Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženýrství

Bakalářská práce

Archivace dat v datovém skladu

Jan Šnaur

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Jan Šnaur

Informatika

Název práce

Archivace dat v datovém skladu

Název anglicky

Data backup in a datawarehouse

Cíle práce

Bakalářská práce je zaměřena na problematiku návrhu, instalaci a spravování datového skladu. Hlavní cíl je seznámení se s možnostmi reálného návrhu a realizace datového skladu pomocí různých softwarových prostředků a jejich využití. Zjištěné možnosti budou využity pro sestavení modelu a následné realizaci modelového prototypu.

Metodika

Metodika řešené problematiky bakalářské práce je založena na studiu a analýze odborných informačních zdrojů. Teoretické poznatky budou převedeny na návrh a popis modelového řešení. Pomocí návrhu bude dále modelové řešení realizováno.

Doporučený rozsah práce

30-50 stran

Klíčová slova

Data warehousing, Data archiving, Datové sklady

Doporučené zdroje informací

Database Archiving: How to Keep Lots of Data for a Very Long Time (Autor: Jack E. Olson)
Databáze, datové sklady, analýza OLAP a dolování dat (Autor: Luboslav Lacko; Výrobce/vydavatel: Computer Press)
Datové sklady – Agilní metody a business intelligence (Autor: Robert Laberge; Výrobce/vydavatel: Computer Press)
The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Autor: Ralph Kimball a Margy Ross)

Předběžný termín obhajoby

2017/18 LS – PEF

Vedoucí práce

doc. Ing. Vojtěch Merunka, Ph.D.

Garantující pracoviště

Katedra informačního inženýrství

Elektronicky schváleno dne 7. 3. 2018

Ing. Martin Pelikán, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 7. 3. 2018

Ing. Martin Pelikán, Ph.D.

Děkan

V Praze dne 14. 03. 2018

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Archivace dat v datovém skladu" jsem vypracoval(a) samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor(ka) uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil(a) autorská práva třetích osob.

V Praze dne 15.3.2018 _____

Poděkování

Rád bych touto cestou poděkoval panu doc. Ing. Vojtěchu Merunkovi, Ph.D. za odborné vedení mé práce a za rady při zpracovávání mé bakalářské práce. V neposlední řadě také děkuji své rodině za velkou podporu.

Abstrakt (česky):

Bakalářská práce se zaměřuje na problematiku využití dostupných technologií pro vybudování datového skladu. Teoretická část práce stručně rozebírá význam jednotlivých komponent datového skladu. V praktické části práce je prováděna analýza technologií datových skladů od různých výrobců, které jsou bodově hodnoceny na základě požadavků fiktivní společnosti. V závěru práce jsou shrnuty výsledky analýzy jednotlivých výrobců a je provedeno porovnání. Z výsledků porovnání vyplývá, že konkurence v oblasti datových skladů a technologií s tím souvisejících je velmi vyrovnaná.

Klíčová slova: datové sklady, databáze, OLAP, business intelligence, dolování dat, technologie datových skladů

Abstract (in English):

This thesis focuses on the issue of use of available technologies to build a data warehouse. The theoretical part briefly discusses the importance of individual components of the data warehouse. The practical part is carried out analysis of data warehouse technology from various manufacturers that are scored based on the requirements of a fictitious company. The conclusion summarizes the results of the analysis of individual manufacturers and the comparison is made. The results of the comparison shows that competition in the field of data warehousing and related technologies is very balanced.

Keywords: data warehouse, databases, OLAP, business intelligence, data mining, data warehousing technology

Obsah

1 Úvod 10
2 Datové sklady 11
2.1 Definice 11
2.1.1 William Inmon 11
2.1.2 Ralph Kimball 13
2.2 Schéma hvězda „star“ 15
2.3 Schéma sněhová vločka „snowflake“ 17
3 Online Analytical Processing – OLAP 19
3.1 Data cubes 19
3.2 Druhy OLAP systémů 21
3.2.1 MOLAP 21
3.2.2 ROLAP 22
3.2.3 HOLAP 23
4 Business Intelligence (BI) 24
4.1 Definice BI 24
4.2 Složení a fungování BI 25
4.2.1 Zdrojové systémy 26
4.2.2 Vrstva transformace dat 26
4.2.3 Databázová vrstva 27
4.2.4 Analytická vrstva 28
4.2.5 Prezentační vrstva 29
5 Strategie budování datového skladu 31
6 Analytická část 33
6.1 Gartner & Forrester 34
6.2 Analýza produktů 37
6.2.1 HP 37
6.2.2 IBM 38
6.2.3 Amazon Web Services 39
6.2.4 SAP 41
6.2.5 Microsoft 42
6.2.6 Oracle 43

6.2.7 Teradata 44
6.3 Diskuze 45
7 Závěr 47
8 Seznam literatury 48

Seznam obrázků

Obrázek 1 - Datová struktura "top-down" 12
Obrázek 2 - Datová struktura "bottom-up" 14
Obrázek 3 – Schéma hvězda 16
Obrázek 4 – Schéma sněhová vločka 17
Obrázek 5 – Obecné zobrazení datové kostky 19
Obrázek 6 – Využití datové kostky 20
Obrázek 7 - Princip složení BI 25
Obrázek 8 – Srovnání řešení datových skladů Gartner 35
Obrázek 9 – Srovnání řešení datových skladů Forrester 36

Seznam grafů

Graf č. 1 – Hodnocení HP 38
Graf č. 2 – Hodnocení IBM 39
Graf č. 3 – Hodnocení Amazon Web Services 40
Graf č. 4 – Hodnocení SAP 41
Graf č. 5 – Hodnocení Microsoft 42
Graf č. 6 – Hodnocení Oracle 43
Graf č. 7 – Hodnocení Teradata 44

Seznam tabulek

Tabulka 1- Výsledné bodové hodnocení 45

1 Úvod

Potřeba realizace datového skladu je v dnešní době u společností operující se svými daty ve smyslu provádění analýz a reportingu téměř nutností. Datové sklady se už netýkají jen velkých společností, které disponují velkým množstvím dat, ale i manažeři v malých a středních firmách potřebují svá rozhodnutí opřít o data z provozních systémů a o jejich analýzy. Drtivá většina společností uchovává svá data, ať už v klasických relačních databázích, nebo v úložištích založených na cloudové technologii. Součástí řešení datových skladů však není jen datové úložiště, ale celá řada dalších funkcí, které umožňují z těchto dat vytěžit cenné informace. Tyto informace pak mohou zefektivnit rozhodování v klíčových momentech a představovat tak možnou konkurenční výhodu. Existuje velké množství produktů od různých společností, které tato řešení nabízí, ne všechny však poskytují stejně kvalitní služby a aplikace pro realizaci datového skladu.

Tato bakalářská práce analyzuje a hodnotí možnosti realizace datového skladu podle různých poskytovatelů těchto řešení. Čerpá z dostupných informací jednotlivých výrobců a dále z informací nestranných společností dlouhodobě provádějící praktický výzkum, analýzu a testování datových skladů od distributorů těchto řešení. Tyto informační zdroje budou na příkladu fiktivní společnosti se stanovenými výběrovými kritérii dále analyzovány a bodovány, a na základě těchto výsledků bude vybrán výrobce s nejlepším hodnocením.

2 Datové sklady

2.1 Definice

Bezpochyby dva z největších průkopníků v oblasti datových skladů, z anglického termínu Data Warehouse, byli William Inmon a Ralph Kimball. V následujících kapitolách si blíže představíme samotné autory a především jejich rozdílný pohled na pojetí datových skladů.

2.1.1 William Inmon

Jako první z této dvojice v roce 1991 popsal tento pojem William Inmon, který je proto označován za "Father of Data Warehousing". (Anupindi, 2005) Definice však není jediný přínos, jenž do dané problematiky vnesl, tomuto tématu se neúnavně věnoval i nadále. Publikoval více než 50 knih, 650 článků a každý měsíc psal sloupky pro Business Intelligence Network. (Corporate Information Factory, 2007)

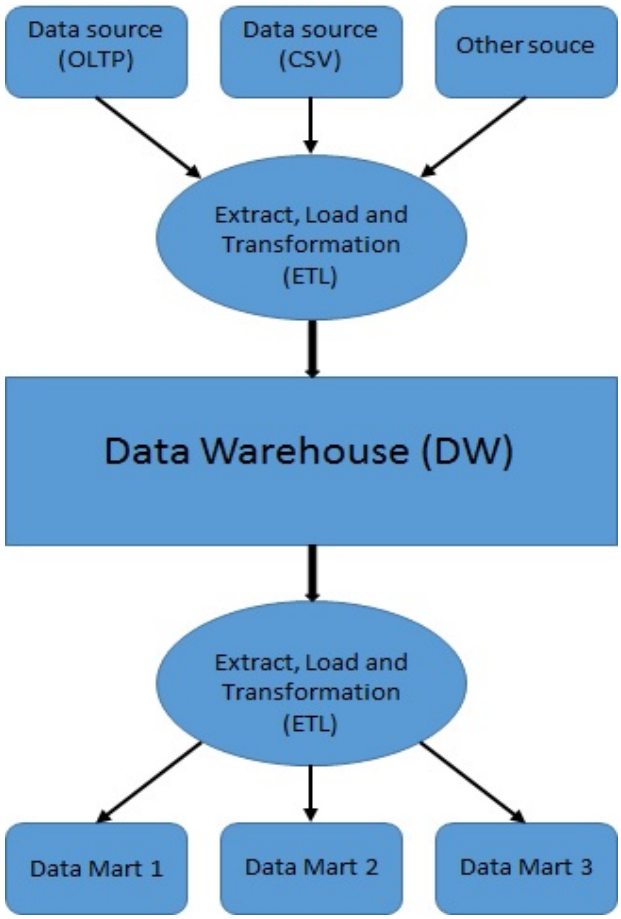
„Datový sklad je subjektově orientovaná, integrovaná, neměnná a časově rozlišitelná kolekce dat sloužící pro podporu rozhodování. Datový sklad obsahuje granulární korporátní data."(Inmon, 2002, str. 31)

Z definice plyne, že autor klade velký důraz na pojmy subjektová orientace, integrovanost, stálost, časová rozlišitelnost a granularita. Pro potřeby této práce považuji za důležité objasnit si každý pojem autorovy formulace z důvodu správného pochopení Inmonova výkladu.

- **Subjektová orientace** – údaje jsou rozdělovány podle typu a ne podle zdrojů těchto dat.
- **Integrovanost** – do úložiště mohou vstupovat data z různých podnikových zdrojů v různých formátech a také v různých jednotkách. Tato vlastnost klade důraz na převod a sjednocení formátů a jednotek.

- **Stálost** – data se po uložení už žádným způsobem nemění, datové sklady jsou budovány pouze pro čtení.
- **Granularita** – udává úroveň detailnosti dat. S rostoucí detailností dat klesá úroveň granularity.

Obrázek 1 - Datová struktura "top-down"



Zdroj: vlastní tvorba podle Danish, 2010.

Rozdílnost pojetí datového skladu Imona se od Kimbalova liší jak v jeho složení, tak i v následném vývoji. William Inmon zastává „top-down“ přístup a jeho myšlenka vychází z vytvoření centralizovaného datového skladu, který zaštiťuje celý podnik. A až po jeho vytvoření budovat konkrétní databáze přizpůsobené individuálním požadavkům jednotlivých částí podniku. Tyto jednotlivé databáze se nazývají datová tržiště, anglicky „data marts“. (Danish, 2010)

Data marts se vyznačují zaměřením na jedinou funkční oblast, tedy pro jedno oddělení jako naříklad Sales, Consumer nebo Finance. (A Data Mart Concepts, 2007) Tím dochází k usnadnění práce datového skladu, který je následně schopen lépe vyhovět potřebám jednotlivých uživatelů.

Schéma datové struktury „top-down“ velmi jednoduchým způsobem popisuje postup zpracování informací od jejich získání ve zdrojových systémech a aplikacích až po vytvoření jednotlivých datových tržišť. Při čtení diagramu postupujeme, jak už napovídá sám název, od shora dolů. Po získání dat ze zdrojových systémů se dále přesouvají ke zpracování, kde se čistí, konsolidují a ověřují, aby se zajistila jejich správnost. Dalším krokem je proces sjednocení, jelikož jednotlivá data pocházejí z různých provozních systémů, proto mohou být uchovávána v různých formátech. Poté již putují do samotného datového skladu. Z tohoto umístění jsou buď přímo distribuovány do jednotlivých datových tržišť, nebo nastává ještě mezikrok, během kterého probíhají operace s těmito daty, a až poté jsou řazeny do data martů. (Danish, 2010)

Realizace data warehousu podle této struktury je složitá a také nákladná. Tato finanční a technická náročnost způsobuje, že podniky tedy mohou na své výsledné řešení mnohdy čekat velmi dlouho.

2.1.2 Ralph Kimball

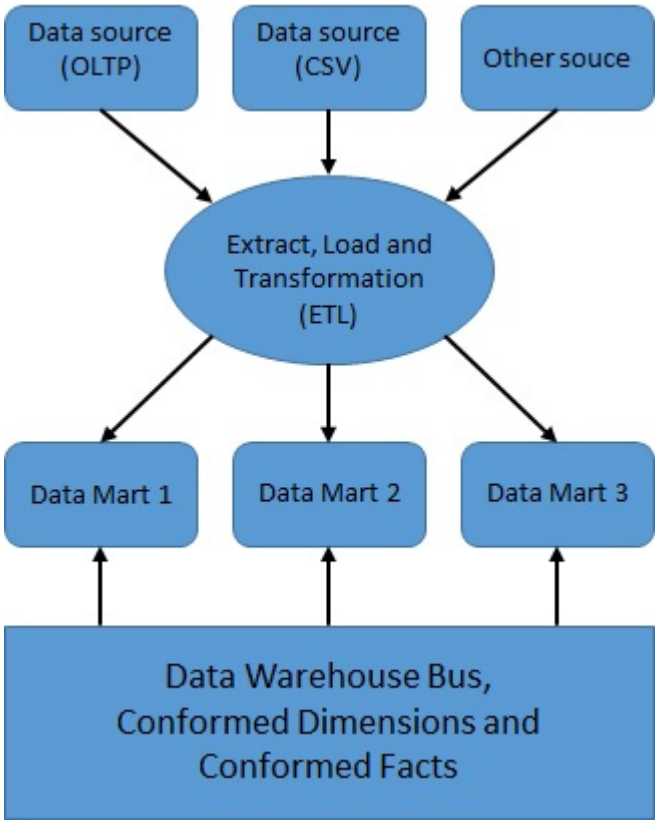
Druhou neméně důležitou osobou zabývající se touto problematikou je Ralph Kimball. Popsal využití a implementaci datové struktury „star“ a „snowflake“. Na datové sklady pohlížel spíše z hlediska využívání informací, což vedlo ke zrodu „Business intelligence“. (Kimball, 2010)

Kimball si představuje datové sklady jako:

“DWH je kopie transakčních dat speciálně strukturovaných pro dotazování a reportování.” (Gála, 2006, str. 103)

Definice je o mnoho stručnější a jednodušší, než jak tomu bylo v předchozím případě. Autor na první místo při tvorbě datového skladu dává potřebu vytvořit „datová tržiště“, která budou lépe splňovat potřeby jednotlivých funkčních částí společnosti. Jedná se o funkční pohled na data warehouse. Nezabývá se tolik tím, jakým způsobem je datový sklad postaven, jak tomu bylo u Inmona, a spíše se soustředí na funkčnost datového skladu a na správné využívání těchto dat. (Gála, 2006)

Obrázek 2 - Datová struktura "bottom-up"



Zdroj: vlastní tvorba dle Danish, 2010.

Kimball se soustředil na strukturu skladu nazývanou „bottom-up“. Jako první se v tomto případě vytváří data marty, které jsou již modelovány tak, aby splňovaly dílčí potřeby jednotlivých částí podniku. Tento model budování začíná tím, co předchozí model „top-down“ vytvářel naposledy. (Kimball, 2010)

Po vytvoření datových tržišť nastává fáze integrace a kombinování, jejímž výsledkem je samotný datový sklad. Kombinování data martů probíhá podle tabulky dimenzí a faktů, jak je znázorněno v dolní části modelu. Obě tabulky si důkladněji popíšeme. (Kimball, 2010)

Tabulky faktů

Tabulky faktů obsahují ukazatele, metriky a „fakta“ konkrétního objektu v procesu a cizí klíče propojující tyto tabulky s tabulkami dimenzí. Tvořeny jsou velkým množstvím číselných údajů a utváří jádro každého schématu. Mohou to být například počty vyrobených zařízení, počet kusů zboží ve skladu, nebo například firemní finanční zisk. (Designing Star Schema, 2015)

Tabulky dimenzí

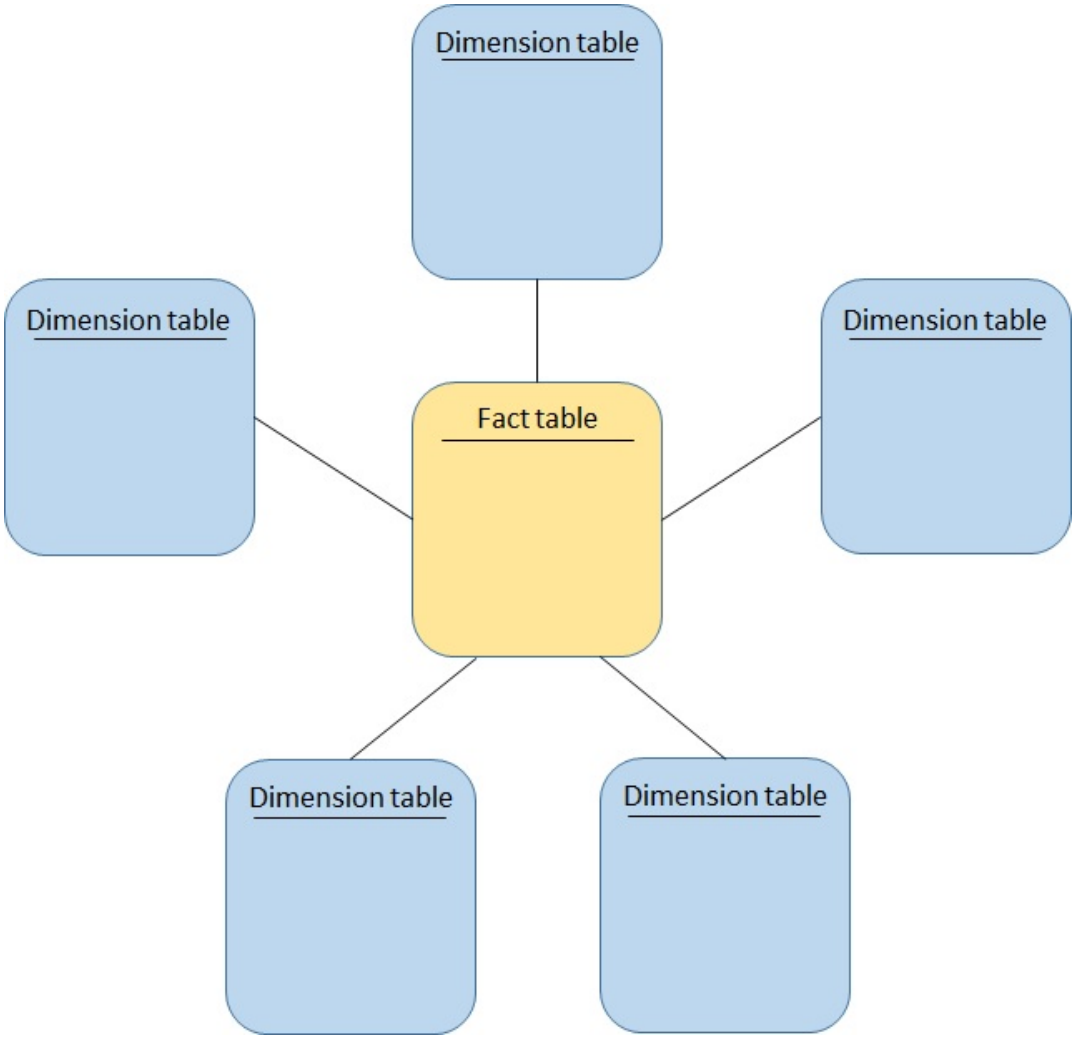
Tyto tabulky jsou hierarchicky seřazeny a skládají se z atributů, které omezují, seskupují a třídí data ve faktových tabulkách. Pro lepší představu toho, jak spolu tyto tabulky souvisí, si uvedeme příklad. (Model Rational Insight Data Warehouse, 2010)

Představme si firmu vyrábějící například svítidla. O každém svítidle eviduje cenu, výrobní číslo, rozměry, hmotnost a počet kusů. Všechny tyto atributy budou obsaženy právě ve faktové tabulce jako číselné hodnoty. Naopak atributy jako typ svítidla, třída svítidla, nebo datum budou uloženy v tabulkách dimenzí. Pod typem mohou uvádět, zda jde o stojací nebo nástěnné svítidlo, a třída může označovat, jestli je svítidlo určeno do interiéru či exteriéru. Tabulky propojuje „cizí klíč“, který definuje jejich vzájemný vztah. (Model Rational Insight Data Warehouse, 2010) Na první pohled je zřetelný rozdíl mezi obsahem tabulky faktů, anglicky „Fact table“ a obsahem tabulek dimenzí, anglicky „Dimensions tables“. Dále se zaměříme na schémata propojení „Fact table“ a „Dimensions tables“, konkrétněji hvězdicové schéma a schéma sněhové vločky. (Designing Star Schema, 2015)

2.2 Schéma hvězda „star“

Název tohoto schématu vyplývá, jak už sám název napovídá, z jednotlivých prvků tvořících hvězdicový tvar. V samotném středu schématu je již několikrát zmiňovaná faktová tabulka obklopená dimenzemi, se kterými je s každou zvlášť propojena cizím klíčem. Každá dimenze je reprezentována jednou tabulkou. (Star Schema, 2015)

Obrázek 3 – Schéma hvězda

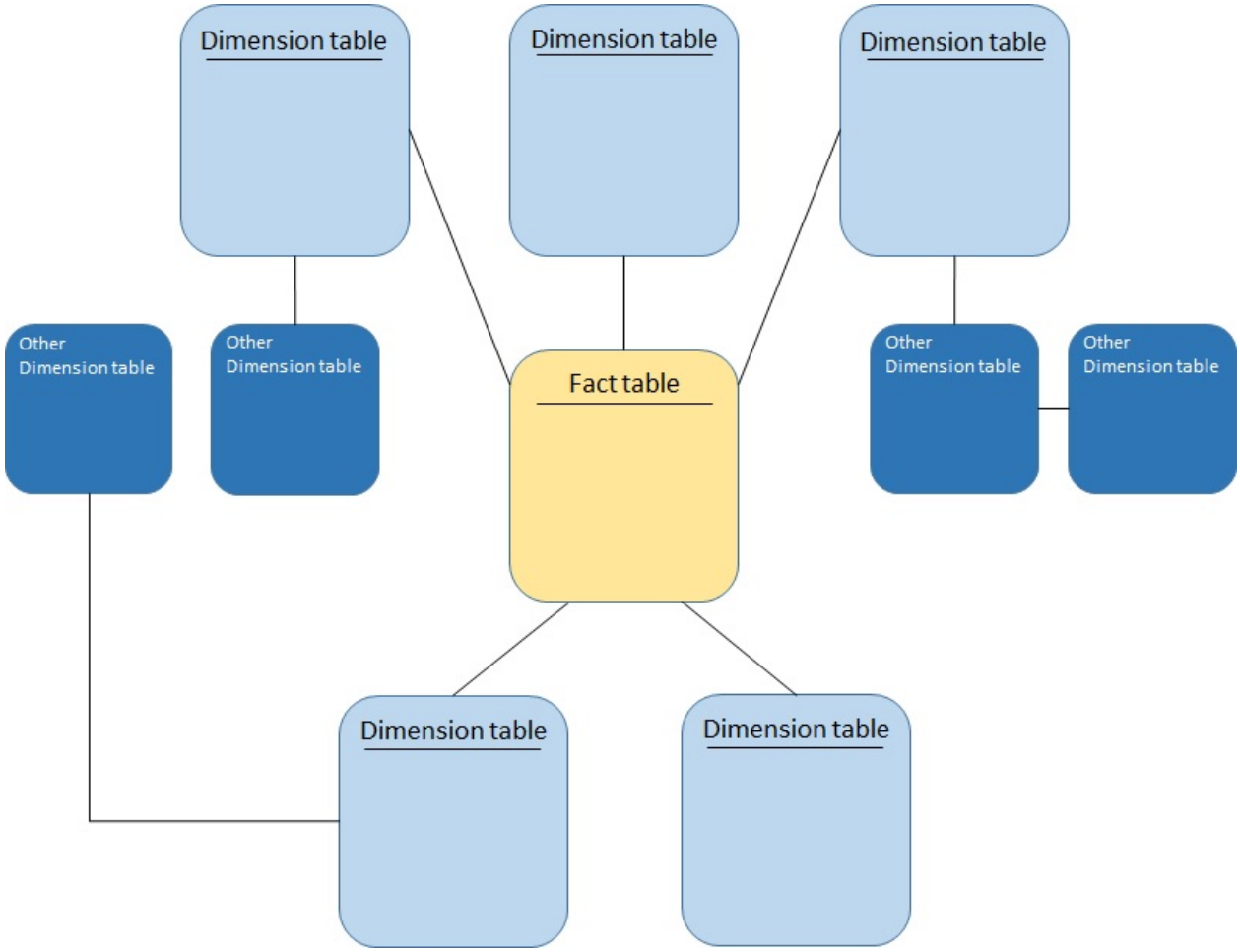


Zdroj: vlastní tvorba dle Novotný, 2005.

Při realizaci tohoto schématu však může nastat situace, kdy se některá data ukládají duplicitně. Duplicita je zapříčiněna hierarchickým řazením dat v každé z dimenzí, které totiž nejsou normalizovány. Z toho plyne jeho velká nevýhoda, což je značně pomalé vytváření tohoto modelu. Naopak jako výhodu lze označit rychlé provádění dotazů. (Lacko, 2011)

2.3 Schéma sněhová vločka „snowflake“

Obrázek 4 – Schéma sněhová vločka



Zdroj: vlastní tvorba dle Novotný, 2005.

Schéma sněhové vločky je o něco složitější než předchozí schéma hvězdy. A to proto, že všechny dimenze nemusí být přímo provázány s faktovou tabulkou, ale mohou se propojovat mezi sebou a tvořit tak hierarchickou strukturu na dimenzionální úrovni, a nikoli pouze na úrovni dat, jak tomu bylo u schématu hvězda. Tabulky jsou zde již normalizovány z důvodu snížení duplicity dat. (Snowflake Schema, 2015)

Za velkou výhodu zle považovat snížení velikosti v důsledku eliminace duplicity. Přesto že je schéma složitější, tak některé aplikace provádějící analýzu s tímto schématem pracují daleko lépe.

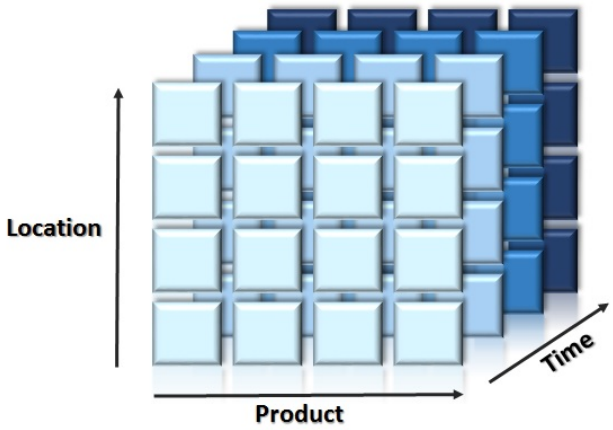
William Inmon se přiklání k organizování dat právě pomocí tohoto modelu. Sice jej pokládá za složitější, ale co do velikosti je díky normalizaci menší, a tím tato struktura usnadňuje procházení záznamů v data warehousu. (Inmon, 2002) Ralph Kimball naopak tvrdí, že právě struktura snowflake schématu je sama sobě určitou překážkou. Zdůrazňuje, že používání normalizovaných schémat nepřináší žádné výhody, ale naopak nevýhody. Vzhledem ke složitosti celého modelu se snižuje možnost jeho využití. S ohledem na velikost databáze s dimenzemi, je vetší množství dimenzí a dvojí uložení některých atributů v modelu hvězda oproti sněhové vločce zanedbatelné a je tedy bezvýsledné používat normalizované modely. (Kimball, 2010)

3 Online Analytical Processing – OLAP

Technologie OLAP poskytuje datům v datovém skladu určitou strukturu tak, aby byla lépe pochopitelná a dosažitelná pro uživatele zabývající se jejich hodnocením a analýzou. OLAP vykonává multidimenzionální analýzu dat a umožňuje tak provádět složité výpočty, které jsou potřebné pro analýzu a datové modelování. To vše slouží jako základ pro aplikace business intelligence, které jsou koncipované například pro plánování, finanční a datový reporting, provádění předpovědi trendů a analýz, pro vytváření simulačních modelů a především pro získávání znalostí a vykazování údajů z data warehouse. Díky této technologii můžou koncoví uživatelé provádět ad hoc reporty v několika rozměrech, což jim výrazně usnadňuje akutní rozhodování a eliminuje možnost špatného rozhodnutí z důvodu špatného porozumění, nebo interpretace dat. (Easy OLAP Definition, 2015)

3.1 Data cubes

Obrázek 5 – Obecné zobrazení datové kostky



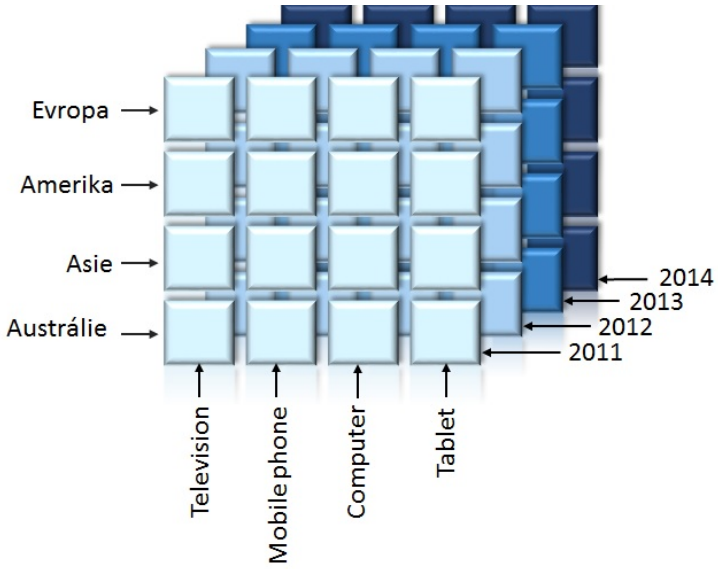
Zdroj: vlastní tvorba.

Data cubes, neboli datové kostky jsou základními stavebními kameny Online analytical processing. Datová kostka je množina dat zkonstruovaná jako podmnožina datového skladu a je organizována a sumarizována do multidimenzionální struktury definované konkrétním počtem dimenzí. Poskytuje jednoduše použitelný mechanismus s rychlou odezvou, založený

na principu dotazování nad daty. Pro dotazování do kostek se používají klientské aplikace schopny připojit se na analytický server a podle příkazů hledat v těchto kostkách. Velká většina aplikací pracujících s data cubes umožňuje jednoduchou manipulaci pomocí ovládacích prvků, které následně určují a generují obsah dotazu směřujícího na server do konkrétní datové kostky. (Introduction to Cubes, 2002)

Značně tím snižují požadavky na uživatele, kteří již nemusí mít obsáhlé znalosti syntaxe v oblasti tvoření dotazů, a velmi tím usnadňují práci. Před připojením do kostky se vytváří souhrnné záznamy zvané agregace. Ty poskytují předem vypočítané údaje zrychlující provedení dotazu. Výsledek dotazu je převzat právě z agregací předpočtených ze zdroje datové kostky v datovém skladu, v klientově cache paměti nebo kombinací těchto zdrojů. Data cube má určité schéma, které si můžeme představit jako soubor spojených tabulek, čerpajících svá data ze zdrojového umístění. (Introduction to Cubes, 2002)

Obrázek 6 – Využití datové kostky



Zdroj: vlastní tvorba.

Centrální tabulka ve schématu je už výše zmiňovaná tabulka faktů a všechny ostatní tabulky jsou tabulky dimenzí. Tyto dimenze mohou být hierarchicky řazeny pro větší přehlednost koncovým uživatelům. Datová kostka může obsahovat libovolný počet polí,

nicméně tvůrci se snaží najít rovnováhu mezi potřebami uživatelů a logickým omezením daného modelu. (Introduction to Cubes, 2002)

3.2 Druhy OLAP systémů

Technologie OLAP je strukturou na jejíž základu vnikly další odlišné modifikace, které se mezi sebou výrazně liší. Každá modifikace nebo druh obecně doplňuje název základní struktury OLAP o další písmeno, určující zaměření a charakterizující jejich odlišnost od ostatních. (Major OLAP Technology Types, 2015)

Základní druhy jsou MOLAP a ROLAP, které jsou také nejvíce rozšířeny a využívány. Dále existují ještě například SOLAP a DOLAP, jenž představují spíše rozdílnost marketingových programů ze strany distributorů. My si popíšeme první dva a jejich vzájemnou kombinaci HOLAP.

3.2.1 MOLAP

Multidimensional OLAP umožňuje uživatelům modelování dat v multidimenzionálním prostředí a nikoliv jen multidimenzionální pohled na relační data, jako je tomu u technologie ROLAP. Základem této technologie jsou tedy datové kostky. MOLAP poskytuje nejrychlejší a nejflexibilnější metody pro zpracování vícerozměrných požadavků. (Major OLAP Technology Types, 2015)

Struktura tohoto modelu není jen řada tabulek, jak je tomu v relačních databázích, ale tvoří ho již výše rozebrané data cubes, neboli datové kostky. Výhody vyplývají z kostek samotných, což je neomezenost počtu členů v jedné dimenzi, zatímco v obyčejné tabulce jsme omezeni na průsečík pouze dvou členů. V multidimenzionální databázi tedy můžeme přidávat celé další dimenze, a ne jen tabulky, což je velká výhoda. MOLAP cube umožňuje mimořádně rychlé a pružné výpočty a datové modelování, protože místo hledání indexu celého datového modelu pomocí SQL příkazů, jako v relační databázi, je aplikace na základě průsečíků

dimenzí v datové kostce schopna dohledat a identifikovat umístění podle názvu. (Major OLAP Technology Types, 2015)

Nevýhodou je velký prostor, který tato organizace dat zabírá na disku. S tímto modelem lze sice pracovat off-line, ale z důvodu aktuálnosti zpracovávaných dat je téměř vždy vyžadováno on-line připojení, dále se příliš nehodí pro větší objemy dat a dimenzí, protože při větších objemech dochází k řídnutí datových kostek. Výhodou MOLAP je naopak možnost vstupování dat z odlišných datových zdrojů a velmi rychlé dotazování díky předpočítaným agregacím. (Major OLAP Technology Types, 2015)

3.2.2 ROLAP

Technologie typu Relational OLAP se vyznačují přímým přístupem k datům uložených v relačních databázích, před zaváděním tohoto systému tedy není nutná úprava relační databáze a zachovává se stávající. Tato struktura je vhodná pro velké objemy dat, ke kterým se nepřistupuje příliš často, to proto že provádění SQL dotazů je časově náročné. Při přístupu k datům se sice agregace vytváří, ale nikam se neukládají, jako v předchozím případě, a vše se musí počítat až při vykonávání dotazu. V případě, že bychom chtěli agregace uchovávat pro zrychlení výpočtů, je nutné je uchovávat ve zvláštních tabulkách, ke kterým se přistupuje při provádění dotazu. (Major OLAP Technology Types, 2015)

Výhoda této technologie spočívá v podstatě v neomezené velikost dat, data jsou omezeny pouze kapacitou primárního systému a díky nutnosti dotazování jsou také aktuální, protože po každém dotazu dostáváme aktualizovaná data přímo z databáze. Za nevýhody se dá považovat již výše zmíněná absence agregací způsobující pomalé výpočty, které probíhají až při vykonávání dotazu a podmínkou tohoto řešení je rovněž existence databáze primárního systému. (Major OLAP Technology Types, 2015)

3.2.3 HOLAP

HOLAP, neboli hybridní OLAP, je výsledek snahy o spojení nejlepších vlastností MOLAP a ROLAP do jedné funkční struktury. Využívá jak multidimenzionální databáze, tak i relační databázový systém, a tím umožňuje ukládat větší množství detailnějších dat v relačních tabulkách, přičemž předpočtené agregace jsou uloženy v kostkách. Objevuje se zde určitá duplicita, ve které je část dat z relační databáze uložena i v kostkách. Tato skutečnost zajišťuje rychlost početního výkonu, jako u technologie MOLAP, a velikost relační databáze je omezena pouze kapacitou primárního systému, stejně jako je tomu v ROLAP. (Major OLAP Technology Types, 2015)

4 Business Intelligence (BI)

Termín Business Intelligence je neodmyslitelně spojena s datovým skladem a v dnešní době je téměř jeho nedílnou součástí. V této kapitole si popíšeme co tento pojem znamená, jakým způsobem funguje a jak je svázán s datovými sklady.

4.1 Definice BI

Pojem business inteligence, který se dostal do popředí zájmu až na konci 20. století, již byl prvně použit mnohem dříve. Už v roce 1958 ho zmínil ve svém článku výzkumník společnosti IBM Hans Peter Luhn. První definice byla takováto:

"The ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal." (Grimes, 2008, str. 1)

Tedy jako schopnost vnímat vzájemné vztahy uvedených skutečností tak, aby vedly k vytouženému cíli. Další definici nabízí Howard J. Dresner, pracovník společnosti Gartner Group v souvislosti s rostoucí potřebou sběru a tím i navyšování množství skladovaných dat. (Gartner IT Glossary, 2013)

“Business intelligence (BI) is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.” (Gartner IT Glossary, 2013, str. 1)

V českém překladu to znamená, že business intelligence je zastřešující pojem, který zahrnuje aplikace, infrastrukturu, nástroje a osvědčené postupy umožňující přístup k analýze informací, které mají zlepšit a optimalizovat rozhodování a výkon. A nakonec definice z posledních let. (Novotný, 2005)

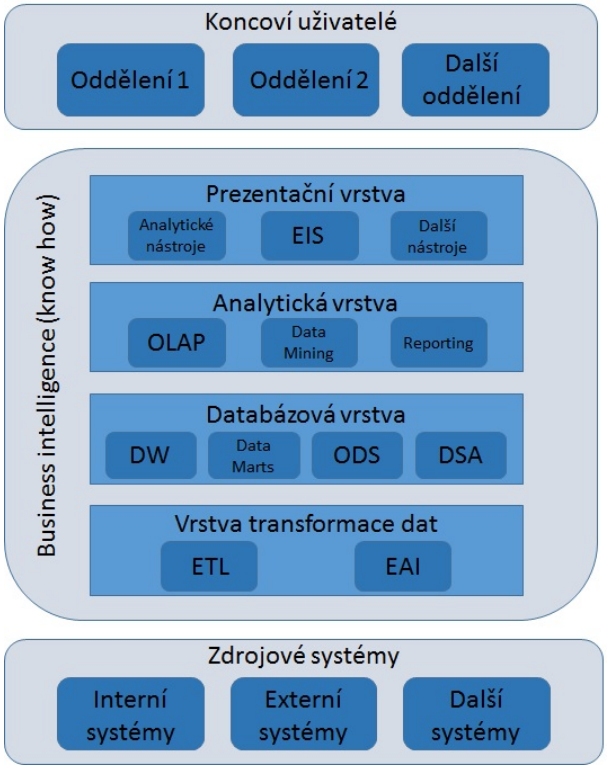
"Business Inteligence je sada procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat rozhodovací procesy ve firmě. Podporují analytické a plánovací činnosti podniků a

organizací a jsou postaveny na principech multidimenzionálních pohledů na podniková data." (Novotný, 2005, str. 19)

Z předchozích definic lze obecně říci, že business intelligence velmi usnadňuje uživatelům těchto systémů, přistupujících do datového skladu, pochopení, interpretaci a především využívání konkrétních dat při důležitých firemních rozhodnutích. Systémy BI se nevztahují pouze k technologiím, ale také ke znalostem procesů, postupů a aplikací, které napomáhají rozhodování. Pracují se spektrem dat od historických až po nejaktuálnější a pomáhají přijímat rozhodnutí důležitá pro budoucnost. (Panec, 2003)

4.2 Složení a fungování BI

Obrázek 7 - Princip složení BI



Zdroj: vlastní tvorba Novotný, 2005.

Konkrétní složení a fungování můžeme označit za tak zvané „know how“, které si tvůrci těchto systémů velmi střeží. Proto se i uspořádání jednotlivých komponent v různých řešeních

BI zpravidla výrazně liší, a to v závislosti na potřebách a situaci daného zákazníka. Strukturu nebo složení z hlediska jednotlivých vrstev však můžeme obecně popsat jako vrstvu transformace dat, databázovou vrstvu, analytickou vrstvu, prezentační vrstvu, kde všechny uvedené části zastřešuje vrstva „know how“. Ilustraci této struktury nám znázorňuje následující Obrázek 7. (Novotný, 2005)

4.2.1 Zdrojové systémy

Tyto systémy se můžou označovat také jako „primární“ nebo „transakční“. Jak už Obrázek 7 napovídá, tak zdrojové systémy nepatří do aplikací business intelligence, avšak slouží jako zdroj těchto aplikací. Primární systémy pracují v reálném čase a jsou koncipovány tak, aby podporovaly modifikaci a ukládání dat, nemohou však sloužit analytickým úlohám. Jejich složení je rozmanité, nemusí se jednat pouze o vnitřní systémy dané firmy, ale zdrojem dat mohou být i externí systémy. Příkladem externích zdrojových systémů je například veřejný telefonní seznam, databáze Českého statistického úřadu a dalších. (Novotný, 2005)

4.2.2 Vrstva transformace dat

Vzhledem k různorodosti zdrojů a dat, je naprosto klíčové data analyzovat, vybrat významná a rozhodující a v posledním kroku je sjednotit, a to ještě před uložením do datového skladu. To je úkolem vrstvy zabývající se transformací a následnou integrací dat. Jejími hlavními pomocníky jsou ETL, neboli „Extraction, Transformation and Loading“ a EAI, což je zkratka „Enterprise Application Integration“. Obě si více objasníme. (Novotný, 2005)

ETL

Tento systém je označován také jako „datová pumpa“ a je naprosto nepostradatelný v celém systému BI. Prvním krokem při práci datové pumpy je získat a vybrat data ze zdrojových systémů, ty se následně upravují a čistí a v posledním kroku se nahrávají do struktur data warehousu. ETL nepracuje v reálném čase, ale v určitých intervalech. Intervaly bývají většinou denní, týdenní a měsíční. (Schiller, 2003)

EAI

„Enterprise Application Integration“ je dalším důležitým prvkem ve vrstvě transformace dat. Zabývá se primárně integrací firemních systémů na úrovni dat i aplikací. Na datové úrovni nejde jen o integraci, ale také o distribuci, a na aplikační úrovni se jedná především o distribuci určitých funkcí systému. Z EAI putují data do datových tržišť a celý proces probíhá v reálném čase. (EAI (enterprise application integration) definition, 2009)

4.2.3 Databázová vrstva

Tato vrstva, jak už název prozrazuje, obsahuje databázové komponenty. Ty se starají o ukládání dat, aktualizaci a správu. Mezi databázové komponenty patří datový sklad a data marty popisované v kapitole 2, a dále operativní datové úložiště zkratkou ODS, a dočasné datové úložiště uváděné zkratkou DSA.

ODS

Operativní datové úložiště je jednotné místo integrace dat ze zdrojových systémů, kde sledování dat sice neprobíhá v reálném čase, ale velmi se tomu blíží. Vyznačuje se krátkou dobou odezvy a zpravidla slouží jako centrální databáze základních číselníků, to mohou být například číselníky produktové. Data v tomto úložišti zůstávají jen do doby nahrání nových, tudíž zde neexistuje žádná historizace, na druhou stranu data obsažená v ODS jsou sjednocená, konzistentní a subjektově orientovaná. (Operational Data Store, 2013)

DSA

Do dočasného úložiště dat se dostávají extrahovaná, ale nekonzistentní data ze zdrojových systémů, u nichž neproběhla transformace ani agregace. Stejně jako v předchozím případě zůstávají jen do doby nahrání nových, proto ani zde není žádná historizace. Struktura těchto dat je naprosto totožná s tou, v jaké byly uloženy ve zdrojových systémech. DSA nepatří mezi povinné prvky řešení BI a používá se v případě potřeby data před zpracováním konvertovat do databázového formátu. (DW Staging area, 2008)

4.2.4 Analytická vrstva

Zde jsou obsaženy komponenty sloužící přímé analýze dat uložených v databázové vrstvě. Mezi prvky přímé analýzy patří OLAP, rozebraný v kapitole 3, reporting a data mining, přeloženo jako dolování dat. Reporting je spojený s dotazováním do databází, například SQL příkazy, pomocí klasických rozhraní. Reporty mohou být buď standardní, spouštěny v periodách, nebo ad hoc reporty, které probíhají jednorázově a bývají zadávány uživatelem.

Data Mining

V českém překladu dolování dat, je značně propracovaný proces, který pomocí speciálních algoritmů umožňuje objevovat v datech další souvislosti a informace. Tento proces probíhá nad velmi rozsáhlými databázemi extrahováním relevantních informací, které nejsou předem definované. (Pilař, 2006)

Je mnoho typů nástrojů pro data mining, ale všechny splňují stejnou vlastnost a tou je analýza odvozená z obsahu předem nespecifikovaných dat. Úkolem dolování je objevování nových faktů, pomáhající vedoucím pracovníkům rozpoznat skryté vztahy mezi proměnnými se kterými společnost pracuje na základě matematických a statistických technik. Jejich cílem je poskytovat informace co možná nejširšímu okruhu pracovníků, a proto jsou tyto techniky z velké části realizovány automaticky podle stanovených algoritmů, není tedy nutnost odborných znalostí statistiky, a analyzovat je může i vedoucí pracovník vypracovávající následné reporty. Některé z těchto technik si zde uvedeme. (Pilař, 2006)

Neuronové sítě

Existuje celá řada obměn neuronových sítí, liší se od sebe algoritmy používanými pro tvorbu modelů, které poskytují možnou predikci. Jsou založeny na systému neuronů, napodobující způsob chování a organizaci lidského mozku. Algoritmy obsažené v těchto neuronových sítích jsou schopny se samy učit, a tím i lépe odhalovat případné skryté vazby, a obvykle se používají pro vytváření modelů poskytujících predikci. (Civín, 2007)

Clustering a klasifikace

Technika clusteringu a klasifikace slouží k identifikování a charakterizování různých segmentů v datech. Umožňuje především rozdělovat a klasifikovat data s podobnými charakteristikami a definovat důležité atributy skupin ve formě klasifikačních kritérií. (Lacko, 2003)

Rozhodovací stromy

Jsou velmi oblíbenou a častou technikou, především protože znázornění a interpretace výsledků je velmi snadná. Tento komplexní model zobrazuje data ve formě stromu, v němž každý uzel udává kritérium pro další dělení dat do jednotlivých větví. To znamená, že se všechna zdrojová data rozdělují do segmentů, ve kterých každý list odpovídá segmentu definovanému předchozími uzly. Každý segment obsahuje data, která se vyznačují stejnými vlastnostmi a výhodou rozhodovacích stromů je používání většího množství algoritmů. (Lacko, 2003)

4.2.5 Prezentační vrstva

Je zaměřena na komunikaci mezi jednotlivými komponentami řešení business intelligence a uživateli, pracujícími právě s prvky prezentační vrstvy. Zajišťuje sběr požadavků na analytické operace a dále pak zobrazení výsledků. Výsledky mohou být prezentovány v různých analytických aplikacích, nebo v systémech ESI, což je zkratka Executive Information Systems.

Executive Information Systems

Jsou podnikové informační systémy sloužící jako podpora řízení. Cílem systémů EIS je podpora procesů používaných koncovými uživateli, což jsou například různé podnikové analýzy, plánování, nebo rozhodování. Oproti reportingu, který má za úkol především poskytovat pravidelné reporty, které jsou stavěny přímo nad daty z datového skladu, vytvářejí

systémy EIS multidimenzionální vrstvu a jejím prostřednictvím přistupují k datům analýz. Definice Executive Information Systems je následující. (Novotný, 2005)

„Manažerské aplikace EIS jsou typem aplikací, které v sobě integrují všechny nejdůležitější datové zdroje systému, významné pro řízení organizace jako celku. S tím jsou spojeny i specifické nároky na prezentace informací a jejich zpřístupnění vedoucím pracovníkům firmy. EIS je tak především analytický a prezentační nástroj.“ (Novotný, 2005, str. 34)

Tyto aplikace umožňují uživatelům online přístup k informacím ve formě, která je jednoduchá a srozumitelná. Jsou totiž navrhovány tak, aby i uživatelé s menšími znalostmi v oblasti používání počítačů, tyto systémy využívali bez jakýchkoliv problémů. (Novotný, 2005)

5 Strategie budování datového skladu

Při návrhu datového skladu musíme vybrat vhodnou metodu pro jeho tvorbu, to je pravděpodobně nejdůležitější krok. Při budování musíme počítat s možnými potížemi, které se při této tvorbě nevyhnutelně objeví. Z hlediska metody budování můžeme volit z několika možností, nejpoužívanější je metoda „Big bang“, neboli metoda velkého třesku, a metoda přírůstková. Níže rozvedeme výhody a nevýhody každé z nich. (Lacko, 2003)

Metoda „Big bang“

Jedinou nespornou výhodu, kterou tato metoda nabízí je, že si můžeme celý projekt naplánovat ještě před jeho začátkem. Je tomu tak proto, že počítá s tím, že celý datový sklad vybudujeme během jednoho jediného projektu. Tento přístup však není nejšťastnější, protože vývoj datového skladu je dynamický proces, nedá se tedy vyřešit najednou, a je časově velmi náročný. Jeho největší nevýhodou je právě časová náročnost, a to jak z důvodu zastarání plánovaných technologií, tak i změn požadavků uživatelů. Metoda velkého třesu se skládá z několika etap, a sice (Lacko, 2003):

- Analýza požadavků
- Vytvoření datového skladu
- Vytvoření přístupu do datového skladu

Metoda přírůstková

Tato metoda předpokládá budování datového skladu po jednotlivých etapách, tedy přibývání jednotlivých přírůstkových řešení, které jsou ve výsledku součástí celkové architektury datového skladu. Začíná se vytvářením několika dílčích částí, například datových trhů, které se následně poskytnou uživatelům. To slouží k otestování funkčnosti a v případě bezproblémového fungování je možné tuto část implementovat jako prvek datového skladu. Tento proces se opakuje, samozřejmě s jinými prvky až do vytvoření kompletního data warehouse. (Lacko, 2003)

Nespornou výhodou této metody je, že datový sklad je neustále přizpůsobován potřebám uživatelů. Oproti metodě velkého třesku, kde realizace probíhá v rámci měsíců až let, je tato metoda z hlediska jednotlivých částí funkční již po relativně krátké době.

6 Analytická část

V praktické části bakalářské práce budu provádět analýzu jednotlivých platforem datových skladů na základě bodového hodnocení zvolených specifických atributů na fiktivní společnosti. Jako fiktivní společnost je zvolena společnost Satax s.r.o, podnikající na trhu mobilních aplikací. Portfolio této firmy zahrnuje aplikace zaměřené na bezpečnost, zábavu a dále údržbu mobilních zařízení napříč celým spektrem mobilních platforem. Zákazníci si aplikace mohou zakoupit a stáhnout z aplikačních marketů, které jsou přímo integrované v mobilním zařízení v závislosti na konkrétním výrobci daného zařízení. Společnost Satax s.r.o zpracovává a eviduje data o zákaznících a objednávkách v informačním systému, který ukládá veškerá data do relační databáze.

Společnost se rozhodla zlepšit aktuální kvalitu svých dat, konsolidovat je a lépe využívat pro přesnější a účinnější zacílení svých marketingových kampaní, ad-hoc reportů a statistik. Proto je nutné provést analýzu dostupných řešení, jejíž výsledek bude realizace nejlépe hodnocené varianty. Analýza vhodných řešení bude prováděna na základě výsledků společností dlouhodobě provádějících výzkum a následné hodnocení v oblasti data warehousingu, konkrétně se jedná o společnosti Gartner a Forrester uznávané autority v tomto oboru. Gartner vydal výsledky své analýzy v článku „Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics“ (2015) a Forrester v obdobné práci „The Forrester Wave: Enterprise Data Warehouse, Q4 2013“ (2013).

Firma nechce data warehouse stavět na službách založených na cloud technologiích, je pro ni velmi důležitá možnost budoucího rozšíření funkcí v podobě nových upgradů, technologie Hadoop umožňující zpracovávání dat velikostí pentabytů až exabytů. Dále chce integrovaný systém pro ukládání a přístup k datům podporující analytické funkce a SQL funkcionalitu.

6.1 Gartner & Forrester

Tyto celosvětově uznávané společnosti provádí každý rok analýzy a hodnocení nástrojů a systémů, které operují s datovými sklady, a výsledné hodnocení zobrazují v intuitivních grafech. Gartner v tak zvaném magickém čtyřúhelníku, viz Obrázek 8 a Forrester v obdobném grafu.

Výzkumný soubor data warehouse produktů vybraly na základě určitých kritérií:

- Do vzorku byly zařazeny pouze produkty v plné verzi a neobsahují tedy beta ani jiné neúplné verze. Zároveň se vybralo to nejaktuálnější vydání.
- Vybraní prodejci museli mít database management system software, který byl obecně součástí licence nebo bylo dostupné stažení zhruba jeden rok.
- Dodavatel musel poskytovat podporu pro datový sklad a database management system. Zároveň musel prokázat svou schopnost poskytovat potřebné služby pro podporu datového skladu prostřednictvím zřízení a poskytování podpůrných procesů.

Výzkum probíhal v podobě provozního režimu a podrobil datové sklady komplexní analýze, která odhalila všechny slabiny a přednosti daného provedení. Výsledkům a charakteristikám jednotlivých společností pak odpovídá konkrétní umístění v grafu.

Osa Ability to execute udává vyspělost produktu a dodavatele. Kritéria pod tímto názvem také hodnotí flexibilitu produktů, zda jde o produkt nebo službu, a jeho schopnost běhu v různých provozních prostředích (dává zákazníkovi řadu možností), a tím pokrýt různé požadavky trhu. Roli zde hrají prvky jako je finanční stabilita, investice do výzkumu a vývoje, celkové vedení společnosti a poměr cena/výkon.

Obrázek 8 – Srovnání řešení datových skladů Gartner

Zdroj: Gartner, 2017.

Osa completeness of vision hodnotí schopnosti dodavatele chápat funkce potřebné k vytvoření produktové strategie splňující požadavky trhu, chápe celkové trendy na trhu. Posuzuje marketingovou a obchodní strategii, business model a hlavně inovace. To vše je nezbytné pro dlouhodobou životaschopnost produktu a společnosti. Graf je dále členěn do čtyř kvadrantů, které blíže charakterizují kvalitu jednotlivých produktů.

Do kvadrantu Leaders se řadí společnosti jejichž řešení datových skladů jsou komplexní a jsou schopny pokrýt poptávku celého trhu. Nabízí velké množství nástrojů pro práci s daty jako data mining, nebo propracované nástroje pro analýzu a pokročilou správu dat. Velmi pružně reagují na měnící se poptávku trhu a díky investicím do inovací a jasným vizím jsou v

dobré pozici do budoucna. Do Challengers spadají společnosti mající stejně technicky vyspělá řešení s tím rozdílem, že nedostatečně investují do inovací a nemají jasné vize dalších období a nebo špatně reagují na poptávku trhu. Visionaries velmi dobře chápou a mají velmi dobře analyzovanou poptávku trhu, vědí kam se bude trh ubírat a mají pokročilé vize ale technické řešení zaostávají za produkty z kategorií Leaders a Challengers. Poslední kategorií jsou Niche players, zde se umísťují především začínající společnosti s nejasnými představami do budoucna a s méně vyspělými technologiemi. Stejně tak i společnost Forrester ve svém grafu zobrazuje výsledky analýz datových skladů a to také ve čtyřech částech.

Obrázek 9 – Srovnání řešení datových skladů Forrester

Zdroj: Forrester, 2017.

Obdobně jako u Gartner nejlépe hodnocené řádí do skupiny Leaders, ti mají nejkomplexnější systémy a cílí na poptávku celého trhu a nabízí velké množství nástrojů pro práci s daty. Gartnerovským Challengers zde odpovídá část zvaná Strong Performers mající velmi propracovaná technická řešení, ale nejsou schopni pokrýt celou poptávku trhu a mají spíše specificky zaměřené produkty pro konkrétní sektor, jako například zdravotnictví. Contenders je sekce obsahující spíše začínající společnosti v oblasti datových skladů, nemají jasné vize pro příští období a jejich produkty technicky velmi výrazně zaostávají za ostatními.

6.2 Analýza produktů

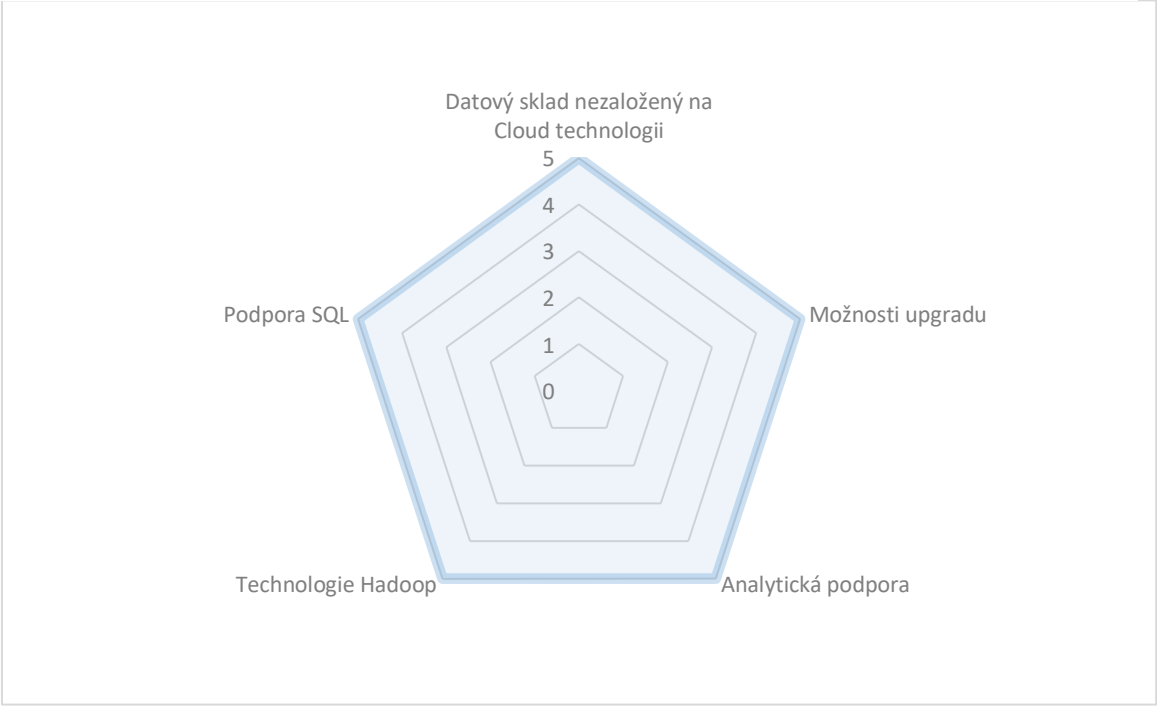
Analyzovat budeme produkty ze segmentu Leaders z důvodu selekce řešení s nejlepším hodnocením podle Gartner a Forrester. Specifikovaná kritéria budou u jednotlivých produktů bodově klasifikována, a to pěti body za splnění a jedním bodem za nesplnění.

6.2.1 HP

Základem nabídky HP je produkt s názvem Vertica. Ten je postaven na samostatném analytickém DBMS, poskytujícím konektory pro propojení s technologií Hadoop, a podporuje pokročilé analytické nástroje, a to včetně těch prediktivních. Vertica je dodávána jako software pro standardní platformy, kromě systému Windows. Pro tento software HP dále nabízí předdefinovanou certifikovanou konfiguraci HP Factory Express.

HP Vertica umožňuje rozšířit svou základní platformu v rámci správy dat, analytiky a možností nasazení. Tento produkt je možné požívat i pro nestrukturovaná data, jako jsou data ze sociálních médií, multimediální data, strojová data a dokumenty. HP Vertica dále poskytuje díky technologii Hadoop velmi rychlé odpovědi na dotazy do databáze, pokročilé komprese dat, podporu pro hybridní úložiště, shared-nothing architekturu, indexování, partitioning a systém pro optimalizace dotazů a rozdělování pracovního zatížení. Podporovány jsou také strukturované dotazovací jazyky (SQL). Podle uživatelských recenzí však ne všechny zmíněné funkce fungují bezproblémově. Problémy se vyskytly s optimalizacemi dotazů a se stabilitou celého systému.

Graf č. 1 – Hodnocení HP

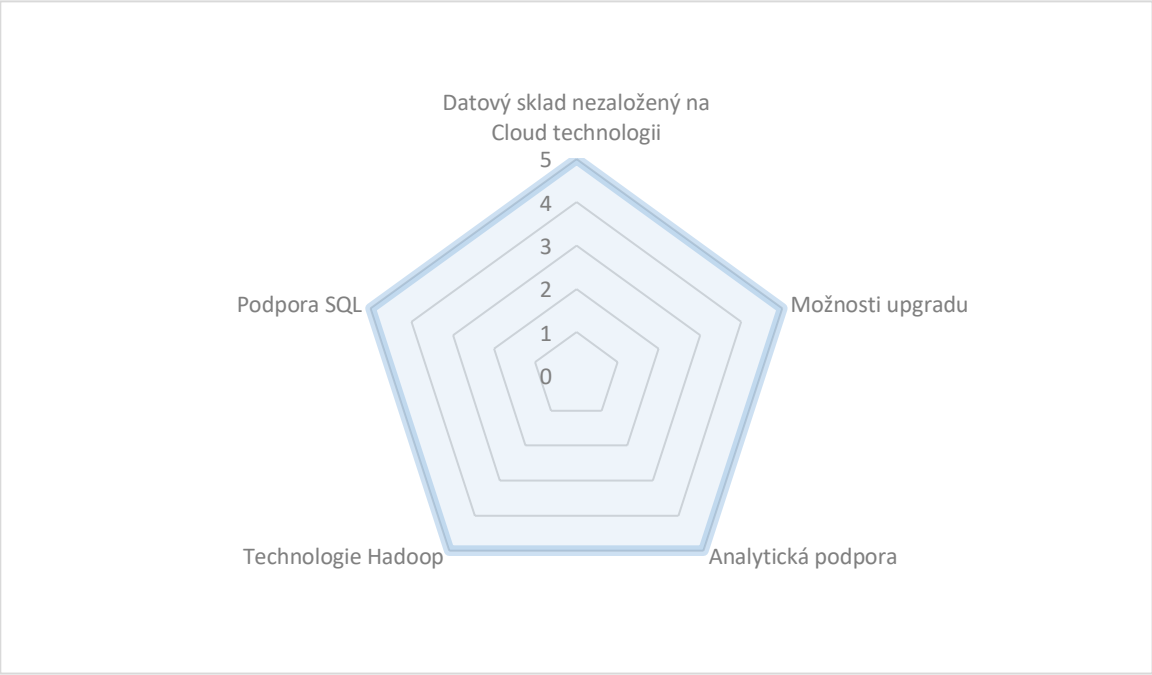


6.2.2 IBM

IBM nabízí jako základ svého datového skladu samostatné DBMS řešení zvané IBM DB2, které je rozšiřováno o další technologie. DB2 disponuje technologií BLU Acceleration, jejímž základem je dynamická sloupcová paměťová architektura, paralelní vektorové zpracování dat a podporuje také SQL. Tato technologie umožňuje snížit nároky na velikost datového úložiště. Další nespornou výhodou je možnost optimalizace přístupu k datům bez nutnosti ladění databáze a indexování. Ve svém portfoliu také nabízí cloudové služby zvané dashDB, umožňující řešení datového skladu v cloudu. Díky širokým možnostem integrace nástrojů pro provozní analýzu v reálném čase a prediktivní analýzu dat, jako je PureData, IBM InfoSphere Optim, Cognos a WebSphere, tedy přímo konkuruje největším společnostem v rámci tohoto odvětví, kterými jsou Oracle, Microsoft a Teradata. Z hlediska funkce, řešení IBM podporuje pokročilé komprese dat, vertikální datové modely, automatizované řízení zdrojů a má nativní podporu pro MapReduce a Hadoop.

Dle recenzí je pro stávající zákazníky IBM složité najít správnou cestu od klasických datových skladů ke cloudovým technologiím. Byť tyto inovativní metody odpovídají na

Graf č. 2 – Hodnocení IBM



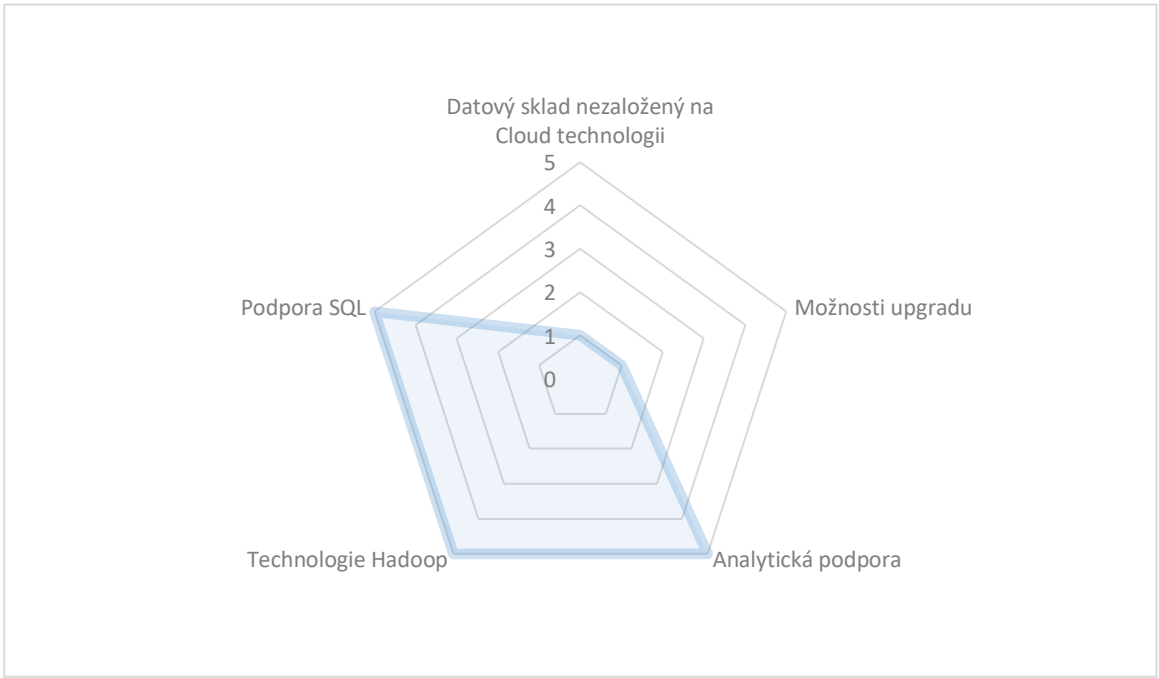
vzniklé požadavky pro lepší zprávu uchovávaných dat.

6.2.3 Amazon Web Services

Společnost Amazoon Web Services je jedním z průkopníků v oblasti řešení datových skladů do velikostí řádů desítek pentabytů. V svém portfoliu nabízí produkty, jako jsou: Amazon Redshift data warehouse v cloudu s podporou dotazování pomocí SQL. Amazon Elastic MapReduce technologie umožňující zpracování dat velikostí až desítek pentabytů. Amazon Simple Storage Service S3, což je rozsáhlé souborové úložiště členěné do stromové struktury a dále dotazovací službou Amazon Athena, která umožňuje pomocí dotazovacího jazyka prohledávat stromovou strukturu a obsah souborů v ní uložených.

Amazon Web Services je jedním z dominantních společností řešících problematiku datových skladů pomocí cloudových technologií. Jako výhodu tohoto řešení můžeme považovat přesné dimenzování konkrétních produktů, které chce společnost využívat a tím i konkrétnější a přesnější cenovou kalkulaci. Produkty i velikost datového úložiště v cloudu lze dále rozšiřovat, ale ovšem za cenu toho že firemní data nejsou fyzicky uložena na severech společnosti, ale jsou ukládána a zpracovávána mimo společnost na serverech poskytovatele služeb. Šířka záběru tohoto řešení v rámci kapacity datového úložiště umí poskytnout optimální řešení jak pro malé firmy, tak také pro firmy pracující s daty o velikosti od stovek terabytů až po desítky petabytů. Amazon využívá technologii Hadoop pomocí jednoduše konfigurovatelného clustru Amazon EC2, tato technologie využívá Apache Hadoop. Problém této technologie může spočívat jednak v upgradu na nejnovější technologické novinky z důvodu smluvních, kde může být možnost upgradu omezena nebo limitována, a dále pak složitější kontrolou nad poskytovanými technologiemi.

Graf č. 3 – Hodnocení Amazon Web Services

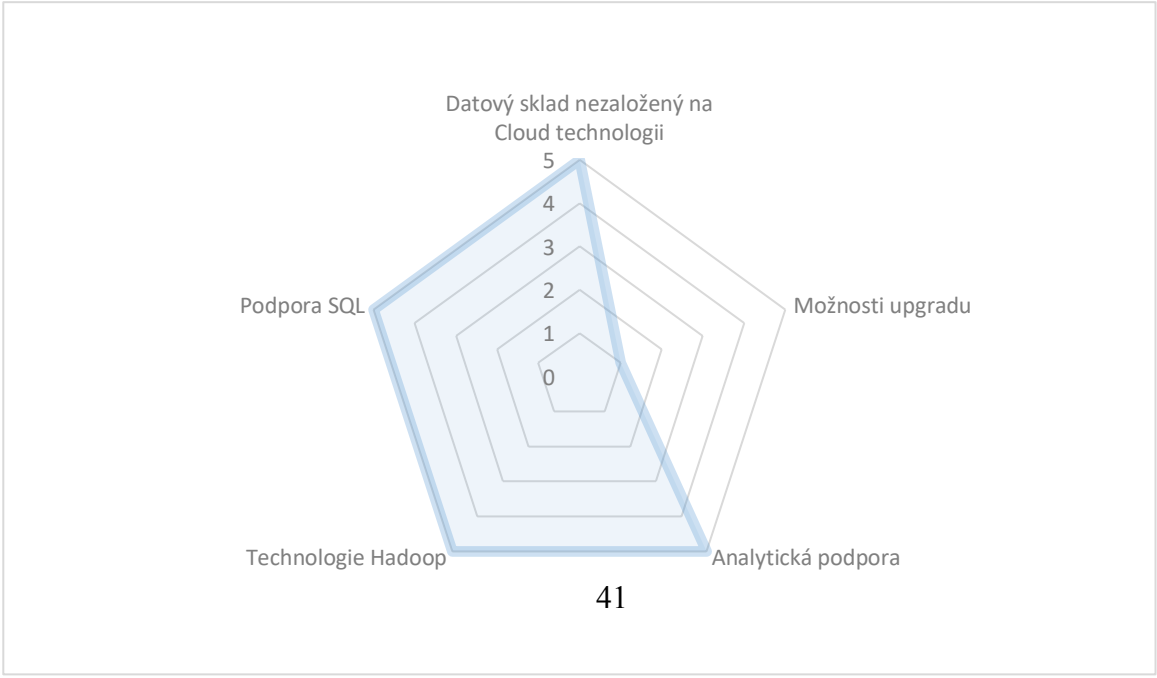


6.2.4 SAP

SAP nabízí dva produkty, které mohou tvořit základ data warehousu, jsou to SAP IQ a SAP Hana. SAP IQ byl první sloupcové úložiště DBMS. IQ je dostupný jako samostatný DBMS a nebo ve formě OEM skrze systémové integrátory. Jednou z klíčových vlastností SAP HANA je umístění celé databáze pouze v operační paměti serveru. Data se již z disků průběžně nenačítají a ani se na ně nezapisují, což zásadně zrychluje přístup k datům a veškeré operace s nimi. Další stěžejní vlastností SAP HANA je způsob ukládání informací. Na rozdíl od běžných SQL databází, se primárně ukládají sloupcově a nikoli řádkově, byť řádkové ukládání SAP HANA také podporuje. Díky podpoře analýz v reálném čase je možné docílit velmi rychlé odezvy na dotazy, jejichž výpočet by na běžné SQL databázi trval mnohonásobně déle. Dále je zaměřena na shared-nothing architekturu, optimalizovaný streaming dat, cache-koherentní programování pro optimalizaci mezipaměťového zpracování, analytiku v reálném čase a pokročilé komprese dat.

Technologie SAP Sybase Replication Server poskytuje vysokorychlostní přenos dat. Pro podniky, které chtějí využít datový sklad na cloud platformě, je SAP Hana a SAP IQ k dispozici jako služba na Amazon Web Services. SAP stejně jako jeho konkurence podporuje technologii Hadoop.

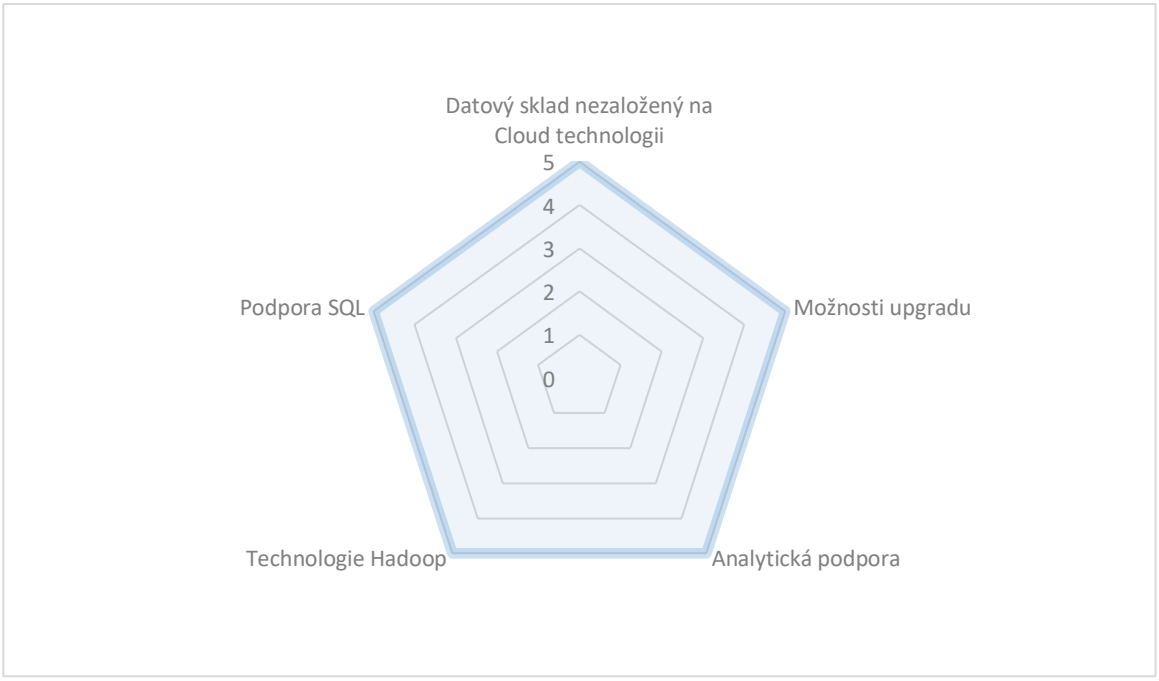
Graf č. 4 – Hodnocení SAP



6.2.5 Microsoft

Microsoft na trhu s datovými sklady uvedl jako základ své nabídky Microsoft SQL Server, Microsoft Analytics Platform, který kombinuje SQL Server Parallel Data Warehouse a HDInsight. V neposlední řadě Azure HDInsight pro Hadoop. Microsoft patří mezi nejznámější a nejrozšířenější společnosti v oblasti databázových produktů, datových skladů a Business intelligence. Microsoft mezi konkurencí vyniká především technologiemi OLAP a tabulkovými procesory jako například PowerPivot, sloužícími pro analytiku dat.

Graf č. 5 – Hodnocení Microsoft



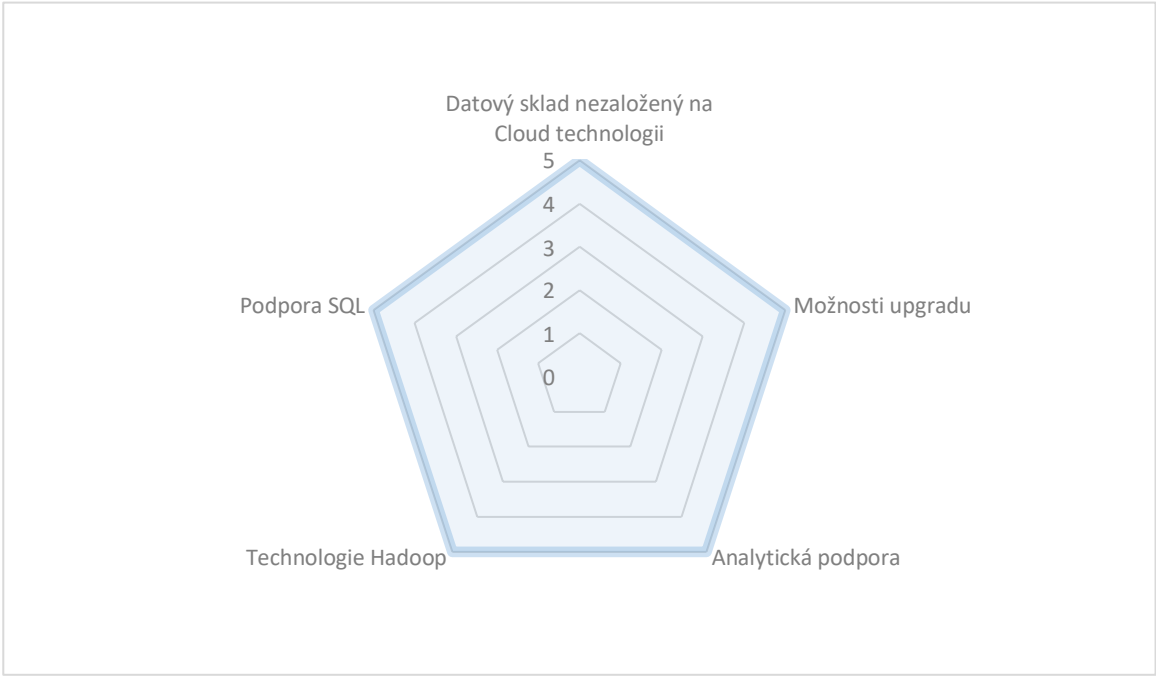
Microsoft SQL Server podporuje různé druhy OLAP systémů, aplikace pro Business intelligence, má nástroje pro prediktivní analytiku a analytické vzory pro analýzu v reálném čase. Microsoft neustále zlepšuje a inovuje svá řešení datových skladů v oblasti ukládání do mezipaměti, běhu databáze v operační paměti, do sloupcovitého ukládání dat, indexování, do pokročilých kompresí a integrace dat, a do analytiky v reálném čase.

6.2.6 Oracle

Oracle má dominantní postavení na trhu s databázemi, jeho rostoucí množství řešení datových skladů a rozsáhlé portfolio aplikačních a middleware řešení dává velkou konkurenční výhodu. Oracle poskytuje své DBMS spolu s technologií Hadoop ve více variantách, souhrnně pojmenovaných jako Oracle Big data Appliance. Zákazníci si z nabídky Oracle mohou vybrat a vytvořit vlastní sklad z certifikovaných konfigurací produktů na Oraclem doporučeném hardwaru. To umožňuje používat databázový software přímo na hardwaru od Oracle zvaném Exadata. Cloudové služby, jsou stejně jako je tomu u společnosti SAP, k dispozici na Amazon Web Services. Exadata využívá inteligentní datové úložiště, hybridní sloupcové komprese, paralelní databázovou analýzu, inteligentní datové cache, SSD technologie a hardwarovou podporu pro analýzu a virtualizaci dat v paměti.

Mimo jiné, Oracle poskytuje optimalizované indexování, flexibilní partitioning a optimalizace dotazů, což z něj dělá vysoce výkonný data warehouse. Na základě finančních investic do výše zmíněných technologií, Oracle pokračuje v rozšiřování nabídky svých

Graf č. 6 – Hodnocení Oracle

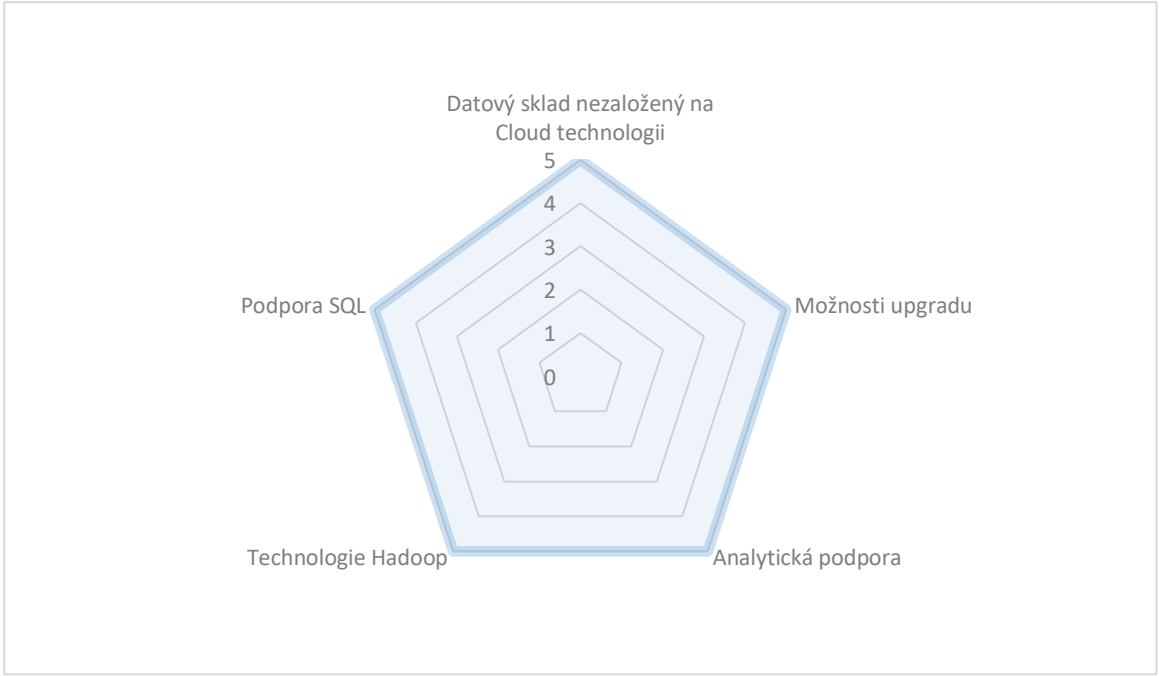


produktů.

6.2.7 Teradata

Teradata operuje na trhu datových skladů již více než 30 let. Nabídka Teradaty zahrnuje jak klasické databázové systémy licencí, tak nabízí také řešení datového skladu prostřednictvím cloudu. Tradiční řešení datového skladu Teradata je známé pod názvem Unified Data Architecture (UDA). Svým zákazníkům nabízí v rámci klasických skladů kombinaci speciálně navrženého hardwaru a databázového softwaru poskytujícího prostředky pro analytiky, Aster databáze a technologii Hadoop, certifikované integrace produktů a širokou škálu partnerských aplikací. Teradata nabízí nejkomplexnější a nejvíce škálovatelnou platformu datových skladů. Propracované funkce pro hloubkovou analýzu databáze, pokročilé komprese dat, partitioning, indexování, optimalizace pracovního vytížení a optimalizace dotazů do databáze z ní dělají téměř nepřekonatelnou konkurenci. Teradata disponuje konektory pro import a export dat z Hadoop a dále také SQL-H a SQL-MapReduce ve dvou variantách pro provádění dotazů. Teradata nabízí různé konfigurace systému, a tím poskytuje velkou flexibilitu nejen v přizpůsobitelnosti produktu, ale také ve stanovení optimální ceny.

Graf č. 7 – Hodnocení Teradata



6.3 Diskuze

Výsledné hodnocení jednotlivých společností ukazuje velkou vyrovnanost na trhu datových skladů, viz Tabulka 1. Téměř všechny analyzované společnosti splnily požadavky fiktivní firmy Satax s.r.o, která aktuálně eviduje data o zákaznících a objednávkách v informačním systému, jenž ukládá veškerá data do relační databáze.

Tabulka 1- Výsledné bodové hodnocení

	Datový sklad nezaložený na Cloud technologii	Možnosti upgradu	Analytická podpora	Technologie Hadoop	Podpora SQL	Celkové hodnocení
Teradata	5	5	5	5	5	**25**
Oracle	5	5	5	5	5	**25**
Microsoft	5	5	5	5	5	**25**
IBM	5	5	5	5	5	**25**
HP	5	5	5	5	5	**25**
SAP	5	1	5	5	5	**21**
Amazon Web Services	1	1	5	5	5	**17**

Satax klade důraz na tradiční řešení datového skladu a nechce využívat cloudové služby, jež také umožňují realizaci datového skladu. Dále vyžaduje podporu technologie Hadoop z důvodu možného rozšíření společnosti a zvětšení datové základny, a dále podporu SQL dotazování. V těchto bodech vyhovělo všech sedm analyzovaných společností. V neposlední řadě bylo požadováno, aby existovaly možnosti rozšíření datového skladu o další funkce, jako například Machine learning, jenž je součástí prediktivní analýzy pro zlepšování kvality svých aplikací. V tomto bodě nevyhověly společnosti SAP a Amazon Web Services, jenž neposkytují dostatečné možnosti pro upgrade datového skladu o nové funkce. Vzhledem ke shodnosti výsledků hodnocených kriterií u společností Teradata, Oracle, Microsoft, IBM a HP

bude konečná volba jedné ze společností záležet primárně na cenách realizace jejich řešení. To musí být vykomunikováno s obchodním oddělením každého poskytovatele.

7 Závěr

Čím dál větší množství společností má potřebu využívat svá data pro lepší zacílení svých marketingových kampaní a také jako oporu pro rozhodování. Tuto výsadu měli dříve jen manažeři velkých společností. Tyto společnosti svá data spravovaly v datových skladech, jenž byly součástí business intelligence, která umožňovala těžit z dat potřebné informace. Tato skutečnost dnes není výsadou jen velkých společností, dostupná jsou již i technologická řešení pro malé a střední firmy, které chtějí využívat těchto technologií pro zlepšení a zvýšení konkurenceschopnosti.

Primárním cílem této práce bylo analyzovat možnosti realizace datového skladu podle různých poskytovatelů těchto řešení. Jako uživatel datového skladu byla zvolena fiktivní společnost Satax, která si na základě svých požadavků stanovila kritéria výběru vhodného řešení datového skladu. Tato kritéria byla aplikována na informace od nestranných společností dlouhodobě provádějící praktický výzkum, analýzu a testování datových skladů od distributorů těchto řešení, a dále na informace poskytované jednotlivými společnostmi v rámci možností a technických specifikací svých produktů. Tato kritéria byla hodnocena základní bodovou škálou a výsledné hodnocení zobrazeno ve formě tabulky, viz Diskuze. V předcházejících kapitolách byla rozebírána také technologie datových skladů jako základ systémů business intelligence.

8 Seznam literatury

A Data Mart Concepts. 2007. *Oracle® Business Intelligence Standard Edition One Tutorial Release 10g* [online]. © 1979, 2007, Oracle [cit. 2015-11-29]. Dostupné z: https://docs.oracle.com/cd/E10352_01/doc/bi.1013/e10312/dm_concepts.htm

ANUPINDI, Nagesh V. 2005. Inmon vs. Kimball - An Analysis. *Nagesh.com: Publications* [online]. © 2015 Nagesh V. Anupindi, 2005, s. 6 [cit. 2015-11-29]. Dostupné z: http://www.nagesh.com/publications/technology/173-inmon-vs-kimball-an-analysis.html

BEYER, Mark A. a Roxane EDJLALI. 2017. Magic Quadrant for Data Management Solutions for Analytics. *Gartner* [online]. Gartner, Inc. © 2017 Gartner, s. 1-36 [cit. 2018-03-11]. Dostupné z: https://myleadcorner.files.wordpress.com/2017/07/magic-quadrant-for-data-management-solutions-for-analytics-feb-2017.pdf

Corporate Information Factory: About Bill. 2007. *Corporate Information Factory* [online]. Castle Rock: Inmon Consulting Services © 2007, 2007 [cit. 2015-11-29]. Dostupné z: http://www.inmoncif.com/about/

DANISH, Wahab. 2010. Indroduction to Data warehousing. *Database Centre* [online]. [cit. 2015-11-29]. Dostupné z: http://database-centre.blogspot.cz/2010/06/introduction-to-data-warehousing.html

Database Centre: Introduction to Data warehousing [online]. 2010. 13.6.2010 [cit. 2015-11-29]. Dostupné z: http://database-centre.blogspot.cz/2010/06/introduction-to-data-warehousing.html

Designing Star Schema: Star Schema: General Information. 2015. *LearnDataModeling.com: Tutorial on Data Modeling, Data Warehouse & Business Intelligence!* [online]. [cit. 2015-11-29]. Dostupné z: http://learndatamodeling.com/blog/designing-star-schema/

DW Staging area. 2008. *Data-Warehouses* [online]. [cit. 2015-11-29]. Dostupné z: http://data-warehouses.net/architecture/staging.html

EAI (enterprise application integration) definition. 2009. *TechTarget* [online]. [cit. 2015-11-29]. Dostupné z: http://searchsoa.techtarget.com/definition/EAI

Easy OLAP Definition. 2015. *OLAP* [online]. [cit. 2015-11-29]. Dostupné z: http://olap.com/olap-definition/

GÁLA, Libor, Jan POUR a Prokop TOMAN. 2006. *Podniková informatika: počítačové aplikace v podnikové a mezipodnikové praxi, technologie informačních systémů, řízení a rozvoj podnikové informatiky*. 1. vyd. Praha: Grada, 482 s. Management v informační společnosti. ISBN 80-247-1278-4.

Gartner IT Glossary: Business Intelligence (BI). 2013. *Gartner.* [online]. © 2013 Gartner [cit. 2015-11-29]. Dostupné z: http://www.gartner.com/it-glossary/business-intelligence-bi/

GRIMES, Seth. 2008. BI at 50 Turns Back to the Future. *InformationWeek* [online]. [cit. 2015-11-29]. Dostupné z: http://www.informationweek.com/software/information-management/bi-at-50-turns-back-to-the-future/d/d-id/1073576?

How Markets and Vendors Are Evaluated in Gartner Magic Quadrants. 2014. *Gartner.* [online]. © 2014 Gartner [cit. 2015-11-29]. Dostupné z: https://www.gartner.com/doc/2804921?ref=SiteSearch&sthkw=magic%20quadrant%20graph&fnl=search&srcId=1-3478922254

INMON, William H. 2002. *Building the data warehouse*. 3rd ed. New York: J. Wiley, xx, 412 p. ISBN 04-710-8130-2.

Introduction to Cubes. 2002. *Microsoft: TechNet* [online]. © 2015 Microsoft [cit. 2015-11-29]. Dostupné z: https://technet.microsoft.com/en-us/library/aa216365%28v=sql.80%29.aspx

KAY, Russell. 2004. Data Cubes. *Computerworld* [online]. © 1994 - 2015 Computerworld, Inc [cit. 2015-11-29]. Dostupné z: http://www.computerworld.com/article/2564238/business-intelligence/data-cubes.html

KIMBALL, Ralph a Margy ROSS. c2010. *The Kimball Group reader: relentlessly practical tools for data warehousing and business intelligence*. Indianpolis, IN: Wiley, xxiv, 718 p. ISBN 978-0-470-56310-6.

LACKO, Ľuboslav. 2011. *1001 tipů a triků pro SQL*. Vyd. 1. Brno: Computer Press, 416 s. ISBN 978-80-251-3010-0.

Major OLAP Technology Types. 2015. *OLAP* [online]. [cit. 2015-11-29]. Dostupné z: http://olap.com/types-of-olap-systems/

Model Rational Insight Data Warehouse. 2010. *IBM Knowledge Center* [online]. [cit. 2015-11-29]. Dostupné z: http://www-01.ibm.com/support/knowledgecenter/SSRL5J_1.1.0/com.ibm.rational.raer.help.doc/topics/c_datawarehousemodel.html?lang=cs

NOVOTNÝ, Ota, Jan POUR a David SLÁNSKÝ. 2005. *Business intelligence: jak využít bohatství ve vašich datech* [online]. 1. vyd. Praha: Grada, 254 s. Management v informační společnosti. ISBN 80-247-1094-3.

Operational Data Store. 2013. *Gartner.* [online]. [cit. 2015-11-29]. Dostupné z: http://www.gartner.com/it-glossary/ods-operational-data-store

PANEC, Zdeněk. 2003. Co je to Business intelligence? *IT Systems* [online]. © CCB 2015, 2003(6) [cit. 2015-11-29]. ISSN 1802-615X. Dostupné z: http://www.systemonline.cz/clanky/co-je-to-business-intelligence.htm

PILAŘ, Pavel. 2006. Data mining - přeměna dat v hodnotné informace. *IT Systems* [online]. [cit. 2015-11-29]. Dostupné z: http://www.systemonline.cz/business-intelligence/data-mining-premena-dat-v-hodnotne-informace.htm

SCHILLER, Martin. 2003. Co se skrývá pod zkratkou ETL? *IT Systems* [online]. [cit. 2015-11-29]. Dostupné z: http://www.systemonline.cz/clanky/co-se-skryva-pod-zkratkou-etl.htm

Snowflake Schema. 2015. *Zentut* [online]. © 2015 by ZenTut Website [cit. 2015-11-29]. Dostupné z: http://www.zentut.com/data-warehouse/snowflake-schema/

Star Schema. 2015. *Zentut* [online]. 2015 by ZenTut Website [cit. 2015-11-29]. Dostupné z: http://www.zentut.com/data-warehouse/star-schema/

The Forrester Wave™: Big Data Warehouse, Q2 2017.2017. *Forrester* [online]. [cit. 2018-03-11]. Dostupné z: https://myleadcorner.files.wordpress.com/2017/07/the-forrester-wave-big-data-warehouse-q2-2017-june-2017.pdf