PROJECT REPORT – GROUP C

INDEXING WIKI DUMPS ON CLOUD USING ELK STACK & SOLR

RACHIT CHOKSI
RAJENDRA JADI
SAI ESHWAR
VENKATESH UMAMAHESWARAN

TABLE OF CONTENTS:

PROJECT DESCRIPTION:

The project agenda was loading Wiki Datasets and Indexing them on following Cloud Platforms using Elasticsearch , Logstash and Kibana Stack & SOLR

- Amazon Web Services
- Google Cloud Platform
- Microsoft Azure

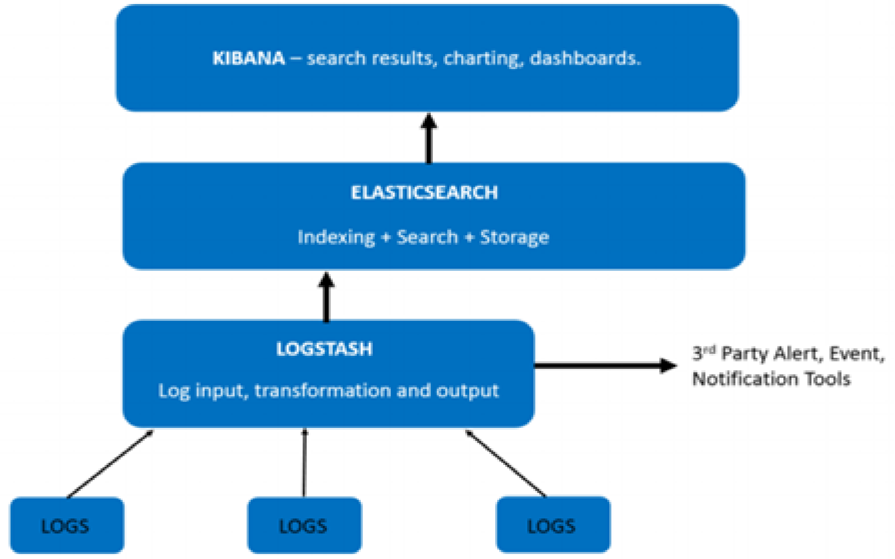
INDEXING:

Indexing is a way to optimize performance of a database by minimizing the number of disk accesses required when a query is processed. An index or database index is a data structure which is used to quickly locate and access the data in a database table.

INDEXING WIKI DUMPS On GOOGLE CLOUD PLATFORM using ELK Stack:

ELASTICSEARCH:

Elastic search is an open source, broadly-distributable, readily-scalable, enterprise-grade search engine based on Lucene and released under the terms of the Apache License. It is Java-based and designed to operate in real time. It can search and index document files in diverse formats. It was designed to be used in distributed environments by providing flexibility and scalability. Now, elastic search is the most popular enterprise search engine followed by Apache Solr, also based on Lucene.


KIBANA – search results, charting, dashboards.
ELASTICSEARCH
Indexing + Search + Storage
LOGSTASH
Log input, transformation and output
3rd Party Alert, Event, Notification Tools
LOGS
LOGS
LOGS

ADVANTAGES OF ELASTIC SEARCH OVER SOLR

Build on top of lucene

Elastic search is built on top of Lucene, which is a full-featured information retrieval library, so it provides the most powerful full-text search capabilities of any open source product.

Document- oriented

Elastic search is document-oriented. It stores real world complex entities as structured JSON documents and indexes all fields by default, with a higher performance result.

Speed

Elastic search is able to execute complex queries extremely fast. It also caches almost all of the structured queries commonly used as a filter for the result set and executes them only once. For every other request which contains a cached filter, it checks the result from the cache. This saves the time parsing and executing the query improving the speed.

Structured search

Elastic Search is schema free, it accepts JSON documents, as well as tries to detect the data structure, index the data, and make it searchable.

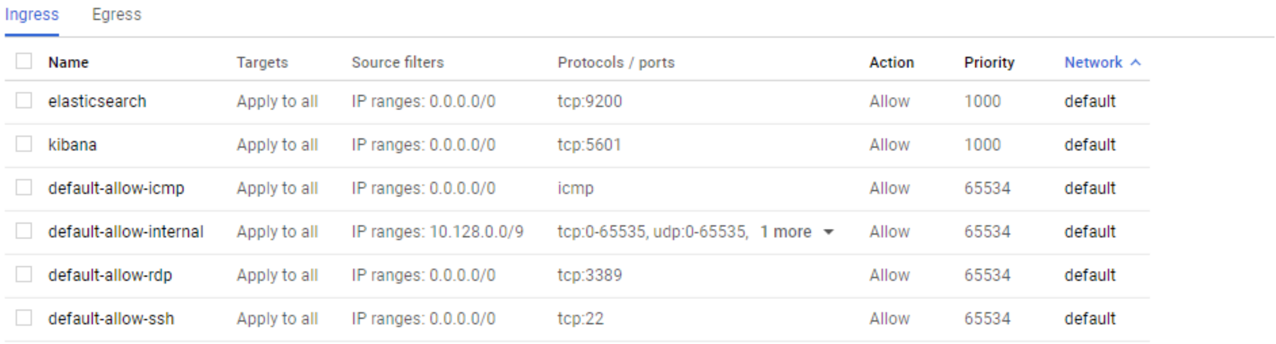
Data record

Elastics earch records any changes made in transactions logs on multiple nodes in the cluster to minimize the chance of data loss.

INSTALLATION OF ELASTICSEARCH:

Firewall Creation:

Creating Firewall rules for access to ports 9200 and 5601 for Elastic search and Kibana.



□	Name	Targets	Source filters	Protocols / ports	Action	Priority	Network ^
□	elasticsearch	Apply to all	IP ranges: 0.0.0.0/0	tcp:9200	Allow	1000	default
□	kibana	Apply to all	IP ranges: 0.0.0.0/0	tcp:5601	Allow	1000	default
□	default-allow-icmp	Apply to all	IP ranges: 0.0.0.0/0	icmp	Allow	65534	default
□	default-allow-internal	Apply to all	IP ranges: 10.128.0.0/9	tcp:0-65535, udp:0-65535, 1 more ▾	Allow	65534	default
□	default-allow-rdp	Apply to all	IP ranges: 0.0.0.0/0	tcp:3389	Allow	65534	default
□	default-allow-ssh	Apply to all	IP ranges: 0.0.0.0/0	tcp:22	Allow	65534	default

To install Java

$ sudo apt-get install default-jre

This will fetch the latest ElasticSearch Version for us

$ wget -qO - https://packages.elastic.co/GPG-KEY-elasticsearch | sudo apt-key add -

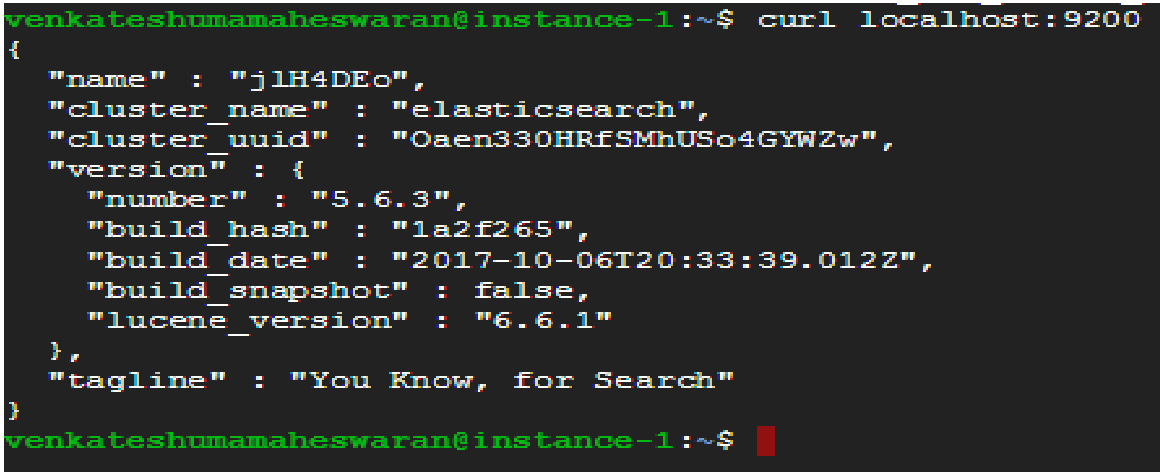
This will complete the installation

$ sudo apt-get install elasticsearch

Find the line referring to the network.host portion. It will be commented out. Uncomment the file and make it read network.host "0.0.0.0"

$ sudo vi /etc/elasticsearch/elasticsearch.yml

ELASTIC SEARCH is installed and running successfully.



```
venkateshumamaheswaran@instance-1:~$ curl localhost:9200
{
  "name" : "jlH4DEo",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "Oaen330HRfSMhUSo4GYWZw",
  "version" : {
    "number" : "5.6.3",
    "build_hash" : "1a2f265",
    "build_date" : "2017-10-06T20:33:39.012Z",
    "build_snapshot" : false,
    "lucene_version" : "6.6.1"
  },
  "tagline" : "You Know, for Search"
}
venkateshumamaheswaran@instance-1:~$
```

KIBANA

Kibana is an open source analytics and visualization platform designed to work with Elasticsearch. You use Kibana to search, view, and interact with data stored in Elasticsearch indices. You can easily perform advanced data analysis and visualize your data in a variety of charts, tables, and maps.

Kibana makes it easy to understand large volumes of data. Its simple, browser-based interface enables you to quickly create and share dynamic dashboards that display changes to Elasticsearch queries in real time.

The different views in Kibana are as follows:

The **discover** view is used to view a list of documents and search for specific documents.

The **visualize** view is used to create visualizations like graphs from the data.

We can add those visualization to **dashboards** to have an overview of you data at a glance.

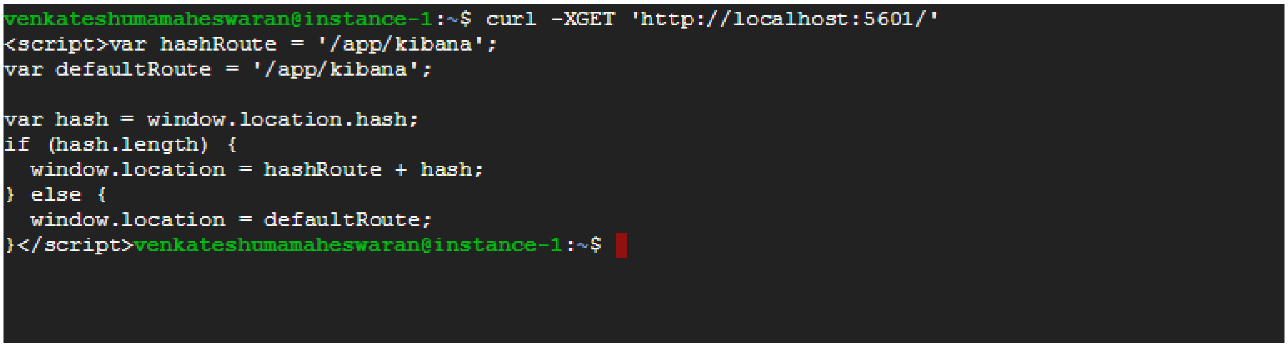
Timelion was formerly a plugin and is now build in. It's used to make advanced timeseries analysis.

The **management** tab are the settings of Kibana where we can add index patterns and tune some advanced settings.
The **Dev Tools** currently only contain the so called Console, which was formerly known as the Sense plugin in Elasticsearch. We can use it to send JSON directly to Elasticsearch and more meant for developers or advanced users.

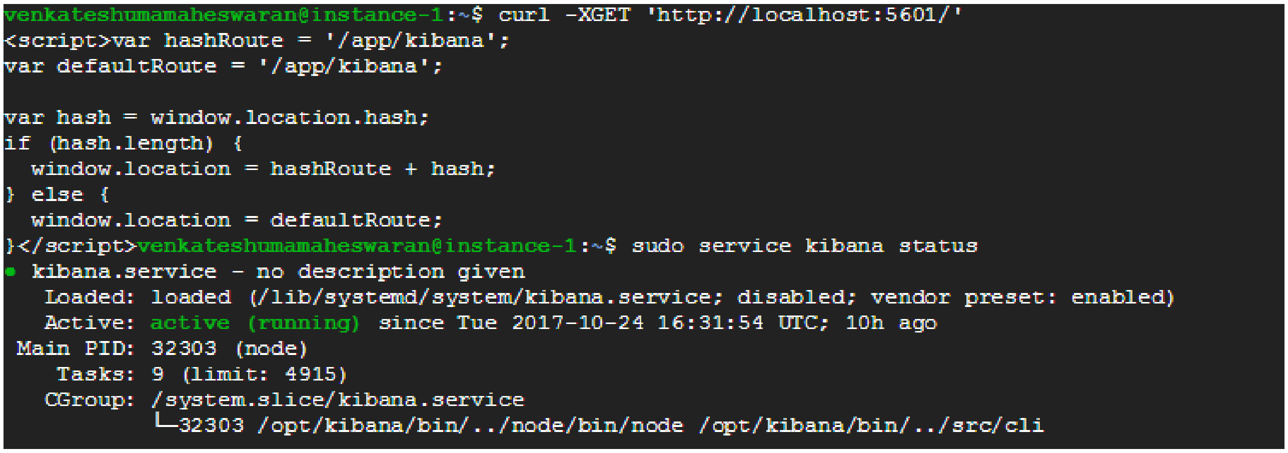
INSTALLATION OF KIBANA

This will establish the source for Kibana
$ echo "deb http://packages.elastic.co/kibana/5.3/debian stable main" | sudo tee -a /etc/apt/sources.list.d/kibana-5.3.x.list

Setting up for Kibana installation
$ sudo apt-get update
$ sudo apt-get install kibana

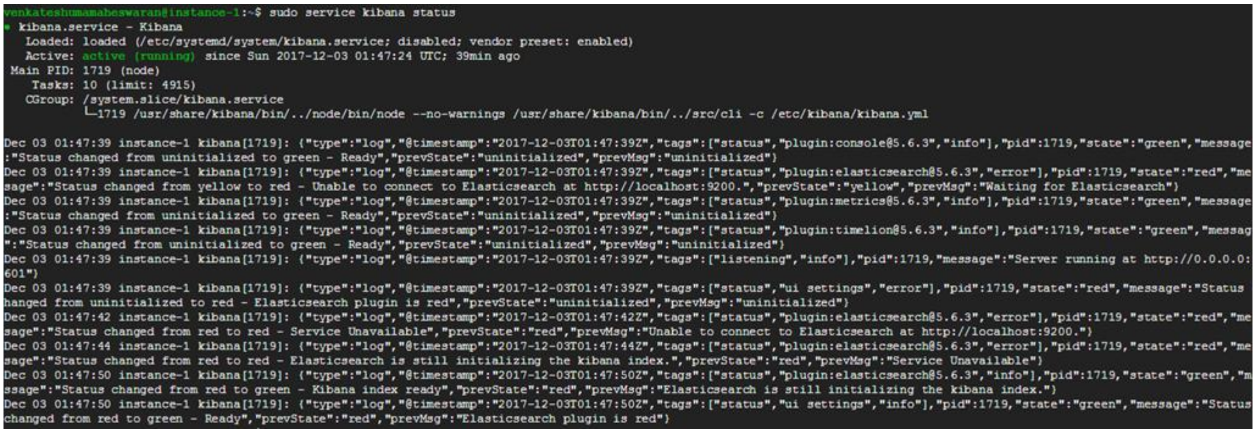


```
venkateshumamaheswaran@instance-1:~$ curl -XGET 'http://localhost:5601/'
<script>var hashRoute = '/app/kibana';
var defaultRoute = '/app/kibana';

var hash = window.location.hash;
if (hash.length) {
  window.location = hashRoute + hash;
} else {
  window.location = defaultRoute;
}</script>venkateshumamaheswaran@instance-1:~$
```



```
venkateshumamaheswaran@instance-1:~$ curl -XGET 'http://localhost:5601/'
<script>var hashRoute = '/app/kibana';
var defaultRoute = '/app/kibana';

var hash = window.location.hash;
if (hash.length) {
  window.location = hashRoute + hash;
} else {
  window.location = defaultRoute;
}</script>venkateshumamaheswaran@instance-1:~$ sudo service kibana status
● kibana.service - no description given
   Loaded: loaded (/lib/systemd/system/kibana.service; disabled; vendor preset: enabled)
   Active: active (running) since Tue 2017-10-24 16:31:54 UTC; 10h ago
 Main PID: 32303 (node)
    Tasks: 9 (limit: 4915)
   CGroup: /system.slice/kibana.service
           └─32303 /opt/kibana/bin/../node/bin/node /opt/kibana/bin/../src/cli
```

$ sudo service kibana start



```
venkateshumamaheswaran@instance-1:~$ sudo service kibana status
● kibana.service - Kibana
   Loaded: loaded (/etc/systemd/system/kibana.service; disabled; vendor preset: enabled)
   Active: active (running) since Sun 2017-12-03 01:47:24 UTC; 39min ago
 Main PID: 1719 (node)
    Tasks: 10 (limit: 4915)
   CGroup: /system.slice/kibana.service
           └─1719 /usr/share/kibana/bin/../node/bin/node --no-warnings /usr/share/kibana/bin/../src/cli -c /etc/kibana/kibana.yml

Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["status","plugin:console@5.6.3","info"],"pid":1719,"state":"green","message
:"Status changed from uninitialized to green - Ready","prevState":"uninitialized","prevMsg":"uninitialized"}
Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["status","plugin:elasticsearch@5.6.3","error"],"pid":1719,"state":"red","me
sage":"Status changed from yellow to red - Unable to connect to Elasticsearch at http://localhost:9200.","prevState":"yellow","prevMsg":"Waiting for Elasticsearch"}
Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["status","plugin:metrics@5.6.3","info"],"pid":1719,"state":"green","message
:"Status changed from uninitialized to green - Ready","prevState":"uninitialized","prevMsg":"uninitialized"}
Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["status","plugin:timelion@5.6.3","info"],"pid":1719,"state":"green","messag
":"Status changed from uninitialized to green - Ready","prevState":"uninitialized","prevMsg":"uninitialized"}
Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["listening","info"],"pid":1719,"message":"Server running at http://0.0.0.0:
601"}
Dec 03 01:47:39 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:39Z","tags":["status","ui settings","error"],"pid":1719,"state":"red","message":"Status
hanged from uninitialized to red - Elasticsearch plugin is red","prevState":"uninitialized","prevMsg":"uninitialized"}
Dec 03 01:47:42 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:42Z","tags":["status","plugin:elasticsearch@5.6.3","error"],"pid":1719,"state":"red","me
sage":"Status changed from red to red - Service Unavailable","prevState":"red","prevMsg":"Unable to connect to Elasticsearch at http://localhost:9200."}
Dec 03 01:47:44 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:44Z","tags":["status","plugin:elasticsearch@5.6.3","error"],"pid":1719,"state":"red","me
sage":"Status changed from red to red - Elasticsearch is still initializing the kibana index.","prevState":"red","prevMsg":"Service Unavailable"}
Dec 03 01:47:50 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:50Z","tags":["status","plugin:elasticsearch@5.6.3","info"],"pid":1719,"state":"green","m
ssage":"Status changed from red to green - Kibana index ready","prevState":"red","prevMsg":"Elasticsearch is still initializing the kibana index."}
Dec 03 01:47:50 instance-1 kibana[1719]: {"type":"log","@timestamp":"2017-12-03T01:47:50Z","tags":["status","ui settings","info"],"pid":1719,"state":"green","message":"Status
changed from red to green - Ready","prevState":"red","prevMsg":"Elasticsearch plugin is red"}
```

LOGSTASH:

Logstash is a tool for managing events and logs.The purpose of Logstash is to get events from any number of inputs (could be from a file, a queue, another Logstash instance, etc), apply filters (parse, modify, or perform any number of processing tasks), and finally output to any number of destinations.

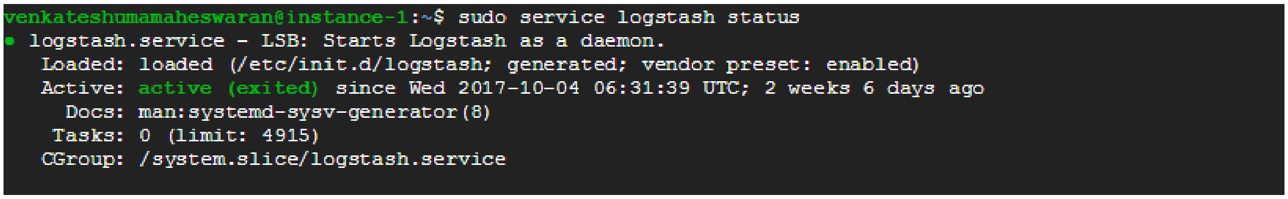
INSTALLATION OF LOGSTASH:

This setups installs for logstash in your system
$ sudo apt-get install apt-transport-https

This will establish the source for Logstash
$ echo "deb https://artifacts.elastic.co/packages/5.x/apt stable main" | sudo tee -a /etc/apt/sources.list.d/elastic-5.x.list

Setting up for Logstash installation
$ sudo apt-get update
$ sudo apt-get install logstash

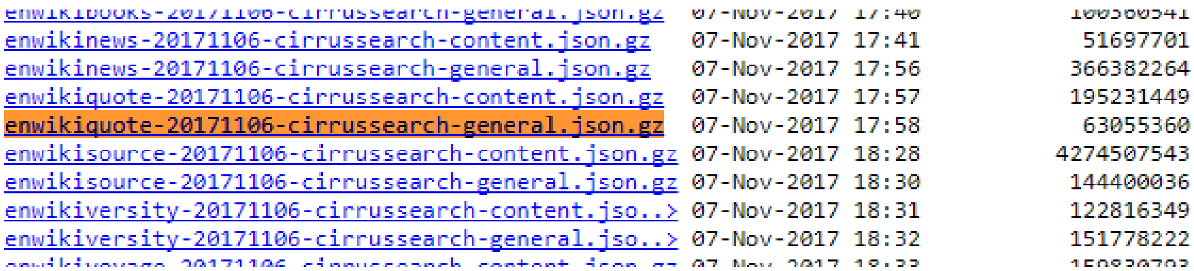
Start the logstash service so we can start shipping logs
$ sudo service logstash start



```
venkateshumamaheswaran@instance-1:~$ sudo service logstash status
● logstash.service - LSB: Starts Logstash as a daemon.
   Loaded: loaded (/etc/init.d/logstash; generated; vendor preset: enabled)
   Active: active (exited) since Wed 2017-10-04 06:31:39 UTC; 2 weeks 6 days ago
     Docs: man:systemd-sysv-generator(8)
    Tasks: 0 (limit: 4915)
   CGroup: /system.slice/logstash.service
```

SIMPLE WIKIPEDIA DATASET:

Elasticsearch only supports JSON documents. We have chosen simplewiki document to index into our Elasticsearch instance.
This is the link for Simple Wikipedia data we are using:
https://dumps.wikimedia.org/other/cirrussearch/20171106/enwikiquote-20171106-cirrussearch-general.json.gz



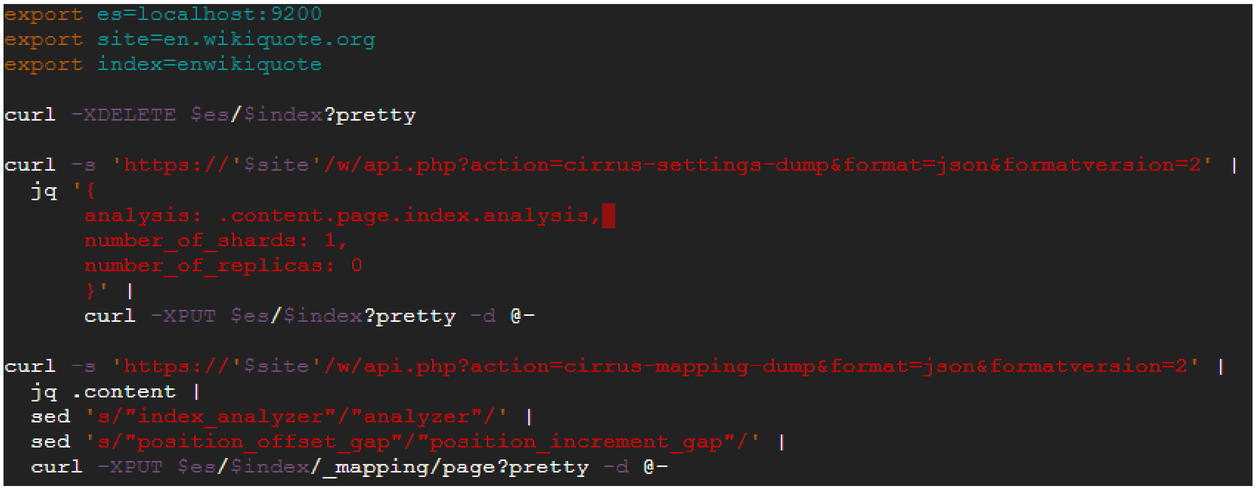
```
enwikibooks-20171106-cirrussearch-general.json.gz  07-Nov-2017 17:40            100360541
enwikinews-20171106-cirrussearch-content.json.gz   07-Nov-2017 17:41             51697701
enwikinews-20171106-cirrussearch-general.json.gz   07-Nov-2017 17:56            366382264
enwikiquote-20171106-cirrussearch-content.json.gz  07-Nov-2017 17:57            195231449
enwikiquote-20171106-cirrussearch-general.json.gz  07-Nov-2017 17:58             63055360
enwikisource-20171106-cirrussearch-content.json.gz 07-Nov-2017 18:28           4274507543
enwikisource-20171106-cirrussearch-general.json.gz 07-Nov-2017 18:30            144400036
enwikiversity-20171106-cirrussearch-content.jso..> 07-Nov-2017 18:31            122816349
enwikiversity-20171106-cirrussearch-general.jso..> 07-Nov-2017 18:32            151778222
enwikivoyage-20171106-cirrussearch-content.json.gz 07-Nov-2017 18:33            [illegible]
```

LOADING WIKI DATA AND INDEXING:

Step 1: Download a wiki dump
Step 2: Get the index ready
Step 3: Prepare the wiki for loading
Step 4: Load the wiki

INDEXING WIKI QUOTE DATA:

We need analysis-icu plugin for Elasticsearch to handle it the index.
bin/plugin install analysis-icu
Then we need jq for some of the json-foo we do next.
sudo apt-get install jq
Then we have to create 3 vim files createindex.sh, chunker.sh, uploader.sh
CREATEINDEX.SH

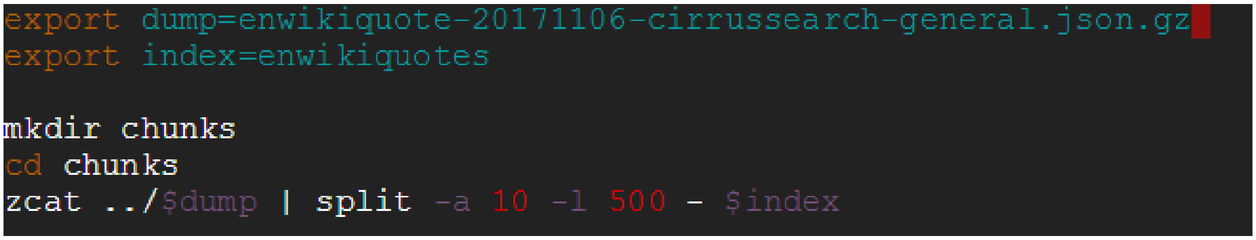


```
export es=localhost:9200
export site=en.wikiquote.org
export index=enwikiquote

curl -XDELETE $es/$index?pretty

curl -s 'https://'$site'/w/api.php?action=cirrus-settings-dump&format=json&formatversion=2' |
  jq '{
      analysis: .content.page.index.analysis,
      number_of_shards: 1,
      number_of_replicas: 0
      }' |
      curl -XPUT $es/$index?pretty -d @-

curl -s 'https://'$site'/w/api.php?action=cirrus-mapping-dump&format=json&formatversion=2' |
  jq .content |
  sed 's/"index_analyzer"/"analyzer"/' |
  sed 's/"position_offset_gap"/"position_increment_gap"/' |
  curl -XPUT $es/$index/_mapping/page?pretty -d @-
```

CODE EXPLANATION FOR CREATEINDEX.SH:

export es=localhost:9200 sets up $es to be Elasticsearch's address.
export site=en.wikiquote.org sets up $site to be the hostname of the MediaWiki instance that you want to use.
export index=enwikiquote just sets $index to the name of the index you'll be loading.
curl -XDELETE $es/$index?pretty deletes the index if it already exists.

CHUNKER.SH

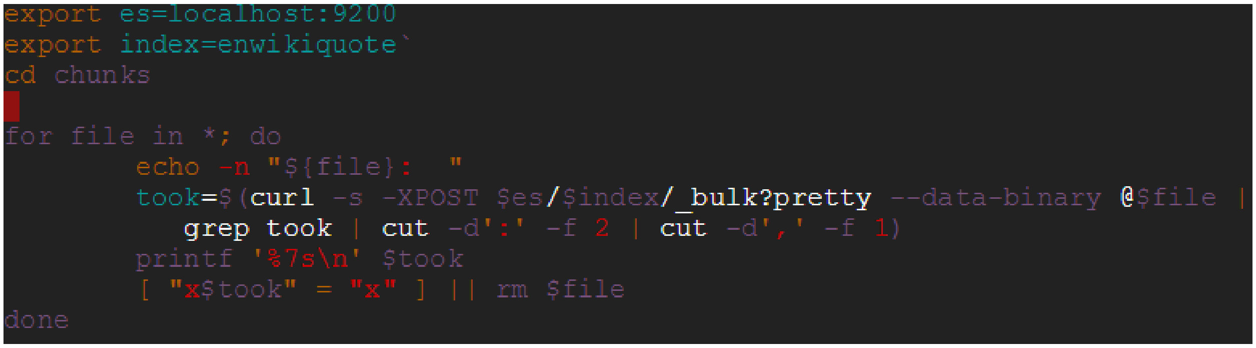


```
export dump=enwikiquote-20171106-cirrussearch-general.json.gz
export index=enwikiquotes

mkdir chunks
cd chunks
zcat ../$dump | split -a 10 -l 500 - $index
```

CODE EXPLANATION FOR CHUNKER.SH:

The first export line just names the file that you downloaded.
The mkdir and cd lines make a directory to hold the files.
The last line cuts the file into 500 line chunks. 250 of those lines are metadata lines for the _bulk api. 250 lines are the actual documents.

UPLOADER.SH:



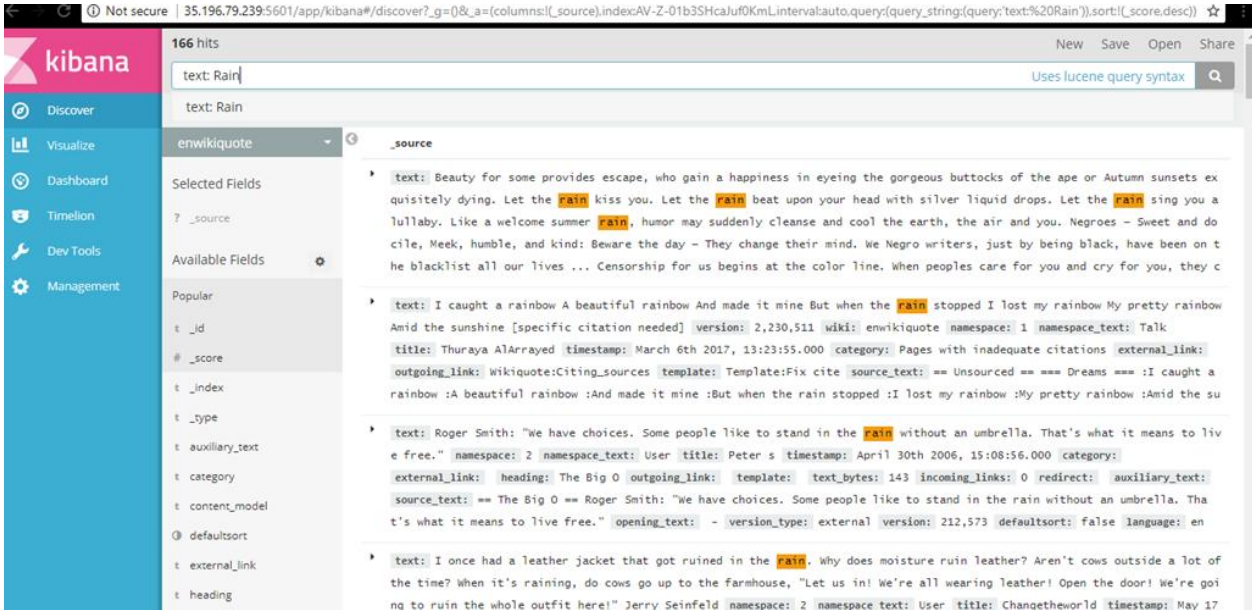
```
export es=localhost:9200
export index=enwikiquote`
cd chunks

for file in *; do
        echo -n "${file}:  "
        took=$(curl -s -XPOST $es/$index/_bulk?pretty --data-binary @$file |
           grep took | cut -d':' -f 2 | cut -d',' -f 1)
        printf '%7s\n' $took
        [ "x$took" = "x" ] || rm $file
done
```

CODE EXPLANATION FOR UPLOADER.SH:

The first three lines should be familiar from above. The loop loads each file and deletes it after it's loaded.
If the file fails to load it isn't deleted and the loop moves on to the next file.

KIBANA VISUALIZATION:

Here are some of the example kibana visualizations we have come up with

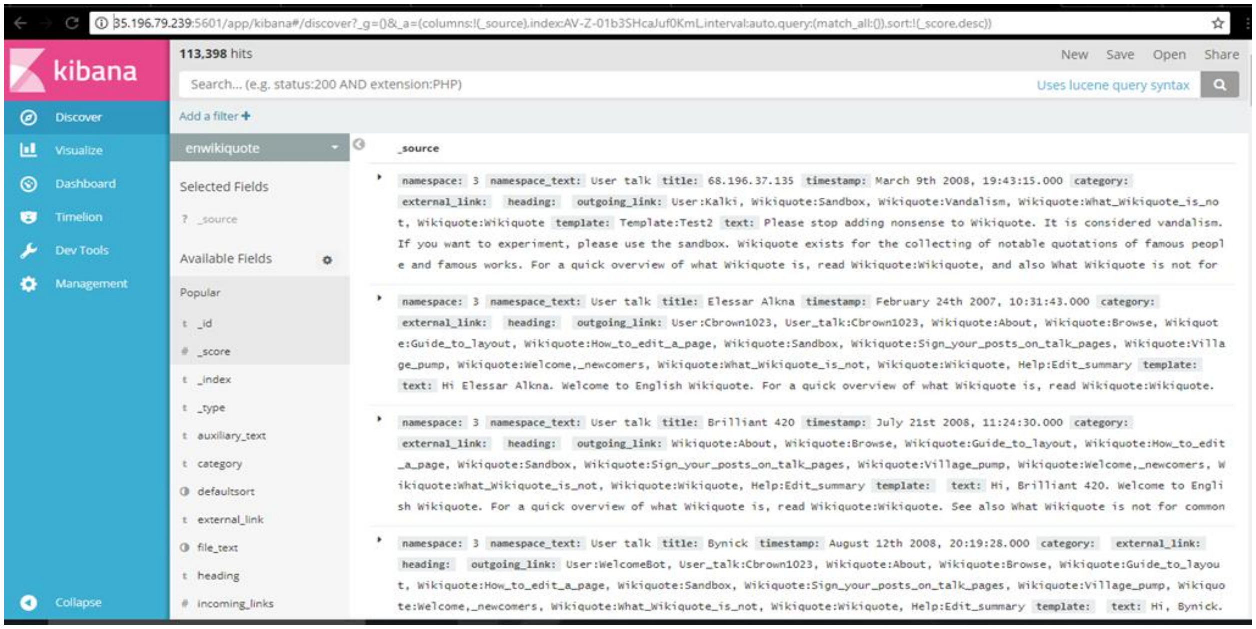


text: Beauty for some provides escape, who gain a happiness in eyeing the gorgeous buttocks of the ape or Autumn sunsets exquisitely dying. Let the rain kiss you. Let the rain beat upon your head with silver liquid drops. Let the rain sing you a lullaby. Like a welcome summer rain, humor may suddenly cleanse and cool the earth, the air and you. Negroes – Sweet and docile, Meek, humble, and kind: Beware the day – They change their mind. We Negro writers, just by being black, have been on the blacklist all our lives ... Censorship for us begins at the color line. When peoples care for you and cry for you, they c

text: I caught a rainbow A beautiful rainbow And made it mine But when the rain stopped I lost my rainbow My pretty rainbow Amid the sunshine [specific citation needed] version: 2,230,511 wiki: enwikiquote namespace: 1 namespace_text: Talk title: Thuraya AlArrayed timestamp: March 6th 2017, 13:23:55.000 category: Pages with inadequate citations external_link: outgoing_link: Wikiquote:Citing_sources template: Template:Fix cite source_text: == Unsourced == === Dreams === :I caught a rainbow :A beautiful rainbow :And made it mine :But when the rain stopped :I lost my rainbow :My pretty rainbow :Amid the su

text: Roger Smith: "We have choices. Some people like to stand in the rain without an umbrella. That's what it means to live free." namespace: 2 namespace_text: User title: Peter s timestamp: April 30th 2006, 15:08:56.000 category: external_link: heading: The Big O outgoing_link: template: text_bytes: 143 incoming_links: 0 redirect: auxiliary_text: source_text: == The Big O == Roger Smith: "We have choices. Some people like to stand in the rain without an umbrella. That's what it means to live free." opening_text: - version_type: external version: 212,573 defaultsort: false language: en

text: I once had a leather jacket that got ruined in the rain. Why does moisture ruin leather? Aren't cows outside a lot of the time? When it's raining, do cows go up to the farmhouse, "Let us in! We're all wearing leather! Open the door! We're going to ruin the whole outfit here!" Jerry Seinfeld namespace: 2 namespace text: User title: Changetheworld timestamp: May 17

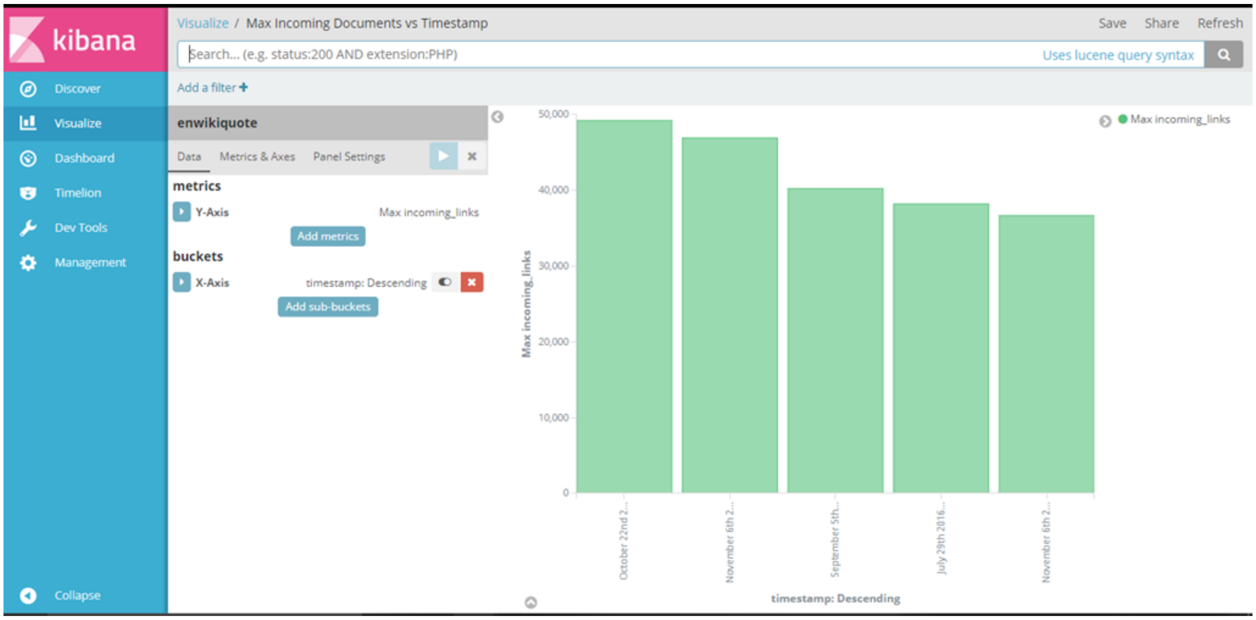


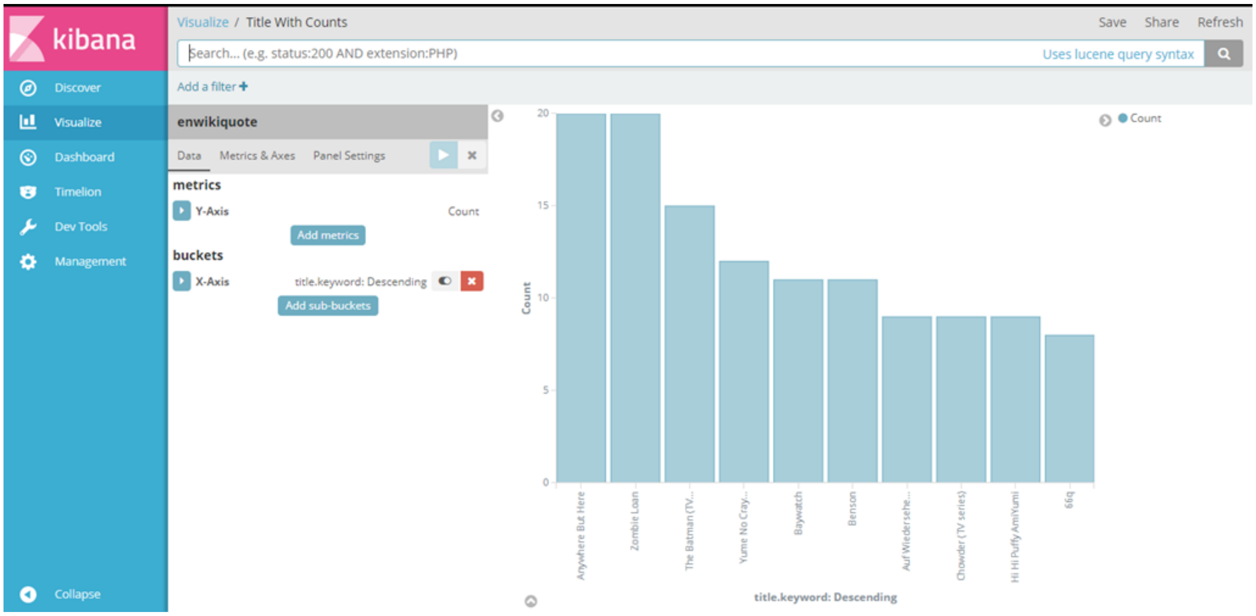
namespace: 3 namespace_text: User talk title: 68.196.37.135 timestamp: March 9th 2008, 19:43:15.000 category: external_link: heading: outgoing_link: User:Kalki, Wikiquote:Sandbox, Wikiquote:Vandalism, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote template: Template:Test2 text: Please stop adding nonsense to Wikiquote. It is considered vandalism. If you want to experiment, please use the sandbox. Wikiquote exists for the collecting of notable quotations of famous people and famous works. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote, and also What Wikiquote is not for

namespace: 3 namespace_text: User talk title: Elessar Alkna timestamp: February 24th 2007, 10:31:43.000 category: external_link: heading: outgoing_link: User:Cbrown1023, User_talk:Cbrown1023, Wikiquote:About, Wikiquote:Browse, Wikiquote:Guide_to_layout, Wikiquote:How_to_edit_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Village_pump, Wikiquote:Welcome,_newcomers, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template: text: Hi Elessar Alkna. Welcome to English Wikiquote. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote.

namespace: 3 namespace_text: User talk title: Brilliant 420 timestamp: July 21st 2008, 11:24:30.000 category: external_link: heading: outgoing_link: Wikiquote:About, Wikiquote:Browse, Wikiquote:Guide_to_layout, Wikiquote:How_to_edit_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Village_pump, Wikiquote:Welcome,_newcomers, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template: text: Hi, Brilliant 420. Welcome to English Wikiquote. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote. See also What Wikiquote is not for common

namespace: 3 namespace_text: User talk title: Bynick timestamp: August 12th 2008, 20:19:28.000 category: external_link: heading: outgoing_link: User:WelcomeBot, User_talk:Cbrown1023, Wikiquote:About, Wikiquote:Browse, Wikiquote:Guide_to_layout, Wikiquote:How_to_edit_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Village_pump, Wikiquote:Welcome,_newcomers, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template: text: Hi, Bynick.

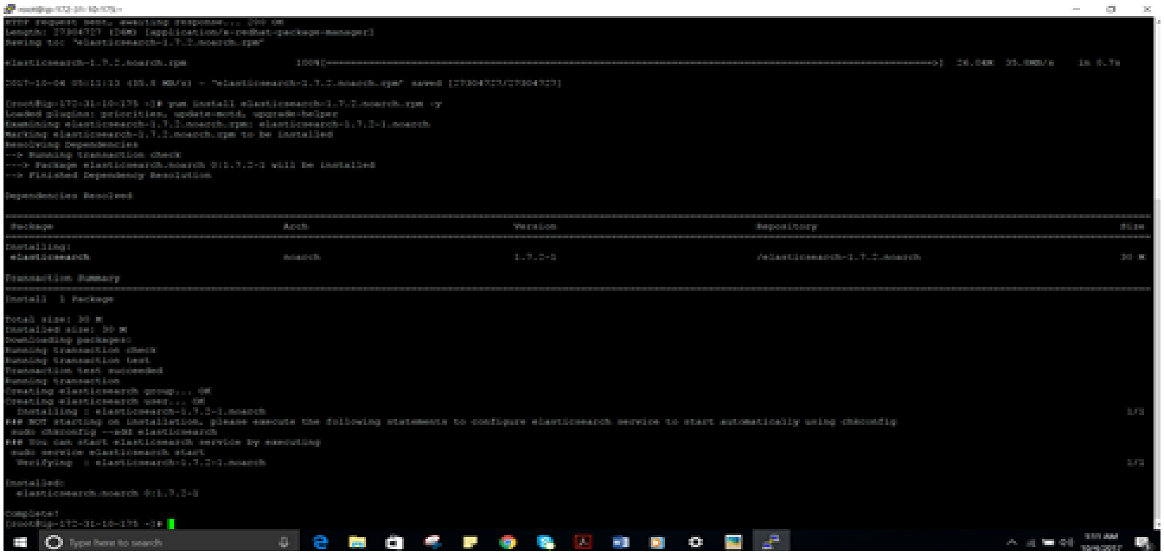




INDEXING WIKI DUMPS On AMAZON WEB SERVICES using ELK Stack:

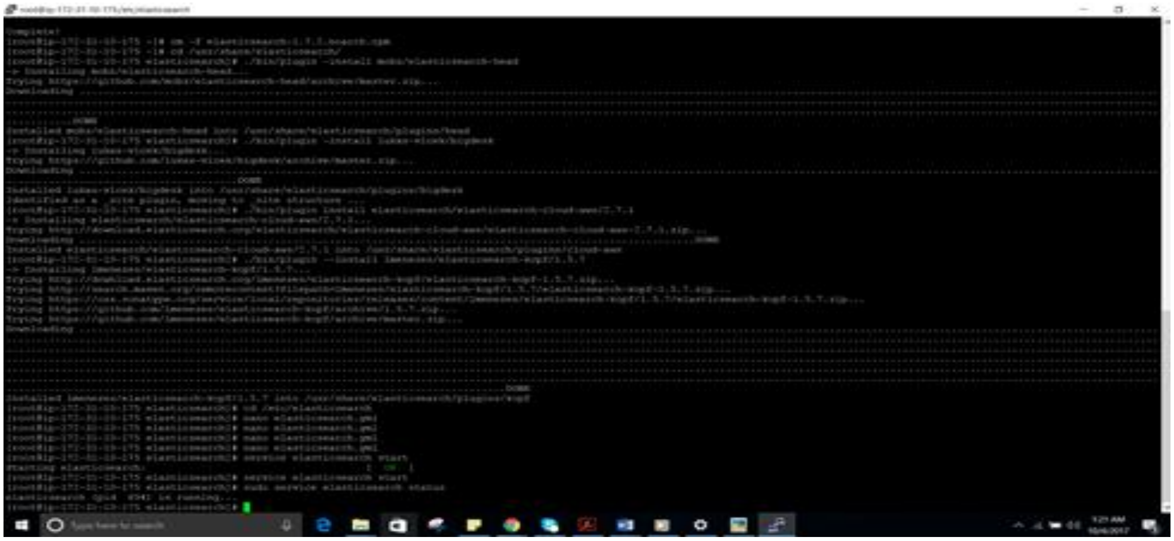
The same commands are to be executed for installation of Elasticsearch, Logstash and Kibana as used for Google Cloud Platform. Since the commands are the very same, you can see the screenshots of successful installation of ELK stack on AMAZON WEB SERVICES.

Elasticsearch Installation:



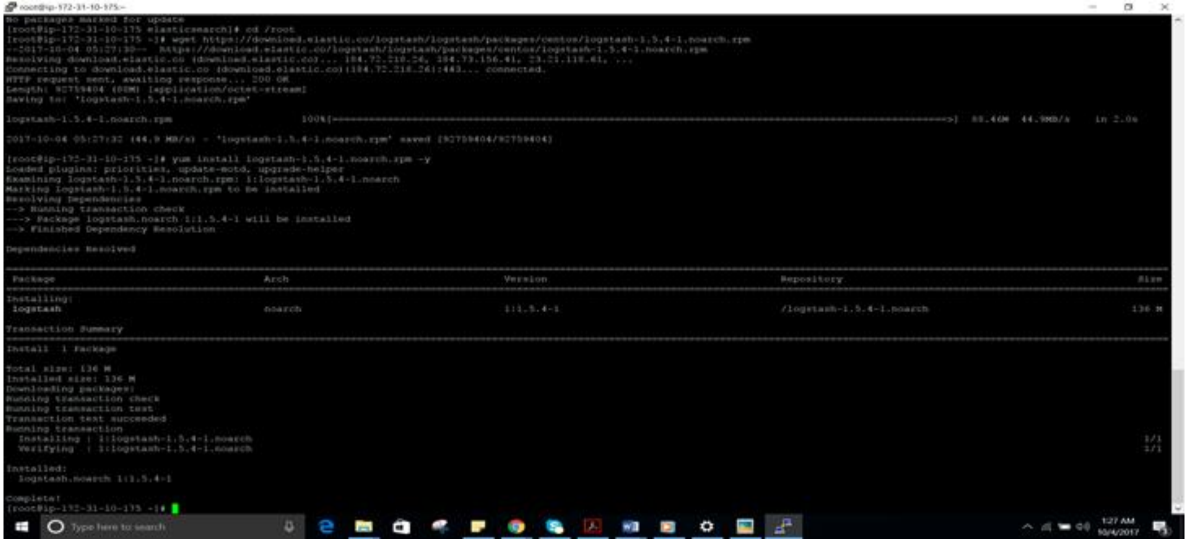
```
HTTP request sent, awaiting response... 200 OK
Length: 27304727 (26M) [application/x-redhat-package-manager]
Saving to: 'elasticsearch-1.7.2.noarch.rpm'

elasticsearch-1.7.2.noarch.rpm              100%[=====================================================================>] 26.04M  35.0MB/s    in 0.7s

2017-10-06 05:11:13 (35.0 MB/s) - 'elasticsearch-1.7.2.noarch.rpm' saved [27304727/27304727]

[root@ip-172-31-10-175 ~]# yum install elasticsearch-1.7.2.noarch.rpm -y
Loaded plugins: priorities, update-motd, upgrade-helper
Examining elasticsearch-1.7.2.noarch.rpm: elasticsearch-1.7.2-1.noarch
Marking elasticsearch-1.7.2.noarch.rpm to be installed
Resolving Dependencies
--> Running transaction check
---> Package elasticsearch.noarch 0:1.7.2-1 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================================================================================
 Package                              Arch                        Version                        Repository                                 Size
================================================================================================================================================
Installing:
 elasticsearch                        noarch                      1.7.2-1                        /elasticsearch-1.7.2.noarch                30 M

Transaction Summary
================================================================================================================================================
Install  1 Package

Total size: 30 M
Installed size: 30 M
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Creating elasticsearch group... OK
Creating elasticsearch user... OK
  Installing : elasticsearch-1.7.2-1.noarch                                                                                                 1/1
### NOT starting on installation, please execute the following statements to configure elasticsearch service to start automatically using chkconfig
 sudo chkconfig --add elasticsearch
### You can start elasticsearch service by executing
 sudo service elasticsearch start
  Verifying  : elasticsearch-1.7.2-1.noarch                                                                                                 1/1

Installed:
  elasticsearch.noarch 0:1.7.2-1

Complete!
[root@ip-172-31-10-175 ~]#
```

Kibana Installation:



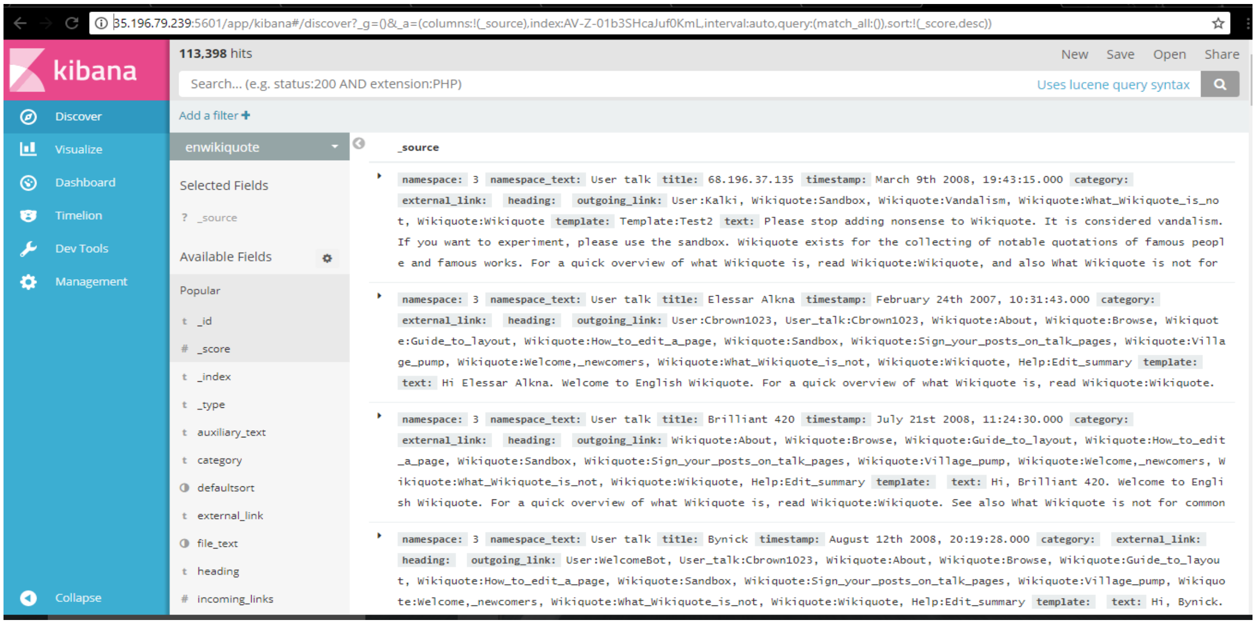
```
Complete!
[root@ip-172-31-10-175 ~]# rm -f elasticsearch-1.7.2.noarch.rpm
[root@ip-172-31-10-175 ~]# cd /usr/share/elasticsearch/
[root@ip-172-31-10-175 elasticsearch]# ./bin/plugin -install mobz/elasticsearch-head
-> Installing mobz/elasticsearch-head...
Trying https://github.com/mobz/elasticsearch-head/archive/master.zip...
Downloading ........................................................................................................................
........................................................................................................................................
........................................................................................................................................
............DONE
Installed mobz/elasticsearch-head into /usr/share/elasticsearch/plugins/head
[root@ip-172-31-10-175 elasticsearch]# ./bin/plugin -install lukas-vlcek/bigdesk
-> Installing lukas-vlcek/bigdesk...
Trying https://github.com/lukas-vlcek/bigdesk/archive/master.zip...
Downloading ........................................................................................................................
.............................................DONE
Installed lukas-vlcek/bigdesk into /usr/share/elasticsearch/plugins/bigdesk
Identified as a _site plugin, moving to _site structure ...
[root@ip-172-31-10-175 elasticsearch]# ./bin/plugin install elasticsearch/elasticsearch-cloud-aws/2.7.1
-> Installing elasticsearch/elasticsearch-cloud-aws/2.7.1...
Trying http://download.elasticsearch.org/elasticsearch/elasticsearch-cloud-aws/elasticsearch-cloud-aws-2.7.1.zip...
Downloading ........................................................................................................................DONE
Installed elasticsearch/elasticsearch-cloud-aws/2.7.1 into /usr/share/elasticsearch/plugins/cloud-aws
[root@ip-172-31-10-175 elasticsearch]# ./bin/plugin --install lmenezes/elasticsearch-kopf/1.5.7
-> Installing lmenezes/elasticsearch-kopf/1.5.7...
Trying http://download.elasticsearch.org/lmenezes/elasticsearch-kopf/elasticsearch-kopf-1.5.7.zip...
Trying http://search.maven.org/remotecontent?filepath=lmenezes/elasticsearch-kopf/1.5.7/elasticsearch-kopf-1.5.7.zip...
Trying https://oss.sonatype.org/service/local/repositories/releases/content/lmenezes/elasticsearch-kopf/1.5.7/elasticsearch-kopf-1.5.7.zip...
Trying https://github.com/lmenezes/elasticsearch-kopf/archive/1.5.7.zip...
Trying https://github.com/lmenezes/elasticsearch-kopf/archive/master.zip...
Downloading ........................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
...........................................................................................................................DONE
Installed lmenezes/elasticsearch-kopf/1.5.7 into /usr/share/elasticsearch/plugins/kopf
[root@ip-172-31-10-175 elasticsearch]# cd /etc/elasticsearch
[root@ip-172-31-10-175 elasticsearch]# nano elasticsearch.yml
[root@ip-172-31-10-175 elasticsearch]# nano elasticsearch.yml
[root@ip-172-31-10-175 elasticsearch]# nano elasticsearch.yml
[root@ip-172-31-10-175 elasticsearch]# nano elasticsearch.yml
[root@ip-172-31-10-175 elasticsearch]# service elasticsearch start
Starting elasticsearch:                                    [  OK  ]
[root@ip-172-31-10-175 elasticsearch]# service elasticsearch start
[root@ip-172-31-10-175 elasticsearch]# sudo service elasticsearch status
elasticsearch (pid  8941) is running...
[root@ip-172-31-10-175 elasticsearch]#
```

Logstash Installation:



```
No packages marked for update
[root@ip-172-31-10-175 elasticsearch]# cd /root
[root@ip-172-31-10-175 ~]# wget https://download.elastic.co/logstash/logstash/packages/centos/logstash-1.5.4-1.noarch.rpm
--2017-10-04 05:27:30--  https://download.elastic.co/logstash/logstash/packages/centos/logstash-1.5.4-1.noarch.rpm
Resolving download.elastic.co (download.elastic.co)... 184.72.218.26, 184.73.156.41, 23.21.118.61, ...
Connecting to download.elastic.co (download.elastic.co)|184.72.218.26|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 92759404 (88M) [application/octet-stream]
Saving to: 'logstash-1.5.4-1.noarch.rpm'

logstash-1.5.4-1.noarch.rpm                100%[==================================================================>] 88.46M  44.9MB/s    in 2.0s

2017-10-04 05:27:32 (44.9 MB/s) - 'logstash-1.5.4-1.noarch.rpm' saved [92759404/92759404]

[root@ip-172-31-10-175 ~]# yum install logstash-1.5.4-1.noarch.rpm -y
Loaded plugins: priorities, update-motd, upgrade-helper
Examining logstash-1.5.4-1.noarch.rpm: 1:logstash-1.5.4-1.noarch
Marking logstash-1.5.4-1.noarch.rpm to be installed
Resolving Dependencies
--> Running transaction check
---> Package logstash.noarch 1:1.5.4-1 will be installed
--> Finished Dependency Resolution

Dependencies Resolved

================================================================================================================================
 Package                          Arch                       Version                      Repository                        Size
================================================================================================================================
Installing:
 logstash                         noarch                     1:1.5.4-1                    /logstash-1.5.4-1.noarch         136 M

Transaction Summary
================================================================================================================================
Install  1 Package

Total size: 136 M
Installed size: 136 M
Downloading packages:
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Installing : 1:logstash-1.5.4-1.noarch                                                                                    1/1
  Verifying  : 1:logstash-1.5.4-1.noarch                                                                                    1/1

Installed:
  logstash.noarch 1:1.5.4-1

Complete!
[root@ip-172-31-10-175 ~]#
```

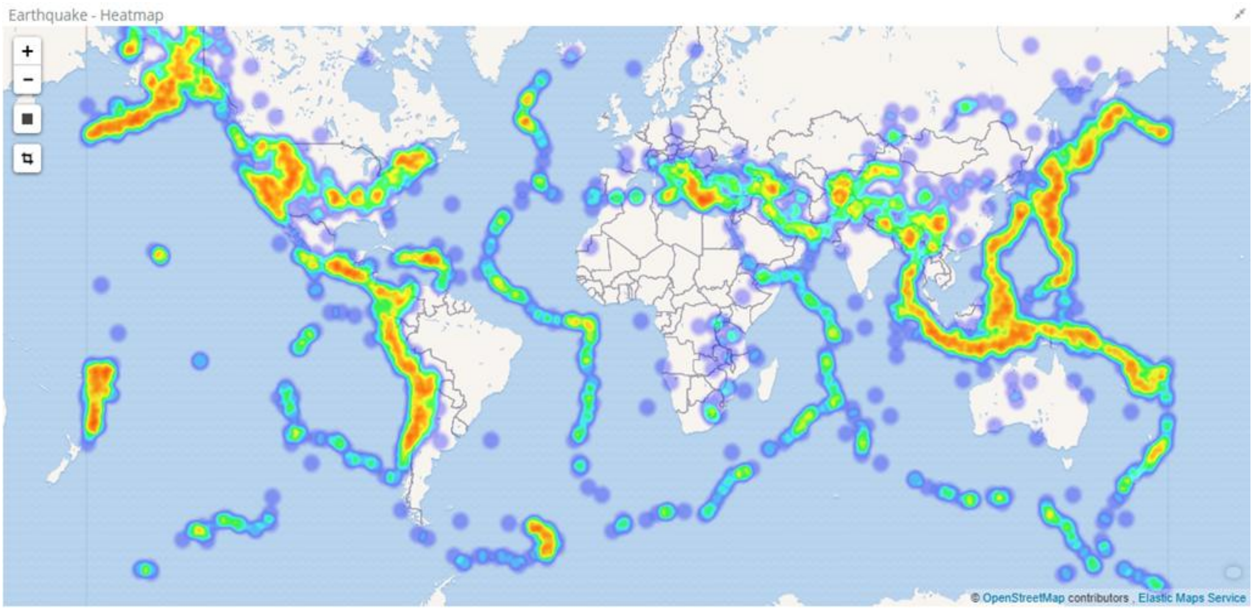
KIBANA VISUALIZATIONS THROUGH AWS:



```
namespace: 3 namespace_text: User talk title: 68.196.37.135 timestamp: March 9th 2008, 19:43:15.000 category:
external_link: heading: outgoing_link: User:Kalki, Wikiquote:Sandbox, Wikiquote:Vandalism, Wikiquote:What_Wikiquote_is_no
t, Wikiquote:Wikiquote template: Template:Test2 text: Please stop adding nonsense to Wikiquote. It is considered vandalism.
If you want to experiment, please use the sandbox. Wikiquote exists for the collecting of notable quotations of famous peopl
e and famous works. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote, and also What Wikiquote is not for

namespace: 3 namespace_text: User talk title: Elessar Alkna timestamp: February 24th 2007, 10:31:43.000 category:
external_link: heading: outgoing_link: User:Cbrown1023, User_talk:Cbrown1023, Wikiquote:About, Wikiquote:Browse, Wikiquot
e:Guide_to_layout, Wikiquote:How_to_edit_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Villa
ge_pump, Wikiquote:Welcome,_newcomers, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template:
text: Hi Elessar Alkna. Welcome to English Wikiquote. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote.

namespace: 3 namespace_text: User talk title: Brilliant 420 timestamp: July 21st 2008, 11:24:30.000 category:
external_link: heading: outgoing_link: Wikiquote:About, Wikiquote:Browse, Wikiquote:Guide_to_layout, Wikiquote:How_to_edit
_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Village_pump, Wikiquote:Welcome,_newcomers, W
ikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template: text: Hi, Brilliant 420. Welcome to Engli
sh Wikiquote. For a quick overview of what Wikiquote is, read Wikiquote:Wikiquote. See also What Wikiquote is not for common

namespace: 3 namespace_text: User talk title: Bynick timestamp: August 12th 2008, 20:19:28.000 category: external_link:
heading: outgoing_link: User:WelcomeBot, User_talk:Cbrown1023, Wikiquote:About, Wikiquote:Browse, Wikiquote:Guide_to_layou
t, Wikiquote:How_to_edit_a_page, Wikiquote:Sandbox, Wikiquote:Sign_your_posts_on_talk_pages, Wikiquote:Village_pump, Wikiquo
te:Welcome,_newcomers, Wikiquote:What_Wikiquote_is_not, Wikiquote:Wikiquote, Help:Edit_summary template: text: Hi, Bynick.
```

Earthquakes Data

For more sample visualizations and better hands- on experience on Elasticsearch and Kibana, we tried this data available on github https://github.com/elastic/examples/tree/master/Exploring%20Public%20Datasets/earthquakes



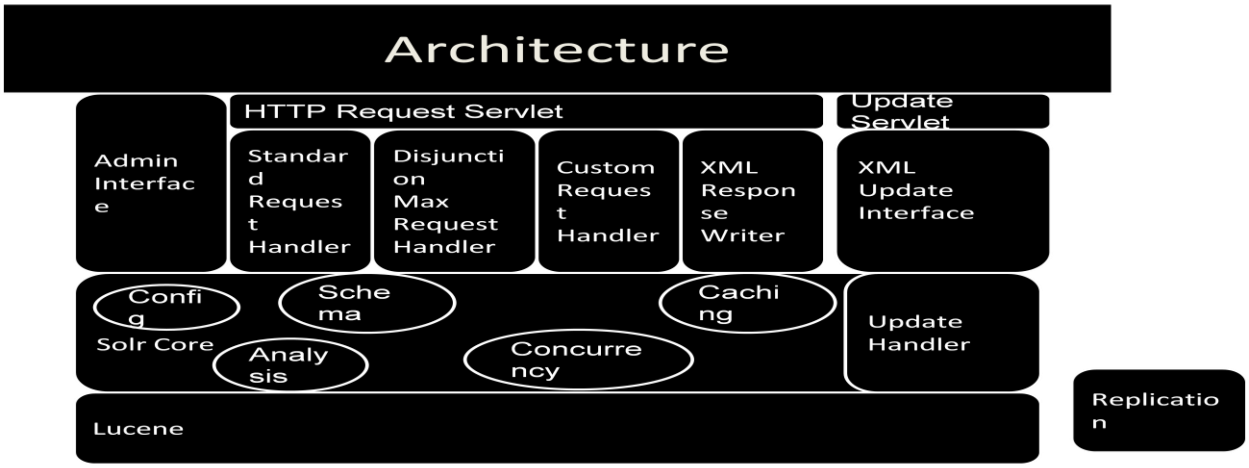
INDEXING WIKI DUMPS ON MICROSOFT AZURE USING SOLR:

SOLR:

Solr is powered by Lucene, a powerful open-source full-text search library, under the hood. Solr is designed for scalability and fault tolerance. Solr is widely used for enterprise search and analytics use cases.

- XML/HTTP Interfaces
- Loose Schema to define types and fields
- Web Administration Interface
- Extensive Caching
- Index Replication
- Extensible Open Architecture

ARCHITECTURE OF SOLR:



INSTALLATION STEPS OF SOLR:

Below are the steps for the installation steps of SOLR clearly explained step by step:

1. Download solr-3.4
2. Download wikipedia dump
3. data-config.xml was used to index Wikipedia dump.

```
<dataConfig>
<dataSource type="FileDataSource" encoding="UTF-8" />
<document>
<entity name="page"
processor="XPathEntityProcessor"
stream="true"
forEach="/mediawiki/page/"
```

4. The relevant portion of schema.xml is below:

```
<field name="id" type="string"  indexed="true" stored="true" required="true"/>
<field name="title" type="string"  indexed="true" stored="false"/>
<field name="revision" type="sint" indexed="true" stored="true"/>
<field name="user" type="string" indexed="true" stored="true"/>
<field name="userId" type="int" indexed="true" stored="true"/>
<field name="text" type="text" indexed="true" stored="false"/>
<uniqueKey>id</uniqueKey>
<copyField source="title" dest="titleText"/>
```

5. Add Dih request handler in solrconfig.xml file

```
<requestHandler name="/update/dih" startup="lazy">
<lst name="defaults">
<str name="config">dih-config.xml</str>
</lst>
```

6. Restart solr
7. Index some documents using below command

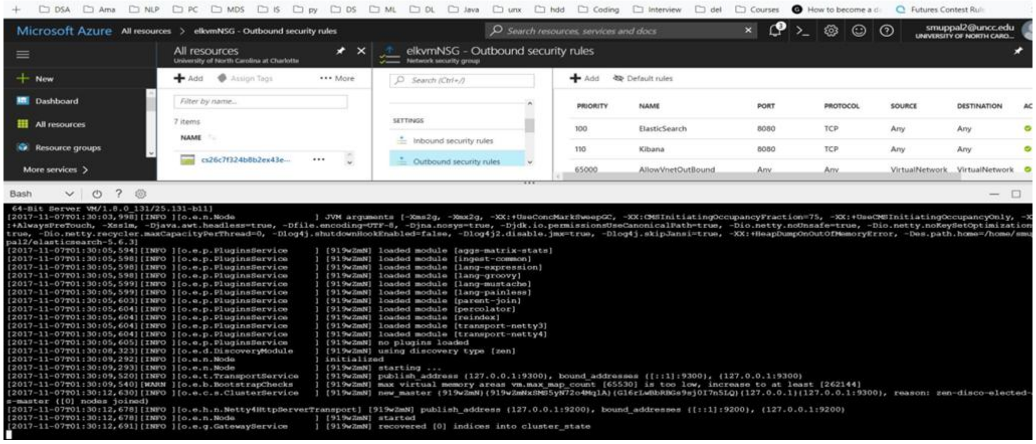
http://localhost:8983/solr/update/dih?command=full-import

Schema: Analyzers:

```
<fieldtype name="nametext" class="solr.TextField">
        <analyzer class="org.apache.lucene.analysis.WhitespaceAnalyzer"/>
</fieldtype>
<fieldtype name="text" class="solr.TextField">
        <analyzer>
                <tokenizer class="solr.StandardTokenizerFactory"/>
                <filter class="solr.StandardFilterFactory"/>
                <filter class="solr.LowerCaseFilterFactory"/>
                <filter class="solr.StopFilterFactory"/>
                <filter class="solr.PorterStemFilterFactory"/>
        </analyzer>
</fieldtype>
<fieldtype name="myfieldtype" class="solr.TextField">
        <analyzer>
                <tokenizer class="solr.WhitespaceTokenizerFactory"/>
                <filter class="solr.SnowballPorterFilterFactory" language="German" />
        </analyzer>
</fieldtype>
```

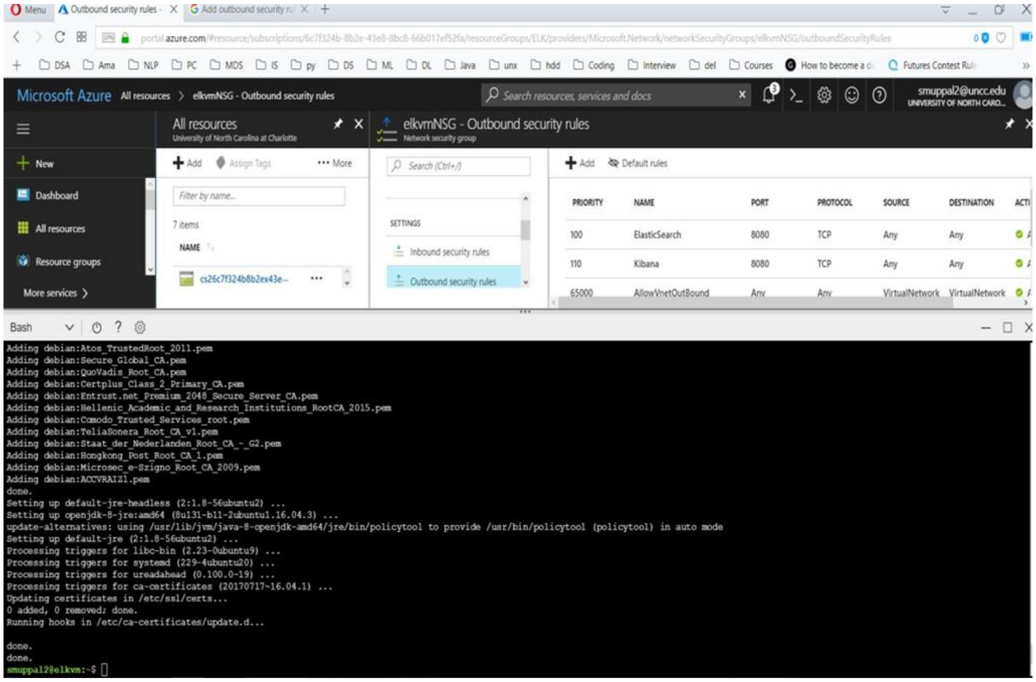
INSTALLATION OF SOLR IN AZURE:

Below are the screenshots for installation of SOLR using AZURE



PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	AC
100	ElasticSearch	8080	TCP	Any	Any	●
110	Kibana	8080	TCP	Any	Any	●
65000	AllowVnetOutBound	Any	Any	VirtualNetwork	VirtualNetwork	●

```
 64-Bit Server VM/1.8.0_131/25.131-b11]
[2017-11-07T01:30:03,998][INFO ][o.e.n.Node               ] JVM arguments [-Xms2g, -Xmx2g, -XX:+UseConcMarkSweepGC, -XX:CMSInitiatingOccupancyFraction=75, -XX:+UseCMSInitiatingOccupancyOnly, -X
:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=true, -Dfile.encoding=UTF-8, -Djna.nosys=true, -Djdk.io.permissionsUseCanonicalPath=true, -Dio.netty.noUnsafe=true, -Dio.netty.noKeySetOptimization=
true, -Dio.netty.recycler.maxCapacityPerThread=0, -Dlog4j.shutdownHookEnabled=false, -Dlog4j2.disable.jmx=true, -Dlog4j.skipJansi=true, -XX:+HeapDumpOnOutOfMemoryError, -Des.path.home=/home/smu
pal2/elasticsearch-5.6.3]
[2017-11-07T01:30:05,594][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [aggs-matrix-stats]
[2017-11-07T01:30:05,598][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [ingest-common]
[2017-11-07T01:30:05,598][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [lang-expression]
[2017-11-07T01:30:05,598][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [lang-groovy]
[2017-11-07T01:30:05,599][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [lang-mustache]
[2017-11-07T01:30:05,599][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [lang-painless]
[2017-11-07T01:30:05,603][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [parent-join]
[2017-11-07T01:30:05,604][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [percolator]
[2017-11-07T01:30:05,604][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [reindex]
[2017-11-07T01:30:05,604][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [transport-netty3]
[2017-11-07T01:30:05,604][INFO ][o.e.p.PluginsService     ] [919wZmN] loaded module [transport-netty4]
[2017-11-07T01:30:05,605][INFO ][o.e.p.PluginsService     ] [919wZmN] no plugins loaded
[2017-11-07T01:30:08,323][INFO ][o.e.d.DiscoveryModule    ] [919wZmN] using discovery type [zen]
[2017-11-07T01:30:09,292][INFO ][o.e.n.Node               ] initialized
[2017-11-07T01:30:09,293][INFO ][o.e.n.Node               ] [919wZmN] starting ...
[2017-11-07T01:30:09,520][INFO ][o.e.t.TransportService   ] [919wZmN] publish_address {127.0.0.1:9300}, bound_addresses {[::1]:9300}, {127.0.0.1:9300}
[2017-11-07T01:30:09,540][WARN ][o.e.b.BootstrapChecks    ] [919wZmN] max virtual memory areas vm.max_map_count [65530] is too low, increase to at least [262144]
[2017-11-07T01:30:12,630][INFO ][o.e.c.s.ClusterService   ] [919wZmN] new_master {919wZmN}{919wZmNxSMS5yN72o4MqlA}{G16rLwBbRBGs9sj0I7n5LQ}{127.0.0.1}{127.0.0.1:9300}, reason: zen-disco-elected-
s-master ([0] nodes joined)
[2017-11-07T01:30:12,678][INFO ][o.e.h.n.Netty4HttpServerTransport] [919wZmN] publish_address {127.0.0.1:9200}, bound_addresses {[::1]:9200}, {127.0.0.1:9200}
[2017-11-07T01:30:12,678][INFO ][o.e.n.Node               ] [919wZmN] started
[2017-11-07T01:30:12,691][INFO ][o.e.g.GatewayService     ] [919wZmN] recovered [0] indices into cluster_state
```

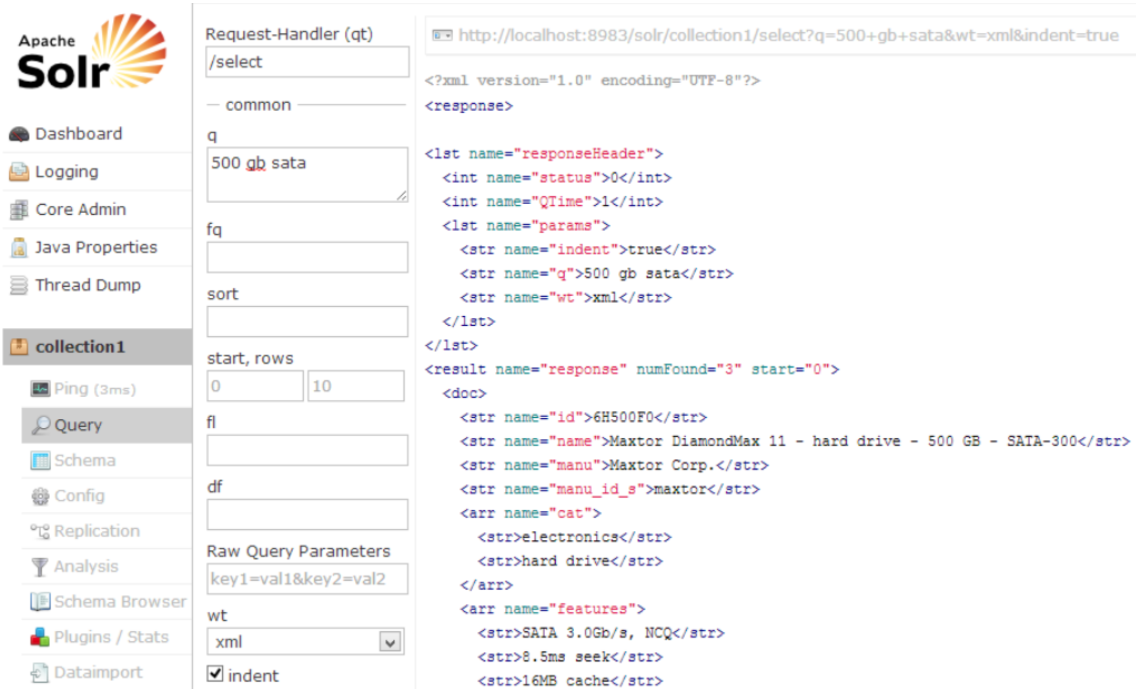


PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTI
100	ElasticSearch	8080	TCP	Any	Any	●
110	Kibana	8080	TCP	Any	Any	●
65000	AllowVnetOutBound	Any	Any	VirtualNetwork	VirtualNetwork	●

```
Adding debian:Atos_TrustedRoot_2011.pem
Adding debian:Secure_Global_CA.pem
Adding debian:QuoVadis_Root_CA.pem
Adding debian:Certplus_Class_2_Primary_CA.pem
Adding debian:Entrust.net_Premium_2048_Secure_Server_CA.pem
Adding debian:Hellenic_Academic_and_Research_Institutions_RootCA_2015.pem
Adding debian:Comodo_Trusted_Services_root.pem
Adding debian:TeliaSonera_Root_CA_v1.pem
Adding debian:Staat_der_Nederlanden_Root_CA_-_G2.pem
Adding debian:Hongkong_Post_Root_CA_1.pem
Adding debian:Microsec_e-Szigno_Root_CA_2009.pem
Adding debian:ACCVRAIZ1.pem
done.
Setting up default-jre-headless (2:1.8-56ubuntu2) ...
Setting up openjdk-8-jre:amd64 (8u131-b11-2ubuntu1.16.04.3) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/policytool to provide /usr/bin/policytool (policytool) in auto mode
Setting up default-jre (2:1.8-56ubuntu2) ...
Processing triggers for libc-bin (2.23-0ubuntu9) ...
Processing triggers for systemd (229-4ubuntu20) ...
Processing triggers for ureadahead (0.100.0-19) ...
Processing triggers for ca-certificates (20170717~16.04.1) ...
Updating certificates in /etc/ssl/certs...
0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...

done.
done.
smuppal2@elkvm:~$ 
```

INDEXED WIKIPEDIA DATA IN SOLR:

Below is the indexed wikipedia data using SOLR



Request-Handler (qt): /select

common

q: 500 gb sata

fq:

sort:

start, rows: 0 | 10

fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: xml

☑ indent

http://localhost:8983/solr/collection1/select?q=500+gb+sata&wt=xml&indent=true

```
<?xml version="1.0" encoding="UTF-8"?>
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">1</int>
  <lst name="params">
    <str name="indent">true</str>
    <str name="q">500 gb sata</str>
    <str name="wt">xml</str>
  </lst>
</lst>
<result name="response" numFound="3" start="0">
  <doc>
    <str name="id">6H500F0</str>
    <str name="name">Maxtor DiamondMax 11 - hard drive - 500 GB - SATA-300</str>
    <str name="manu">Maxtor Corp.</str>
    <str name="manu_id_s">maxtor</str>
    <arr name="cat">
      <str>electronics</str>
      <str>hard drive</str>
    </arr>
    <arr name="features">
      <str>SATA 3.0Gb/s, NCQ</str>
      <str>8.5ms seek</str>
      <str>16MB cache</str>
```

CREDITS:

https://logz.io/blog/elk-stack-google-cloud/
https://dzone.com/articles/how-to-install-the-elk-stack-on-google-cloud-platf-1
https://www.elastic.co/blog/loading-wikipedia
https://dumps.wikimedia.org/other/cirrussearch/current/
https://www.digitalocean.com/community/tutorials/how-to-use-kibana-dashboards-and-visualIzations
https://github.com/elastic/examples/tree/master/Exploring%20Public%20Datasets/earthquakes
https://stackoverflow.com/questions/3846793/running-solr-on-azure