Project BD2

- Proiectul își propune realizarea unui Bechmark pentru Sisteme de Gestiune a Bazelor de Date NewSQL (SGBD NewSQL)
- Un Bechmark reprezintă o metodologie bine definită pentru a evalua performanțele unui sistem

- Obiective:
    - Înțelegerea funcționării unei baze de date distribuite
    - Înțelegerea mecanismelor de replicare și fragmentare
    - Însușirea cunoștințelor necesare pentru a lucra cu noi tipuri de baze de date
    - Dezvoltarea abilităților de a prezenta și discuta într-un mod obiectiv rezultatele obținute
    - Dezvoltarea abilităților de a face o comparație și o evaluare critică folosind date empirice a mai multor sisteme distribuite de baze date
    - Învățarea să gestioneze eficient timpul

- Cum se realizează un benchmark:

1. Se alege un set de date care are un număr mare de înregistrări (de ordinul sutelor de mii/milioanelor)
2. Se propune o schema logică pentru stocarea datelor (Diagrama entitate relație)
3. Se creează schema fizică pentru sistemul de gestiune ales care să reflecte schema logică
4. Se propun un set de interogări/operații prezentate formal (de obicei în algebră relațională): în general operații de tip INSERT, SELECT, UPDATE, DELETE
5. Se definește factorul de scalare (adică se împarte setul mare de date în subseturi, de obicei se folosește pentru testare o optime, un sfert, o jumătate și tot setul de date
6. Se scriu interogările în limbajul nativ al fiecărui sistem de gestiune de baze de date pentru a nu avea latență
7. Se execută fiecare cerere pe fiecare subset din setul de date de 10 ori și se calculează / notează timpul de execuție
8. Se compară timpul mediu și deviația standard

Ce trebuie sa faceti:

1. Alegeti un SGBD NewSQL din urmatoarele (puteti veni si voi cu o propunere care trebuie sa primeasca ok-ul de la mine): (Nota: un SGBD poate fi ales doar de un student)
    a. CockroachDB
    b. VoltDB
    c. NuoDB
    d. YugabyteDB
    e. TokuDB

2. Pentru SGBD-ul NewSQL scrieti o descriere cu urmatoarele informatii (exemplul de descriere in articolul [2]- Setiunea 3) :
    a. Ce fel de baza de date este, evidentiind caracteristicile acesteia
    b. In ce limbaj a fost dezvoltat
    c. Limbajul (limbajele) de interogare care poate fi folosit pentru a accesa datele stocare
    d. Tipuri de indecși
    e. Tipul de replicare
    f. Tipul de distribuție suportat
3. Completati un tabel precum urmatorul, care are urmatoarele linii (exemplul tabel din articolul [2])

Table 1
DODBMS comparison.

	BaseX	eXist-db	Sedna	MongoDB	CouchDB	Couchbase
DBMS type	XDBMS	XDBMS	XDBMS	JDBMS	JDBMS	JDBMS
Data format	XML	XML	XML	BSON (Binary JSON)	JSON	JSON
Implementation	Java	Java	C	C++	Erlang	C/C++, Go, Erlang
Transaction	ACID	Isolation safe	ACID	BASE Multi-document isolation	Document-level ACID with MVCC	ACID
Consistency	Transaction Consistency	Automatic consistency Sanity checker	Transaction Consistency	Causal Consistency	Eventual Consistency	Eventual Consistency Immediate Consistency
In-memory	Yes	Yes	Yes	Yes	No	Yes
Replication	No	Primary-Secondary	No	Primary-Secondary	Primary-Primary Primary-Secondary	Primary-Primary Primary-Secondary
Partitioning	No	No	No	Sharding	Sharding	Sharding
Ad-hoc queries	XQuery 3.1 XPath 3.1	XQuery 3.1 XPath 3.1	XQuery 1.0 XPath 2.0	JavaScript	Mango	N1QL JavaScript
MapReduce	No	No	No	Yes	Yes	Yes
Secondary indices	Yes	Yes	Yes	Yes	Yes	Yes
Geospatial indices	No	No	Yes	Yes	Yes	Yes
Text indices	Yes	Yes	Yes	Yes	Yes	Yes

4. Instalati SGBD-ul NewSQL intr-un mediu cu o singura instanta si intr-un mediu cu 3 instante distribuite. (recomand in docker)
    a. Se va documentat instalarea pas cu pas
    b. Se va atasa la documnetatia finala 2 scripturi cu comenzile complete pentru a instala SGBD-ul intr-un mediu cu o singura instanta precum si in mediul distribuit.
5. Luati setul de date de la https://drive.google.com/drive/folders/1ofKgkw62gm2edPu9E3ysYbztQYTrshQr?usp=sharing
6. Precesati setul de date astfel incat sa poata fi inserat in urmatoarele 2 schemele 1 si 2 (referinta articolul [3]):

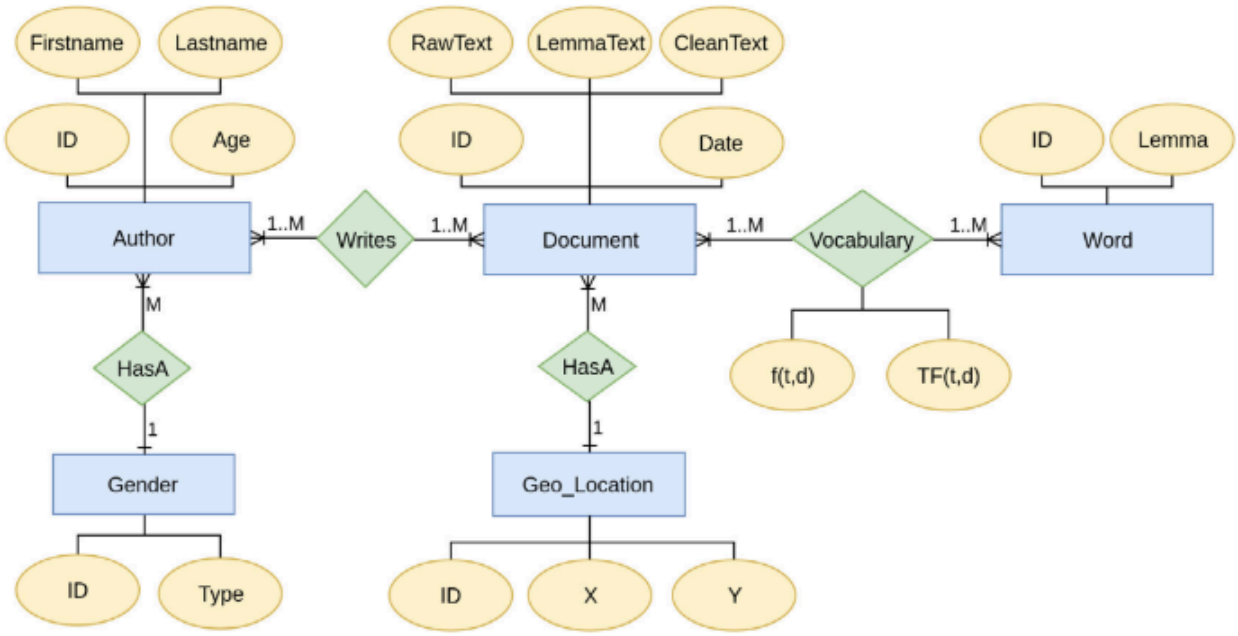

Fig. 1. T^2K^2 conceptual data model.

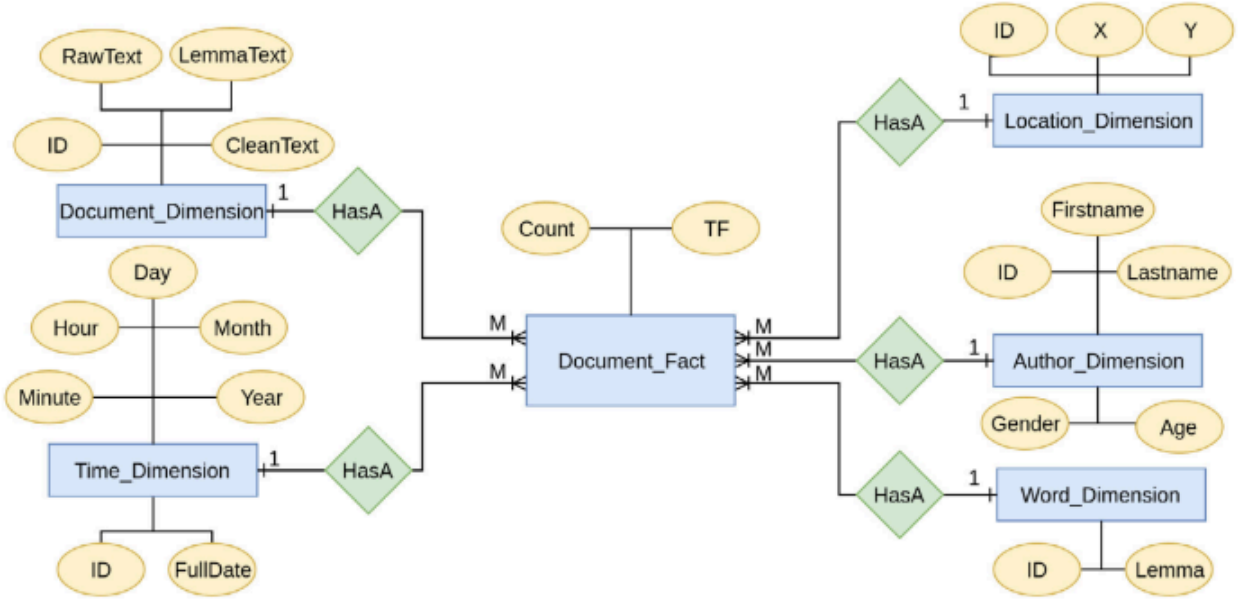

Fig. 2. $T^2K^2D^2$ conceptual data model.

7. Creati schemele fizice la nivelul SGBD-ului ales (referinta articolul [3] figurile 3 si 5).
8. Scrieti cererile din articolul [3] in limbajul nativ al SGBD-ului ales. Gasiti aici https://github.com/cipriantruica/T2K2D2_Benchmark versiunile SQL ale acestora pentru Oracle si PostgreSQL, trebuie doar adaptate sa functioneze cu SGBD-ul ales.
    a. Se for scrie cererile pentru Top-K Keywords (in directorul cu TopK_Keywords)

    b. Se for scrie cererile pentru Top-K Documents (in directorul cu TopK_Documents)
    c. Se va atasa la documnetatia finala scripturile cu comenzile pentru selectie.
    d. Cererile pentru schema 1 au prexiful DB_
    e. Cererile pentru schema 2 au prexiful OLAP_
9. Se vor face 2 seturi de experimente, unul pe mediul cu o instanta si unul pe mediul distribuit, astfel:
    a. Setul de date este deja impartit in in 5 subseturi (i.e., documents_clean500K.json pana la documents_clean2500K.json)
    b. Fiecare cerere se va executa de 10 ori pe fiecare subset in parte si se va inregistra timpul de executie
    c. Se va crea un tabel cu timpul mediu de executie si deviatia standard a timpului mediu de executie pentru fiecare cerere si subset in parte (tabelel se gasesc aici aici https://drive.google.com/drive/folders/1ofKgkw62gm2edPu9E3ysYbztQYTrshQr?usp=sharing).
10. Documentatie cu urmatoarele:
    a. Descrierea SGBD-ului NewSQL ales (punctul 2)
    b. Tabelul cu sumarizarea characteristicilor SGBD-ului (punctul 3)
    c. Instalarea detaliata a SGBD-ului (punctul 4)
    d. Cererile pentru crearea bazelor de date (punctul 7)
    e. Cererile in limbajul nativ cu scripturile (punctul 8)
    f. Timpi de executie pentru fiecare cerere (tabelul de la punctul 9) impreuna cu un excel in care sunt inregistrati toti timpii.
11. Deadline proiect **26.08.2024**

Referințe:

[1] Ciprian-Octavian Truică, Elena-Simona Apostol, Jérôme Darmont, Ira Assent. *TextBenDS: a generic Textual data Benchmark for Distributed Systems*. Information Systems Frontiers, 23:81-100. ISSN 1387-3326, Springer. February 2021 (published online March 2020). DOI: 10.1007/s10796-020-09999-y (**Q1** Journal) [pdf]

[2] Ciprian-Octavian Truică, Elena-Simona Apostol, Jérôme Darmont, Torben Bach Pedersen. *The Forgotten Document-Oriented Database Management Systems: An Overview and Benchmark of Native XML DODBMSes in Comparison with JSON DODBMSes.* Big Data Research, 25:1-14(100205), ISSN 2214-5796, July 2021 DOI: 10.1016/j.bdr.2021.100205 (**Q1** Journal) [pdf]

[3] Ciprian-Octavian Truică, Jérôme Darmont, Alexandru Boicea, Florin Rădulescu. *Benchmarking Top-K Keyword and Top-K Document Processing with T2K2 and T2K2D2*. Future Generation Computer Systems, 85:60-75, ISSN 0167-739X, August 2018. DOI: 10.1016/j.future.2018.02.037 (**Q1** Journal) [pdf]