Datový sklad na technologiích IBM a jeho možnosti

Autor: Jiří Snítil
Vedoucí: doc. Ing. Jan Pour, CSc.
Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky

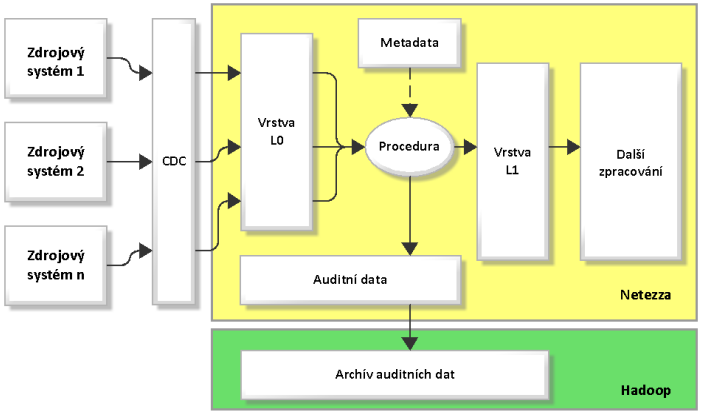
Úvod

Diplomová práce se zabývá analýzou rozšiřujících konceptů použitelných v datových skladech. V práci jsou k analýze vybrány tři rozšiřující koncepty a je zdůvodněn jejich výběr.

1) Zachycení změn ve zdrojových systémech (CDC)
2) Historizace dat po CDC
3) Analytické funkce v datovém skladu

Použité metody

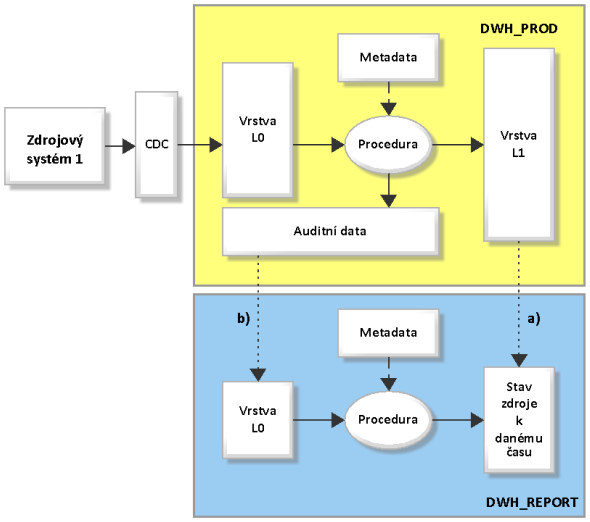
Bylo vytvořeno komplexní testovací prostředí, které se skládalo z několika propojených systémů a technologií. V tomto prostředí byly rozšiřující koncepty vyzkoušeny a analyzovány. Na základě výsledků a zjištění z testovacího prostředí byly jednotlivé koncepty zhodnoceny, a to zejména vzhledem k přínosům, nevýhodám, rizikům a možnostem uplatnění těchto konceptů při jejich použití v datovém skladu.



Obr. 1 – Navržené generické řešení zpracování dat ze zdrojových systémů

Přínosy práce a zjištění na základě testů

1) Při vhodném použití konceptu CDC lze jednoznačně určit stav dat zdrojového systému v libovolném minulém časovém bodě.
2) Data ze zdrojových systémů lze zpracovávat do historické kolekce dat i v případě, kdy jsou ze zdrojového systému získávány pouze změny, které v datech nastaly.
3) Bylo navrženo generické řešení zpracování dat ze zdrojových systémů (obrázek č. 1).
4) Bylo navrženo a vyzkoušeno zpracování dat k určitému času, včetně tzv. zpětného zpracování dat k určitému minulému času s využitím auditních dat (obrázek č. 2).
5) Byl demonstrován přenos analytického výpočtu k datům uloženým v datovém skladu, bez nutnosti migrace těchto dat do analytického prostředí (pomocí jazyka R a C++).
6) Byl demonstrován příklad použití nových analytických funkcí vyvinutých v jazyce C++ pro specifické analytické účely v rámci jazyka SQL (ukázka na obrázku č. 3).



Obr. 2 – Zpracování dat k určitému času

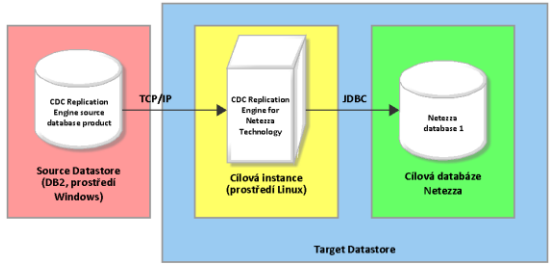
```
select vysledek.*
from data_fce2, table(vytvor_tab(id, atr1,atr2,atr3, atr4)) vysledek
order by id, typ;
```

Output | Result 1

Drag a column header here to group by that column.

	REF	I	J	TYP
1	1	4	8	odecteno
2	1	8	12	secteno
3	1	3	5	vydeleno
4	1	12	20	vynasobeno
5	2	0	0	odecteno
6	2	20	40	secteno
7	2	1	1	vydeleno
8	2	100	400	vynasobeno
9	3	5	98	odecteno
10	3	15	102	secteno
11	3	2	50	vydeleno
12	3	50	200	vynasobeno

Obr. 3 – Ukázka nové analytické funkce volané z SQL



Obr. 4 – Znázornění části testovacího prostředí určeného pro koncept CDC

Přínosy práce pro praxi

Výstupy diplomové práce mohou sloužit jako jeden z podkladů k rozhodnutí, zda jednotlivé rozšiřující koncepty použít v konkrétních podmínkách datového skladu, protože u všech tří rozšiřujících konceptů jsou analyzovány možné přínosy, nevýhody a rizika, které s těmito koncepty souvisí. Například první a třetí koncept značně omezuje potřebu migrace dat mezi systémy, což vzhledem k obecnému růstu objemů dat může přinést finanční úspory a značné zvýšení flexibility celého řešení. Druhý koncept přináší například možnosti zpětného zpracování dat k různým časovým bodům a s jeho pomocí lze tak lépe poskytovat data pro další vrstvy datového skladu, nebo pro uživatele, kteří s těmito daty pracují.