**ТЕМА 3. ИНФОРМАЦИОННИ ТЕХНОЛОГИИ ЗА ОБРАБОТКА НА ИНФОРМАЦИЯ В БИЗНЕС ИНТЕЛИГЕНТНИТЕ СИСТЕМИ**

**Съдържание на темата:**

3.1. Бази от данни и складове от данни. Технологии OLTP и OLAP.

3.2. Интелигентни технологии в БИС.

**3.1. Бази от данни и складове от данни.** **Технологии OLTP и OLAP.**

Данните са основен източник на информация в БИС. За това проблемите, свързани с тяхната организация и администриране са сред най-важните при създаване на БИС.

Базата от данни (БД) и складът от данни са видове организация на данните. Eдно работно определение на БД е: „организирана колекция от логически свързани данни”.

Според Дж.Мартин (James Martin), един от водещите учени в областта на базите от данни, този вид организация на данните се дефинира както следва: “Базата от данни може да се определи като съвкупност от взаимосвързани, съхранявани заедно данни, при наличието на такава минимална излишност, която допуска тяхното използване по оптимален начин за едно или няколко приложения; данните се запомнят така, че да бъдат независми от програмите, които ги използват. За добавяне на нови или модификация на съществуващите данни, а така също за търсене на данните в базата от данни се използва общ управляем способ”. Базата от данни е общодостъпно множество от данни, което е специално организирано да задоволява потребностите на различни потребители на информационната система, в зависимост от определени правила на достъп. Към базата от данни се поддържа програмно осигуряване за обработка и администриране на данните, а също и езици за описанието им, за формиране на структурирани по определни правила запитвания, за предоставяне на потребителски интерфейс към базата. Тази специална система от програми се нарича **система за управление на базата данни** (СУБД).

Чрез СУБД се осигуряват възможности за създаване структурата на БД и различни модели на данните. Обезпечава се и независимост на данните в два аспекта- логически и физически. Логическият аспект е свързан с това как данните са организирани и достъпвани от приложните програми и потребителите. Физическият аспект се отнася до това как данните се съхраняват физически върху външната памет.

**Схема на данните** е описанието на логическата структура на данните, специфицирано на език за описание на данните и обработваемо от СУБД.

Към БД и СУБД се поставят следните изисквания:

* високо бързодействие;
* простота на обновяване на данните;
* независимост на данните. Независимостта на данните е свързана с отделянето на описанието на данните от програмите за тяхното управление и обработка. Това е възможност за изменение на логическата и физическата структура на БД без изменение на представянето за потребителя.
* възможност за многопотребителско използване на данните;
* безопасност на данните. Това е защита от преднамерено и непреднамерено нарушение на секретността, изменение или разрушение. Безопасността включва 2 компонента- **цялостност и защита на данните** от несанкциониран достъп. Цялостността е устойчивост на съхраняваните данни на разрушение и унищожение поради неизправни технически средства, системни грешки и погрешни действия на потребителите. Защитата предполага ограничение на достъпа до данните, разграничаване на правата на достъп, използване на пароли и т.н.
* стандартизация на построяването и експлоатацията на БД. Стандартизацията обезпечава приемственост между поколенията на конкретна СУБД, опростява взаимодействията на БД на едно поколение СУБД с еднакви и различни модели на данните (релационни, многомерни, мрежови или йерархични).
* дружелюбен интерфейс.

Описанието на данните е представяне на структурата на данните (техния модел) в определена база от данни. Един от най- разпространените модели е **релационният**. Тъй като той е приложим за всяко представяне на данните и притежава редица предимства, ще опишем накратко същността му. При релационния модел, базата от данни се приема като съвкупност от двумерни таблици, които са свързани помежду си на основата на общи полета. Елементите в даден стълб на таблицата са от един и същ вид (числов, дата, текст). Стълбовете имат уникални (неповтарящи се) имена, които са имена на полета. Наборът от значения на елементите, разположени в отделните стълбове, но в рамките на един ред от таблицата, изразяват едно отношение и представляват един запис в базата от данни. Всяко отношение има първичен ключ (често и алтернативни ключове), по който се извършва достъпа до записа.

Като модели за описание на данните се използват също мрежови, йерархични, обектно- ориентирани и други.

Има няколко начина на класифициране на променливите (величините) в БД. Променливите могат да бъдат числови или текстови. Те могат да бъдат непрекъснати (continuous- приемат някаква реална числова стойност, обикновено в даден диапазон), целочислени (integer), или категорийни (categorical- ако приемaт, една от ограничен брой стойности). Категорийните променливи могат да бъдат или числови (1, 2, 3) или текстови (плащания, неплащания, несъстоятелност). Категорийните променливи могат да бъдат и неподредените (наречени номинални променливи), напр. като категориите Америка, Европа и Азия; или да бъдат подредени (наречени редни променливи), като напр. категориите: висока стойност, ниска стойност, нулева стойност.

**Складовете от данни** (СД) са съвременна форма на организация на данните. Те представляват огромни обеми от строго структурирани, дългосрочно натрупвани данни, предоставящи коректна и систематизирана представа за даден обект. William, Bill и Inmon H. предлагат много полезна дефиниция за склад от данни: “Складът от данни е предметно ориентирана, интегрирана, зависима от времето и непроменлива колекция от данни за поддръжка на управленския процес на вземане на решения” ([http://infogoal.com/datawarehousing/ overview.htm](http://infogoal.com/datawarehousing/%20overview.htm)).

В. Кисимов oпределя понятието „склад от данни“ по-широко: “Складът от данни може да се разположи в няколко компютъра и може да съдържа няколко бази от данни, както и да съдържа информация от многобройни източници в различни формати. Достъпът до склада от данни е прозрачен за потребителя (бизнес анализатора), който може да използва прости команди, за да открие и анализира необходимата информация. Складът от данни съдържа също и метаданни - информация за това как е организиран самият склад от данни, къде може да се открие информацията в него и какви са връзките между данните. Метаданните в складовете от данни осигуряват въведените данни да са в правилния формат и да притежават съответните зависимости с другите данни в складовете от данни. Складът от данни обикновено съдържа огромно количество данни, организирани така, че да подпомагат вземането на решения“.

Складовете от данни могат да се характеризират с няколко типични качества.Такива са:

* тематично ориентиране. Складовете за данни обобщават данни, които са организирани около една основна тема или обект и имат връзка помежду си. Дадена организация може да поддържа и повече от един склад за данни, за да се удовлетвори това изискване.
* интегритет. Складът за данни събира данни от всички оперативни приложения в организацията и се грижи тези данни да са консистентни (съвместими). Складовете за данни уеднаквяват входящите данни и по този начин улесняват достъпа до там през клиентски приложения. Пример: Представете си, че отдел продажби във Вашата организация поддържа Excel-ски документ, в който вписва всеки клиент и срещу името му удостоверява и от коя държава е той, напр. “Bulgaria”. ERP системата ви за обработка на фактури, обаче, използва номенклатура „BG” за същия този клиент. При обединяване на записите от тези два източника в склада за данни следва да се наложат правила за консистентност на данните, така че складът за данни да зададе „единната истина” за тези записи и да предотврати грешни отчети, които вземат само “Bulgaria” или само „BG”, когато искате справка за продажбите си от страната.
* постоянство във времето. Данните в склада никога не се презаписват или изтриват. След като веднъж са попаднали в склада, данните стават статични и достъпни за четене. Всички промени върху тези данни се записват чрез определени техники и могат да бъдат отразени при бъдещо достъпване и анализ на данните.

Данните в СД може да се класифицират в следните групи:

* първични оперативни данни. Напр. продажба на стока за определена цена в даден град и определено време)
* агрегирани данни. Преминали са обработка по натрупване или обединение. Пресметнати са бавните изчисления. Напр. сумата на продажбите за 1 месец.
* метаданни. Поддържат описание на свойствата и характеристиките на други данни. Напр. информацията за продажбите се състои от 3 таблици със задължителни параметри номер на продажба и цена.
* бизнес правила. Примерно представляват ограничения за диапазон на стойностите и начин на записването им. Например, цената трябва да е в определен диапазон.
* схеми на базите от данни. Напр., таблици за цени със специафикация на колоните, първичните ключове и връзките между тях.

Съдържанието на СД се обновява през определени периоди от време или в режим реално време (оn-line). Зареждането на данните в СД е чрез процеса ETL (еxtract, transform, load)-извличане, трансформиране и зареждане. ETL включва:

* Извличане на данни от хомогенни или хетерогенни източници;
* Трансформиране на данните, за да могат да бъдат запазени в подходящ формат или структура, с цел след това от тях да се извличат справки и да се правят анализи;
* Зареждане на данните в крайната цел.

Обикновено тези три фази се изпълняват паралелно. Тъй като извличането на данни отнема време, докато то протича, друг процес на трансформация се изпълнява. Последният обработва вече извлечените данни и ги подготвя за зареждане. Веднага щом има готови данни, тяхното зареждане започва, без да се чака приключването на предишните фази.

Основни понятия в многомерния модел на данните са измерение, факт, n- мерен куб.Измеренията имат качествен характер и категоризират данните по определени критерии, напр. време, регион, видове услуги. Стойностите на измеренията се организират в йерархии, с помощта на които могат да се правят анализи на различни нива на агрегация. Фактите имат количествен характер и съхраняват стойностите на анализираните показатели. Справките при много мерния модел се представят като n- мерни кубове, по осите на които се разполагат съответните измерения, които може да бъдат йерерхично изградени. Например, за измерението „време“ йерархията може да включва година, месец, седмица, ден. Точките на така зададеното n- мерно пространство представят фактите. Потребителят има възможност да визуализира различни разрези и сечения на дефинирания куб.

Многомерните анализи са насочени към въпроса „Какво става?“ а не толкова към „Защо е така?“

**Витрините от данни** (Data Marts) са екстракти от СД, фокусирани върху една тема – отдел на фирмата, продажби, финанси, маркетинг или др.

Някои от причините за създаване на витрините от данни са:

* Улеснен и по-бърз достъп до често използвани данни.
* Осигурява се защита на данните, тъй като определена група потребители ползва дадена витрина от данни и има оторизиран достъп до нея.
* Използване на определена схема на данните за дадени анализи.

Обикновено складовете от данни се поддържат отделно от оперативните бази данни. Една от причните за това е, че при първите се използва технологията OLAP (оn-line analytical processing)- аналитична обработка в реално време, чиито функционални изисквания са напълно различни от on-line транзакционната обработка- ОLTP.

OLTP се характеризира с голям брой малки он-лайн транзакции (вмъкване, актуализиране, изтриване). Основният акцент при OLTP се поставя върху много бързата обработка на заявки, поддържане целостта на данните в среда с множествен достъп и на ефективността, измерена чрез броя на заявките в секунда. Данните в OLTP системите постоянно се обновяват. Но те не може директно „да се излеят“ от системата за OLTP в многомерна база данни, защото OLTP данните е лесно да бъдат замърсени от непълни транзакции. Освен това директният достъп до данните за OLTP може да наруши работата и намали производителността.

OLAP (он-лайн аналитична обработка) се характеризира със сравнително малък обем на тразакциите. Заявките често са много комплексни и включват големи съвкупности от данни. OLAP кубът е агрегиран набор от данни, който подпомага бързия анализ. За това, обикновено анализът върху OLAP куб взема 0.1 от времето, което би било необходимо, ако се ползват данни от оперативните БД[[1]](#footnote-1).

OLAP приложенията са широко използвани от Data Mining техники за извличане на зависимости от данни.

Сравнение между технологиите OLТP и OLAP е направено в следващата таблица.

| **Категория** | **Online Transactional Processing** | **Online Analytical Processing** |
| --- | --- | --- |
| Източници на данни | Oперативна информация за системите. Тези процеси са източниците на данни. | Исторически и архивни данни |
| Фокус | Актуализиране на данни. | Отчитане и извличане на информация. |
| Приложения | Оперативно управление, web услуги, client-server. | Управленски системи. Вземане на решения. |
| Потребители | Служители и персонал на фирмата. | Mенажери, изследователни на данни, маркетолози. |
| Задачи | Oперативни. Вмъкване, актуализация и операции по изтриване. | Отчети и анализи на данни. |
| Обновяване на данни | Изпълнява се бързо. Незабавни резултати. | Обновяване на данни с огромни множества от данни. Отнема време. |
| Moдел на данните | Връзка между същности в БД. | Едно или мулти дименсионни данни. |
| Схема | Много таблици и връзки. | Звезда, снежинка и съзвездие. По-малко таблици. |
| Backup | Редовно архивиране.Данните са критични и не може да бъдат загубени. | Архивиране или ново зареждане на механизма, който поддържа добавянето на данни. |
| Хоризонт | Дневен, седмичен, месечен. | Дългосрочни данни. |
| Заявки | Прости, връщат резултатите от активността на системата. | Комплексни заявки на данни с цел да се обедини информация за докладване. |
| Скорост | Висока. | Ниска. Зависи от обема на данните. |
| Пространство | Запомнени оперативни данни, типично малко. | Големи множества от историческа информация. Необходима е голяма памет. |

Съществуват 3 начина за съхраняване на данните в OLAP системите или 3 архитектури OLAP сървъри:

* MOLAP (Мultidimensional OLAP);
* ROLAP (Relational OLAP);
* HOLAP (Hybrid OLAP).

При MOLAP изходните и многомерни данни се съхраняват в многомерна БД или многомерен локален куб. Този начин на съхранение осигурява висока скорост на изпълнение на OLAP операциите. Но многомерната база в този случай е с излишества. Кубът, построен на тази основа, зависи от броя на измеренията. При увеличаване на броя им, обемът на куба нараства експоненциално.

При ROLAP изходните данни се съхраняват в релационни БД или в плоски локални таблици. Агрегатните данни може да се разположат в служебни таблици в същата БД. Преобразуването на данните от релационни БД в многомерни кубове е при поискване на OLAP средствата. Скоростта на построяване на куба силно зависи от източника на данни и за това времето за отговор може да стане доста голямо.

При използване на хибридна архитектура (НOLAP), изходните данни остават в релационна база, а агрегатите се разполагат в многомерна. Построяване на OLAP куба се изпълнява при поискване на OLAP средствата на основата на релационните и многомерни данни. Така се избягва големия ръст на данните и може да се постигне оптимално време за отговор на клиентски запитвания.

Други варианти на архитектури са:

* Web OLAP. Отнася се до OLAP приложение, достъпно чрез уеб браузър. Използва се 3 слойна архитектура с компоненти: клиент, мидълуеър и сървър за данни.
* DOLAP. Базиран е на идеята, че потребителят може да зареди част от данните от БД или от друг източник и да работи с тях локално или върху desktop-a. Toва е лесно постижимо и евтино, но е с по-малка функционалност в сравнение с други OLAP приложения.

Мобилен OLAP се отнася до OLAP функционалност върху мобилни устройства.

Основните OLAP операции над данните са:

* Обединяване (roll-up)- увеличаване на нивото на обобщаване;
* Детайлизиране (drill-down)- намаляване на нивото на обобщаване или увеличаване на подробностите;
* Създаване на „резени“ и „зарове“ (slice and dice). Резенът е двумерна матрица, при която всички останали измерения имат само 1 стойност. (Избират се 2 измерения за създаването на резена, при фиксиране на определена стойност за всяко от останалите измерения). Зарът се изгражда от множество резени, при които се използват данни за 1 или няколко от измеренията, които са имали фиксирани стойности при създаването на резена.
* Филтриране (filtering)- извеждане на данни, които отговарят на определено условие.
* Пренасочване (pivot) на многомерния изглед на данните.

**3.2. Интелигентни технологии в БИС**

Интелигентните технологии са информационни технологии от областта на изкуствения интелект. В БИС те се прилагат в процеса на извличане на зависимости (модели) от данните – Data Mining, при прогнозиране и при логическа обработка на знания, предимно във форма на правила. Ще разгледаме общите характеристики на интелигентните технологии, тъй като имат принципни разлики от традиционните технологии за обработка на данни.

Основна характеристика на интелигентните технологии е, че обект на обработка (интерпретация) са знанията.

С конвенционалните информационни технологии се решават предимно добре структурирани задачи. При тях обработката на информация е по определени алгоритми и при точно задаване на данните. Задължителен е принципът на **детерминираност**, при който на всяка стъпка от алгоритъма е точно определена следващата. Примери за структурирани задачи са тези по оперативното управление на фирмата- разпределение на бюджета, обработка на клиентските изисквания, продажби, доставки и др.

Обаче, реално съществува голям клас от задачи, от неструктуриран тип, които според Нюел (Nеwell) имат някои от изброените особености:

- нямат алгоритмично решение;

- величините, с които се оперира не могат да се зададат в числова форма;

- величините са нееднозначни, непълни и противоречиви и динамично се променят;

- целите не могат да се изразят чрез точна целева функция.

За решението им е необходим човешки опит под форма на знания. Решението се приема като най- добрият възможен вариант при наличната информация. Поради невъзможността да се създаде аналитичен модел за решаване на задачата, последователността от действия за решаването й не е известна предварително и трябва да се генерира в процеса на работа. Примери за такива задачи са тези по стратегическото управление на фирмите, вземане на решение в условия на неопределеност и динамика на средата, разсъждения по аналогия и др.

Неструктурираните задачи се решават при различна степен на разбиране на проблема и с използване на методите и средствата на изкуствения интелект.

В зависимост от особеностите на задачата се прилага различен подход. Например наличната информация за задачата да се представи на понятийно (смислово) ниво в рамките на избран формализъм, който да позволява автоматизирани разсъждения чрез търсене на решения стъпка по стъпка. Този подход не изключва използване на формален аналитичен модел, стига да е възможно да бъде намерен.

В зависимост от подхода, на който са основани, интелигентните технологии са три общи вида- символни, числови и хибридни (комбинация от предходните два вида). Чрез символните технологии се моделира съзнателното поведение на хората. Използва се и когато знанията съществуват единствено в езикова форма. Знанията при символните интелигентни технологии са организирани в бази от знания. Представянето на знания в явна форма дава възможност за провеждане на разсъждения върху тях. В резултат на „разсъждения” от БЗ може да се генерира нова, неприсъстваща явно в нея информация. При символните технологии се прилагат нови подходи за намиране на решения като: логически извод (дедукция), търсене и др. Информационните технологии, в чиято основа са залегнали принципите на символния подход, са силно развити и практически широко използвани.

За възпроизвеждане на подсъзнателното поведение и еволюцията при човека се използват числовите интелигентни технологии. При този подход се използват идеи, взаимствани от:

* работата на централната нервна система на човека (невронните мрежи);
* еволюционни процеси (генетичните алгоритми).

## На невронните мрежи е посветена следваща тема.

## Интелигентните технологии са съвместими с конвенционалните. Всеки от двата вида използва различни подходи за обработка на отделни страни от информацията (знания и данни) и е предназначен за решаване на различен клас задачи- неструктурирани и структурирани. Интеграцията между двата вида технологии дава едни от най- ефективните приложения.

При БИС има възможност да се използват интелигентни и традиционни технологии за решаване на различни по вид задачи.

1. http://www.unwe.bg/uploads/ResearchPapers/Research%20Papers\_vol1\_2010\_No6\_V%20Kisimov,%20K%20Stefanova.pdf [↑](#footnote-ref-1)