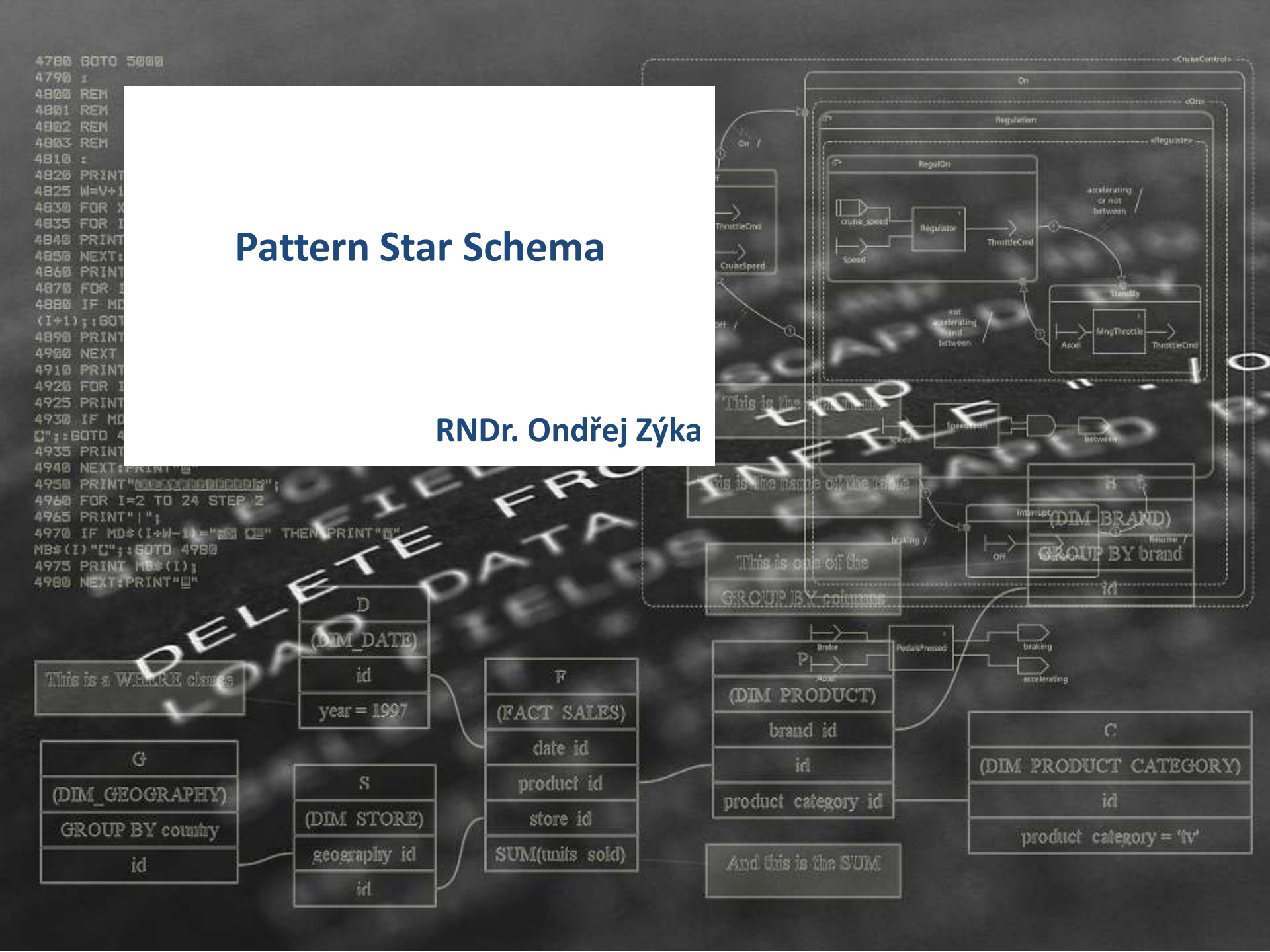
Pattern Star Schema

RNDr. Ondřej Zýka



Star Schema

- Pattern spojený s datovými sklady a systémy pro reporting
- Modely používané pro Dimenzionální modelování

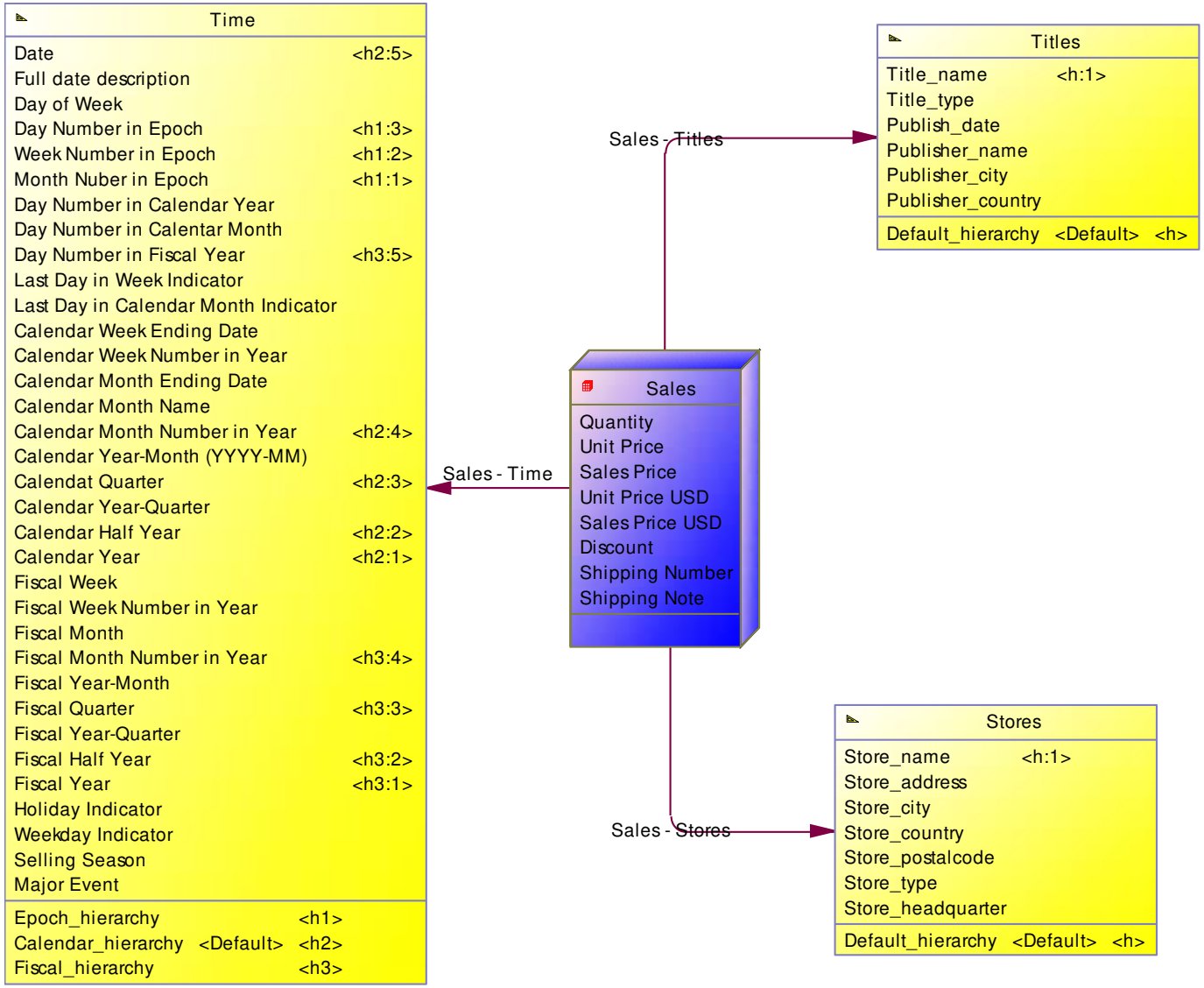
- Něco historie
  - Bill Inmon – Corporate Information Factory – používá pro Data Marty
  - Ralph Komball (1997) – Dimensional Data Warehouse – používá pro celé řešení
  - Stand-Alone Data Marts – doporučený pattern

- Pattern odděluje do tabulek
  - Fakta – hodnoty, které se mají analyzovat (ceny, počty, ..)
  - Dimenze - číselníky, podle kterých se má analyzovat (zákazníci, čas, produkty, pobočky, ...)

Star Schema

- Silně denormalizovaný model
- Pochopitelné pro netechnicky orientované uživatele (byznys uživatele)
- Orientované na analytické dotazy
- Umožňují analyzovat extrémní objemy dat
- Podporuje historizaci a dosažitelnost historických dat
- Podporované datovými servery a analytickými nástroji (OLAP)
- Výhodný pro dotazovací servery, které jsou specializované na dotazy a mají špatný výkon pro update a delete operace.

Příklad – Star schéma



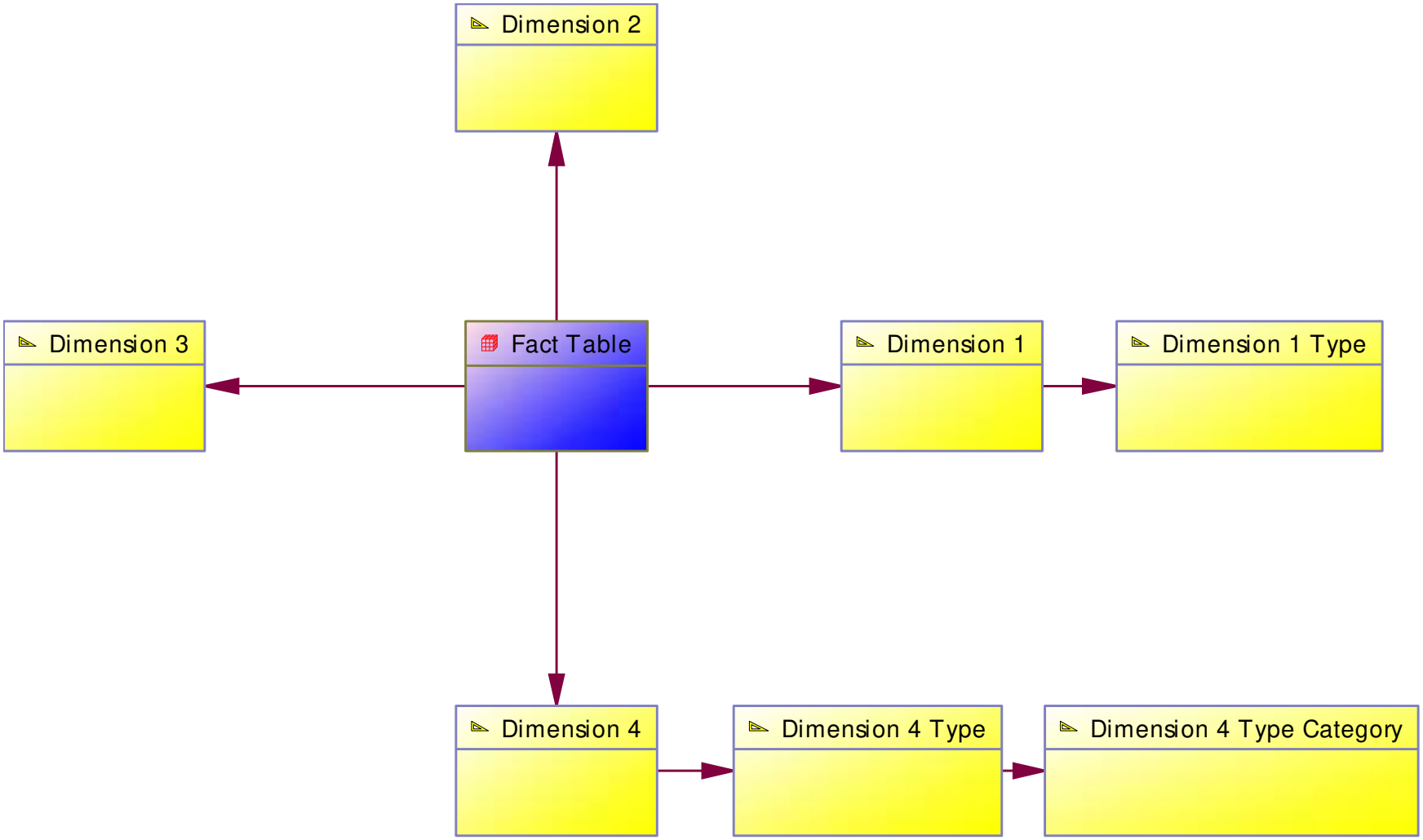
Standardní dotaz

```
select SUM(qty) from
   F_SALES,D_TIME,D_TITLES,D_STORES
where
  F_SALES.TITLES_KEY = D_TITLES.TITLES_KEY
  and F_SALES.STORES_KEY = D_STORES.STORES_KEY
  and F_SALES.DATE_KEY = D_DATE. DATE_KEY

  and podminky na D_TITLES
  and podminky na D_STORES
  and podminky na D_DATE

group by
  pozadovana granularita vysledku
```

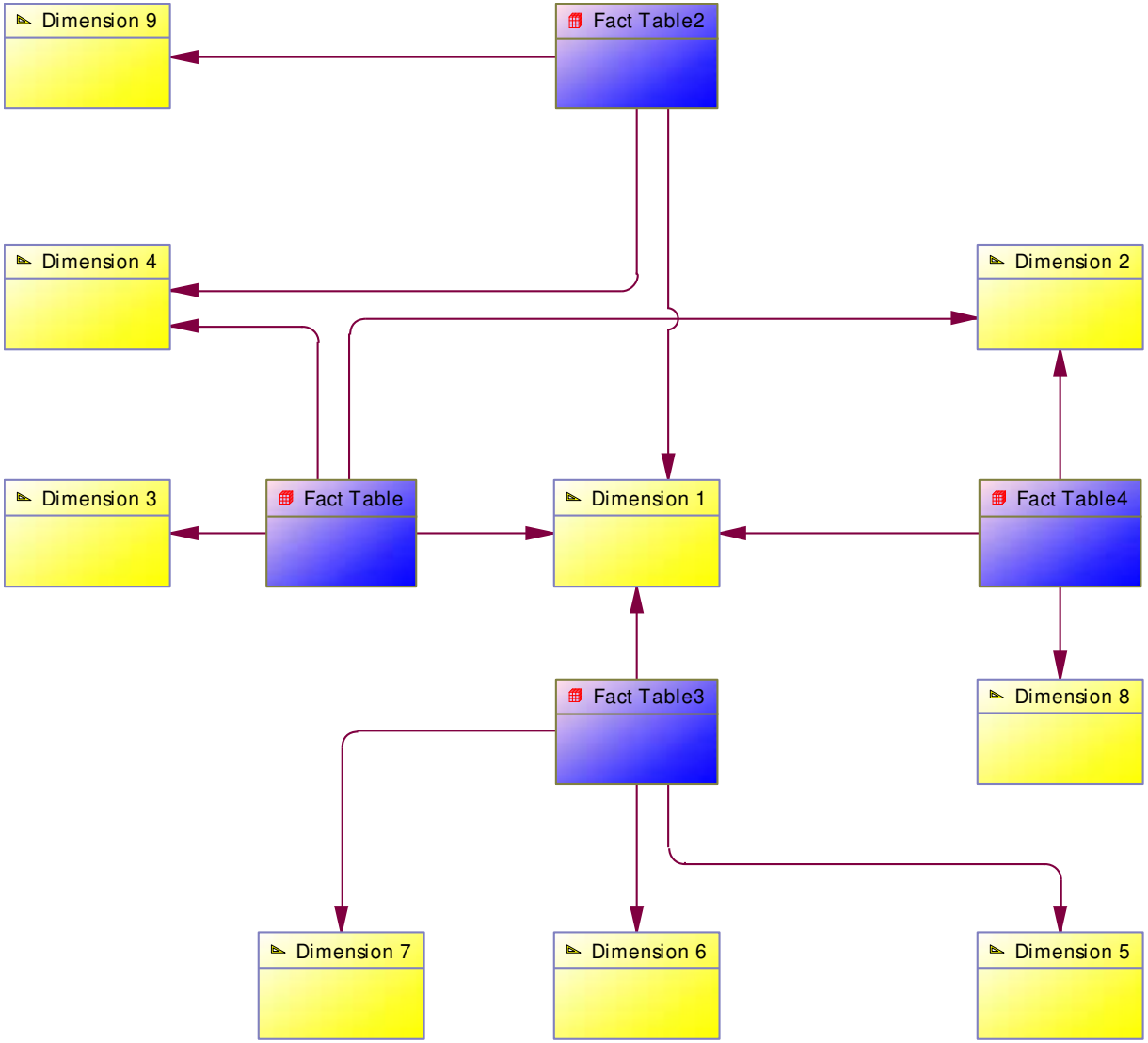
Snowflake schéma



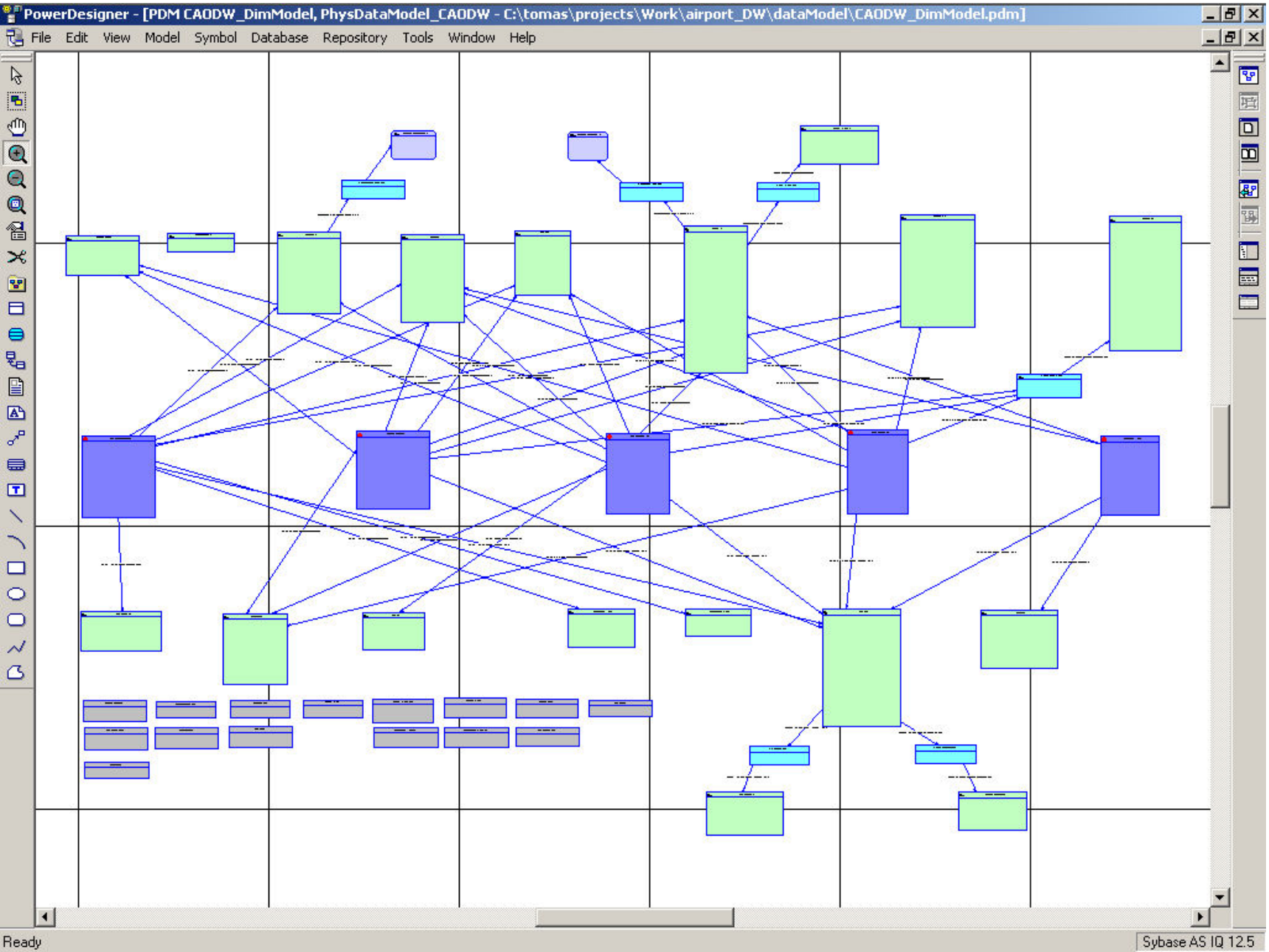
Snowflake model

- Výhody
  - Minimální redundance dat v rámci dimenzí
  - Úspora místa v databázi
  - Větší flexibilita pro modelování
  - Užitečný pro dimenze se složitou strukturou
- Nevýhody
  - Složitější konstrukce dotazů, mnoho joinů
  - Nižší výkonnost
  - Komplikovaný snowflake model může odradit uživatele od přímého přístupu k datům
    - uživatelské nástroje zpravidla zavádějí sémantickou vrstvu, která uživatele odstíní od datového modelu
  - Možný konflikt s bitmapovými indexy
  - Úspora místa je většinou převážena nižší výkonností a složitější administrací

Constellation schema



Snowstorm schema



Vytvořen dimenzionálního modelu

- Výběr Sledovaných procesů
- Definice Metrik
- Definice Dimenzí
- Definice Hierarchií
- Definice Granularity

- Plnění dimenzionálního modelu
- Technologie
  - Relační databáze
  - OLAP technologie

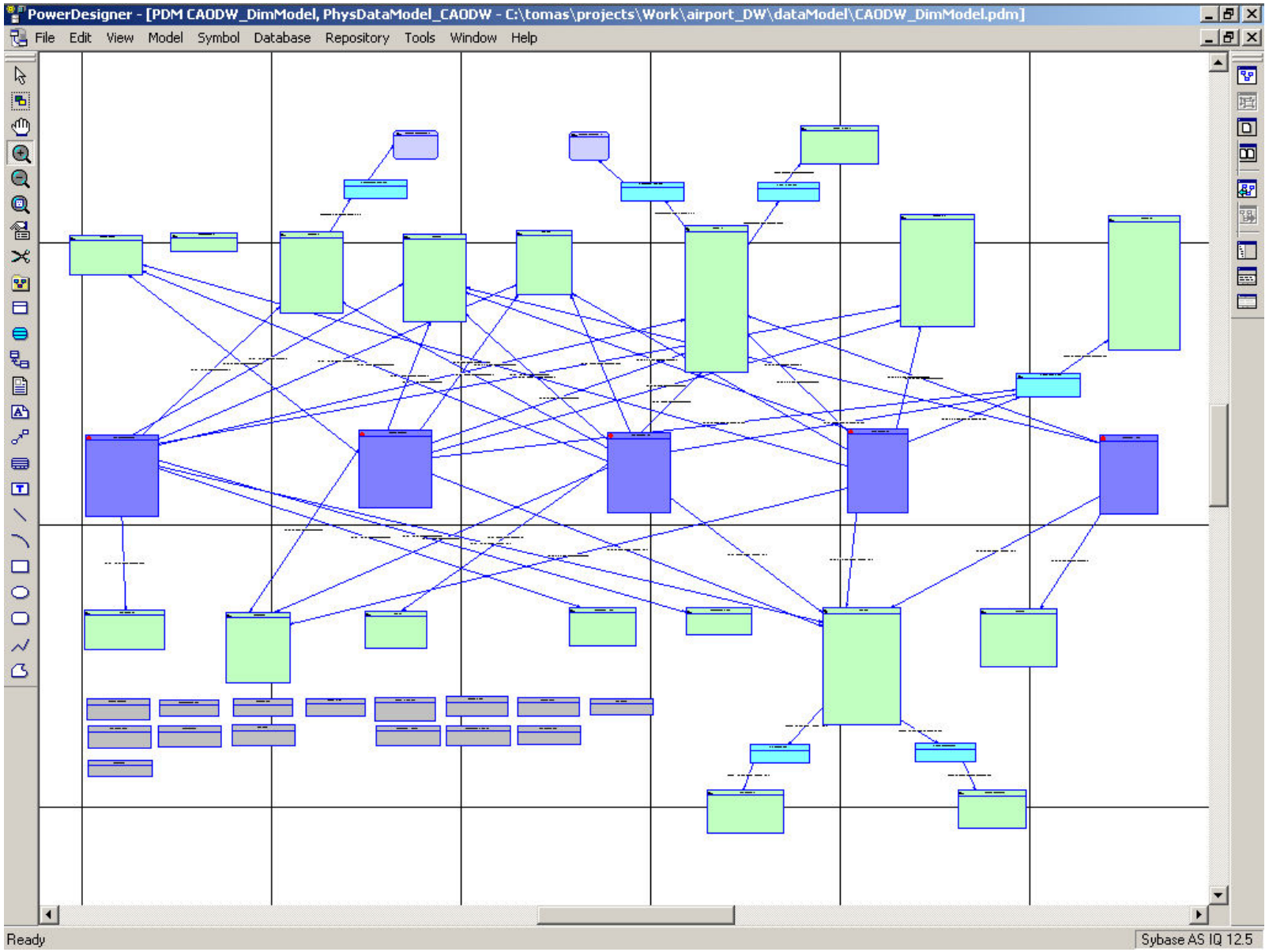
Výběr sledovaných procesů

- Seznam procesů, které chceme analyzovat
  - Od jednodušších ke složitějším
- Bus matrix
  - Matice: Business procesy x Dimenze
- Často odpovídá jeden business proces ≈ jeden datamart ≈ jedna hvězda

Buss matrix – příklad Letiště

Data Marts	dimDate	dimTime	dimCarrier	dimAirport	dimAircraft	dimPAX	dimHandling	dimDelay	dimFlight	dimPartner	dimCancelOfFlight	dimSource	dimFlightAttrib (snowflake)	dimStateAttrib (snowflake)	dimCarrierAttrib (snowflake)	dimSourceSystem (snowflake)	dimLandedFee	dimShareCode
Flight movement	X	X	X	X	X				X	X	X						X	
Source usage	X	X	X		X		X		X			X						X
Passenger (PAX)	X	X	X	X		X	X		X			X						X
Goods (CARGO)	X	X	X		X		X		X			X						X
Delay of flight	X		X					X	X									

Bus matrix – příklad implementace



Bus matrix – příklad telco operátor

	Date	Customer	Service	Rate Category	Local Svc Provider	Calling Party	Called Party	Long Dist Provider	Internal Organization	Employee	Location	Equipment Type	Supplier	Item Shipped	Weather	Account status
Custommer Billing	×	×	×	×	×			×			×					×
Service Orders	×	×	×		×			×	×	×	×	×			×	×
Trouble Reports	×	×	×		×	×		×	×	×	×	×	×	×	×	×
Yellow Page Ads	×	×		×		×			×	×	×					×
Customer Inquiries	×	×	×	×	×	×		×	×	×	×				×	×
Promotion	×	×	×	×	×	×		×	×	×	×	×	×	×		×
Billing Call Detail	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Network Call Detail	×	×	×	×	×	×	×	×	×		×	×	×	×	×	×
Customer Inventory	×	×	×	×	×			×	×		×	×	×	×		×
Network Inventory	×		×						×	×	×	×	×	×		
Real eastate	×								×	×	×	×				
Labor & Payroll	×								×	×	×					
Computer Charges	×	×	×		×			×	×	×	×	×	×	×		
Purchase Orders	×								×	×	×	×	×	×		
Supplier Deliverables	×								×	×	×	×	×	×		

Tabulky faktů

- Transaction - co řádek to transakce (například obchody)
  - Proces může obsahovat více typů transakcí, rozhodnutí zda jedna nebo více tabulek není jednoduché
- Snapshots - každý den se udělá celý snímek
  - State model – celé denní snímky
  - Event model – každý den pouze změněné záznamy
  - Možnost dopočítání dalších hodnot ke každému snímku
- Akumulujíce se shapshoty (sklad)
  - Id výrobku jako primární klíč a doplňují/updatují se hodnoty pro události popisující životní cyklus
  - Do daného řádku se doplní datum expedice, fakturace, dodání, vyúčtování, …
  - Pozor - update v tabulce faktů
- (Fact tables bez faktů – slouží jako n:n vazba mezi dimenzemi)

Fact tables

- Fakta
  - aditivní - počet, cena v transakčních fact tabulkách
    - Význam pro všaechny dimenze
    - Nejlépe se s nimi pracuje
    - Cílem je převést na aditivní fakta maximum
      - Discount -> ceníková cena, prodejní cena
  - semiaditivní - počet cena v snapshot tabulkách
    - součet za produkty má význam, za čas nemá význam
    - Obecně význam pouze pro některé dimenze
  - nonadditive - procentuální profit
    - Často text
    - Někdy možné přenést do dimenzí (degenerované dimenze)
- Factless fact table – pouze cizí klíče, žádná fakta
  - Příznak existence (účast v kampani)

Určení dimenzí

- Konformní dimenze
  - Jedna nejpodrobnější dimenze, ostatní jsou jejich agregací
  - Jednotné dimenze pro všechny business procesy
- Jeden sloupec primárního klíče
- Hodně sloupců popisů, často přes 30, čím více tím lépe
- Atributy spíše textové (srozumitelnost)
- Hierarchie pro analýzy
- Časová dimenze
- Degenerovaná dimenze – nemá popis (číslo faktury)
- Dimenze jsou denormalizované (jedna široká tabulka)
  - Normalizace – vločkové schéma
- Umělé klíče pro odstínění změn
- Řádek s hodnotu „Not applicable“, „Uknown“

Časová dimenze

- V každém datovém skladu
- Často mnoho hierarchií
  - Provozní rok
  - Fiskální rok
  - Kalendářní rok
- Mnoho sloupců
  - Textová informace
  - Číselná informace
  - Konce a začátky období
  - …
- Umožňuje dotazy na
  - Kalendářní i fiskální kalendář
  - Volné dny
  - Dny v týdnu
  - …

Time	
Date	<h2:5>
Full date description	
Day of Week	
Day Number in Epoch	<h1:3>
Week Number in Epoch	<h1:2>
Month Nuber in Epoch	<h1:1>
Day Number in Calendar Year	
Day Number in Calentar Month	
Day Number in Fiscal Year	<h3:5>
Last Day in Week Indicator	
Last Day in Calendar Month Indicator	
Calendar Week Ending Date	
Calendar Week Number in Year	
Calendar Month Ending Date	
Calendar Month Name	
Calendar Month Number in Year	<h2:4>
Calendar Year-Month (YYYY-MM)	
Calendat Quarter	<h2:3>
Calendar Year-Quarter	
Calendar Half Year	<h2:2>
Calendar Year	<h2:1>
Fiscal Week	
Fiscal Week Number in Year	
Fiscal Month	
Fiscal Month Number in Year	<h3:4>
Fiscal Year-Month	

Časová dimenze

- Jména anglicky, česky nebo alternativní jména
- Definice prvního dne týdne
- Definice fiskálního roku
- Definice formátu datumu (Date)
- Definice svátků a prázdnin

Name	Data type	Example
Date Key	Integer	20180215
Date	String	15.02.2018
Full date description	String	Tue Feb 13 2018
Day of Week	String	Thursday
Dey Number in Week	Integer	4
Day Number in Epoch	Integer	6621
Week Number in Epoch	Integer	946
Month Nuber in Epoch	Integer	217
Day Number in Calendar Year	Integer	43
Day Number in Calentar Month	Integer	15
Day Number in Fiscal Year	Integer	43
Last Day in Week Indicator	Char	N
Last Day in Calendar Month Indicator	Char	N
Calendar Week Ending Date	Integer	20180218
Calendar Week Number in Year	Integer	6
Calendar Month Ending Date	Integer	20180228
Calendar Month Name	String	February
Calendar Month Number in Year	Integer	2
Calendar Year-Month (YYYY-MM)	String	201802
Calendat Quarter	String	Q1
Calendar Year-Quarter	String	2018Q1
Calendar Half Year	String	2018H1
Calendar Year	String	2018
Fiscal Week	String	F Week 6 2018
Fiscal Week Number in Year	Integer	6
Fiscal Month	String	F Month 2 2018
Fiscal Month Number in Year	Integer	2
Fiscal Year-Month	String	F201802
Fiscal Quarter	String	FQ1
Fiscal Year-Quarter	String	F2018Q1
Fiscal Half Year	String	F2018H1
Fiscal Year	String	F2018
Holiday Indicator	Char	N
Weekday Indicator	Char	Y
Selling Season	String	S201801

Typy dimenzí z pohledu změn

- Statické
  - Žádné ošetření změn
  - Změna hodnot znamená změnu řešení
- Rostoucí dimenze
  - Přidávají se nové záznamy
  - Změna záznamu je chyba zpracování
- Rychle rostoucí dimenze – změny několikrát denně
  - Nutné speciální řešení
  - Oddělení rychle se měnících atributů do vlastní dimenze
  - (Jako Slowly changed dimension Type 2)
- Slowly changing dimenze
  - Změny maximálně jednou denně
  - Mnoho definic různých typů

Změny v dimenzích

- Identifikace záznamu – Musí být definován identifikátor záznamu (přirozený klíč, kód)
- Je nutné rozlišovat Umělý klíč záznamu a Identifikátor řádku.
- Je možné uchovávat historické hodnoty v oddělené tabulce.
- Historizace musí řešit
  - Vznik nového záznamu
  - Změnu záznamu
  - Zrušení záznamu
  - Opětný vznik zrušeného záznamu se stejným identifikátorem
- Řešení může podporovat různé typy historizace pro různé sloupce tabulky. To vede ke změně modelu.

Slowly changing dimension

- Typ 0 – ignorování změn
- Typ 1 – přepis hodnot
  - Žádná historie
- Typ 2 – přidávání řádků, vždy jeden platný
  - Přidané pole: DWH_START_DATE, DWH_END_DATE, DWH_CURRENT_FLAG
  - Kompletní historie
- Typ 3 – přidání sloupců s historickými hodnotami
  - Současné a předchozí hodnoty uchovávány v různých sloupcích (omezená délka historie)

- Redundance nebývá problém
  - Dimenze zabírají cca 5% místa v DWH

SCD2

Nový záznam 01.01.2017

USER_CODE	USER_NAME	USER_LOCATION
peterh	Peter Horffer	Prague

USER_KEY	USER_CODE	USER_NAME	USER_LOCATION	DWH_START_DATE	DWH_END_DATE	DWH_CURRENT_FLAG
100	peterh	Peter Horffer	Prague	01.01.2017	31.12.2999	Y

- Technologické sloupce s identifikovatelnými jmény
- DWH_END_DATE – definovaná maximální hodnota, nepoužívat null
- DWH_START_DATE, DWH_END_DATE – datum dávky, identifikace aktuálnosti dat v datovém skladu (nemusí odpovídat datu zpracování), nedá se vztahovat k byznys platnosti záznamu.
- Hodnoty jako Recorded date, Load date, Transaction date, Effective date, Booking date a podobně musí být přesně definovány na analytické úrovni a zpracovávány jako standardní atributy entity.

SCD2

Změna záznamu 5.5.2017

USER_CODE	USER_NAME	USER_LOCATION
peterh	Peter Horffer	London

USER_KEY	USER_CODE	USER_NAME	USER_LOCATION	DWH_START_DATE	DWH_END_DATE	DWH_CURRENT_FLAG
100	peterh	Peter Horffer	Prague	01.01.2017	04.05.2017	N
556	peterh	Peter Horffer	London	05.05.2017	31.12.2999	Y

- User_key je identifikace řádky (entita s verzí), nikoliv samotné entity.
- DWH_END_DATE je o den menší, než DWH_START_DATE

Zrušení záznamu
10.10.2017

USER_KEY	USER_CODE	USER_NAME	USER_LOCATION	DWH_START_DATE	DWH_END_DATE	DWH_CURRENT_FLAG
100	peterh	Peter Horffer	Prague	01.01.2017	04.05.2017	N
556	peterh	Peter Horffer	London	05.05.2017	09.10.2017	N

- Pouze ukončení platnosti záznamu

SCD2

Nový záznam se stejným kódem 2.1.2018

USER_CODE	USER_NAME	USER_LOCATION
peterh	Peter Horffer	London

USER_KEY	USER_CODE	USER_NAME	USER_LOCATION	DWH_START_DATE	DWH_END_DATE	DWH_CURRENT_FLAG
100	peterh	Peter Horffer	Prague	01.01.2017	04.05.2017	N
556	peterh	Peter Horffer	London	05.05.2017	09.10.2017	N
892	peterh	Peter Horffer	London	02.01.2018	31.12.2999	Y

- Není zachována nepřerušená řada ve v auditních sloupcích.

SCD2 - poznámky

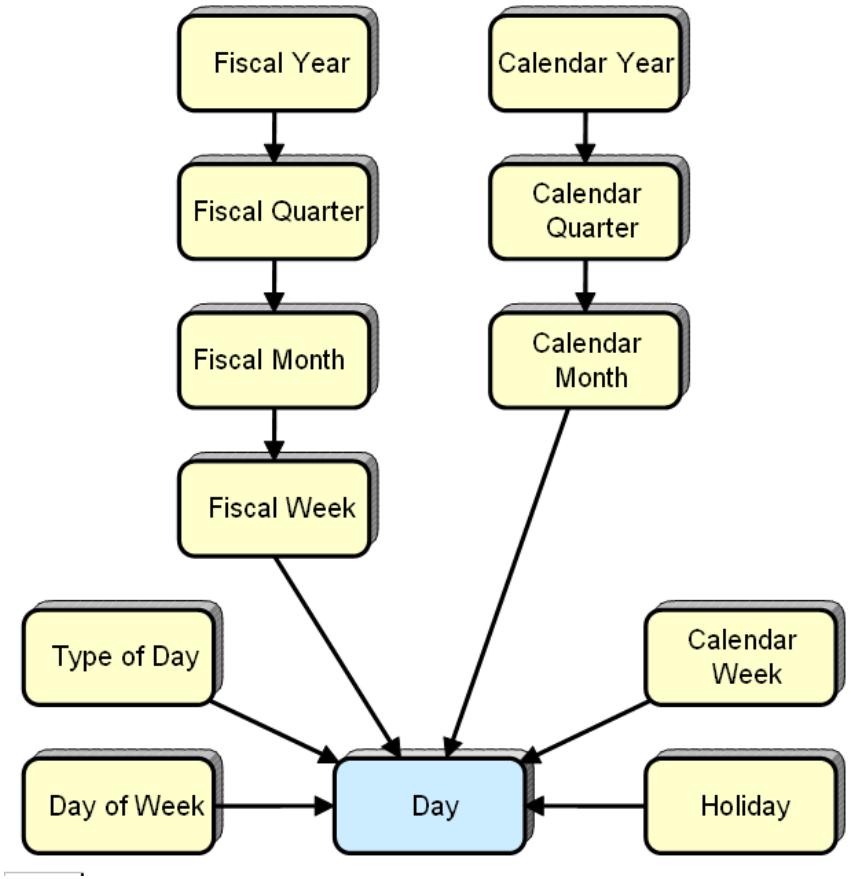
- Většinou více auditních sloupců
  - DWH_DELETED_FLAG
  - DWH_TRAMSFORMATION_ID
  - DWH_PROCESS_ID
  - DWH_ROW_HASH
  - DWH_DATA_QUALITY_STATUS
  - …
- Neměnící se USER_KEY, USER_KEY odpovídá USER_CODE.
  - USER_KEY není primární klíč tabulky
- Možnost více časových os
  - DWH_START_TIME, DWH_END_TIME – odpovídá přesnému času zpracování
  - Údaj typu TIMESTAMP, DWH_START_TIME = DWH_END_TIME předchozího řádku
  - Umožňuje vytvářet dotazy k danému okamžiku v historii skladu.

- Nutnost přesné definice pro konkrétní řešení a generování historizace pomocí odzkoušených template.

Další typy dimenzí

- Konformní
  - Pro celý podnik
  - Ostatní dimenze jako podimenze konformních dimenzí
- Minidimenze a sběrné dimenze
  - Číselníky
  - Stavové a textové atributy
  - Možné sloučit do sběrných dimenzí
- Degenerované dimenze
  - Přímo v tabulce faktů (číslo objednávky)

Hierarchie

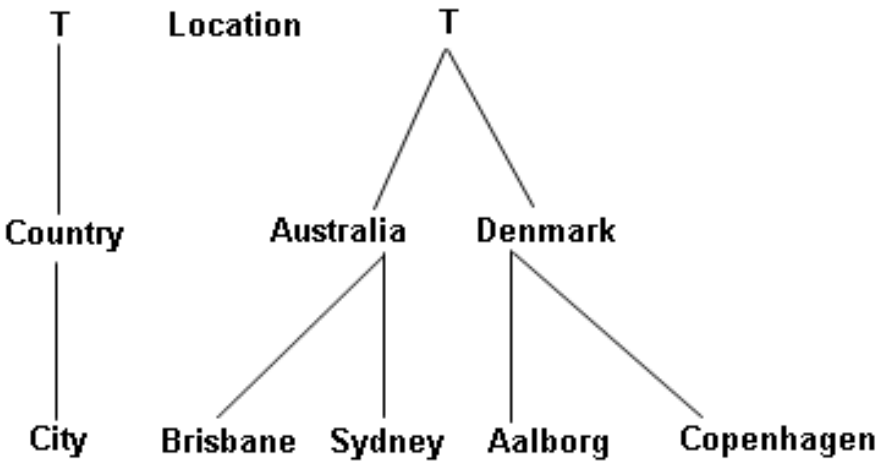


- Popis jak agregovat hodnoty jedné dimenze
- Může existovat několik nezávislých hierarchií na jedné dimenzi
- Drill-down podle dimenzí

- Dimenze času
- Nejmenší granularita – den
- 6 nezávislých dimenzí

Hierarchie

- Schéma a instance dimenze lokace



- Použití srozumitelných dat, texty
- Často odvozeno z jiných zdrojů (i externích)
- Redundance dat je pouze v dimenzích (nikoliv ve faktových tabulkách)
- Umožňuje vybírat a agregovat data po úrovních
- Hierarchie by měli mít konstatní hloubku
  - (nedoplňovat regiony jenom někde)
- Hierarchie jsou obsaženy v metadatech o dimenzích

Určení granularity

- Každý řádek faktové tabulky odpovídá hodnotám průniku všech dimenzí
- Všechny řádky musí mít stejnou granularitu
- Řádky s hodnotou nula se nezapisují
- Pokud zdroje neobsahují dostatečně detailní data, provádí se realokace dat na několik řádek tak, aby výsledky odpovídaly zdrojovým datům

- Granularita malá
  - Jeden řádek ≈ jedno měření
  - Velký objem dat
- Granularita velká (pouze sumy za měsíce, regiony, …)
  - Malé databáze
  - Omezená možnost analýz

Technologie

- Plnění schématu
  - Dávkové plnění pomocí ETL nebo ELT procesů z primárních systémů

- Požadavky na persistence dat
  - Podpora stár schématu
  - Podpora historizace dimenzí
  - Rychlý výpočet star-like dotazu
  - Agregace přes dimenze a zejména přes hierarchie dimenzí

Multidimenzionální databáze

- Nutné rozlišit
  - Princip
    - Práce s dimenzemi
    - Práce s hierarchiemi
  - Skutečný způsob uložení dat
    - Relační model
    - Speciální úložiště se speciálními indexy

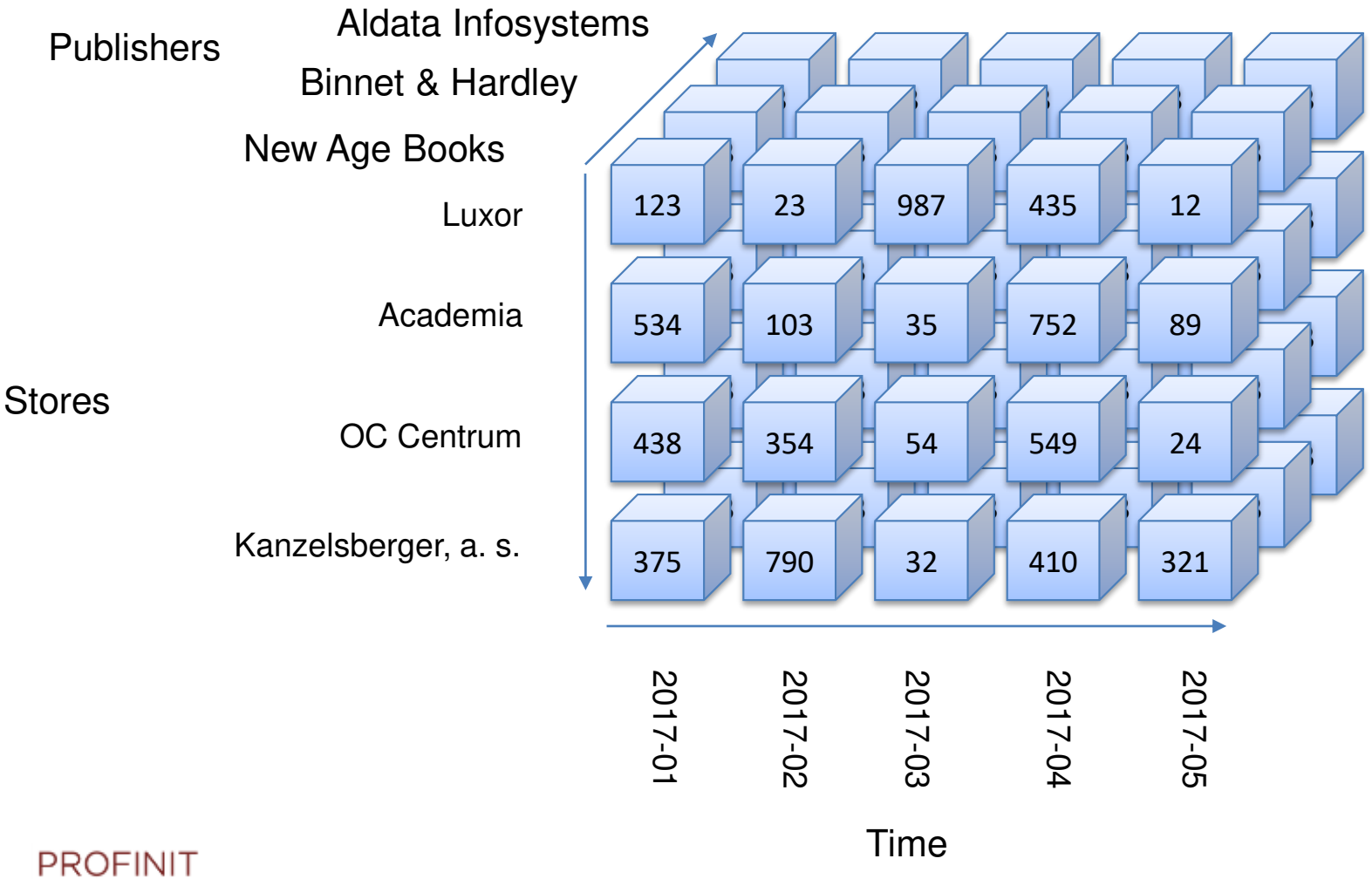
- Kategorizace dle místa uložení dat a agregací
  - MOLAP– veškerá data uložená v multidimenzionální databázi
  - ROLAP – veškerá data uložená v relační
  - HOLAP – hybrid

- Další typy
  - RTOLAP – real time, data pouze v RAM
  - DOLAP – desktop OLAP, data uložená na klientském počítači

OLAP technologie

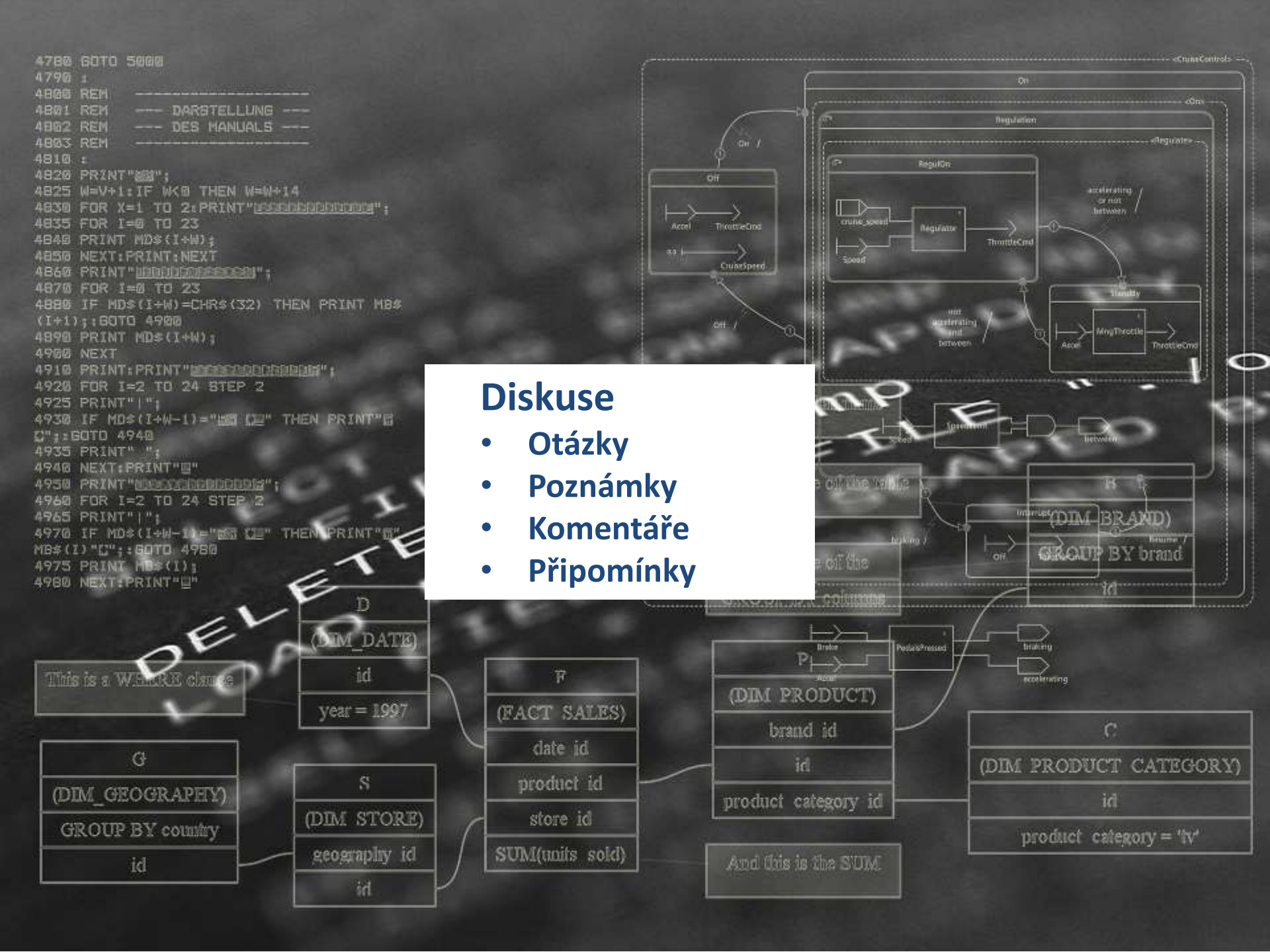
- Uložení a zpracování dat podporující určité druhy analýz
  - *parameterized static reporting*
  - *slicing and dicing with drill down*
  - *'what if?' analysis*
  - *goal seeking models*
- Způsob uložení předpočítaných hodnot (denormalizace)
  - Uložení agregovaných hodnot vyžadovaných analýzami podle zadaných
    - Metrik
    - Dimenzí
    - Hierarchií na dimenzích

Příklad uložení



Co si zapamatovat

- K čemu slouží dimenzionální datové modely
- Jaké jsou hlavní rozdíly relačního a dimenzionálního modelování
- Jaké jsou rozdíly mezi modely typu hvězda, souhvězdí, vločka nebo sněhová bouře
- Co to je Buss Matrix, k čemu slouží
- Jaké typy faktových tabulek se používají
- Co to je aditivní, semiaditivní a neaditivní metrika
- Jaké typy dimenzí se používají
- Co to je "Slowly changing dimension of type 2"
- Co to jsou Hierarchie a k čemu slouží
- Co to je OLAP databáze


4780 GOTO 5000
4790 :
4800 REM ----------
4801 REM --- DARSTELLUNG ---
4802 REM --- DES MANUALS ---
4803 REM ----------
4810 :
4820 PRINT"
4825 W=V+1:IF W<0 THEN W=W+14
4830 FOR X=1 TO 2:PRINT"
4835 FOR I=0 TO 23
4840 PRINT MD$(I+W);
4850 NEXT:PRINT:NEXT
4860 PRINT"
4870 FOR I=0 TO 23
4880 IF MD$(I+W)=CHR$(32) THEN PRINT MB$
(I+1);:GOTO 4900
4890 PRINT MD$(I+W);
4900 NEXT
4910 PRINT:PRINT"
4920 FOR I=2 TO 24 STEP 2
4925 PRINT"|";
4930 IF MD$(I+W-1)="
4935 PRINT" ";
4940 NEXT:PRINT"
4950 PRINT"
4960 FOR I=2 TO 24 STEP 2
4965 PRINT"|";
4970 IF MD$(I+W-1)="
MB$(I)"
4975 PRINT MB$(I);
4980 NEXT:PRINT"
Diskuse
Otázky
Poznámky
Komentáře
Připomínky
D
(DIM_DATE)
id
year = 1997
This is a WHERE clause
G
(DIM_GEOGRAPHY)
GROUP BY country
id
S
(DIM_STORE)
geography_id
id
F
(FACT_SALES)
date_id
product_id
store_id
SUM(units_sold)
P
(DIM_PRODUCT)
brand_id
id
product_category_id
And this is the SUM
C
(DIM_PRODUCT_CATEGORY)
id
product_category = 'tv'
(DIM_BRAND)
GROUP BY brand
id