FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Analýza možností využití Datových skladů v bankovnictví
Student:	Bc. Petr Antoš
Vedoucí:	Ing. Jaromír Mataj
Studijní program:	Informatika
Studijní obor:	Webové a softwarové inženýrství
Katedra:	Katedra softwarového inženýrství
Platnost zadání:	Do konce letního semestru 2018/19

Pokyny pro vypracování

1. Nastudujte problematiku datových skladů vzhledem k použitelnosti v bankovním sektoru.
2. Analyzujte používané struktury datových skladů a postupy tvorby DWH.
3. Analyzujte současné poskytovatele DWH na českém trhu.
4. Vytvořte průvodce pro rozhodnutí o výběru poskytovatele DWH pro bankovní sektor.
5. Vytvořte případovou studii (po dohodě s vedoucím práce) pro implementaci a nasazení DWH ve zvolené instituci.
6. Na případové studii otestujte vytvořeného průvodce.
7. Vyhodnoťte přínosy DWH s náklady na jeho implementaci a provoz.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Michal Valenta, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 10. ledna 2018

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Analýza možností využití Datových skladů v bankovnictví

Bc. Petr Antoš

Katedra softwarového inženýrství
Vedoucí práce: Ing. Jaromír Mataj

10. ledna 2019

Poděkování

Rád bych poděkoval vedoucímu své práce Ing. Jaromíru Matajovi za ochotu, trpělivost a drahocenné rady. Můj hlavní a obrovský dík patří mým rodičům za podporu a možnost studovat ČVUT v Praze.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 10. ledna 2019 .....

Odkaz na tuto práci

Antoš, Petr. *Analýza možností využití Datových skladů v bankovnictví.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2019.

Abstrakt

Tato diplomová práce se zabývá problematikou datových skladů se zaměřením na bankovní sektor, jejich architekturou a využívanými nástroji. Dále obsahuje případovou studii zaměřenou hlavně na finanční náročnost daného projektu. Součástí práce je také vytvoření průvodce, sloužícího pro rozhodování jakého dodavatele využít na základě naimplementovaných algoritmů na odhad vah.

Klíčová slova Datový sklad, BI, Databáze, Datová integrace, Datové modelování, Case nastroje, BI nástroje, ETL, ELT

Abstract

This diploma thesis deals with banking sector data warehousing, data warehouse architecture concepts and data warehouse implementation tools. Furthermore, it includes a case study analyzing data warehouse project financial aspects. Besides, this thesis introduces a decision-making tool using statistics and algorithms for the best supplier selection.

Keywords Data Warehouse, BI, Database, Date Integration, Date Modeling, Case Tools, BI Tools, ETL, ELT

Obsah

Seznam obrázků

Úvod

Tato diplomová práce se zabývá problematikou datových skladů a jejich využití v bankovním sektoru. Skoro všechny banky v ČR jsou na datových skladech už velice závislé. Investice do DWH jsou nemalé, pohybují se v milionech korun, proto je rovněž důležitá otázka, kdo DWH postaví, jakou architektůru preferuje, zda má zkušenosti s poptávanými nástroji apod. Před vypracováním této práce byly stanoveny následující cíle:

- Přehledně zpracovat problematiku DWH se zaměřením na bankovní sektor.
- Analyzovat používané struktury DWH a postupy tvorby.
- Analyzovat současné poskytovatele DWH na českém trhu
- Vytvořit průvodce, který napomůže při rozhodování o výběru dodavatele.
- Vytvořit případovou studii pro implementaci a nasazení DWH.
- Otestovat průvodce.

Stanoveným cílům rovněž odpovídá i struktura diplomové práce. Práce je rozdělena do dvou částí. Teoretická část se zabývá problematikou DWH, analýzou používaných řešení a poskytovatelů. Druhá část práce, tj. praktická část, která se zabývá vytvořením průvodce nad poznatky načerpanými při tvorbě teoretické části. Vytvořením případové studie pro implementaci a nasazení a následným otestováním průvodce. Závěr této práce patří zhodnocení kladů a záporů implementace datového skladu.

Kapitola 1

Úvod do problematiky DWH

V této kapitole se čtenář seznámí s tím, co je datový sklad a co je banka. Dále se seznámí se základními pojmy z problematiky datových skladů, jako jsou: databáze, normální formy, BI, dimenzionální model, datová integrace a životní cyklus datového skladu.

1.1 Co je to DWH

Pojem datový sklad se v IT používá už od 90.tých let minulého století. Definic datového skladu lze v dnešní době najít mnoho, avšak za základní definice jsou považovány dvě a to od Williamem H. Inmona a Ralpha Kimball. Definice od Inmona zní takto:
Datový sklad je subjektově orientovaná, integrovaná, časově proměnná a stálá kolekce dat pro podporu rozhodování managementu. [3]
A definice datového skladu od Ralpha Kimballa zní takto:
Datový sklad je kopií transakčních dat speciálně strukturovaných pro dotazy a analýzu.[8]
Podle mého názoru je pro laika, který nikdy o datových skladech nic neslyšel, lepší definice od Kimballa, která jednodušeji vystihuje základní podstatu datového skladu.

Datové sklady jsou v dnešní době využívány v mnoha odvětvích. Jako jsou například energetický, automobilový, pojišťovnický, ale i třeba nábytkářský. Tyto segmenty firem využívají především datové sklady ke konsolidaci a vyčištění dat, dále pak k využití pro analytické a reportovací nástroje, které se v dnešní době používají napříč všemi odděleními ve firmách. Odborníci pracující na vývoji datových skladů by měli ideálně rozumnět jak businessu, tak i struktuře modelu a technologiím relačních databazí. Základem datového skladu je databáze, dále je pak vhodné

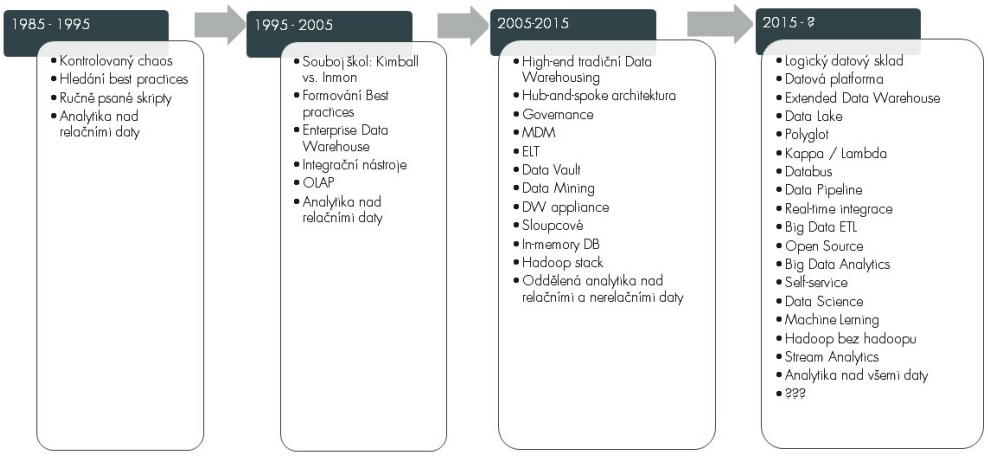
využívat nástroje, které pomáhají s integrací, modelováním a následnou prezentací dat.

1.2 Co je to Business Inteligence

Pojem Business Inteligence (BI) se používá už od 60.tých let minulého století. BI vzniklo s cílem podporovat rozhodování. „Business intelligence můžeme chápat jako ucelený a efektivní přístup k práci s firemními daty, který má vliv na správnost strategických rozhodnutí, a tím i na obchodní úspěch společnosti.“[9] BI poskytuje firmě, která ho využívá, možnost analyzovat historická a současná data, ale i predikovat budoucnost. Tato funkcionalita by byla velice obtížná, pokud by BI nestála na pevném základu. Za pevný základ lze považovat právě datový sklad, který je základním kamenem pro správné fungování BI. Protože pokud nemá BI naprosto validní data, nemůže být ani výstup z BI validní.

1.3 Historický vývoj datových skladů

Vývoj datových skladů můžeme rozdělit do několika etap. Do první etapy můžeme zařadit léta 1985 - 1995, kdy se začaly formovat datové sklady. V tomto období nebyl žádný pevně daný přístup pro tvorbu DWH, takže se spíše v této době objevovaly best practices pro jeho tvorbu. Další dekáda se vyznačuje soupeřením přístupů Inmona a Kimballa. Tyto přístupy jsou více popsány v kapitolách níže. V tomto období už byly zformovány best practices, dále se klade velký důraz na integrační nástroje. V další dekádě se zaměřujeme zejména na ETL procesy, Data Vault, Data Mining a High-end tradiční DWH (tyto pojmy budou popsány v následujících kapitolách). Poslední etapu můžeme pomyslně brát od roku 2015, kdy se začínají objevovat Machine Learning nad DWH a pojmy jako Big Data, Data Lake se začínají ve velkém využívat i v praxi.



Obrázek 1.1: Vývoj DWH převzato[2]

1.4 Datový sklad v bankovním sektoru

Bankovní sektor je velmi specifický tím, že na něj dohlíží mnoho regulátorů od státních až po evropské. Další věcí je, že jsou zde vysoké požadavky na bezpečnost dat a hlavně na její 100 procentní validitu.

1.4.1 Co je banka

Banka je na českém území definována Českou národní bankou takto: „Banka je akciová společnost se sídlem na území ČR, která je na základě udělené bankovní licence oprávněna vykonávat bankovní činnost na území ČR a při dodržení postupů stanovených právem ES (viz též pasportizace, notifikace), též na území jiných členských států EU.“[10] Základní činnosti banky jsou: [10]

- Poskytování úvěrů
- Přijímání vkladů od klientů
- Platební styk a zúčtování
- Vydávání a správa platebních prostředků
- A mnoho dalších.

1.4.2 Co je bankovní produkt

Bankovní produkt je základní pojem v bankovnictví. Tyto produkty jsou přesně definovány takto: „Bankovní produkty jsou jednotlivé služby, které mohou banky samostatně nabízet svým klientům a zpravidla za úplatu provádět“[11] Tyto produkty se dále dělí například na:

- Finančně úvěrové
- Vkladové
- Platební
- Produkty investičního bankovnictví
- Pokladní a směnárenské

1.5 Databáze

Datové sklady musejí na něčem fungovat, proto je potřeba vhodná databáze, která bude vyhovovat specifickým potřebám datového skladu. „Databáze je pojem, jehož význam je chápán mnoha způsoby. Pro někoho to může být obecný výraz pro jakýkoliv zdroj informací, např. Zlaté stránky, pro někoho jiného je to především technologická platforma pro ukládání strukturovaných dat, a jindy to může ještě navíc znamenat celý komplex administrátorských a vývojářských nástrojů. “[12] Předchůdcem dnešních databází byly tzv. papírové kartotéky, které jsme znali například od lékaře. Již tyto kartotéky splňovaly základní myšlenku databází a to uspořádat data podle kategorií. Všechny operace s nimi prováděl přímo člověk. Správa takových kartoték byla velmi náročná. Mezi první elektronické databáze se považuje souborová databáze. Souborová databáze neobsahuje model ukládání dat. Data jsou uložena v souborech. Následníkem souborové databáze jsou databáze hierarchické a síťové. Tyto databáze už mají model, který popisuje formát a strukturu uložení a jejich vzájemné propojení. Model má tvar sítě nebo stromu. Dnes už se tento přístup nevyužívá. Dnes jsou nejvíce využívané relační databáze. Objektové databáze jsou vázané hlavně na objekt. Jako například objektové programování. V podkapitolách si rozebereme nejpoužívanější druhy databazí a to jsou: relační a objektově orientované.

1.5.1 Relační databáze

Termín relační databáze jako první definoval Edgar F. Codd už roku 1970. Je to skoro čtvrt století před tím, než Inmon a Kimball definovali

datový sklad. Relační databáze je databáze, která je založena na relačním modelu. Stavebním kamenem relační databáze je tabulka. V tabulce jsou řádky, tyto řádky označujeme jako záznamy. Mezi největší zástupce těchto databází můžeme zařadit Oracle DB a Microsoft SQL, které jsou popsány v následujících bodech.

1.5.2 Objektově orientované databáze

Stejně jako u programování lidé nejprve používali struktury a pak až objekty, tak je to stejné i u databází. Takže z myšlenky ukládání dat do tabulek a relací se přesunujeme pomalu k myšlence ukládat data v objektech, tak jak s nimi pracují následné aplikace. Každý takto vytvořený objekt je jednoznačně identifikován OID, které na logické úrovni odpovídá ukazateli do virtuální paměti počítače. Objektové databáze také nabízejí možnost vícenásobné dědičnosti, zapouzdření a polymorfizmu. Navíc atributy objektů nemusí být jen primitivního typu, ale mohou být dále strukturované jako například objekt (pomocí reference), množina nebo seznam.[13]

1.6 Normalizace databáze

Normalizace databáze definuje v jaké formě mají být data uložena. Slouží k odstranění redundancí dat a k tomu, aby databáze působila pružněji.

- První normální forma

  „Relace je v první normální formě, pokud každý její atribut obsahuje jen atomické hodnoty. Tedy hodnoty z pohledu databáze již nedělitelné.“[14]

- Druhá normální forma

  „Relace se nachází v druhé normální formě, jestliže je v první normální formě a každý neklíčový atribut je plně závislý na primárním klíči, a to na celém klíči a nejen na nějaké jeho podmnožině. Z čehož vyplívá, že druhou normální formu musíme řešit pouze v případě, že máme vícehodnotový primární klíč.“[14]

- Třetí normální forma

  „Relace se nachází ve třetí normální formě, je-li ve 2.NF a žádný jejich atribut nevykazuje tranzitivní závislost, tzn. že všechny neklíčové atributy jsou navzájem nezávislé. Tedy tranzitivní závislost je taková závislost, mezi minimálně dvěma atributy a klíčem, kde

jeden atribut je funkčně závislý na klíči a druhý atribut je funkčně závislý na prvním atributu.“ [14]

- Boyceho–Coddova normální forma

  Relace R je v BCNF tehdy a jen tehdy, když pro každou netriviální závislost X -> Y, kde X a Y jsou množiny atributů a zároveň Y není podmnožinou X, platí, že X je nadmnožinou nějakého klíče, nebo X je klíčem relace R. [15]

- Čtvrtá normální forma

  „Čtvrtá normální forma se zabývá vztahy uvnitř složeného primárního klíče. Pokud je v tabulce složený primární klíč, může se stát, že některé hodnoty tohoto klíče jsou na sobě nezávislé, ale tím, že spolu tvoří klíč, vzniká falešná souvislost mezi těmito hodnotami a nemohou existovat nezávisle na sobě, což není v souladu s modelovanou realitou. 4.NF proto vyžaduje, aby klíč tvořily jen ty hodnoty, které mají skutečnou vzájemnou souvislost.“ [16]

1.7 Datová integrace

Integrace dat je kombinací technických a business procesů používaných ke spojení dat z různých zdrojů do smysluplných a cenných informací. Kompletní řešení pro integraci dat poskytuje důvěryhodné údaje z různých zdrojů. To znamená, že vstupní data jsou zpracovávána podle předem daných technických a business pravidel a to podle různých transformačních algoritmů. [17] V této podkapitole budou dále popsány ETL a ELT procesy. Tyto ETL a ELT procesy jsou podmonožinou datové integrace a primárně v datovém skladu slouží k přenosu a transformaci dat mezi jednotlivými vrstvami.

1.7.1 ETL (Extract, Transform, Load)

ETL je jedním z procesů využívaných při tvorbě datových skladů. Jeho úkolem je přemístit data ze zdrojových systémů do datového skladu v požadované formě a kvalitě. Tento proces má následující 3 podprocesy:[18]

- Extract

  Účelem tohoto podprocesu je získat data ze zdrojových systémů.

- Transform

  Dalším podprocesem je transformace dat. Tato fáze je důležitá pro datovou kvalitu celého řešení. Proto je nutné se tomuto procesu pečlivě věnovat. Tento proces obsahuje několik další podprocesů.

  - Čištění dat

    Pod tímto pojmem se skrývají procesy, které nám zajistí správnou kvalitu dat.

  - Sledování chybových stavů

    Úlohou sledování chybových stavů je upozornit na chyby, které vznikly v průběhu zpracování dat, ale i na chyby v datové kvalitě.

  - Deduplikace

    Protože data jsou získávána z více zdrojových systémů, dochází k duplikaci záznamu. Může nastat situace, kdy pro jednoho klienta máme například dvě adresy. Proto musíme rozhodnout, která data použít a která naopak zahodit, abychom dosáhli co největší datové kvality.

- Load

  Poslední procesem v ETL je Load, který slouží k nahrání očištěných dat v požadované kvalitě do cílového systému.

1.7.2 ELT(Extract, Load, Transform)

Postupem času s nárůstem výpočetního výkonu a většího objemu dat, se začaly objevovat modifikace ETL procesu v datové integraci. Jak název napovídá, byla prohozena předposlední s poslední fází. tj. fáze loadování s fazí transformace.

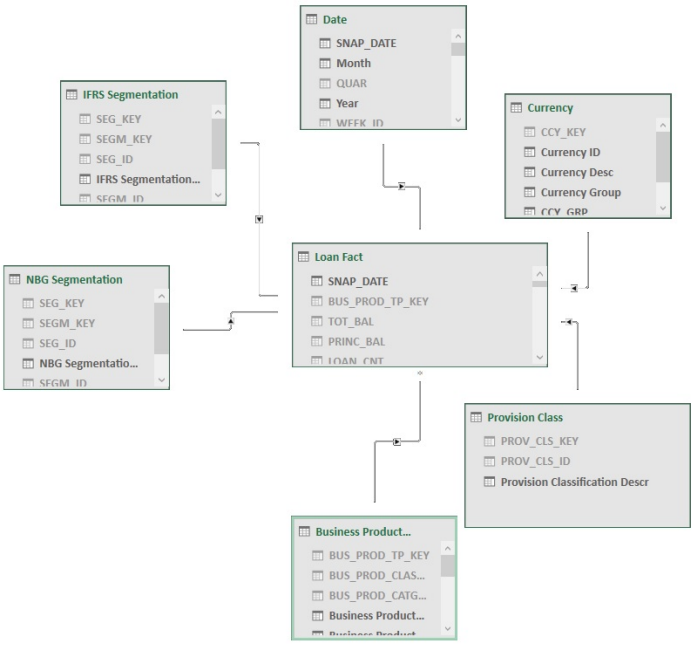
1.8 Datové modelování

Datové modelování je proces, při němž se definují základní požadavky na strukturu dat. „Základním principem datového modelování je centrální a standardizovaný návrh (nebo schéma) databáze. Bez takového schématu nemůže existovat žádná robustní architektura a tomuto schématu musí rozumět všichni, kdo na projektu datové architektury pracují, a to jak obchodníci, tak i technici, obchodní uživatelé, datoví architekti, analytici, návrháři databází, projektoví manažeři, vývojáři i databázoví administrátoři.“[19]

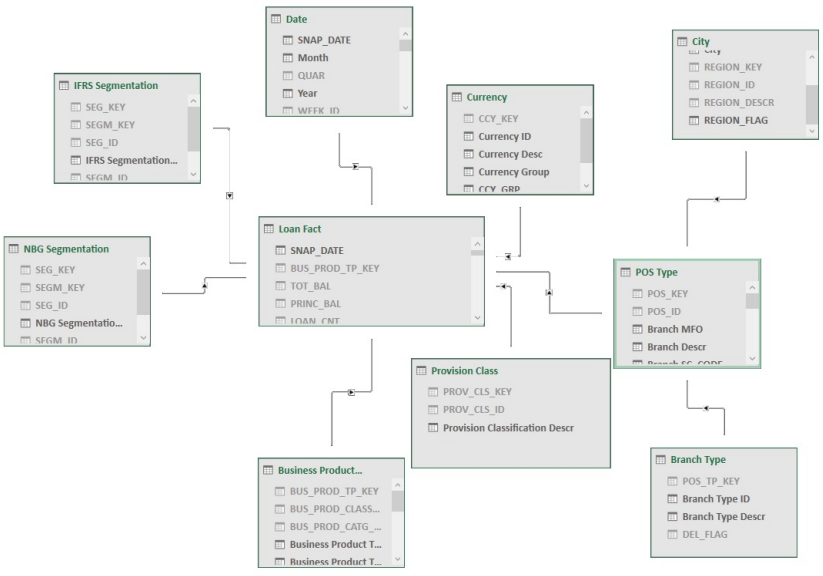
Datový model je základ databáze a tedy i základní kámen datových skladů. Cílem datového modelování je tedy snaha o co nejlepší zachycení reality, kterou se snažíme vymodelovat.

1.8.1 Dimenzionální model

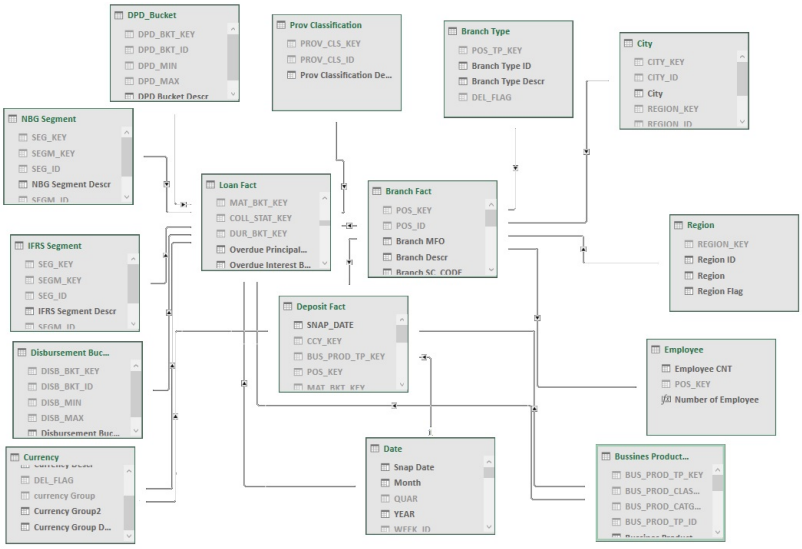
V datovém skladu se vedle 3NF využívá i přístup takzvaného dimenzionálního modelování. Tento model je vhodný zejména na složité analytické dotazy nad velkým množstvím dat. Kromě toho je tento přístup více intuitivní a srozumitelný i pro uživatele, kteří vytvářejí reporty nad datovým skladem. Struktura dat je následující. Data jsou rozdělena do faktových a dimenzionálních tabulek. Faktové tabulky obsahují obchodní metriky jako jsou třeba zůstatky, následně obsahuje klíče do tabulek dimenzí. Tabulky dimenzí obsahují data pro kategorizaci, jako jsou například míra zajištění nebo kategorizaci zákazníků. Základní typy modelů jsou star, který je koncipován tak, že ve středu je faktová tabulka a na tuto tabulku jsou napojeny dimenzionální tabulky. A snowflake schéma je rozšíření star schématu, kde jsou dimenze ještě více normalizovány. Z těchto modelů se pak používají složitější struktury jako je například fact constellation, jak už z názvu vypovídá je to schéma, kde více faktových tabulek sdílí jednu dimenzi. Podle Kimballa je vrstva Data warehouse v dimenzionální podobě, podle Inmona je v 3NF a až vrstva Data Access je v dimenzionální podobě.



Obrázek 1.2: Star schéma



Obrázek 1.3: Snowflake schéma



Obrázek 1.4: Fact constellation schéma

1.9 Životní cyklus vývoje datových skladů

V této kapitole si rozeberem cyklus vývoje datových skladů.

1.9.1 Road mapa

Road mapa datového skladu upřesňuje předběžné zaměření první verze datového skladu a to na základě podrobností o uživatelských požadavcích na dotazy a výstupy. Vytvořením předběžného návrhu schématu datového skladu, splňující tyto požadavky a zjištěním potřebných zdrojů dat. Díky tomu zjistí realizační tým požadavky spojené s implementací. Plánování trvá v závislosti na zaměření datového skladu cca 5 až 8 týdnů.

1.9.2 Nabídka

Nabídka obsahuje konkrétní řešení na implementaci a případný support, kde firma nabízí mimo jiné za jak dlouho a za jakou částku bude datový sklad postaven a supportován. Nabídka vychází z vypracované Road mapy. Poté následuje ve většině případů výběrové řízení na dodavatele datového skladu.

1.9.3 Implementace a vývoj datového skladu

Pokud firma zahajuje implementaci, musí mít už správně vytvořený tým, který implementaci bude provádět. Tým dodavatele musí být vyvážen a

měl by obsahovat tyto role:

- Project Manager

  „Osoba odpovědná za řízení projektu od jeho začátku po jeho konec a to k zajištění úspěšné implementace projektu v rámci dohodnutých nákladů, harmonogramu a kvalitativních ukazatelů.“ [20]

- Architect

  Člověk zodpovědný za použitá technická řešení, která jsou použita při implementaci datového skladu.

- DWH Developer

  Člověk, který se zabývá aspekty procesu vývoje datového skladu, včetně výzkumu, návrhu a programování.

- BI/Business Consultant

  Člověk zodpovědný za správnou analýzu požadavků, který má také na starosti komunikaci se zákazníkem a finální prezentování výsledků vytvořených například pomocí BI nástrojů.

- Tester

  Člověk jehož náplní je testování výsledků a potažmo hledání důvodů proč a kde jsou chyby.

Toto jsou hlavní role v týmu. V týmu mohou být i další role jako technologický specialista a ETL specialista, BI specialista a monoho dalších.

Tým na straně zadavatele by měl obsahovat:

- Sponsor (Projekt leader)

  Je to člověk který je vlastník projektu a je zodpovědný za finanční stránku projektu.

- Business vlastník produktu (Product owner)

  Toto jsou lidé, kteří zodpovídají za danou businessovou část projektu. Například vedoucí riskového oddělení, pod kterým jsou specialisté, kteří mimo jiné mohou vytvářet reporty pro národní banky.

- IT vlastník produktu (Product owner)

  Člověk, který je hlavní komunikační partner ohledně všech otázek z IT.

Na záčátku projektu se musí rovněž určit konkrétní zodpovědné osoby na straně zákazníka i dodavatele, které budou zodpovídat za součinnost z jejich strany. Tzn. vydefinovat si takzvaný komunikační model.

1.9.3.1 Metodika vývoje

Metodika vývoje slouží k definování jak bude vývoj daného datového skladu probíhat.

Agilní model

Agilní model rozděluje projekt na menší části. Myšlenka spočívá v tom, že jakmile bude iterace dokončena, vývojář by měl být schopen odeslat produkt uživateli k testování. Tento proces vytváří software, který je mnohem lépe vyhovující uživatelským potřebám. Díky tomuto přístupu se mohou objevovat nové požadavky, které jsou objeveny během testování iterace zákazníkem. Doba trvání každé iterace je vždy pevná, aby bylo zajištěno, že vývojáři a jejich uživatelé budou schopni pravidelně kontrolovat směr vývoje. [21] Mezi nejznámější metody agilního vývoje patří například Scrum.

Vodopádový model

Model vodopádu je ten, ve kterém každá fáze životního cyklu výrobku probíhá postupně, takže vývoj postupně proudí směrem dolů přes tyto fáze projektu jako vodopád. [22]

Pro vývoj datového skladu se podle mého názoru používá kombinace těchto dvou přístupů. A to taková, že nejprve proběhne analýza, a pak následný proces vývoje a testování ve sprintech.

1.9.4 Testování výstupů z DWH

Datový sklad se dá testovat na několika úrovních. Nejpoužívanější testování je v data martech, kdy se buď porovnávají surová data nebo se porovnávají výstupy vytvořené v reportech z DWH s reporty, které dodá zákazník.

1.9.5 Support

Volbu druhu supportu datového skladu je vhodné vyřešit už na počátku životního cyklu datového skladu, protože se jedná o velice důležitou věc. Existují pouze 3 možné varianty. Firma, která implemementuje datový sklad, ho bude následně i podporovat. Druhá varianta je, že support datového skladu bude dělat interní team v dané organizaci. A třetí varianta, která si myslím je nejméně obvyklá je, že supportovat datový sklad bude jiná externí firma. Je také možnost kombinace jednotlivých řešení

a to například taková, že support bude probíhat pod hlavičkou interního teamu banky, která si najme jednotlivé lidi z dodavatelské firmy, která implementovala datový sklad. Toto řešení se mi zdá asi nejlepší, protože banka není tak moc vázaná na dodavatele, ale zároveň s ním nadále spolupracuje.

Support je obvykle smluvně řešen pomocí SLA (Service-level agreement). SLA je podle definice:„ Dohoda mezi poskytovatelem služeb IT a zákazníkem. Dohoda o úrovních služeb (SLA) popisuje službu IT, dokumentuje cíle úrovní služeb a specifikuje odpovědnosti poskytovatele služeb IT a zákazníka. Jedna dohoda o úrovni služeb může pokrývat řadu služeb IT nebo více zákazníků.“[23]

1.10 Závěr kapitoly

V této kapitole se čtenář seznámil s tím, co je datový sklad a co je banka, dále pak se základními pojmy z problematiky datových skladů, jako jsou: databáze, BI, dimenzionální model a datová integrace. A následně s životním cyklem datového skladu.

Kapitola 2

Architektura datových skladů

V této kapitole se čtenář seznámí se základními typy architektur podle Kimballa, Inmona a Lindstedta, které se využívají při tvorbě datových skladů.

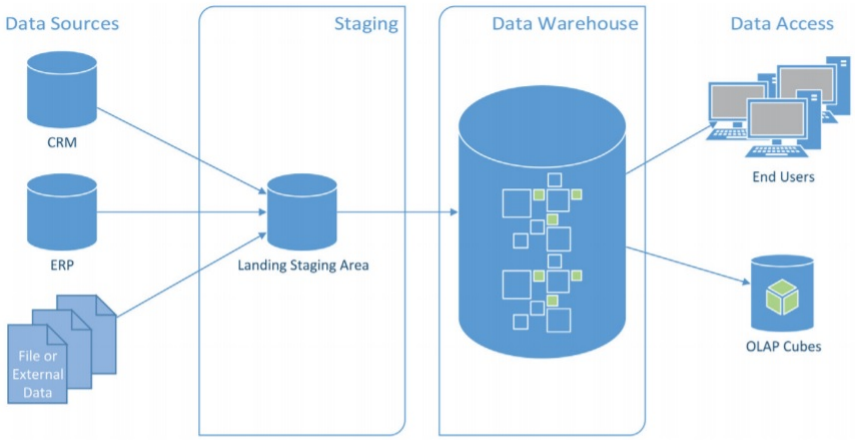
2.1 Datový sklad podle Ralpha Kimballa

Ralph Kimball nahlíží na datový sklad od uživatelů. Ve svém přístupu dává největší důraz na dimenzionální modelování a datové sběrnice. Datový sklad podle Kimballa je tvořen datovými tržišti. Datová tržiště neboli data marty jsou založena na procesech ve firmě. Data mart je podmnožina datového skladu, která je přímo orientovaná na konkrétní oddělení. Aby datový sklad fungoval jako prvek, existují v architektuře tzv. datové sběrnice (datová sběrnice je prvek, který propojuje jednotlivá datová tržiště prostřednictvím sdílených dimenzí). V reálném životě to znamená, že každé datové tržiště musí obsahovat tzv. konformní dimenzi. Dimenze jsou konformní, pokud jejich atributy jsou stejně pojmenované a obsahují stejná data. Díky tomu, můžeme pro jeden finální report kombinovat data z různých datových tržišť. [24]

2.1.1 Architektura DWH podle Ralpha Kimballa

Podle Ralpha Kimballa by měl být datový sklad modelován od businessu. „Jeho princip spočívá v relativně nezávislém vytváření jednotlivých datových tržišť pro specifické útvary podniku (divize, oddělení, pobočky, závody). Každé takové tržiště je typicky kompletně života schopné, tzn. obsahuje veškeré vrstvy a komponenty, které umožní získat data z primárních systémů, zpracovat je, uložit v datovém tržišti, případně analyzovat pomocí OLAP aplikací či Data Mining komponent a prezentovat

je uživateli“[25] Data jsou modelována ve star schématu pro optimalizaci použitelnosti a výkonu dotazu. Každý data mart je tvořen z dimenzí a faktový tabulek, které odpovídají skutečnosti.[8]



Obrázek 2.1: Architektura DWH podle Ralpha Kimballa převzato[3]

2.1.1.1 Data Sources

Zdroje dat, také nazývané operační zdrojové systémy, mohou být považovány za interní nebo externí datové systémy, ve kterých má uživatel velmi omezenou nebo žádnou kontrolu nad obsahem a datovým formátem. Hlavní požadavky na tyto systémy spočívají ve zpracování výkonnosti a dostupnosti. Dotazy proti zdrojovým systémům jsou úzce specifikované. Historizace údajů v těchto zdrojových systémech bývají minimální. Kimball ve své knize zmiňuje, že data ve zdrojových systémech nejsou obvykle stavěná na další integraci. Ačkoli se velmi často domníváme, že vytváření datového skladu zahrnuje pouze kopírování dat z operačních systémů. Nic by nemohlo být dále od pravdy, kvůli odlišné architektuře zdrojových systémů. Datové sklady a provozní systémy mají velmi odlišné požadavky na pracovní zátěž, úpravy dat a historické zpracování dat. Jedním z hlavních rozdílů je, že datové sklady jsou částečně denormalizovány, aby optimalizovaly dotazování a analytický výkon na rozdíl od operačních zdrojových systémů, které jsou navrženy normalizovaným způsobem pro optimalizaci výkonu update-insert-delete. [8]

2.1.1.2 Staging Area

Staging Area je místo, kde jsou načtena zdrojová data ze zdrojových systémů, aby byla transformována a následně zavedena do datového skladu. Hlavním úkolem v této vrstvě je snížit počet operací na zdrojových systémech a čas, kdy se z nich stahují data. Struktura tabulek v této vrstvě je totožná jako ve zdrojových systémech. [3]

2.1.1.3 Data Warehouse Layer

Data Warehouse Layer je to vrstva, kde jsou data organizována, uložena a zpřístupněna pro přímý dotaz uživatelů. Data byla extrahována a transformována a proto je možné data načíst do této vrstvy. Hlavní výhodou dvouvrstvého přístupu je jednoduchost vytváření dimenzionálního modelu ze zdrojových dat ve srovnání s jinými architekturami. [3]

Mezi hlavní výhody konceptu patří:

- První hmatatelné výsledky jsou vidět relativně brzo a to s malým rozpočtem.
- Opravování a tvorba změn je relativně jednoduchá.
- Je uživatelsky přívětivější, jelikož vychází z požadavků menší skupiny lidí konkrétního oddělení.

2.2 Datový sklad podle William Inmona

Inmonův přístup k tvorbě datových skladů je od zdrojových dat. „Tento přístup spočívá v jednorázovém vybudování celkového řešení, tedy všech potřebných komponent pokrývajících současně všechny definované analytické potřeby zadavatelské organizace. “[25].

Podle definice datového skladu Inmona jsou nejdůležitější následující body:

- Subject–oriented

  Subjektivní orientaci můžeme vysvětlit jako organizaci kolem hlavních subjektů sledované domény, jako například: zákazník, produkt. Data jsou organizována spíše dle subjektu než dle aplikace.

- Integrated

  Do datového skladu vstupují data z mnoha primárních systémů, které se mohou lišit způsobem zaznamenávání jednotlivých údajů.

Integrovatelnost tedy zajišťuje, že data v datovém skladu budou uložena ve sjednoceném stavu.

- Novolatile

  Data která jsou nahraná do datového skladu se dále už manuálně nemění. Na rozdíl od provozních systémů, kde se data mohou časem libovolně měnit, jsou data do datových skladů nahrávána a dál už nijak nemodifikována.

- Time–variant

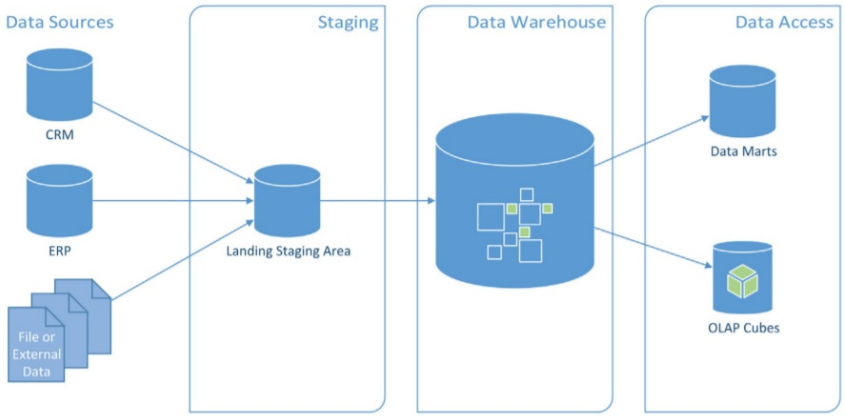
  Z DWH můžeme získat nejen aktuální data, protože v datovém skladu probíhá historizace dat a to tak, že do datového skladu jsou data nalévána ve snímcích. Dále pak v DWH mají data buď časovou známku, kdy jsou validní, například zůstek k 31.8.2018 nebo obsahují záznam, který značí od kdy do kdy jsou validní. Pokud například klient změní adresu. Stará adresa dostane platnost do dne změny a nová ode dne vytvoření do nekonečna.

- Granularita dat

  Podle Inmona má být v DWH uložena informace v co nejmenší granularitě.

2.2.1 Architektura DWH podle Billa Inmona

Bill Inmon přistupuje k problematice tak, že datový sklad by měl být modelován jako jedno velké integrované schéma skladu, které definuje jako centralizované uložiště pro celý podnik. Mělo by být navrženo tak, aby odpovídalo třetí normální formě a bylo odpovědné za uchovávání "atomových"údajů na nejnižší úrovni detailu. Dimenzionální data marty se vytvářejí až po vytvoření úplného datového skladu. Tyto data marty obsahují údaje požadované pro konkrétní obchodní procesy nebo konkrétní oddělení. Za tímto účelem je přístup Inmon běžně označován jako přístup shora dolů.[3]



Obrázek 2.2: Architektura DWH podle Inmona převzato[3]

Hlavní rozdíl oproti dvouvrstvé architektuře je ten, že ve vrstvě datového skladu jsou data uložena ve třetí normální formě a až třetí vrstva obsahuje jednotlivé data marty, kam přistupují uživatelé. Kimbalův přístup má data marty už v Date warehouse vrstvě. Data Source a Staging Area mají stejnou funkcionalitu.

Mezi hlavní výhody konceptu patří:

- Design je komplexní a proto jednodušeji udržovatelný.
- Všechna data jsou pohromadě na jednom místě, proto je možné lépe hledat souvislosti napříč celým podnikem.
- Pokud je datový sklad připravený, vytvářet libovolná datová tržiště je pak relativně jednoduché.

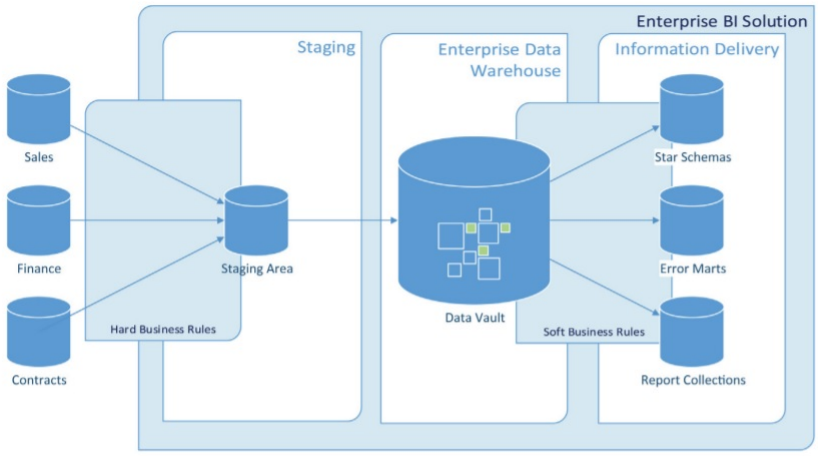
Z počátku si datové sklady kvůli investici mohly dovolit jen velké korporáty, protože se k vývoji datového skladu přistupovalo podle Inmona, který byl i první, kdo přišel s myšlenkou datových skladů. Na přelomu století se začal prosazovat i přístup Ralpha Kimballa, který je méně nákladný, proto si ho mohly dovolit i menší podniky a výsledky byly vidět v kratším časovém horizontu. Nesmíme ale zapomínat, že tyto přístupy popisují pouze doporučení jak datové sklady stavět. V praxi se často ukazuje, že se využije kombinace obou těchto postupů. Tuto kombinaci popsal právě Dan Lindstetd, o kterém je následující kapitola.

2.3 Datový sklad podle Dana Lindstedta

Po Kimballovy a Inmonovy přišel Dan Lindstedt a jeho definice Data Vault. Data Vault je detailně orientované (má tedy nejvyšší možnou granularitu), historicky sledováná a jedinečně propojená sada normalizovaných tabulek, která podporuje jednu nebo více funkčních oblastí podnikání. Jedná se o hybridní přístup zahrnující kombinaci 3. normální formy (3NF) a dimenzionálního modelování. Návrh je flexibilní, škálovatelný, konzistentní a přizpůsobitelný potřebám podniku. Jedná se o datový model, který je navržen speciálně pro potřeby současných podnikových datových skladů.[26]

2.3.1 Architektura DWH podle Dana Lindstedta

Architektura podle Dana Lindstedta, která je znázorněna na obrázku 2.3, je kombinací architektur od Kimballa a Inmona, které byly popsány v předešlém textu.



Obrázek 2.3: Architektura DWH podle Dana Lindstedta převzato[3]

Data Source a Staging Area mají stejnou funkcionalitu. Návrh se liší hlavně v dalších vrstvách.

2.3.1.1 Business Rules

Tato architektura rozlišuje tvrdé a měkké business pravidla. Obecně řečeno, tyto pravidla jsou omezením příchozích údajů, aby vyhovovaly požadavkům podniku. Rozdíl mezi tvrdými a měkkými pravidly spočívá v tom, že tvrdá business pravidla nikdy nemění význam příchozích dat,

nechá je uložena tak jak jsou. Jinými slovy, týkají se pouze prosazování datových typů. Naopak měkká business pravidla mění význam příchozích údajů, například změnou úrovně detailů nebo interpretace. Typickým příkladem je agregace v kategoriích nebo konsolidace dat z více zdrojů. Určují také, jak se data transformují tak, aby splňovala business pravidla. Tato architektura podporuje změnu dat při načítání do staging vrstvy na rozdíl od dvou a třívrstvých architektur. Tato včasná implementace zlepšuje aplikaci pravidel a kvalitu dat. Existují však problémy při změnách obchodních pravidel z důvodu závislosti na vyšších vrstvách datového skladu. [3]

2.3.1.2 Enterprise Date Warehouse

Tato vrstva, kterou pojmenoval Linstedt jako Data Vault, ukládá všechna historická data. Obsahuje nezpracované údaje, které jsou modifikovány pouze tvrdými business pravidly. Na rozdíl od Kimballovy architektury není tato vrstva datového skladu přímo přístupná koncovým uživatelům. Koncoví uživatelé obvykle získají přístup pouze k data martům, kde jsou data pro ně intuitivněji uložená. Uložiště dat může být také definováno jako detailně orientovaná, historicky sledováná a jedinečně propojená sada normalizovaných tabulek, která podporuje jednu nebo více funkčních oblastí podnikání. Jedná se o hybridní přístup, jehož cílem je spojit vlastnosti dvou a třívrstvých architektur využívajících třetí normální formu (3NF) a dimenzionální modelování. [3]

2.3.1.3 Information Delivery Layer

Lindstedt názvem této vrstvy zdůrazňuje, že data byla doručena na správná místa. Jak již bylo řečeno, koncoví uživatelé získají přístup pouze k této vrstvě. Často se řídí principy dimenzionálního modelování a tvoří základ jak pro modelování dimenzí, tak pro zpracování online analýzy. Data marty v této vrstvě však mohou být také modelovány ve 3NF, aby vyhovovaly preferencím koncových uživatelů. V této vrstvě existují mimo jiné dvě speciální tržiště, které se liší od ostatních. Jsou to centrální uložiště chyb a metadat. Koncoví uživatelé, například správci, je používají k analýze řady problémů v datovém skladu.[3]

Mezi hlavní výhody konceptu patří:[27]

- Model jde snadno upravovat a rozšiřovat
- Podporuje relativně jednoduše změny v business pravidlech

2.4 Porovnání přístupů

Pro porovnání přístupů jsem zvolil následujicí metriky.

- Struktura uložení dat ve vrstvě datového skladu:
  - Kimball

    Data jsou uložena ve formátu finální potřeby optimalizované pro reportování. Rovněž zde dochází k historizaci záznamů.
  - Inmon

    Data jsou uložena v 3NF. Struktura není optimalizována pro přímé dotazy a proto vyžaduje vytvoření data martů pro reportování.
  - Lindstedt

    Struktura dat není optimalizována pro přímé dotazy a proto vyžaduje vytvoření data martů pro reportování.
- Cena
  - Kimball

    Z variant by mělo být nejméně nákladné.
  - Inmon

    Nákladnější než varianta od Kimballa.
  - Lindstedt

    Náklady jsou srovnatelné s Inmonovým přístupem.
- Modelování
  - Kimball

    Složitost modelování se liší podle oboru. Vzhledem k tomu, že samotný model je postaven pro finální podobu dat pro reporting, nepotřebuje další modelování pro zachycení zdrojových dat, jako v přístupech Inmona a Lindstedta
  - Inmon

    Modelování není obecně složité, ale vyžaduje modelování do dvou vrstev.
  - Lindstedt

    Modelování by mělo být podobně složité jako u Inmonovy architektury. Vyžaduje další modelování data martů.

- Dotazování

  Složitost dotazování by měla být u všech přístupů podobná, protože jsou všechny nejvrchnější vrstvy orientovány na uživatele.

Přístupy Lindstedta a Inmona jsou v dosti věcech velice podobné. Je zřejmě nemožné rozhodnout, jaký přístup je obecně nejlepší. Každý má své klady a zápory.

2.5 Závěr kapitoly o architektuře datových skladů

V této kapitole se čtenář seznámil se základními typy architektur podle Kimballa, Inmona a Lindstedta, které se využívají při tvorbě datových skladů.

Kapitola 3

Technologie a nástroje pro DWH

V této kapitole se čtenář seznámí se základními technologiemi a nástroji, které jsou potřeba a které se využívají při tvorbě a používání datových skladů.

3.1 HW datového skladu

„Hardware označuje veškeré fyzicky existující technické vybavení počítače.“[28] HW datového skladu můžeme pomyslně rozdělit do dvou kategorií a to:

- Nákup vlastního hardwaru
- Využití cloudu

V dnešní době v bankovním sektoru je z 99% zastoupena první varianta, ale o cloudu se mluví jako o budoucím řešení.

3.1.0.1 Vlastní hardware

Na začátcích warehousingu stačila nasazená databáze na bežném serveru. Nyní banky preferují nákup specifického zařízení pro tuto funkčnost. Na trhu existuje několik tzv. krabicových řešení. Já jsem si vybral jako zástupce těchto řešení Exadatu a Teradata Data Warehouse Appliance.

Exadata

Oracle Exadata Database Machine (Exadata) je integrované řešení třídy Database Appliance pro transakční i analytické úlohy. Alternativně lze

Exadatu využít jako platformu pro konsolidaci více instancí Oracle Database. Jedná se o jeden z „engineered systémů“ společnosti Oracle určených pro databázové úlohy. Exadata integruje software a hardware Oracle do jednoho balení a vytváří dokonale vyváženou platformu s vysokou dostupností dosahovaných výkonů, kterých nelze na běžném hardwaru dosáhnout ani teoreticky. Exadata využívá řadu softwarových a hardwarových prvků, které nelze na běžném hardwaru použít z licenčních nebo technologických příčin.Nejdůležitější vlastností Exadaty je schopnost přesunout část databázové logiky přímo na storage servery, které část SQL operací realizují samostatně bez nutnosti přenášet data z diskového subsystému do paměti databázového serveru. Škálovatelnost Exadaty je teoreticky neomezená, díky možnosti spojovat jednotlivé skříně Exadaty do jednoho clusteru pomocí InfiniBandu. Oracle Exadata Database Machine je nejdokonalejší formou Oracle Database. Zanedbat nelze ani skutečnost, že všechny komponenty řešení jsou od jednoho dodavatele, který centrálně zajišťuje podporu hardwaru i softwaru včetně důsledného testování softwarových záplat všech vrstev vůči sobě. Za zmínku stojí i možnost využití stávajících Oracle Database licencí pro Exadatu. [29]

Teradata Data Warehouse Appliance

Teradata Data Warehouse Appliance, je vysoce výkonné krabicové řešení, které může být používáno i jako platforma pro datový sklad. Toto zařízení je kompletní, integrované řešení na klíč, které zahrnuje hardware a Teradata Database řízené servery a volitelnou záložní technologii, vše v jediné skříňce. Teradata Data Warehouse Appliance poskytuje rychlejší zpracování v paměti a vyšší průchodnost dotazů využíváním technologie Intel Haswell a nejnovějšími paměťové komponenty pro rychlejší přístup k datům uloženým v paměti. [30]

3.1.0.2 Využití Cloudu

„Podle definice je cloud computing virtualizací výpočetních prostředků, což znamená, že výpočetní procesy a jimi poskytované služby informačních a komunikačních technologií jsou provozovány jinde než u jejich odběratele, jemuž jsou poskytovány zpravidla za úplatu jako služba.“[31] Problematika datových skladů je tu už dlouho, oproti tomu cloudové služby jsou relativně nová věc. Cloud dělen podle modelu nasazení a to na:

- Veřejný Cloud

  „Někdy je označován jako klasický model Cloud computingu. Jedná se o model, kdy je poskytnuta a nabídnuta široké veřejnosti výpočetní služba.“[32]

- Hybridni Cloud

  „Hybridní Cloudy kombinují jak veřejné, tak soukromé Cloudy. Navenek vystupují jako jeden Cloud, ale jsou propojeny pomocí standardizačních technologií.“[32]

- Privátní Cloud

  „Oblak je v tomto případě provozován pouze pro organizaci, a to buď organizací samotnou, nebo třetí stranou.“[32]

- Komunitní Cloud

  „Jedná se o model, kdy je infrastruktura Cloudu sdílena mezi několika organizacemi, skupinou lidí, kteří ji využívají. Tyto organizace může spojovat bezpečnostní politika, stejný obor zájmu.“[32]

Pro datový sklad je zajímavá varianta například hybridního Cloudu, kdy může být produkční prostředí nasazeno v privátní části a vývojové naopak ve veřejném Cloudu.

Další rozdělení Cloudu je pomocí distribučního modelu a to na:

- IAAS

  „infrastruktura jako služba (z „Infrastructure as a Service") — v tomto případě se poskytovatel služeb zavazuje poskytnout infrastrukturu. Hlavní výhodou tohoto přístupu je to, že se o veškeré problémy s hardwarem stará poskytovatel. Model IaaS je vhodný pro ty, kteří vlastní software (či jejich licence) a nechtějí se starat o hardware. Příkladem IaaS jsou Amazon WS, Rackspace nebo Windows Azure.“[32]

- PAAS

  „platforma jako služba (z „Platform as a Service“) — poskytovatel v modelu PaaS poskytuje kompletní prostředky pro podporu celého životního cyklu tvorby a poskytování webových aplikací. To zahrnuje různé prostředky pro vývoj aplikace jako IDE nebo API, ale také např. pro údržbu. Nevýhodou tohoto přístupu je proprietární uzamčení, kdy může každý poskytovatel používat např. jiný programovací jazyk. Příkladem poskytovatelů PaaS jsou Google App Engine nebo Force.com (Salesforce.com).“[32]

- SAAS

  „software jako služba (ze „Software as a Service") — aplikace je licencována jako služba, která je pronajímána uživateli. Uživatelé si tedy kupují přístup k aplikaci. SaaS je ideální pro ty, kteří požadují přístup k aplikaci odkudkoliv. Příkladem může být známá sada aplikací Google Apps.“[32]

Výhody a nevýhody DWH v Cloudu[33]

- Výhody
  - Minimální nebo malé počáteční investice.
  - Snadné a rychlé získání datového skladu, pokud zákazník žádný nemá nebo mu stávající řešení nevyhovuje.
  - Transparentní finanční model, kdy lze snadno vyčíslit celkové náklady na datový sklad.
  - Vysoká elasticita kapacit pro výpočty i uložení dat, kdy lze šetřit na některých typech kapacit průběžně (např. nastavení pro běžný provoz vs. nastavení pro měsíční uzávěrku).
  - Pravidelné aktualizace.
  - Maximální automatizace řešení včetně proaktivního monitoringu a optimalizací.
- Nevýhody
  - Silná závislost na kvalitním připojení k internetu u čistě cloudových služeb.
  - Některé velmi specifické funkcionality lze zrealizovat velmi obtížně nebo vůbec.
  - Možné pochybnosti o bezpečnosti.
  - Fixní softwarové komponenty s těžko modifikovatelným vývojovým cyklem.

Jak můžeme vidět, je velký počet výhod i nevýhod pro variantu datového skladu v Cloudu. Podle mého názoru, ale v bankovním sektoru nepřijde doba, kdy by se banky vrhly do používání veřejného Cloudu, protože je v lidech zakořeněný názor, že ty data, která nemám fyzicky u sebe, nejsou v takovém bezpečí, jako ta, které mám uložena na vlastních serverech. Proto v případě datových skladů připadá v úvahu pouze privátní a částečně hybridní model nasazení. V rámci distribučního modelu mohou být využity pro určité využití všechny tři distribuční modely.

3.2 Zástupci databází

Jako vhodné zástupce databází jsem si vybral tyto:

- Oracle Database
- Microsoft SQL server
- Teradata

Obrázek 3.1: Magic Quadrant for Data Management Solutions for Analytics převzato[4]

Všichni tři jsou zástupci relačních databází (RDBMS). Vybral jsem si je, protože dnes jsou datové sklady stavěny hlavně na relačních databázích a tito tři zástupci jsou podle mě jedni z nejčastěji vybraných databazí pro datový sklad právě v bankovním sektoru. Mimo jiné jsou

všichni tři podle Magic Quadrantu od firmy Gartner v pozicích největších leaderů na trhu v řešení správy dat.

3.2.1 Oracle Database

Mezi nejsilnější stránky této databáze patří schopnost automaticky najít vhodný způsob přístupu k datům. Mezi další silné stránky mimo jiné patří velká dobře fungující komunita uživatelů, vývojářů a rovněž také velké množství literatury. Samotná databáze je řádkově orientovaná, avšak v posledních verzích přibývá podpora i pro sloupcové uložení dat (hybridní sloupcová komprese, in-memory option). Oracle DB velmi dobře pracuje s malým, ale i rovněž s velkým objemem dat. Podporuje i různé přístupy k tvorbě relačního modelu, což z ní činí zcela univerzální platformu s velkým potenciálem pro škálování. Databáze používá vlastní dialekt SQL. Pro složitější úlohy lze využít například PL/SQL. Oracle Database podporuje operativní i analytické úlohy typu datový sklad, a to včetně OLAP. Oracle Database podporuje vysokou dostupnost i řadu různých vysoce robustních technik zálohování pro obnovu v případě selhání. [34]

3.2.2 Microsoft SQL server

„Microsoft SQL Server představuje ucelené řešení databázového serveru a informační platformy, nabízející úplnou sadu technologií a nástrojů připravených pro podnikové prostředí, které uživatelům pomáhá vytěžit z informací maximální hodnotu, a to při nejnižších celkových nákladech na vlastnictví. SQL Server se vyznačuje vysokou dostupností, výkonem, škálovatelností, jednoduchou migrovatelností a spolehlivostí. Samozřejmostí je podpora standardu jazyka SQL, rozšířeného o T-SQL (Transact-SQL) pro psaní funkcí, uložených procedur, triggerů nebo využití v SQL kódu. Kromě databázového enginu pro ukládání dat a transakčně orientovaného zpracování dat obsahuje i enginy pro Business Intelligence (BI), které jsou součástí instalace a jsou v ceně licence: analytický engine pro OLAP databáze a datové kostky, reportovací engine pro online reporting a integrační engine pro integraci dat (ETL) mezi různými zdrojovými i cílovými databázemi. Tím se stává součástí integrované platformy pro Business Intelligence od společnosti Microsoft a zahrnuje kompletní funkce pro datové sklady, analýzy a generování sestav či přehledy výkonnostních metrik s rozpadem do podrobných datových analýz a Data Mining modelů. SQL Server se nachází v kvadrantu lídrů v hodnoceních organizace Gartner Magic Quadrant for BI Platforms a Magic Quadrant

for Data Warehousing a svou pozici na rozdíl od konkurentů neustále vylepšuje.“ [29]

3.2.3 Teradata

Jako dalšího zástupce jsem si vybral Teradatu, které se ale budu věnovat jen okrajově, protože i v rámci Magic Quadrantu od firmy Gartner lehce zaostává před dvěmi předešlímy zástupci. Firma Teradata působí na trhu od svého založení v roce 1979. Koncept systému vznikl už o 3 roky dříve v Californském technologickém institutu(CALTECH). Teradata symbolizuje schopnost spravovat extrémní množství dat. Databáze od Teradaty je primárně určena pro datové sklady a BI aplikace. Je založena na „shared nothing“ architektuře umožňující lineární rozšiřitelnost. V současnosti využívána v největších datových skladech v Česku a to konkretně v KB, České pojišťovně. [35]

3.2.4 Porovnání

Všechny 3 zástupci jsou zástupci relačních databází. Databáze se liší například v partitioningu (logické rozdělení databáze nebo jejích prvků do odlišných nezávislých částí. Rozdělení databáze se obvykle provádí pro účely většího výkonu, dostupnosti nebo pro vyvažování zátěže) nebo v metodách replikace.[36]. Pro detailnější porovnání jsem si vybral pouze DB od Oraclu a Microsoftu.

3.2.4.1 Jazyk

Ačkoli oba systémy používají verzi jazyka Structured Query neboli SQL, MS SQL Server používá Transakční SQL neboli T-SQL, což je rozšíření SQL, který byl původně vyvinut společností Sybase. Oracle mezitím používá PL/SQL neboli procedurální jazyk SQL. Jazyky mají odlišnou syntaxi. Hlavním rozdílem mezi těmito dvěma jazyky je, jak řídí proměnné, uložené procedury a vestavěné funkce. PL/SQL v Oracle může také seskupovat procedury dohromady do balíčků, které nelze provést v MS SQL Serveru. PL/SQL je podle mnohých složitější a potenciálně výkonnější.[37]

3.2.4.2 Transakce

Dalším rozdílem mezi Oracle a MS SQL Server je řízení transakcí. Transakci lze definovat jako skupinu operací nebo úkolů, které by měly být považovány za jednu jednotku. Například sada dotazů SQL upravujících záznamy, které musí být aktualizovány současně, kde například selhání

aktualizace jednoho záznamu v sadě by nemělo vést k aktualizaci žádné z těchto záznamů. Ve výchozím nastavení MS SQL Server spustí každý příkaz jednotlivě a bude obtížné nebo nemožné vrátit zpět změny, pokud dojde k nějakým chybám. Je ale možné řádně seskupit příkazy pomocí Begin Transaction, který se používá k deklarování začátku transakce a na konci je použit příkaz commit. Tento příkaz commit zapíše změněné údaje na disk a ukončí transakci. Při správném použití a manipulaci s chybami umožňuje rollback určitý stupeň ochrany před poškozedním dat, ale po příkazu commit už nelze vratit zpět upravené záznamy. V Oraclu se na druhé straně každé nové připojení k databázi považuje za novou transakci. Při provádění dotazů, se změny provádějí pouze v paměti a nic se neukládá do databáze, dokud se neobjeví explicitní příkaz commit (s několika výjímkami souvisejícími s příkazy DDL, které zahrnují implicitní závazky a jsou okamžitě závazné). Po příkazu commit vydává další příkaz, který v podstatě iniciuje novou transakci a proces začíná znovu. To poskytuje větší flexibilitu a také pomáhá při kontrole chyb, neboť se žádné změny data nezmění, dokud DB explicitně nevydá tento příkaz.[37]

3.2.4.3 Organizace databáze

MS SQL Server organizuje všechny názvy objektů, jako jsou tabulky, pohledy a procedury podle databázového jména. Ve službě Oracle jsou všechny databázové objekty seskupeny podle schémat, které jsou kolekcí podmnožin databázových objektů a všechny databázové objekty jsou sdíleny mezi všemi schématy a uživateli. I když je vše sdílené, každý uživatel může být omezen prostřednictvím rolí a oprávnění.[37]

3.3 Nástroje pro datovou integraci

Existuje velká škála nástrojů pro datovou integraci ať už z řad komerečních, tak i open source nástrojů. Pro vhodný výběr nástrojů jsem použil Gartner Magic Quadrant for Data Integration Tools 2018, který můžete vidět na obrázku 3.2, od společnosti Gartner, který zobrazuje aktuální stav nástrojů pro datovou integraci na světovém trhu.

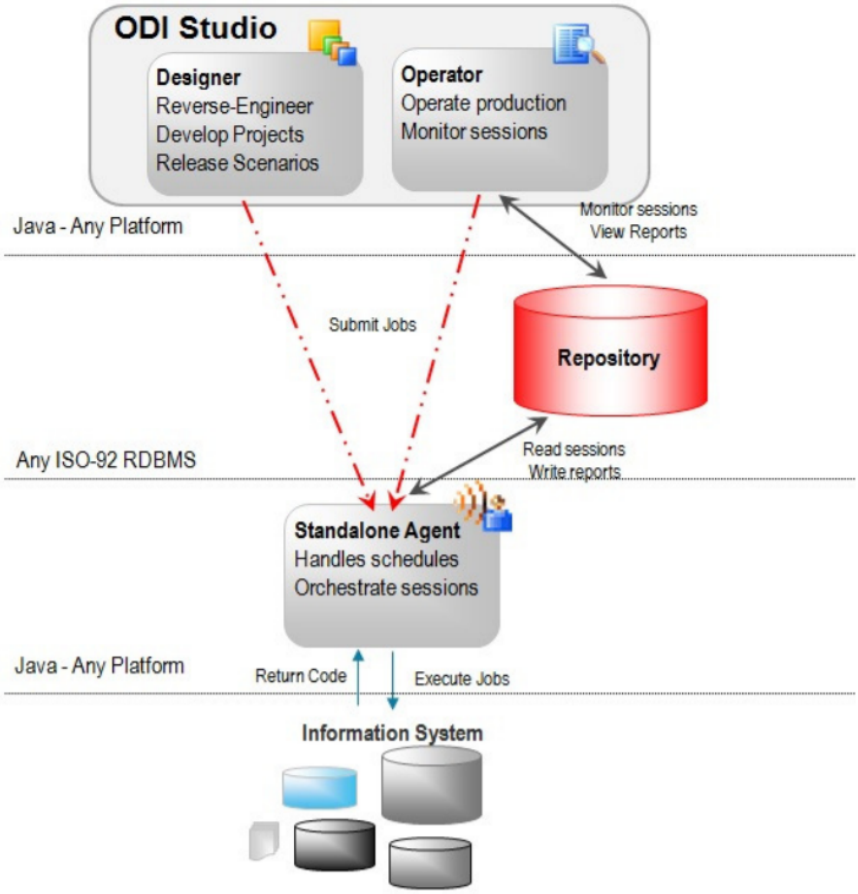
Obrázek 3.2: Magic Quadrant for Data Integration Tools 2018 převzato[5]

Pro hlubší analýzu nástrojů jsem vybral: Oracle Data integrator (ODI), MS Integration Services(MSIS) a Hitachi Vantara neboli Pentaho. ODI a MSIS jsem si vybral protože jsou to zástupci integračních nástrojů od firem, které jsou leadři v databázovém světě. Hitachi jsem si vybral protože nabízí zajímavé řešení v podobě community edice, která je zdarma.

3.3.1 Oracle Data integrator

Oracle Data Integrator je široce používaný softwarový produkt pro integraci dat. Poskytuje nový přístup k definování procesů, transformaci a integraci dat, což vede k rychlejšímu a jednoduššímu vývoji a údržbě. Oracle Data Integrator je založený na architektuře ELT (Extract, Load, Transform).[38]

3.3.1.1 Komponenty Oracle Data Integrator



Obrázek 3.3: Architektura Oracle Data Integrator převzato[6]

Architektura Oracle Data Integratoru se skládá z několika komponent: [6]

- ODI Studio
  - Designer

    Zde jsou definována pravidla pro transformaci a integritu dat. Jsou zde také metadata databází a aplikací. Tento modul je hlavní pro vývojáře a správce metadat.
  - Operátor

    Operátor řídí a monitoruje procesy integrace dat. Zobrazuje protokoly o spuštění s chybami, počet zpracovaných řádků,

statistiky výkonu, skutečný kód, který se provádí atd. V době návrhu mohou vývojáři také použít modul Operátor pro účely odstraňování chyb.

  - Topologie

    Topologie není zobrazena na grafu architektury, definuje fyzickou a logickou architekturu infrastruktury.

  - Security

    Security neni zobrazena na grafu architektury, řídí uživatelské profily, role a jejich oprávnění. Security může také přiřadit přístup k autorizaci objektů a funkcí.

- Repository

  Repository se skládá obvykle z jednoho nebo více master a z několika pracovních repositářů. Tyto repository jsou množina tabulek uložených v relační databázi. Jsou zde uloženy všehny objekty, které jsou využívány nebo vyvíjeny.

- Agent

  Agent řídí provádění jednotlivých úloh. Získává kód uložený v repository, připojí se k cílové a zdrojové databázi a provede celý proces datové integrace.

Mezi silné stránky Oracle Data Integrator patří například:

- Rychlejší a jednodušší vývoj a údržba

  Přístup k integraci dat založený na deklarativních pravidlech výrazně snižuje křivku učení produktu a zvyšuje produktivitu vývojářů a současně usnadňuje průběžnou údržbu.

- Data kvality firewall

  Zajišťuje, že chybné údaje jsou automaticky detekovány a recyklovány před vložením do cílové aplikace.

- Nezávislost platformy

  Oracle Data Integrator podporuje všechny platformy, hardware a operační systémy se stejným softwarem.

- Znalostní (Knowledge) moduly

  „Best practices“ řešení pro různé oblasti datové integrace a použité technologie.

3.3.2 Hitachi Vantara

Pentaho Data Integration (PDI, také nazývaná Kettle) nyní vystupuje pod značkou Hitachi Vantara, jehož součástí je nástroj Pentaho zodpovědného za procesy ETL.

3.3.2.1 Komponenty Pentaho Data Integration

Pentaho data integratiion obsahuje tyto komponenty: [39]

- Data Integration Server
  - Execution
    Provádí ETL joby a transformaci pomocí Pentaho Data Integration engine.
  - Security
    Umožňuje spravovat uživatelské role, přidělovat práva apod.
  - Content Management
    Centrální repozitory umožňuje správu ETL jobů a transformací. Obsahuje i úplnou historii vytvořeného obsahu.
  - Scheduling
    Poskytuje služby, které umožňují naplánovat a sledovat aktivity na Data Integration Serveru v prostředí Spoon.
- Spoon

  Je desktopová aplikace, která využívá grafické rozhraní pro editování a vytváření transformací a jobů. Spoon poskytuje způsob, jak vytvořit komplexní úlohy ETL bez nutnosti číst nebo psát kód.
- Pan

  Je proces, který spouští transformace a joby vytvořené ve Spoonu. Transformační mechanismus umožňuje číst data z více zdrojových systémů a zapisovat je do různých cílových systémů.
- Kitchen

  Je proces, který má stejnou funkcionalitu jako Pan. Joby jsou obvykle spouštěny v dávkách v pravidelných intervalech.
- Carte

  Carte je odlehčený webový kontejner, který umožňuje nastavit vyhrazený vzdálený ETL server. Ten poskytuje podobné funkce vzdáleného zpracování jako server Data Integration Server, ale neposkytuje scheduling ,security a content management.

Mezi silné stránky Hitachi Vantara patří například: [40]

- Grafické rozhraní

  Hitachi Vantara se snadno používá. Každý proces je vytvořen grafickým nástrojem, v němž se specifikuje, co se má dělat bez psaní kódu.

- Licenční politika

  Hitachi Vantara lze pořídit i pod community edicí, která je zdarma, ale neobsahuje některé funkcionality.

- Modularita

  Hitachi Vantara lze použít jako součást většího Pentaho Suite nebo jako samostatnou aplikaci.

3.3.3 SQL Server Integration Services

SQL Server Integration Services je nástroj pro datovou integraci mezi jednotlivými systémy. Zdrojem i cílem datového toku může být jak relační i nerelační databáze, tak textové soubory, webové služby a další. V případě nutnosti si můžeme vlastní konektor napsat v MS .NET Frameworku, na němž MS Integration Services běží. Integrace datových toků se designuje v plně grafickém prostředí, jehož součástí je i debugger umožňující zobrazení tekoucích dat. SQL Server Integration Services pokrývá stejnou funkcionalitu jako oba předchozí integrační nástroje. [41]

3.3.3.1 Komponenty SQL Server Integration Services

SQL Server Integration Services obsahuje mimo jiné tyto komponenty:[42]

- Průvodce importem a exportem

  Průvodce importem a exportem jednoduše přenáší data ze zdroje na cíl, ale nezahrnuje možnosti transformace dat.

- SSIS Designer

  Grafický nástroj.

- SSIS API

  Programovací modul SSIS API umožňuje kódovat balíčky SSIS pomocí různých programovacích jazyků.

Mezi silné stránky SQL Server Integration Services patří například:

- Snadné použití.
- Možnost pořízení s DB.

Porovnání

Výše zmíněné nástroje na datovou integraci, mají v kostce stejnou funkčnost a použitelnost. Nástroje se odlišují například v tom, že integrační nástroj od Microsoftu nepodporuje sdílené uložiště, nebo Pentaho ve své community edici nepodporuje nastavení rozvrhů spouštění. Další rozdíl je v licenční politice, kde Hitachi je nabízeno i ve verzi s community edition, která je zdarma a splňuje v omezené míře požadavky na vývoj datového skladu.

3.4 CASE nástroje

CASE nástroje neboli Computer Aided Software Engineering je „skupina počítačových nástrojů pro podporu analýzy, návrhu a implementace informačních systémů a informačních a komunikačních technologií, ale i dalších činností, souvisejících s vývojem IS/ICT2. Nástroje typu CASE umožňují zachycení modelů zkoumaného světa ideálně na třech úrovních – konceptuální, technologické a implementační. Dalšími výstupy CASE systémů jsou dokumentace IS/ICT a programové kódy a skripty pro definici obsahu datové základny.“ [43]

Z definice vyplývá, že to jsou takové nástroje, které mají za úkol mimo jiné nám pomoci s vytvořením modelů, který pro tvorbu datového skladu potřebujeme. Ale to je jen jedna z mála funkcionalit těchto nástrojů. Tyto nástroje neslouží pouze k vizuálnímu zobrazení modelu, ale velmi často i ke generování skriptů, které vytvoří stejné schéma tabulek i následně v databázi. Rovněž nám pomáhájí s tvorbou dokumentace, technickou kontrolou a i třeba k samotnému generování ETL, které jsou jedním ze stavebních kamenů datových skladů. Na trhu je velké množství těchto nástrojů. Často nabízejí více druhů licencí a to například licence na uživatele a nebo takzvanou floating licenci (licence, která není pevně vázaná na konkrétního uživatele). Mezi zástupce těchto nástrojů patří například:

3.4.1 SAP - PowerDesigner

PowerDesigner je nástroj pro vytváření modelů business procesů a koncepčních, logických a fyzických datových modelů. PowerDesigner dokáže

vytvořit z fyzického modelu DDL scripty, pro vytvoření konkrétní databáze. Rovněž dokáže opačný proces a to z DDL scriptů vytvořit model. Mezi hlavní výstupy nástroje patří schémata entity-relationship (ER), zprávy o analýze dopadu změn návrhu a standardní nebo vlastní zprávy o všech objektech v návrhu (tabulky, pole, vztahy).

- Cena za uživatele je 62 984 Kč [44]
- Cena za floating licence 157 450 Kč [44]

3.4.2 Sparx Systems - Enterprise Architect

Sparx Systems Enterprise Architect je založen na UML. Platforma podporuje návrh a konstrukci systémů, modelování podnikových procesů a datové modelování průmyslových domén. Vedle modelování dat tento nástroj pokrývá základní aspekty životního cyklu vývoje aplikací od řízení požadavků až po fáze návrhu, konstrukce, testování a údržby. Poskytuje také podporu sledovatelnosti, řízení projektů a řízení změn, jakož i nástroj pro vývoj kódu.[45]

- Cena za uživatele je 15 888 Kč [46]
- Cena za floating licence 20 434 Kč [46]

3.4.3 Embarcadero - ER/Studio Data Architect

Tento nástroj pracuje napříč více databázovými platformami. Nástroj využívají například vývojáři, architekti databází a obchodní analytici. ER/Studio Data Architect se zejména používá k vytváření a správě databází, tvorbě dokumentací atd. ER/Studio je jedním z nástrojů komplexního modelování dat, který kombinuje datové podnikání a modelování aplikací do víceúrovňového návrhového prostředí. ER/Studio XE3 obsahuje kromě ER/Studio architekturu ER/Studio (je to nástroj pro modelování procesů, který dokumentuje obchodní procesy a umožňuje společnostem pochopit a zlepšit vztah mezi obchodními procesy a daty.) ER/Studio repository (Serverový systém pro správu modelů, který pomáhá firmám ušetřit každodenní obtíže s modelování dat v týmovém prostředí).[45]

- Cena za uživatele je 33 413 Kč [47]

Porovnání

Výše zmíněné nástroje pokrývají funkčnost, kterou po nich můžeme vyžadovat při tvorbě datových skladů.

3.5 BI nástroje

Existuje mnoho BI nástrojů, které pomahají firmám s analytickými problémy. Firma Gartner ve svém grafu 3.4 zobrazuje nejdůležitější zástupce z řad těchto nástrojů.

Obrázek 3.4: 2018 Gartner Magic Quadrant for BI and Analytics převzato[7]

Pro svoji práci jsem si vybral Microsoft Power BI a Tableau protože patří podle Gartnerova quadrantu mezi nejlepší na poli BI aplikací.

3.5.1 Microsoft Power BI

Microsoft Power BI je sada analytických a vizualizačních nástrojů pro analýzu a vizualizaci obchodních dat v reálném čase. Rovněž dokáže načítat data ze všech možných druhů databází.[48]

3.5.2 Komponenty Microsoft Power BI

Microsoft Power BI obsahuje následující komponenty:

- Power Pivot

  „Power Pivot umožňuje tvorbu a správu relačního datového modelu, obsahuje též DAX jazyk. Je součástí aplikace Excel i Power BI Desktop a je velmi zajímavou a pro většinu uživatelů, bohužel, skrytou možností.“[49]

- Power Query

  „Power Query pro Excel 2010 až 2016 (Get & Transform) a také pro Power BI Desktop může řešit spousty každodenních problémů s načítáním dat z různých zdrojů, jejich očišťování o nepotřebné znaky, sloupce, atd. Pohodlně a hlavně rychle můžeme data nachystat pro Excel, případně Power Pivot datový model. Vše lze udělat elegantně klikáním nebo s přímým využitím M jazyka. Nástroj umí využívat Azure prostředí (mimo jiné machine learning funkce).“ [49]

- Power view & Power Map

  „Nad již vydolovanými a následně takzvaně učesanými daty je někdy potřeba udělat tu správnou vizualizaci a obyčejné grafy v Excelu mnohdy nestačí. Co do interaktivity a přehlednosti vedou právě Power View a Power Map (3D Map v Excel 2016).“[49]

- Power BI Desktop

  „samostatná desktopová aplikace, která „integruje“ všechny již zmíněné nástroje“[49]

- Power BI Server

  „on-premises řešení pro firemní reporting“[49]

3.5.2.1 Licence Microsoft Power BI:[48]

Author	**Share and collaborate**	**Scale large deployments**
Power BI Desktop	Power BI Pro	Power BI Premium
0	227 Kč/měsic	Cena na míru

„Power BI report Server je zdarma pro ty, kteří mají SQL Server Enterprise se Software Assurance nebo využívají Power BI Premium službu.“[49]

3.5.3 Tableau

Tableau BI je nástroj pro vizualizaci a analýzu dat pro všechny typy organizací a firemních uživatelů. Díky jednoduchým funkcím mohou uživatelé snadno analyzovat klíčová data, sdílet kritické informace v podniku a vytvářet nové vizualizace a reporty. Navíc Tableau nabízí možnost vložit dashboardy do stávajících podnikových aplikací, jako je SharePoint, pro rychlou analýzu.[50]

3.5.3.1 Komponenty Tableau:

Tableau má následující komponenty:

- Tableau Desktop

  „je výchozí produkt pro firemní analytiky, controllery či kohokoliv, kdo se potřebuje napojit na data a vytvářet reporty či dashboardy. Jedná se o klientskou aplikaci, jejíž součástí jsou předpřipravené datové konektory pro více než 50 datových zdrojů“[51]

- Tableau Server

  „Tableau Server umožňuje sdílet a řídit přístup k dashboardům, které vytvořili jiní uživatelé v desktop aplikaci. Nabízí automatickou publikaci dashboardů do PDF formátu a zejména publikaci do html prostředí včetně podpory přístupů z mobilních zařízení“[51]

- Tableau Online

  „Tableau Online umožňuje stejně jako Tableau Server sdílet a řídit přístup k dashboardům, které vytvořili jiní uživatelé v Tableau Desktop. Jedná se o cloudovou službu, tj. Tableau Server hostovaný přímo společností Tableau.“[51]

3.5.3.2 Licence Tableau:

- Creator - 19 098 Kč bez DPH/rok

  „Uživatel, který má na přístup do databází, zná strukturu dat v nich a vytváří buďto datový model pro další uživatele (pospojuje správné tabulky, vyčistí data a nahraje model na sdílený prostor) nebo tvoří i finální analýzy, reporty, dashboardy (které ukládá na sdílený prostor). Pracuje v desktopovém či webovém prostředí. Licence na jméno.“[51]

- Explore - 9 546 Kč bez DPH/rok

  „Uživatel, který tvoří vlastní analýzy z již vytvořeného datového modelu nebo může pouze „konzumovat“ obsah vytvořený někým jiným. Pracuje pouze ve webovém prostředí. Licence na jméno, přičemž minimální počet je 5 uživatelů.“[51]

- Viewer - 3 273 Kč bez DPH/rok

  „Uživatel, který „konzumuje“ obsah vytvořený někým jiným. Licence na jméno, přičemž minimální počet je 100 uživatelů.“[51]

Porovnání

Porovnávání těchto nástrojů je velmi složité, protože oba nástroje, mají podobnou funkcionalitu. Při finálním výběru záleží nejvíc na aktuální ceně produktu a na to od jakého technologického gigantu banka chce tento software.

3.6 Závěr kapitoly o technologiích a nástrojích

V této kapitole byl řečen úvod do technologií a nástrojů, které se využívají v datových skladech. Nejprve byly řečeny podrobnosti o hardwaru pro datový sklad a možných variantách. Následně byl popsán aktuální stav na poli databází a několik zástupců vybraných pomocí Gartnerovského Magic Quadrantu. Dále pak jsem se zaobíral zástupci z řad nástrojů pro datovou integraci, case a BI nástrojů.

Kapitola 4

Přehled poskytovatelů DWH

V následující kapitole se čtenář seznámí se stručným popisem firem, které podnikají na trhu s datovými sklady. Popis bude obsahovat informace o historii těchto firem a technoligiích, které využivají. Na světě existuje mnoho firem, které se zabývají problematikou datových skladů a BI. Proto se ve své práci zaměřím pouze na poskytovatele působící na českém trhu.

Všechny níže zmíněně firmy se zaměřují v oblasti datových skladů zejména na: [52]

- Kompletní vývoj nových datových skladů, operačních databází a specializovaných datových tržišť.
- Analýzu (Assesment) stávajících řešení na úrovni využití technologií, analýzu datové kvality stávajících dat a nebo procesů spojených s provozováním řešení.
- Služby spojené s rozvojem Information a Data Governance v organizaci, včetně návrhu a zavedení potřebných metodik a procesů.
- Vysoce odborné technologické služby pro Datawarehouse a BI oblasti – optimalizace řešení, migrace na nové verze, dodávky technologií, školení, podporu provozu v kritických situacích aj.
- Projekty datové migrace mezi systémy.
- Projekty integrace dat a Master Data Management.
- Realizace a rozvoj reportingových řešení.

4.1 Profinit

Firma Profinit působí na českém trhu už od roku 1998. Profinit působí na poli aplication outsourcing a information management. Zaměřuje se zejména na oblast vývoje softwaru na zakázku, datových skladů a business inteligence pro zákazníky po celé Evropě ale i v USA. V teamu profinit paracuje přes 400 zaměstnanců. Obrat profinitu v roce 2016 činil necelých 500 mil. Kč. Zaměřuje se na segment Finance a Telco.

Firma Profinit se zaměřuje hlavně na technologie:[52]

- IBM
- Oracle
- Informatica
- Microsoft
- Teradata

4.2 Adastra

Adastra je mezinárodní konzultační společnost, která dodává funkční odvětvová IT řešení. Od svého vzniku v roce 2000 se zaměřuje na zpracování dat, jejich analýzu a budování datových skladů. Hlavní kompetence se postupně rozšířily o oblasti Internetu věcí (IoT), Big Dat, umělou inteligenci nebo vývoj mobilních aplikací. Cílem společnosti je prostřednictvím chytrých datových řešení přispívat k rozvoji byznysu zákazníků. Patří mezi ně přední globální i lokální společnosti z financí a bankovnictví, pojišťovnictví, telekomunikací, obchodu, automobilového průmyslu a řady dalších odvětví včetně státní správy. Na 1500 konzultantů Adastry realizuje projekty po celém světě z 12 kanceláří. Obrat společnosti v roce 2017 překročil 3 miliardy korun.

Firma Adastra se zaměřuje hlavně na technologie:[53]

- Oracle DB
- Miscrosoft SQL Server
- Sybase
- Teradata
- PostgreSQL
- Exdata

- Oracle TimesTen
- Hp Vertica
- MongoDB
- Netezza

4.3 Sophia Solutions

Firma Sophia Solutions byla založena roku 2002 Janem Kadlecem. V porovnání s dvěma přechozími je to relativně malá firma, kde počet zaměstnanců je v řádech desítek. Jedna část této firmy se rovněž zabývá datovými sklady.

Firma Sophia Solutions se zaměřuje hlavně na technologie:[54]

- SAP
- Oracle
- Informatica
- Microsoft

4.4 Závěr analýzy trhu

Po analýze trhu můžeme řici, že většina firem nabízí vesměs stejné technologie, rozdíl bude pouze ve zkušenostech, které tyto firmy mají. Je jasné, že menší firmy si mohou jen těžko dovolit jít do velkých DWH projektů, které jsou velkou zátěží na cashflow neboli tok hotovosti(pokud má projekt pevnou cenu, kterou zákazník zaplatí na konci nebo v určité části projektu, musí mít dodavatel dostatek svých zdrojů, ze kterých zaplatí práci teamu). Pro banku je tedy důležité najít vhodné metriky, které budou rozhodovat, jakou konkrétní nabídku si vyberou. Proto je i výstupem mé práce průvodce, který by měl toto rozhodování ulehčit.

Kapitola 5

Vytvoření průvodce

V této kapitole se čtenář seznámí s popisem vytvořeného průvodce. Dále pak s kladenými otázkami, jejich odpověďmi a algoritmi pro stanovení vah kritérií.

5.1 Funkcionalita průvodce

Průvodce má následující funkcionalitu:

- Poradce, který porovná nabídky a určí jejich výhodnost.
- Ruční změna koeficientů jednotlivých odpovědí.
- Ruční změna vah otázek.
- Změna vah pomocí Saatyho algoritmu.
- Změna vah pomocí Fullerova algoritmu.

Tato funkcionalita je naimplemetována na http://dp2019.borec.cz.

5.2 Otázky v průvodci

Otázky jsou rozděleny do dvou sekcí. Obě sekce vyplňuje banka sama. Otázky v první sekci se týkají požadavků banky na projekt. Druhá sekce se týká už samotných dodavatelů a jejich nabídek.

5.2.1 První sekce

Tato sekce je zaměřena na banky.

- Délka projektu: (v měsících)

  Odhad délky projektu ze strany zadavatele.

- Rozpočet projektu

  Jaký je rozpočet projektu.

5.2.2 Druhá sekce

Druhá sekce otázek obsahuje otázky, které se týkají konkrétních firem a jejich nabídek na realizaci.

- Název firmy
- Jaké zkušenosti máte s danou firmou
  - vynikající
  - dobré
  - bez problému
  - špatné
  - žádné

  Pokud má daná organizace s jednou dodavatelskou firmou lepší zkušenost než s jinou, je jasné, že ji bude ve výsledku více preferovat než druhou, se kterou nemá tak dobré zkušenosti.

- Kolik má firma odborníků na DWH

  Vytvoření datového skladu je finančně nákladné. Pokud se jedná o fix price nabídku, firma potřebuje relativně dost velký interní budget, aby byla schopná ufinancovat vývoj. Jeli to menší firma, je tu možnost, že ji dojdou peníze a projekt nedokáže dodat.

- Má firma zkušenosti už s podobným projektem
  - ano
  - ne

  Pokud má dodavatelská firma zkušenosti s podobným projektem, má i více know how a je pravděpodobnější, že lépe odhadne úskalí a rizika projektu.

- Nabídka lidí se zkušenostmi s danou technologií DB
  - ano

  - ne

  Je lepší, když na projektu pracují lidé, kteří už pracovali s danou technologií.

- Nabídka lidí se zkušenostmi s danou technologií datové integrace
  - ano
  - ne

  Je lepší, když na projektu pracují lidé, kteří už pracovali s danou technologií.

- Nabídka lidí se zkušenostmi s danou technologií BI
  - ano
  - ne
  - nevím

  Je lepší, když na projektu pracují lidé, kteří už pracovali s danou technologií.

- Model připravený z jiných projektů
  - ano
  - ne
  - nevím

  Otázka trochu navazuje na jednu z předchozích, která se ptá na zkušenost firmy s podobným projektem. Jeli databázový model připravený a vyladěný z předešlých projektů, je menší pravděpodobnost chyb. Pokud se například jedná o tří vrstvou architekturu, je jasné, že když použijeme model z předešlého projektu, bude se nejspodnější a nejsvrchnější vrstva lišit, ale právě prostřední vrstva by měla být téměř totožná, protože core business banky je všude stejný.

- Fix price nabídka
  - ano
  - ne

  Pokud dodavatelská firma nabídne svoji dodávku za fixní cenu, značí to, že firma si je jistá délkou a náročností projektu. A je zde nižší riziko zvýšení nákladů na projekt.

- Nabídka na support
  - ano
  - ne

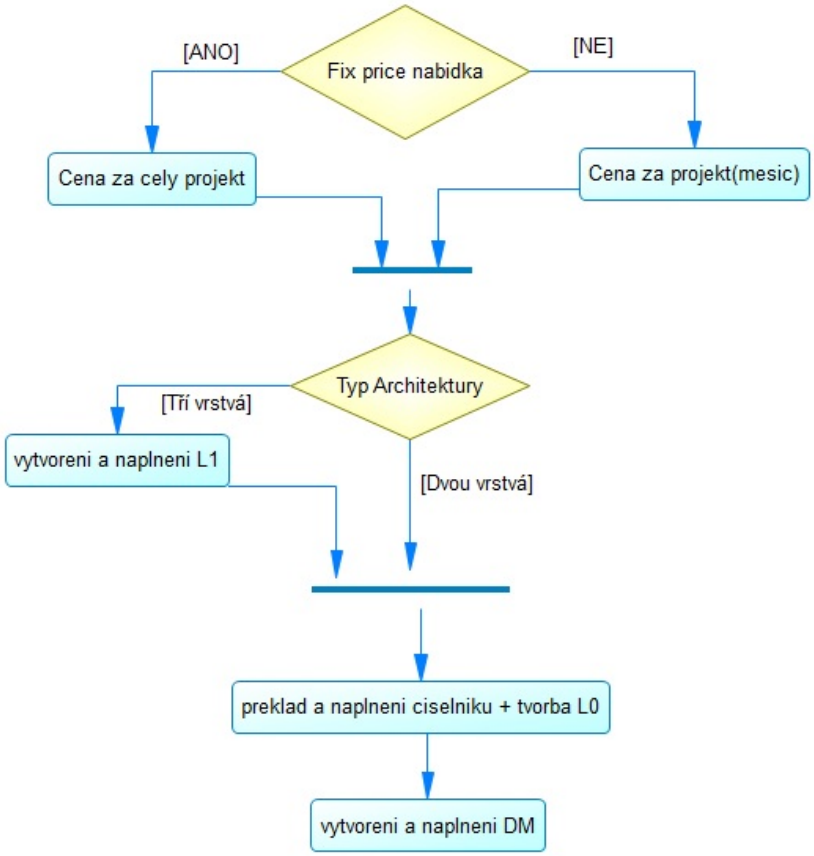
  Pokud firma nabízí po dodání řešení i možnost supportu, je to pro zadavatele lepší, než pokud nabízí pouze vývoj a nasazení. Pak si následný support bude muset firma zajišťovat sama nebo následně hledat dalšího dodavatele.
- Typ architektury
  - dvouvrstvá
  - třívrstvá

  Otázka pro výběr dalších otázek v průvodci.

Na další straně jsou otázky ohledně odhadované pracnosti projektu.

- Předpokladaný počet lidí na projektu
- Odhadovaná pracnost překladu, naplnění číselníků a celková tvorba nejspodnější vrstvy
  - MD
- Odhadovaná pracnost na vytvoření a naplnění prostřední vrstvy(L1) - pokud se jedná o 3 vrstvou architekturu
  - MD
- Odhadovaná pracnost na vytvoření a naplnění DM
  - MD
- Odhadovaný potřebný čas na první relevantní výstup dat z DM
  - v měsících
- Odhadovaná délka projektu bez supportu
  - v měsících
- Odhadovaná cena za projekt
  - cena za celý projekt pokud se jedná o fix price
  - cena za měsíc pokud se nejedná o fix price

V průvodci jsou občas implementované otázky, které později slouží k správně zvolené podotázce, tyto otázky jsou vidět na diagramu 5.1.



Obrázek 5.1: Základní rozhodovaní o podotázkách v průvodci

5.3 Další vstupy do průvodce

Průvodce nyní používá konfiguranční soubor: konf.txt, který se nachází v kořenovém adresáři. Tento textový soubor obsahuje mnou definované váhy jednotlivých odpovědí, které budou rozebrány v kapitole 5.5. Pokud zadavatel chce změnit jednotlivé váhy podle svých preferencí, je to možné na stránce http://dp2019.borec.cz/koeficient.php, kde se mohou změnit jednotlivé váhy daných odpovědí pomocí webového formuláře. Hodnoty jsou pozměněny pouze v jednom běhu programu. Pokud se program pustí znovu, hodnoty budou opět defaultní.

5.4 Technologie použité k naprogramování průvodce

Rozhodoval jsem se nad několika variantami, jak pojmout daného průvodce. Varianty byly následující:

- Desktopová aplikace

  Průvodce naprogramovat jako desktopovou aplikace buď pomocí jazyka Java nebo C++.

- Průvodce v Microsoft Excel

  Průvodce koncipovat jako Excel file, kde by byly predvyplněny otázky a pomocí vzorců by se napočítávaly různé koeficienty.

- Webový formulář

  Průvodce pojmout jako jednoduchý webový formulář vytvořený v HTML a pro výpočty použít php.

Po konzultaci s vedoucím práce jsme vybrali variantu webového formuláře, který je uživatelsky asi nejpřívětivější a rovněž lehce přenositelný. Jako programovací jazyk jsem si vybral HTML a logiku jsem zapouzdřil pomocí PHP skriptů. Výsledné grafy jsou pak zobrazeny pomocí Javascriptu.

5.5 Stanovení počátečních vah[1]

Existuje mnoho metod jak stanovit váhy jednotlivých kritérií. Já ve svém průvodci mám naimplementovány tyto metody dvě a to:

- Fullerovu metodu
- Saatyho metodu

Tyto metody jsem si vybral, protože obě jsou postavené na párovém porovnání tzn. porovnávám vždy pouze 2 kritéria a vybírám, které má pro mě větší váhu. Fullerova metoda je pro uživatele, který daný test vyplňuje jednoduší, protože pouze označí kritérium, které je pro něj hodnotnější. Saatyho metoda je jakousi nadstavbou Fullerovy, protože uživatel udává i jak hodně je dané kritérium pro něj důležitější. Saatyho metoda by měla být tedy mnohem přesnější než Fullerova, ale za cenu složitějšího a časově náročnějšího vyplňování testu.

5.5.1 Fullerova matoda párového srovnání

Princip této metody je vytvoření všech různých dvojic kritérií, kde se vždy vybere jedno kritérium, které je důležitější. Celkový počet dvojic je tedy:

$$pocet = \frac{k * (k-1)}{2}$$

Výsledkem tohoto porovnání je čtvercová matice $k*k$, kde k je počet kritérií. Na diagonále této matice jsou jedničky. Pokud kritérium K_i je důležitější než kritérium K_j. Potom hodnota na pozici $K_{i,j}$ v matici se rovná jedné a hodnota na pozici $K_{j,i}$ se rovná nule.

Následně v této matici vypočítám počet bodů pro dané kritérium, dále pak pro každé toto kritérium vypočítám jeho váhu:

$$vaha = \frac{krit}{celkem}$$

Kde krit je počet bodů, které kritérium získalo a celkem je celkový počet bodů všech kritérií. Díky tomu získám konkrétní váhu daného kritéria a to v hodnotách od 0 do 1, kde součet vah všech kritérií je roven 1.

5.5.2 Saatyho matoda kvantitativního párového srovnání

Saatyho metoda patří k nejčastěji používaným metodám pro volbu vah jednotlivých kritérií. Tato metoda je stejně jako předešlá, postavena na porovnávání dvou kritérií. Následné hodnoty se ukládají do tzv. Saatyho matice. Hodnocení dvou porovnávaných kritérií je následující:

- 1 - kritéria jsou stejně významná
- 3 - první kritérium je slabě významnější než druhé
- 5 - první kritérium je silně významnější než druhé
- 7 - první kritérium je velmi silně významnější než druhé
- 9 - první kritérium je absolutně významnější než druhé
- 3 - druhé kritérium je slabě významnější než první
- 5 - druhé kritérium je silně významnější než první
- 7 - druhé kritérium je velmi silně významnější než první
- 9 - druhé kritérium je absolutně významnější než první

Saatyho matice se následně vyplňuje tak, že v matici do příslušné pozice např. pokud kritérium K_i je slabě významější než kritérium K_j, pak v matici na $pozici_{i,j}$ je hodnota 3 a na $pozici_{i,j}$ je hodnota $\frac{1}{x}$, kde x je v tomto případě hodnota 3. Na diagonále opět jako v předešlém případě jsou jedničky.

Následně pak v této matici vynásobím každý člen v řádku $s_i{=}\prod_{j=1}^{n} s_{i,j}$. Dále pak tuto hodnotu odmocním $R_i{=}\sqrt[k]{s_i}$, kde K je počet kriterií. Z této hodnoty už mohu vypočítat výslednou váhu pro dané kritérium: $v_i{=}\frac{R_i}{\sum_{i=1}^{k} R_i}$.

s_{ij}	f1	f2	f3	f4	f5	$s_i = \prod_{j=1}^{5} s_{ij}$	$R_i = (s_i)^{1/5}$	$v_i = R_i / \sum_{i=1}^{5} R_i$
f1	1	1/5	1/5	1/3	1/2	1/150	0.367	0.054
f2	5	1	1	4	5	100	2.512	0.373
f3	5	1	1	4	5	100	2.512	0.373
f4	3	1/4	1/4	1	2	3/8	0.822	0.122
f5	2	1/5	1/5	1/2	1	1/25	0.525	0.078
						součet	6.738	1

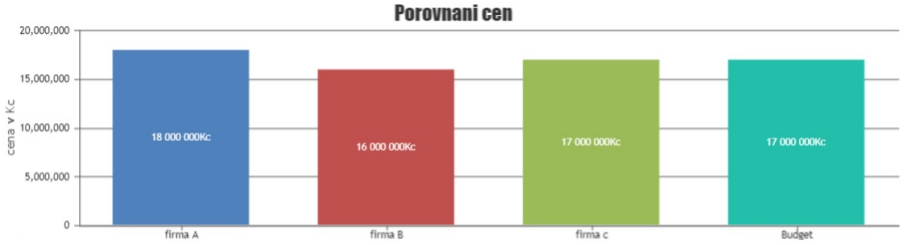
Obrázek 5.2: Ukázka Saatyho matice převzato[1]

5.6 Výstupy z průvodce

Za nejvhodnější výstup průvodce jsem zvolil několik grafů zobrazujících různé metriky, za kterými následuje stručný popis toho co zobrazuje graf.

5.6.1 Porovnání cen

Tento graf porovnává jednotlivé cenové nabídky od potencionálních dodavatelů. Pokud není nabídka od dodavatele jako fix price, ale je formou měsíční sazby, vypočítám si výslednou cenu projektu tak, že vynásobím cenu za měsíc očekávanou délkou projektu a následně tuto hodnotu zvýším o 10%. Protože průměrné zvýšení nákladů na projekt je okolo 10%[55].



Obrázek 5.3: Porovnání cenových nabídek z poradce

5.6.2 Celková pracnost projektu

Tento graf zobrazuje celkovou pracnost projektu, tj. součet všech dílčích pracností, která firma rozepsala ve své nabídce.

Obrázek 5.4: Porovnání odhadované pracnosti jednotlivých dodávek

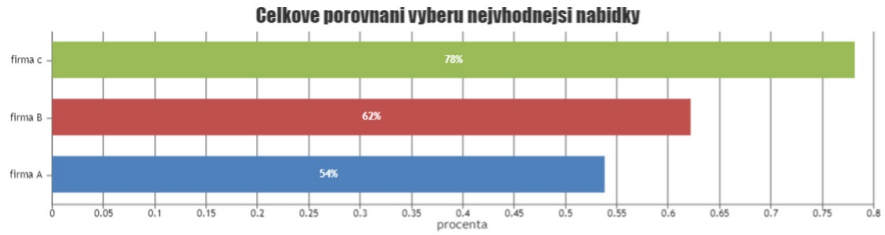
5.6.3 Kdy budou první relevantní výsledky z data martů

Tento graf vizuálně zobrazuje odhadovanou délku práce, která je potřeba k výstupu relevantních výsledků z jednotlivých data martů.

Obrázek 5.5: Kdy budou první relevantní výsledky z data martů

5.6.4 Celkové porovnání nabídek

Tento graf zobrazuje porovnání celkové výhodnosti jednotlivých nabídek. Hodnoty grafu vychází z několika zdrojů.



Obrázek 5.6: Graf vhodnosti jendotlivých nabídek

Hodnoty v tomto grafu pro jednotlivé dodavatele značí, kolik procent z maximálního počtu bodů získá jejich nabídka. Jedním zdrojem jsou odpovědi na tyto otázky s počtem bodů, které jednotlivé otázky mají.

- Zkušenosti s firmou
  - vynikající- 1
  - dobré - 0,8
  - bez problému - 0,5
  - špatné - 0
  - žádné - 0,4
- Zkušenosti firmy s podobným projektem
  - ano - 1
  - ne - 0
  - nevím - 0,5
- Připravený model z jiných projektů
  - ano - 1
  - ne - 0
  - nevím - 0,5
- Nabídka lidí se zkušenostmi s danou technologií DB
  - ano - 1
  - ne - 0

- Nabídka lidí se zkušenostmi s danou technologií datové integrace
  - ano - 1
  - ne - 0
- Nabídka lidí se zkušenostmi s danou technologií BI
  - ano - 1
  - ne - 0
- Nabídka supportu
  - ano - 1
  - ne - 0
- Nabídka fix price
  - ano - 1
  - ne - 0
- Kolik má firma odborníků na datové sklady

  $$\frac{pocet}{max}$$

  Pod proměnnou pocet se skrývá počet zaměstnanců dodavatele se zkušenostmi s datovým skladem. Max je maximalní počet z nabídek.
- Cenová nabídka

  $$\frac{min}{nabidka}$$

  Pod proměnnou min se skrývá minimální cena nabídky a pod proměnnou nabidka cena.

Takto jsou nastavené defaultní hodnoty jednotlivých odpovědí. Tyto hodnoty, až na vzorce, se mohou měnit pomocí webového formuláře na: http://dp2019.borec.cz/koeficient.php a následně se tyto pozměněné hodnoty použijí pro výpočet vhodnosti nabídky.

Dalším vstupem jsou jednotlivé váhy daných otázek. Tyto hodnoty jsou defaultně dané v průvodci. Hodnoty vychází z testování několika respondentů. Jednotlivé výsledky konkrétních testů jsou na přiloženém CD.

- Zkušenosti s firmou 0,151
- Kolik má firma zaměstnanců se zkušenostmi s DWH 0,033
- Zkušenosti firmy s podobnými projekty 0,142
- Připravený model z jiných projektů 0,082
- Nabídka supportu 0,078
- Nabídka lidí se zkušenostmi s danou technologií DB 0,111
- Nabídka lidí se zkušenostmi s danou technologií datové integrace 0,104
- Nabídka lidí se zkušenostmi s danou technologií BI 0,078
- Nabídka fix price 0,074
- Celková cena nabídky 0,147

Pokud uživatel průvodce bude chtít měnit jednotlivé váhy podle svých preferencí, jsou v průvodci naimplementované dva algoritmy a to Fullerův a Saatyho. Lze je využít pro vygenerování nových vah podle konkrétního uživatele a jeho preferencí.

5.7 Závěr kapitoly o vytvoření průvodce

V této kapitole se čtenář seznámil se mnou vytvořeným průvodcem, s jeho funkcionalitou a popisem výstupů. Průvodce byl vytvořen na základě poznatků získaných během tvorby teoretické části diplomové práce. Průvodce bude následně otestován v kapitole 7.

Kapitola 6

Případová studie

Tato kapitola se bude zabývat případovou studií. Ceny jsou získány z veřejných ceníků. Pokud jsou originální ceny v cizí měně, používám kurz podle ČNB platný k 5. listopadu 2018 a to 22,73 Kč za 1 USD.

6.1 Co je případová studie

Případová studie je: „detailní studium jednoho případu nebo několika málo případů“ [56]

6.2 Případová studie o datovém skladu

V této kapitole se budu zabývat vytvořením případové studie, zejména technologické a finanční části. Pro výpočet finanční náročnosti jsem použil TCO neboli celkové náklady na vlastnictví.

6.2.1 O zákazníkovi

Tato banka je soukromá, netvoří žádné konzorcium s dalšími finančními subjekty. Ve své zemi je 5. největší a má podíl na trhu něco okolo 10 procent. Banka vznikla transformací státní banky. Hlavním businessem pro tuto společnost jsou takzvané mikroúvěry tj. menší půjčky bez ručení na kratší dobu. Další klíčovou složkou jsou půjčky, které jsou kryté statními dávkami, jako jsou například důchody. Tato finanční instituce uvažuje o implementaci datového skladu nejprve pro oddělení Risk managementu, který posílá reporty do Národní banky. A také pro oddělení Centrál Reportingu, které slouží pro reportování aktuálního stavu členům Boardu. Tato banka má okolo 70 produktů a cca 1 mil. klientů, kteří využívají uvěrové protfolio.

6.2.2 Aktuální stav

V následujicích podkapitolách bude popsán aktuální stav, ve kterém se banka nachází. Datový sklad bude stavěn na tzv. zelené louce, protože banka nemá žádný datový sklad.

6.2.2.1 Businessový stav

Banka v dnešní době připraví cca 1000 reportů ročně. Na těchto reportech pracuje 20 lidí. Drtivá většína těchto reportů je připravována manuálně, bez použití nějaké větší automatizace. Následná kontrola dat je už velice obtížná a proto přichází po vytvoření reportů dlouhá diskuze nad správností předložených dat. Velká část těchto reportů je následně posílána k regulátorovi trhu, tedy k místní národní bance. Napříč odděleními není zaveden standardizovaný slovník tzn. každá metrika se může v jiném oddělení počítat trochu jinak. Výsledkem toho je, že pokud porovnám 2 reporty na stejnou věc z oddělení Centrálního reportingu a například Risku, mohu získat jiná čísla. Minimální unifikace klienta. Jeden klient je v DB pod více záznamy. Například pokud do systému byl jednou zadán jako Jan Novak a podruhé jako Novak Jan apod. Špatná komunikace mezi lidmi z IT oddělení a lidmi z Bussinesu. Mnoho IT specialistů moc nerozumí faktickému významu dat, se kterými pracují, proto nastávají situace, že IT specialista něco v dobré víře změní, ale businessově to nedává smysl.

Datová kvalita

Není nastavený žádný proces na kontrolu kvality dat. Data jsou uložena v mnoha různých systémech nijak neošetřená.

6.2.2.2 Technologický stav

V této podkapitole bude popsán současný technologický stav věcí, který se přímo týká budoucího datového skladu.

Hardware

Aktuálně má banka několik databazových serverů, kde má poukládaná data. Servery má od různých dodavatelů.

Databáze

Banka nepoužívá jeden databázový systém. Můžeme zde vidět různé databáze:

- Microsoft SQL Server 2016
- Oracle 12.1. Enterprise Edition
- Oracle 12.1. SE

Prezentační nástroje

Banka nyní používá pro formu prezentaci dat v drtivé případě Microsoft Excel. Konkrétně verzi z roku 2016, která má následujicí funkcionalitu:

- Power Pivot
- Power View a Power Map
- Power Query

6.2.3 Plánovaný stav

V následujicí kapitole budou popsány varianty i s finanční náročností pro budoucí stav v bance.

6.2.3.1 Businessové požadavky

- Shromažďovat data ze zdrojů/primárních systémů na jednom místě. Konsolidace dat z těchto zdrojových systémů a vylepšení procesu vytváření reportu.
- Připravit data marty, ve kterých budou shromážděny všechny informace potřebné pro tvorby reportů riskového, central reportingového a finančního oddělení.
- Zkrátit čas a úsilí při vytváření pravidelných reportů.
- Vytvořit unifikovaného klienta.

6.2.3.2 Plánovaný technologický stav

V této kapitole budou rozepsány jednotlivé možné varianty týkající se použití technologií pro datový sklad.

Plánovaný HW pro DWH

Banka plánuje nákup IT infrastruktury od Oraclu. Konkrétně Exadata Database Machine X7-2 High Capacity (HC) Eighth Rack. Cena tohoto HW je 5 500 600 Kč.[57].

Licence pro HW

Tento stroj má 48 procesorů. Licence na jedno jádro stojí 397 775 Kč.[58].

Náklady na HW + licence na HW

Celkové náklady na hardware a s ním spojené databázové licence se pohybují okolo 24,5 mil Kč.

Varianty a ceny integračních nástrojů

Každý vývojář datového skladu bude potřebovat licenci na daný integrační nástroj. Banka preferuje využít Oracle Data Integrator EE. Cena licence je následující:

- Oracle Data Integrator EE
  - 20 457 Kč na uživatele[58]

Tento nástroj budou potřebovat v určité míře všichni členové projektu až na PM. Celková cena 9 licencí je tedy 184 113 Kč.

Varianty a ceny Case nástrojů

V úvahu pro využití Case nástrojů připadají všechny tři zmíněné varianty v kapitole 3.4. V této kategorii banka preferuje využít Case nástroj od firmy SAP. Tento nástroj bude opět potřeba pro všechny členy DWH teamu až na PM.

- Floating - 314 901 Kč
- Uživatel - 440 894 Kč

Celkem 755 795 Kč. Rozhodl jsem se zde kalkulovat se sedmi licencencemi na uživatele a zbytek floating.

Varianty a ceny prezentačních nástrojů

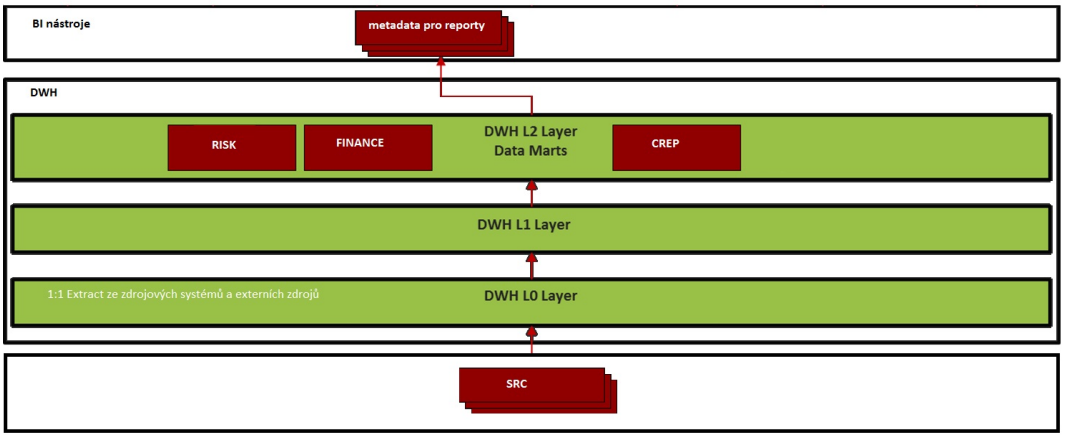
Banka chce zautomatizovat tvorbu reportů a také plánuje využít BI nástroje. Pro toto jsou vhodné všechny tři varianty nástrojů, které jsou rozebrány v kapitole o BI nástrojích. Je tu také možnost využít i stávající řešení ve formě Microsoft Excel 2016. Funkcionalita MS Excel by měla pokrýt nejzákladnější požadavky, ale nepřináší žádný bonus navíc.

Banka má, jak je již zmíněno v předchozí kapitole, servery, kde běží Microsoft SQL server 2016, jehož součástí je mimo jiné i Reporting Services, který lze využít pro prezentaci a nahlížení na výsledné reporty. Pro

vytváření reportů je potřeba mít nainstalovaný Power BI desktop, který je zdarma. V této kategorii se tedy banka rozhodla jít cestou Microsoft Power BI. Což díky aktuálnímu stavu v bance lze využívat zdarma.

6.2.4 Finance a pracnost

Banka ve svém projektu počítá s tím, že bude DWH sloužit pro odělení risku, financí a centrál reportingu. Při zaměřování velikosti datového skladu bylo zjištěno, že bude potřeba propojit cca 60 zdrojových systémů, ve kterých se nachází cca 180 různých databázových modelů. Bude také potřeba napojit a vyspecifikovat okolo 160 číselníků. To znamená, že v nejspodnější vrstvě datového skladu, tedy ve vrstvě, která je označována jako L0 neboli Staging Area, bude něco okolo 340 tabulek. Finální výstup bude ve 3 data martech. Bude implementována 3 vrstvá architektura:



Obrázek 6.1: Architektura datového skladu

- L0 - Staging

  Vrstva obsahuje denní snímky všech potřebných zdrojů 1:1.

- L1 - Date Warehouse

  Tato vrstva obsahuje standardizovanou referenční databázi. Zabezpečuje konsolidaci dat do jednotlivých struktur. Toto schéma může být převzato z jiných projektů, protože referenční modely jsou v bankovnictví velice podobné.

- L2 - Data Marts

  Tato vrstva obsahuje jednotlivé data marty, kam budou přistupovat uživatelé a odkud se berou data i pro BI nástroje.

6.2.4.1 Složení teamu

Team bude obsahovat tyto následující role i s průměrnou superhrubou mzdou [59]:

- Projekt Manager

  Náklady (superhrubá mzda za měsic): 107 000 Kč

- Architekt

  Náklady (superhrubá mzda za měsic): 134 000 Kč

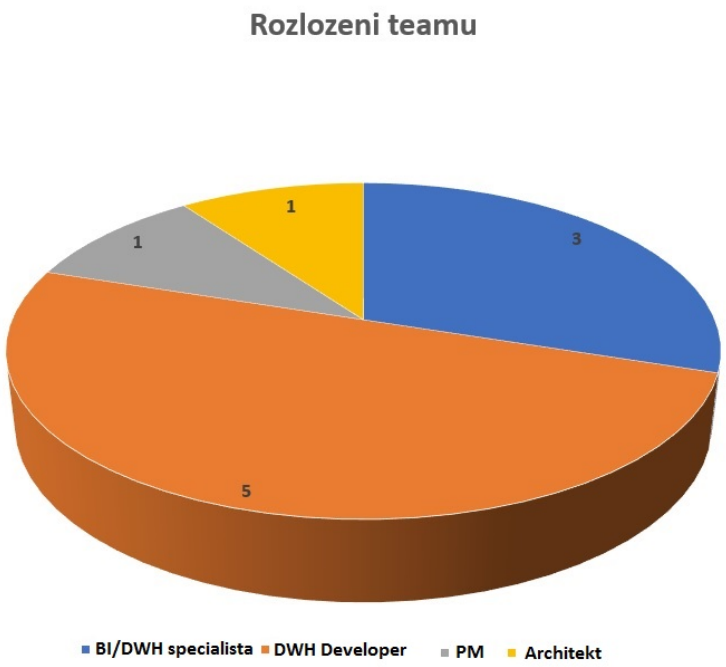
- BI/DWH Specialista

  Náklady (superhrubá mzda za měsic): 94 000 Kč

- DWH Developer

  Náklady (superhrubá mzda za měsic): 94 000 Kč

BI/DWH specialista se liší od DWH Developera tím, že pracuje na vrstvě L2 + vytváří BI reporty. DWH Developer pracuje na vrstvě L0 a L1.



Obrázek 6.2: Rozložení teamu

6.2.4.2 Úkony potřebné k implementaci datového skladu

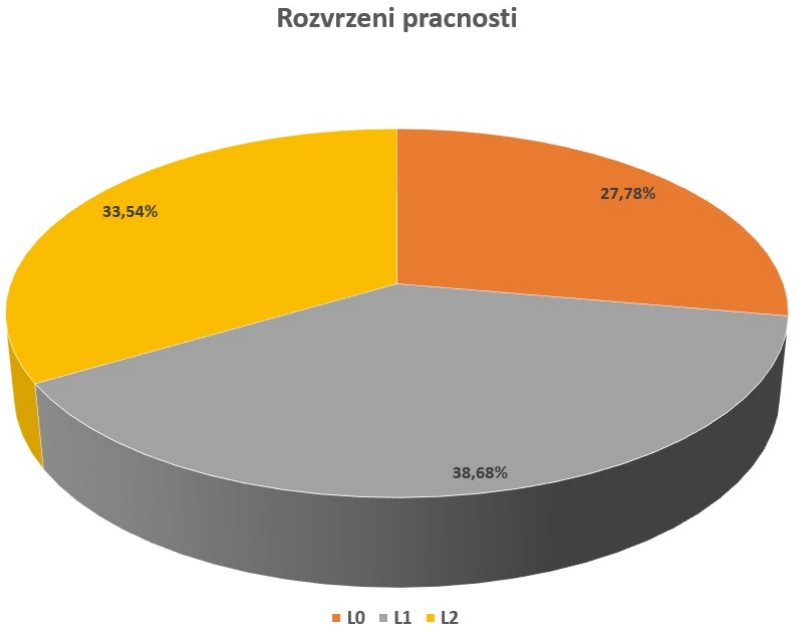
Zde jsou popsané základní úkony, které je potřeba udělat při vývoji datového skladu se zvolenou architekturou.

Vrstva	**Cinnost**	**Kdo**
Analýza zdrojů	L0	BI/DWH Spec + DWH Dev
Příprava vstupů do L0	L0	DWH Dev
Překlad a naplnění číselníků	L0	DWH Dev
Vytvoření tabulek v L0 + maping	L0	DWH Dev
Vytvoření ETL do L0	L0	DWH Dev
Vytvoření fakt. tabulek L1 + maping	L1	DWH Dev
Vytvoření dim. tabulek + maping	L1	DWH Dev
Vytvoření ETL do L1	L1	DWH Dev
Analýza výstupů z DM	L2	BI/DWH Spec
Vytvoření fakt. tabulek v L2 + maping	L2	BI/DWH Spec
Vytvoření dim. tabulek v L2 + maping	L2	BI/DWH Spec
Vytvoření ETL do L2	L2	BI/DWH Spec
Testování	L1+L2	BI/DWH Spec + DWH Dev

6.2.4.3 Pracnost jednotlivých vrstev

V tabulce 6.3 jsou uvedeny pouze odhadované pracnosti vrstev. Podrobný popis jednotlivých činností je vidět v přiloženém excel souboru kalkulace.xls na CD.

Vrstva	MD
L0	734
L1	1022
L2	886



Obrázek 6.3: Pracnost jednotlivých vrstev k celé náročnosti projektu

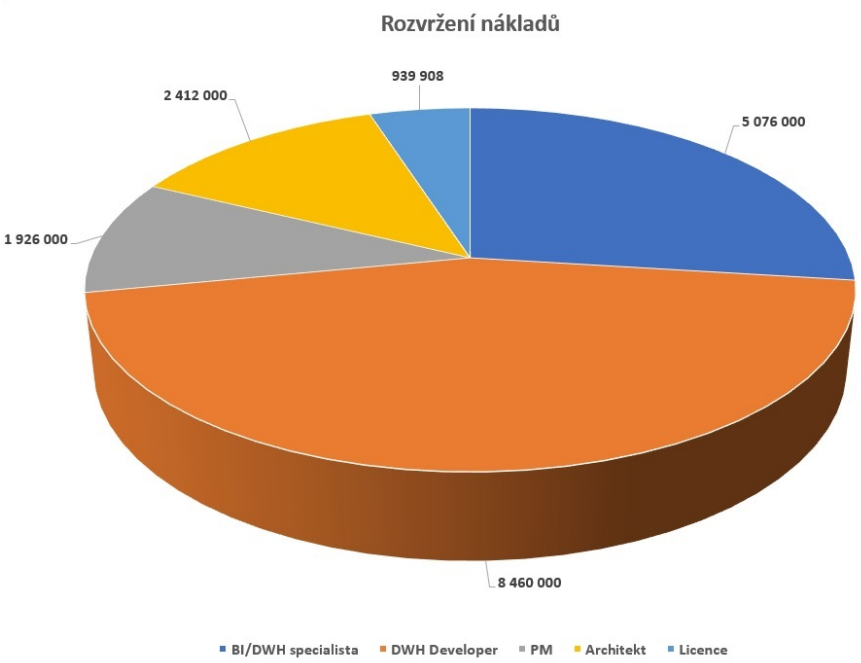
Celkový odhad pracnosti se pohybuje okolo 2650 MD což je pro 9-ti členný team plus projekt manažera práce na cca 15 měsíců, pokud počítáme, že průměrný pracovní měsíc má 20 dní. Odhadované náklady na pracovní sílu jsou tedy:

role	cena
Projekt Manager	1 605 000 Kč
Architekt	2 010 000 Kč
BI/DWH Specialisti	4 230 000 Kč
DWH Developeři	7 050 000 Kč

Celkové odhadované náklady na pracovní sílu jsou: 14 895 000 Kč.

6.2.4.4 Finální cena a délka projektu

Finální cena je započtáná bez HW a licencí na DB, které budou stát cca 24,5 mil Kč. Jako Case nástroj preferuje banka Power Designer od firmy SAP. Celková částka bude obsahovat ještě rezervu, což budu brát jako 20% z odhadované ceny práce, která je 14 895 000 Kč. Rezerva je tudíž 2 979 000 Kč. V této sumě se skrývá riziko možného prodloužení projektu, které jsem zvolil vyšší než při výpočtu ceny v průvodci, protože tento konkrétní projekt se zdá rizikovější než ostatní. Náklady na licence pro nástroje na datovou integraci, Case a prezentační nástroje jsou 939 908 Kč. Výsledná cena je tedy 18 813 908 Kč, s odhadovanou délkou projektu 15 měsíců. Do výsledné ceny nepočítám nákup hardwaru, protože není součástí dodávky.



Obrázek 6.4: Rozvržení nákladů

V tomto grafu je rozpuštěna rezerva do jednotlivých nákladů na pracovníky.

6.3 Závěr kapitoly o případové studii

V této kapitole se čtenář seznámil hlavně s finanční stránkou případové studie. S preferovanými variantami jednotlivých nástrojů i s jejich cenou, rovněž také s detailním rozpadem pracnosti jednotlivých úkonů potřebných k implementaci datového skladu.

Kapitola 7

Otestování průvodce nad daty získanými případovou studií a dodanými vedoucím práce

V této kapitole se čtenář seznámí s praktickým otestováním průvodce. Nejprve budou v kapitole popsány jednotlivé nabídky firem a poté budou následovat konkrétní výstupy z průvodce.

7.1 Nabídky

Zde jsou popsány nabídky i se základními charakteristikami. Tyto nabídky jsou uměle vytvořené pro otestování průvodce.

7.1.1 Firma Alpha

Firma Apha, která vytvořila nabídku na tento datový sklad, je firma o 50 zaměstnancích, kde se DWH zabývá 30 z nich. Banka s touto společností měla bezproblémovou spolupráci. Tato firma má zkušenosti s podobným projektem v jiné bance. Nabídka obsahuje také konkrétní osoby, které mají zkušenosti s danou technologií DB a BI, ale nikoliv s ODI, který je vybrán jako nástroj pro datovou integraci. Firma vvyužívá 2 vrstvou architekturu. Firma Alpha nemá z jiných projektů plně připravený model. Nabídka je koncipována jako fix price. Délka projektu je odhadnuta na 17 měsíců s tím, že na projektu bude pracovat 8 lidí. Předpokládaný počet MD na naplnění a vytvoření L0 je 800. Náklady na vytvoření požadovaných DM se odhadují na 1900 MD. Firma nenabízí následný support. První relevantní výstup z DM bude podle harmonogramu po 12-ti měsících. Celková cena za projekt je 17 523 000 Kč.

7.1.2 Firma Beta

Firma Beta má cca 20 zaměstnanců, kteří se zabývají problematikou DWH. Banka už jednou s touto firmou spolupracovala a má s ní vynikající zkušenosti. Rovněž tato firma, stejně tak jako předcházející, má zkušenosti s podobným projektem v jiné bance. Nabídka obsahuje osoby, které mají zkušenosti se všemi potřebnými technologiemi. Firma má připravený model z jiných projektů a používá v něm tří vrstvou architekturu. Nabídka je koncipována jako fix price. Délka projektu je odhadnuta na 13 měsíců s tím, že na projektu bude pracovat 10 lidí. Předpokládaný počet MD na vytvoření a naplnění L0 je 750 MD, L1 je 1000 MD a L2 je 800 MD. Firma nenabízí následný support. První relevantní výstup z DM bude podle harmonogramu po 10 měsících. Cena za celý projekt je odhadnuta na 22 035 000 Kč.

7.1.3 Firma Gama

Firma Gama má cca 30 zaměstnanců, kteří se zabývají problematikou DWH. Banka už jednou s touto firmou spolupracovala a má s ní špatné zkušenosti. Firma má zkušenosti s podobným projektem. Nabídka obsahuje osoby s potřebnými znalostmi daných technologií. Používá dvou vrstvý model. Nabídku nekoncipuje jako fix price. Předpokládaný počet lidí na projektu je 9 a předpokládaná délka projektu je 15 měsíců. Předpokládaný počet MD na naplnění a vytvoření L0 je 780. Náklady na vytvoření požadovaných DM se odhaduji na 1800 MD. Firma nenabízí následný support. První relevantní výstupy z DM se plánují v horizontu 11 měsíců. Navržené platby jsou v měsíčních cyklech a to v hodnotě 1 400 000 Kč.

7.1.4 Firma Delta

Firma Delta je disponuje 100 odborníky na DWH. Banka s touto firmou nemá žádné zkušenosti. Firma má zkušenosti s podobnými projekty a má experty na vybrané technologie. Firma využívá 3 vrstvou architekturu. Firma využije model z předešlých projektů. Podrobnosti k této nabídce jsou popsány v případové studii v předcházející kapitole. Odhad pro vytvoření a naplnění vrstvy L0 je 734 MD. Pro vrstvu L1 je to 1022 MD a pro vrstvu L2 neboli konkrétní DM je to 886 MD. První relevantní výstupy z DM se plánují v horizontu 12 měsíců. Firma nabízí nasledný support. Odhadovaná délka projektu je 15 měsiců a rozpočet je 18 813 908 Kč.

7.2 Vstupní data banky

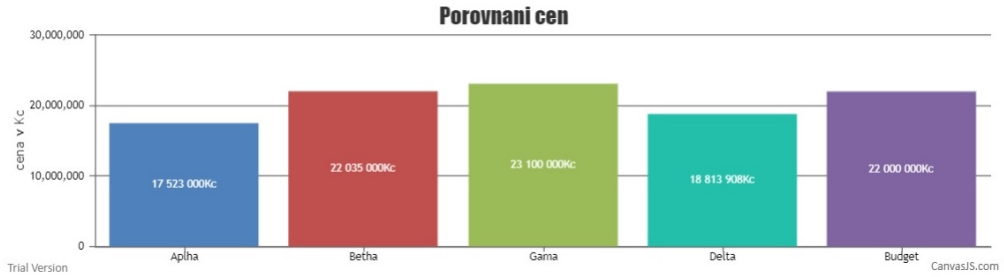
Banka plánuje na tento projekt vyčlenit rozpočet ve vyši 22 000 000 Kč a očekává ukončení tohoto projektu za 17 měsíců.

7.3 Výstupy z průvodce

V této kapitole budou popsány výstupy z vytvořeného průvodce.

7.3.1 Dílčí metriky

Jako první výstup z průvodce je graf finanční náročnosti jednotlivých řešení.



Obrázek 7.1: Finanční náročnost

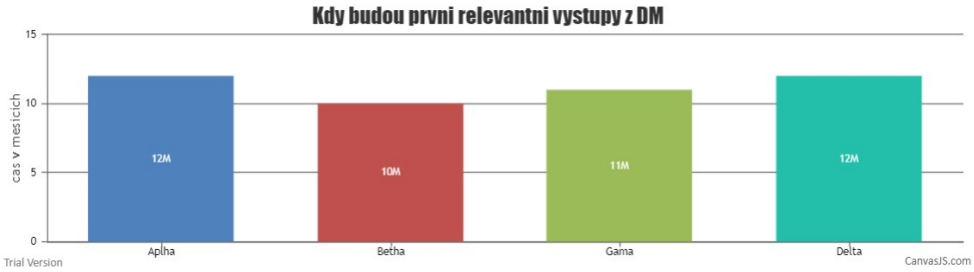
Na tomto grafu je vidět, že finanční náročnost nabídky od firmy Gama je nejdražší. Na druhém pólu je finační náročnost firmy Alpha. Firmy Betha a Gama překračují odhadovaný rozpočet o 35 000 respektivě 1 100 000 Kč.

Jako druhý výstup z průvodce je graf odhadu počtu MD na realizaci projektu.

Obrázek 7.2: Počet MD na realizaci

Na tomto grafu je vidět, že firma Aplha počítá s větší pracovní náročností než ostatní, přesto má firma nejnižší nabízenou cenu. S nejnižší pracností naopak počítá firma Betha, která odhaduje pracnost projektu na 2 550 MD, těsně v závěsu je firma Gama se svým odhadem 2 580 MD.

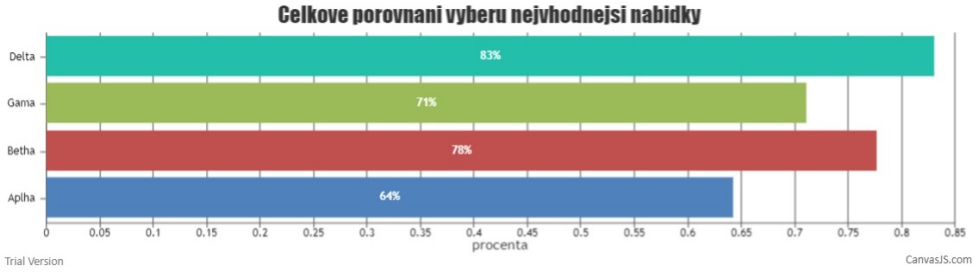
Jako třetí výstup z průvodce je graf, který ukazuje kdy budou první relevantní výstupy z DM.



Obrázek 7.3: Kdy budou relevantní výstupy z DM

Na tomto grafu můžeme vidět, že rozptyl je ve 2 měsících. Kdy jako první vychází řešení od firmy Betha. Naopak poslední jsou v tomto ukazateli Delta a Alpha.

7.3.2 Výsledná metrika podle defaultních hodnot



Obrázek 7.4: Finální vyhodnocení z průvodce

Na tomto grafu můžeme vidět, že nejvyšší hodnocení má nabídka od firmy Delta - 83%. Následují nabídky od firmy Betha 78%, Gama 71% a Alpha 64%. Průvodce by tedy podle defaultního nastavení vah jednotlivých otázek nejvíce upřednostnil nabídku od firmy Delta, i když z hlediska ceny není nejlevnější. V závěsu za touto nabídkou je nabídka od firmy Betha, která dokonce překračuje plánovaný rozpočet o 35 tisíc. Na posledním místě se umístila nabídka od firmy Alpha, která je nejlevnější,

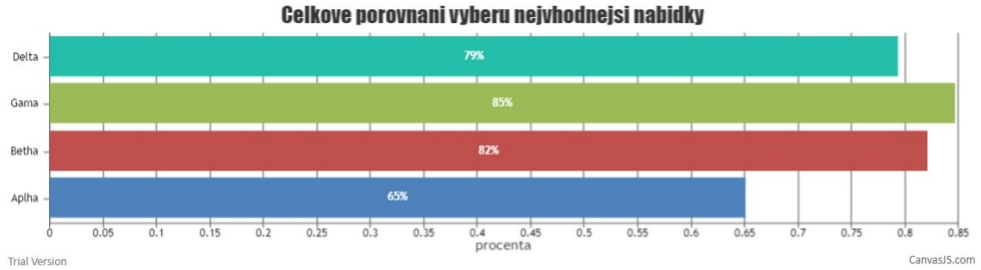
protože v ostatních metrikách, jako třeba nabídka lidí se zkušenostmi s danou technoligií, zaostává za ostatními. Dokonce se i před tuto nabídku dostala nabídka od firmy Gama, se kterou má banka špatnou zkušenost.

7.3.3 Výsledná metrika podle konkrétních hodnot zadaných testovacím uživatelem

V předešlém testu se přidělovaly váhy podle defaultně daných vah, které vzešly z testování. Následně jsem dal tohoto průvodce otestovat ještě jednomu uživateli. Tento uživatel se rozhodl nepoužít defaultně dané hodnoty, ale přistoupil k přiřazení vah pomocí odpovídání na otázky a vyhodnocení podle Fullerova algoritmu, který je popsaný v kapitole o vytvoření průvodce. Z tohoto postupu byly získány následující váhy:

- Zkušenosti s firmou 0.16364
- Kolik má firma zaměstnanců se zkussnostmi s DWH 0.05455
- Zkušenosti s podobnými projekty 0.18182
- Připravený model z jiných projektů 0.03636
- Nabídka supportu 0.05455
- Nabídka lidí se zkušenostmi s danou technologií DB 0.12727
- Nabídka lidí se zkušenostmi s danou technologií DI 0.12727
- Nabídka lidí se zkušenostmi s danou technologií BI 0.12727
- Nabídka fix price 0.03636
- Celková cena nabídky 0.09091

Váhy jsou ve výpisu zaokrouhleny na 5 destinných míst. Z takto vygenerovaných vah je vidět, že pro tohoto konkrétního uživatele hraje nejvyšší roli, jestli dodavatel má už zkušenosti s podobným projektem. Naopak nejméně důležité je, zda má už připravený model a jestli je jeho nabídka ve formě fix price. Díky takto nastaveným vahám je finální výstup z průvodce tento:



Obrázek 7.5: Finální vyhodnocení z průvodce

Na tomto grafu můžeme vidět změnu v pořadí. Na první místo se dostává nabídka od firmy Gama, která si polepšila o 14% oproti defaultním vahám. Mírně si polepšily firmy Aplha a Delta na 65% respektivě na 79%. Naopak propad zaznamenala firma Delta a to o 4%. Je to dáno tím, že tento uživatel preferuje tochu odlišné vlastnosti nabídky.

7.4 Závěr testování

Na zadaných datech bylo provedeno otestování tohoto průvodce. Na průvodci můžeme vidět, že pracuje flexibilně s cenou nabídky a to tak, že automaticky nezavrhuje dražší nabídky a ani vyloženě neupřednostňuje nabídky, které mají nejnižší cenu. Snaží se najít určitý kompromis mezi jednotlivými hodnotami. Také je zde vidět, že pokud má uživatel odlišné preference než testované subjekty, ze kterých jsem vytvořil defaultní hodnoty vah, může se výpočet výhodnosti nabídky naprosto změnit.

Kapitola 8

Vyhodnocení přínosu DWH a nákladů na jeho implementaci a provoz

V této kapitole se čtenář seznámí s finálním zhodnocením DWH pro bankovní sektor.

8.1 Finance

Datový sklad je relativně velmi finančně náročná věc. Implementace konkrétního datového skladu v případové studii je investice v řádech desítek milónů korun jen na vývoj. Ve své práci jsem neřešil support, který potřebuje po imlepentaci datového skladu také velký finanční rozpočet. Pro datový sklad který je popsán v případové studii, to může být okolo 2 FTE (počet osob na plný úvazek) pouze pro udržování chodu naimplementovaného řešení.

8.2 Výhody

Mezi výhody patří:

- Kontrola dat

  Nasazením a používáním datového skladu získá daná organizace jakousi kontrolu dat, které používá pro přípravu svých reportů a ze kterých dělá následné analýzy.

- Zprůhlednění procesu přípravy dat pro reporting

  DWH přináší zprůhlednění procesu zpracování dat od zdroje až k výslednému reportu.

- Úspora lidí

  Díky nasazení a využívání datového skladu nastane i v tomto případě úspora z řad lidí, kteří vytvářejí v aktuálním stavu reporty.

- Automatizace kontroly dat

  Automatizace kontroly a čištění dat se díky datovému skladu dostane na vyšší úroveň.

- Centralizace dat

  Díky centralizaci dat v datovém skladu je možná lepší unifikace klienta a celkově i lepší zaměření klienta.

- Lepší vstup pro další analytické procesy

  Datový sklad je rovněž lepším vstupem na další analytické procesy, jako je například Data mining.

- Historizace dat

  V datovém skladu vzniká rovněž historizace. Díky této funkčnosti je možné vytvářet jak aktuální reporty, tak i historické.

- Rychlé nalezení dat

  Pokud jsou data uložena v datovém skladu, není problém je rychle najít.

- Identifikace a korekce chyb

  Díky nasazení datového skladu se mohou opravit i chybné procesy v bance.

- Jedna verze pravdy - sjednocený slovník

  Jeden pojem znamená jednu věc. Nenastane situace, že jeden pojem bude v oddělení Risku znamenat něco jiného než v oddělení Central reportingu.

8.3 Nevýhody

Mezi nevýhody patří: [60]

- Počáteční náklady

  Velké počáteční náklady za na první pohled malý přínos pro společnost.

- Závislost

  Pokud banka přejde na řešení datového skladu trvale a nechá si datový sklad supportovat od dodavatelské firmy, stane se banka velice závislou na dané firmě.

- Práce navíc

  Může docházet k tomu, že při přípravě určitých dat pro reporting bude vyšší pracnost než bez datového skladu. Například při použití dat, které se získají manuálně.

- Sdílení citlivých údajů

  Pokud banka přistoupí k implementaci datového skladu externí firmou, musí rovněž zpřístupnit svá citlivá data 3. straně.

8.4 Vyhodnocení

Datový sklad má jak pozitiva tak negativa. Nicméně jeho pozitiva přesahují negativa, a proto v dnešní době datové sklady promlouvají do mnoha segmentů podnikání, kde bankovnictví není vyjímkou.

Závěr

Hlavními cíli této práce bylo přehledně zpracovat problematiku datových skladů a analyzovat používané architektonické struktury. Dále pak vytvoření průvodce pro rozhodování o výběru poskytovatele-dodavatele DWH, vytvoření případové studie a následné otestování vytvořeného průvodce. Všechny tyto vytyčené cíle se mi podle mého názoru podařilo adekvátně naplnit.

První cíle byly vypracovány v prvních kapitolách práce, kde popisuji, co jsou datové sklady, jaké postupy a architektury se používají. Následně navazují příklady technologií, které jsou potřeba nebo přímo pracují s datovým skladem.

Následné cíle jsou vypracovány v druhé polovině diplomové práce, která se nesla v praktickém duchu. Nejprve vytvořením vhodného průvodce, který má pomoci při rozhodování o výběru dodavatele. Následně pak vytvořením části případové studie, zaobírající se zejména finanční stránkou daného projektu. A závěrečné otestování mnou vytvořeného průvodce.

Mezi hlavní přínosy mé práce řadím přehledně zpracovanou problematiku datových skladů, vytvoření průvodce a vyformulování nejdůležitějších otázek pro rozhodování a následné vytvoření případové studie pro konkrétní implementaci datového skladu v bance.

Osobně jsem si na této práci velice prohloubil znalosti problematiky datových skladů a používaných postupů. Oblast datových skladů je nedílnou součástí bank působících na našem uzemí.

Literatura

[1] Sekničková, J.: *Vícekriteriální hodnocení variant – VHV [online]*. [cit. 2018-12-12]. Dostupné z: `http://jana.kalcev.cz/vyuka/kestazeni/EKO422-Vahy.pdf`

[2] Bém, M.: DW Klub Resuscitace. [cit. 2018-10-10]. Dostupné z: `http://www.dwklub.cz`

[3] Linstedt, D.; Olschimke, M.: *Building a Scalable Data Warehouse with Data Vault 2.0*. Todd Green, třetí vydání, 2013, ISBN 978-0-12-802510-9.

[4] Ronthal, A.; Edjlali, R.; Greenwald, R.: Magic Quadrant for Data Management Solutions for Analytics. *Gartner [online]*, únor 2018, [cit. 2018-12-16]. Dostupné z: `https://www.gartner.com/doc/reprints?id=1-3U1LC65&ct=170222&st=sb`

[5] Beyer, M. A.; Thoo, E.; Zaidi, E.: *Gartner Magic Quadrant for Data Integration Tools*. New York : John Wiley and Sons Ltd, 2018.

[6] Oracle: *Oracle Data Integrator 12c Architecture Overview [online]*. [cit. 2019-01-02]. Dostupné z: `https://www.oracle.com/technetwork/middleware/data-integrator/overview/oracledi-architecture-1-129425.pdf`

[7] Howson, C.; Sallam, R.; Richardson, J.; aj.: *Magic Quadrant for Analytics and Business Intelligence Platforms*. 2018.

[8] Kimball, R.; Margy, R.: *The data warehouse toolkit: the definitive guide to dimensional modeling*. Indianopolis: John Wiley, třetí vydání, 2013, ISBN 11-185-3080-2.

[9] Panec, Z.: Co je to Business intelligence? *IT Systems [online]*, červen 2003, [cit. 2019-12-10]. Dostupné z: https://www.systemonline.cz/clanky/co-je-to-business-intelligence.htm

[10] Slovník pojmů. [cit. 2018-12-11]. Dostupné z: https://www.cnb.cz/cs/obecne/slovnik/b.html

[11] Dvořák, P.: *Bankovnictví pro bankéře a klienty.* Praha: Linde, třetí vydání, 2005, ISBN 80-7201-515-X.

[12] Polášek, M.: Databáze z hlediska podnikových informačních systémů. *IT SYSTEM [online]*, srpen 2003, [cit. 2018-12-10]. Dostupné z: https://www.systemonline.cz/clanky/co-je-to-business-intelligence.htm

[13] Pokorný, J.: Objektově-orientovaná databáze. *Gomputer World [online]*, duben 1998, [cit. 2018-12-16]. Dostupné z: https://computerworld.cz/archiv/objektove-orientovana-databaze-9390

[14] Otte, L.: Databázové systemy. [cit. 2018-12-11]. Dostupné z: http://projekty.fs.vsb.cz/463/edubase/VY_01_044/Datab%C3%A1zov%C3%A9%20syst%C3%A9my/02%20Text%20pro%20e-learning/Datab%C3%A1zov%C3%A9%20syst%C3%A9my%2005.%20Normalizace%20dat.pdf

[15] Heath, I.: Unacceptable File Operations in a Relational Database. listopad 1971, [cit. 2018-12-16].

[16] Zechmeister, J.: Databázové systémy I. [cit. 2018-12-20]. Dostupné z: https://slideplayer.cz/slide/2980967/

[17] DYCHÉ, J.; LEVY, E.: *Customer Data Integration: Reaching a Single Version of the Truth.* New York : John Wiley and Sons Lt, 2006, ISBN 978-0-471-91697-0.

[18] Schiller, M.: Co se skrývá pod zkratkou ETL? *IT Systems [online]*, květen 2003, [cit. 2018-12-16]. Dostupné z: https://www.systemonline.cz/clanky/co-se-skryva-pod-zkratkou-etl.htm

[19] Brunet, T.; Noska, M.: Datové modelování: Základ datové architektury. *computerworld [online]*, březen 2009, [cit. 2018-11-14]. Dostupné z: https://computerworld.cz/software/datove-modelovani-zaklad-datove-architektury-3745

[20] Volko, V.: PROJECT MANAGEMENT. [cit. 2018-12-26]. Dostupné z: http://www.dtostrava.cz/volko/downloads/PM_Volko_new.pdf

[21] Jones, C.: What is agile development? *IT PRO [online]*, listopad 2018, [cit. 2018-12-29]. Dostupné z: https://www.gartner.com/doc/reprints?id=1-3U1LC65\&ct=170222\&st=sb

[22] Bowes, J.: Agile vs Waterfall: Comparing project management methods. *TPX MANIFESTO [online]*, červenec 2014, [cit. 2018-12-29]. Dostupné z: https://manifesto.co.uk/agile-vs-waterfall-comparing-project-management-methodologies/

[23] ITIL slovník pojmů. [cit. 2018-12-26]. Dostupné z: http://www.tcox.cz/slovnik/itil

[24] Kimball, R.: *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. New York: John Wiley, třetí vydání, 1998, ISBN 04-712-5547-5.

[25] Novotný, O.; Pour, J.; Slánský, D.: *Business Intelligence - jak využít bohatstvi ve vašich datech*. Grada Publishing, 2005, ISBN 80-247-1094-3.

[26] Linstedt, D.: Data Vault Series 1 – Data Vault Overview. *The Data Aministration Newsletter [online]*, červenec 2002, [cit. 2018-12-12]. Dostupné z: http://tdan.com/data-vault-series-1-data-vault-overview/5054#

[27] Is Data Vault Modeling a Good Choice for Your Organization? *Inside BIG DATA [online]*, červenec 2017, [cit. 2018-12-12]. Dostupné z: https://insidebigdata.com/2017/07/28/data-vault-modeling-good-choice-organization/

[28] Dostál, J.: *Hardware moderního počítače*. Olomouc: UP, 2011, ISBN 978-80-244-2787-4.

[29] DATABÁZE UMÍME A MÁME JE RÁDI. [cit. 2018-11-15]. Dostupné z: http://www.adastra.cz/technologie/databaze

[30] Teradata: Newest Teradata Data Warehouse Appliance is a Powerhouse for the Most Demanding Analytics. duben 2015, [cit. 2018-12-29]. Dostupné z: https://www.teradata.com/Press-Releases/2015/Newest-Teradata-Data-Warehouse-Appliance-is-a

[31] Co je to cloud computing. *CIO [online]*, duben 1999, [cit. 2018-12-16]. Dostupné z: https://businessworld.cz/ostatni/co-je-to-cloud-computing-7159

[32] Mácha, P.: Historie a základní principy cloud computingu. *SystemOnline [online]*, 2015, [cit. 2018-12-30]. Dostupné z: https://www.systemonline.cz/virtualizace/historie-a-zakladni-principy-cloud-computingu.htm

[33] Bém, M.; Ludvikova, D.: Data Warehouse as a Service jako budoucnost datových skladů. *IT Systems [online]*, duben 2018, [cit. 2018-10-31]. Dostupné z: https://m.systemonline.cz/business-intelligence/data-warehouse-as-a-service.htm

[34] Oracle: *Oracle Database Administrator's Guide*. 2015.

[35] Zíka, O.: Databáze v praxi. [cit. 2018-12-26]. Dostupné z: https://docplayer.cz/3428324-Databaze-v-praxi-rndr-ondrej-zyka-principal-consultant.html

[36] DB-Engines: Knowledge Base of Relational and NoSQL Database Management Systemsy. 2018, [cit. 2018-10-31]. Dostupné z: https://db-engines.com/en/system/Microsoft+SQL+Server%3BOracle%3BTeradata

[37] Stansfield, J.: Microsoft SQL Server vs. Oracle: The Same, But Different? *Segue Technologies [online]*, květen 2014, [cit. 2018-12-29]. Dostupné z: https://www.seguetech.com/microsoft-sql-server-vs-oracle-same-different/

[38] Oracle: *Oracle Data Integrator [online]*. [cit. 2018-07-07]. Dostupné z: https://docs.oracle.com/cd/E17904_01/integrate.1111/e12641/overview.htm#ODIGS113

[39] Hitachi Vantara: *Pentaho Data Integration Architecture [online]*. [cit. 2019-01-02]. Dostupné z: https://help.pentaho.com/Documentation/5.1/0L0/0Y0/010

[40] Pentaho Data Integration. 2018, [cit. 2018-10-31]. Dostupné z: https://www.hitachivantara.com/en-in/products/big-data-integration-analytics/pentaho-data-integration.html

[41] Microsoft: *SQL Server Integration Services [online]*. [cit. 2018-09-09]. Dostupné z: https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-servicesOTAZNIKview=sql-server-2017

[42] Uwujaren, J.: What Is the Difference Between Microsoft SSRS, SSIS and SSAS? [cit. 2018-12-26]. Dostupné z: https://smallbusiness.chron.com/difference-between-microsoft-ssrs-ssis-ssas-34689.html

[43] Chlapek, D.; Řepa, V.; Stanovská, I.: *Analýza a návrh informačních systémů.* Praha: Oeconomica, 2011, ISBN 978-80-245-1782-7.

[44] PowerDesigner. [cit. 2018-12-11]. Dostupné z: http://powerdesigner.de/en/pricing/

[45] EDUCBA: 9 Tools to Become Successful In Data Modeling. [cit. 2018-12-26]. Dostupné z: https://www.educba.com/9-best-data-modeling-tools/

[46] ENTERPRISE ARCHITECT - Ultimate Edition. [cit. 2018-12-11]. Dostupné z: https://sparxsystems.com/products/ea/editions/ultimate.html

[47] ER/Studio Data Architect - Multi-Platform. [cit. 2018-12-11]. Dostupné z: https://www.idera.com/buynow/onlinestore?ptid=\protect\T1\textbraceleft1d1561c7-0759-4d83-babf-d2fd13125317\protect\T1\textbraceright#ERSDataArchitect

[48] Power BI. [cit. 2018-12-12]. Dostupné z: https://powerbi.microsoft.com/en-us/

[49] Haman, M.: POWER BI NÁSTROJE. [cit. 2018-10-31]. Dostupné z: http://martinhaman.com/cs/power-bi-nastroje/

[50] Tableau. [cit. 2018-12-12]. Dostupné z: https://www.tableau.com/products/server

[51] Tableau. [cit. 2018-12-11]. Dostupné z: http://www.inekon-systems.cz/tableau/produkty/ceny-licenci/

[52] Profinit: BI a DWH. [cit. 2018-11-15]. Dostupné z: https://profinit.eu/sluzby/bi-dwh/

[53] Adastra: DWH. [cit. 2018-11-15]. Dostupné z: http://www.adastra.cz/ict-reseni/data-warehousing

[54] Solutions, S.: Datové sklady. [cit. 2018-11-15]. Dostupné z: http://www.sophias.cz/cz/sluzby_a_reseni/dwh/technologie.php

[55] Young, E. .: Jak řídíme v Česku projekty? *CFO World [online]*, srpen 2010, [cit. 2018-12-10]. Dostupné z: https://cfoworld.cz/ostatni/jak-ridime-v-cesku-projekty-456

[56] Hendl, J.: *Úvod do kvalitativního výzkumu*. Praha: Karolinum, 1997, ISBN 80-7184-549-3.

[57] *Oracle Engineered Systems Price List[online]*. [cit. 2018-12-12]. Dostupné z: https://www.oracle.com/assets/exadata-pricelist-070598.pdf

[58] *Oracle Technology Global Price List [online]*. [cit. 2018-12-12]. Dostupné z: https://www.oracle.com/assets/technology-price-list-070617.pdf

[59] Trendy na pracovním trhu. [cit. 2018-12-11]. Dostupné z: https://www.hays.cz/cs/groups/hays_common/@cz/@content/documents/digitalasset/hays_1854406.pdf

[60] Burnside, K.: The Disadvantages of a Data Warehouse. *Chron [online]*, [cit. 2018-12-10]. Dostupné z: https://smallbusiness.chron.com/disadvantages-data-warehouse-73584.html

Příloha **A**

Seznam použitých zkratek

DWH Datový sklad

ETL Extrakt, Transform, Load

ELT Extrakt, Load, Transform

DB Databáze

BI Business Inteligence

DI Datová integrace

ODI Oracle Data Integrator

PDI Pentaho Data Integrator

L0 Staging Layer

L1 Date Warehouse Layer

L2 Data Marts Layer

TCO Celkové náklady na vlastnictví

PM Projekt manager

DM Data mart

MD Man-day

Příloha B

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `img` ..... adresář s obrázky
  - `antospe1` ..... zdrojová forma práce ve formátu LaTeX
  - `antospe1.pdf` ..... text práce ve formátu PDF
  - `priloha` ..... adresář s přílohami