Posudek oponenta diplomové práce

Autor práce: **Bc. Vojtěch Danišík**

Název práce: **Tvorba rozsáhlých úložišť patentových dat**

Diplomová práce se na 72 stranách zabývá problematikou tvorby úložišť pro patentová data a jejich vytěžování. V analytické části dokumentu diplomant seznamuje čtenáře s problematikou patentů, jejich klasifikací a kde a jaká patentová data zadarmo získat. V další kapitole obsáhle představuje databázovou problematiku. V této kapitole se bohužel vyskytují chybné informace. Např. výhodnou vlastností relační databáze je normalizace chybně definována takto: *„Normalizace je metoda, pomocí které lze rozdělit jednu informaci do několika bloků za účelem snížení velikosti.“* To rozhodně není hlavní účel normalizace. Analytická část je zakončena analýzou patentových dat jednotlivých zemí a rozhodnutím, jaké typy databází budou použity pro návrh datového úložiště.

Samotná implementace datového úložiště je popsána v realizační části dokumentu. Do tohoto úložiště jsou ukládána jen „očištěná“ patentová data sadou jednorázových programových nástrojů, které vznikly v rámci řešení této práce a jejich zdrojové kódy jsou dostupné v přiloženém archivu. Dále se diplomant zabýval možnostmi přidávání nových patentových dat a vysvětlil, proč v úložišti nejsou obsažena patentová data České republiky. Ověřením funkčnosti vytvořeného úložiště komplexními dotazy je zakončena realizační část textu.

Text dokumentu na mě působí, jako by byl dokončen ve spěchu a neprošel jazykovou kontrolou. Obsahuje velké množství překlepů, chybných větných formulací a také gramatické chyby. Velmi nízkou vypovídací hodnotu mají sejmuté obrazovky v kapitole 7, které mají bílý text na černém pozadí a mají ukazovat výsledky full-textových dotazů nad patentovými daty. Poslední výtka se týká zkratek. Vadí mi, že většina zkratek není v textu nikde rozepsána, pouze v seznamu zkratek na konci dokumentu.

Pro realizaci úložiště patentových dat diplomant využil celou řadu programových produktů. Očištěná patentová data z 10 národních patentových institucí ukládá do dokumentové databáze *MongoDB*, kde pro rychlé full-textové vyhledávání data indexoval vyhledávač *Elasticsearch*. O jejich propojení se postaral nástroj *Apache Kafka*. Metadata o patentech jsou uložena do relační databáze *MySQL* zahrnující 5 propojených relačních tabulek. Bezproblémové nasazení výše uvedených nástrojů diplomant vyřešil použitím kontejnerové platformy *Docker*.

Odevzdaný archiv kromě exportovaných patentových dat obsahuje příslušné konfigurace za účelem vytvoření a nasazení jednotlivých softwarových kontejnerů a jednorázové naplnění *MongoDB* a *MySQL* databází patentovými daty. Uživatel následně může v internetovém prohlížeči nástroji *Mongo-express*, resp. *phpMyAdmin* formulovat dotazy nad patentovými daty, resp. jejich metadaty.

Diplomant použil pro napsání práce 5 knižních a 37 elektronických zdrojů, v závěru práce zmínil jednu diplomovou práci. Kladně hodnotím použití 4 knižních zdrojů zabývající se databázemi a jeden knižní zdroj, který se zabývá problematikou patentů. Také souhlasím s využitím webových stránek produktů, které jsou v textu představeny nebo jsou použity pro vlastní realizaci. Úplně se mi nelíbí použití 12 různých databázově zaměřených webových stránek mající podobu blogu, které převážně sloužily k sepsání kapitoly o databázích. Nevhodným použitím těchto zdrojů jsou v kapitole o databázích zaneseny chybné informace. Všechny prameny jsou aktuálně dostupné a v dokumentu řádně citované.

Dotazy k práci

1. V textu práce uvádíte, že patentová data všech zemí jsou uložena v jediné kolekci `patents`. Uvažoval jste možnost patentová data rozdělit do více kolekcí, např. pro každou zemi vytvořit vlastní kolekci?
2. Odevzdaný archiv obsahuje exporty patentových dat a metadat o velikostech 787 MB a 547 MB. Jsou v těchto exportech uvedena všechna získaná patentová data a metadata anebo se jedná o reprezentativní vzorek?

Zadání a zásady pro vypracování diplomové práce student splnil bez výhrad.

Vzhledem k výše uvedeným nedostatkům navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 2.6.2022 Ing. Martin Zíma, Ph.D.