2. Vežbe

Prikupljanje zahteva, projektovanje dimenzionog modela, projektovanje šeme DW baze podataka

Izvođači laboratorijskih vežbi

- Marko Knežević (kancelarija TMD 9b)
- Nikola Obrenović

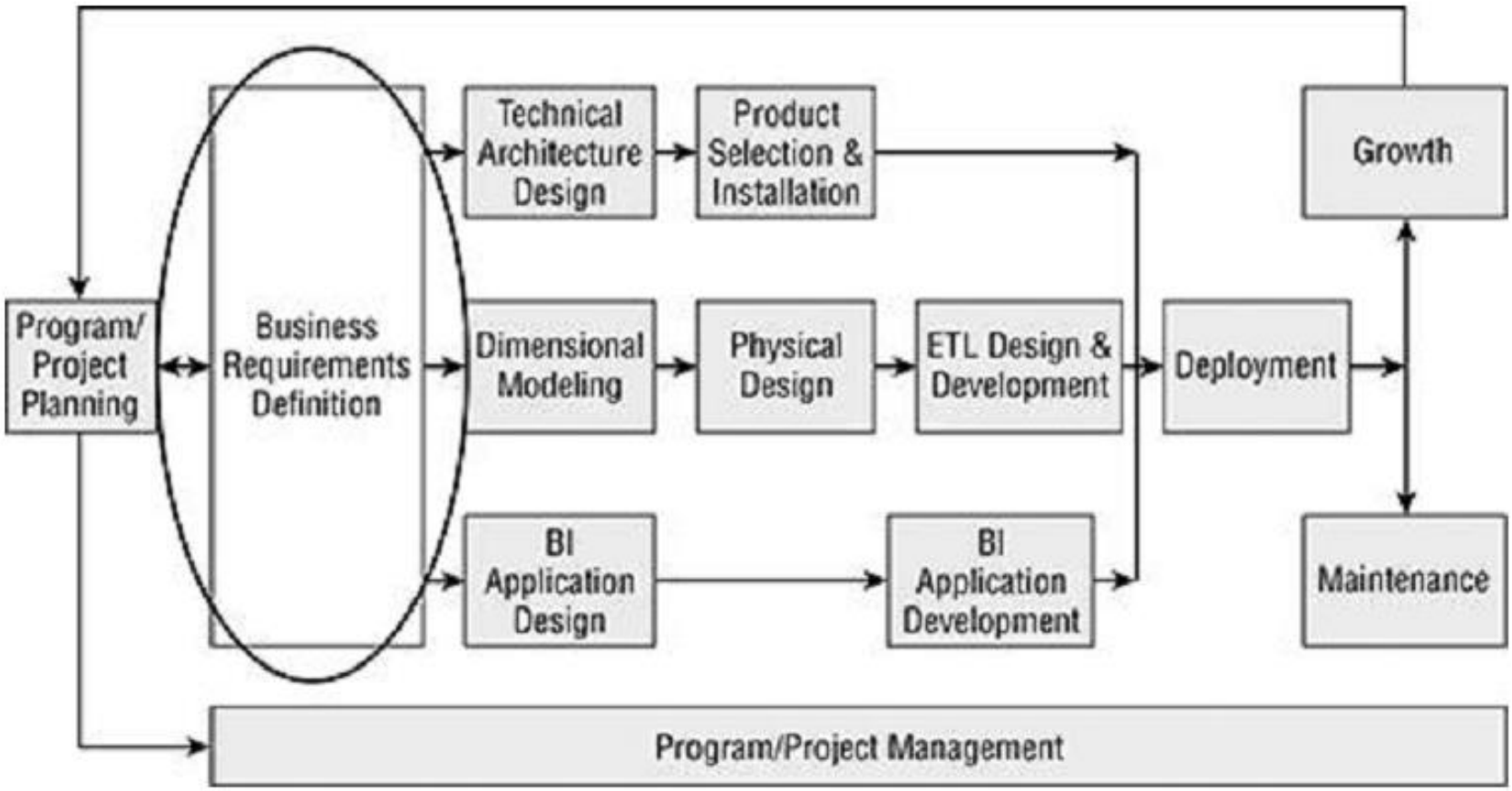
Termin konsultacija

- Marko Knežević: petak 15:00 TMD 9b
  marko.knezevic(AT)uns.ac.rs
- Nikola Obrenović
  nikob(AT)uns.ac.rs

Prikupljanje zahteva

- Identifikacija zahteva je preduslov za uspeh prilikom projektovanja DW sistema
- Sprovodi se kroz:
  - Intevjue sa klijentom, tj. budućim korisnicima
  - Analizom dokumentacije koja opisuje klijentove poslovne procese
  - Analizom postojećih izvora podataka (eng. data profiling)
- Suština je što bolje upoznati se sa klijentovim poslovnim procesima, potrebama i postojećom informacionom infrastrukturom

Kimball's Lifecycle



Prikupljanje zahteva

- Loš pristup: “Šta želite?”
- Dobar pristup: “Opišite mi Vaš posao.”
- Na osnovu sprovedenih intervjua identifikovati poslovne zahteve

Intevju: Brian Welker, VP of Sales

- Prodaja:
  - Praćenje narudžbenica po mesecima i teritoriji
  - Praćenje narudžbenica po predstavniku na mesečnom nivou
  - Broj narudžbenica po potrošaču, po teritoriji
- Ažuriranje cenovnika između AWC i predstavnika – **nije DW problem**
- Specijalne ponude
  - Identifikovati potrošače koji kupuju najviše posebne ponude
  - Identifikovati i potrošače koji ne čekaju posebne ponude jer donose veći profit

Intevju: Brian Welker, VP of Sales

- Odnos prema kupcima
  - Praćenje pritužbi prema tipu žalbe, proizvodu, predstavniku
  - Praćenje vraćenih proizvoda prema predstavniku i proizvodu
- Prevođenje opisa proizvoda na druge jezike – **nije DW problem**

Poslovni procesi

- Mapirati zahteve na poslovne procese
- Poslovni procesi će identifikovati izvore podataka potrebne za ispunjenje zahteva
- Za svaki zahtev identifikovati potencijalne probleme u izvorima podataka
- Identifikovaće se poslovni procesi koji pokrivaju najviše poslovnih zahteva

Poslovni procesi

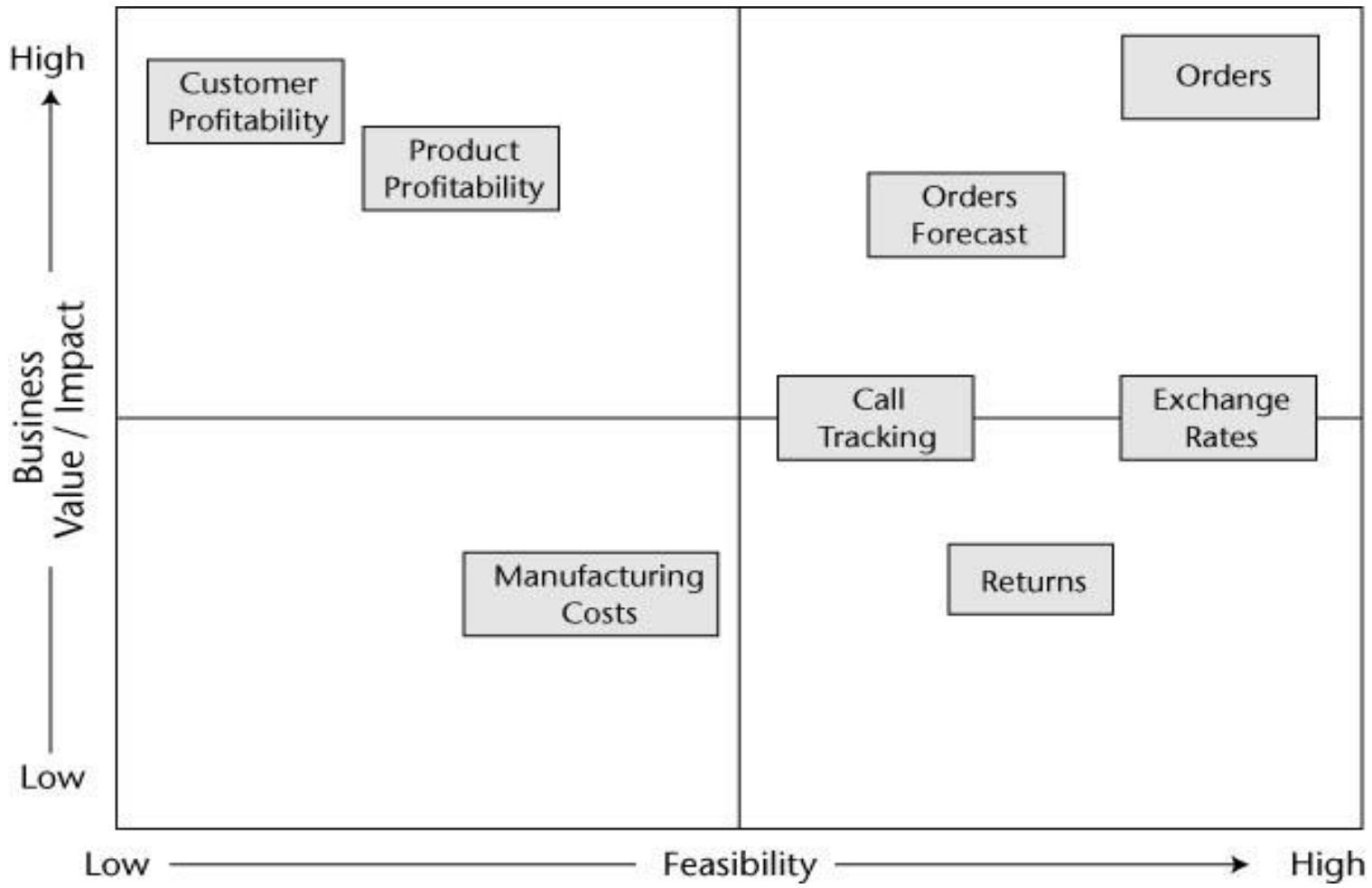
Business Requirements Category	Inferred or Requested Analyses	Supporting Business Process	Comments
Sales Planning	- Reseller historical orders analyses	- orders	By customer, by territory, by sales region (from state)
	- Sales forecast	- orders	Forecast is a business process that uses orders data as an input
Sales Performance	- Orders by current territory	- orders	
	- Orders by original territory	- orders	
	- Sales rep performance report	- orders - forecast	Orders and forecast by sales rep
Sales Reporting	- Resellers ranked by orders in a given territory	- orders	
	- Churned customer list	- orders	Customers who have not ordered in X months
Price Lists	- Current price list	- orders	This is a connectivity issue, not a data warehouse issue
Special Offers	- Relevant customers by territory based on orders history	- orders	
	- Inventory status (out of stock)	- inventory	
Customer (Reseller) Satisfaction	Customer Satisfaction Dashboard	Multiple	This is a compound requirement based on several underlying business processes
	- Calls by complaint type, product and customer attributes	- call tracking	
	- Order metrics of satisfaction	- orders	e.g. due date versus ship date
	- Returns by reseller by return reason	- returns	
International Support	- Local language translations of product descriptions	- n/a (product dimension)	This is a transaction system problem. We need to make sure we can handle multiple languages in the DW/BI system, but the source system has to capture them when new products are created.

Poslovni procesi

- Jedinice razvoja DW sistema
- Bus matrix mapira poslovne procese na dimenzije

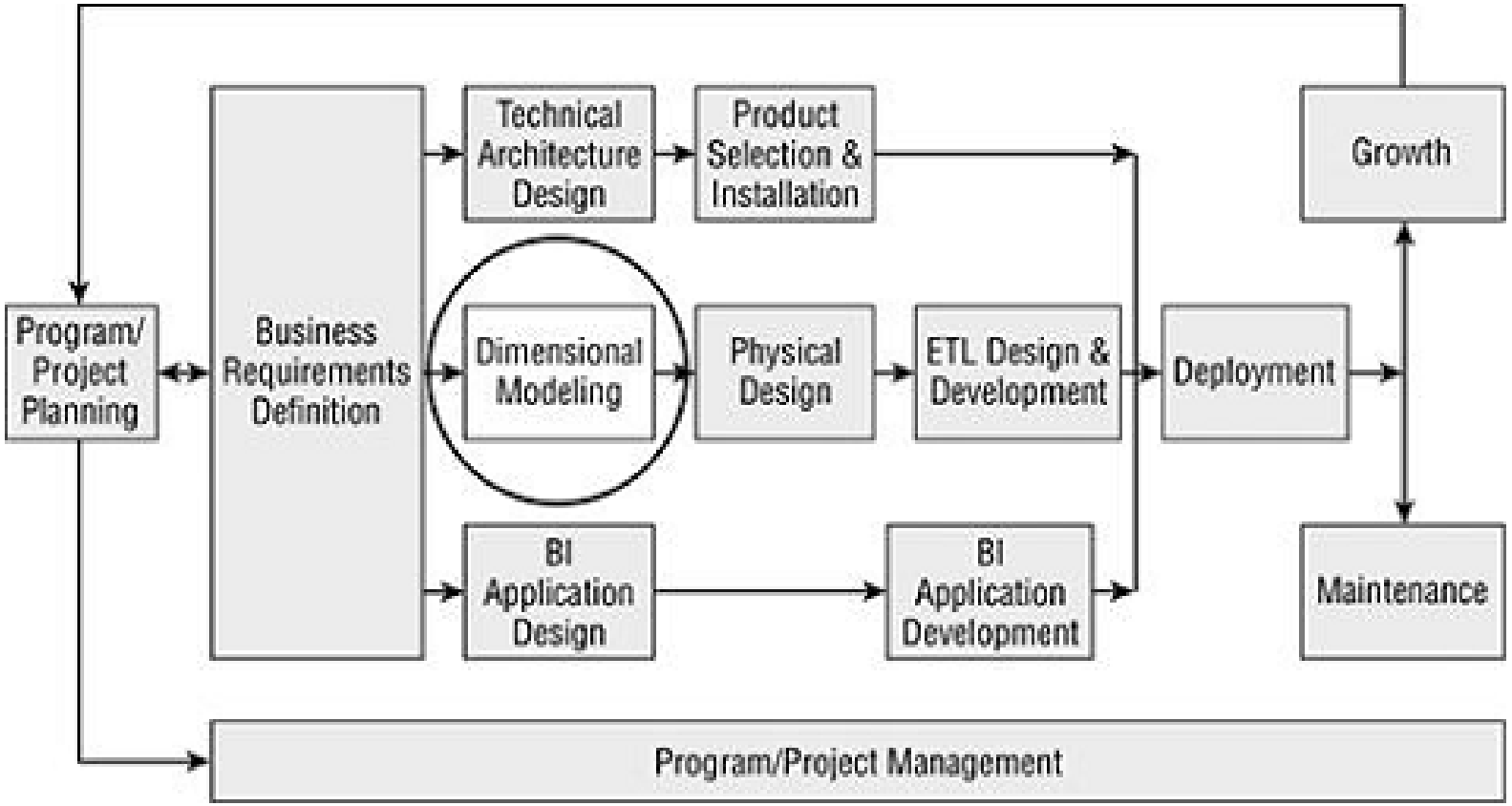
	Dimensions										
Business Process	Date	Product	Employee	Customer (Reseller)	Customer (Internet)	Sales Territory	Currency	Channel	Promotion	Call Reason	Facility
Sales Forecasting	X	X	X	X	X	X	X				
Orders	X	X	X	X	X	X	X	X	X		
Call tracking	X	X	X	X	X	X				X	
Returns	X	X		X	X	X	X		X		X

Prioritetizacija poslovnih procesa



Projektovanje dimenzionog modela

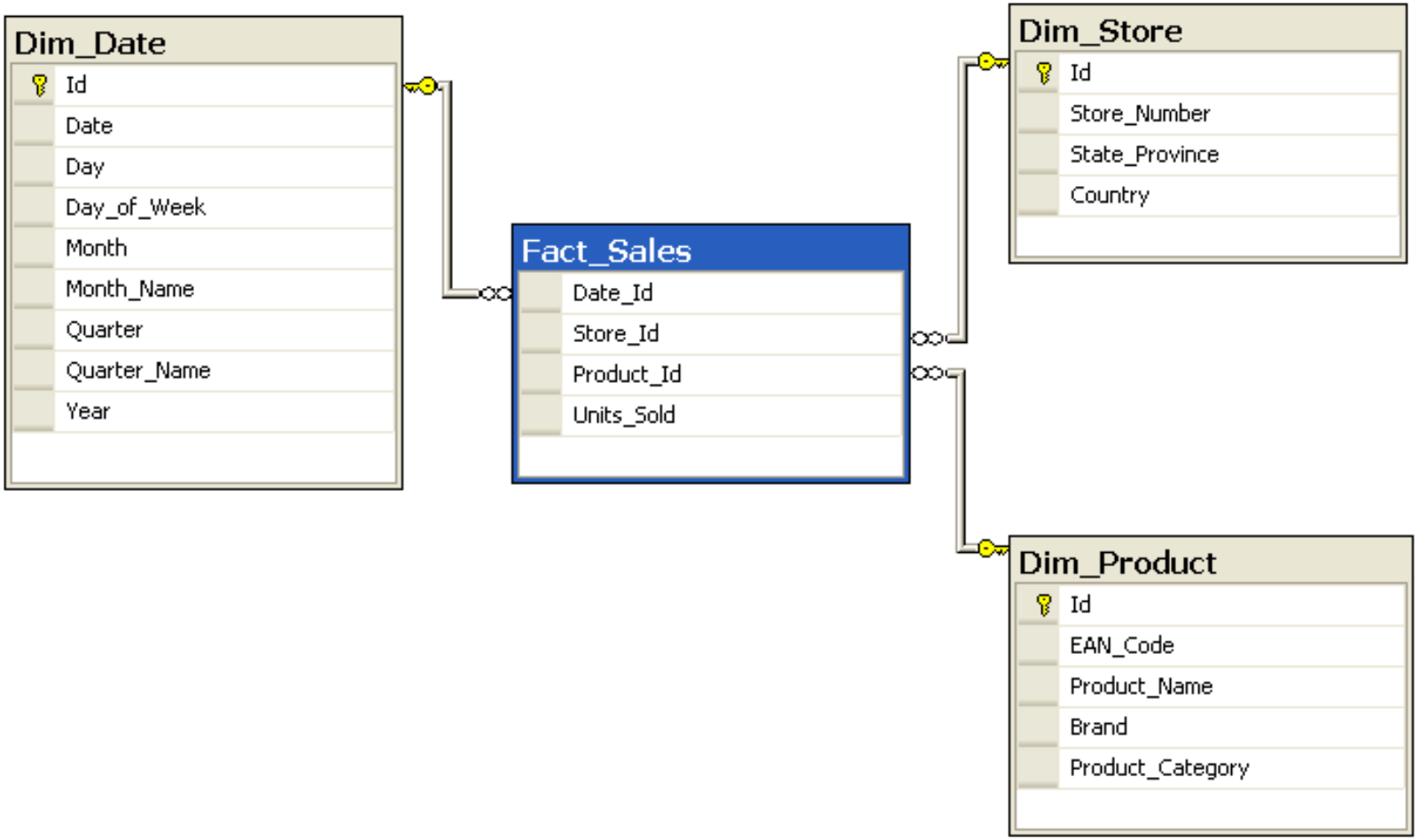
- Šema DW baze podataka



Dimenzioni model

- Razumljiviji za poslovne korisnike od OLTP šeme baze podataka
- Sadržana informacija više odgovara poslovnim procesima
- Bolje performanse upita zbog denormalizacije i unapred agregiranih podataka
- Činjenice i dimenzije
- Zvezdasta šema

Primer zvezdaste šeme



Činjenice

- Činjenica sadrži mere koje opisuju neki poslovni proces. Npr. prosečan broj kupovina nekog artikla na dnevnom nivou, prosečan broj položenih ispita po smeru, godini studija i školskoj godini.
- Daju odgovore na poslovne zahteve
- Normalizovana
- Red u činjenici je identifikovan jedinstvenom kombinacijom dimenzija
- Granularnost činjenice treba da bude bar za jedan stepen manja od korisničkih zahteva
- Činjenice sadrže mere za koje je poželjno da budu aditivne ili bar semi-aditivne

Činjenice

- Periodični preseci
  - Mere reprezentuju period između trenutnog i poslednjeg preseka
- Akumulirani preseci
  - Mere predstavljaju akumuliranu vrednost od prvog preseka do trenutnog
- Prazne činjenice (*Factless facts*)
  - Pojave tipa 0 ili 1 (npr. prisustvo vežbama)
  - Činjenicu modeluje postojanje reda u činjeničnoj tabeli
  - Samo strani ključevi koji referenciraju dimenzije
- Izvedene činjenice
  - Izračunavaju se iz dve ili više drugi činjenice ili mera
  - Činjenična tabela ili pogled

Dimenzije

- Predstavljaju entitete koji opisuju šta se meri činjenicom. Npr. proizvod, radnik, student, boja, vreme...
- Dimenzije – nezavisne promenjive, činjenica – zavisna promenjiva
- Jedna dimenzija može opisivati više činjenica
- Dimenzije mogu sačinjavati hijerarhiju.
  Npr. Grad->Pokrajina->Država
- Denormalizovane
- Bus matrix omogućava da se ne dupliraju dimenzije
- Conformed dimenzija opisuje više od jedne činjenice

Surogatni ključevi

- Dimenzije se identifikuju surogatnim ključevima
- Prednosti:
  - Ne zavise od promena prirodnih ključeva u izvoru
  - Omogućavaju da se isti podaci importuju iz različitih sistema gde imaju različite ključeve
  - Omogućavaju da se doda red u dimenziji sa specijalinim značenjem poput ,,nepoznat proizvod'', ,,ostalo'' i sl.
  - Omogućavaju da se prati promena vrednosti dimenzije po vremenu
  - Upiti nad celobrojnim surogatnim ključevima su uglavnom brži od upita nad prirodnim ključevima
- Mane:
  - Referenciranje dimenzija iz činjenica prilikom ETL procesa – zahteva look-up surogatnog ključa
- Slavica Aleksić, Milan Čeliković, Sebastian Link, Ivan Luković, Pavle Mogin: Faceoff: Surrogate vs. Natural Keys. ADBIS 2010

Sporo promenjive dimenzije

- Tehnike izmene atributa dimenzije tokom vremena:
  - Tip 1: vrednost atributa se prepisuje novm vrednošću, tj. ne čuva se stara vrednost
  - Tip 2: dimenziji se dodaju dva timestamp atributa koji označavaju početak i kraj važenja sloga. Ukoliko se promeni vrednost atributa dimezije unosi se novi slog u dimenziju
    - Čuva više informacija
    - Komplikovaniji ETL proces

Vremenska dimenzija

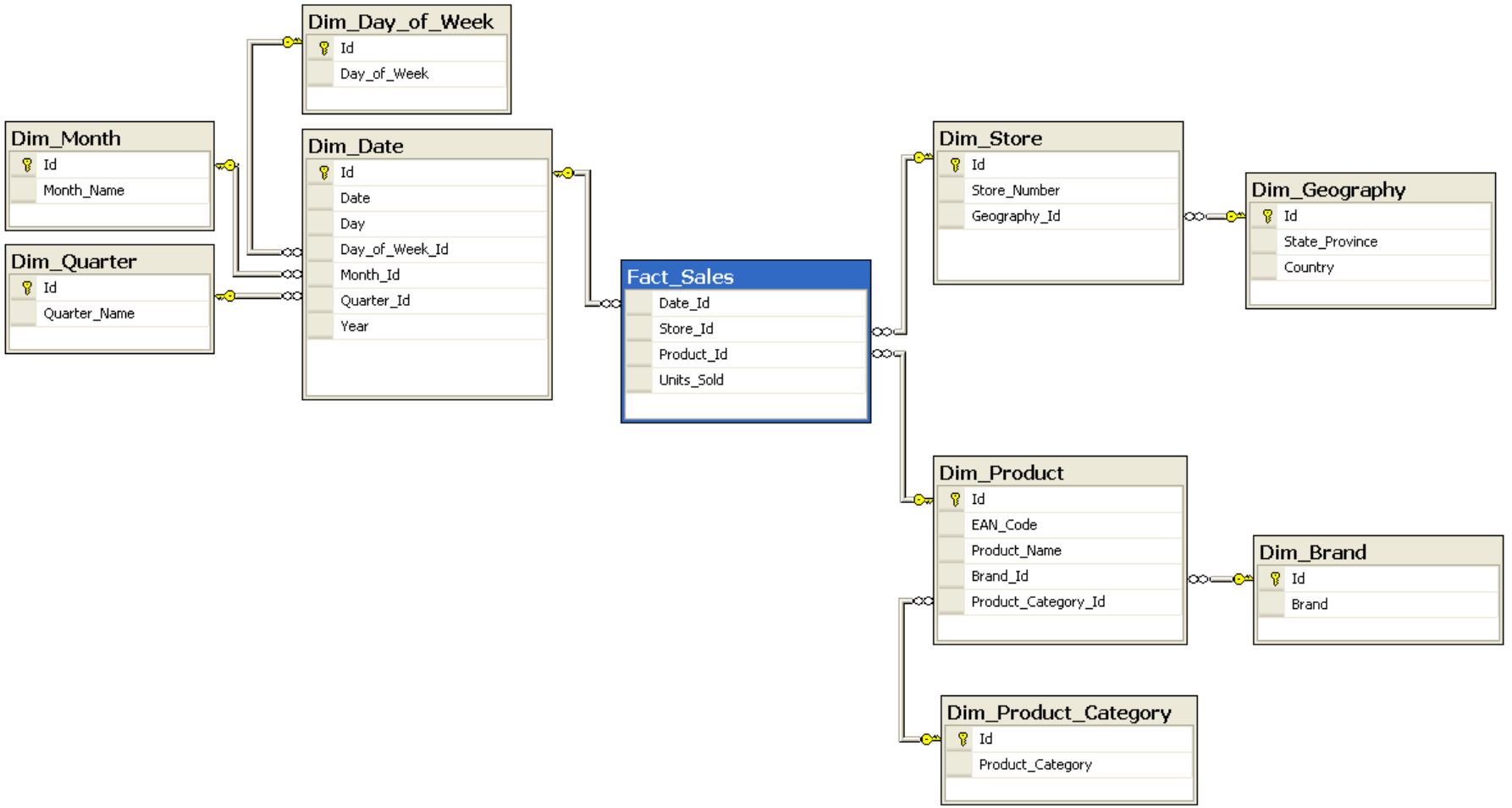
- Obavezna dimenzija u DW sistemima
- Granularnost zavisi od namene, najčešće na nivou dana
- Dodatna dimenzija: DayPeriod
- Surogatni ključ uglavnom ima oblik *<godina><mesec><dan>*, npr. 20150924
  - Bolje performanse upita ako se tabela particioniše po ključu

Degenerativna dimenzija

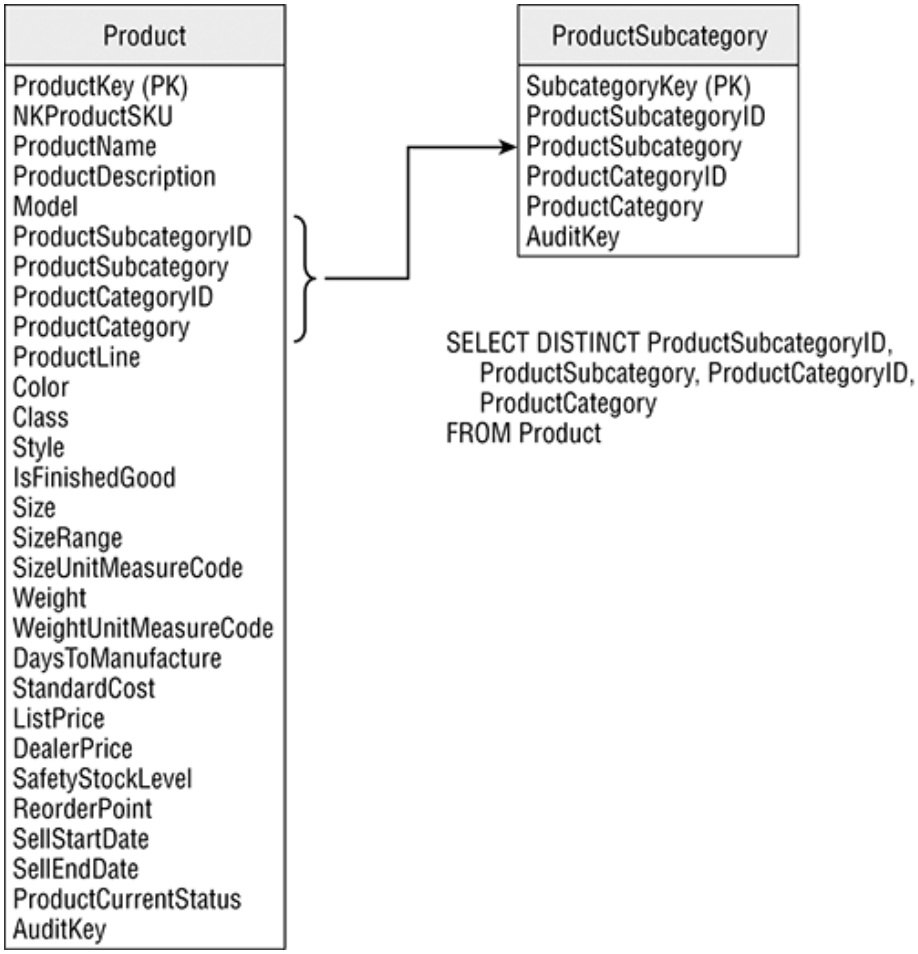
- Dimenzija koja ima samo ključ
- Svi potencijalni atributi su već raspoređeni po drugim dimenzijama
- Npr. ključ narudžbenice

Šema pahulje

- Ponovna normalizacija dimenzija
- Ubrzava punjenje DW baze ali usporava upite



Agregirana dimenzija



- Izdvajanje dela atributa dimenzije u novu dimenziju
- Postižu se brži upiti jer nema agregacije kroz SQL

Junk dimenzija

- Spajanje nekoliko malih dimenzija u jednu zajedničku iako originalne dimenzije nemaju prirodnih sličnosti
- Razlog je smanjivanje broja dimenzija
- Dekartov proizvod osnovnih dimenzija ili kombinacije koje postoje u stvarnosti

Admit_Type_Source

Admit_Type_ID	Admit_Type_Descr
1	Walk-in
2	Appointment
3	ER
4	Transfer

Care_Level_Source

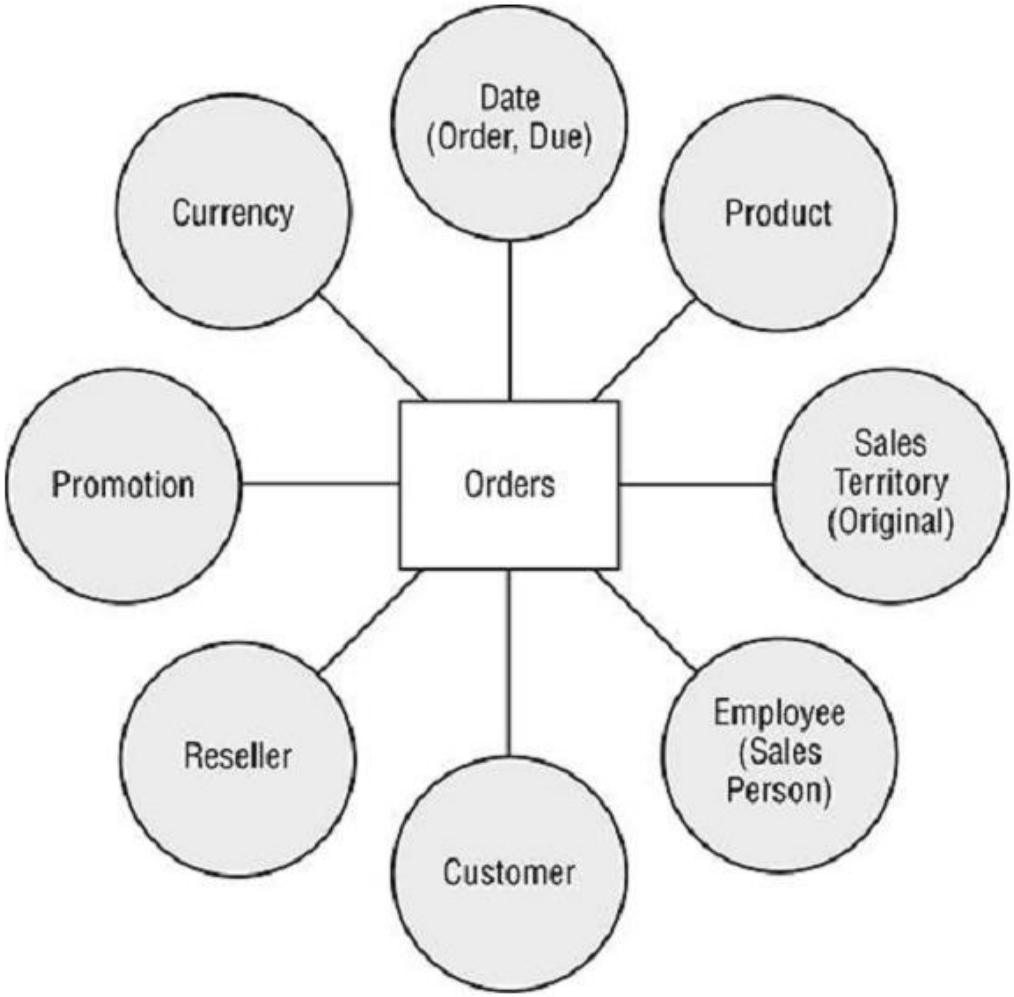
Care_Level_ID	Care_Level_Descr
1	ICU
2	Pediatric ICU
3	Medical Floor

Admission_Info_Key	Admit_Type_ID	Admit_Type_Descr	Care_Level_ID	Care_Level_Descr
1	1	Walk-In	1	ICU
2	2	Appointment	1	ICU
3	2	Appointment	2	Pediatric ICU
4	4	Transfer	3	Medical Floor

Proces modelovanja

- Izvor za inicijalnu verziju modela je bus matrix
- Iterativni proces
- Činjenice i dimenzije se formiraju kroz analizu poslovnih zahteva i analizu OLTP izvora podataka
- Identifikovati atribute činjenica i dimenzija i kako se prate atributi dimenzija tokom vremena
- Sa strane klijenta obavezne su osobe koje poznaju postojeće podatke i poslovne procese
- Identifikacija i ispravljanje nedostajućih i nevalidnih podataka (*data remediation*)
- Na kraju svake iteracije potrebno je proveriti da li su zahtevi zadovoljeni

Primer AWC – Inicijalni bubble chart



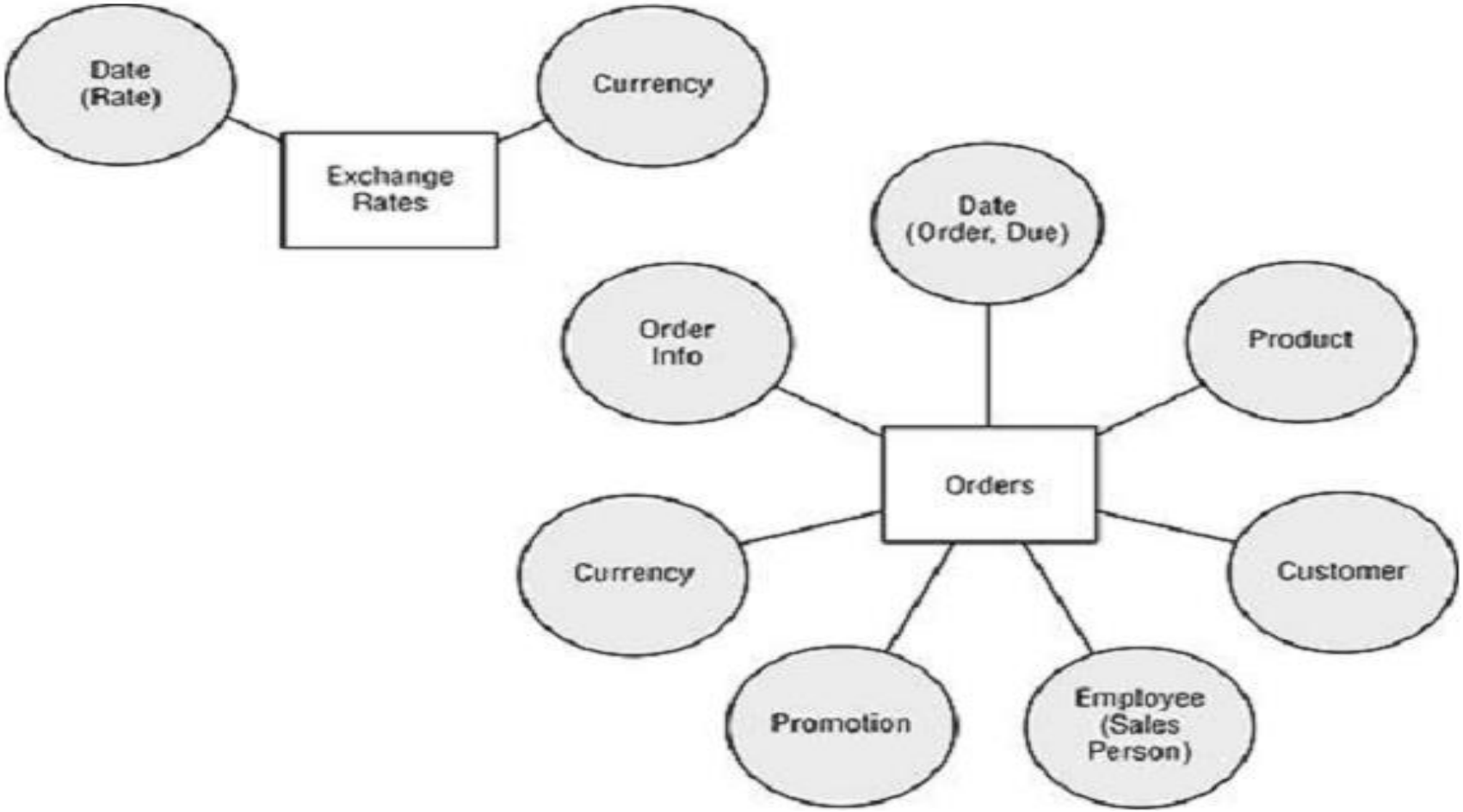
Primer AWC - Modelovanje dimenzija

- Postoji potreba da se promene kod zaposlenih prate kroz vreme => većina atributa Employee dimenzije su tip 2 – dodati atributi RowStartDate, RowEndDate i RowIsCurrent
- Veći broj izveštaja će zahtevati sumirane podatke o predstavnicima i internet kupcima, nego odvojene => Predstavnici i internet kupci su spojeni u jedinstvenu Customer dimenziju i razlikuju se preko CustomerType obeležja

Primer AWC - Modelovanje dimenzija

- U činjenici Orders je potrebno izraziti cene i u USD i u lokalnoj valuti => potrebno je importovati u DW bazu kursne liste
- Time su importovani svi podaci potrebni za još jedan poslovni proces “Exchange Rates” pa je dodata još jedna činjenica FactExchangeRates
- OrderInfo je junk dimenzija koja modeluje razlog prodaje i kanal prodaje
- Paralelno se identifikuju atributi dimenzija

Primer AWC – sledeća iteracija dimenzionog modela



Primer AWC – modelovanje činjenica

- Mere se definišu na osnovu zahteva
- Npr. prosečna cena, frekvencija događaja, maksimalno merenje, udeo tipa pojave u skupu pojava...

Fizičko projektovanje relacione šeme DW

- Šeme relacija koje će implementirati činjenice i dimenzije su u velikoj meri definisane kroz dimenziono modelovanje: skupovi atributa i njihovi tipovi, primarni ključevi, strani ključevi, ograničenja...
- Izabrati veličinu surogatnog ključa prema očekivanoj veličini dimenzije
  - Brži upiti i manja veličina tabela (posebno činjenice)
- Odrediti mehanizam za generisanje surogatnih ključeva:
  - sekvence u Oracle-u
  - IDENTITY kolona u MS SQL Serveru
  - okidač ili stored procedura u slučaju kompleksnih surogatnih ključeva

Fizičko projektovanje relacione šeme DW

- Odrediti dužinu string kolona
- Odrediti preciznost float kolona
- Dodavanje Audit dimenzije za praćenje izmena u podacima
- Kreirati indekse po primarnim ključevima u dimenzijama (u SQL serveru to se uradi automatski kroz definisanje primarnog ključa)
- Kreirati indekse po primarnim ključevima dimenzija u činjenici, za dimenzije sa mnogo redova
- Izbaciti primarni ključ činjenice i strane ključeve ka dimenzijama
- Ako veličina diska dozvoljava, kreirati indekse po prirodnim ključevima kako bi se ubrzao look-up tokom ETL procesa
- Opšti savet: kreirati indekse po svim kolonama po kojima se očekuje pretraga tabela

Fizičko projektovanje relacione šeme DW

- Kreirati poglede nad tabelama činjenica i dimenzija
  - pristup podacima za SSAS i korisnike
  - sakrivanje pomoćnih kolona, npr. RowStartDate ili AuditKey
  - Kreiranje jedne tabele dimenzije u šemi pahulje
- Particionisati velike tabele

Staging tabele

- Pomoćne tabele koje služe da privremeno čuvaju podatke tokom ETL procesa
- Olakšavaju transformaciju podataka. Npr, mapiranje prirodnog na surogatni ključ