UNIVERZITA PARDUBICE   
Fakulta elektrotechniky a informatiky

*MongoDB Cluster*

*Andrii Streblychenko 3 ročník*

V …..dne….

Obsah

[Úvod 4](#_Toc195037598)

[1 Architektura 5](#_Toc195037599)

[1.1 Schéma a popis architektury 5](#_Toc195037600)

[1.2 Specifika konfigurace 6](#_Toc195037601)

[1.2.1 CAP teorém 6](#_Toc195037602)

[1.2.2 Cluster 6](#_Toc195037603)

[1.2.3 Uzly 6](#_Toc195037604)

[1.2.4 Sharding 6](#_Toc195037605)

[1.2.5 Replikace 7](#_Toc195037606)

[1.2.6 Perzistence dat 7](#_Toc195037607)

[1.2.7 Distribuce dat 7](#_Toc195037608)

[1.2.8 Zabezpečení 7](#_Toc195037609)

[2 Funkční řešení 8](#_Toc195037610)

[2.1 Struktura 8](#_Toc195037611)

[2.1.1 docker-compose.yml 9](#_Toc195037612)

[2.2 Instalace 10](#_Toc195037613)

[3 Případy užití a případové studie 11](#_Toc195037614)

[3.1 Účely použití databáze MongoDB 11](#_Toc195037615)

[3.2 Důvody výběru databáze MongoDB pro tento projekt 11](#_Toc195037616)

[3.3 Případové studie 11](#_Toc195037617)

[3.3.1 Vývoj e-commerce platformy pomocí MERN stacku 11](#_Toc195037618)

[3.3.2 Využití MongoDB v IBM pro správu různorodých dat 12](#_Toc195037619)

[3.3.3 Definice a vlastnosti MongoDB podle TechTarget 12](#_Toc195037620)

[4 Výhody a nevýhody 13](#_Toc195037621)

[4.1 Výhody a nevýhody MongoDB 13](#_Toc195037622)

[4.2 Výhody a nevýhody navrženého řešení 13](#_Toc195037623)

[5 Další specifika 15](#_Toc195037624)

[6 Data 16](#_Toc195037625)

[6.1 Použité datové soubory 16](#_Toc195037626)

[6.2 Struktura a typy dat 16](#_Toc195037627)

[6.3 Rozsah dat 16](#_Toc195037628)

[6.4 Úpravy dat 17](#_Toc195037629)

[6.5 Zdroj dat 17](#_Toc195037630)

[6.6 Analýza v Pythonu 17](#_Toc195037631)

[6.6.1 Netflix (netflix\_films.csv) 18](#_Toc195037632)

[6.6.2 Indian Traffic Violations (indian\_traffic\_violations.csv) 19](#_Toc195037633)

[6.6.3 Mastercard Stock History (mastercard\_stock\_history.csv) 21](#_Toc195037634)

[7 DOTAZY 23](#_Toc195037635)

[7.1 Práce s daty (Filtrace a vyhledávání) 23](#_Toc195037636)

[7.1.1 Netflix 23](#_Toc195037637)

[7.1.2 Indian Traffic Violations 24](#_Toc195037638)

[7.1.3 Stock Prices 25](#_Toc195037639)

[7.2 Agregační dotazy 26](#_Toc195037640)

[7.2.1 Netflix 26](#_Toc195037641)

[7.2.2 Indian Traffic Violations 27](#_Toc195037642)

[7.2.3 Stock Prices 28](#_Toc195037643)

[7.3 Aktualizace dokumentů 29](#_Toc195037644)

[7.3.1 Netflix 29](#_Toc195037645)

[7.3.2 Indian Traffic Violations 30](#_Toc195037646)

[7.3.3 Stock Prices 31](#_Toc195037647)

[7.4 Mazání záznamů 32](#_Toc195037648)

[7.5 Indexace 33](#_Toc195037649)

[7.6 Testování replikace: výpadek jednoho shardu při zapnutém mongos routeru 34](#_Toc195037650)

[Závěr 35](#_Toc195037651)

[Zdroje 36](#_Toc195037652)

[Přílohy 37](#_Toc195037653)

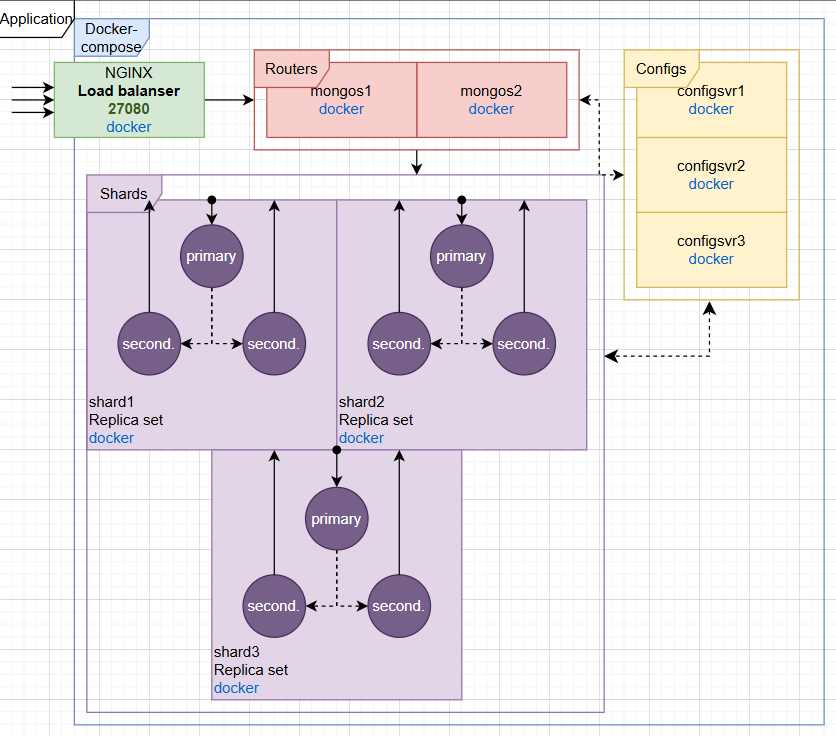
Úvod

xxxxxxxxx

# Architektura

Cílem této kapitoly je detailně popsat architekturu navrženého řešení pomocí dokumentové NoSQL databáze **MongoDB** ve formě plně shardovaného a replikovaného clusteru. Tento systém je nasazen pomocí **Docker Compose** a využívá více kontejnerů pro různé role v clusteru, jako jsou **config servery, datové shardy, mongos routovací uzly** a **NGINX load balancer**. Řešení je navrženo s důrazem na škálovatelnost, vysokou dostupnost a bezpečnost komunikace mezi jednotlivými komponentami. Architektura reflektuje doporučené praktiky MongoDB

## Schéma a popis architektury



Celý systém je navržen jako plně škálovatelný MongoDB sharded cluster, nasazený prostřednictvím docker-compose. Architektura zahrnuje tyto komponenty:

NGINX Load Balancer (27080). NGINX slouží jako vstupní bod zvenku. Rozkládá zátěž mezi dvě instance mongos routeru, čímž zajišťuje vysokou dostupnost a vyrovnávání zátěže.

Routers – mongos1, mongos2. Mongos je komponenta MongoDB, která směruje dotazy na správné shardy podle metadata uložených v config servers. Dvě instance mongos zajišťují redundanci a lepší výkon.

Config Servers – configsvr1, configsvr2, configsvr3. Tyto tři uzly tvoří replikový set configReplSet, který uchovává metadata o rozmístění dat mezi shardami. Tato komponenta je nezbytná pro správnou funkčnost sharded clusteru.

Shards – shard1, shard2, shard3. Každý shard je samostatný replikový set o třech uzlech – jeden hlavní (PRIMARY) a dva sekundární (SECONDARY). Tato architektura zajišťuje vysokou dostupnost a replikaci dat pro každý shard.

Docker síť. Všechny komponenty běží v izolované Docker síti, což umožňuje vzájemnou komunikaci, ale omezuje přístup zvenku pouze na port 27080 skrze NGINX.

Odlišnosti od doporučené architektury MongoDB:

Byla přidána NGINX vrstva jako externí přístupový bod s vyrovnáváním zátěže.

## Specifika konfigurace

Architektura tohoto MongoDB řešení je navržena jako plně shardovaný a replikovaný cluster s důrazem na škálovatelnost, vysokou dostupnost, bezpečnost a rozšiřitelnost. Všechny komponenty jsou nasazeny prostřednictvím Docker Compose a komunikují spolu ve vnitřní síti Dockeru.

### CAP teorém

MongoDB ve sharded architektuře se přiklání ke kombinaci CP (Consistency + Partition Tolerance)

Consistency: MongoDB replikové sety zaručují, že zápisy jsou konzistentní a dostupné na PRIMARY uzlu.

Partition Tolerance: systém je odolný vůči výpadkům částí sítě nebo uzlů díky shardingu a replikaci.

### Cluster

Používá se jeden cluster MongoDB, postaveny na «*docker-compose*», který obsahuje všechny komponenty (config servery, mongos, shardy, load balancer). Tento přístup odpovídá standardnímu návrhu distribuované MongoDB infrastruktury.

### Uzly

Celkem je nasazeno 14 uzlů MongoDB:

**3 config servery + 2 mongos routery + (3 shardy × 3 repliky) = 14 uzlů**

Každý replikový set má jednoho PRIMARY a dva SECONDARY uzly.

### Sharding

Použity jsou 3 shardy: *shard1, shard2, shard3* Každý shard obsahuje 3 repliky. Výběr byl učiněn kvůli práci s rozsáhlými daty (např. tisíce záznamů o přestupcích a transakcích), kde je důležité rovnoměrné rozložení zátěže a výkon.

### Replikace

Každý shard je implementován jako replikový set se třemi uzly – *PRIMARY* a dvěma *SECONDARY*. Tim se dava ochranu před výpadkem jednoho nebo dvou uzlů, možnost čtení z replik, pokud je potřeba a automatické přepnutí PRIMARY uzlu v případě výpadku.

### Perzistence dat

Data se ukládají do Docker volume (mapované na /data/dbX) na každém uzlu. MongoDB používá WiredTiger storage engine, který zajišťuje zápis na disk při každé operaci (write concern) a journaling pro zotavení při selhání. Data jsou perzistentní i při restartu kontejnerů díky svázaným svazkům (volumes).

### Distribuce dat

MongoDB rozděluje data na základě shard key, které určujeme při vytváření kolekcí. *Config servery* drží metadata o rozmístění chunků (částí kolekcí). *Mongos* router rozhoduje, který shard má být dotazován. Díky replikaci jsou zápisy zajištěny i v případě výpadku jednotlivého uzlu.

### Zabezpečení

Používá se keyfile-based internal authentication – všechny uzly (config servery, mongos, shard) sdílejí jeden *keyfile*. Dále je nastaven uživatel *admin* s heslem přes .*env* soubor, vyžadovaný pro všechny externí přístupy. Všechny porty kromě *27080* (NGINX) jsou uzavřeny pro vnější přístup, což chrání před neautorizovanými přístupy.

# Funkční řešení

Cílem projektu bylo vytvořit plně automatizovaný a škálovatelný MongoDB cluster, který by fungoval „z krabice“ – tedy bez nutnosti manuálního nastavování jednotlivých komponent. Díky tomu celý cluster lze spustit jediným příkazem přes Docker Compose. Toho bylo dosaženo kombinací konfigurací v Docker Compose souboru (docker-compose.yml), Dockerfile-ů a Bash skriptů (entrypoint skripty, inicializační skripty shardů a config serverů, a automatická tvorba validačních schémat kolekcí). V případě potřeby rozšíření clusteru stačí přidat nové uzly do existujících skriptů a Docker konfigurací.

## Struktura

* **mongodb\_cluster**
  + **config-server**
    - Dockerfile – definuje obraz Docker kontejneru config serveru.
    - configsvr.conf – konfigurační soubor MongoDB config serveru.
    - entrypoint.sh – skript, který inicializuje config server.
    - keyfile – sdílený soubor pro interní autentizaci clusteru.
  + **mongos**
    - Dockerfile – definuje obraz Docker kontejneru pro mongos router.
    - mongos.conf – konfigurační soubor pro mongos router.
    - entrypoint.sh – automaticky registruje shardy v clusteru.
    - schema-collections-init.sh – automatická inicializace validačních schémat kolekcí.
    - keyfile – sdílený autentizační soubor.
  + **shard**
    - Dockerfile – definuje obraz Docker kontejneru shardu s replikační sadou.
    - entrypoint.sh – inicializační skript shardu, který startuje všechny instance shardu.
    - init-shard.sh – nastavuje MongoDB replikační sadu pro shard.
    - mongod{1-3}.conf.template – šablony konfigurací pro jednotlivé instance shardu.
    - supervisord.conf.template – konfigurace Supervisora pro správu instancí shardu.
    - keyfile – sdílený autentizační soubor.
  + **nginx**
    - nginx.conf – konfigurace load balanceru, který směruje požadavky na mongos routery.
  + .env – definuje přístupové údaje a společné proměnné prostředí.
  + docker-compose.yml – centrální soubor konfigurující celý cluster MongoDB.

### docker-compose.yml

Soubor docker-compose.yml definuje strukturu a konfiguraci celého clusteru MongoDB v Docker kontejnerech. Cluster je složen z:

**Config servery** (3 uzly: configsvr1, configsvr2, configsvr3)

* Uložení metadat a konfigurace celého sharded clusteru.
* Nastavují se v režimu replikační sady nazvané configReplSet.
* Používají se sdílené proměnné prostředí (.env) a keyfile pro zabezpečenou interní komunikaci.
* Uzly sdílí volume pro perzistenci konfiguračních dat.

**Shardy** (3 replikované shardy: shard1, shard2, shard3)

* Každý shard sestává ze 3 instancí (hlavní uzel + 2 repliky).
* Každý shard má vlastní replikační sadu (např. shard1).
* Každá instance shardu běží pod kontrolou Supervisoru.
* Data jsou persistována přes oddělené Docker volumes.

**Mongos routery** (2 routery: mongos1, mongos2)

* Přijímají a směrují dotazy aplikací na konkrétní shardy.
* První router automaticky registruje shardy při spuštění (REGISTER\_SHARDS=true).
* Komunikují s config servery a shardy přes interní Docker síť.

**NGINX load balancer**

* Veřejně přístupný jediný bod vstupu (port 27080).
* Směruje požadavky na oba mongos routery pro zajištění vysoké dostupnosti a škálovatelnosti.
* Konfigurační soubor NGINX se připojuje přes volume.

Díky této konfiguraci všechny služby běží izolovaně a bezpečně. Jediný veřejně dostupný port je NGINX na localhost:27080.

## Instalace

Instalace řešení byla vytvořena tak, aby probíhala zcela automaticky, s minimální nutností manuálních zásahů. Celý cluster MongoDB lze spustit jediným příkazem z kořenového adresáře projektu:

**docker-compose up --build**

Parametr --build zajistí automatické sestavení všech kontejnerů a spuštění jejich inicializačních skriptů, čímž nastaví replikační sady, autentizaci a registraci shardů v clusteru.

Cluster bude po úspěšném spuštění dostupný přes load balancer na:

**mongodb://clusterAdmin:YourStrongPassword@localhost:27080**

* Uživatel: clusterAdmin
* Heslo: YourStrongPassword
* Připojení: Port 27080, databáze: main

Databáze main se zakládá automaticky při spuštění clusteru, spolu s kolekcemi a jejich validačními schématy.

Testovací data je možné nahrát následně jednoduchým spuštěním připravených Python skriptů, které jsou umístěné v adresáři noSQL/Data/scripts. Tyto skripty automaticky nahrají CSV datasety do odpovídajících MongoDB kolekcí:

**python upload\_netflix\_films.py**

**python upload\_indian\_traffic.py**

**python upload\_mastercard\_stock.py**

# Případy užití a případové studie

## Účely použití databáze MongoDB

MongoDB je optimální volbou pro aplikace vyžadující vysokou flexibilitu při práci s daty, rychlé dotazování a snadnou rozšiřitelnost. Tyto schopnosti plynou z její dokumentově orientované architektury, která používá formát BSON – binární rozšíření JSON. Díky této struktuře není nutné předem definovat pevné schéma a data lze upravovat i během provozu.

Systémy pracující s obsahem, profitují ze schopnosti MongoDB efektivně ukládat nestrukturované i polostrukturované informace. V oblasti e-commerce umožňuje databáze rychlé operace se sklady, produkty i zákaznickými daty. IoT řešení a mobilní aplikace získávají možnost dynamického ukládání senzorických dat bez nutnosti konverze do tabulkového formátu. Webové platformy ukládají profily, komentáře či hodnocení s minimální latencí. MongoDB také nachází uplatnění při analýze dat v reálném čase – například v monitorovacích nástrojích a vizualizačních panelech typu dashboard.

## Důvody výběru databáze MongoDB pro tento projekt

MongoDB byla zvolena jako hlavní databázová platforma kvůli potřebě pracovat s rozsáhlými, různorodými daty bez nutnosti pevně definované struktury. Projekt vyžaduje schopnost snadno importovat a analyzovat tisíce záznamů z různých domén – například dopravní přestupky, historické burzovní hodnoty nebo filmové záznamy.

Volba padla na MongoDB i kvůli osobní preferenci pracovat s moderní technologií, která je široce využívána v praxi a poptávaná na trhu práce. Apache Cassandra je orientovaná na sloupcová data a její model je zaměřený především na velmi rychlé zápisy a čtení v masivně distribuovaném prostředí a v tomto případě přinášelo omezení kvůli nedostatku flexibility při práci s dynamickými schématy. Databáze typu klíč-hodnota (Redis, Valkey) jsou vhodné spíše pro krátkodobá data nebo cache, nikoli pro komplexní datové modely. Taky s Redis jsem už pracoval.

Řešení bylo navrženo tak, aby podporovalo horizontální škálování, automatizaci nasazení a rychlou odezvu systému. MongoDB umožňuje efektivní využití shardingu, replikace i indexování. V praxi by tento systém mohl být nasazen například v analytických firmách zaměřených na business intelligence, kde je prioritou rychlá transformace surových dat do využitelných informací.

## Případové studie

### eBay – Vysoce dostupné aplikace s MongoDB

Společnost eBay, globální lídr v oblasti online obchodování, obsluhuje více než 170 milionů aktivních uživatelů a spravuje přes miliardu aktivních nabídek ve 190 zemích. Pro zajištění vysoké dostupnosti a odolnosti svých systémů se eBay rozhodla implementovat MongoDB jako jednu z hlavních databázových platforem.​

V rámci této strategie eBay vyvinula architektonické vzory pro odolné aplikace, které zahrnují sedmičlenné replikové sady MongoDB rozprostřené napříč třemi datovými centry v USA. Tato konfigurace zajišťuje, že i v případě výpadku jednoho datového centra zůstávají aplikace dostupné díky automatickému přepnutí na jiný uzel. MongoDB umožňuje eBay nastavit priority pro volbu primárního uzlu, což zajišťuje konzistenci a minimalizuje dobu obnovy.​

Pro čtenářsky náročné aplikace, jako je katalog produktů, eBay škáluje replikové sady až na 50 členů, čímž dosahuje vysoké dostupnosti a výkonu. Pro zápisově náročné aplikace využívá eBay shardované klustery MongoDB, které jsou distribuovány napříč datovými centry, což umožňuje efektivní rozložení zátěže a zajišťuje odolnost vůči výpadkům.​

Díky těmto opatřením eBay dosahuje vysoké dostupnosti svých služeb a zajišťuje spolehlivý provoz svých aplikací i v případě neočekávaných událostí. [1]

### Autodesk – Flexibilní datová platforma pro návrhářské nástroje

Společnost Autodesk, známá svými CAD a 3D návrhovými programy, čelila výzvě efektivního shromažďování a analýzy uživatelských dat z různých produktů. Cílem bylo zlepšit uživatelskou zkušenost prostřednictvím personalizovaných doporučení a zefektivnění pracovních postupů. Původní řešení, kombinující objektové a metadatové úložiště, nedokázalo držet krok s rostoucími požadavky na flexibilitu a škálovatelnost.​

Přechodem na MongoDB Atlas získal Autodesk možnost ukládat analytická data do jediného dokumentu, což usnadnilo jejich načítání a poskytování poznatků napříč různými softwarovými produkty. Tato změna umožnila týmu doručovat miliony poznatků týdně a zároveň snížit provozní náklady. Flexibilita MongoDB Atlas umožnila Autodesku rychle reagovat na měnící se požadavky a datové struktury bez nutnosti rozsáhlých migrací.​

Integrace s GraphQL přinesla zlepšení výkonu API a dosažení subsekundových latencí, což je klíčové pro poskytování personalizovaných doporučení v reálném čase. MongoDB Atlas se tak stal centrálním prvkem v infrastruktuře Autodesku, podporujícím inovace a zlepšující uživatelskou zkušenost napříč produkty. [2]

### ****Digitalizace Toyoty s využitím MongoDB – propojení Toyota Connected a Toyota Financial Services****

Toyota Financial Services (TFS) a Toyota Connected (TC) úspěšně integrovaly MongoDB do svých digitálních ekosystémů, čímž výrazně zvýšily efektivitu a flexibilitu svých operací.

TFS nasadila MongoDB již v roce 2017 jako základ nového úložiště dat. Toto rozhodnutí se stalo klíčovým pro rozvoj agilní kultury vývoje v rámci společnosti. MongoDB umožnila TFS efektivně pracovat s různorodými daty, což vedlo k rychlejšímu vývoji finančních produktů a lepšímu rozhodování na základě dat.

Toyota Connected, která se zaměřuje na vývoj telematických a mobilních řešení, zvolila cloudovou službu MongoDB Atlas, kterou provozuje na platformě AWS. Díky tomu bylo možné zajistit vysokou dostupnost služeb, jako je například Safety Connect, a bezproblémový provoz aplikací v režimu 24/7. MongoDB Atlas zároveň nabídla potřebnou škálovatelnost pro zpracování velkých objemů dat v reálném čase.

Obě společnosti uvedly, že MongoDB přispěla k lepší spolupráci mezi vývojovými a provozními týmy, urychlila nasazování nových funkcí a zvýšila celkovou výkonnost. MongoDB se tak stala nedílnou součástí digitální transformace Toyoty, podporující inovace a dlouhodobý růst v rychle se měnícím technologickém prostředí. [3] [4]

# Výhody a nevýhody

## Výhody a nevýhody MongoDB

***Výhody***

1. Práce s různorodými a měnícími se datovými strukturami je v MongoDB bezproblémová díky volnému schématu a formátu BSON.
2. Vysoká rychlost vývoje je umožněna flexibilitou dotazovacího jazyka, který nabízí agregace, vnořené výrazy a dynamické filtry.
3. Rozšiřování úložiště je snadno dosažitelné pomocí shardingu, který rozděluje data horizontálně mezi více serverů.
4. Široká škála případů použití je umožněna podporou více datových modelů (dokumentový, key-value, grafový) v jednom systému.

***Nevýhody***

1. Zpracování komplexních relací mezi daty je neefektivní kvůli absenci plnohodnotných JOIN operací.
2. Nepřesně navržená validace může vést k nesourodým strukturám, protože MongoDB nevyžaduje explicitní schéma.
3. Při masivních zápisech dochází ke kolísání výkonu, pokud nejsou optimálně nakonfigurovány indexy a distribuce dat.
4. Obnova a zálohování dat z více shardovaných a replikovaných nodů je složitější než v monolitických databázových systémech.

## Výhody a nevýhody navrženého řešení

***Výhody***

1. Nasazení celého clusteru je okamžité díky automatické konfiguraci a připravené infrastruktuře v Docker Compose.
2. Migrace mezi prostředími je plně podporována díky použití volume svazků, které uchovávají persistentní data odděleně od kontejnerů.
3. Koncová bezpečnost systému je vysoká, protože všechny služby využívají interní autentizaci pomocí keyfile a žádný port kromě NGINX není vystaven veřejnosti.
4. Přidání nových uzlů do clusteru je rychlé a nevyžaduje zásah do běžící infrastruktury, stačí doplnit konfiguraci a spustit nové kontejnery.
5. Vytváření nových kolekcí se specifickými validačními schématy probíhá automaticky při startu systému, bez manuálních zásahů.
6. Ladění a sledování stavu systému je intuitivní díky přehlednému výstupu logů každé služby v konzoli.
7. Řešení je připraveno i pro produkční nasazení, protože obsahuje všechny klíčové komponenty MongoDB clusteru s důrazem na modularitu a bezpečnost.

***Nevýhody***

1. Úprava logiky inicializačních skriptů je náročná kvůli jejich složitosti a vzájemným závislostem mezi jednotlivými částmi.
2. Diagnostika chyb je obtížná v případě pádu některého z kontejnerů, protože jejich restart neprobíhá automaticky.
3. Spuštění Python skriptů pro naplnění databáze vyžaduje předchozí ruční instalaci knihoven na lokálním systému.
4. Údržba rozsáhlého clusteru je náročná na technické znalosti, protože nesprávná konfigurace shardingu nebo replikace může vést k nekonzistenci nebo ztrátě dat.

# Další specifika

Nejvýraznějším rozšířením je integrace NGINX jako reverzní proxy a load balanceru. Díky tomuto komponentu je možné směrovat veškerou komunikaci mezi klientem a MongoDB clusterem výhradně přes jeden veřejně dostupný port 27080. Veškeré požadavky jsou následně rovnoměrně distribuovány mezi dvě instance mongos, čímž je zajištěno:

* vyvážení zátěže při více současných připojeních,
* zvýšená dostupnost v případě výpadku jedné instance mongos,
* jasně definovaný a bezpečný vstupní bod do infrastruktury,
* možnost budoucí implementace HTTPS, rate limiting nebo ochrany pomocí firewallu.

Tato vrstva navíc umožňuje jednodušší škálování směrovací části systému, protože nové instance mongos mohou být přidány bez zásahu do konfigurace klientů.

# DATA

V rámci tohoto semestrálního projektu byly použity tři samostatné datové soubory ve formátu CSV. Každý z nich představuje zcela odlišnou oblast – streamovací média, dopravní přestupky a finanční trhy. Jeden z nich obsahuje více než 5 000 záznamů, jak vyžaduje zadání. Všechna data byla nahrána do dokumentové databáze MongoDB, kde byla dále validována, analyzována a vizualizována pomocí knihoven jazyka Python – zejména pandas, numpy, matplotlib a seaborn. Veškeré výstupy a poznatky jsou zaznamenány ve skriptu data\_analysis\_notebook.ipynb, který je součástí odevzdávané dokumentace.

## Použité datové soubory

**Netflix\_films.csv**  
Datová sada obsahuje informace o více než 8800 filmech a televizních pořadech streamovaných na platformě Netflix. Každý záznam obsahuje textová metadata jako název, režisér, obsazení, žánry, zemi původu, datum přidání do katalogu, rok vydání, hodnocení a popis obsahu. Typickým rysem této sady je variabilita dat (např. některé filmy nemají uvedeného režiséra) a přítomnost více hodnot v jednom poli (např. více žánrů nebo herců). Tato struktura je ideální pro MongoDB, která umožňuje záznamy s volitelnými a opakovanými poli.

**Indian\_Traffic\_Violations.csv**  
Tato sada čítá 4 000 záznamů o dopravních přestupcích z různých oblastí Indie. Obsahuje detailní údaje o typu přestupku, pokutě, lokalitě, věku a pohlaví řidiče, typu vozidla, podmínkách na silnici a ve vzduchu (počasí), úrovni alkoholu v krvi, přítomnosti bezpečnostních prvků (helma, pás), a dalších. Vzhledem k časté nepřítomnosti některých údajů (např. barva vozidla, stav semaforu, poznámka) se tato nestrukturovaná data skvěle hodí do dokumentové databáze, kde nejsou prázdné hodnoty problémem.

**Mastercard\_stock\_history.csv**  
Tato datová sada obsahuje přes 5 000 historických záznamů o denních cenách akcií společnosti Mastercard. Zahrnuje sloupce jako datum, otevírací, maximální, minimální a uzavírací cena akcie, obchodovaný objem, vyplacené dividendy a případné dělení akcií. Všechna data jsou numerická nebo datového typu, což umožňuje efektivní agregace a výpočty. Díky MongoDB bylo možné snadno doplnit o nové pole (např. legacy\_data pro záznamy staršího data).

## Struktura a typy dat

Data byla původně ve formátu CSV a po importu převedena do formátu BSON, který MongoDB interně používá. Tento formát umožňuje flexibilní strukturu bez nutnosti předem definovaného schématu. Právě proto byla zvolena dokumentová databáze – umožňuje snadné přidávání, úpravu a filtrování záznamů.

Alternativní databázové přístupy, jako relační SQL databáze nebo sloupcové NoSQL databáze typu Cassandra, by vyžadovaly pevně definovanou schématickou strukturu a nebyly by tak vhodné pro různorodá data s volitelnými poli. MongoDB navíc poskytuje možnost validace přes $jsonSchema, bylo v tomto projektu použito u všech tří kolekcí.

## Rozsah a kvalita dat

* **Netflix:** 8 807 záznamů
* **Indian Traffic Violations:** 4 000 záznamů
* **Mastercard Stock History:** 5 026 záznamů

Ve všech sadách byly identifikovány chybějící hodnoty. Například u dopravních přestupků často chybí údaje o modelu vozidla nebo ID policisty. Tyto hodnoty byly v MongoDB uloženy jako null a při analýze byly zohledněny pomocí funkcí isnull() a fillna() v pandas. U finančních dat byly odstraněny záznamy s nulovým objemem obchodů. Netflix dataset vykazoval rozdílnou hloubku informací napříč záznamy, zejména u starších filmů a regionálně specifického obsahu.

## Úpravy a transformace dat

* **Konverze typů:** Datum byl převeden na datetime, čísla jako Fine\_Amount či Volume na float nebo int, podle potřeby.
* **Odstranění irelevantních záznamů:** Např. záznamy s nulovým objemem nebo bez datumu přidání.
* **Doplnění nových polí:** Například legacy\_data pro záznamy starší roku 2010 nebo big\_dividend pro dividendy vyšší než 0.2.
* **Normalizace:** Pole s více hodnotami (např. listed\_in v Netflix datasetu) byly rozdělena pomocí $split a $unwind pro agregace.

## Zdroj dat

* **Netflix:** [Kaggle - Netflix Movies and TV Shows](https://www.kaggle.com/datasets/shivamb/netflix-shows)
* **Indian Traffic Violations:** [Kaggle - Indian Traffic Violation Dataset](https://www.kaggle.com/datasets/khushikyad001/indian-traffic-violation)
* **Mastercard Stock History:** Interní dataset získaný z platformy Yahoo Finance (spolubydlící měl na počítače 😊)

## Analýza v Pythonu

Analytická část projektu probíhala v Jupyter Notebooku data\_analysis\_notebook.ipynb. Notebook obsahuje přehledné části věnované každému z datasetů. Notebook využívá připojení přes knihovnu PyMongo, transformuje dokumenty na pandas DataFrame a následně provádí vizualizace pomocí matplotlib a seaborn. Níže je ukázka hlavního kódu:

# Imports

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from pymongo import MongoClient

import warnings

warnings.filterwarnings('ignore')

sns.set(style='whitegrid')

plt.rcParams["figure.figsize"] = (12, 6)

# Connect to MongoDB

client = MongoClient(mongodb://clusterAdmin:YourStrongPassword@localhost:27080/?authSource=admin&readPreference=nearest")

db = client['main']

# --- Netflix Dataset ---

netflix = pd.DataFrame(list(db.netflix.find()))

netflix.drop('\_id', axis=1, inplace=True)

# Graf 1: Typ obsahu (TV Show / Movie)

sns.countplot(data=netflix, x='type')

plt.title('Rozdělení obsahu podle typu')

plt.show()

# Graf 2: Rok vydání

netflix['release\_year'].value\_counts().sort\_index().plot(kind='bar')

plt.title('Počet titulů podle roku vydání')

plt.xlabel('Rok')

plt.ylabel('Počet')

plt.show()

# Graf 3: Top 10 zemí

netflix['country'].value\_counts().head(10).plot(kind='barh')

plt.title('Top 10 zemí původu')

plt.show()

# --- Indian Traffic Violations ---

traffic = pd.DataFrame(list(db.Indian\_Traffic\_Violations.find()))

traffic.drop('\_id', axis=1, inplace=True)

# Graf 1: Top 10 typů přestupků

traffic['Violation\_Type'].value\_counts().head(10).plot(kind='bar')

plt.title('Top 10 typů dopravních přestupků')

plt.show()

# Graf 2: Věk řidičů

sns.histplot(data=traffic, x='Driver\_Age', bins=20, kde=True)

plt.title('Distribuce věku řidičů')

plt.show()

# Graf 3: Trestné body podle pohlaví

sns.boxplot(data=traffic, x='Driver\_Gender', y='Penalty\_Points')

plt.title('Rozdělení trestných bodů podle pohlaví')

plt.show()

# --- Mastercard Stock History ---

stock = pd.DataFrame(list(db.stock\_prices.find()))

stock.drop('\_id', axis=1, inplace=True)

stock['Date'] = pd.to\_datetime(stock['Date'], errors='coerce')

stock = stock.dropna(subset=['Date'])

# Graf 1: Vývoj závěrečné ceny

stock.sort\_values('Date').plot(x='Date',y='Close', title='Závěrečná cena akcií v čase')

plt.show()

# Graf 2: Distribuce objemu obchodování

sns.histplot(stock['Volume'], bins=30)

plt.title('Distribuce objemu obchodování')

plt.show()

# Graf 3: Vývoj dividend

sns.lineplot(data=stock, x='Date', y='Dividends')

plt.title('Vývoj dividend v čase')

plt.show()

Pro každý dataset byly použity následující techniky:

* Popisná statistika: počet záznamů, rozsah hodnot, průměr, minimum, maximum, směrodatná odchylka
* Analýza chybějících hodnot: pomocí isnull() .sum() a dropna()
* Vizualizace: histogramy, boxploty, časové řady, párové korelace, žebříčky nejčastějších hodnot
* Specifické metriky: např. rozdíl mezi rychlostí a limitem u přestupků, změna ceny akcie během dne, distribuce žánrů a herců
* Ukázky grafů: jsou vloženy do dokumentace jako ilustrace výstupů a interpretace dat (viz kapitola 6.6 v plném textu)

Notebook je klíčovým nástrojem ověření kvality dat, výběru správných metrik a návrhu struktury indexů a dotazů v MongoDB.

### Netflix (netflix\_films.csv)

Изображение выглядит как снимок экрана, текст, График, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.

**Rozdělení obsahu podle typu (Movie vs TV Show):**  
První graf ukazuje, že většina obsahu tvoří filmy. Tato dominance je důležitým poznatkem při profilování obsahu dostupného uživatelům.

Изображение выглядит как текст, линия, График, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

**Počet titulů podle roku vydání:**  
Historický přehled počtu vydaných titulů odhalil prudký nárůst po roce 2010. Tento trend koresponduje s expanzí streamovacích platforem.

**Изображение выглядит как текст, снимок экрана, линия, число

Контент, сгенерированный ИИ, может содержать ошибки.Top 10 zemí původu:**  
Třetí graf analyzuje geografické rozložení produkce. Nejvíce titulů pochází ze Spojených států, následovaných Indií a neidentifikovanými položkami.

### Indian Traffic Violations (indian\_traffic\_violations.csv)

Изображение выглядит как текст, снимок экрана, Параллельный, линия

Контент, сгенерированный ИИ, может содержать ошибки.

**Top 10 typů dopravních přestupků:**  
Mezi nejčastější přestupky patří řízení pod vlivem alkoholu, absence helmy a jízda bez řidičského oprávnění. Tyto informace lze využít pro návrh preventivních kampaní.

**Изображение выглядит как График, диаграмма, линия, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.Distribuce věku řidičů:**  
Věková distribuce ukazuje vyvážené zastoupení napříč věkovými kategoriemi, s mírně vyšší frekvencí u starších řidičů (60–70 let).

**Изображение выглядит как снимок экрана, Прямоугольник, диаграмма, прямоугольный

Контент, сгенерированный ИИ, может содержать ошибки.Rozdělení trestných bodů podle pohlaví:**  
Pomocí boxplotu bylo zjištěno, že rozdělení bodových postihů je podobné napříč všemi kategoriemi pohlaví, bez výraznějších odchylek.

### Mastercard Stock History (mastercard\_stock\_history.csv)

**Изображение выглядит как линия, График, текст, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.Závěrečná cena akcií v čase:**  
Graf znázorňuje dlouhodobý růst cen akcií Mastercard od roku 2006 do roku 2022. Lze pozorovat prudký růst zejména po roce 2016.

**Изображение выглядит как текст, линия, снимок экрана, График

Контент, сгенерированный ИИ, может содержать ошибки.Distribuce objemu obchodování:**  
Histogram ukazuje, že většina obchodních objemů byla soustředěna v nízkých hodnotách, s několika výjimečně vysokými výkyvy.

**Изображение выглядит как текст, линия, График, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.Vývoj dividend v čase:**  
Vývoj dividend potvrzuje stabilní růst a ziskovost společnosti Mastercard, s pravidelným navyšováním výplat od roku 2012.

# DOTAZY

Tato kapitola obsahuje 30 netriviálních příkladů dotazů rozdělených do několika kategorií. Každý dotaz je okomentován a vysvětlen, včetně jeho chování v rámci aktuálně použitých datových sad. Dotazy zahrnují práci s daty, agregační operace, mazání, aktualizace a indexaci. Ve všech případech dotazy operují nad shardovanými kolekcemi databáze main a demonstrují schopnosti MongoDB při práci v distribuovaném prostředí.

## Práce s daty (Filtrace a vyhledávání)

### Netflix

Vyhledání filmů s hodnocením PG-13 z posledních 5 let a délkou 90–120 minut:

db.netflix.find({

type: "Movie",

rating: "PG-13",

date\_added: {

$gte: new Date(new Date().setFullYear(new Date().getFullYear() - 5))

},

duration: { $regex: /^(9\d|1[01]\d|120) min$/ }

}).limit(10)

Vyhledání kriminálních seriálů z USA po roce 2010:

db.netflix.find({

type: "TV Show",

country: "United States",

release\_year: { $gt: 2010 },

listed\_in: /Crime TV Shows/

}).limit(10)

Vyhledání starších filmů s neznámým režisérem:

db.netflix.find({

type: "Movie",

director: "Unknown",

release\_year: { $lt: 2015 }

}).limit(10)

### Indian Traffic Violations

Přestupky starších řidičů (60+) s vysokou hladinou alkoholu a špatným počasím:

db.Indian\_Traffic\_Violations.find({

Driver\_Age: { $gt: 60 },

Alcohol\_Level: { $gt: 0.08 },

Weather\_Condition: { $in: ["Rainy", "Foggy", "Snowy"] }

}).limit(10)

Noční přestupky s povinnou soudní účastí ve státě Maharashtra:

db.Indian\_Traffic\_Violations.find({

Registration\_State: "Maharashtra",

Court\_Appearance\_Required: true,

Time: { $regex: /^(22|23|00|01|02|03|04|05):/ }

}).limit(10)

Řidiči překračující rychlostní limit o více než 30 km/h:

db.Indian\_Traffic\_Violations.find({

Recorded\_Speed: { $ne: null, $exists: true },

Speed\_Limit: { $ne: null, $exists: true },

$expr: {

$gt: [

{ $subtract: ["$Recorded\_Speed", "$Speed\_Limit"] },

30

]

}

}).limit(10)

### Stock Prices

Dny, kdy cena akcie vzrostla o více než 10 jednotek:

db.stock\_prices.aggregate([

{

$project: {

Date: 1,

Open: 1,

Close: 1,

Change: { $subtract: ["$Close", "$Open"] }

}

},

{ $match: { Change: { $gt: 10 } } },

{ $sort: { Change: -1 } }

])

Vyhledání nadprůměrně obchodovaných dnů:

const avgVolume = db.stock\_prices.aggregate([

{ $group: { \_id: null, avgVol: { $avg: "$Volume" } } }

]).toArray()[0].avgVol;

db.stock\_prices.find({ Volume: { $gt: avgVolume } })

Pokles ceny při vyplacení dividend:

db.stock\_prices.aggregate([

{ $match: { Dividends: { $gt: 0 } } },

{ $project: {

Date: 1,

Open: 1,

Close: 1,

DropPercent: {

$multiply: [

{ $divide: [{ $subtract: ["$Close", "$Open"] }, "$Open"] },

100

]

}

}},

{ $match: { DropPercent: { $lt: -2 } } },

{ $sort: { DropPercent: 1 } }

])

## Agregační dotazy

### Netflix

Top 5 žánrů přidaných za poslední 2 roky:

db.netflix.aggregate([

{ $match: { date\_added: { $gte: new Date(new Date().setFullYear(new Date().getFullYear() - 2)) } } },

{ $unwind: { path: { $split: ["$listed\_in", ", "] } } },

{ $group: { \_id: "$listed\_in", total\_titles: { $sum: 1 } } },

{ $sort: { total\_titles: -1 } },

{ $limit: 5 }

])

Průměrný věk filmů podle země:

db.netflix.aggregate([

{ $match: { type: "Movie", country: { $ne: null } } },

{ $group: {

\_id: "$country",

avg\_movie\_age: { $avg: { $subtract: [new Date().getFullYear(), "$release\_year"] } },

total\_movies: { $sum: 1 }

}},

{ $match: { total\_movies: { $gt: 10 } } },

{ $sort: { avg\_movie\_age: 1 } }

])

Počet unikátních herců v žánrech Komedie a Horory:

db.netflix.aggregate([

{ $match: { listed\_in: { $in: ["Comedies", "Horror Movies"] }, cast: { $ne: null } } },

{ $unwind: { path: { $split: ["$cast", ", "] } } },

{ $group: { \_id: "$listed\_in", unique\_actors: { $addToSet: "$cast" } } },

{ $project: { genre: "$\_id", actors\_count: { $size: "$unique\_actors" }, \_id: 0 } }

])

### Indian Traffic Violations

Nejčastější typy přestupků s průměrnou výší pokuty:

db.Indian\_Traffic\_Violations.aggregate([

{ $group: {

\_id: "$Violation\_Type",

avg\_fine: { $avg: "$Fine\_Amount" },

total\_violations: { $sum: 1 }

}},

{ $sort: { total\_violations: -1 } },

{ $limit: 5 }

])

Průměrná hladina alkoholu u mladých řidičů do 30 let:

db.Indian\_Traffic\_Violations.aggregate([

{ $match: { Driver\_Age: { $lt: 30 }, Alcohol\_Level: { $ne: null } } },

{ $group: {

\_id: "$Vehicle\_Type",

avg\_driver\_age: { $avg: "$Driver\_Age" },

avg\_alcohol\_level: { $avg: "$Alcohol\_Level" }

}},

{ $sort: { avg\_alcohol\_level: -1 } }

])

Státy s nejvyšším procentem neuhrazených pokut:

db.Indian\_Traffic\_Violations.aggregate([

{ $group: {

\_id: "$Registration\_State",

total\_violations: { $sum: 1 },

unpaid\_fines: { $sum: { $cond: [{ $eq: ["$Fine\_Paid", false] }, 1, 0] } }

}},

{ $project: {

unpaid\_fine\_percentage: {

$multiply: [{ $divide: ["$unpaid\_fines", "$total\_violations"] }, 100]

},

total\_violations: 1

}},

{ $sort: { unpaid\_fine\_percentage: -1 } },

{ $limit: 10 }

])

### Stock Prices

Průměrná závěrečná cena podle roku:

db.stock\_prices.aggregate([

{ $addFields: { Year: { $year: "$Date" } } },

{ $group: { \_id: "$Year", avgClose: { $avg: "$Close" } } },

{ $sort: { \_id: 1 } }

])

Dny s rozdělením akcií:

db.stock\_prices.aggregate([

{ $match: { Stock\_Splits: { $gt: 0 } } },

{ $count: "days\_with\_splits" }

])

Největší procentuální nárůst hodnoty akcie během dne:

db.stock\_prices.aggregate([

{

$project: {

Date: 1,

percChange: {

$multiply: [

{ $divide: [{ $subtract: ["$Close", "$Open"] }, "$Open"] },

100

]

}

}

},

{ $sort: { percChange: -1 } },

{ $limit: 5 }

])

## Aktualizace dokumentů

### Netflix

Převod ratingu starších TV pořadů na nový systém „Teen“:

db.netflix.updateMany(

{

type: "TV Show",

rating: "TV-14",

date\_added: { $lt: ISODate("2019-01-01") }

},

{ $set: { rating: "Teen" } }

)

Označení populárních filmů:

db.netflix.updateMany(

{ type: "Movie", rating: "R", release\_year: { $gt: 2020 } },

{ $set: { popular: true } }

)

### Indian Traffic Violations

Zneplatnění řidičských průkazů u recidivistů:

db.Indian\_Traffic\_Violations.updateMany(

{

Breathalyzer\_Result: true,

Previous\_Violations: { $gt: 2 }

},

{ $set: { License\_Validity: false } }

)

Přidání komentáře u vážných překročení rychlosti s alkoholem:

db.Indian\_Traffic\_Violations.updateMany(

{

$expr: {

$gt: [

"$Recorded\_Speed",

{ $add: ["$Speed\_Limit", 20] }

]

},

Alcohol\_Level: { $gt: 0.05 }

},

{ $set: { Comments: "Severe violation" } }

)

### Stock Prices

Označení starých záznamů jako historických:

db.stock\_prices.updateMany(

{ Date: { $lt: ISODate("2010-01-01") } },

{ $set: { legacy\_data: true } }

)

Označení vysokých dividend:

db.stock\_prices.updateMany(

{ Dividends: { $gt: 0.2 } },

{ $set: { big\_dividend: true } }

)

## Mazání záznamů

Mazání zastaralých TV pořadů:

db.netflix.deleteMany({

type: "TV Show",

release\_year: { $lt: 2000 },

rating: "TV-MA"

})

Odstranění dopravních přestupků s nízkou závažností z minulých let:

db.Indian\_Traffic\_Violations.deleteMany({

Date: { $lt: ISODate("2024-01-01") },

Penalty\_Points: { $lt: 5 }

})

Odstranění irelevantních obchodních záznamů:

db.stock\_prices.deleteMany({ Volume: { $lt: 1000000 } })

## Indexace

Složený index pro Netflix:

db.netflix.createIndex(

{ type: 1, country: 1, release\_year: -1, rating: 1 },

{ name: "type\_country\_year\_rating\_idx" }

)

Složený index pro dopravní přestupky:

db.Indian\_Traffic\_Violations.createIndex(

{

Violation\_Type: 1,

Registration\_State: 1,

Driver\_Age: -1,

Alcohol\_Level: -1

},

{ name: "ViolationType\_State\_Age\_Alcohol\_idx" }

)

Index pro stock\_prices podle data a objemu:

db.stock\_prices.createIndex({ Date: 1, Volume: -1 })

## Testování replikace: výpadek jednoho shardu při zapnutém mongos routeru

Tento test byl proveden tak, že byl simulován výpadek kontejneru shard2. Přes mongos byl poté otestován fallback přístup ke zbývajícím shardům – dotazy pokračovaly bez přerušení -> potvrzuje správné nastavení replikace a shardingu.

Závěr  
Tato semestrální práce se zaměřila na návrh, implementaci a analýzu plně funkčního shardingového clusteru dokumentové databáze MongoDB. Cílem bylo vytvořit automatizované a přenositelné řešení, které lze spustit jediným příkazem a které bude zároveň snadno škálovatelné, bezpečné a připravené pro reálné použití v datově náročném prostředí.

Z technického hlediska bylo dosaženo vysoké úrovně komplexnosti – řešení zahrnuje tři shardy s replikacemi, centrální konfigurační servery, dvojici mongos routerů a NGINX load balancer. Veškeré komponenty jsou nakonfigurovány prostřednictvím Docker kontejnerů, jejichž logiku zajišťují samostatné Dockerfile soubory, shellové skripty a konfigurační šablony. Nasazení systému probíhá zcela automaticky pomocí docker-compose, bez nutnosti ručních zásahů.

Z pohledu datové vrstvy bylo do systému integrováno několik reálných datasetů z různých domén (streaming, finance, doprava), které byly analyzovány pomocí Pythonu a knihoven pandas, matplotlib a seaborn. Byla provedena validační kontrola schémat, vytvořeny indexy, realizována základní i pokročilá filtrace a agregace dat. Systém tak slouží nejen pro ukládání, ale i pro analytické zpracování.

Mezi hlavní přínosy patří:

Modularita a rozšiřitelnost – přidání nového shardu nebo uzlu je otázkou úpravy několika řádků v konfiguračních souborech;

Automatizace – kompletní nasazení a vytvoření schémat probíhá bez ručního zásahu;

Bezpečnost – systém je zabezpečen pomocí autentizace a keyfile, přístup je směrován výhradně přes load balancer;

Reálné využití – řešení je vhodné pro prostředí BI/analytiky, kde se pracuje s rozsáhlými datovými sadami.

Kriticky je třeba zmínit, že složitost inicializačních skriptů a konfigurací může při dalším rozšiřování systému způsobovat problémy. Dále by bylo vhodné vytvořit separovaný datový init kontejner pro naplnění databáze, čímž by se zvýšila přenositelnost i robustnost systému.

Celkově lze práci hodnotit jako hluboce zpracovanou a prakticky orientovanou. Byla ověřena funkčnost jak v oblasti databázové architektury, tak v oblasti reálné analytické práce s daty. Systém může sloužit jako základ pro další rozšíření – například přidání rozhraní REST API, front-end dashboardu, nebo pokročilejších analytických algoritmů.

Zdroje

[1] MongoDB. eBay: Building mission-critical multi-data center applications with MongoDB. Dostupné z: <https://www.mongodb.com/blog/post/ebay-building-mission-critical-multi-data-center-applications-with-mongodb>

[2] MongoDB. Autodesk: Transforming design and engineering with MongoDB Atlas. Dostupné z: <https://www.mongodb.com/solutions/customer-case-studies/autodesk>  
[3] MongoDB. Toyota Connected: Building the future of mobility services on MongoDB Atlas. Dostupné z: <https://www.mongodb.com/solutions/customer-case-studies/toyota-connected>  
[4] MongoDB. Toyota Financial Services: Accelerating Digital Transformation with MongoDB Atlas. Dostupné z: <https://www.mongodb.com/pt-br/library/customer-case-studies/toyota-financial-ser>