MongoDB

21 November 2025 09:09

- **Wired Tiger** is the storage engine in MongoDB

- it manager all storage.
- It makes mongoDB faster.

Indexes

In MongoDB, **indexes** are data structures that improve the speed and efficiency of queries
MongoDB uses indexes to quickly locate the data it needs.

Exercise :

From file Indexes-Operations file.txt run commands.

- How to check if index is working ?
  - first create the index
  - run any query that includes that index
  - use .explain() method on the last run query to see a stage called IXSCAN means indexscan inside the query planner.
  - another key index name and many more things.

Compound Index

- Using multiple keys together.

Multikey Index :

- If we create an index on array, it becomes multikey index.

Text Index :

- If we want to search based on any text, we can add test index.

This index is based on weight.

Ex. It'll calculate frequency of that word appearing .

Hashed Index :

Ex. The name is bond, James bond.

Just make a hash key for this sentence, means no other sentence can have same value for this,
so to search this sentence we can use that hashkey to directly get it.

Connect MongoDB from python :

Install jupyter in pycharm.
Install pymongo in pycharm.
Copy paste pyMongo.ipynb file that has code to connect to mongoDB into our project folder.
Open terminal in pychar, type : jupyter notebook.
This will open a web page of jupyter notebook.
Run the commands in that notebook.

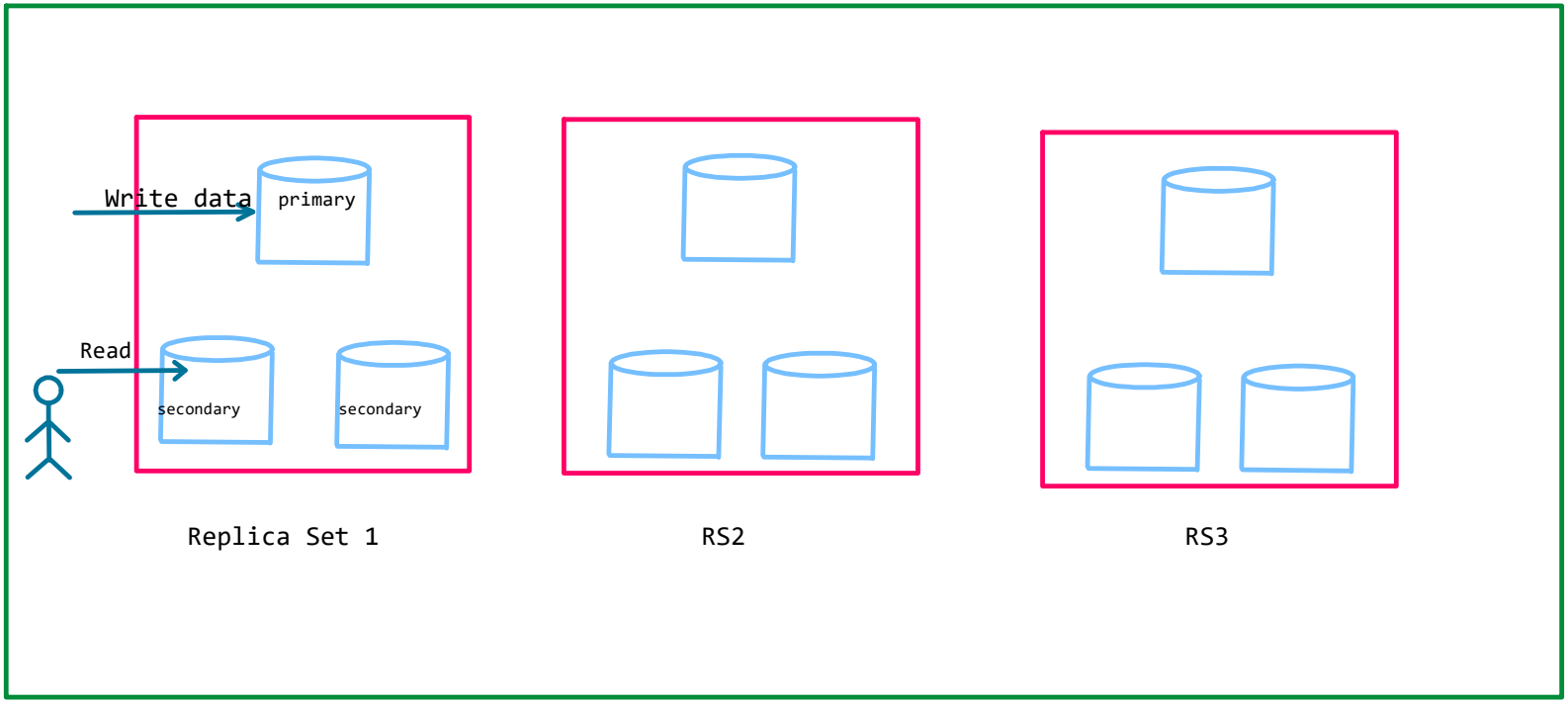
Sharding :
Sharding is when we store the data in distributed manner, and it is distributed in replica sets.

Replica set in mongoDB :

- The data is always written into primary machine.
- The data is replicated into secondary machines.
- Clients can read from secondary machines.

- This is done for fault tolerance.

- Faster retrievals.
- Availability.
- Consistency.



Aggregation in MongoDB :

Create a collection named pizza.
Run commands from aggregation.txt

Go on Atlas, load the Sample_training dataset into the cluster.

Graph Databases:

In MongoDB, **graphs** are used to model and query highly connected data—data where relationships between documents matter as much as the documents themselves.

Social media companies use graph APIs to show their data.
Ex. LinkedIn uses it for showing connections as 1st, 2nd, 3rd connection.
Ex. Instagram use it for recommending mutual connections.

Graph Components

Node :

- it's any obejct/person in a graph.

Labels :

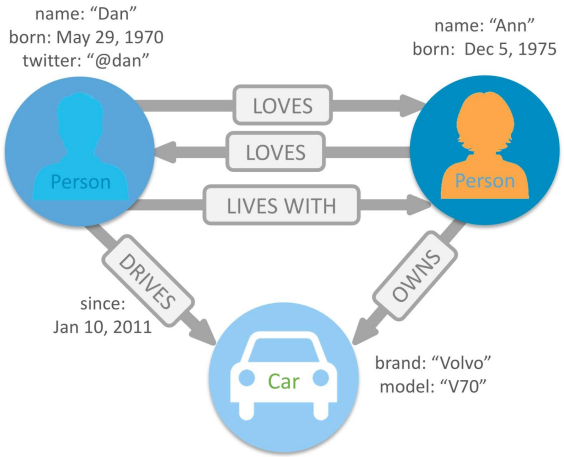
- Nodes can have labels ex. Car / employee /manager

Relationships :

- Relationships connects the nodes.
- Shown using arrows
- These are directional.
- Directions changed means meaning changed.

Properties :

- Properties can define nodes and relations.
- Ex. This employee is driving car since 2 years.
- Year of Birth



What are people using Neo4j for?

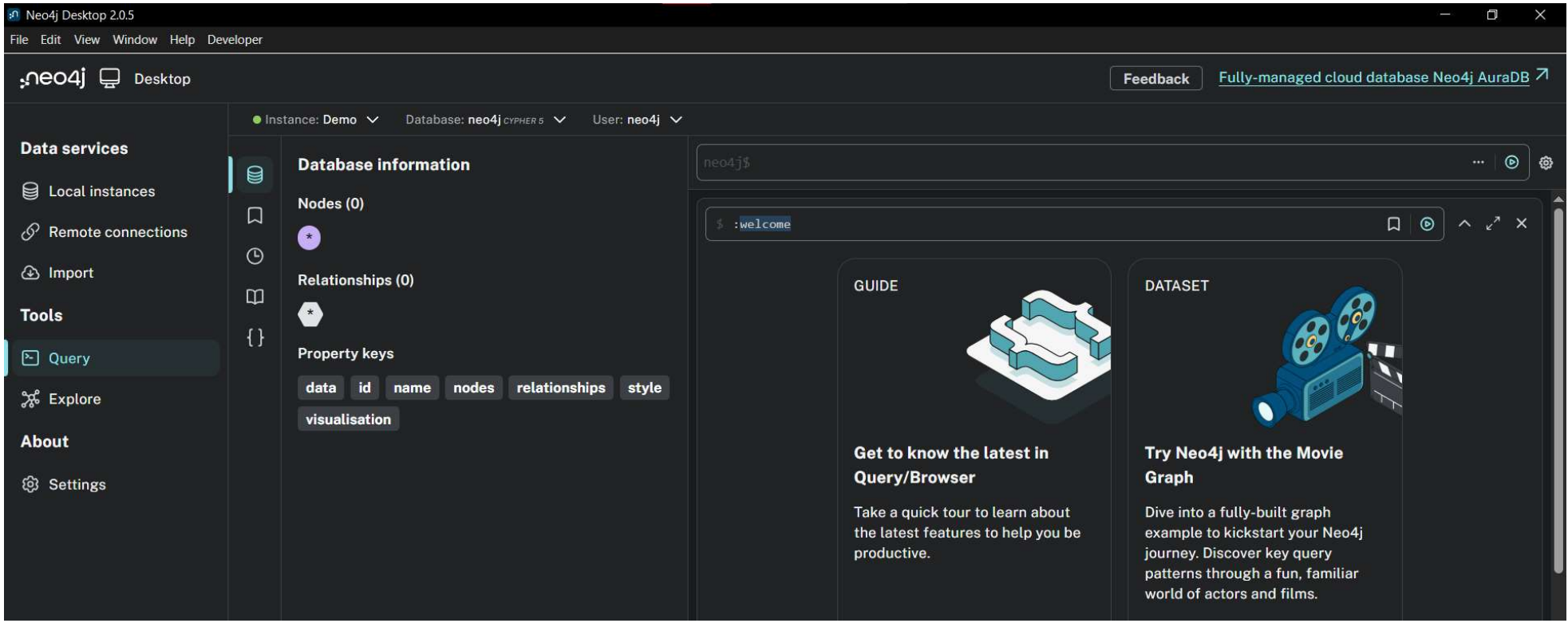
- Realtime promotion recommendation.
- Realtime pricing engine

- Realtime Handling of package routing
- WhiteBoard friendlyness
    - Means we can draw a graph and convert it into Neo4j graph

Open Neo4j, create instance.
Give Name, password
Click on connect -> query.

write create query at neo4j$ placeholder above $:welcome
From neo4j things.txt file

Run button is there at the end of query box.



Click on any labels from database information sections.
We'll see the graph, right click on the graph and select show all relations.
Explore……..

AuraDB is a cloud database for Neo4j

MapReduce Streaming.

MapReduce Streaming is a **utility included with Hadoop** that allows you to create MapReduce jobs using **standard input (stdin)** and **standard output (stdout).**
This means:

- Your **mapper** reads input lines from stdin
- Your **mapper** writes key/value pairs to stdout
- Your **reducer** reads mapper output from stdin
- Your **reducer** writes final results to stdout

You can write mapper/reducer in:

- Python
- Bash
- Perl
- Ruby
- C/C++
- Any language that can read from stdin and write to stdout.

Why MapReduce Streaming Is Useful

- Easy to prototype MapReduce jobs
- Allows using any language you're comfortable with
- Great for text processing
- Keeps the Hadoop cluster usage but simplifies development

https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

Activity

Topic : traffic sensor data processing .

WE have to come up with data sources, storage system,
design how the analysis will work ? In half an hour.

Draw a diagram on paper. With all these details.