Experiment report: Document DataBases

Done by student: Bylkova Kristina (伯汀娜), 1820249049

Course: Big Data Analysis Technology

Date: September 2024

Summary

Document databases are a type of NoSQL database that store data in the form of documents. Unlike traditional relational databases that store data in rows and columns, document databases use a flexible, schema-less model, allowing data to be stored in structures such as JSON, BSON, or XML. Each document can have a unique structure, making document databases ideal for handling unstructured or semi-structured data.

This report shows the MindMap of Document DB and the use of the Mongo DataBase. MongoDB is a document database designed for ease of application development and scaling. Its main key features are horizontal scalability, high availability, flexibility and high performance with rich query mechanism.

MindMap

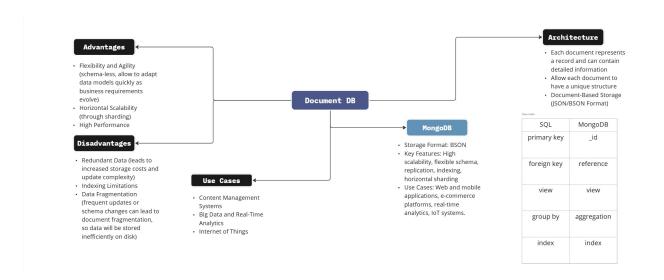


Figure 1: MindMap for DocumentDB

Installation

I installed and configured MongoDB and Compass via the guide from the team that performed.

Figure 2: Starting the MongoDB service



Figure 3: Result of starting

Also I configured logs for starting MongoDB.

```
C:\mongoDB\bin>mongod dbpath "C:\mongoDB\data\db" -logpath "C:\mongoDB\data\log" -install -serv iceName "MongoDB" ("t":("$\date":"2024-10-01T18:26:47.728Z"},"s":"F", "c":"CONTROL", "id":20574, "ctx":"thread 1","msg":"Error during global initialization","attr":{"error":{"code":38,"codeName":FileNotOpe n","errmsg":"logpath \"C:\\mongoDB\\data\\log\" should name a file, not a directory."}}}
C:\mongoDB\bin>net start MongoDB
The MongoDB Server (MongoDB) service is starting.
The MongoDB Server (MongoDB) service was started successfully.
```

Figure 4: Start via net start MongoDB

And connected via Compass.

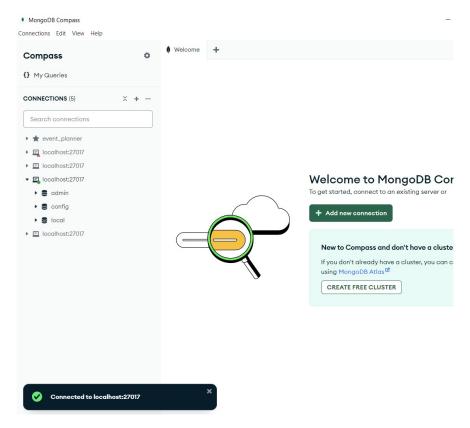


Figure 5: Connected in Compass

CRUD operations

I used basic commands in MongoDB shell here.

```
>>MONGOSH

> db.createCollection("sessions")

< { ok: 1 }

> db.sessions.insertOne({ "user": "Alice", "createdAt": new Date() })

< {
    acknowledged: true,
    insertedId: ObjectId('66fd0661fc5f7dd48be6943d')
}

> db.sessions.find().pretty()

< {
    _id: ObjectId('66fd0661fc5f7dd48be6943d'),
    user: 'Alice',
    createdAt: 2024-10-02T08:37:53.993Z
}

test>
```

Figure 6: Create and Insert One

```
>_MONGOSH

> db.sessions.deleteOne({ "user": "Alice Smith" })

< {
    acknowledged: true,
    deletedCount: 1
}

> db.sessions.find().pretty()

< {
    _id: ObjectId('66fd068bee1e1f5e519f2501'),
    user: 'Bob',
    createdAt: 2024-10-02T08:38:35.810Z,
    status: 'active'
}

{
    _id: ObjectId('66fd068bee1e1f5e519f2502'),
    user: 'Charlie',
    createdAt: 2024-10-02T08:38:35.810Z
}

test>
```

Figure 7: Delete One

Figure 8: Insert Many

```
>_MONGOSH

> db.sessions.deleteMany({})

< {
    acknowledged: true,
    deletedCount: 2
  }

> db.sessions.find().pretty()
```

Figure 9: Delete Many

```
>_MONGOSH
> db.sessions.updateOne(
     { "user": "Alice" }, // Query condition
     { $set: { "user": "Alice Smith" } } // Update content
    acknowledged: true,
> db.sessions.find().pretty()
   _id: ObjectId('66fd0661fc5f7dd48be6943d'),
   user: 'Alice Smith',
   createdAt: 2024-10-02T08:37:53.993Z
   _id: ObjectId('66fd068bee1e1f5e519f2501'),
   user: 'Bob',
   createdAt: 2024-10-02T08:38:35.810Z
    _id: ObjectId('66fd068bee1e1f5e519f2502'),
   user: 'Charlie',
   createdAt: 2024-10-02T08:38:35.810Z
test>
```

Figure 10: Update One

Code

Test 1: Generate and insert data

```
1 import random
2 from datetime import datetime, timedelta
3 from faker import Faker
4 from pymongo import MongoClient
6 fake = Faker()
7 fake.seed instance(42)
8 random.seed(42)
9 # MongoDB connection settings
10 mongo_client = MongoClient('mongodb://localhost:27017/')
db = mongo_client['test5']
12 collection = db['test5']
13 collection.delete_many({})
14
15
16
  def generate_data(num_records):
       categories = ['Category A', 'Category B', 'Category C', 'Category D']
17
       tags = ["tag1", "tag2", "tag3", "tag4", "tag5"]
18
       data = []
19
      for _ in range(num_records):
21
22
          record = {
23
               "_id": fake.uuid4(),
               "name": fake.name(),
24
               "description": fake.text(max_nb_chars=200),
25
               "category": random.choice(categories),
26
               "price": round(random.uniform(10.0, 500.0), 2),
27
               "location": {
28
                   "type": "Point",
30
                   "coordinates": [
31
                       round(random.uniform(-180, 180), 6), # Longitude
                       round(random.uniform(-90, 90), 6) # Latitude
32
                   ]
33
               "tags": [random.choice(tags) for _ in range(random.randint(1, 5))],
35
36
               "ratings": [
                   {
37
                       "userId": random.randint(100, 1000),
39
                       "rating": round(random.uniform(1.0, 5.0), 1),
                       "timestamp": datetime.strftime(
40
                           datetime.utcnow() - timedelta(days=random.randint(0, 365)),
41
                           '%Y-%m-%dT%H:%M:%SZ'
42
                   } for _ in range(random.randint(1, 5))
44
               ]
45
           }
46
47
           # data.append(record)
           # Bulk insert data
           collection.insert_one(record)
```

```
if _ % 1000 == 0:
    print(f"Inserted {_} data.")

4  # Generate 100,000 records and insert them into MongoDB

num_records = 100000
data_set = generate_data(num_records)

print(f"Successfully inserted {num_records} data into MongoDB.")
```

Test 2: Aggregation

```
1 import random
2 from datetime import datetime, timedelta
3 from faker import Faker
4 from pymongo import MongoClient
6 fake = Faker()
7 fake.seed_instance(42)
8 random.seed(42)
9 # MongoDB connection settings
10 mongo_client = MongoClient('mongodb://localhost:27017/')
db = mongo_client['test5']
collection = db['test5']
14 pipeline = [
15 {
16
          "$group": {
17
              "_id": "$category",
              "count": {"$sum": 1}
18
          }
19
     }
20
21 ]
23 result = list(collection.aggregate(pipeline))
24 for item in result:
print(f"Category: {item['_id']}, Count: {item['count']}")
```

Test 3: Calculate Average Price per Category

```
import random
from datetime import datetime, timedelta
from faker import Faker
from pymongo import MongoClient

fake = Faker()
fake.seed_instance(42)
random.seed(42)
# MongoDB connection settings
mongo_client = MongoClient('mongodb://localhost:27017/')
db = mongo_client['test5']
collection = db['test5']

pipeline = [
```

Test 4: Measure the time with no index

```
1 import random
2 from datetime import datetime, timedelta
3 from faker import Faker
4 from pymongo import MongoClient
6 fake = Faker()
7 fake.seed_instance(42)
8 random.seed(42)
9 # MongoDB connection settings
10 mongo_client = MongoClient('mongodb://localhost:27017/')
db = mongo_client['test5']
collection = db['test5']
14 search_point = {
      "type": "Point",
16
      "coordinates": [-50.781527, -32.869214]
17 }
18 near_query = {
     "location": {
19
          "$geoWithin": {
               "$centerSphere": [search_point['coordinates'], 1000 / 6378.1]
21
22
          }
     }
23
24 }
25 import time
26 start_time = time.time()
27 cursor = collection.find(near_query)
28 for doc in cursor:
      print(doc)
30 end_time = time.time()
31 print(f"measure the time with no index: {end_time - start_time:.4f}s")
33 collection.create_index([('location', '2dsphere')])
```

Test 5: Create index of text and measure the time after creating index

```
import random
from datetime import datetime, timedelta
from faker import Faker
from pymongo import MongoClient
```

```
6 fake = Faker()
7 fake.seed_instance(42)
8 random.seed(42)
9 # MongoDB connection settings
10 mongo_client = MongoClient('mongodb://localhost:27017/')
db = mongo_client['test5']
12 collection = db['test5']
14 # create index of text
collection.create_index([('description', 'text')])
17\ \mbox{\#} measure the time after creating index
18 import time
19 start_time = time.time()
20 search_query = {'$text': {'$search': 'administration'}}
cursor = collection.find(search_query)
22 for doc in cursor:
print(doc)
24 end_time = time.time()
25 print(f"measure the time after creating index: {end_time - start_time:.4f}s")
27 print("current Index: ")
28 for index_info in collection.list_indexes():
print(index_info)
```

Code Execution

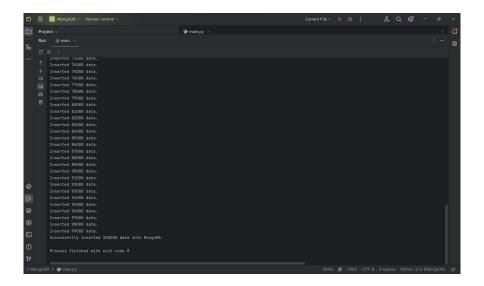


Figure 11: Test 1

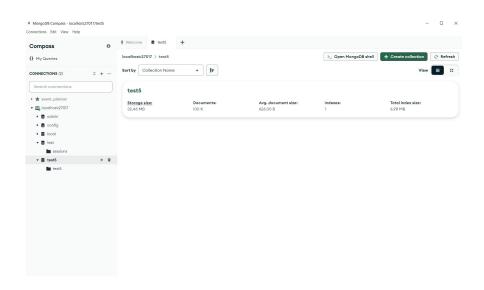


Figure 12: Test 1 (Compass)

Figure 13: Test 1 (MongoDB Shell)

```
Run test2 ×

C:\Users\Admin\MINE\Python_Rep\MongoDB\.venv\Scripts\python.exe C:\Users\Admin\MINE\Python_Rep\MongoDB\test2.py

C:\Users\Admin\MINE\Python_Rep\MongoDB\test2.py

Category: Category C, Count: 25018

Category: Category B, Count: 25063

Category: Category D, Count: 24970

Category: Category A, Count: 24949

Process finished with exit code 0
```

Figure 14: Test 2

Figure 15: Test 3

Figure 16: Test 4

```
**Comparison of the Comparison of the Comparison
```

Figure 17: Test 5