Životný cyklus dátových skladov*

Tomáš Tisovský
Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
xtisovskyt@stuba.sk

4. november 2021

Abstrakt

Témou tohto článku je proces tvorby dátových skladov vrámci oblasti Business Intelligence. Vzhľadom na pomerne vysokú komplexnosť DW/BI systémov je využitie agilných praktík v tejto oblasti nutnosťou pre efektívnu tvorbu daných projektov. Pri znalosti metód a postupov pri vývoji dátových skladov je práca na projekte organizovana a môže dôjsť k výraznému ušetreniu času, či finančných prostriedkov. Článok má za cieľ predstaviť najlepšie techniky, praktiky a prístupy pre tento obor. Článok sa venuje rôznym etapám životného cyklu dátoveho skladu ako plánovanie a riadenie projektu, definícia a zber požiadaviek, technologická fáza, dátová fáza, aplikačná fáza a údržba. Tieto ciele článok spĺňa opísaním kompletného životného cyklu dátových skladov.

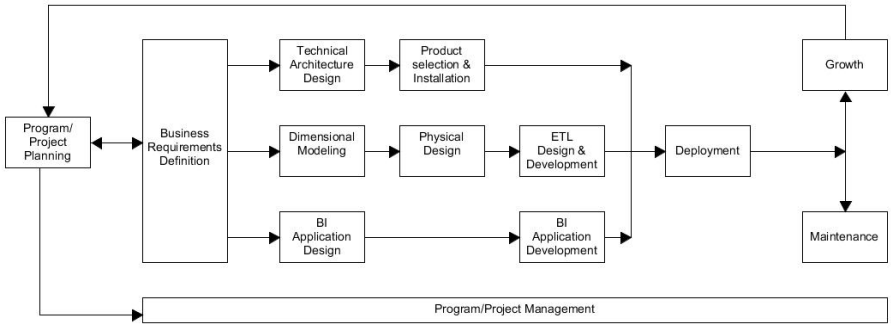
1 Úvod

Tento článok sa venuje životnému cyklu dátových skladov. Technológia dátových skladov predstavuje v súčasnosti jeden z najvýznamnejších trendov v rozvoji podnikových informačných systémov. Dátový sklad (Data Warehouse) možno definovať mnohými spôsobmi. Za základ však budeme považovať definície jedného zo zakladateľov DWH, Williama Inmon.„Dátový sklad je integrovaný, subjektovo orientovaný, stály a časovo rozlíšený súhrn dát, usporiadaný pre podporu potrieb manažmentu“ [5]

Životný cyklus prebieha v niekolkých etapách. Poznanie etáp projektu je dôležité pre všetkých účastníkov projektu, teda manažérov, analytikov, návrhárov, či vývojárov na vykonanie správnych úloh v správny čas. Pri vytvorení takéhoto softvéru sa kladie hlavný dôraz na požiadavky užívateľov, iteratívnosť a dimenzionalitu v poňatí štruktúrovaných dát. Za štandard v tejto oblasti sa považuje životný cyklus od Ralpha Kimballa. Tento model využíva agilný prístup pre jeho vyššiu efektivitu a úspešnosť. Tento model je zobrazený v Kimball Lifecycle diagram (obr. nižšie).

Tento diagram poskytuje celkový plán znázornujúci postupnosť úloh na vysokej úrovni potrebnných pre úspešné DW/BI projekty.

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2021/22, vedenie: Ing. Vladimír Mlynarovič, PhD.



Obr. 1: Diagram životného cyklu podľa Kimballa

Model podľa Kimballa začína plánovaním. V tejto fáze je potrebné určiť rozsah projektu a potrebné zdroje. V tejto fáze sa začínajú aj riadiace povinnosti, ktoré pretrvávajú počas celého zvyšku projektu. Na fázu plánovania nadväzuje fáza definovania uživateľských požiadaviek. Táto fáza je pre projekt klúčová pretože ovplyvnuje všetky nasledovné fázy projektu. Medzi fázou plánovania a fázou zberu požiadaviek je potreba úzkej vzájomnej spolupráce. Horná časť diagramu sa venuje technologickej stránke projektu kde prebieha technologická architektúra. Po dokončení architektúri sa vyberú vhodné nástroje na tvorbu sofvéru. Stredná časť diagramu sa zaoberá dátovej stránke projektu. V tejto fáze sa vývojári venujú dátam, operáciam s nimi. Vzniká tu multidimenzionálny model, ktorý je podstatou DW/BI projekov. Taktiež sa tu tvorí fyzický model a prebehne ETL proces(extract, transform, load). Spodná časť diagramu sa sustredí na výstupy pre uživateľov vo forme multidimenzionalnej aplikácie. V sekcii s názvom rast sú vyjadrené praktiky inkrementálnosti. Táto sekcia hovorí o tom , že pri každom prírastku dát by sa vývojári mali vracať k plánu a držať sa požiadaviek uživateľov. Nasledujúca časť článku je venovaná jednotlivím etapám životného cyklu. [3]

2 Plánovanie a riadenie projektu

Dlhodobý cieľ projektu dátového skladu počíta nielen s jeho vybudovaním, ale definuje aj stratégiu správy dátového skladu, pričom počíta s dokumentáciou dátového skladu a školením jeho užívateľov. V tejto fáze sú definované aj základy architektúry podnikového dátového skladu. Behom fázy definície sa definuje rozsah a cieľ prírastkového vývoja. Vytvorí sa počiatočný prírastok, konceptuálny model, zdokumentujú sa zdroje dát a presne sa vymedzí rozsah kvality týchto dát. Je navrhnutá ako aj architektúra dátového skladu, tak aj architektúra technických prostriedkov. V tejto fáze máme najlepšiu príležitosť zamerať sa na pochopenie štruktúry operačných a externých zdrojov dát. Stanovia sa krátkodobé a dlhodobé obchodné ciele, pre podporu ktorých je dátový sklad budovaný. [1]

Proces riadenia projektov sa týka koordinácie ľudských, finančných a materiálnych zdrojov, je zameraný na dosiahnutie dopredu stanovených cieľov v danom rozsahu, čase, nákladoch, kvalite a spokojnosti účastníkov projektu. [6]

Tak ako pri každom projekte aj pri DW/BI projekte je kľúčový vedúci projektu. Pre projekt je tiež dôležitá úloha biznis analytika, ktorý by mal mať dobré povedomie o spolupráci IT a biznisu. Úlohou dátového analytika je analyzovať kvalitu dát, ich kvalitu a granularitu. Úlohou externého konzultanta je obyčajne transfer poznatkov podniku, vyškolenie a riadenie ľudí, vedenie projektov. Konzultačné služby v oblasti dátových skladov sa spravidla týkajú výberu hardvéru a softvéru, návrhu architektúry a optimálnej zostavy softvérových technológií a zaistenie využitia najnovších informačných technológií v rôznych oblastiach podnikania. Konzultanti poskytujú služby tiež v oblasti systémovej integrácie, konvergencie služieb a technológií, služby v oblasti on-line bezpečnosti, vytváranie webových stránok, podnikové informačné systémy. IT konzultant všeobecne pomáha firmám pochopiť, akým spôsobom môžu využiť technológie pre svoj prospech. [1]

Pre projekt je kľúčové vymedzenie nákladov:

- Hardvér - náklady na hardvér nie sú zanedbateľné, ale z hľadiska filozofie dátového skladu sa jedná o technické prostriedky, ktoré sú nahraditeľné. Jedná sa o technické prostriedky, nie o dáta ktoré sú v nich uložené. Pre rýchly prístup k obrovskému množstvu dát je potrebné mať výkonné servery, alebo dátové centrá. [5]

- Softvér - Nástroje pre vytváranie dátových skladov a analýzu dát sú veľmi drahou záležitosťou. Čím ďalej viac sa presadzuje trend integrácie analytických služieb priamo do inštalácií databázových serverov. Poplatok za analytické služby je buď zahrnutý priamo v cene databázového serveru, alebo sa licenčné poplatky platia zvlášť. [4]

Taktiež je dôležité aj určenie benefitov, ktoré daný systém prinesie, teda určiť ktoré konkrétne rozhodovacie procesy v spoločnosti budú podporené.

3 Definícia a zber požiadaviek

V tejto etape životného cyklu je dôležitá komunikácia so zákazníkmi, a zisťovanie ich požiadaviek pre daný softvér. Technické otázky ako architektúra, granularita a multidimenzionalita na tejto úrovni nie sú komunikované. Hlavným nástrojom tímu vývojárov na získanie požiadaviek je interview, kde je možné získať užitočné a detailné informacie. Pre vytvorenie úspešného DW/BI systému je potrebné komunikovať so zástupcami naprieč celou organizáciou a osvojiť si ich používané systémy a dáta. Zároveň je výhodné ak na sa stretnutiach so zákaznkmi zúčastnujú aj dátový analyticy, ktorý do hĺbky rozumrjú uživateľským dátam a môžu prispieť k správnemu smerovaniu zberu požiadaviek. Pri zbere požiadaviek je dôležitá dokumentácia, zhrnutie a zaznamenanie kľúčových informácii a analýza požiadaviek. Dôležitou súčasťou je tiež zistenie kritérií úspechu od užívateľov, ktoré očakávajú od DW/BI systému. Pri veľkom množstve požiadaviek je dôležitá prioritizácia, ktorá by mala brať do úvahy potenciálnu hodnotu požiadavky pre organizáciu. [3]

4 Technologická fáza

Táto kapitola sa venuje technologickej architektúre a možnostiach implementácie, ktoré sú k dispozícii pre dátové sklady. Cieľom je transformovať požiadavky zákazníkov do detailných podmienok návrhu dátového skladu. Zvolená architektúra zavisí od rozhodnutia manažmentu, ktoré bude založené na faktoroch, ako je súčasná infraštruktúra, podnikateľské prostredie, požadovaná štruktúra riadenia a kontroly, angažovanosť a rozsah implementačného úsilia, technické prostredie organizácie a dostupné zdroje. Technologická architektúra sa delí na tri hlavné zložky: architektúra dát, architektúra aplikácie a infraštruktúra. Návrh architektúry by sa mal vykonať pred začatím implementácie. Architektúru je možné meniť alebo upraviť aj po začatí implementácie, avšak takýto postup často vyžaduje prepracovanie projektu a tým aj oneskorenie projektu. Zvolený implementačný prístup je tiež rozhodnutím manažmentu ktoré môže mať vplyv na úspech projektu dátového skladu. Premenné ovplyvnené touto voľbou sú čas na dokončenie, návratnosť investície, rýchlosť realizácie benefitov a spokojnosť užívateľov. Výber architektúry určí, alebo bude určený tým, kde budú dátové sklady alebo dátové trhoviská sídliť a kde bude sídliť kontrola. Údaje sa môžu napríklad nachádzať v centrálnom umiestnení, ktoré je spravované centrálne, alebo sa údaje môžu nachádzať v distribuovaných miestnych vovzdialených lokalitách, ktoré sú riadené centrálne. Pri výbere implementácie prichádzajú do úvahy možnosti ako implementácia zhora nadol, zdola nahor alebo kombinácia oboch. Technologická architektúra je tiež dôležitá pre dokumentáciu projektu. Technologická architektúra sa tiež využíva pri komunikácií medzi organizáciou, implementátorom a inými dodávateľmi. V technologickej architektúre sa tiež špecifikujú vhodné nástroje, techniky, služby a platformy potrebné na realizáciu projektu, podľa kritérií zvolených počas analýzy požiadaviek. [2] [4]

5 Dátová fáza

Hlavnou zložkou tejto časti projektu je dimenzionálne modelovanie. Dimenzionálne modelovanie je proces tvorby dátového modelu. Vo všeobecnosti je model abstrakcia a reflexia reálneho sveta. Modelovanie nám dáva možnosť vizualizovať to, čo si ešte nedokážeme uvedomiť a predstaviť. Rovnako je to s dátovým modelovaním. Dimenzionálne modelovanie používa tri základné pojmy: miery(premenné), fakty a dimenzie. Dimenzionálne modelovanie je účinné pri reprezentácií požiadaviek obchodného používateľa v kontexte databázových tabuliek. Najpopulárnejším spôsobom vizualizácie dimenzionálneho modelu je nakreslenie kocky. Pomocou kocky môžme reprezentovať trojrozmerný model. Zvyčajne model pozostáva z viac ako troch rozmerov a označuje sa ako hyperkocka. Hyperkocku je však ťažké si predstaviť, preto je kocka bežnejší výraz. Ďalsia etapa je proces tvorby fyzického modelu. Účelom tejto etapy je navrhúť model pre skutočnú fyzickú implementáciu. Návrh fyzického modelu musí brať do úvahy fyzické obmedzenia ako je priestor, výkonnosť a fyzickú distribúciu dát. Jeden z najdôležitejších aspektov návrhu fyzického modelu súvisí s granularitou dát. Granularita dát v dátovom sklade sa týka úrovne sumarizácie dátových prvkov, teda úrovne detailov dostupných v dátových prvkoch. Čím podrobnejšie údaje sú k dispozícii, tým nižšia úroveň granularity. Naopak, čím nižšia úroveň detailov, tým vyššia úroveň granularity dát. Ďalej nasleduje porces ETL

(extract, transform, load). Nástroje a postupy procesu ETL sú velmi dôležitou súčastou každého projektu dátového skladu. Celý proces ETL je veľmi komplexný a vo väčšine prípadov časovo náročný. V niektorých implementáciach môže zabrať viac ako polovicu celkového času a velkú časť nákladov potrebných na vytvorenie dátového skladu. Hlavným cielom etapy ETL je centralizácia údajov, teda ich zhromaždenie z viacerých rôznorodých zdrojov a naplnenie dátového skladu určenými údajmi v požadovanom čase. Proces ETL sa skladá z troch etáp:

- Extrakcia (extract) – výber dát prostredníctvom rôznych metód
- Transformácia (transform) – overenie, čistenie, integrovanie a časove označenie dát
- Načítanie/naplnenie (Load) – premiestnenie dát do dátového skladu

Proces ETL neprebehne vždy úspešne a preto je dôležité etapu ETL dôkladne otestovať. [2] [4]

6 Aplikačná fáza

Táto fáza je zameraná na tvorbu reportov a aplikácií. Aplikácie a reporty kladú dotazy do databáz dátových skladov, testujú a overujú hypotézy s informáciami, ktoré vyťahujú z dátového skladu, rekonštruujú reťazce udalostí, ktoré potom analyzujú s cieľom odhaliť vzory alebo trendy a tvoria prognózy do budúcnosti a tak podporujú rozhodovacie procesy v organizácií. Klientské aplikácie môžeme v BI rozlišovať do dvoch skupín a to sú reporting a analytické aplikácie. Reporting pozostáva z analytických tabuliek a prehľadov, realizovaných na základe dotazov do databáz dátových skladov. Analytické aplikácie sú navrhované špeciálne pre poskytovanie manažérskych informácií umožňujúce sledovať firemné procesy, plnenie cieľov organizácie atď. Analytické aplikácie poskytujú vyššiu flexibilitu vzhľadom k momentálnym požiadavkám užívateľa. Aplikácie dátových skladov skôr zamerané na údaje ako na procesy. Koncoví používatelia narábajú s údajmi takmer priamo a nexistujú žiadne pevné pracovné postupy. Koncoví používatelia aplikácií iba používajú údaje z dátového skladu, nezaznamenávajú žiadne údaje do skladu. Ďalšiou zložkou tejto fázy je analýza dát. Existuje niekoľko techník analýzy dát, ktoré sa dnes bežne používajú ako napríklad multidimenzionálna analýza a dolovanie dát. Používajú sa na formulovanie a zobrazenie výsledkov dotazov a na analýzu údajov. Táto fáza by mala začať ešte pred dokončením ETL procesu, pretože aplikační vývojári často odhalujú chyby v dátovej kvalite. [2] [4]

7 Nasadenie, údržba a rast

8 Zhrnutie

Literatúra

[1] D. ARNOŠT. *Business Intelligence příručka manažera.* Praha : TATE International, s.r.o., 2007.

[2] Chuck Ballard, Dirk Herreman, Don Schau, Rhonda Bell, Eunsaeng Kim, and Ann Valencic. Data modeling techniques for data warehousing. Technical Report SG24-2238-00, International Technical Support Organization, San Jose, California, USA, February 1998.

[3] R. KIMBALL and M. ROSS. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Third Edition.* Indianopolis: John Wiley, 2013.

[4] Ľ. LACKO. *Datové sklady analýza OLAP a dolování dat s příklady v SQL Serveru a Oracle.* Brno : Computer Press, 2003.

[5] O. NOVOTNÝ, J. POUR, and D. SLÁNSKY. *Business Intelligence Jak využít bohatství ve vašich datech.* Praha : Grada Publishing, a.s., 2005.

[6] M. TVRDÍKOVÁ. *Aplikace moderních informačních technologií v řízení firmy. Nástroje ke zvyšování kvality informačních systémů.* Praha : Grada Publishing, a.s., 2008.