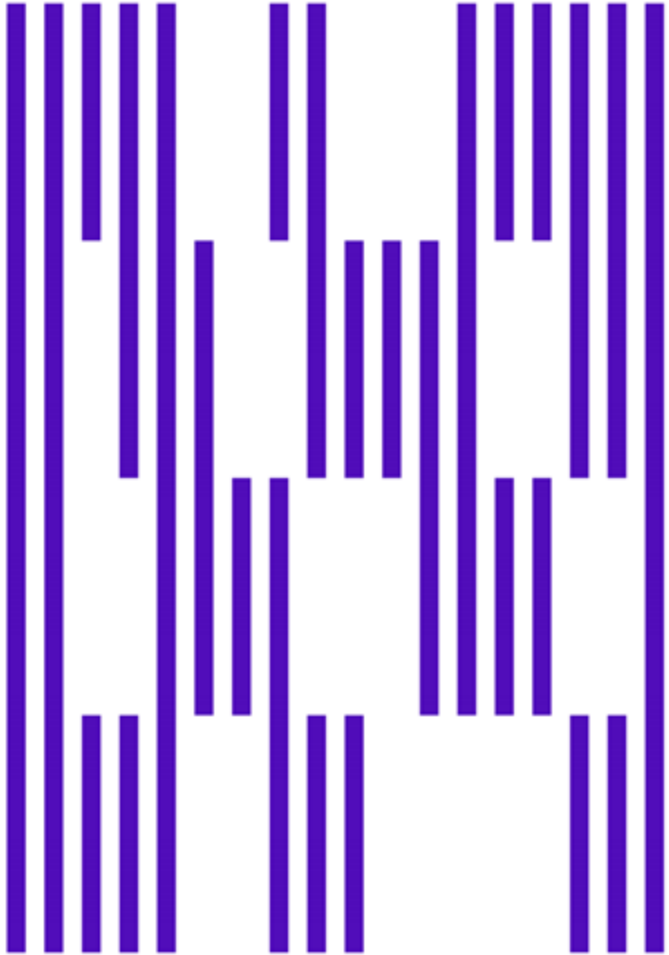
Chatujte s vašimi daty: Praktická ukázka RAG

S využitím interně provozovaných jazykových modelů

www.datera.cz

Kdo jsme?

- Zpracování velkých objemů nestrukturovaných dat
- Infrastruktura + software (LISSA, EMMA)
- Bezpečnostní složky, státní správa, pojišťovny



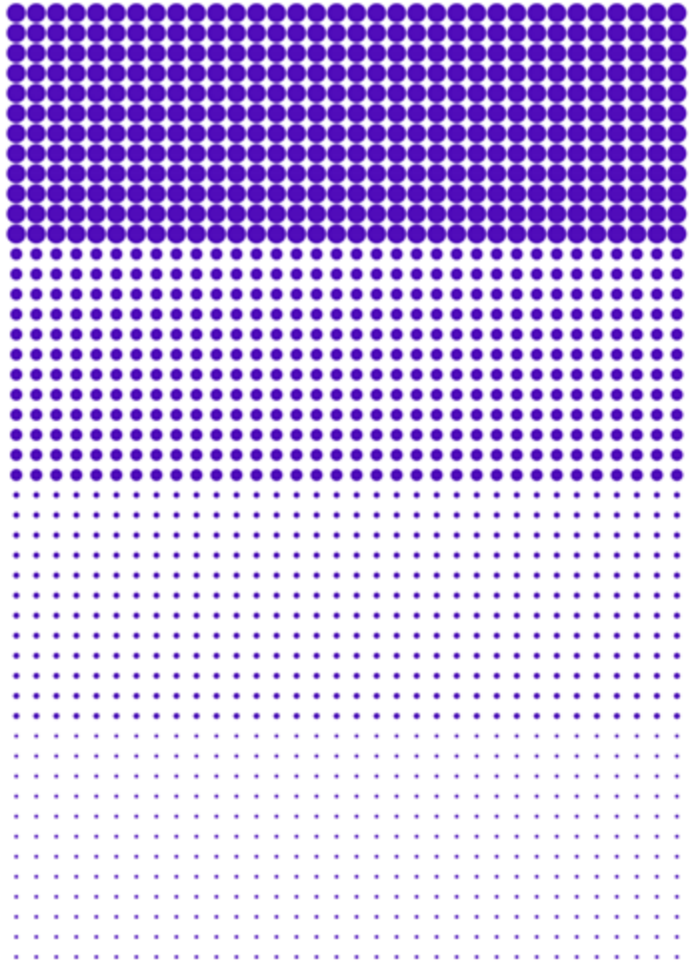
Agenda

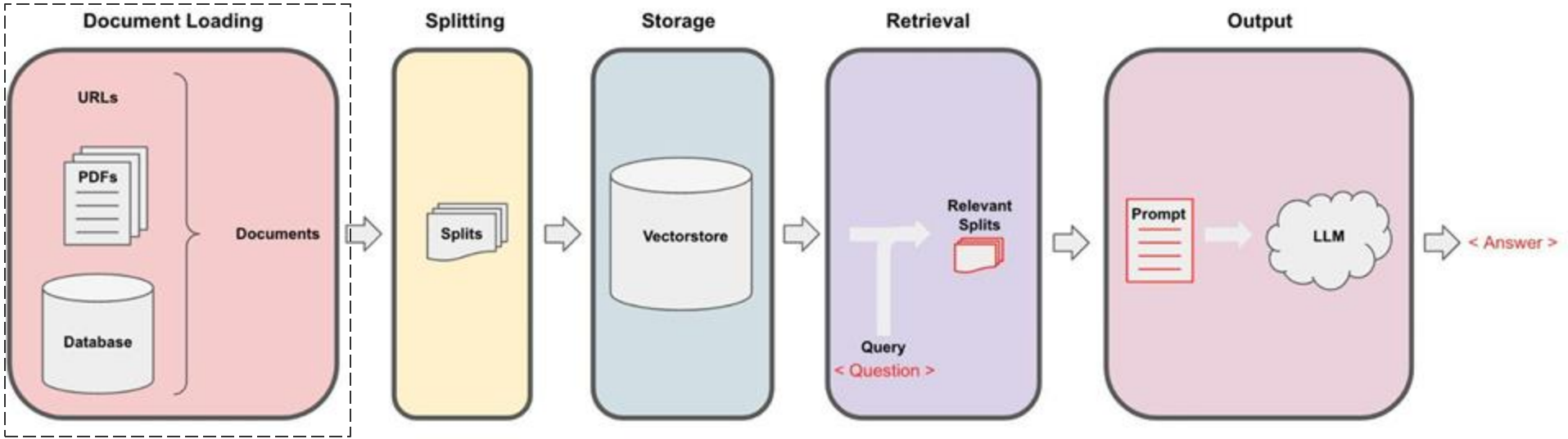
1. Úvod do RAG
2. Ukázka RAG nad pojistnými podmínkami
3. Technologie
4. Benchmarky

Úvod

Knowledge check

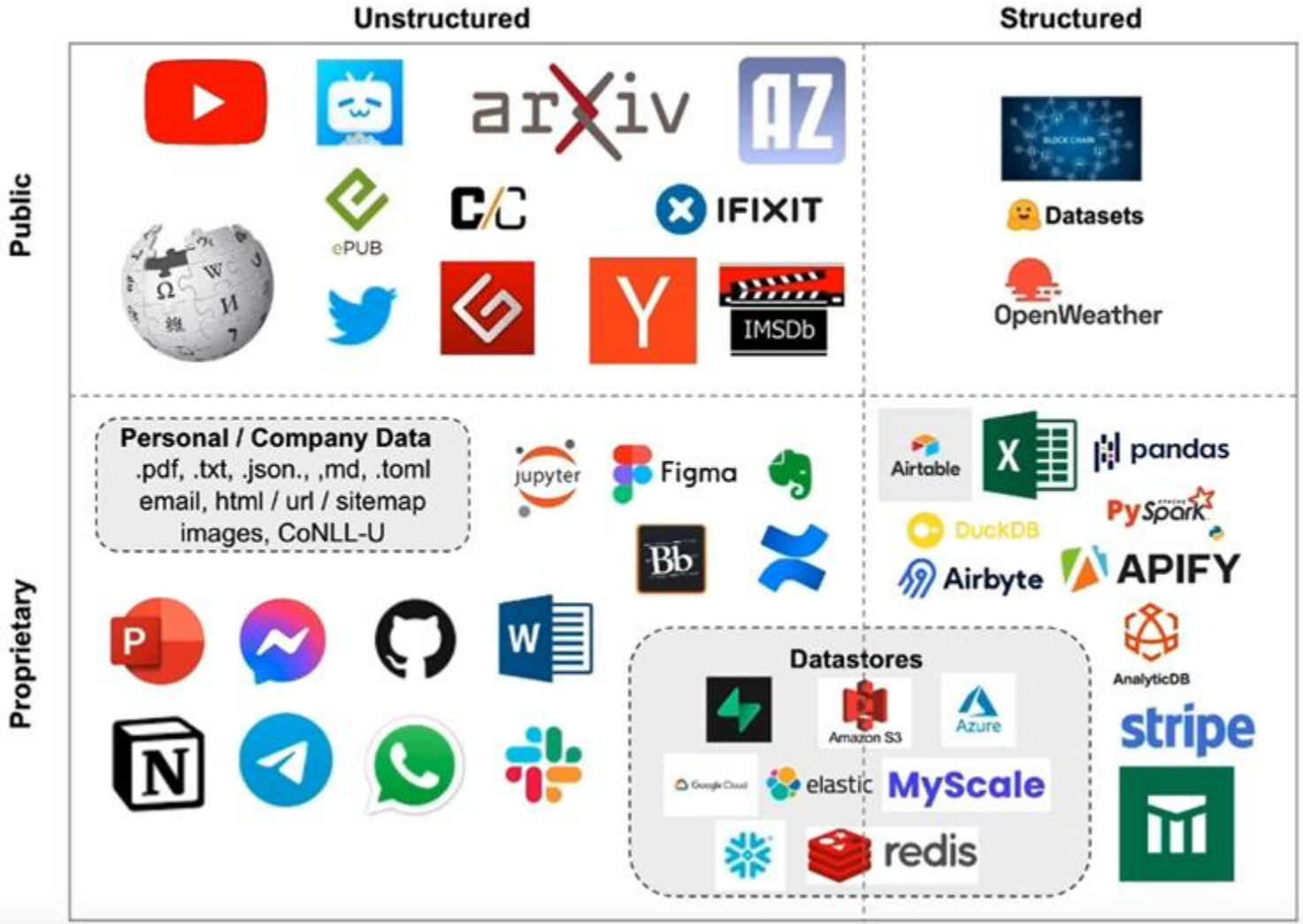
- Kdo ví, co je RAG?
- Kdo zvládne vysvětlit RAG klientovi?
- Kdo má RAG aplikaci v produkci?

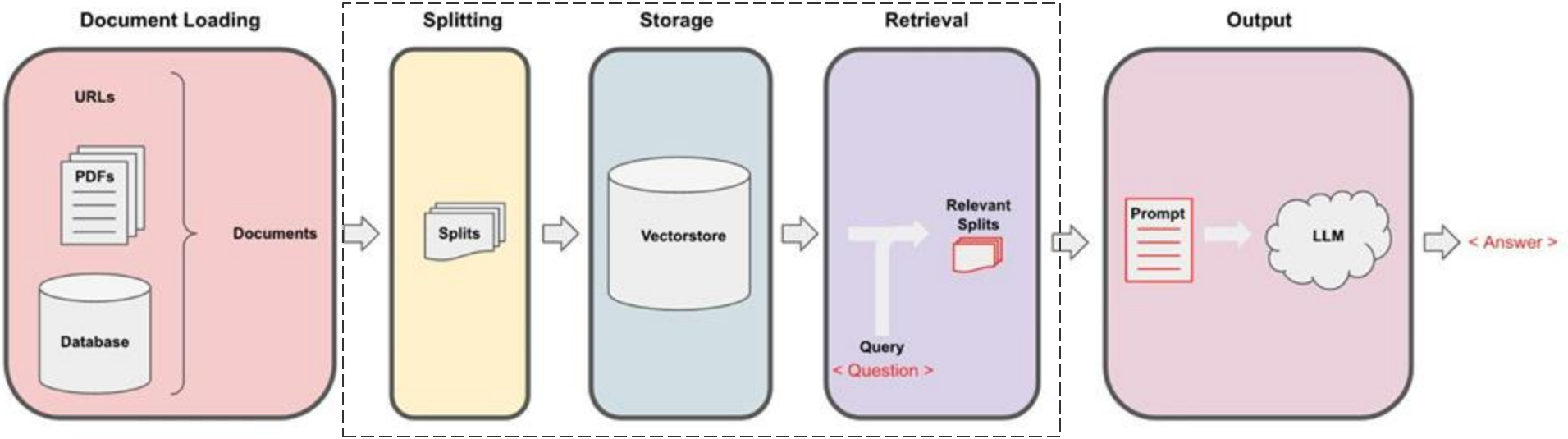




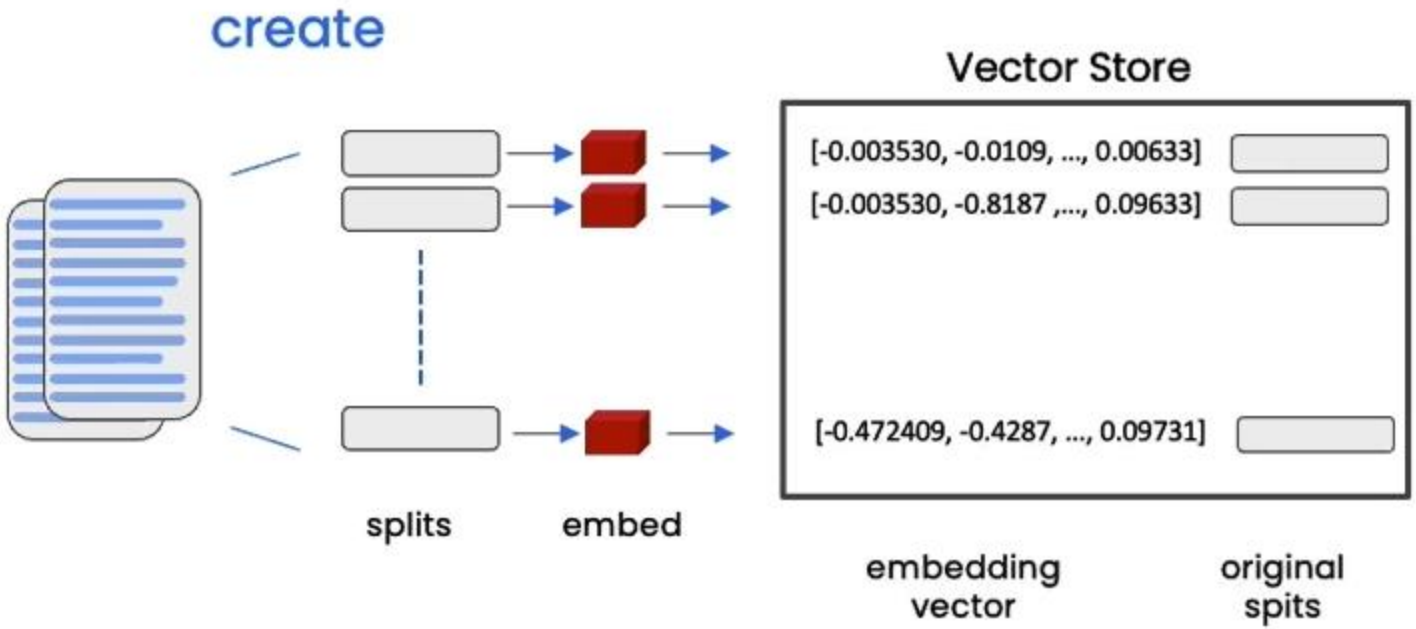
Source: https://www.deeplearning.ai/short-courses/langchain-chat-with-your-data/

datera

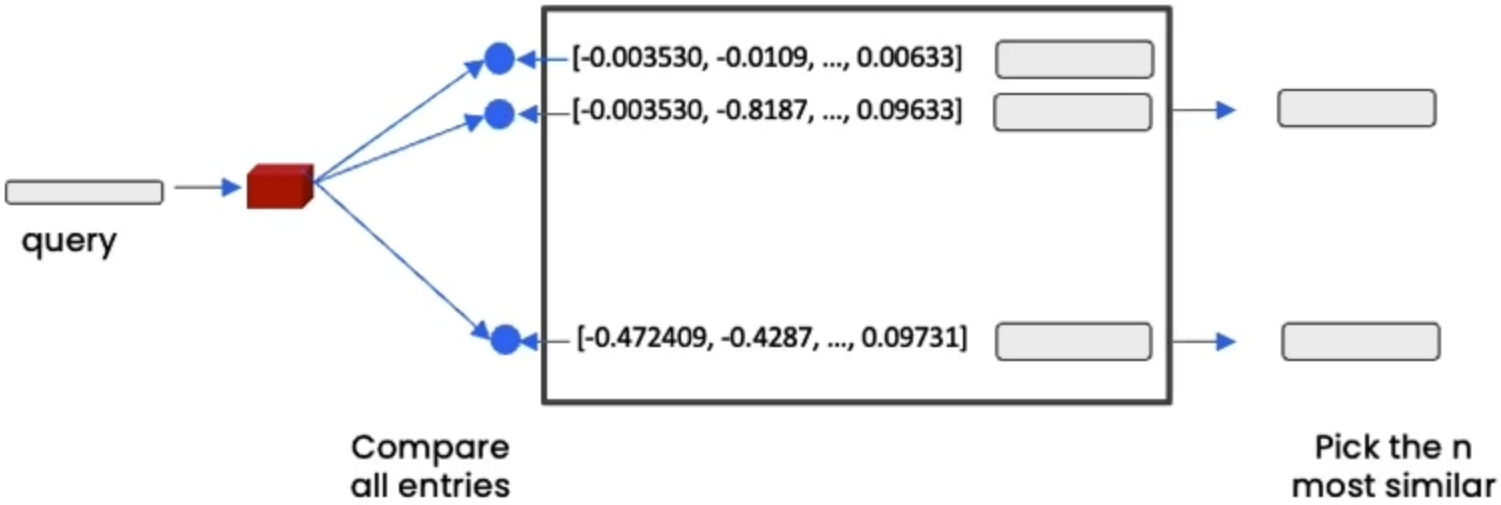

Unstructured
Structured
Public
Proprietary
arXiv
AZ
ePUB
C/C
IFIXIT
IMSDb
Datasets
OpenWeather
Personal / Company Data
.pdf, .txt, .json., ,md, .toml
email, html / url / sitemap
images, CoNLL-U
jupyter
Figma
Bb
Airtable
pandas
PySpark
DuckDB
Airbyte
APIFY
AnalyticDB
Datastores
Amazon S3
Azure
elastic
MyScale
redis
stripe

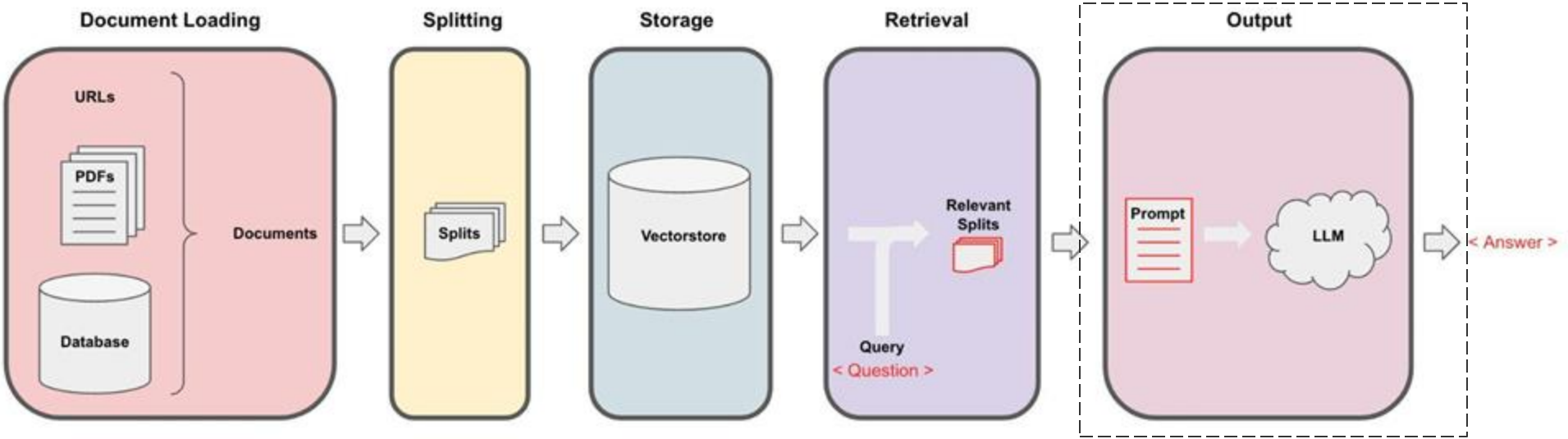

Document Loading
URLs
PDFs
Database
Documents
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant Splits
Query
< Question >
Output
Prompt
LLM
< Answer >

datera


create
Vector Store
[-0.003530, -0.0109, ..., 0.00633]
[-0.003530, -0.8187 ,..., 0.09633]
[-0.472409, -0.4287, ..., 0.09731]
splits
embed
embedding vector
original spits

index




Document Loading
URLs
PDFs
Database
Documents
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant Splits
Query
< Question >
Output
Prompt
LLM
< Answer >

datera

Ukázka

PROJEKT
nvidia-rag

RAG

Zapnout editaci

Omezit časový rozsah

SÉMANTICKÉ

Doplňte dotaz

Žádné aktivní filtry

Vyhledáno **2** výsledků Řadit dle project.fields.author.keyword

0.92 PDF 19.05.2023 09:20

cestovni_poj.pdf

`cestovni_poj.pdf`

Soubor pojistných podmínek pro cestovní pojištění
!
Co je důležité vědět Dovolujeme si Vás upozornit, že ne všechny části tohoto souboru pro Vás musí být relevantní. Vždy se prosím řiďte především tím, jaká konkrétní pojištění máte v pojistné smlouvě sjednána. V určitých případech můžeme sníži…

0.13 PDF 18.03.2024 10:22

vozidla_poj.pdf

`vozidla_poj.pdf`

A-001 (platné od 1. 4. 2024)
Pojištění vozidel Pojistné podmínky
Pro lepší orientaci a rychlejší vyhledávání využijte interaktivní prvky:
Prostřednictvím horní zelené lišty se rychle dostanete na vybranou část předsmluvních informací....

Stránkování 10

1

Dokumenty **1** až **2** z **2**

Čekám na váš dotaz, stačí se jen zeptat!

v0.14.2

DATERA s.r.o.

Technologie

Přehled možností pro LLM enterprise řešení

- **vLLM**
  - Jednoduché, funkční, výkonné
  - Široký výběr kompatibilních modelů
  - OpenAI kompatibilní API
- **NVIDIA Triton Inference Server**
  - Vyžaduje znalost architektury daného modelu pro tvorbu konfigurace
  - Pro spuštění modelu používá tzv. backendy (TensorRT-LLM, vLLM, PyTorch, TensorFlow, Python, ...)
  - Zatím nemá OpenAI kompatibilní API
- **NVIDIA NIM**
  - Kontejnerizovaná microservice
  - Konkrétní model (např. llama-3.1-70b-instruct), optimalizovaný pro daný HW
  - Pod kapotou používa NVIDIA Triton Inference Server
  - Jednoduché, funkční, výkonné
  - OpenAI kompatibilní API

Benchmarky

Testovací prostředí

- Srovnáváme **llama-3.1-70B-instruct** přes:

    1. **NIM**
        1. Triton + vLLM backend, profil vllm-bf16-tp4
    2. **vLLM**
    3. **DIM**
        1. Vlastní kontejner s Triton + TensorRT-LLM backend

- Na následujícím HW:
    - **4x** NVIDIA **L40S** (4 x 46,068 MiB = ~**184 GiB VRAM**)
    - Intel Xeon Gold 6548Y+ (128 cores @ 2.5 GHz)
    - 1007 GiB RAM

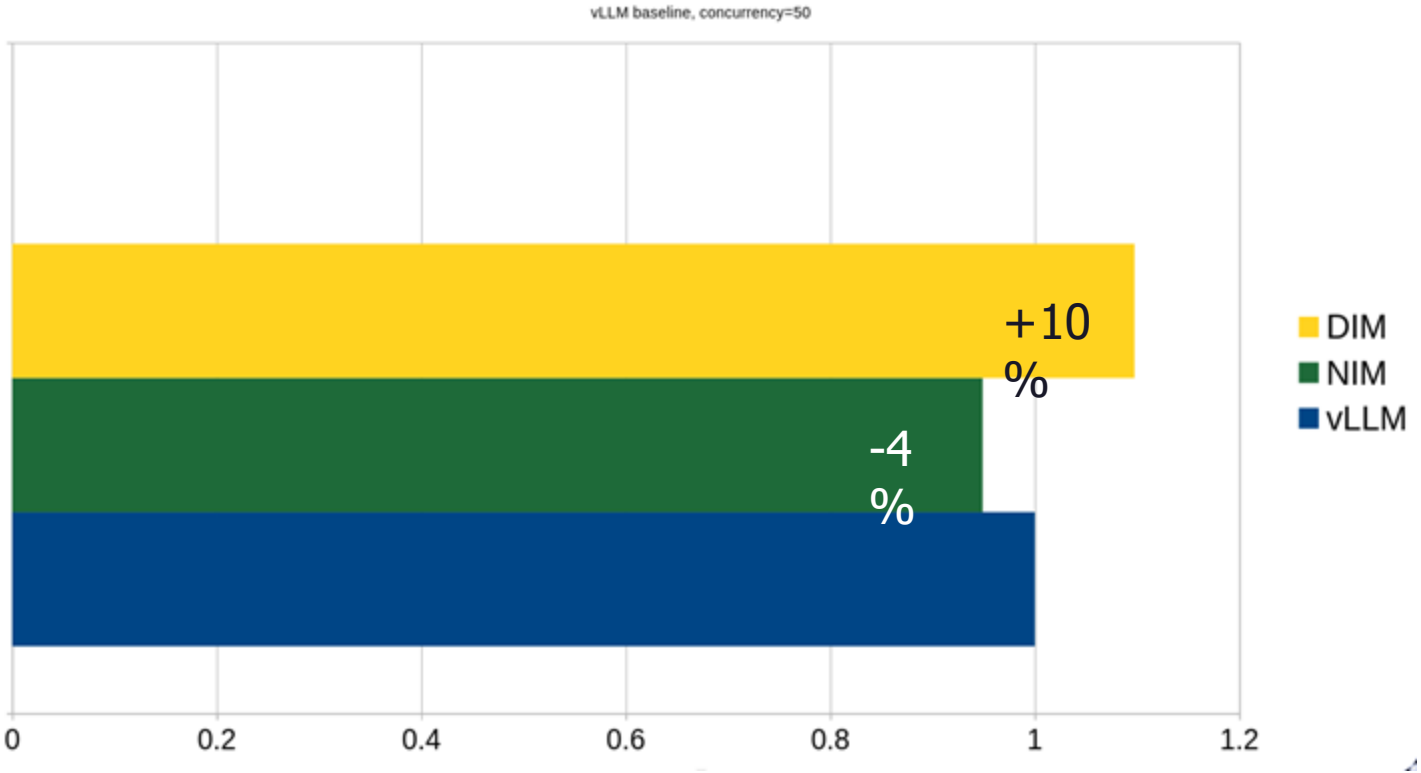
datera

Testovací scénáře

- Simulace několika typických scénářů:
  1. **Klasifikace** (input=200, output=5)
  2. **Překlad** (input=200, output=200)
  3. **Sumarizace** (input=1000, output=200)
- Pro každý scénář testovány různé úrovně concurrency
- Testování pomocí nástroje Genai-perf

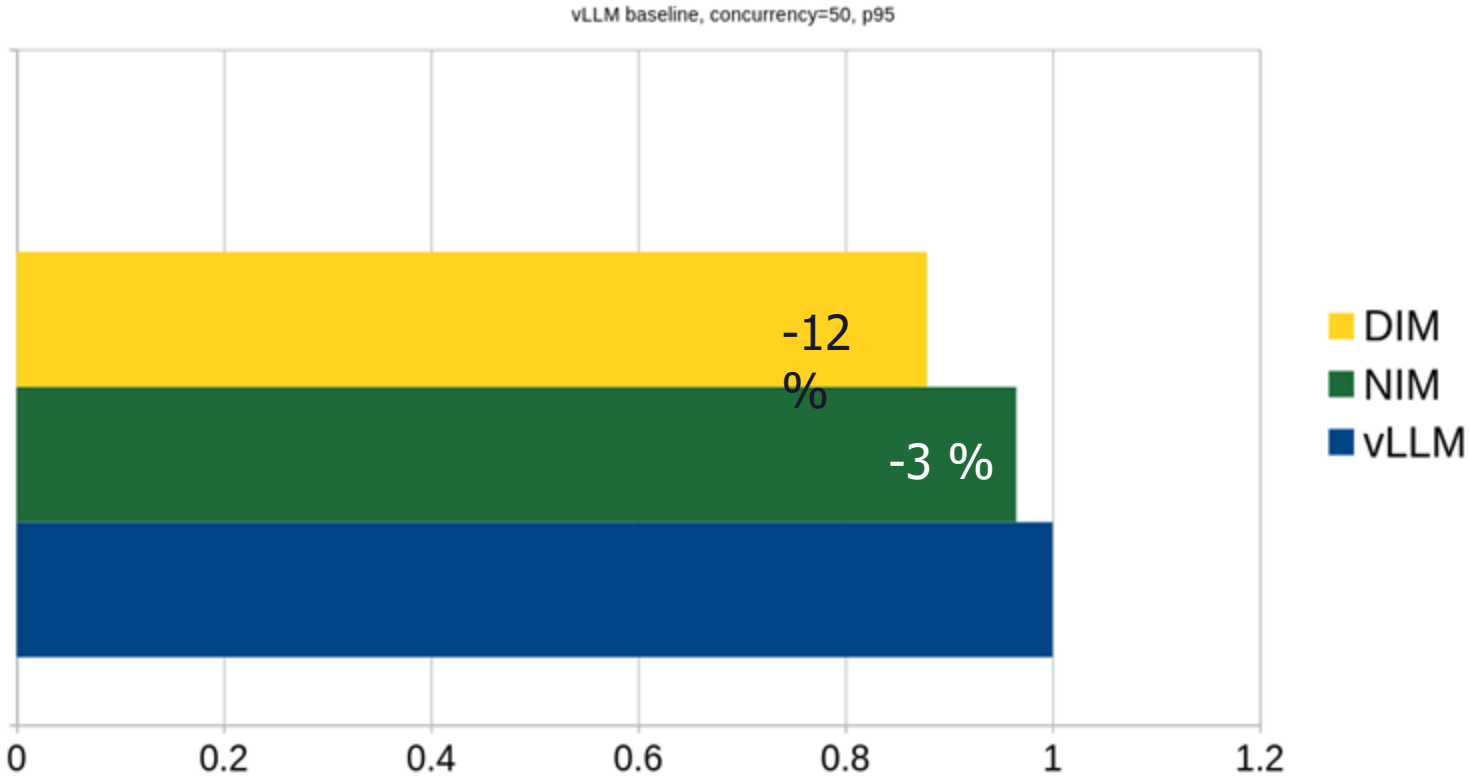
datera

Výsledky - Output Token Throughput (%)



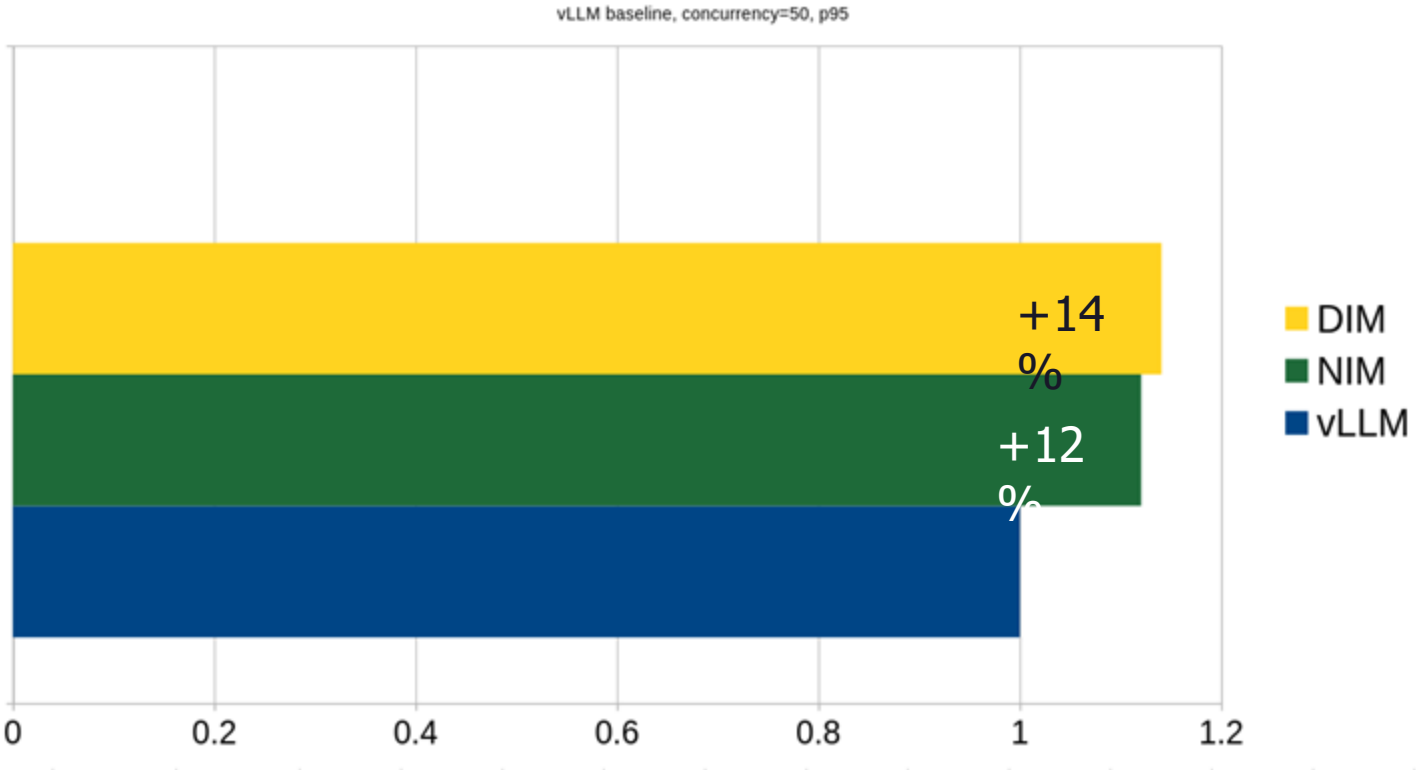
datera

Výsledky - Time To First Token (%)



datera

Výsledky - Inter Token Latency (%)



datera

otázky?

datera

Děkujeme za pozornost!

Matouš Eibich
matous.eibich@datera.cz
linkedin.com/in/matous-eibich

Tomáš Ptáček
tomas.ptacek@datera.cz

DATERA s.r.o.
Hráského 25, 148 00 Praha 4
www.datera.cz