ZÁKLADNÉ ZÁSADY TESTOVANIA DÁTOVÝCH SKLADOV

THE FUNDAMENTAL PRINCIPLES OF DATA WAREHOUSE TESTING

Pavol TANUŠKA, Andrej TRNKA

Autori: ***Doc. Ing. Pavol Tanuška, PhD.*** [1]***, Ing. Andrej Trnka*** [2]
Pracovisko: ***Ústav aplikovanej informatiky, automatizácie a matematiky, Katedra informatiky, Materiálovotechnologická fakulta STU,***
[2]***Fakulta prírodných vied, Univerzita sv. Cyrila a Metoda,***
Adresa: [1]***Paulínska 16, 917 24 Trnava***
[2]***Nám. J. Herdu 1, Trnava***
Email: ***pavol.tanuska@stuba.sk, andrej.trnka@ucm.sk***

Abstract

Článok popisuje základné zásady, ktoré sa odporúča dodržovať počas testovania aplikácií dátových skladov. Proces testovania prechádza viacerými fázami a dodržiavanie popisovaných odporúčaní môže zabezpečiť lepšiu produkciu alebo môže zabrániť určitým chybám, ktorých odstránenie je veľmi nákladné.

This contribution deals with utilization of fundamental principles that are recommended to use in the phases of Data Warehouse applications testing. The testing process goes through a various phases. The adherence of testing recommendations can prevent existence of mistakes and errors that are very expensive to remove it.

Key words

dátový sklad, testovanie, proces ETL, V- procesný model

data warehouse, testing, ETL (Extraction –Transformation - Load), V- process model

Úvod

Dátový sklad je kolekcia zjednotených, predmetovo orientovaných databáz, ktoré sú navrhnuté za účelom poskytovania informácií požadovaných pre rozhodovanie.

Dátový sklad obsahuje dáta získané z viacerých prevádzkových systémov a môže byť naplnený tiež externými dátami [1].

Každý prevádzkový systém nemusí zaznamenávať tie isté typy operácií. V reáli môže nastať situácia, že každý prevádzkový systém bol vytvorený na zákazku alebo nebol implementovaný integrovaný systém. To znamená, že tieto prevádzkové systémy nie sú

zjednotené. Dátový sklad zjednocuje dáta z rôznorodých prevádzkových systémov a poskytuje integrovaný pohľad.

Dátový sklad prináša tradičné informačné pohľady na subjekty. Jadrom každého dátového skladu je rozsiahla databáza, ktorá obsahuje integrované dáta získané z interných a externých zdrojov dát. Pojem interné dáta zahŕňa všetky dáta, ktoré sú získané z prevádzkových systémov. Externé dáta sú dáta poskytnuté tretími stranami.

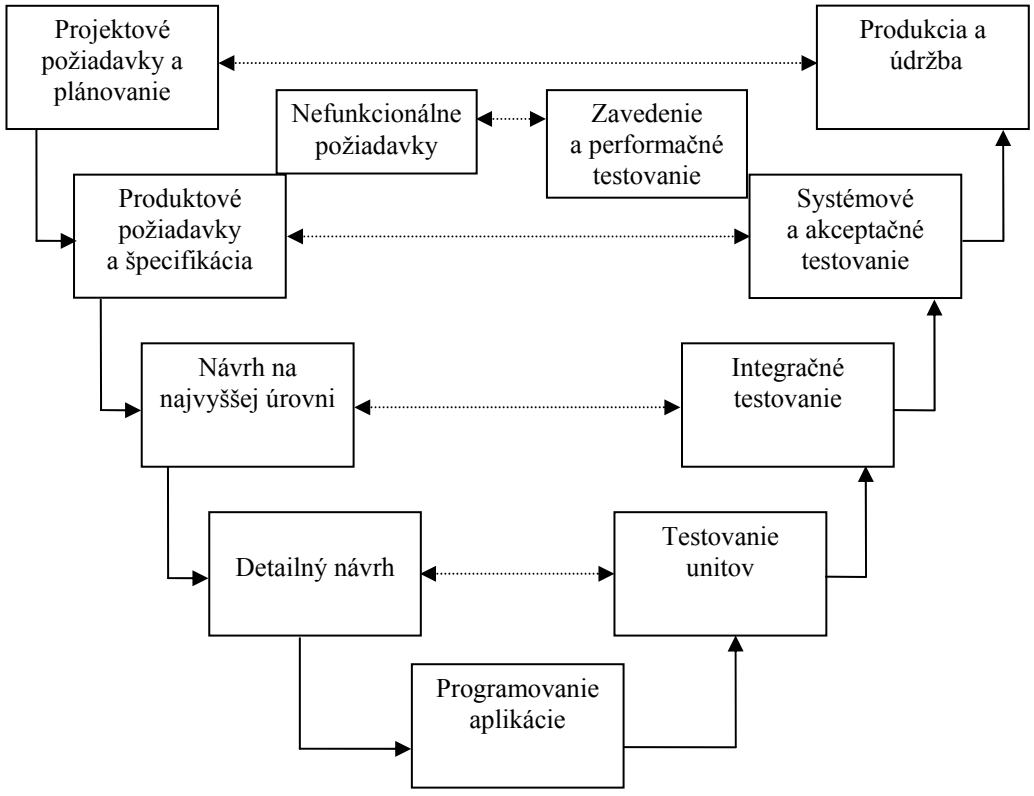
Riadenie sa stále viac a viac sústreďuje na zber a organizáciu dát pre strategické rozhodovanie. Schopnosť posúdiť historické trendy a kontrolovať dáta spracovávané v reálnom čase sa stala kľúčom konkurenčnej výhody.

Článok poskytuje praktické odporúčanie pre testovanie, transformovanie a zavedenie ETL aplikácií.

ETL (Extract-Transform-Load) je súbor procesov, pomocou ktorých sa zdrojové dáta pripravujú pre dátový sklad. Pozostáva z extrahovania dát zo zdrojovej aplikácie, ich transformácie, zavedenia, indexovania, zaistenia kvality a ich publikovania. [1, 2, 4].

Proces testovania

Proces testovania sa dá zobraziť pomocou tzv. V- procesného modelu (obr. 1). Tento model zobrazuje jednotlivé fázy vývoja a testovania aplikácie [6].



Obr. 1. *V- procesný model*

Ciele testovania

Vývojárske tímy vynakladajú veľa úsilia a nákladov (finančných a materiálnych), aby sa eliminovali a vyhľadali pokiaľ možno všetky chyby v softvérových systémoch. V dátových skladoch je to obdobné, pretože použitie nesprávnych dát môže vyvolať kritické rozhodnutia s katastrofálnymi následkami.

Medzi hlavné atribúty testovania ETL aplikácií (ako primárna súčasť DW - Datawarehouse, niekedy tiež označovaná ako ETT alebo dátová pumpa) môžeme zaradiť nasledovné:

- Kompletnosť dát. Zaisťuje sa, či všetky očakávané dáta sú zavedené.
- Transformácia dát. Zaisťuje sa, či všetky dáta sú korektné podľa pravidiel a špecifikácií návrhu.
- Kvalita dát. Zaisťuje sa, či ETL aplikácia korektne odmietne, nahradí implicitné hodnoty, opraví alebo ignoruje chybné dáta a poskytne správu.
- Výkonnosť. Zaisťuje sa, či sa správne zavedú dáta a vykonajú jednotlivé dotazy v očakávaných časových intervaloch.
- Integračné testovanie. Zaisťuje sa, či ETL proces funguje dobre s ostatnými procesmi.
- Testovanie akceptácie užívateľom. Zaisťuje sa riešenie, ktoré vyhovuje požiadavkám užívateľov a predpovedá ich budúce požiadavky.
- Regresné testovanie. Zaisťuje sa funkčnosť vždy, keď je dokončená nová verzia aplikácie [3].

Kompletnosť dát

Jedným zo základných testov kompletnosti dát je verifikovať, či všetky očakávané dáta sú korektne uložené (zavedené) do dátového skladu. Tento test obsahuje potvrdenie, že záznamy, všetky polia a celý obsah každého poľa je uložený.

Stratégie, ktoré treba vziať do úvahy zahŕňajú:

- Porovnávanie počtu záznamov medzi zdrojovými dátami, dátami zavedených do dátového skladu a odmietnutými záznamami.
- Porovnávanie jedinečných hodnôt kľúčových polí medzi zdrojovými dátami a dátami zavedenými do dátového skladu. Toto je technika, ktorá upozorňuje na rôznorodosť možných chýb v dátach bez nutnosti robenia plnej validácie všetkých dátových polí.
- Používanie nástrojov profilujúcich dáta, ktoré zobrazujú rozsah a hodnoty polí v skupine dát. Toto môže byť použité počas testovania a na porovnanie zdrojových a cieľových dátových setov a k upozorneniu na dátové anomálie zo zdrojového systému, ktoré môžu chýbať aj vtedy, keď presun dát bol korektný.
- Overenie plného obsahu každého poľa, t.j. aby sa validovalo, že nenastane žiadne „osekanie“ v nejakom kroku transformácie. Napr. ak je pole zdrojových dát typu string [3], je dôležité, aby sa testovalo s 30 znakmi.
- Testovanie okrajových podmienok každého poľa, aby sa našli databázové obmedzenia. Napr. pre pole decimal [3], ktoré obsahuje hodnoty -99 a 999 a pole s dátami, ktoré obsahuje celý rozsah očakávaných dát. V závislosti na type databázy a jej indexovania nie je vylúčené, že rozsah hodnôt, ktoré databáza akceptuje, je príliš malý [3].

Transformácia dát

Validovanie (dáta sú transformované správne), ktoré je založené na pravidlách, môže byť najkomplexnejšou časťou testovania ETL aplikácií. Jednou typickou metódou je vybrať nejakú vzorku záznamu a validovať transformáciu dát ručne. Táto metóda môže byť užitočná, ale vyžaduje ručné testovacie kroky a testovacieho pracovníka, ktorý rozumie ETL logike. Kombinácia automatizovaného profilovania dát a automatizovanej validácie pohybu dát je z dlhodobej stratégie lepšia.

Niektoré techniky jednoduchej automatickej transformácie dát:

- Vytvoriť tabuľkový kalkulátor scenárov vstupných dát a očakávaných výsledkov a validovať ich so zákazníkom. Je to dobrá požiadavka, ktorá môže byť použitá v čase návrhu, ale tiež behom testovania.
- Vytvoriť testovacie dáta, ktoré obsahujú všetky scenáre. Vyžiadať si pomoc vývojára ETL k tomu, aby automatizoval proces tvorby dátových setov s tabuľkovým kalkulátorom scenárov.
- Využiť výsledky profilovania dát k porovnaniu rozsahu hodnôt medzi zdrojovými a cieľovými dátami.
- Validovať správne spracovanie polí generovaných ETL, ako sú napr. náhradné kľúče.
- Validovať, či sú dátové typy v dátovom sklade špecifikované v návrhu a v dátovom modeli.
- Vytvoriť dátové scenáre, ktoré testujú referenčnú integritu medzi tabuľkami. Napr., čo sa stane, keď dáta obsahujú hodnotu cudzieho kľúča, ktorá nie je v rodičovskej tabuľke.
- Validovať vzťahy rodič-dieťa v dátach. Vytvoriť scenáre, ktoré testujú ako sú ovládané záznamy, ktoré nemajú rodiča [3].

Kvalita dát

Kvalita dát je v tomto zmysle definovaná ako spôsob, akým ETL systém obsluhuje zamietnutie dát, substitúciu, korekciu a notifikáciu bez modifikácie dát. Pre zaistenie úspechu testovania kvality dát, je potrebné zahrnúť čo najväčší počet dát.

Základné pravidlá, ktoré majú vplyv na kvalitu dát a sú definované počas návrhu môžu vyzerať nasledovne:

- Odmietnuť záznam, ak niektoré celočíselné pole (decimal) obsahuje nečíselné znaky.
- Nahradiť hodnotou NULL, ak niektoré celočíselné pole (decimal) obsahuje nečíselné znaky.
- Validovať a opraviť stav poľa, pokiaľ je nutné, aby bol založený napr. na PSČ.
- Porovnať hodnoty produktových kódov vo vyhľadávacej tabuľke a ak sa nenájde žiadna zhoda, treba to ohlásiť užívateľovi.
- Odstrániť problém viacerých možností zápisu toho istého údaju (obr. 2).

Nové Mesto nad Váhom	Nové Mesto n/Váhom
N. Mesto nad Váhom	Nové Mesto n/V
Nové M. nad Váhom	N. Mesto n. Váhom
N. M. nad Váhom	N. Mesto n/Váhom
Nové Mesto n. Váhom	Nové Mesto
Nové Mesto n. V.	.....

Obr. 2. *Príklad problému viacerých možností zápisu toho istého údaju*

V závislosti na kvalite dát testovanej aplikácie, testovacie scenáre môžu obsahovať kľúčové hodnoty NULL, duplicitné záznamy v zdrojových dátach a nekorektné dátové typy v poliach (napr. textové znaky v číselnom poli). Je potrebné posúdenie detailných testovacích scenárov s užívateľmi a technickými návrhármi k zaisteniu ich vzájomnej dohody. Pravidlá kvality aplikované na dáta budú obvykle pre užívateľa neviditeľné, keď bude aplikácia vo vývoji, t.j. užívatelia uvidia to, čo je zavedené až v databáze. Z tohto dôvodu je dôležité zaistiť informácie o tom, čo sa urobilo s nekorektnými dátami. Tieto tzv. reporty kvality dát prezentujú hodnotné dáta, ktoré niekedy odhaľujú systematické problémy so zdrojovými dátami [3].

Výkonnosť a rozšíriteľnosť

S rastúcim objemom dát v dátovom sklade, sa môžu zvyšovať časy potrebné na jeho naplnenie a tým aj výkon jednotlivých dopytov sa môže znížiť. Tento stav môže byť eliminovaný dobrým návrhom ETL. Cieľom výkonnostného (performance testing) testovania je upozorniť na potencionálne slabé stránky návrhu ETL ako napr. viacnásobné čítanie súboru alebo vytváranie nadbytočných spolu komunikujúcich súborov.

Nasledovné stratégie pomáhajú objaviť problémy s výkonom [3]:

- Naplnenie databázy s predpokladaným objemom údajov k zaisteniu, že tento objem dát môže byť zavedený ETL procesom.
- Porovnanie týchto ETL zavádzacích časov s menším množstvom dát k tomu, aby sa dali predvídať problémy rozšíriteľnosti. Porovnanie ETL času spracovania od komponentu ku komponentu k tomu, aby sa upozornilo na slabé miesta.
- Monitorovať odmietnuté procesy a zvažovať, s akým veľkým objemom odmietnutých dát sa bude pracovať.
- Vykonávať jednoduché a viacnásobné spojené dopyty k tomu, aby sa validoval výkon.

Integračné testovanie

Systémové testovanie obsahuje len testovanie ETL aplikácie. Koncovými bodmi systémového testovania sú vstupy a výstupy ETL kódu, ktorý je testovaný. Integračné testovanie alebo aj test úplnosti ukazuje, ako aplikácia zapadá do celkového toku údajov vo všetkých aplikáciách. Keď sa vytvára scenár integračného testovania, je potrebné zvážiť, ako môže byť celkový proces narušený a zamerať sa na rozhrania medzi aplikáciami. Tak isto

treba zvážiť, ako by malo byť v každom kroku ovládané zlyhanie jednotlivých procesov a ak to je nevyhnutné, ako by mali byť obnovené alebo vymazané dáta.

Integračné testovanie sa zameriava na testovanie rozhraní, cez ktoré prechádzajú dáta medzi systémami. Integračné testovanie sa robí potom, čo bola aplikácia úspešne otestovaná testom unitov (jednotiek) a systémovým testovaním.

Pri integračnom testovaní je dôležité vykonať testy, ktoré skúšajú hranice rozhrania – maximálne a minimálne hodnoty, rovnako ako aj hodnoty, ktoré systém prijíma k tomu, aby vykonal špecifické výpočty.

Integračné testovanie vo vývoji dátových skladov zahŕňa celý cyklus testovania unitov a súbežne zahŕňa beh viacnásobných aplikácií alebo dávok.

Očakávané a aktuálne výsledky pre integračné testovanie by mali byť dokumentované.

Väčšina problémov nájdená počas integračného testovania je buď relácia k dátam alebo výsledky z nesprávneho návrhu inej aplikácie. Preto je dôležité, aby sa integračné testovanie uskutočňovalo s dátami podobnými reálnym dátam. Najvhodnejšie by bolo použitie reálnych dát, ale v závislosti na obsahu dát sa môže jednať o súkromné alebo bezpečnostné dáta. Tiež netreba zabúdať na význam dobrej komunikácie medzi testovacím a vývojovým tímom všetkých zahrnutých systémov. Na preklenutie tejto komunikačnej medzery je dobré zhromaždiť členov tímu zo všetkých úrovní za účelom formulácie testového scenára a diskusie o tom, čo by mohlo byť v reáli nesprávne. Procesy by sa mali riadiť štýlom „end to end“, v tom istom poradí a s tými istými závislosťami ako v reáli. Integračné testovanie by malo byť kombinované úsilie a nie len zodpovednosť testovacieho tímu ETL aplikácie [3, 5].

Akceptačné testovanie užívateľom

Akceptačné testovanie užívateľom by malo zabezpečiť akceptáciu celkového výkonu aplikácie.

Hlavným dôvodom pre budovanie aplikácií dátových skladov je sprístupnenie dát užívateľom. Užívatelia poznajú dáta najlepšie a ich účasť na testovaní je kľúčovým komponentom úspechu implementácie dátového skladu. Testovanie akceptácie užívateľom je založené na princípe zavedenia dát do dátového skladu bez toho, aby užívateľ vedel ako pracuje ETL aplikácia. Musia však byť zvážené nasledovné stratégie:

- Použiť dáta, ktoré sú buď z reálu alebo dáta, ktoré sú reálu čo najbližšie. Užívatelia objavia nezrovnalosti, keď vidia reálne dáta. Niekedy to môže viesť až k zmenám architektúry a návrhu.
- Testovať dátové pohľady porovnávaním s očakávaným výsledkom. Je dôležité či užívatelia jasne pochopia ako sú pohľady vytvorené.
- Plán aktivít pre tím systémového testovania k tomu, aby podporoval užívateľov počas testovania akceptácie. Užívatelia budú mať pravdepodobne otázky k uloženiu dát a budú chcieť porozumieť tomu, ako pracuje ETL.
- Zvážiť ako často budú dáta obnovované alebo dopĺňané [3, 5].

Regresné testovanie

Regresné testovanie je potvrdenie funkčnosti s každou novou verziou aplikácie. Pri budovaní testovacích prípadov je dôležité zapamätať si, že budú pravdepodobne spúšťané viackrát, aj keď už je nová verzia vytvorená. Testovacie prípady môžu byť zoradené podľa

rizikovosti za účelom pomoci určenia, ktoré potrebujú byť opakované vždy pri novej verzii. Jednoduchá a stále účinná a efektívna stratégia je uchovávať zdrojové dátové sety a výsledky z úspešných spúšťaní kódu a porovnávať nové výsledky s predchádzajúcimi behmi aplikácie. Pri realizácii regresného testovania je omnoho rýchlejšie porovnávať výsledky predchádzajúceho behu, než robiť znova celú dátovú validáciu [3].

Záver

Akceptovanie týchto základných zásad počas návrhu a testovania dátových skladov zaistí, že aplikácie sú vytvárané kvalitnejšie a tiež zabránia drahým chybám, ktoré by mohli byť objavené pri nasadení a prevádzke.

Tento príspevok bol podporovaný **grantovou agentúrou VEGA v rámci projektu číslo 1/4078/07**, za čo všetci autori vyslovujú poďakovanie.

Zoznam bibliografických odkazov:

[1] INMON, W. H. *Building The Data Warehouse*. Wiley Computer Publishing, 2002. ISBN 0-471-08130-2
[2] HUMPRIES, M. *Data warehousing: návrh a implementace*. Computer Press, 2002. ISBN 80-7226-560-1
[3] THEOBALD, J. Strategies for Testing Data Warehouse Application. [cit. 18.2.2008]. Dostupné na internete <http://www.dmreview.com>
[4] KIMBALL, R., ROSS, M. *The Data Warehouse Toolkit: The complete guide to dimensional modeling.* Wiley Computer Publishing, 2002. ISBN 0-471-20024-7
[5] STEVENSON, D. Integration Testing. [cit. 18.2.2008]. Dostupné na internete <http://wiki.ittoolbox.com>
[6] SHARATH, R. BHAT. Data Warehouse Testing : Practical. [cit. 18.2.2008]. Dostupné na internete <http://www.stickyminds.com>