Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra informačních technologií

Studijní program: Aplikovaná informatika
Obor: Informační systémy a technologie

Diplomant: Bc. Sandra Müllerová
Vedoucí diplomové práce: doc. Ing. Ota Novotný, Ph.D.
Oponent diplomové práce: Ing. Barbora Tomášková

METODIKA BUDOVÁNÍ A ÚDRŽBY ZÁVISLÝCH DATOVÝCH TRŽIŠŤ

školní rok 2011/2012

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně a že jsem uvedla všechny použité prameny a literaturu, ze kterých jsem čerpala.

V Praze dne:

…….

Poděkování

Ráda bych poděkovala vedoucímu své diplomové práce panu doc. Ing. Otovi Novotnému, Ph.D. za podnětné a odborné rady, nápady a ochotu.

Abstrakt

Práce je zaměřena na oblast integrovaných datových skladů a zejména pak na jejich podmnožinu – datová tržiště. Hlavními cíli této práce je návrh metodiky vývoje a údržby závislých datových tržišť a ověření využitelnosti metodiky v praxi v reálné organizaci.

První část práce se zabývá teoretickým vymezením pojmů. Je zaměřena zejména na vymezení pojmu Business Intelligence, data warehousing a datová tržiště. Každý z pojmů je pak detailněji rozebrán v samostatných kapitolách. V popisu oblasti Business Inteligence se zaměřuji na popis jednotlivých komponent. V oblasti data warehousingu popisuji koncepce datového skladu a obsah jednotlivých vrstev v datovém skladu. Konečně v části zaměřené na datová tržiště popisuji závislá a nezávislá datová tržiště a navíc i „speciální“ případy řešení datového tržiště, konkrétně sémantickou vrstvu a sandboxy.

Druhá část se zaměřuje na samotný návrh metodiky. Na začátku této části jsem provedla analýzu existujících metodik a zhodnotila jejich využitelnost v mé práci. Následuje popis aktuální situace v přístupu k vývoji a údržbě závislých datových tržišť v reálné organizaci. V závěru druhé části je navržena vlastní metodika, která z části vychází právě z analýzy metodik a z části z analýzy současného stavu.

Třetí část práce se zaměřuje na zhodnocení využitelnosti metodiky v praxi. Zhodnocení je provedeno na základě kritiky metodiky ze strany pracovníků v organizaci, kteří se přímo věnují budování a správě datových tržišť.

Konečně čtvrtá a závěrečná část se zaměří na popis alternativního řešení, které by mohlo být uvažováno jako jedna z cest k trvale udržitelnému rozvoji datového skladu v organizaci. Jedná se o použití architektury, která oproti „klasické-Inmonovské“ architektuře navíc využívá sémantickou vrstvu. Výstupem bude zhodnocení alternativního řešení vůči současnému řešení.

Klíčová slova

integrovaný datový sklad, závislé datové tržiště, Business Intelligence, data mart, data warehouse, sandbox, sémantická vrstva

Abstract

The thesis is primary focused on the integrated data warehouse, particularly on a subset – dependent data marts. The main objectives of this thesis are design the methodology of development and maintenance of dependent data marts and verification methodology usefulness in the real organization.

The first part deals with the theoretical definition of terms. It focuses particularly on the definition of terms from Business Intelligence area especially data warehousing and data marts. Each of the terms is described in detail in separate chapters. Business Intelligence area puts emphasis on description of individual components. In data warehousing area are described the data warehouse concepts and content of layers in a data warehouse. Finally, the data mart area is designed to describe dependent and independent data marts and also "special "cases of data marts, likely the semantic layer, and a sandbox.

The second part focuses on the design methodology itself. At the beginning of this part is analysis of the existing methodologies and assess their usefulness with connection to designed methodology. The following part describes the current situation in the approach to the development and maintenance of dependent data marts in the organization. At the end of the second part is designed own methodology, which is based in part on the analysis methodology and in part on the analysis of current situation.

The third part focuses on usability and usefulness evaluation of methodology in the organization. Evaluation is based on the methodology of criticism from employees in the organization who are directly engaged in designing and maintaining dependent data marts.

Finally, the fourth and final part will focus on the description of an alternative solution that could be considered as one of the ways to sustainable development of data warehouse in the organization. It´s about comparison architecture based on utilization of semantic layer as oppose to the "three layers" concept of data warehouse by Bill Inmon, which is implemented in the organization. The output evaluates alternative solutions to the current solution.

Keywords

Integrated data warehouse, dependent data marts, Business Intelligence, sandbox, semantic layer

Obsah

1. Úvod ... 1
   1.1. Důvod výběru tématu ... 1
   1.2. Cíle a přínosy ... 1
   1.3. Vstupy práce ... 2
   1.4. Struktura a obsah práce ... 2
      1.4.1. Úvodní část ... 2
      1.4.2. Návrh metodiky ... 2
      1.4.3. Aplikace metodiky v praxi ... 3
      1.4.4. Porovnání stávající architektury s architekturou využívající sémantickou vrstvu ... 3
   1.5. Výstup práce ... 3
   1.6. Omezení práce ... 3
2. Rešerše závěrečných prací ... 4
3. Business Intelligence ... 5
   3.1. Jednotlivé komponenty BI architektury ... 6
      3.1.1. Datový zdroj – „*Source system*“ ... 6
      3.1.2. Dočasné úložiště dat –„*Data Staging Area*“ (DSA) ... 7
      3.1.3. Operativní datový sklad- „*Operational Data Store*“ (ODS) ... 7
      3.1.4. Datový sklad „*Data Warehouse*“ ... 7
      3.1.5. Datové tržiště „*Data Mart*“ ... 8
      3.1.6. OLAP kostka (whatis.com, 2011) ... 9
4. Data Warehousing ... 10
   4.1. Koncepce Williama Inmona – „*Corporate Information Factory*“ (CIF) (Inmon, 2005) ... 10
      4.1.1. Datový sklad – „*Enterprise Data Warehouse*“ ... 11
   4.2. Koncepce Raplha Kimballa – „*Data Warehouse Bus*“ (Kimball, a další, 2002) ... 11
      4.2.1. Dočasné úložiště dat – „*Data Staging Area*“ ... 12
      4.2.2. Datový sklad – „*Data Presentation*“ ... 12
      4.2.3. Přístupové nástroje – „*Data Access Tools*“ ... 12
5. Datové tržiště ... 13
   5.1. Nezávislé datové tržiště ... 13
      5.1.1. Tabulka faktů (Kimball, a další, 2002) ... 13
      5.1.2. Tabulka dimenzí (Kimball, a další, 2002) ... 14
      5.1.3. Schémata ... 15
   5.2. Závislé datové tržiště ... 17

5.3. Speciální případy „datových tržišť“ ... 18
5.3.1. Sandboxy „*Analytical Sandbox*“ ... 18
5.3.2. Sémantická vrstva ... 20
6. Metodiky řízení BI projektů a vývoje BI řešení ... 23
6.1. Řízení BI projektu dle (Moss, a další, 2003) ... 23
6.1.1. Fáze Obhájení („*Justification*“) ... 23
6.1.2. Fáze Plánování („*Planning*“) ... 23
6.1.3. Fáze Byznys analýza („*Business analysis*“) ... 24
6.1.4. Fáze Návrh („*Design*“) ... 24
6.1.5. Fáze Budování („*Construction*“) ... 24
6.1.6. Fáze Údržba („*Deployment*“) ... 25
6.2. Metodika vývoje dle Inmona - „*Spiral Methodology*“ ... 25
6.2.1. „Spirálový“ přístup ... 26
6.2.2. Heuristický přístup (Inmon, 2004) ... 28
6.3. Metodika vývoje dle Kimballa - „*Kimball Lifecycle*“ ... 29
6.4. Metodika vývoje datového tržiště dle společnosti Talus – „PrimeMover Life cycle“ (Talus) 31
6.5. Zhodnocení metodik ... 32
7. Analýza a popis současného stavu ... 34
7.1. Vysvětlení pojmů používaných v organizaci ... 34
7.2. Všeobecný popis procesu správy tržiště z procesního hlediska ... 37
7.3. Všeobecný popis procesu správy tržiště z pohledu architektury datového skladu ... 39
7.4. Detailní popis procesu správy datového tržiště ... 42
7.4.1. Přijetí požadavku na úpravu/vývoj datového tržiště ... 42
7.4.2. Příprava tvorby/úpravy datového tržiště ... 43
7.4.3. Proces změnového řízení ... 44
7.4.4. Nasazení a údržba datového tržiště ... 48
7.4.5. Vyřazení tržiště z provozu ... 49
7.5. Využití reportovací vrstvy a virtuálních tržišť – sandboxů ... 49
7.6. Zhodnocení současného stavu ... 50
8. Návrh metodiky ... 52
8.1. Analýza požadavků aneb definice rozsahu projektu ... 53
8.2. Analýza dostupnosti dat ... 55
8.3. Sandboxing ... 56
8.4. Příprava dat do vrstvy integrovaného datového skladu ... 57

8.5. Vývoj prototypu ... 58
8.6. Návrh datového tržiště ... 58
8.7. Vývoj ... 61
8.8. Test ... 62
8.9. Nasazení ... 64
8.9.1. Dohoda o úrovni poskytovaných služeb – „SLA“ ... 64
8.9.2. Nastavení datových kontrol ... 64
8.9.3. Dokumentace datového tržiště ... 64
8.10. Provoz a údržba ... 65
8.11. Vyřazení z provozu ... 65
8.12. Shrnutí – „*checklist*“ ... 66
9. Zhodnocení využití sémantické vrstvy v prostředí organizace ... 68
10. Závěr ... 69
11. Zdroje ... 70
13. Terminologický slovník ... 72
14. Rejstřík ... 73
15. Seznam obrázků a tabulek ... 74

1. Úvod

Relativně značná část publikací se věnuje problematice budování datových skladů. Také ve vybrané organizaci je formalizován postup rozvoje datového skladu. Ale pokud jde o budování a údržbu datových tržišť, je tento proces neřízený a zaměstnanci na něj reagují dle svých zkušeností a znalostí. Lze namítnout, že datové tržiště je ve své podstatě založené na stejných principech jako datový sklad. Tím se nabízí otázka, proč vlastně navrhovat metodiku budování datového tržiště? Na vybudování datového tržiště ve většině případů není třeba tak robustní metodiky, která může odrazovat svou obsáhlostí.

Přesto je třeba proces budování tržiště formalizovat. Jinak se může stát, že budou vznikat datová tržiště s duplicitním obsahem, nevyužívaná tržiště, případně se proces vývoje prodlouží na neúměrně dlouhou dobu s vynaložením nemalých finančních nákladů a přetížením lidských zdrojů, protože se bude „objevovat Amerika“.

1.1.Důvod výběru tématu

Důvodem výběru tématu je motivace napsat práci, která bude reagovat na současné potřeby organizace, pro kterou pracuji. A zároveň přispěje k udržitelnému rozvoji datového skladu potažmo datových tržišť v organizaci.

1.2.Cíle a přínosy

Hlavní cíl práce:

- Návrh metodiky budování a údržby závislého datového tržiště, která porývá celý životní cyklus tržiště od sběru požadavků až po vyřazení ze zpracování.
- Analýza a popis současného stavu v organizaci, který zatím není zdokumentován.

Dílčí cíle

- zhodnocení využitelnosti metodiky v praxi a zhodnocení přesnosti popisu současného stavu – splnění cíle bude hodnoceno kolegy, kteří se přímo zabývají správou datových tržišť,
- analýza odborné literatury, která se věnuje architektuře využívající sémantickou vrstvu a popis této architektury,
- zhodnocení alternativního řešení s využitím sémantické vrstvy, jako cesty k dlouhodobě udržitelnému rozvoji datového skladu v organizaci.

1.3. Vstupy práce

V práci využiji především zahraniční literaturu věnující se problematice budování datových skladů. Důvodem je fakt, že v češtině jsem kromě práce „Business Intelligence. Jak využít bohatství ve vašich datech“ od Oty Novotného, Jana Poura a Davida Slánského, která je ale vyprodaná, nenašla žádnou ucelenou publikaci věnující se problematice datových skladů. A zároveň také proto, že pokud to bylo možné, tak jsem se snažila využívat přímo „původní“ publikace.

Jelikož se ve své práci zaměřím zejména na integrované datové sklady, využiji jako hlavní zdroj informací publikace a články vytvořené „otcem“ datových skladů – Williamem H. Inmonem. Jako elektronický zdroj využiji odborné články umístěné na stránkách The Data Warehousing Institute www.tdwi.org a komunitního serveru www.b-eye-network.com.

Dalšími zdroji budou konzultace s pracovníky organizace, kteří se účastní projektů návrhu a následné údržby datových tržišť.

Primární zdroje:

- Inmon, W. H.: Building the Data Warehouse, Fourth Edition. Wiley Publishing, 2005. ISBN 0-7645-9944-5.
- Inmon, W.H.: DW 2.0 - Architecture for the Next Generation of Data Warehousing. Elsevier, 2008. ISBN 978-0-12-374319-0.
- Imhoff, C.; Galemmo, N.; Geiger, J. G.: Mastering Data Warehouse Design Relational and Dimensional Techniques. Wiley Publishing, 2003. ISBN 0-471-32421-3.

1.4. Struktura a obsah práce

1.4.1. Úvodní část

Tato teoreticky zaměřená část práce se bude věnovat vymezení základních pojmů, které jsou pak využity v dalších částech práce. Zejména se jedná o vymezení pojmu integrovaného datového skladu, popis jednotlivých vrstev a jednotlivých komponent v koncepcích data warehousingu a detailní popis jednotlivých druhů datových tržišť.

1.4.2. Návrh metodiky

V první části této kapitoly analyzuji existující metodiky. Následuje analýza procesu vývoje a údržby závislého datového tržiště v organizaci. Na základě těchto analýz pak navrhnu metodiku, ve které zároveň identifikuji části, které jsou kritické při budování a údržbě závislých datových tržišť.

1.4.3. Aplikace metodiky v praxi

Tato kapitola se bude zabývat použitelností metodiky v praxi a to prostřednictvím kritiky ze strany pracovníků, kteří přímo řeší správu a vývoj datových tržišť. Vzhledem k citlivosti dat, zde ovšem nebudou žádné ukázky výstupů z reálného projektu.

1.4.4. Porovnání stávající architektury s architekturou využívající sémantickou vrstvu

Kapitola se zaměří na popis architektury, která bude využívat sémantickou vrstvu a bude předpokládat existenci business slovníku. Tuto architekturu pak porovnám se stávající architekturou datového skladu v organizaci.

1.5. Výstup práce

Výstupem práce bude metodika zaměřující se na budování a údržbu závislých datových tržišť, která bude zhodnocena svou využitelností v organizaci.

1.6. Omezení práce

Navržená metodika bude zaměřená pouze na integrovaný datový sklad a závislá datová tržiště. Práce bude sloužit jako podklad pro formalizaci postupu budování datových tržišť. Předpokládá se další zpřesňování a prohlubování metodiky jejím používáním v praxi.

Práce neřeší technologickou stránku řešení, je zaměřená spíše procesně a soustředí se na předání nejlepších postupů tzv. „best practices“, které jsem načerpala během svého působení v organizaci a studiem odborné literatury.

2. Rešerše závěrečných prací

Pro výběr závěrečných prací, které se zabývají tématem Business Intelligence a zejména pak datovými tržišti, jsem využila katalog knihovny VŠE dostupný na http://library.vse.cz/aleph/ a databázi kvalifikačních prací http://theses.cz/. V podstatě jsem nenašla práci, která by se zabývala přímo procesem vývoje a správy datových tržišť. Práce se zaměřují především na technický návrh a obecně neřeší postupy, jak tyto tržiště budovat a spravovat.

Ing. Pavel Kopecký - DP: Rozšíření datového tržiště – přenos, ukládání a analýza dat – Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky

Práce se zabývá zdokumentováním a realizací postupu rozšíření existujícího datového tržiště na MS SQL server, zabývá se pouze technickou stránkou řešení. Řeší nastavení ETL procesu, vytvoření OLAP kostek a následný návrh reportů v nástroji MS Excel. Z hlediska mé práce ne příliš využitelná práce, protože je příliš technologicky a úzce zaměřená.

Ing. Michal Juhás – DP: Riešenie dátových trhovísk a DWH v bankovníctve – 2008/2009 – VŠE, fakulta informatiky a statistiky

Práce se kromě porovnání přístupů podle Inmona a Kimballa soustředí na analýzu průběhu zavedení datového tržiště a jeho následnou realizaci vzniklého v rámci AML (Asset Liability Management) projektu. Z hlediska zaměření mé práce je zajímává část věnující se analýze business požadavků a analýze projektového managementu.

Ing. David Slánský, Ph.D. – Disertační práce: Řešení úloh Business Intelligence se zaměřením na prostředí telekomunikačních společností – 2004 – VŠE, fakulta informatiky a statistiky

Práce se zabývá vymezením pojmu BI v celé jeho šíři, sleduje a srovnává vývoj architektury BI a související procesy. Dále se zaměřuje na proces řízení BI a na implementaci tohoto procesu v prostředí telekomunikací. Z pohledu mé práce je zde zajímavý zejména proces řízení BI, který by se dal aplikovat i na metodiku datových tržišť.

3. Business Intelligence

Úvodní kapitola teoretické části je zaměřena na definování a vymezení pojmu Business Intelligence . Co všechno může tento pojem zahrnovat, tedy popis architektury Business Intelligence a definice jednotlivých komponent této architektury je rovněž součástí této kapitoly. Vzhledem k zaměření mé práce bude oblasti datových skladů a datových tržišť věnována samostatná kapitola.

Business Intelligence (BI) není jen o technologiích, ale i o použité metodice a lidech. Tedy pod BI si lze představit nejen konkrétní technologie, řešení či architekturu, ale i procesy, které souvisí se zaváděním a používáním BI. Cílem BI je pomocí těchto postupů transformovat data na informace, informace na znalosti a ty pak výhodně využít v podnikání. Získání těchto informací probíhá prostřednictvím sběru, čištění, integrace, analýzy a prezentace dat.

Definice pojmu BI se různí, zde uvádím definici podle knihy *Business Intelligence Roadmap* (Moss, a další, 2003), která popisuje, že BI je třeba chápat jako na celek, tedy že nejde jen o jeden konkrétní „BI produkt“ [1] :

„BI není ani produkt ani systém. BI je architektura a kolekce integrovaných provozních aplikací, systémů pro podporu rozhodování a databází, která poskytuje odběratelům dat snadný přístup k jejich datům.“ (Moss, a další, 2003 str. 2)

Další, dle mého názoru ještě výstižnější definice podle The Data Warehousing Institute, dále jen TDWI, která se navíc zaměřuje i na výše zmíněné procesy[2]:

„Sada procesů, technologií a nástrojů potřebných k transformaci dat na informace, informací na znalosti a znalostí na plány tak, aby tyto plány vedly k efektivním obchodním rozhodnutím. Business Intelligence v sobě zahrnuje oblast datových skladů, reportovací nástroje a znalostního managementu.“ (Loshin, 2003 str. 6) (Loshin, 2003 str. 6) (Loshin, 2003 str. 6)

Definice zmiňuje jako součást BI i procesy, což je důležitá součást, neboť pokud jsou nevhodně nastavené procesy, ani výkonné nástroje toho mnoho nevyřeší. Zároveň z definice vyplývá,

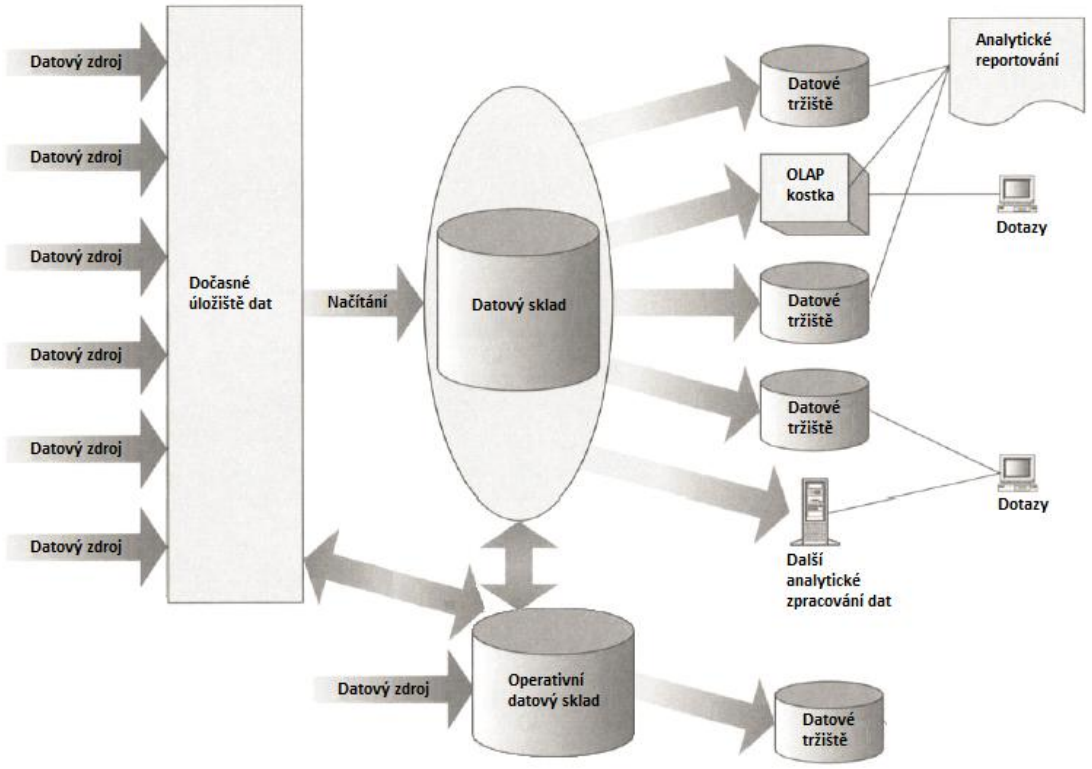
---

[1] Originální definice v angličtině: *„BI is neither a product nor a system. It is an architecture and a collection of integrated operational as well as decision-support applications and databases that provide the business community easy access to business data.“*

[2] Originální definice v angličtině: *„The processes, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action. Business intelligence encompasses data warehousing, business analytic tools, and content/knowledge management.“*

že jedním z účelů BI je přeměna dat na informace, tak aby tyto informace vedly k efektivnímu rozhodování jejich odběratelů.

Jedna z možných variant BI architektury na obrázku č. 1, znázorňuje, jakým způsobem je možné tuto transformaci provést. Oproti zdroji jsem obrázek doplnila o operativní datový sklad (ODS – Operational Data Store), protože odběratelé dat si nárokují stále rychlejší přístup k informacím, a ODS je možnost, jak tuto potřebu uspokojit. Umístění ODS v architektuře BI je rovněž variabilní a vždy záleží na konkrétní implementaci. V následující části této kapitoly se zaměřím na popis jednotlivých komponent BI architektury.



Obrázek 1 - Architektura BI; Doplněno a přeloženo z: (Loshin, 2003 str. 41) (Loshin, 2003 str. 41) (Loshin, 2003 str. 41)

3.1. Jednotlivé komponenty BI architektury

3.1.1. Datový zdroj – *„Source system“*

Datové zdroje představují data ze zdrojových (transakčních) systémů, které jsou využívány v každodenní praxi v organizaci. Může se jednat například o aplikace typu CRM, která obsahuje data o zákaznících, ERP obsahující data o zakázkách, výrobě či třeba účetní aplikace a jiné. Jako datový zdroj mohou sloužit nejen databáze, ale i textové soubory[3].

[3] V angličtině označované jako „flat files”

V běžném zdrojovém systému se data uchovávají po dobu 60-90 dnů, v datovém skladu i 5 let a více. Zdrojové systémy tedy obsahují data, která jsou v daném okamžiku platná. Naproti tomu, v datovém skladu můžeme porovnávat aktuální stav se stavem platným např. před rokem, to je možné díky historizaci dat, kterou blíže popisuji v samostatné kapitole věnující se datovým skladům.

3.1.2. Dočasné úložiště dat –„*Data Staging Area*“ (DSA)

Dočasné úložiště dat (DSA – Data Staging Area) slouží, jak už název napovídá ke sběru dat z datových zdrojů a jejich následnému dočasnému uložení předtím, než budou data nahrána pomocí ETL/ELT (ETL – Extract Transform Load/ ELT – Extract Load Transaform)[4] procesu do datového skladu. Oproti datovému skladu zde neprobíhá historizace dat a nekontroluje se jejich konzistence. Integrace dat probíhá až na úrovni DSA – datový sklad.

3.1.3. Operativní datový sklad- „*Operational Data Store*“ (ODS)

(Inmon, 2005) uvádí, že datový sklad a ODS se mohou v některých oblastech prolínat. Obě zmíněná datová úložiště poskytují podporu systémům pro podporu rozhodování, stojí až za zdrojovými systémy a obě řešení využívají integrovaná data. Je i více možností jak propojit ODS a datový sklad a vždy záleží na konkrétní implementaci. Datový sklad může být zdrojem dat v ODS, někdy je tomu naopak, případně přenos dat funguje obousměrně anebo nejsou propojeny vůbec[5].

ODS na rozdíl od datového skladu podporuje zpracování v požadovaném čase, je stavěna na pravidelné aktualizace, má rychlejší dobu odezvy a historická data se udržují jen v omezené míře.[6]

ODS na rozdíl od datového skladu nemusí nutně být v BI architektuře použita. Rozhodnutí o implementaci ODS záleží na potřebách organizace.

3.1.4. Datový sklad „*Data Warehouse*“

Datový sklad je podmnožinou Business Intelligence. Existují dvě hlavní koncepce řešení datového skladu. Koncepce datového skladu dle Inmona (označována jako „*Corporate Information Factory*“), jejíž základ je založen na třívrstvé architektuře, obsahující DSA, centrální integrovanou datovou vrstvu – datový sklad (označovaný jako „*Enterprise Data Warehouse*“) a z té vycházející vrstva závislých datových tržišť. Tato architektura může být doplněna ještě o individuální vrstvu, která slouží k ad hoc dotazování.

---

[4] ETL/ELT proces je proces, který zajišťuje sběr dat ze zdrojových systémů a jejich integraci v datovém skladu a případně i plnění z datové skladu do datového tržiště

[5] (Inmon, 2005) popisuje na str. 434-435 druhy nebo spíše možnosti využití ODS

[6] Pro zajímavost, v článku uvedeném zde: http://www.b-eye-network.com/view/9017 Inmon mimo jiné srovnává ODS s datovým tržištěm

Koncepce dle Kimballa je založena na dvouvrstvé architektuře, obsahující DSA a nezávislá integrovaná datová tržiště, která jsou navzájem propojena prostřednictvím konformních dimenzí. Ve své knize (Kimball, a další, 2002) Kimball představuje vrstvu nezávislých integrovaných tržišť - datový sklad termínem „D*ata Presentation*".

Tyto koncepce blíže popíši a porovnám v samostatné kapitole. V rámci této podkapitoly uvádím pro úplnost definice výše zmíněných autorů. Zde je třeba také zmínit, že pojem datový sklad je v této práci používán ve smyslu centrální integrované datové vrstvy v případě Inmona nebo vrstvy nezávislých datových tržišť v případě Kimballa. DSA a datová tržiště tedy stojí mimo datový sklad.

Definice datové skladu podle Williama Inmona[7]:

„Datový sklad je subjektivně orientovaná, integrovaná, stálá a časově proměnná kolekce dat, určená k podpoře manažerského rozhodování." (Inmon, 2005 str. 29)

Definice datového skladu dle Ralpha Kimballa[8]:

„Vrstva integrovaných datových tržišť je místo, kde jsou data řízena, ukládána a zpřístupňována uživatelským dotazům, tvůrcům sestav a ostatním analytickým aplikacím. Jelikož dočasné úložiště je mimo dostupnost běžných uživatelů, je tato vrstva pro odběratele dat synonymem datového skladu. … Tato vrstva je tvořena souborem integrovaných datových tržišť." (Kimball, a další, 2002 str. 10)

3.1.5. Datové tržiště „*Data Mart*"

Obdobně jako v podkapitole výše i zde se potkávají koncepce Inmona a Kimballa. Vrstva datových tržišť je v pojetí Inmona chápána jako samostatná vrstva umístěná a vycházející pouze z datového skladu. Na rozdíl od datového skladu jsou data v tržištích „šitá na míru" konkrétnímu odběrateli dat.

Naproti tomu v koncepci Kimballa je datové tržiště chápáno jako jedna z částí samotného datového skladu.

Definice datového tržiště dle Inmona[9]:

---

[7] Originální definice v angličtině: *„A data warehouse is a subject-oriented, integrated, collection of data in support of management's decisions."*

[8] Originální definice v angličtině: *„The data presentation area is where data is organized, stored, and made available for direct querying by users, report writers, and other analytical applications. Since the backroom staging area is off-limits, the presentation area is the data warehouse as far as the business community is concerned. …We typically refer to the presentation area as a series of integrated data marts"*

[9] Originální definice v angličtině: *„Data mart is a data structure that is dedicated to serving the analytical needs of one group of people, such as the accounting department or the finance department."*

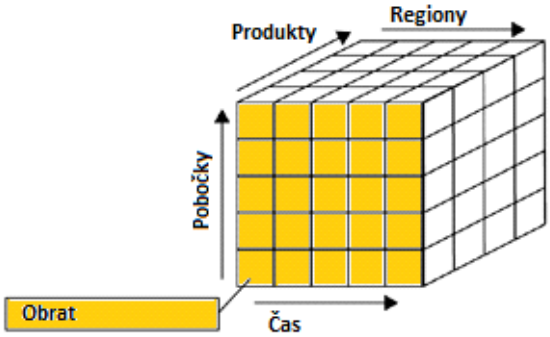
„Datové tržiště je datová struktura určená k uspokojení analytických potřeb určité skupiny odběratelů dat, jako třeba účetní nebo finanční oddělení.“ (Inmon, 2005 str. 370)

Definice datového tržiště dle Kimballa[10]:

„… datové tržiště je definováno jako procesně orientovaná podmnožina všech dat v organizaci, která je založena na nejvyšším možném detailu…“ (Kimball, a další, 2004 str. 25)

3.1.6. OLAP kostka (whatis.com, 2011)

OLAP kostka je metoda uložení dat v multidimensionální formě, určená především pro účely reportování a uzpůsobená pro konkrétního odběratele. OLAP kostka umožňuje nahlížet na data - fakta z různých pohledů – dimenzí, např. sledovat vývoj obratu z pohledu času, jednotlivých poboček, produktů, regionů aj, viz obrázek č. 2.



Obrázek 2 - OLAP kostka; Zdroj: autor

Vzhledem k velkému objemu dat, kdy nejsou výjimkou miliony záznamů, se data v kostce před použitím napočítávají napříč dimenzemi.

[10] Originální definice v angličtině: *„…we have consistently defined a data mart as a process-oriented subset of the overall organization’s data based on a foundation of atomic data…“*

4. Data Warehousing

4.1. Koncepce Williama Inmona – „*Corporate Information Factory*“ (CIF) (Inmon, 2005)

„Corporate Information Factory je široce uznávaná konceptuální architektura, která popisuje a kategorizuje datová úložiště, která jsou využita k řízení a provozu BI infrastruktury.“ [11] (Imhoff, a další, 2003 str. 11)

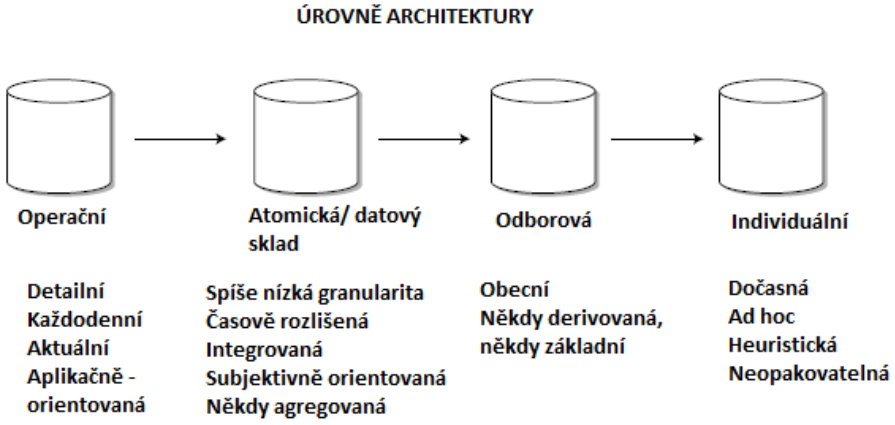
Obrázek č. 3 popisuje architekturu Inmonovy koncepce. Jednotlivé úrovně spolu tvoří základ pro širší architekturu - CIF:

Na operační úrovni - DSA probíhá sběr dat ze zdrojových systémů. Zde jsou tři možnosti jak nahrávat data do skladu. Buď historická data, aktuální data nebo data, u kterých došlo od posledního načítání ke změně, někdy označované jako inkrementy. Data do této úrovně bývají nahrávána ve struktuře, která odpovídá struktuře dat ve zdrojovém systému.

Na atomické úrovni – datový sklad dochází k již zmíněné integraci dat, a jejich historizaci. Data jsou ukládána pokud možno s nízkou úrovní granularity. Na této úrovni jsou data transformována a ukládána podle jednotlivých subjektů.

Na odborové úrovni – závislá datová tržiště, se data přizpůsobují požadavkům jednotlivých odběratelů. Tato úroveň přebírá data pouze z datového skladu, nikoliv z DSA.

Individuální úroveň představuje ad hoc uživatelské dotazování.



Obrázek 3 – Jednotlivé úrovně CIFu; Přeloženo ze zdroje: (Inmon, 2005 str. 16)

[11] Originální definice v angličtině: *„The Corporate Information Factory (CIF) is a widely accepted conceptual architecture that describes and categorizes the information stores used to operate and manage a successful and robust BI infrastructure.“*

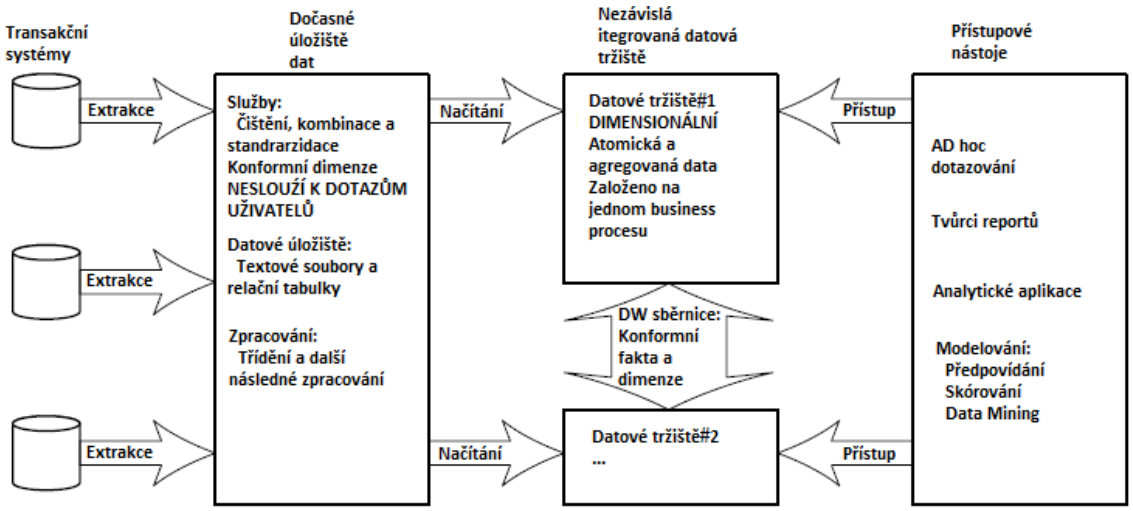
4.1.1. Datový sklad – *„Enterprise Data Warehouse“*

Jak bylo uvedeno v předchozí kapitole, datový sklad je dle definice Inmona (Inmon, 2005):

- **subjektivní** – data jsou ukládána podle subjektů, na které je organizace zaměřena např. banka se zaměřuje na klienty, smlouvy, produkty. Na rozdíl od transakčních systémů, kde jsou data ukládána s ohledem na transakční zpracování.
- **integrovaný** – jedna z nejdůležitějších vlastností datového skladu. Protože je datový sklad plněn z různých zdrojů, je třeba data vyčistit, konsolidovat, někdy agregovat tak, aby vznikla „jedna verze pravdy“.
- **neměnný** – v datovém skladu na rozdíl od zdrojového systému neprobíhá aktualizace.[12]
- **časově rozlišený** – každý záznam v datovém skladu je časově vymezený. Tím je možné zjistit v jakém časovém rozmezí, v případě zůstatků/transakcí k jakému datu, byl záznam platný. V datovém skladu se historie uchovává v řádech let, v transakčním systému je to otázka týdnů maximálně měsíců.

4.2. Koncepce Raplha Kimballa – *„Data Warehouse Bus“* (Kimball, a další, 2002)

Datový sklad dle konceptu Kimballa, viz č. 4, je založen na integraci jednotlivých datových tržišť. Jednotlivé komponenty této koncepce budou popsány dále.



Obrázek 4 - Koncepce datového skladu dle Kimballa; Přeloženo ze zdroje: (Kimball, a další, 2002 str. 7)

[12] Aktualizace v tom smyslu, že v e zdrojovém systému se záznamy přepisují nebo se rovnou smažou. Naproti tomu v datovém skladu zůstávají jak nezměněné původní záznamy, tak i záznamy aktualizované.

4.2.1. Dočasné úložiště dat – „*Data Staging Area*“

První vrstva – DSA, dle Kimballa je DSA cokoliv co leží mezi zdrojovými systémy a datovým skladem. Kromě toho, že slouží jako úložiště dat ze zdrojových systémů, tak je i součástí ETL/ELT procesu:

- **Extrakce (Extract)** - První krok procesu, ve kterém probíhá nahrávání dat ze zdrojových systémů.
- **Transformace (Transform)** - V dalším kroku jsou data čištěna a kombinována z různých zdrojů. Odstraňují se duplicity a přiřazují se zde jednotné identifikátory v rámci datového skladu.
- **Nahrávání (Loading)** - Konečně v posledním kroku se data nahrávají do datového skladu.

Přístup do této vrstvy má jen omezený počet uživatelů, protože data zde uložená nejsou integrována do „jedné verze pravdy“ jako je tomu v datovém skladu. Toto platí jak pro Kimballovu koncepci, tak i pro Inmovu.

4.2.2. Datový sklad – „*Data Presentation*“

Datový sklad je tvořen souborem integrovaných datových tržišť. Data v těchto tržištích jsou ukládána a zobrazována dimensionálně.

Data v této vrstvě musí být v každém případě atomická. Navíc zde mohou být uložena i agregovaná data. Nicméně pro zajištění případných požadavků odběratelů, kteří mohou chtít vidět i detailní data, je třeba data ukládat i na detailní úrovni.

Zároveň všechny tržiště musí být propojeny prostřednictvím konformních dimenzí a faktů, jejichž význam je stejný napříč všemi datovými tržišti. Tato charakteristika datového skladu je nejdůležitější, a je tím „lepidlem“, které stmeluje jednotlivá datová tržiště do jednoho celku – datového skladu.

4.2.3. Přístupové nástroje – „*Data Access Tools*“

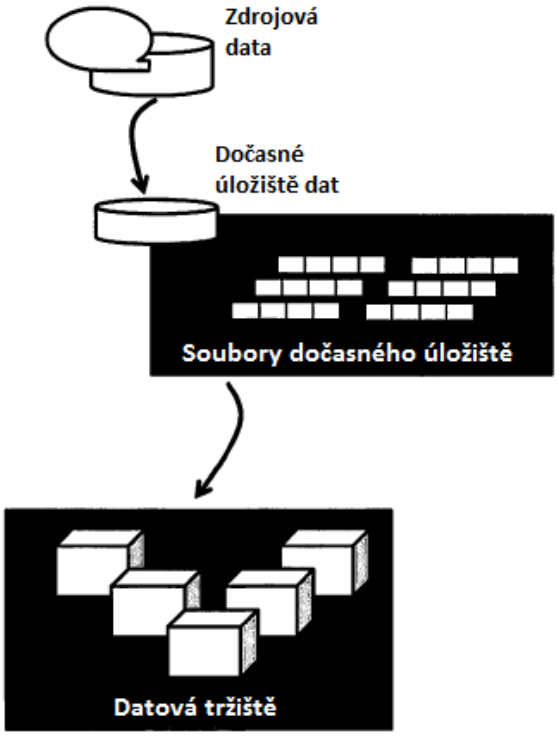
Přístupové nástroje slouží k přístupu odběratelů a aplikací do datového skladu. Přístupovým nástrojem může být jak jednoduchá dotazovací aplikace, tak i komplexní nástroj data-miningu. Některé z těchto nástrojů mohu své výsledky nahrávat zpět do DSA a případně zdrojových systémů.

5. Datové tržiště

Datové tržiště je podmnožina datového skladu a slouží jako výchozí úložiště dat pro odběratele. Tato věta platí pro oba přístupy k realizaci datových tržišť, tedy závislá a nezávislá datové tržiště, které vychází z koncepce datových skladů. Jinak jsou ale oba přístupy založené na různých principech. Nejen popisem těchto přístupů se zabývá právě tato kapitola. Vzhledem k tématu mé práce se více zaměřím na závislá datová tržiště.

5.1.Nezávislé datové tržiště

Nezávislé datové tržiště představuje jeden konkrétní business proces. Data do nezávislého datového tržiště se načítají přímo z DSA, jak je vidět na obrázku č. 5. Data v tomto tržišti jsou uspořádána podle dimensionálního modelu. Dimensionální model obsahuje dva typy tabulek. Tabulku faktů a tabulku dimenzí. Tyto tabulky mohou být uspořádány buď ve schématu hvězda, nebo sněhová vločka.



Obrázek 5 - Nezávislá datová tržiště; Přeloženo ze zdroje: (Wells, 2008 stránky 3-20)

5.1.1. Tabulka faktů (Kimball, a další, 2002)

Tabulka faktů je hlavní tabulkou v dimensionálním modelu. Každé datové tržiště má svou vlastní a právě jednu tabulku faktů. Zároveň představuje zdaleka nejobsáhlejší část datového tržiště. Tabulka obsahuje business ukazatele, jejichž charakteristikou je proměnlivost v čase a cizí klíče, které odkazují do tabulek dimenzí . Primární klíč „*composite key*“ může být složený buď z cizích klíčů, anebo

pokud je třeba, se vytvoří vlastní primární klíč. Nicméně Kimball příliš nedoporučuje druhý postup, a to z důvodu dalšího rozšiřování tabulky. Přičemž index na tomto atributu zůstane nevyužitý.

Hodnota ukazatele neboli fakta je vymezena dimenzemi, které jsou na tento ukazatel navázány, např. výše utracené částky (fakt) se odehrála v určitém čase a v určité prodejně (dimenze). Kimball uvádí, že nejvíce užitečné bývají numerické a aditivní ukazatele, jako třeba objem prodeje. Kromě numerických je možné ukládat i textové ukazatele, nicméně to není obvyklé a je doporučeno tyto ukazatele ukládat spíše do tabulky dimenzí.

Možnost vytvářet aditivní ukazatele, tedy agregovat fakta je důležitou vlastností, protože BI aplikace často využívají agregované hodnoty, spíše než by pracovaly s tisíci či miliony jednotlivých faktů. Ovšem ne všechny ukazatele lze agregovat. Kromě aditivních ukazatelů, existují semiaditivní, u kterých má agregace smysl jen pro některé dimenze a neaditivní, u kterých nemá agregace smysl vůbec. Příkladem aditivního ukazatele je např. obrat na jednotlivých pobočkách za určité časové období. Semiaditivního např. zásoby ve skladu na jednotlivých pobočkách, kde má smysl sčítat zásoby za pobočky, ale nikoliv v čase. Neaditivním ukazatelem je např. průměrný počet uskutečněných zakázek na pobočkách za měsíc.

Další vlastností faktové tabulky je, že fakta jsou uložena na stejné úrovni granularity.

5.1.2. Tabulka dimenzí (Kimball, a další, 2002)

Tabulka dimenzí obsahuje především textové hodnoty, tedy tyto tabulky často představují číselníky. Každá tabulka má vlastní primární klíč „*surrogate key*“. Tento klíč jednoznačně identifikuje dimenzi a je používán k zajištění referenční integrity mezi tabulkou dimenzí a tabulkou faktů.

Je vcelku obvyklé, že tyto tabulky mívají menší počet záznamů než tabulka faktů, zato často obsahují více atributů (sloupců). Což znamená, že tento typ tabulek může být vysoce denormalizovaný. Nicméně míra denormalizace záleží na typu použitého schématu. Obecně ve sněhové vločce jsou tabulky dimenzí více normalizované než ve hvězdě.

Dimenze slouží jako základ pro seskupování a třídění faktů[13]. Proto by se jejich návrhu měla věnovat velká pozornost. Dimenze se také mohou seskupovat do hierarchií, např. u geografie jedna tabulka dimenzí obsahuje města, další tabulka seskupuje města do okresů a další okresy do krajů. Tomuto uspořádání se říká sněhová vločka.

Hodnoty v tabulce dimenzí sice nezávisí na čase jako fakta, na druhou stranu nejsou ani neměnné, a čas od času se vyskytne potřeba tyto hodnoty změnit. V literatuře se tento proces změny

[13] Tato skutečnost je využívána k podpoře analytické techniky „*drill down/up*“, která umožňuje pohybovat se napříč různými úrovněmi detailu. Např. z celkového objemu prodeje se dostat na prodej na jednotlivých pobočkách, až třeba na samotný detail jednotlivých transakcí.

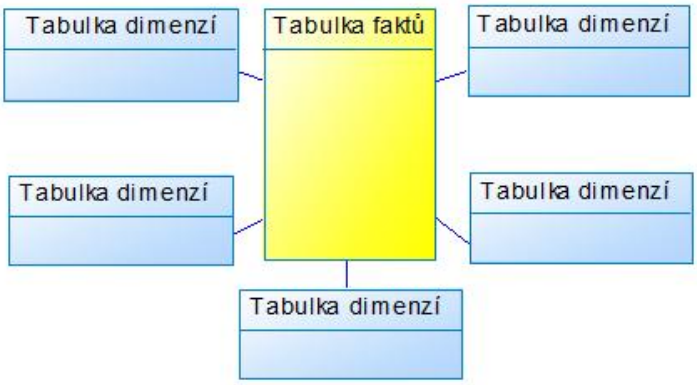
označuje termínem „*slowly changing dimension*“ (SCD). Tyto změny je třeba nějakým způsobem promítnou do tabulky. A existuje více možností – typů SCD jak se s těmito změnami vypořádat:

- **Typ 1** „Přepsání hodnoty“ – Představuje prosté nahrazení původní hodnoty v záznamu tabulky. Výhodou tohoto typu je nenáročnost a rychlost. Nevýhodou je, že nemáme přehled o změnách – nelze sledovat historii.
- **Typ 2** „Přidání záznamu“ – V tomto typu SCD se zachová původní záznam a zároveň se vloží záznam nový s novým primárním klíčem. Záznamy jsou od sebe odlišeny časovou platností, tedy od kdy do kdy je/byl záznam platný. Výhodou je, že máme přehled o historii. Nevýhodou, především u častěji měnících se dimenzí, je nárůst na objemu tabulky.
- **Typ 3** „Přidání sloupce“ – SCD 3 stojí někde na pomezí mezi předchozími typy. Zachováme původní záznam a zároveň vidíme změny. Princip je založen na přidání nového sloupce, kam se přesune původní hodnota. Do původního sloupce se přepíše nová hodnota k příslušnému záznamu. Výhodnou je možnost vidět v jednom záznamu původní i novou hodnotu. Nevýhoda je zřejmá, pokud se budou hodnoty často měnit, tabulka bude neúměrně široká. Z tohoto důvodu se tento typ využívá jen výjimečně.

Často měnící se dimenze, pro které se nehodí SCD 2, Kimball označuje jako „*rapidly changing monster dimension*“. Typicky se toto týká dimenze obsahující data o zákaznících. Tento proces je založen na oddělení často se měnících atributů do samostatné tabulky – mini dimenze. Tabulka faktů bude pak obsahovat jak cizí klíč původní dimenze, tak i klíč do mini dimenze.

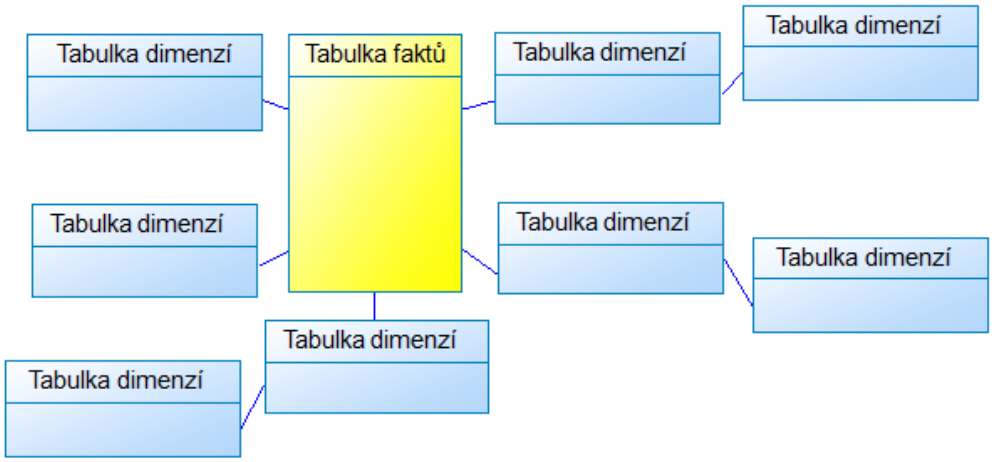
5.1.3. Schémata

Jak se výše uvedené tabulky propojují, zobrazují následující obrázky. Ve hvězdě stojí v centru modelu tabulka faktů, viz obrázek č. 6. Tato tabulka je obklopena dimenzemi, které popisují fakta. Výhodou tohoto provedení je relativní jednoduchost, přehlednost a rychlost. Nevýhoda se projeví u častěji měnících se dimenzí, viz výše „*rapidly changing monster dimension*“.



Obrázek 6 - Dimensionální model: Schéma hvězdy Zdroj: Autor

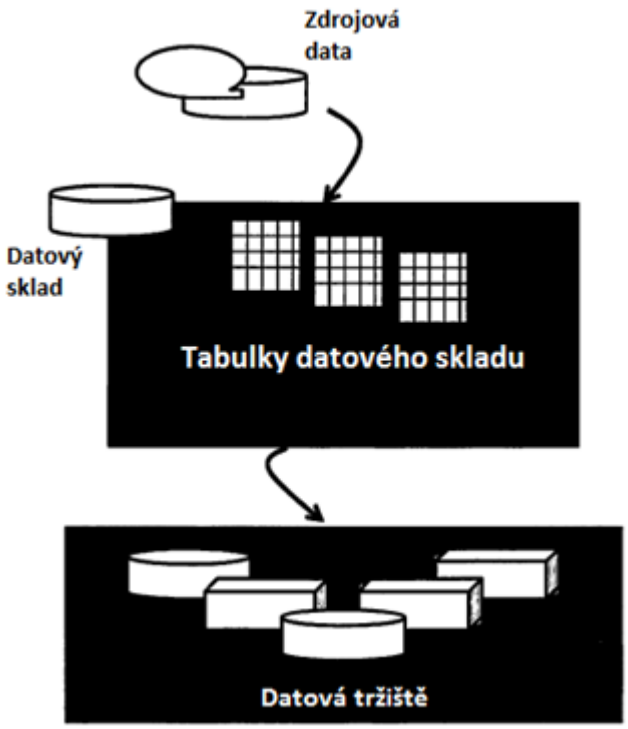
Ve sněhové vločce je rovněž v centru jediná tabulka faktů, která je obklopena dimenzemi, viz obrázek č. 7. Tyto dimenze mohou být navíc napojeny na další dimenze, a to prostřednictvím vztahu n:1. Tímto způsobem je možné optimálně zachytit hierarchie – normalizovat tabulky, a předejít tím duplikaci dat. Výhodou je možnost normalizace, nevýhodou relativní nepřehlednost, možné snížení výkonu z důvodu vzájemného propojování tabulek dimenzí a složitější dotazování.



Obrázek 7 - Dimensionální model: Schéma sněhové vločky Zdroj: Autor

5.2.Závislé datové tržiště

Závislé datové tržiště je primárně plněno z datového skladu[14], viz obrázek č. 8. To znamená, že při budování tohoto typu tržiště musí data projít i přes integrovanou vrstvu. Což může, ale nemusí být finančně a časově náročnější, než v případě nezávislých datových tržišť. Samozřejmě záleží na konkrétním případu. Pokud jsou data uložena již v datovém skladu, je to něco jiného, než když je třeba připojit nový datový zdroj do DSA, následně tyto data „přelít“ do datového skladu a pak teprve tvořit tržiště. Výjimkou kromě zmíněných „Oper Martů“ je plnění datového tržiště ze sémantické vrstvy (kterou popisuji dále), samozřejmě za předpokladu, že je tato vrstva implementována.



Obrázek 8 - Závislá datová tržiště; Přeloženo a upraveno ze zdroje: (Wells, 2008 stránky 3-24)

V článku (Inmon, 2011) Inmon uvádí výhody závislého datového tržiště:

- existence další vrstvy nad datovým skladem, odlehčuje datovému skladu, protože zpracování dat pro odběratele a jejich dotazování zajišťuje právě vrstva datových tržišť
- závislé tržiště se přizpůsobuje potřebám odběratele a je možné tržiště těmto potřebám flexibilně upravovat

[14] Zvláštním případem je plnění tržiště přímo z ODS. Takové tržiště se nazývá „O*per Mart*“ a je „*…podmnožinou ODS, která slouží odběratelům k dimenzionálnímu pohledu na transakční data.“* (Imhoff, a další, 2003 str. 8). Originální definice v angličtině: „*The oper marts are derivatives of the ODS used to provide the business*

- lze snadno vysledovat původ dat v tržišti a vysledovat, kde všude se tyto data využívají
- umístěním tržiště na oddělenou databázi od datového skladu se snižují náklady na zpracování. Protože pokud je třeba opakované načtení tržiště, je výhodnější jej spouštět na vlastním stroji než na centrálním, kde se zpracovává datový sklad.
- všechna datová tržiště vycházejí z jednoho zdroje dat – datového skladu. Pokud je tedy problém s vykazováním v jednom tržišti oproti jinému, je možné se vrátit do datového skladu, kde jsou data vykazována shodně.

S uvedenými výhodami souhlasím, protože jsem toho názoru, že z hlediska výkonnosti, efektivnosti a údržby je klíčové oddělení datového skladu od datových tržišť. A to, i když jsou obě vrstvy zpracovávány na stejném stroji, je zde výhoda, že všechny tržiště přebírají shodná data z jednoho zdroje, v datovém skladu je shodná úroveň detailu dat a je relativně snadné přizpůsobit a měnit podobu tržiště dle potřeb odběratele. Navíc právě díky tomuto oddělení je možné zpracovávat data a obsluhovat dotazy odběratelů odděleně v jiných databázích, což má pozitivní dopad výkonnost v datovém skladu. Proto se v této práci budu dále soustředit na závislá datová tržiště.

Kromě toho, závislé datové tržiště nemusí být nutně navrženo v podobě normalizovaného modelu jako datový sklad, ale může být využito i dimensionální modelování z nezávislých tržišť. Forma uložení dat v tržišti tedy závisí především na aplikaci/odběrateli, který bude s daty pracovat.

5.3. Speciální případy „datových tržišť“

5.3.1. Sandboxy „*Analytical Sandbox*“

Sandbox je „*specializované prostředí, které umožňuje analytikům vytvářet tabulky, nahrávat, seskupovat a agregovat data, kombinovat pole a spouštět dotazy – bez toho, aby tyto činnosti ovlivňovaly ostatní odběratele v datovém skladu*.“[15] (Eckerson, 2008)

Jak napovídá anglický název, jedná se skutečně o „pískoviště“. Tedy o dočasný prostor sloužící odběratelům k zjišťování a ujasnění, jaká data vlastně potřebují. Mají zde možnost kombinovat data z datového skladu i z externích zdrojů, přeskupovat, přepočítávat aniž by tím ovlivňovali samotný datový sklad. Jakmile si tedy odběratelé ověří, jaká data potřebuje, je toto řešení implementováno do „skutečného“ tržiště, potažmo datového skladu případně DSA.

[15] Originální definice v angličtině: „*These are dedicated environments that permit analysts to do all the things analysts love to do-such as creating tables, loading data, merging and aggregating, combining fields, and running queries-without impinging on other users of the data warehousing environment*.“

Vzhledem k výše uvedenému není tento typ tržiště určen k reportování a vykazování získaných dat, protože je zde narušen princip „jedné verze pravdy“. Jak je uvedeno v článku (Eckerson, 2010), tyto tržiště neprosazují informační politiku. Odběratelům tedy v podstatě nic nebrání takto získaná data vykazovat. Proto je třeba stanovit pravidla a poučit odběratele, jak s tímto tržištěm a jeho výstupy zacházet.

(Eckerson, 2010) uvádí tři typy realizace sandboxů, které se mohou mezi sebou kombinovat:

Sandbox jako součást datového skladu

Tradiční forma, která si pro sebe zabírá prostor přímo v databázi datového skladu. Zde si mohou odběratelé dat kombinovat data z datového skladu se svými „vlastními“ daty. Výhodou je, že se nemusí přenášet data ze skladu do jiného prostoru a je zde možnost sledovat činnost odběratelů. A v případě implementace řešení je výstup uložen v datovém skladu, což usnadní práci BI týmu. Nevýhodou je, že se musí nastavit práva a prostor, ve kterém se bude pracovat. Vzhledem k zabíranému prostoru je třeba vyvážit výkon mezi uživateli skladu a tržiště. Navíc při upgradu databázového stroje, je možné, že bude třeba toto tržiště upravit.

Zdvojený sandbox

Další forma je založena na oddělení virtuálního tržiště od datového skladu. A to zkopírováním datového skladu na další samostatný stroj, který bude určen pouze uživatelům virtuálních tržišť. Tím se omezí výše zmíněné nevýhody. Nevýhody jsou u toho řešení zřejmé. Pokud se pomine duplikace dat, je třeba zajistit a udržovat další prostředí navíc, což si vyžaduje zvýšené finanční i personální kapacity.

Sandbox na vlastním počítači

Tato forma realizace virtuálního tržiště využívá počítače samotného odběratele. Některé analytické nástroje si ukládají tabulky do své paměti. Uživatel má tak takřka neomezený přístup k datům a možnosti jak s těmito daty pracovat. Pomocí těchto nástrojů je pak možné tyto výstupy i publikovat do firemní sítě. Pro detailnější informace k těmto nástrojům doporučuji projít zmíněný článek. Relativní nevýhodou je, že výsledky práce na tomto tržišti uvidí pouze samotný uživatel. Také zde může být problém s výkonností, zejména při procesu nahrávání dat do paměti aplikace.

5.3.2. Sémantická vrstva

Jedním z problémů, se kterým se mimo jiné potýká i datový sklad, potažmo datová tržiště a reporting je propojení atributů v datovém skladu s pojmy, kterým rozumí odběratelé. Či ještě lépe sjednocení významu jednotlivých pojmů. Svým způsobem se tento problém snaží řešit sémantická vrstva:

„Sémantická vrstva je interpretace podnikových dat v pojmech známých business uživatelům. Cílem vrstvy je odstínit uživatele od technologického řešení a poskytnout jim možnost provádět dotazy nad jim známými pojmy.“[16] (Daniels, 2010)

Tento pojem původně pochází z Business Objects. Jak uvádí definice, smyslem sémantické vrstvy je přizpůsobit uživatelům dotazování v pojmech, kterým rozumí. A v podobě, kterou si sami zvolí. Pokud se tato technika u uživatelů ujme, budou si uživatelé schopni zajistit velkou část výstupů vlastními silami. Což odlehčí práci vývojářům a analytikům v datovém skladu a uživatelé budou mít možnost vytvořit si výstup kdykoliv bude třeba.

Tato myšlenka zní lákavě, ale zde je ta podmínka „pokud se ujme“. Mohou se vyskytnout tací uživatelé, které tato technika příliš nenadchne. Protože očekávají, že přístup pro ně připraví někdo jiný. Na druhou stranu se nevylučuje, aby tuto vrstvu využívali i vývojáři. Pak je ale otázkou, zda vývojář raději nepoužije „starý známý“ skriptovací jazyk. Navíc ne všechny požadavky na vstupy bude možné pokrýt z této vrstvy. Kromě toho sémantická vrstva, jakožto obecně platná vrstva, vyžaduje sjednocení významu dat, což předpokládá existenci podnikového byznys slovníku. Právě tato skutečnost pravděpodobně hraje největší roli při rozhodování, zda zavádět sémantickou vrstvu.

Přes výše uvedené, může tato vrstva najít své využití. Třeba při ujasňování podoby o obsahu výstupu je možné poměrně rychle a interaktivně vytvořit prototyp výsledné sestavy. Nezanedbatelný přínos je i v zjednodušení tvorby sestavy. Pokud by se tato technika ujala v širším měřítku, byla by zároveň i zajištěna konzistence mezi výstupy získanými z této vrstvy. Dalším přínosem je dodání jednotného bussinesového významu atributům z datového skladu. A v neposlední řadě je zde i snazší údržba sestav než u sestav získaných pomocí skriptu. Zde stačí provést změnu na jednom místě, kdežto v druhém případě je třeba upravit každý skript zvlášť.

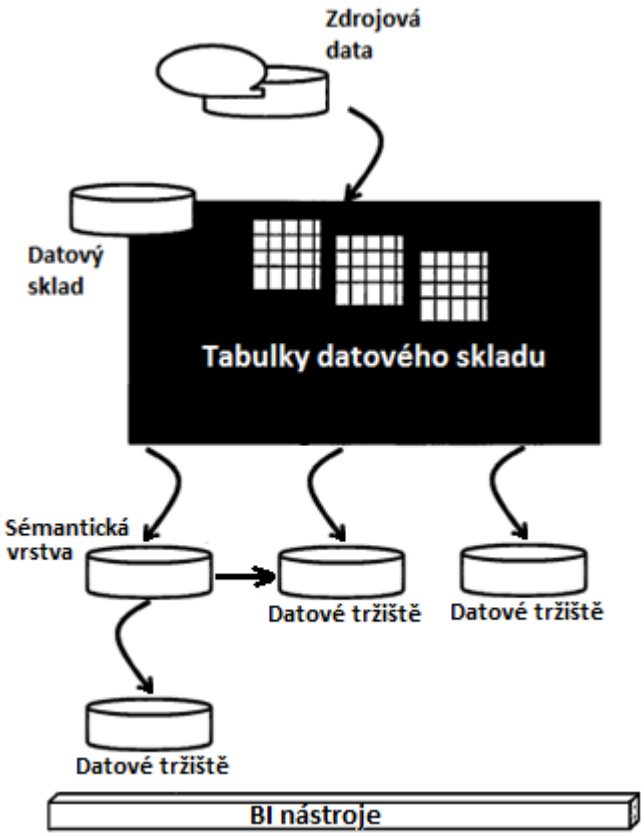
Tato vrstva tedy může sloužit jako úložiště, kde budou vystavována data jak uživatelům a to nejen z businessu, tak i samotným aplikacím a BI nástrojům. A je zde případně možné aplikovat business pravidla, transformace, výpočty. Podmínkou je, že tato vrstva bude stejně jako závislá

---

[16] Originální definice v angličtině: *„A semantic layer is a business representation of corporate data that helps end users access data using common business terms. The aim is to insulate users from the technical details of the data store and allow them to create queries in terms that are familiar and meaningful.“*

datová tržiště čerpat data z datového skladu. Díky tomuto bude zajištěn jednotný zdroj již integrovaných a ověřených dat. S tím rozdílem, že těmto datům bude dodán business rozměr a budou zde možné takové transformace, které z nejrůznějších důvodů nelze provést přímo v datovém skladu. Jak už jsem psala v kapitole věnující se závislým datovým tržištím, datové tržiště se plní z datového skladu, nicméně je zde jedna výjimka, a to plnění tržišť ze sémantické vrstvy.

Jakým způsobem zapadá sémantická vrstva do architektury datového skladu, ukazuje obrázek č. 9. Je zde patrné, že sémantická vrstva se plní z datového skladu. Pak zde mohou být tržiště, která se plní pouze ze sémantické vrstvy, pouze z datového skladu, případně z obou zdrojů. Jak sémantická vrstva, tak i datová tržiště, dokonce i datový sklad mohou sloužit jako zdroje dat pro BI nástroje.



Obrázek 9 - Sémantická vrstva; Přeloženo a doplněno autorem ze zdroje: (Wells, 2008 stránky 3-24)

Aby bylo možné s touto vrstvou pracovat, je tedy třeba vyřešit její plnění. Zde se nabízí k řešení otázka, kdy a jak často plnit sémantickou vrstvu? Vzhledem k dnešním, stále se zvyšujícím se nárokům odběratelů dat, kdy nejsou výjimkou datová tržiště plnící se s denní frekvencí a pravidelné denní reporty, doporučuji denní plnění. Nicméně alespoň s jednodenním zpožděním a to proto, aby plnění sémantické vrstvy nezasahovalo do bussinesového okna (zpracování uživatelských dotazů) datového skladu. A to ať už důvodu pozdního dodání dat na straně zdrojového systému, tak případného pomalejšího zpracování v datovém skladu.

Další skutečností, na kterou je třeba myslet, je, že musí být dohledatelné, z jakých dat pochází data v sémantické vrstvě. Otázkou také je, na jakých časových osách se budou data nejčastěji propojovat. Tedy jestli bude odběratele zajímat aktuální stav, nebo budou chtít vidět např. aktuální údaje o klientovi, ale zůstatky na účtu za poslední čtvrtletí.

Dále je třeba zvážit, jakým způsobem se budou uchovávat časové snímky, tedy historizace dat v sémantické vrstvě. Zde stejně jako v předchozích případech opět záleží na požadavcích odběratelů. Například je možné držet historii denních snímků za poslední tři měsíce, nebo jeden aktuální denní snímek, snímek za ultimo měsíce, čtvrtletí, atd.

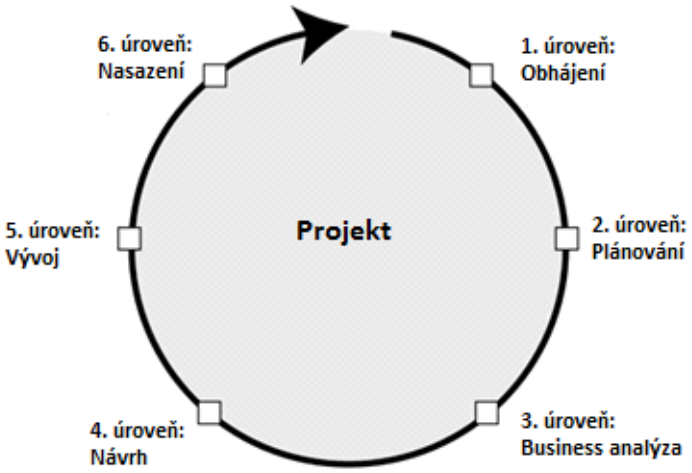
Pokud jde o formu ukládání dat, tedy jestli využít dimensionální nebo relační modelování, tak toto vždy záleží na použité technologii a požadavcích odběratelů. Nicméně bych zde vyloučila tu variantu, při které by sémantickou vrstvu představovala jedna „dlouhá“ tabulka. A to jak z důvodu nepřehlednosti, obtížné manipulaci s touto tabulkou, tak i s výkonnostními problémy vznikajícími při plnění této tabulky. Spíše si lze představit sémantickou vrstvu jako „pomyslný soubor datových tržišť“, které obsahují data platná napříč celou organizací, nejsou tedy přizpůsobeny jednomu konkrétnímu odběrateli.

6. Metodiky řízení BI projektů a vývoje BI řešení

Až do této části práce jsem se věnovala především architektuře, tedy popisu jednotlivých komponent BI potažmo datového skladu, funkcionalitě těchto komponent a jakým způsobem mezi sebou tyto komponenty fungují. V této části práce se budu zabývat tím, jakým způsobem tyto komponenty vyvíjet a zavádět, tedy na metodiku. A jak už napovídá název práce, zaměřím se především na vývoj a údržbu závislých datových tržišť.

6.1.Řízení BI projektu dle (Moss, a další, 2003)

V první „metodické“ části práce popisuji BI projekt, protože ať už jde o zavedení nového BI nástroje, tržiště či přímo datového skladu, musí každý projekt týkající se BI projít jednotlivými fázemi vývoje. Samozřejmě nemusí všemi, záleží na konkrétním projektu a některé fáze je možno vynechat. Tyto vývojové fáze jsou pro všechny komponenty víceméně společné, tedy platí i na datová tržiště. Obrázek č. 10 a jeho popis znázorňuje jednotlivé fáze BI projektu. Metodika obsahuje celkem šest fází, které se dále rozpadají na šestnáct kroků, které popisuji dále.



Obrázek 10 – Vývojové úrovně projektu; Přeloženo ze zdroje: (Moss, a další, 2003)

6.1.1. Fáze Obhájení („*Justification*“)

Podnět ke vzniku projektu na základě posouzení potřeb odběratelů dat. V této fázi projektu se hodnotí přínosy, případně výhody, které vzniknou zavedením řešení (krok „*Business Case Assessment*“).

6.1.2. Fáze Plánování („*Planning*“)

Fáze plánování řeší návrh strategických a taktických plánů, na základě kterých bude projekt vyvíjen a posuzován. Plánování v sobě zahrnuje plánování projektu (krok „*Project Planning*“) a zhodnocení podnikové architektury (krok „*Enterprise Infrastructure Evaluation*“).

Zhodnocení enteprise architektury je třeba proto, že BI zasahuje takřka do všech částí organizace. A k tomu, aby mohlo být efektivně využíváno, je třeba, aby BI mělo v enterprise architektuře podporu. A to nejen technologickou, ale i v podobě standardizace meta dat, jmenných konvencí, enterprise logického datového modelu, metodik, procesů a dalších.

6.1.3. Fáze Byznys analýza („*Business analysis*“)

Představuje detailní analýzu byznys problému s cílem získat požadavky odběratelů dat na vznikající řešení. V první části probíhá definice rozsahu projektu (krok „*Project Requirements Definition*“).

Následuje datová analýza (krok „*Data Analysis*“), kde je klíčové vyřešit kvalitu zdrojových dat a význam dat. Vzhledem k tomu, že zvyšovaní kvality dat často znamená změnit staré návyky, zjistit a sjednotit význam dat napříč celou organizací, je třeba počítat s tím, že datová analýza zabere velkou část v celém projektu.

V dalším kroku probíhá vývoj prototypu řešení (krok „*Application Prototyping*“). Cílem tohoto kroku je ukázat odběratelům výhody i omezení daného řešení. Navíc v této fázi, pokud by bylo třeba, je relativně snadné přizpůsobit řešení požadavkům. Spíše než když je dodáno již finální řešení.

Posledním krokem fáze byznys analýzy je analýza úložiště meta dat (krok „*Meta Data Repository Analysis“).* Technická meta data vznikající v průběhu projektu by měla být mapována na business meta data. A všechny meta data by měly být uloženy v úložišti. Požadavek na to, která meta data ukládat specifikuje logický meta model.

6.1.4. Fáze Návrh („*Design*“)

Představuje návrh takového řešení, které bude splňovat požadavky odběratelů dat. V prvním kroku fáze návrhu probíhá návrh databáze (krok „*Database Design*“). Při návrhu databáze je důležité brát v úvahu požadavky odběratelů, tedy zda chtějí/nechtějí detailní či agregovaná data, zda je možné využít dimensionální modelování a tyto požadavky promítnout do struktury databáze.

Dále probíhá návrh ETL procesů (krok „*Extract/Transform/Load Design*“). Pravděpodobně nejsložitější krok z celého BI projektu a zároveň i ten méně zajímavý. Složitost se projeví zejména tam, kde je nízká kvalita zdrojových dat.

Závěrečný krok této fáze je návrh úložiště meta dat (krok „*Meta Data Repository Design*“). A to proto, že ne vždy splňuje vybraný produkt úložiště meta dat všechny podmínky definované v logickém modelu meta dat.

6.1.5. Fáze Budování („*Construction*“)

Fáze budování začíná vývojem ETL procesu (krok „*Extract/Transform/Load Development*“).

Následuje vývoj řešení, kdy se po skončení vývoje prototypu začne pracovat na analýze a vývoji již finálního řešení (krok *„Application Development“*). Toto může představovat jak poměrně jednoduché dopracování řešení z funkčního prototypu, tak i předělávání prototypu na jiný robustnější nástroj.

Jako další krok autoři uvádějí Data Mining a vývoj úložiště meta dat (krok *„Meta Data Repository Development“*), to v případě, kdy by se organizace rozhodla vyvinout vlastní úložiště meta dat.

6.1.6. Fáze Údržba (*„Deployment“*)

Představuje zavedení nebo prodej výsledného řešení a jeho následné zhodnocení, zda splňuje předpokládanou návratnost investic. Jakmile je vyvíjené řešení otestováno začíná se s krokem zavádění (krok *„Implementation“*). Dále se školí uživatelé, spouští se podpora uživatelů a nastavují se procesy údržby řešení.

Konečně po nasazení začíná zhodnocení, a to nejen finálního řešení, ale i celého projektu (krok *„Release Evaluation“*).

6.2. Metodika vývoje dle Inmona - *„Spiral Methodology“*[17]

Inmon se ve své metodice vývoje datového skladu inspiroval „spirálovým“ přístupem využívaným při softwarovém vývoji, s kterým poprvé přišel Barry W. Boehm v roce 1986[18]. Na první pohled se „spirálový“ přístup podobá „klasickému“ přístupu vývoje v IT, a to „SLDC – System Development Lifecycle“ nebo také „vodopádový“ přístup. Z hlediska obsahu jednotlivých fází, jsou si oba přístupy podobné, liší se způsobem získávání požadavků a iteracemi vývoje řešení.

„Vodopádový“ přístup se využívá k vývoji a rozvoji v transakčních systémech, pro BI řešení není úplně vhodný. To proto, že „vodopádový“ přístup počítá s jasně definovanými a pokud možno neměnnými požadavky, pokud se změní požadavek, je třeba upravit celé řešení. Přičemž postup jednotlivých fází probíhá sekvenčně, tedy následující fáze čeká na výstup fáze předchozí.

„Spirálový“ přístup začíná s daty – proto také Inmon označuje metodiku jako metodiku řízenou daty. A končí identifikací požadavků, přičemž se nevylučuje, aby probíhala zároveň např. analýza a návrh. Přístup je vhodný pro BI řešení zejména proto, že není výjimkou, když se požadavky odběratelů v průběhu BI projektu mění. Metodika počítá s tím, že odběratelé nevědí, co chtějí. Tedy

---

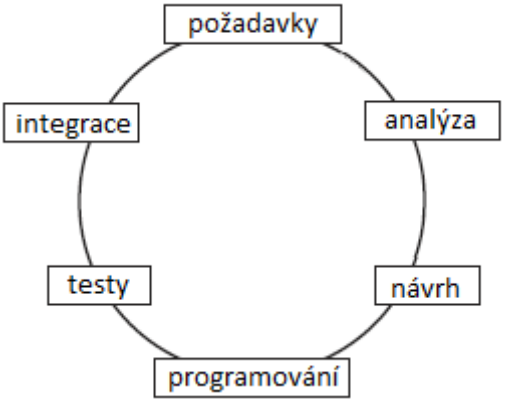
[17] Zdrojem pro tuto kapitolu byly následující prameny: (Inmon, a další, 2008), (Inmon, 2004)

[18] Pro zajímavost uvádím odkaz na článek „A Spiral Model on Software Development and Enchancement“, ve kterém Boehm rozvíjí spirálovou metodu vývoje softwaru: http://www.cs.umd.edu/class/spring2003/cmsc838p/Process/spiral.pdf

mají „nějakou“ představu o tom, co chtějí, ale když pak dostanou to, co požadovali, zjistí, že vlastně potřebují něco jiného. A následuje další vývoj řešení, a pokračuje se jednotlivými iteracemi až do té doby, kdy budou uspokojeni odběratelé a budou známy požadavky na BI řešení.

6.2.1. „Spirálový“ přístup

Cílem metodiky tedy je rychlý vývoj malých, pokud možno znovupoužitelných částí řešení v rychlém sledu za sebou. Vývoj řešení probíhá v jednotlivých iteracích, kdy se malá část řešení dokončuje, zatímco jiná část už byla dokončena. Jednotlivé fáze metodiky (viz obrázek č. 11) mohou v iteraci probíhat v libovolném pořadí, takže na sobě nejsou zcela závislé. Počítá se s tím, že první iterace není konečná a bude dále upřesňována. Z čehož plyne, že metodika klade důraz na prototypování.



Obrázek 11 – Jednotlivé fáze „spirálového“ přístupu; Přeloženo ze zdroje: (Inmon, a další, 2008 str. 124)

Aby měl „spirálový“ přístup úspěch, je třeba přijmout k podnikové datové architektuře novou strategii. Inmon tuto strategii nazývá jako sedmi proudový přístup. Každý proud je zaměřen na nějakou jinou část datové architektury a je současně řízen, koordinován a monitorován. Není předem dané pořadí, v jakém se mají proudy vykonávat. Zaměření jednotlivých proudů je následující (Inmon, a další, 2008):

Proud podnikového modelu - „Enterprise reference model stream“

Proud zaměřený na vývoj a údržbu podnikového datového modelu. Přičemž datový model je budován inkrementálně, po jednotlivých samostatných oblastech (např. produkty, segmenty, osoby…). Tento proud představuje klasické techniky, používané při vývoji databází, jen je aplikován v mnohem větším měřítku a postupně.

Proud koordinace podnikových znalostí: „Enterprise knowledge coordination stream“

Tento proud spojuje tři různé výstupy, a to z podnikového datového modelování, vývoje CIFu a z profilace dat. Účelem je poskytnutí znovupoužitelných a spolehlivých znalostí o podnikových datech.

Proud vývoje CIFu „Information Factory development stream“

V tomto proudu je vyvíjen samotný CIF, a to po jednotlivých tématech (např. ziskovost, růst organizace…). V tématech se pak řeší jednotlivé business požadavky tzv. „burning questions“, které se třídí podle priority, a na které odběratelé potřebují znát odpověď, a to proto, aby se mohli vhodně rozhodovat. Každé z témat může zasahovat do více subjektů, např. při řešení ziskovosti se budeme dívat na produkty, segmenty a osoby (osobami není myšleno pouze fyzická osoba, ale třeba i pobočka). Priority zadané v tomto proudu řídí i priority ve všech ostatních proudech. Tedy na základě priorit je pak určováno, která entita se bude nasazovat v jaké iteraci. Z čehož také plyne, že právě požadavky odběratelů dodávají obsah datovému modelu. Datový model pak slouží jako propojovací bod mezi jednotlivými iteracemi. Právě v tomto proudu se uplatňuje „spirálový“ přístup.

Proud datové profilace a mapování „Data profiling and mapping stream“

Zde probíhá vyhodnocování dat ze zdrojových systémů, zejména se hodnotí úplnost a kvalita dat. Účelem je poskytnutí podkladů pro mapování zdrojových dat do CIFu.

Proud datové korekce „Data correction and stream“

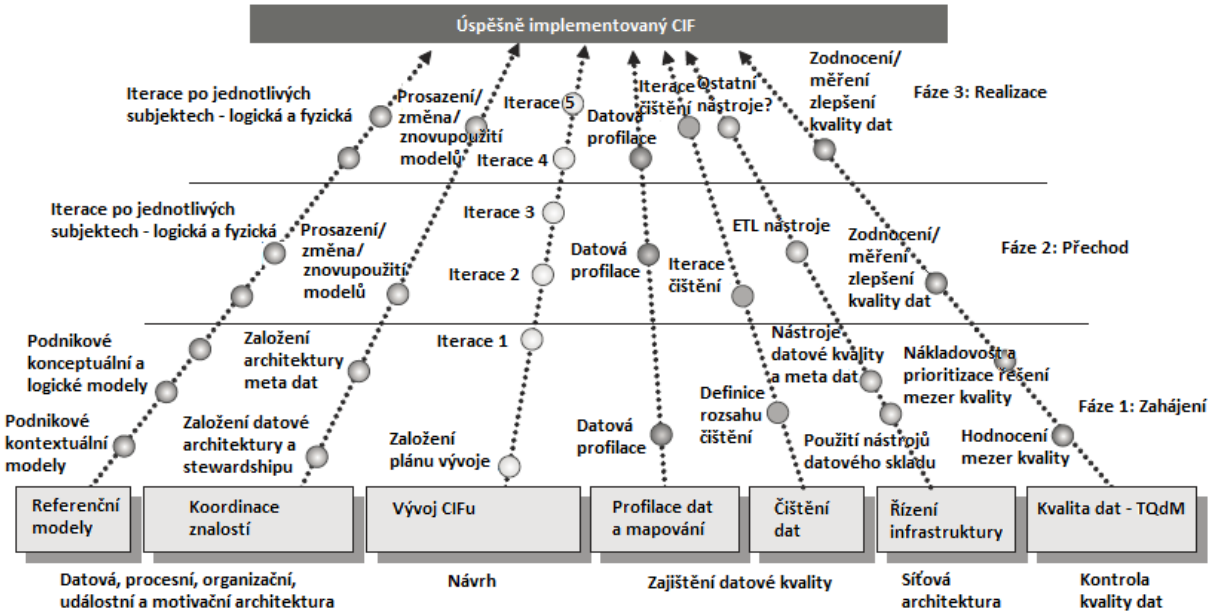
V tomto proudu probíhá čištění, kompletace a případná oprava zdrojových dat, jinak řečeno právě zde probíhá proces ETL/ELT.

Proud infrastruktury „Infrastructure stream“

Zde probíhá podpora infrastruktury, tak aby byla zajištěna budoucí rozšiřitelnost CIFu. A prostřednictvím definování a následné aktualizace zdrojů, nástrojů, standardů, osob a restrikcí.

Proud řízení kvality dat „Total information quality management stream“

Proud zaměřený na průběžné sledování kvality dat. Což představuje jak proces monitorování kvality dat, tak i hledání a eliminaci dat s nízkou kvalitou a odhad nákladů vzniklých v důsledku „nekvalitních“ dat.



Obrázek 12 - Znázornění průběhu sedmi proudového přístupu; Přeloženo ze zdroje: (Inmon, a další, 2008 str. 138)

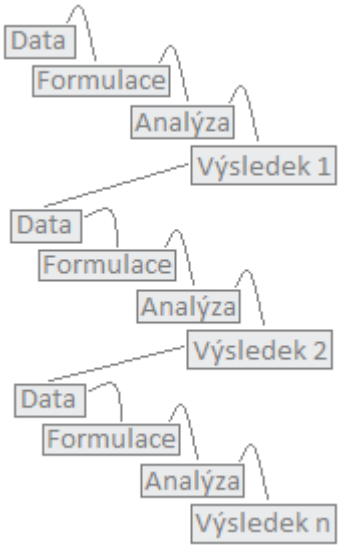
Obrázek č. 12 detailněji znázorňuje průběh výše zmíněných proudů. Zároveň je zde vidět, v jaké fázi se současně provádí aktivity v jednotlivých proudech.

Spojení „spirálového“ přístupu se sedmi proudovým tak zajistí, že při vývoji a údržbě datového skladu bude k dispozici datový model, budou definovaná business pravidla a definice dat, známa kvalita zdrojových dat a bude připravena odpovídající infrastruktura.

Jako další variantu způsobu vývoje a tedy i upřesňování požadavků uvádí Inmon ve svém článku (Inmon, 2004) heuristický přístup.

6.2.2. Heuristický přístup (Inmon, 2004)

V následujícím obrázku č. 13 je znázorněn princip heuristického přístupu. Výsledek jedné iterace slouží jako vstup – formulace požadavku pro další iteraci. Jedná se o značně neformální metodu, u které jsou na začátku jen minimálně definovány požadavky. Ze všech uvedených přístupů v této podkapitole podává, vzhledem ke své neformálnosti, výsledky nejrychleji.



Obrázek 13 - Heuristický přístup; Upraveno a přeloženo ze zdroje: (Inmon, 2004)

Jedním z cílů tohoto přístupu je tedy právě zjištění, jaké jsou vlastně požadavky odběratelů. Inmon uvádí typický příklad použití například při zjišťování, proč poklesl prodej v posledním měsíci. Při zjišťování důvodu je na data pohlíženo z různých stran, mění se požadavky na to, jaká data jsou vlastně třeba a v jaké formě.

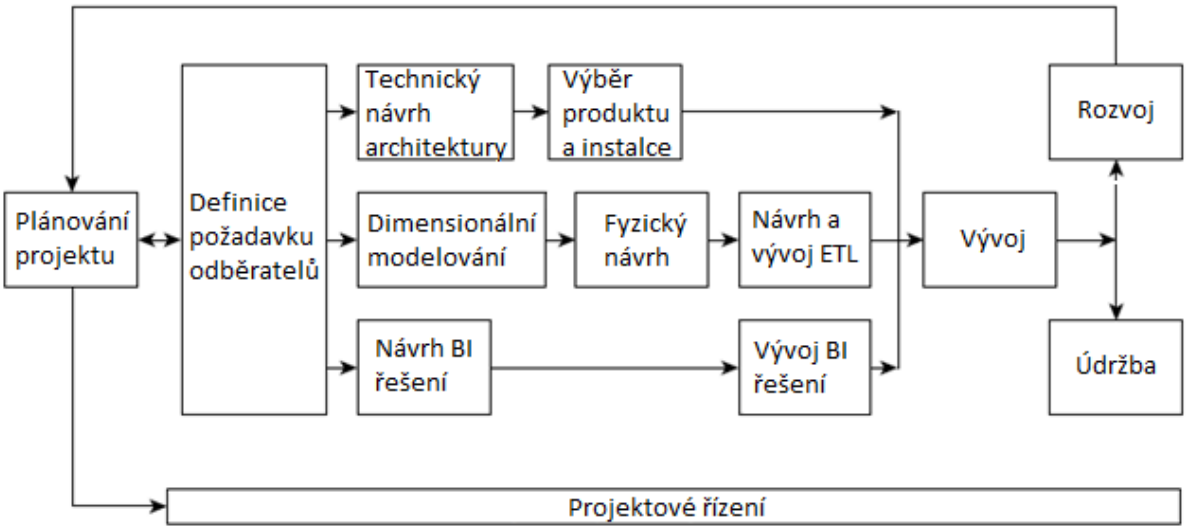
V rámci tohoto přístupu není zaručeno, že zadaný problém má řešení. V podstatě, je obtížné i naplánovat harmonogram, protože nedokážeme říct, kolik iterací bude třeba a jakého výsledku se má dosáhnout. Tento přístup najde své využití v analytických a DSS systémech. Výstupy získané pomocí tohoto přístupu je možné použít jednorázově a jen pro daný konkrétní problém.

6.3.Metodika vývoje dle Kimballa - *„Kimball Lifecycle“*[19]

Obrázek č. 14 ukazuje posloupnost jednotlivých fází nutných pro vývoj BI řešení dle Kimballa. Šipky naznačují závislosti mezi fázemi, jejichž posloupnost je nutné respektovat.

Kimball zdůrazňuje, že cílem je především dodání takového řešení, které uspokojí odběratele. A také, že vývoj BI řešení nelze brát jako jednorázovou záležitost, ale třeba počítat s neustálým přizpůsobováním řešení odběratelům. Na druhou stranu je třeba myslet na to, že každý projekt má jasně vymezený začátek a konec. Ostatně toto platí i pro všechny ostatní metodiky. Navíc jak už vyplývá z povahy metodiky, je tato metodika zaměřena na dodání dat v dimensionální podobě.

[19] Zdrojem pro tuto kapitolu byly následující prameny: (Kimball, a další, 2010), (Kimball, a další, 2002)



Obrázek 14 – Metodika vývoje dle Kimballa; Přeloženo ze zdroje: (Kimball, a další, 2010 str. 97)

Metodika začíná plánováním projektu a získáváním požadavků odběratelů. Po těchto dvou krocích následují tři paralelní větve – technologická, datová a BI větev.

Technologická větev – představuje soubor činností, které zajistí technické zázemí pro vývoj a provoz datového skladu. Patří sem definování požadavků, které musí technologie - produkt splňovat a výběr a instalace technologie.

Datová větev – obsahuje činnosti, jejichž cílem je poskytnout požadovaná data odběratelům. Je zde zahrnuto dimensionální modelování, které bylo probráno v kapitole věnující se datovým tržištím. Fyzický model, který se generuje z dimensionálního modelu. V poslední části probíhá návrh a vývoj ETL.

BI větev – tato větev se soustředí na vývoj BI aplikací, pomocí kterých budou odběratelé přistupovat k datům. Pod pojmem BI aplikace se skrývá tvorba sestav, interaktivních reportů, analytických modelů či data miningových aplikací.

Tyto tři větve se spojují do jednoho bodu – vývoje. Další fází je pak údržba vyvinutého řešení, nebo pokud je třeba, je možné pokračovat další iterací vývoje řešení.

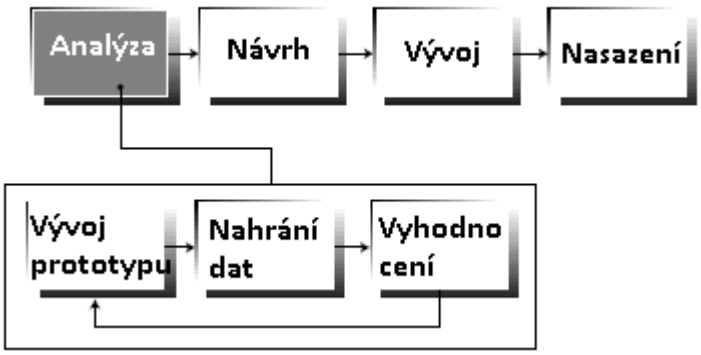
Napříč všemi uvedenými fázemi probíhá projektové řízení, které zajišťuje synchronizaci jednotlivých fází. Probíhá zde kontrola průběhu projektu, řeší se zde případné problémy a sledují se požadavky, aby nebyl překročen rozsah projektu. Zároveň projektový management zajišťuje komunikaci mezi odběrateli a vývojáři.

6.4. Metodika vývoje datového tržiště dle společnosti Talus – „PrimeMover Life cycle“ (Talus)

PrimeMover metodika je určena především pro IT manažery, kteří plánují svůj první projekt výstavby datového tržiště. Primárně je zaměřena na vývoj nezávislých datových tržišť a za použití techniky dimensionálního modelování. Nicméně lze některé principy aplikovat i na vývoj závislých datových tržišť, potažmo datového skladu. Metodika používá inkrementální vývoj datového tržiště. A zároveň využívá prototypizaci a iterativní vývoj. V metodice jsou zdůrazněny dvě pravidla, na která je třeba brát zřetel. Jedná se o pravidlo složitého nahrávání dat a pravidlo nejistoty při návrhu.

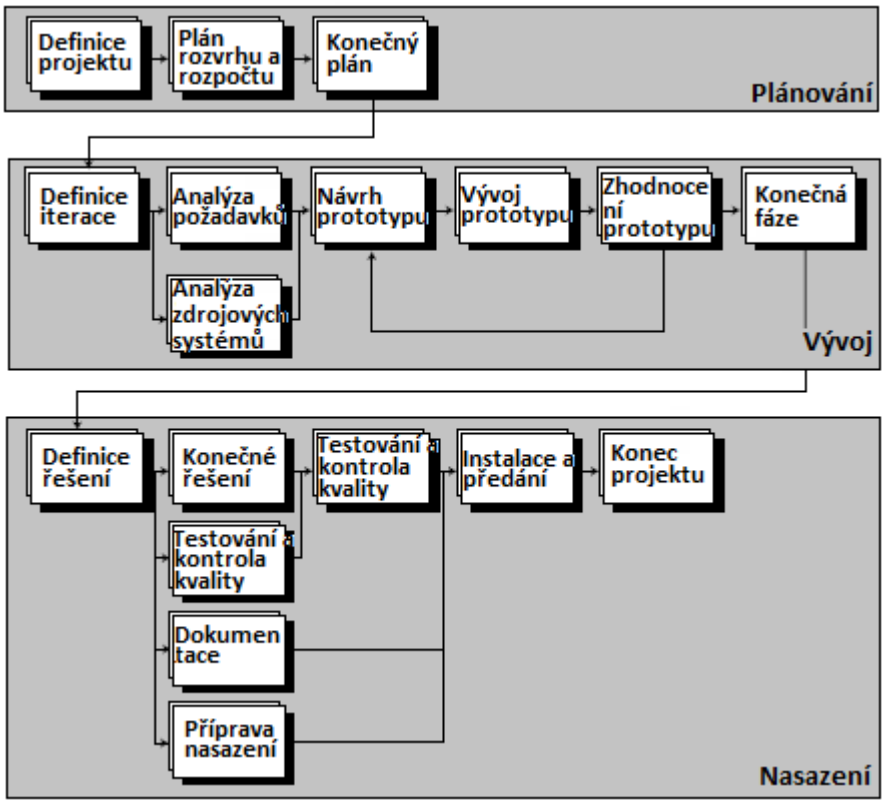
Pravidlo složitého nahrávání dat (v podstatě se jedná o ETL/ELT proces) sestává ze získání dat ze zdrojového systému, jejich transformace a agregace a nahrání dat do datového skladu. Rizika, která mohou při zpracování ETL/ELT procesu nastat, jsou nízká kvalita dat, chybějící nebo neúplná dokumentace zdrojového systému, načítání dat z více systémů či výskyt redundantních dat. Metodika upozorňuje, že ETL/ELT proces je ta nejsložitější, časově nejnáročnější a časově nejvíce riziková část výstavby datového tržiště.

Pravidlo nejistoty návrhu upozorňuje, že *„je nemožné vyhodnotit návrh datového tržiště, aniž by byla nahrána data ze zdrojových systémů a byla předvedena odběratelům“* (Talus str. 5). Proto je vhodné poskytnout koncovým odběratelům data získaná pomocí ETL/ELT procesu a nechat je tyto data vyhodnotit. Teprve pak se může přistoupit k návrhu tržiště. Tato interakce s odběrateli je velmi důležitá, protože předchází riziku, že uživatelé odmítnou výstup z projektu trvajícího v řádech měsíců, který stál nemálo časových, peněžních a lidských zdrojů. Jednoduše proto, že neobsahuje taková data, jaká odběratelé požadují. Jakým způsobem této situaci předejít ukazuje obrázek č. 15. Na obrázku je vidět, že vyhodnocení dat odběrateli probíhá v začátcích projektu. Díky tomu, že se získá zpětná vazba od odběratelů co nejdříve, a je využit prototyp a iterativní vývoj, není problém upravovat výslednou podobu tržiště ke spokojenosti odběratelů.



Obrázek 15 - Znázornění, jakým způsobem předcházet riziku neakceptace tržiště odběrateli; Přeloženo ze zdroje: (Talus str. 7)

Obrázek č. 16 znázorňuje detailní kroky každé z jednotlivých fází budování datového tržiště – plánování, vývoje a nasazení.



Obrázek 16 - Detailní přehled jednotlivých kroků a fází PrimeOver metodiky; Přeloženo ze zdroje: (Talus str. 14)

Plánování je zaměřeno na definici rozsahu projektu. Vývoj se zaměřuje na konstrukci prototypu, který vyhovuje požadavkům odběratelů dat a slouží k naplánování fáze nasazení. Výstupem fáze nasazení je konstrukce výsledného řešení a dokumentace k řešení. Přičemž každá z fází respektuje dvě výše uvedená pravidla.

6.5.Zhodnocení metodik

Metodika řízení BI projektu je robustní metodika zaměřená jak na řízení projektu, tak i na obecný popis jakými způsoby lze jednotlivé kroky vyřešit. Využívá prototypizaci a myslí i na meta data. Vzhledem ke své robustnosti se dá aplikovat jak na vývoj datového skladu, tak i na vývoj datového tržiště. Výhodou také je, že ne všechny kroky uvedené v metodice je nutné vykonat. Na konci popisu každého kroku je uvedeno riziko, které hrozí při přeskočení daného kroku a odkazy na doplňující literaturu, což je velmi vítané. Relativní nevýhodou může být právě robustnost metodiky. U této metodiky jsem se inspirovala v tom, že u každého kroku jsem uvedla za jaké situace je vhodné krok použít a jaká jsou rizika v případě, kdy dojde k přeskočení kroku.

Inmonova metodika je rovněž relativně robustní, pokud uvažuji celou publikaci jako metodiku, ale primárně je zaměřená především na vývoj datového skladu. Detailně popisuje technický způsob vývoje řešení, ale méně se zaměřuje na projektové řízení. Na první pohled díky propojení „spirálového“ přístupu a sedmi proudého přístupu může působit chaoticky. Kromě toho se odvolává i na Metodiku řízení BI. Kladné je využití iterativního vývoje a protypizace a využití „spirálového“ přístupu. Zde jsem se inspirovala možností využití iterativního vývoje, prototypizace a také principem, kterým je možné získat požadavky od odběratelů.

Kimballova metodika je obdobně jako Inmonova – zaměření na CIF, zaměřena na Kimballův koncept - nezávislá datová tržiště. Nepočítá s prototypizací a iterativním vývojem, vždy je třeba dojít k fázi vývoje, kdy pak může nastat buď údržba, nebo další rozvoj řešení. Sice zmiňuje projektové řízení, ale ani v jednom ze zdrojů uvedených v kapitole jsem nenašla, konkrétní postupy k této fázi. Z toho soudím, že obdobně jako předchozí metodika, je zaměřená spíše na technickou stránku řešení. Relativní nevýhodou může být zaměření na vývoj podle dimensionálního modelu. Jako výhoda je možnost zpracovávat všechny tři větve paralelně. Kladný je rovněž popis způsobu získávání požadavků od odběratelů. Tuto metodiku jsem při návrh své metodiky nakonec nevyužila a dala přednost Inmonově metodice. Především proto, že v organizaci je použita architektura, která vychází z CIFu.

Na metodice PrimeMover hodnotím kladně její přehlednost, srozumitelnost a možnost využít její principy nejen při návrhu nezávislého datového tržiště. Dále využití iterativního vývoje za pomocí prototypů, zapojení odběratelů do procesu vývoje a popis výstupů jednotlivých fázi. Ze všech uvedených metodik se jeví jako nejvíc „odlehčená“, tedy zaměřená pouze na vývoj a dodání řešení v rámci projektu. Z mého pohledu je metodika určena především pro projektového manažera, tedy říká, co je třeba udělat, ale neřeší jakým způsobem. Jako nevýhodu vidím to, že metodika neuvažuje o údržbě a explicitně nezmiňuje způsob získávání požadavků od odběratelů. Z této metodiky jsem převzala její dva hlavní principy, a to počítání s náročností při přípravě podkladových dat. A zároveň pravidlo nejistoty návrhu, kde jsem ve své metodice využila jednak prototypizace a jednak využití sandboxů (což souvisí i s prvním pravidlem).

Ani jedna z uvedených metodik neřeší vyřazení řešení z provozu. Na to je třeba myslet v praktické části, při samotném návrhu metodiky.

7. Analýza a popis současného stavu

Následující analýza vychází z pohledu datového stewarda a zkušeností, které jsem v organizaci získala při vývoji datových tržišť. První část kapitoly je věnována všeobecnému popisu procesu správy tržiště a to z procesního a architektonického pohledu. Následuje detailní popis, který začíná popisem procesu přijetí požadavku na vývoj či úpravu datového tržiště v organizaci. Následně je popsán proces zavedení tržiště do provozu a detailnější popis rolí, které v procesu vystupují. Na to navazuje popis nasazení, údržby a vyřazení tržiště z provozu. Navíc jsem do této kapitoly doplnila i využití sandboxů a reportovací vrstvy v organizaci.

V závěru kapitoly, na základě výše popsaných procesů, jsem zhodnotila, co je v organizaci ošetřeno a co by bylo třeba do metodiky doplnit. Z tohoto pak budu vycházet v kapitole, která se zabývá vlastním návrhem metodiky. Na konec procesu změnového řízení jsem pro lepší názornost zakreslila detailní průběh procesu spolu s rolemi, které se tohoto procesu účastní. Viz obrázek č. 19. Tento proces v sobě zahrnuje proces vývoje/úpravy tržiště od přijetí požadavku po nasazení. Tedy kromě pravidelné údržby a odstavení z provozu obsahuje vše, co je v této kapitole popsáno.

7.1. Vysvětlení pojmů používaných v organizaci

Na začátek je zařazeno vysvětlení pojmů, kdekteré budou používány v této kapitole a kapitole obsahující metodiku. Důvodem je přiblížit text zaměstnancům v organizaci, tak aby se v něm mohli rychle zorientovat, a proto je použito „názvosloví“, které je běžně v organizaci využíváno.

- **CDS (Central Data Store)** – jedná se o ekvivalent pojmu Corporate Information Factory- CIF a to doslova. Architektura datového skladu v organizaci vychází právě z této koncepce.
- **Stage** - vrstva dočasného úložiště neboli DSA (pozor neplést s aplikací DSA = Development Support Application)
- **Target** - vrstva integrovaného datového skladu
- **Interní datové tržiště** – tržiště, ze kterého se data nenahrávají do externích aplikací. Zpravidla bývá plněno SQL skriptem.
- **Externí datové tržiště** – tržiště, které slouží jako zdroj dat pro externí aplikace. Zpravidla bývá plněno ETL procesem, ale může být i SQL skriptem.
- **DSA (Development Support Application)** - aplikace určená pro ukládání specifikací datových struktur a mapování. V rámci změnového řízení je specifikacím a mapováním přiřazeno číslo změnového řízení.
- **Bikipedie** - interní wikipedie, která je využívána k dokumentaci a předávání znalostí

- **Specifikace** – definice datové struktury, zpravidla uložená v aplikaci DSA
- **Mapování** – znamená definování odkud, kam a za jakých podmínek budou probíhat datové transformace. Děje se na dvou úrovních a to zdroj-target nebo target-datové tržiště. Zpravidla bývá uložené v aplikaci DSA, ale může být i v MS Excell.
- **ETL proces** – představuje realizaci mapování.
- Pro účely nasazování požadavků do CDS jsou v organizaci zavedena tři prostředí – vývojové, testovací a produkční:
    - **Vývojové prostředí** - je určeno především pro technicky orientované pracovníky, tedy zejména pro úpravu datového modelu targetu a vývoj DDL skriptu a ETL procesu.
    - **Testovací prostředí** - je určeno jak k technickým tak uživatelským testům při nasazování změn v CDS.
    - **Produkční prostředí** - slouží k reportingu a k přístupu odběratelů k datům.

Popis rolí vyskytujících se v procesu:

- **Odběratel dat** – je iniciátorem požadavků do CDS.
- **Business partner** – vystupuje jako primární kontaktní osoba pro odběratele dat a tím zajišťuje sběr uživatelských požadavků. Tyto požadavky pak předává do oddělení CDS. Zároveň ve spolupráci s oddělením CDS prioritizuje požadavky a pomáhá odběratelům s definicí jejich požadavků.
- **Datový steward** – vystupuje jako kontaktní osoba za svěřenou oblast v datovém skladu. Může mít na starosti nejen datová tržiště, ale i specifikaci zdrojových extraktů a jejich mapování do targetu. Zároveň vytváří a podílí se na realizaci požadavků na úpravy v CDS ve svých svěřených oblastech. Požadavky je myšlena na tvorba/změna specifikací datových struktur a datových transformací.
- **Koordinátor změnového řízení** – řídí proces změnového řízení a požadavky na kapacity ETL vývojářů a DDL exekutorů.
- **Datový architekt** – zajišťuje rozvoj targetu (rozšiřováním datového modelu), a podílí se na upgradech SW i HW komponent v CDS.
- **DDL exekutor** – zajišťuje generování DDL skriptů. V procesu se vyskytuje ve chvíli, kdy se zakládají/mění datové struktury ve stage, targetu a datovém tržišti.

- **ETL vývojář** – má na starosti vývoj ETL procesu a nastavování pravidelných plnících úloh – jobů[20]. Případně verzování plnících SQL skriptů.
- **Koordinátor nasazování** – řídí nasazování změnového řízení do testovacího a produkčního prostředí.

Dokumenty využívané v procesu vývoje a správy datových tržišť:

- **Plán mapování** – evidence požadavků změn v CDS. Vyplněním plánu mapování žádá datový steward o otevření změnového řízení.
- **Průvodka zavedením rozhraní** – slouží k vytvoření kanálu mezi zdrojovým systém a datovým skladem nebo mezi datovým skladem a externí aplikací.
- **Průvodka změnovým řízením** - průvodka zajišťuje synchronizaci všech činností spojených s realizací požadavků na změny v CDS.
- **Metodika mapování do datového tržiště** – určuje jak postupovat při mapování z targetu do datového tržiště.
- **Metodika mapování do targetu** – určuje jakým způsobem provést mapování do targetu a jak si toto mapování otestovat.
- **Metodika psaní SQL** – definuje nejen formální požadavky na plnící SQL skript datového tržiště, ale i z jakých databází je možno data do tržiště nahrávat.
- **Metodika odstraňování implementovaných požadavků v CDS** – popisuje proces odstraňování nejen tržiště z CDS.
- **Šablona k nastavení datových kontrol** – na základě vyplněných údajů o tom, co a jak se bude sledovat, se automaticky vygeneruje SQL skript. Tento skript se pak zařadí do automatického zpracování a zároveň se vytvoří reporty ke sledování datových kontrol v aportovacím nástroji.
- **Šablona SLA (= Service Level Agreement)** – dohoda o úrovni poskytovaných služeb mezi oddělením datového skladu a odběratelem.
- **Šablona dokumentace k datovému tržišti** (existují dvě, a to v Excelu a na Bikipedii)
- **Žádost o službu MIS** – slouží odběratelům k zadávání požadavků do CDS

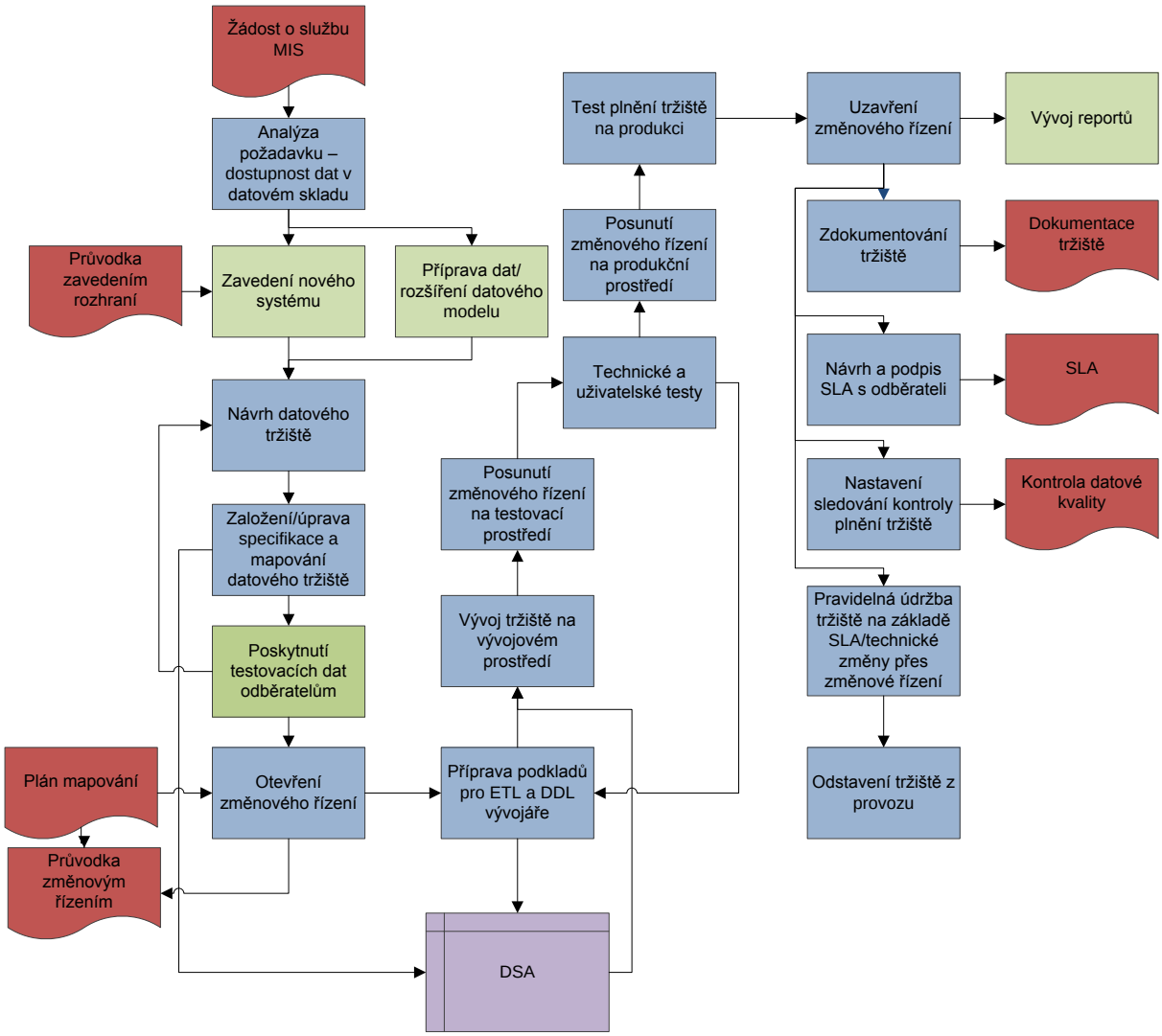
[20] Joby jsou automatické úlohy, které zajišťují automatické pravidelné plnění tabulek v stage, targetu a datových tržištích.

7.2. Všeobecný popis procesu správy tržiště z procesního hlediska

Obrázek č. 17 a č. 18 ukazuje pohled datového stewarda na správu datového tržiště, č. 17 z hlediska procesního a č. 18 z hlediska architektury CDS. Jedná se o zjednodušený popis. Na konci této kapitoly na obrázku č. 19 je pak detailnější popis i s jednotlivými rolemi.

Zeleně označené procesy na obrázku č. 17 mohou být vynechány, záleží na konkrétní situaci. Oproti tomu modře označené procesy by měly být vykonány vždy. Některé procesy mohou být vykonávány paralelně (např. „Zavedení nového systému“ a „Příprava podkladových dat pro tržiště“) a některé se mohou i vícekrát opakovat (cyklus „Příprava podkladů pro ETL vývojáře a DDL exekutora“ – „Technické a uživatelské testy“). Některé procesy mají vazbu na standardizované dokumenty např. proces „Zavedení nového systému“ na dokument „Průvodka zavedením rozhraní“. Aby bylo zřetelné, jak se které dokumenty promítají do jednotlivých vrstev datového skladu, vypracovala jsem ještě obrázek č. 18. K obrázku č. 18 se v následujícím textu ještě vrátím. Zde uvádím jen pro vysvětlení vztahu mezi obrázky č. 17 a č. 18.

Vzhledem k tomu, že celý proces jako takový není v organizaci popsán, je pořadí některých kroků a rozhodnutí zda je vůbec vykonávat na libovůli a zkušenostech datového stewarda a také v závislosti na rozsahu datového tržiště. Upozorňuji, že popis procesu správy datového tržiště a obrázky v této kapitole jsou mým dílem a byly vytvořeny na základě mých zkušeností a pozorování v organizaci. Tudíž se mohou lišit v tom, jak to vidí zkušenější kolegové. Proto jsem na závěr práce zařadila kapitolu, kde jsou zaznamenány reakce kolegů na popis současného stavu a na samotnou metodiku.



Obrázek 17 - Zjednodušený proces vývoje datového tržiště v organizaci; Zdroj: autor

V první fázi procesu probíhá zadání požadavku odběratelem. Na základě tohoto požadavku se provádí analýza dat v CDS. Dle výsledků analýzy se pak přikročí buď k doplnění dat do CDS, což může i nemusí ovlivnit datový model targetu. Nebo se zjistí, že všechna data jsou už k dispozici, a tak není třeba nic doplňovat. Každopádně, pokud se vytváří nové tržiště, datový steward ve spolupráci s odběratelem vyplní „Průvodku zavedením rozhraní“.

Za předpokladu, že jsou všechna požadovaná data v targetu, nastává samotný návrh datového tržiště. Další krok je pak volitelný a záleží na uvážení datového stewarda a rozsahu tržiště, zda v této fázi poskytne vzorky dat odběratelům. Jestliže je připravený návrh tržiště, datový steward žádá o otevření změnového řízení prostřednictvím vyplnění plánu mapování. Na základě vyplněného plánu mapování koordinátor změnového řízení vyplňuje průvodku změnovým řízením.

Po vytvoření průvodky změnovým řízením datový steward připraví specifikace tabulek, mapování a případně SQL skript. Na základě těchto podkladů začínají DDL exekutor a ETL vývojář s vývojem datového tržiště.

Vývoj DDL a ETL skriptů probíhá vždy na vývojovém prostředí. Po ukončení vývoje následuje posunutí změnového řízení na testovací prostředí.

Testovací prostředí slouží k ověření funkčnosti změnového řízení, a to jak z pohledu odběratelů dat, tak i z pohledu datového stewarda a samotného plnění datového tržiště. Za předpokladu, že testy dopadly v pořádku, se změnové řízení nasadí do produkčního prostředí. Pokud testy nedopadly korektně, datový steward vrací změnové řízení zpět na vývojové prostředí. Další postup závisí na tom, zda je chyba v zadání datového stewarda (datový steward upraví mapování/specifikaci datových struktur) nebo v realizaci zadání, tedy ETL procesu (ETL vývojář upraví ETL proces/plnící joby).

V případě, že je vše v pořádku, probíhá nasazení změnového řízení do produkčního prostředí a postimplementační kontrola.

Kroky nastavení datových kontrol, dokumentace a návrh a podpis SLA mohou probíhat paralelně. Vývoj reportů v reportovacím nástroji je volitelný na základě potřeb odběratele. Samotný vývoj repotů je rovněž velmi komplexní proces, který je dle mého názoru dobře zdokumentován, a proto se mu v této práci podrobněji nebudu věnovat.

Po nasazení nového tržiště do produkčního prostředí následuje pravidelná údržba a sledování plnění tržiště datovým stewardem. Pravidelnou údržbou se myslí zavádění změn v tržišti. Změny mohou být vyvolány jak potřebami odběratele, tak i z důvodu technické údržby na straně CDS (např. při úpravě datového modelu targetu, změny ve zdrojových extraktech ve stage). Pokud máme informaci od odběratele dat, že tržiště již nevyužívá, dochází k vypnutí plnění tržiště a jeho smazání z databáze.

7.3. Všeobecný popis procesu správy tržiště z pohledu architektury datového skladu

Na obrázku č. 18 jsou zobrazené stejné dokumenty jako na obrázku č. 17. Rozdíl je v tom, že na tomto obrázku je vidět, do kterých částí CDS, dané dokumenty zasahují. Navíc jsou zde vyznačeny možnosti přístupu k datům uloženým v CDS z pohledu odběratele dat. Pro lepší přehlednost je v diagramu znázorněn pouze odběratel dat, datový steward se podílí více či méně na vyplnění všech dokumentů. Detailní znázornění všech rolí a procesů je na konci kapitoly na obrázku č. 19.

Pokud tedy propojím obrázek č. 18 s obrázkem č. 17, tak proces začíná zasláním požadavku od odběratele dat k datovému stewardovi. Na základě analýzy dat, pokud se zakládá nové tržiště,

nebo pokud je třeba vytvořit nový zdrojový systém ve stage, datový steward doplňuje dokument „Průvodka zavedením rozhraní“. V případě tržiště se na vyplnění podílí i odběratel. Na obrázku č. 18 je tedy znázorněno, že zavedení rozhraní se může vztahovat jak k stage, tak k vrstvě datových tržišť.

Pokud jsou všechna data připravena v targetu, a je vytvořený systém pro datové tržiště, vyplňuje datový steward dokument „Plán mapování“. Kromě změn ve vrstvě datových tržišť slouží plán mapování i ke změnám v targetu a stage.

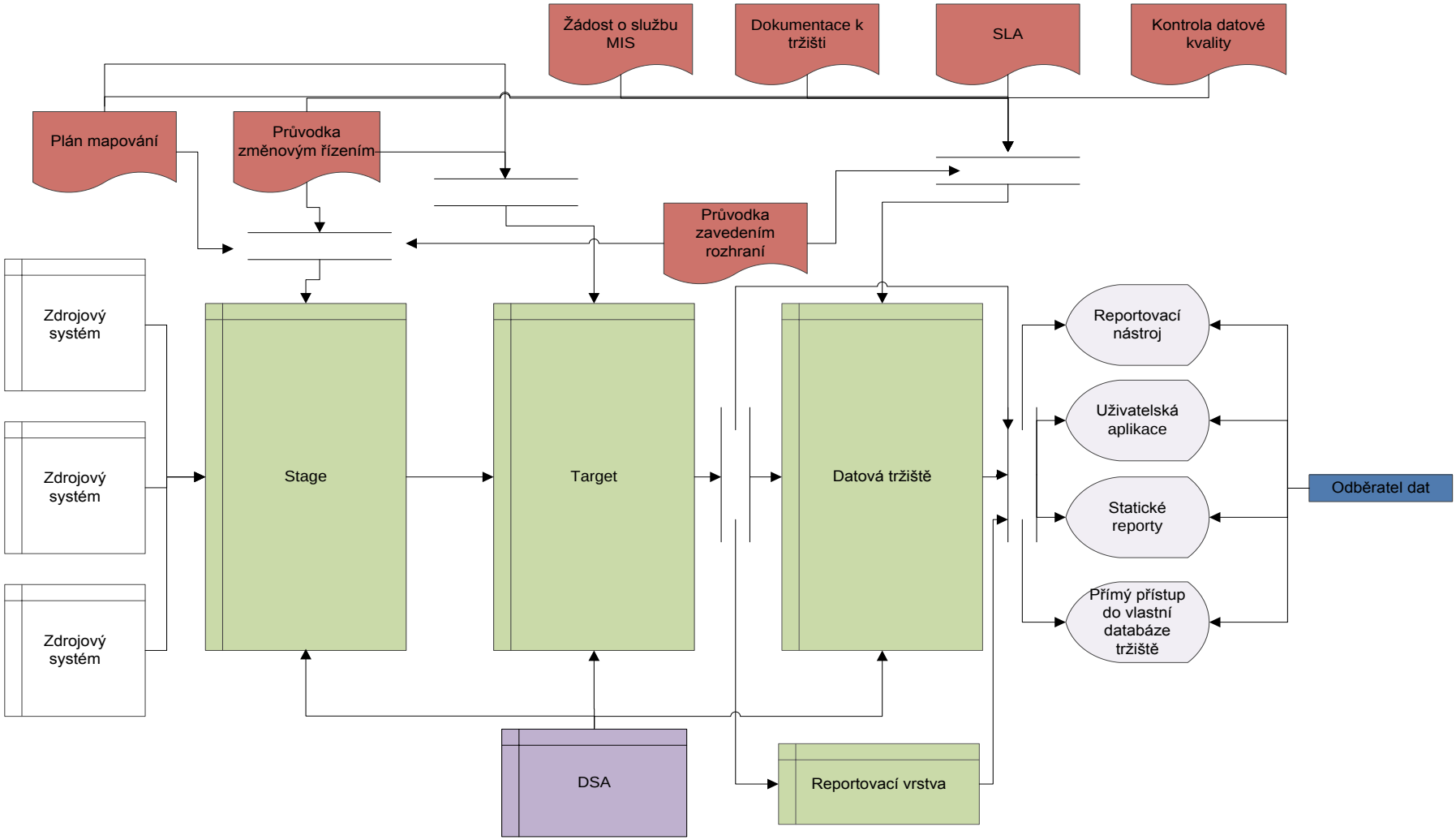
Na základě vyplněného plánu mapování zakládá a vyplňuje koordinátor změnového řízení „Průvodku změnovým řízením“. Tato průvodka může zasahovat do všech vrstev v CDS, jak je vidět na obrázku č. 18.

Na ukládání specifikací a mapování je vytvořena speciální aplikace „DSA“. Tato aplikace je zařazena do procesu změnového řízení. Na obrázku č. 18 je vidět, že aplikace slouží nejen k ukládání mapování z targetu do tržiště a specifikaci datových struktur tržišť. Kromě toho je zde uložena specifikace datových struktur ve stage a targetu, a zároveň i mapování ze stage do targetu.

Po uzavření průvodky změnovým řízením, datový steward nastavuje kontroly sledování kvality dat. Na obrázku č. 18 je znázorněno, že datová kvalita se sleduje jak na vstupu v stage, tak na výstupu ve vrstvě datových tržišť. Spolu s tímto procesem probíhá dokumentace k tržišti a návrh a podpis SLA s odběratelem.

Na obrázku č. 18 jsou také znázorněny možnosti přístupu odběratele k datům uloženým v datovém skladu. V případě, že se data z tržiště nahrávají do uživatelské aplikace, mluvíme o externím datovém tržišti, v ostatních případech se jedná o interní tržiště.

Kromě toho je na obrázku č. 18 znázorněna reportovací vrstva. Tato vrstva byla vytvořena pro účely reportingu odběratelům dat. Reportovací vrstva je blíže popsána v samostatné podkapitole.



Obrázek 18 - Architektura CDS a související dokumenty; Zdroj: autor

7.4.Detailní popis procesu správy datového tržiště

7.4.1. Přijetí požadavku na úpravu/vývoj datového tržiště

V současné době přijímá požadavky na úpravu, rozvoj či přímo vývoj nového tržiště business partner případně datový steward. V případě vývoje nového datového tržiště je zaveden nový projekt. Tento projekt je pak řešen v koordinaci s business partnerem, datovým stewardem, týmovým vedoucím a případně manažerem útvaru.

V případě úpravy datového tržiště si odběratel prostřednictvím formalizované šablony „Žádost o službu MIS“ zadá požadavek na úpravu či vývoj datového tržiště. V případě, že se jedná o drobnou úpravu, je tato úprava realizována v rámci rutinní činnosti datového stewarda. Jestliže se jedná o výraznější zásahy či nové tržiště, je třeba předem zamluvit kapacity oddělení datového skladu. Záleží pak na rozsahu změny, jestli se úprava provede v rámci standardního SLA nebo se vypíše nový projekt. Nový projekt se vypisuje vždy, když je vyvíjeno nové tržiště.

Ve všech zmíněných případech probíhá na počátku analýza, jaká data jsou v CDS k dispozici. Pokud nejsou, je třeba si tyto data předpřipravit. Což znamená, že datový steward vytvoří nové mapování ze stage do targetu, případně pokud data nejsou ani v stage, založí/rozšíří nový extrakt do stage V obou případech je třeba, aby si datový steward podal žádost o změnové řízení a to prostřednictvím vyplnění dokumentu „Plán mapování“. Změnové řízení popisuji v následné kapitole.

Samostatnou kapitolou je pak zavedení nového zdrojového systému na vstupu do CDS či vytvoření systému pro nové tržiště. Toto se pak řeší samostatnou průvodkou „Průvodka zavedením rozhraní“. Viz obrázek č. 18. a kapitola všeobecný popis z architektonického pohledu.

Dodatečné dokumenty, které zde nejsou zmíněné, ale je pravděpodobné, že je datový steward v této fázi využije:

- návod k vyplnění průvodky zavedením rozhraní – v případě zavádění nového zdrojového/výstupního systému
- metodika mapování do targetu – použije se v případě, kdy nejsou v této vrstvě všechna data požadovaná odběratelem

Co je definováno:

- šablona k požadavku o vývoj/rozšíření datového tržiště neboli „Žádost o službu MIS“

Co není definováno:

- způsob získávání požadavků od odběratelů
- postup provedení analýzy dat v datovém skladu

7.4.2. Příprava tvorby/úpravy datového tržiště

V organizaci není blíže specifikován proces, jakým způsobem probíhá samotný vývoj či rozvoj datového tržiště. Tedy není jasně vymezené, jestli provést mapování a přípravu skriptu, a pak ručně exportovat data pro odběratele, počkat si na odezvu odběratele a pak až s finální podobou zadání vyplnit žádost o změnové řízení. Anebo si připravit mapování a skripty, vyplnit žádost, a pak pokračovat dále ve změnovém řízení. Obecně dobrým zvykem je, v případě, že si odběratel není jistý zadáním a technicky nic nebrání tomu ručně exportovat data, zvolit první přístup. Pokud je zadání komplexnější, či není technicky možné si připravit skript, nebo je třeba nahrát data do externí aplikace, zvolit druhý přístup. V obou případech je možné využít sandbox.

Při návrhu záleží na tom, zda se jedná o interní či externí datové tržiště a na jeho složitosti a rozsahu. Přičemž interní datové tržiště je v organizaci definováno jako tržiště, ze kterého se data nikam dále neexportují. Odběratelé mají na stejném databázovém stroji jako je umístěna CDS vytvořenou vlastní databází, která je realizována pomocí views, a kde si odběratelé sami mohou analyzovat data z tržiště. Interní datové tržiště se zpravidla plní SQL skriptem.

Oproti tomu data z externího tržiště se dále exportují do externích aplikací. Samozřejmě, ani v tomto případě nic nebrání tomu, poskytnout uživatelům vlastní databázi. Pokud jde o složitost a rozsáhlost, jen těžko si lze představit, že budeme ručně spouštět SQL skripty a exportovat data z tržiště, které má rozsah dvaceti tabulek, plní se 3 hodiny a tabulky jsou mezi sebou provázané. Externí datové tržiště se zpravidla plní ETL skriptem, ale není tomu tak vždy. Pokud nahráváme data přímo do externí aplikace ani jiná možnost není. Pokud se ale data exportují do souboru a je technicky možné[21] plnit tržiště SQL skriptem, pak je možné zvolit i plnění pomocí SQL skriptu.

Dále není určeno, jak navrhovat datové modely tržiště a zda vůbec. Na druhou stranu, je velmi dobře popsáno jak zapisovat mapování a jak konstruovat plnící SQL skript.

V případě plnění tržiště SQL skriptem, si steward „odladí“ SQL skript ve všech prostředích CDS, a zároveň zajistí, aby skript dobíhal v rozumném čase. Co je rozumný čas, samozřejmě závisí na objemu zpracovaných dat. U ETL skriptu tato možnost samozřejmě chybí, zde odladění skriptu závisí na ETL vývojářovi.

Mapování je zároveň i dokumentací k datovému tržišti, je zde popsáno, co, odkud, kam a za jakých podmínek bude plněno.

Realizaci požadavku na vývoj/úpravu tržiště zajistí datový steward prostřednictvím vyplnění plánu mapování, čímž podává žádost o otevření změnového řízení.

---

[21] V organizaci existují datová tržiště, u nichž jsou takové podmínky na plnění, že je nelze zaznamenat pomocí SQL skriptu. Pak je nutné realizovat plnění ETL skriptem.

Co je definováno:

- metodika mapování do datových tržišť
- metodika psaní SQL
- plán mapování, který slouží jako plánování kapacit pro oddělení provozu datového skladu

Co není definováno:

- návrh a tvorba datového modelu tržiště
- proces vývoje/rozvoje tržiště ve spolupráci s odběrateli

7.4.3. Proces změnového řízení

Žádost o změnové řízení neboli vyplnění plánu mapování, slouží k alokaci kapacit v oddělení provozu datového skladu, tzn. ETL vývojáře, DDL vývojáře a případně i datového architekta. Při vyplňování žádosti se vyplňuje co, kam a kdy bude mapováno.

Změnové řízení je řízeno pomocí formalizovaného dokumentu „Průvodka změnovým řízením“. V tomto dokumentu je definováno, kdy a co se bude nasazovat a jaké kroky je k tomu třeba provést. V každém kroku je vyplněno, kdo má krok vykonat a kdy a kdo ve skutečnosti vykonal daný krok. Po vykonání pak vykonavatel odešle avízo následujícím alokovaným pracovníkům. Krok představuje jednu konkrétní činnost, např. vygenerování DDL skriptu, vývoj ETL procesu, testování dat atd.

Příprava zadání pro DDL exekutora a ETL vývojáře

Je realizované datovým stewardem, prostřednictvím mapování a specifikací datových struktur v již zmíněné aplikaci DSA a přiřazením čísla změnového řízení v této aplikaci. Přiřazení slouží k tomu, aby bylo poznat, co se v rámci změnového řízení bude nasazovat. Pokud se datové tržiště plní SQL skriptem, připraví datový steward tento skript pro ETL vývojáře.

Vývoj datového tržiště

V této fázi DDL exekutor připraví DDL skript, tedy za předpokladu, že datový steward vytváří či mění datové struktury tržiště. Jen pro úplnost uvádím, protože to nemusí být zřejmé, že datový steward může upravovat/zakládat specifikace datových struktur ve stage a tržišti a nikoliv targetu. Na to má právo pouze datový architekt.

Po úpravě DDL skriptu přichází na řadu práce ETL vývojáře. Jeho úkolem je zaverzovat SQL skript eventuelně vytvořit či změnit ETL skript, a to na základě mapování. Pro vývoj ETL skriptů je v organizaci zaveden specializovaný ETL nástroj. Na základě těchto skriptů pak ETL vývojář nastaví

automatické úlohy – joby. V jobech je definováno, z jakých zdrojů a jak často se daná tabulka bude plnit.

Nasazení změnového řízení do testovacího prostředí

Po ukončení ETL vývoje následuje nasazení změnového řízení na testovací prostředí. Nasazení do testovacího prostředí provádí koordinátor nasazování, který zároveň řídí i nasazení do produkčního prostředí a i případné vrácení změnového řízení do vývojového prostředí.

Koordinátor nasazování zároveň kontroluje, že se všechny joby z technického hlediska zpracovaly v pořádku. Pokud ne, následuje komunikace s datovým stewardem, který spolu s koordinátorem nasazování a případně i ETL vývojářem identifikuje příčinu chyby.

Testování změnového řízení

Za předpokladu, že vše proběhnulo korektně, následuje proces testování. A to jak uživatelské, tak i technické testování. Technické testování slouží k tomu, aby si datový steward ověřil, že se datové tržiště plní tak, jak bylo předpokládáno.

Uživatelské testy jsou pak určeny pro odběratele, kteří mají možnost si data „osahat“ a vyzkoušet, zda dostali požadovaná data. Pokud některý z těchto testů neprojde, následuje vrácení změnového řízení na vývojové prostředí. A to až ke kroku, kdy datový steward označoval mapování, eventuelně připravoval skript. Toto se opakuje do doby, kdy odběratel akceptuje výsledky získané z datového tržiště. Případně do doby, kdy datový steward vyhodnotí, že se tržiště plní požadovaným způsobem.

Podle příčiny neúspěchu testu, se pak upravuje mapování, skripty či se kontroluje nastavení jobů. Zde je opět třeba uvést, že v organizaci je výborně ošetřeno testování při nahrávání nových zdrojů do DSA a při mapování dat do datového skladu. Ovšem testování datového tržiště zatím blíže není popsáno. Takže je na uvážení datového stewarda, jakým způsobem si ověří správnost plnění datového tržiště.

Nasazení změnového řízení do produkčního prostředí

Jestliže testy na testovacím prostředí dopadnou v pořádku, posouvá se změnové řízení na produkční prostředí. Na produkčním prostředí rovněž probíhá kontrola správnosti plnění.

Postimplementační testy na produkčním prostředí slouží pro vyhodnocení nasazených změn v CDS na „ostrých“ produkčních datech. Ovšem vracet nasazené změny z produkčního na testovací prostředí je velmi náročný proces a děje se tak pouze ve výjimečných případech. Proto pokud je třeba provést dodatečné změny, musí datový steward otevřít nové změnové řízení.

Poté je už jen poslední krok a to uzavření změnového řízení koordinátor změnového řízení.

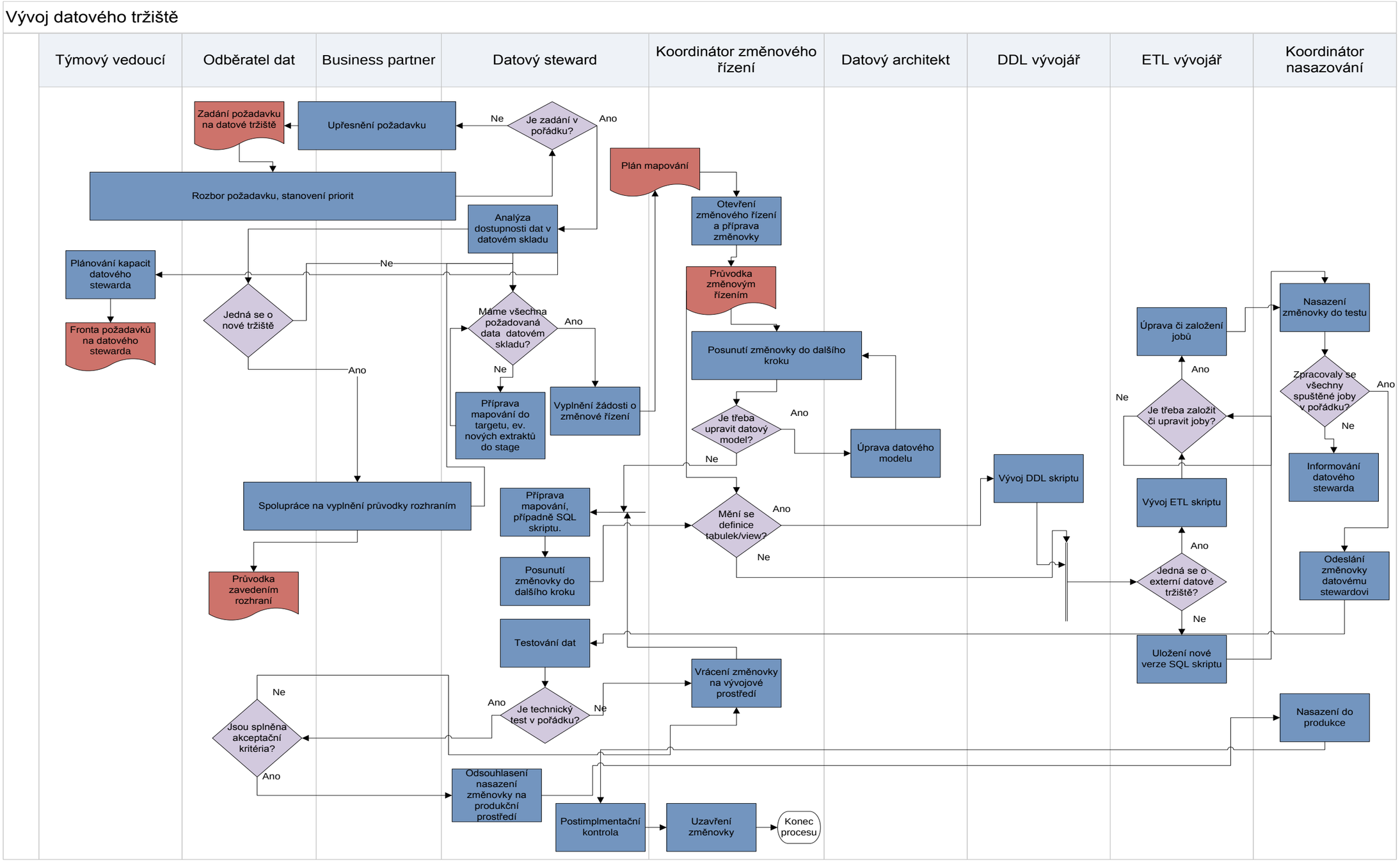
Výše popsaný proces od přijetí požadavku až po ukončení změnového řízení jsem pro lepší názornost zaznamenala diagramem, který je znázorněn na následujícím obrázku č. 19, který je pro svou velikost umístěn na samostatné stránce.

Co je definováno:

- dokumentace k aplikaci pro ukládání specifikací a mapování
- samotná průvodka změnovým řízením je velmi intuitivní a je z ní zřejmé, co má který účastník změnového řízení udělat.

Co není definováno:

- metodika testování datového tržiště



Obrázek 19 - Současný proces vývoje datového tržiště - od přijetí požadavku ke změnovému řízení; Zdroj: autor

7.4.4. Nasazení a údržba datového tržiště

Tvorba a odsouhlasení SLA

Po nasazení nového tržiště do produkčního prostředí je s odběratelem podepisováno SLA. V SLA je definováno, jakým způsobem a jak často budou probíhat úpravy v datovém tržišti. Zároveň je zde uvedena kontaktní osoba za datový sklad – datový steward a kontaktní osoba za odběratele. Dále se zde řeší, kdy a za jakých podmínek se budou data dodávat a co dělat v případě havárie dat v datovém skladu. Jak jsem psala na počátku této kapitoly, úpravy v datovém tržišti se realizují buď v rámci SLA, nebo v rámci nového projektu. V obou případech je třeba projít celým popsaným cyklem vývoje datového tržiště.

Datové kontroly

Po nasazení datový steward nastavuje kontroly datové kvality, ovšem není to pravidlem. Na kontroly je vytvořena šablona a zároveň je vytvořena metodika, která popisuje práci s datovými kontrolami. Výstupy z datových kontrol se sledují v reportovacím nástroji. Přičemž princip datových kontrol spočívá v tom, že se sledují hodnoty atributů v aktuálním měsíci a porovnávají se s hodnotami v předchozích měsících. Například se kontroluje výskyt prázdných hodnot, objemy číselných hodnot, četnost výskytu záznamů, atd. Pokud jsou hodnoty výrazně vyšší či nižší může to být signál, že je něco v nepořádku. Posouzení co je a co není v normě, pak závisí na individuálním posouzení datového stewarda, který má znalost dat. Samozřejmě, pokud něco není dle kontrol v pořádku, konzultuje datový steward výsledky kontrol s odběratelem dat.

Dokumentace

Zároveň je vhodné zdokumentovat tržiště. Pro tyto účely byla vytvořena šablona v Bikipedii a kolegyně vytvořila seznam kontrolních otázek, které by měl za své tržiště datový steward doplnit. Kromě toho ještě existuje zvláštní dokument obsahující kontrolní seznam datových tržišť, datových stewardů a kontaktních osob na straně odběratele.

Údržba

Záleží na domluvě datového stewarda s odběratelem a samozřejmě na SLA. Zpravidla je realizována prostřednictvím vedení seznamu požadavků, tzv. „wish listu“. Tyto požadavky jsou pak zadávány formou žádosti o službu MIS. Úpravy v tržišti mohou být vyvolány nejen odběratelem, ale i z důvodu změn probíhajících v CDS. Při realizaci požadavků je pak třeba projít již zmíněným změnovým řízením.

Co je definováno:

- šablona SLA
- datové kontroly
- co by mělo být k datovému tržišti dokumentováno
- způsob evidence požadavku odběratelů

Co není definováno:

- dokumentace k tržišti je stále ve vývoji, a ne každý z kolegů, který má ve správě tržiště, o této šabloně k dokumentaci ví

7.4.5. Vyřazení tržiště z provozu

Pokud jde o vyřazení z provozu, existuje metodika jak toto řešit. Tato metodika je více zaměřena na ETL vývojáře, ale i datový steward tam může nalézt informace, jak postupovat při odstraňování implementovaných požadavků v CDS. Nicméně není definováno, kdy má být tržiště odstraněno.

Co je definováno:

- metodika odstraňovaní implementovaných požadavků v CDS

Co není definováno:

- proces rozpoznání, kdy vyřadit tržiště z provozu

7.5. Využití reportovací vrstvy a virtuálních tržišť – sandboxů

Reportovací vrstva

V organizaci existují tabulky nad targetem, které se dají považovat za obdobu sémantické vrstvy. Plní se pouze za vybrané zdrojové systémy a doba historizace je snímek na konci roku, dva měsíční snímky, a denní snímky za poslední dva týdny. Tato vrstva je určena především pro účely vývoje statických jednorázových a pravidelných sestav.

Obecně je její využití pro datová tržiště definováno jako problematické. A to z toho důvodu, že při plnění tržiště se musí počkat, až se tabulky v reportovací vrstvě naplní všemi systémy. Což u tabulek v targetu nehrozí. Pokud tedy některý zdrojový systém dodá data později a přitom datové tržiště data z tohoto systému nevyužívá, dochází tak ke zbytečnému zpoždění při plnění datového tržiště. Dalším omezením reportovací vrstvy je délka historizace dat, která nemusí být pro odběratele akceptovatelná. Zároveň není možné, aby došlo k rozšíření reportovací vrstvy okamžitě.

Protože tato vrstva je využívána i při reportování a v jiných tržištích, je třeba zjistit dopady a získat součinnost od ostatních odběratelů. To proto, aby si otestovali, že změna nemá na „jejich“ data negativní dopad. A naopak, odběratel musí počítat s tím, že se bude muset podílet na rozvoji reportovací vrstvy. Byť i třeba jen pasivně, formou prohlášení, že souhlasí s úpravou reportovací vrstvy. Obecně se její použití pro datová tržiště nedoporučuje.

Abych zde neuváděla jen nevýhody, tak reportovací vrstva má samozřejmě i své výhody. Je zde větší kontrola nad plněním, a pokud dojde k rozsáhlejší změně, stačí změnu provést na jednom místě. Uvedu příklad z nedávné doby, kdy se rozšiřovala pobočková hierarchie a kde bylo třeba pracně vyhledávat, kam má změna dopad. Následně probíhala časově náročná úprava v řádu desítek sestav a změna se nevyhnula ani datovým tržištím. Další výhodou je rychlejší vývoj, protože data jsou zde uložena po subjektech, tedy v jedné tabulce najdeme údaje o klientech, v další o produktech, kontraktech atd. Tím odpadá časově náročnější propojování jednotlivých tabulek z datového skladu.

Kromě reportovací vrstvy se v organizaci využívá i sandboxů.

Sandboxy

Sandboxy se používají jako jeden z nástrojů při vývoji i úpravě stávajících datových tržišť. Cílem je, aby si odběratel vyzkoušel, co a zda vůbec potřebuje a vyžaduje od datového tržiště. Sandboxy jsou umístěné na stejném databázovém stroji jako datový sklad, ale mají svou vlastní databázi. Životnost sandboxu je omezena na dobu jednoho roku a za správnost dat vykazovaných ze sandboxu oddělení CDS neručí.

Dokumenty, související s reportovací vrstvou a sandboxy:

- pravidla využití reportovací vrstvy při plnění datových tržišť
- pravidla pro užívání sandboxů

7.6.Zhodnocení současného stavu

V organizaci jsou dobře popsány jednotlivé úkony, konkrétně je vypracovaná metodika mapování, psaní SQL skriptu, využití reportovací vrstvy a sandboxů, šablona SLA smlouvy a kontrola datové kvality. Samotné změnové řízení je také velmi návodné.

Přesto zde chybí ucelený popis a postup co dělat, v rámci celého životního cyklu datového tržiště. Neřeší, jakým způsobem analyzovat uživatelské požadavky, není ošetřen návrh datové modelu tržiště a jakým způsobem při návrhu vůbec postupovat ve spolupráci s odběrateli dat. A jak provést analýzu dostupných dat v datovém skladu. Neřeší se analýza, zda něco podobného už nebylo v minulosti realizováno. Zřejmě z toho důvodu, že je obtížné se „dohodnout s odběrateli, aby se spolu dohodnuli“. Jde zejména o dohodu o tom, z kterého rozpočtu se budou hradit změny v tržišti, a dohoda o změnách v tržišti vůbec.

Je částečně určeno, co by měla obsahovat dokumentace k datovému tržišti. Není specifikován postup testování. Způsob vyřazení tržiště z provozu je částečně ošetřen pomocí metodiky odstraňování implementovaných požadavků v CDS.

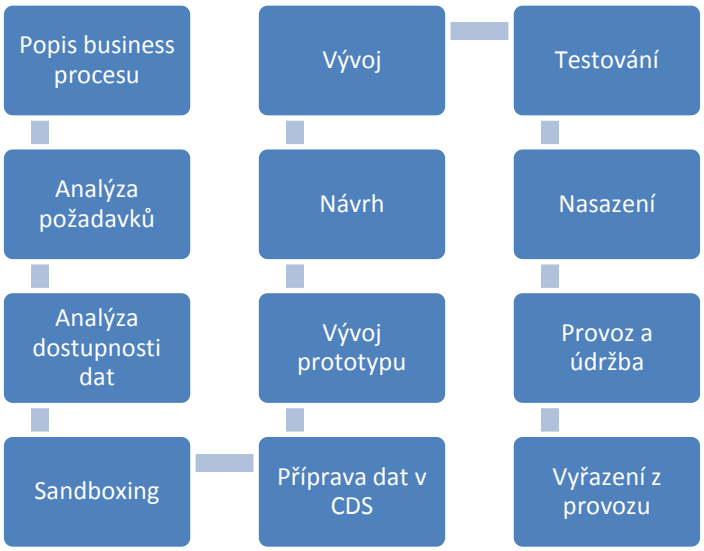
8. Návrh metodiky

Metodika popisuje celý životní cyklus datového tržiště. Obrázek č. 20 ukazuje návaznost jednotlivých kroků v metodice. Některé z kroků je možno při „pouhé“ úpravě tržiště vynechat, záleží na rozsáhlosti úpravy a rozsahu datového tržiště. Pokud by se daný krok vynechal, uvedu možná rizika, která plynou z vynechání.

Kroky na obrázku č. 20 jsou tedy následující:

1) **Popis business procesu** - V případě zcela nového tržiště se postupuje od identifikace a popisu řešeného business procesu, tedy problému, který odběratel potřebuje vyřešit.
2) **Analýza požadavků** – Cílem analýzy požadavků je získat alespoň obecnou představu o řešeném problému. Požadavky odběratelů dat jsou pak upřesňovány v dalších krocích.
3) **Analýza dostupnosti dat** - Na základě požadavků pak probíhá analýza, jaká data vlastně máme v datovém skladu. Tento krok se dá přirovnat k tzv. „prestudy“. Výstupy z analýzy jsou důležité pro následné plánování projektu.
4) **Sanboxing –** Slouží k tomu, abychom si ověřili, že data požadovaná odběratelem jsou skutečně ta data, která odběratel potřebuje.
5) **Příprava dat v CDS** – v této fázi probíhá příprava podkladových dat pro datové tržiště
6) **Vývoj prototypu** - Krok vývoje prototypu je klíčový pro detailní upřesnění požadavku odběratelů. Na základě schváleného prototypu probíhá návrh tržiště. K vývoji prototypu může být využit sandboxing.
7) **Návrh** – Tento krok představuje návrh datového modelu tržiště, specifikace tabulek či pohledů a mapování. V případě externího tržiště je třeba návrh přizpůsobit koncové aplikaci, jiné požadavky má OLAP kostka, jiné analytická aplikace a jiné dataminingový nástroj.
8) **Vývoj** - Ve vývoji probíhá psaní SQL skriptu či návrh ETL skriptu. Zároveň se zde nastavuje způsob doručení dat v případě externího datového tržiště.
9) **Test** - Po vývoji následuje nasazení na testovací prostředí. Kde se testuje funkcionalita a plnění datového tržiště. Jestliže je vše v pořádku, přistoupí se k nasazení tržiště na produkční prostředí. Zároveň se nastavují datové kontroly, tvoří dokumentace a probíhá návrh SLA.
10) **Nasazení** - Po akceptaci následuje nasazení datového tržiště na produkci, a případně školení uživatelů. A podpis SLA dokumentu.

11) **Provoz a údržba –** znamená pravidelné provádění změn v tržišti, a to jak z iniciativy odběratele, tak i nutnosti technické údržby (úprava datového modelu ve vrstvě integrovaného datového skladu, změna zdrojového extraktu…)
12) **Vyřazení z provozu –** znamená vypnutí jobů a smazání tržiště z databáze



Obrázek 20 - Procesy v metodice; Zdroj: autor

8.1.Analýza požadavků aneb definice rozsahu projektu

Fáze definování rozsahu projektu je ideální místo k získávání business požadavků. Výstupem této fáze by měl být soupis obecných požadavků odběratelů dat na datové tržiště.

Analýza požadavků je klíčový proces, který by měl předcházet každému projektu. Protože datové tržiště vytváříme přímo pro odběratele, a bude to právě odběratel, kdo bude akceptovat výstupy z projektu, je velmi důležité věnovat tomuto procesu pozornost. Cílem tohoto procesu tedy je zjistit, co uživatelé vlastně chtějí a očekávají. A k tomu se nabízí jako zdroj informací buď stávající řešení, pokud existuje, nebo samotní odběratelé.

Odběratelé jako zdroj požadavků

Ovšem jít se přímo ptát odběratelů s tím, co vlastně chtějí, není zase až tak dobrý nápad. Už proto, že odběratel zpravidla nemá úplně jasnou představu o tom, co chce. Navíc pokud je to jeho první zkušenost s datovým skladem, tak neví a ani nemůže vědět, jaké možnosti mu datový sklad může nabídnout. Důsledkem pak je, že se vyvine datové tržiště a odběratel zjistí, že to vlastně není to, co chtěl. To je možné i tehdy, pokud se požadavkům věnuje malá pozornost.

Proto je vhodné pro analýzu požadavků použít princip ze sedmi proudového přístupu, konkrétně proud „Analýza CIFu“. Kde se analýza požadavků bude probírat po jednotlivých tématech, např. vyhodnocení kampaně či hodnocení a motivace pracovníků a jiné. V tomto případě, pojem

téma odpovídá pojmu proces. Z těchto procesů pak vyplynou otázky, na které potřebuje odběratel znát odpověď. V případě kampaně třeba „Kolik a kteří klienti si v daném období sjednali daný produkt?“, v případě hodnocení „Kolik kontraktů a v jaké výši se uzavřel na jednotlivých pobočkách?“. Tedy zjišťujeme, jak vlastně funguje řešený business proces a jaká je pracovní rutina odběratele v tomto procesu.

A pak teprve je třeba si klást otázku, jak může tyto otázky datový sklad vyřešit? V úvahu je také třeba vzít, kdo se na tyto otázky ptá. Jiný požadavek na detail dat bude mít pobočkový manažer a jiný obchodní ředitel.

Stávající řešení jako zdroj požadavků

Kromě toho se ještě nabízí analýza stávajícího řešení. Je zřejmé, že nemá smysl „objevovat Ameriku“. Na druhou stranu je třeba zvážit, proč není možné stávající řešení použít. Odběratelé o něm nevědí, nebo jim nevyhovuje? Pokud nevyhovuje, pak nemá smysl nechat se ovlivňovat stávajícím a je lepší začít od začátku.

Výstupem části analýzy požadavků tedy bude odpověď na otázky:

- Jaký business proces bude datové tržiště řešit?
- Na jaké otázky potřebuje odběratel znát odpověď?
- Jak tyto otázky může zodpovědět datový sklad potažmo datové tržiště?
- Pro koho bude datové tržiště určeno? Kdo je klíčovým odběratelem?
- Je možné použít stávající řešení?

Zachycení architektury řešení

Pokud známe požadavky, můžeme přikročit k jednání o architektuře řešení. V první řadě je třeba s odběratelem probrat, zda se bude jednat o interní či externí datové tržiště. U interního tržiště hraje roli, zda budou uživatelé přistupovat přímo do databáze, nebo budou využívat reportovací nástroj.

V případě externího tržiště zda bude tržiště řešeno jedna k jedné k uživatelské aplikaci. Jestli se budou data exportovat do souboru nebo nahrávat přímo do aplikace. V případě, že do souboru, tak zda si bude odběratel odebírat data sám nebo mu budou data posílána do adresáře.

Toto je pak třeba zohlednit ve fázi návrhu při navrhování datového modelu tržiště. A v případě externího tržiště hrají tyto informace roli při vyplňování průvodky zavedením rozhraní. U nového interního tržiště se průvodka zavedením rozhraní také vyplňuje. Ovšem vzhledem k tomu, že se data nebudou nikam exportovat, v podstatě „jenom“ řešíme, kdo bude gestorem za business a za oddělení datového skladu.

V rámci návrhu architektury řešení je na datovém stewardovi volba, zda zvolí realizaci pomocí view nad targetem/reportovací vrstvou, nebo zda bude tabulky materializovat. Obecně je zažitá praxe vytvářet fyzické tabulky. View se nevyužívají z výkonnostních důvodů (při každém otevření view se opětovně spouští SQL dotaz), což u často využívaných tržišť, u tržišť se složitými transformacemi a u nebo u tržišť s velkým objemem dat způsobuje zátěž databázového stroje.

Co se stane, pokud se tento krok přeskočí?

- Je zřejmé, že tento krok je naprosto kritický a nemá smysl ho vynechávat. V opačném případě se zdroje na vývoj vynaloží na něco, co s největší pravděpodobnosti bude nevyužito. Pokud vůbec bude možné začít s vývojem.

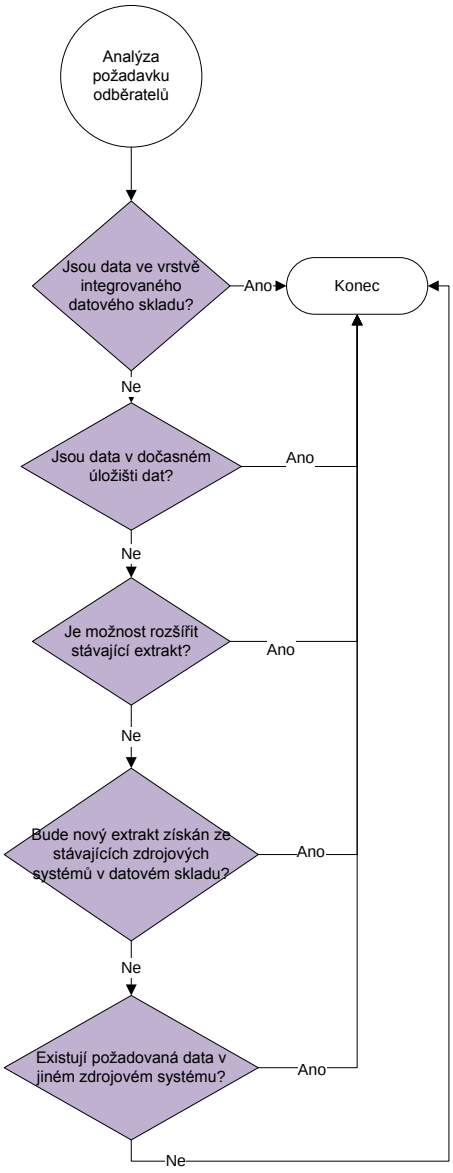
8.2.Analýza dostupnosti dat

Analýza dostupnosti dat by měla předcházet každému většímu projektu. V rámci této analýzy se analyzuje, jaká data, která jsou požadována v rozsahu projektu, jsou v datovém skladu k dispozici. Tato analýza je důležitá z hlediska plánování projektu. A to proto, že je obtížné, ne-li nemožné, naplánovat projekt, když nevíme, jaká data máme k dispozici.

Jde o to, že dobu trvání projektu může významně prodloužit mapování nových dat do CDS. Navíc pokud tato data nejsou ani v stage, může proces přípravy dat významně prodloužit dobu trvání projektu. Proto při požadavku na nové tržiště či jeho úpravu musí předcházet analýza dostupnosti dat.

Z této analýzy by mělo být zřejmé, jaká data jsou v targetu k dispozici, jaká jsou v dočasném úložišti dat a případně odkud potřebná data získat. To platí v tom lepším případě. V horším se také může zjistit, že požadovaná data ve zdrojovém systému vůbec nejsou. Zjištění zdroje nových dat je důležité z toho důvodu, že rozšíření existující specifikace o nové pole je méně pracné. Oproti tomu, když je třeba vytvořit zcela novou specifikaci, nebo rovnou připojit zcela nový zdrojový systém je pracnost daleko vyšší.

Pro lepší přehlednost jsem sestavila posloupnost otázek, viz obrázek č. 21, jejichž zodpovězení spolu se seznamem atributů a zdrojů odkud tyto atributy získáme, výstupem této analýzy.



Obrázek 21 - Posloupnost otázek v analýze dostupnosti dat; Zdroj: autor

Co se stane, pokud se tento krok přeskočí?

- Tento krok je rovněž jedním z těch kritických. Pokud se tato analýza neprovede, je zde velké riziko, že budou provedeny nesprávné odhady pracnosti. A v případě většího zpoždění projektu může odběratel prohlásit, že data už nepotřebuje. Případně se může dojít k závěru, že požadovaná data nejsou, čímž jsme skončili.

8.3.Sandboxing

Předtím než se na základě analýzy dostupnosti dat přistoupí k rozšiřování či vývoji nových extraktů do dočasného úložiště dat, či k mapování do targetu, je třeba nechat odběratele si tato data „osahat“. Aby se ujistili, že právě ta data, která požadují, jsou opravdu ta data, která potřebují pro

svou práci (v angličtině existuje výstižný termín „prove the value of data“). A právě k tomu jsou určené sandboxy. V podstatě jde o „pískoviště“, kde si odběratel může „hrát“ s experimentálními daty získanými z CDS i z jiných zdrojů.

Využití sandboxů má tři podmínky – data se neobnovují, doba trvání sandboxu je časově ohraničená a data získaná tímto způsobem lze vykazovat jen omezeně. A vždy musí být uvedeno, že data pocházejí z tohoto zdroje, a oddělení datového skladu tak nemůže ručit za jejich správnost.

Je zřejmé, že sandboxy mají své opodstatnění ve chvíli, kdy máme nejednoznačné zadání, anebo v případě, kdy je zadání velmi složité. V případě, kdy se jedná o jednoznačné zadání, není třeba sandoboxu využít.

Co se stane, pokud se tento krok přeskočí?

- V případě, kdy nemá odběratel úplně jasno v tom, co potřebuje, je zde jistá pravděpodobnost, že po pracném zavádění nových zdrojových systémů a mapování do vrstvy integrovaného datového skladu se zjistí, že odběratel nedostal taková data, jaká původně očekával. Takže vynaložená námaha na přípravu podkladových byla vcelku zbytečná.

8.4.Příprava dat do vrstvy integrovaného datového skladu

V případě, že víme, kde jsou uložená zdrojová data, a víme, že tato data jsou požadována odběrateli, může se v závislosti na výsledku analýzy dostupnosti dat začít s přípravou podkladových dat. Postup v rámci tohoto kroku tedy bude následující:

1) Vytvoření nového zdrojového systému pomocí „Průvodky zavedením rozhraní“ – pokud je třeba
2) Příprava datových struktur v dočasném úložišti dat v aplikaci DSA – pokud je třeba
3) Mapování do targetu – pomůže existující metodika mapování do targetu
4) Vyplnění plánu mapování – na Bikipedii je umístěn návod k vyplnění
5) Projít procesem změnového řízení na základě průvodky změnovým řízením, která byla vytvořena na základě údajů v plánu mapování

Co se stane, pokud se tento krok přeskočí?

- Tento krok je za předpokladu, že máme všechna data v targetu dostupná bezpředmětný. V opačném případě nebude na čem vyvíjet, vzhledem k tomu, že ctíme Inmonovu podmínku, že data do tržiště můžeme brát pouze z targetu.

V kontextu organizace je možné ještě uvažovat o reportovací vrstvě. Ale i ta bere data z targetu, takže to nic neřeší.

8.5. Vývoj prototypu

Pokud máme ověřeno, že požadovaná data jsou opravdu ta, která odběratel chce, může se přikročit k přípravě prototypu. K tomu je možné využít i zmíněných sandboxů, především v případě interních tržišť. U externích tržišť je výsledná podoba definována cílovou aplikací.

Prototyp slouží mimo jiné k ověření, že zamýšlené řešení bude vůbec fungovat. Toto je důležité především u externích datových tržišť, která slouží k nahrávání dat do uživatelských aplikací. U interních tržišť si zase ověříme, že uživatelům vyhovuje podoba tržiště a zvolená úroveň detailu.

Principem je poskytnutí dat odběratelům v takové podobě, jaká bude použita při finálním řešení. Tento krok je možné opakovat vícekrát až do té doby, kdy dojde k akceptaci navrženého řešení odběrateli.

Co se stane, pokud se tento krok přeskočí?

- Zde záleží na rozsáhlosti změny a technickém řešení tržiště. V případě nového tržiště, a pokud je to technicky možné, tento krok doporučuji provést. Jinak se může stát, že se při uživatelských testech zjistí, že navržená podoba tržiště nevyhovuje. Což znamená předělání návrhu a opětovnou práci ETL a případně i DDL vývojáře.

8.6. Návrh datového tržiště

1) Návrh datového modelu

Návrhu datového modelu tržiště organizace se dle mého názoru nevěnuje přílišná pozornost. Pravděpodobně proto, že není dostatek licencí v modelovacím nástroji, je to pracné anebo na to není jen čas. Připouštím, že ve chvíli, kdy máme tržiště o třech tabulkách to zase, až tak palčivé není, ale ve chvíli kdy máme tržiště o rozsahu dvaceti tabulek, to už může být problém.

Důvodem proč si myslím, že datový model má svůj význam, je několik. Za prvé máme přehled o celé struktuře tržiště, a pokud je model dobře navržený, vyčteme z něj i problematiku řešeného procesu. Tohle dle mého názoru jen z mapování nejde vyčíst, prostě proto, že mapování je příliš detailní. A pokud se máme bavit s odběratelem nad následným rozvojem tržiště a přidáváním nových tabulek, tak potřebujeme mít náhled na celou strukturu a proces, který tržiště řeší. Abychom dokázali vhodně začlenit nové tabulky či atributy. Další důvodem je, že datový model představuje dokumentaci k tržišti. A pokud si má tržiště přebírat někdo, kdo o tomto tržišti nic neví, je dle mého názoru snazší pochopit principy z datového modelu, kdy vidíme celek a můžeme si probrat význam

jednotlivých tabulek, proč jsou spolu takto propojené a co vlastně která tabulka řeší. Než když jsou k dispozici definice datových struktur, mapování z targetu do tržiště v DSA a případně znázorněné vazby pro plnění.

Navíc v datovém modelu na první pohled vidíme referenční integritu. Takže ve chvíli, kdy provádíme zásahy do tržiště, si můžeme na první pohled ověřit, jak nám musí sedět počty záznamů v jednotlivých tabulkách.

Cílem této podkapitoly není dát návod, jak přistoupit k datovému modelování tržišť, ale obhájit a vyzdvihnout význam datového modelu, aby se v organizaci vůbec začalo o modelování uvažovat. Kromě toho, že to přesahuje rámec této práce, je každé tržiště specifické a záleží na účelu, za kterým bude využito. Kromě toho je třeba brát v úvahu i technologii, na které bude tržiště zpracováno. Na druhou stranu je třeba říct, že bez ohledu na použitou technologii si data v tržišti musí zachovat konzistenci a význam. Takže pokud mluvím o modelu, myslím tím konceptuální a logický model, kde jsou popsány entity a vztahy mezi nimi a následně popis a význam atributů v jednotlivých entitách. Způsob implementace modelu pak už záleží na konkrétním využití.

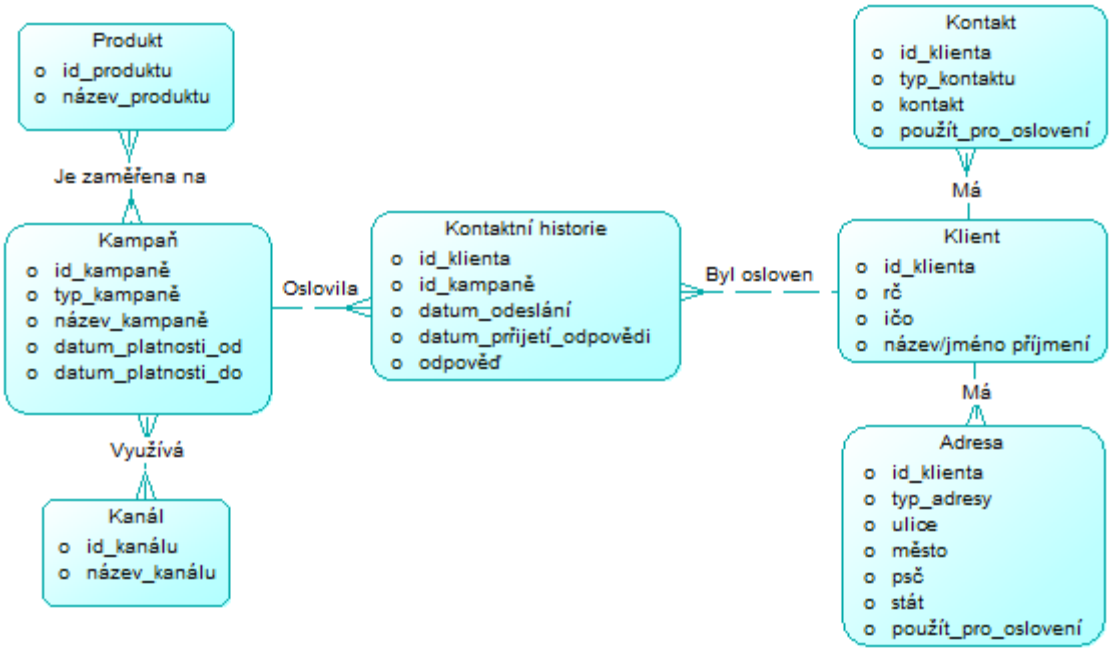
Pokud to shrnu, tak datový model tržiště by měl principielně splňovat tyto podmínky:

- model musí zohledňovat business požadavky
- v modelu musí být zachycena řešená problematika
- z modelu musí být možné vyčíst business logiku řešení

Na základě návrhu datového modelu pak probíhá specifikace tabulek/view, mapování a případně vývoj SQL skriptu.

Na obrázku č. 22 je znázorněn příklad konceptuálního modelu datového tržiště pro účely vyhodnocování kampaní. Kde je požadavek na to, aby bylo možno sledovat, v rámci kterých kampaní byly oslovení klienti organizace, prostřednictvím jakých kanálů byla kampaň distribuována a na jaké produkty byla zaměřena. Zároveň je požadavek na to, aby byla možnost kontaktování klienta.

Z příkladu je zřejmé, že globální pohled prostřednictvím datového modelu dá lepší představu o business procesu, které tržiště řeší, než když je k dispozici pouze mapování, které je příliš detailní na to, aby z něj bylo možné vyčíst business proces.



Obrázek 22 - Příklad konceptuálního modelu datového tržiště; Zdroj: autor

2) Návrh datových struktur datového tržiště

Znamená specifikaci datových struktur tržiště v DSA. Zde je třeba mít rozmyšlenou frekvenci plnění (denní, týdenní, měsíční…), datové typy a business popisky atributů. V případě externího tržiště by mělo být už z fáze analýzy požadavků známo formátování souboru a způsob předávání dat s odběratelem. Zároveň je třeba si s odběratelem odsouhlasit ještě jednou specifikaci datových struktur.

3) Mapování

Představuje mapování z vrstvy targetu, případně z reportovací vrstvy do datového tržiště. Mapování provádí datový steward dle metodiky mapování do datových tržišť. Eviduje se v aplikaci pro ukládání specifikací a mapování.

4) Vývoj SQL skriptu

Vývoj plnícího skriptu tržiště na základě metodiky psaní SQL. Toto platí jen v případě, že se tržiště nerealizuje pomocí ETL skriptu. Skript je třeba odladit na produkčním i testovacím prostředí.

5) Sestavený havarijní scénář

Zároveň je třeba si naplánovat, jaký postup bude v případě, že dojde k výpadku dat ze zdrojového systému. Havarijní scénář se sestavuje ve spolupráci s odběratelem. Kdy si odběratel určí, jaké postupy v rámci havarijního scénáře jsou pro něj akceptovatelné.

Co se stane, pokud se tento krok přeskočí?

- V případě, že se nevytvoří datový model, je ztížen proces předávání agendy. Rovněž při změnách v tržišti je riziko, že provedené změny způsobí nekonzistenci v datech. Kromě toho, v případě plánování kam přidat novou tabulku či nový atribut do tržiště, se stráví podstatně více času, protože je třeba projít mapování a složitě analyzovat kam změnu přidat. A to proto, že nám chybí celkový náhled na řešenou problematiku a návaznost mezi entitami.
- Je zřejmé, že návrh datových struktur a mapování nemá smysl vynechávat.
- Vývoj SQL skriptu má význam i tehdy, kdy se tržiště plní ETL procesem. A to pro usnadnění testování, kdy si můžeme snadno ověřit správnost ETL. Když se porovnají výsledky získané ETL skriptem oproti výsledkům „našeho“ SQL skriptu.
- Vynecháním havarijního scénáře následuje docíleno toho, že není definován postup, co dělat když CDS není schopen dodat data odběrateli. Což má za následek nelibost odběratele, v tom lepším případě, a v tom horším i pokuty organizaci, pokud se data vykazují třetí straně.

8.7. Vývoj

V případě, že je připraven návrh datového tržiště, nastává samotný vývoj. Ten probíhá v následujících krocích:

1) Vytvoření nového systému pro tržiště pomocí průvodky zavedením rozhraní – jen v případě nového tržiště
2) Zadání požadavku na změnové řízení do plánu mapování – na Bikipedii je umístěn návod k vyplnění
3) Doplnění průvodky změnovým řízením, což v této fázi znamená vytvoření podkladů pro ETL vývojáře a DDL exekutora – ty by měly být předpřipravené z kroku návrhu. V tomto případě, je jen třeba přiřadit číslo změnového řízení specifikacím a mapování v aplikaci pro ukládání specifikací a mapování.
4) Vývoj DDL a případně ETL skriptu DDL exekutorem a ETL vývojářem na vývojovém prostředí. Zároveň ETL vývojář nastavuje plnící úlohy – joby.

Co se stane, pokud se tento krok přeskočí?

- Opět je zřejmé, že nemá smysl tento krok přeskakovat. V opačném případě byly veškeré vynaložené kapacity vynaloženy zbytečně.

8.8. Test

Ve fázi, kdy je změna v tržišti či nové tržiště vytvořeno ve vývojovém prostředí, nastává nasazení do testovacího prostředí. To slouží jednak k ověření funkčnosti změnového řízení, tak i k poskytnutí dat odběratelům. Rozlišujeme tedy dva druhy testů a to uživatelské a technické.

Na začátek ještě doporučení. V případě nového tržiště, které se plní ETL skriptem, je vhodné si vytvořit vlastní SQL skript, který vychází z mapování, jak už jsem zmiňovala ve fázi návrhu tržiště. A navíc doporučuji, před samotným začátkem testování si připravit přibližné odhady objemů a počtu záznamů.

Technický test

Na začátku probíhá základní kontrola, zda se vůbec naplnily tabulky, případně zda pohledy dávají nějaké výsledky. Pokud se některé nenaplnily, přichází kontrola, zda máme na testu všechna potřebná data. Další možností je, nevhodně nastavené business datum, tzn., nemáme na testu data k danému business datu. Nebo jsou nevhodně nastavené závislosti, a tabulky se naplnily dřív, než byly naplněné tabulky v targetu. Anebo v případě tržiště plněného ETL nástrojem došlo k chybě při vývoji ETL.

V případě externího tržiště probíhá kontrola formátování dat v souboru. Dále pokud není nastaveno generování logů, je vhodné si zkontrolovat počet záznamů v tabulce a v souboru. Pokud jde o formátování dat, je vhodné provést kontrolu, zda se neořízly textové řetězce případně číselné hodnoty, to platí pro externí i interní tržiště.

Pokud jde o samotný obsah dat, je možnost testovat buď celý výběr záznamů, nebo jen vybranou množinu a v případě agregátu celkové objemy a počty záznamů podle kategorií. Pokud je to možné (v tržišti jsou data na stejné úrovni detailu jako v integrované vrstvě, v tržišti nejsou příliš složité transformace), je možné porovnat celý vzorek dat a to za pomocí příkazu EXCEPT. To je možné například u tabulek, které jsou kopií tabulek uložených v dočasném úložišti dat. Nejčastějším takovým případem jsou číselníky.

V opačném případě se vybere pouze malá množina dat, která se bude podrobněji testovat. Je dobré, aby vybrané testované záznamy obsahovaly nejen naplněné hodnoty, ale i ty prázdné. Ještě před testováním vybrané množiny, jako základní test se provede porovnání počtu záznamů. V testované množině se sleduje, jestli záznamy splňují business požadavky odběratele. Například při segmentaci probíhá kontrola, že se klienti přiřadili do segmentů podle pravidel.

V případě agregátu je možné si porovnat objemy oproti integrované vrstvě. A v případě, že jsou v tržišti sloupce, které obsahují jen několik málo hodnot (třeba ta segmentace), spočítáme si počet záznamů v jednotlivých kategoriích.

V případě, že jde pouze o změnu v tržišti, je vhodné si porovnat počet záznamů před a po změně. Pokud jsou v tržišti tabulky, které jsou na sobě závislé, je vhodné zkontrolovat si referenční integritu. Ke kontrole referenční integrity pomůže datový model tržiště.

Pokud výsledky nesedí, jsou tyto možné příčiny:

Příčina	**Řešení**
Nemáme naplněné všechny tabulky na testu	Zkontrolovat si naplněnost tabulek v integrované vrstvě za zdrojové systémy, které se plní do tržiště. A rovněž si zkontrolovat databázi, ve které testujeme.
Nedoběhly plnící joby	Zkontrolovat si doběhnutí jobů v aplikaci pro správu jobů.
Vycházíme z jiného business data	Zkontrolovat si, k jakému business datu se tržiště plnilo a jaké je použito v testovacím skriptu.
Nevhodně zvolené datové typy	Zkontrolovat si, zda se neořezávají textové řetězce, případně objemy.
Množící se záznamy	Zkontrolovat si zda máme všude omezení na business datum. A zda jsme nezapomněli někde nějakou podmínku.
Nevhodně zvolené podmínky	Zkontrolovat si podmínky

Tabulka 1 - Příčiny chybných výsledků při testování a jejich řešení; Zdroj: autor

Pokud žádná z výše uvedených příčin nenastala, zřejmě jsme odhalili chybu. V tom případě, jsou zřejmě chybně nastavené závislosti, nebo v případě, že se tržiště plní ETL nástrojem je chyba v ETL. V tom případě se obrátíme na ETL vývojáře a změnové řízení vrátíme v průvodce zpět do vývojového prostředí.

Uživatelský test

Pokud dopadly technické testy v pořádku, je možné postoupit data uživatelům. Účelem je akceptace podoby a obsahu datového tržiště. V případě externích tržišť si tím zároveň ošetříme, že funguje komunikační kanál mezi datovým skladem a uživatelskou aplikací. A že uživatelská aplikace je schopná se zaslanými daty pracovat.

Co se stane, pokud se tento krok přeskočí?

- Riskujeme tím, že vývoj neproběhl tak jak bylo zamýšleno. A po nasazení na produkční prostředí s tím už nic neuděláme. Také se vystavujeme riziku, že odběratel neakceptuje výstupy z tržiště, pokud je vůbec bude schopen zpracovat (v případě externího tržiště).

8.9.Nasazení

Nasazením do produkčního prostředí se ukončuje změnové řízení. Vracení změnového řízení z produkčního na testovací potažmo vývojové prostředí se děje jen výjimečně. Pro další změny je třeba si otevřít nové změnové řízení. Proto testování na produkčním je určeno spíše pro ujištění se, že se tržiště plní tak, jak bylo předpokládáno a ověřeno na testovacím prostředí.

8.9.1. Dohoda o úrovni poskytovaných služeb – „SLA“

Potom, co se tržiště nasadí do produkčního prostředí, se podepisuje SLA. V rámci SLA jsou definovány odpovědné osoby za stranu odběratele a za stranu oddělení datového skladu. Je zde definována frekvence poskytování dat, v jakém formátu budou data dodávána, případná historizace, jak bude probíhat pravidelná údržba, krizový scénář a proces sledování kvality dat. V organizaci je vytvořená šablona k SLA. V rámci podpisu SLA se dohodnout na způsobu evidence požadavků na změnu v datovém tržišti, čili vytvoření tzv. „issue listu“. K „issue listu“ je rovněž vytvořena šablona, jejíž odkaz je umístěn na Bikipedii.

8.9.2. Nastavení datových kontrol

Zároveň se nastavuje kontrola datové kvality, která byla blíže popsána v kapitole věnující se analýze současného stavu. Pro datové kontroly je rovněž vytvořená šablona a na Bikipedii je popsán přesný postup jak s touto šablonou pracovat, a jakým způsobem zajistit nasazení datových kontrol do prostředí datového skladu.

8.9.3. Dokumentace datového tržiště

V rámci nasazení zbývá ještě jeden krok, a tím je dokumentace k tržišti. Na Bikipedii je umístěna šablona, která obsahuje název tržiště, popis tabulek, seznam kontaktních osob, specifické transformace pro tržiště, za které zdrojové systémy se tržiště plní a jiné. Kromě toho ještě kolegyně vytvořila okruhy informací, obsahující např. technické informace o tržišti, jako počet tabulek, objem dat v tabulkách, délka historie, dále odkazy na projektovou dokumentaci, aktuální status a fáze projektu a jiné.

Co se stane, pokud se tento krok přeskočí?

- Podepsané SLA s odběratelem je klíčové. Jde o smlouvu mezi poskytovatelem dat – oddělením CDS a odběratelem dat o úrovni poskytovaných služeb. Tím, že s odběratelem podepíšeme smlouvu SLA, máme jistotu, že odběratel akceptoval a je srozuměn s podmínkami poskytování dat. V opačném případě můžeme získat nespokojeného odběratele, což není účelem.

- Kontrola datové kvality slouží k tomu, abychom si ověřili, že odběrateli budeme vykazovat správná data. Vykazováním nesprávných dat můžeme poškodit nejen odběratele, ale i samotnou organizaci.
- Chybějící dokumentace zapříčiní, že se po čase pravděpodobně zapomene, co se vlastně v tržišti řeší, což znesnadní následnou údržbu a schopnost odpovídat na dotazy odběratelů k tržišti. Nemluvě o ztíženém předávání agendy k tržišti jinému pracovníkovi.

8.10. Provoz a údržba

V rámci provozu se provádí pravidelná údržba datového tržiště. Četnost změn závisí na SLA. Úpravy v tržišti také znamenají aktualizaci dokumentace. Kromě toho se může vyskytnout požadavek na provedení dopadové analýzy. To platí v případě, kdy se upravují data ve zdrojovém systému, případně se rozšiřují tabulky v targetu anebo se zasahuje do stávajícího mapování v targetu.

V rámci pravidelného provozu rovněž sledujeme již zmíněnou kvalitu dat.

Co se stane, pokud se tento krok přeskočí?

- V případě, že se nebude brát ohled na úpravu tržiště dle zadání odběratele, hrozí, že odběratel přestane tržiště využívat. Což znamená, že vynaložené kapacity na vývoj tržiště byly zbytečné. Tržiště, které nikdo nevyužívá, nemá smysl v databázi udržovat. Obdobně pokud se nebudou provádět dopadové analýzy, případně úpravy v tržišti, které byly zapříčiněné změnami v datovém skladu a zdrojovém systému, začnou se vykazovat nesprávná data. To samé platí i pro datové kontroly, kde pokud nebudou sledovány, tak může dojít k vykazování nesprávných dat.

8.11. Vyřazení z provozu

Zde je otázka datové tržiště z provozu vyřadit. V případě, kdy sami odběratelé podají podnět, že tržiště jim už netřeba. Dále v případě, že vyvíjíme zcela novou verzi tržiště. Nebo když zjistíme, že data z tržiště nebyla využívána déle než určité časové období a vypnutí nám odsouhlasí odběratel dat. Ještě před samotným vypnutím je vhodné probrat s odběratelem, proč stávající řešení nepoužívá. A zároveň odběratele informovat o možnosti upravit tržiště dle aktuálních potřeb odběratele.

Odstavení z provozu znamená, že je třeba provést dopadovou analýzu. Následuje vypnutí plnících jobů, a poté vymazání datového tržiště a ETL procesů z databáze.

Co se stane, pokud se tento krok přeskočí?

- Platí zde to samé, co je uvedeno u kroku provozu a údržby. Pokud máme signál, že je tržiště nevyužívané, nemá smysl zatěžovat databázový stroj s jeho plněním.

8.12. Shrnutí – *„checklist“*

Zde je uveden seznam kroků, které má datový steward provést, jestliže chce mít jistotu, že tržiště bude odběrateli akceptováno. A zároveň bude tržiště navrženo tak, aby byl pokud možno zajištěn jeho bezproblémový provoz:

- Pomoci odběrateli vyplnit žádost o službu MIS. Případně poskytnout součinnost při definování požadavků. Při schůzkách nad požadavky nezapomenout na obstarání/pořízení zápisu ze schůzky, který bude odsouhlasený všemi účastníky.
- Provést analýzu dostupnosti dat v CDS na základě požadavku odběratele
- V případě požadavku na úpravu v existujícím tržišti provést dopadovou analýzu na stávající řešení
- Připravit pro uživatele data z CDS či zdrojových systémů -> sandboxing. Toto platí zejména v případě, kdy je složité či nejednoznačné zadání a odběratel si potřebuje data „vyzkoušet“.
- Připravit podkladová data pro tržiště, což znamená:
    - založit nový zdrojový systém nebo jen rozšířit stávající zdrojové extrakty – pomocí průvodky zavedením rozhraní
    - provést mapování z stage do targetu – k tomu využít metodiky mapování do targetu
    - v případě nejasnosti k mapování či potřebě rozšířit datový model targetu se obrátit na architekta datového skladu
- Před samotným vývojem prototypu a návrhem tržiště se ujistit, že už neexistuje podobné řešení
- Vytvořit pro odběratele dat prototyp datového tržiště. Toto opět platí hlavně pro komplikované či nejednoznačné zadání. Kde je předpoklad, že odběratel bude svůj požadavek měnit podle toho, co uvidí v datech.
- Provést návrh tržiště:
    - vytvořit datový model tržiště
    - vyplnit průvodku zavedením rozhraní ve spolupráci s odběratelem – v případě nového tržiště
    - vyplnit plán mapování

    - vytvořit/upravit definice tabulek v aplikaci DSA
    - vytvořit/upravit mapování v aplikaci DSA
    - nechat si odsouhlasit strukturu tabulek/view odběratelem – toto se týká především externích datových tržišť. Kde je třeba si odsouhlasit, v jakém formátu budou data poskytována.
- Projít změnovým řízením
    - přiřadit číslo změnového řízení specifikacím a mapováním v aplikaci DSA
    - předat změnové řízení k vývoji ETL a DDL vývojáři
    - vyhodnotit změnové řízení na testovacím prostředí – uživatelské i technické testy. V případě akceptace odběratelem a za předpokladu, že technické testy dopadly v pořádku, provést implementační kontrolu. Jinak se vrátit zpět k návrhu tržiště a vrátit změnové řízení na vývoj.
    - provést postimplementační kontrolu a tím uzavřít změnové řízení
- Vytvořit datové kontroly – na Bikipedii je umístěný návod spolu s odkazy na příslušné dokumenty
- Dokumentovat tržiště na Bikipedii a zároveň vyplnit šablonu dokumentace v Excelu. Na tuto šablonu bude umístěn odkaz v Bikipedii.
- Navrhnout a podepsat SLA s odběrateli.
- Dohodnout se na evidenci požadavků tzv. „issue list“

9. Zhodnocení využití sémantické vrstvy v prostředí organizace

Jak už jsem zmiňovala v analýze současného stavu, v organizaci je obdoba sémantické vrstvy zvaná reportovací vrstva. Tato vrstva není z důvodů omezení, které klade na odběratele příliš využívána. Navíc, jak jsem uváděla v kapitole věnující se přímo tomuto tématu, její existence předpokládá i existenci business slovníku. A to je hlavní důvod, který brání jejímu využití, protože v organizaci nic takového jako business slovník není. V současnosti probíhá prestudy na výběr nástroje pro ukládání metadat a poté se do budoucna „možná“ počítá s využitím business slovníku.

Pokud by měla být reportovací vrstva hojněji využívaná pro plnění datových tržišť, musely by se v ní provést zásadní změny. Provést tyto změny by bylo velmi náročné, vzhledem k tomu, že existují stovky statických reportů, které čerpají data z této vrstvy. Z toho by vyplývaly i změny v těchto reportech, a na to bohužel nejsou kapacity. Navíc by tato změna měla ještě dalekosáhlejší důsledky, a to proto, že existují další stovky reportů, které vykazují obdobná data jako reporty v předchozím případě, ale čerpají data nikoliv z reportovací vrstvy, ale přímo z targetu. A lze si jen těžko představit, že by se podařilo udržet vykazování stejných hodnot. Je třeba říct, že už teď je problém tyto výstupy udržet. A tak se stává, že se potkají dva výstupy z různých oddělení (a je jedno jestli byly získané z tržiště nebo reportů), které by měly dle zadání obsahovat stejná data, a liší se.

Namísto toho, aby se řešilo, co výsledky v reportech znamenají, se řeší, který výstup je „správnější“. Právě této situaci by zabránil business slovník, kde by byly jednoznačně definované pojmy, uvedené vzorce jak počítat objemové hodnoty a tyto pojmy by byly známé napříč organizací. A zde je „kámen úrazu“, tvorba slovníku by znamenala, že se na ní musí dohodnout jednotlivé útvary. A na to, aby se tyto útvary dokázaly dohodnout, by bylo třeba, aby se touto problematikou začalo zabývat vedení organizace. Bez podpory vedení je malá pravděpodobnost, že by se oddělení datového skladu podařilo business slovník prosadit.

Za těchto podmínek je využití sémantické vrstvy velmi nereálné, a to i v blízké budoucnosti. Stejně tak rozšíření reportovací vrstvy, protože je pravděpodobné, že by se u odběratelů nepodařilo prosadit jednotný pohled na význam dat, o která by se vrstva rozšiřovala. Pokud se pomine náročnost provedení dopadové analýzy a následných změn v reportech.

10. Závěr

Hlavním cílem práce bylo analyzovat a dokumentovat současný přístup k vývoji a údržbě datového tržiště v organizaci. Zároveň v případě zjištění nedostatků v současném řešení navrhnout metodiku, která by zaštiťovala celý proces správy datových tržišť v organizaci.

V případě analýzy a dokumentace současného stavu, jsem vytvořila seznam rolí vystupujících v procesu, činnosti těchto rolí a seznam dokumentů souvisejících s procesem. Zároveň jsem popsala proces vývoje a údržby tržiště, kde jsem vycházela ze svých zkušeností datového stewarda v organizaci. Tyto své zkušenosti jsem pak pro lepší přehlednost zachytila v procesních diagramech. Na základě těchto zjištění jsem pak identifikovala neošetřené oblasti, čímž byl vytvořen základ pro návrh metodiky. Na základě zpětné vazby od kolegů, nebyly v této části práce zjištění žádné závažné nedostatky a popis se víceméně shoduje s aktuálním stavem v organizaci. Z čehož lze odvodit, že tento cíl práce se mi podařilo splnit.

Při návrhu metodiky jsem kromě výsledků zmíněné analýzy vycházela i z analýzy existujících metodik. Zde jsem se soustředila na předání svých zkušeností s vývojem a na sepsání tzv. nejlepších přístupů k vývoji a údržbě tržišť. Tedy nikoliv na přesný popis, jakým konkrétním způsobem řešit jednotlivé části procesu. Vzhledem k rozsahu práce na toto nebyl prostor. Zde vidím potenciál v dalším rozvoji metodiky, kde budou rozepsány konkrétní postupy. Pokud jde o využitelnost, k tomu jsem nedostala od kolegů bližší vyjádření. Nicméně metodik zjišťoval, zda bude práce přístupná i v rámci organizace, z čehož odvozuji, že bude návrh metodiky používán a dále rozvíjen. Jako zakončení návrhu jsem sepsala dohromady „checklist“ obsahující stručný popis kroků a souvisejících dokumentů, kterými je třeba při návrhu, vývoji a údržbě tržiště projít.

Kromě výše uvedeného jsem si dala za cíl i zjistit význam sémantické vrstvy, jaké jsou její omezení a výhody a jak by se tato vrstva dala využít v organizaci. K tomu jsem využila zahraniční literatury a znalosti získané využíváním reportovací vrstvy v organizaci. Zde jsem došla k závěru, že v současné situaci není možné využití této vrstvy při vývoji datových tržišť. Tomu brání jak neexistence business slovníku, a tedy časová a finanční náročnost potřebná k jeho vytvoření, tak náročnost dopadové analýzy v případě rozvoje reportovací vrstvy.

Přínos mé práce vidím v dokumentaci současného stavu, což může pomoct nový kolegům a externistům při orientaci v procesu správy datových tržišť. Kde se budou moct soustředit na samotný návrh, vývoj a údržbu a nikoliv na zjišťování činností, které je třeba provést. Zároveň jako přínosný hodnotím návrh metodiky, kde bylo mým cílem minimalizovat riziko, že se tržiště bude vyvíjet neúměrně dlouho či nebude vůbec odběrateli využito.

11. Zdroje

Daniels, Chris. 2010. What is a Semantic Layer and Why Would I Want One? *Beye Network.* [Online] 3. Březen 2010. [Citace: 25. Říjen 2011.] http://www.b-eye-network.co.uk/view/12763.

Eckerson, Wayne. 2008. Bridging the Divide: Aligning Analytical Modelers and IT Administrators. *TERADATA.* [Online] Červenec 2008. [Citace: 18. Říjen 2011.] http://www.teradata.com/white-papers/Bridging-the-Divide-Aligning-Analytical-Modelers-and-IT-Administrators-eb5782/.

—. 2010. Three Tiers of Analytic Sandboxes: New Techniques to Empower Business Analysts. *TDWI: The Premier Source for Business Intelligence and Data Warehousing.* [Online] 22. Březen 2010. [Citace: 18. Říjen 2011.] http://tdwi.org/blogs/wayne-eckerson/2010/03/analytic-sandboxes.aspx.

Imhoff, Claudia, Galemmo, Nicholas a Geiger, Jonathan G. 2003. *Mastering Data Warehouse Design: Relational and Dimensional Techniques.* Indianapolis : Wiley Publishing, Inc., 2003. 0-471-32421-3.

Inmon, Bill. 2004. Methodology and The Government Information Factory (GIF). *Beye Network.* [Online] 01. Červen 2004. [Citace: 10. Leden 2012.] http://www.beyenetwork.de/view/184.

—. 2011. The Good Data Mart. *Beye Network.* [Online] 7. Duben 2011. [Citace: 18. Říjen 2011.] http://www.b-eye-network.com/view/15040.

Inmon, Wiliam H. 2005. *Building The Data Warehouse, Fourth Edition.* Indianapolis : Wiley Publishing, Inc., 2005. 978-0-7645-9944-6.

Inmon, William H., Strauss, Derek a Neushloss, Genia. 2008. *DW 2.0 - Architecture for the Next Generation of Data Warehousing.* Burlington : Morgan Kaufmann: Elsevier, 2008. 978-0-12-374319-0.

Kimball, Ralph a Caserta, Joe. 2004. *The Data warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Indianapolis : Wiley Publishing, Inc., 2004. 0-7645-7923 -1.

Kimball, Ralph a Ross, Margy. 2002. *The data warehouse toolkit: the complete guide to dimensional modeling, Second Edition.* místo neznámé : Wiley Computer Publishing, 2002. 0-471-20024-7.

Kimball, Raplh a Ross, Margy. 2010. *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence.* Indianapolis : Wiley Publishing, Inc., 2010. 978-0-470-56310-6.

Loshin, David. 2003. *Business Intelligence: The Savvy Manager's Guide.* San Francisco : Morgan Kaufmann Publishers, 2003. 78-1-55860-916-7.

Moss, Larissa T. a Atre, Shaku. 2003. *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications.* místo neznámé : Addison Wesley, 2003. 0-201-78420-3.

2011. Musing about semantics in BI. *esjewett.com.* [Online] 28. Únor 2011. [Citace: 26. Říjen 2011.] http://www.esjewett.com/blog/musing-about-semantics-in-bi.

Talus. A practical guide to building data marts. *ftp.wiley.com.* [Online] [Citace: 09. Leden 2012.] ftp://ftp.wiley.com/public/computer_books/updates/guide.pdf.

Wells, David L. 2006. *Choosing the Right Data Warehousing Approach.* Mnichov : The Data Warehousing Institute, 2006.

—. 2008. *TDWI Business Intelligence Fundamentals: From Data Warehousing To Business Impact.* Mnichov : autor neznámý, 2008.

whatis.com. 2011. OLAP cube. *SearchDataManagement.com.* [Online] Květen 2011. [Citace: 10. Říjen 2011.] http://searchdatamanagement.techtarget.com/definition/OLAP-cube.

13. Terminologický slovník

Termín (CZ)	Termín (AJ)	Zkratka	Význam
	Business Intelligence	BI	Je architektura a soubor integrovaných provozních aplikací, aplikací pro podporu rozhodování a databází, které poskytují odběratelům dat snadný přístup k jejich datům.
Centrální datový sklad	Central Data Store	CDS	Označení koncepce datového skladu v organizaci. Tato koncepce vychází z koncepce Williama Inmona = CIFu.
	Corporate Information Factory	CIF	Označení koncepce datového skladu dle Williama Inmona.
Dočasné úložiště dat	Data Staging Area	DSA	Datové úložiště sloužící ke sběru dat, před jejich nahráním do integrované vrstvy.
Datový sklad	Data Warehouse	DWH	Představuje soubor podnikových dat.
	Extract, Transformation, Loading	ETL/ELT	Nástroje získávání, transformace a nahrávání dat mezi DSA – integrovaná vrstva a integrovaná vrstva – datová tržiště
Operativní datový sklad	Operational Data Store	ODS	Datové úložiště sloužící k poskytování dat v požadovaném čase.
Reportovací vrstva			Obdoba sémantické vrstvy, ale bez business slovníku a určená pouze pro reporting.
	Sandbox		Specializované prostředí umožňující odběratelům kombinovat data získaná z datového skladu a experimentální dat. Data ze sandboxu nejsou určená k vykazování.
Sémantická vrstva	Semantic layers		Vrstva umístěná nad datovým skladem. Účelem je poskytnutí dat odběratelům z datového skladu v jim známých pojmech.
Smlouva o úrovni poskytovaných služeb	Service Level Agreement	SLA	Definuje kvalitu a rozsah služeb poskytovaných informačním systémem.

14. Rejstřík

B

Business Intelligence · 6, *viz* 6-9

C

CIF · 10, 34

D

Datové tržiště · 8, 13, 34
- Nezávislé datové tržiště · 13
- Sandbox · 18
- Závislé datové tržiště · 10, 17

Datový sklad · 7, 10, 11, 12, 34, 39
Dočasné úložiště dat · 7, 10, 12, 34

I

Inmon · 8, 10, 25

K

Kimball · 8, 9, 11, 29

M

Metodika · 23, 32
- Metodika dle Inmona · 25
- Metodika Metodika vývoje dle Kimballa · 29
- Metodika vývoje dle spol. Talus · 31
- Řízení BI projektu dle Mosse · 23

O

OLAP kostka · 9
Operativní datový sklad · 7

P

Proces změnového řízení · *viz* 44-47

S

Sandbox · 49, 55
Sémantická vrstva · 67, *viz* 20-22
- Reportovací vrstva · 49

15. Seznam obrázků a tabulek

Obrázek 1 - Architektura BI; Doplněno a přeloženo z: (Loshin, 2003 str. 41) (Loshin, 2003 str. 41) (Loshin, 2003 str. 41) 6
Obrázek 2 - OLAP kostka; Zdroj: autor 9
Obrázek 3 – Jednotlivé úrovně CIFu; Přeloženo ze zdroje: (Inmon, 2005 str. 16) 10
Obrázek 4 - Koncepce datového skladu dle Kimballa; Přeloženo ze zdroje: (Kimball, a další, 2002 str. 7) 11
Obrázek 5 - Nezávislá datová tržiště; Přeloženo ze zdroje: (Wells, 2008 stránky 3-20) 13
Obrázek 6 - Dimensionální model: Schéma hvězdy Zdroj: Autor 16
Obrázek 7 - Dimensionální model: Schéma sněhové vločky Zdroj: Autor 16
Obrázek 8 - Závislá datová tržiště; Přeloženo a upraveno ze zdroje: (Wells, 2008 stránky 3-24) 17
Obrázek 9 - Sémantická vrstva; Přeloženo a doplněno autorem ze zdroje: (Wells, 2008 stránky 3-24) 21
Obrázek 10 – Vývojové úrovně projektu; Přeloženo ze zdroje: (Moss, a další, 2003) 23
Obrázek 11 – Jednotlivé fáze „spirálového“ přístupu; Přeloženo ze zdroje: (Inmon, a další, 2008 str. 124) 26
Obrázek 12 - Znázornění průběhu sedmi proudového přístupu; Přeloženo ze zdroje: (Inmon, a další, 2008 str. 138) 28
Obrázek 13 - Heuristický přístup; Upraveno a přeloženo ze zdroje: (Inmon, 2004) 29
Obrázek 14 – Metodika vývoje dle Kimballa; Přeloženo ze zdroje: (Kimball, a další, 2010 str. 97) 30
Obrázek 15 - Znázornění, jakým způsobem předcházet riziku neakceptace tržiště odběrateli; Přeloženo ze zdroje: (Talus str. 7) 31
Obrázek 16 - Detailní přehled jednotlivých kroků a fází PrimeOver metodiky; Přeloženo ze zdroje: (Talus str. 14) 32
Obrázek 17 - Zjednodušený proces vývoje datového tržiště v organizaci; Zdroj: autor 38
Obrázek 18 - Architektura CDS a související dokumenty; Zdroj: autor 41
Obrázek 19 - Současný proces vývoje datového tržiště - od přijetí požadavku ke změnovému řízení; Zdroj: autor 47
Obrázek 20 - Procesy v metodice; Zdroj: autor 53
Obrázek 21 - Posloupnost otázek v analýze dostupnosti dat; Zdroj: autor 56
Obrázek 22 - Příklad konceptuálního modelu datového tržiště; Zdroj: autor 60

Tabulka 1 - Příčiny chybných výsledků při testování a jejich řešení; Zdroj: autor 63