ОБУЧЕНИЕ ПО ИЗВЛИЧАНЕ НА ЗНАНИЯ И СКЛАДОВЕ ЗА ДАННИ*

н.с. III ст. Вера Маринова – Бончева

ИИТ – БАН, ул. "акад. Г. Бончев" бл. 2
vera_boncheva@yahoo.com

Този доклад има за цел да представи за дискусия програми за обучение по извличане на знания и складове за данни като технологии и инструменти, които се прилагат ефективно в процесите на анализ и откриване на знания за подпомагане взимането на решения.

Ключови думи: *извличане на знания, откриване на знания, складове за данни, центрове за данни*

1. **История**

През последните две десетилетия и особено през 90-те години организациите складират огромни количества от данни като са построяват OLTP (Online Transaction Processing) системи, ERP (Enterprise Resource Planning) системи, call центрове и Интернет. В търсене на по-добро управление на предприятията се построяват складове за данни, центрове за данни и се инсталират ETL (Extract Transform Load) инструменти за работа със складовете за данни. Но много малка част от тези данни са били превърнати в информация и са се използвали в системи за поддръжка на решенията поради липсата на начини за достъп и анализ на данните от бизнес потребителите. С напредъка на новите технологии и приложения бизнес интелигентността достига сегашното си състояние и продължава да се развива, предоставяйки аналитични системи. [1]

Днес бизнес интелигентността обхваща процесите по събиране, управление и анализиране на големи обеми от данни с цел подпомагане на стратегическите бизнес решения на една компания, както и улесняване на служителите от различни отдели да реагират гъвкаво на пазарните промени. Всяка бизнес интелигентна система изисква наличие на добра инфраструктура за работа с големи обеми от данни, организирани в складове за данни, центрове за данни, извличане на знания, web майнинг или текст майнинг. [4] Необходими са и OLAP (Online Analytical Processing) средства, с помощта на които данните се трансформират в полезна информация и подпомагат взимането на решения.[2]

* Тази работа е изпълнена по договор с вх. № МУ-106 с Министерството на образованието и науката

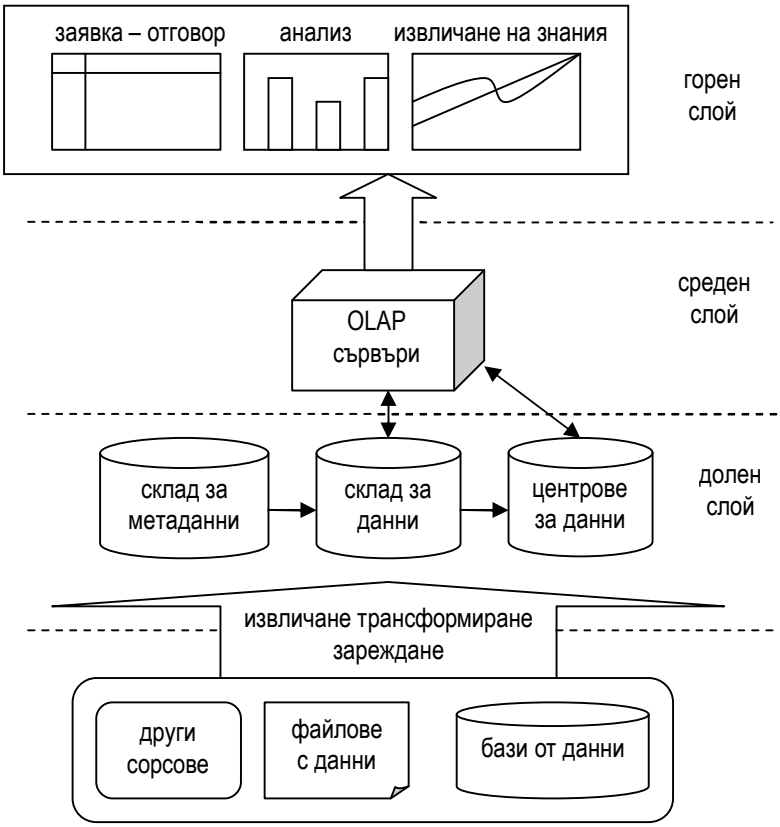
Съвременните бизнес системи изискват добро познаване на извличането на знания и складовете за данни като обща методология на архитектури, технологии и инструменти за подпомагане на процеса за взимане на решения.

Това налага изграждането на високо компонентни специалисти със знания и умения по извличане на знания и складове за данни. Естествено място на обучение на такива специалисти могат да бъдат магистърските програми не само по информационни технологии. Такава практика има в редица чужди университети. За да бъдат накратко аргументирани учебните програми, може да се маркират двете основни архитектури на склад за данни и на системата за извличане на знания.

2. Архитектура на склад за данни

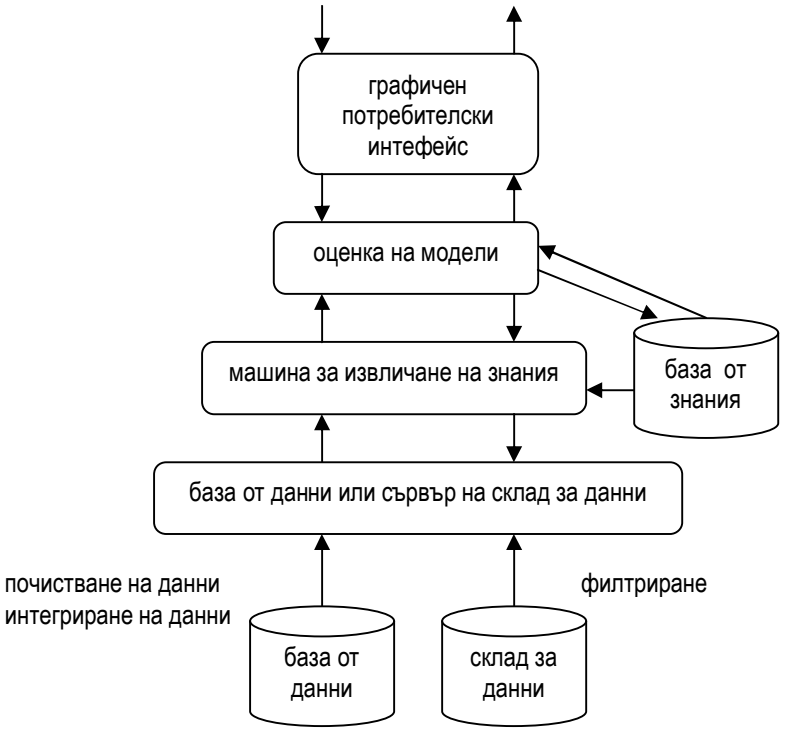
Архитектурата на един склад за данни е изградена от определен брой свързани части и най-често представлява трислойна архитектура, като представената на фиг. 1. [3]

Фиг. 1 Архитектура на склад за данни



След като се премине през различните етапи като: разбиране на извлечените данни, трансформация и почистване на данните, може да се смята, че данните са доставени за съхранение в склада за данни. С помощта на аналитичните услуги, които представляват инструменти за организиране на данните от склада в мултидименсионни кубове, и OLAP инструментите може да се създадат големи масиви от данни по начин, по който клиентските инструменти и приложения да могат лесно и бързо да отправят запитвания и да анализират данните. Най-накрая върху извлечените данни може да се приложат методите на извличане на знания от тези данни, което знание е необходимо, за да се разкрият модели, скрити в огромни множества с данни. [2]

3. Архитектура на система за извличане на знания



Фиг. 2 Архитектура на система за извличане на знания

Архитектурата на една типична система за извличане на знания (виж. Фиг. 2) се състои от следните компоненти:

- База от данни, склад за данни или друго информационно хранилище – тук се изпълняват техники по почистване на данните и тяхната интеграция, а също така и трансформиране и зареждане на данните.
- Бази от данни или сървър на склад за данни – отговорни са за прихващането на съответната данна на базата на потребителската заявка.
- База от знания – има се предвид знание за област, което се използва, за да подпомогне изследването и да оцени доколко са интересни крайните модели. Такова едно знание може да представлява концептуалните йерархии, потребителското доверие, допълнителните ограничения или прагове на интерес и метаданните (т.е. данни, извлечени от данните).
- Машина за извличане на знания – това е "мислещата част" на системата и в най-добрия случай се състои от множество от функционални модули, изпълняващи задачи като характеристика, асоциация, класификация, клъстърен анализ, анализ на тенденции, предсказване и други техники за извличане на знания.
- Модул за оценка на модели – този компонент използва мерки на заинтересованост и взаимодейства с модулите за извличане на знания, така че да се фокусира върху търсенето на интересни модели. Този модул може да се интегрира с извличащия модул, в зависимост от имплементирането на използвания метод за извличане на знания. За да има ефективено извличане на знания се препоръчва да се извърши оценка на интересните модели колкото се може по-навътре в извличащия процес.
- Графичен потребителски интерфейс – този модул осъществява взаимодействие между потребителя и системата за извличане на знания, като позволява на потребителя да преглежда базите от данни и схемите на складовете за данни или структурите от данни, да оценява и да визуализира моделите по различен начин, напр. чрез таблици, схеми, диаграми, графики и други. [3]

4. Примерни учебни програми

4.1. Обучение по складове за данни с общ хорариум от 30 учебни часа - лекции и 15 учебни часа - упражнения.

Програмата за обучение по учебната дисциплина складове за данни може да включва следния списък от примерни теми:

Тема 1 Склад за данни - въведение.

Тема 2 Архитектура на склад за данни. Инфраструктура на склад за данни. Метаданни.

Тема 3 Проектиране и подготовка на данните в склад за данни. Мултидименсионно моделиране.

Тема 4 Центрове за данни.

Тема 5 Информационен достъп и доставка. OLAP технология в склад за данни.
Тема 6 Изпълнение и поддръжка на склад за данни.
Тема 7 Тенденции за развитие.

Упражненията, свързани с изучаването на складове за данни, могат да започнат с разглеждане на съпоставки между релационните бази от данни и складовете за данни. Студентите имат възможност да придобият практически умения за разработване и дизайн на склад за данни или център за данни, чрез избор на подходяща схема – звезда, снежинка или съзвездие, с помощта на SQL Server 2000, които предоставя Enterprise Manager - среда за мултидименсионно моделиране на данни във вид на факт таблици и димeнсии (димeнсионни таблици) и Analysis Manager - за построяване на димeнсионни кубове и осъществяване на анализ на данните. Върху тези мултидименсионни кубове се осъществява анализ на данните с помощта на OLAP инструменти, които чрез операции като: slice and dice, drill down/across/through, roll up и pivot могат да извършат детайлизиране или обобщаване на данните и тяхното разглеждане от различен ъгъл. Тук се съпоставят OLAP и OLTP системите. [2]

4.2. Обучение по извличане на знания с общ хорариум от 30 учебни часа – лекции и 15 учебни часа – упражнения.

Обучението по извличане на знания може да обхваща следните примерни теми:
Тема 1 Основи на извличането на знания.
Тема 2 Складове за данни и OLAP технология за извличане на знания.
Тема 3 Архитектура на извличането на знания.
Тема 4 Подготовка на данните за извличане на знания.
Тема 5 Техники за описание на концепции – характеризиране и разграничаване. Начини за представяне на знания и визуализация.
Тема 6 Модели и концепции за извличане на знания.
Тема 7 Класификация и предсказване.
Тема 8 Клъстерен анализ.
Тема 9 Откриване на знания в сложни типове от данни.
Тема 10 Приложение и бъдещи тенденции в извличането на знания.

В упражненията по извличане на знания може да се включи запознаване със специализиран DMQL (Data Mining Query Language), който представлява SQL базиран език. В DMQL една заявка за извличане на знания се дефинира в термините на следните примитиви: множество от данни, свързани с дефинирана задача; вид на знанието, което се търси; историческо знание, което се използва в откривателския процес; мерки на интерес и прагове за оценка на модела; очаквано представяне и визуализация на откритите модели.

След това заявката се подава на системата за извличане на знания. По нататък може да се включи и изучаване на OLE DB for Data Mining на Microsoft, който предоставя езикови примитиви за опериране с виртуален обект, т.н.

модел за извличане на знания. За разлика от OLAP, чрез който се изпълнява анализ на минали и текущи данни, то този модел осигурява структура, която позволява да се съхранява „наученото" от данните, например вероятности и информация и на тяхна база да се прогнозира поведението на нови данни, т.е. да се прави прогнозен анализ, с който да се подпомага взимането на решения. Analysis Manager на SQL Server 2000 предоставя инструменти за използване на извличащи техники като: дървета на решения и клъстериране. [2]

Във ФМИ на СУ "Св. Климент Охридски" в магистърската програма по информатика, спец. "Софтуерни технологии" е включена дисциплина „Data warehouse", а в Стопанския факултет-СУ в магистърска програма "Управленски информационни системи" се изучава дисциплина "складове от данни". [5]

Обучението по извличане на знания и складове за данни трябва да завърши с изработване на курсова работа.

5. Цели на програмите за обучение

Тези програми са предназначени за магистърски програми за студенти, които са изучавали бази от данни и имат програмни умения, например от специалност информационни технологии. По тези програми може да се обучават и магистри, които са от не информационни специалности след съответно подходящо адаптиране.

Целите на тези програми са да се запознаят обучаваните с основните концепции в съвременните информационни технологии за аналитична обработка и подпомагане изработването на управленски решения като складове за данни, центрове за данни, структури от данни, които осигуряват възможности за мултидименсионно моделиране, принципи на онлайн аналитична обработка, извличане на знания, както и взаимната връзка между тях като компоненти на аналитичните системи.

Литература

[1] А. Розева, „Компютърни технологии в помощ на управлението на бизнеса", София, 2001

[2] Джефри Шапиро, "SQL Server 2000 / Пълно ръководство за програмиране", Инфодар - София, 2001

[3] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, ISBN 1558609016, 2006.

[4] Daniel T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, ISBN: 0471666572, John Wiley, 2004.

[5] http://www.fmi.uni-sofia.bg/education/magisters/uchebni-planove-magistrski-programi/informatica/