Towards improving ‘Search’

Author: Anindita Nath

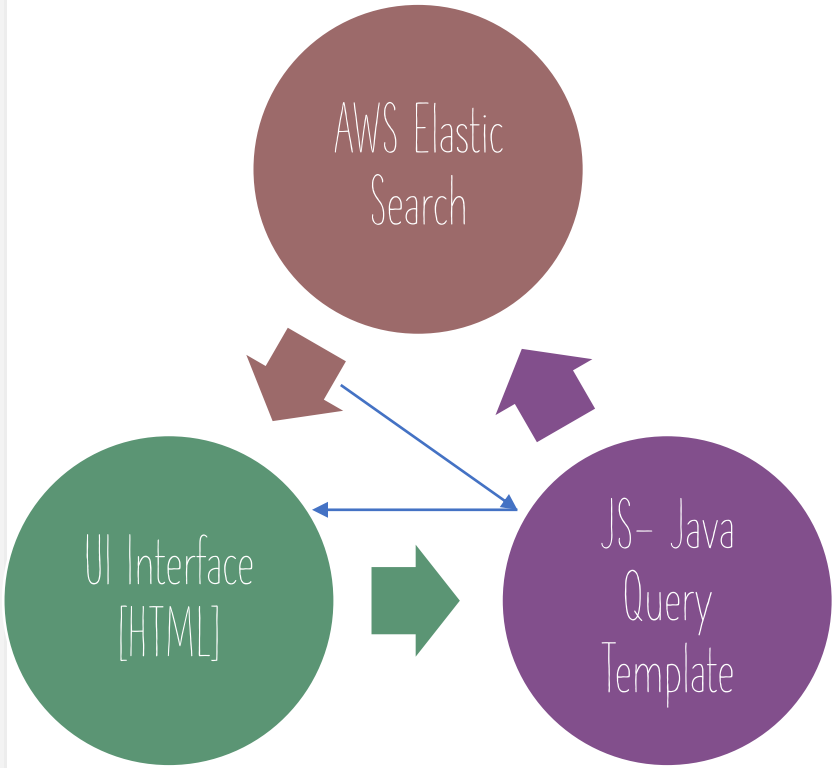
Dated: 08.09.21

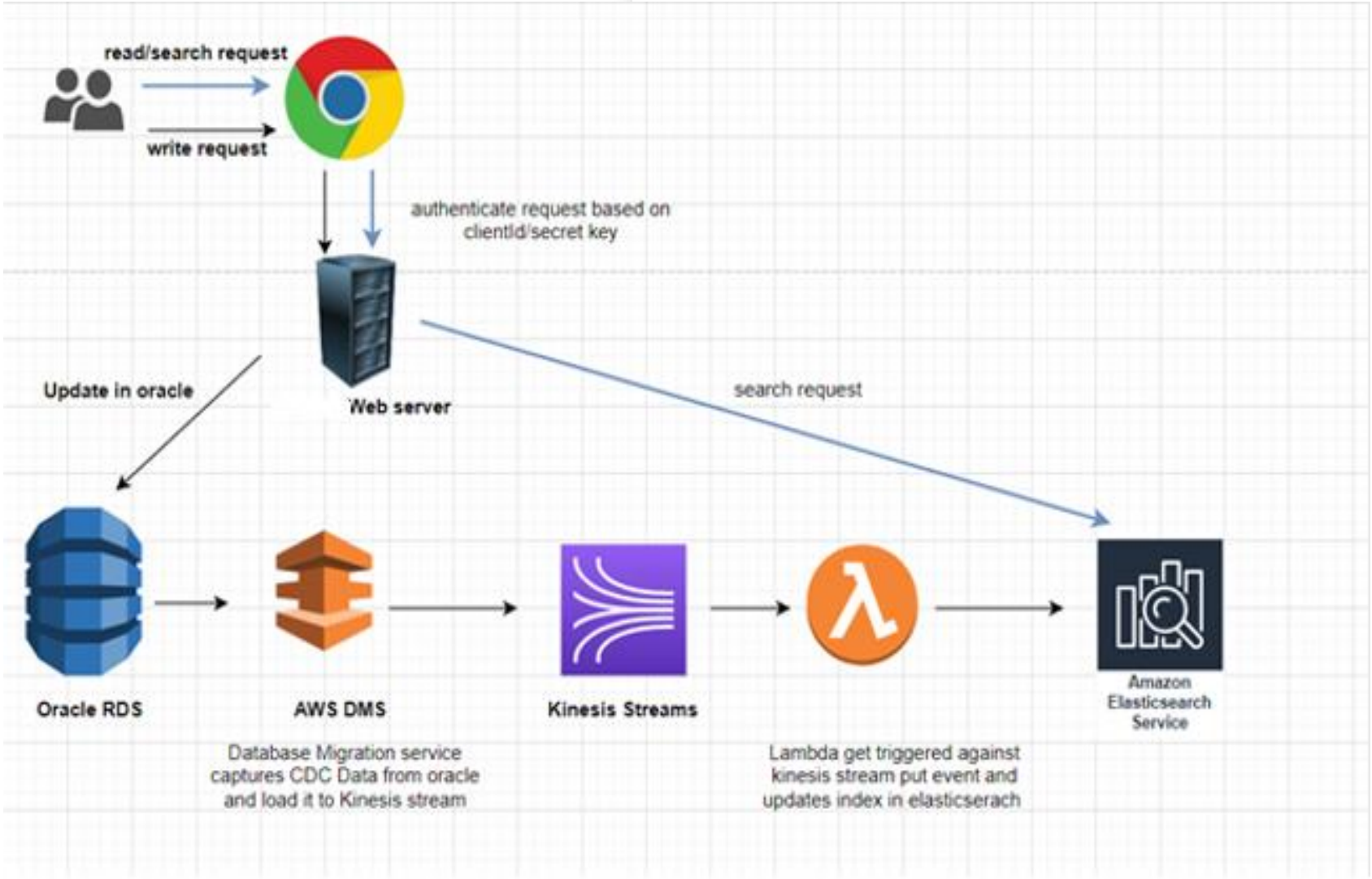
Data Search

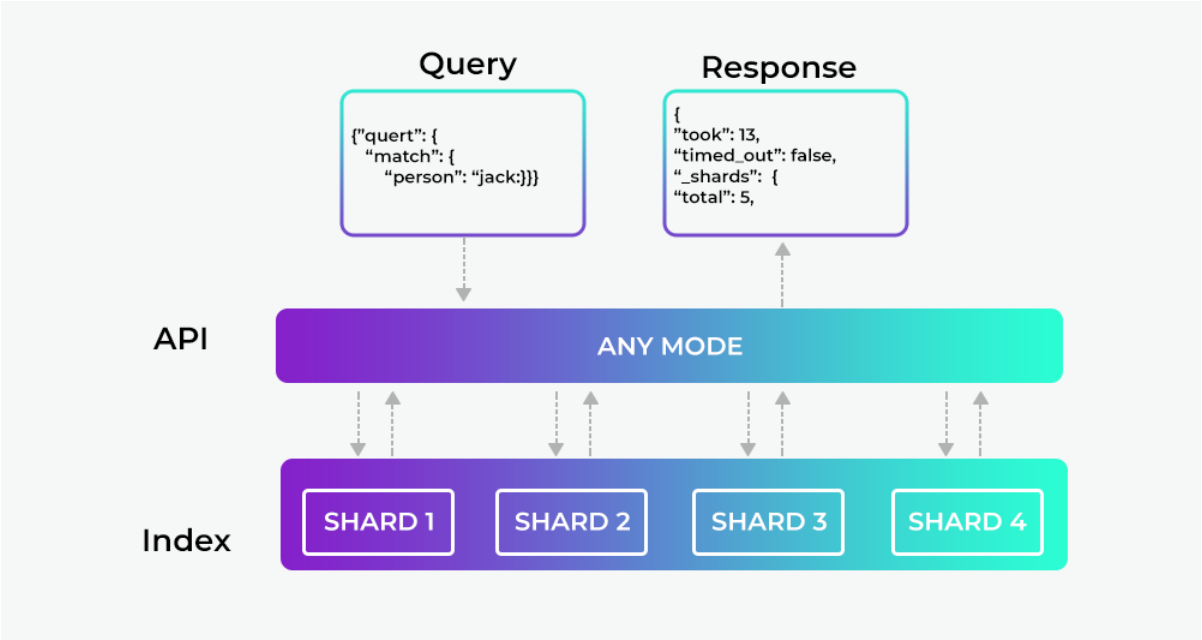
- Aim
  - Develop a fast and automated enterprise-wide search engine deployed in cloud architecture
- Scope
  - Augment Relational Database Search with Elastic Search.
  - Improve search, include semantic/NLU search
  - Searches only for meta data, raw data is still RDBMS
  - Eventually, replace RDBMS with ES and Graph DB search.
- Outcome
  - Developed prototype engine.
  - Improved certain functionalities.
  - Showed improvements through Kibana console

RDBMS	ES
- Slow on huge data sets - Slower fetching of search results through queries - Every field cannot be indexed - Updating rows to heavily indexed tables - lengthy and excruciating.	- Faster - NoSQL Distributed Database - Document-oriented search engine - Lucene Standard Analyzer - Store and retrieve data in JSON document form. - Schema-less

Search Engine Architecture (Conceptual)




read/search request
write request
authenticate request based on clientId/secret key
Update in oracle
Web server
search request
Oracle RDS
AWS DMS
Kinesis Streams
Amazon Elasticsearch Service
Database Migration service captures CDC Data from oracle and load it to Kinesis stream
Lambda get triggered against kinesis stream put event and updates index in elasticserach



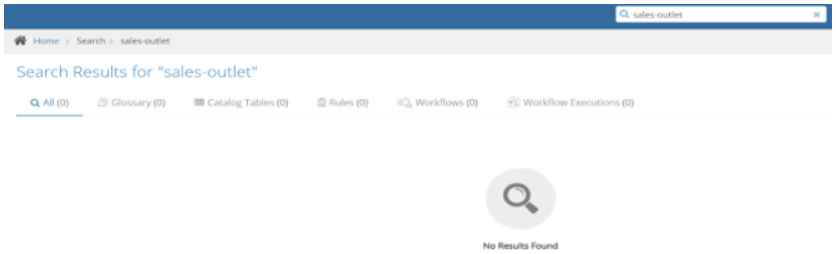
ES Search Query (In Action)

Use-Cases to improve

Problem

Spelling error, delimitation error

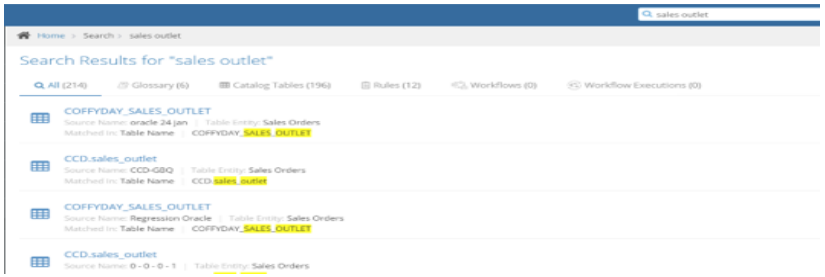
e.g., sales-outlet, sles outlet : both should match to sales_ outlet



Improve ranking, better matching

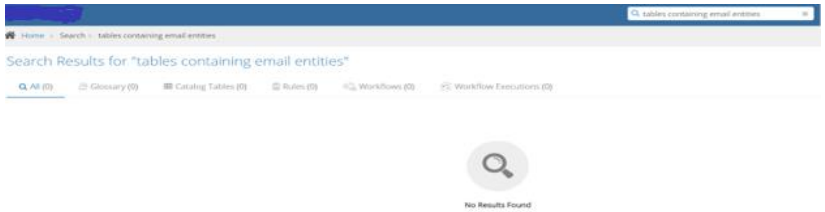
e.g., "CCD.sales_outlet" seems to be closer sales outlet.

Direct match to "sales_outlet" exists, does not show in top ranks



Semantic/NLU Search

e.g., search is for the index type "tables" and field type "email" in "tables"



Solution

Word-delimiter graph; fuzziness; "porter-stemmer"; n-gram tokenizer settings

most_fields

Combines scores form all documents matching the query, pushing relevant one first, analyzes each different form of query.

Text-embedding

- **Google's Universal Sentence Encoder**
- download, create the embedding model in TensorFlow
- **create the Elasticsearch index**
- Convert documents to vector representations : **doc_vector**
- Convert query to vector representations : **query_vector**
- Cosine similarity between doc_vector and query_vector for ranking

Changes in Elastic Search

```
PUT /cloned-ctables/_settings
 # index analyzer, ngram settings changed to improve query
#auto-completion
{
        "analysis" : {
         "analyzer" : {
         "ngram_tokenizer_analyzer" : {
            "filter" : [ "lowercase","stop"],
            "type" : "custom",
            "tokenizer" : "ngram_tokenizer"
          }
         },
         "tokenizer" : {
          "ngram_tokenizer" : {
           "type" : "nGram",
           "min_gram" : "3",
           "max_gram" : "20"
          }
         }

}


       }
```

```
PUT /cloned-ctables/_settings # search analyzer
{
        "analysis" : {
         "analyzer" : {
         "lowercase_space_analyzer" : {

           "tokenizer" : "standard",
            "type":"custom",
           "filter" : [ "lowercase" , "stop","porter_stem",
"my_custom_word_delimiter_graph_filter"]

          }
         },
          "filter": {
       "my_custom_word_delimiter_graph_filter": {
        "type": "word_delimiter",
       "split_on_case_change": true,
       "split_on_numerics": true,
        "stem_english_possessive": true
       }

       }

       }



       }
```

Use Cases Improved

```
"query": {
      "bool": {
        "must": [{
            "multi_match": {
              "query": "sales-outlet”,
                       # “sales outlet”,
                       #“sales?outlet”,
                       #“sles outlet”,
                       #“saling outlet”,
,               "fuzziness": "AUTO",
              "type": "most_fields",
              "fields": "*",
              "operator": "and"
            }
          },
```

```
"hits" : {  "total" : {    "value" : 1000,  "max_score" : 316.31873,
"hits" : [    {      "_index" : "ctables",      "_type" : "_doc",      "_id" : "60396361",      "_score" : 316.31873,
"_source" : {       "TABLE_NAME" : "SALES_OUTLET"      },……..}
```

Use Cases Improved [contd.]

```
"query": {
    "bool": {
      "must": [{
          "multi_match": {
              "query": “tables with email”,
            "query": “tables containing email”,
            "query": “tables email”,
            "query": “show me tables with email”,

              "fuzziness": "AUTO",
              "type": "most_fields",
              "fields": "*",
              "operator": "and"
          }
        },
"hits" : {  "total" : {  "value" : 10000,  "relation" : "gte"  },  "max_score" : 7.2697496,  "hits" : [
{  "_index" : "ctables",  "_type" : "_doc",  "_id" : "5319660",  "_score" : 7.2697496,
"_source" : {  "TABLE_COLUMNS" : [  {
},  {  "COLUMN_NAME" : "EMAIL" {  "COLUMN_NAME" : "CUST_EMAIL"  },
"COLUMN_NAME" : "FIRSTNAME"  },  {  "COLUMN_NAME" : "LASTNAME"  },
{  "COLUMN_NAME" : "RANK"  },  {  "COLUMN_NAME" : "CUST_FIRST_NAME"
}, "COLUMN_NAME" : "CUSTOMER_ID"  }  ]  },
```

Q & A time