Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženyrství

Diplomová práce

Řešení datové kvality v podnikové praxi

Filipp Podriadchikov

Řešení datové kvality v podnikové praxi

Abstrakt

Hlavním cílem této práce je provést zhodnocení a navrhnout opatření pro zlepšení datové kvality v rámci podnikového prostředí, s důrazem na oblast business intelligence (BI). Případová studie je zvláště zaměřena na procesy čištění, sjednocení a deduplikace dat uložených v datovém skladu. Druhotným cílem je provést analýzu fungování BI, datových skladů a dalších komponent BI, s ohledem na řízení kvality dat.

Klíčová slova: čištění dat, data, datová kvalita, business intelligence, unifikace, deduplikace

Cíl práce

Diplomová práce je tematicky zaměřena na problematiku řešení datové kvality v podnikové praxi.

Hlavním cílem práce je návrh řešení, které povede ke zlepšení datové kvality v podnikové praxi.

Za účelem dosažení hlavního cíle jsou stanoveny následující dílčí cíle:

- Prozkoumat a zmapovat oblasti a principy Business Intelligence, dat a datové kvality.
- Provést analýzu a zhodnocení nástrojů pro řízení datové kvality.
- Aplikovat vybrané metodiky/techniky zabývající se unifikací dat v podnikové praxi a demonstrovat je na případové studii.

Metodika

Pro dosažení stanovených cílů se budeme řídit několika důležitými kroky a postupy. Naše práce bude zahrnovat pečlivou analýzu odborné literatury, jak zahraniční, tak české, a také elektronických zdrojů, které se týkají oblasti datové kvality a nástrojů pro její řízení.

Tato teoretická část nám poskytne hlubší porozumění problematice datové kvality a přístupů k jejímu zlepšení.

Náš další významný krok zahrnuje praktickou fázi naší výzkumné práce, která je zaměřena na provedení několika klíčových procesů. Konkrétně se jedná o procesy čištění dat, standardizace a deduplikace. K řešení těchto úkolů jsme se rozhodli využít metodu vytvoření vlastní aplikace s využitím programovacího jazyka Python. Tato aplikace bude schopna provádět důkladnou validaci dat a následně korigovat identifikované chyby. Pro spuštění aplikace využijeme prostředí Jupyter Notebook, a tato aplikace bude konfigurována tak, aby prováděla operace přímo nad MySQL databází umístěnou v rámci podnikové infrastruktury.

Vývoj vlastní aplikace v jazyce Python nám umožňuje přizpůsobit proces zlepšování kvality dat konkrétním potřebám a složitostem datového souboru podniku. Tato aplikace bude poskytovat flexibilní a přizpůsobitelné řešení pro identifikaci a následnou opravu problémů spojených s kvalitou dat, což nakonec přispěje k zvýšení celkových standardů kvality dat v organizaci. Použití Jupyter Notebook nabízí uživatelsky přívětivé a interaktivní prostředí pro provádění a dokumentaci procedur zlepšování kvality dat.

Dále je důležité zdůraznit, že provádění těchto operací přímo v MySQL databázi organizace zajišťuje, že vylepšení dat jsou implementována přímo v jádru datového repozitáře, což přináší prospěch všem následným procesům a analýzám závislým na tomto datovém zdroji.

Teoretická část

V úvodu teoretické části jsou představeny klíčové koncepty a zásady spojené s oblastí business intelligence, datových skladů a samotných dat. Následně jsou definovány relevantní pojmy a charakteristiky týkající se kvality dat, a to včetně úkolů a klíčových aktivit nezbytných pro úspěšné řízení datové kvality. Zvláštní pozornost je věnována faktorům, které přispívají ke vzniku nedostatečné kvality dat, opatřením vedoucím k její prevenci a důsledkům, které špatná kvalita dat s sebou nese.

Poslední část teoretické části se zaměřuje na vysvětlení základních operací prováděných nad daty a jejich vztah k celkové kvalitě dat.

Praktická část

Praktická část studie se věnuje implementaci procesů čištění, sjednocení a deduplikace dat v rámci podnikového datového skladu. V průběhu této fáze jsou navrhována konkrétní opatření a metody, které budou aplikovány na skutečná data s cílem zvýšit jejich kvalitu. Tato část zahrnuje detailní postupy a techniky, které mají potenciál efektivně odstranit nedostatky spojené s datovou kvalitou a přispět k celkovému zlepšení podnikového informačního prostředí.

Závěr

Kvalita dat je klíčovým faktorem v rámci všech organizací, bez ohledu na jejich velikost. Zlepšení kvality dat v malých a středních podnicích přináší mnoho významných a dosažitelných výhod. Vysoká kvalita dat umožňuje podnikům lépe informovaná a kvalifikovaná rozhodnutí, identifikaci trendů, příležitostí a výzev, a vytvoření strategií a plánů, které jsou založeny na konkrétních datech spíše než na domněnkách a odhadech. Kvalitní data také přispívají ke zvýšení efektivity a produktivity, což vede ke snížení časových ztrát, omezení chyb a minimalizaci ztrát. Díky kvalitním datům je snazší jejich zpracování, správa a analýza, což umožňuje automatizaci a zefektivnění klíčových obchodních procesů.

V konečném důsledku poskytují kvalitní data konkurenční výhodu na trhu. Pro zlepšení kvality dat v malých a středních firmách je klíčové implementovat standardní procesy a systémy pro sběr, ukládání a analýzu dat, a to i v menším měřítku. Důležitou součástí je také implementace strategie správy dat, která zahrnuje definování rolí a odpovědností za řízení dat, stanovení procesů pro ukládání dat a zavedení kontrolních opatření ke zvýšení kvality dat.

Myslím si, že tato práce bude užitečná všem, kteří mají zájem o oblast datové kvality, zejména těm, kteří se zabývají jejím řízením. Tato práce poskytuje návod, jak postupovat při řízení kvality dat, a přináší metodiku, kterou lze prakticky využít při řešení otázek týkajících se datové kvality v podnicích. Doufám, že se mi podařilo dosáhnout všech cílů, které jsem si pro tuto práci stanovil. Přesto věřím, že existuje široký prostor pro další rozvoj a prohlubování této problematiky.