

Метаподатоците во складиштата на податоци

Виолета Маневска¹, Снежана Савоска²

¹⁾ violeta.manevska@uklo.edu.mk, ²⁾ savoskasnezana@yahoo.com

Апстракт

Новиот приод во размислување на компаниите насочено кон опстанок и позиционирање на глобалните пазари, неминовно побарува ново архитектонско окружување кое треба да се добие од оперативните бази со потполна интеграција на компаниските податоци во физичка смисла во складишта на податоци. За да се дефинираат експлицитно, потребно е да се користи компанискиот речник на податоци чија заедничка основа се метаподатоците на складиштето на податоци. Метаподатоците ги документираат податоците за податочните елементи или атрибутите, податоците за записите или структурата на податоците како и податоци за податоците. Метаподатоците го овозможуваат контекстот за податоците и се користат за да го олеснат разбирањето, користењето и управувањето со податоците од страна на луѓето. Без метаподатоците ниеден податок нема да има значење.

Клучни зборови: Метаподатоци, Складишта на податоци, Метаподатоци на складиштето на податоци

ВОВЕД

Новиот приод во размислување на денешните компании е насочен во правец на опстанок и позиционирање на глобалните пазари и неминовно побарува ново архитектонско окружување кое треба да се добие од оперативните бази со потполна интеграција на компаниските податоци во физичка смисла – во складиште на податоци. Ова е важен пресврт во користењето на компаниските податоци во интегрирана форма која е погодна за поддршка на одлучувањето¹. Во тој момент било потребно да се стандардизираат поимите и дефинициите за СП со користење на компаниски речник на податоци. Во литературата се среќаваат различни дефиниции за складиштето и за неговите делови. Некои дефиниции во речникот на

¹ Inmon W.H., *Building the Data Warehouse*, Fourth edition, Wiley Publishing, Inc., 2005

податоци се фокусирани на податоците, други се фокусирани на луѓето, софтверот, орудјата и бизнис продуктите. Заедничките особини на сите тие податоци се опишани со метаподатоците.

Метаподатоци или „метаинформации“ се „податоци за податоците“. Метаподатоците ги документираат податоците за податочните елементи или атрибутите (имиња, големини, типови на податоци и др.) и податоците за записите или структурата на податоците (должина, полиња, колони и др.) како и податоци за податоците (каде се лоцирани, како да се пристапи до нив, сопственост). Метаподатоците може да вклучат описни информации за контекстот, квалитетот, условите или карактеристиките на податоците. Може да се снимени со висока или ниска грануларност. Тие го овозможуваат контекстот за податоците и се користат за да се олесне разбирањето, користењето и управувањето со податоците. Ги опишуваат податоците концептуално за да другите луѓе (корисниците и администраторите) може да ги разберат и тоа во аналитичка и во агрегирана форма.

Софистицирани дефиниции за метаподатоците

Метаподатоците овозможуваат ефективна работа со различни типови податоци, контексти на користење и нивни намери. При користењето на базите на податоци, корисниците може да имаат пристап до различни метаподатоци, со чија помош можат да постигнат различни цели. Самиот поим крајни корисници значи луѓе или други пресметковни системи. Поради тоа, изречени се повеќе софистицирани дефиниции за метаподатоците²:

"Метаподатоците се структурирани, кодирани податоци кои ги опишуваат карактеристиките на информационите однесувања на ентитетите за да се помогне во идентификацијата, откривањето, проценката и управувањето со опишаните ентитети"³.

"Метаподатоците се множество на опционално структурирани описи кои се јавно достапни за експлицитна помош во лоцирање на објектите"⁴.

Michael Bracket ги дефинира метаподатоците (што ги вика "Податоци за ресурсите податоци") како "било кој податок за организациските податочни ресурси"⁵.

² Elmasri R., Navathe S.B., Fundamentals of Database system, 3-th edition, 2000

³ American Library Association, Task Force on Metadata Summary Report, June 1999

⁴ D. C. A. Bulterman, Is It Time For a Moratorium on Metadata?, IEEE MultiMedia, Oct-Dec 2004]

Adrienne Tannenbaum ги дефинира метаподатоците како "детални описи за примерите на податоците". Форматот и карактеристиките на „населените“ (вчитани, поставени) примери на податоци треба да се такви да обезбедуваат информации и да се зависни од улогата на примателот на метаподатоците⁶.

Метаподатоци на складиштето на податоци

Метаподатоците на складиштето на податоци може да се поделат на два дела: *back* (позадински) метаподатоци и *front* (челни) метаподатоци. Првите се користат за екстракција, трансформација и полнење на податоците од OLTP⁷ во складиштето, а вторите за означување на екраните и креирање на извештаи.

Kimball ги предложил следните типови на метаподатоци во складиштето:

- Метаподатоци за изворните системи (source system metadata);
- Спецификација на изворите како репозиториуми и изворни логички шеми;
- Описни информации за изворите, како опис на сопственоста, фреквенција на ажурирање и методи на пристап;
- Процесни информации како планот, задачите и кодовите за екстракција;
- Метаподатоци за состојбата на податоците (data staging metadata);
- Информации за собирањето на податоците, како план за трансмисија на податоците и користење на фајловите;
- Управување со димензиите табели, како дефиниција за димензиите и доделување на сурогат клучеви;
- Трансформација и агрегација, како унапредување на податоците и означување (мапирање), DBMS скрипти за полнење, агрегирање на дефинирање на податоците;
- Audit, job logs и документирање (потекло на податочните записи, logs за трансформација и слично).

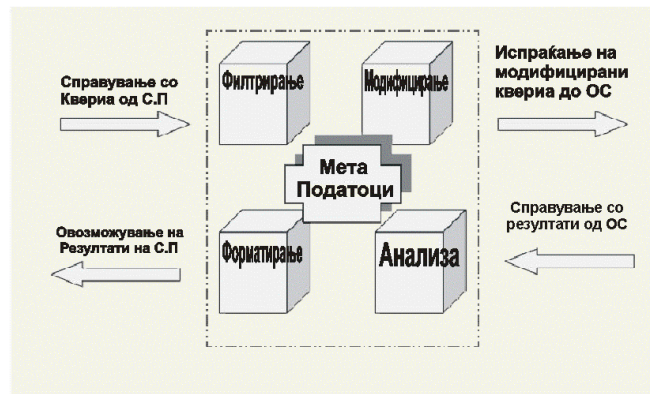
Метаподатоците треба да ги дефинираат и заедничките единечни мерки за сите елементи (ентитети) во различните табели. Задничките единечни мерки не

⁵ Gay V Tozer, Metadata Management for Information Control and Business Success, Arch House, 1999, ISBN 0-89006-280-3

⁶ Adrienne Tannenbaum, Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand, Addison-Wesley, 2002, ISBN 0-201-71976-2.

⁷ On Line Transactional processes

служат само за интеграција на различните мерни атрибути кои треба да бидат сместени во складиштето на податоци, туку и како глобално прифатлива единица мерка за крајниот податок. Треба да се постигне консензус во компанијата за единечните мерки кои ќе се користат во СП исто како и за метаподатоците во СП. Тие треба да се конвертираат при подготовката на податоците за полнење во складиштето на податоци. Користењето на единечните мерки и одлуката кои мерки ќе се користат може да биде ограничена и попречена од корисниците. Тие може да имаат широк опсег на барања и намери за користење на различни единечни мерки во кверијата. Во овие случаи, станува неопходно податокот да се конвертира во некои „нестандардни“ единечни мерки кои ќе содржат „најмал заеднички именител“ на користените единечни мерки и кои ќе дозволат пофлексибилна шема за креирање на различни единечни мерки во различни кверија.



Слика 1 Метаподатоците се податоци за податоците и централна точка за СП⁸

За да се обезбеди универзален пристап до податоците, неопходно е да се управува со одредени форми на директориуми со податоци или репозиториуми на информации од метаподатоците. Метаподатоците вклучуваат директориуми каде податоците се складирали со нивните имиња и значења, правила за користење на сумаризација и „прочистување“ т.е. овозможува увид во тоа како се подготвени податоците и како се складирали во складиштето на податоци, од каде се земени, т.е. „снимање“ на оперативните податочни извори.

Посебно интересни се метаподатоците во базите на модели и агрегираните табели во складиштата на податоци. Тие метаподатоци, покрај податоците за

⁸ Peter H., Greenidge C., *Aligning the Warehouse and the Web*; Wang J., *Data Warehousing and mining*, Second edition, USA, 2008

складираните податоци во нив, потребно е да содржат и дефиниции за алгоритми на моделите сместени во нив, период на важност и услови за нивна примена. Форматот на ваквите податоци најчесто е сложен, со униваријантни, биваријантни или мултиваријантни форми⁹.

Местото на метаподатоците во компаниите и нивното користење

Метаподатоците се мошне важен компаниски ресурс. За да се разберат детално метаподатоците, мора да се разбере синцирот на апстракција во компанијата. Метаподатоците може да ги споредиме со објект во објектното програмирање – секој има атрибути кој го дефинираат, класи и инстанци. Софтверот за управување со метаподатоците е поврзан со синцир на апстракција. Метаподатоците се едноставно апстракција на податоците и се опишани на концизен начин. Всушност, секој податок, како и секој збор во стварноста е полн со метаподатоци. Еве неколку примери:

- Автор, презиме, име, ISBN број, изведувач
- Наслови на написи во весници
- Бизнес дефиниции за сите податочни елементи во складиштето на податоци
- Карти на патиштата
- Дијаграм на проток на податоци во CASE репозиториуми со алатки и слично

Метаподатоците се суштински важни во трансформацијата на суровите податоци во знаење. Ако податоците бидат измешани, ако не се знае кој податок што значи и претставува, во која единечна мерка, од каде потекнува, тогаш податоците нема да имаат смисол. Значи, секој податок треба да одговара на метаподатокот за него, документот и граfiците на нивните соодветни податоци. Во спротивно, сè ќе биде измешано и невидливо за нашите сетила, па може да се заклучи дека без метаподатоците, податоците во складиштето се безначајни и нема да знаеме како да ги искористиме. Едно од предложените решенија за метаподатоците во СП е тие да се чуваат во самата база за да бидат секогаш достапни¹⁰.

⁹ Adrienne Tannenbaum, Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand, Addison-Wesley, 2002, ISBN 0-201-71976-2

¹⁰ George M. Marakas, Modern Data Warehousing, Mining and Visualization, 2003, Indiana University, New Jersey;

За да се прикаже како се користи складиштето на податоци, потребно е да се обезбедат метаподатоците, да им бидат достапни на аналитичарите на СП. Аналитичарите и менаџерите секогаш се во потрага по нови и интересни модели на податоци кои се репликација на оперативните или внесените (импортираните) екстерни податоци. Поради ова, потребно е шемите на метаподатоците секогаш да се ажурни и да се обезбеди нивниот интегритет. Управувањето со метаподатоците на СП (Слика 2) се врши со визуелен интерфејс што претставува огромно олеснување за администраторите но и за крајните корисници кои се помош на Help-фајлови може да ги дознаат податоците за податоците.



Слика 2 Управување со компонентите на складиштето на податоци¹¹

За да се добијат конзистентни и валидни метаподатоци за СП, потребно е да се добијат одговори на прашањата кои се дадени во Табела 1¹².

Екстракцијата на метаподатоците, како и екстракцијата на информациите кои се однесуваат на врските што се содржат во складиштето, мора да е управувана. Метаподатоците може да припаѓаат на поврзани табели кои имаат некои ограничувања (constraints) и треба да се управувани според нив и според описите во записите за сопственост. Сите овие информации, како и историјата на промени, се многу важни за донесувачите на одлуки бидејќи можат да им овозможат разновидни анализи од аналитичките податоци или од повеќекратните агрегации на податоците. Алгоритмите за агрегација на деталните податоци се многу важни како

¹¹ Превод од Bellatreche L., Mohania M., Physical Data Warehousing Design; Wang J., Data Warehousing and mining, USA, 2008

¹² George M. Marakas, Modern Data Warehousing, Mining and Visualization, 2003, Indiana University, New Jersey;

поради фактот што тие се користат во креирањето на бази на модели и знаење, така и поради тоа што на тој начин збирните податоци се достапни веднаш. Ако се знаат алгоритмите за сумаризација (агрегација), тие може да се употребат и во друг контекст и во друго време, т.е. да се бараат во базата на модели на складиштето на податоци.

Табела 1- Потребни податоци при креирање на метаподатоци на складиштата

- Кои табели, атрибути и клучеви ги содржи складиштето на податоци?
- Од каде доаѓа секое множество на податоци?
- Кои логички трансформации треба да се направат на податоците за да се наполнат во складиштето на податоци?
- Како се менуваат метаподатоците низ времето?
- Дали постојат ознаки за alias-и, како се поврзани со другите елементи на складиштето?
- Кои се заеднички врски помеѓу техничките поими, ентитетите и бизнис поимите?
- Колку често треба да се „полнат“ податоците во складиштето на податоци?
- Колку податочни елементи има во складиштето на податоци?

Механизмите за трансформација на податоците – метаподатоците за нив, „запамтуваат“ како се трансформирани податоците од оперативните трансакциски бази и од екстерните извори, по кои алгоритми и на кој начин се ставени и наполнети во складиштето на податоци. Со користење на современите орудија за складишта на податоци, процесот на означување на метаподатоците станува неодоив од процесот на трансформација, се одвива на визуелен начин и се креираат дијаграми со табели и врски помеѓу нив, како и примарни и надворешни клучеви на табелите во базата на податоци. Во Табела 2 се наведени неколку типови на мапирање (означување) на метаподатоците во складиштето на податоци.

Оперативните бази на податоци најчесто се сопственост на различни бизнис единици или сектори во организацијата. Во окружувањето на складиштето на податоци на компанијата, сите податоци се зачувани во заеднички формат и се достапни за овластените корисници. Поради ова, корисно е да се разграничи „сопствеништвото“ над податоците од оперативното окружување и нивните наследени особини да бидат пренесени над складиштето на податоци. Бидејќи одговорен за полнење, филтрирање, сумаризација и дисеминација на податоците во

складиштето на податоци е администраторот на податоци, тој е исто така одговорен за спецификацијата на потеклото на податоците во складиштето. За точноста на податоците во оперативните системи е одговорен администраторот на оперативните податоци.

Табела 2

Како точно да се означуваат (мапираат) метаподатоците во складиштето на податоци

- Идентификација на оригиналните извори на податоци со изворните полиња
- Еднакво атрибут-атрибут означување (мапирање)
- Конверзија на атрибути
- Конверзија на физичките карактеристики
- Конверзија на енкодирањето – креирање на референтна табела за кодови
- Измени на именувањата и табела на измените
- Зададени вредности кои се дозволени за складиштето
- Дефинирање на логика за избор од содржината на различни извори
- Промена на алгоритмите и документирање на оригиналните и променетите

За да се подесат и оптимизираат параметрите на складиштето на податоци, треба да се одреди моделот на пристап до складиштето на податоци и потребната оптимизација на параметрите¹³. Треба да се дефинираат и идентификуваат и некои помали, но значајни типови метаподатоци кои обично се користат за опис на модели, бизнис логика, а кои ги идентификуваат ентитетите, врските и правилата на секое поединечно бизнис опкружување¹⁴.

Друга категорија на метаподатоци која помага во лесно користење на складиштата на податоци се „alias“¹⁵ метаподатоците. Alias-ите се корисни кога различни бизнис единици ги користат своите сопствени имиња за да се обрратат до основните податоци. Внимателно обележување на alias-ите за податочните елементи е важен дел од метаподатоците во складиштето на податоци. Се користат и недизајнирани табели во фазата на тестирање или кога складиштето на податоци е сèуште неактивно.

¹³ Ако се знае до кои табели треба да се пристапи, колку често, кога да се бара упростување на менија за корисникот, кои се поретко користени податоци за да мигрираат на поефтин медиум, како да се забрза пристапот до почесто користените податоци и слично

¹⁴ Тие помагаат да се решат двосмисленостите и несигурностите при известувањето.

¹⁵ Се користат за скратување на описите и имаат форма на кратки „прекари“ на апсолутните и релативните имиња на описите на деловите на складиштето на податоци

Source	ground	rail	road	sea
nonground				
air				

Route	190	215	160	240
Africa	Feb-17-99	Apr-22-99	Sep-07-99	Dec-01-
Asia	550	600	520	780
Eastern Hemisphere	Mar-19-99	May-31-99	Sep-19-99	Dec-22-
Australia	212	240	300	410
Europe	Mar-05-99	May-19-99	Aug-09-99	Nov-27-
Western Hemisphere	500	470	464	696
North America	Mar-07-99	Jun-20-99	Sep-11-99	Dec-15-
South America	3056	4050	4360	5112
	Mar-30-99	Jun-28-99	Sep-30-99	Dec-29-
	600	490	315	580
	Feb-27-99	Jun-03-99	Aug-21-99	Nov-30-

Слика 3 Метаподатоците на OLAP мултидимензионалните коцки се дефинираат во фазата на креирање на ROLAP или MOLAP

Кога се работи за метаподатоците на мултидимензионалните коцки од типот ROLAP и MOLAP, карактеристично е тоа што метаподатоците за нив се дефинираат во самиот процес на нивно креирање. Тоа подразбира дефинирање на изворите на податоци, димензиите и мерките на мултидимензионалните коцки и одредиштето на податоците. Алатките кои вршат екстракција, трансформација и полнење побаруваат интерактивно внесување на димензиите кои ќе бидат предмет на анализа на сервисот, како и мерките кои ќе бидат анализирани¹⁶.

Заклучок

Метаподатоците се неопходен и суштински елемент на архитектурата на складиштето на податоци. И двата вида на метаподатоци за СП, позадински и челни се особено важни при неговото креирање, како поради дефинирањето на точното значење на секој податок во складиштето, така и поради неговата употреба од страна на крајните корисници. Поради тоа, тие се чуваат во посебни директориуми кои треба да се соодветно управувани и достапни до администраторот на складиштето. Тој има должност да ги одржува ажурни и да направи Help-екрани кои ќе им ги појаснуваат значењата на податоците во складиштето. Ова е посебно важно кога се извршени декодирања, трансформации или агрегирање на податоците, како и креирање на бази на знаење од податоците со

¹⁶ Софтверот кој сето ова го овозможува е OLTP Analysis services (Microsoft), Oracle Data Warehouse Builder, SAS или други алатки за аналитичко процесирање на податоците.

примена на соодветни алгоритми. Добро дефинираните и документираните метаподатоци во складиштето обезбедуваат непречено користење на складиштето од страна на крајните корисници.

Користена литература

- [1] Inmon W.H., Building the Data Warehouse, Fourth edition, Wiley Publishing. Inc., 2005
- [2] George M. Marakas, Modern Data Warehousing, Mining and Visualization, 2003, Indiana University , New Jersey;
- [3] Bellatreche L., Mohania M., Physical Data Warehousing Design; Wang J., Data Warehousing and mining, USA, 2008
- [4] Elmasri R., Navathe S.B., Fundaments of Database system, 3-th edition, 2000
- [5] American Library Association, Task Force on Metadata Summary Report, June 1999
- [6] D. C. A. Bulterman, Is It Time For a Moratorium on Metadata?, IEEE MultiMedia, Oct-Dec 2004
- [7] Gay V Tozer, Metadata Management for Information Control and Business Success, Arech House, 1999, ISBN 0-89006-280-3
- [8] Adrienne Tannenbaum, Metadata Solutions: Using Metamodels, Repositories, XML, and Enterprise Portals to Generate Information on Demand, Addison-Wesley, 2002, ISBN 0-201-71976-2
- [9] Peter H., Greenidge C., Aligning the Warehouse and the Web; Wang J., Data Warehousing and mining, Second edition, USA, 2008

Metadata and Data Warehouse

Manevska V., Savoska S.

Abstract

The new approach of company's view for concurency and positioning of the global market obviously demand a new architecture frame for databases which have to be created from operational company's databases. It have to integrate completly the company's data in the physical maner in company's data warehouse. To explicit define all entities and issue, it is necessary to use a company's data dictionary. The main base of data dictionary for data warehouse is metadata. Metadata documents data about data or attributes, data about the data structure and data about locations, positioning and physical data place. Metadata show the context of data and can be used to facilitate the human undestanding, usage and data managing. The data without metadata is uneligable and without sense.