache/nutch/collection/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/crawl/
apache-nutch-1.18/docs/api/org/apache/nutch/crawl/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/exchange/
apache-nutch-1.18/docs/api/org/apache/nutch/exchange/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/exchange/jexl/
apache-nutch-1.18/docs/api/org/apache/nutch/exchange/jexl/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/fetcher/
apache-nutch-1.18/docs/api/org/apache/nutch/fetcher/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/hostdb/
apache-nutch-1.18/docs/api/org/apache/nutch/hostdb/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/indexer/
apache-nutch-1.18/docs/api/org/apache/nutch/indexer/anchor/
apache-nutch-1.18/docs/api/org/apache/nutch/indexer/anchor/class-use/
apache-nutch-1.18/docs/api/org/apache/nutch/indexer/basic/
apache-nutch-1.18/docs/api/org/apache/nutch/indexer/basic/class-use/
```

## 2.3 Checking for correctly Installed and running fine

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ cd apache-nutch-1.18/
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch
nutch 1.18
Usage: nutch COMMAND [-Dproperty=value]... [command-specific args]...
where COMMAND is one of:
  readdb          read / dump crawl db
  mergedb         merge crawl db-s, with optional filtering
  readlinkdb      read / dump link db
  inject          inject new urls into the database
  generate         generate new segments to fetch from crawl db
  freegen         generate new segments to fetch from text files
  fetch           fetch a segment's pages
  parse           parse a segment's pages
  readseg         read / dump segment data
  mergesegs       merge several segments, with optional filtering and slicing
  updatedb        update crawl db from segments after fetching
  invertlinks     create a linkdb from parsed segments
  mergelinkdb     merge linkdb-s, with optional filtering
  index           run the plugin-based indexer on parsed segments and linkdb
  dedup           deduplicate entries in the crawl db and give them a special status
  dump            exports crawled data from segments into files
  commoncrawl_dump exports crawled data from segments into common crawl data format encoded as CBOR
  solrindex       run the solr indexer on parsed segments and linkdb - DEPRECATED use the index command instead
  solrdedup       remove duplicates from solr - DEPRECATED use the dedup command instead
  solrclean       remove HTTP 301 and 404 documents from solr - DEPRECATED use the clean command instead
  clean           remove HTTP 301 and 404 documents and duplicates from indexing backends configured via plugins
  parserchecker   check the parser for a given url
  indexchecker    check the indexing filters for a given url
  filterchecker   check url filters for a given url
  normalizerchecker check url normalizers for a given url
  domainstats     calculate domain statistics from crawl db
  protocolstats   calculate protocol status code stats from crawl db
  crawlcomplete   calculate crawl completion stats from crawl db
```

## 2.4 Configuring NUTCH and Crawling URLs:

### (2.4.1) Configuring our **nutch-site.xml** file located in the **apache-nutch-1.18/conf** Directory

D:\crawler\apache-nutch-1.18\conf\nutch-site.xml (DC) - Sublime Text (UNREGISTERED)

File Edit Selection Find View Goto Tools Project Preferences Help

```
nutch-site.xml — apache-nutch-1.18\conf x import java.util.*; provider.py x 01.html x untitled x todo.ejs x schema.xml x solrconfig.xml
1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3
4 <!-- Put site-specific property overrides in this file. -->
5
6 <configuration>
7   <property>
8     <name>http.agent.name</name>
9     <value>MY DC CRAWLER</value>
10   </property>
11   <property>
12     <name>plugin.includes</name>
13     <value>protocol-http|urlfilter-regex|parse-(html|tika)|index-(basic|anchor)|urlnormalizer-(pass|regex|basic)|scoring-opic|indexer-elastic</value>
14   </property>
15   <property>
16     <name>db.ignore.external.links</name>
17     <value>>false</value>
18     <description>If true, outlinks leading from a page to external hosts or domain
19       will be ignored. This is an effective way to limit the crawl to include
20       only initially injected hosts or domains, without creating complex URLFilters.
21       See 'db.ignore.external.links.mode'.
22     </description>
23   </property>
24   <property>
25     <name>elastic.host</name>
26     <value>localhost</value>
27     <description>The hostname to send documents to using TransportClient.
28       Either host and port must be defined or cluster.
29     </description>
30   </property>
31   <property>
32     <name>elastic.port</name>
33     <value>9300</value>
34     <description>
35       The port to connect to using TransportClient.
36     </description>
37   </property>
38   <property>
39     <name>elastic.cluster</name>
40     <value>elasticsearch</value>
41     <description>The cluster name to discover. Either host and port must
42       be defined.
43     </description>
44   </property>
45   <property>
46     <name>elastic.index</name>
47     <value>nutch</value>
48     <description>
49       The name of the elasticsearch index. Will normally be autocreated if it
50       doesn't exist.
51     </description>
52   </property>
53 </configuration>
54
```

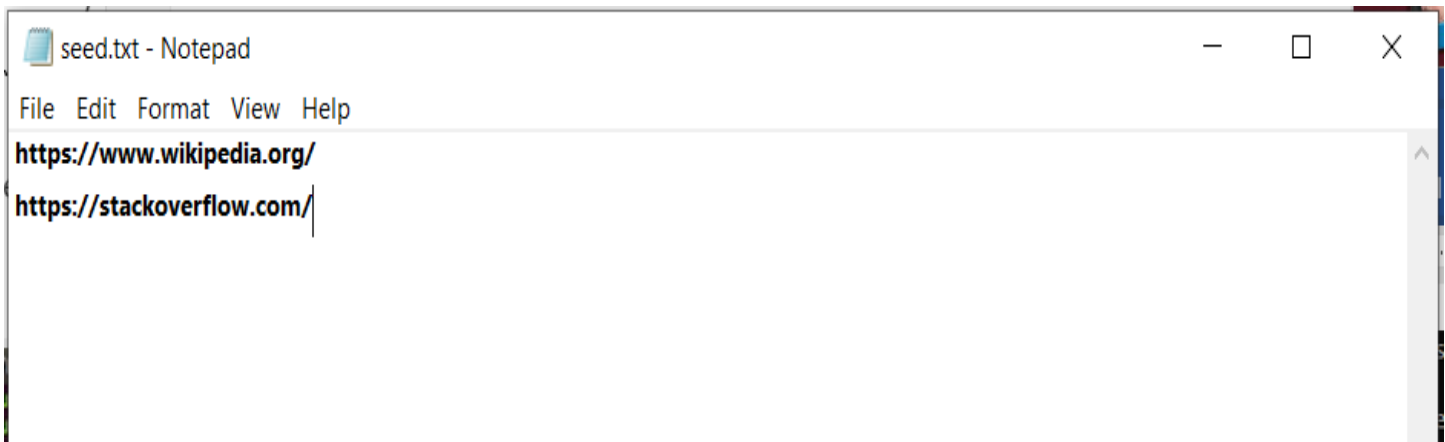
## 2.5 Creating URL seed List to crawl our very-first URLs

- Create a directory called **urls**
- Inside it , create a text file name **seed.txt**

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18/urls

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ mkdir -p urls
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ cd urls
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18/urls$ touch seed.txt
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18/urls$ ls
seed.txt
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18/urls$
```

- I am going to crawl **2 websites** given below.



### 3. Crawling the Websites :

- Injecting seeds into apache **crawldb** of Apache Nutch

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch inject crawl/crawldb urls
Injector: starting at 2021-03-31 20:42:36
Injector: crawlDb: crawl/crawldb
Injector: urlDir: urls
Injector: Converting injected urls to crawl db entries.
Injecting seed URL file:/mnt/d/crawler/apache-nutch-1.18/urls/seed.txt
Injector: overwrite: false
Injector: update: false
Injector: Total urls rejected by filters: 0
Injector: Total urls injected after normalization and filtering: 2
Injector: Total urls injected but already in CrawlDb: 0
Injector: Total new urls injected: 2
Injector: finished at 2021-03-31 20:42:50, elapsed: 00:00:13
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- Generating a list of pages (segment) to be fetched from the seed URLs. (Generation Process)

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch generate crawl/crawldb crawl/segments
Generator: starting at 2021-03-31 20:51:10
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: running in local mode, generating exactly one partition.
Generator: number of items rejected during selection:
Generator: Partitioning selected urls for politeness.
Generator: segment: crawl/segments/20210331205120
Generator: finished at 2021-03-31 20:51:23, elapsed: 00:00:13
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- Create a shell temporary variable name **s1** for easy fetching

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ s1=crawl/segments/20210331205120
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ echo $s1
crawl/segments/20210331205120
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

#### 4. Fetching The Content :

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch fetch $s1
```

Fetcher: starting at 2021-03-31 21:05:02

Fetcher: segment: crawl/segments/20210331205120

Fetcher: threads: 10

Fetcher: time-out divisor: 2

QueueFeeder finished: total 2 records

QueueFeeder queuing status:

2	SUCCESSFULLY_QUEUED
0	ERROR_CREATE_FETCH_ITEM
0	ABOVE_EXCEPTION_THRESHOLD
0	HIT_BY_TIMELIMIT

FetcherThread 44 Using queue mode : byHost

FetcherThread 50 fetching https://stackoverflow.com/ (queue crawl delay=5000ms)

FetcherThread 44 Using queue mode : byHost

FetcherThread 51 fetching https://www.wikipedia.org/ (queue crawl delay=5000ms)

FetcherThread 44 Using queue mode : byHost

FetcherThread 52 has no more work available

FetcherThread 52 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 53 has no more work available

FetcherThread 53 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 54 has no more work available

FetcherThread 54 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 55 has no more work available

FetcherThread 55 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 56 has no more work available

FetcherThread 56 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 57 has no more work available

FetcherThread 57 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

FetcherThread 58 has no more work available

FetcherThread 58 -finishing thread FetcherThread, activeThreads=2

FetcherThread 44 Using queue mode : byHost

Fetcher: throughput threshold: -1

FetcherThread 59 has no more work available

FetcherThread 59 -finishing thread FetcherThread, activeThreads=2

Fetcher: throughput threshold retries: 5

-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2

-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2

-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2

-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2

-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2

FetcherThread 50 has no more work available

FetcherThread 50 -finishing thread FetcherThread, activeThreads=1

FetcherThread 51 has no more work available

FetcherThread 51 -finishing thread FetcherThread, activeThreads=0

Select akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
FetcherThread 59 -finishing thread FetcherThread, activeThreads=2
Fetcher: throughput threshold retries: 5
-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2
-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2
-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2
-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2
-activeThreads=2, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=2
FetcherThread 50 has no more work available
FetcherThread 50 -finishing thread FetcherThread, activeThreads=1
FetcherThread 51 has no more work available
FetcherThread 51 -finishing thread FetcherThread, activeThreads=0
-activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=0
-activeThreads=0
Fetcher: finished at 2021-03-31 21:05:21, elapsed: 00:00:18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

## 5. Parsing :

- Generating Useful Information and more URLs for fetching

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch parse $s1
ParseSegment: starting at 2021-03-31 21:12:18
ParseSegment: segment: crawl/segments/20210331205120
Parsed (915ms): https://stackoverflow.com/
Parsed (59ms): https://www.wikipedia.org/
ParseSegment: finished at 2021-03-31 21:12:28, elapsed: 00:00:10
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- 
- 
- 
- 
- Updating **CrawlDb** with the new Information.



akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch updatedb crawl/crawldb $s1
CrawlDb update: starting at 2021-03-31 21:14:59
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20210331205120]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2021-03-31 21:15:09, elapsed: 00:00:10
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- Repeating the Whole process from crawling 1 more time by taking into account all the new URLs and creating the segment with top 1000 to select best scoring URLs due for fetch.

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch generate crawl/crawldb crawl/segments -topN 1000
Generator: starting at 2021-03-31 21:18:56
Generator: Selecting best-scoring urls due for fetch.
Generator: filtering: true
Generator: normalizing: true
Generator: topN: 1000
Generator: running in local mode, generating exactly one partition.
Generator: number of items rejected during selection:
Generator:      2  SCHEDULE_REJECTED
Generator: Partitioning selected urls for politeness.
Generator: segment: crawl/segments/20210331211906
Generator: finished at 2021-03-31 21:19:08, elapsed: 00:00:11
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ s2=`ls -d crawl/segments/2* | tail -1`
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ echo $s2
crawl/segments/20210331211906
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- Fetching again ...

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch fetch $s2
Fetcher: starting at 2021-03-31 21:23:34
Fetcher: segment: crawl/segments/20210331211906
Fetcher: threads: 10
Fetcher: time-out divisor: 2
QueueFeeder finished: total 140 records
QueueFeeder queuing status:
    140    SUCCESSFULLY_QUEUED
     0     ERROR_CREATE_FETCH_ITEM
     0     ABOVE_EXCEPTION_THRESHOLD
     0     HIT_BY_TIMELIMIT
FetcherThread 44 Using queue mode : byHost
FetcherThread 50 fetching https://be-tarask.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 51 fetching https://gl.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 52 fetching https://gamedev.stackexchange.com/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 53 fetching https://apple.stackexchange.com/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 54 fetching https://eu.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 55 fetching https://tt.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 56 fetching https://nl.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 57 fetching https://sh.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
FetcherThread 58 fetching https://my.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 44 Using queue mode : byHost
Fetcher: throughput threshold: -1
FetcherThread 59 fetching https://ceb.wikipedia.org/ (queue crawl delay=5000ms)
Fetcher: throughput threshold retries: 5
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=130, fetchQueues.getQueueCount=107
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=130, fetchQueues.getQueueCount=107
FetcherThread 51 fetching https://serverfault.com/ (queue crawl delay=5000ms)
FetcherThread 50 fetching https://an.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=128, fetchQueues.getQueueCount=107
FetcherThread 54 fetching https://ba.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=127, fetchQueues.getQueueCount=107
FetcherThread 56 fetching https://pt.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 53 fetching https://ru.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 52 fetching https://zh.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 55 fetching https://hu.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 58 fetching https://la.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 57 fetching https://fy.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=121, fetchQueues.getQueueCount=107
FetcherThread 50 fetching https://es.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 59 fetching https://fo.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=119, fetchQueues.getQueueCount=107
FetcherThread 51 fetching https://askubuntu.com/ (queue crawl delay=5000ms)
FetcherThread 56 fetching https://war.wikipedia.org/ (queue crawl delay=5000ms)
```

🟢 akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
FetcherThread 57 fetching https://unix.stackexchange.com/ (queue crawl delay=5000ms)
FetcherThread 54 fetching https://bcl.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=100, fetchQueues.getQueueCount=107
FetcherThread 52 fetching https://zh-min-nan.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=99, fetchQueues.getQueueCount=107
FetcherThread 50 fetching https://af.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=98, fetchQueues.getQueueCount=107
FetcherThread 55 fetching https://cy.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=97, fetchQueues.getQueueCount=107
FetcherThread 56 fetching https://upload.wikimedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=96, fetchQueues.getQueueCount=107
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=96, fetchQueues.getQueueCount=107
FetcherThread 58 fetching https://als.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 51 fetching https://fi.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 53 fetching https://bg.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=93, fetchQueues.getQueueCount=107
FetcherThread 54 fetching https://dba.stackexchange.com/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=92, fetchQueues.getQueueCount=107
FetcherThread 59 fetching https://quantumcomputing.stackexchange.com/ (queue crawl delay=5000ms)
FetcherThread 50 fetching https://ro.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 55 fetching https://min.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 52 fetching https://he.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 57 fetching https://stackexchange.com/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=87, fetchQueues.getQueueCount=107
FetcherThread 56 fetching https://stackoverflowbusiness.com/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=86, fetchQueues.getQueueCount=107
FetcherThread 53 fetching https://creativecommons.org/licenses/by-sa/3.0/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=85, fetchQueues.getQueueCount=107
FetcherThread 51 fetching https://hy.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 58 fetching https://ta.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=83, fetchQueues.getQueueCount=107
FetcherThread 54 fetching https://ar.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 55 fetching https://ja.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=81, fetchQueues.getQueueCount=107
FetcherThread 50 fetching https://ur.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=80, fetchQueues.getQueueCount=107
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=80, fetchQueues.getQueueCount=107
FetcherThread 59 fetching https://lv.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 51 fetching https://softwareengineering.stackexchange.com/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=78, fetchQueues.getQueueCount=107
FetcherThread 53 fetching https://mk.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 50 fetching https://sv.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=76, fetchQueues.getQueueCount=107
FetcherThread 58 fetching https://ca.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 52 fetching https://cs.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 55 fetching https://et.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 54 fetching https://de.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 59 fetching https://id.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=71, fetchQueues.getQueueCount=107
FetcherThread 51 fetching https://no.wikipedia.org/ (queue crawl delay=5000ms)
FetcherThread 53 fetching https://ms.wikipedia.org/ (queue crawl delay=5000ms)
-activeThreads=10, spinWaiting=0, fetchQueues.totalSize=69, fetchQueues.getQueueCount=107
```

[illegible]

```
--activeThreads=10, spinWaiting=10, fetchQueues.totalSize=2, fetchQueues.getQueueCount=1
FetcherThread 50 fetching https://stackoverflow.com/teams (queue crawl delay=5000ms)
--activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=9, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
--activeThreads=10, spinWaiting=10, fetchQueues.totalSize=1, fetchQueues.getQueueCount=1
FetcherThread 50 fetching https://stackoverflow.com/teams/create/free (queue crawl delay=5000ms)
--activeThreads=10, spinWaiting=9, fetchQueues.totalSize=0, fetchQueues.getQueueCount=1
FetcherThread 52 has no more work available
FetcherThread 52 -finishing thread FetcherThread, activeThreads=9
FetcherThread 58 has no more work available
FetcherThread 58 -finishing thread FetcherThread, activeThreads=8
FetcherThread 55 has no more work available
FetcherThread 55 -finishing thread FetcherThread, activeThreads=7
FetcherThread 57 has no more work available
FetcherThread 57 -finishing thread FetcherThread, activeThreads=6
FetcherThread 59 has no more work available
FetcherThread 59 -finishing thread FetcherThread, activeThreads=5
FetcherThread 53 has no more work available
FetcherThread 53 -finishing thread FetcherThread, activeThreads=4
FetcherThread 54 has no more work available
FetcherThread 54 -finishing thread FetcherThread, activeThreads=3
FetcherThread 56 has no more work available
FetcherThread 56 -finishing thread FetcherThread, activeThreads=2
FetcherThread 51 has no more work available
FetcherThread 51 -finishing thread FetcherThread, activeThreads=1
--activeThreads=1, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=1
FetcherThread 50 has no more work available
FetcherThread 50 -finishing thread FetcherThread, activeThreads=0
--activeThreads=0, spinWaiting=0, fetchQueues.totalSize=0, fetchQueues.getQueueCount=0
--activeThreads=0
Fetcher: finished at 2021-03-31 21:27:43, elapsed: 00:04:08
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```



- Parsing again

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch parse $s2
ParseSegment: starting at 2021-03-31 21:37:51
ParseSegment: segment: crawl/segments/20210331211906
Parsed (840ms): https://ai.stackexchange.com/
Parsed (126ms): https://apple.stackexchange.com/
Parsed (81ms): https://askubuntu.com/
Parsed (51ms): https://creativecommons.org/licenses/by-sa/3.0/
Parsed (122ms): https://dba.stackexchange.com/
Parsed (89ms): https://gamedev.stackexchange.com/
Parsed (80ms): https://meta.stackoverflow.com/
Parsed (54ms): https://networkengineering.stackexchange.com/
Parsed (78ms): https://quantumcomputing.stackexchange.com/
Parsed (64ms): https://salesforce.stackexchange.com/
Parsed (48ms): https://serverfault.com/
Parsed (51ms): https://softwareengineering.stackexchange.com/
Parsed (59ms): https://stackexchange.com/
Parsed (246ms): https://stackexchange.com/sites
Mar 31, 2021 9:38:04 PM org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
WARNING: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

Mar 31, 2021 9:38:04 PM org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
WARNING: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Error parsing: https://stackoverflow.blog/: failed(2,0): Can't retrieve Tika parser for mime-type application/o
Parsed (4070ms): https://stackoverflow.blog/
Parsed (58ms): https://stackoverflow.com/advertising
Parsed (45ms): https://stackoverflow.com/company
Parsed (30ms): https://stackoverflow.com/company/contact
Parsed (50ms): https://stackoverflow.com/company/press
Parsed (38ms): https://stackoverflow.com/company/work-here
Parsed (35ms): https://stackoverflow.com/enterprise
Parsed (33ms): https://stackoverflow.com/enterprise/get-started
Parsed (618ms): https://stackoverflow.com/feeds
Parsed (37ms): https://stackoverflow.com/help
Parsed (51ms): https://stackoverflow.com/jobs
Parsed (24ms): https://stackoverflow.com/jobs/directory/developer-jobs
Parsed (29ms): https://stackoverflow.com/jobs/salary
Parsed (25ms): https://stackoverflow.com/legal/cookie-policy
Parsed (35ms): https://stackoverflow.com/legal/privacy-policy
Parsed (32ms): https://stackoverflow.com/legal/terms-of-service
Parsed (34ms): https://stackoverflow.com/opensearch.xml
Parsed (57ms): https://stackoverflow.com/questions
Parsed (54ms): https://stackoverflow.com/tags
Parsed (49ms): https://stackoverflow.com/teams
Parsed (22ms): https://stackoverflow.com/teams/calculate/attract-and-retain-talent
Parsed (7ms): https://stackoverflow.com/teams/create/basic
Parsed (8ms): https://stackoverflow.com/teams/create/business
Parsed (7ms): https://stackoverflow.com/teams/create/free
Parsed (19ms): https://stackoverflow.com/teams/integrations/github
```

```

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
Parsed (19ms): https://stackoverflow.com/teams/integrations/github
Parsed (21ms): https://stackoverflow.com/teams/integrations/jira
Parsed (19ms): https://stackoverflow.com/teams/integrations/microsoft-teams
Parsed (20ms): https://stackoverflow.com/teams/integrations/okta
Parsed (24ms): https://stackoverflow.com/teams/integrations/slack
Parsed (58ms): https://stackoverflow.com/teams/tour
Parsed (48ms): https://superuser.com/
Parsed (45ms): https://unix.stackexchange.com/
ParseSegment: finished at 2021-03-31 21:38:09, elapsed: 00:00:18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$

```

- Update the database again

```

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch updatedb crawl/crawldb $s2
CrawlDb update: starting at 2021-03-31 21:40:21
CrawlDb update: db: crawl/crawldb
CrawlDb update: segments: [crawl/segments/20210331211906]
CrawlDb update: additions allowed: true
CrawlDb update: URL normalizing: false
CrawlDb update: URL filtering: false
CrawlDb update: 404 purging: false
CrawlDb update: Merging segment data into db.
CrawlDb update: finished at 2021-03-31 21:40:31, elapsed: 00:00:10
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$

```

- CHECK THE unique segments created in crawl directory.

The screenshot shows a Windows File Explorer window with the address bar set to `This PC > New Volume (D:) > crawler > apache-nutch-1.18 > crawl > segments`. The ribbon includes tabs for File, Home, Share, and View. The Home tab is active, showing various file management icons. The file list displays two folders:

Name	Date modified	Type	Size
20210331205120	3/31/2021 8:51 PM	File folder	
20210331211906	3/31/2021 9:38 PM	File folder	

A red arrow points to the folder `20210331211906`.

6. Updating the database **Linkdb** (Contains the list of known links to each URL, including both source URL and anchor text.
- If we use something like **Apache SOLR**, it can index incoming anchor text with the pages.

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
```

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch invertlinks crawl/linkdb -dir crawl/segments
LinkDb: starting at 2021-03-31 21:49:29
LinkDb: linkdb: crawl/linkdb
LinkDb: URL normalize: true
LinkDb: URL filter: true
LinkDb: internal links will be ignored.
LinkDb: adding segment: file:/mnt/d/crawler/apache-nutch-1.18/crawl/segments/20210331205120
LinkDb: adding segment: file:/mnt/d/crawler/apache-nutch-1.18/crawl/segments/20210331211906
LinkDb: finished at 2021-03-31 21:49:36, elapsed: 00:00:07
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

- Now , lets Index all the contents we just crawled with Nutch into Elastic search .( For this Elasticsearch service must be running)

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/apache-nutch-1.18
```

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$ bin/nutch index crawl/crawldb/ -linkdb crawl/linkdb/ $s2 -filter -normalize -deleteGone
Segment dir is complete: crawl/segments/20210331211906.
Indexer: starting at 2021-03-31 23:28:59
Indexer: deleting gone documents: true
Indexer: URL filtering: true
Indexer: URL normalizing: true
No exchange was configured. The documents will be routed to all index writers.
ERROR StatusLogger Log4j2 could not find a logging implementation. Please add log4j-core to the classpath. Using SimpleLogger to log to the console...
Active IndexWriters :
ElasticIndexWriter:
```

host	Comma-separated list of hostnames	localhost
port	The port to connect to elastic server.	9200
index	Default index to send documents to.	nutch
username	Username for auth credentials	elastic
password	Password for auth credentials	
max.bulk.docs	Maximum size of the bulk in number of documents.	250
max.bulk.size	Maximum size of the bulk in bytes.	2500500
exponential.backoff.millis	Initial delay for the BulkProcessor exponential backoff policy.	100
exponential.backoff.retries	Number of times the BulkProcessor exponential backoff policy should retry bulk operations.	10
bulk.close.timeout	Number of seconds allowed for the BulkProcessor to complete its last operation.	600

```
Indexer: number of documents indexed, deleted, or skipped:
Indexer:      1 deleted (gone)
Indexer:     93 deleted (redirects)
Indexer:     45 indexed (add/update)
Indexer: finished at 2021-03-31 23:29:18, elapsed: 00:00:19
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/apache-nutch-1.18$
```

## 7. Indexing with Elasticsearch and searching with Kibana

- Download and extract Elasticsearch from the official site
- Let's test whether elastic search is running on **port 9200** or not.

akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/elasticsearch-7.4.2

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/elasticsearch-7.4.2$ cd ..
```

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ cd elasticsearch-7.4.2/
```

```
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/elasticsearch-7.4.2$ bin/elasticsearch
```

OpenJDK 64-Bit Server VM warning: Option UseConcMarkSweepGC was deprecated in version 9.0 and will likely be removed in a future release

[2021-03-31T21:56:25,744][INFO ][o.e.e.NodeEnvironment ] [DESKTOP-3LU8PAC] using [1] data paths, mounts [[/mnt/d (D:\)]], n

[2021-03-31T21:56:25,784][INFO ][o.e.e.NodeEnvironment ] [DESKTOP-3LU8PAC] heap size [990.7mb], compressed ordinary object

[2021-03-31T21:56:25,821][INFO ][o.e.n.Node ] [DESKTOP-3LU8PAC] node name [DESKTOP-3LU8PAC], node ID [UPLHdDAQX

[2021-03-31T21:56:25,824][INFO ][o.e.n.Node ] [DESKTOP-3LU8PAC] version[7.4.2], pid[18417], build[default/tar/2f

8-microsoft-standard/amd64], JVM[Ubuntu/OpenJDK 64-Bit Server VM/11.0.10/11.0.10+9-Ubuntu-0ubuntu1.20.04]

[2021-03-31T21:56:25,826][INFO ][o.e.n.Node ] [DESKTOP-3LU8PAC] JVM home [/usr/lib/jvm/java-11-openjdk-amd64]

[2021-03-31T21:56:25,829][INFO ][o.e.n.Node ] [DESKTOP-3LU8PAC] JVM arguments [-Xms1g, -Xmx1g, -XX:+UseConcMarkS

es.networkaddress.cache.ttl=60, -Des.networkaddress.cache.negative.ttl=10, -XX:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=tr

ty.noUnsafe=true, -Dio.netty.noKeySetOptimization=true, -Dio.netty.recycler.maxCapacityPerThread=0, -Dio.netty allocator.numDi

.tmpdir=/tmp/elasticsearch-6774551292795198609, -XX:+HeapDumpOnOutOfMemoryError, -XX:HeapDumpPath=data, -XX:ErrorFile=logs/hs\_

ecount=32,filesize=64m, -Djava.locale.providers=COMPAT, -Dio.netty.allocator.type=unpooled, -XX:MaxDirectMemorySize=536870912,

sticsearch-7.4.2/config, -Des.distribution.flavor=default, -Des.distribution.type=tar, -Des.bundled\_jdk=true]

[2021-03-31T21:56:51,976][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [aggs-matrix-stats]

[2021-03-31T21:56:51,978][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [analysis-common]

[2021-03-31T21:56:51,980][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [data-frame]

[2021-03-31T21:56:51,981][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [flattened]

[2021-03-31T21:56:51,983][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [frozen-indices]

[2021-03-31T21:56:51,985][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [ingest-common]

[2021-03-31T21:56:51,987][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [ingest-geoip]

[2021-03-31T21:56:51,988][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [ingest-user-agent]

[2021-03-31T21:56:51,990][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [lang-expression]

[2021-03-31T21:56:51,992][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [lang-mustache]

[2021-03-31T21:56:51,995][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [lang-painless]

[2021-03-31T21:56:51,996][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [mapper-extras]

[2021-03-31T21:56:51,998][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [parent-join]

[2021-03-31T21:56:52,000][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [percolator]

[2021-03-31T21:56:52,002][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [rank-eval]

[2021-03-31T21:56:52,003][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [reindex]

[2021-03-31T21:56:52,005][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [repository-url]

[2021-03-31T21:56:52,007][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [search-business-rules]

[2021-03-31T21:56:52,009][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [spatial]

[2021-03-31T21:56:52,011][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [transport-netty4]

[2021-03-31T21:56:52,013][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [vectors]

[2021-03-31T21:56:52,015][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-analytics]

[2021-03-31T21:56:52,016][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-ccr]

[2021-03-31T21:56:52,018][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-core]

[2021-03-31T21:56:52,020][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-deprecation]

[2021-03-31T21:56:52,021][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-graph]

[2021-03-31T21:56:52,023][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-ilm]

[2021-03-31T21:56:52,025][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-logstash]

[2021-03-31T21:56:52,027][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-ml]

[2021-03-31T21:56:52,029][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-monitoring]

[2021-03-31T21:56:52,036][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-rollup]

[2021-03-31T21:56:52,038][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-security]

[2021-03-31T21:56:52,039][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-sql]

[2021-03-31T21:56:52,041][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-voting-only-node]

[2021-03-31T21:56:52,043][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] loaded module [x-pack-watcher]

[2021-03-31T21:56:52,045][INFO ][o.e.p.PluginsService ] [DESKTOP-3LU8PAC] no plugins loaded

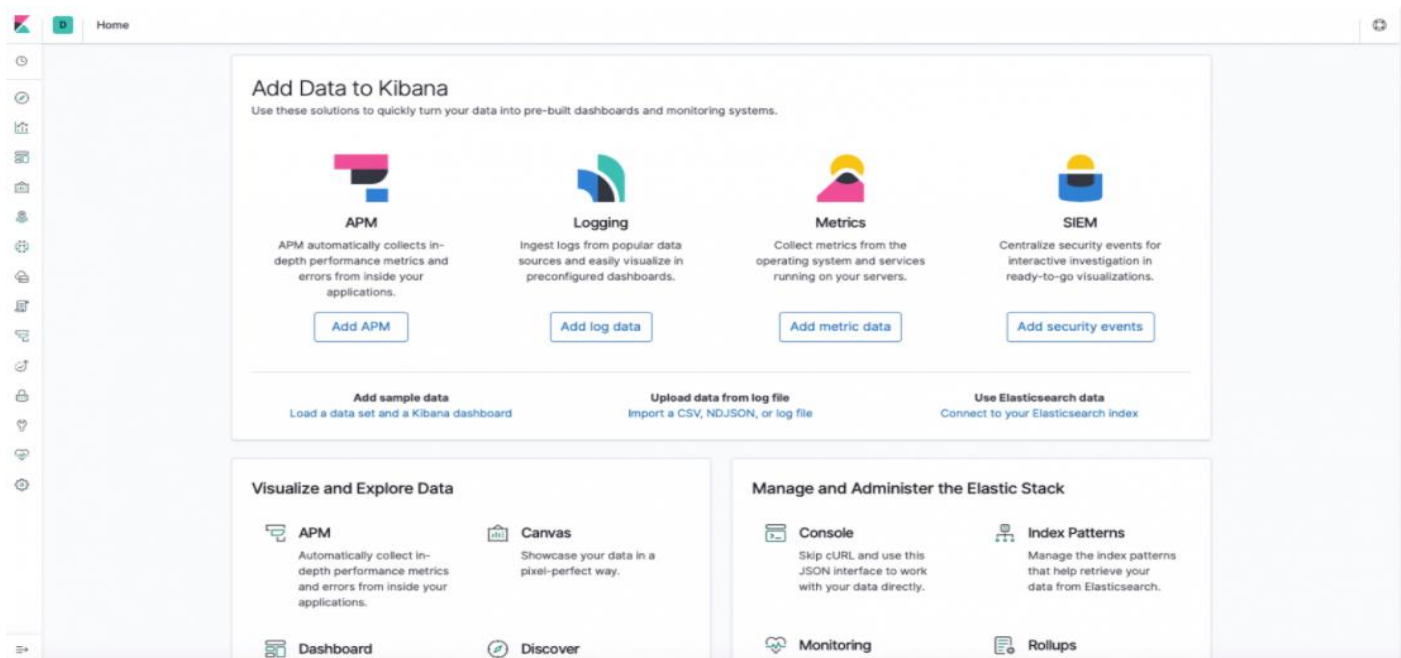


```
{
  "name" : "DESKTOP-3LU8PAC",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "0Acla6pNR8KisurSrPcmCQ",
  "version" : {
    "number" : "7.4.2",
    "build_flavor" : "default",
    "build_type" : "tar",
    "build_hash" : "2f90bbf7b93631e52bafb59b3b049cb44ec25e96",
    "build_date" : "2019-10-28T20:40:44.881551Z",
    "build_snapshot" : false,
    "lucene_version" : "8.2.0",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

## Checking Kibana Installation

```
akhil@DESKTOP-3LU8PAC: /mnt/d/crawler/kibana-6.8.15-linux-x86_64
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler$ cd kibana-6.8.15-linux-x86_64/
akhil@DESKTOP-3LU8PAC:/mnt/d/crawler/kibana-6.8.15-linux-x86_64$ bin/kibana
```

- Go to <http://localhost:5601/>



- Select **Index Management** under Elasticsearch for our newly created **Nutch Index**.

The screenshot shows the Kibana Index Management page. The left sidebar contains the 'Elasticsearch' section with 'Index Management' selected, and the 'Kibana' section with 'Index Patterns' selected. The main content area is titled 'Index Management' and has two tabs: 'Indices' (active) and 'Index Templates'. Below the tabs, there's a search bar and a 'Reload indices' button. A table lists the indices, with one index named 'nutch' shown. The table has columns for Name, Health, Status, Primaries, Replicas, Docs count, and Storage size. The 'nutch' index has a health of 'yellow', status of 'open', 1 primary, 1 replica, 1751 docs, and a storage size of 13.6mb. At the bottom, it says 'Rows per page: 10'.

Name	Health	Status	Primaries	Replicas	Docs count	Storage size
nutch	yellow	open	1	1	1751	13.6mb

- Creating a **New Index** pattern **nutch\***

The screenshot shows the 'Create index pattern' page in Kibana. The title is 'Create index pattern'. Below the title, it says 'Kibana uses index patterns to retrieve data from Elasticsearch indices for things like visualizations.' There's a toggle switch for 'Include system indices' which is turned off. The main content area is titled 'Step 1 of 2: Define index pattern'. Under the heading 'Index pattern', there's a text input field containing 'nutch\*'. Below the input field, there's a message: 'You can use a \* as a wildcard in your index pattern. You can't use spaces or the characters \, /, ?, ", <, >, |.' To the right of this message is a 'Next step' button. Below the message, there's a green checkmark and the text 'Success! Your index pattern matches 1 index.' At the bottom, there's a text input field containing 'nutch' and a 'Rows per page: 10' dropdown menu.

nutch\*

★

↺

🗑

Time Filter field name: tstamp

This page lists every field in the **nutch\*** index and the field's associated core type as recorded by Elasticsearch. To change a field type, use the Elasticsearch [Mapping API](#)

Fields (24)

Scripted fields (0)

Source filters (0)

🔍 Filter

All field types ▾

Name	Type	Format	Searchable	Aggregatable	Excluded
_id	string		●	●	<div>✎</div>
_index	string		●	●	<div>✎</div>
_score	number				<div>✎</div>
_source	_source				<div>✎</div>
_type	string		●	●	<div>✎</div>
boost	string		●		<div>✎</div>
boost.keyword	string		●	●	<div>✎</div>
content	string		●		<div>✎</div>
content.keyword	string		●	●	<div>✎</div>
digest	string		●		<div>✎</div>

Rows per page: 10 ▾

<

1

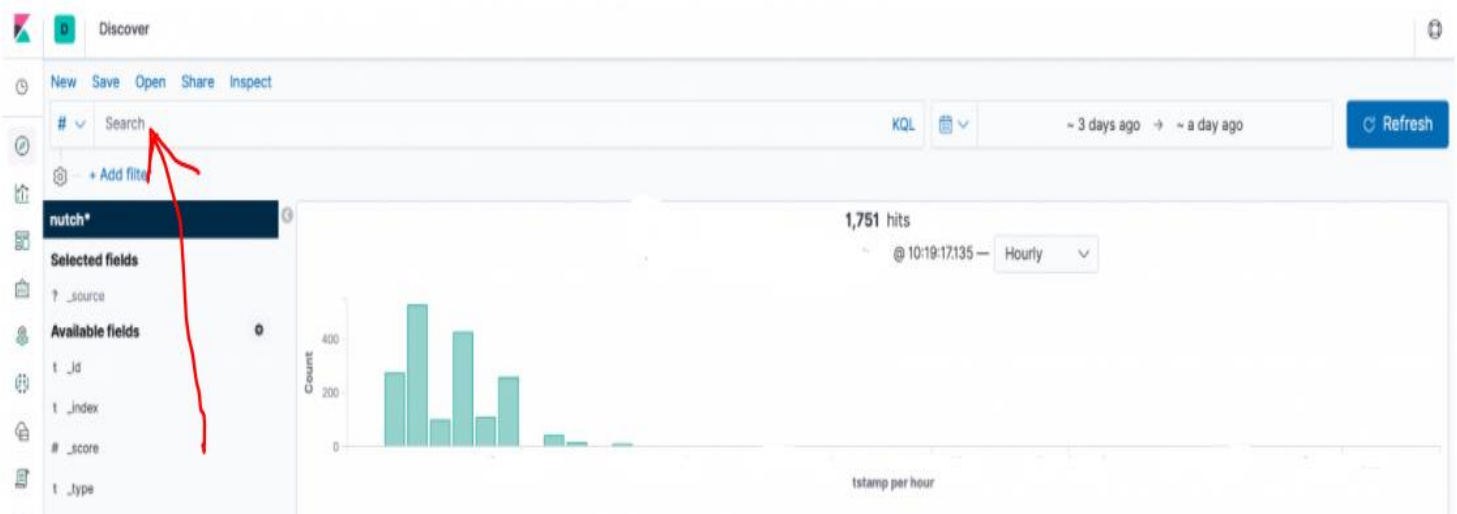
2

3

>

- Click on **Discover Compass icon** to search the documents indexed in Elasticsearch

Now, We can enter our “**queries or questions**” in the search bar to get the desired result



--- THANK YOU 😊 ---