Senac

Fecomércio Sesc

Big Data

Prof. Marco Mialaret

Abril

2024

Big Data

Onde me encontrar:

https://www.linkedin.com/in/marco-mialaret-junior/

e

https://github.com/MatmJr

NoSQL

Big Data

Historicamente, sistemas de gerenciamento de bancos de dados relacionais dominavam o processamento de dados, necessitando de dados estruturados em tabelas precisas. Com o aumento dos volumes de dados e complexidade das relações, esses sistemas tornaram-se menos eficientes.

Big Data

Em resposta, surgiram os bancos de dados NoSQL e NewSQL, projetados para atender as exigências de armazenamento e processamento de grandes volumes de dados, muitas vezes distribuídos globalmente em data centers. Atualmente, há mais de 8 milhões de centros de dados pelo mundo.

Big Data

NoSQL, inicialmente um acrônimo para "Não SQL", evoluiu para significar "Não Apenas SQL" devido à integração do SQL em plataformas de grandes dados como Hadoop e Spark SQL. Esses bancos de dados são ideais para gerenciar dados não estruturados, como imagens, vídeos e textos, e dados semi-estruturados como JSON e XML.

Big Data

Dados semi-estruturados contêm metadados que fornecem estrutura adicional, como os encontrados nos vídeos do YouTube, que incluem informações como autor, título e tags, tornando-os mais organizados e acessíveis.

Tipos de NoSQL

Big Data

NoSQL chave-valor

Os bancos de dados NoSQL do tipo chave-valor, semelhantes a dicionários em Python, armazenam pares de chave-valor e são otimizadas para sistemas distribuídos e processamento de grandes volumes de dados. Para garantir a confiabilidade, esses bancos costumam replicar dados em múltiplos nós do cluster.

Big Data

Alguns, como o Redis, são implementadas em memória para melhorar o desempenho, enquanto outros, como o HBase, armazenam dados em disco e operam sobre o sistema de arquivos distribuídos HDFS do Hadoop.

Big Data

Outros bancos de dados chave-valor populares incluem o Amazon DynamoDB, o Google Cloud Datastore e o Couchbase. Tanto o DynamoDB quanto o Couchbase são bancos de dados multi-modelo que também suportam documentos, e o HBase é orientado a colunas.

Big Data

Bancos de dados de Documentos

Os bancos de dados de documentos NoSQL, como o MongoDB, são especializados em armazenar dados semi-estruturados em formatos como JSON ou XML. Eles utilizam índices para melhorar a eficiência na localização e manipulação de documentos, especialmente úteis para grandes volumes de dados, como os gerados por dispositivos IoT ou redes sociais como o Twitter.

Big Data

O MongoDB é notável por sua capacidade de armazenar "grandes" volumes de dados. Outros bancos de dados de documentos incluem Amazon DynamoDB, que também suporta a modelagem chave-valor, Microsoft Azure Cosmos DB e Apache CouchDB. Essas bases permitem armazenar dados diretamente de APIs e facilitam análises subsequentes usando ferramentas como pandas e Folium para visualização de dados.

Big Data

Banco de dados colunar

Em um banco de dados relacional, uma operação comum de consulta é obter o valor de uma coluna específica para cada linha. Como os dados são organizados em linhas, uma consulta que seleciona uma coluna específica pode ter desempenho ruim. O sistema do banco de dados precisa obter cada linha correspondente, localizar a coluna requerida e descartar o restante das informações da linha.

Big Data

Um banco de dados colunar, também conhecido como banco de dados orientado a colunas, é semelhante a um banco de dados relacional, mas armazena dados estruturados em colunas em vez de linhas. Como todos os elementos de uma coluna são armazenados juntos, selecionar todos os dados de uma coluna específica é mais eficiente.

Big Data

Os elementos em cada coluna são mantidos na ordem das linhas, de modo que o valor em um determinado índice em cada coluna pertence à mesma linha. Bancos de dados colunares populares incluem MariaDB ColumnStore e HBase.

Big Data

Bancos de dados de grafos

Os bancos de dados de grafos NoSQL modelam relações entre objetos, chamados de nós (ou vértices), e as relações são denominadas arestas, que são direcionais. Por exemplo, uma aresta que representa um voo de companhia aérea aponta da cidade de origem para a cidade de destino, mas não o inverso. Esses bancos de dados armazenam nós, arestas e seus atributos.

Big Data

Se você usa redes sociais como Instagram, Snapchat, Twitter e Facebook, pode considerar seu "grafo social", que inclui as pessoas que você conhece (nós) e as relações entre elas (arestas). Cada pessoa tem seu próprio grafo social, e estes estão interconectados. O famoso problema dos "seis graus de separação" sugere que quaisquer duas pessoas no mundo estão conectadas por no máximo seis arestas no grafo social mundial.

Big Data

Os algoritmos do Facebook utilizam os grafos sociais de seus bilhões de usuários ativos mensalmente para determinar quais histórias aparecerão no feed de notícias de cada usuário, baseando-se em interesses pessoais e conexões.

Big Data

Muitas empresas utilizam técnicas semelhantes para criar motores de recomendação. Por exemplo, a Amazon usa um grafo de usuários e produtos para sugerir produtos comparáveis que outras pessoas navegaram antes de fazer uma compra. A Netflix utiliza um grafo de usuários e filmes que gostaram para sugerir filmes que possam ser do seu interesse.

Big Data

Um dos bancos de dados de grafos mais populares é o Neo4j. Muitos casos de uso reais para bancos de dados de grafos estão disponíveis em:

https://neo4j.com/graphgists/

Na maioria dos casos, são mostrados diagramas de grafos amostrais produzidos pelo Neo4j, que visualizam as relações entre os nós do grafo. Confira também o livro gratuito em PDF da Neo4j, "Graph Databases".

Big Data

Bancos de dados NewSQL

As vantagens dos bancos de dados relacionais incluem sua segurança e suporte a transações. Em particular, os bancos de dados relacionais tipicamente usam transações ACID (Atomicidade, Consistência, Isolamento, Durabilidade):

Big Data

- **Atomicidade** garante que o banco de dados só é modificado se todas as etapas de uma transação forem bem-sucedidas. Por exemplo, se você for a um caixa eletrônico para sacar $100, esse dinheiro não será retirado da sua conta a menos que você tenha dinheiro suficiente para cobrir o saque e haja dinheiro suficiente no caixa eletrônico para atender sua solicitação.

Big Data

- **Consistência** garante que o estado do banco de dados seja sempre válido. No exemplo do saque, o novo saldo da sua conta após a transação refletirá exatamente o que você sacou da sua conta (e possíveis taxas do caixa eletrônico).

Big Data

- **Isolamento** garante que transações concorrentes ocorram como se fossem realizadas sequencialmente. Por exemplo, se duas pessoas compartilham uma conta bancária conjunta e ambas tentam sacar dinheiro ao mesmo tempo em caixas eletrônicos separados, uma transação deve esperar até que a outra seja concluída.

- **Durabilidade** garante que as alterações no banco de dados sobrevivam mesmo a falhas de hardware.

Big Data

Ao pesquisar benefícios e desvantagens dos bancos de dados NoSQL, você verá que, geralmente, eles não oferecem suporte ACID. Os tipos de aplicações que usam bancos de dados NoSQL tipicamente não requerem as garantias que os bancos de dados compatíveis com ACID proporcionam.

Big Data

Muitos bancos de dados NoSQL aderem ao modelo BASE (Disponibilidade Básica, Estado Suave, Consistência Eventual), que foca mais na disponibilidade do banco de dados. Enquanto os bancos de dados ACID garantem consistência quando você escreve no banco de dados, os bancos de dados BASE proporcionam consistência em algum momento posterior no tempo.

Big Data

Bancos de dados NewSQL combinam os benefícios de ambos os bancos de dados relacionais e NoSQL para tarefas de processamento de grandes dados. Alguns bancos de dados NewSQL populares incluem VoltDB, MemSQL, Apache Ignite e Google Spanner.

Configurando o MongoDB

Big Data

1 - Baixar o Docker desktop em
https://www.docker.com/products/docker-desktop/

Big Data

2 - Obter a imagem do Mongo que será o molde para criarmos nossos containers. Para isso, executamos o comando abaixo.

```
> docker pull mongo
```

Big Data

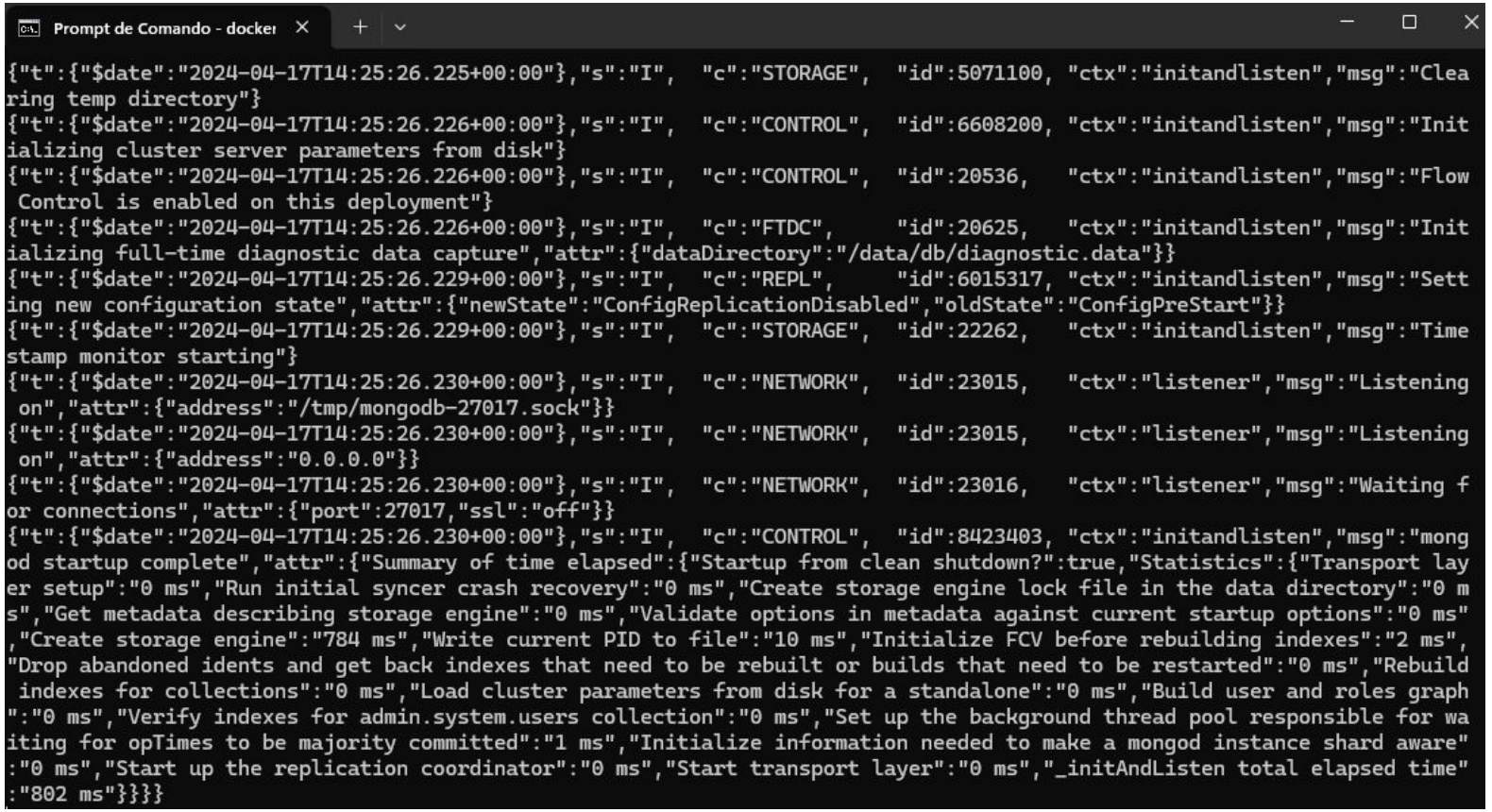
```
C:\Users\Marco>docker pull mongo
Using default tag: latest
latest: Pulling from library/mongo
3c645031de29: Pull complete
bfa196f67a92: Pull complete
2eb794c1d8eb: Pull complete
3a415fad0b1a: Pull complete
eb0ca0d8db42: Pull complete
7341c0351e26: Pull complete
11062490c406: Pull complete
e953fad04d18: Pull complete
Digest: sha256:a70130d7c7ad8dea76a48ac27ddebede0bb24122dd38ad83c229afea03f0f279
Status: Downloaded newer image for mongo:latest
docker.io/library/mongo:latest

What's Next?
  View a summary of image vulnerabilities and recommendations → docker scout quickview mongo
```

Big Data

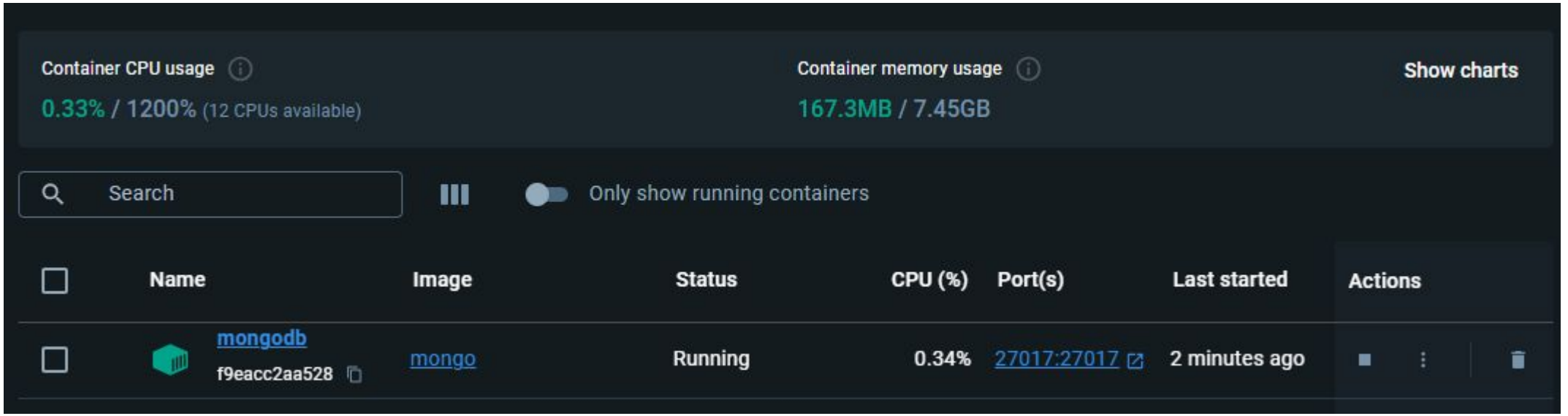
3 - Para executar esta imagem você pode usar a linha abaixo.

```
> docker run --name mongodb -p 27017:27017 -e
MONGO_INITDB_ROOT_USERNAME=root -e
MONGO_INITDB_ROOT_PASSWORD=root mongo
```

Big Data

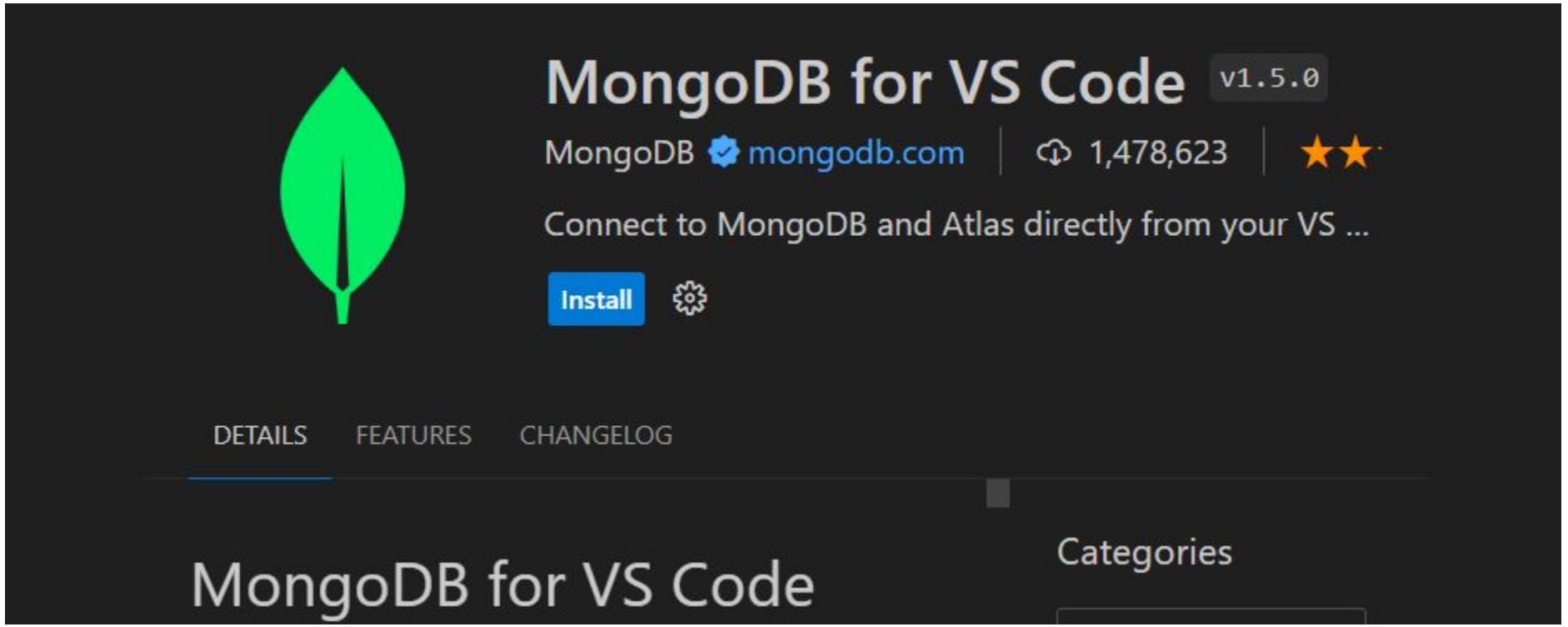

Prompt de Comando - docker
{"t":{"$date":"2024-04-17T14:25:26.225+00:00"},"s":"I", "c":"STORAGE", "id":5071100, "ctx":"initandlisten","msg":"Clearing temp directory"}
{"t":{"$date":"2024-04-17T14:25:26.226+00:00"},"s":"I", "c":"CONTROL", "id":6608200, "ctx":"initandlisten","msg":"Initializing cluster server parameters from disk"}
{"t":{"$date":"2024-04-17T14:25:26.226+00:00"},"s":"I", "c":"CONTROL", "id":20536, "ctx":"initandlisten","msg":"Flow Control is enabled on this deployment"}
{"t":{"$date":"2024-04-17T14:25:26.226+00:00"},"s":"I", "c":"FTDC", "id":20625, "ctx":"initandlisten","msg":"Initializing full-time diagnostic data capture","attr":{"dataDirectory":"/data/db/diagnostic.data"}}
{"t":{"$date":"2024-04-17T14:25:26.229+00:00"},"s":"I", "c":"REPL", "id":6015317, "ctx":"initandlisten","msg":"Setting new configuration state","attr":{"newState":"ConfigReplicationDisabled","oldState":"ConfigPreStart"}}
{"t":{"$date":"2024-04-17T14:25:26.229+00:00"},"s":"I", "c":"STORAGE", "id":22262, "ctx":"initandlisten","msg":"Timestamp monitor starting"}
{"t":{"$date":"2024-04-17T14:25:26.230+00:00"},"s":"I", "c":"NETWORK", "id":23015, "ctx":"listener","msg":"Listening on","attr":{"address":"/tmp/mongodb-27017.sock"}}
{"t":{"$date":"2024-04-17T14:25:26.230+00:00"},"s":"I", "c":"NETWORK", "id":23015, "ctx":"listener","msg":"Listening on","attr":{"address":"0.0.0.0"}}
{"t":{"$date":"2024-04-17T14:25:26.230+00:00"},"s":"I", "c":"NETWORK", "id":23016, "ctx":"listener","msg":"Waiting for connections","attr":{"port":27017,"ssl":"off"}}
{"t":{"$date":"2024-04-17T14:25:26.230+00:00"},"s":"I", "c":"CONTROL", "id":8423403, "ctx":"initandlisten","msg":"mongod startup complete","attr":{"Summary of time elapsed":{"Startup from clean shutdown?":true,"Statistics":{"Transport layer setup":"0 ms","Run initial syncer crash recovery":"0 ms","Create storage engine lock file in the data directory":"0 ms","Get metadata describing storage engine":"0 ms","Validate options in metadata against current startup options":"0 ms","Create storage engine":"784 ms","Write current PID to file":"10 ms","Initialize FCV before rebuilding indexes":"2 ms","Drop abandoned idents and get back indexes that need to be rebuilt or builds that need to be restarted":"0 ms","Rebuild indexes for collections":"0 ms","Load cluster parameters from disk for a standalone":"0 ms","Build user and roles graph":"0 ms","Verify indexes for admin.system.users collection":"0 ms","Set up the background thread pool responsible for waiting for opTimes to be majority committed":"1 ms","Initialize information needed to make a mongod instance shard aware":"0 ms","Start up the replication coordinator":"0 ms","Start transport layer":"0 ms","_initAndListen total elapsed time":"802 ms"}}}}

Big Data


Container CPU usage
0.33% / 1200% (12 CPUs available)
Container memory usage
167.3MB / 7.45GB
Show charts
Search
Only show running containers
Name
Image
Status
CPU (%)
Port(s)
Last started
Actions
mongodb
f9eacc2aa528
mongo
Running
0.34%
27017:27017
2 minutes ago

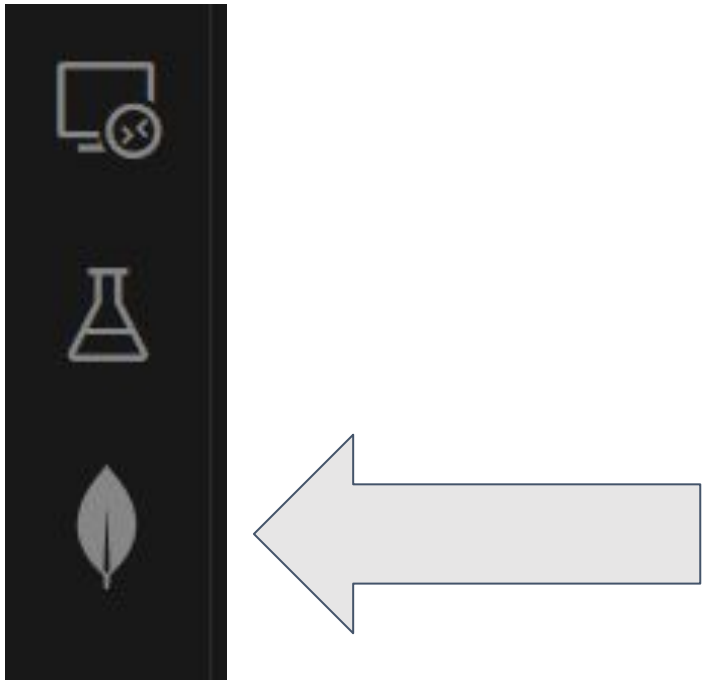
Big Data

4 - Instale a extensão MongoDB for VS Code.

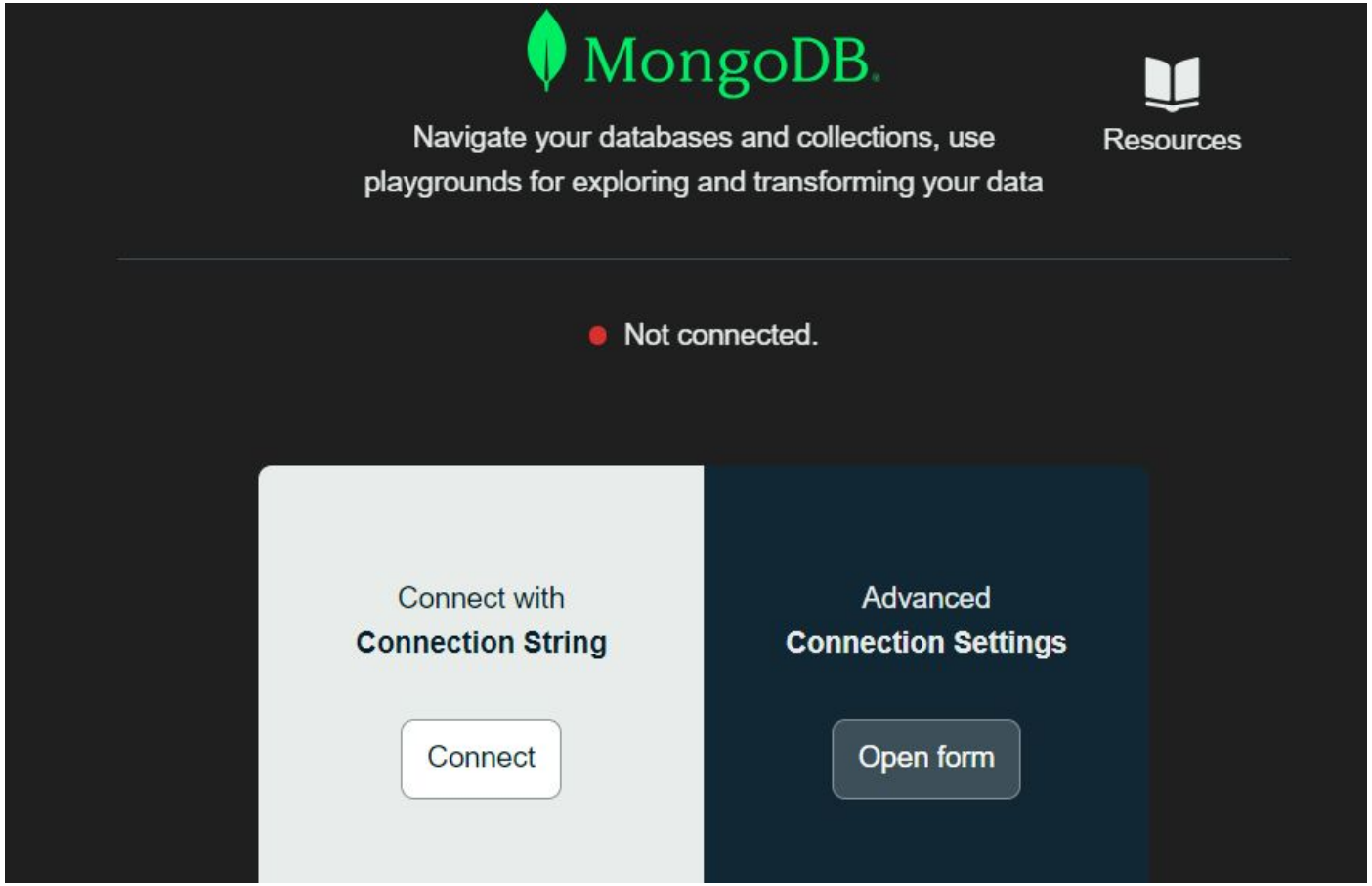


Big Data

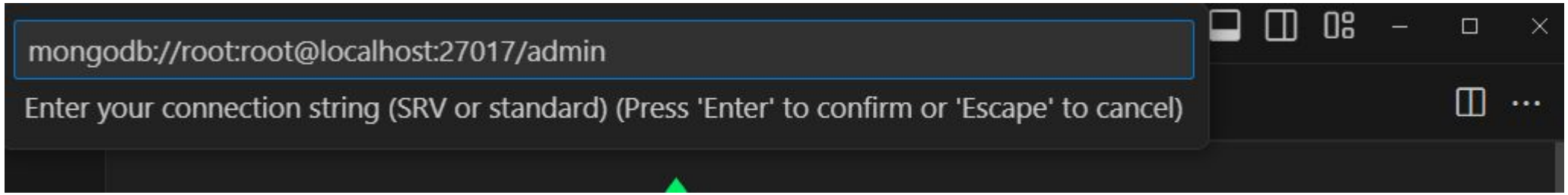
5 - Acesse a extensão (uma folha na barra lateral), clique em Add Connection e use as configurações:

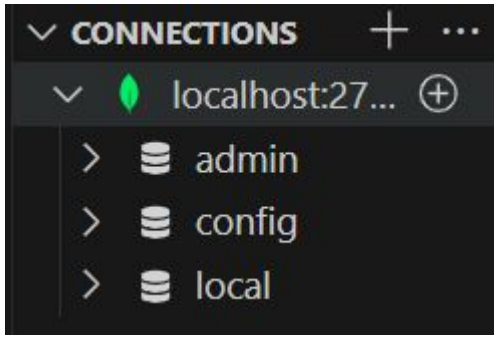


Big Data


MongoDB.
Navigate your databases and collections, use playgrounds for exploring and transforming your data
Resources
Not connected.
Connect with
Connection String
Connect
Advanced
Connection Settings
Open form

Big Data


mongodb://root:root@localhost:27017/admin
Enter your connection string (SRV or standard) (Press 'Enter' to confirm or 'Escape' to cancel)


CONNECTIONS
localhost:27...
admin
config
local

Big Data

Mais detalhes:

https://balta.io/blog/mongodb-docker

Acessando o banco com python

Big Data

Criar um venv com o requirements.txt da semana 07. Acesse o diretório no qual o requirements se encontra e execute o comando

```
python -m venv venv && .\venv\Scripts\activate
&& pip install -r requirements.txt
```

Big Data

```python
from pymongo import MongoClient

# Crie uma conexão com o MongoDB (ajuste o host e a porta conforme
# necessário)
client = MongoClient("mongodb://root:root@localhost:27017/admin")
```

✓ 2.7s Python

Big Data

```python
# Seleciona o banco de dados
db = client["aula07"]

# Seleciona a coleção
collection = db["cotacoes"]
```

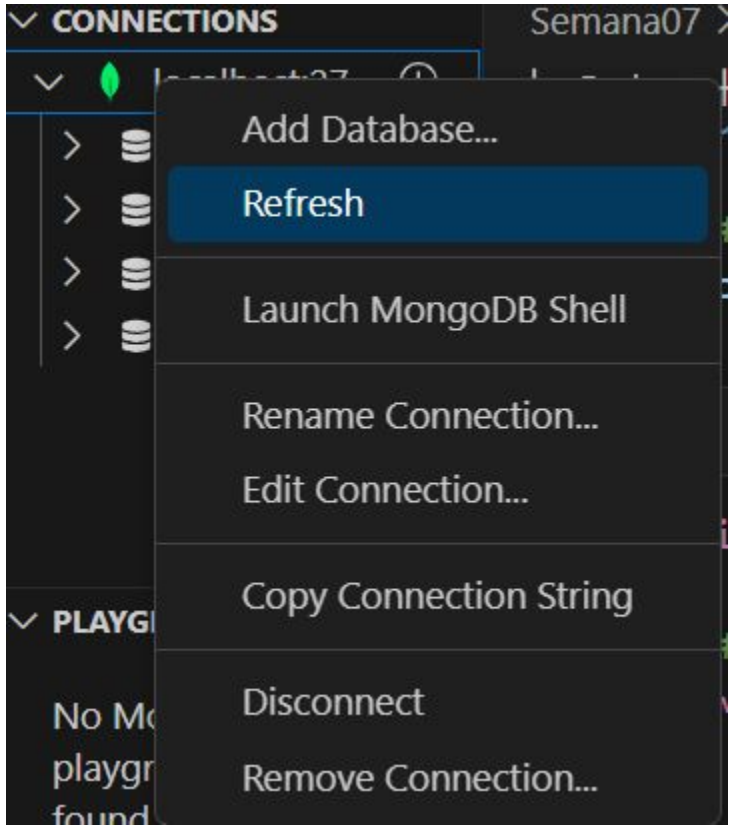
✓ 0.0s Python

Big Data

```python
import json

# Carrega os dados do arquivo JSON
with open("cotacoes.json", "r") as file:
    data = json.load(file)
```

✓ 0.0s Python

Big Data


CONNECTIONS
Semana07
Add Database...
Refresh
Launch MongoDB Shell
Rename Connection...
Edit Connection...
Copy Connection String
Disconnect
Remove Connection...

Big Data


CONNECTIONS
localhost:27017...
admin
aula07
cotacoes
Documents 1
"661fe1414d0...
Schema
Indexes
{} aula07.cotacoes:{"$oid":"661fe1414d0d69f8d2e883c1"}.json > ...
{
"_id": {
"$oid": "661fe1414d0d69f8d2e883c1"
},
"USDBRL=X": {
"Open": {
"2023-04-17": 4.908699989318848,
"2023-04-18": 4.939799785614014,
"2023-04-19": 4.984300136566162,
"2023-04-20": 5.0742998123168945,

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com