Posudek oponenta diplomové práce

Autor/autorka práce: **Lukáš Moučka**

Název práce: **Koncept Data Lakehouse pro zpracování medicínských dat**

Obsah práce

Práce se zabývá relativně novým přístupem ke způsobu uložení dat za účelem analytického zpracování - Data Lakehouse. V textu je nejprve představen tento koncept v kontrastu s předchozími přístupy (Data Warehouse, Data Lake). V další části následuje popis aktuálního stavu MRE platformy, která je na KIV provozována za účelem uchování medicínských dat a kterou student používá ke srovnání a zhodnocení navrhovaného řešení. V kapitole 4 student srovnává tři alternativní technologie, které lze použít pro základ nového řešení, a zkoumá u nich vhodnost dle několika sledovaných kritérií. Následuje kapitola implementace vlastního řešení a kapitola popisující jednotlivé use-case. Student dále poskytuje několik námětů na zlepšení řešení a vyhodnocuje, zda řešení je vhodné pro oblast medicínských dat.

Zkoumaná problematika je rozsáhlé téma a je obtížné sestavit dokument, aby zaznělo vše podstatné a zároveň text zůstal koherentní. Z mého pohledu je nešťastné uvádět kapitolu zabývající se procesem ETL až poté, co je tento klíčový koncept několikrát uváděn při formulaci architektur jednotlivých přístupů k uložení dat. Dále kapitola 2.5 FAIR data představuje důležitý koncept pro sdílení a zpracování dat, ale nepozoroval jsem, že by s ním student dále nějak pracoval.

Kvalita řešení a dosažených výsledků

Student vytvořil funkční prototyp aplikace dobře demonstrující zkoumaný přístup. Aplikace byla navrhována a testována v kontextu medicínských dat v XML formátu, nicméně řešení je obecné a bylo by možné ho použít i v jiných oborech.

Zvolená architektura aplikace dává v kontextu úlohy smysl. Aplikace momentálně obsahuje část obsluhující úložiště a část řešící administrativní obsluhu. Interně mezi sebou části komunikují přes RESP API a je tedy možné v budoucnu části oddělit bez větších komplikací.

Zdrojové kódy jsou vhodně strukturované. V dokumentu je popsána implementovaná funkcionalita a jako příloha práce je dodána uživatelská příručka pro nasazení a používání aplikace.

Drobným nedostatkem je momentálně používaná logika pro rozeznávání typů vstupních XML dat, která spoléhá na název root tagu dokumentu. Pro další použití bude potřeba použít komplexnější přístup.

Formální úroveň

V textu jsem narazil na několik překlepů, které by bylo možné nalézt běžnými nástroji pro kontrolu pravopisu. Text je vhodně prokládán ilustračními obrázky a výpisy.

V dokumentu je sice přiložen seznam zkratek, nicméně není lexikograficky řazen a je tedy obtížné ho použit. Zkratky jsou v textu vhodně vysvětleny.

Práce s literaturou

Autor používá 43 zdrojů, které jsou k řešené problematice relevantní a aktuální. Literatura je řádně odkazované v textu. Bylo by vhodnější číselné značky řadit dle prvního výskytu v textu.

Splnění zadání

Zadání bylo splněno bez výhrad.

Dotazy k práci

V textu (str. 42) je uvedeno, že tři **nejlepší** open-source projekty pro budování Data Lakehouse jsou Delta Lake, Apache Hudi, Apache Iceberg. Dle jakého kritéria nebo jakým způsobem je toto hodnoceno?

Dle kapitoly 5.3.1 (str. 58) používáte root tag XML dokumentu k třídění vstupních dat. Jak byste řešil situaci, kdy různé typy dokumentů mají shodný root tag?

Jakým způsobem by do řešení mohla být implementována podpora pro jiná vstupní data než XML, například JSON nebo CSV?

Navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

V Plzni 5.6.2023 Ing. Martin Kryl