OSNOVE POSLOVNE INTELIGENCE

Iztok Savnik

Podatkovna skladišča

- Podatkovno skladišče je velik repozitorij zgodovinskih podatkov, ki je narejen za podporo pri odločanju.
  - Uporaba podatkovnega skladišča je različna od dnevne organizacije dela
  - Tabele v dnevni organizaciji dela se ohranjajo majhne tako, da se občasno brišejo stari podatki

Podatkovna skladišča

- Podatkovna skladišča delujejo obratno
  - periodično sprejemajo zgodovinske podatke in rastejo s časom
  - Velikost PS sčasoma postane na stotine GB
- Cilj PS hitra obdelava velikih količin podatkov
  - Kontrast med dnevno organizacijo dela in podatkovnimi skladišči kreira način načrtovanja PS

Podatkovna skladišča

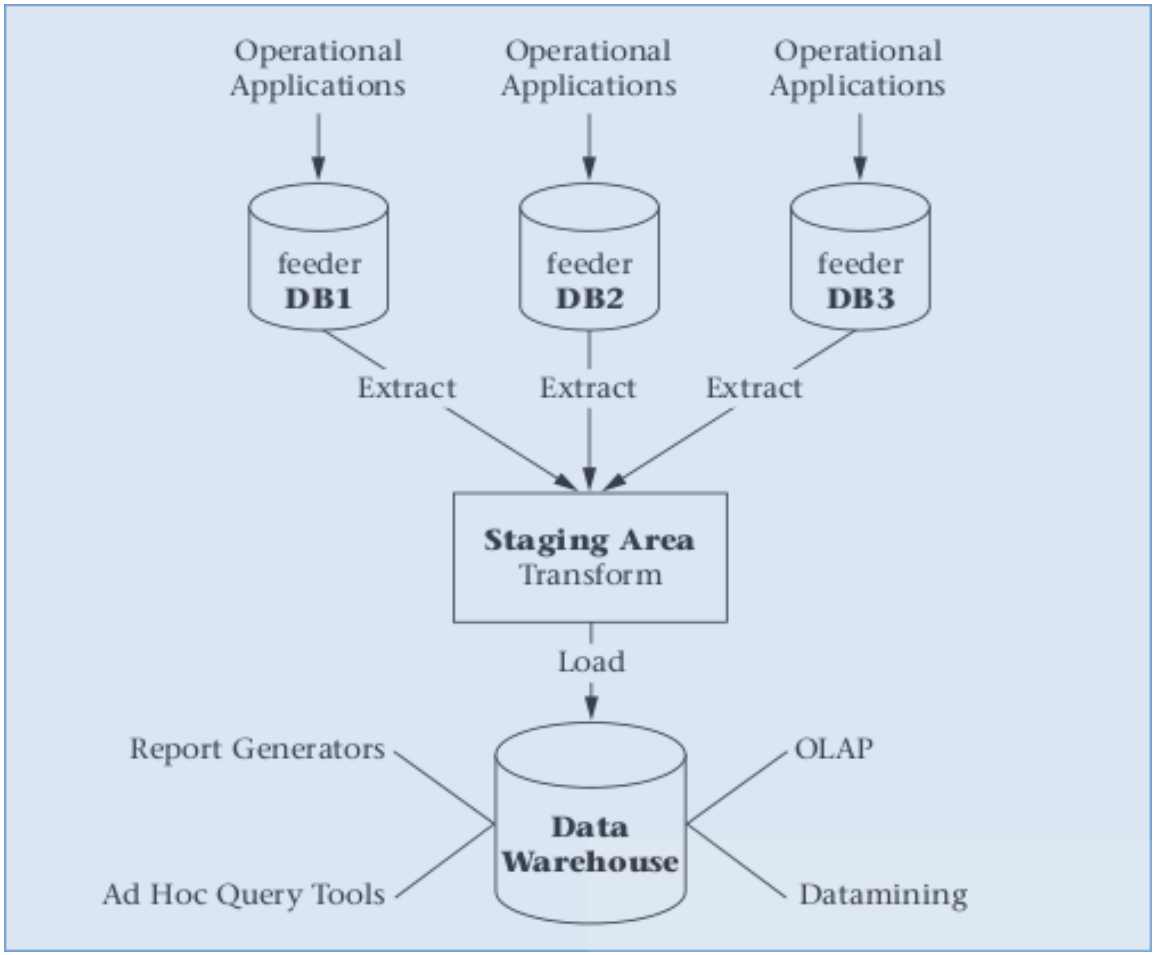
- PS vsebuje množico orodij za podporo odločanju
  - Področje je izšlo iz tehnologije sistemov za podporo odločanju (DSS) in upravnih informacijskih sistemov (executive IS)
  - DSS se uporablja za študij podatkovnih virov in in izdelavo poročil
  - Poročila niso časovo kritična

Podatkovna skladišča

Primerjava med OLTP in podatkovnimi skladišči

OLTP	Podatkovna skladišča
Transakcijsko orientirano	Orientirano glede na poslovni proces
Tisoče uporabnikov	Nekaj uporabnikov
Majhna baza (MB-> GB)	Velika baza (100GB-> večTB)
Trenutni podatki	Historični podatki
Normalizirani podatki	Denormalizirani podatki
Konstantni popravki	Popravki v svežnjih
Od enostavnih do kompleksnih vprašanj	Običajno zelo kompleksna vprašanja

Osnovna arhitektura PS



Principi načrtovanja PS

1. PS so organizirana okoli tematskih področij
   - Podoben koncept funkcijskem področju
     - Prodaja, vodenje projektov, zaposleni, ...
   - Vsako tematsko področje ima svojo konceptualno shemo
     - Predstavljeno z eno ali večimi razredi | entitetami
   - Tematska področja so tipično neodvisna od posameznih transakcij (insert, update, delete)
   - Meta-podatkovni repozitoriji opisujejo
     - podatkovne baze, PS objekte, transformacije podatkov,...

Principi načrtovanja PS

2. PS imajo nekatere zmožnosti integracije podatkov
   - Definira se skupna predstavitev podatkov, ki predstavlja vse individualne predstavitve
3. Uporabljeni podatki so persistentni in se nalagajo v svežnjih
   - Naj ekstrakcija deluje na nivoju n-teric ali v svežnjih?
   - Potrebna so orodja za čiščenje podatkov
     - Analiza konsistentnosti, homonimov, sinonimov, ...
     - Migracija, kontrola sprememb, ...
     - Orodja podobna tistim v RDBMS

Principi načrtovanja PS

4. Podatki obstajajo na večih nivojih podrobnosti in so vezani na čas
   - Podrobnosti so vezane na več dimenzij, ne samo časovno ampak tudi geografsko, po tipih produkta, po tipih firme, itd.
   - Velikost PS je kritična še posebaj za nekatera vprašanja in popravke podatkov

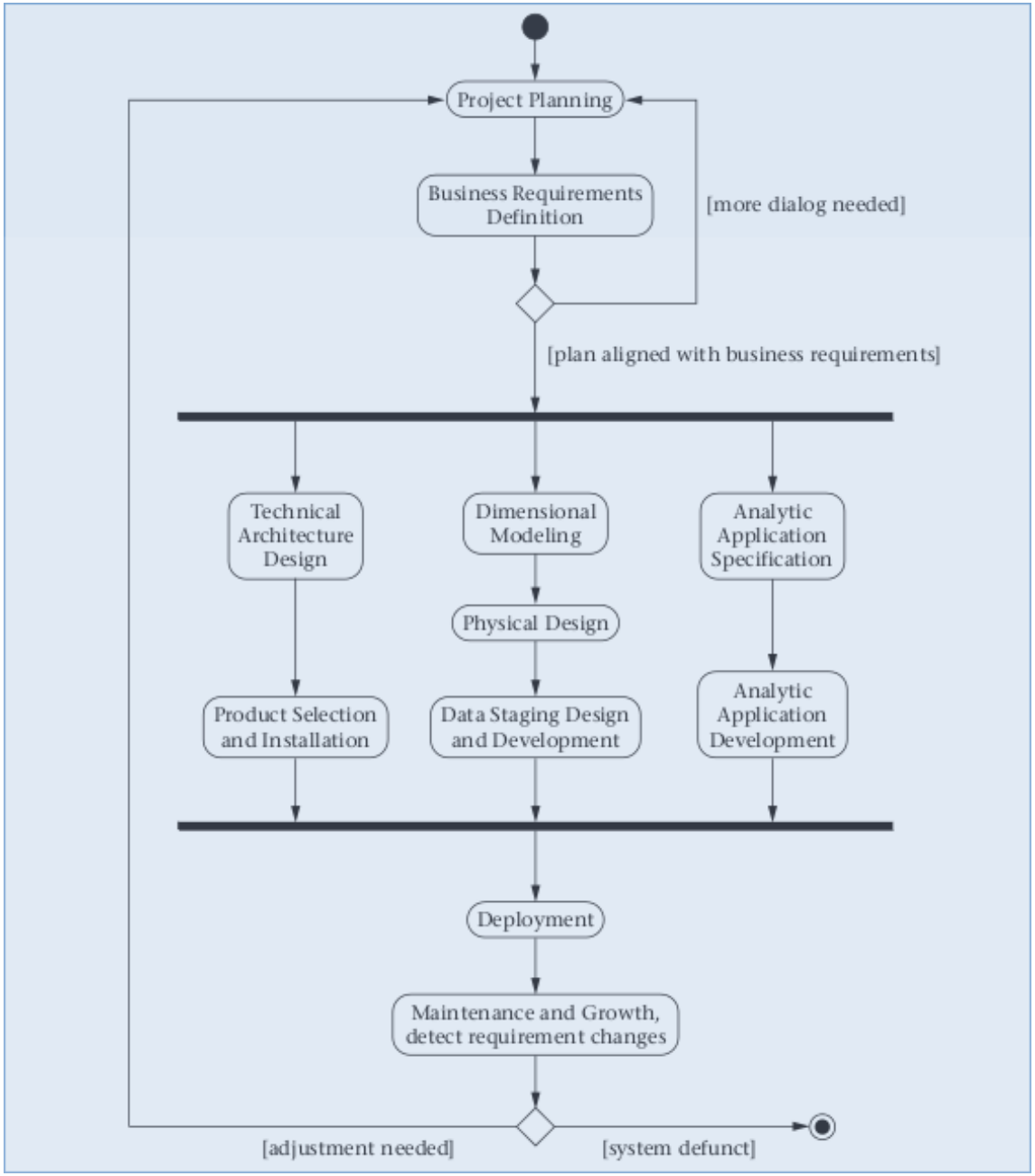
Principi načrtovanja PS

5. PS morajo biti zadosti flksibilna za spremembe
   - Dodajanje novih dimenzij, spremembe dimenzij, spremembe nivoja podrobnosti, ...
6. PS ima zmožnosti prepisovanja zgodovine
   - »Kaj-če« analiza
   - Po končani analizi vrnemo podatke v prejšnje stanje

7., 8., ...

   - Izbrati je potrebno vmesnik: SQL ali več-dimenzionalni vmesnik.
   - Podatki so bodisi centralizirani ali fizično por

Življenski cikel PS



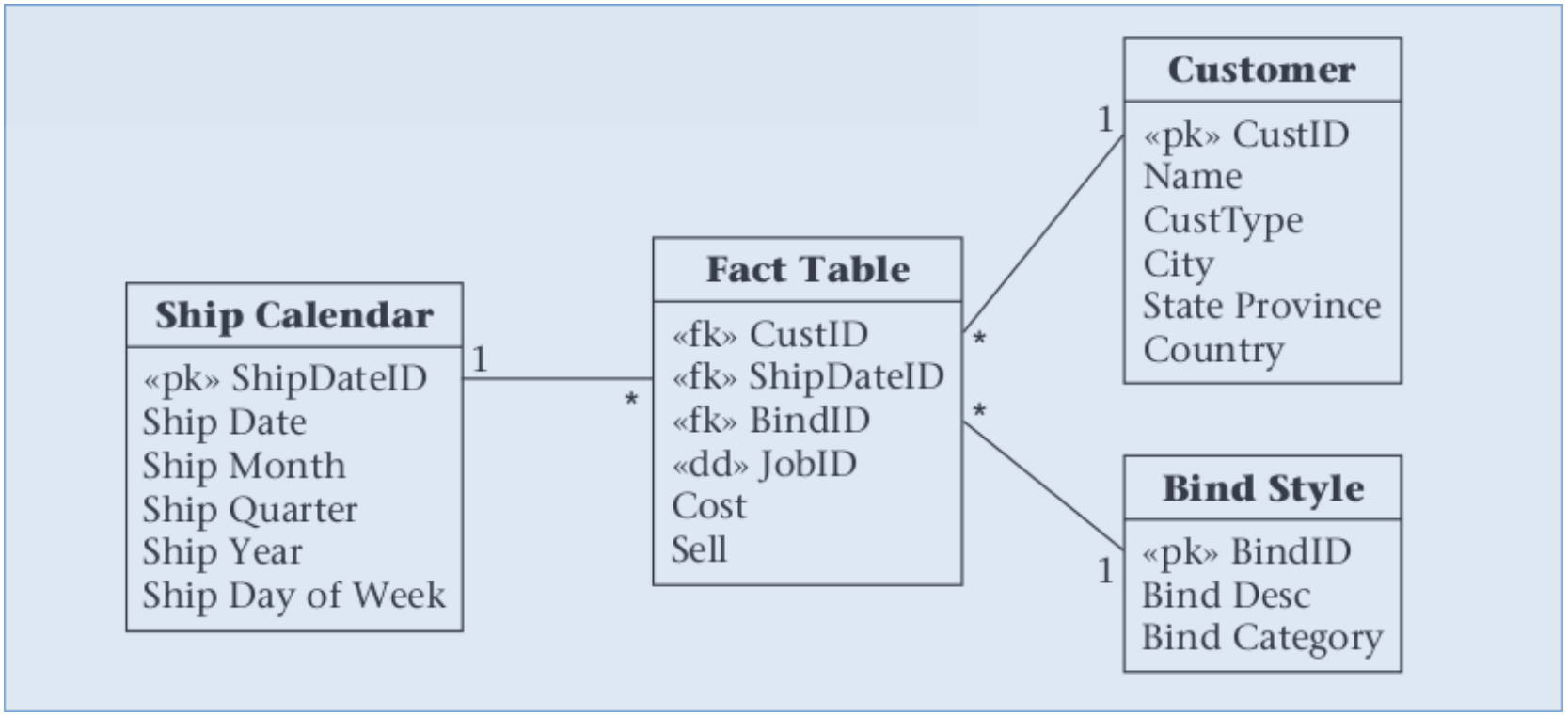
Logično načrtovanje

- Dimenzionalno podatkovno modeliranje
- Imamo različne tipe shem
  - Zvezda
  - Snežinka

Shema zvezda

- PS je organizirana okoli centralne tabele in več manjših dimenzijskih tabel
- Centralna tabela ima dve vrsti atributov
  - Dimenzijski atributi + meritve
  - Večina dimenzijskih atributov so tuji ključi
  - Včasih obstaja dimenzijski atribut brez tabel; degenerirana dimenzija
  - Meritve so vrednosti, ki jih agregiramo oz. vrednosti nad katerimi delamo analizo

Primer zvezdaste sheme



Poizvedbe nad zvezdasto shemo

- Atributi v dimenzijskih tabelah so uporabljeni za selekcijo vrstic v centralni tabeli
- Dimenzijski atributi so uporabljeni za grupiranje vrstic v centralni tabeli
- Dimenzijska tabela ima na razpolago več nivojev podrobnosti, ki jih uporabnik lahko uporabi
- Uporabnik se lahko pomika navzdol ali navzgor po hierarhiji podrobnosti
  - Vrtanje-navzdol in dviganje-navzgor

Poizvedbe nad zvezdasto shemo (2)

- Dimenzijski atributi skupaj tvorijo kandidatni ključ v centralni tabeli
- Nivo podrobnosti, ki jo nudijo dimenzijski atributi predstavlja zrnatost centralne tabele dejstev
- Zrnatost je določena z najbolj podrobnimi zahtevami po podatkih v dani dimenziji
  - Uporabnik želi razločevati v tabeli dejstev vrstice glede na izbrane dimenzijske atribute

Poizvedbe nad zvezdasto shemo (3)

- Normalizacija ni vodilni princip načrtovanja PS
- Osnovni namen PS je omogočiti hiter odziv na vprašanja nad velikimi količinami historičnih podatkov
  - Zvezdasta shema to omogoča
  - Tabela dejstev zbira najbolj pomembne podrobnosti, dimenzijske tabele pa omogočajo več podrobnosti za večjo ceno
- Dimenzijske tabele niso v 3NO
  - Normalizacijski proces bi razbil dimenzijsko tabelo na več tabel

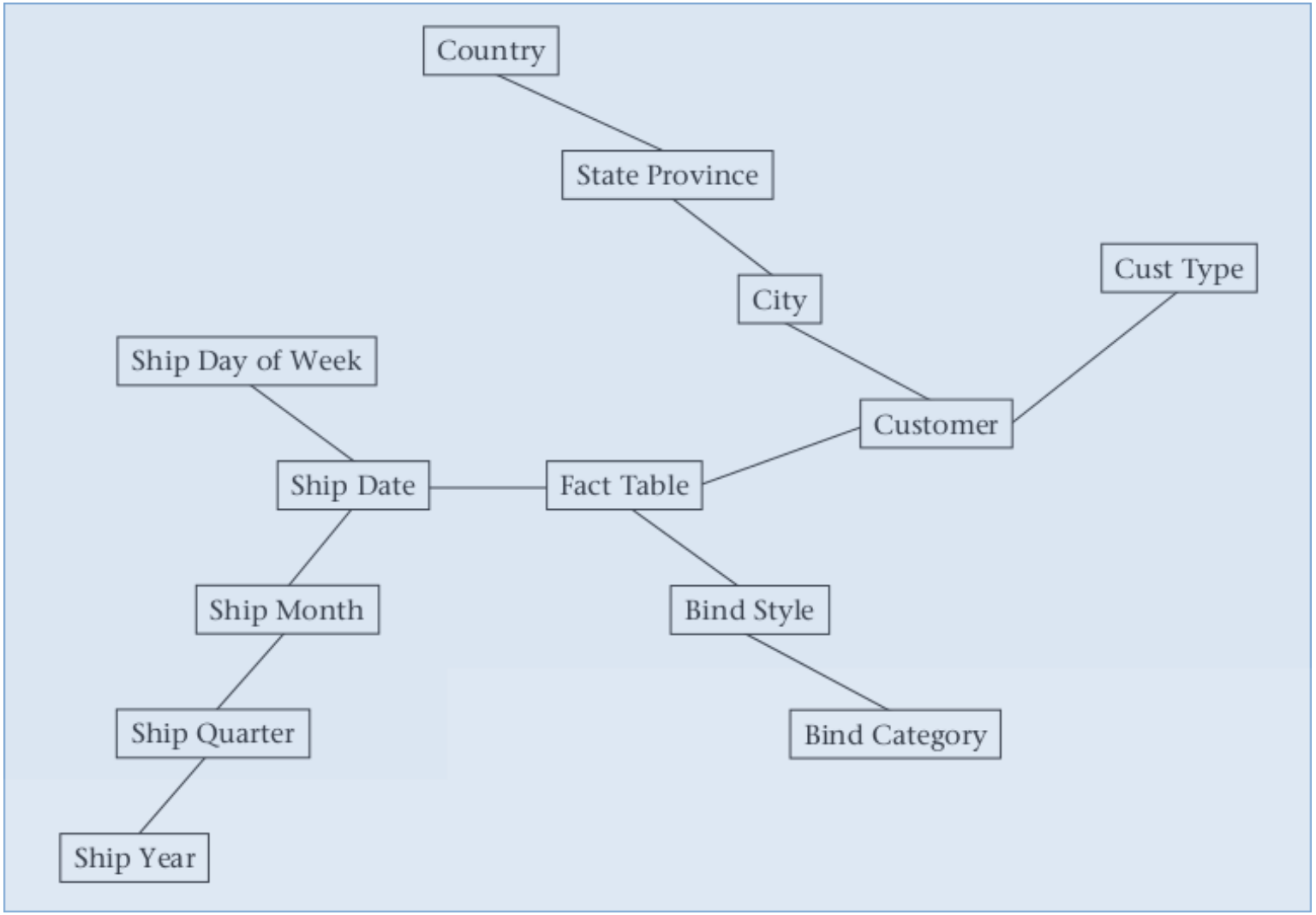
Poizvedbe nad zvezdasto shemo (4)

- Dimenzijske tabele so tipično precej manjše od tabele dejstev
- Dimenzijske tabele se tipično ne spreminjajo pogosto
- Večina operacij nad PS je bralnih
- Prednosti hitrejšega dostopa (brez stikov) enostavnejših vprašanj odtehtajo prednosti normalizacije
- ==> imamo drugačno tehniko načrtovanja PS

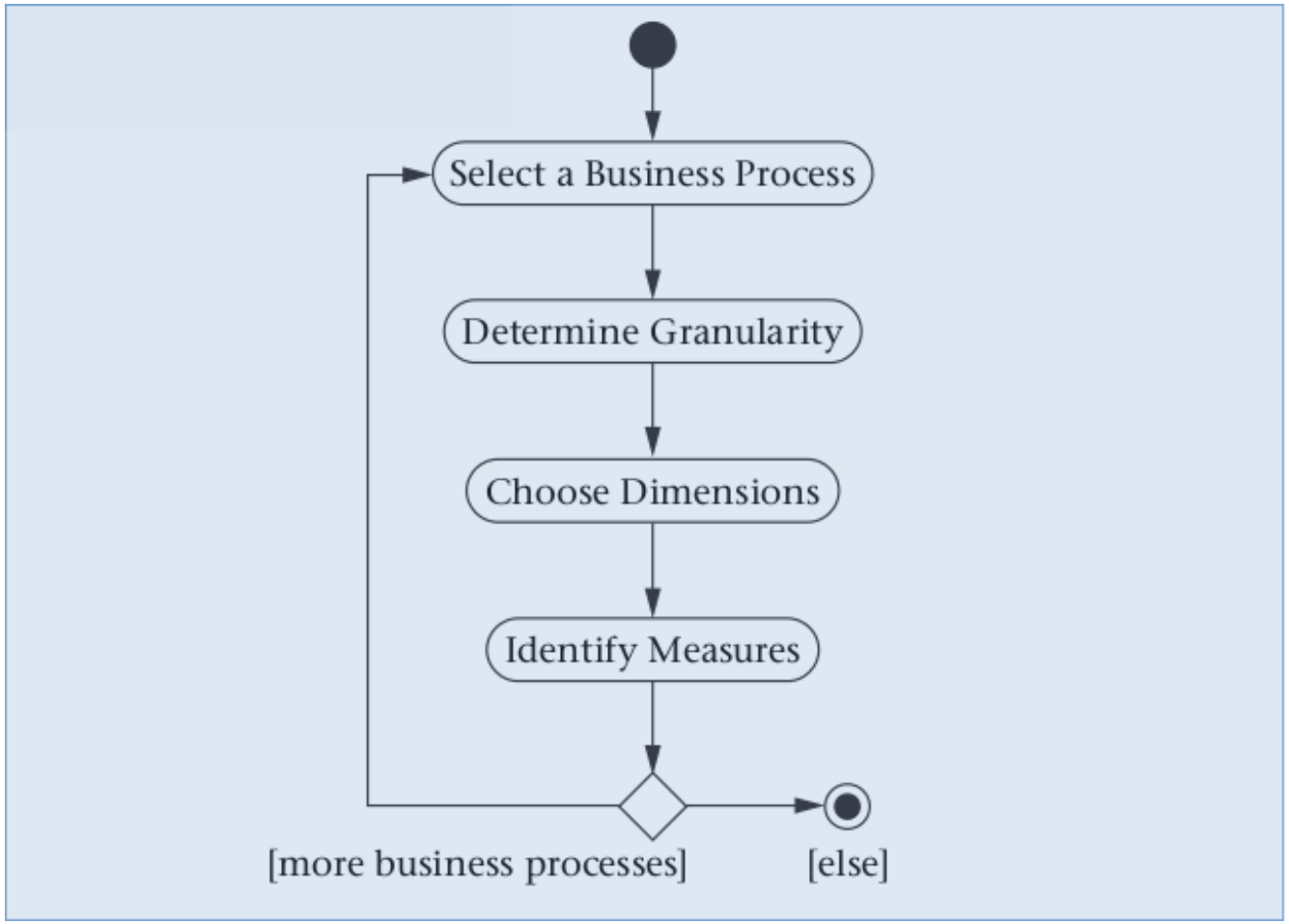
Shema snežinka

- Variacija sheme zvezda je shema snežinka
- Normalizacija dimenzijskih tabel vodi do snežinkaste sheme
- Vsak hierarhični nivo v zvezdasti shemi postane svoja tabela
- Podpira se zvezdasto shemo
  - Enostavnost sheme in principa hierarhičnih nivojev dimenzionalnih tabel
  - Omogoča hitro izvedbo vprašanj

Primer snežinkaste sheme



Dimenzijski načrtovalski proces



Primer dimenzijskega načrtovanja

- XYZ Widget Company - Wish List

XYZ Widget – seznam želja

1. Kakšni so trendi za različne produkte v obliki prihodka od prodaje, profit, itd.?
2. Za produkte, ki niso profitni: lahko izvrtamo zakaj niso profitni?
3. Kako natančno se napovedane cene ujemajo z dejanskimi cenami?
4. Ko spremenimo napovedi: kako to vpliva na prodajo in profit?
5. Kakšni so trendi v procentih poslov, ki so opravljeni pravočasno?

XYZ Widget – seznam želja

6. Kakšni so trendi v produktivnosti po oddelkih, za vsak stroj, in za vsakega zaposlenega?

7. Kakšni so trendi v izpolnjevanju rokov za vsak oddelek in za vsak stroj?

8. Kako učinkovito se je izvedla prenovitev na stroju 123?

9. Katere stranke prinesejo najbolj profitno delo?

10. Kako stanje računa za promocijo vpliva na prodajo in profit?

Poslovni procesi

1. Napovedovanje
2. Razporejanje
3. Spremljanje produktivnosti
4. Cena dela

Napovedovanje

- Vnese se specifikacija naprave
  - Tip naprave določa stroje na katerih se izdeluje
  - Program za napovedovanje oceni koliko časa se bo izdelovala naprava na posameznem stroju
  - Ocena se pomnoži s faktorjem, ki da ceno dela
- Stroj: čas nastavitve + hitrost
- Vsaka ocena vsebuje
  - Specifikacija naprave
  - Sestavne dele z ocenami dela na posameznih strojih
  - Pribitek, popust in cena

Napovedovanje

- Potrditev stranke
  - Če stranka sprejme kvoto se kvota poveže s št. posla
- Specifikacije se izpišejo kot *izpisek posla*
  - Izpisek posla gre naprej v razporejanje

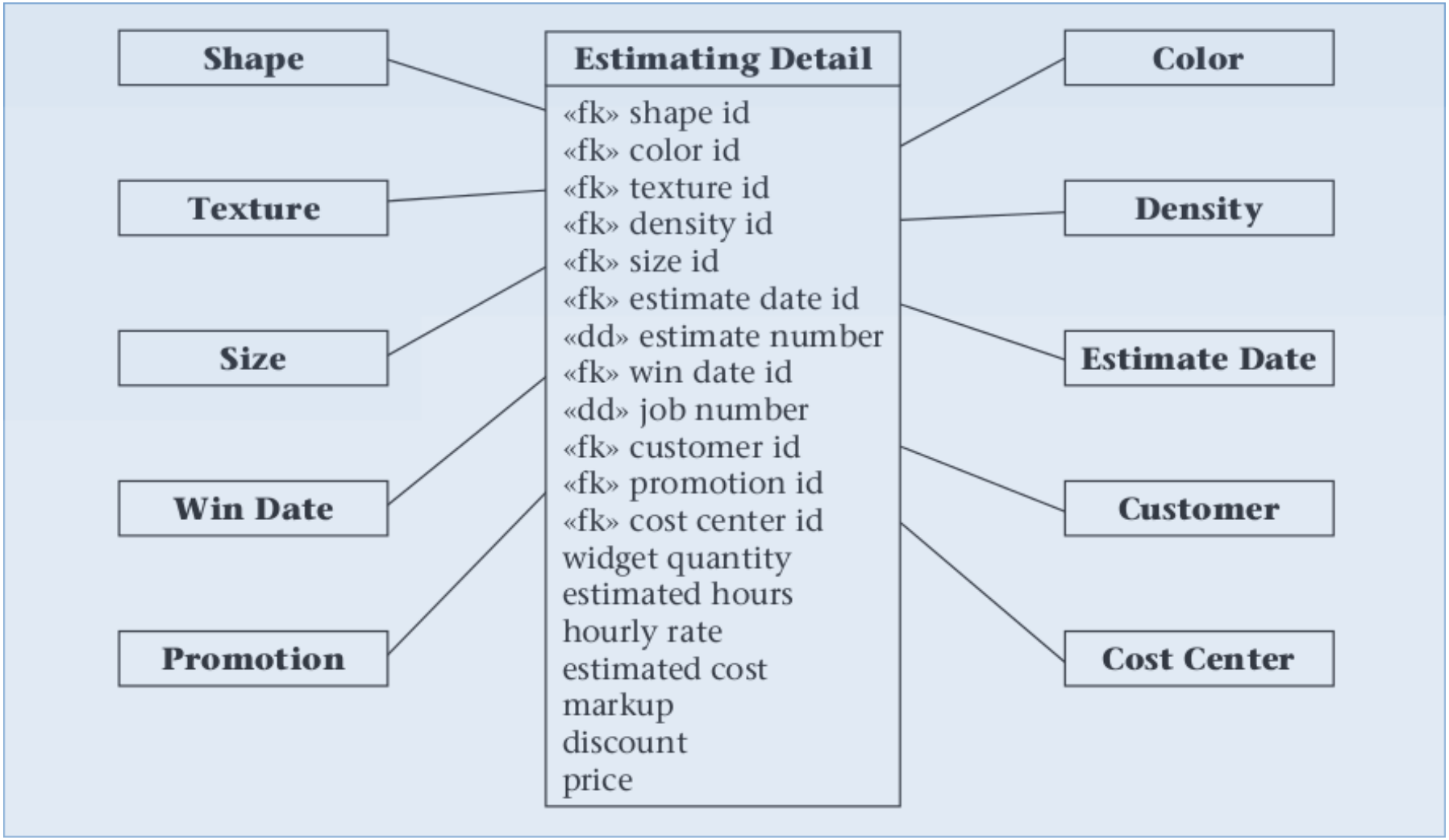
Nivo podrobnosti

- Nivo podrobnosti dimenzij je potrebno planirati
- Najbolj podroben nivo, ki omogoča vrtanje za preučevanje podatkov
- Najbolj podrobni nivoji napovedovanja
  - Podrobnosti v zapisu napovedovanja za center cen

Dimenizije

- Naslednji korak je definicija dimenzij
  - *Atributi:* specifikacija posla, predvideno število in datum, št.posla in datum pridobitve, stranka, reklama, center za cene, količina, ocenjene ure, urna postavka, cena, pribitek in cena.
- Dimenzije so tisti atributi po katerih bi uporabniki želeli imeti grupiranje
  - Različne specifikacije posla, center za cene, datumi, reklame, stranke.
- Izbrani atributi postanejo dimenzije zvezdaste sheme

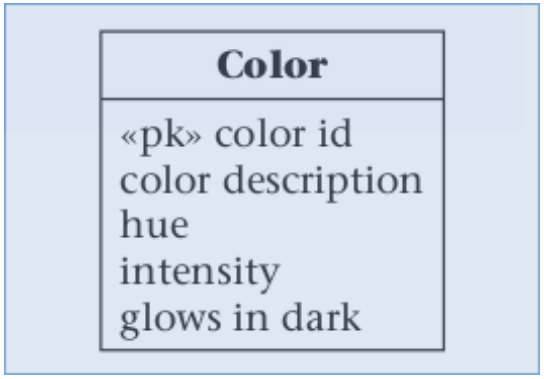
Zvezdasta shema za proces napovedovanja

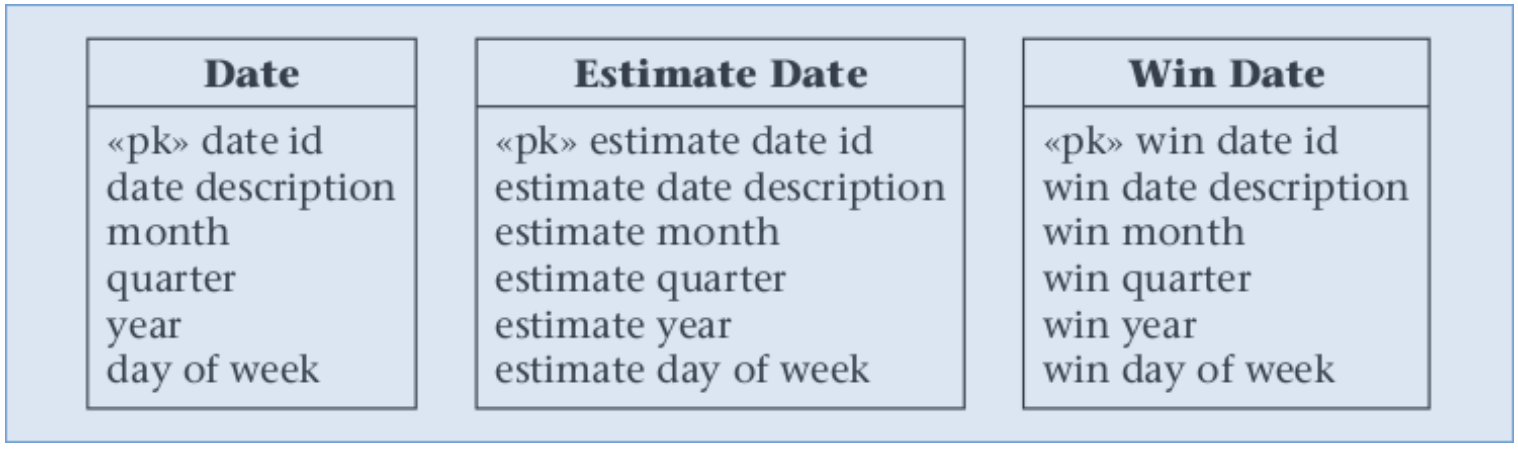


Dimenzije

- Dimenzijska tabela vsebuje vrednosti, ki so lahko pomembne za analizo
  - Hierarhije uporabne za analizo
- Dimenzijske vrednosti so pogosto kategorične
  - Npr. S,L,XL
  - Meritve so običajno numerične
- Dimenzije datumov so zelo pogoste
  - Predlagajo standardizacijo (znotraj firme)

Dimenzije sheme





Razporejanje

- Izpisek posla se uporablja za dodelitev posla vsakem posameznem stroju
  - Dolčijo se ciljni časi
- Izpisek posla se po razporejanju na koncu premakne v proizvodnjo
- Firma ima stroj za avtomatsko zbiranje podatkov
  - Vsak izpisek posla ima bar kodo za posel
  - Vsak delavec ima tudi bar kodo
  - Vsak stroj ima bar kode za vse operacije

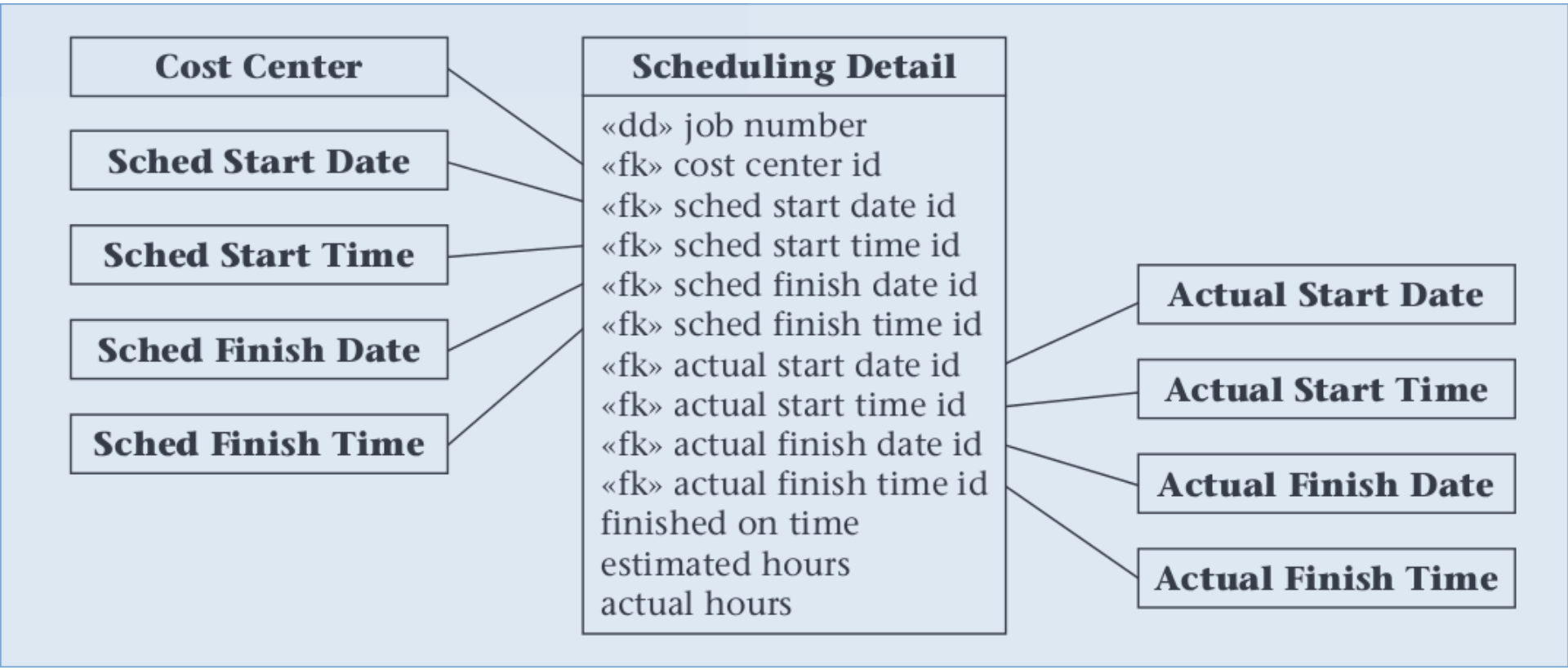
Razporejanje

- Ko se začne določena operacija
  - Preberejo se bar kode operacije, posla in delavca
  - Trenuten čas je čas začetka operacije
  - Ko se operacija začne se konča prejšnja operacija
  - Na koncu delavec preko stroja za vnos podatkov zabeleži zaključek dela
- Avtomatski vnos podatkov se uporablja za
  - Podatke razporejanja, delo in obremenitev zaposlenega in sledenje obremenitvi stroja

Razporejanje

- Nivo podrobnosti zapisa razporejanja
  - Zapis posla gre v center za cene
  - Uporabniki so zainteresirani za vrtanje do tega nivoja
  - Primeren nivo podrobnosti za zvezdasto shemo se določi na osnovi št.posla in centra za cene
- Naprej določimo nivo podrobnosti za krake sheme razporejanja
  - Začetni/končni predviden čas, začetek/konec dejanske izvedbe, predvidno in dejansko trajanje operacije, center za cene, zastavica, ki pove glede pravočasnosti opravila, ...
  - Št.posla je degenrirana dimenzija

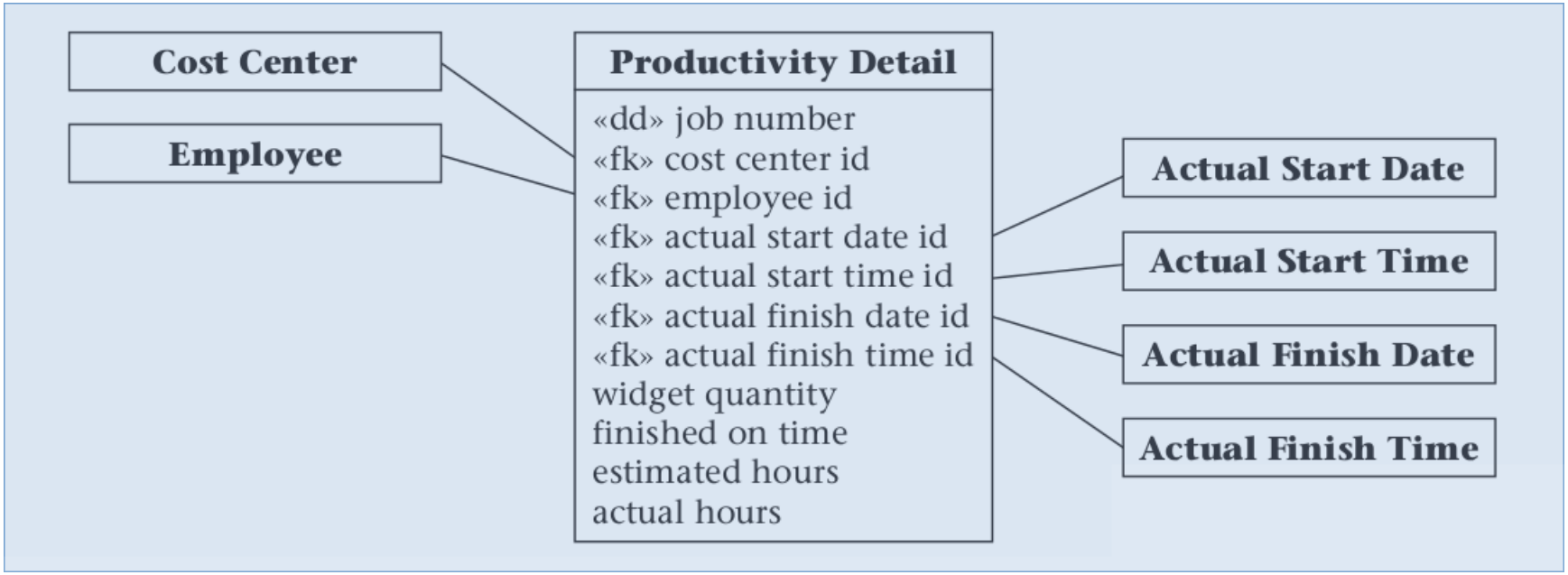
Shema za proces razporejanja



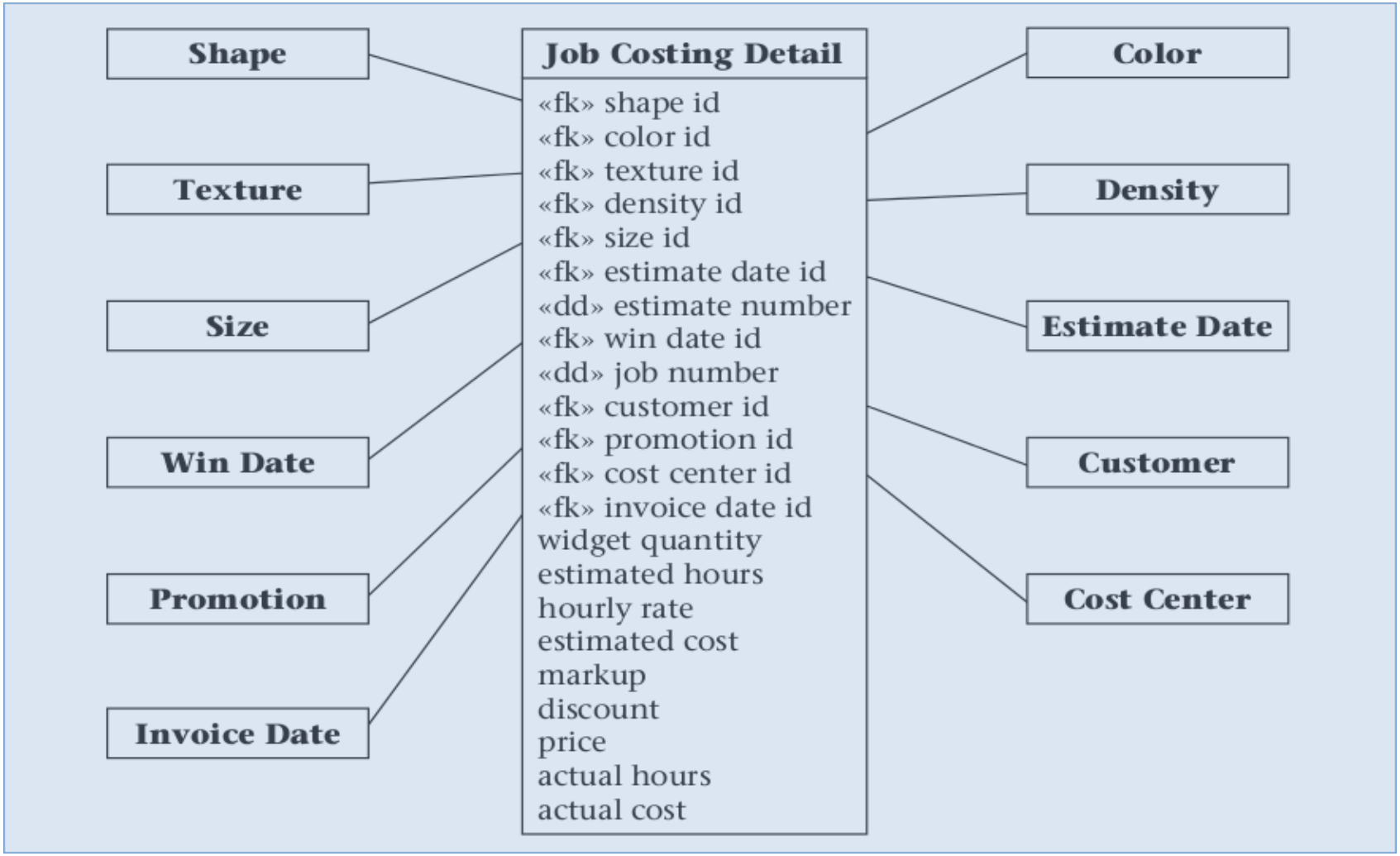
Dimenzije

- Imamo precej časovnih zapisov
  - Dobro je imeti standardizirane časovne zapise
- Dimenzije se lahko uporabljajo v več shemah
  - Dimenzija centra za cene je v dveh shemah
  - Predviden čas začetka/konca operacije je v obeh shemah
  - Dimenzije z istim pomenom imajo isto ime !

Shema za spremljanje produktivnosti



Shema za spremljanje cene dela



Dimenzije PS

	Shape	*Color*	*Texture*	*Density*	*Size*	*Estimate Date*	*Win Date*	*Customer*	*Promotion*	*Cost Center*	*Sched Start Date*	*Sched Start Time*	*Sched Finish Date*	*Sched Finish Time*	*Actual Start Date*	*Actual Start Time*	*Actual Finish Date*	*Actual Finish Time*	*Employee*	*Invoice Date*
Estimating	X	X	X	X	X	X	X	X	X	X										
Scheduling										X	X	X	X	X	X	X	X	X		
Productivity Tracking										X					X	X	X	X	X	
Job Costing	X	X	X	X	X			X	X	X										X

Sumarna vprašanja

- Tabela dejstev je primerna za pregledovanje podrobnosti
- Uporabniki pogosto želijo povzetke
  - Vse se lahko dobi iz celotne tabele dejstev
  - Sumarna vprašanja se izvedejo nad tabelo dejstev
  - Tabela dejstev je lahko zelo velika (>> nekaj M)
  - Sumarna vprašanja so tako zelo draga (cpu)
- Sumarne tabele lahko pripravimo vnaprej

OLAP

- Uporaba sumarnih tabel
  - Načrtovanje in implementacija strateških sumarnih tabel je dober pristop, če imamo majhno število pogostih sumarnih vprašanj
  - Velikokrat je potreben ad-hoc način preiskovanja podatkov, ki onemogoči pripravo sumarnih tabel
- OLAP nudi alternativo
  - Avtomatično izbere strateško množico sumarnih oken in shrani avtomatične sumarne tabele (AST) kot materializirana okna
  - Okna se popravljajo ob popravkih tabele dejstev

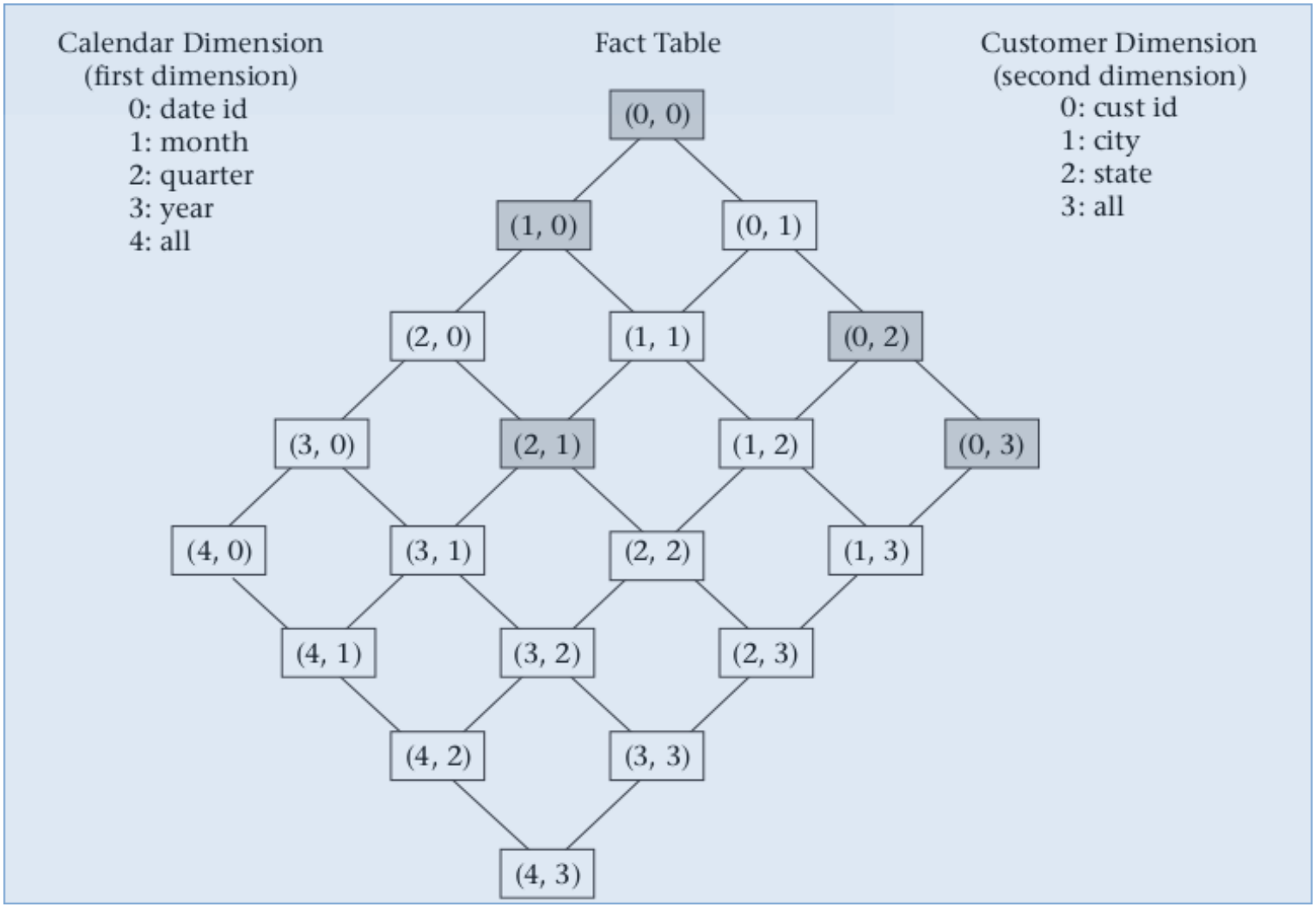
OLAP (2)

- Če uporabnik potrebuje sumarne podatke
  - Sistem odkrije katere AST potrebuje
- OLAP sistemi so dobra rešitev za ad-hoc preiskovanje sumarnih podatkov osnovanih na veliki količini podatkov PS

Eksplozija oken

- Materializirana okna agregirana iz tabele dejstev
  - Hierarhija po dimenziji: 0, 1, 2, ...
  - Lahko enolično identificiramo glede na agregacijski nivo
  - Datum Računa: 0 – racID, 1 – mesec, 2 – kvartal, 3 – leto, 4 - „vse“
  - Dimenzija nima hierarhije: 0 – brez agregacije, 1 – agregirano po celotni dimenziji
- Dobimo več-dimenzionalen prostor glede na hierarhijo dimenzij

Graf produkta dveh dimenzij



Produkt dveh dimenzij

- Na vrhu (0,0) imamo tabelo dejstev
- Vsako vozlišče (i,j) predstavlja okno z agregacijskimi nivoji i in j
- Relacija vozlišča do nižjih vozlišč je agregacija
- Vprašanje za (i,j) se lahko odgovori iz nadrejenih vozlišč (<i,<j), ki so materializirana z agregacijo
- Kvartali se integrirajo v leta in mesta v države

Eksplozija oken

- Graf produkta dimenzij označen z nivoji agregacije
- Št. možnih pogledov: $\prod h_i$, za i $\in$ [1..d]
- Če je g povprečna dimenzija dobimo: g^d
- Dimenzionalnost se povečuje linearno, št pogledov eksponentno
- $5^{10} = 9765625$
  - OLAP ne more vzdrževati vseh pogledov
  - Tipično se izbere podmnožica oken, ki se jih materializira

Pregled OLAP