Posudek oponenta bakalářské práce

Autor/autorka práce: **Dominik Jež**

Název práce: **Datová kostka pro analýzy výzkumu a vývoje inovací pro datový sklad ZČU**

Obsah práce

Text se skládá z pěti velkých kapitol. V první z nich autor vysvětluje pojmy související s informačními systémy a dodává stručný historický přehled databázových modelů a popis normálních forem relačních databází. Následně uvádí teorii týkající se budování datových skladů. Ve třetí kapitole je popsán registr CEP a jeho API, které je hlavním datovým zdrojem práce. V následující kapitole autor popisuje vlastní zpracování dat a stavbu datové kostky. V poslední kapitole je popsána možnost využití kostky v analytickém nástroji Power BI a popsán testovací scénář k ověření správnosti výstupů.

Z mého pohledu první kapitola příliš nezapadá do kontextu práce, resp. poznatky zde uváděné autor později v textu nevyužívá. Například po definici normálních forem je v textu jen strohá zmínka, že datové sklady nemusí splňovat normální formy, protože se s tím lépe pracuje (s. 10). Kapitola ani nepůsobí uceleně, protože zásadní pojmy k definici normálních forem, jako klíčový atribut nebo kandidátní klíč, nejsou vysvětleny.

V textu v některých místech postrádám diskuzi zvoleného řešení – například zamyšlení u návrhu struktury datové kostky, zda je užitečné uchovávat texty cílů a názvy projektů, či zda by nebylo vhodnější dimenze oborů a oborových skupin zpracovat hierarchicky.

Kvalita řešení a dosažených výsledků

Hlavním výstupem je sada PL/SQL skriptů sloužící jako datová pumpa pro zpracování dat z registru a následnou transformaci do datové kostky. Skripty jsou dostatečně komentované a vhodně členěné do souborů.

Dalším výstupem je Power BI model, do kterého byla nahrána datová kostka bez dalších transformací a následně nad ní sestaveny dva dashboardy. Tento výstup považuji za kvalitně zpracovaný.

Posledním identifikovaným výstupem je sada dotazů do původních dat, které mají sloužit k ověření, že data jsou po transformaci stále v souladu s původními daty. Výsledky autor srovnává s výstupy v Power BI projektu. Z neznámého důvodu není ve skriptech implementován test na některé položky jako celkový počet projektů, či počet kategorií.

V práci bych uvítal přidání přehledu mapování vstupních dat na atributy datové kostky a případný způsob transformace atributu (např. sloučení klíčových slov do jednoho řetězce).

Formální úroveň

V textu jsem narazil na několik překlepů a pro mě zvláštních českých ekvivalentů k anglickým termínům, ale celkově je text dobře čitelný. Použité číslování citačních značek nikoliv podle pořadí výskytu, ale dle abecedního řazení záznamu považuji za nevhodné.

Práce s literaturou

Použitá literatura je relevantní k řešenému problému.

Mám výhradu k tabulce 3.2: OLTP vs OLAP (s. 14), kterou autor značí jako přejatou. V původním zdroji jsem nalezl podobnou tabulku, nicméně její obsah se liší. Autor uvádí, že odezva dotazu je pomalá u OLTP a rychlá u OLAP. V původním zdroji je informace opačná, OLTP má odezvu v řádech milisekund, pro OLAP to jsou „sekundy, minuty nebo hodiny dle množství dat“. Dále v tabulce autor přidal položku počty záznamů, kde pro OLTP uvádí desítky. Běžná OLTP ale bude pracovat s o několik řádů vyšším množstvím záznamů.

Splnění zadání

Splněno bez výhrad.

Dotazy k práci

S ohledem na plánované využití datové kostky, považujete za vhodné udržovat klíčová slova vámi navrženým způsobem, tj. sloučené do jednoho textového atributu?

Co je myšleno srovnávací metrikou „Počty záznamů“ v tabulce 3.2?

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

V Plzni 30. 5. 2022 Ing. Martin Kryl