

UNIVERZITET DŽEMAL BIJEDIĆ U MOSTARU
FAKULTET INFORMACIJSKIH TEHNOLOGIJA MOSTAR

SEMINAR

Poslovna inteligencija bazirana na "open source" software-u

Student: *Ernad Husremović, DL 2792*

Mentor: *prof.dr Vanja Bevanda*

ver: 1.9.7

Mostar, februar 2012.

SADRŽAJ

1. Uvod	1
1.1. Poslovna inteligencija (BI)	1
1.1.1. Operativni podaci, skladišta podataka	1
1.2. Poslovni motivi za realizaciju sistema BI-a	2
1.2.1. OLAP kocka	3
1.2.2. Alati za integraciju podataka (DI)	3
1.2.3. Rudarenje podataka	4
2. ‘Open source’ BI software	5
2.1. Pentaho BI	5
2.2. Pentaho komponente	6
2.2.1. Mondrian OLAP	6
2.2.2. Mondrian OLAP schema	6
2.3. Pentaho ETL dizajner ‘Spoon’	7
2.3.1. OALP Analyzer	7
3. OLAP Case study	9
4. Iza case study-ja ?	18
4.1. Ne znam	18
4.1.1. dimension table	18
4.1.2. facts table	18
4.1.3. ETL (Extract Transform Load)	18
4.2. Analiza podataka	18
5. Zaključak	22
5.0.1. Ekspert	22
6. Literatura	23

A. Izvorni kod, dostupni resursi	25
B. Bilješke	26

1. Uvod

1.1. Poslovna inteligencija (BI)

Poslovna inteligencija (nadalje BI¹) se primarno odnosi na računarski bazirane tehnike identificiranja, ekstrakcije i analize poslovnih podataka. BI tehnologije obezbeđuju pregled ranijih i tekućih poslovnih operacije, kao i predviđanje budućih trendova u poslovanju² Wikipedia (2012a).

Standardne funkcije BI-a su:

1. izvještavanje (reporting)
2. analitičko procesiranje (analytical processing)
3. rudaranje podataka (data mining)
4. prediktivne analize (predictive analytics)

Sve ove funkcije pomažu poslovnom odlučivanju i planiranju - kako strateškom³ tako i operativnom planiranju⁴.

1.1.1. Operativni podaci, skladišta podataka

BI kao izvor podataka koristi skladišta podataka.

Skladišta podataka treba razlikovati od operativnih podataka. Operativni podaci - podaci o tekućem poslovanju nalaze se unutar poslovnih aplikacija (ERP)⁵.

ERP sistemi pohranjuju poslovne transakcije (poslovne dokumente) u realnom vremenu.

¹BI Business intelligence

²prediktivna analiza

³Odlučivanje 'top' managera

⁴'department' manageri

⁵ERP - Enterprise Resource Planning software, software za podršku tekućem poslovanju

ERP sistemi sadrže sisteme izvještavanje⁶ koji su primarno usmjereni na davanje podataka o tekućem poslovanju⁷.

Operativni podaci su glavni izvor za gradnju skladišta podataka. Međutim, skladišta podataka se najčešće grade iz heterogenih izvora.

Iako slični po načinu konstrukcije, u BI terminologije se pravi distinkcija između 'data mart' i 'data warehouse' skladišta podataka⁸:

Data mart (DMart) sadrži informacije o jednom dijelu organizacije (npr. prodaja, ljudski resursi),

Data warehouse (DW) sadrži informacije iz više područja - obrađuje organizaciju globalno.

DW je stoga usmjeren na podršku 'top' menadžmenta, dok 'datamart' obezbjeđuje informacije za upravljanje i operativno planiranje pojedinih dijelova organizacije (Roldan, 2010, str. 391).

1.2. Poslovni motivi za realizaciju sistema BI-a

Analiza poslovnih podataka radi kvalitetnog poslovnog odlučivanja je postojala i prije informatičke podrške poslovanju.

Slijedeća poslovna pitanja su postavljana i prije informatičke ere:

- Kako se kretala prodaja određene grupe artikala u predhodnom periodu ?
- Koja grupa artikala se najviše zadržava na lageru ?
- Kakva je struktura prodaje po regijama ?
- Unutar koje grupacije artikala / proizvoda je najbolja marža ?

Sva ova pitanja praktično vrše analizu efekata poslovanja usljed različitih uticaja (multidimenzionalna analiza).

Iz svih gornjih pitanja mogu se uočiti dva tipa informacija:

- mjere - poslovni indikatori (prodaja, vrijednost zalihe, visina marže)
- dimenzije - atributi poslovanja (geografske dimenzije - grad, region, vremenske dimenzije, grupe artikala, rang cijena ...)

Ovo je bilo polazište za konstrukciju multidimenzionalnih skladišta podataka (OLAP cube).

⁶traditional reporting

⁷period tekuće poslovne godine

⁸U domaćoj literaturi se najčešće za oba pojma koristi termin 'skladište podataka'

1.2.1. OLAP kocka

OLAP kocka (OLAP cube⁹) je set podataka organizovanih na taj način da omogućavaju *nedeterminirane* upite nad agregiranim podacima, odnosno online analitičko procesiranje podataka Wikipedia (2012b).

Ovakva organizacija podataka omogućava OLAP klijentima pregled podataka u različitim varijantama.

Ono što RDBMS¹⁰ predstavlja za ERP sistem, OLAP kocka predstavlja za BI sistem.

Analogija postoji i u dijelu pretrage podataka:

- RDBMS <-> SQL structured query language
- OLAP cube <-> MDX - multidimensional query language

Današnje implementacije OLAP kocki:

- ROLAP - podaci smješteni u relacijske baze podataka
- MOLAP - podaci su u proprietary formatu prilagođenom procesiranju multidimenzionalnih struktura podataka

Pored gornjih postoje i hibridne implementacije OLAP kocki koje kombinuju obje tehnologije.

1.2.2. Alati za integraciju podataka (DI)

DI¹¹ alati omogućavaju vezu BI sistema sa "vanjskim" svijetom (vidi 1.1.1 operativni podaci)

Glavni dio DI 'toolset'-a su je ETL softver¹².

ETL softver obavlja sljedeće funkcije:

1. *Extract*: uzimanje podataka iz vanjskih izvora
2. *Transform*: izvrši transformaciju podataka u format koji je pogodan za pohranu u DW odnosno DMart (vidi 1.1.1)
3. *Load*: konačno snimi podatke 'prečišćene' u predhodnom koraku podatke u DW/DMart

⁹OLAP - online analytical processing

¹⁰RDBMS relational database management system

¹¹DI - data integration

¹²ETL - Extract/Transform/Load

1.2.3. Rudarenje podataka

Pojam rudaranje podataka se može definisati kao pronalaženje zakonitosti među podacima.

Jedna od definicija rudaranje glasi: rudarenje podataka je sistematičan, interaktivan i iterativan proces izvođenja i prikazivanja korisnika, implicitnog i inovativnog *znanja* iz podataka (Mršić, 2004, str. 40).

Unutar ovog rada nećemo se baviti ovim dijelom BI-a.

Open source software koji pokriva ove oblasti:

- Data mining ‘Weka’ projekat: University of Waikato (2012), Pentaho Community (2012)
- ‘R’ statistički paket foundation (2012)

Treba uočiti da je ‘Weka’ jedan od podprojekata ‘Pentaho BI suite’-a (vidi 2.2).

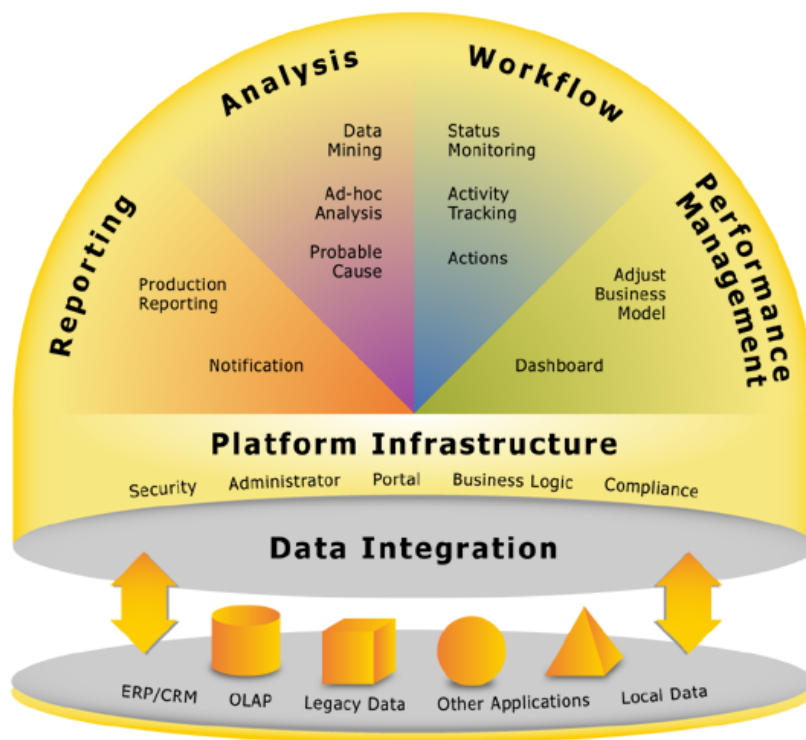
2. 'Open source' BI software

'Open source' software pokriva kompletan opseg alata

2.1. Pentaho BI

Pentaho BI obuhvata praktično funkcije BI-a (Roldan, 2010, str. 7):

1. multidimenzionalna analiza
2. reporting
3. 'dashboards' - prikaz glavnih poslovnih indikatora



Slika 2.1: Pentaho arhitektura (Bimonte i Wehrle (2007))

2.2. Pentaho komponente

2.2.1. Mondrian OLAP

Pentaho implementacija OLAP kocke naziva se 'Mondrian'. Mondrian je kocka ROLAP tipa.

Bitno je naglasiti da je 'Mondrian' XMLA kompatibilan provajder¹

2.2.2. Mondrian OLAP schema

Kao ROLAP implementacija, podaci se nalaze u relacijskog bazi podataka²

Kod konstrukcije sheme koriste se sljedeći pojmovi:

- 'dimension table' - tabela u kojoj su pohranjene dimenzije
- SCD slow changing dimension
- SCD Type I - čuva se samo jedna vrijednost dimenzije
- SCD Type II - čuva se istorija vrijednosti dimenzije kroz vremenski period ³
- 'facts table' - tabela u kojoj su mjere - poslovni indikatori koje analiziramo
- 'business key' (bk) - ključ koji koriste aplikacije za rukovanje operativnim podacima (ERP software)
- 'surogat key' (id) - ključ u bazi podataka koji koristi OLAP storage
- 'snowflake' schema - šema u kojoj su dimenzije u sopstvenim tabelama, dimenzije su visokog stepena normalizacije podataka⁴ Pentaho (2012)
- Type II

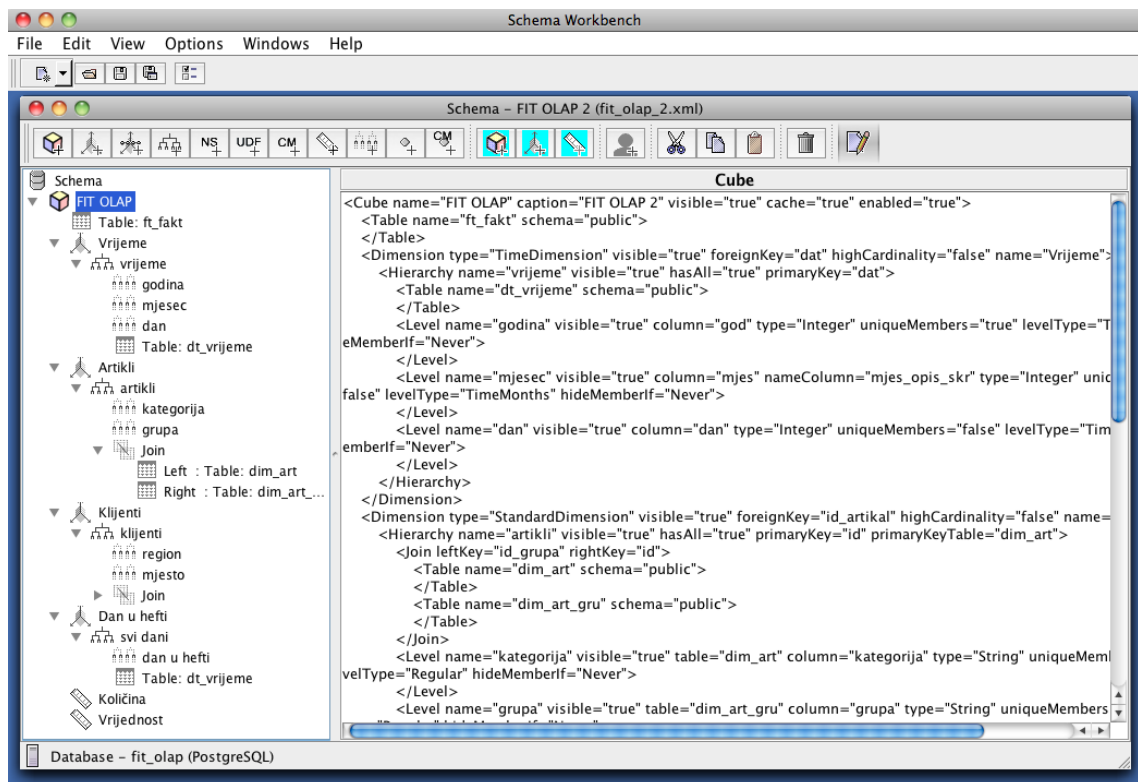
Pogledajmo jednu gotovu mondrian schemu (vidi 3 dobijenu u narednom *case study*-ju):

¹XML za Analizu (XMLA) je industrijski standard za pristup podacima u analitičkim sistemima kao što su OLAP i 'data mining'. Baziran je na drugim industrijskim standardima kao što su XML, SOAP i HTML Wikipedia (2012c).

²Podržane su sve JDBC podržane relacijske baze: PostgreSQL, MySQL, MSSQL, Oracle

³npr. cijena artikla se mijenja tokom vremena, SCD type II dimenzija za svaku novu cijenu pravi novi zapis u dimension tabeli

⁴normalizacija podataka u relacijskog bazi podataka



Slika 2.2: Mondrian schema OLAP 2 cube

2.3. Pentaho ETL dizajner ‘Spoon’

‘Spoon’ GUI aplikacija u kome se vrši definicija i testiranje ETL transformacija i ‘job’-ova.

Kod ETL operacija bitno je poznavati sljedeće pojmove:

- ‘cleansing’ - “čišćenje” podataka - ispravka (ili izbacivanje) netačnih podataka

2.3.1. OALP Analyzer

Pentaho sadrži dva ‘OLAP analyzer’⁵ rješenja:

- JPilot - starije rješenje, napušta se njegov razvoj⁶
- novi ‘OLAP Analyzer’ koji se nalazi samo u ‘Enterprise’ (plaćenj verziji)⁷

Kao ‘OLAP Analysis’ sofver korišten je Saiku.

Saiku je modularni open-soruce analitički sotver koji nudi jednostavnu OLAP analizu podataka analytical labs (2012)

⁵OLAP cube consumer

⁶označeno od Pentaho razvojnog tima kao ‘depreated’

⁷znači ova komponenta je ‘closed source’ software tako da se u ovom radu neće dalje razmatrati.

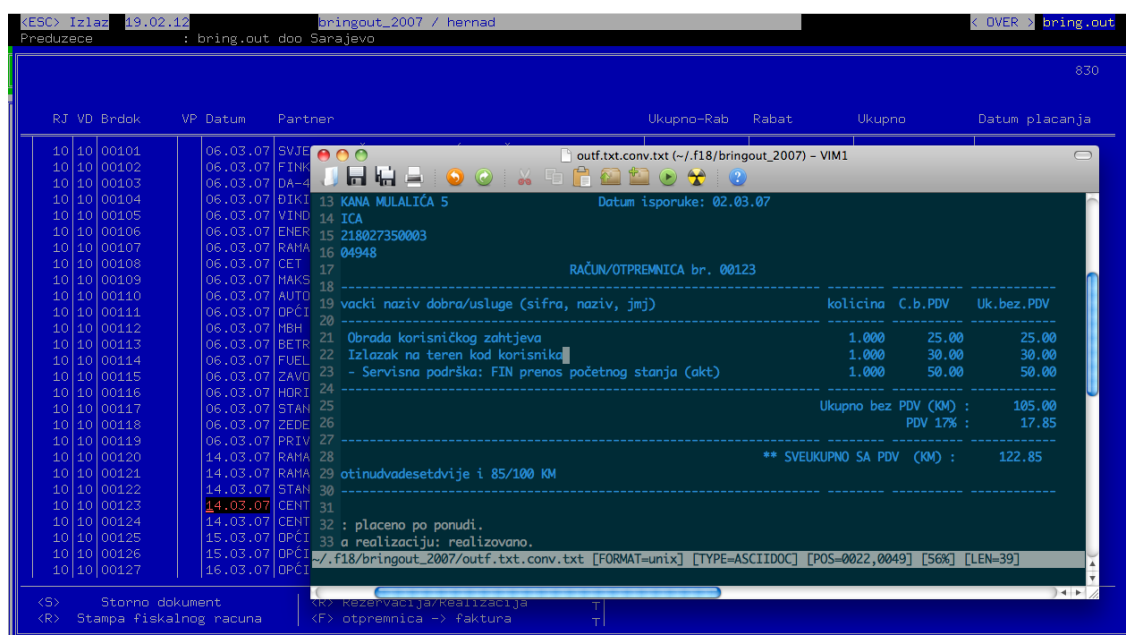
OLAP analitički softver barata sa sljedećim pojmovima:

- redovi - sadrži jednu ili više mjera ili dimenzija koje se prikazuju u redovima kod prikaza podataka
- kolona - sadrži jednu ili više mjera ili dimenzija koji se prikazuju u kolonama prikaza podataka
- filteri - ograničenje podataka po određenim vrijednostima dimenzija

3. OLAP Case study

U ovom 'case study'-ju ćemo izvršiti formiranje OLAP kocke za operativne podatke ERP aplikacije 'F18 knowhow'. U DMart ćemo staviti podatke za sve aktivne poslovne godine firme "bring.out" (1996-2011).

3.1. F18 knowhow ERP



Slika 3.1: ERP aplikacija, F18 klijent

3.2. Poslovni cilj analize

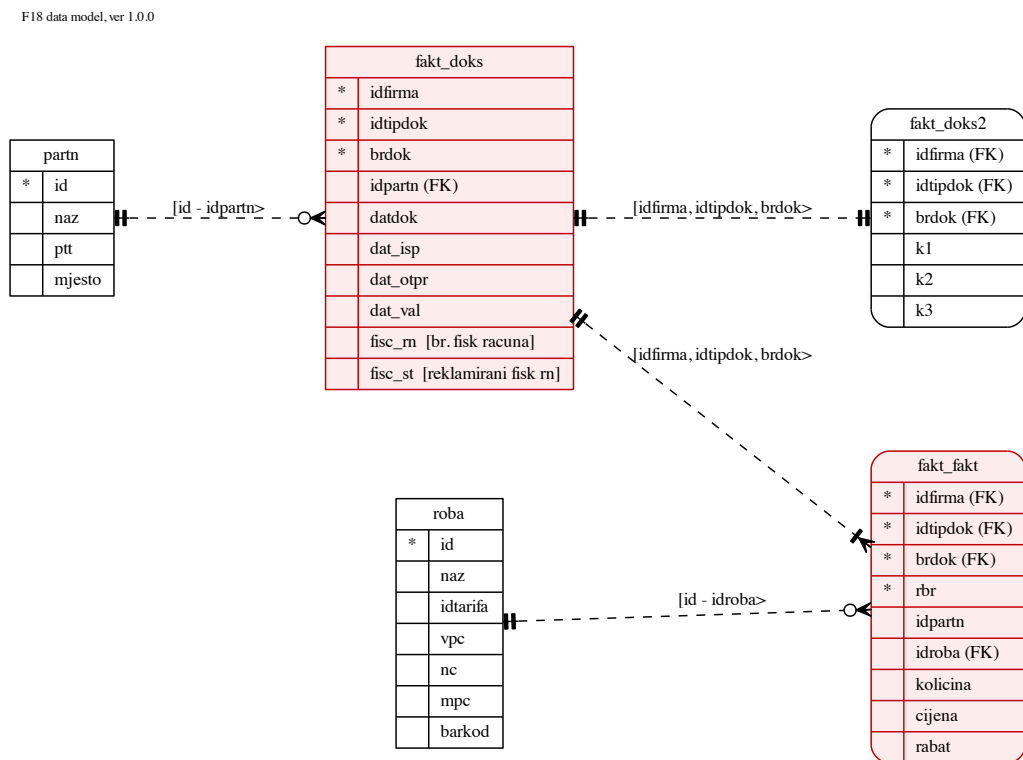
Cilj je analizirati prodaju firme po godinama pri čemu nas interesuje struktura klijenata po gradovima i regionima, te prodaja artikala po određenim kategorijama i grupama.

Glavni Operativni podaci nalaze se u ERP aplikaciji, u PostreSQL relacijskog bazi. Dio potrebnih dimenzija (regioni, kategorije i grupe artikala) nisu implementirani

unutar operativnih podataka, tako da analitičar mora unutar ETL procesa ove informacije dodati.

Podaci svake poslovne godine nalaze se u posebnoj bazi podataka.

Operativni podaci ‘F18 knowhow’ smješteni su u sljedeći relacijski model:



Slika 3.2: F18 transakcijski db model (relevantni dio)

F18 ‘cleansing’ podaci (Dodatak A, olap_cleansing ‘spreadsheet’ dokument)

Klasificiranje izvornih podataka - šifarnik artikala

	A	B	C	D	E	F	G	H
1	idroba4	kategorija	grupa					
2	7ADV	SW	3-rd party software					
3	7LXV	SW	3-rd party software					
4	7LXW	SW	3-rd party software					
5	7MSO	SW	3-rd party software					
6	9ADV	SW	3-rd party software					
7	9AV0	SW	3-rd party software					
8	9DOD	SW	3-rd party software					
9	9MSO	SW	3-rd party software					
10	9XBA	SW	3-rd party software					
11	MSWI	SW	3-rd party software					
12	WSER	HW	3-rd party software					
13	WXPB	SW	3-rd party software					
14	WXPB	SW	3-rd party software					
15	NSC-	HW	fiskalni uređaji					
16	6OBR	SW	fmk software					
17	9FMK	SW	fmk software					
18	9MOD	SW	fmk software					

Slika 3.3: F18 klasificiranje - šifarski sistem artikala

	A	B	C	D	E	F	G	H
1	Mjesto	Mjesto						
2	BEGOV HAN	Begov Han						
3	ŽELJEZNO POLJE	Begov Han						
4	BIHAĆ	Bihać						
5	BUGOJNO	Bugojno						
6	BUSOVAČA	Busovača						
7	DONJI VAKUF	Donji Vakuf						
8	GORA@DE	Goražde						
9	GORAŽDE	Goražde						
10	GORAZDE	Goražde						
11	U.S.A	inostranstvo						
12	U.S.A	inostranstvo						
13	DEUTSCHLAND	inostranstvo						
14	USA	inostranstvo						
15	KAKANJ	Kakanj						
16	KISELJAK	Kiseljak						
17	KONJIC	Konjic						
18	MAGLAJ	Maglaj						

Slika 3.4: 'cleansing' F18 podataka - klijenti - mjesta/gradovi

olap_cleansing.xls - LibreOffice Calc

Arial 10 B I U

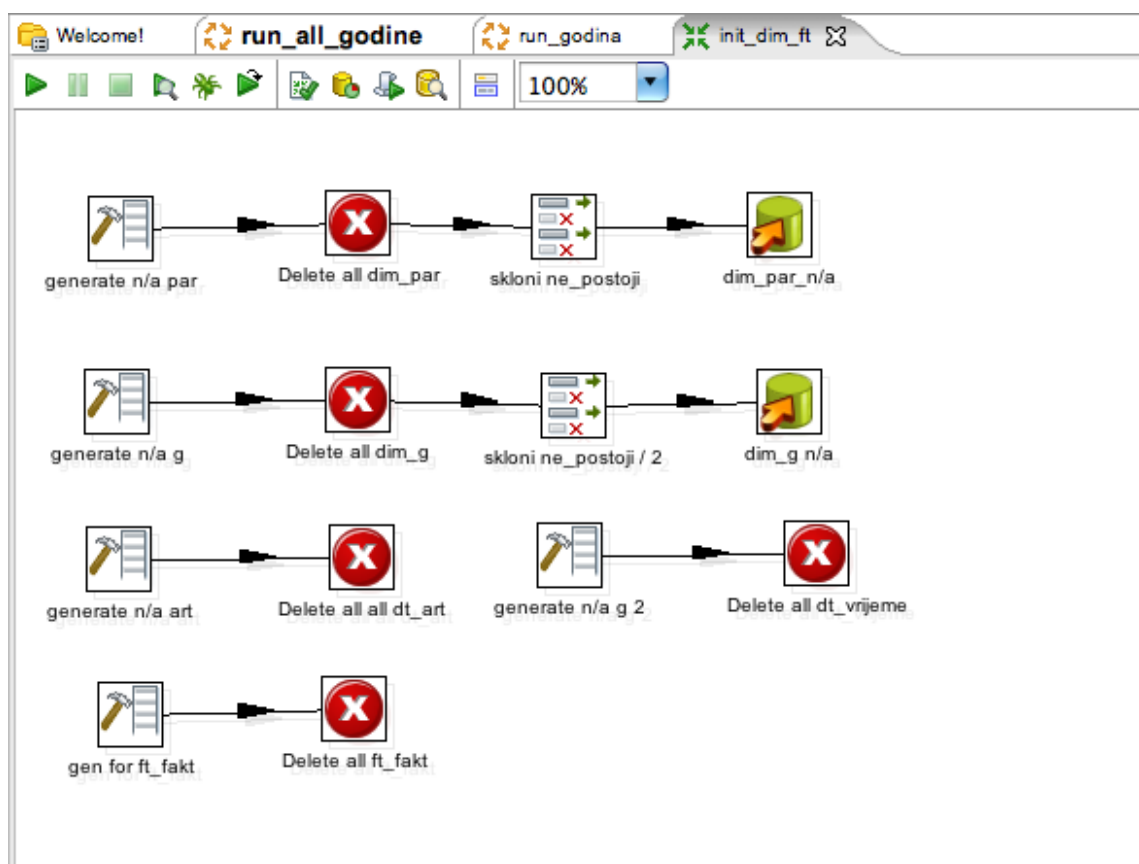
A1 Mjesto

	A	B	C	D	E	F	G	H
1	Mjesto	Oznaka						
2	Begov Han	ŽEP						
3	Bihać	BI						
4	Bugojno	BUG						
5	Busovača	VIT						
6	Donji Vakuf	BUG						
7	Goražde	GOR						
8	inostranstvo	INO						
9	Kakanj	ZE						
10	Kiseljak	KIS						
11	Konjic	KO						
12	Maglaj	ZAV						
13	n/a	XX						
14	Olovo	OLO						
15	Sanski Most	BI						
16	Sarajevo	SA						
17	Tešanj	TEŠ						
18	Travnik	TRA						

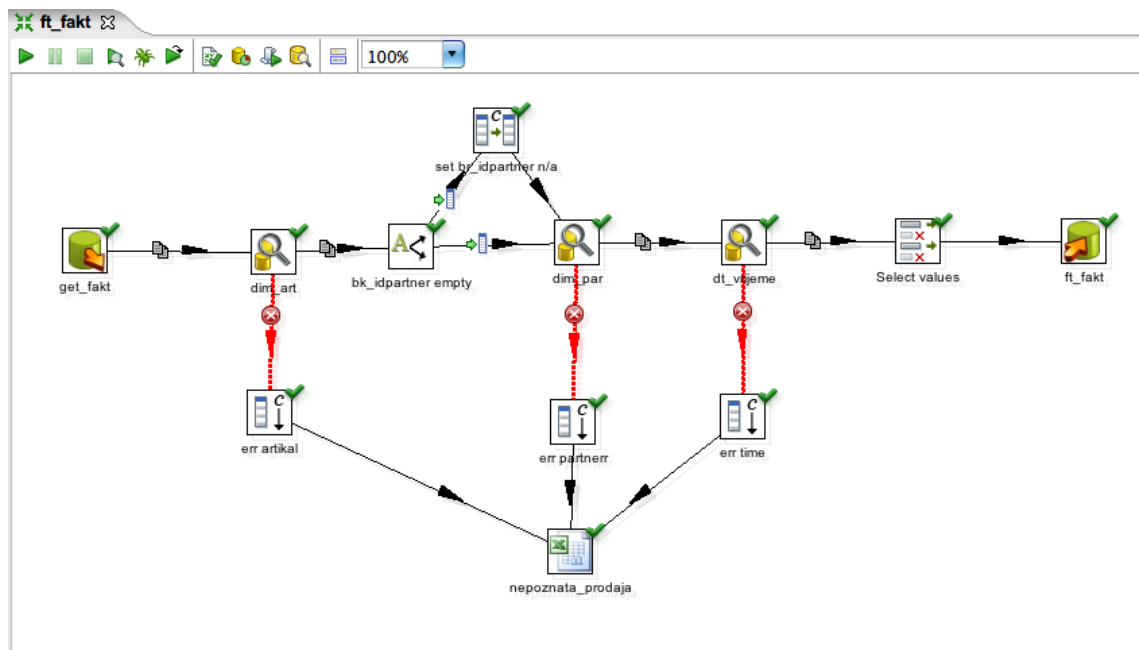
Find

Sheet 2 / 3 PageStyle_mjesto_kod STD Sum=0 100%

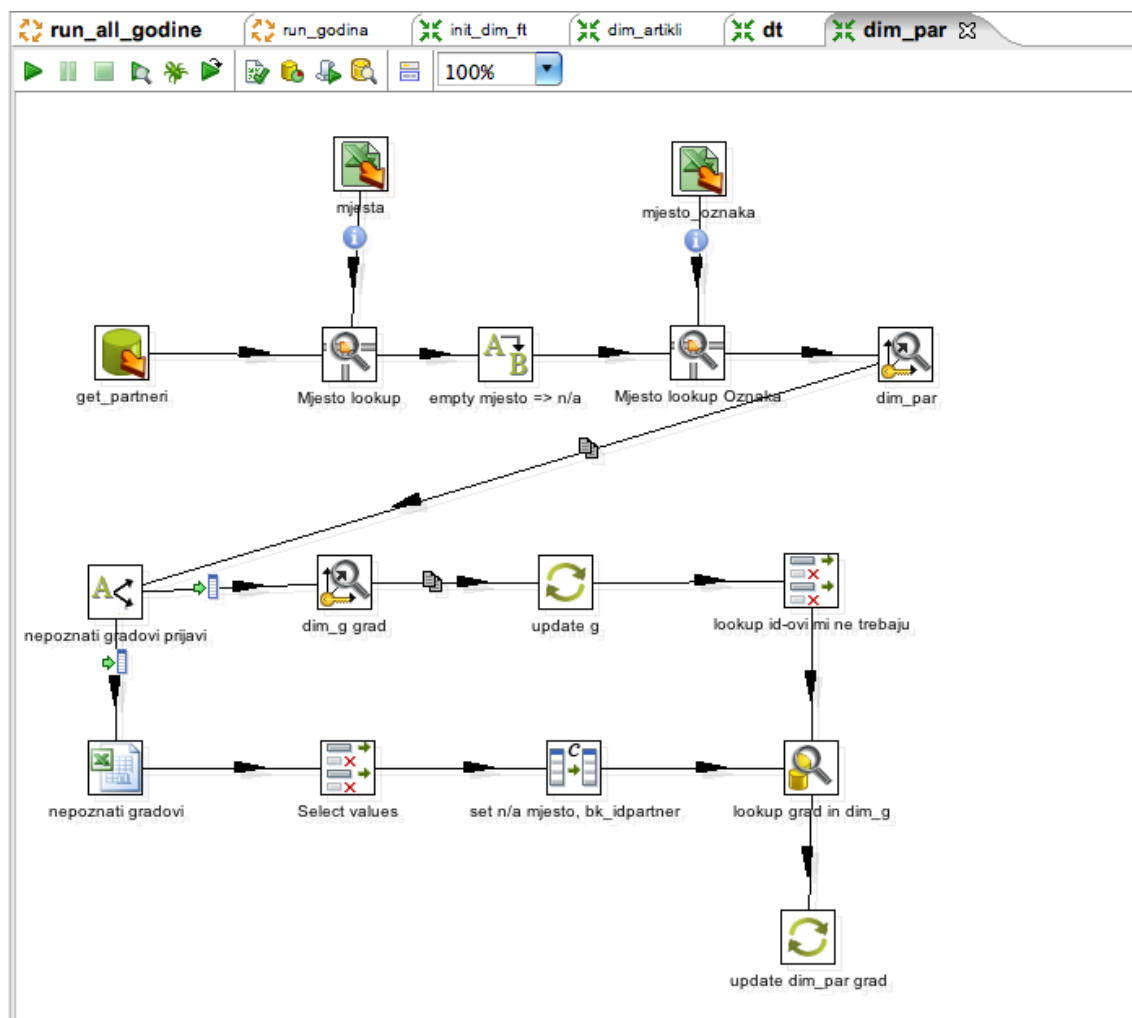
Slika 3.5: F18 kodiranje regiona - klasifikacija mjesta/gradova



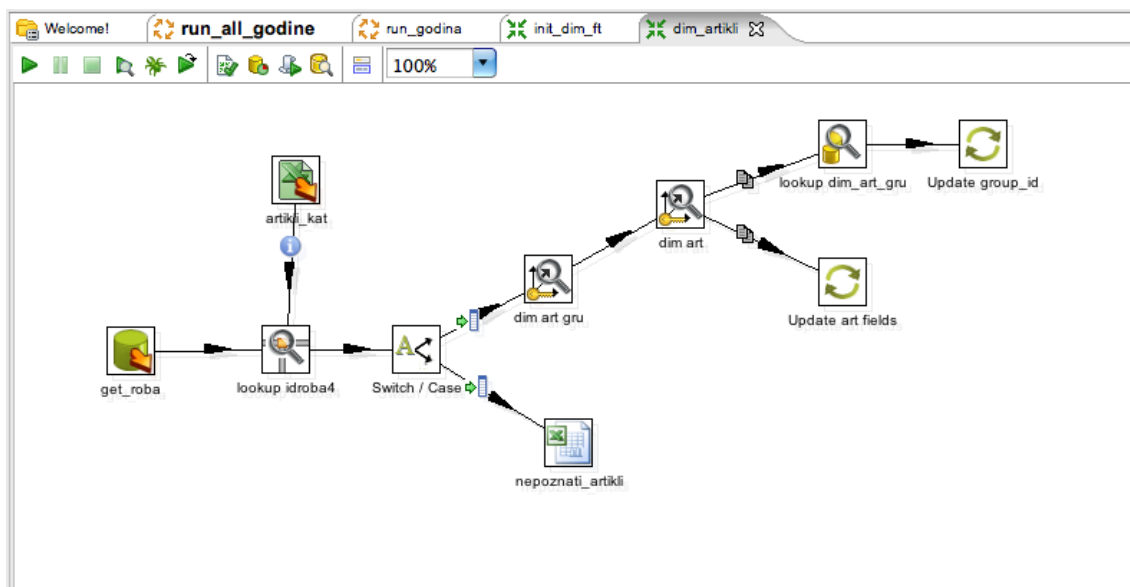
Slika 3.6: Inicijalizacija 'dimension' i 'facts' tabela



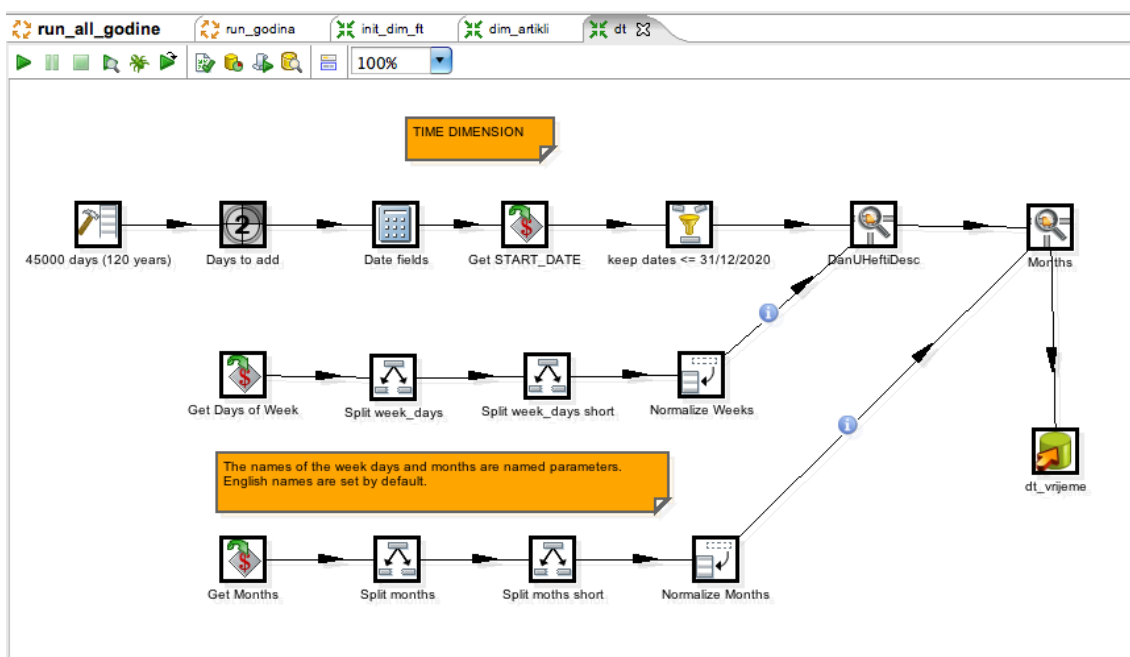
Slika 3.7: Generacija "ft_fakt" 'facts' tabele za određenu poslovnu godinu



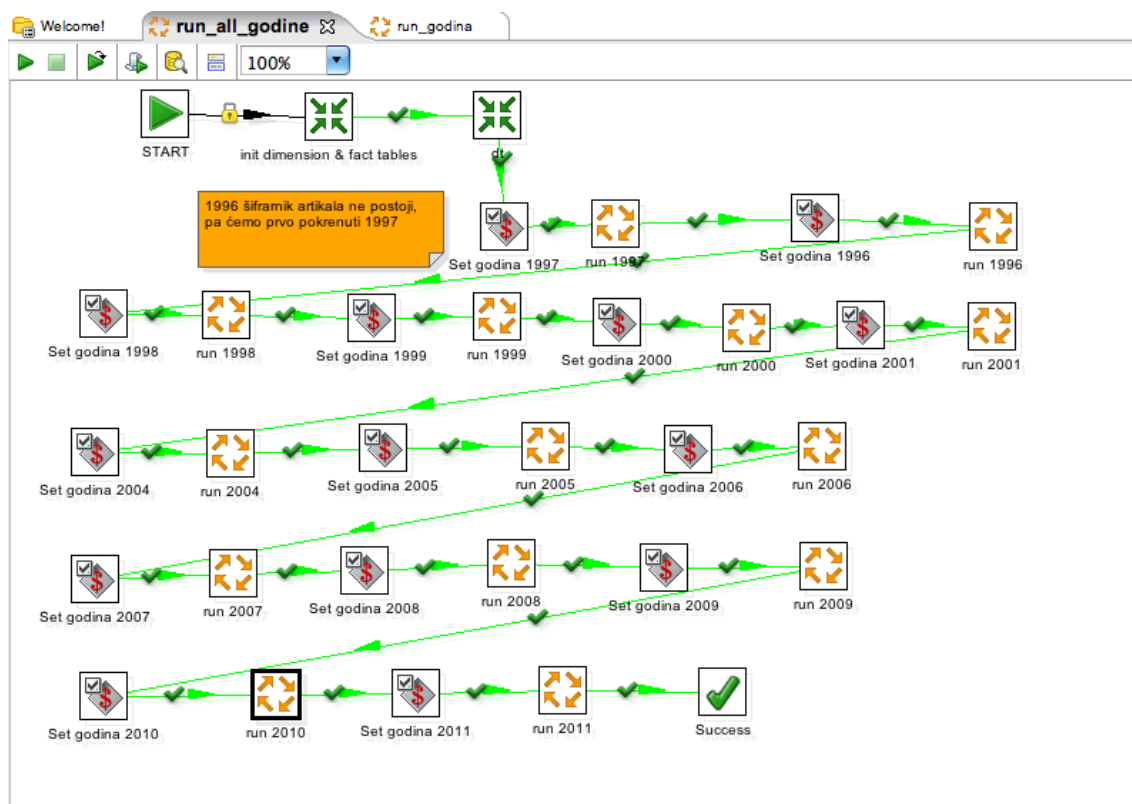
Slika 3.8: Kettle transformacija: Generacija "dim_par" i "dim_g" 'dimension' tabela za određenu poslovnu godinu



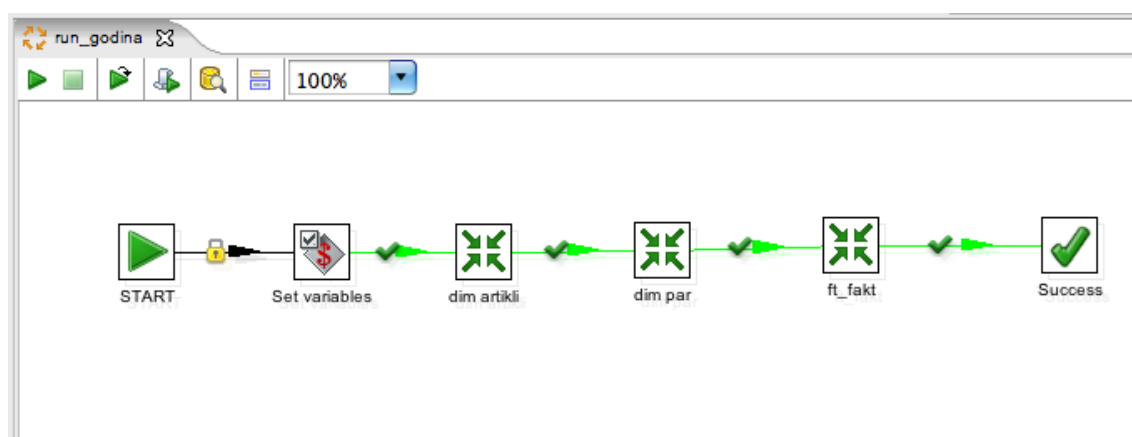
Slika 3.9: Kettle transformacija: Generacija "dim_art" i "dim_art_gru" 'dimension' tabela za određenu poslovnu godinu



Slika 3.10: Kettle transformacija: Generacija "dim_dt" 'dimension' tabele - vremenska dimenzija



Slika 3.11: Kettle job: inicijalizacija OLAP tabela, te generacija OLAP pdoataka za sve poslovne godine 1996-2011



Slika 3.12: Kettle job: Generacija OLAP podataka iz F18 ERP izvora za jednu poslovnu godinu

	A	B	C	D	E	F	G	H	I
8	2NAP001	2NAP	NAPOJNA JEDINICA L MODEL *		40				
9	3	3							
10	3	3 ..	ŠTAMPAČI, KOPIR APARATI						
11	4	4 ..	OSTALI PERIFERALI/DODATNA OPREMA						
12	5	5 ..	MREŽNA OPREMA						
13	6	6 ..	POTROŠNI MATERIJAL						
14	7	7 ..	SOFTVER DRUGIH PROIZVOĐAČA						
15	8	8 ..	OSTALO						
16	9	9 ..	SOFTVER SIGMA-COMA						
17	9OST.....	9OST	OSTALI PROGRAMI						
18	9OST0001	9OST	SCC EXPLORER - MANAGER PROJEKATA		1465				
19	S.....	S...	NOMENKLATURA ZA SERVIS						
20	SEMON	SEMO	SERVIS MONITOR						
21	SEOST	SEOS	OSTALI DIJELOVI						
22	SERAC	SERA	RAČUNARI SERVIS						
23	SESTA	SEST	ŠTAMPAČI SERVIS						
24									

Slika 3.13: Error reporting putem 'spreadsheet' dokumenata - artikli za koje nisu definisani kodovi u olap_cleansing tabelama

	A	B	C	D	E	F	G	H	I
1	ERROR	idtpdok	brdok	dat	kolicina	cijena	rabat	vrijednost	bk_idroba
2	ERR_ART	10	00004	2008-01-04	1,00	166,00		166	OS-DI
3	ERR_ART	10	00005	2008-01-04	1,00	50,00		50	PA-1
4	ERR_ART	10	00022	2008-02-08	2,00	50,00		100	PA-1
5	ERR_ART	10	00023	2008-02-08	3,00	40,00	20,00	120	PA-1
6	ERR_ART	10	00031	2008-02-08	2,00	167,40	7,00	334,8	KALK-DI
7	ERR_ART	10	00066	2008-02-25	1,00	318,50	30,00	318,5	VIRM
8	ERR_ART	10	00067	2008-02-27	12,00	85,00		1020	UPS600
9	ERR_ART	10	00135	2008-03-10	1,00	513,00	10,00	513	EPDV
10	ERR_ART	10	00144	2008-03-25	1,00	110,70	10,00	110,7	KADEV-MI
11	ERR_ART	10	00152	2008-03-25	1,00	454,00		454	DGSIGASW
12	ERR_ART	10	00152	2008-03-25	20,00	20,00		400	RJ45UTDUP
13	ERR_ART	10	00159	2008-04-08	1,00	50,00		50	PA-1
14	ERR_ART	10	00160	2008-04-11	50,00	1,00		50	KABUTP05

Slika 3.14: Dokumenti prodaje u kojima su neispravni podaci potrebni za popunjavanje dimension tabela (datum, klijent, roba)

4. Iza case study-ja ?

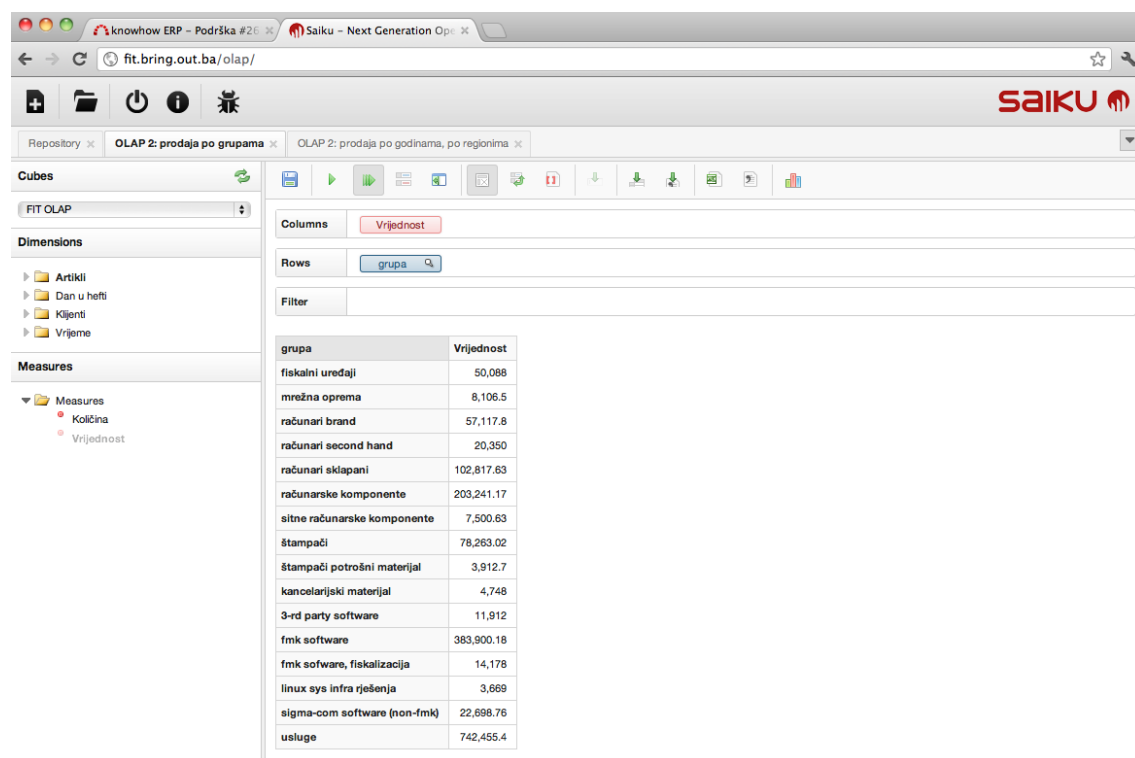
4.1. Ne znam

4.1.1. dimension table

4.1.2. facts table

4.1.3. ETL (Extract Transform Load)

4.2. Analiza podataka



The screenshot shows the Saiku web application interface. The browser tabs include 'knowhow ERP - Podrška #26' and 'Saiku - Next Generation Op...'. The address bar shows 'fit.bring.out.ba/olap/'. The interface has a sidebar on the left with sections: 'Cubes' (showing 'FIT OLAP'), 'Dimensions' (listing 'Artikli', 'Dan u hefti', 'Klijenti', 'Vrijeme'), and 'Measures' (listing 'Measures', 'Količina', 'Vrijednost'). The main area displays a table with columns 'grupa' and 'Vrijednost'. The table contains the following data:

grupa	Vrijednost
fiskalni uređaji	50,088
mrežna oprema	8,106.5
računari brand	57,117.8
računari second hand	20,350
računari sklapani	102,817.63
računarske komponente	203,241.17
sitne računarske komponente	7,500.63
štampači	78,263.02
štampači potrošni materijal	3,912.7
kancelarijski materijal	4,748
3-rd party software	11,912
fmk software	383,900.18
fmk software, fiskalizacija	14,178
linux sys infra rješenja	3,669
sigma-com software (non-fmk)	22,698.76
usluge	742,455.4

Slika 4.2: Pregled prodaje po grupama artikala

```

1 SELECT
2 NON EMPTY { Hierarchize ({ [ Measures ]. [ Vrijednost ] }) }
3 ON COLUMNS,
4   NON EMPTY { Hierarchize ({ [ Artikli . artikli ]. [ grupa ]. Members }) }
5 ON ROWS
6 FROM [ FIT OLAP ]
7 WHERE { Hierarchize ({ [ Vrijeme . vrijeme ]. [ All Vrijeme . vrijeme ] }) }

```

Listing 4.1: Pregled prodaje po grupama artikala

godina	region	mjesto	Vrijednost
1996	XX	n/a	7,970.73
1997	XX	n/a	49,270.19
1998	XX	n/a	158,515.59
1999	XX	n/a	171,123.46
2000	ZE	Zenica	1,550
	BI	Bihać	1,327
		Sanski Most	20
	GOR	Goražde	20
	KO	Konjic	614
	SA	Sarajevo	10,788.2
	TEŠ	Tešanj	85
	XX	n/a	159,547.88
	ZE	Zenica	6,511.2
	ŽEP	Žepče	75
2001	BI	Bihać	5,693.7

Slika 4.3: Pregled prodaje po regionima, po godinama

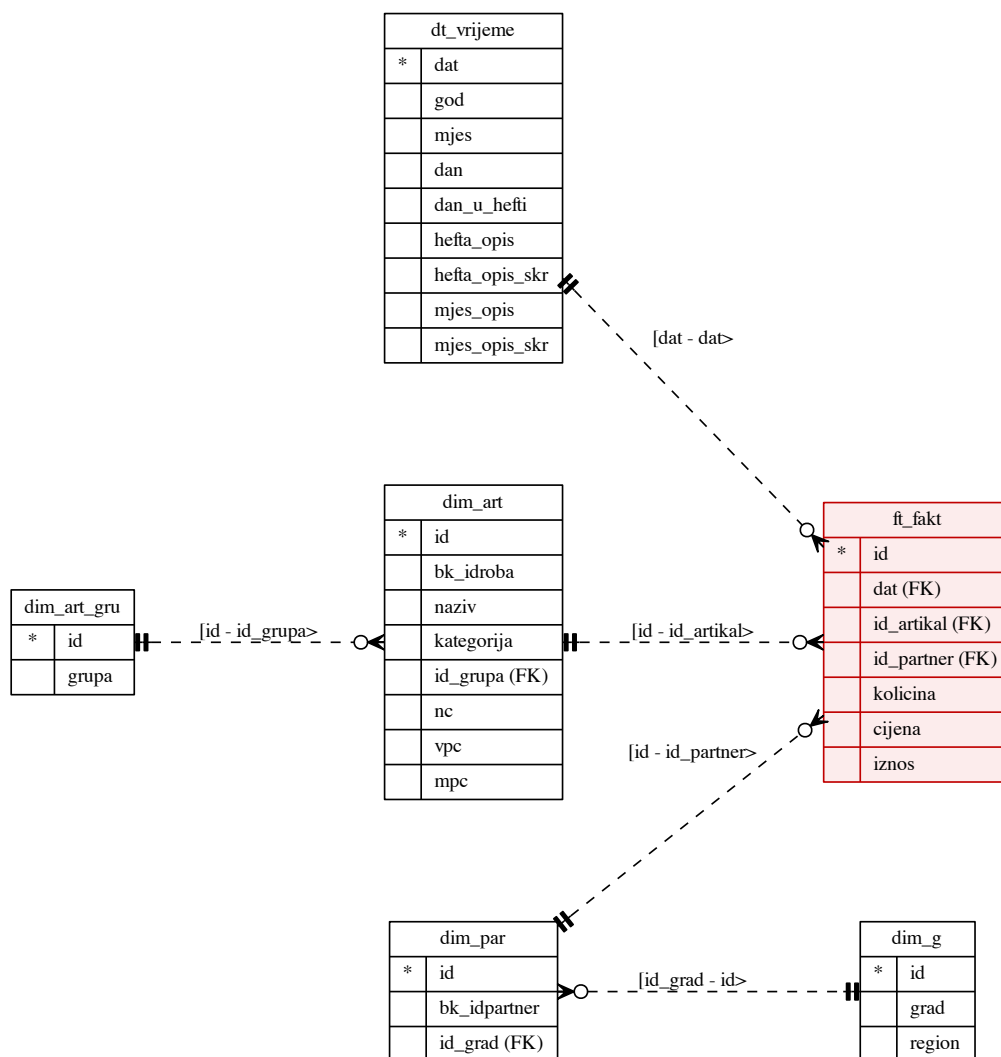
```

1 SELECT
2   NON EMPTY { Hierarchize ({ [ Measures ]. [ Vrijednost ] }) }
3 ON COLUMNS,
4   NON EMPTY
5     Hierarchize (
6       Union ( CrossJoin ( [ Vrijeme . vrijeme ]. [ godina ]. Members ,
7         [ Klijenti . klijenti ]. [ region ]. Members ) , CrossJoin ( [
8         Vrijeme . vrijeme ]. [ godina ]. Members ,
9         [ Klijenti . klijenti ]. [ mjesto ]. Members ) )

```

```
9      )  
10 ON ROWS  
11 FROM [FIT OLAP]
```

Listing 4.2: Pregled prodaje po regionima



Slika 4.1: OLAP schema

5. Zaključak

5.0.1. Ekspert

Poznavanje sadržaja i postojećih struktura podataka.

6. Literatura

The analytical labs. Saiku analytics software, Februar 2012. URL <http://analytical-labs.com>.

Sandro Bimonte i Pascal Wehrle. An olap solution using mondrian and jpivot, 2007. URL http://eric.univ-lyon2.fr/~sbimonte/doc/presentation_2007-02.pps.

R foundation. The r project for statistical computing, Februar 2012. URL <http://www.r-project.org>.

Leo Mršić. *Primjena metoda rudarenja podataka u trgovini tekstilnim i srodnim proizvodima*. Magistarski rad, Ekonomski fakultet u Zagrebu, November 2004.

Pentaho. Mondrian snowflake schema, Februar 2012. URL http://mondrian.pentaho.com/documentation/schema.php#Star_schemas.

Pentaho Community. Pentaho weka project, Februar 2012. URL <http://weka.pentaho.com/>.

Maria Carina Roldan. *Pentaho 3.2 Data Integration: Beginner's Guide*. Packt Publishing, 2010. URL <http://www.packtpub.com/pentaho-32-data-integration-beginners-guide/book>.

Machine Learning Group University of Waikato. The weka data mining software: An update, Februar 2012. URL <http://www.cs.waikato.ac.nz/ml/weka>.

Wikipedia. Business intelligence, Februar 2012a. URL http://en.wikipedia.org/wiki/Business_intelligence.

Wikipedia. Olap cube, Februar 2012b. URL http://en.wikipedia.org/wiki/OLAP_cube.

Wikipedia. Xml for analysis, Februar 2012c. URL http://en.wikipedia.org/wiki/XML_for_Analysis.

Dodatak A

Izvorni kod, dostupni resursi

1. OLAP mondrian, kettle transformacije i job-ovi, erviz modeli: https://github.com/hernad/hello_bi
2. Latex kod ovog dokumenta <https://github.com/hernad/MIS/tree/master/latex>
3. olap_cleansing 'spreadsheet' dokument https://github.com/hernad/hello_bi/raw/master/olap_cleansing.xls
4. Saiku demo server online: <http://fit.bring.out.ba/olap/#>

Dodatak B

Bilješke

1. Prva verzija ovog seminarskog rada, neuspješno https://github.com/hernad/MIS/raw/master/knowhowERP_OLAP_blog_style.pdf
2. FIT OLAP 2 cube: <http://redmine.bring.out.ba/issues/26711>