FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Martin Oharek
Student:	Vítězslav Hušek
Název práce:	Tvorba stage vrstvy datového skladu (DWH) na bázi metadat v prostředí datové platformy Databricks v cloudu - případová studie
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	9. června 2024

Hodnotící kritéria

1. Splnění zadání

- ▸ [1] **zadání splněno**
  [2] zadání splněno s menšími výhradami
  [3] zadání splněno s většími výhradami
  [4] zadání nesplněno

Zadání považuji za splněno bez výhrad.

2. Písemná část práce 80 /100 (B)

Předložená práce je dobře a přehledně členěna do jednotlivých kapitol, které na sebe logicky navazují. Práce obsahuje teoretickou a praktickou část, doplněnou o "nepovinné" části zahrnující subjektivní zamyšlení autora nad danou problematikou (např. kapitola Diskuze). Jazyková i typografická stránka odpovídá normě. Jako mírný nedostatek shledávám drobné faktické nepřesnosti a překlepy (např. typo v kapitole 1.3 - místo data lake by mělo být delta lake, dále např. označení Unity Catalogu jako uložiště v kapitole 4.4) a občasné nekonzistence mezi používání anglicismů a jejich přeložených ekvivalentů (např. data marts - datová tržiště).

3. Nepísemná část, přílohy 90 /100 (A)

Kladně hodnotím, že autor zpracoval doplňující diagramy sám, než aby je pouze převzal z existujících zdrojů, což napomáhá si lépe osvojit studovanou oblast. Je skvělé, že i přes vedlejší zátěž, která vždy souvisí s experimentováním v cloudových prostředích (založení účtů, monitorování nákladů,...), se toho student nezalekl a vyzkoušel si kromě implementace samotné logiky i potřebnou administrativní část, jako např. skriptování infrastruktury v CloudFormation. Praktická část práce obsahuje různorodé a důkladné srovnání dvou frameworků pro zpracování dat, které autor porovnává v úloze

implementace stage vrstvy datového skladu. Nejedná se pouze o strohý výčet vlastností, ale autor se také zamýšlí nad posazením frameworků do různých fází vývojového cyklu datového skladu a jejich vhodnost pro jednotlivé činnosti.

4. Hodnocení výsledků, jejich využitelnost 85 /100 (B)

Téma migrace do cloudu v oblasti DWH je aktuálně velmi probírané téma a pravdou je, že není toto území proti klasickému řešení datových skladů dostatečně prozkoumáno. Závěry práce jsou legitimní a rozhodně mohou pomoci firmám k rozhodnutí při formování správné metodiky a SOTA technologického stacku pro migraci datových skladů. K perfektnímu porovnání mi ještě trochu chybí zvážení nedostatků výpočetního engine samotné platformy Databricks (Apache Spark) při porovnání s klasickými relačními databázemi. To beru jako jeden z bodů, na které se dá navázat a rozšířit závěry zmíněné zde v rámci DP.

Celkové hodnocení 87 /100 (B)

Z obsahově bohaté práce je patrné, že student věnoval jejímu vypracování mnoho času. Přes mírné nedostatky týkající se drobných nepřesností je předložená práce nadprůměrně kvalitní.

Otázky k obhajobě

1) Věděl byste, zdali dbt framework dokáže generovat více modelů z jediného "vzoru" (template) jen na základě metadat? Např. mám jedno SQL pro model a chci jej spustit n-krát, pokaždé s jinými sloupci a datovými typy.
2) Napadla by Vás nějaká nevýhoda/vlastnost Spark/Spark SQL (v rámci Databricks), kterou má v porovnání s klasickými relačními databázemi (např. PostgreSQL)?

Instrukce

Splnění zadání

Posuďte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.