



Posudek oponenta závěrečné práce

Oponent práce: Ing. Petr Košvanec
Student: Jan Chybík
Název práce: Analýza datových toků v Azure Data Factory
Obor / specializace: Teoretická informatika
Vytvořeno dne: 6. června 2022

Hodnotící kritéria

1. Splnění zadání

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

Cíle jsou vydefinovány v souladu se zadáním a splněny bez výhrad. Zadání považuji za těžší, komplexita práce a množství dodaného kódu tomu také odpovídá.

2. Písemná část práce

80/100 (B)

Práce je trochu delší, než musela být. Zkrátil bych některé definice (výčet gramatik a jazyků jimi generovaných) a kapitoly (5.1 Ways of generating lineage). Práce také působí dojmem, že úvodním definicím bylo věnováno nepoměrně více času, než posledním dvou kapitolám, které trochu kvapí (např. zapojení do platformy Manta je komentováno stroze, přínos DataflowQueryService taktéž, použité technologie kromě ANTLR nejsou komentovány). Jinak je její obsah ale velmi přínosný. Oceňuji korektnost definic, užité značení a praktické ukázky pro rychlejší pochopení a lepší čtivost. V definicích jsem nenašel výrazné faktické chyby. Pár faktických nepřesností se vyskytlo v kapitolách 5 a 7. Stejně tak práce obsahuje nadprůměrné množství překlepů (desítky) a občasných chyb z nepozornosti a místy působí dojmem, že další autorovo čtení nebo použití nástrojů pro korekci by jí prospělo. Úroveň angličtiny a gramatiky je slušná a výběr jazyka nepůsobí rušivě. Práce je tématicky vystavená velmi dobře, kapitoly rozumně navazují. Práci by prospěla trochu lepší grafika, některé diagramy by byly ve vhodném barevném provedení čitelnější a i ukázky kódu mohly být vyobrazeny v editoru, který podporuje syntaxi Javy/ANTLR, aby nepůsobily tak jednoduše. Práce je jinak typograficky v pořádku a zmíněné diagramy jsou hodnotné. Oceňuji množství tištěných zdrojů, které autor použil i konkrétní volbu titulů. Citace jsou v pořádku a i v práci je jasně odděleno, co je vlastním přínosem autora.

3. Nepísemná část, přílohy

90 /100 (A)

Přiložený kód je rozsáhlý a ukazuje na složitost a časovou náročnost implementace. Kód je rozumně členěn do tříd respektujících Single responsibility pattern a velikost metod je ve většině případů vhodná. Kód se opírá o poctivý design, na němž je v několika místech znát informované rozhodování nad architekturou. Kognitivní složitost tak není příliš vysoká. Oceňuji výskyt designových vzorů (např. Visitor) a zajímavých algoritmů (např. Topologické uspořádání). Také kód ANTLR je vhodně strukturovaný a volba názvů usnadňuje čtení. Pokrytí kódu testy je více než 70%, testy jsou vhodně zvoleny (např. porovnávání výstupních grafů vůči očekávaným). V řadě tříd včetně metod rozhraní chybí Javadoc dokumentace a jako celek působí nekompletně. Některé třídy mají nekonzistentní formátování.

4. Hodnocení výsledků, jejich využitelnost

87 /100 (B)

Výsledkem práce je poctivá analýza nástroje ADF, která poskytuje solidní vhled do většiny definičních souborů ADF objektů. Na tomto základu je vystaven rozumný design a implementován prototyp, který se poměrně brzo dočká produkčního nasazení. Cítím ale, že bude třeba na skeneru ještě vytrvale pracovat a pokrýt řadu dalších transformací, proprietárních funkcí, zlepšit interpretaci výrazů a doimplementovat podporu některým chybějícím ADF objektům. Současná implementace je kvalitní a rozšiřitelný základ.

Celkové hodnocení

87 /100 (B)

Práce byla komplikovanější než srovnatelné práce zaštitěné firmou Manta a jako taková si vyžádala i více úsilí a času. Písemná práce byla obsahově velmi kvalitní, mezi její kvality patří především solidní základ definic, uvedení do problému a analýza nástroje ADF. Výtky jsem měl především k jednodloti zpracování a množství překlepů, které práci zbytečně ubírají na kvalitě. Dodané SW dílo je rozsáhlé a poskytuje velmi dobrý a využitelný základ pro další implementaci.

Otázky k obhajobě

- 1) V sekci ANTLR uvádíte příklad volání funkce hostitelského jazyka `{isB(getCurrentToken())}`. Kde je specifikováno, jaké třídy obsahují tyto metody? Kdybychom chtěli zavolat metodu nějaké specifické Java knihovny, jak bychom to zařídili?
- 2) Podporuje nástroj ADF dynamické (tedy za běhu programu) určení sloupců, se kterými Aktivity nebo Transformace pracují? Pokud ano, jak to ovlivňuje metadata, která z nástroje může skener získat? Jaké možnosti má skener obecně, pokud nejsou k dispozici příslušná metadata?
- 3) Jaký je vztah ADF aktivity a transformace?
- 4) V popisu generování datových toků pro sloupce transformací je uvedeno "If there is ColumnDefinition, a new column is created, and it is not connected to anything.". Jakým způsobem je tedy například vytvořen (a v metadatach reflektován) nový sloupec, jenž vznikl libovolnými operacemi ze dvou jiných vstupních sloupců?

Instrukce

Splnění zadání

Posudte, 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, posudte 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 posudte, 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. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, 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.