Pavel Dytrych - 527565

ISK12 Organizace znalostí

Datové sklady – historie, rozdělení, funkce a využití

Autor: Pavel Dytrych, UČO: 527565

Ročník: 2.

Typ studia: kombinované

Počet znaků: 13800

Obsah

Pavel Dytrych - 527565

Úvod

Dnešní společnost vygeneruje každým dnem ohromné množství dat a informací a toto množství stále roste. Zatímco ve starém Říme dosahovaly, při přepočtu na dnešní jednotky, největší sbírky svitků velikosti okolo 100Mb (Smil, 2021, p. 157), v roce 2016 již lidstvo dokázalo vygenerovat 16Zb dat každý rok (Smil, 2021, p. 160), což představuje nárůst o celých 14 řádů.

Spolu s velkým množstvím nově vygenerovaných informací se také ale mění jejich forma, zatímco ve zmíněných římských sbírkách bychom nalezli zejména písemnosti, v současné době představují písemnosti spíše marginální část uložených informací. Například v repositářích Kongresové knihovny ve Spojených státech amerických, zabírají písemnosti pouze 1 % ze všech uchovaných informačních artefaktů. (Smil, 2021, p. 158)

Vzhledem k takto velkému množství dostupných informací, je jejich zpracovaní velmi komplikované, a to i na úrovni menších celků, jako jsou například společnosti, univerzity anebo jim podobné organizace. Abychom mohli základě informací dělat jakékoliv rozhodnutí nebo analýzu, je nejprve nutné, je ideálně konsolidovat do jednoho uložiště, provést základní očištění a tím data standardizovat.

Jednou z možností, jak tohoto docílit, je vytvoření datového skladu. Pojem *datový sklad* může mít vícero definic a je možné na něj pohlížet i pouze jako na proces, který má zjednodušit přístup k datům, technické řešení je pak pouze podpůrný prvek celého procesu. Nejčastější pohled na datový sklad je však jako na systém, který může být definován jako:

> Systém, který extrahuje, čistí, upravuje a dodává zdrojová data do dimenzionálního datového uložiště a poté podporuje a zavádí dotazování a analýzu dat za účelem rozhodování. (Kimball & Caserta, 2004, p. 23)

Z této definice jasně vyplývá, že primárním účelem datového skladu je centralizace a standardizace dat, díky, které je pak přístup k datům velmi usnadněn. Zároveň je zde důležité uvést, že každý datový sklad se skládá z několika komponent, z nichž nejdůležitější je proces extrakce, čištění a úprava dat. Tento proces se ve spojitosti s datovými sklady nazývá ETL proces (z angl. Extract, Transform and Load) a pojem datový sklad tak nelze zaměňovat s pouhým datový uložištěm. (Kimball & Caserta, 2004, p. 24)

Využití datových skladů je pak velmi rozmanité, jednou z nejčastějších aplikací je ale jejich využití ve spojení se systémy podpory rozhodování (zkr. DSS z angl. Decision support systém), kde často slouží jako primární znalostní báze. (Inmon, 2005, p. 2)

Pavel Dytrych - 527565

Historický vývoj datových skladů

Datové sklady se spolu se systémy podpory rozhodování začali vyvíjet v 60. letech, kdy z důvodu ukládání dat na magnetické pásky bylo problematické data zpětně procházet a analyzovat, data se tak postupně začala pročišťovat a ukládat v již zpracované formě, což by se dalo označit za prvopočátek datových skladů. (Inmon, 2005, p. 2)

V polovině 60. let, se postupně začala rozšiřovat technologie diskových uložišť, která na rozdíl od magnetický pásků umožňovala přímý přístup k datům, bez nutnosti jejich kompletního nahrání do pamětí počítače. Díky této technologii začali vznikat první systémy, které lze označit za DBMS (z angl. Database management system), které umožňovali lepší správu dat. (Foote, 2018) Vzhledem k primitivnímu systému uložení a strukturalizace dat však tyto systémy neposkytovali data v takové kvalitě jako je známe dnes.

Velkým milníkem pak byly relační databáze, které začali vznikat v osmdesátých letech a které pomocí dotazovacích jazyků, jako je například SQL (z angl. Structured Query Language) umožňovali velmi snadnou práci s daty, v porovnání s předchozím stavem. (Foote, 2018)

Spolu s nástupem osobních počítačů a rostoucí velikostí firemních sítí, stoupala i potřeba skutečného datového skladu. Vzájemně propojené osobní počítače měli oproti architektuře centrálních počítačů velkou nevýhodu v roztříštěnosti dat, kdy data byla uložena na velké množství míst a bylo velmi složité udržet data aktuální a konzistentní, souhrnně se tento problém nazývá *spider web* (Inmon, 2005, p. 6). Dalším problémem bylo zvětšující se množství formátů, ve kterých byla data uložena, jelikož s osobními počítači se začali objevovat specializované formáty a aplikace, jako je například MS Excel, MS Access atp. (Foote, 2018)

V devadesátých letech, tak všechny tyto pohnutky vyvrcholili v potřebu vzniku jednotného uložiště, které by umožňovalo data snadno udržet v aktualizované a konzistentní podobě. Za tímto účelem vzniklo několik konceptů, z nich nejrozšířenější jsou již zmíněný datový sklad, data lake a data cube.

> Koncept *data lake*, se od datového sklad odlišuje převážně tím, že data nepřevádí do jednotné struktury, ale udržuje je v jejich původní formě. Tzn jedná se primárně o jednotné uložiště pro všechny možné formáty souborů a dat a jejich detailní anotaci, pro usnadnění jejich zpracování. (Foote, 2018)

> Více odlišný je pak koncept *data cube*, který ukládá data ve třech a více rozměrných maticích, kde každá matice představuje jednu datovou dimenzi. Když data v dimenzích následně relačně propojíme, je možné snadno filtrovat data dle jednotlivých dimenzí, bez nutnosti propojovat tabulky mezi sebou. (Foote, 2018)

V této práci se dále již zabývám pouze konceptem datového skladu tak, jak byl definován v úvodní kapitole.

Pavel Dytrych - 527565

Typy datových skladů

Datové sklady se dají rozdělit na několik druhů dle různých kritérií. William H. Inmon v knize *Building the Data warehouse* (2005) dělí datové sklady dle způsobu v jaké rozsahu jsou shromažďovaná data a dle prostředí, ve kterém je datový sklad nasazen.

Pro první rozdělovací kritérium, tedy dle rozsahu shromažďovaných dat zavádí Inmon postupně tři kategorie, a to hluboký datový sklad (angl. Deep Data Warehouse), široký datový sklad (angl. Wide Data Warehouse) a hybridní datový sklad (angl. Hybrid Data Warehouse). Tyto jednotlivé kategorie jsou pak definovány takto:

- Hluboký datový sklad
  - Shromažďovaná data jsou převážně historická a velmi detailní, slouží primárně k analýze vývoje a trendů.
- Široký datový sklad
  - Zaměřuje se na velké spektrum informací, přičemž neukládá data v takovém detailu, využívá se především pro tvorbu přehledů a rychlé vyhledávání dat.
- Hybridní datový sklad
  - Jedná se o kombinaci obou předchozích kategorií

V případě druhého kritéria, které dělí datové sklady dle prostředí, ve kterém je nasazujeme, pak Inmon definuje čtyři kategorie, a to Operational Data Stores, Data marts, Enterprise data warehouse a Virtual data warehouse. Toto rozřazení je pak Inmon knize definuje následovně

- Operational Data Stores
  - Datový sklad, který určen k provozním potřebám organizace, integruje a centralizuje data z firemních procesů.
- Data marts
  - Malé datové sklady, typicky se budují pro jednotlivá oddělení organizací a na základě jejich specifických potřeb.
- Enterprise data warehouse
  - Datový sklad, který centralizuje všechna data z celé organizace a slouží jako hlavní znalostní báze pro její řízení.
- Virtual data warehouse
  - Specifický typ datového skladu, který funguje primárně jako vrstva nad různými datovými zdroji a pro uživatele tak vytváří jeden přístupový bod.

Každý z těchto typů datových skladů vyžaduje specifický přístup, a to jak k architektuře samotného skladu, zpracování dat, tak i k technickému vybavení. Z tohoto důvodu je důležité zvolit správný typ datového skladu v časné fázi jeho návrhu, neboť možnost konverze mezi jednotlivými druhy nemusí být vždy možná.

Pavel Dytrych - 527565

Zpracování dat v datových skladech

Vzhledem k tomu, že datové sklady nejsou primárním cílem nově vygenerovaných dat, je potřeba data do datových skladů nejprve nahrát a předtím patřičně upravit. Celý tento proces bývá zpravidla automatický, byť to není nikterak podmíněno.

Proces nahrávání dat do datového skladu se nazývá ETL (z angl. Exract, transform and Load) a jedná se o esenciální součást datového skladu. Celý proces se skládá, jak již název napovídá, ze tří fází, a to fáze extrakce, transformace a nahrání dat.

Kimball a Caserta (2004) celý proces rozšiřují do fází čtyř, a to extrakce, čištění, přizpůsobení (angl. Conforming) a doručení. V další části se tedy budu zbývat tímto rozdělením ETL procesu, a ne pouze základními třemi fázemi.

ETL proces dle Kimballa a Casertyho

Extract

První operací, kterou je potřeba pro nahrání dat udělat, je jejich extrahování z původních uložišť, to probíhá většinou v přesně definovaných cyklech. Během této operace dojde k nahrání všech požadovaných dat do přechodného uložiště, odkud budou dále zpracovávány v dalších krocích ETL procesu. Data jsou nahrávána v surové formě a v původních formátech a souborech, které mohou představovat například data z relačních databází, XML, csv, JSON nebo XLS soubory. (Kimball & Caserta, 2004, p. 18)

Surová data bývají po zpracování smazána, ale v některých případech se ponechávají jako dlouhodobá záloha, někdy se též nechávají do dalšího nahrávacího cyklu, aby bylo možné porovnat změny mezi jednotlivými průběhy. (Kimball & Caserta, 2004, p. 18)

V některých případech je možné využít jako zdroj dat pro datový sklad uložiště jiného typu, typicky například Data Lake.

Cleaning

Poté, co jsou požadovaná data vyextrahována je potřeba je vyčistit. Během této fáze se provádí kontrola integrity dat, odstraňují se duplicity a provádějí další operace s cílem dosáhnutí požadované kvality dat. (Kimball & Caserta, 2004, pp. 18-19)

Během této fáze také dobré zvolit požadovanou granularitu neboli míru detailu uchovávaných informací (Inmon, 2005, p. 41), a případně data na požadovanou úroveň zdecimovat. Granularita patří mezi základní parametry využívané při návrhu datového skladu, a tak by měl i její hodnota z tohoto návrhu vycházet. (Inmon, 2005, p. 41)

Přizpůsobení (Conforming)

Fáze potvrzování má význam zejména v případech, kdy extrahujeme data z více datových zdrojů. Jejím primárním úkolem je data z různých zdrojů propojit dohromady tak, aby bylo možné se dotazovat napříč všemi těmito zdroji (Kimball & Caserta, 2004, p. 19). Dalším důležitým úkolem pak je kontrola konfliktů mezi názvy jednotlivých dimenzí, neboť napříč datovými zdroji nemusí mít vždy stejný význam (Kimball & Caserta, 2004, p. xxx).

Doručení (Delivering)

Během fáze doručení pak dochází k samotné transformaci dat do stejného formátu a následně se z dat vytvoří dimenzionální model nebo požadované schéma, jako jsou například hvězdicovité schéma či datová krychle, což značně zkracuje čas zpracování dotazu (Kimball & Caserta, 2004, p. 19).

Výstupem této fáze je tedy již patřičně setříděný a provázaný soubor dat, nad kterým je již možné provádět dotazy, jako nad celkem, tedy bez ohledu na datové zdroje a jejich formát. Tento datový soubor se pak následně nahraje do databáze samotného datového skladu.

Díky ETL procesu jsou ve výsledku v datovém skladu data uloženy v jednoduché, snadno dotazovatelné formě, což značně usnadňuje přístup a práci s těmito daty.

Pavel Dytrych - 527565

Využití datových skladů

Datové sklady nacházejí velké využití zejména u velkých organizací a společností, ale i u tak velkých celků, jako jsou například samostatné státy. Velkou výhodou datových skladů je jejich perzistentní struktura, kdy data můžeme shromažďovat po dlouhou dobu a jejich analýzu provést až ve chvíli, kdy nastane její potřeba. Díky tomu se velmi hodí na analýzu dlouhodobých dat.

Jedním z typických zástupců tohoto typu nasazení je pak například sdílení dat mezi vědeckými pracovišti. Například v případové studii *Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer* (Seneviratne, Seto, Blayney, Brooks & Hernandez-Boussard, 2018) se autoři zabývali propojením vědeckých pracovišť Stanford University, Stanford Cancer Institute a California Cancer Registry za účelem vytvoření společného datového skladu obsahujícího elektronické zdravotní záznamy pacientů postižených rakovinou prostaty. Díky tomuto datovému skladu tak vědci přístup mají přístup k tisícům reálných případů, napříč těmito třemi institucemi.

Další velmi zajímavou aplikací, je například optimalizace a analýza mléčné farmy. Tato aplikace popsaná v článku *Building an active semantic data warehouse for precision dairy farming* (Schuetz, Schausberger & Schrefl, 2018) se zabývá návrhem a implementací datového skladu, který má za cíl pomocí různých senzorů, sledujících například pohyb dojnic či samotné dojení, zoptimalizovat a vylepšit výkon této mléčné farmy. Celý projekt je velmi zajímavý zejména z toho důvodu, že se jedná o velmi specifickou oblast, kde ještě nebyl koncept datového skladu nasazen.

Zajímavým příkladem jednoúčelově zaměřeného datového skladu je projekt, týkající se nedávno proběhlé pandemie viru SARS-COV-2. V článku *COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data* (Agapito, Zucco & Cannataro, 2020) , autoři popisují implementaci datového skladu vytvořeného za účelem monitorování šíření viru v Itálii. Zajímavé je, že datový sklad zahrnuje kromě informací týkajících se přímo samotného viru i další informace, které zahrnují klima, znečištění ovzduší, ale i například sílu a směr větru. Výzkumné výstupy z tohoto projektu se tak mohou zabývat i vlivem počasí na šíření nákazy.

Závěr

Cílem této práce bylo poskytnout základní vhled do historie, fungování a aplikace datových skladů. V práci jsem se snažil zachytit dle mého názoru nejpodstatnější body týkající se problematiky datových skladů, avšak vzhledem k omezenému rozsahu je množství zahrnutých informací opravdu jen velmi základní.

V práci jsem postupně prošel definic datového skladu, historii jeho vývoje a rozdělení. Poněkud větší část je pak věnována samotnému procesu zpracování dat ETL procesem, ta to část je však velmi zkrácená a bylo by vhodné ji patřičně rozšířit.

V poslední kapitole se pak věnuji několik aplikacím datového skladu v reálném světě, které mi přišli zajímavé a pro dané téma relevantní.

Bibliografie

Agapito, G., Zucco, C., & Cannataro, M. (2020). COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data. *International Journal of Environmental Research and Public Health*, *vol. 17*(issue 15). https://doi.org/10.3390/ijerph17155596

Foote, K. (2018). A Brief History of the Data Warehouse [Online]. Retrieved from https://www.dataversity.net/brief-history-data-warehouse/

Inmon, W. (2005). *Building the Data Warehouse*. Hoboken (New Jersey): Wiley.

Kimball, R., & Caserta, J. (2004). *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. Indianapolis, IN: Wiley.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Hoboken (New Jersey): Wiley.

Seneviratne, M., Seto, T., Blayney, D., Brooks, J., & Hernandez-Boussard, T. (2018). Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer, *vol. 6*(issue 1). https://doi.org/10.5334/egems.234

Schuetz, C., Schausberger, S., & Schrefl, M. (2018). Building an active semantic data warehouse for precision dairy farming. *Journal of Organizational Computing and Electronic Commerce*, *vol. 28*(issue 2), 122-141. https://doi.org/10.1080/10919392.2018.1444344

Smil, V. (2021). *Čísla nelžou: 71 věcí, které byste měli vědět o světě*. Praha: Kniha Zlin.