Universidade Federal da Paraíba – Campus I
Centro de Informática
Departamento de Informática

Big Data: conceitos e aplicações

Laboratório 2: Modelo de Dados Orientado a Documentos

Aluna: Emmanuella Faustino Albuquerque

1) Criar cluster no MongoAtlas e fazer conexão com mongo shell ou Compass.

Setup do Ambiente

- ☐ Debian 11
- ☐ MongoDB Community Server (https://www.mongodb.com/docs/v4.4/mongo/)
- ☐ Instalação MongoDB 5.0

```
~$ wget -qO - https://www.mongodb.org/static/pgp/server-5.0.asc |
sudo apt-key add -

~$ echo "deb [ arch=amd64,arm64 ]
https://repo.mongodb.org/apt/ubuntu focal/mongodb-org/5.0
multiverse" | sudo tee
/etc/apt/sources.list.d/mongodb-org-5.0.list

~$ sudo apt update

~$ sudo apt install mongodb-org
```

2) Criar conta e cluster no MongoAtlas

Connect with the MongoDB Shell

- ☐ Database Access, create Username and Password (MongoDB Atlas)
- ☐ Connection string:
- ☐ mongosh "mongodb+srv://<uri>/myFirstDatabase" --apiVersion 1 --username manu
- ☐ Connection URL: mongodb+srv://<credentials>@<uri>/myFirstDatabase?appName=mongosh+1.6.0

3) Fazer o Download das amostras de dados

> https://github.com/ozlerhakan/mongodb-json-files

Comandos para importar dados em json e bson.

a) JSON

```
~$ mongoimport --db earth --collection countries --drop --file
countries.json

or

~$ mongoimport --uri
"mongodb+srv://<credentials>@<uri>/earth?appName=mongosh+1.6.0"
earth --collection countries --drop --file countries.json
```

b) BSON

```
~$ mongorestore --db twitter --collection tweet
twitter/tweets.bson

or

~$ mongorestore --uri
"mongodb+srv://<credentials>@<uri>/myFirstDatabase?appName=mongosh
+1.6.0" twitter --collection tweet dump/twitter/tweets.bson
```

4) Ler tutoriais do MongoDB sobre MapReduce e Sharding.

https://docs.mongodb.com/manual/core/map-reduce/
https://docs.mongodb.com/manual/sharding/

EXERCÍCIOS

Sharding

1. O que é Sharding no MongoDB?

```
Sharding é um método para distribuir dados entre várias máquinas.
O MongoDB usa sharding para dar suporte a implantações com
conjuntos de dados muito grandes e operações de alto rendimento.
```

2. Quais são os diferentes componentes necessários para implementar a Sharding?

Um cluster fragmentado do MongoDB, necessita de três componentes, o shard, o mongos e o config servers.

O componente shard, contém um subconjunto dos dados fragmentados (conjunto de réplicas).

O componente mongos, atua como um roteador de consulta, fornecendo uma interface entre os aplicativos cliente e o cluster fragmentado.

E o componente config servers, servidores de configuração, que armazenam metadados e definições de configuração para o cluster.

3. Explicar a arquitetura de Sharding no MongoDB?

(Sharding Architecture): O aspecto mais importante de um Sharded Cluster é que podemos adicionar qualquer número de shards. É por esse motivo que as client applications, não vão se conectar diretamente com os shards. Ao invés disso, é definido um tipo de processo de roteamento chamado Mongos, assim, o cliente pode se conectar ao Mongos e as rotas do Mongos requerem os shards corretos.

Como os dados são divididos de uma maneira específica em cada um dos shards, o Mongos precisa entender exatamente como os dados são distribuídos para realizar as consultas para o cliente no shard correto. Nós também podemos ter múltiplos processos Mongos de alta disponibilidade, ou para atender vários aplicativos ao mesmo tempo.

Os processos do Mongos vão utilizar metadados em torno das coleções que foram fragmentadas para descobrir exatamente para onde rotear as consultas. Os metadados da coleção são armazenados em servidores de configuração (config servers), que monitoram constantemente onde cada parte dos dados reside no cluster.
Então, o Mongos consulta frequentemente a configuração dos servidores de configuração, no caso de um pedaço de dado ser movido (isso ocorre, caso a distribuição de dados não seja uniforme - desproporcional)

MapReduce

MapReduce com mongoDB (primeiro comando)

Como primeiro exercício, você deve carregar os dados do reddit do link mencionado na sessão de configuração do ambiente. Com a ajuda do mapreduce, você precisa encontrar as 10 principais "lang" (languages) dos documentos no reddit.

```
Database: earth
Collection: countries

Document Example:
{
      "_id": {
            "$oid":"55a0f1d420a4d760b5fbdc16"
      },
      "Country Name": "North America",
      "Language": "en",
      "ISO": {
            "$numberInt":"0"
      }
}
```

4. Fornecer implementação da função MapReduce

```
// Acess MongoDB shell version v4.4.17

~$ mongo

~$ use earth

// Função de Mapeamento
function map() { emit( this.Language, 1 ); }

// Função de Redução: soma todos os valores com a mesma key
(Language)
function reduce(key, value) { return Array.sum(value); }

// Função MapReduce
let res = db.countries.mapReduce(map, reduce, { out: { inline: 1 }
});
```

```
// Função de Ordenação (Decrescente)
res.results.sort((a, b) => {
      return b.value - a.value;
});

// Função Filter: retorna as 10 principais linguagens
let top10 = res.results.filter((item, index)=>{
  return index >= 0 && index < 10 ;
})
```

5. Fornecer o resultado do comando de execução para executar o MapReduce

```
> resultado se encontra no arquivo
src/assets/json/languages.earth.json
```

6. Forneça os 10 primeiros registrados do resultado classificado. (dica: use sort no resultado retornado pelo MapReduce)

```
> top10
[
        {
                "_id" : "en",
                "value" : 300
        },
        {
                "_id" : "te",
                "value" : 284
        },
        {
                "_id" : "mr",
                "value" : 284
        },
        {
                "_id" : "ta",
                "value" : 284
        },
        {
                "_id" : "kn",
                "value" : 284
        },
        {
```

```
                "_id" : "ml",
                "value" : 284
        },
        {
                "_id" : "sk",
                "value" : 284
        },
        {
                "_id" : "or",
                "value" : 284
        },
        {
                "_id" : "hi",
                "value" : 284
        },
        {
                "_id" : "cs",
                "value" : 284
        }
]
```

MapReduce com mongoDB (hashtag query)

Para esta tarefa, você precisa baixar o conjunto de dados do Twitter no link mencionado na sessão de configuração do ambiente. Desta vez, você deve responder à pergunta "quais são as 10 principais hashtags usadas nos tweets fornecidos". Para responder isso, você precisa usar MapReduce. Você pode ver o esquema da coleção usando db.collection.findOne(). Ele imprimirá um registro com informações do esquema. Além disso, você pode usar uma função como this.hasOwnProperty('field_name') para verificar se existe um campo no registro. (se o campo não existir, você receberá um erro).

```
Database: tweets
Collection: tweet

Document Example: (esquema da coleção - db.collection.findOne())
{
   "_id":{
      "$oid":"5545e52c23877707ea64d540"
   },
```

```
   "created_at":"Sun May 03 06:15:01 +0000 2015",
   "id":{
      "$numberDouble":"5.9474696802852E+17"
   },
   "id_str":"594746968028516355",
   "text":"There were 18 goals in the #BPL on Saturday - a look at
where they landed... http://t.co/6I0mkuW2bt",
      ...
   "lang":"en"
}
```

7. Fornecer implementação da função MapReduce

```
// Acess MongoDB shell version v4.4.17

~$ mongo

~$ use twitter

// Função de Mapeamento
function map() {

     //  \s - whitespace ; \n,\r - new line
      var hashtag_array = this.text.split(/[\s\r\n]/)
).filter((word) => word.startsWith('#'));

      for (let hashtag of hashtag_array) {
            emit(hashtag, 1);
      }

}

// Função de Redução: soma todos os valores com a mesma key
(Hashtag)
function reduce(key, value) { return Array.sum(value); }

// Função MapReduce
let res = db.tweet.mapReduce(map, reduce, { out: { inline: 1 } });

// Função de Ordenação (Decrescente)
res.results.sort((a, b) => {
      return b.value - a.value;
```

```
});

// Função Filter: retorna às 10 Hashtags mais utilizadas.
let top10 = res.results.filter((item, index)=>{
  return index >= 0 && index < 10 ;
})
```

8. Fornecer o resultado do comando de execução para executar o MapReduce

```
> resultado se encontra no arquivo
src/assets/json/hashtag.twitter.json
```

9. Forneça os 10 primeiros registrados do resultado classificado. (dica: use sort no resultado retornado pelo MapReduce)

```
> top10
[
        {
                "_id" : "#FCBLive",
                "value" : 27
        },
        {
                "_id" : "#AngularJS",
                "value" : 26
        },
        {
                "_id" : "#nodejs",
                "value" : 21
        },
        {
                "_id" : "#EspanyolFCB",
                "value" : 18
        },
        {
                "_id" : "#LFC",
                "value" : 17
        },
        {
                "_id" : "#webinar",
                "value" : 16
        },
```

```
        {
                "_id" : "#CFCLive",
                "value" : 15
        },
        {
                "_id" : "#CFC",
                "value" : 15
        },
        {
                "_id" : "#RedBizUK",
                "value" : 15
        },
        {
                "_id" : "#javascript",
                "value" : 13
        }
]
```

Referências Bibliográficas

[1] Como conectar um banco MongoDB via terminal Shell. Disponível em: https://king.host/wiki/artigo/mongodb-via-terminal-shell/. Acesso em: 04 de outubro de 2022.

[2] How To Enable Copy Paste (Shared Clipboard) Between VirtualBox Host And Guest OS. Disponível em: https://www.dev2qa.com/how-to-enable-copy-paste-shared-clipboard-between-virtualbox-host-and-guest-os/. Acesso em: 04 de outubro de 2022.

[3] m103 Sharding Architecture. Disponível em: https://university.mongodb.com/videos/y/6cCL4-3gF8o. Acesso em: 04 de outubro de 2022.

[4] How To Install MongoDB 4.4 / 4.2 / 4.0 on Debian 10. Disponível em: https://www.itzgeek.com/post/how-to-install-mongodb-4-2-4-0-on-debian-10/. Acesso em: 04 de outubro de 2022.

[5] Instalar MongoDB 4.4 no Debian 10 Buster. Disponível em: https://www.vivaolinux.com.br/dica/Instalar-MongoDB-44-no-Debian-10-Buster. Acesso em: 04 de outubro de 2022.

[6] BigData: conceitos e aplicações. Disponível em: https://sig-arq.ufpb.br/arquivos/2022214028dd4746347326e2954e3923e/BigData_06_ModeloDocumentos.pdf. Acesso em: 05 de outubro de 2022.

[7] Couldn't connect to server 127.0.0.1:27017. Disponível em: https://stackoverflow.com/a/34835813. Acesso em: 05 de outubro de 2022.