Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženýrství

Bakalářská práce na téma

Datové sklady

Vedoucí práce:	doc. Ing. Vojtěch Merunka, Ph.D.
Autor bakalářské práce:	Karel Smíšovský

PROHLÁŠENÍ

Prohlašuji, že jsem bakalářskou práci na téma Datové sklady vypracoval samostatně a využil jsem zdrojů uvedených v přehledu použité literatury.

V Praze, dne .....

Podpis

Děkuji doc. Ing. Vojtěchovi Merunkovi, Ph.D. za vedení práce, vstřícnost a odborné rady poskytnuté při zpracování bakalářské práce.

Datové sklady

Datawarehouse

Souhrn

Pojem "datový sklad" je v dnešním světě plným informačních technologií fenoménem a velmi často používaným termínem. Stejným, jakým byl ještě před několika málo lety pojem "informační systém", v dnešní terminologii "primární" nebo "provozní" informační systém. Zdaleka ne každý si však dostatečně dobře uvědomuje, jaký je rozdíl mezi těmito dvěma typy systémů, jaké jsou jejich hlavní charakteristiky, výhody, nevýhody a účel.

Řešení datového skladu je v praxi značně odlišné od standardních provozních systémů, u nichž se vychází ze strategických záměrů a formulování požadavků na budoucí systém. V případě datového skladu je situace poněkud odlišná. Na počátku jsou data. Existují ve strukturách daných jejich využitím v provozních transakčních systémech. Je tedy nutné respektovat základní principy tvorby řešení, zejména pak rozdílné potřeby uživatelů, orientaci na rozdílná data a v neposlední řadě možnosti získání z dat informace.

Klíčová slova: datový sklad, databáze, data, informace

Summary

The term „Datawarehouse“ has become a phenomenon in today's world full of information technologies. Ten years ago was the same phenomenon term „information system“, nowadays called „operation“ or „primary“ system. But there are still lot of people who have not come to realize the difference between those two natively different systems and their characteristics, advantages, disadvantages and purpose.

Datawarehouse solutions vary in praxis from standard operational systems, where the purpose is company strategy and well defined requirements on building system. Datawarehouses are different. Data is first. They exist in structures native to operational systems and therefore we need to respect basic principles of solution building, especially different needs of users and stakeholders, their orientation on different data and last but not least a possibility of obtaining information from data.

Key words: datawarehouse, database, data, information

OBSAH

1 ÚVOD

Cílem všech podniků je generování zisku. Aby podniky zisku dosáhli, snaží se být maximálně efektivní, flexibilní a včasně reagovat na trh. V současném turbulentním prostředí potřebují mít dostatek informací pro rozhodování. A to nejen o konkurenci a trhu, ale zejména informace o činnosti a fungování vlastního podniku nebo dceřiných společností. Všechny významné společnosti v současnosti využívají bohatého SW vybavení pro správu a chod svých činností. Tyto systémy evidují značné množství dat o jednotlivých obchodních transakcích, ale data v nich uložená jsou použitelná pouze na úrovni jednotlivých systémů i přes veškerou snahu o systémovou integraci. Získáváním dat a jejich přetvářením na informace a případně znalosti se zabývá problematika datových skladů. V posledních letech tak podniky investují značné částky do budování a spravování těchto datových skladů.

Při rozhodování o nasazení datového skladu je nutné si uvědomit, že skutečný datový sklad nemá řešit požadavek oddělení IT na vlastnictví datového skladu (což se v podnicích často děje), ale řeší původní problém obchodních organizačních jednotek podniku, které také nesou náklady na budování a správu řešení. Tyto jednotky tak jsou přímými zákazníky IT spravujícím datový sklad na náklady obchodní organizační jednotky.

2 CÍL PRÁCE

Cílem předložené práce je charakteristika datových skladů a návrhu modelu datového skladu konkrétního podniku. Pro tento cíl je nutné zhodnocení možností využívání datových skladů v podnikové sféře. Je zde tak nastíněn jak teoretický, tak i praktický rámec návrhu datových skladů pro získávání dat z transakčních podnikových systémů.

V neposlední řadě je cílem této bakalářské práce hodnocení různých přístupů při výstavbě datových skladů a obecný postup při jejich implementaci.

3 METODIKA

Předkládaná práce je rozdělena do dvou částí a zaměření těchto částí odpovídá použitá metodika.

První část je věnována literární rešerši, ve které je charakteristika datových skladů a zhodnocení různých architektonických přístupů k tvorbě datových skladů. Obsahuje popis hlavních typů systémů, ze kterých se získávají data pro následné analýzy, a zaměřuje se na služby poskytované datovými sklady.

Druhá část práce vypracovává návrh datového skladu konkrétního podniku. Pro tuto část bylo zapotřebí vybrat vhodný podnik. V něm bylo třeba dále analyzovat strategii a obchodní cíle společnosti, strategii IT oddělení, provést analýzu architektury IS a využívání prostředků ICT.

Teoretický podklad pro tuto práci tvoří publikace Merunka V.: Objektové modelování, Alfa Publishing 2008, ISBN 978-80-87197-04-2; Vaníček J. a kolektiv: Teoretické základy informatiky, Alfa Publishing 2007, ISBN 80-903962-4-1; Vrana I., Richta K.: Zásady a postupy zavádění podnikových informačních systémů, Grada 2005, ISBN 8024711036; Business Intelligence v SQL Serveru 2005 (Lacko L.). Jako podpůrné publikace byly použity The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Kimball R.), The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouse (Kimball R.), Oracle® DBA Guide to Data Warehousing and Star Schemas (Scalzo B.).

Jako zdroj aktuálních informací byly použity internetové stránky firmy Oracle, Gartner, Ness Logos, internetových periodik zabývajících se problematikou IS/IT a dalších, které jsou uvedeny v seznamu literatury.

I. Literární rešerše

4 DATOVÉ SKLADY JAKO KOMODITA

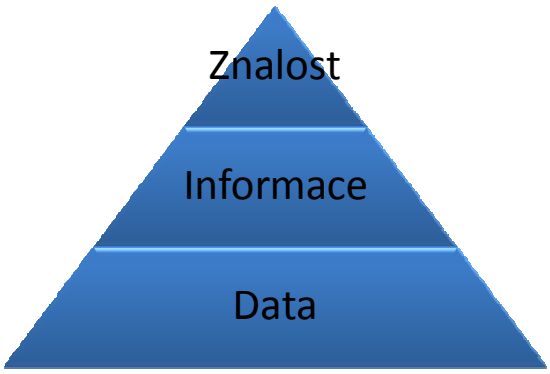
4.1 Data a informace obsažené v databázích

[4]Při intenzivním nasazování informačních technologií pro sběr a zpracování údajů do rozdílných odvětví lidské činnosti, hlavně do procesů odehrávajících se v podnikové oblasti, dochází ke shromažďování velkého množství různých údajů, hlavně z technologických, administrativních a obchodních procesů. Ke shromažďování údajů dochází i v obchodě. Data vznikají a mění se i v procesu provozu obchodu, například z elektronických pokladen a podobně. Výsledek je lehce předvídatelný. Za kratší, či delší dobu se podaří shromáždit obrovské množství dat. Moderní databázové servery umožňují nejen bezpečnou a rychlou práci s takovým množstvím údajů, ale umožňují nám z těchto dat získat i informace. Na první pohled by se mohlo zdát, že mezi pojmy data a informace můžeme položit znaménko rovnosti. Ale jen na první ohled.

Data se stávají informacemi, pokud:

- máme data,
- víme, že máme data,
- víme, kde tato data máme,
- máme k nim přístup,
- zdroji dat můžeme důvěřovat.

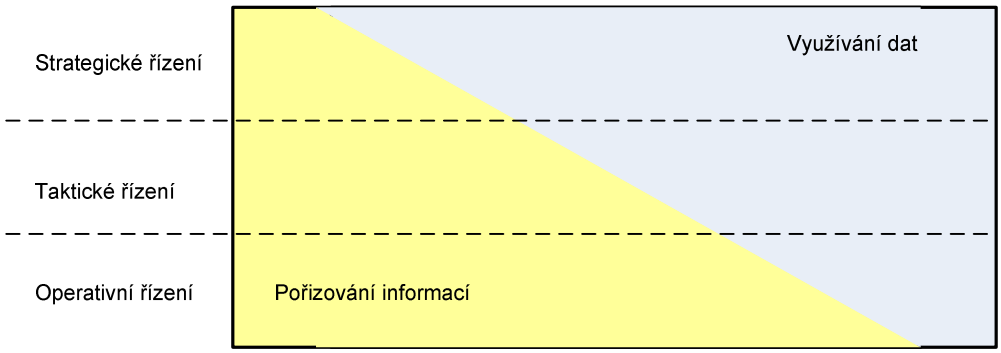
Přeměnu dat na informace, informací na znalosti a budování „moudrosti“ na základě znalostí můžeme zobrazit na hierarchické pyramidě informačních úrovní. Základem všeho jsou data. Data obsahují jen jednoduchá fakta, přičemž tušíme, že někde uvnitř množiny dat jsou ukryté uvnitř informace. Tyto informace ale odhalíme až tehdy, pokud přidáme k datům souvislosti. Když do hry vstoupí kromě informací i tvořivá inteligence, získáme znalosti. Pokud tyto znalosti zobecníme, získáme „moudrost“, to znamená schopnost přesného zhodnocení znalostí a jejich následné uplatnění v reálné praxi.



Obrázek 1 - Hierarchie znalostí, informací a dat

Často je potřeba sledovat trend nějaké veličiny, například při obchodování s cennými papíry, nebo najít mezi údaji určité závislosti. Proto moderní databázové servery obsahují rozsáhlou podporu pro budování datových skladů (**datawarehouse**), **OLAP** (Online Analytical Processing) analýzy a **data mining** (dolování, odkrývání dat).

Dle různých průzkumů (viz obrázek 2) 5 až 10 procent uživatelů používá výsledky analýz, 15 až 25 procent uživatelů tyto informace zkoumá a hledá v nich souvislosti. Největší skupina uživatelů informace používá ve formě různých výpisů a reportů.



Obrázek 2 - Podíl využití a pořizování dat

Na podnikové úrovni se generují různé druhy reportů například pro obchodní oddělení, finanční oddělení, oddělení lidských zdrojů, ve sféře CRM a podobně. Data jsou buď v podnikových databázích, nebo v datových skladech. Výhodou je, že data jsou už předzpracovaná a přetransformovaná v etapě ETL a přenesená z produkčních systémů do datových skladů (data warehouse), případně datových trhů (data mart). Reporty z reportovacích služeb potom vhodně doplňují data z analytických aplikací business intellingence. Nasazení

reportovacích služeb je v tomto případě převážně na úrovni podnikových portálů, takže koncoví uživatelé k nim přistupují v rámci podnikového intranetu.

4.2 Používané druhy databází

[4]Data jsou zpravidla ukládána do transakčních OLTP (On-line Transaction Processing) databází. Ty jsou určeny pro vykonávání velkého množství online transakcí, například bankovních, obchodních a podobně. Takové databáze jsou propojeny na IT systémy, jejichž cílem je automatizace každodenních činností, které jsou předmětem podnikání, například skladové hospodářství, mzdy, nákup a prodej, případně řízení a monitorování technologických procesů v reálném čase. Transakční systémy jsou v některých oblastech firemní informatiky téměř nezastupitelné a kromě výhod, které z nich vyplývají z jejich principu, jsou v podnikové praxi preferované z důvodu existence množství specialistů, ať už administrátorů, nebo vývojářů. Dochází samozřejmě k průnikům systémů. V případě, že transakční databázový systém s příslušnými aplikačními nadstavbami pokrývá většinu podnikových aktivit, nazýváme ho systémem ERP (Enterprise Resource Planning). Ke zdroji údajů tedy ve stejném čase přistupuje velké množství uživatelů, kteří data z databáze čtou, jiní do něj zapisují, případně někteří vykonávají i analýzy. Navzdory některým teoretickým předpokladům, které zdánlivě dokazují nevhodnost systému OLTP pro analýzy, pomocí výkonných technologií a vhodné architektury můžeme tyto analýzy efektivně vytvářet.

Rozhraní mezi relačními a analytickými systémy není ani ostré, ani jednoznačné. Od klasických transakčních systémů se plynule dostáváme k systémům pro podporu řízení a rozhodování nebo jinak řečeno z úrovně okamžitých transakcí se dostáváme na jakousi firemní „operačně taktickou úroveň“.

Systémy MIS (Management Information Systems), do kterých vstupují data z transakčních systémů, už poskytují řídícím pracovníkům různé komplexní přehledy a sestavy, agregované podle různých hledisek, například časových, geografický, organizačních a jiných. Do systémů MIS vstupují data transakčních systémů. Nevýhoduo je poměrně velká režie. Požadavky na sestavu se odeslaly vývojovému týmu MIS, který vytvořil sestavu a poskytnul ji manažerům až po určité době, zpravidla po několika dnech, týdnech nebo dokonce až po několika měsících.

Operativnější výsledky poskytují systémy DSS – (Decision–Support Systems), kde už v názvu je naznačeno jejich určení pro podporu rozhodování.

Na rozdíl od systémů typu MIS, které se nasazují na operativní taktické úrovni, systémy DSS jsou už na rozhraní taktického a strategického rozhodování. Poskytují řídícím pracovníkům například výsledky poměrně složitých analýz. Přes operačně taktickou úroveň jsme se dostali v našem pohledu až na strategickou úroveň k informačním systémům pro vrcholové řízení.

Ty jsou někdy označovány jako EIS – (Executive Information Systems), ale mnohem častěji se setkáme s termínem Business Intelligence. Tento pojem můžeme definovat jako proces transformace dat na informace a převod těchto informací na poznatky prostřednictvím objevování. Účelem Business Intelligence je tedy konverze velkých objemů dat na poznatky, které jsou potřebné pro koncové uživatele. Tyto poznatky můžeme potom efektivně využít například v procesu rozhodování. Nejčastějšími uživatelskými nástroji pro Business intelligence jsou od firmy Oracle a Microsoft[viz Příloha 1][9].

Pod pojmem informace nerozumíme jen konkrétní záznam, nebo množinu záznamů. Často potřebujeme sledovat trend nějaké veličiny, například při obchodování s cennými papíry, nebo potřebujeme najít mezi údaji určité závislosti.

Moderní databázové servery s podporou pro vytváření datových skladů obsahují rozsáhlou podporu pro OLAP (Online Analytical Processing), Data Mining (dolování, odkrývání dat), Data Warehouse (datové sklady) a Reportingu.

4.2.1 Kvalita údajů pro analýzy

Firmy využívají pro svou činnost různé druhy ekonomického softwaru, například účetnictví, skladové hospodářství, evidence pohybu zboží a podobně, přičemž samozřejmě shromažďují data. Zčásti jsou možná bezcenná, ale možná i velmi cenná, ale zůstávají nevyužita, protože jsou uložena ve formě, která je činí nedostupnými pro účely získávání informací. Existence dat totiž vůbec neznamená dostupnost informací. Představme si firmu třeba i jen střední velikosti, která prodává 1000 druhů výrobků, prostřednictvím 10 prodejních kanálů, 100 odběratelům. Lehko se dopočítáme miliónů možných kombinací uvedených položek. A kdybychom chtěli sledovat obchodní život firmy po měsících, dostáváme dvanáct miliónů možných kombinací. A pokud bychom chtěli sledovat více ukazatelů, například zisky, výsledky kampaní a podobně, museli bychom uvedených dvanáct miliónů ještě vynásobit počtem ukazatelů. Zdánlivě obrovské tabulky bez možnosti orientace. Vzpomeňme si ale, jak jsme se k tomuto vysokému číslu, v našem případě 12 miliónů, dopracovali. Násobením poměrně malých čísel, tedy dimenzí 10 x 100 x 1000 x 12. Pokud data uspořádáme do multidimenzionální struktury, budeme pracovat s mnohem

menšími rozměry dimenzí a data potom budou mít mnohem větší vypovídací schopnost, ale přes dimenze se dokážeme dostat k příslušnému faktu, který leží na průsečíku dimenzí. Snažíme se, aby multidimenzionální informace byla uložena tak, aby byla orientována na předmět podnikání, a ne vázaná na konkrétní systém pro sběr údajů, odkud předmětný údaj pochází.

4.2.2 Nevhodnost transakčních databází pro analýzy

Transakční databáze označované i jako OLTP databáze, tak jak je známe z běžné podnikové praxe, jsou určeny pro ukládání operačních údajů. Výsledkem dotazování jsou databázové tabulky, souhrny získané pomocí agregačních funkcí, různé sestavy a podobně. OLTP databáze jsou z důvodu jednoduchého dotazování a vyloučení redundance zpravidla normalizované, to znamená, že transakční databáze vyhovují pravidlům tzv. normálních forem. Struktura operačních údajů v OLTP databázích je ve většině případů v komplexních transakcích než při složitých analýzách, které jsou velmi náročné na výpočetní kapacitu procesorů. Komplexní analýza vyžaduje jiné techniky návrhu databází, například použití multidimenzionálních a hvězdicových schémat s tabulkami faktů, které obsahují měřitelné jednotky obchodování a vysoce denormalizované tabulky dimenzí.

Databázové systémy typu OLTP jsou optimalizované pro obchodní transakce, ale pro získání informací pro podporu rozhodování nejsou příliš vhodné.

4.2.3 Decentralizovanost systémů OLTP

Největší překážkou použití databázových systémů OLTP pro analýzy je skutečnost, že tyto systémy nemají k dispozici integrovaný zdroj údajů ze všech operačních systémů v rámci podniku tak, aby umožnily tvorbu komplexních analýz, to znamená, že potřebná data, nebo data, která by měla sloužit jako podklady pro analýzy, jsou roztroušena v různých zpravidla heterogenních OLTP systémech a musí se pokaždé pracně integrovat dříve, než je možné získat požadované informace. Časová náročnost případných analýz, a nemusí jít ani o příliš složité ani příliš komplexní analýzy, je proto poměrně vysoká. Někdy se dokonce ani nepodaří konsolidovat data mezi jednotlivými systémy, takže vlastně ani nemůžeme získat globální obraz o stavu podnikání.

Z technického hlediska nic nebrání dělat analýzy dat z transakčních databází a nové systémy se o to v mnohých případech i snaží, využívají hlavně paralelismus zpracování, modelování a podobně. Důležitým předpokladem je i vhodný návrh architektury. Analýza decentralizovaných heterogenních dat

z heterogenních zdrojů propojených ne příliš rychlých síťových spojením určitě nebude dosahovat výsledků v požadovaném čase. Pokud používáme pro on-line zpracování transakcí a analýzy pro podporu rozhodování stejné počítače, tento přístup degraduje výkon použitého hardwaru i operačního systému. Důsledkem je prodloužení času odezvy uživatelům a to jednak uživateli vykonávajícímu transakci a také i uživateli čekajícímu na výsledek analýzy. Problémy mohou být i s výkonem sítě. Výpočet potřebných agregací, souhrnů, predikcí a podobně v transakčních systémech trvá velmi dlouho a v mnoha případech neúměrně zatěžuje databázový stroj produkční databáze, takže dochází ke značné degradaci výkonu produkčního systému. Prodloužení doby odezvy transakčního systému je mnohem větší nevýhoda, než by se na první pohled mohlo zdát, protože transakční systém je primární a přímo životně důležitý. Pokud by tento systém selhal, nebo byl zatížen nad únosnou míru, potom by už nebyl žádný důvod pro analýzu údajů. Jednoduše by nebylo z čeho data analyzovat. Kromě uvedených problémů v obecné rovině vstupují více nebo méně do hry i jiné faktory:

- **transakční systém neuchovává historická data**

  Ne vždy systémy OLTP umožňují uchovávání dat po delší dobu, takže v mnoha případech chybí historická data potřebná na komplexní analýzu nebo predikci. Buď na to nestačí disková kapacita produkčních počítačů pro sběr dat, nebo nehomogenita historických dat může vzniknout i tím, že nekoncepčně navržený systém některá data neuchovává vůbec, například období platnosti valutových kurzů, ceníků a podobně.

- **nehomogenní struktura dat**

  Ta samá data mohou být v různých decentralizovaných systémech uložena v různých tvarech a formátech. Například záznam o telefonním čísle zákazníka může být v jednom systému uložen jako 15místné číslo, v jiném jako 18znakový řetězec, a to nehovoříme o předvolbách ať už místních nebo mezinárodních, případně s mezerami mezi skupinami číslic, pomlčkou za předvolbou a podobně. Podobné problémy bývají s rodným číslem, někde je uvedeno jako desetimístný řetězec, jinde je po prvních šesti číslicích lomítko a až za ním následují poslední čtyři číslice.

- **dlouhý čas přípravy dat**

  Ve většině případů data, která jsou potřebná jako vstupní pro různé analýzy, se nacházejí v různých, zpravidla nehomogenních zdrojích. Postupné připojení k nim není sice ve většině případů z technického hlediska velký problém, ale vyžaduje to nemálo času a námahy. Ovšem analyzovat taková data

z více transakčních systémů najednou je z technického hlediska velký problém. A to jsme stále jen u technického hlediska. Kolik špičkových analytiků a marketingových odborníků ovládá jazyk SQL? Narážíme na další bariéru dostupnosti dat, tentokrát způsobenou lidským faktorem. Analytici proto musí při vypracovávání analýzy spolupracovat s databázovými specialisty. A to trvá dlouho a také se významně podílí na zvýšení nákladů. I když se to podaří, po změně struktury některé z transakčních databází jsme tam, kde jsme byli na začátku a musíme definice příslušných dotazů přeprogramovat.

- **není jednoduché najít příčiny a vysvětlení problému a obtížně hledat závislosti jednotlivých veličin**

Ani po překonání všech naznačených problémů stále nemáme vyhráno. Je totiž problém najít konfliktní hodnoty sledovaných veličin, hlavně pokud závisí na více faktorech. U složitějších jevů není vždy jasné, na čem a v jaké míře je sledovaná veličina závislá. Platí totéž jako u Data Miningu – nezahrnutí důležité závislosti může často vést k podstatnému nebo úplnému zkreslení výsledků analýzy.

4.3 Datový sklad

Pokud se zamyslíme nad tím, co je to datový sklad, tak odpověď na tuto otázku je jednoduchá. Je to určitým způsobem strukturované úložiště dat. Pokud bychom ale hledali nějaké analogie mezi klasickým skladem a datovým skladem, tak to není tak jednoduché, jak to na první pohled vypadá. V klasickém skladu skladujeme buď materiály, součástky a polotovary, které vstupují do výrobního procesu, nebo naopak skladujeme výrobky, před tím, než se budou expedovat. Nikdo totiž nemá zájem skladovat dlouhou dobu polotovary a už vůbec ne hotové výrobky. Čím rychleji je dokážeme vyexpedovat a prodat, tím lépe pro ekonomiku firmy. V datovém skladu naproti tomu chceme shromaždovat a uchovávat informační bohatství firmy za co nejdelší období. Spíše než ke klasickým skladům můžeme datové sklady přirovnat k depozitářům muzeí. I v tomto případě se snažíme shromažďovat exponáty, třídit je jednak časově, geograficky, podle druhů a podobně.

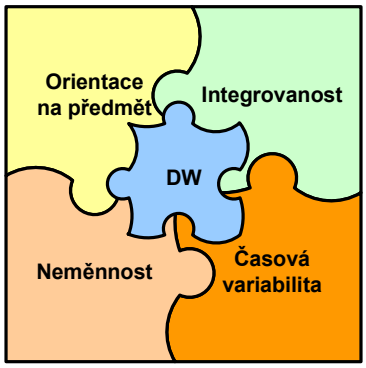
Data se získávají a ukládají do produkčních (operačních) databází, které mohou být v různých odděleních firem, nebo dokonce v rozličných geografických lokalitách. Tato data v pravidelných intervalech sesbíráme, předzpracujeme a zavedeme do datového skladu. Datový sklad je v podstatě také databáze, ale je organizovaná podle trochu jiných pravidel, tabulky například nemusí být normalizované a podobně. Datový sklad je tedy soubor technologií pro efektivní

skladování dat tak, aby tato data po jejich přeměně na informace sloužila pro podporu rozhodování.

4.3.1 Architektura dle Inmona

Nejznámější definice datového skladu pochází od Williama Inmona: „Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných , integrovaných, časově proměnných, historických dat použitých na získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data.“

Definice podle Williama Inmona je velmi stručná a výstižná. Pravděpodobně ale bude nutné tuto definici přečíst několikrát a zamyslet se nad jednotlivými pojmy, které definici tvoří.



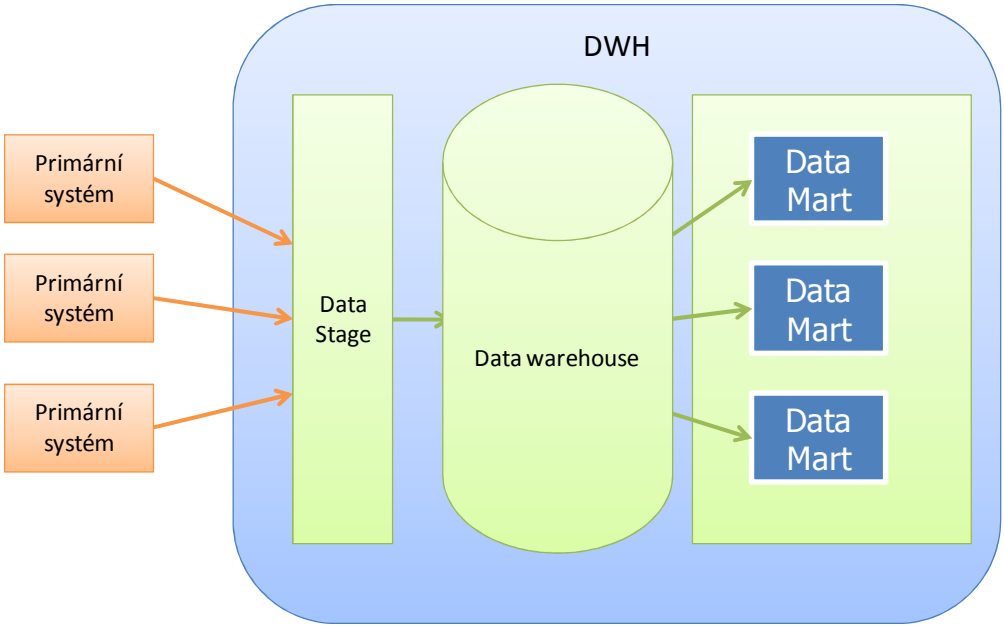
Obrázek 3 - Grafické vyjádření definice datového skladu

- **orientace na předmět (subjektová orientace):** Data se do datového skladu zapisují spíše podle předmětu zájmu, než podle aplikace, ve které byla vytvořena. Při orientaci na subjekt jsou data v datovém skladu kategorizována podle subjektu, kterým může být např. zákazník, dodavatel, zaměstnanec, výrobek a podobně. Orientace na aplikaci naproti tomu znamená, že data jsou v systému uložena podle jednotlivých aplikací, například data aplikace pro odbyt, data aplikace pro fakturaci, data aplikace pro personalistiku.
- **integrovanost:** Datový sklad musí být jednotný a integrovaný. To znamená, že data týkající se konkrétního předmětu se do datového skladu ukládají jen jednou. Proto musíme zavést jednotnou terminologii, jednotné a konzistentní jednotky veličin. Ale ne vždy je tato úloha jednoduchá, protože data přicházejí do datového skladu z nekonzistentního a neintegrovaného operačního prostředí. Proto musí být data v etapě přípravy a zavedení

upravena, vyčištěna a sjednocena. Pokud data nejsou konzistentní a důvěryhodná, tak datový sklad ztrácí význam.

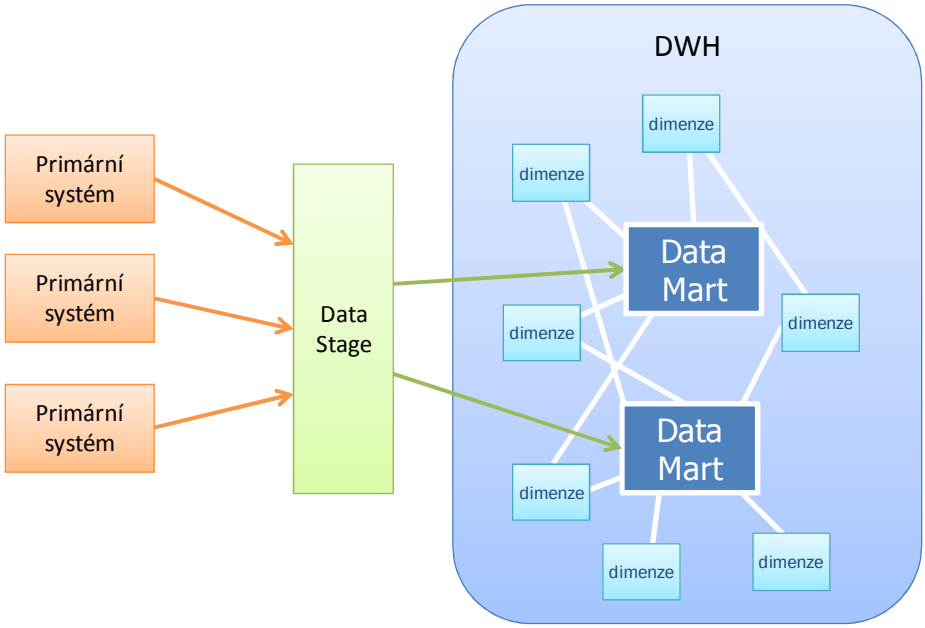
- **časová variabilita.** Data se ukládají do datového skladu jako série snímků, ze kterých každý reprezentuje určitý časový úsek. Na rozdíl od operačního prostředí, kde jsou data platná v okamžiku přístupu, v datových skladech jsou data platná pro určitý časový moment, časový snímek. Zatímco v operačním databázovém prostředí se data ukládají za kratší časové období dní, maximálně měsíců, v datovém skladu obsahují čas, který v operačních databázích nemusí být uveden. Jakmile je v datovém skladu zaznamenaný konkrétní snímek dat z operativní databáze, nemohou být už tato data v datovém skladu modifikována.
- **neměnnost:** V operačních transakčních databázích jsou data do databáze jednak vkládána, modifikována ale i mazána. Data v datovém skladu se obvykle nemění ani neodstraňují, jen se v pravidelných intervalech přidávají nová data. Proto je manipulace s daty daleko jednodušší v datových skladech. V zásadě můžeme povolit jen dva typy operací. Zavedení dat do datového skladu a přístup k těmto datům. Žádné změny dat nejsou povoleny. Z toho vyplývá, že většina metod pro optimalizaci a normalizaci dat a transakční přístup k datům je v datovém skladu nepotřebná.

Tato architektura minimalizuje redundantní data a zároveň minimalizuje počet interface mezi produkčními systémy a datovým skladem. Ne méně významným požadavkem je snadnost monitorování aktivit. Minimalizace redundancí přivedla Billa Inmona ke konceptu centrálního datového skladu. Co je centrální datový sklad? Pod pojmem centrální datový sklad (též celopodnikový datový sklad) rozumíme integrovanou, předmětově orientovanou, nepodléhajícím změnám, časově proměnnou kolekci detailních dat[16].



4.3.2 Architektura dle Kimballa

Tento přístup je od předchozího odlišný. Nebuduje DWH jako jeden ucelený a uzavřený blok, ale buduje ho jako sjednocení data martů. Logickým sjednocením těchto data martů pak rozumíme datový sklad.



Obrázek 4 - Architektura DWH dle Kimballa

Kimballova myšlenka budování datového skladu má velké výhody zejména v:

- Rychlosti implementace data martu, kdy se nezdržuje s důslednou datovou analýzou do celkového datového modelu datového skladu
- A tím i nižších počátečních investic.

Při budování datového skladu se setkáváme se skutečností, že jednotlivé data marty můžou být vždy budovány na základě požadavků jednotlivých oddělení společnosti. Každé z těchto oddělení má však malinko odlišné potřeby a požadavky a tak dochází k redundanci pojmů a dále i redundanci dat. Z toho vyplývá, že jednotlivé data marty obsahují i odlišné dimenze a odlišná fakt na stejnou reálnou skutečnost obsaženou v primárních systémech. Tímto způsobem vybudované prostředí pro podporu rozhodování však neposkytuje celopodnikový pohled na informace. Podíváme-li se na schematické znázornění architektury, odpovídající zmiňovanému přístupu, pak tato nám může připomínat „spaghetti architecture“.

Tento přístup tak ve větších datových skladech, a v prostředí, kde datový sklad využívá více oddělení, vede paradoxně přes nízké počáteční náklady k vysokým nákladům na udržení řešení v chodu a jeho využívání přes celý podnik.

4.3.3 Získávání dat

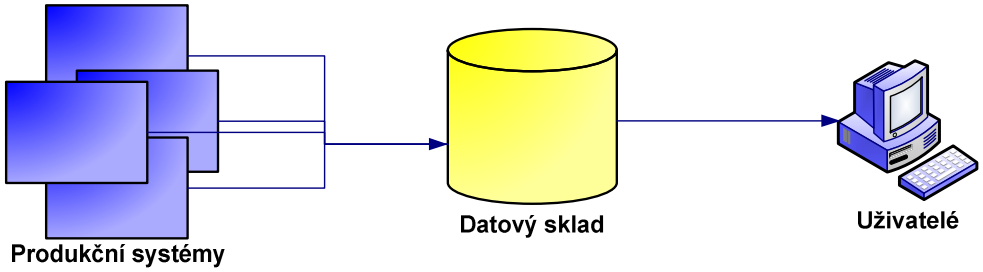
Data se získávají a ukládají do produkčních (operačních) databází, které mohou být v různých odděleních firem, nebo dokonce v rozdílných geografických lokalitách. Tato data v pravidelných intervalech sesbíráme, předzpracujeme a zavedeme do datového skladu. Datový sklad je v podstatě také databáze, jen je organizovaná podle uvedených pravidel. Rozdíly mezi produkční databází a datovým skladem přehledně zobrazuje následující tabulka.

Tab. 1 Rozdíly mezi produkční databází a datovým skladem

Vlastnost	**Produkční databáze**	**Datový sklad**
Čas odezvy	Zlomky sekund až sekundy	Sekundy až hodiny
Operace	DML (data manipulation language)	Primárně jen na čtení
Původ dat	30 – 60 dní	Série snímků za časový úsek
Organizace dat	Podle aplikace	Podle předmětu, času ...
Velikost	Malá až velká	Velká až velmi velká
Zdroje dat	Operační, interní	Operační, interní, externí

Činnosti	Procesy	Analýza

V datovém skladu můžeme vykonávat různé analýzy pro potřeby rozhodování manažerů, obchodníků a dalších uživatelů. Nástroje pro budování a provoz datových skladů ale představují poměrně velkou počáteční investici za hardware a software, takže datové sklady využívají většinou banky, pojišťovny, mobilní a telekomunikační operátoři, velké obchodní řetězce a podobně. K datům v datovém skladu a výsledkům analýz nad těmito daty mohou mít přístup prostřednictvím webu manažeři firmy po celém světě a obchodní partneři na základě přidělených přístupových práv. Informace jsou totiž zbožím. Tím, že je můžeme poskytovat i svým obchodním partnerům, se nám může postupně část vynaložených nákladů na datový sklad vrátit. Například dodavatelé velkého obchodního domu určitě ocení informace o tom, jaké zboží jde nejvíce na odbyt, případně o jaké zboží bude pravděpodobně zájem během nejbližších týdnů.



Obrázek 5 - Procesní schéma datového skladu

4.4 Datové trhy

Datové trhy (anglicky datamart) jsou určité přesně specifikované podmnožiny datového skladu, které jsou určeny pro menší organizační složky firmy.

Datový sklad je z hlediska investic i objemu prací velmi náročný projekt. Proto se v některých případech přistupuje k budování datového skladu po částech, což znamená, že pro některé důležité organizační složky se vytvoří jakési podmnožiny datového skladu – datové trhy. Kromě ekonomického efektu má tento postup i psychologický efekt, protože fungující podmnožina datového skladu prohlubuje důvěru v úspěšnost a potřebnost datového skladu jako celku.

Datové trhy mohou vzniknout i opačným postupem, tedy tak, že nejdříve se vytvoří centrální integrovaný datový sklad a z něj se potom vytvoří několik datových trhů. Takové řešení je flexibilnější a klade menší nároky na provoz a údržbu. Datové trhy tedy mohou existovat jako subsystémy datových skladů nebo i samostatně jako jednoduchý datový sklad.

4.5 Zpracování dat z produkčního prostředí

Data je samozřejmě možné „přetavit“ na informace a následně analyzovat i v operačním prostředí, kde tato data vznikají. Takový postup je však možný jen u málo vytížených transakčních systémů. Jinak dochází k neúměrnému snižování výkonu těchto systémů. Problém snižování výkonu je možné částečně vyřešit výběrem (extrakcí) a přenesením dat z jednoho prostředí do prostředí jiného. Problém se částečně vyřešil použitím technik zpracování extraktu, které vybírají data z jednoho prostředí a přenášejí je do jiného prostředí, kde se na jiném hardwaru zpracují.

Data pro extrakci se vybírají podle určitých kritérií a následně se umístí zpravidla do souborů nebo databází v jiném operačním prostředí. Tato vyextrahovaná data a výsledky analýz získaných nad těmito daty jsou potom k dispozici analytikům a pracovníkům, kteří řídí a rozhodují.

Proces extrakce byl logickým krokem od systémů OLTP k systémům na podporu rozhodování. Data se přesouvají z transakčních systémů do klientských systémů určených pro analýzu. Zdálo by se, že extrakce dat a zpracování takto získaných extraktů je ideální řešení, ale dochází k četným problémům. Jednak může docházet k mnohonásobnému větvení tím způsobem, že extrahovaná data se stanou zdrojem pro další extrakci. Extrakce dat může úplně zaměstnat kapacitu IT oddělení podniku, což je také nežádoucí. Dochází také k duplicitám, kdy se při extrakci a zpracovávání extraktů pokaždé přistupuje ke stejným datům. Také flexibilita extrakce je velmi omezená. Protože extrakty obsahují jen data, a ne metadata, tedy data o datech, nebo jiným způsobem řešeno data o tom, jakým způsobem byla data získána, je těžké přizpůsobit extrakci změnám v předmětu a způsobu podnikání. Po přečtení kapitol o datových skladech bychom našli i další nevýhody extrakce dat. Především chybí jednotná časová základna, jednotné algoritmy pro transformaci dat a výpočet požadovaných hodnot. Přístup k externím údajům pravděpodobně bude nekonzistentní a nebude správně definovaná ani granularita externích údajů. Sestavy vygenerované na základě extrahovaných dat tak ve většině případů obsahují spíše data než informace.

5 POSKYTOVANÉ SLUŽBY DATOVÝCH SKLADŮ

5.1 Integrační služby

S informačními systémy úzce souvisí nejen ukládání dat, ale i jejich sběr, přesuny, import a export. Tuto problematiku musíme řešet skoro u každého databázového systému. Fáze zavádění dat je také neodmyslitelnou součástí každého datového skladu.

Hlavní rozdíl mezi importem dat do informačních systémů a zaváděním dat do datového skladu spočívá v tom, že import je záležitostí jednorázovou, ale zavádění dat do datového skladu se odehrává periodicky v určitých, například 24 hodinových intervalech. I u těchto dvou scénářů je začátek velmi podobný, import dat do informačního systému a prvotní naplnění datového skladu. A aby to nebylo tak jednoduché, nemůžeme nevzpomenout archivní data obsahující historická data. Hlavní rozdíl mezi datovým skladem a archivem je v tom, že data v datovém skladu se pravidelně obnovují. Data z archivů jsou nezastupitelným zdrojem historických dat při prvotním naplnění datového skladu.

Import-Export

Import a export dat se ve většině případů jeví jako spojené nádoby a to, o jakou operaci v konkrétním případě jde, závisí na úhlu pohledu. Máme-li záměr importovat data do databáze v nějakém formátu, je důležité na druhé straně ve „zdrojovém“ systému zařídit jejich export. Pomocí této nepřímé metody, pokud využijeme například textový formát s daty oddělenými čárkou, tak dokážeme importovat a exportovat data prakticky mezi libovolnými heterogenními systémy.

5.1.1 Extrakce, transformace a přenos

Procesy integračních služeb mohou být navrženy pro jednorázovou akci, nebo pro akce periodicky se opakující. Jednorázovou akcí může být například migrace dat z jedné databázové platformy do jiné, nebo přenos dat ze souborů dokumentů kancelářských balíků do cílových databází. U periodicky se opakujících úloh, například při každodenním zavádění dat z produkčních databází do datových skladů je důležité, aby tyto operace proběhly v požadovaném čase.

Jednotlivé etapy procesu ETL jsou:

- **Extrakce** – výběr dat prostřednictvím různých metod
- **Transformace** – ověření, čistění, integrování a časové označení dat
- **Loading** – přemístění dat do datového skladu

Správně zvládnutá etapa ETL je nevyhnutelná v obou popisovaných scénářích, tedy pro rychlou a úspěšnou migraci dat v databázových projektech i pro nasazování a provozování projektů Business Intelligence a v datových skladech. V obou případech se proces ETL zapojí do určitého stavu informačního systému. K dispozici jsou různé sestavy, dokumenty a data z primárních transakčních systémů OLTP (On-line Transaction Processing).

Ani proces Business Intelligence zpravidla nezačíná jak se říká na „zelené louce“, tedy způsobem, že do této doby tu nebylo nic a naším cílem je vybudovat BI systém pro podporu rozhodování. Zpravidla i v těchto případech používá zákazník pro získávání informací pro podporu rozhodování nějaké provizorní řešení, například se generují výstupní sestavy z transakčních systémů a ty jsou potom buď ručně, nebo pomocí automatizačních prvků softwaru typu Office zpracovávány do manažerských pokladů poskytovaných manažerům pro podporu rozhodování.

Obecně tedy data, po kterých chceme, aby vstupovala do procesu business intelligence, pocházejí z různých nehomogenních zdrojů. Mohou to být data ze souborových databází (MS Access, dBase, ...), data z databází spravovaných některým databázovým serverem (Oracle, Informix, Microsoft SQL Server, Sybase, Interbase, Ingres ...), nebo data vyexportovaná nějakou databázovou platformou nebo informačním systémem, například pobočkou telefonní ústřednou do tzv. flat file, dokumentu XML a podobně.

Data z operačního prostředí je důležité před zavedením do datového skladu vyextrahovat, vyčistit, upravit a až následně ve vhodné formě do datového skladu zavést.

5.2 OLAP analýzy

5.2.1 Multidimenzionální databáze

Východiskem pro překonání dvou hlavních omezení relačních databází je zavedení organizace dat do multidimenzionálních struktur. Takto vytvořené databáze slouží jako podklad pro získání sumarizovaných a agregovaných dat, tedy vlastně informací. Jak uvidíme později, do multidimenzionálních databází

ukládáme upravená a „vyčištěná“ data. Na rozdíl od relačních databází používáme převážně nenormalizované tabulky, které můžeme rozdělit na dva druhy: na tabulky faktů a tabulky dimenzí.

Multidimenzionální databáze mají svoje výhody a nevýhody. K hlavním výhodám patří:

- rychlý komplexní přístup k velkému objemu dat
- přístup k multidimenzionálním a relačním datovým strukturám
- možnost komplexních analýz
- silné schopnosti pro modelování a prognózy

Nevýhodou jsou například vyšší nároky na kapacitu úložiště, problémy při změně dimenzí, bez přizpůsobení časové dimenze a podobně. Analytické databáze označujeme i pojmem OLAP (Online Analytical Processing) – tato zkratka zahrnuje struktury dat a analytické služby, které slouží pro analýzu velkého množství dat.

5.2.2 Multidimenzionální databázový model

Převážná většina dat je organizovaná v relační databázi v dvourozměrných relačních tabulkách. Každý řádek takové tabulky se vztahuje k nějakému předmětu, události nebo jejich částí. Výsledkem agregace a analýzy dat bývá obvykle multidimenzionální datová struktura – **kostka**. Zjednodušeně by se dalo říci, že kostka je v multidimenzionálním datovém modelu jakýmsi ekvivalentem tabulky v relační databázi. Typické využití systémů OLAP je pro analýzu velkého množství dat. Výsledkem analýzy jsou souhrny a reporty, které slouží manažerům jako podklady pro jejich rozhodnutí, ať už v oblasti řízení firmy, řízení ekonomických a technologických procesů a podobně. Pro výpočet OLAP kostek je nutné vykonat velké množství výpočtů a agregací, a to skoro v reálním čase. Každá OLAP kostka má několik dimenzí. Na rozdíl od geometrické kostky může mít multidimenzionální databázový model mnohem více dimenzí.

- čas
- region
- produkt

5.3 Dolování dat

Data mining je momentálně jednoznačně nejrychleji rostoucí segment Business Intelligence a v současnosti podobně jako OLAP analýzy mají tuto

technologii implementovanou všechny významné komerčně dodávané databázové servery. S trochou nadsázky by se dalo říci, že se jedná o odvětví někde na rozhraní vědy a magie. Termín data mining můžeme volně přeložit jako způsob získávání, dolování, odkrývání dat a informací pro podporu rozhodování z existujících datových zdrojů.

V úvodní fázi obvykle na základě určitých indicií a nekomplexních poznatků jen předpokládáme, že se v definovaném vzorku dat požadované informace nacházejí. Na základě statistického pohledu jsme vyslovili hypotézu. Hypotézu musíme následně na vybraném vzorku průzkumem ověřit a na základě výsledku průzkumu zamítnout nebo nezamítnout. Ze statistických definicí z oblasti testování totiž vyplývá, že hypotéza se testováním nedá potvrdit. Může se buď zamítnout, nebo nezamítnout.

Data mining je proces analýzy dat z různých perspektiv a jejich přeměna na užitečné informace. Z matematického a statistického hlediska jde o hledání korelací, tedy vzájemných vztahů nebo vzorů v datech. Pomáhá sledovat a analyzovat trendy a předvídat události. Může se využívat v bankovnictví při analýze a predikci úvěrového rizika, predikci rizika při vydávání kreditních karet, u operátorů telekomunikačních sítí, ve zdravotnictví pro analýzu laboratorních vzorků.

II. PROJEKT

6 NÁVRH DATOVÉHO SKLADU PRO KONKRÉTNÍ PODNIK

6.1 Základní charakteristika podniku

Z důvodu nutnosti utajení bylo použito fiktivní jméno společnosti a upraveny citlivé údaje.

Fincom je společnost poskytující komplexní finanční služby podnikatelským i nepodnikatelským subjektům za účelem financování všech druhů nových i ojetých dopravních prostředků, strojů, zařízení, investičních celků a výpočetní techniky. Společnost Fincom je co do počtu smluv menším hráčem na trhu, avšak kvalitou služeb patří mezi nejlepší společností v tomto segmentu v ČR.

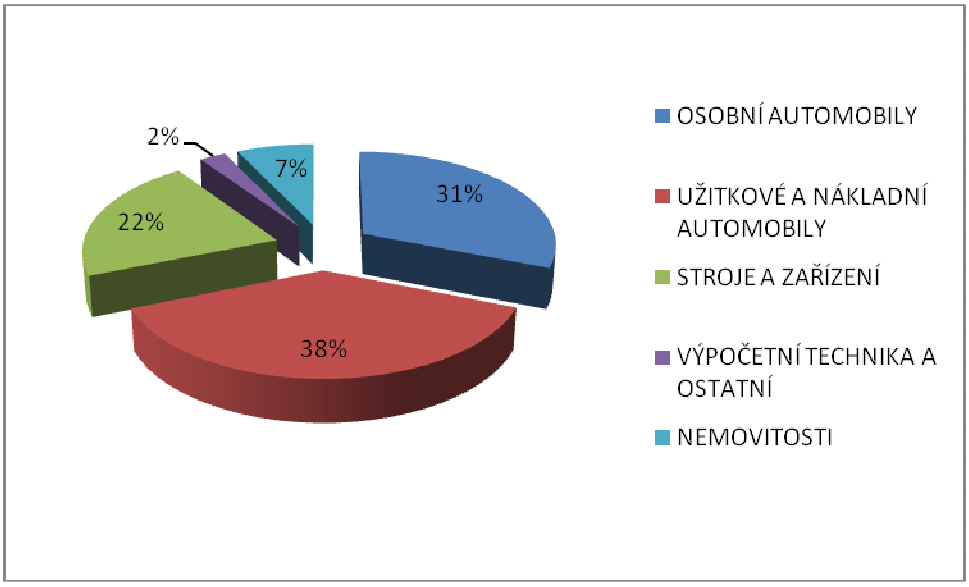
Společnost je většinovým dílem vlastněná zahraničním investorem.

6.1.1 Ekonomická struktura

ZÁKLADNÍ FINANČNÍ ÚDAJE	**v mil. Kč (pokud není uvedeno jinak)**
AKTIVA CELKEM	5185
ZISK PŘED ZDANĚNÍM	120
ZISK PO ZDANĚNÍ	88
NOVÉ OBCHODY V POŘIZOVACÍ CENĚ	4184
POČET NOVĚ UZAVŘENÝCH SMLUV	7271
POČET AKTIVNÍCH SMLUV	22426
PODÍL NA LEASINGOVÉM TRHU ČR V MOVITOSTECH (%)	3
PODÍL NA LEASINGOVÉM TRHU ČR V NEMOVITOSTECH (%)	2
PRŮMĚRNÝ POČET ZAMĚSTNANCŮ	85

6.1.2 Výrobní struktura

Společnost je zaměřena zejména na poskytování služeb pro financování automobilů (osobních a užitkových) a strojů.



Obrázek 6 - Struktura výrobního programu

6.1.3 Řízení informatiky

Velikostí, počtem organizačních jednotek a počtem zaměstnanců se jedná o menší až středně velkou firmu. Tomu odpovídá i řízení IS/IT. Oddělení IT má několik zaměstnanců a jsou řízení Manažerem IT, který není součástí top managementu společnosti. Vedení firmy si však uvědomuje strategický význam IT, proto se vedení IT často účastní strategických rozhodnutí nejvyššího vedení. Vytvořením tohoto štábního oddělení je snaha o jednotné řízení projektů a o větší kontrolu z pohledu celkové architektury systému.

6.1.4 Strategie IS

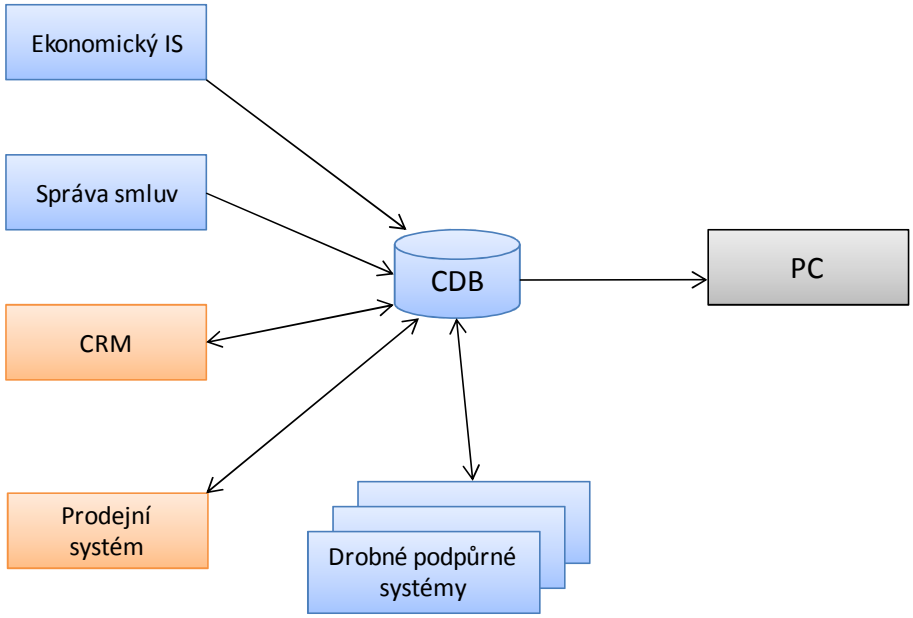
Z pohledu strategie informačních technologií na následující dva roky je hlavním cílem zavedení jednotného řešení pro reporting a tvorbu marketingových analýz. V současnosti využívá společnost v oblasti reportingu pouze OLTP operační databázi a dále sadu dynamických reportů vyvinuté interními kapacitami. Stávající řešení je shledáno jako nevyhovující a je plánováno pořízení odpovídajícího řešení pro podporu strategických cílů společnosti. Jedná se o vytvoření nové integrační platformy pro konsolidaci dat pořizovaných v primárních systémech.

Dalším významnou oblastí řešení je pořízení a integrace systému DMS (Dokument management systém – zpracování a digitalizace papírových dokumentů) a integrace s klíčovými systémy, kde tisky z tohoto systému jsou

přímo zařazovány do DMS. Tato část však není předmětem této práce, tudíž je zde zcela opomíjená.

6.1.5 Architektura IS

Architektura IS je dělí na Business critical systémy (označeny žlutou barvou), podpůrné systémy (modrá barva), osobní počítače (černou barvou) a ostatní prostředky ICT (nejsou znázorněny).



Obrázek 7 - Stávající architektura IS

Aplikace jsou buď koupené krabicové softwary, nebo řešení na zakázku.

Strategií IT v oblasti architektury je udržet všechny aplikace na jednotné platformě. Kvůli nutnosti rychle reagovat na změny a požadavkům agilního vývoje byla zvolena platforma Microsoft, která umožňuje rychlý vývoj a flexibilní architekturu aplikací postavené na SOA a snadnou integraci při zachování vysokých nároků na bezpečnost všech systémů. Podpůrné aplikace jsou vytvářeny v MS Access nebo MS .Net, což vede k jednotnosti systémů, rychlejšímu vývoji a snadné správě existujících systémů.

6.1.6 Investice do ICT a náklady na provoz ICT

Náklady na pořízení a správu IT (investice do úprav nebo pořízení IT/IS a náklady na provoz) ročně sahají až k částce 50 mil. Kč. Do těchto nákladů jsou započítávány i prostředky vynaložené na komunikační technologie (LAN, WAN, konektivita). Nejsou zde započítány provozní náklady na telefonní a datové spojení (pevné i mobilní).

6.1.7 Přínosy DWH

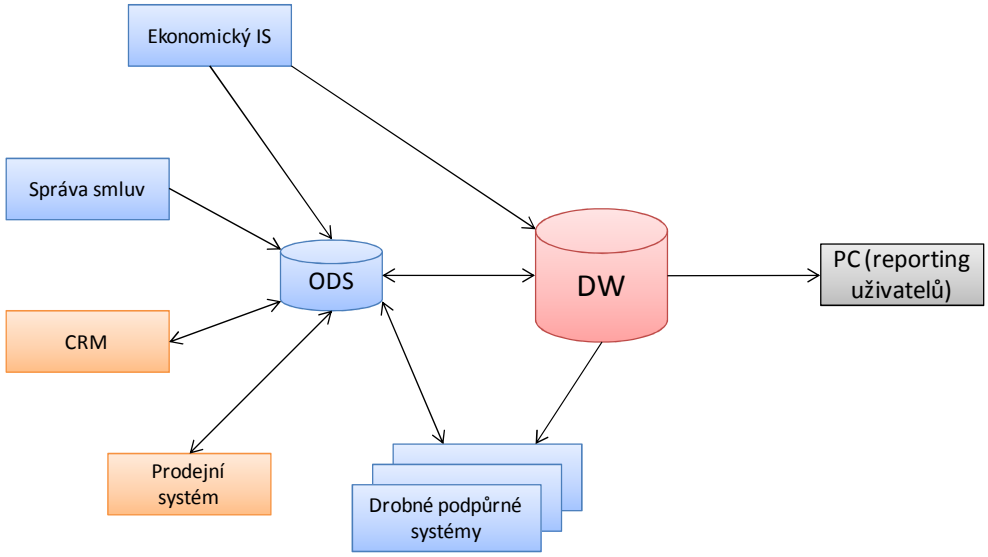
Dle strategie IT oddělení je plánováno pořízení nového DW řešení pro podporu reportingu. Datový sklad bude dodán externím dodavatelem a spravován a dále rozvíjen vlastními kapacitami. Vybraná technologie musí být v souladu s již zaběhlými standardy a používanými platformami v IT.

Očekávané přínosy nasazení takového řešení jsou:

- Konsolidace, čistota a jednotnost dat
- Snadná dostupnost historických dat
- Odstranění problémů s výkonem při ad-hoc reportech nad produkční OLTP databází (CDB)
- Přehledný a jasný reporting pro operativní řízení a manažerské rozhodování
- Využití business inteligence v oblasti marketingu a strategického řízení

6.2 Plánovaná architektura IS

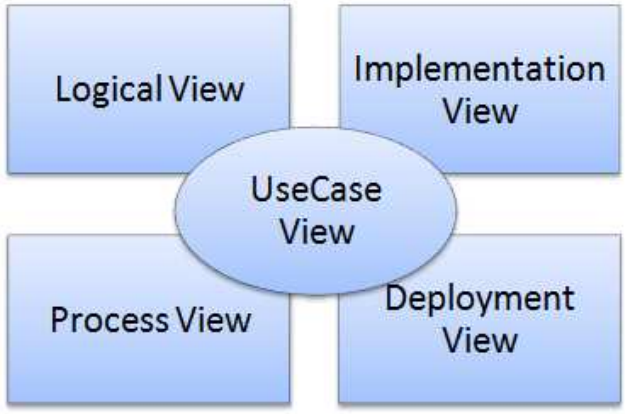
Implementováním datového skladu do podniku dojde k přesměrování uživatelů pro reporting do datového sklady z centrální databáze, čímž bude tato databáze ušetřena nevhodných dotazů v průběhu pracovního dne a obchod ohrožujícím výpadkům v business critical systémech závislých na této databázi. Datový sklad bude čerpat data z CDB v době, kdy nejsou obchodní hodiny (přes noc) a následně je bude zpracovávat do svých struktur a poskytovat okolním systémům. Změny v architektuře vidíme na následujícím schématu.



Obrázek 8 – Plánovaná architektura informačních systémů podniku

6.3 Model datového skladu

Při navrhování modelu datového skladu jsem vycházel z osvědčené praktické metody Modelu 4+1 pohledů na softwarovou architekturu [18], kde v každém z pohledů se popisuje budovaný systém z jiného úhlu.



Obrázek 9 - Model 4+1 pohledů dle [18]

6.3.1 Pohled užití

V tomto pohledu popisuji použití systému uživateli, tedy jak budou systém používat pro svoji práci. Jedná se o klíčový pohled, který vychází z požadavků na systém a který významně ovlivňuje všechny ostatní pohledy. Případy užití (UseCase) na této úrovni jsou identifikovány tři - Vytvoření reportu, Spuštění reportu a Stažení dat. Nicméně případy užití nejsou pro DW systém klíčové, klíčová jsou data a přeměna dat na informace, proto je můžeme pro další rozpracování nebrat v úvahu.

Požadavky na reporting lze shrnout do kategorií dle typů uživatelů ze dvou hledisek. Za prvé, jak k datům budou přistupovat, zde jsou uživatelé členěni dle oddělení společnosti na:

- Management
- Finance
- Risk
- Marketing
- Sales
- Ostatní

Za druhé, jak s daty budou pracovat a jakých cílů či oblastí zájmu se data týkají. Zde jsou identifikovány následující oblasti:

Operativní reporting

Tito uživatelé jsou nejpočetnější skupinou uživatelů. Spouštějí a prohlížejí si již předdefinované reporty v systému dle svých požadavků. Těmto uživatelům je v systému k dispozici standardní nástroj pro zobrazení reportů a případné extrahování dat z reportů do kancelářského SW vybavení.

Manažerský reporting

Jedná se o uživatele, kteří ke svému operativnímu i strategickému rozhodování potřebují jednoduchou formou znázornit důležité podnikové informace. Na formu je zde kladen velký důraz, proto je těmto uživatelům vytvářen tzv. Manažerský dashboard, viz ukázka níže.

Obrázek 10 - Ukázka manažerského dashboardu v systému Oracle [17]

Analýzy uživatelů

Jedná se o skupinu uživatelů, kteří mají přístup přímo k vytvořeným data marts či přímo do datového skladu. Znají dotazovací jazyky a vytvářejí adhoc reporty. Typicky se jedná o pracovníky IT či marketingové nebo riskové specialisty.

6.3.2 Pohled nasazení

Tento pohled se zaměřuje na to, jak bude systém zakomponován do svého prostředí a jak bude rozčleněn.

Abychom mohli správně určit požadované prostředky pro běh každého SW serverového řešení, je klíčové určit zejména:

- Množství dat, které bude uchovávat a zpracovávat,
- Počty konkurentních uživatelů pracujících se systémem.

Počty konkurentních uživatelů jsou v uvažovaném případě jednotky, maximálně desítky, proto je tento údaj pro další úvahy zanedbatelný.

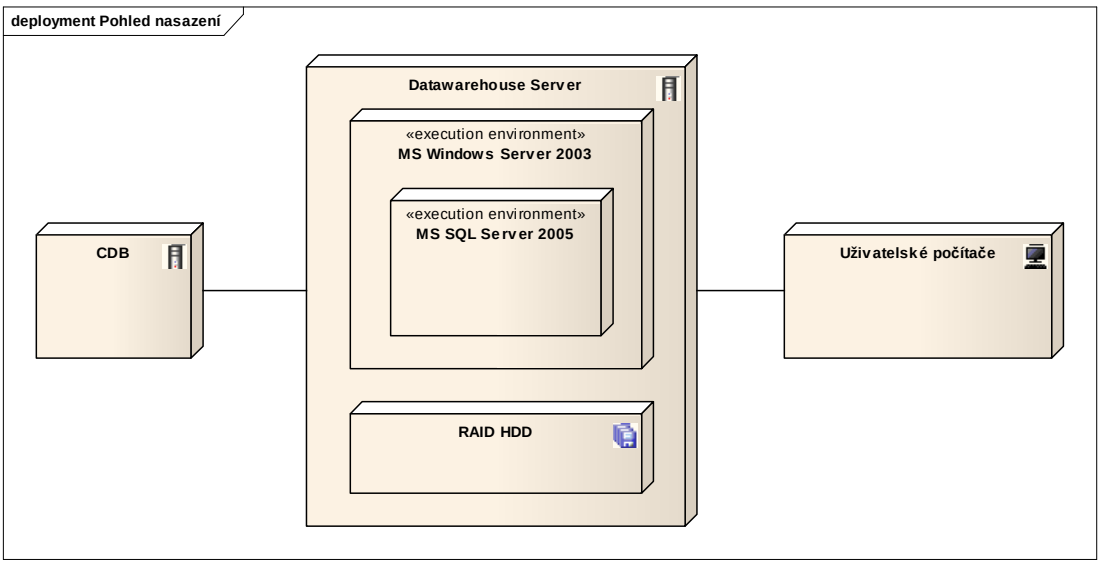
Objem dat a nárůsty

Pro výpočet množství dat vycházím z navržených logických entit (viz Logický pohled) a určuji odhad počtu záznamů v každé entitě, její velikost (odhad velikost surových dat) a velikost alokovaného místa na disku (násobeno konstantou 6). Pro DW je důležité držet i historická data a tak je nutno počítat s nárůstem množství dat, jako dobu odhadu jsem zvolil 5 let. Za tuto dobu již bude výhodnější pořídit silnější HW a systém zmigrovat než nyní pořizovat předimenzovaný a velmi nákladný HW. Roční přírůstek dat je počítán z plánované strategie společnosti. Odhad není 100%ně přesný, slouží ale k nástinu množství zpracovávaných dat. Tvorbou datových kostek a datamartů se bude alokované místo zvětšovat, neočekávám však nárůst větší než na 100 GB dat.

Logická entita	Roční přírůstek	Velikost záznamu (kB)	Velikost místa na disku (kB)	Počet záznamů (rok 2009)	Počet záznamů (rok 2014)	Místo na disku (MB)
klient	13%	0,5	3	18 300	33 717	101
smlouva	11%	0,5	3	22 426	37 789	113
splátka	11%	0,5	3	269 112	453 469	1360
prodejce	3%	0,5	3	206	239	1
pobočka	1%	0,5	3	20	21	0,5
region	0%	0,1	0,6	10	10	0,1
segment	0%	0,1	0,6	5	5	0,1
produkt	0%	0,3	1,8	15	15	0,3
čas	100%	0,1	0,6	730	23 360	14
Velikost CELKEM						1 590

Popis HW řešení

Vzhledem k očekávané velikosti datového skladu a počtu uživatelů čerpající data je navržen jednoprocesorový server Intel s OS MS Windows Server 2003. Dle [18] sice není shledána jako optimální, ale pokud přihlédneme k nárůstu výkonů HW za stejnou cenu od doby vydání knihy ke dnešnímu dni a plánovanému využití, je výkon dostačující, platforma odpovídá firemní IT strategii a je vyhodnocena jako nejlepší poměr cena/výkon.



Obrázek 11 - Diagram nasazení DWH řešení

Data budou do DW čerpána z CDB (centrální databáze). Uživatelé budou přistupovat ze svých PC k datovému skladu a to k portálu s vystavenými reporty a manažerskými dashboardy nebo přímo do databáze, dle způsobu užití (viz Pohled užití).

6.3.3 Implementační pohled

V implementačním pohledu se zaměřím zejména na rozvrstvení jednotlivých částí datového skladu a vnitřní architekturu řešení.

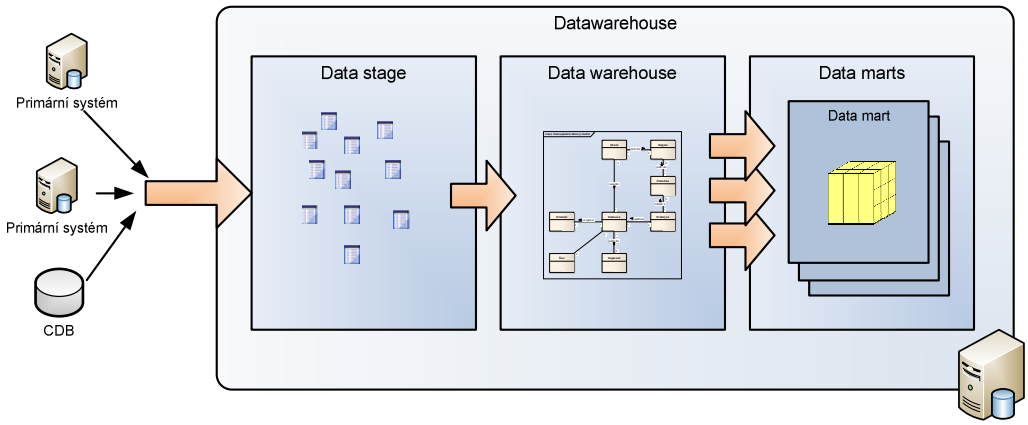
Z porovnání výhod a nevýhod architektonických přístupů uvedených v kapitole Datový sklad byla zvolena architektura dle Inmona. Důvody jsou následující:

- Strategií podniku je plánovaný růst a rozšíření podnikatelských aktivit
- Strategií IT je mít DWH jako cílovou platformu pro reporting a datové analýzy, tedy je plánován jeho další rozvoj pro celopodnikovou sféru
- Vyšší počáteční investice tak bude vykoupena nižšími náklady na další vývoj a udržení konzistence dat v datovém skladu.

V datovém skladu tak budou vybudovány tři samostatné logické celky:

- **Datastage**, do které budou importovány všechna data z primárních systémů a z CDB

- **Datový sklad**, jako zdroj jediné pravdy v podniku. Bude obsahovat očištěná a konzistentní data včetně historie
- **Datové marty**, které budou načítat data z datového skladu anebo využívat pohledů do DWH a budou sloužit pro jednotlivá oddělení/uživatele jako zdroj dat.



Obrázek 12 - Architektura datového skladu

6.3.4 Logický pohled

V této kapitole je popsán interní datový model datového skladu. Model je vytvořen jako konceptuální datový model [viz 11].

V rámci Konceptuálního datového modelu jsou klíčové pro tuto oblast (finanční produkty) zejména informace o smlouvách a splátkách, klientech a jejich segmentech a prodejích.

Entity jsou do tabulky faktů a tabulek dimenzí rozděleny následovně:

6.3.4.1 Entita faktů

Splátka

Tato entita je entitou faktů. Tzn. je to klíčová entita, ve které jsou všechny atributy, u kterých požadujeme výsledky. Splátky jsou standardně měsíční, ale pro potřeby fraud&risk managementu a marketingu potřebujeme členění na denní bázi.

6.3.4.2 Entity dimenzí

Smlouva

Jedná se o entitu obsahující informace o smlouvě, typicky kdy byla uzavřena a podmínky uzavřené smlouvy.

Prodejce, Pobočka, Region

Tyto entity jsou třeba pro získávání informací o tom, kdo a kde (ve které geografické oblasti) uzavřel danou smlouvu. Tyto dimenze jsou potřebné pro geografické analýzy prodejů a výnosů, výpočty výkonnosti regionů a prodejců. Tyto entity by bylo možné sloučit do jedné denormalizované entity, což by mohlo vést k vyšší výkonnosti při zpracování dotazů. Pro lepší práci s regionálními daty a častým změnám jsou navrženy takto separátně.

Produkt

Dimenze obsahuje informace o typu produktu, tedy se jedná o číselník produktů.

Čas

Dimenze obsahuje časové údaje. Nejmenší sledovaná časová jednotka je jeden kalendářní den.

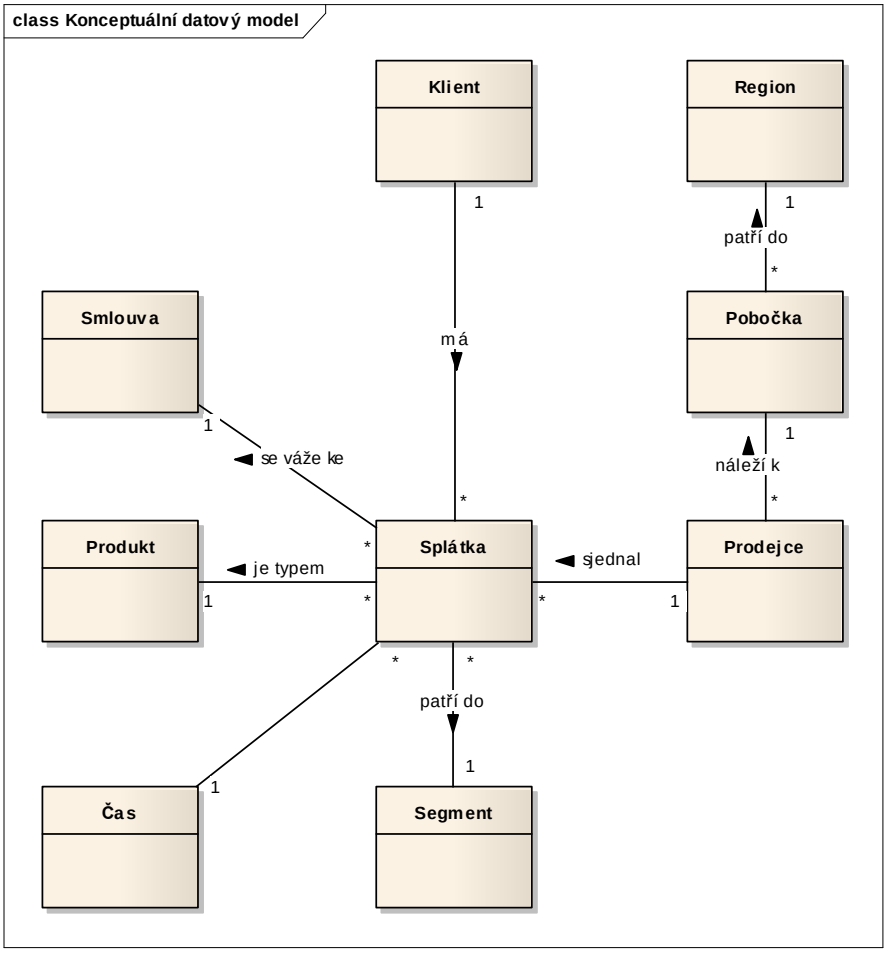
Segment

Dimenze segment obsahuje informace o segmentu klienta, typicky jeho rozčlenění na retail, SME, Enterprise. V rámci těchto segmentů je dále členěn na Top, Affluent a Standard.

Klient

V této dimenzi jsou zachyceny informace o klientech.

Grafické znázornění těchto entit je na následujícím diagramu.



Obrázek 13 - Konceptuální datový model, oblast klientských smluv a splátek

7 ZÁVĚR

Datové sklady jsou nedílnou součástí architektury moderních podnikových informačních systémů. Zavádění a přínosný provoz datového skladu patří mezi náročné oblasti podnikových informačních technologií. Všechny významné společnosti neustále budují a zdokonalují své datové sklady o nové informace a nové analýzy. Hledají nové souvislosti v datech z transakčních systémů, a to jim umožňuje hledat nové prostory na trhu, přesně hromadně oslovovat určité skupiny klientů s nabídkou dalších služeb a činit efektivní rozhodnutí.

Jejich význam však není vhodné ani přeceňovat ani podceňovat. Obdobně jako v případě klasických systémů je i v případě datových skladů nezbytně nutná odpovídající potřeba zadavatelů a uživatelů v podniku pro získání podpory ze strany zodpovědného managementu. Velmi nebezpečná jsou přehnaná očekávání, nerealistické představy nebo poptávání datového skladu jiným útvarem než uživateli (IT, neznalí uživatelé s velkými pravomocemi, ...). Datové sklady jsou prostředkem umožňujícím zefektivnění a zdokonalení rozhodovacích procesů, nikoliv cílem. Nedostatečně kvalitní či dokonce chybná data mohou způsobit přijetí chybných závěrů. Proto je na kvalitu vytvářených řešení v oblasti datových skladů dbán velký důraz.

Podmínek pro úspěšné nasazení a provoz datového skladu je pochopitelně celá řada. Důležité je si uvědomit, že nasazením starost o datový sklad zdaleka nekončí, následná podpora a rozvoj je v případě technologie datových skladů podstatně důležitější, než například u běžných provozních systémů.

V této práci se podařilo rozebrat pojem datový sklad, vydefinovat používané architektonické přístupy a definovat nejčastěji používané služby datových skladů.

Díky těmto znalostem se podařilo navrhnout model datového skladu pro konkrétní podnik, který naplňuje očekávání uživatelů, zapadá do architektury informačních systémů podniku a podporuje strategii společnosti pro další rozvoj. Požadavky zadavatelů (uživatelů) jsou tedy tímto návrhem naplněny.

8 LITERATURA

[1] Merunka V.: Objektové modelování, Alfa Publishing 2008, ISBN 978-80-87197-04-2

[2] Vaníček J. a kolektiv: Teoretické základy informatiky, Alfa Publishing 2007, ISBN 80-903962-4-1

[3] Vrana I., Richta K.: Zásady a postupy zavádění podnikových informačních systémů, Grada 2005, ISBN 8024711036

[4] LACKO, L.: Business Intelligence v SQL Serveru 2005. 1. vyd. Brno, Computer Press, a.s. 2006, ISBN 80-251-1110-5

[5] KIMBALL, R.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2. vyd., Wiley & Sons 2002, ISBN 0471200247

[6] KIMBALL, R.: The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouse, Wiley & Sons 1998, ISBN 0471255475

[7] Hľadanie nových foriem reprezentácie [online]. [cit 12.2.2009]
dostupné z:
http://bidwcz.blogspot.com/search/label/BI%20Trendy

[8] Oracle Positioned in Leader's Quadrant in Latest Business Intelligence Platform Magic Quadrant [online]. [cit 20.6.2007]
dostupné z:
http://www.oracle.com/corporate/press/2007_jan/012907-BIMQ.html

[9] Magic Quadrant for Business Intelligence Platforms, 1Q07 [cit 12.2.2009]
dostupné z:
http://mediaproducts.gartner.com/reprints/hyperion/145507.html

[10] SCALZO, B.: Oracle® DBA Guide to Data Warehousing and Star Schemas, Prentice Hall PTR 2003, ISBN 0-13-032584-8

[11] Webová prezentace společnosti Ness Logos - DWH/BI [cit 15.12.2008]
dostupné z:
http://www.logos.cz/reseni/business-inteligence/

[12] Otevřená encyklopedie Wikipedia – Datové sklady [cit 24.4.2009]
dostupné z:
http://cs.wikipedia.org/wiki/Datov%C3%BD_sklad

[13] Datové sklady [cit 24.6.2007]
dostupné z:
http://www.dbsvet.cz/view.php?cisloclanku=2002051501

[14] Many Companies Plan To Increase BI Spending [cit 15.2.2009]
dostupné z:
http://www.informationweek.com/showArticle.jhtml;jsessionid=SKLTDDFRY
MPPEQSNDLQSKIKCJUNN2JVN?articleID=198001258&queryText=preferred
+vendor+bi

[15] A UML Profile for Data Modeling [cit 27.4.2009]
dostupné z:
http://www.agiledata.org/essays/umlDataModelingProfile.html

[16] Dva způsoby budování datového skladu [cit 27.4.2009]
dostupné z:

http://www.systemonline.cz/clanky/dva-zpusoby-budovani-datoveho-skladu.htm

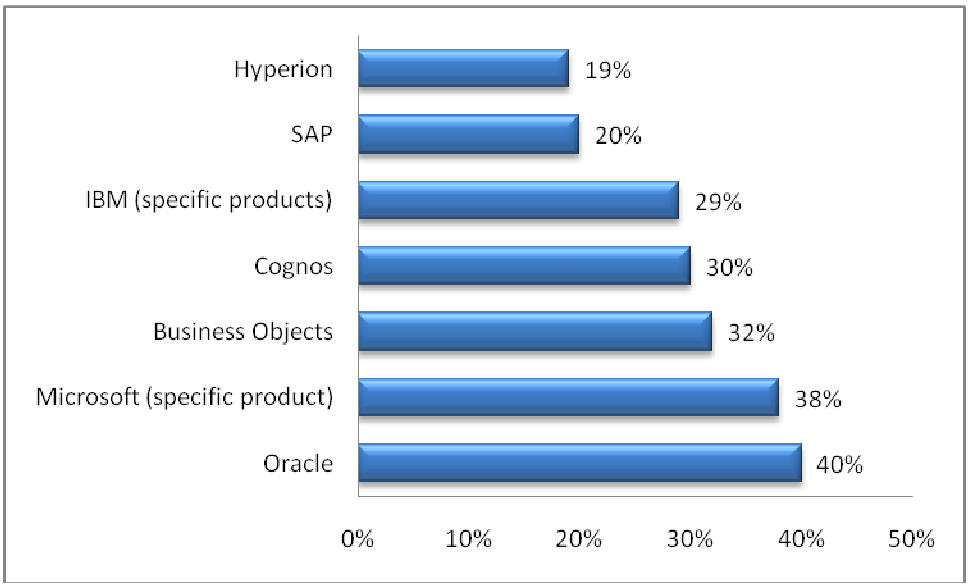
[17] Oracle Czech BI/DW Blog [27.4.2009]
dostupné z:
http://bidwcz.blogspot.com/search/label/Sout%C4%9B%C5%BE%20s%20BI%2FDW

[18] KRUCHTEN, P.: Rational Unified Process, The: An Introduction, Third Edition, Addison Wesley 2003, ISBN 0-321-19770-4

III. Přílohy

9 PŘÍLOHA 1

Internetový časopis InformationWeek uveřejnil v březnu 2007 průzkum o používání BI nástrojů. Průzkumu se zúčastnilo 500 BI profesionálů. Z výsledků vyplývá, že téměř 40% procent zúčastněných používá nebo plánuje nákup BI řešení od společnosti Oracle nebo Microsoft jako od preferovaného poskytovatele. Respondenti měli možnost vícenásobné odpovědi. Výsledky průzkumu jsou uvedeny v následujícím grafu představujícím preferovanou platformu řešení.



Zdroj: Many Companies Plan To Increase BI Spending http://www.informationweek.com/showArticle.jhtml;jsessionid=SKLTDDFRYMPPEQSNDLQSKIKCJUNN2JVN?articleID=198001258&queryText=preferred+vendor+bi