Koncept Data Lakehouse pro zpracování medicínských dat

Lukáš Moučka[1]

1 Úvod

Diplomová práce se zabývá problematikou nového typu úložiště Data Lakehouse. V teoretické části je zmíněna evoluce jednotlivých typů úložišť. Hlavním cílem je ověření vhodnosti tohoto úložiště pro oblast medicínských dat, kterou aktuálně řeší MRE platforma – provozována a vyvíjena na FAV ZČU. Pro ověření vhodnosti byla vyvinuta aplikace implementující úložiště Data Lakehouse za pomocí open-source projektu Delta Lake. Aplikace byla úspěšně dokončena a na základě testování ukazuje vhodnost použití úložiště pro heterogenní medicínská data.

2 Analytická část

Pro pochopení konceptu úložiště Data Lakehouse jsou zevrubně analyzovány typy úložišť, které tomuto typu předcházely. Jedná se o datové sklady a datová jezera. Datové sklady jsou efektivní pro ukládání strukturovaných dat, ale postupem času se ukázalo, že je nutné ukládat a následně pracovat se semi-strukturovanými a nestrukturovanými daty. Ukládání těchto heterogenních dat řeší datová jezera, ale zároveň neposkytují požadované klíčové vlastnosti. Hlavním úskalím je absence transakčního zpracování. Následuje zmiňovaný Data Lakehouse, který umí zpracovávat heterogenní data a poskytuje meta-transakční vrstvu, díky které podporuje transakce typu ACID a z toho plynoucí výhody (Michael Armbrust et al. (2020)). Mezi další výhody patří např. vývoj schéma (schema evolution), cestování v čase (time-travel) a podpora open-source sloupcově orientovaných formátů (Parquet, ORC nebo AVRO).

3 Výběr vhodné technologie pro budování úložiště

Pro ověření vhodnosti úložiště bylo nutné vyvinout aplikaci implementující Data Lakehouse za použití technologie, která poskytuje následující vlastnosti:

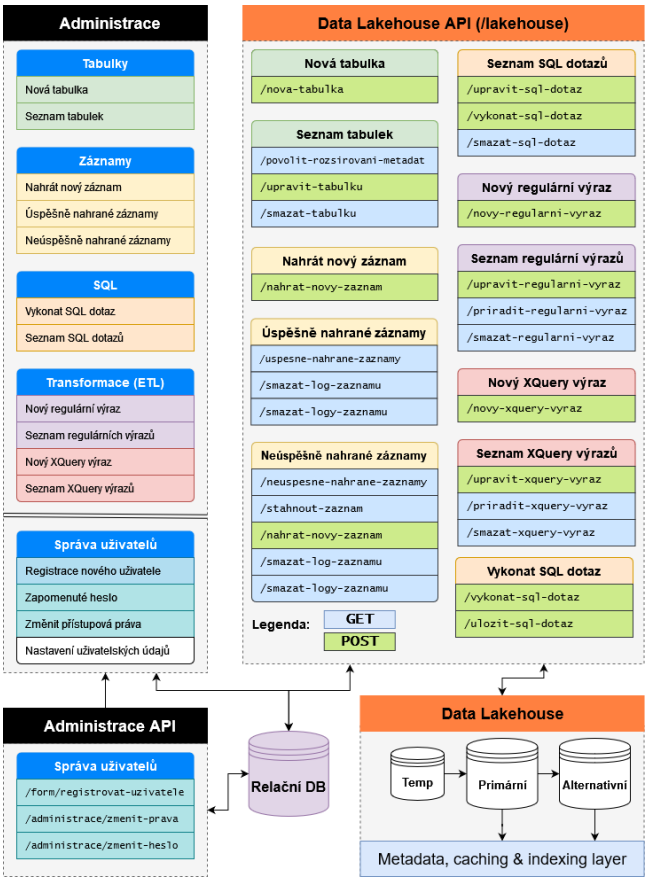
- extrakce metadat,
- transakční zpracování (ACID),
- možnost používání dotazovacího jazyka pro získávání informací o pacientech.

Do užšího výběru se dostaly tři open-source projekty: Delta Lake, Apache Hudi a Apache Iceberg. Každý z těchto projektů musel pokrývat zmíněné klíčové vlastnosti. Následně proběhlo zhodnocení projektů v několika aspektech a byl vybrán Delta Lake – vznikla tedy komplexní Spring Boot aplikace využívající Delta Lake. Aplikace je logicky rozdělena na dvě části. Administrace poskytuje kompletní správu uživatelů a sekce pro správu úložiště. Část úložiště poskytuje skrze REST API veškeré funkce pro jeho správu.

[1] student navazujícího studijního programu Inženýrská informatika, obor Informační systémy,
e-mail: lmoucka@students.zcu.cz

4 Architektura aplikace

Na obrázku 1 jsou viditelné obě části aplikace, poskytované REST API a komunikace s úložištěm Data Lakehouse. Relační databáze zde slouží jako datový slovník.



Obrázek 1: Schéma architektury aplikace.

5 Dosažené výsledky

Testováním aplikace byly ověřeny všechny požadované vlastnosti a vhodnost úložiště pro oblast medicínských dat. V těchto klíčových vlastnostech se vyrovná aktuálnímu řešení, ale poskytuje některé funkce navíc. Umožňuje aktualizaci schématu – např. rozšiřování struktury metadat dat (schéma) tabulky na základě vkládaných záznamů. Je možné agregovat informace o pacientech – vytváření reportů nebo podkladů pro souhrnné statistiky. Pro strojové učení je zde funkce pro cestování v čase (time-travel) a nechybí ani možnost vytváření klonů tabulek (shallow a deep klony).

Technologie byly zvoleny v souladu se současným řešením, takže by bylo potenciálně možné aplikaci integrovat do MRE platformy. V průběhu vývoje a testování vzniklo několik námětů na zlepšení implementace a v práci jsou také řádně popsány.

Literatura

Michael Armbrust and Ali Ghodsi and Reynold Xin and Matei Zaharia, C. (2020) *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics*, Databricks, UC Berkeley, Stanford University.