Univerzita Hradec Králové

Fakulta informatiky a managementu

Katedra informačních technologií

Výběr ETL nástroje pro cloudové řešení BI

Diplomová práce

Autor: Michal Zima
Studijní obor: Informační management 5

Vedoucí práce: Ing. Karel Mls, Ph.D.

Odborný konzultant: Ing. Jiří Tobolka
GoodData

Hradec Králové Srpen 2015

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracoval samostatně a s použitím uvedené literatury.

V Hradci Králové dne 17. 8. 2015 Michal Zima

Poděkování:

Děkuji vedoucímu diplomové práce Ing. Karlu Mlsovi, Ph.D. za metodické vedení práce, podnětné rady a čas, který mi věnoval.

Anotace

Diplomová práce se zabývá výběrem ETL nástroje pro Business Intelligence řešení od společnosti GoodData. Práce pojednává mimo jiné o postupech a trendech v daném odvětví, s ohledem na využití technologie cloud. Popisuje metodiku a sestavení AHP modelu a řeší praktické použití ETL nástrojů na reálných datových zdrojích. Na základě nabytých poznatků při práci s nástroji je vytvořeno hodnocení modelu, jehož výsledky pak slouží jako podklad pro doporučení čtenáři.

Klíčová slova: Business Intelligence, GoodData, cloud technologie, ETL, AHP model

Annotation

Title: Choosing ETL tool for cloud-based BI

This diploma thesis deals with choosing an ETL tool for BI solution from GoodData. The thesis discusses the procedures and trends in the sector, with regard to the use of cloud technology. It describes the methodology and compilation of AHP model, and talks about practical usage of ETL tools on real data sources. Acquired knowledge from practical usage of tools and the results of model evaluation serve as a basis for recommendations to readers.

Keywords: Business Intelligence, GoodData, Cloud computing, ETL, AHP model

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Velké množství podniků přichází na to, že v dnešní době je dostatečně rychlá a kvalitní analýza podnikových dat klíčová. Bez získávání nových informací, predikování budoucnosti a prezentace dat, je jen velmi obtížné získat konkurenční výhody. Jak tedy docílit co nejefektivnější analýzy dat, která jsou uložena na mnoha místech a v různých formách? Pro tuto potřebu je na trhu s informačními technologiemi řešení BI (Business Intelligence). To si klade za cíl vyřešit právě tuto otázku.

S rozvojem IT (Informační Technologie) a příchodem nových technologií se neustále zvyšují nároky na část BI řešení, kterou obstarávají nástroje ETL (Extract-Transform-Load) a která se zabývá nahráváním vstupních dat do dalších částí řešení. Vhodný výběr takových nástrojů je tedy důležitou součástí celé strategie tvorby BI a měli bychom mu věnovat dostatečnou pozornost. Děláme to tak, abychom docílili řešení, které bude nejen funkční a vyhovující, ale také do budoucna snadno upravitelné a znovupoužitelné. Trendem se pak stává přesunutí dat a logiky BI do tzv. cloudu. To s sebou přináší výhodu přístupu k řešení pomocí internetu.

V první, teoretické části, se práce zabývá popisem BI technologií a uvedením pojmů do širších souvislostí. Důraz je kladen především na část o ETL nástrojích a přínos cloud computingu do dané problematiky. V druhé, praktické části, se práce věnuje již samotnému použití vybraných nástrojů. Zkoumána jsou reálná data poskytnutá telefonickou společností 1188 a vytvořeno kompletní řešení BI v platformě GoodData. Všechny postupy v popsaných nástrojích jsou dostatečně dokumentovány a mohou sloužit jako manuál pro práci s nimi. Na závěr jsou vybrané nástroje hodnoceny na základě zkušeností a získaných poznatků a jsou vyslovena doporučení pro čtenáře.

2 Formulace řešeného problému a cíl práce

2.1 Důvod výběru tématu

Toto téma jsem si pro svoji závěrečnou práci vybral z důvodů zkušeností, které jsem nabyl během studia ve firmě GiST s. r. o. V této společnosti jsem se mimo jiné věnoval návrhu a tvorbě datových skladů na platformě MS SQL Server a podílel jsem se na vývoji BI nástroje GiST Intelligence.

Zkušenosti s těmito technologiemi a vlastní zájem o tento dynamicky se rozvíjející obor se staly základem tohoto rozhodnutí. Věřím také, že nově získané dovednosti a poznatky s tvorbou této diplomové práce budou přínosem v mém budoucím profesním životě.

2.2 Popis problému

Vhodný výběr ETL nástroje je důležitou součástí celého řešení BI. V této diplomové práci se tedy budu zabývat výběrem nástroje pro BI aplikaci od společnosti GoodData, která byla založena v roce 2007 panem Romanem Staňkem. Toto řešení jsem si vybral po předchozím vyzkoušení technologie na akci Enterprise Data Hackathon 2014 v Praze. Ta byla, mimo jiné, zaměřena na praktické testování analytických nástrojů na datových zdrojích, poskytnutých několika českými firmami.

Právě zde se objevil problém, jakým způsobem a jaký nástroj bude nejvhodnější pro nahrání datového zdroje do aplikace od společnosti GoodData. Pro nedostatek času a kapacit, se nikdo nemohl příliš zabývat rozsáhlejším zpracováním této otázky.

2.3 Cíle práce

Hlavním cílem diplomové práce je **vytvoření doporučení** čtenáři, jaký ETL nástroj je nejvhodnější pro nahrání dat do platformy GoodData.

Vedlejším cílem je pak podrobné prozkoumání, **návrh a implementace řešení v platformě GoodData**, kdy výstupem vedle BI projektu bude také podrobná dokumentace řešení a **analýza dat firmy 1188**.

Dalším postupným cílem je **definovat AHP model** pro hodnocení jednotlivých ETL nástrojů a za pomoci nabytých zkušeností během implementace praktické úlohy tento model vhodně ohodnotit, a získat tak podklady pro vytvoření závěrečného doporučení.

2.4 Metodika dosažení cíle

Pro dosažení cílů této diplomové práce byla vypracována podrobná **rešerše** akademických prací, literatury a článků na dané či podobné téma. Výběr nástrojů byl poté diskutován s odborným konzultantem a omezen pouze na dva vhodné nástroje. Ty byly podrobeny následnému hlubšímu a detailnějšímu zkoumání.

Bylo využito také **vícekriteriálního rozhodovacího modelu AHP** (Analytický hierarchický proces). Pomocí párového srovnání a ohodnocení vah jednotlivých kritérií byl vytvořen model, který se stal nástrojem pro zpracování požadovaného doporučení.

V neposlední řadě byla použita také **poskytnutá data** od společnosti 1188, která se stala datovým zdrojem pro řešení praktické úlohy pomocí obou zkoumaných ETL nástrojů. Tímto způsobem došlo k požadovanému praktickému testování.

2.5 Struktura práce

Celá práce je rozdělena na dvě části; teoretickou část, která je shrnutím teoretického bádaní a praktickou, která zaštituje popis věcného zkoumání a jeho poznatky.

Dále je k práci připojen **terminologický slovník pojmů**, který by měl čtenáři poskytnout dostatečné vysvětlení některých odborných názvů často přebíraných z anglického jazyka.

Přiloženy jsou také seznamy použité literatury, obrázků, tabulek a v poslední řadě přílohy, které slouží především k dokumentaci řešení praktické části.

2.6 Výstupy a očekávané přínosy práce

Hlavním výstupem této práce by mělo být **doporučení čtenáři**, který z ETL nástrojů je nejvhodnější pro řešení BI v cloudu.

Vedlejším cílem bude **představení jednotlivých nástrojů** a platforem a také stručný uživatelský **návod pro implementaci ETL nástrojů** ve spojení s konkrétní platformou.

V neposlední řadě bude přínosem **analýza dat společnosti 1188**, která díky získaným poznatkům může výsledky a přístup aplikovat v praxi.

3 Rešerše

Pro potřebu diplomové práce jsem provedl rešerši akademických prací na dané nebo společné téma a také rešerši knižních a internetových zdrojů. Na výsledcích tohoto pátrání a přešetřování byla vytvořena teoretická část práce tak, aby se nepřekrývala s pracemi mých kolegů a byla dostatečným podkladem pro vypracování praktické části.

3.1 Akademické práce

Za nejvíce tematicky podobnou akademickou práci považuji *„Nástroje Business Intelligence jako open source“* (Filipčík, 2013), která se zabývá **výběrem open source** nástrojů pro jednotlivé části řešení BI. V práci je věnován dostatečný prostor pro zkoumání open source **ETL nástrojů**, které jsou následně hodnoceny podle autorových kritérií. Jako nedostatek vidím, že autor nástroje porovnává velmi povrchně a nezachází příliš do detailů. Navíc je omezen tématem a vybral pouze nástroje open source, což má za následek vynechání velké škály komerčních nástrojů.

Diplomová práce *„Srovnání cloud BI řešení a faktory ovlivňující jejich nasazení“* (Černý, 2014) velice dobře analyzuje **přínosy cloud computingu** ve spojení s BI a podrobně pak popisuje toto řešení. Autor také stručně rozebírá možnosti ETL nástrojů u takového řešení, ale téma už dále podrobněji nerozvádí. Zde jsem nalezl příležitost na navázání a podrobnější rozpracování autorových myšlenek.

Další diplomová práce *„Návrh a implementace Business Intelligence řešení“* (Kocábek, 2012) je zaměřená, jak již název napovídá, na celkový **popis řešení BI**. Autor odvedl dobrou práci při popisu **historie a vývoje** BI, jen stručně zmiňuje možnosti cloud computingu v této oblasti. Přesto si myslím, že se jedná o velmi dobrou práci, která může čtenáři pomoci osvojit si a lépe pochopit principy BI.

Dále už jen zmiňuji bakalářskou práci *„ETL nástroje“* (Kubán, 2013), která **hodnotí** tři komerční a tři open source **nástroje pro práci s relačními databázemi**, ale svým rozsahem nezabíhá do detailnějších rozborů a je spíše představením autorem zmíněných aplikací.

3.2 Literární zdroje a odborné články

Tato část rešerše vychází převážně z knihy *„Business Intelligence - Jak využít bohatství ve vašich datech“* (Novotný a kol., 2005), která patří mezi zdroji v českém jazyce na dané téma k jedné z nejvýznamnějších. Kniha je velmi dobře a podrobně zpracována a poskytuje tak čtenáři možnost vytvořit si komplexní obraz o dané tématice. Celá **architektura a principy BI** řešení jsou zde velmi podrobně popsány a titul doporučuji jako doplňkovou četbu k této diplomové práci.

Dalším zdrojem je *„The Data Warehouse ETL Toolkit“* (Kimball a Caserta, 2004). Tato kniha popisuje detailně **principy** a techniku **práce s datovými zdroji**. Bohužel se jedná o starší knihu, a v tomto rychle rozvíjejícím odvětví by si žádala přepracování s přihlédnutím k novějším technologiím a trendům. Přesto mi byla dobrým vodítkem při zpracování kapitoly o ETL nástrojích a poté v praktické části při práci se zdroji a tvorbě datových pump.

Dobrým doplňkem k předchozí knize je *„Business Intelligence. A Managerial Approach“* (Turban, 2011). Kniha nabízí pohled na danou tématiku trochu jinýma očima a to konkrétně z pohledu manažerů. Dílo je o několik let mladší než Kimballova kniha, což mi pomohlo k uvědomění trendů v oboru.

V odborném článku *„Bussines Intelligence in the Cloud?“* (Baars a Kemper, 2010) autoři velmi podrobně popisují výhody a **směřování Cloud BI**. Věnují také dostatečnou pozornost využití webových ETL nástrojů a rozebírají výhody a nevýhody tohoto přístupu.

Pro potřebu literární rešerše bylo zapotřebí také pročíst mnoho dalších zdrojů, ať už méně či více významných. Tato rešerše je shrnutím nejdůležitějších pramenů, které se staly základem pro zpracování práce. Ostatní použité zdroje jsou pak uvedeny v kapitole Seznam použité literatury.

4 Teoretické vymezení dané problematiky

4.1 Vymezení základních pojmů

Pro snazší porozumění této práce si definujeme několik důležitých pojmů, které budou často používány. Popíšeme si také jejich kontext, ve kterém budou po celou dobu vnímány. Některé další použité termíny ze zkoumané oblasti jsou vysvětleny v Terminologickém slovníku pojmů na konci práce.

4.1.1 Business Intelligence

S tímto pojmem, který nemá v češtině svůj vlastní ekvivalent, se setkáváme v praxi běžně a překlad by byl spíše na škodu. Podle knihy (Novotný a kol., 2005, s. 13) označuje tento termín celý **komplex činností, úloh a technologií**, které dnes stále častěji tvoří součást řízení podniků a jejich informačních systémů.

Volně přeložená definice tohoto pojmu z vědeckého článku (Negash, 2004), která nejlépe vyhovuje našemu dalšímu použití, by zněla: „Business Intelligence je systém, který kombinuje sběr dat, ukládání dat a řízení znalostí pomocí analytických nástrojů pro prezentování informací, pro osoby s rozhodovacími a plánovacími pravomocemi ve firmě."

Takového vnímaní se přidržíme, v dalším textu bude použita pouze zkratka BI.

4.1.2 ETL

Se sběrem a ukládáním dat souvisí další pojem z informatiky **Extract - Transform - Load** neboli ETL. Tak, jak popisuje Novotný (Novotný a kol., 2005, s. 29), se jedná o transformační nástroj, který je hlavním stavebním kamenem celého BI řešení, a jeho úkoly jsou; data ze zdrojových systémů získat a vybrat (Extraction), upravit do požadované formy a vyčistit (Transformation) a nakonec nahrát do datového skladu (Loading).

V praxi se také běžně používá termín „datová pumpa", což je český pojem pro nástroj, který slangově řečeno pumpuje data do připravených struktur.

4.1.3 Cloud

Cloud neboli **cloud computing** zjednodušeně znamená přistupování k datům a programům pomocí internetu, bez využití pevných disků počítačů. Ve světě byznysu se objevují v této souvislosti další termíny, jako je **SaaS** (Software as a Service), tedy aplikace, ke které přistupujeme přes internet, nebo **PaaS** (Platform as a Service), kde se vytváří přímo celá aplikace, nebo platforma, která je umístěna mimo vnitřní síť (Jamsa, 2013, s. 5). V dnešní době se, ale používá místo zmíněných pojmů naprosto běžně pojem cloud (Griffith, 2015). Toho se přidržíme i my a výraz bude použit jako slovo přejaté a běžně skloňované. Podrobněji jsou tyto technologie a principy popsány v následujících kapitolách.

4.2 Historie a vývoj BI

Dalo by se říci, že historie BI je poměrně dlouhá, protože podle článku (Power, 2007) byly základy položeny již koncem 60. let, kdy se objevovaly první debaty o **Decision Support Systems (DSS).** Tak, jak se píše v literatuře (Khosrow-pour, 2009, s. 1753), se jedná, volně přeloženo, o na modelech založený soubor procedur, na zpracování dat a usuzování o nich, který je určený k pomoci manažerovi při jeho rozhodování. Tedy přesně tak, jak bychom mohli vnímat i BI.

Dlouhou dobu jsme si vystačili s tímto označením, dokud zaměstnanec společnosti Gartner[1], **Howard Dressner** v roce 1989 nepoužil dnes dobře známý název Business Intelligence. Tento uznávaný odborník a analytik mluví o BI podle Powera (Power, 2007), volně přeloženo, jako o **sadě konceptů a metod ke zlepšení tvorby manažerských rozhodnutí,** za použití na faktech založených systémů.

Automatizací dalších procesů a příchodem nových paměťových medií došlo k nárůstu objemu dat, která v sobě ukrývala pro management zajímavé informace. Dřívější aplikace, které měly každá svoje vlastní provozní databáze, bylo potřeba propojit. Koncem osmdesátých a začátkem devadesátých let se v USA začal velmi silně prosazovat

[1] Společnost Gartner, Inc. patří mezi přední poradenské firmy a zabývá se výzkumem IS/ICT technologií. Jejím zakladatelem je Gideon Gartner.

nový trend. Použití multidimenzionálního modelu, datových skladů (Data Warehouse) a datových tržišť (Data Marts). Tak, jak je zdůrazněno v knize (Novotný a kol., 2005, s. 17), za rozvojem těchto přístupů stáli především **Ralph Kimball** a **Bill Inmon**. Právě prvně jmenovaný Ralph Kimball je uznávaný jako jeden z průkopníků v oblasti ukládání dat. Jeho metody a myšlenky jsou velmi dobře známy jako dimenzionální modelování.

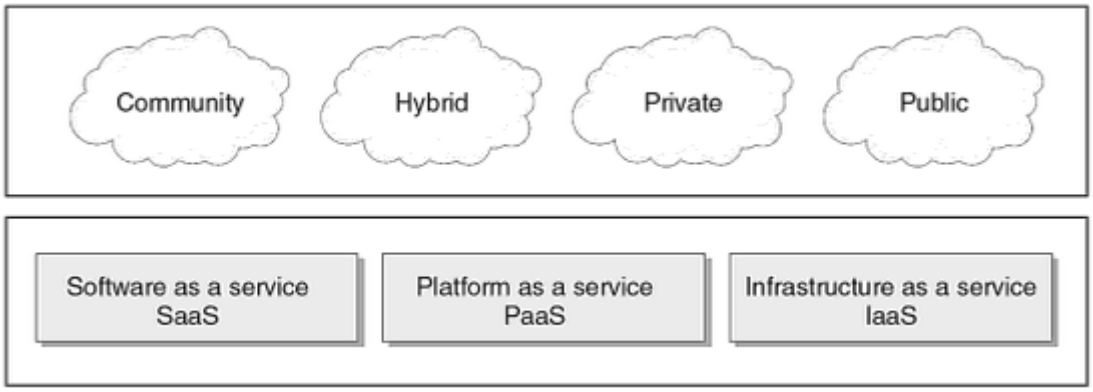
Dimenzionální model je populární způsob ukládání dat, který odděluje fakta od dimenzí do samostatných tabulek, a tím přispívá ke stabilitě při měnícím se prostředí vstupních dat a rychlosti zpracování dotazů do datového uložiště (Kimball, 2004). Mnoho let popsaná pravidla v knihách a pracích Ralpha Kimballa jsou bezpochyby základem většiny BI řešení. S jeho myšlenkami se setkáváme v této oblasti neustále. Přestože by se dalo říci, že vše už bylo napsáno a praxí mnohokrát ověřeno, stále dochází v tomto odvětví k značnému vývoji. Hlavními důvody jsou měnící se požadavky uživatelů a nástup nových technologií. Takovou technologií potom může být například cloud a cloudové řešení BI.

4.2.1 Přínos cloud computingu

S nástupem cloud computingu se mění zažitá pravidla a stereotypy v BI. Vše díky technologii, která je atraktivní pro uživatele a má potenciál změnit také velkou část IT průmyslu. **Cloud computing** byl již od začátku zajímavým tématem, které plnilo množství magazínů, blogů a periodik. To zapříčinilo, že uživatelé začali chápat jako cloud vše, co je nějakým způsobem spojeno s internetem (Armbrust, 2010), viz Obr. 1.

Obr. 1 Často chybně vnímaný pojem cloud.
Zdroj: (OneMetric, 2015).

Takový přístup není zcela správný, a proto si cloud computing lépe představíme a popíšeme. V praxi se dnes setkáváme již s velkou škálou cloudových řešení, která mají za úkol uspokojit množství rozličných zákaznických přání. Abychom mohli analyzovat a popsat tyto systémy, mnoho lidí, včetně pracovníků NIST[2], dělí přístup do dvou modelů podle jejich funkce (Jamsa, 2013, s. 4). Takové rozdělení na Model publikování (Deployment Model) a Model služeb (Services Model) je patrné i na Obr. 2.



Obr. 2 Rozdělení vnímaní cloudu podle modelů použití.
Zdroj: (Jamsa, 2013, s. 4).

Zmíněného rozdělení se přidržíme i v této diplomové práci a za pomocí knihy Cloud computing (Jamsa, 2013) si oba modely podrobněji rozebereme.

Rozdělení podle způsobu publikování a sdílení dat

- **Private cloud** - vlastněn specifickou skupinou a využíván jen pro potřeby této skupiny nebo jejích uživatelů.
- **Public cloud** - dostupný pro používání veřejností. Často je vlastněn velkou organizací nebo firmou poskytující cloudové služby. Díky vyšší dostupnosti může být méně bezpečný.

[2] National Institute of Standards and Technology (NIST) je uznávaná laboratoř Ministerstva obchodu Spojených států amerických, jejímž cílem je zlepšování vědeckých měření, podpora inovací a další činnosti.

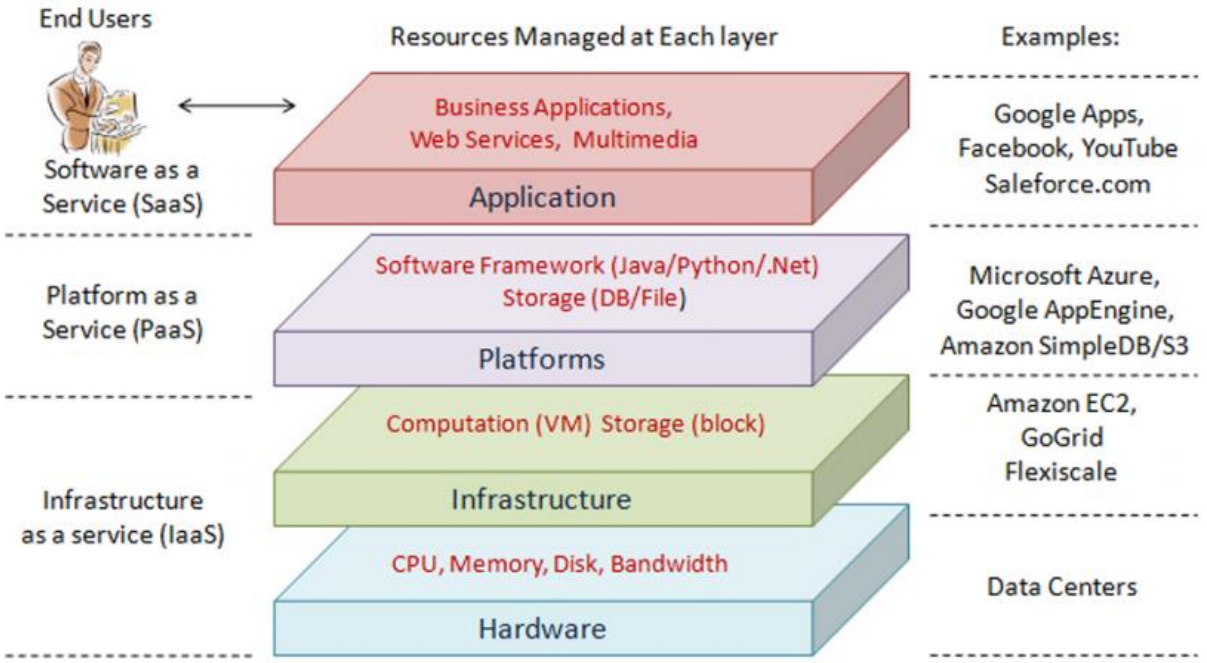
- **Community cloud** - sdílen dvěma nebo více organizacemi. Jedná se o typické řešení pro organizace, které se dělí na menší subjekty a které mezi sebou potřebují spolupracovat (např. fakulty spadající pod jednu univerzitu).
- **Hybrid cloud** – spojení dvou nebo více předchozích typů do jednoho sdruženého celku.

Rozdělení podle způsobu poskytovaných služeb

Cloud může také komunikovat s uživateli pomocí tzv. služeb. Tímto rozdělením dostáváme tři základní způsoby:

- **Software as a Service (SaaS) –** kompletní aplikace, která má vlastní uživatelské rozhraní.
- **Platform as a Service (PaaS)** – platforma, díky které mohou vývojáři vytvářet vlastní aplikace. Toto řešení nabízí cloudové operační systémy, vývojové nebo administrační nástroje a také hardware v podobě serverů a disků.
- **Infrastructure as a Service (IaaS)** – poskytuje virtuální stroje, uložiště a síťové zdroje, které mohou vývojáři využít pro instalování vlastních operačních systémů nebo aplikací.

Podrobnější rozdělení můžeme vidět na Obr. 3, kde v pravé části Examples jsou uvedeny i konkrétní příklady známých služeb pro jednotlivé typy řešení v cloudu.



Obr. 3 Rozdělení cloudu podle způsobu poskytovaných služeb.
Zdroj: (Zhang a kol., 2010).

Přesun IT služeb do cloudu se stává stále častějším zvykem. Nutno podotknout, že mnohá BI řešení využívají například cloudový přístup jen pro některé svoje části. Dokladem toho, že BI v cloudu má svůj potenciál, je i následující kapitola o současných trendech.

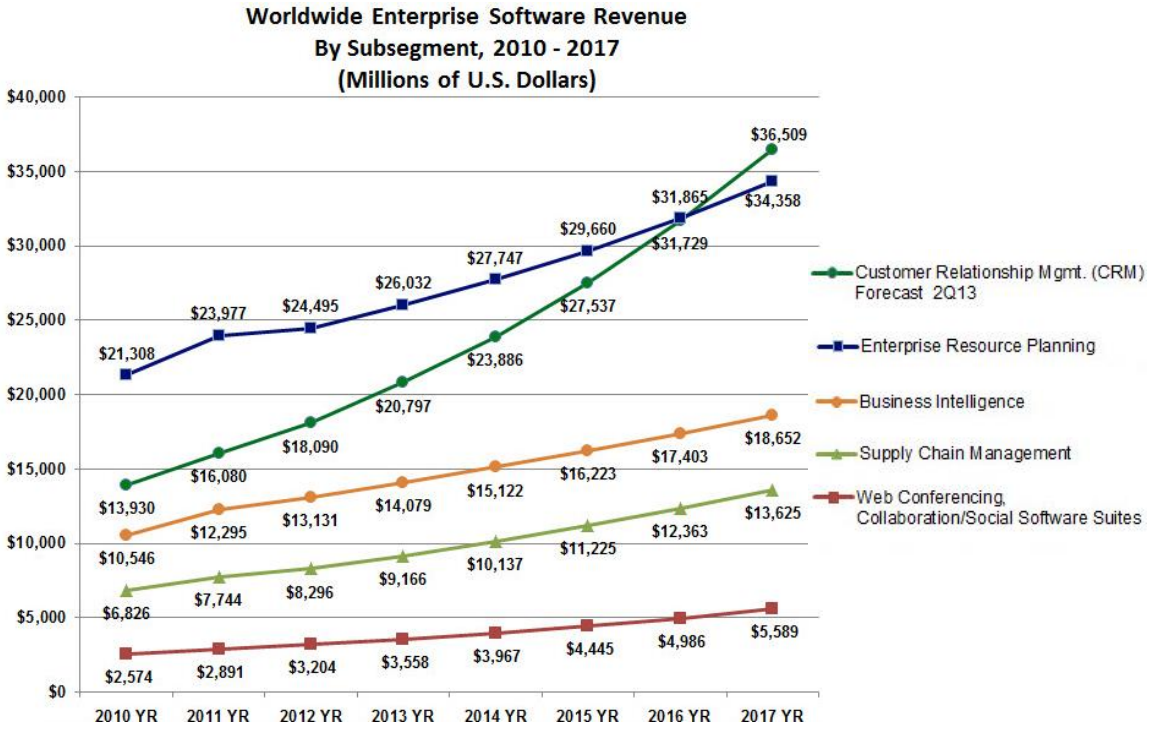
4.2.2 Současné trendy

Nároky uživatelů v oblasti informačních technologií a také v oblasti BI, se neustále zvyšují. Vyžadují nejen okamžitý a dostatečně detailní přístup k datům v informačních systémech, ale kladou důraz také na vysokou míru bezpečnosti a schopnost přizpůsobení analytického rozhraní nástroji, se kterým uživatel pracuje (Zikmunda, 2014).

To potvrzuje i Gartner ve své predikci o vývoji BI v dalších letech. Podle výzkumu předních analytiků této společnosti bude mít v roce 2017 většina uživatelů a analytiků v organizacích přístup k nástrojům, které jim umožní samostatně si připravit data pro analýzu. To by znamenalo, že vývoj BI neztrácí na tempu a v budoucnu bude na tuto technologii kladen ještě větší důraz (Laney a Zaidi, 2015).

Toto dokládá i vývoj a predikce světových příjmů z oblasti BI, který Gartner uveřejnil na svých webových stránkách v roce 2013. Přestože výsledky nejsou volně dostupné a jejich cena je pro běžného čtenáře vysoká, některé servery uveřejnily alespoň dva roky starý graf na Obr. 4 a jako zdroj uvádí právě zmíněnou hodnotnou studii.

Tento graf znázorňuje příjem z jednotlivých segmentů informačních technologií, včetně námi zkoumaného BI. Můžeme si také všimnout, že objem investic v daném odvětví se zvyšuje. Po dosazení ročních výnosů do vzorce pro výpočet průměrného meziročního růstu nám vychází, že tento růst je 8,1%.



Obr. 4 Graf vývoje světových příjmů[3] za jednotlivé segmenty.
Zdroj: (Forbes, 2013a).

Podle další studie společnosti Gartner, jejíž výsledky zveřejnil opět Forbes (Forbes, 2013b), je tomu i u řešení BI v cloudu. V Tab. 1 vidíme údaje zachycující tento vývoj s predikcí do roku 2016. Výpočet průměrného meziročního růstu zde pak vychází na 30,5%. Je tedy jasné, že potenciál růstu tohoto přístupu je obrovský a jen potvrzuje význam cloudu ve spojení s BI.

[3] Uvedené příjmy jsou v milionech amerických dolarů.

Tabulka 1 Predikce pro vývoj celosvětového množství výdajů[4] na cloudová řešení.

Cloud Application Services (SaaS)	2010	2011	2012	2013	2014	2015	2016
Business Intelligence Applications	0.14	0.22	0.29	0.37	0.48	0.59	0.73
CRM	3.39	4.21	5.03	5.92	6.87	7.88	8.96
Digital Content Creation	0.10	0.22	0.27	0.37	0.48	0.71	0.93
Enterprise Content Management	0.20	0.26	0.37	0.51	0.62	0.72	0.82
ERP	1.50	1.97	2.51	3.18	3.95	4.74	5.65
Office Suites	0.11	0.23	0.41	0.73	1.11	1.39	1.72

Zdroj: (Forbes, 2013b).

Díky vyšší zralosti cloudových služeb, snižování nákladů a využití modelu pay-as-you-go se předpokládá neustálé zvyšování počtu těchto řešení (Noctuint, 2014). Model **pay-as-you-go** umožňuje firmě ušetřit nemalé prostředky, protože ta platí pouze za výkon, který opravdu používá, a nemusí hradit náklady na pořízení vlastních serverů, jejich správu a údržbu (Chaudhuri, 2011). To přináší společnostem také více prostoru pro interpretaci a využívání informací z dat, oproti starostem s technickým zabezpečením řešení.

4.3 Architektura BI řešení

Rozmanitost problémů řešených pomocí nástrojů BI, stejně jako rozmanitost nástrojů a dostupných technologií vede k tomu, že obecná architektura má několik vývojových větví a její konkrétní aplikace v reálných situacích se podstatně liší. Přesto lze podle Novotného (Novotný a kol., 2005, s. 26) identifikovat tyto čtyři základní vrstvy:

- **Zdrojové systémy a data**
- **Vrstva pro extrakci, transformaci, čištění a nahrávání dat**
- **Vrstva pro ukládání dat**
- **Prezentační vrstva**

Postupně si tyto vrstvy představíme a podrobněji rozebereme.

[4] Uvedené výdaje jsou v miliardách amerických dolarů.

Nesmíme však zapomenout připomenout důležitý faktor, a tím je velikost a komplexnost projektu. V praxi to znamená, že u malých projektů dochází k vynechání některých částí, a naopak u projektů větších se model komplikuje o další prvky řešení.

4.3.1 Zdrojové systémy a data

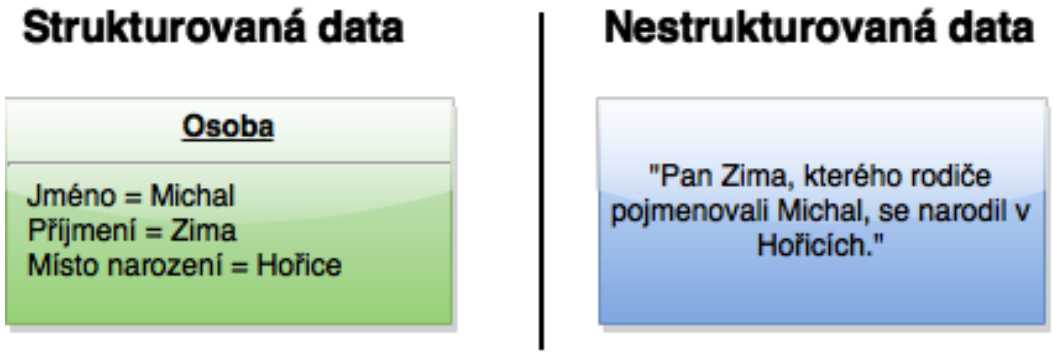
Na úplném začátku celého procesu řešení jsou data. Každý systém generující datové zdroje využívá pro svoje účely jinou organizaci dat a také jiný formát. Data se dělí do třech základních kategorií:

Strukturovaná data

Jedná se o data, která mají vysoký stupeň organizace a pevně dané schéma pro generování. Hlavní výhodou je snadné vstupovaní, ukládání a pozdější analýza. Takovým příkladem dobře strukturovaných dat mohou být např. zákaznická data nebo obchodní data uložená třeba v databázi.

Nestrukturovaná data

Struktura těchto dat není zřejmá a to ani s opakujícím se počtem datových instancí. Data tedy postrádají jakékoliv schéma. Pro ilustraci takového rozdílu oproti datům strukturovaným slouží schéma na Obr. 5.



Obr. 5 Ukázka rozdílu mezi strukturovanými a nestrukturovanými daty.
Zdroj: Vlastní práce autora.

Příkladem takovým dat, kromě textu a dokumentů, mohou být i obrázky, videa, profily sociálních sítí, emaily, hlasové záznamy, kontakty, kalendáře, noviny, webové stránky a další.

Semi-strukturovaná data

Jedná se o formu strukturovaných dat, která však není v souladu s formální strukturou datového modelu. Nejčastěji se jedná také o data z ERP a CRM systémů nebo o různé samostatné soubory s vlastní strukturou, jako jsou generovaná metadata[5]. Vhodným příkladem může pak být formát XML nebo JSON. Názorná ukázka takového souboru je vidět na Obr. 6.

```
<? XML VERSION = "1.0" STANDALONE = "yes" ?>
<PRIKLAD>
    <OSOBA><JMENO>Michal</JMENO>
        <PRIJMENI>Zima</PRIJMENI>
        <ADRESA><ULICE>Vysoká 123</ULICE>
            <MESTO>Hořice</MESTO>
        <ADRESA><ULICE>U Letců 321</ULICE>
            <MESTO>Brno</MESTO>
        </ADRESA>
    </OSOBA>
    <OSOBA><JMENO>Jan</JMENO>
        <PRIJMENI>Novák</PRIJMENI>
        <POHLAVI>muž</POHLAVI>
        <VEK>31<VEK>
    </OSOBA>
</PRIKLAD>
```

Obr. 6 Ukázka semi-strukturovaného souboru XML.
Zdroj: Vlastní práce autora.

Podle odborného článku (Rahm a Do., 2000) se u zdrojových souborů musíme často vypořádat se špatnou kvalitou, jako jsou duplicitní záznamy, překlepy a chybějící údaje. V takovém případě je zapotřebí data tzv. vyčistit a při práci dodržovat několik následujících pravidel a kroků:

- **Analýza dat** – proces, sloužící k zjištění, které chyby a nekonzistence v datech budou odstraňovány.

[5] Metadata jsou strukturované informace, která popisují nebo vysvětlují použití datového zdroje. Někdy se zjednodušeně uvádí, že jde o data o datech. (NISO, 2004)

- **Definice pravidel transformace a mapování na zdrojích** - součást ETL procesů, která určuje, jak budou data zpracována a transformována, aby se předešlo případným duplicitám a nekonzistencím.
- **Ověření** – proces, který na vzorku nebo části dat kontroluje správnost nastavení datové pumpy.
- **Transformace** - jedna z částí ETL procesu, která se provádí nad načtenými zdroji a jejímž úkolem je data upravit do potřebné podoby neboli transformovat.
- **Zpětné čištění dat –** proces, který opravuje a nahrazuje data i na původním datovém zdroji, aby se chyby na zdroji již neopakovaly. Dochází tím k usnadnění práce při dalším použití.

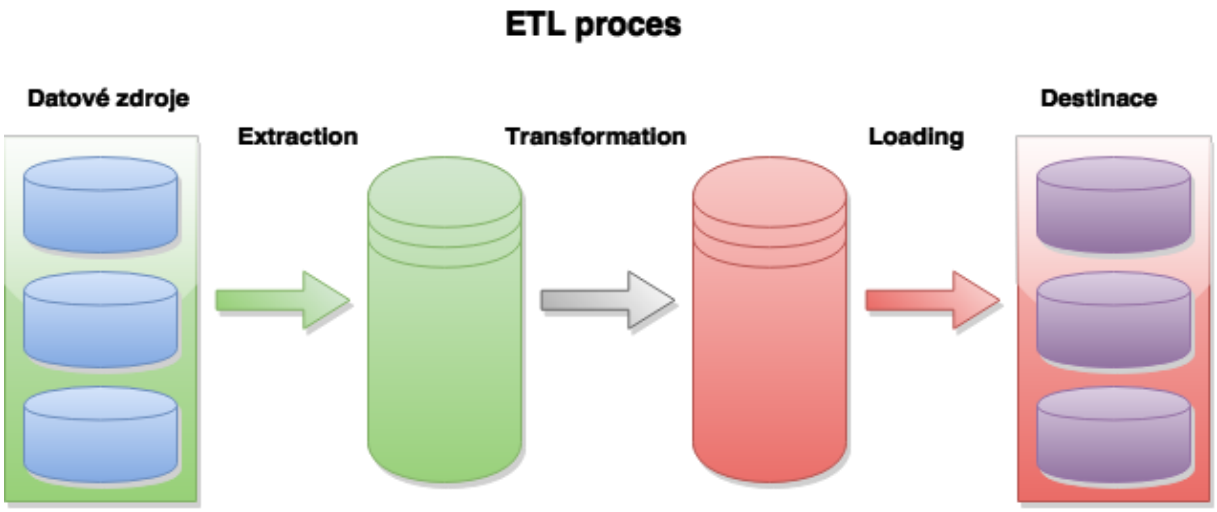
4.3.2 ETL proces

Následující kapitola je, pokud tomu není uvedeno jinak, zpracována na základě Novotného knihy (Novotný a kol. 2005). ETL proces často běží na pozadí celého BI řešení, a přesto je na něj kladen velký důraz, protože výrazně ovlivňuje kvalitu projektu. Základními úlohami takového procesu jsou:

- **Udržování konzistence dat**, tedy doplňování chybějících údajů nebo odstraňování chyb, které vznikají při samotné transformaci či selháním systému.
- **Sběr informací vyjadřujících míru důvěryhodnosti těchto dat.**
- **Doručit data co nejefektivněji do dalších vrstev řešení.**
- **Úprava dat do podoby pro jejich napojení s dalšími datovými zdroji.**
- **Úprava dat pro potřeby dalších systémů,** které k datům přistupují, tedy funkce rozhraní.
- **Sběr metadat,** která vznikají, pokud se mění struktura nebo formát původních vstupů.

Životní cyklus ETL procesu závisí vždy na konkrétní implementaci, a hledat obecně platná pravidla je jen velmi obtížné, tak jako modelování tohoto cyklu. Příčinou je vlastní

přístup dodavatelů softwaru pro řešení těchto otázek. (Vaisman, 2014, s. 285) I přes tento fakt se v každém takovém řešení setkáme minimálně se třemi základními fázemi, viz Obr. 7.



Obr. 7: Schéma ETL procesu.
Zdroj: Vlastní práce autora podle článku (Rahm a Do., 2000) .

Extraction

První fází celého procesu je Extraction, neboli extrakce dat ze zdroje. Tato fáze je velmi důležitá pro dodržování předem daných konvencí, protože je velmi náchylná k chybám ze strany poskytovatelů dat. Řešitel by ji měl přikládat dostatečnou pozornost a popsat procesy do takových podrobností, aby při opětovném použití pumpy byla poskytnutá data ve správném formátu. Cílem je tedy **správně přečíst zdrojové soubory**, podle předem připraveného nebo tzv. namapovaného schématu, a umožnit tak přechod do další částí ETL procesu.

Jak již bylo předesláno, k tomu je zapotřebí provést správnou **analýzu zdrojových dat**, protože se v praxi stává, že data jsou duplicitně uložena na více místech, nebo je potřeba objevit jejich další skrytá napojení.

Data jsou v průběhu také kontrolována, neboli validována. Pokud při validaci dojde k chybě, není tedy možné data z nějakého důvodu uložit do předpřipravené struktury, dochází k zahození celého bloku. Může dojít i k zahození všech rozbalených dat, a celý proces se tak musí opakovat.

Transformation

Druhá fáze má za úkol načtená data transformovat do takového stavu, který bude vyhovovat dalším operacím s nimi. Jedná se o proces, který upravuje data z načtené formy do formy cílové. Zde jsou uvedeny operace, ke kterým může docházet:

- **Konverze datových typů, vzdáleností a jednotek**
- **Změna údajů pomocí matematických operací**
- **Denormalizace dat**
- **Vytvoření multidimenzionálních struktur**
- **Vygenerování pomocných identifikátorů**
- **Shlukování dat**
- **Odvození nových hodnot**

Loading

Poslední etapou je nahrání dat, kdy se informace ukládají v již upravené formě do centrálního uložiště BI řešení. U složitějších ETL procesů se můžeme často setkat s tzv. dočasným úložištěm dat, označovaným také jako staging area. Tady probíhá část transformací a data jsou zde pro potřeby procesu dočasně uložena. Hlavní výhodou je pak možnost jednou načtená data opětovně v transformacích využít, bez nutnosti dalšího zatížení zdrojového systému. (Peterka, 2010)

4.3.3 ETL nástroje v cloudu

Můžeme se také setkat s ETL nástroji, které fungují v cloudu, a uživatel tak nemusí mít nainstalovaný program na svém počítači, ale přistupuje k němu přes webové rozhraní. Je nutno podotknout, že tento trend je celkem populární a následující výčet nástrojů (Lane, 2013) je toho jen důkazem:

- **Cloudwork** – služba, která umožňuje uživatelům zautomatizovat přesun dat mezi Google Apps, Salesforce, Evernote, Zoho, Twitter, Freshbooks, MailChimp, Zendesk, Dropbox, WordPress a dalšími.

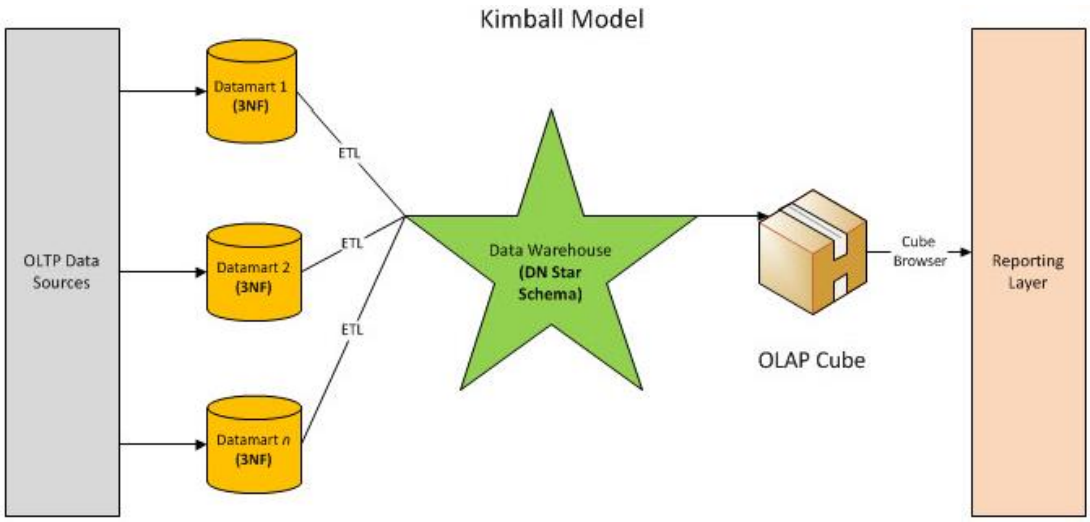
- **Elastic.io** - vysoce škálovatelná platforma, která pomáhá zautomatizovat provoz a připojit mezi sebou další aplikace v cloudu.
- **Foxweave** – tento nástroj umožňuje snadno přenést a synchronizovat data napříč cloudovými a on-premise aplikacemi a databázemi.
- **itDuzzit** - tato platforma dokáže synchronizovat data mezi službami v cloudu. Vyznačuje se vysokým výkonem, bez potřeby větších technických znalostí uživatele.
- **CloverETL** - tato open source platforma vytvořená v Javě poskytuje uživateli možnost migrace dat, jejich čištění a případnou transformaci, a jako open source je základem pro řešení CloudConnect, který využívá i firma GoodData.
- **AutomateIt** - je open source nástroj pro automatizaci nastavení a údržbu serverů, aplikací a jejich závislostí, poskytuje způsob, jak spravovat soubory, balíčky služeb, sítě, účty, role, šablony a další.
- **Pentaho** - pomocí metadat řízený a dobře škálovatelný ETL nástroj, který je znám díky intuitivnímu a graficky příjemnému prostředí.
- **Talend** - další z open source řešení pro správu podnikových dat fungujících v cloudu.

4.3.4 Datový sklad

Další částí BI řešení je tzv. datový sklad. Řešení vždy záleží na konkrétních požadavcích zadavatelů, a proto i metody tvorby datového skladu se v praxi často liší. Nejčastěji se však setkáváme s jednou z metod popsanou Raplhem Kimballem nebo Billem Inmonem. Obě metody si za dobu své existence vytvořily svoje vlastní tábory zastánců, a diskuze na téma, který model je lepší, se vedou neustále. Pokusíme si tedy vysvětlit alespoň základní rozdíly obou variant.

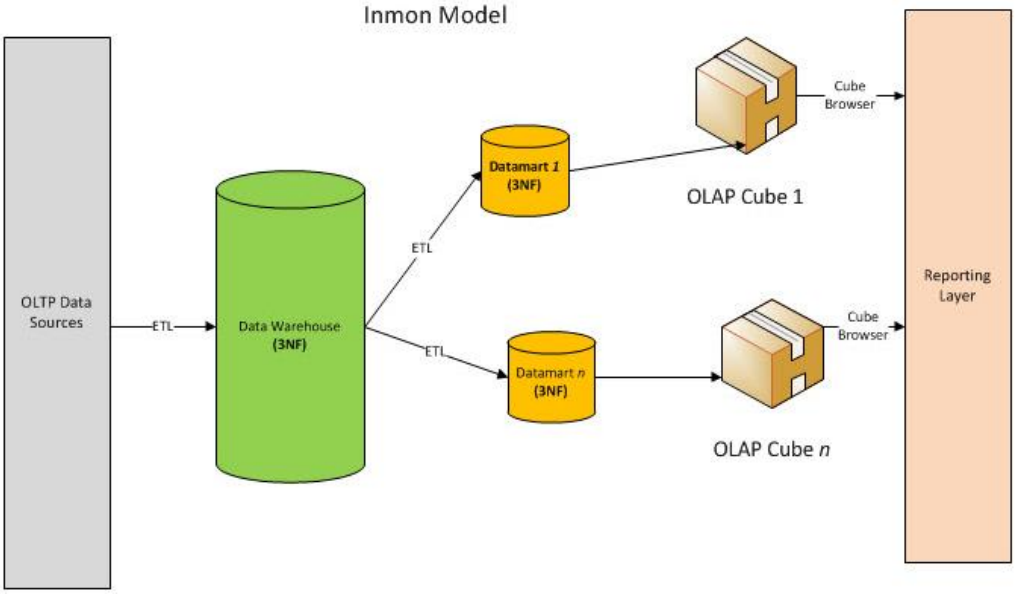
Jak je vidět na Obr. 8, Kimball zvolil přístup zdola nahoru. V této metodě je potřeba nejprve vytvořit tzv. datová tržiště. Tržiště si můžeme pro jednoduchost představit jako podmnožinu tabulek z databáze, která poskytují náhled do datových struktur a dají se podle potřeby kombinovat. Volně přeložená definice datového skladu podle Kimballa

zní: „Datový sklad je kopie transakčních dat, speciálně strukturovaných pro dotazování a analyzování.“(George, 2012).



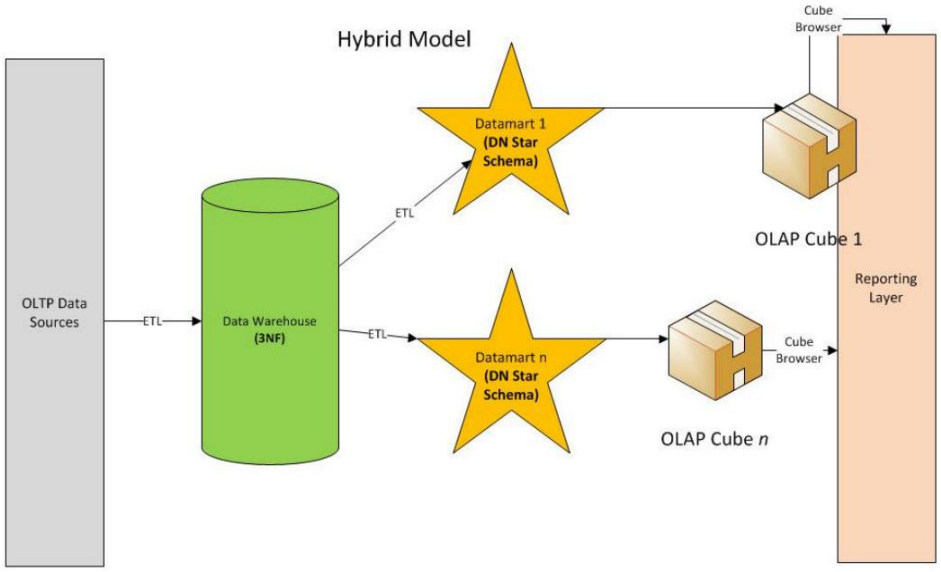
Obr. 8 Schéma řešení datového skladu podle Kimballa.
Zdroj: (Austin, 2010).

Jak je vidět na Obr. 9, druhá varianta, popsaná Inmonem, využívá opačného přístupu, tedy postupu shora dolů. Inmon definuje datový sklad jako centrální uložiště pro celý podnik, který je umístěn ve středu celého řešení a ukládá data do co možná největší úrovně detailů. Datová tržiště jsou vytvořena až po samotném datovém skladu, tedy přesně opačně oproti předchozímu řešení.



Obr. 9 Schéma řešení datového skladu podle Immona.
Zdroj: (Austin, 2010).

Oba tyto modely se opírají o společný fakt, že datový sklad je centrálním zdrojem dat, což bylo demonstrováno v obou předchozích modelech. Dále pak oba autoři využívají OLAP kostek, které jsou vytvářeny v denormalizovaném hvězdicovém schématu. To vede k zamyšlení vytvořit hybridní model, který je průnikem obou přístupů (Austin, 2010), viz Obr. 10.



Obr. 10 Schéma řešení datového skladu hybridním přístupem.
Zdroj: (Austin, 2010).

4.3.5 Dolování dat

Tento pojem vznikl z anglického "data mining", definice tohoto pojmu by se dala volně přeložit jako **umění extrakce užitečných informací** z velkého zdroje dat. Setkáváme se s tím v dnešním světě prakticky denně, od našeho spamového filtru u emailového klienta až po nabídku podobného zboží v internetovém obchodě (Shumeli a kol., 2010, s. 1). Všechny tyto aktivity využívají dolovaní dat a kladou si za cíl co nejrychleji zprostředkovat relevantní informace z rozsáhlého datového zdroje, v našem případě datového skladu.

Protože by zpracování některých dotazů do datového skladu mohlo trvat velmi dlouho, využívá se pomoci OLAP kostky. Pojem OLAP pochází z anglického jazyka "on-line analytical processing" a je to pojmenování pro další vrstvu, která v řešení BI dopomáhá k pružnému a rychlému zpracování dotazů a analýz nad datovým skladem (Solutions, 2002). Vše je založeno na multidimenzionálním datovém modelu, který umožňuje data

zobrazovat ve formě datové kostky, neboli OLAP kostky. Taková kostka pak může být prohlížena **pohledem mnoha dimenzí**, kdy každá z dimenzí je reprezentována svojí dimenzionální tabulkou v datovém skladu. Tento přístup umožňuje uživatelům zobrazovat potřebná data bez znalosti, kde jsou konkrétně uložena v datovém skladu (Sethi, 2012).

Nevýhodou tohoto řešení je tzv. **přepočítávání datové kostky**, ke kterému dochází podle nastavené frekvence nebo nárazově. Přepočet pak může trvat i velmi dlouhou dobu, podle množství dat a výkonu stroje, na kterém k přepočtu dochází. Díky neustálému tlaku ze strany uživatelů na prezenční logiku BI řešení s co možná nejmenšími prodlevami a potřebou aktuálních dat nahrazuje stále častěji řešení pomocí OLAP kostky tzv**. in-memory řešení**. (Lachlan, 2010) Tento postup pracuje přímo s datovým skladem, který je načten do vnitřní paměti, a umožnuje tak zpracovávat dotazy s minimální prodlevou.

Dále se v oblasti dolování dat setkáváme s pojmem **reporting**. Vyjadřuje soubor nástrojů a metod, které mají za úkol data uložená v řešení BI zobrazit ve formě interaktivních reportů[6]. Jak již bylo řečeno v kapitole věnující se současným trendům v BI, budoucnost se nachází v umožnění uživatelům proaktivního sestavování reportů a vytváření vlastních analýz. To však klade také nároky na uživatele v nutnosti znalosti terminologie a postupů s tím spojených.

4.3.6 Prezentace dat

Poslední částí řešení jsou klientské aplikace, které zajišťují komunikaci koncových uživatelů s ostatními komponentami řešení BI, tedy zejména sběr požadavků na analytické operace a následnou prezentaci výsledků (Novotný a kol., 2005, s. 27). Forma prezentace může být různorodá a liší se podle koncového zařízení nebo použité technologie.

[6] Report - sestavený dotaz do komponent řešení BI, který zobrazuje data ve formě informací. (Turban, 2011)

- **Analytické aplikace**

Tyto aplikace jsou navržené speciálně pro **poskytování informací** získaných z předchozích částí BI řešení. Dále zahrnují nástroje umožňující uživateli operace prohlížení dat (drill up, drill down, slice and dice)[7] a identifikaci výjimek (Gála a kol., 2009, s. 230).

- **Systémy EIS – Executive Information Systems**

Systém EIS je část celkového IS, zajišťující mechanismus vícekriteriální analýzy. Tyto nástroje jsou velmi vhodné prostředky pro přehlednou prezentaci dat pomocí grafů a tabulek. Typické pro tyto nástroje je také to, že jsou multidimenzionální, což umožňuje rychlé a jednoduché vytvoření pohledu na data, řazení do nových souvislostí a identifikaci odchylek klíčových ukazatelů od plánovaných hodnot (Novotný a kol., 2005, s. 34).

Jak již bylo zmíněno, uživatelé kladou na prezenční vrstvu stále větší a větší nároky. V roce 2014 nejvýznamnější společnosti na poli BI (SAP, Oracle, IBM, Microsoft, Tableau, GoodData nebo QlikView) představily nová rozhraní a snaží se neustále inovovat přístup k prezentaci dat, tak aby se přizpůsobili moderním trendům a hlavně samotným uživatelům. (Noctuint, 2014)

4.4 Cloudové řešení GoodData

Cloud je bez pochyby budoucnost celého BI. Pokud tedy chce dnes firma uspět na trhu BI technologií, musí s tímto trendem počítat a dát uživatelům to co si žádají.

Společnost Gartner vydává každý rok studii o **postavení IT firem** na daném trhu, **Magický kvadrant (Magic Quadrant).** Studie hodnotí oblast BI podle vlastních hodnotících kritérií, její výsledný graf je vidět na Obr. 11.

Kvadrant je pak rozdělen do těchto čtyř částí: lídři (leaders), vyzyvatelé (challangers), vizionáři (visionaries) a těžko přeložitelná část (niche players), která by se dala vysvětlit

[7] Výčet operací, pro které nemá čeština své překlady a které umožnují v prezentační vrstvě rozdělovat, spojovat a porovnávat data.

jako čtverec určený pro společnosti, které se snaží zaplnit mezeru na trhu. Vidíme zde, že mezi lídry jsou firmy, které se dostatečně zavedly již v jiných odvětvích a že celková poloha všech zmíněných řešení je koncentrovaná do centra diagramu. Takový výsledek bychom mohli interpretovat jako **vyrovnanou situaci** mezi jednotlivými soutěžiteli.

Nás však bude zajímat především společnost GoodData, protože její BI řešení bylo vybráno pro zpracování praktické části této práce. Společnost nakonec skončila ve čtvrté části Magického kvadrantu a řadí se tak nakonec mezi firmy, které se stále ještě nestaly lídry, ale zároveň **poskytují něco, co na trhu chybí** (Sallam a kol., 2015).

Obr. 11 Magický kvadrant pro oblast BI společnosti Gartner pro rok 2015.
Zdroj: (Sallam a kol., 2015).

GoodData je **portfolio nástrojů a knihoven**, které umožňují uživateli vytvořit vlastní funkční BI aplikaci podle svých potřeb. Je tedy kompletně podpořen sběr, uložení, analýza a vizualizace dat. Celé řešení se pak nachází v cloudu jako koncový produkt. Díky otevřenému API (Application Programming Interface) je možné integrovat nové datové zdroje, měnit datové transformace nebo např. aktualizovat datový model. Oproti

tradičním BI řešením GoodData nevyužívá OLAP kostek, ale pro zvýšení výkonu multi-level caching[8].

K roku 2014 spravovala firma 30 tisíc datových skladů a řadí se tak mezi nejvýznačnější poskytovatele BI řešení na světovém trhu. Výhodou celého řešení je rychlost realizace, protože zákazníku odpadá nutnost vytvořit prostředí pro běh datového skladu. Toho je dosaženo propojením jednotlivých komponent a uložením dat v cloudu. Potvrzuje to i výrok firmy, který ve volném překladu zní: „To, co ostatní poskytovatelé slibují, že bude hotovo do několika měsíců, s GoodData se dá vytvořit za několik dní.“ (GoodData, 2015a).

4.5 Metodika výběru nástroje

Pro výběr využijeme multikriteriální hodnocení variant, konkrétně metodu založenou na párovém srovnání variant neboli Analytický hierarchický proces (AHP).

4.5.1 Analytický hierarchický proces (AHP)

Tato metoda vícekriteriálního rozhodování by se, volně přeloženo, dala definovat, podle jejího autora T. L. Saatyho jako teorie relativního měření na absolutních stupnicích hmotného i nehmotného kritéria, založených na úsudku informovaných odborníků s využitím stávajících měření a statistických údajů potřebných k rozhodnutí (Saaty, 2005).

AHP model dopomáhá členit řešený problém do hierarchií a porovnávat mezi sebou vždy pouze dva prvky. Hodnocení je potom syntéza výsledků na základě vah kritérií a dílčích hodnocení alternativ.

[8] Jedná se o moderní přístup načítání velkých množství objektů do paměti v několika úrovních. Tím je umožněno se přiblížit svojí rychlostí odezvě, kterou známe např. při využití CPU v našem počítači (Gill, 2008).

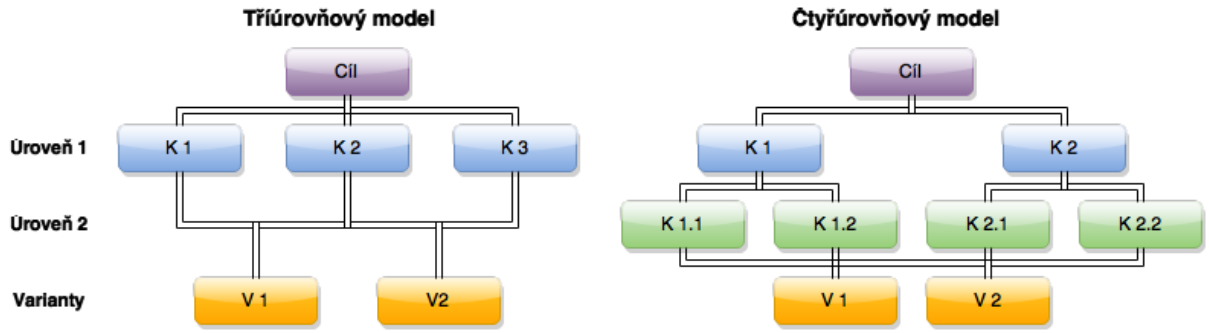
Celý proces se skládá z několika na sobě závislých fází:

- **Určení a formulace cíle**
- **Určení souboru kritérií**
- **Výběr variant řešících stanovený problém**
- **Syntéza výsledků a výběr nejvhodnější varianty**

4.5.2 Metoda stanovení kritérií a jejich vah

Základním parametrem modelu jsou kritéria hodnocení. Ta volíme tak, aby odrážela cíl, který si hodnotitel určil, a zároveň musíme zabezpečit, aby při větším počtu kritérií nebyla mezi sebou vzájemně redundantní.

Kritéria mohou být rozdělena do více úrovní nebo mohou být pouze v jedné úrovni, takové rozdělení je patrné na Obr. 12, kde naleznete porovnání tří a čtyřúrovňového modelu AHP. Rozdíl je v použití jedné úrovně kritérií, navíc u čtyřúrovňového modelu, která uskupuje kritéria na základě jejich věcné podobnosti a příbuznosti do vlastních skupin. Tímto způsobem je možné pokračovat a přidávat úrovně.



Obr. 12 Porovnání AHP modelů.
Zdroj: Vlastní práce autora.

K vytvoření takového modelu můžeme dojít několika způsoby. Například postup shora dolů, kdy nejprve vytvoříme skupiny kritérií, na jejichž základě hledáme konkrétní kritéria. Nebo postup opačný, kdy vytvoříme soubor konkrétních kritérií, která se poté snažíme sloučit do logických prvků, tedy skupin.

Po rozdělení kritérií do skupin je zapotřebí stanovit jejich váhy. K tomu se používá podle Saatyho (Saaty, 2005) metoda párového srovnání, která je založena na principu srovnání každého kritéria s každým. Jedná se o zjištění preferencí vzhledem ke všem ostatním

kritériím obsaženým v souboru. Pro vyjádření velikosti preferencí Saaty doporučuje použití devítibodové stupnice, viz volně přeložená Tab. 2.

Tabulka 2 Saatyho devítibodová stupnice.

Počet bodů	Porovnání prvků A a B (A je … než B)
1	**stejně** významný (equal importance)
3	**mírně** významnější (moderate importance)
5	**silně** významnější (strong importance)
7	**velmi silně** významnější (very strong importance)
9	**extrémně** významnější (extreme importance)

Zdroj: (Saaty, 1977).

Stupnice je základním kamenem celé metody, protože ta se právě díky tomuto přístupu odlišuje od ostatních metod vícekriteriálního rozhodování.

Saatyho matice

Máme skupinu *k* prvků, která je homogenní (aby bylo možné prvky mezi sebou srovnávat) a ve které hodnotitel porovnává každou dvojici vzhledem k nadřazenému prvku za použití Saatyho stupnice. Vniká matice párových srovnání, kterou označíme písmenem ***S***. Její prvky $\mathbf{s_{ij}}$ jsou poměrem důležitosti prvků *i* a *j* vzhledem k nadřazenému prvku (Saaty, 1987). Platí tyto vztahy:

$$s_{ij} \cong \frac{v_i}{v_j}, \quad i = 1,2,\dots,k;\ j = 1,2,\dots,k$$

$$s_{ij} = 1, \quad i = 1,2,\dots,k$$

$$s_{ij} = \frac{1}{s_{ij}}, \quad i = 1,2,\dots,k;\ j = 1,2,\dots,k$$

kde *k* je počet porovnávaných prvků (řád matice).

Sestavená Saatyho matice:

$$S = \begin{pmatrix} 1 & S_{12} & \cdots & S_{1j} \\ \frac{1}{S_{12}} & 1 & & S_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{S_{1j}} & \frac{1}{S_{2j}} & \cdots & 1 \end{pmatrix}$$

Z takto sestavené matice pak pomocí vlastního vektoru matice lze normalizováním spočítat vektor priorit prvků vzhledem k jejich nadřazenému prvku. Protože by náročnost hledání vlastních čísel matice, především u matic vyššího řádu, byla náročnější, Saaty popisuje jako jednu z možných metod výpočet pomocí normalizovaného geometrického průměru řádků (Saaty, 1977):

$$s_i = \prod_{j=1}^{k} s_{ij} \ , \qquad r_i = \sqrt[k]{s_i} \ , \qquad w_i = \frac{r_i}{\sum_{i=1}^{k} r_i}$$

Pro výpočet geometrického průměru nejprve provedeme mezikroky **s** a **r**. Pomocí těchto kroků se dostaneme k vektoru **w**. Jedná se o normalizovaný vektor vah, který určuje vliv jednotlivých kritérií ve vztahu k nadřazenému prvku. Tímto způsobem dostáváme Tab. 3, která bude využita pro ohodnocení vah kritérií modelu v praktické části práce.

Tabulka 3 Výpočet vah kritérií.

Kritérium	**K_1**	**K_2**	**...**	**K_j**	s_i	r_i	w_i
K_1	1	s_{12}	...	s_{1j}	$\prod_{j=1}^{k} s_{1j}$	$\sqrt[k]{s_1}$	$\frac{r_1}{\sum_{i=1}^{k} r_1}$
K_2	$\frac{1}{s_{12}}$	1	...	s_{2k}	$\prod_{j=1}^{k} s_{2j}$	$\sqrt[k]{s_2}$	$\frac{r_2}{\sum_{i=1}^{k} r_2}$
...	...	...	⋱	...	...	...	..
K_i	$\frac{1}{s_{1j}}$	$\frac{1}{s_{2j}}$	...	1	$\prod_{j=1}^{k} s_{ij}$	$\sqrt[k]{s_i}$	$\frac{r_i}{\sum_{i=1}^{k} r_i}$

Zdroj: Vlastní práce autora podle článku (Saaty, 1977).

Nejprve vyřešíme část matice vpravo od diagonály. Pokud je kritérium v řádku důležitější než to ve sloupci, zapíšeme příslušný počet bodů podle stupnice. Pokud je kritérium v řádku méně významné, tak se vyplní jeho převrácená hodnota. Levou část vyplníme pomocí převrácených hodnot z pravé části. Na diagonále vyplníme samé jedničky, protože při porovnání kritérií sama se sebou jsou stejně významná (Saaty, 1977). Dále dopočítáme hodnoty a určíme váhy kritérií.

Konzistence párových srovnání

Jak uvádí Saaty (Saaty, 1987), při metodě párového srovnání nesmíme opomenout důležitý faktor, a to je konzistence modelu. Vzhledem k tomu, že používáme vlastní úsudek, tak prakticky vždy dochází k určité míře nekonzistence.

V praxi to znamená, že pokud máme prvek A a tvrdíme, že je 5-krát lepší než prvek B, a stejně tak tvrdíme, že prvek B je 5-krát lepší než prvek C, pak, abychom předešli nekonzistenci, musí být prvek A 25-krát lepší než prvek C.

Pro potřeby ověření nekonzistence modelu použijeme program Expert Choice, který disponuje funkcí výpočtu hodnoty konzistence modelu, a ověříme tak, že model neztrácí výpovědní hodnotu. Saaty (Saaty, 1987) uvádí, že pokud je míra nekonzistence menší nebo rovna 0,1, je model brán jako dostatečně konzistentní. Pokud bychom totiž tento proces neprovedli a míra nekonzistence by byla vysoká, model by neměl požadovanou výpovědní hodnotu.

5 Popis výzkumu

V následující kapitole se budeme zabývat samotným výběrem vhodného ETL nástroje. Vytvoříme AHP model, jehož váhy ohodnotíme pomocí párového srovnávání. Následně popíšeme praktické testování ETL nástrojů a na jeho základě ohodnotíme předem připravený model.

5.1 Vytvoření AHP modelu

Fáze tvorby modelu jsme si již vyjmenovali v předešlé kapitole, na to nyní navážeme a v pořadí těchto fází tvorbu popíšeme.

5.1.1 Určení a formulace cíle

V první fázi celého postupu určíme cíl, kterým je **výběr vhodného ETL nástroje** pro nahrání souboru zdrojových dat s ohledem na platformu GoodData.

5.1.2 Určení souboru kritérií

Pro určení souboru kritérií byla vybrána čtyř úrovňová hierarchie. Bylo postupováno podle popsané metody shora dolů. Vytvořeny byly nejprve tři základní skupiny kritérií, a následně umístěny další kritéria do těchto skupin:

Funkční kritéria

- **Náročnost použití** – hodnocení a porovnání náročnosti dosažení cíle s daným nástrojem. Cílem je pak myšleno nahrání dat do platformy GD.
- **Funkcionalita** – hodnocení komponent použitých pro splnění úkolu, jejich propracovanost a práce s nimi.
- **Odstraňování chyb** – hodnocení připravenosti nástroje na řešení chyb, tzv. debugging a jejich odstraňování.
- **Řešení práv** – přístup více uživatelů, možnost omezení přístupu k určitým částem řešení a jejich ochrana.

- **Přívětivost uživatelského rozhraní** – náročnost zorientování v daném prostředí a celkový uživatelský dojem z rozhraní.

Technická kritéria

- **Architektura** – vnitřní architektura nástroje a podpora paralelních procesů pro zrychlení nahrávání.
- **Podpora formátů** – porovnání pestrosti možných nahrávaných formátů zdrojových souborů.
- **Příprava dat a čištění** – podpora filtrovaní dat, odstraňování nekonzistencí vstupních dat, jejich převádění a doplňování.
- **Rychlost nahrání** – porovnání rychlosti nahrávání datových zdrojů do platformy GoodData.
- **Konektivita** – podpora přímého napojení dalších aplikací a služeb.

Obecná kritéria

- **Možnost opětovného použití** – znovupoužití celého projektu jako řešení nebo jeho části při tvorbě datových pump.
- **Bezpečnost a stabilita** – celková stabilita programu, ochrana dat a procesů.
- **Cena** – porovnání cen jednotlivých řešení.
- **Rozšiřitelnost** – podpora propojení s dalšími aplikacemi a službami.
- **Podpora** – hodnoceno množství dostupných materiálů, online podpora a zabudovaná nápověda v nástroji

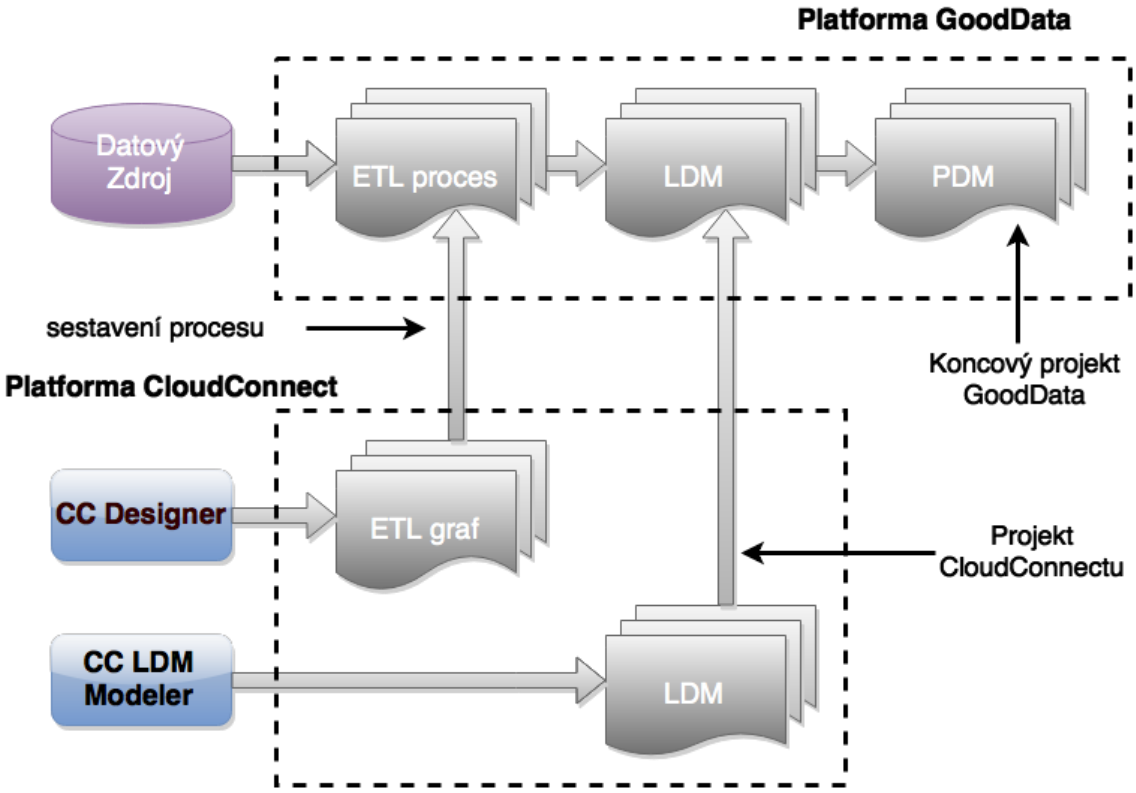
5.1.3 Výběr variant pro stanovený rozhodovací problém

Po diskuzi s odborným konzultantem ze společnosti GD, výběr variant pro stanovený rozhodovací problém se omezil pouze na dvě možnosti, **CloudConnect** a **Keboola Connection**. Oba tyto nástroje si před samotnou implementací představíme.

CloudConnect (CC)

Tento nástroj byl speciálně navržen, aby vytvořil propojení mezi daty a řešením GoodData. Mezi jeho hlavní úkoly patří: **umožnění centrálního přístupu** ke správě více připojených projektů, **paralelní spouštění procesů** a **plánování nebo sledování jejich běhů** (GoodData, 2012). Celé řešení je pak rozděleno na dvě části, CC LDM Modeler a CC Designer. LDM Modeler umožňuje uživateli vytvářet vlastní logický datový model (LDM) pro jednotlivá řešení. CC Designer je ETL nástroj, který funguje jako desktopová aplikace vytvořená v programovacím jazyce Java. Jde o upravenou open source verzi programu CloverETL. Tvůrci GD je pak doporučován jako jednoduchý nástroj s mnoha předdefinovanými komponentami a příjemným grafickým rozhraním (GoodData, 2015a).

Pro snazší pochopení popisované architektury nám poslouží schéma řešení na Obr. 13, vypracované podle zdrojů z dokumentace k tomuto produktu. Můžeme si všimnout rozdílu mezi oběma platformami a způsobu zprostředkování jejich propojení.



Obr. 13 Propojení platforem GoodData a CloudConnect.
Zdroj: Vlastní práce autora podle článku (GoodData, 2015a).

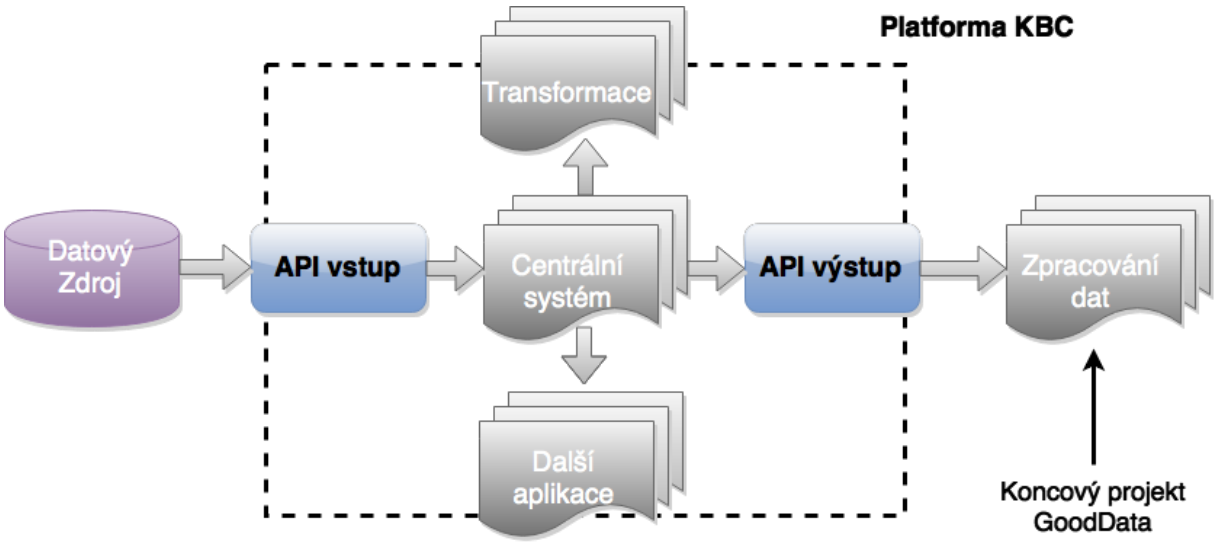
Ze schématu je také na první pohled patrné, že CC Designer a CC LDM Modeler mají rozdílné úlohy. Zatímco Designer se zabývá úlohami klasického ETL, Modeler má za úkol **vytvořit logický datový model**. Ten je nahrán do platformy GD a tím umožní Modeleru se namapovat na předpřipravenou datovou strukturu.

Firma také uvádí, že nezáleží, jestli jsou data uložena v SaaS aplikaci, nebo v lokálním prostředí, tento nástroj si dokáže poradit s oběma přístupy. To je zřejmé i z Obr. 13, kde datový zdroj stojí mimo obě platformy a nahrávání zdrojů se provádí až při spuštění procesu ETL.

Keboola Connection (KBC)

KBC je méně známý ETL nástroj české společnosti Keboola, jehož hlavní komponentou je uložiště dat, které umožňuje nahrávání souborů přímo do cloudových serverů společnosti Keboola (Keboola, 2014).

Tento server je pak možné pomocí vlastního rozhraní propojit s projektem GD a vytvořit aplikaci, která se do tohoto serveru na data dotazuje. Pro ilustraci takového řešení je na Obr. 14 přiložené schéma celé platformy.



Obr. 14 Schéma platformy KBC.
Zdroj: Vlastní práce autora podle článku (Keboola, 2014).

V levé části schématu vidíme datový zdroj, pod kterým si můžeme představit širokou škálu možných zdrojů dat, např. data z GA[9], Facebooku, on-premise databází, IoT[10]zařízení, až po různé analytické aplikace. Ta jsou nejprve nahrána do platformy KBC, kde jsou extrahována na vstupu, a teprve poté mohou být podrobena transformacím nebo dalším operacím.

Centrální systém pak obstarává ukládání dat a přístup k nim. Výstupem jsou tzv. writters, česky bychom tyto komponenty mohli nazývat zapisovače. Jejich úkolem je dopravit data z KBC do aplikace, která tato data dále zpracovává, v našem případě je to projekt GoodData.

5.1.4 Vlastní AHP model

Po definici kritérií a alternativ nám nic nebrání v sestavení vlastního AHP modelu, který bude vycházet z výše zmíněných poznatků. Aby byl model kompletní, schází nám již jen **určit váhy** jednotlivých kritérií a jejich skupin. Použijeme popisovanou metodu párového srovnání a uspořádaný model ohodnotíme. Hlavní cíl je ohodnocen vahou 1. Tuto váhu rozdělíme pomocí devítibodové stupnice pro vyjádření velikosti preferencí mezi skupiny kritérií na první úrovni. Stejně budeme postupovat i u další úrovně. V Tab. 4 je pro přehlednost uvedeno pouze shrnutí výsledků, detailní tabulky hodnocení jsou uveřejněny na konci práce v kapitole Přílohy.

[9] Google Analytics – analytický nástroj od společnosti Google, který umožňuje uživateli získávat statistická data o provozu na jeho webových stránkách.

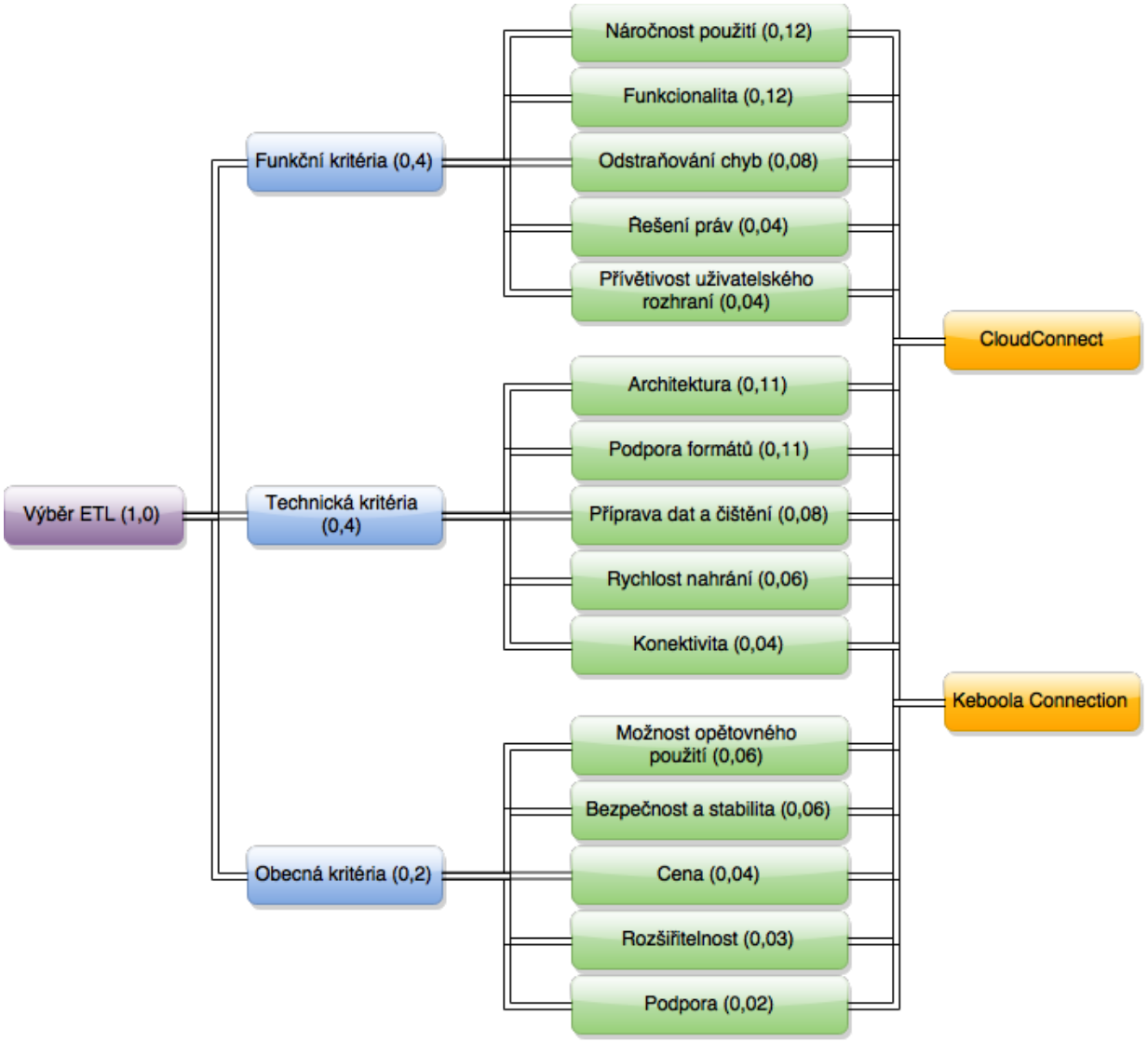
[10] Internet of Things – Jedná se o vestavěná zařízení, která jsou napojena na internet a přináší tak možnost interakce mezi různými systémy. (Evans, 2011)

Tabulka 4 Shrnutí výsledku párového hodnocení kritérií a jejich skupin.

Skupina kritérií	Váha skupin kritérií	Název kritéria	Váha ve skupině
Funkční kritéria	0.40	Náročnost použití	0.31
		Funkcionalita	0.31
		Odstraňování chyb	0.20
		Řešení práv	0.09
		Přívětivost uživatelské rozhraní	0.09
Technická kritéria	0.40	Architektura	0.28
		Podpora formátů	0.28
		Příprava dat a čištění	0.19
		Rychlost nahrání	0.16
		Konektivita	0.09
Obecná kritéria	0.20	Možnost opětovného použití	0.28
		Bezpečnost a stabilita	0.28
		Cena	0.18
		Rozšiřitelnost	0.17
		Podpora	0.08

Zdroj: Vlastní práce autora.

Vhodné bude také váhy kritérií ve skupinách převést na váhy napříč skupinami. Toho docílíme znormováním vah kritérií podle postupu zmíněného v kapitole 4.5.2 Metoda stanovení kritérií a jejich vah. Kompletní grafické znázornění modelu po znormování vah kritérií je vidět na Obr. 15.



Obr. 15 Model AHP po ohodnocení a znormování vah jednotlivých kritérií.
Zdroj: Vlastní práce autora.

5.2 Praktické testování ETL nástrojů

Pro lepší demonstraci a pozorování chování jednotlivých ETL nástrojů budou tyto nástroje podrobeny praktické úloze. Ta vznikla jako požadavek společnosti 1188[11] při události Enterprise Data Hackathon v Praze v roce 2014. Pomocí obou vybraných nástrojů importujeme data do projektu GD a odpovíme na několik otázek, které firma

[11] 1188, Informační linky s.r.o., sídlo firmy: Praha 1 – Nové Město, Opletalova 1015/55, předmět podnikání: Poskytovaní informačních a asistenčních služeb

1188 vznesla. Našim cílem tedy není odpovědět na všechny otázky společnosti a prezentovat její data, ale praktické testování nástrojů.

5.2.1 Datový zdroj a praktická úloha

Na výše zmíněné události v roce 2014 poskytla společnost 1188 anonymizované[12] datové zdroje. Soubory ve formátu CSV jsou upravené výsledky činnosti informační linky, které obsahují mnohé informace, ale bez potřebné analýzy a transformace jsou jen s velkými obtížemi čitelné. Na dalších řádcích je přiložena Tabulka 5, která obsahuje seznam zdrojových soborů, včetně jejich analýzy.

Tabulka 5 Seznam zdrojových souborů, včetně jejich analýzy.

Název souboru	Stručný popis obsahu a struktury souboru
ope_calls.csv	Tento soubor obsahuje celkem deset atributů, z nichž pro analýzu významné jsou: *ID* (jednoznačný identifikátor záznamu), *phone number* (z bezpečnostních důvodů zašifrované telefonní číslo zákazníka), *operator* (jednoznačný identifikátor operátora, který hovor obsluhoval), přesné časy: *answered_at, created_at, updated_at* (jedná se o začátek hovoru a v mnoha případech se časy nijak nelišší), ostatní sloupce jsou prázdné.
ope_wub_logs.csv	Soubor obsahuje těchto šest atributů: *ID* (jednoznačný identifikátor záznamu), *ope_call_id* (vnitřní cizí klíč ve vztahu 1:1 k záznamům v souboru **ope_calls.csv**), *ope_wup_question_id* (jednoznačný identifikátor typu otázky, který vytváří napojení na soubor **out.c-production.wup-questions.csv**),

[12] Anonymizace proběhla z důvodů ochrany osobních údajů. Telefonní čísla a další citlivé údaje byly firmou pro potřebu události zašifrovány.

	ope_wup_answer_id (jednoznačný identifikátor typu odpovědi vytvářející napojení na **out.c-production.wup-answers.csv**), *created_at* (datum a čas vytvoření záznamu), *updated_at* (datum a čas poslední úpravy záznamu, který je ve většině případů shodný s atributem vytvoření)
out.c-production.hovory.csv	Tento soubor obsahuje celkem 14 atributů, ale bohužel velká část z nich neobsahuje data. Pro další práci budou tedy použitelné pouze: *primary_key* (jednoznačný identifikátor záznamu), *disposition* (způsob přijetí hovoru, název kategorie slovy), *agent_id* (jednoznačný identifikátor agenta, který obsloužil hovor), *delay* (doba čekání v sekundách na přijetí hovoru), *time_block* (datum a čas hovoru), *handling* (délka trvání hovoru v sekundách), *originator* (zašifrované telefonní číslo volajícího) a *service_id* (identifikátor kategorie služeb, které se hovor týkal),
out.c-production.operatori.csv	Jedná se o seznam, který je z bezpečnostních důvodů anonymizovaný. Soubor obsahuje tedy pouze dva atributy: *ID* (jednoznačný identifikátor operátora), *name* (anonymizované jméno operátora)
out.c-production.tarifikace.csv	Tento soubor obsahuje celkem 21 atributů. Bohužel kvalita dat a napojení na ostatní soubory je špatná. Atribut *call_id*, který je primárním klíčem celého souboru obsahuje záznamy, které se nedají s ostatními soubory propojit, protože mají rozdílné délky primárních klíčů. Informace slouží k zjištění délky hovorů a jejich tarifikaci.

out.c-production.tv-spoty.csv	Tento soubor obsahuje 13 atributů, které souvisí s reklamou v televizi. *Primary_key* (jednoznačný identifikátor záznamu), *date_time_start* (datum a čas začátku běhu reklamy), *date_time_end* (datum a čas konce běhu reklamy), *footage* (délka reklamního spotu v sekundách), *day_of_spot* (den v týdnu, kdy běžela reklama), *prime_time* (údaj, jestli se jedná o hlavní vysílací čas), *station* (název televizní stanice), *program_before* (název programu, který běžel před reklamou), *program_after* (název programu, který běžel za reklamou) a další údaje o charakteru televizního spotu.
out.c-production.vnejsi-prepojeni.csv	Jedná se o soubor se sedmi atributy. *Call_id* (jednoznačný identifikátor), *length* (délka v sekundách) a další. Soubor dokresluje historii vývoje hovoru, tedy přináší provozní informace o přepojení.
out.c-production.vnitrni-prepojeni.csv	Tento soubor obsahuje 14 atributů. Svou strukturou a významem je velmi podobný předchozímu souboru **out.c-production.vnejsi-prepojeni.csv**.
out.c-production.wup-answers.csv	Seznam kategorií odpovědí na zákazníkovi otázky. Obsahuje atributy: *id* (jednoznačný identifikátor), *title* (název kategorie)
out.c-production.wup-questions.csv	Seznam kategorií otázek zákazníka. Obsahuje atributy: *id* (jednoznačný identifikátor), *title* (název kategorie)

Zdroj: Vlastní práce autora podle společnosti 1188.

Tato analýza poukázala na to, že datové zdroje mají mezi sebou jen málo napojení, a tak se předpokládá, že logický datový model nebude tolik konzistentní. Dalším zjištěním bylo také to, že obsahují sloupce bez hodnot, několik sloupců, které jsou pouze nepotřebné logovací informace firemních systémů, a atributy, které je nutné transformovat.

Jelikož je naším cílem zkoumat především ETL nástroje, pokusíme se vytěžit z těchto dat co možná nejvíce užitečných informací tak, aby byl celý výzkum přínosem i pro společnost 1188. Ta vznesla na zmíněné události několik okruhů otázek. Budeme hledat odpovědi na témata:

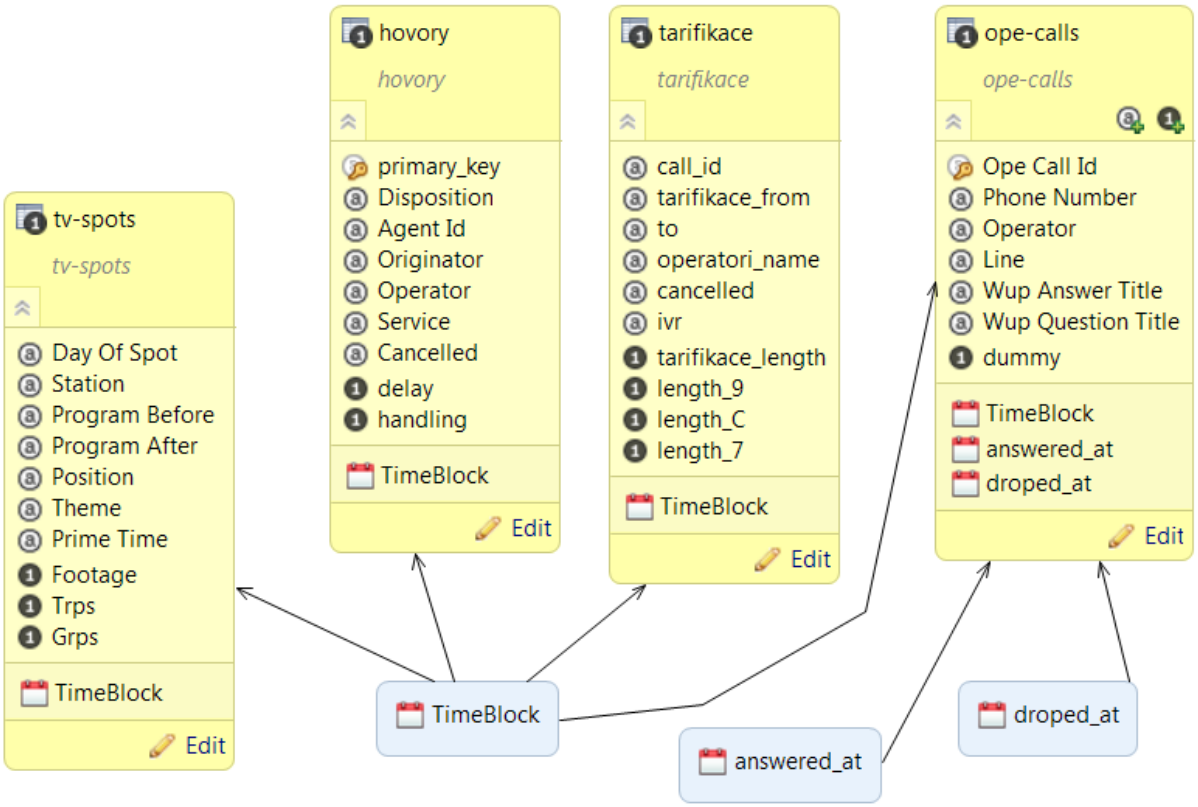
1. **Chování zákazníků podle dní v týdnu.**

2. **Hodnocení agentů podle dostupných ukazatelů.**

3. **Vliv TV spotů na provoz linky.**

5.2.2 Logický datový model

Po definování okruhů otázek jsou vybrány zdrojové soubory, které budou potřeba pro jejich zodpovězení. Na Obr. 16 je vlastní logický datový model pro toto řešení. Jedná se o čtyři faktické tabulky propojené mezi sebou časovou dimenzí. Tabulka *ope-calls* byla doplněna o prázdný sloupec *dummy* tak, aby se chovala také jako faktická a nebyla zařazena v aplikaci GD mezi dimenze. Můžeme si povšimnout, že došlo také ke sjednocení u tabulky *ope-calls*, a to z důvodů rychlejšího nahrávání a zjednodušení řešení, což bude podrobněji popsáno v následující kapitole. Model byl nahrán do platformy GD jako nezbytná součást řešení pomocí nástroje CC.



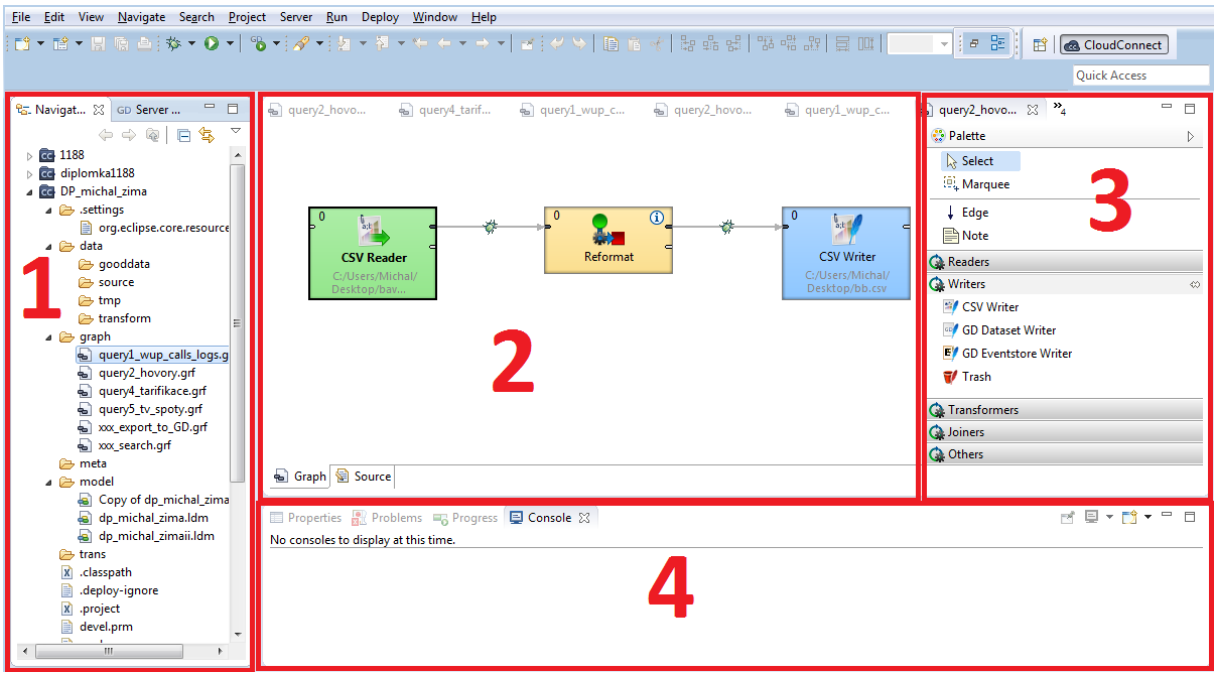
Obr. 16 Logický datový model pro řešení otázek společnosti 1188.
Zdroj: Vlastní práce autora (prostředí CC).

5.2.3 Popis implementace úlohy v nástroji CloudConnect

Nástroj CloudConnect je desktopový program a jeho prostředí působí díky své architektuře velmi povědomě. Nejprve si představíme základní obrazovku nástroje, rozdělenou do několika částí, viz Obr. 17.

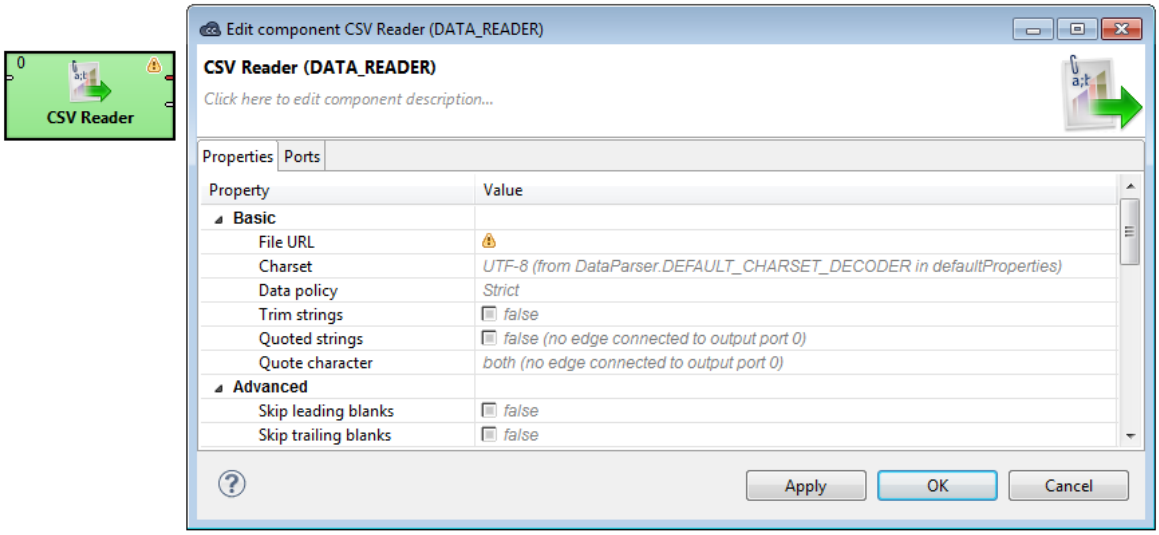
1. **Strom řešení** – Po otevření nového projektu se nám vytvoří struktura řešení, ve které se nachází např. složka pro logický datový model (model) nebo složka pro jednotlivé datové pumpy (graph).
2. **Okno řešení** – V centru obrazovky se nachází prostor pro otevření konkrétní zpracovávané úlohy. V našem případě je to datová pumpa.
3. **Nabídka komponent** – Nabídka nástrojů, tzv. toolbar, obsahuje všechny potřebné komponenty pro vytvoření datové pumpy. Pro přehlednost je rozdělena podle funkčnosti na Readers (čtení dat), Writers (zapisování dat), Transformers (transformace dat), Joiners (propojování dat) a Others (ostatní funkce).

4. **Konzole** – V této části se nachází prostor pro vypisování zpráv o běhu programu, včetně stavů a chybových hlášení.



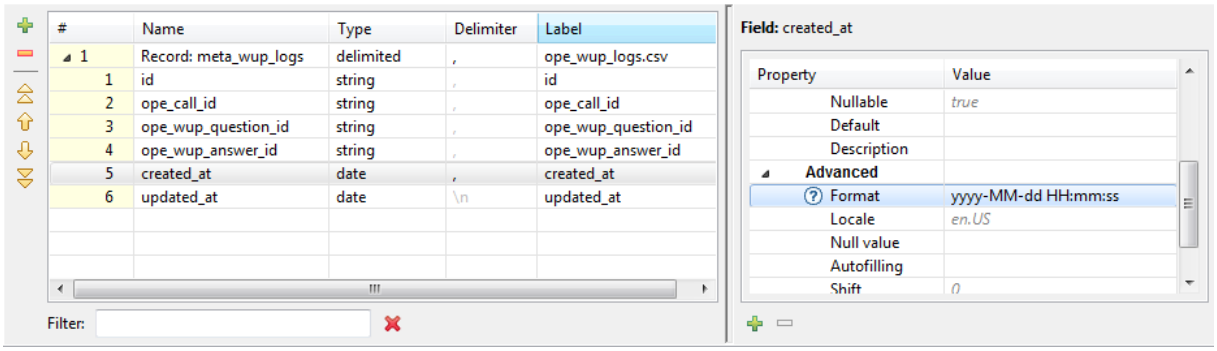
Obr. 17 Základní obrazovka nástroje CloudConnect, rozdělená na jednotlivé části.
Zdroj: Vlastní práce autora (prostředí CC).

Po vytvoření projektu a logického datového modelu můžeme začít se samotným sestavováním ETL procesu. Začneme s novým grafem a přidáme do něj komponentu CSV Reader, která slouží ke **čtení dat** ze souborů typu CSV, viz Obr. 18. Vybereme cestu k souboru (File URL) a nastavíme kódování řetězce znaků. V našem případě se jedná o způsob kódování UTF-8.



Obr. 18 Komponenta CSV Reader a její nastavení.
Zdroj: Prostředí CC.

Abychom mohli se zdrojem dále pracovat, je potřeba nastavit správně metadata. V okně na Obr. 19 zvolíme nejprve oddělovač sloupců (delimiter). To je v našem případě čárka, která je použita i u ostatních CSV souborů. Následně **definujeme datové typy** jednotlivých sloupců, a kde je to nutné, tak i jejich formát.

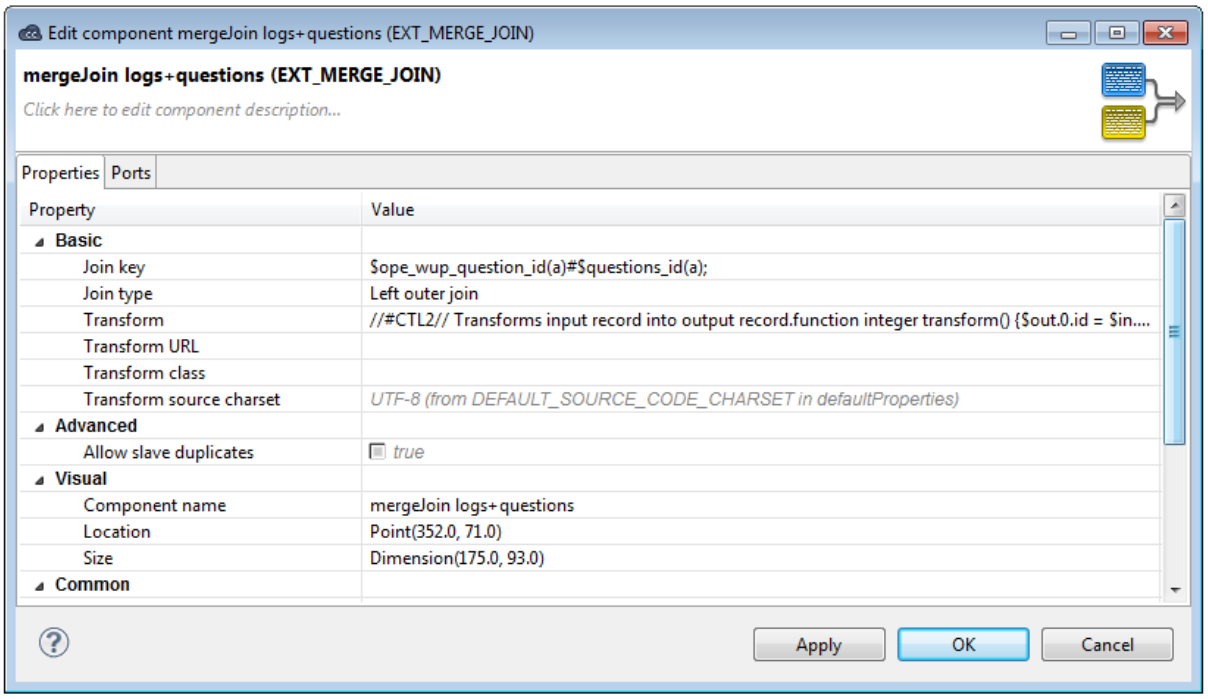


Obr. 19 Nastavení metadat pro CSV soubor.
Zdroj: Vlastní práce autora (prostředí CC).

Po takto připraveném souboru přichází na řadu transformace. Popíšeme si z celého řešení tu nejsložitější, a to propojení **ope_calls.csv, ope_wub_logs.csv, wup-answers.csv** a **wup-questions.csv** do jednoho datového uskupení.

K tomu bude potřeba pro každý soubor z nabídky komponent vybrat tzv. ExtSort, který se postará o seřazení podle určitého sloupce. To je pravidlo, které si vyžaduje komponenta ExtMergeJoin, je nutné, aby napojované zdroje, byly **řazeny podle řídících sloupců**. Celý krok není obtížný a stačí pouze nastavit klíč pro řazení (sort key)

z nabídky sloupců. Po těchto úpravách přidáme do grafu zmíněnou komponentu ExtMergeJoin a nastavíme ji podle Obr. 20. Začneme s parametrem Join key, tedy klíč ke spojení, a poté vybereme typ spojení. Zde jsou v nabídce klasické typy, jak je známe z jazyka SQL (Inner join, Left outer join a Full outer join). V našem případě použijeme vnější napojení zleva a doplníme tak **ope_wub_logs.csv** o názvy otázek z **wup-questions.csv.**



Obr. 20: Nastavení komponenty ExtMergeJoin.
Zdroj: Vlastní práce autora (prostředí CC).

Dalším krokem je vytvoření transformace (Transform), kde je zapotřebí pomocí skriptovacího jazyka CTL2[13] sestavit kód, který bude sloužit jako instrukce pro komponentu. Zajímavostí je, že tento kód je možné přeložit také do jazyka Java. Na následujících řádcích je uveden kód pro zmíněnou transformaci.

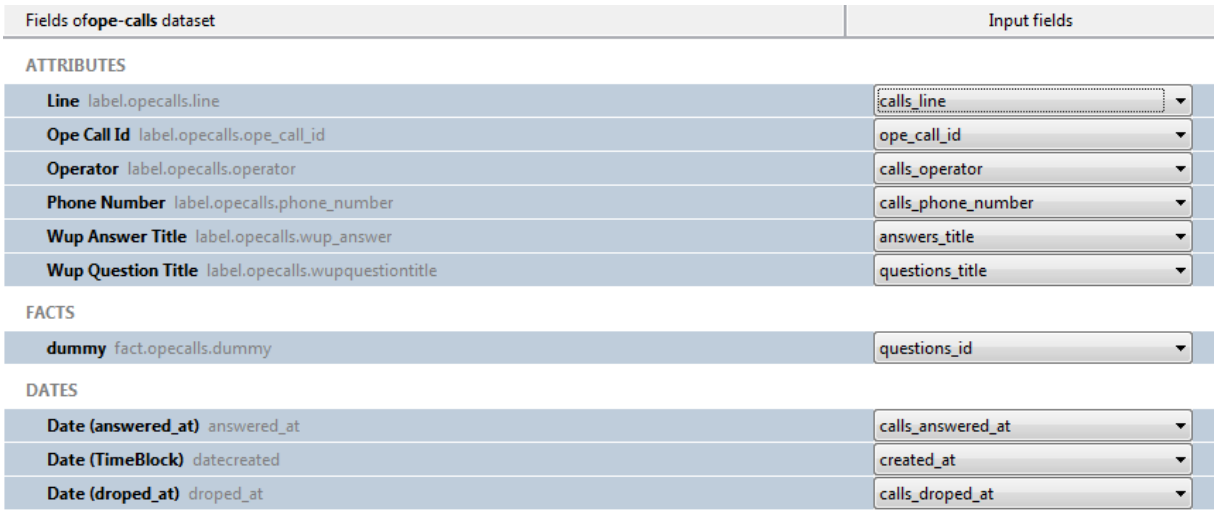
```
function integer transform() {
	$out.0.id = $in.0.id;
	$out.0.ope_call_id = $in.0.ope_call_id;
	$out.0.ope_wup_question_id = $in.0.ope_wup_question_id;
```

[13] Jedná se skriptovací jazyk vyvinutý společností CloverETL pro vytváření transformací. Více o tomto jazyku je možné dohledat v jeho dokumentaci, dostupné na webu: doc.cloveretl.com.

```
        $out.0.ope_wup_answer_id = $in.0.ope_wup_answer_id;
        $out.0.created_at = $in.0.created_at;
        $out.0.updated_at = $in.0.updated_at;
        $out.0.questions_id = $in.1.questions_id;
        $out.0.questions_title = $in.1.questions_title;

        return ALL;
}
```

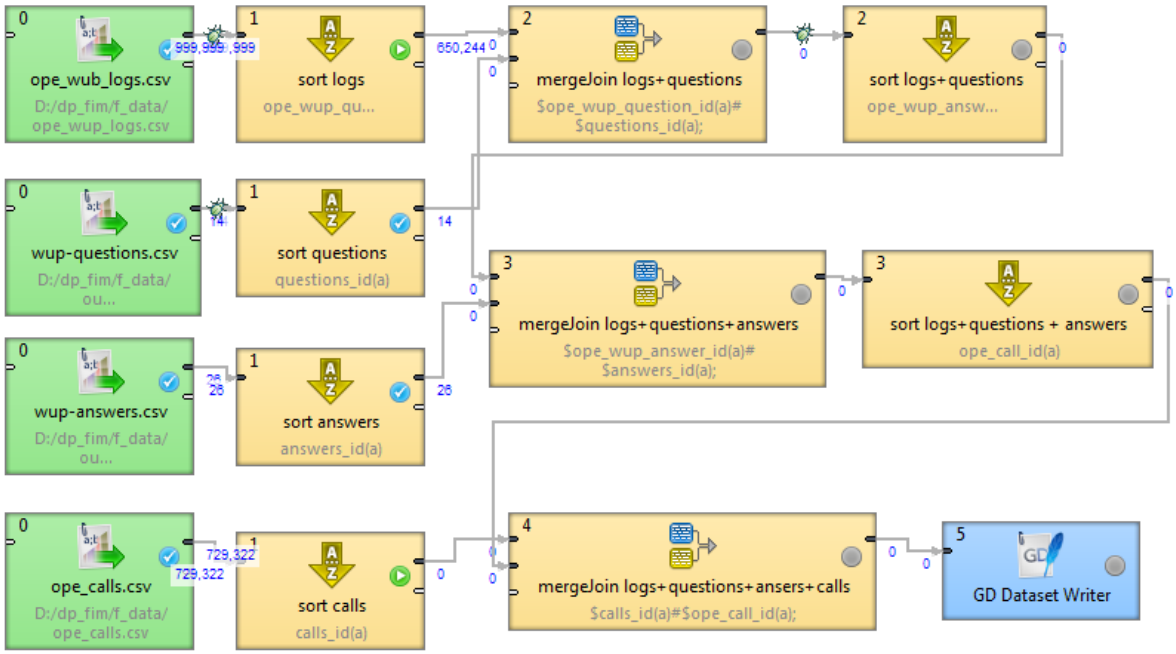
Tímto způsobem budeme pokračovat i u dalších souborů a postupně propojíme všechny zdrojové soubory do jednoho toku dat.

K dokončení celého procesu bude potřeba komponenta GD Dataset Writer, která umožňuje **nahrání transformovaných dat** do platformy GoodData. V průvodci nastavení vyplníme nejdříve jednoznačný identifikátor projektu GD, na jehož základě se vyplní nabídka datových setů (faktické tabulky z nahraného datového modelu), a vybereme ten, který odpovídá nahrávaným datům. Následně zvolíme jeden ze způsobů nahrávání do platformy GD, plné nahrání (Full load) nebo přírůstkové (Incremental). Jako poslední krok nastavíme v komponentě propojení mezi zdrojovými sloupci a cílovými, viz Obr. 21.



Obr. 21 Nastavení napojení datového toku a komponenty GD Dataset Writer.
Zdroj: Vlastní práce autora (prostředí CC).

Takto jsme vytvořili celý proces nahrání, který si můžeme prohlédnout na Obr. 22. Celá datová pumpa je rozdělena do několika fází, které jsou znázorněny číselnými indexy u jednotlivých komponent. Toto rozdělení je především vhodné z důvodu **ladění procesu nahrávaní**, kdy konzole vypisuje logové zprávy v pořadí jednotlivých fází a snadněji se řeší případné nesrovnalosti nebo chyby.



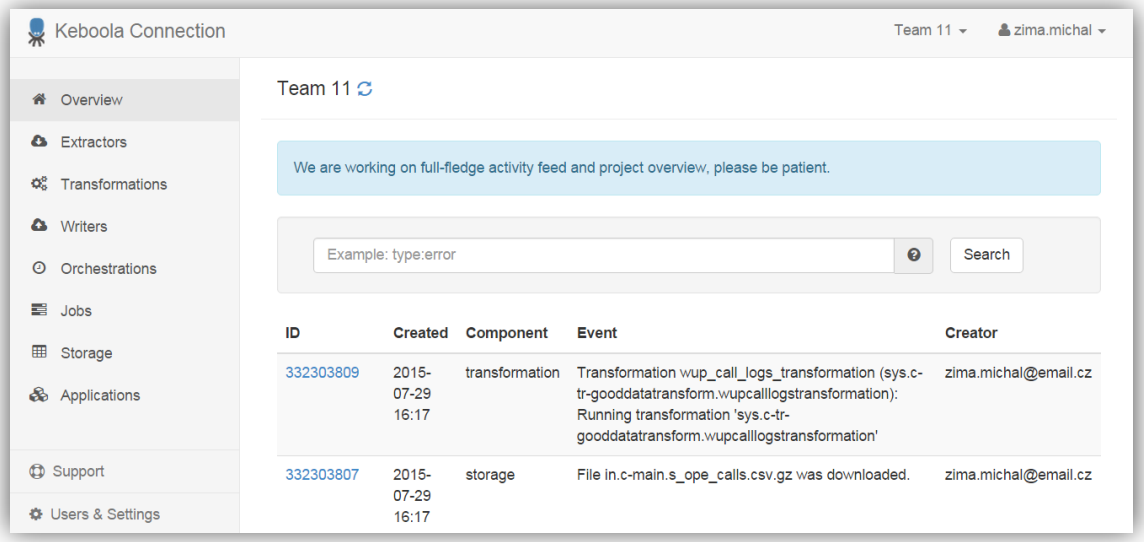
Obr. 22: Datová pumpa nahrávání dat do platformy GoodData.
Zdroj: Vlastní práce autora (prostředí CC).

Při spuštění datové pumpy vidíme u šipek znázorňujících napojení čísla, která uvádějí počet předávaných řádků na vstupech a výstupech.

Obdobně vytvoříme i ostatní datové pumpy, jejichž grafy si můžeme prohlédnout na konci práce v kapitole Přílohy. Z hlediska komplikovanosti jsou však jednodušší než toto námi popisované řešení.

5.2.4 Popis implementace úlohy v nástroji Keboola Connection

Keboola Connection je webová aplikace, proto je na první pohled odlišná od předchozího nástroje. Webové rozhraní standardně neumožňuje vykreslování na obrazovku nebo techniku drag and drop[14], kterou známe z desktopových ETL nástrojů. Po přihlášení do KBC se zobrazí uživateli seznam projektů, ke kterým má přístupová práva, a také menu pro správu účtu. Po otevření příslušného projektu se načte úvodní obrazovka, kterou si můžeme prohlédnout na Obr. 23.



Obr. 23 Úvodní obrazovka projektu v nástroji Keboola Connection.
Zdroj: Vlastní práce autora (prostředí KBC).

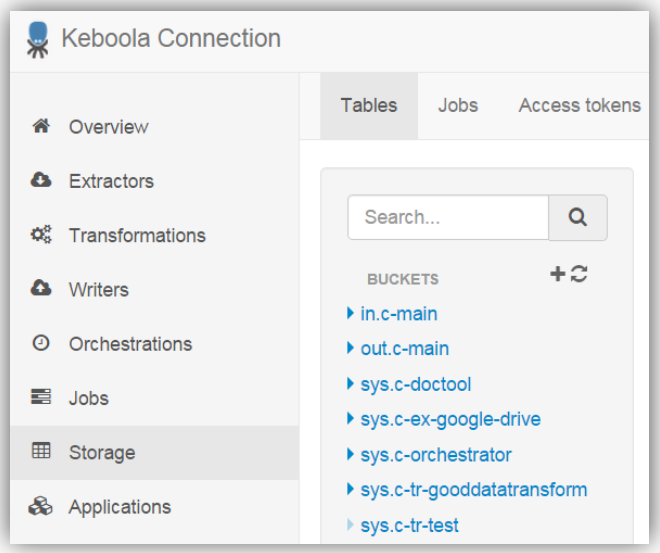
Aplikace má vedle horní lišty pouze dvě části. Navigační menu vlevo a prostor pro vlastní obsah jednotlivých položek vpravo. Strukturu aplikace lépe pochopíme po prozkoumání jednotlivých položek menu:

- **Overview –** úvodní obrazovka sloužící také k prohlížení logových zpráv systému.
- **Extractors –** komponenty obsluhující načtení dat do platformy.
- **Transformations –** transformace dat a vytváření propojení.

[14] Tento pojem se používá v informačních technologiích jako pojmenování způsobu přesouvání objektů na obrazovce, kdy uživatel vezme objekt pomocí kurzoru myši a umístí ho na jiné místo.

- **Writers –** komponenty, které nahrávají data do dalších aplikací např. Dropbox, GoogleDrive, Tableau, GoodData, různé databázové systémy a další.
- **Orchestrations –** plánování pravidelných aktualizací dat, včetně kontroly běhů.
- **Jobs –** seznamy proběhlých operací, včetně výpisu detailů k jejich běhu.
- **Storage –** uložiště dat, které umožňuje jejich správu včetně nahrávání, mazání a dalších operací.
- **Applications –** Vlastní aplikace pro analýzu dat ve webovém rozhraní KBC

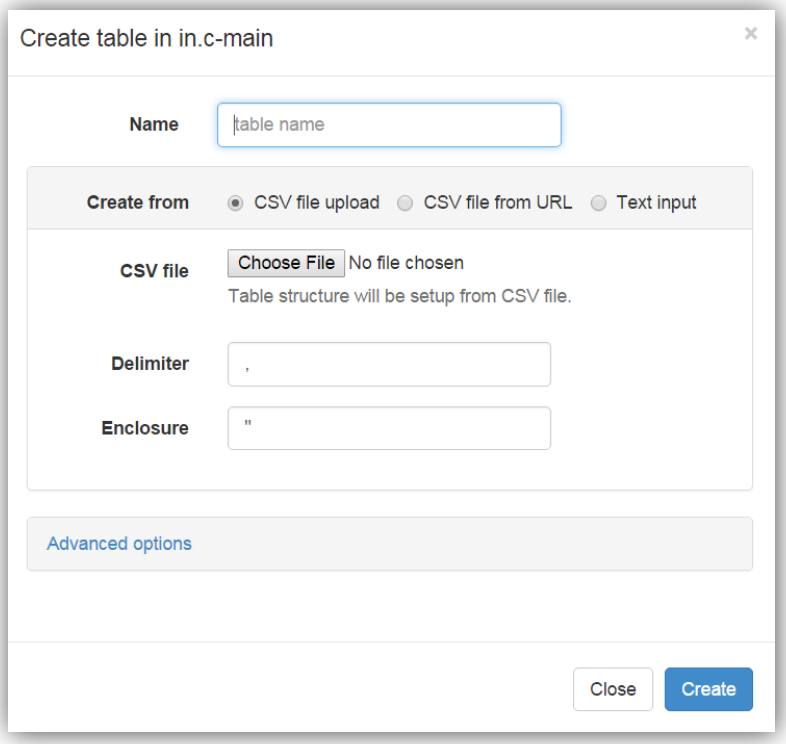
Po úspěšném přihlášení do aplikace a vytvoření projektu můžeme začít s řešením praktické úlohy. Nejprve je nutné zdrojová data nahrát do uložiště dat, tak aby s nimi mohla aplikace dále pracovat. V menu vybereme položku Storage (uložiště) a stejně jako na Obr. 24 se přepneme na kartu Tables (tabulky).



Obr. 24: Uložiště dat v aplikaci Keboola Connection.
Zdroj: Vlastní práce autora (prostředí KBC).

Na výběr máme několik složek, nazývaných v aplikaci buckets, a v nich jsou uloženy konkrétní tabulky. V našem případě použijeme pouze první dvě složky **in.c-main**, pro nahrání dat do KBC, a **out.c-main**, pro nahrání dat do GD. Anglická slova **in** (dovnitř) a **out** (ven) mají v tomto případě tedy svoje opodstatnění. Právě tabulky ze složky typu in jsou zobrazovány při transformacích a dalších úkonech na zdroji, a pokud máme

tabulky ve složce typu out, jsou tabulky destinace pro výstup. Vybereme tedy složku **in.c-main** a přidáme do ní novou tabulku pomocí průvodce na Obr. 25.

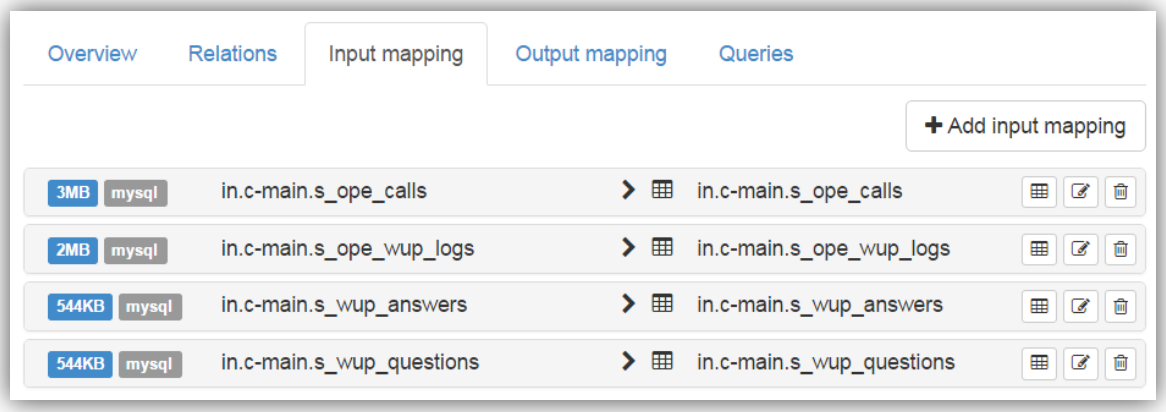


Obr. 25 Vytváření nové tabulky ze souboru CSV v KBC.
Zdroj: Vlastní práce autora (prostředí KBC).

Tabulku vytvoříme nahráním souboru CSV. Nejprve zvolíme název tabulky, poté cestu k souboru, následně oddělovač pro soubor CSV a nakonec nastavíme znak, který uvozuje hodnoty. Po tomto postupu se začnou nahrávat data do KBC a tabulka se objeví ve vybrané složce. Tímto způsobem nahrajeme všechny ostatní potřebné soubory.

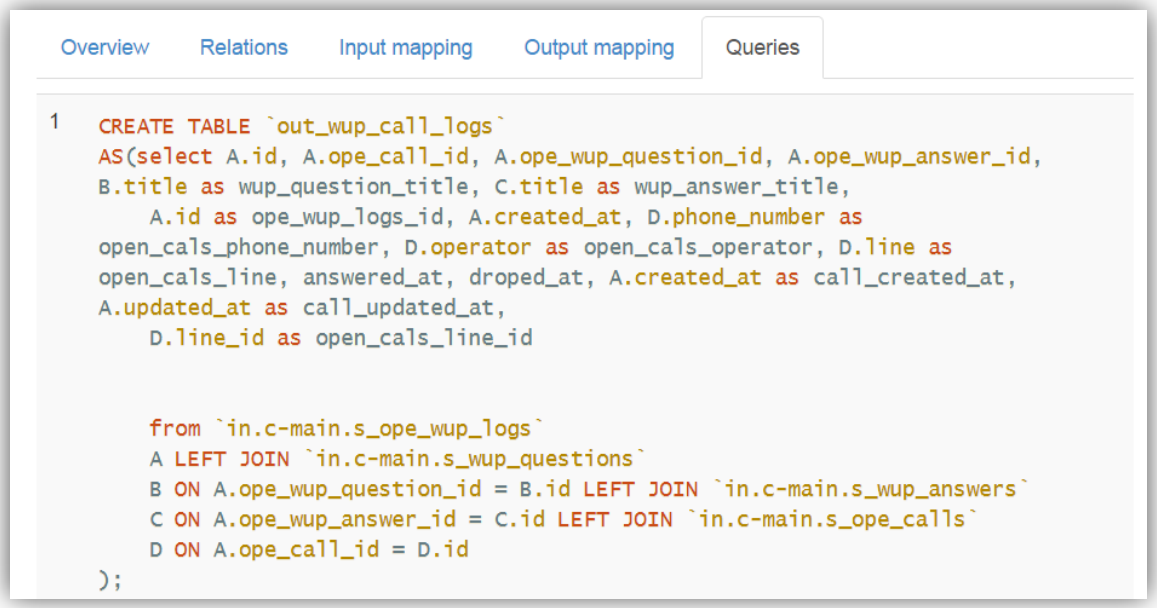
Dalším krokem bude příprava tabulky pro export. Tabulku tvoříme obdobným způsobem pomocí průvodce jako v předchozím případě s tím rozdílem, že ji umístíme do složky pro export do GD. Po takto vytvořených tabulkách se dostáváme k přípravě transformací. Pro lepší porovnání použijeme stejný příklad jako u předchozího nástroje.

V záložce transformací přidáme novou položku a přepneme se do karty **Input mapping**, do které přes průvodce postupně vložíme již nahrané tabulky, viz Obr. 26. U každé tabulky můžeme pomocí průvodce také změnit datové typy sloupců nebo odfiltrovat některé řádky, a to vše pomocí jednoduchých průvodců.



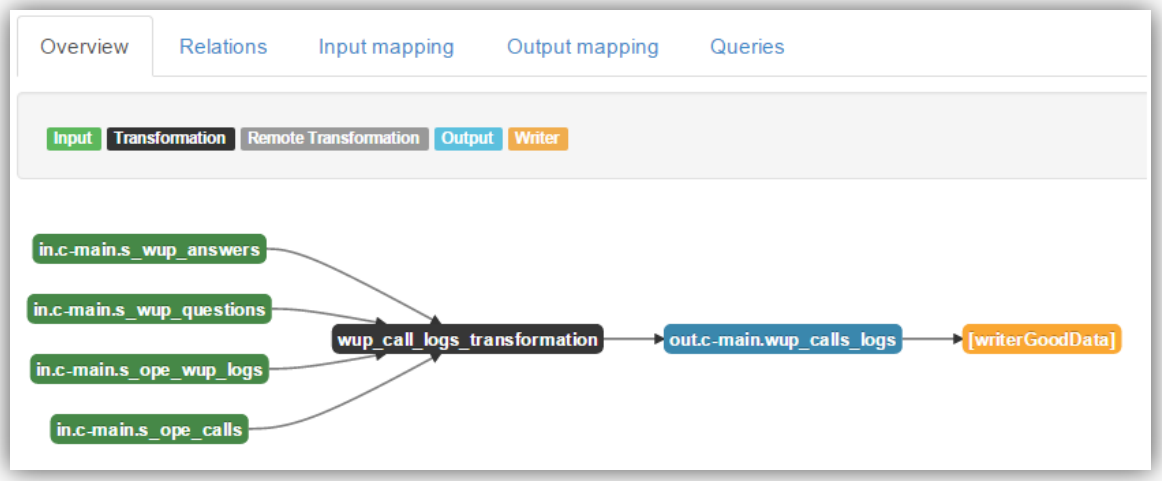
Obr. 26 Příprava transformace – vložení zdrojových tabulek.
Zdroj: Vlastní práce autora (prostředí KBC).

Obdobně pracujeme i s tabulkou v záložce **Output mapping**, kterou si připravíme pro nahrání výsledku transformace. Ty vytvoříme v záložce **Queries**. Při tvorbě transformace je možné si také vybrat, na jaké platformě bude probíhat. V nabídce je MySQL, Redshift a R. Z důvodů autorových zkušeností právě s prvně jmenovaným databázovým systémem bude kód vytvořen v jazyce SQL nad touto platformou. Celý kód napojení tabulek **ope_calls, ope_wup_logs, wup_question** a **wup_answers** je vidět na Obr. 27.



Obr. 27 SQL kód napojení tabulek v prostředí KBC.
Zdroj: Vlastní práce autora (prostředí KBC).

Tímto postupem jsme připravili transformaci, která je zároveň datovou pumpou stejně jako v nástroji CC. Na Obr. 28 je již vykreslený graf transformace z karty Overview.



Obr. 28: Schéma datové pumpy v nástroji KBC.
Zdroj: Vlastní práce autora.

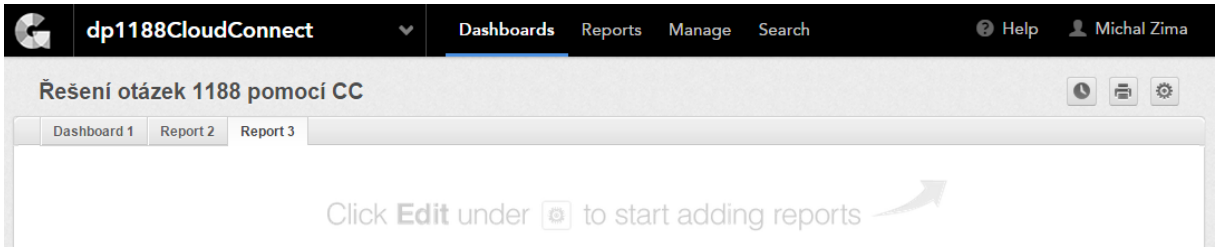
Stejným způsobem pokračujeme i u zbylých datových zdrojů a připravíme ostatní datové pumpy, které jsou co do složitosti jednodušší. Jejich řešení je dokumentováno na konci práce v kapitole Přílohy.

Posledním krokem je už jen nahrání dat do platformy GD. To provedeme v záložce Writers tím, že si nadefinujeme nový zapisovač pro náš projekt a vybereme tabulky, které budeme nahrávat. Nastavit můžeme také **automatické opakované nahrávání** v určitý čas nebo jednoduše nahrát pouze jeden datový set.

5.2.5 Prezentace praktické úlohy v GoodData

Práci s oběma ETL nástroji jsme si již představili a nyní zbývá už jen ověřit, jestli jimi nahraná data korespondují i v projektech GD.

Po přihlášení do webového rozhraní GoodData se ocitneme na úvodní obrazovce s horizontálním menu a prostorem pro obsah jednotlivých záložek, viz Obr. 29.



Obr. 29: Úvodní obrazovka řešení GD.
Zdroj: Vlastní práce autora (prostředí GD).

Menu, které je hlavním ovládacím prvkem celého řešení, je rozděleno do několika částí. Vlevo vidíme logo společnosti a nabídku projektů, kterých je uživatel součástí. Uprostřed panelu pak vidíme tyto záložky:

- **Dashboards –** tento v BI zavedený název by se dal chápat jako nástěnka, přestože doslovně přeložený název je přístrojová deska. Každý dashboard totiž může obsahovat reporty, různé ukazatele a ovládací prvky. Jedná se také o finální výstup řešení.
- **Reports –** záložka pro tvorbu reportů, které mohou být poté vloženy do dashboardu, exportovány nebo jinak publikovány.
- **Manage –** část aplikace, která spravuje většinu logických a analytických operací, včetně vytváření metrik a proměnných. Obsluhuje také správu uživatelů a automatického odesílání emailů, obsahujících dashboardy a reporty, odběratelům.
- **Search –** pole pro vyhledávání v projektu GD.

Poslední částí na pravé straně menu je odkaz na nápovědu a **správa vlastního profilu** uživatele, která umožňuje uživateli spravovat lokální nastavení nebo se odhlásit z aplikace.

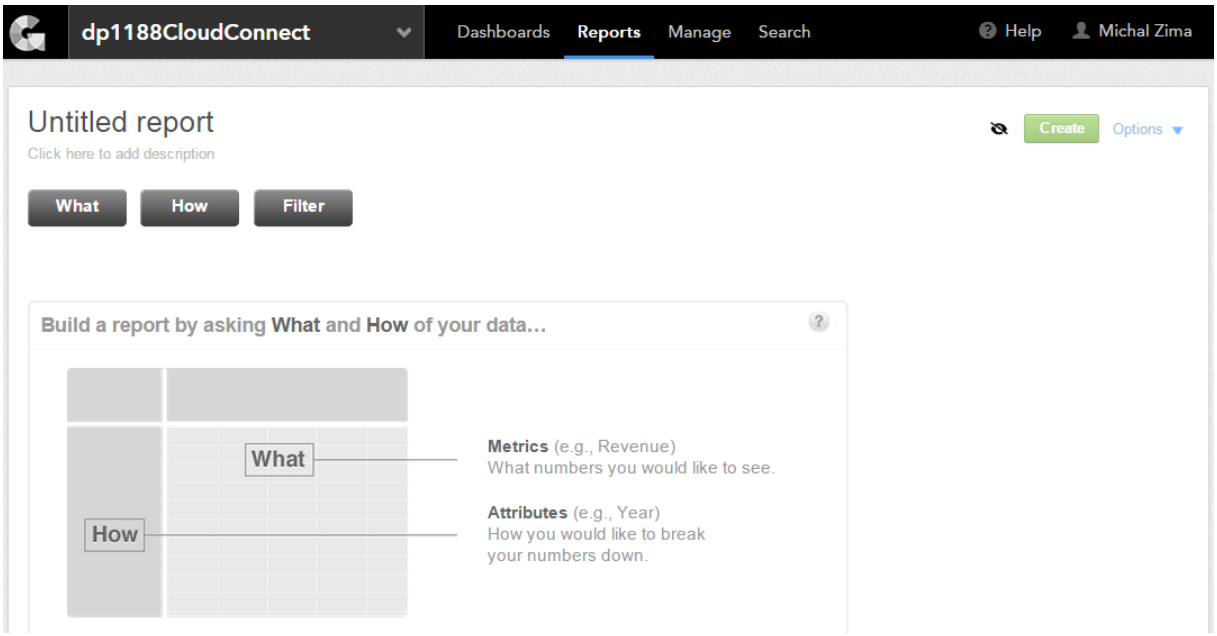
Pro přehlednost byly vytvořeny pro každý ETL nástroj vlastní projekty tak, aby se nemísily datové modely. Obsah obou projektů je ale stejný tak, aby na oba nástroje byly kladeny totožné nároky. Na dalších řádcích si tedy popíšeme jak pracovat s nástrojem GD.

Porovnání logických datových modelů

V nástroji CC jsme vytvořili datový model, který byl odeslán do projektu GD, a nyní si ho můžeme porovnat i s modelem, který generuje automaticky i nástroj KBC. Díky odlišnému přístupu k tvorbě modelu v nich můžeme naleznout odchylky. Tyto rozdíly jsou způsobeny například odlišným vytvářením časové dimenze. Nástroj KBC vytváří časovou dimenzi automaticky. Naproti tomu u nástroje CC je potřeba ji vytvořit ručně. Tyto rozdíly tedy není nutné dále rozvádět, oba modely jsou pak uvedeny v kapitole Přílohy.

Tvorba reportu v GD

Data máme nahrána v cloudu a nyní je na řadě je vhodně prezentovat. Přepneme se do karty Reports a tlačítkem vytvoříme nový report, jehož průvodce je vidět na Obr. 30.



Obr. 30 Tvorba nového reportu v aplikaci GD.
Zdroj: (prostředí GD).

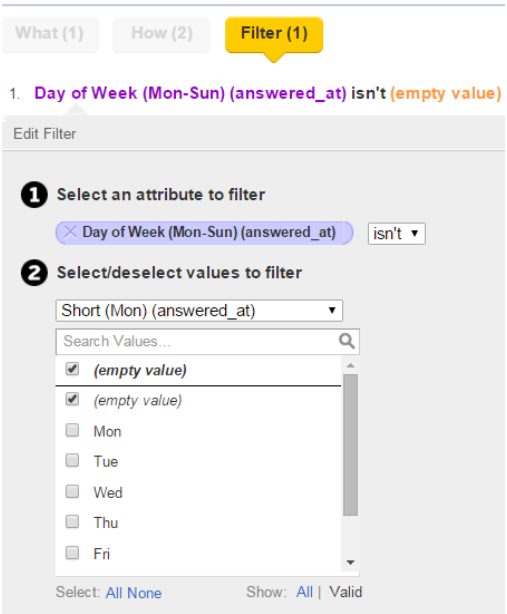
Pomocí průvodce si odpovíme na dvě základní otázky, Co? (What) a Jak? (How). V případě první otázky budeme vybírat fakta a vytvářet tzv. metriky. Ty umožňují definovat vzorce nad daty a provádět tak agregace nebo numerické a logické operace. Ukázka kódu jedné takové metriky je na následujících řádcích a byla vytvořena za účelem nalezení odpovědi na otázku: „Jak se projevuje den v týdnu na téma hovoru?“

```
SELECT ope_calls_count /
       (
              SELECT ope_calls_count BY day OF week (mon-sun) (answered_at),
                     ALL other without pf)
```

Kód je psán v jazyce MAQL[15], se kterým se čtenář mohl setkat pouze u řešení GD. Přestože se nejedná o jazyk SQL, tak jeho syntaxe je podobná. Po projití několika ukázkových kódů, které jsou dostupné přímo v editoru, by pro čtenáře, s předchozí zkušeností s SQL, nemělo být obtížné si vytvořit vlastní jednoduchou metriku.

Dalším krokem je výběr dimenze, podle které budou data zobrazována. V našem případě je to čas a konkrétně dny v týdnu v kombinaci s parametrem *wup_question_title*, tedy názvem otázky.

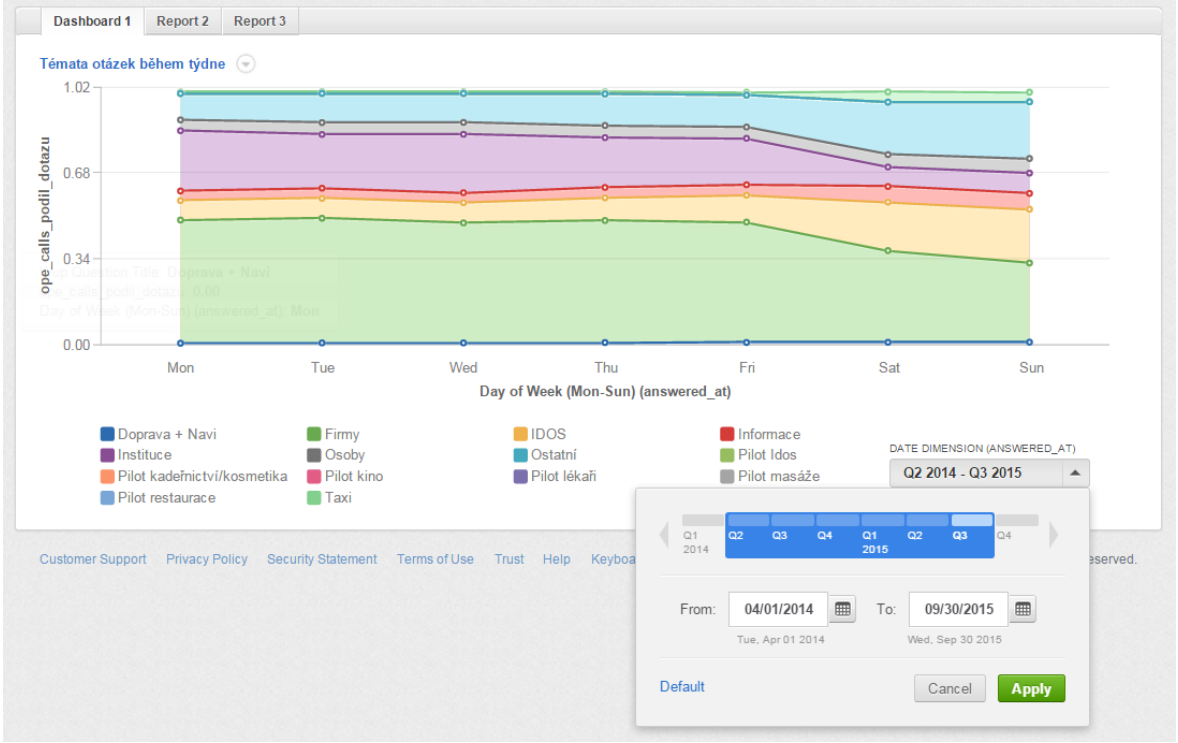
Posledním krokem je vytvoření filtru. Filtr má za úkol nezobrazovat v grafu hodnoty, u kterých není uveden atribut *answered_at,* a vyloučit tak případný osmý sloupec, viz Obr. 31.



Obr. 31 Nastavení filtru při tvorbě reportu.
Zdroj: Vlastní práce autora (prostředí GD).

[15] Multi-Dimension Analytical Query Language je jazyk z dílny společnosti GoodData a je navržen pro čtení dat v platformě této společnosti. Dokumentace je dostupná z: http://help.gooddata.com/doc/public/pdf/MAQL%20Reference%20Guide.pdf?

Takto jsme vytvořili všechna potřebná nastavení pro nový report. Po nastavení viditelnosti reportu pro zvolenou skupinu uživatelů ho přidáme do nového dashboardu, viz Obr. 32.



Obr. 32 Ukázkový report „Témata otázek během týdne“ v aplikaci GD.
Zdroj: Vlastní práce autora.

Za pozornost také stojí možnost přidat filtr přímo do dashboardu a tím vytvořit dynamický report, který se mění podle například zvoleného období, viz předchozí obrázek.

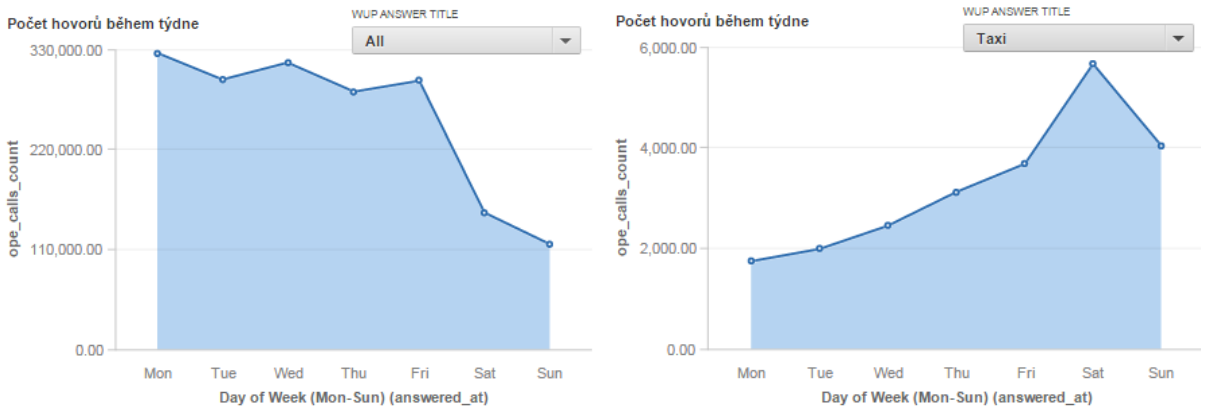
Tímto postupem pokračujeme i při tvorbě dalších reportů tak, abychom dokázali odpovědět na otázky od společnosti 1188.

5.2.6 Odpovědi na praktické otázky od společnosti 1188

Protože by popis samotných výsledků přesahoval rozsah této diplomové práce, omezíme se pouze na stručný popis dashboardů pro jednotlivé otázky, včetně komentářů k některým zjištěním. Celé řešení pak bude předáno této společnosti, která může využít dynamické reporty pro vytváření konkrétních dotazů.

1. Chování zákazníků podle dní v týdnu.

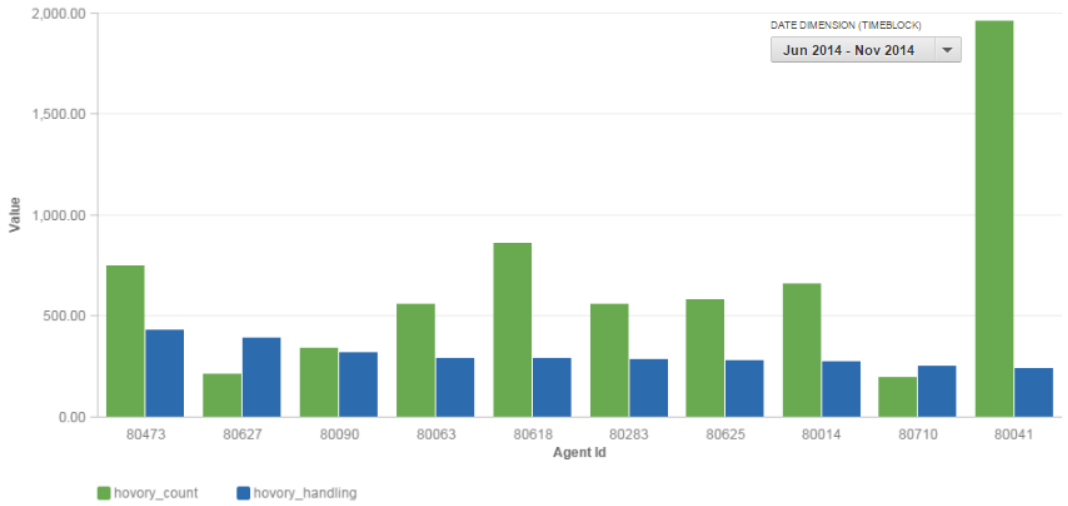
Pro tuto otázku byly vytvořeny dva grafy. Ten první na Obr. 32 jsme již využili pro demonstraci řešení v GD. Z něho vyplývá, že příchodem víkendu se snižuje četnost otázek na „Firmy“ či „Osoby“, a naopak vzrůstá četnost otázek typu „Informace“. Jako doplněk k této skutečnosti nám poslouží porovnání dvou verzí grafu na Obr. 33 ukazující počet hovorů během týdne s možností výběru typu odpovědi, kterou zákazník dostal. V prvním případě jsme vybrali všechny odpovědi a tím zjistili, že u linky dochází o víkendu k celkovému úbytku hovorů. V případě odpovědi typu „Taxi“ naopak dochází k nárůstu. Na základě takových zjištění můžeme například lépe cílit reklamu nebo připravit operátory na předpokládanou tématiku hovoru.



Obr. 33 Porovnání počtu hovorů během týdne na základě typu odpovědi.
Zdroj: Vlastní práce autora, data 1188 společnosti (prostředí GD).

2. Hodnocení agentů podle dostupných ukazatelů.

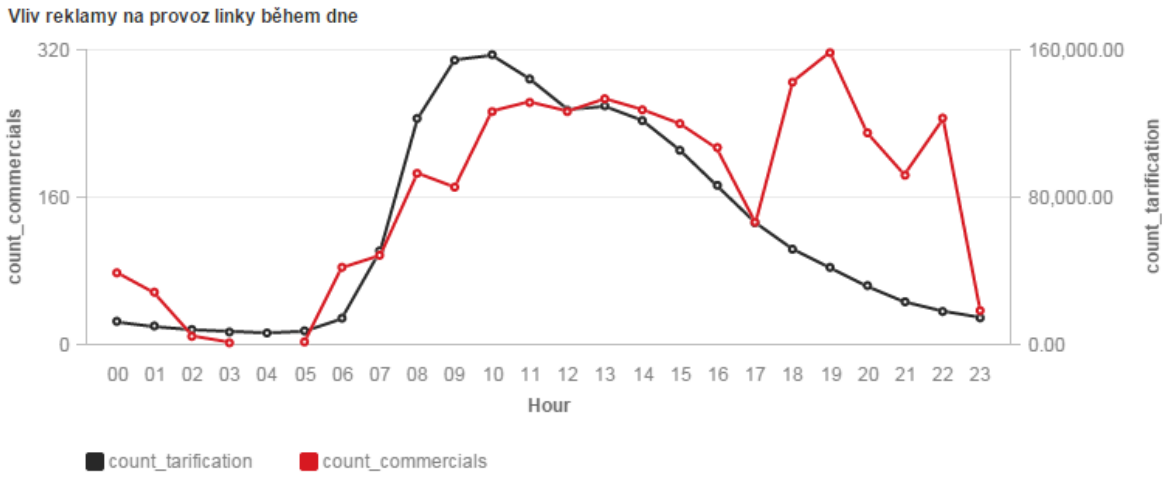
V tomto případě jsme vytvořili tři grafy a umístili je do dashboardu, viz kapitola Přílohy. První ukazuje deset operátorů s nejdelší délkou hovoru, druhý ukazuje desítku operátorů s nejvíce hovory a poslední je kombinací obou předešlých. První dva zmíněné grafy jsou pouze seřazeným seznamem agentů podle určitého kritéria a mohou sloužit například k finančnímu ohodnocování jednotlivých operátorů na konci měsíce podle jejich výkonu. Třetí graf na Obr. 34 je o něco složitější a je z něho patrné, že zvýšený počet hovorů nemá nejspíš přímý vliv na délku hovorů. Mezi desítkou operátorů s nejdelší délkou hovorů najdeme i případy s velmi nízkým počtem obsloužených zákazníků, tak i s velmi vysokým počtem za zkoumané období.



Obr. 34: Graf deseti agentů s počtem hovorů, řazených podle délky hovoru
Zdroj: Vlastní práce autora, data společnosti 1188 (prostředí GD).

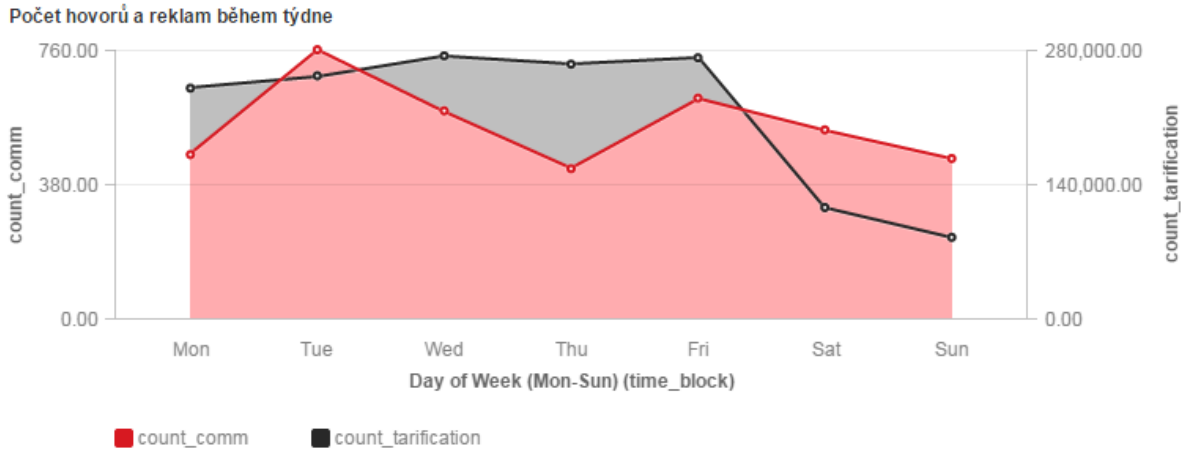
3. Vliv TV spotů na provoz linky.

Pro zodpovězení této otázky byl vytvořen graf na Obr. 35, který porovnává četnost hovorů a četnost reklam během dne. Z grafu je patrné, že ve večerních hodinách, kdy ubývá hovorů na linku, nemá pravděpodobně zvýšení četnosti reklamy vliv na počet hovorů.



Obr. 35 Vliv reklamy na provoz linky během dne.
Zdroj: Vlastní práce autora, data společnosti 1188 (prostředí GD).

Pokud bychom porovnávali četnost hovorů a četnost reklam během týdne, viz Obr 36, dalo by se říci, že výrazným snížením četnosti reklam během týdne dochází i ke snížení počtu hovorů během dne. Naopak z grafu se zdá, že úterní navýšení počtu reklam vede k drobnému navýšení počtu hovorů. Obě tato tvrzení jsou samozřejmě velmi odvážná a pro potvrzení bychom potřebovali mnohem více informací o chování zákazníků, které nemáme k dispozici. Hodnotit tedy dopad reklamy na chování zákazníků pouze podle počtu hovorů během dne není nejvhodnějším řešením.



Obr. 36 Počet hovorů a reklam během dne.
Zdroj: Vlastní práce autora, data společnosti 1188 (prostředí GD).

Kompletní dashboardy jsou obsaženy v kapitole Přílohy na konci práce. Reporty vznikly v obou projektech, jen s drobnými kosmetickými rozdíly, které byly zapříčiněny odlišnostmi logických datových modelů. Konkrétně v názvosloví faktických tabulek a některých atributů. Pro jednoduchost a zamezení opakování jsou uvedeny jednotlivé dashboardy pouze jednou.

5.3 Hodnocení AHP modelu

Po praktickém vyzkoušení nástrojů můžeme přistoupit k ohodnocení vytvořeného AHP modelu. Jak již bylo řečeno, využijeme metodu párového srovnání a Saatyho stupnici. Bodové hodnocení nástrojů podle jednotlivých kritérií je k naleznutí v Tab. 6, za slovním popisem porovnání nástrojů, které obsahuje mimo jiné i důvody autorova hodnocení.

Funkční kritéria

- **Náročnost použití** – Dosažení cíle, tedy nahrání dat do aplikace GD, bylo o něco snadnější s nástrojem KBC. Důvodem bylo automatické generování logického modelu na základě nastavení výstupní tabulky. Také v nástroji CC bylo nutné často upravovat ručně metadata. V ostatních ohledech byly nástroje na stejné úrovni náročnosti, a proto je hodnocení jen mírně v neprospěch nástroje CC.
- **Funkcionalita** – Propracovanost použitých funkcionalit a jejich nabídka se nějak výrazně nelišila. Pro práci s daty byla k dispozici dostatečná paleta komponent a ani jednomu z nástrojů nic nechybělo. V tomto ohledu je tedy hodnocení neutrální.
- **Odstraňování chyb** – Nástroj CC nabízí mnohem podrobnější logování při běhu nástroje, které vede ke snadnějšímu odhalení problémů. Nabízí také grafické znázornění probíhajících transformací během debuggingu. Obě tyto funkcionality byly během řešení často používány. Celkově se hledaly chyby hůře v nástroji KBC, který nenabízí tak podrobné logování. Pokud se objevila nějaká chyba, bylo nutné ji často složitěji dohledávat.
- **Řešení práv** – Nástroj KBC umožňuje současnou práci více uživatelů na jednom projektu díky umístění projektu v cloudu. Desktopová aplikace CC toto řešení neumožnuje a práce více uživatelů je tak složitější. Jednou z možností je používat sdílené datové zdroje a předávat si zdrojové soubory projektu, ale to je značně nepohodlné.
- **Přívětivost uživatelského rozhraní** – Nástroj CC je v mnoha ohledech intuitivnější na ovládání a netrvá dlouho se v prostředí zorientovat. Nástroj KBC budí při prvním setkání dojem, že je komplikovanější a bez přečtení návodu je jen velmi obtížně pochopitelné jak s nástrojem pracovat. V tomto ohledu, i díky použití desktopového přístupu, je nástroj CC hodnocen lépe.

Technická kritéria

- **Architektura** – Velkou výhodou nástroje KBC je cloudový přístup, který transformace a práci velmi zrychluje. V našem případě se osvědčil více přístup nahrání dat do aplikace KBC než práce s lokálními soubory v nástroji CC. Oba nástroje umožňují paralelní zpracování procesů a jejich plánování.
- **Podpora formátů** – Oba nástroje podporují velkou paletu zdrojových formátů. Vývojáři nástroje KBC se soustředí hlavně na moderní zdroje a napojení na webové služby. Nástroj CC v tomto ohledu poněkud ztrácí, na druhou stranu podporuje nahrávání zdrojových dat ve formátech, jako je xml, xlsd, a další. Z tohoto důvodu je hodnocení neutrální.
- **Příprava dat a čištění** – V tomto ohledu je nástroj CC propracovanější a to díky množství komponent, které umožňují s daty dělat prakticky cokoliv. Na druhou stranu nástroj KBC je schopen mnoho práce udělat automaticky při nahrávání datových zdrojů do platformy nebo pomocí vytvoření kódu pro transformaci.
- **Rychlost nahrání** – Proces nahrávání dat byl v aplikaci KBC rychlejší. Počet kroků potřebných k nahrání byl také menší. Nedá se říci, že by byl rozdíl propastný, přesto byl dostatečně zřejmý a nástroj CC v porovnání s KBC viditelně ztrácel.
- **Konektivita** – Jednoznačným plusem pro nástroj KBC je podpora množství služeb poskytujících data, ke kterým se může napřímo napojit. Nástroj CC neoperuje takovou paletou možných napojení, důvodem je i celková koncepce nástroje a také poněkud rozdílný účel použití v běžné praxi.

Obecná kritéria

- **Možnost opětovného použití** – V tomto ohledu je jednoznačně výhoda na straně řešení v nástroji CC. Projekt a celé řešení je snadno kopírovatelné a modifikovatelné. KBC nabízí možnost opětovného použití řešení také, ale ne v tolik přívětivé formě jako nástroj CC.
- **Bezpečnost a stabilita** – Stabilitu můžeme hodnotit pouze po dobu testování. Nutno říci, že aplikace CC se chovala často nestabilně. Několikrát bylo třeba

program restartovat a výjimkou nebyla ani náhlá ukončení programu. S nástrojem KBC, jako se službou takovou, nedocházelo k žádným potížím. Tento nástroj je samozřejmě do jisté míry závislý na rychlosti připojení a jeho stabilitě, přesto se po dobu testování choval velmi stabilně. Hodnocení je ve prospěch nástroje KBC.

- **Cena** – Cena nástroje CC je zahrnuta již v nákladech na projekt GD a stejně je tomu i v případě nástroje KBC. Roční poplatek za projekt GD je pak odvislý od datové kapacity projektu. V našem případě, kdy objem dat je do 2GB, hledáme nejúspornější variantu. Podle Pavla Kvasničky z firmy Keboola dokáže společnost nabídnout nejmenší balíček řešení a to až s pětinásobně menší datovou kapacitou než firma GD, a tím výrazně ušetřit. Hodnocení tohoto kritéria je tedy ve prospěch nástroje KBC.
- **Rozšiřitelnost** – KBC poskytuje množství dalších nástrojů, které mohou pracovat s nahranými daty. Vhodné je to například, pokud má uživatel v úmyslu obohatit data o další informace. Takovým příkladem je služba Geneea[16], která dokáže analyzovat texty a její napojení je součástí KBC.
- **Podpora** – Oba nástroje mají dobře propracovaný manuál. Díky většímu povědomí o nástroji CC je i snazší dohledat potřebné rady na internetu. Pokud budu hodnotit online podporu, kterou jsem vyzkoušel u firmy Keboola i firmy GoodData, tak rychlost řešení problémů u firmy Keboola byla vyšší.

[16] Geneea, je služba umožňující analyzovat text, přidávat diakritiku a odhalovat téma v textu. Nabízená je pro jazyky čeština a angličtina (Geneea, 2015).

Tabulka 6 Bodového ohodnocení AHP modelu pro nástroje CC a KBC.

Název kritéria	Váha kritéria	Párové srovnání (Saatyho stupnice)		Výsledné body*	
		CC	KBC	CC	KBC
Náročnost použití	0.12	0.33	3.00	0.04	0.37
Funkcionalita	0.12	1.00	1.00	0.12	0.12
Odstraňování chyb	0.08	3.00	0.33	0.24	0.03
Řešení práv	0.04	0.20	5.00	0.01	0.18
Přívětivost uživatelské rozhraní	0.04	3.00	0.33	0.11	0.01
Architektura	0.11	0.33	3.00	0.04	0.34
Podpora formátů	0.11	1.00	1.00	0.11	0.11
Příprava dat a čištění	0.08	2.00	0.50	0.15	0.04
Rychlost nahrání	0.06	0.33	3.00	0.02	0.19
Konektivita	0.04	0.33	3.00	0.01	0.11
Možnost opětovného použití	0.06	7.00	0.14	0.39	0.01
Bezpečnost a stabilita	0.06	0.33	3.00	0.02	0.17
Cena	0.04	0.14	7.00	0.01	0.25
Rozšiřitelnost	0.03	0.50	2.00	0.02	0.07
Podpora	0.02	1.00	1.00	0.02	0.02
		Součet bodů		**1.30**	**2.01**

Zdroj: Vlastní práce autora. *součin váhy kritéria a párového srovnání

V některých případech byly nástroje obodovány hodnotou, která není slovně popsána v Saatyho stupnici. Došlo k tomu z důvodu potřeby zvětšení citlivosti hodnocení a docílení tak přesnějšího výsledku. Např. kritérium *Příprava dat a čištění* bylo ohodnoceno ve prospěch nástroje CC dvěma body, tedy hodnocení mezi stupni mírně a stejně významný. Takové hodnocení by se dalo popsat jako jen velmi mírně až podobně významný.

6 Shrnutí výsledků

Na základě ohodnocení **AHP modelu** a převedení bodového hodnocení na procentuální podíl jsme dospěli k výsledku hodnocení **60,7% pro nástroj KBC a 39.3% pro nástroj CC**.

Pro ověření míry konzistence párových srovnání jsme celý model vytvořili a ohodnotili i v programu Expert Choice 2000. Tato míra se pohybovala u jednotlivých rozdělení vah kritérií v rozmezí 0 až 0,13, výsledky jsou obsaženy v Přílohách. Model můžeme na základě Saatyho tvrzení o míře nekonzistence (Saaty, 1987) hodnotit jako dostatečně **konzistentní**.

Po doplnění výsledků obodování jednotlivých kritérií také do modelu v programu **Expert Choice 2000**, docházíme k výsledkům; **56% pro KBC a 44% pro CC**. Rozdíl mezi oběma modely vznikl na základě přístupu programu k míře nekonzistence. Zatím co námi vytvořený model hodnotíme, jako by byl stoprocentně konzistentní, model z programu s mírou nekonzistence počítá, a proto jsou u výsledky hodnocení programem **korigovány**.

Oba porovnávané ETL nástroje dokázaly splnit praktickou úlohu nahrání dat do platformy GD. Na základě těchto dat společnosti 1188 byly vytvořeny reporty a dashboardy, které odpovídají na okruhy otázek touto společností vyslovených. Protože jsme pro každý nástroj použili vlastní projekt GD, tak minimálně na výsledných grafech se potvrdilo, že **nahraná data se neliší** a obě vytvořená řešení nahrála data správně.

Odpovědi na jednotlivé okruhy otázek jsou popsány v kapitole 5.2.6 Odpovědi na praktické otázky od společnosti 1188. Jedná se pouze o stručný popis zjištění z vytvořených podkladů. Právě tyto podklady jsou požadovaným výsledkem, protože jen na jejich základě může společnost 1188 vytvářet jednoduchou změnou konfigurace reportů nebo dashboardů svoje **vlastní konkrétní dotazy**.

7 Závěry a doporučení

Hlavním cílem této diplomové práce bylo doporučit čtenáři ETL nástroj pro BI v cloudu. Na základě mnohahodinového praktického testování, hodnocení pomocí patnácti na sobě nezávislých kritérií a především všech popsaných výsledků **doporučuji nástroj KBC**. Moderní přístup, který firma Keboola promítla do jejího nástroje, se zdá jako ten správný. Přesto je důležité zmínit fakt, že oba nástroje splnily úlohu nahrát data do platformy GD, a oba jsou použitelné.

Celé praktické testování ukázalo, že velmi záleží na konkrétním požadavku zadavatele na nástroj. Oba zkoumané nástroje ukázaly, že mají své silné i slabé stránky. V některých ohledech mohlo být moje hodnocení ovlivněno subjektivním vnímáním skutečností. Tento fakt jsem se snažil vhodným **doplněním o konkrétní údaje**, **praktické zkušenosti** nebo vhodnou **konzultací s pověřenými osobami** minimalizovat.

Mezi další cíle, které byly splněny, bylo **vytvoření návodu** pro práci s popisovanými aplikacemi, který by měl čtenáři sloužit jako představení principů, a manuál pro práci v daném prostředí.

Stejně tak byla popsána i platforma GoodData a tím **představen moderní směr BI**, který se více a více stává standardem ve zkoumaném odvětví. Dospěl jsem k uvědomění, že přínos cloud computingu v oblasti řešení BI je zcela neoddiskutovatelný.

Nesmíme také opomenout společnost 1188, jejíž data byla analyzována. Poznatky a kompletní řešení bylo předáno kompetentním osobám a věřím, že i pro ně bude tato práce přínosem.

Všechny v úvodu definované **cíle a výstupy byly** tedy bezezbytku **naplněny**. Díky ucelenému pohledu na tuto zajímavou problematiku jsem dospěl k vlastnímu závěru, že BI se opravdu ubírá správným směrem a zanechává to ve mně osobní pocit těšení se na další vývoj.

Myslím si, že nad rámec této diplomové práce by bylo zajímavé provést podrobnější **prozkoumání platformy GD**, protože se ukázalo, že poskytuje nepřeberné množství zajímavých metod a funkcí. Bylo by jistě přínosné prověřit tyto funkcionality

na datových zdrojích, které by umožnily větší zapojení tohoto komplexního nástroje a ukázaly tak v plném světle jeho sílu.

Terminologický slovník pojmů

Termín	Zkratka	Význam (zdroj)
Debugging	**---**	Krokování a hledaní kritických chyb a varování u zkoumaného procesu. (Autor)
CloudConnect	**CC**	CloudConnect je ETL řešení licencované společností GoodData, používané pro nahrávání dat do této platformy. (GoodData, 2013)
Dashboard	**---**	Nástroj pro poskytnutí grafického uživatelského rozhraní, které umožňuje uživateli vidět klíčové informace nebo metriky. Často obsahuje data obchodního charakteru, která uživatel může využít k vytvoření rozhodnutí obchodního charakteru. (Yan, 2010)
Decision Support System	**DSS**	Na modelech založený soubor procedur na zpracování dat a usuzovaní o nich, určený k pomoci manažerovi při jeho rozhodování. (Khosrow-pour, 2009, s. 1753)
Desktopová aplikace	**---**	Aplikace, která běží na osobním počítači nebo notebooku. (Autor)
Extensible Markup Language	**XML**	Značkovací jazyk a způsob zápisu, určený pro výměnu dat mezi aplikacemi. (Quin, 2015)
Infrastructure as a Service	**IaaS**	Služba, která pomocí internetu poskytuje infrastrukturu, např. servery, uložiště atd. (Jamsa, 2013, s. 6)
Java	**---**	Objektově orientovaný programovací jazyk, vyvinutý společností Sun Microsystems. (Autor)
JavaScript Object Notation	**JSON**	Formát pro výměnu dat, který je založený na textu a je jazykově nezávislý. (Bray, 2014)
metadata	**---**	Data nebo také informace o datech. (Suehring, 2002, s. 516.)
Metrika	**---**	Metrika je indikátor vyjadřující stav určitého systému, například jeho kvality nebo efektivnosti, a nabývá přitom různých hodnot. (Management Mania, 2013)
Multi-Dimension Analytical Query Language	**MAQL**	Jazyk vyvinutý společností GoodData pro definici metrik a vytváření dotazů v její platformě. (GoodData, 2015b)

Termín	Zkratka	Význam (zdroj)
Multi-level caching	**---**	Načítání velkých množství objektů do paměti v několika úrovních. (Gill, 2008)
MySQL	**---**	Relační databázový systém, který se těší velké oblibě a je dostupný pro velké množství operačních systémů.(Suehring, 2002, s. 11-12.)
Online Analytical Processing	**OLAP**	Metoda uchování a prezentace velkého objemu dat, umožňující manažerům a analytikům multidimenzionální analýzu.(Vassiliadis, 1999)
On-premise	**---**	Označení softwaru nebo komponenty, která náleží společnosti (např. vlastní server). (Autor)
Open source	**---**	Typ licencování softwaru, nabízející kód programu volně k dispozici a distribuci. (Dibona a kol., 1999, s. 2)
Platform as a Service	**PaaS**	Služba, která skrze internet poskytuje kompletní platformu pro tvorbu vlastního řešení nebo aplikace. (Jamsa, 2013, s. 6)
Redshift	**---**	Platforma vyvinutá společností Amazon pro správu a uchování dat. (Amazon, 2015)
Software as a Service	**SaaS**	Aplikace umístěná v cloudu s vlastním uživatelským rozhraním, ke které je přistupováno pomocí internetu. (Jamsa, 2013, s. 6)
Validace	**---**	Kontrola vstupních údajů při zadávání dat. (Autor)
R	**---**	Programovací jazyk, který má určité prvky podobnosti s jazykem C. (R Core Team, 2015)

Seznam použité literatury

AMAZON. *Amazon Redshift System Overview* [online]. 2015 [cit. 2015-08-04]. Dostupné z: http://docs.aws.amazon.com/redshift/latest/dg/c_redshift_system_overview.html

ARMBRUST, Michael, Ion STOICA, a kol. A view of cloud computing. *Communications of the ACM* [online]. 2010, **53**(4): 50- [cit. 2015-07-17]. DOI: 10.1145/1721654.1721672. ISSN 00010782. Dostupné také z: http://portal.acm.org/citation.cfm?doid=1721654.1721672

AUSTIN, Benny. Kimball and Inmon DW Models. *Benny Austin* [online]. 2010, 2010-05-02 [cit. 2015-06-22]. Dostupné z: https://bennyaustin.wordpress.com/2010/05/02/kimball-and-inmon-dw-models/

BAARS, Henning; Kemper, Hans-Georg. Business intelligence in the cloud?. In: *PACIS* [online]. 2010, p. 145. [cit. 2015-06-22]. Dostupné z: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1143&context=pacis2010

BRAY, Tim. The JavaScript Object Notation (JSON) Data Interchange Format. [online]. 2014, [cit. 2015-06-22]. Dostupné z: http://tools.ietf.org/html/rfc7159.html

ČERNÝ, Richard. *Srovnání cloud BI řešení a faktory ovlivňující jejich nasazení*. Praha, 2014. Dostupné také z: https://www.vse.cz/vskp/40325_srovnani_cloud_bi_reseni_a%C2%A0faktory_ovlivnujici_jejich_nasazeni. Diplomová práce. Vedoucí práce Pour, Jan.

DIBONA, Chris, Sam OCKMAN a Mark STONE. *Open sources: voices from the open source revolution*. 1st ed. Sebastopol, CA: O'Reilly, c1999, viii, 272 p. ISBN 15-659-2582-3.

EVANS, Dave. The internet of things: How the next evolution of the internet is changing everything. *CISCO white paper*, 2011, 1: 14. Dostupné z: https://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

FILIPČÍK, Zdeněk. *Nástroje Business Intelligence jako Open Source*. Praha, 2013. Dostupné také z: https://www.vse.cz/vskp/34862_nastroje_business_intelligence_jako_open_source. Diplomová práce. Vysoká škola ekonomická v Praze. Vedoucí práce Pour, Jan.

FORBES. Gartner Predicts CRM Will Be A $36B Market By 2017. Forbes [online]. 2013, 2013-6-18 [cit. 2015-07-16]. Dostupné z: http://www.forbes.com/sites/louiscolumbus/2013/06/18/gartner-predicts-crm-will-be-a-36b-market-by-2017/

FORBES. Gartner Predicts Infrastructure Services Will Accelerate Cloud Computing Growth. Forbes [online]. 2013, 2013-2-19 [cit. 2015-07-16]. Dostupné z: http://www.forbes.com/sites/louiscolumbus/2013/02/19/gartner-predicts-infrastructure-services-will-accelerate-cloud-computing-growth/

GÁLA, Libor, Jan POUR a Zuzana ŠEDIVÁ. *Podniková informatika*. 2., přeprac. a aktualiz. vyd. Praha: Grada, 2009, 496 s. Expert (Grada). ISBN 978-80-247-2615-1.

Geneea. [online]. 2015 [cit. 2015-08-10]. Dostupné z: http://www.geneea.com

GEORGE, Sansu. Inmon vs. Kimball: Which approach is suitable for your data warehouse?. *Search Business Intelligence* – Tech Target[online]. 2012, 2012-04-01 [cit. 2015-06-22]. Dostupné z: http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse

GILL, Binny S. On multi-level exclusive caching: offline optimality and why promotions are better than demotions. In: *Proceedings of the 6th USENIX Conference on File and Storage Technologies*. USENIX Association, 2008. p. 4. [cit. 2015-06-22]. Dostupné z: https://www.usenix.org/legacy/event/fast08/tech/full_papers/gill/gill.pdf

GoodData. *CloudConnect Designer User Manual* [online]. 2012 [cit. 2015-07-10]. Dostupné z: https://developer.gooddata.com/cloudconnect/manual/index-frames.html

GoodData. *CloudConnect core concepts* [online]. 2013 [cit. 2015-07-10]. Dostupné z: https://developer.gooddata.com/article/intro

GoodData. *GoodData Platform Overview* [online]. 2015a [cit. 2015-07-20]. Dostupné z: http://info.gooddata.com/rs/gooddata/images/GoodData%20Platform%20Technical%20Brief.pdf

GoodData. *MAQL Reference Guide* [online]. 2015b [cit. 2015-07-05]. Dostupné z: http://help.gooddata.com/doc/public/pdf/MAQL%20Reference%20Guide.pdf?

GRIFFITH, Eric. What Is Cloud Computing? *PCMag.com* [online]. 2015, 2015-04-17 [cit. 2015-06-22]. Dostupné z: http://www.pcmag.com/article2/0,2817,2372163,00.asp

CHAUDHURI, Surajit, Umeshwar DAYAL a Vivek NARASAYYA. An overview of business intelligence technology. *Communications of the ACM* [online]. 2011,54(8): 88- [cit. 2015-06-22]. DOI: 10.1145/1978542.1978562. ISSN 00010782. Dostupné z: http://portal.acm.org/citation.cfm?doid=1978542.1978562

JAMSA, Kris. *Cloud computing: SaaS, PaaS, IaaS, virtualization, business models, mobile, security and more*. 1. vydání. Burlington, MA: Jones, 2013, xix, 322 p. ISBN 978-144-9647-391.

QUIN, Liam. Extensible Markup Language (XML). *W3C* [online]. 2015, 2015-05-19 [cit. 2015-08-18]. Dostupné z: http://www.w3.org/XML/

Keboola. *Manual and automatic uploading of CSV tables to Keboola Connection* [online]. 2014 [cit. 2015-07-10]. Dostupné z: http://wiki.keboola.com/home/keboola-connection/user-space/data-related-tricks/sapi-how-tos/manual-and-automatic-uploading-of-csv-tables-to-keboola-connection

KHOSROW-POUR, Mehdi. *Encyclopedia of information science and technology*. 2. vyd. Hershey, PA: Information Science Reference, 2009, 8 v. ISBN 978-160-5660-271.

KIMBALL, Ralph a Joe CASERTA. *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. 1.vyd. Indianapolis, IN: Wiley, 2004, 491 s. ISBN 07-645-6757-8.

KOCÁBEK, Tomáš. *Návrh a implementace business intelligence řešení*. Praha, 2012. Dostupné také z: http://is.bivs.cz/th/6940/bivs_m/Navrh_a_implementace_BI_reseni_Kocabek_Tomas.doc?so=nx;info=. Diplomová práce. Bankovní institut vysoká škola Praha. Vedoucí práce Michal Valenta.

KUBÁN, Michal. *ETL nástroje*. Brno, 2013. Dostupné také z: http://is.muni.cz/th/373858/fi_b/tlac_BC.pdf. Bakalářská práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Jaroslav Bayer.

LACHLAN, James. OLAP cubes, outdated BI technology? *YellowfinBI* [online]. 2010, 2010-10-14 [cit. 2015-06-22]. Dostupné z: http://www.yellowfinbi.com/YFCommunityNews-OLAP-cubes-outdated-BI-technology-99879

LANE, Kin. Bringing ETL to the Masses with APIs. *API Evangelist* [online]. 2013, 2013-02-10 [cit. 2015-08-15]. Dostupné z: http://apievangelist.com/2013/02/10/bringing-etl-to-the-masses-with-apis/

LANEY, Douglas a Ehtisham ZAIDI. 100 Information and Analytics Predictions Through 2020: Free preview of Gartner research. *Gartner* [online]. 2015, 2015-01-30 [cit. 2015-07-15]. Dostupné z: https://www.gartner.com/doc/2974431?ref=SiteSearch&sthkw=Public%20Cloud%20Services%2C%20Worldwide%2C&fnl=search&srcId=1-3478922254#86757108

MANAGEMENT MANIA. *Metrics* [online]. 2013, 2013-06-06 [cit. 2015-08-05]. Dostupné z: https://managementmania.com/en/metrics

NEGASH, Solomon. Business intelligence. *Communications of the Association for Information Systems* [online]. 2004, (13) [cit. 2015-06-21]. Dostupné z: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3234&context=cais

NISO. *Understanding metadata*. Bethesda, MD: NISO, 2004. ISBN 18-801-2462-9. Dostupné také z: http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Noctuint. Pět trendů v Business Intelligence pro rok 2014. *Noctuint* [online]. 2014, 2014-06-06 [cit. 2015-06-22]. Dostupné z: http://www.noctuint.cz/blog/2014-06-06-bi-trends

NOVOTNÝ, Ota, Slánský, David a Pour, Jan. *Business intelligence: jak využít bohatství ve vašich datech.* 1. vyd. Praha: Grada, 2005, 254 s. ISBN 80-247-1094-3.

OneMetric. Cloud Computing. *OneMetric* [online]. 2015 [cit. 2015-07-17]. Dostupné z: http://www.onemetric.com.au/IT-Services/Cloud-Computing

PETERKA, Miloslav. Seznamte se s BI. *DAQUAS* [online]. 2010, 2010-06-09 [cit. 2015-06-22]. Dostupné z: http://www.daquas.cz/Articles/379-seznamte-se-s-bi.aspx

POWER, D. J. *A Brief History of Decision Support Systems* [online]. 2007, [cit. 2015-06-21]. Dostupné z: http://dssresources.com/history/dsshistory.html

R CORE TEAM. *R Language Definition* [online]. 2015 [cit. 2015-07-07]. Dostupné z: ftp://155.232.191.133/cran/doc/manuals/r-devel/R-lang.pdf

RAHM, Erhard; DO, Hong Hai. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull* [online]. 2000, 23.4: 3-13. [cit. 2015-06-22]. Dostupné z: https://www.informatik.hu-berlin.de/de/forschung/gebiete/ki/mac/lehre/lehrmaterial/Informationsintegration/Rahm00.pdf

SAATY, R.W. The analytic hierarchy process—what it is and how it is used. *Mathematical Modelling* [online]. 1987, **9**(3-5): 161-176 [cit. 2015-08-15]. DOI: 10.1016/0270-0255(87)90473-8. ISSN 02700255. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/0270025587904738

SAATY, Thomas L. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology* [online]. 1977, 15(3): 234-281 [cit. 2015-07-10]. DOI: 10.1016/0022-2496(77)90033-5. Dostupné z: Dostupné z: http://wenku.baidu.com/view/d37a355a804d2b160b4ec05d

SAATY, Thomas L. The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making. *Multiple Criteria Decision Analysis: State of the Art Surveys* [online]. New York: Springer-Verlag, 2005, : 345 [cit. 2015-06-29]. DOI: 10.1007/0-387-23081-5_9. ISBN 0-387-23067-x. Dostupné z: http://link.springer.com/10.1007/0-387-23081-5_9

SALLAM, Rita L., Bill HOSTMANN, Kurt SCHLEGEL, et al. Magic Quadrant for Business Intelligence and Analytics Platforms. *Gartner* [online]. 2015, 2015-02-23 [cit. 2015-07-20]. Dostupné z: http://www.gartner.com/technology/reprints.do?id=1-2ACLP1P&ct=150220&st=sb

SETHI, Manya. Data Warehousing And OLAP Technology. *International Journal of Engineering Research and Applications (IJERA)* [online]. 2012, 2.2: 955-960. [cit. 2015-06-22]. ISSN, 2248-9622. Dostupné z: http://www.ijera.com/papers/Vol2_issue2/FD22955960.pdf

SHUMELI, Galit, Nitin R PATEL a Peter C BRUCE. Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner. Hoboken, N.J.: Wiley, 2010, xxiv, 404 p. ISBN 9780470526828.

Solutions. Datové sklady a OLAP. *Solutions* [online]. 2002, 2002-10-28 [cit. 2015-06-22]. Dostupné z: http://datamining.xf.cz/view.php?cisloclanku=2002102808

SUEHRING, Steve. *MySQL bible*. New York, NY: Wiley Pub., c2002, xxviii, 686 p. ISBN 07-645-4932-4. Dostupné také z: http://www.chettinadtech.ac.in/g_article/Textbook2%20%20MySQL%20Bible.pdf

TURBAN, Efraim. *Business intelligence: a managerial approach*. 2nd ed. Boston: Prentice Hall, c2011, xx, 292 s. ISBN 978-0-13-610066-9.

VAISMAN, Alejandro a Esteban ZIMÁNYI. *Data Warehouse Systems Design and Implementation* [online]. Aufl. 2014. Berlin: Springer Berlin, 2014 [cit. 2015-06-21]. ISBN 36-425-4654-4. Dostupné z: http://link.springer.com/chapter/10.1007/978-3-642-54655-6_8

VASSILIADIS, Panos a Timos SELLIS. A survey of logical models for OLAP databases. *ACM SIGMOD Record* [online]. 1999, 28(4): 64-69 [cit. 2015-08-02]. DOI: 10.1145/344816.344869. ISSN 01635808. Dostupné z: http://portal.acm.org/citation.cfm?doid=344816.344869

YAN, Nancy. *Systems to provide data visualization and business process action in an on-demand enterprise dashboard*. U.S. Patent Application 12/820,810, 2010.

ZHANG, Qi, Lu CHENG a Raouf BOUTABA. Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications* [online]. 2010, **1**(1): 7-18 [cit. 2015-07-17]. DOI: 10.1007/s13174-010-0007-6. ISSN 1867-4828. Dostupné z: http://www.springerlink.com/index/10.1007/s13174-010-0007-6

ZIKMUNDA, Martin. Současné trendy v oboru business intelligence: Samoobslužnost, přirozené dotazování, information discovery, mobilita. *SystemOnLine*[online]. 2014 [cit. 2015-06-22]. Dostupné z: http://www.systemonline.cz/business-intelligence/soucasne-trendy-v-oboru-business-intelligence.htm

Přílohy

Příloha č. 1 - Tabulky hodnocení vah kritérií.
Zdroj: Vlastní práce autora.

Příloha č. 2 – Dokumentace datové pumpy v nástroji CC pro datový set *tarifikace.*
Zdroj: Vlastní práce autora (prostředí CC).

Příloha č. 3 – Dokumentace datové pumpy v nástroji CC pro datový set *tv-spots*.
Zdroj: Vlastní práce autora (prostředí CC).

Příloha č. 4 – Dokumentace datové pumpy v nástroji CC pro datový set *hovory.*
Zdroj: Vlastní práce autora (prostředí CC).

Příloha č. 5 – Dokumentace datové pumpy v nástroji KBC pro datový set *tarifikace.*
Zdroj: Vlastní práce autora (prostředí KBC).

Příloha č. 6 – Dokumentace datové pumpy v nástroji KBC pro datový set *tv-spots*.
Zdroj: Vlastní práce autora (prostředí KBC).

Příloha č. 7 – Dokumentace datové pumpy v nástroji KBC pro datový set *hovory.*
Zdroj: Vlastní práce autora (prostředí KBC).

Příloha č. 8 - LDM řešení pomocí CC.
Zdroj: Vlastní práce autora (prostředí GD).

Příloha č. 9 - LDM řešení pomocí KBC.
Zdroj: Vlastní práce autora (prostředí GD).

Příloha č. 10 - Dashboard k Otázce č. 1.
Zdroj: Vlastní práce autora podle dat 1188 (prostředí GD).

Příloha č. 11 - Dashboard k Otázce č. 2.
Zdroj: Vlastní práce autora podle dat 1188 (prostředí GD).

Příloha č. 12 - Dashboard k Otázce č. 3.

Zdroj: Vlastní práce autora podle dat 1188 (prostředí GD).

Příloha č. 13 – Dokumentace úspěšného nahrání dat do GD pomocí obou nástrojů.
Zdroj: Vlastní práce autora (prostředí GD).

Příloha č. 14 – Výsledky modelování v programu Expert Choice 2000.

Zdroj: Vlastní práce autora (prostředí Expert Choice 2000).

Příloha č. 15 – Zadání diplomové práce

Tabulky hodnocení vah kritérií

Tabulka 7 Párové hodnocení vah skupin kritérií.

Kritérium	K_1	K_2	K_3	s_i	r_i	w_i
K_1	1.00	1.00	2.00	2.00	1.26	0.40
K_2	1.00	1.00	2.00	2.00	1.26	0.40
K_3	0.50	0.50	1.00	0.25	0.63	0.20

Zdroj: Vlastní práce autora. (K_1 = Funkční kritéria, K_2 = Technická kritéria, K_3 = Obecná kritéria)

Tabulka 8 Hodnocení vah skupiny Funkčních kritérií.

Kritérium	K_1	K_2	K_3	K_4	K_5	s_i	r_i	w_i
K_1	1.00	1.00	2.00	3.00	3.00	18.00	1.78	0.31
K_2	1.00	1.00	2.00	3.00	3.00	18.00	1.78	0.31
K_3	0.50	0.50	1.00	3.00	3.00	2.25	1.18	0.20
K_4	0.33	0.33	0.33	1.00	1.00	0.04	0.52	0.09
K_5	0.33	0.33	0.33	1.00	1.00	0.04	0.52	0.09

Zdroj: Vlastní práce autora. (K_1= Náročnost použití, K_2 = Funkcionalita, K_3 = Odstraňování chyb, K_4 = Řešení práv, K_5 = Přívětivost uživatelského rozhraní)

Tabulka 9 Hodnocení vah skupiny Technická kritéria.

Kritérium	K_1	K_2	K_3	K_4	K_5	s_i	r_i	w_i
K_1	1.00	1.00	2.00	1.00	4.00	8.00	1.52	0.28
K_2	1.00	1.00	2.00	2.00	2.00	8.00	1.52	0.28
K_3	0.50	0.50	1.00	2.00	2.00	1.00	1.00	0.19
K_4	1.00	0.50	0.50	1.00	2.00	0.50	0.87	0.16
K_5	0.25	0.50	0.50	0.5	1.00	0.03	0.50	0.09

Zdroj: Vlastní práce autora. (K1= Architektura, K2 = Podpora formátů, K3 = Příprava dat a čištění, K4 = Rychlost nahrání, K5 = Konektivita)

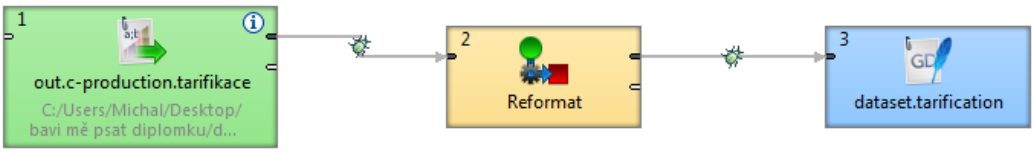
Tabulka 10 Hodnocení vah skupiny Obecná kritéria.

Kritérium	K_1	K_2	K_3	K_4	K_5	s_i	r_i	w_i
K_1	1.00	1.00	3.00	1.00	3.00	9.00	1.55	0.28
K_2	1.00	1.00	3.00	1.00	3.00	9.00	1.55	0.28
K_3	0.33	0.33	1.00	3.00	3.00	1.00	1.00	0.18
K_4	1.00	1.00	0.33	1.00	2.00	0.67	0.92	0.17
K_5	0.33	0.33	0.33	0.50	1.00	0.02	0.45	0.08

Zdroj: Vlastní práce autora. (K1= Možnost opětovného použití, K2 = Bezpečnost a stabilita, K3 = Cena, K4 = Rozšiřitelnost, K5 = Podpora)

Dokumentace datové pumpy v nástroji CC pro datový set *tarifikace*

Graf datové pumpy



Kód transformace

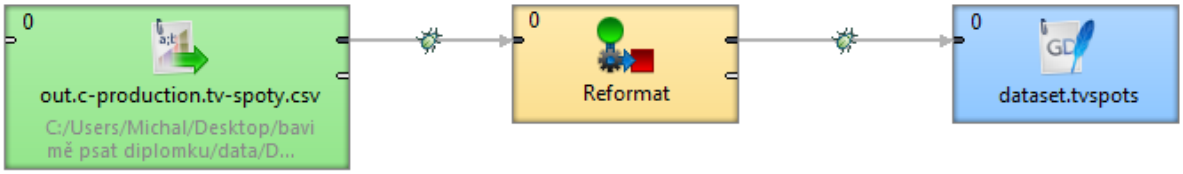
```
//#CTL2

string [] inputDate;

// Transforms input record into output record.
function integer transform() {
    $out.0.call_id = $in.0.call_id;
    $out.0.tarifikace_from = $in.0.tarifikace_from;
    $out.0.to = toString($in.0.to);
    inputDate = split($in.0.time_block, " ");
    $out.0.time_block = str2date(inputDate[0],'yyyy-MM-dd'):null;
    $out.0.time_dimension__time = substring(inputDate[inputDate.length() - 1],0,8):null;
    $out.0.services_name = $in.0.services_name;
    $out.0.length = $in.0.tarifikace_length;
    $out.0.length_9 = str2decimal($in.0.length_9):null;
    $out.0.length_c = str2decimal($in.0.length_C):null;
    $out.0.length_7 = str2decimal($in.0.length_7):null;
    $out.0.operatori_name = $in.0.operatori_name;
    $out.0.cancelled = $in.0.cancelled;
    $out.0.ivr = $in.0.ivr;


    return ALL;
}
```

Mapování datové pumpy do platformy GD

ATTRIBUTES	
call_id label.tarification.call_id	call_id
cancelled label.tarification.cancelled	cancelled
ivr label.tarification.ivr	ivr
operatori_name label.tarification.operatori_name	operatori_name
tarifikace_from label.tarification.tarifikace_from	tarifikace_from
to label.tarification.to	to
FACTS	
length_7 fact.tarification.length_7	length_7
length_9 fact.tarification.length_9	length_9
length_C fact.tarification.length_c	length_c
tarifikace_length fact.tarification.length	length
DATES	
Date (TimeBlock) datecreated	time_block

Dokumentace datové pumpy v nástroji CC pro datový set *tv-spots*

Graf datové pumpy



Kód transformace

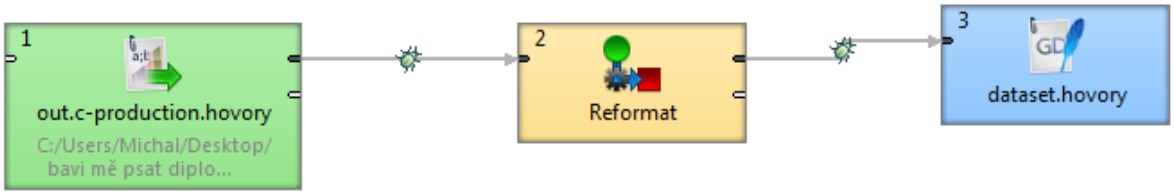
```
//#CTL2

// Transforms input record into output record.
function integer transform() {
    $out.0.primary_key = $in.0.primary_key;
    $out.0.footage = str2decimal($in.0.footage):null;
    $out.0.day_of_spot = $in.0.day_of_spot;
    $out.0.prime_time = $in.0.prime_time;
    $out.0.station = $in.0.station;
    $out.0.program_before = $in.0.program_before;
    $out.0.program_after = $in.0.program_after;
    $out.0.position = $in.0.position;
    $out.0.theme = $in.0.theme;
    $out.0.trps = str2decimal($in.0.trps):null;
    $out.0.grps = str2decimal($in.0.grps):null;
    $out.0.time_block = $in.0.date_start;

    return ALL;
}
```

Mapování datové pumpy do platformy GD

ATTRIBUTES	
Day Of Spot label.tvspots.day_of_spot	day_of_spot
Position label.tvspots.position	position
Prime Time label.tvspots.prime_time	prime_time
Program After label.tvspots.program_after	program_after
Program Before label.tvspots.program_before	program_before
Station label.tvspots.station	station
Theme label.tvspots.theme	theme
FACTS	
Footage fact.tvspots.footage	footage
Grps fact.tvspots.grps	grps
Trps fact.tvspots.trps	trps
DATES	
Date (TimeBlock) datecreated	time_block

Dokumentace datové pumpy v nástroji CC pro datový set *hovory*

Graf datové pumpy.



Kód transformace.

```
1  //#CTL2
2
3  // Transforms input record into output record.
4  function integer transform() {
5      string [] inputDate;
6
7
8      $out.0.primary_key = $in.0.primary_key;
9      $out.0.disposition = $in.0.disposition;
10     $out.0.agent_id = $in.0.agent_id;
11     $out.0.delay = $in.0.delay;
12     $out.0.handling_time = $in.0.handling_time;
13     $out.0.originator = $in.0.originator;
14     $out.0.operator = $in.0.operator;
15     $out.0.service = $in.0.service;
16     $out.0.cancelled = $in.0.cancelled;
17
18     inputDate = split($in.0.time_block, " ");
19     $out.0.time_block = inputDate[0];
20     $out.0.time = substring(inputDate[inputDate.length() - 1],0,8):null;
21
22     return ALL;
23 }
```

Mapování datové pumpy do platformy GD.

ATTRIBUTES	
Agent Id label.hovory.agent_id	agent_id
Cancelled label.hovory.cancelled	cancelled
Disposition label.hovory.disposition	disposition
Operator label.hovory.operator	operator
Originator label.hovory.originator	originator
Service label.hovory.service	service
primary_key label.hovory.primary_key	primary_key
FACTS	
delay fact.hovory.delay	delay
handling fact.hovory.handling_time	handling_time
DATES	
Date (TimeBlock) datecreated	time_block

Dokumentace datové pumpy v nástroji KBC pro datový set *tarifikace*

Graf datové pumpy

Transformace a mapování datové pumpy do platformy GD

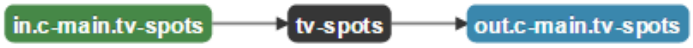
Writers / GoodData - writerGoodData / out.c-main.tarifikace

Column	GoodData Name	Type	
call_id	call_id	CONNECTION_POINT	
tarifikace_length	tarifikace_length	FACT	
tarifikace_from	tarifikace_from	ATTRIBUTE	
to	to	ATTRIBUTE	
time_block	time_block	DATE	yyyy-MM-dd HH:mm:ss DateKBC
services_name	services_name	IGNORE	
length_9	length_9	FACT	DECIMAL(12,2)
length_C	length_C	FACT	DECIMAL(12,2)
length_7	length_7	FACT	DECIMAL(12,2)
operatori_name	operatori_name	ATTRIBUTE	
cancelled	cancelled	ATTRIBUTE	
ivr	ivr	ATTRIBUTE	

Edit Columns

Dokumentace datové pumpy v nástroji KBC pro datový set *tv-spots*

Graf datové pumpy



Transformace a mapování datové pumpy do platformy GD

Writers / GoodData - writerGoodData / out.c-main.tv-spots

Column	GoodData Name	Type	
primary_key	primary_key	IGNORE	
date_time_start	Date	DATE	yyyy-MM-dd HH:mm:ss DateKBC
date_time_end	date_time_end	IGNORE	
footage	footage	FACT	
day_of_spot	Day Of Spot	ATTRIBUTE	
prime_time	Prime Time	ATTRIBUTE	
station	Station	ATTRIBUTE	
program_before	Program Before	ATTRIBUTE	
program_after	Program After	ATTRIBUTE	
position	Position	ATTRIBUTE	
theme	Theme	ATTRIBUTE	
trps	trps	FACT	
grps	grps	FACT	

Edit Columns

Dokumentace datové pumpy v nástroji KBC pro datový set *hovory*

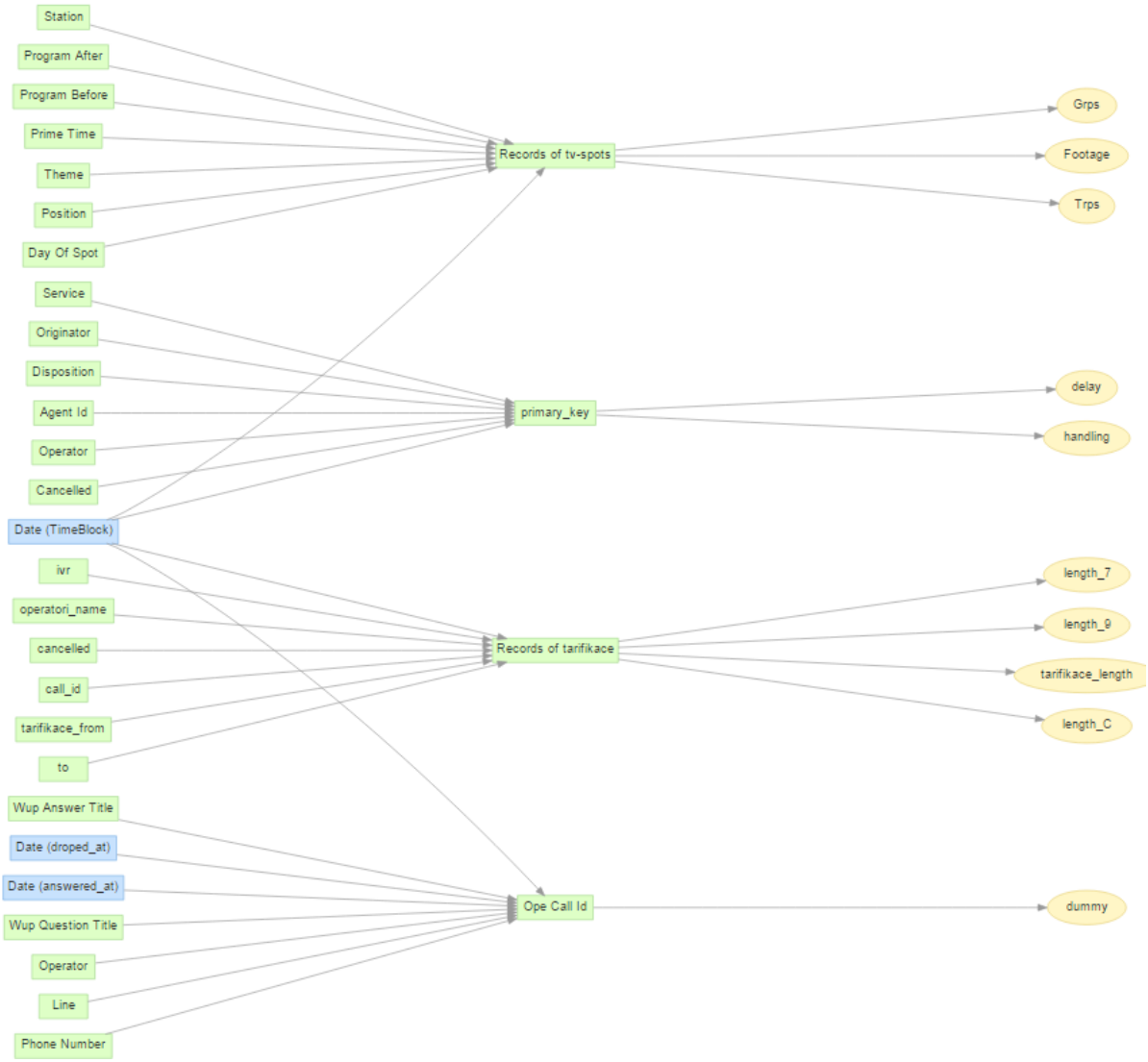
Graf datové pumpy

Transformace a mapování datové pumpy do platformy GD

Column	GoodData Name	Type	
primary_key	primary_key	ATTRIBUTE	
disposition	disposition	IGNORE	
agent_id	Agent Id	ATTRIBUTE	
delay	Delay	FACT	DECIMAL(12,2)
time_block	time_block	DATE	yyyy-MM-dd HH:mm:ss DateKBC
handling_time	Handling	FACT	DECIMAL(12,2)
originator	Originator	ATTRIBUTE	
international	international	IGNORE	
time_segment_id	time_segment_id	IGNORE	
operator	Operator	ATTRIBUTE	
service_id	Service	ATTRIBUTE	
wup_question_id	wup_question_id	IGNORE	
wup_answer_id	wup_answer_id	IGNORE	
cancelled	Cancelled	ATTRIBUTE	

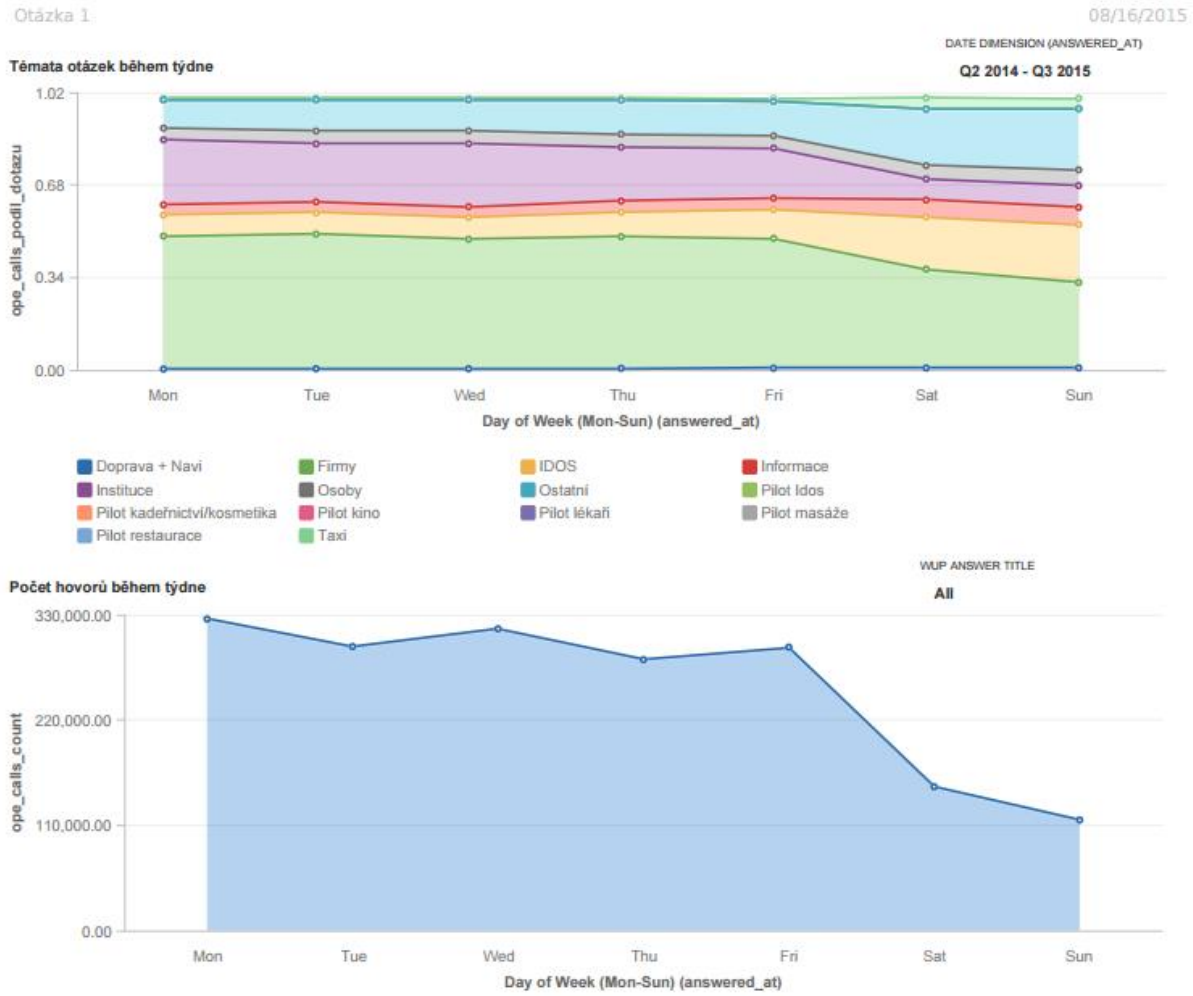
Edit Columns

LDM řešení pomocí CC

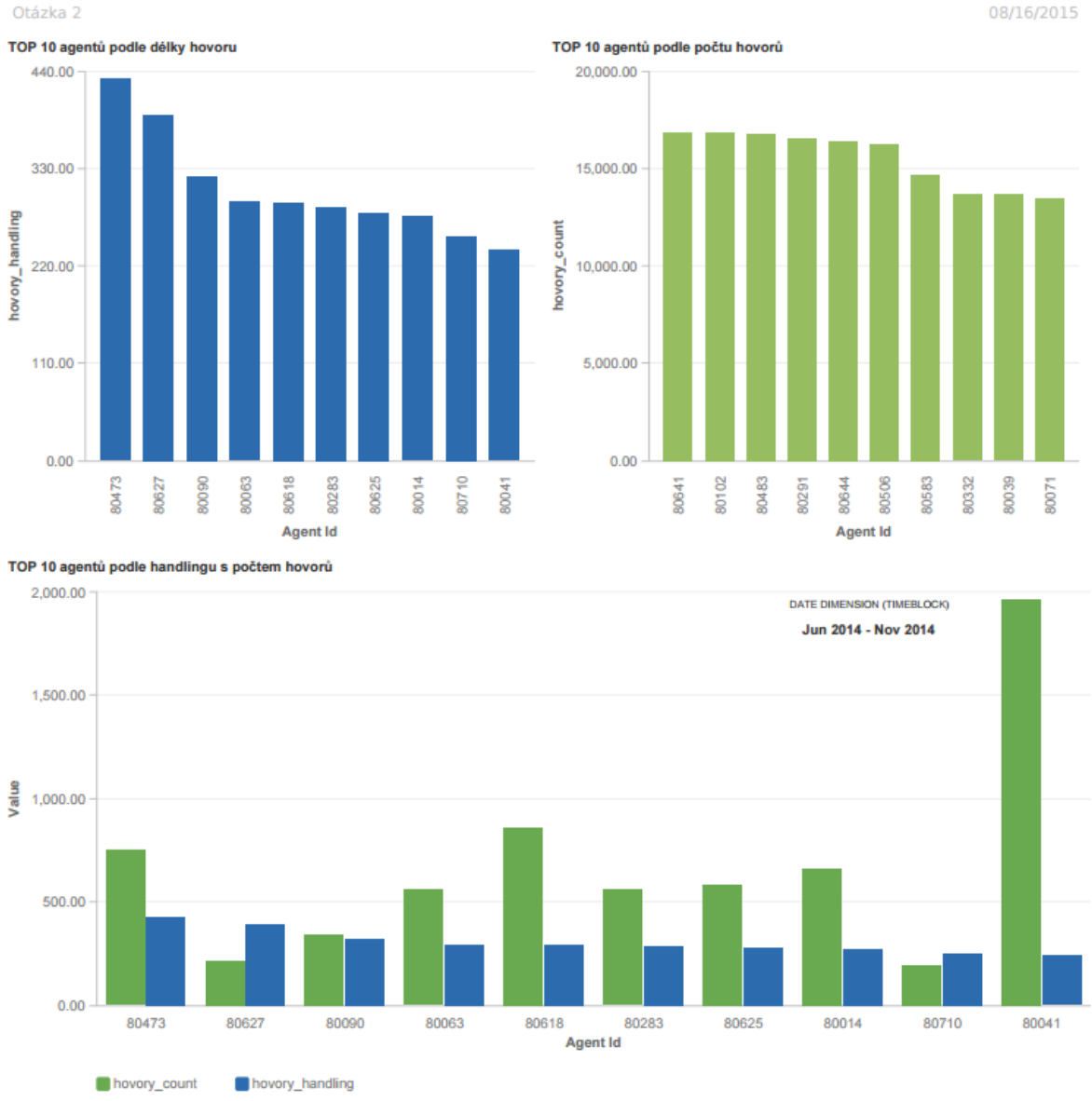


LDM řešení pomocí KBC

Dashboard k Otázce č. 1

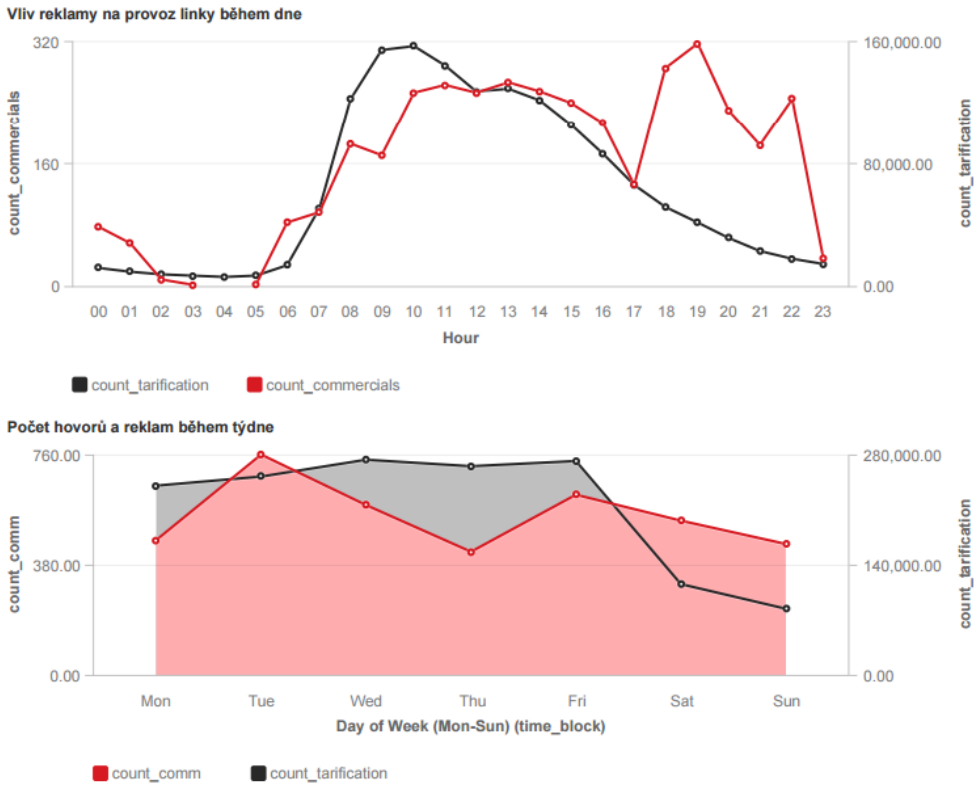


Dashboard k Otázce č. 2



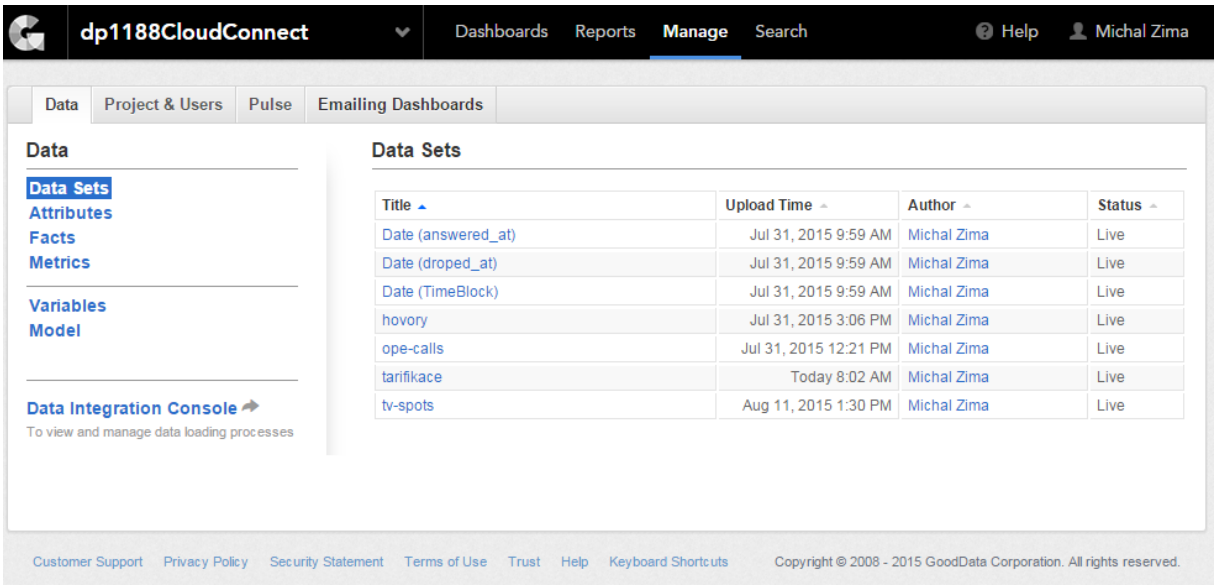
Dashboard k Otázce č. 3

Otázka 3 08/16/2015

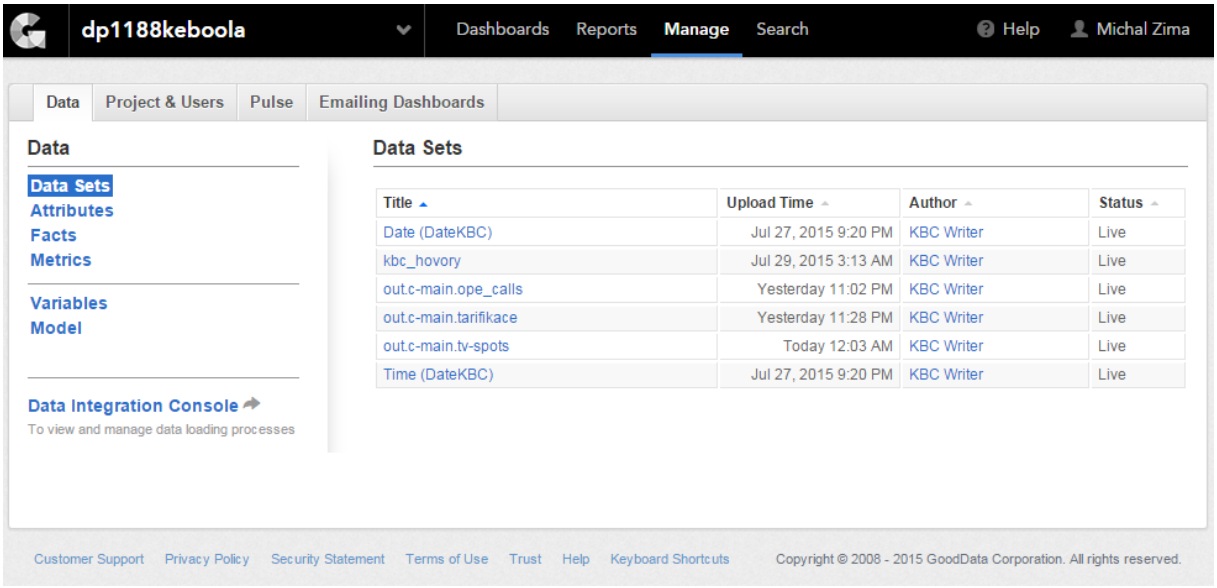


Dokumentace úspěšného nahrání dat do GD pomocí obou nástrojů

Seznam datových setů v řešení pro CC



Seznam datových setů v řešení pro KBC



AHP model pro výběr ETL nástroje v programu Expert Choice 2000

Hodnocení skupiny Funkční kritéria s mírou nekonzistence 0,02

	Náročnost použití	Funkcionalita	Odstraňování	Řešení prá	Přívětivost
Náročnost použití		1,0	2,0	3,0	3,0
Funkcionalita			2,0	3,0	3,0
Odstraňování chyb				3,0	3,0
Řešení práv					1,0
Přívětivost uživatelského rozhraní	Incon: 0,02				

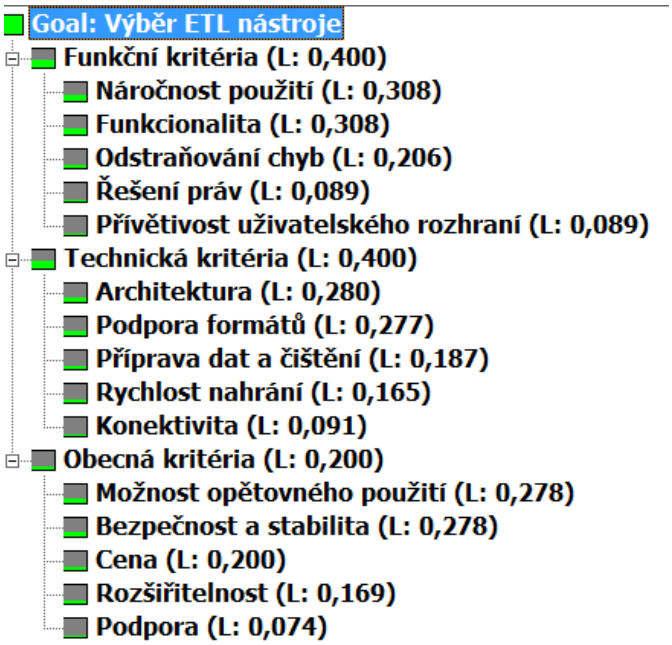
Hodnocení skupiny Technická kritéria s mírou nekonzistence 0,05

	Architektura	Podpora form	Příprava dat a	Rychlost n	Konektivita
Architektura		1,0	2,0	1,0	4,0
Podpora formátů			2,0	2,0	2,0
Příprava dat a čištění				2,0	2,0
Rychlost nahrání					2,0
Konektivita	Incon: 0,05				

Hodnocení skupiny Obecná kritéria s mírou nekonzistence 0,13

	Možnost opětovní	Bezpečnost a	Cena	Rozšiřitelnost	Podpora
Možnost opětovného použití		1,0	3,0	1,0	3,0
Bezpečnost a stabilita			3,0	1,0	3,0
Cena				3,0	3,0
Rozšiřitelnost					2,0
Podpora	Incon: 0,13				

Kompletní AHP model včetně vah kritérií



Zadání diplomové práce

FIM UHK
UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Rokitanského 62, 500 03 Hradec Králové, tel: 493 331 111, fax: 493 332 235

Zadání k závěrečné práci

Jméno a příjmení studenta: **Michal Zima**

Obor studia: Informační management (5)
Jméno a příjmení vedoucího práce: **Karel Mls**

Název práce:
Výběr ETL nástroje pro cloudové řešení BI

Název práce v AJ:
Choosing the ETL tool for cloud-based BI solution

Podtitul práce:

Podtitul práce v AJ:

Cíl práce: Teoretická část má za cíl shrnout možnosti využití ETL nástrojů v závislosti na BI nástrojích a popsat jejich princip. Cílem praktické části je experimentální ověření možností nahrání dat do řešení BI v cloudu.

Osnova práce:
1. Úvod

2. Teoretické vymezení dané problematiky

3. Formulace řešeného problému a cíl práce

4. Popis výzkumu

5. Shrnutí výsledků

6. Závěry a doporučení

7. Seznam použité literatury

8. Přílohy

Projednáno dne: 22. 10. 2014

Podpis studenta

Podpis vedoucího práce