FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Kristián Vadkerti
Student:	Bc. Stanislav Němec
Název práce:	Případová studie využití NoSQL databázi pro část datového skladu VZP
Obor / specializace:	Webové a softwarové inženýrství, zaměření Softwarové inženýrství
Vytvořeno dne:	January 28, 2022

Hodnotící kritéria

1. Splnění zadání

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

Zadání bylo splněno s menšími výhradami.

Seznam uvedených výhrad:

1) V ZP použité faktové tabulky obsahovaly vždy pouze jednu „PARTITION" s určitým faktorem. Každý dotaz provedl načtení vždy všech dat z faktové tabulky. Uvedený případ neodpovídá typickým dotazům prováděným analytiky v rámci práce s BI. Obvykle se pracuje s 1 až N „PARTITIONs" v rámci celkového počtu M „PARTITIONs".
V ZP použitá varianta přistupuje k datům faktové tabulky „FULL TABLE SCAN".

2) Použitá verze databáze IMPALA, nebyla dle certifikace OAS.5.9, podporovanou verzí. Z uvedeného můžou v budoucnu plynout problémy s dotazováním přímo z OAS.
Danou připomínkou bych rád pouze poukázal, že by se mohlo jednat o potencionální problém. Vždy je vhodné provádět integraci nad verzemi, které podporuje daný výrobce.

2. Písemná část práce 82 /100 (B)

Písemnou část práce považuji za přiměřeně provedenou. Kapitoly na sebe logicky navazují a provedou čtenáře problematikou, kterou se práce zabývá.

Seznam uvedených výhrad:

1) V tabulce 5.2, byla zmíněna velikost souborů „Parquet" pro technologii IMPALA. V dané práci mi chybí obdobná informace využitého místa pro technologii ORACLE a případně i

dosahovaný kompresní poměr ORACLE vs IMPALA.
Velikost souborů, ze kterých jsou data čtena, bude mít vliv na celkovou dobu potřebnou pro zpracování samotného dotazu.

2) V ZP použitá metrika měření, doba odezvy, je hlavním kritériem pro srovnání jednotlivých řešení. V uvedené práci mi chybí i jiné - doplňkové kritérium. Například počet výstupních záznamů dotazu.
Na základě vlastního měření nad dodanou sadou dotazů jsem se setkal s problémem, že stejný dotaz v databázi IMPALA, vracel nižší počet záznamů oproti databázi ORACLE.
Uvedená skutečnost měla pochopitelně vliv na celkovou dobu zpracování dotazu a i na správnost samotného výsledku.

3) U zvolené metody měření na základě průměrování doby běhů jednotlivých dotazů, mi chybí zmínka k použité HW specifikaci. Také chybí informace o velikosti alokovaných zdrojů, které může dané řešení pro svoji práci využít. N násobné spouštění jednotlivých dotazů a jejich průměrování, mohlo způsobit u některých dotazů využití, „databaze caching" a tím ovlivnit výsledky měření.
V rámci dané práce byly HW zdroje srovnatelné, ale alokované paměťové zdroje, které mohly jednotlivé řešení pro svoji práci využít, byly výrazně rozdílné ve prospěch databáze ORACLE.

4) U dotazu označeného jako „problematický" pro databázi ORACLE, mi v rámci dané práce chybí bližší analýza důvodů, které mohly vést k danému problému.
V dodané sadě dotazů neměl být žádný neoptimálně fungující dotaz.

V rámci provedené práce, bych rád ocenil konstruktivní přístup k řešením vzniknutých problémů, se kterými se autor musel setkat v rámci části „Realizace vybraných řešení" popsané v kapitole 4. Využití technologie Docker a nástroje pySpark, použitého v rámci transformací, byla hezkou ukázkou reakce na vzniknutý problém.
Citační etika nebyla porušena a zdrojů je dostatečný počet a jsou všechny relevantní tématu práce.

3. Nepísemná část, přílohy 88/100 (B)

Nepísemná část práce, dále jen SW dílo, je svým rozsahem a charakterem postačující k zadání ZP.

Seznam uvedených výhrad:

1) V SW díle, chybí skripty pro napočet statistik pro obě technologie.
V písemné části dokumentace je zmínka o nutnosti napočtu statistik, ale bez skriptů, nelze provést relevantní opakovatelnost experimentu.

2) V rámci dodaných skriptů použitých pro dotazování, jsem narazil na drobné rozdíly u srovnávaných technologií.
Jednalo se zejména o poškozenou češtinu v rámci importu do NoSQL databázi, kterou autor vyřešil použitím upravené podmínky „like" místo „=".

4. Hodnocení výsledků, jejich využitelnost 80/100 (B)

Výsledky této práce splnili očekávání. Daná práce umožnila praktické odzkoušení jiného přístupu ke zpracování dat v rámci aktuální struktury DWH řešení a poukázala na některé silné a slabé stránky jednotlivých řešení.

Uvedené výsledky jistotně poslouží jako základ pro další studium využití NoSQL technologie v prostředí VZP ČR. Po provedení dodatečných testů, zejména prověření vlivu horizontálního škálování na celkovou dobu běhu, může být NoSQL řešení v budoucnu vhodným rozšířením stávající implementace.

Celkové hodnocení

85 /100 (B)

Práci, jako celek, hodnotím kladně a navrhuji k obhajobě známku B.
Rád bych zde také ocenil aktivitu a snahu v rámci řešení dané práce. I přes to, že jsem měl k práci několik drobných připomínek, připisuji je nedostatku praktických zkušeností autora, které zatím neměl možnost nabýt v rámci dosavadní praxe. Pevně věřím, že připomínky zde zmíněné, poslouží jako inspirace k dalšímu rozvoji v rámci budoucího směrování.

Otázky k obhajobě

1) Lze v NoSQL technologii IMPALA doporučit použití UPDATE?
2) Vyjmenujte několik nevýhod použití souborového systému parquet ?
3) Uveďte několik SQL funkcí, u kterých je NoSQL přístup výhodnější?

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.