Posudek oponenta diplomové práce

Autor/autorka práce: **Josef Bozděch**

Název práce: **Možnosti analytického rozšíření úložiště Data Lakehouse**

Obsah práce

Práce se zabývá rozšířením prototypu datového lakehouse (Moučka, 2023). V textu je nejprve představen výchozí stav a vysvětlena role analytických a statistických metod. Následně autor navrhuje několik změn v lakehouse, popisuje jejich implementaci a přidává sadu testovacích scénářů pro ověření funkcionality. Poté jsou diskutovány výsledky a řešení je porovnáno s existujícím řešením typu data lake provozovaném na KIV.

Text práce má vhodnou strukturu odpovídající tématu, nicméně některé kapitoly působí obsahově málo naplněné. Především kapitola *6 – Návrh změn* by z mého pohledu měla detailněji rozebírat jednotlivé body, poukazovat na možná řešení a lépe vymezit, co bude autor vlastně implementovat.

Kvalita řešení a dosažených výsledků

Řešení je funkční, ale obsahuje několik bodů, které považuji do budoucna za problematické.

Původně aplikace umožnila nahrávat XML soubory a rozřazovat je dle root tagu do různých tabulek. Autor přidal možnost nahrávání souborů datových typů CSV a JSON, ale žádný mechanismus třídění není vytvořen. Jako problematické vidím i automatické rozlišování datového typu souboru dle koncovky v názvu souboru (místo důslednějšího rozlišení dle obsahu, nebo volbou uživatelem při nahrání).

Aplikace nyní umožňuje zobrazit statistiky o atributech (viz obr. 7.8, str. 46), ale provedení působí jako narychlo implementované bez řádného navržení a rozmyšlení. Údaje po stranách částečně duplikují údaje v tabulce, některé číselné hodnoty jsou symbolicky vypisované jako texty, histogram hodnot není logicky seřazený a ani není vhodně zvolený typ grafu pro zobrazení charakteru hodnot spojité proměnné.

Výsledky logistické regrese nejsou předány v uživatelsky přívětivém formátu. Jednak není patrné, pro jakou hodnotu vysvětlované proměnné jsou uvedené koeficienty, jednak zobrazení odds ratios do běžného sloupcového grafu je nezvyklé a obtížně interpretovatelné. Zároveň se obávám, že se zvolenou knihovnou pro tvorbu grafů bude problematické sestavit lepší vizuál pro tento případ.

Pro testovací scénáře by bylo potřeba přidat i testovací data a detailněji popsat, jak má uživatel jednotlivé činnosti vykonat v GUI aplikace.

Postrádám testování rychlosti výpočtu statistik na netriviálním množství dat. Když jsem aplikaci zkoušel na svém počítači, doba vyhodnocení statistik skrze GUI se pohybovala v řádu vteřin pro tabulku 20x4. Pokud by většinu tohoto času trval samotný výpočet, jednalo by se o varovný signál pro reálné nasazení.

Formální úroveň

V textu se občas vyskytují krátké pasáže, které působí spíše jako text z propagačního materiálu než technická práce popisující existující řešení - například spojení „*Díky […] řadě integrací s vašimi oblíbenými nástroji […] je Thymeleaf ideální …*“ Autor občas střídá české a anglické ekvivalenty technických pojmů (např. potrubí/pipeline), což může být matoucí.

U vložených kódů by bylo vhodné důkladněji vysvětlit, co tam čtenář má vidět. Ukázka XSD (str. 25) není vůbec komentovaná, či obrázek s rozhraním WEKA Explorer (str. 48) není z textu nijak referován a vysvětlen.

Textová část práce je jako celek srozumitelná a ucelená.

Práce s literaturou

Autor pracuje se zdroji, které jsou aktuální a relevantní k tématu.

Splnění zadání

Zadání bylo splněno.

Dotazy k práci

- Jakým způsobem by bylo mohlo být řešeno nahrání a zpracování DICOM souboru při rozšiřování aplikace?
- Jak složité by bylo přidat funkcionalitu umožňující definovat a počítat odvozené atributy - např. dobu trvání události spočtenou jako rozdíl atributů začátku a konce události?

Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 3.6.2024 Ing. Martin Kryl