

Hodnocení vedoucího závěrečné práce

České vysoké učení technické v Praze

Fakulta informačních technologií

Student: Bc. Jakub Koza
Vedoucí práce: RNDr. Petr Škoda, CSc.
Název práce: Interactive Cloud-Based Platform for Parallelized Machine Learning of Astronomical Big Data
Obor: Webové a softwarové inženýrství

Datum vytvoření: 4. 6. 2017

Hodnotící kritérium:	Způsob hodnocení - následující škálou 1 až 5:
1. Náročnost a další komentář k zadání	<u>1=mimořádně náročné zadání,</u> 2=náročnější zadání, 3=průměrně náročné zadání, 4=lehčí, ale ještě dostatečně náročné zadání, 5=nedostatečně náročné zadání
Popis kritéria: Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)	
Komentář: Zadání práce je velmi náročné, je částí projektu řešeného na Astronomickém ústavu AVČR v Ondřejově v rámci grantu MŠMT pro podporu aktivit COST "Applications of Artificial Intelligence in Astronomy". Student se musel detailně seznámit se základy astronomické spektroskopie a s nástroji a technologiemi Virtuální observatoře. Musel pochopit principy zpracování a analýzy astronomických spekter i základy strojového učení a předzpracování dat pro ně, stejně jako funkci použitých modulů systému VO-CLOUD již dříve vyvinutých v rámci předchozích bakalářských a diplomových prací jiných studentů. Toto vše bylo nezbytné aby mohl realizovat integraci několika samostatných subsystémů navíc používající odlišné programovací jazyky (Java EE, Python 3.0 a SCALA) a mnoho souvisejících SW technologií a knihoven (např. Jupyter, SPARK, Wildfly, UWS, SSAP, Matplotlib, Websocket, YARN, Astropy).	
Hodnotící kritérium:	Způsob hodnocení - následující škálou 1 až 4:
2. Splnění zadání	<u>1=zadání splněno,</u> 2=zadání splněno s menšími výhradami, 3=zadání splněno s většími výhradami, 4=zadání nesplněno
Popis kritéria: Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.	
Komentář: Zadání práce bylo splněno beze zbytku. Za nejcennější výstup považuji úspěšnou integraci Jupyter notebook serveru s individuálními Docker kontejnery a implementaci vizualizace velkého množství spekter na serveru při prohlížení v browseru podporujícím pouze JavaScript. Na rozdíl od podobných řešení vyvíjených např. pro archiv družice Gaia nebo archiv astronomických misí NASA nevyžaduje instalaci speciálního plug-inu na straně klienta a umožňuje tak v principu strojové učení gigantických objemů dat na běžném mobilním telefonu či tabletu. Kromě toho student musel ozkoušet i mnoho testovacích variant modulů s různými řešeními grafických knihoven a docker kontejnerů či HDFS importerů než našel optimální řešení.	
Hodnotící kritérium:	Způsob hodnocení - následující škálou 1 až 4:
3. Rozsah písemné zprávy	<u>1=splňuje požadavky,</u> 2=splňuje požadavky s menšími výhradami, 3=splňuje požadavky s většími výhradami, 4=nesplňuje požadavky
Popis kritéria: Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.	
Komentář: Práce má 63 stran textu, 4 strany bibliografie, 20 stran příloh a 8 stran tiráže a rejstříků. Svým rozsahem tedy plně splňuje požadavky na diplomové práce na FIT ČVUT.	
Hodnotící kritérium:	Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):
4. Věcná a logická úroveň práce	98 (A)
Popis kritéria: 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.	

Komentář:

Práce je logicky velmi dobře členěna, místy psána didaktickým způsobem s ohledem na čtenáře, který nemusí být expert na všechny použité technologie. Proto jsou krátce představeny v první kapitole. Druhá kapitola shrnuje předchozí stav systému VO-CLOUD spolu s popisem současného workflow. Třetí kapitola pak představuje nově použité technologie a shrnuje funkční a nefunkční požadavky na nově realizované vlastnosti systému. Vlastní těžiště práce spočívá v kapitole čtvrté, kde je detailně popisována implementace v zadání požadovaných modulů včetně návrhů na budoucí vylepšení (ty bych osobně dal do samostatné kapitoly a trochu více rozvedl). V příloze jsou pak vedle seznamu zkratk a popisu obsahu DVD ukázkové konfigurační soubory a zejména popis instalace celé infrastruktury systému VO-CLOUD.

Hodnotící kritérium:

Způsob hodnocení - bodové hodnocení 0 až 100 bodů
(známka A až F):

5. Formální úroveň práce

98 (A)

Popis kritéria:

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 14/2015, článek 3.

Komentář:

Formální úroveň práce je velmi vysoká. Srozumitelný výklad je doprovázen názornými schémata, ukázkami klíčových částí kódu a několika kopiemi obrazovek. Všechny termíny a zkratky jsou řádně vysvětleny. Práce je psána dobrou angličtinou bez zjevných překlepů, což svědčí o použití spell-checkeru i několikanásobného pečlivého přečtení. Typografická úroveň je rovněž vysoká, použití odlišných fontů je jasně definované, v pdf verzi jsou i barevně odlišeny klíčové kusy kódu.

Hodnotící kritérium:

Způsob hodnocení - bodové hodnocení 0 až 100 bodů
(známka A až F):

6. Práce se zdroji

98 (A)

Popis kritéria:

Vyjáďte se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:

Všechny použité zdroje jsou řádně citovány a myslím, že analýza dostupných pramenů je dostačující. Vzhledem k novosti problematiky je většina odkazů elektronických. V úvodní kapitole jsou mnohonásobně opakovány odkazy na referenční dokumentaci (např. pro Javu EE) shrnutou do jedné reference. Pokud by se psal odkaz na každou část jako samostatná reference, objem citací by podstatně vzrostl. I tak je citováno 35 pramenů, což svědčí o tom, že student provedl velmi rozsáhlý průzkum dostupné literatury i elektronických zdrojů.

Citace jsou uváděny podle citačních standardů včetně citací elektronických zdrojů, kde je i uveden okamžik čtení či prohlídky daného zdroje. V souladu s moderními trendy jsou uváděny i doi, popř. odkazy na ArXiv, u knih pak i ISBN.

Hodnotící kritérium:

Způsob hodnocení - bodové hodnocení 0 až 100 bodů
(známka A až F):

7. Hodnocení výsledků, publikační výstupy a ocenění

99 (A)

Popis kritéria:

Vyjáďte se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:

Všechny použité software je typu Open Source a celá práce je i k dispozici pod licencí GPL. Dosažené výsledky jsou vynikající. Celá platforma VO-CLOUD je světově unikátní technologie pro strojové učení astronomických spekter a v budoucnu i časových řad v cloudovém prostředí. Realizace vizualizace velkého objemu dat v běžném browseru a zejména integrace Jupyter notebooku běžícího na serveru je průkopnická (pro zadání práce byly školitelem konzultovány nedostatky stávajících řešení se SW experty několika současných projektů - např. archivu družice Gaia, či multi-mission archivu NASA IPAC). Systém VO-CLOUD byl již několikrát prezentován na mezinárodních workshopech o virtuální observatoři a na prestižních konferencích o astronomickém SW ADASS (v Sydney 2015) a Astroinformatics (Sorento 2016), jeho současná podoba - výstup hodnocené DP - se zaměřením na podporu Jupyteru a Sparku pak byla předváděna na zasedání koordinačního workshopu virtuální observatoře v Šanghaji v květnu 2017. Současně je publikována na úložišti Zenodo, ma přiděleno doi. V přípravě je i článek do impaktovaného časopisu Astronomy and Computing v souvislosti s řešením grantu MŠMT COST LD-15113.

Hodnotící kritérium:

Způsob hodnocení - nehodnotí se

8. Komentář o využitelnosti výsledků

Popis kritéria:

Uvedte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uvedte možnosti využití výsledků ZP v praxi.

Komentář:

Výsledkem práce je několik unikátních řešení technologie pro interaktivní analýzu velkých astronomických dat v cloudovém prostředí. VO-CLOUD bude v budoucnu používán pro rutinní strojové učení aplikované na velké archivy astronomických spekter a časem i světelných křivek v rámci řešení několika grantů na Astronomickém ústavu AVČR. Celý systém je publikován na serveru github (několik modulů) a v budoucnu uvažujeme o jeho distribuci formou skriptu a obrazů disků pro Docker Compose, což umožní i snadné nasazení na velké průmyslové cloudy typu Amazon. Kromě astronomického nasazení si lze představit nasazení tohoto systému pro jakoukoli práci s velkými jednodimenzionálními daty - např. obecné časové řady, laboratorní spektra či např. pro analýzu DNA.

Hodnotící kritérium:

Způsob hodnocení - následující škálou 1 až 5:

9. Aktivita a samostatnost studenta v průběhu řešení

9a:

1=výborná aktivita,
2=velmi dobrá aktivita,
3=průměrná aktivita,
4=slabší, ale ještě dostatečná aktivita,
5=nedostatečná aktivita

9b:

1=výborná samostatnost,
2=velmi dobrá samostatnost,
3=průměrná samostatnost,
4=slabší, ale ještě dostatečná samostatnost,
5=nedostatečná samostatnost

Popis kritéria:

Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven (9a). Posuďte schopnost studenta samostatné tvůrčí práce (9b).

Komentář:

Student pracoval velmi samostatně, dohledal si drtivou většinu použité literatury a dokumentace sám a zcela sám si i naprogramoval veškeré programy a skripty, stejně jako několikrát instaloval celý systém SPARK na několika serverech stelárního oddělení ASU AVČR na Ondřejově a udržoval je aktualizované. Pravidelně se účastnil všech konzultací (většinou jednou týdně) a reagoval operativně na e-mailovou korespondenci, Spolupracoval i s dalšími členy grantového projektu a aktivně přicházel s řešeními problematiky i před vlastním formálním zadáním diplomové práce. Několikrát byl i na Ondřejově na setkání širšího grantového týmu a vývoj systému VO-CLOUD konzultoval i s tvůrci jeho modulů A. Paličkou a K. Shakurovou, kteří jej v minulém roce použili na své diplomové práce - strojové učení velkých archivů spekter. Je třeba zdůraznit, že J. Koza s naším týmem spolupracoval již od doby řešení bakalářské práce. Současná diplomová je v podstatě jejím pokračováním řešící problémy v ní kdysi nastíněné, jejichž řešení tehdy nevypadalo reálně.

Hodnotící kritérium:

Způsob hodnocení - bodové hodnocení 0 až 100 bodů
(známka A až F):

10. Celkové hodnocení

100 (A)

Popis kritéria:

Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:

Student již ve své bakalářské práci vytvořil profesionální velmi komplexní aplikaci kterou při řešení předložené diplomové práce doplnil o zásadní funkcionality - vizualizaci velkých dat a interaktivní rozhraní na bázi serverové implementace Jupyter notebooku, to celé napojené na paralelizovanou infrastrukturu - systém HDFS a SPARK, což je problém velmi aktuální nicméně velmi málo dořešený i ve velkých komerčních cloudových systémech. Komplexní platformu zdařile integrující Python, Javu EE a Scalu pro SPARK a HDFS dobře zdokumentoval a publikoval na GitHubu. Připravil i prezentace předváděné školitelem na mezinárodní konferenci o Virtuální observatoři v Šanghaji. Velmi dbal i na bezpečnostní hledisko - kdy právě požadavky šifrované komunikace mezi rozdílnými moduly v různých programovacích jazycích způsobovaly nemalé komplikace při integraci. Student velmi aktivně spolupracoval na řešení grantového projektu a jeho propagaci i dilčích publikacích. Podle mého názoru je celé řešení systému VO-CLOUD světově unikátní s velkým potenciálem dalšího vývoje. Vše perfektně funguje a umozňuje uživatelsky příjemnou práci s velkými objemy dat v cloudovém prostředí. Proto práci uděluji známku A.

Podpis vedoucího práce: