Questions:

Indexing

Basic Indexing:

1- The commands and their output are available in the elasticcommands text file.

2- Explain in one sentence why we might break an index into shards.

Having multiple shards helps taking advantage of parallel processing on a single machine.

But the whole point is that if we start another elasticsearch instance on the same cluster, the shards will be distributed in an even way over the cluster. Because elasticsearch is a distributed search engine and this way we can make use of multiple nodes/machines to manage big amounts of data.

3- Explain in one sentence why we might replicate an index.

Replicas are used to increase search performance and for fail-over.

A replica shard is never going to be allocated on the same node where the related primary is (it would pretty much be like putting a backup on the same disk as the original data). With a setup like that, if a node goes down, we will still have the whole index. The replica shards will automatically become primaries and the cluster will work properly despite the node failure.

4- Explain in one sentence why your cluster health is yellow.

Because the one replica which is created by elastic search (by default) is not still allocated to any node.

Yellow means that some replicas are not (yet) allocated. The reason this happens for this index is because Elasticsearch by default created one replica for this index. Since we only have one node running at the moment, that one replica cannot yet be allocated (for high availability) until a later point in time when another node joins the cluster. Once that replica gets allocated onto a second node, the health status for this index will turn to green.

Indexing Reddit:

1- Ensure that your Elasticsearch instance has no indices in it. Provide the command you used to verify this.

Command:
curl -X GET "localhost:9200/_cat/indices?v&pretty"

Output:

```
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current Dload  Upload
Total   Spent    Left  Speed

100    83  100    83    0     0   8300      0 --:--:-- --:--:-- --:--:--  9222

health status index uuid pri rep docs.count docs.deleted store.size pri.store.size
```

2- Download the test.json file from the OWL assignment page. Run the following command, ensuring that you do this from the same folder that has the test.json file in it.

I used the command below to view the created index:

```
curl -X GET "localhost:9200/_cat/indices?v&pretty"
```

The output is:

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   208  100   208    0     0  16000      0 --:--:-- --:--:-- --:--:-- 17333
health status index    uuid                   pri rep docs.count docs.deleted store.size pri.store.size
yellow open   comments E-NtbOXLTwWbthkkb9kTnA   1   1      61013            0     34.6mb         34.6mb
```

What is the name of the index created by the command?

Comments

3- How large is the index? (pri.store.size)

34.6mb

Search

1. Do the three queries below return different sets of documents? How can you tell?

Query1:

max_score: 10.503805

- "Here is some example code that handles both cat with, and cat without arguments. (and as an added bonus, handles failures in a way similar to standard cat)",
- "* emacs\r\n\r\n* nano\r\n\r\n* cat\r\n\r\n* ?"
- "masturbating cat..so coool"
- "Copy Cat has kittens."
- "This reminds me of Kircher's \"cat piano\""
- "...and the cat and mouse game continues."
- "A giant cat and a giant toilet!"
- "That's easy... [ceiling cat!](http://www.surrenderthesoul.com/ceiling_cat.jpg)"
- "shame the cat one isnt an illusion"
- "My mother's cat used to do that.\r\n\r\nI think she got a new cat after this happened at around 3 AM one morning and she called the police..."

Query 2:

max_score: 11.629243

- http://www.knitemare.org/cats/
- "Stop piping cats!"
- "I thought he had cats."
- "Because cats do funny things."
- "Works on ferrets, dogs, cats, and kids too!"
- "I agree. I've lost 5 cats that way."
- "I'm amazed at the replies to my comment. I just found the phrase funny out of context and repeated it because it sounded funny. Stop piping cats. Stop piping cats."
- "Seems like she should have gotten rid of her cats",
- "fuck you psychology 101 for saying you can't train cats!"
- "What about treeing cats? Would that make for bonsai kittens?"

Query 3:

max_score: 18.84402

- " Lemur photos: The next [funny cat](http://www.knitemare.org/cats/index.php? Type=all) fad?"
- Cats are good jumpers. \r\nI once saw a cat clear a fence.\r\nRichard Norton can teach Zimba I am confident.
- I'm not sure where you're getting that from. Cats can have both hunger and pain. Ask any cat with an empty food dish. Also, it does matter if you have time to look after them properly. However, looking after cats properly requires much less time than kids.\r\n\r\nYou have some strange ideas, sir.",

- > On the other hand, infected women tend to be more outgoing, friendly, more promiscuous, and are considered more attractive to men compared with non-infected controls.\r\n\r\nSo how long before cat shit is classified as a social drug, and cats are banned or heavily taxed?",
- "It's not really about cats. Basically, I said something along the lines of \"If you do X, Y might happen,\" and you said \"But I know of times where X was done and Y didn't happen.\" So your argument really only does anything if it's in response to a claim like \"Whenever X happens, Y *always* happens.\" So it's like me saying that throwing a cat off a roof is a bad idea and that the cat is likely to get hurt; then you say \"But my cousin did it and his cat was fine.\" It's still a risky practice.\r\n\r\nAnd about it being a major contributing factor, I dunno. I'm sure there are studies about this but I'm too lazy to look them up.",
- "I call nearly all of those tags overly picky relics of the day when HTML was trying to be a purely abstract markup in the style of docbook. On today's web they're as dumb as having a <cat> tag for talking about cats.\r\n\r\nOf all of those, only abbr and acronym make sense - because they actually add something new, a mouseover popup of the expansion.",
- "Probably the same stuff europe is doing. Tax refunds, and better/cheaper childcare, yadda yadda yadda. No idea whether it's effective.\r\n\r\nThough, the real thing would be to convince someone that a kid isn't just an expensive pet. Lord knows how to do that though.\r\n\r\nMost yuppies would rather spend money on anything but, and even if they do get lonely would still prefer getting a cat. Cats never wear diapers."

As we can see, these three queries return different sets of documents.

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.

As we can see, for the first query, the output contains only the documents with the exact term “cat” and it does not include any document with the term “cats”. The same goes for query2 too. The output only contains the documents with the exact term “cats” that the query had. This shows there is no stemming applied, because if there was, for both queries, all or a subset of documents which have the terms “cat” or “cats” will be returned. How many of these documents will be returned with stemming enabled, would depend on the stemming algorithm being applied.

2. Give the highest score for each query.

- Query1: max_score: 10.503805
- Query 2: max_score: 11.629243
- Query 3: max_score: 18.84402

3. What can you deduce about the default stemming procedure used by Elasticsearch?

When stemming is not setup on the document field (containing the documents we are using for our query), an Elastic Search query searching on a specific term (for example “cat”) will return only the documents that contain the exact term. Here, Query1 returns documents with “cat” and not “cats” and query2 returns documents with “cats” and not “cat”. While if stemming is set up on the document field, all of the documents containing the stem will be returned as part of the

search result set. As mentioned above, how many of these documents will be returned with stemming enabled depends on the stemming algorithm being applied.

4. Give a query that could be used to check whether or not Elasticsearch removes a common English stopword.

I used the command below to check what tokens it's looking for:

```
mahta@LAPTOP-QM32K7CA MINGW64 /
$ curl -X POST "localhost:9200/comments/_analyze?pretty" -H 'Content-Type: application/json' -d'
> {
>    "text": "the book"
> }
> '
```

The output is:

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   322  100   295  100    27  32777   3000 --:--:-- --:--:-- --:--:-- 40250{
  "tokens" : [
    {
      "token" : "the",
      "start_offset" : 0,
      "end_offset" : 3,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "book",
      "start_offset" : 4,
      "end_offset" : 8,
      "type" : "<ALPHANUM>",
      "position" : 1
    }
  ]
}
```

As we can see, it's looking for both tokens: "the" in position 0 and "book" in position 1. So it's not removing the common stopword which here is "the".

The query below looks for "the":

```
mahta@LAPTOP-QM32K7CA MINGW64 ~/Desktop/second semester/Unstructured Data/Assignments/Assignment1A
$ curl -XGET 'localhost:9200/_search?pretty' -H 'Content-Type: application/json' -d'
> { "query": { "match" : { "body" : { "query" : "the"} } } } '
```

The output is shown:

- "Its Cook Strait, The image also shows the top of the South Island down to the Banks Peninsula in the bottom of the pic (Christchurch is located on the top of the Peninsula) Wellington (NZ's Capital) is obscured by the antenna sticking up over the bit of land at the top of the image."
- "Actually, if the bottom of the fridge trays weren't solid, the gap where the shelved where could retain all the cold air. Only the items inside the fridge would move, not the cold air."
- "1996 was the golden year for me. \r\n\r\nThe best pirates, the best police, the best outer space, and the absolutely best underwater sets. Also, the start of the western, too. The only thing it's missing out on are the explorers."
- "The Jabber client just sends the message to the server and the server is the one that figures it out (same for mail). In the case of mail, the servers asks for the MX record of the domain name. In the Jabber case, the server asks for the SRV record of _jabber._tcp.pupeno.com, that is like asking Where is the JABBER TCP SeRVer of PUPENO.COM."
- "The bears mauled *42* of the youths. Truly, the bible does contain the answer to the great question."
- "\"...O'er the land of the free and the home of the brave.\""
- "Ok, I rate the ones that creep me out the most: \r\n\r\nThe King\r\nThe Noid\r\nThe Nasty Hamster thing"
- "The quote in the title is the only funny line in the piece.\r\n"
- "The reaction of the people in the town kind of proves the point."
- "The article has correct assertions, yet the title is ridiculous. If you define the founding of the US as the signing of the Declaration of Independence, a vast majority of the signers were Christians, and the rest were deists."

Analyzers

1- Delete the all indices in your Elasticsearch instance before beginning this section. Submit the code you use to do so.

First, we check to see all the indices in the cluster.

```
mahta@LAPTOP-QM32K7CA MINGW64 ~/Desktop/second semester/Unstructured Data/Assignments/Assignment1A
$ curl -X GET "localhost:9200/_cat/indices?v&pretty"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   208  100   208    0     0   2736      0 --:--:-- --:--:-- --:--:--  2736
health status index    uuid                   pri rep docs.count docs.deleted store.size pri.store.size
yellow open   comments lo8RNqHYQrW3__gb8t-42w   1   1      61013            0     34.7mb         34.7mb
```

The only index in the cluster is Comments. We delete the index using the command below:

```
mahta@LAPTOP-QM32K7CA MINGW64 /
$ curl -X DELETE "localhost:9200/comments?pretty&pretty"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100    28  100    28    0     0    560      0 --:--:-- --:--:-- --:--:--   571{
  "acknowledged" : true
}
```

We check the indices to make sure there is no index in the cluster.

```
mahta@LAPTOP-QM32K7CA MINGW64 /
$ curl -X GET "localhost:9200/_cat/indices?v&pretty"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100    83  100    83    0     0  13833      0 --:--:-- --:--:-- --:--:-- 13833
health status index uuid pri rep docs.count docs.deleted store.size pri.store.size
```

2- Run the following two (quite large) commands to change the default stemming behavior.

First command to define a new analyzer, my_analyzer, which uses a stemmer

```
mahta@LAPTOP-QM32K7CA MINGW64 ~/Desktop/second semester/Unstructured Data/Assignments/Assignment1A
$ curl -XPUT 'localhost:9200/comments?pretty' -H 'Content-Type: application/json' -d'
> {
>     "settings": {
>         "analysis" : {
>             "analyzer" : {
>                 "my_analyzer" : {
>                     "tokenizer" : "standard",
>                     "filter" : ["lowercase", "my_stemmer"]
>                 }
>             },
>             "filter" : {
>                 "my_stemmer" : {
>                     "type" : "stemmer",
>                     "name" : "english"
>                 }
>             }
>         }
>     }
> }'
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   513  100    84  100   429    608   3108 --:--:-- --:--:-- --:--:--  3744{
  "acknowledged" : true,
  "shards_acknowledged" : true,
  "index" : "comments"
}
```

Second command to inform Elasticsearch use my_analyzer when analyzing the body field of documents going into the comments index.

```
mahta@LAPTOP-QM32K7CA MINGW64 ~/Desktop/second semester/Unstructured Data/Assignments/Assignment1A
$ curl -XPUT 'localhost:9200/comments/_mapping?pretty' -H 'Content-Type: application/json' -d'
> {
>         "properties" : {
>           "body" : {
>             "type" : "text",
>             "analyzer" : "my_analyzer"
>           }
>         }
> }
> '
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   169  100    28  100   141    777   3916 --:--:-- --:--:-- --:--:--  4828{
  "acknowledged" : true
}
```

1- Re-run the three queries from the previous section. In a few sentences, describe how the results have changed.

For all three queries, the outputs (body) were the same and it is shown below.
Because now we have stemming applied for these terms, “cat” and “cats”, the stem which is “cat” is being considered. That’s why the results for these queries are the same. Max score for the first two queries was also the same: 10.429893. Max score for the last query was: 20.859785.

- "Lemur photos: The next [funny cat](http://www.knitemare.org/cats/index.php?type=all) fad?"
- "Here is some example code that handles both cat with, and cat without arguments. (and as an added bonus, handles failures in a way similar to standard cat)"
- http://www.knitemare.org/cats/
- "Stop piping cats!"
- "* emacs\r\n\r\n* nano\r\n\r\n* cat\r\n\r\n* ?"
- "Cats are good jumpers. \r\nI once saw a cat clear a fence.\r\nRichard Norton can teach Zimba I am confident."

- "masturbating cat..so coool"
- "Copy Cat has kittens."
- "I thought he had cats."
- "Because cats do funny things."

2- Has the size of the index become larger or smaller compared with the previous runs? Why do you think this might be?

As we can see below, now the size of index is 27 mb which is smaller than before.

```
mahta@LAPTOP-QM32K7CA MINGW64 ~/Desktop/second semester/Unstructured Data/Assignments/Assignment1A
$ curl -X GET "localhost:9200/_cat/indices?v&pretty"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   208  100   208    0     0  17333      0 --:--:-- --:--:-- --:--:-- 18909
health status index    uuid                   pri rep docs.count docs.deleted store.size pri.store.size
yellow open   comments FY0Xr1naQsWec30rBPwGyQ   1   1      61013            0       27mb           27mb
```

Stemming reduces the size of the index, since it reduces the number of separate terms indexed by "collapsing" multiple word forms into a single base word.