Technologies for Big Data Management

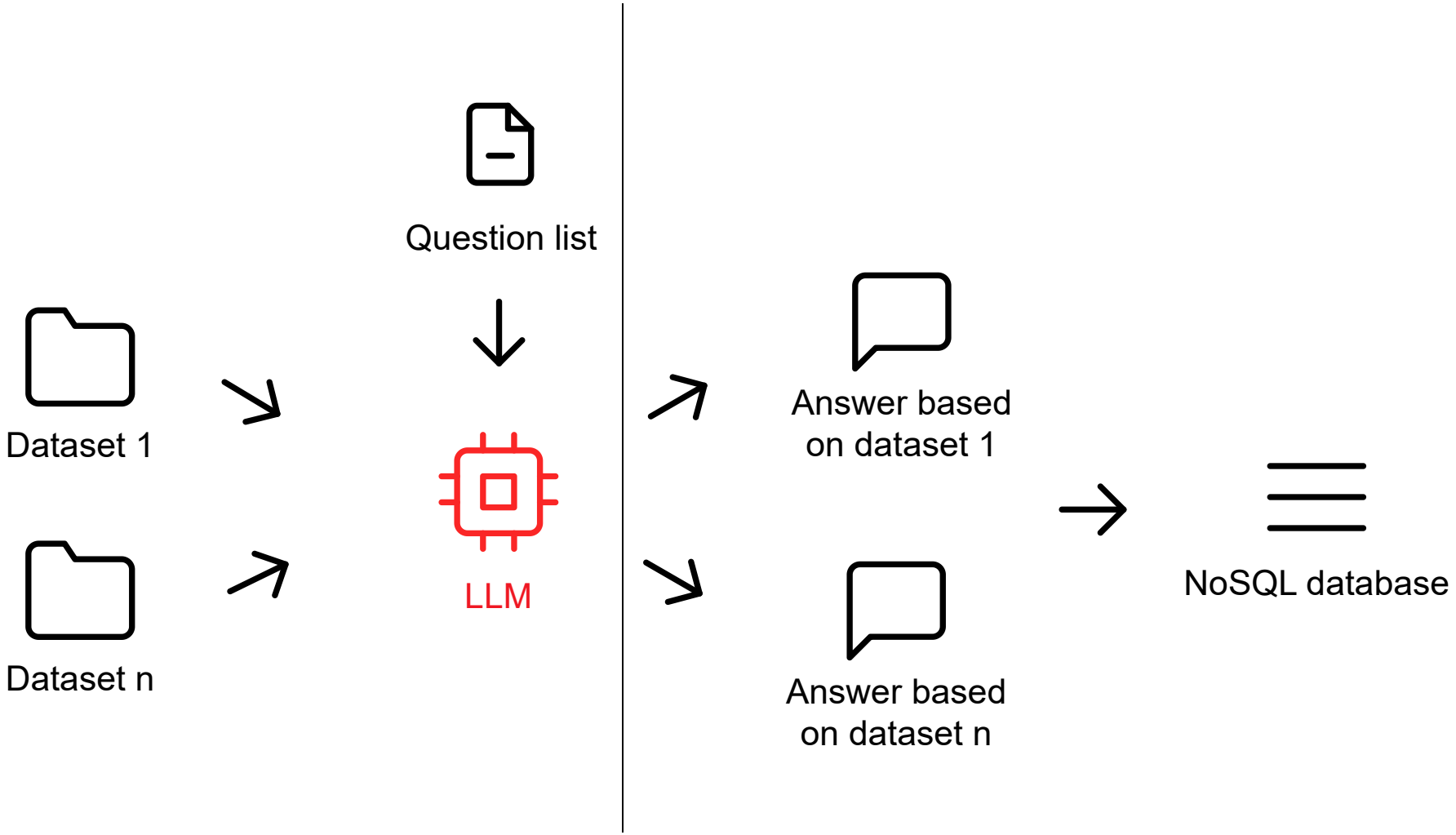
Large language models Analysis

Group: Đoàn Gia Bảo, Marcin Czarnecki, Ilaria Cuccaro
Professor: Massimo Callisto De Donato

AY 2023/24

CONTENTS

1. Project description



1. Project description

Requirements:

1. The LLM model **automatically execute** the question list as default queries.
2. Capability of answering a **wide range of questions** (keyword extraction requests, summarization, degree of consistency with a specific topic, etc.)
3. The results of the queries must be **saved to a NoSQL database**.
4. The answers have to be **grammatically and semantically correct**.

2. Methodology and technology

MongoDB

ANACONDA®

NVIDIA

CUDA®

kafka

Connect

LLAMA C++

PyTorch

jupyter

OnPrem.LLM

> A tool for running large language models on-premises using non-public data

Why MongoDB ?

1. Schema flexibility
2. Scalability
3. High performance
4. Geospatial capabilities
5. Native support for JSON
6. Agile development

Local LLM advantages

01 **Privacy and security**

02 **Reliability**

03 **Processing speed**

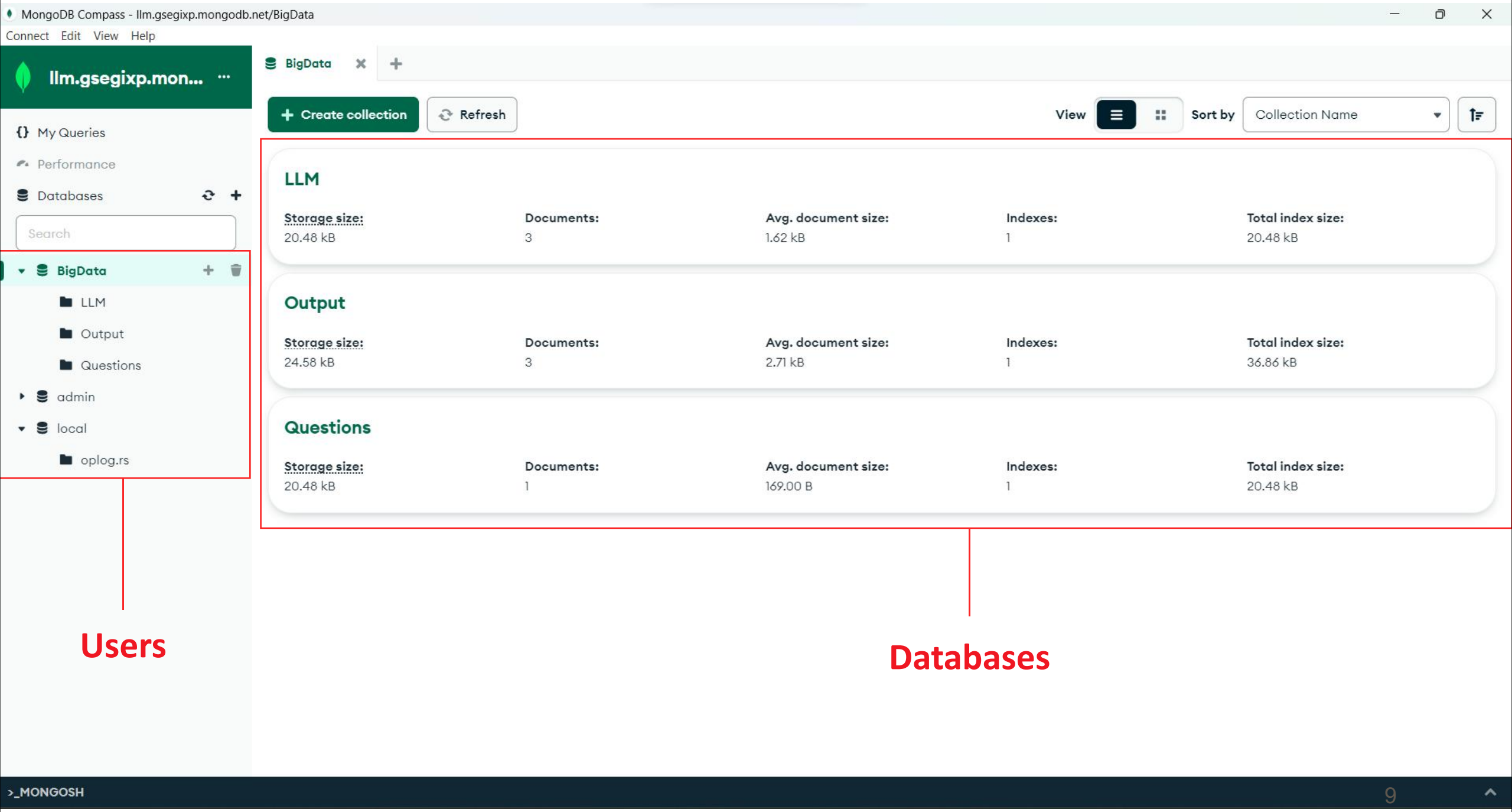
04 **Long-term cost**

3. Technical implementation

- The database has been configured to attach to the **LLM prototype**, along with the **authentication and authorisation** machanism.

- The database has been added to the **cloud** (MongoDb Atlas), allows any user who wants to connect to the database to access it from anywhere.

- Access the database using **MongoDB Compass** with the URL: mongodb+srv://bigdata:ProjectBigData123@llm.gsegixp.mongodb.net/


MongoDB Compass - llm.gsegixp.mongodb.net/BigData
Connect Edit View Help
llm.gsegixp.mon...
My Queries
Performance
Databases
Search
BigData
LLM
Output
Questions
admin
local
oplog.rs
BigData
Create collection
Refresh
View
Sort by
Collection Name
LLM
Storage size: 20.48 kB
Documents: 3
Avg. document size: 1.62 kB
Indexes: 1
Total index size: 20.48 kB
Output
Storage size: 24.58 kB
Documents: 3
Avg. document size: 2.71 kB
Indexes: 1
Total index size: 36.86 kB
Questions
Storage size: 20.48 kB
Documents: 1
Avg. document size: 169.00 B
Indexes: 1
Total index size: 20.48 kB
Users
Databases
>_MONGOSH

3. Technical implementation

Operation

Install

allows using GPUs for computing operations

A **7 billion** parameter model was used

prerequisites for Onprem

3. Technical implementation

Operation

From the core.py file in Onprem:

n_gpu_layers - how many model layers are downloaded to the GPU.
n_batch - how many tokens are processed in parallel.

Optimal value →

24 n_gpu_layer | **800** n_batch

Since the question has to be asked for each section of the document, a loop has been added:

```
answer = []
    for document in docs:
        answer.append(self.combine_documents_chain.run(
            input_documents=document, question=question, callbacks=_run_manager.get_child()
        )
    )
```

3. Technical implementation

Installation

1. Install the requirements packages for Onprem
2. Install Onprem
3. Check Jupyter installation in Anaconda
4. Speeding Up Inference
5. Adapt the model to the proposed problem
6. "Talk to Your Documents"
7. Save results in a JSON file
8. Send the JSON file to the MongoDB database

4. Achieved result


```
"number_article_1": {
    "title": "How artificial intelligence will change the future of marketing",
    "content": " The application of Artificial intelligence in marketing is in order to continuously follow and anticipate the next purchasing decisions of the target consumers. Machine learni
    "question_answer": [
        {
            "question": "What the text is about",
            "answer": " The given text delineates on how Artificial Intelligence (AI) will transform marketing in future, and the three use cases that demonstrate its impact on sales processes
        },
        {
            "question": "What machine learning techniques are used?",
            "answer": " The answer depends on the specific use cases mentioned in the context. However, some commonly used machine learning techniques that could be applied to these scenarios
        },
        {
            "question": "Does the text mention predictive analytics?",
            "answer": " Yes, the text mentions that machine learning allows marketers to understand and draw logical conclusions from large data collections which can be used for predictive an
        }
    ]
},
"number_article_3": {
    "title": "Spotify personalizes audio experiences with machine learning",
    "content": " In 2008, the Spotify platform incorporated machine learning into its business model, transitioning from user curated playlists to Machine Learning driven recommendations. Tony
    "question_answer": [
        {
            "question": "What the text is about",
            "answer": " Spotify personalizes audio experiences with machine learning by utilizing raw data from playlists, listening behaviors of users, and other factors to build ML models th
        },
        {
            "question": "What machine learning techniques are used?",
            "answer": " Collaborative filtering, content-based recommendation and user behavior analysis are some of the Machine Learning techniques that Spotify uses to personalize audio expe
        },
        {
            "question": "Does the text mention predictive analytics?",
            "answer": " No, the text does not explicitly mention predictive analytics."
        }
    ]
},
"number_article_2": {
    "title": "How Does Facebook Use Machine Learning to Deliver Ads?"
```

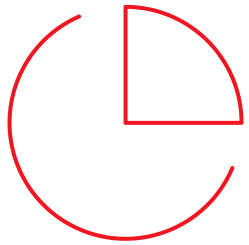
4. Achieved result

- The prototype is able to **read** the input and give **correct answers** both grammatically and semantically.

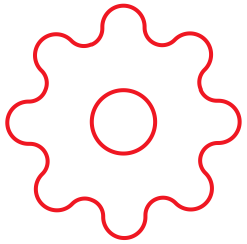
- The output is **automatically saved** into existing MongoDB

- When the file with 3 documents is used, the first answer was given by the prototype in **1,30 minutes**.

5. Possible future improvements



Larger data size



Increase the number of tokens processed in parallel

Less delay in answering time

6. Live demo