FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Pavel Karol
Student:	Kristina Zolochevskaia
Název práce:	Paralelizace ETL procesů DW ČVUT – případová studie
Obor / specializace:	Webové a softwarové inženýrství, zaměření Softwarové inženýrství
Vytvořeno dne:	2. června 2023

Hodnotící kritéria

1. Splnění zadání

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

Cíle této práce jsou částečně společné s konkurenční BP studenta Adama Marhefky.
Společně s druhou prací studentka identifikuje požadavky na paralelizaci ETL procesů Datového skladu ČVUT (dále jen DS) a společný je také výběr části zpracování, která se má optimalizovat do paralelního běhu k porovnání výkonnosti obou řešení.
Rozdílnou částí cílů této BP je vytvoření Proof of Concept (POC) analýzu dostupných open-source řešení na trhu ETL nástrojů a implementace vybraného řešení na základě této analýzy.
Všechny části předložené BP jsou precizně zpracované a zadání práce bylo splněno v plném rozsahu.

2. Písemná část práce 95 /100 (A)

Práce je tvořena 7 kapitolami, úvodem a závěrem. Rozsahem odpovídá náročnému, širokému zadání. Má jazykově odpovídající úroveň a je po formální stránce v pořádku. Práce je psána v angličtině, tudíž kvalitu textu nejsem schopen posoudit. Nicméně je logicky a přehledně strukturována s využitím dostatečného množství relevantních elektronických i knižních zdrojů.
Kapitoly jsou autorkou rozděleny do dvou částí – Teoretická část a praktická, implementační část.
Teoretická část je obsažena v kapitole 1 a 2, ve kterých autorka obecně popisuje datové sklady a pojem ETL.
V praktické části, tvořené zbylými kapitolami 3 až 7, se studentka zabývá postupně současným stavem ETL procesů DS ČVUT, analýzou požadavků zadání práce, průzkumem

trhu ETL nástrojů s výběrem vhodného řešení pro splnění zadání práce, následnou implementací vybraného SW včetně splnění veškerých nutných předpokladů a nastavení SW. V poslední krátké kapitole 7 hodnotí výsledky testování svého řešení, jeho výhody a nevýhody, včetně zmínění potenciálních vylepšení do budoucna. Upozorňuje také na náročnější požadavky na kapacity DS při prvotní implementaci řešení a na nutnost údržby kódu.

3. Nepísemná část, přílohy 95 /100 (A)

Přílohy bakalářské práce obsahují veškeré kódy, které studentka vytvářela a screeny z vybrané aplikace Apache Airflow. Implementované řešení je plně funkční a otestované.

4. Hodnocení výsledků, jejich využitelnost 90 /100 (A)

Hlavním přínosem této práce je POC analýza, průzkum trhu a vytvoření řešení s využitím open source nástroje. Bereme-li v potaz konkurenční BP, tato práce může mít velký pozitivní dopad na DS, potažmo na celou ČVUT. Zrychlení zpracování umožní aktualizovat DS denní frekvencí a tím zpřesnit výstupy pro fakulty – např. sestavy a fakultní weby.

Celkové hodnocení 92 /100 (A)

Práci navrhuji hodnotit klasifikačním stupněm A. Teoretická i praktická část je zpracována velmi kvalitně a bez výhrad. Přínos práce hodnotím jako velký z pohledu DS a jeho odběratelů na úrovni fakult i rektorátu ČVUT.

Otázky k obhajobě

Jak se Apache Airflow rozšiřuje o nové typy datových zdrojů?
V čem se liší knihovna Pandas od Dask, Vaex nebo Modin, zkoušela jste i tyto knihovny?

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.