داده کاوی

فصول چهارم و پنجم

تهیه شده توسط
محمد پورزعفرانی
Pourzaferani.ir
دانشگاه اصفهان

منابع

- اسلایدهای زیر بر گرفته از منابع زیر است
  - کتاب: مفاهیم پایگاه داده سیلبرشاتس (نسخه پنجم، فصل ۱۸)
  - کتاب: مفاهیم داده کاوی هان (نسخه دوم، فصل ۴)
  - اسلایدهای درس سمیه علیزاده از دانشگاه خواجه نصیر

دانشگاه اصفهان

فصل چهارم: انباره داده‌ها

- مفاهیم اساسی انباره داده‌ها
- روش های مدلسازی انباره داده‌ها: Data cube و OLAP
- طراحی انباره داده‌ها و استفاده از آن
- پیاده سازی انباره داده‌ها
- عمومیت بخشی به داده‌ها با استفاده از استنتاج مبتنی بر ویژگی
- خلاصه فصل

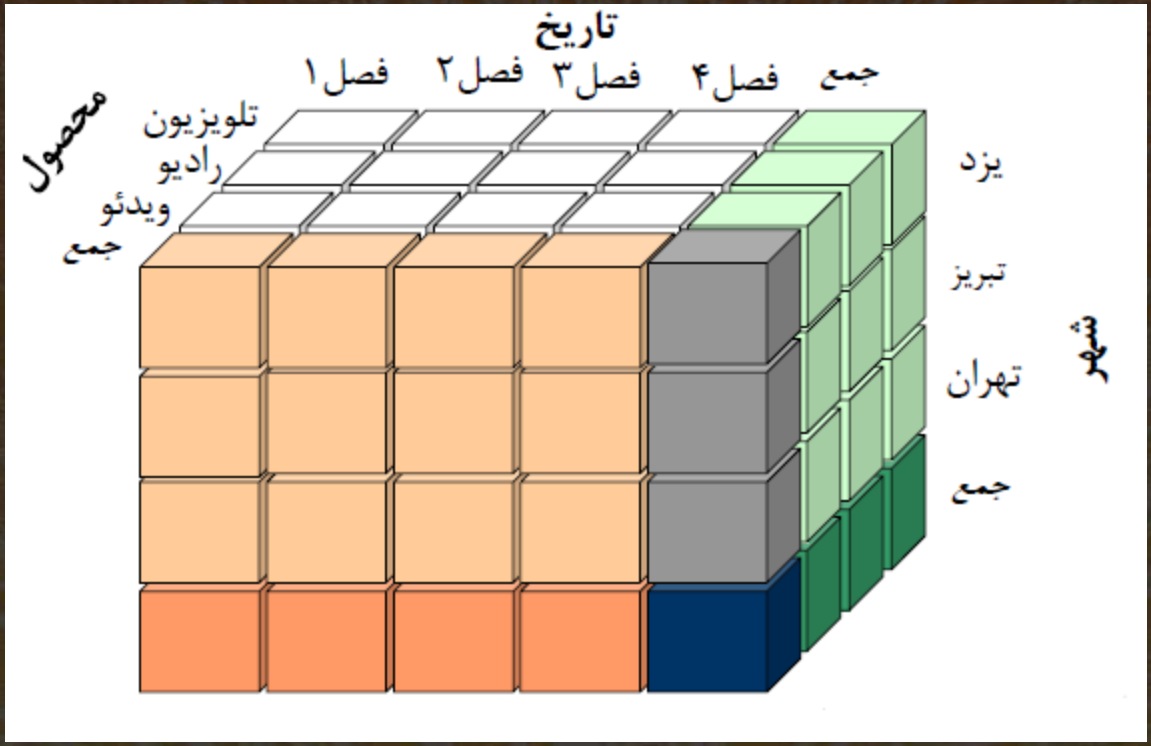
انباره داده‌ها چیست؟

- تعاریف مختلفی برای آن ارائه شده است
  - یک پایگاه داده تصمیم‌یار $(decision\ support\ database)$که به طور مجزا از پایگاه داده عملیاتی سازمان نگهداری می‌شود
  - راهکاری است برای کمک به پردازش اطلاعات از طریق فراهم کردن پلتفورمی برای تجمیع و آنالیز داده‌های پیشین
- تعریف دقیق‌تر: یک انبار داده تجمیعی است از داده‌ها برای فرآیند تصمیم‌سازی که دارای چهار مشخصه زیر باشند
  1) مبتنی بر موضوع $(Subject - oriented)$
  2) تجمیع شده $(integrated)$
  3) وابسته به زمان $(time - variant)$
  4) غیرفرار و دایمی بودن $(nonvolatile)$

مبتنی بر موضوع (Subject-Oriented)

- بر اساس موضوعات مطرح در پروژه سازمان‌دهی می‌شوند
  - به طور مثال : مشتریان، محصولات، فروش‌ها
- به جای تمرکز بر روی عملیات انجام شده در سیستم، بر روی مدلسازی و تحلیل داده‌ها متمرکز است
- یک دید ساده و صحیح نسبت به یک موضوع را با کنار گذاشتن داده‌های نامفید، فراهم می‌کند

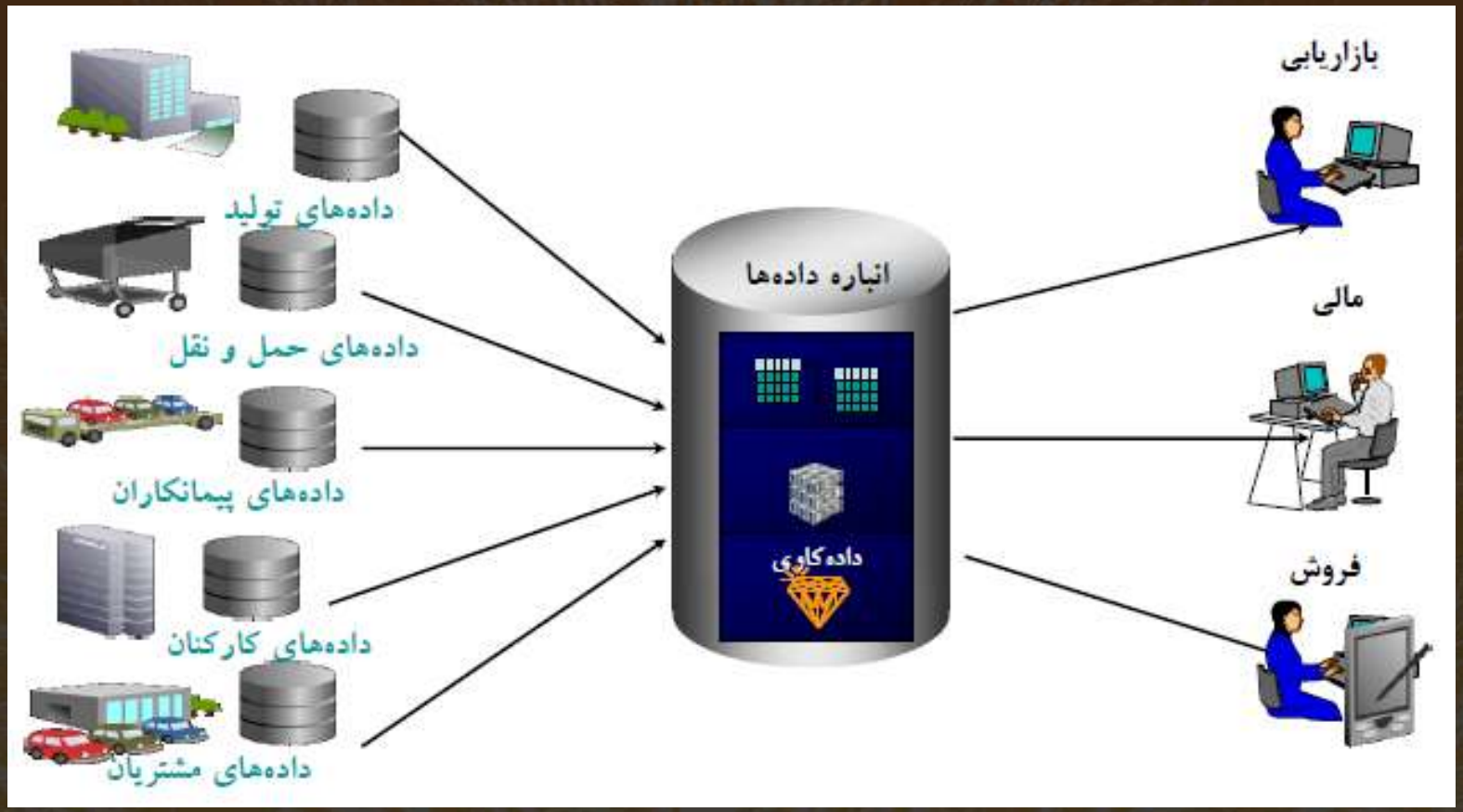
مثال


تاریخ
فصل ۱
فصل ۲
فصل ۳
فصل ۴
جمع
محصول
تلویزیون
رادیو
ویدئو
جمع
یزد
تبریز
تهران
جمع
شهر

تجمیع شده

- انباره داده از منابع متعدد و نامتجانس تجمیع شده است
  - پایگاه داده‌های رابطه‌ای، فایلهای متنی و بدون ساختار و ...
- برای این منظور باید تکنیک‌های پاکسازی داده‌ها و تجمیع داده‌ها اعمال شوند
  - به طور مثال: در مورد روش‌های نام‌گذاری یکسان، معیارهای اندازه‌گیری ویژگی‌ها تصمیم‌گیری شود
- زمانی که داده‌ها به انباره داده‌ها منتقل می شوند به طور طبیعی تبدیل خواهیم داشت

تجمیع داده‌ها



وابستگی به زمان (Time Variant)

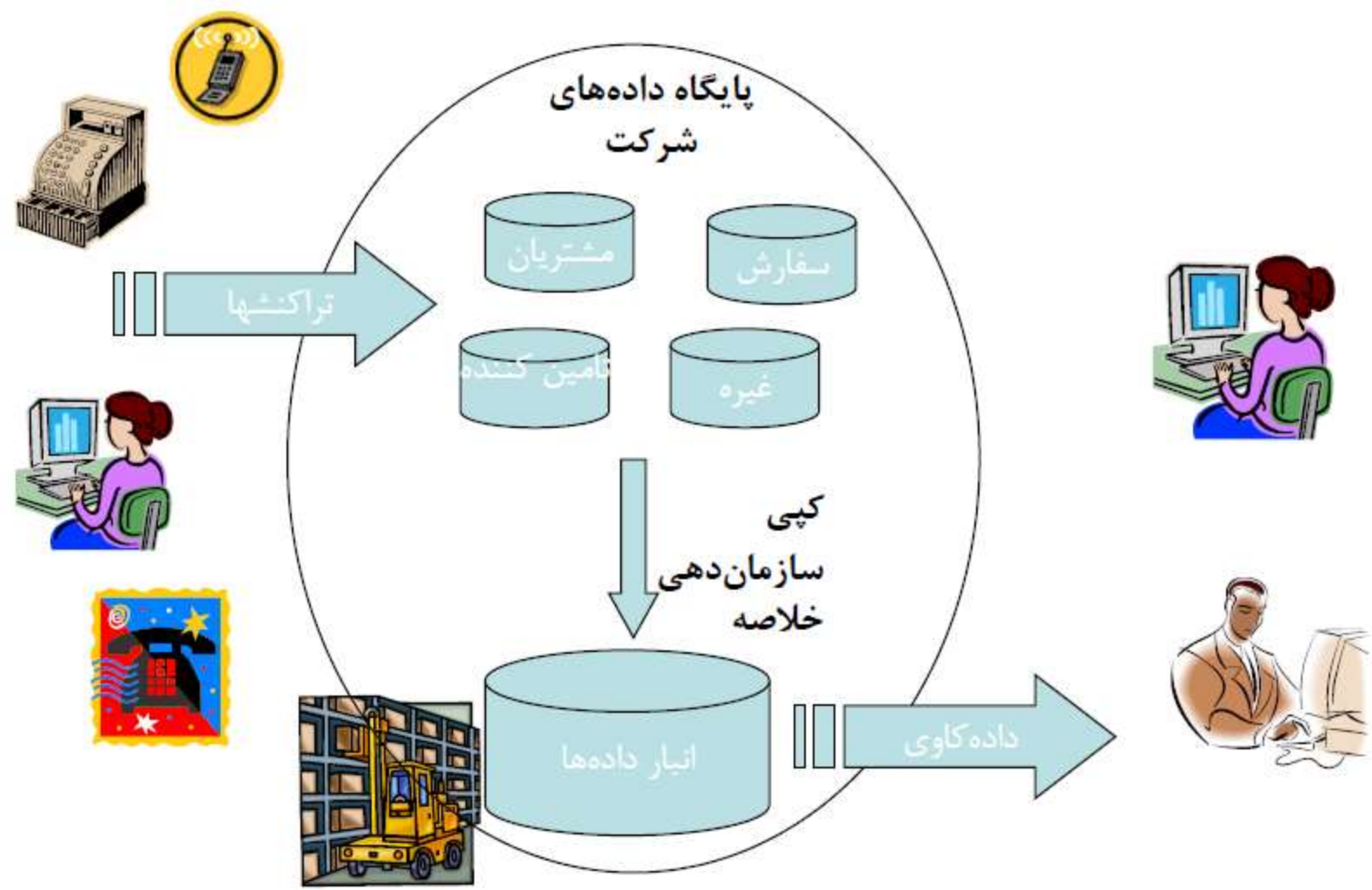
- بازه زمانی انباره داده خیلی وسیع‌تر از سیستم‌های عملیاتی است
  - پایگاه داده‌های عملیاتی: در مورد داده‌های حال حاضر صحبت می‌کنند
  - داده‌های انبار داده: حاوی اطلاعاتی در رابطه با قدمت یک سیستم (به طور مثال ۵-۱۰ سال)

- بنابراین هر داده در انباره داده‌ها چه به صورت صریح، چه ضمنی بعدی از زمان را داراست
- اما در پایگاه داده‌های معمولی زمان به عنوان یک عنصر کلیدی نیست

غیرفرار بودن

- انباره داده یک مکان برای ذخیره‌سازی داده‌ها است که به طور فیزیکی از محیط عملیاتی جداگانه است
- تغییرات (update) انجام گرفته در محیط عملیاتی در انباره داده‌ها اعمال نمی‌شوند بنابراین:
  - نیازی به پردازش تراکنش ها ندارد
  - نیازی به راهکارهای ریکاوری و کنترل همزمانی تراکنشها ندارد تنها به دو عملیات مهم نیازمند است
    - بارگزاری اولیه داده‌ها (*initial loading of data*)
    - روشی برای دسترسی به داده‌ها (*access of data*)

طرح کلی


پایگاه داده‌های
شرکت
مشتریان
سفارش
تامین کننده
غیره
تراکنشها
کپی
سازمان‌دهی
خلاصه
انبار داده‌ها
داده‌کاوی

پایگاه داده عملیاتی (Operational Database)

- در سیستم‌های مبتنی بر پایگاه داده‌های عملیاتی عمده وظایف عبارتند از
  - اجرای تراکنش‌ها به صورت آنلاین
  - پردازش پرس‌وجوها
- به این سیستم‌ها سیستم‌های پردازش تراکنش آنلاین می‌گویند
- OLTP = OnLine Transaction Processing
- تقریباً تمام عملیات روزانه سازمان‌ها مانند خرید، تولید، محاسبات، ثبت، پرداخت را پوشش می‌دهند

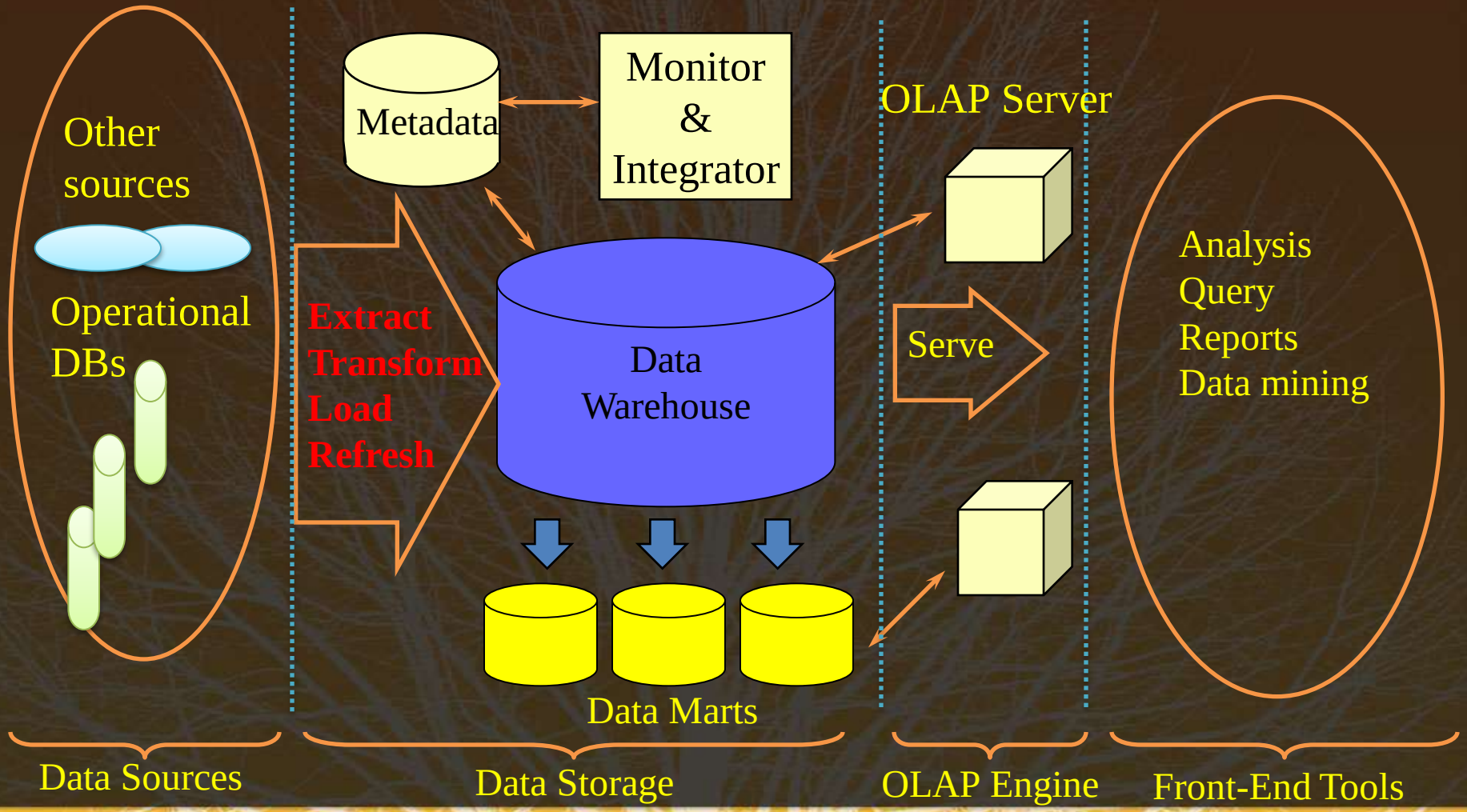
انباره داده‌ها

- در مقابل سیستم‌های مبتنی بر انباره داده تحلیل اطلاعات را ارائه می‌دهند و زمینه را برای تصمیم‌گیری مهیا می‌سازند
- به این سیستم ها، سیستمهای پردازش تحلیلی آنلاین می‌گویند
- OLAP = OnLine Analytical Processing
- در ادامه تحلیلی دقیق‌تر از تفاوت بین این سیستم‌ها ارائه خواهیم نمود

OLTP v.s OLAP

ویژگی	OLTP	OLAP
خصوصیات	پردازش اطلاعات	پردازش اطلاعات
جهت گیری	تراکنش ها	تحلیل
کاربران	فروشندگان،DBA،کاربران پایگاه داده‌ها	عوامل اطلاعاتی (برای مثال: مدیران، تحلیل گران)
وظیفه	عملیات روزانه	مستلزم اطلاعات بلند مدت و حمایت از تصمیم گیری
طراحی DB	در جهت ER، عملی	موضوعی، /starبرف دانه
داده	جاری، اصلاح و بروز شده	تاریخی، با وقت زیاد طی زمان حفظ شده است
خلاصه سازی	ابتدایی، با جزئیات بالا	اطلاعات خلاصه شده، یکی شده
دیدگاه	با جزئیات، ارتباطات همسطح	اطلاعات خلاصه شده، چند بعدی
واحد کاری	کوتاه، تراکنش ساده	پرس و جوی پیچیده
دسترسی	خواندن / نوشتن	عموماً خواندنی
تمرکز	به اطلاعات موجود	استخراج اطلاعات تاریخی
عملیات	شاخص گذاری / کار روی عملیات ابتدایی	بررسی‌های زیاد
تعداد موارد ثبت شده	ده‌ها	میلیون‌ها
تعداد کاربران	هزاران	صدها
اندازه DB	۱۰۰ MGتا GB	۱۰۰GB تا TB
اولویت	عملکرد بالا، دسترسی زیاد	انعطاف زیاد، استقلال کاربر نهایی
معیار	بازده تراکنش‌ها	زمان پاسخ، بازده پرس و جوها

معماری سه لایه‌ای انباره داده‌ها



معماری سه لایه‌ای از نگاهی دیگر

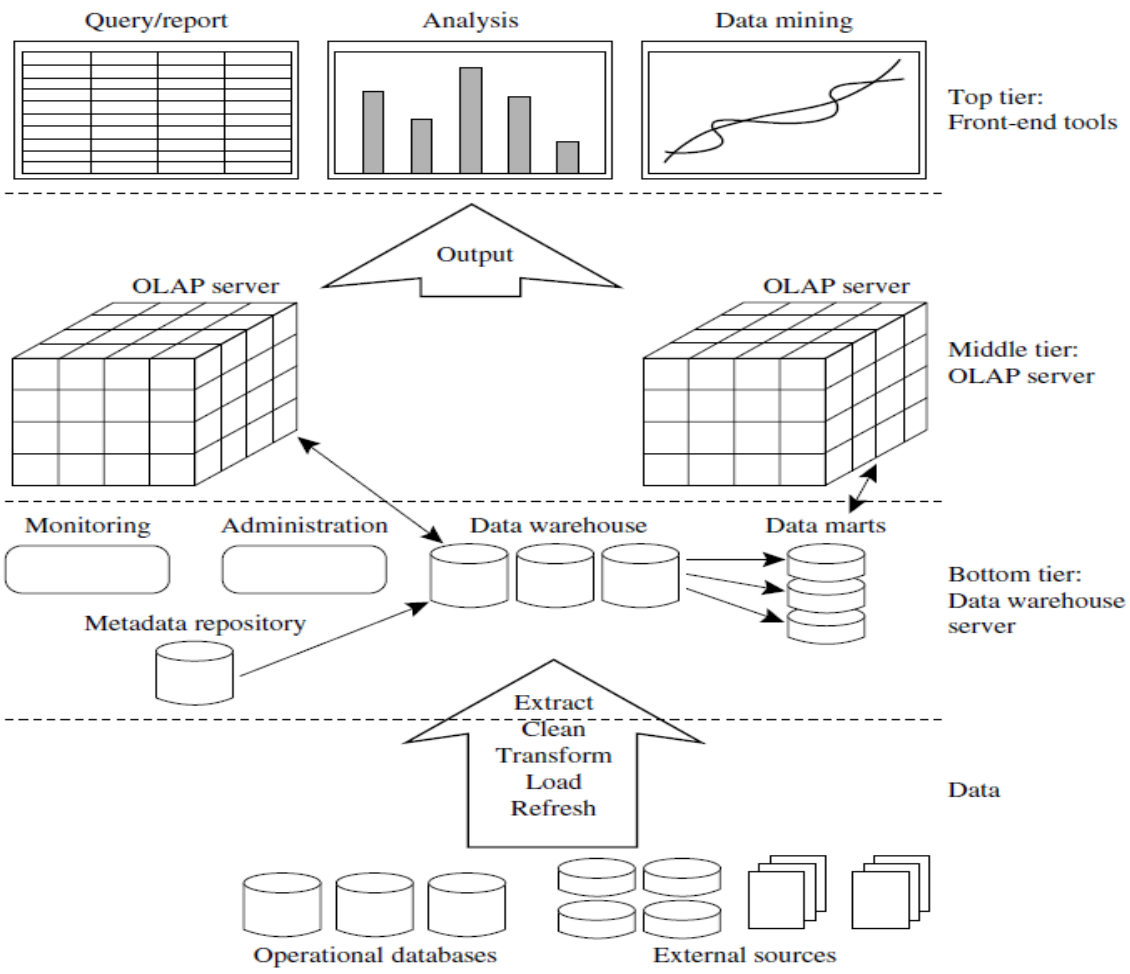


Figure 4.1 A three-tier data warehousing architecture.

لزوم استفاده از انباره داده‌ها در داده کاوی

- دلیل اول:
- انباره داده‌ها، داده‌های خام را برای تحلیل آماده می‌سازد که این آماده‌سازی به شکل‌های مختلفی می‌تواند مفید باشد
- یکی از ویژگی انباره داده‌ها، یکپارچگی داده‌ها هنگام قرارگیری در انباره داده هاست. در واقع دقت بسیاری بکار گرفته می‌شود تا یکنواختی و پیوستگی در درک اهداف کلی سازمان به وجود بیاید
- اگر انباره داده‌ها نباشد، داده کاو بایستی زمان زیادی برای جمع آوری، پاکسازی و یکپارچه‌سازی داده‌ها صرف کند

لزوم استفاده از انباره داده‌ها در داده کاوی

- دلیل دوم:
- یکی از موفقیت‌های داده کاوی به خاطر وجود انباره داده‌ها است. در انباره داده‌ها، داده‌های تاریخی جمع‌آوری و سازماندهی می‌شوند.
- وجود داده‌های تاریخی برای یافتن الگوها و روابطی که سازمان بدنبال آنهاست برای داده کاوی ضرورت دارد
- اگر داده‌های تاریخی وجود نداشته باشد، داده کاو بایستی بدنبال جمع آوری این داده‌ها باشد.

لزوم استفاده از انباره داده‌ها در داده کاوی

- دلیل سوم:
- اهمیت انباره داده‌ها به این دلیل است که انباره داده‌ها شامل داده‌های جزئی و داده‌های کلی در کنار یکدیگر است
- برای تحلیل علاوه بر داده‌های جزئی نیاز به خلاصه شده این داده‌ها نیز داریم
- زمانیکه نسخه خلاصه شده داده‌ها وجود داشته باشد، داده کاو می‌تواند به سرعت وضعیت انباره داده‌ها را بررسی کند

مفاهیم OLAP

- برای روشن شدن موضوع با یک مثال مفاهیم را تشریح خواهیم نمود
- فرض کنید برنامه‌ای می‌خواهیم که می‌خواهیم با استفاده از آن تشخیص دهیم چه نوع لباسی نزد مشتریان محبوب‌تر است؟
- برای این منظور روابط زیر را داریم

- Clothes = {item-name, color, size}
- Sales = (item-name, color, size, number)
- Item-name = {skirt, dress, shirt, pant}
- Color = {dark, pastel, white}
- Size = {small, medium, large}

نمایی از میزان فروش‌ها

name	color	size	quantity
skirt	dark	small	2
skirt	dark	medium	5
skirt	dark	large	1
skirt	pastel	small	11
skirt	pastel	medium	9
skirt	pastel	large	15
skirt	white	small	2
skirt	white	medium	5
skirt	white	large	3
dress	dark	small	2
dress	dark	medium	6
dress	dark	large	12
dress	pastel	small	4
dress	pastel	medium	3
dress	pastel	large	3
dress	white	small	2
dress	white	medium	3
dress	white	large	0
shirt	dark	small	2
shirt	dark	medium	6
shirt	dark	large	6
shirt	pastel	small	4
shirt	pastel	medium	1
shirt	pastel	large	2
shirt	white	small	17
shirt	white	medium	1
shirt	white	large	10
pants	dark	small	14
pants	dark	medium	6
pants	dark	large	0
pants	pastel	small	1
pants	pastel	medium	0
pants	pastel	large	1
pants	white	small	3
pants	white	medium	0
pants	white	large	2

نیاز به خلاصه‌سازی

- هدف ما بدست آوردن اطلاعات کلی فروش است
- جزئیات پایگاه داده برای ما مهم نیست
- بنابراین باید خلاصه سازی کنیم
- راه‌کار:
  - رابطه‌ای که برای هدف ما مهم است را خلاصه‌سازی می‌کنیم
  - تعدادی از ویژگی‌ها را به عنوان معیار (measure attribute) تعیین می‌کنیم
  - این ویژگی‌ها برای اندازه‌گیری هدف ما هستند و با استفاده از آنها می‌توانیم عملیات Aggregation داشته باشیم
  - در مثال ما تعداد فروش measure است چون تعداد واحدهای به فروش رفته را اندازه گیری می‌کند
  - تعدادی (یا همه) ویژگی‌های باقیمانده به عنوان ویژگی‌های بعد (dimension attribute) در نظر گرفته می‌شوند
  - این ویژگی‌ها ابعاد ویژگی measure را مشخص می‌کنند که بر اساس آنها خلاصه‌سازی انجام گیرد

چند تعریف دیگر

- در مثال ما {item-name , color , size} می‌توانند dimension attribute باشند

- داده‌هایی که بتوانند با ویژگی‌های measure و dimension مدل شوند را داده‌های چند بعدی یا multidimensional data می‌گوییم

ادامه مثال

- در این جدول {item-name , color} به عنوان dimensional attributeانتخاب شده اند
- این جدول را با نام‌های زیر نیز می‌شناسند
  - Cross-tabulation (Cross-tab)
  - Pivot-table

clothes_size **all**

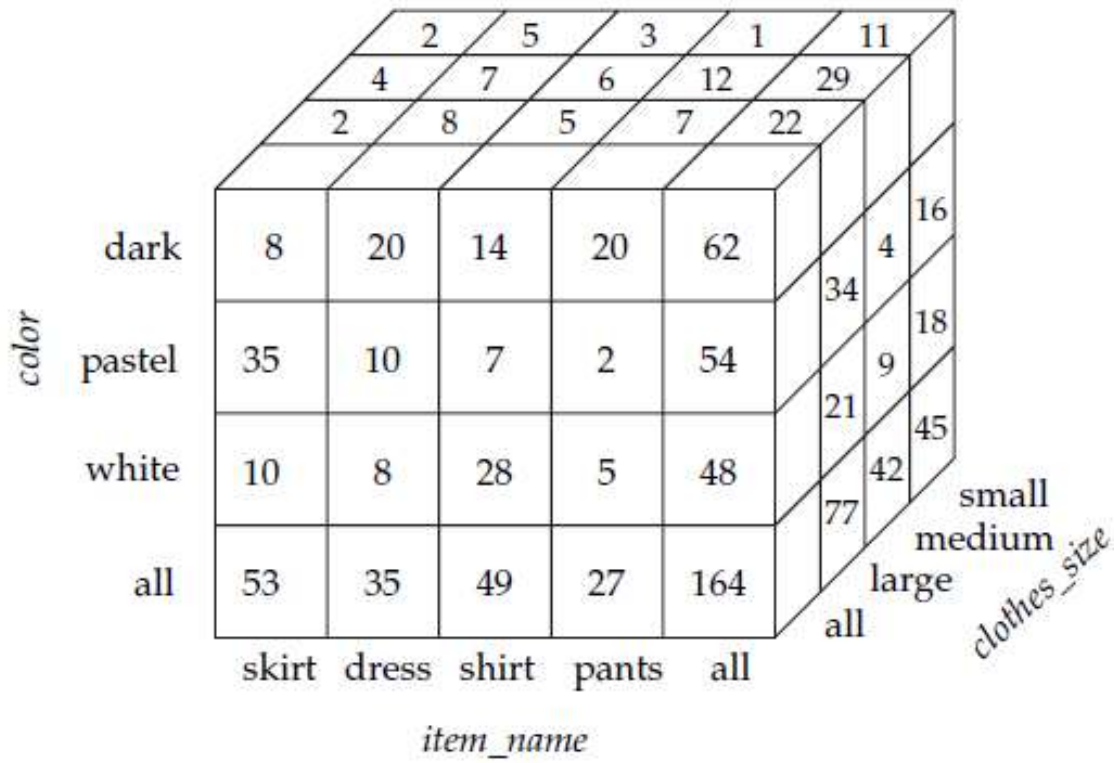
item_name	*color*: dark	pastel	white	total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pants	20	2	5	27
total	62	54	48	164

Cross tabulation of *sales* by *item_name* and *color*.

Cross-tab

- تفاوت Cross-tab با جداول رابطه‌ای چیست؟
  - تعداد ستون‌ها در Cross-tab بستگی به داده‌ها دارد
  - کاملا از دید کاربر اطلاعات واضح است
- صحبت از داده‌های چند بعدی بود اما Cross-tab دارای دو بعد است
- برای مصورسازی n-بعد از n-dimensional cube یا در اصطلاح data cube استفاده میکنیم

Data Cube

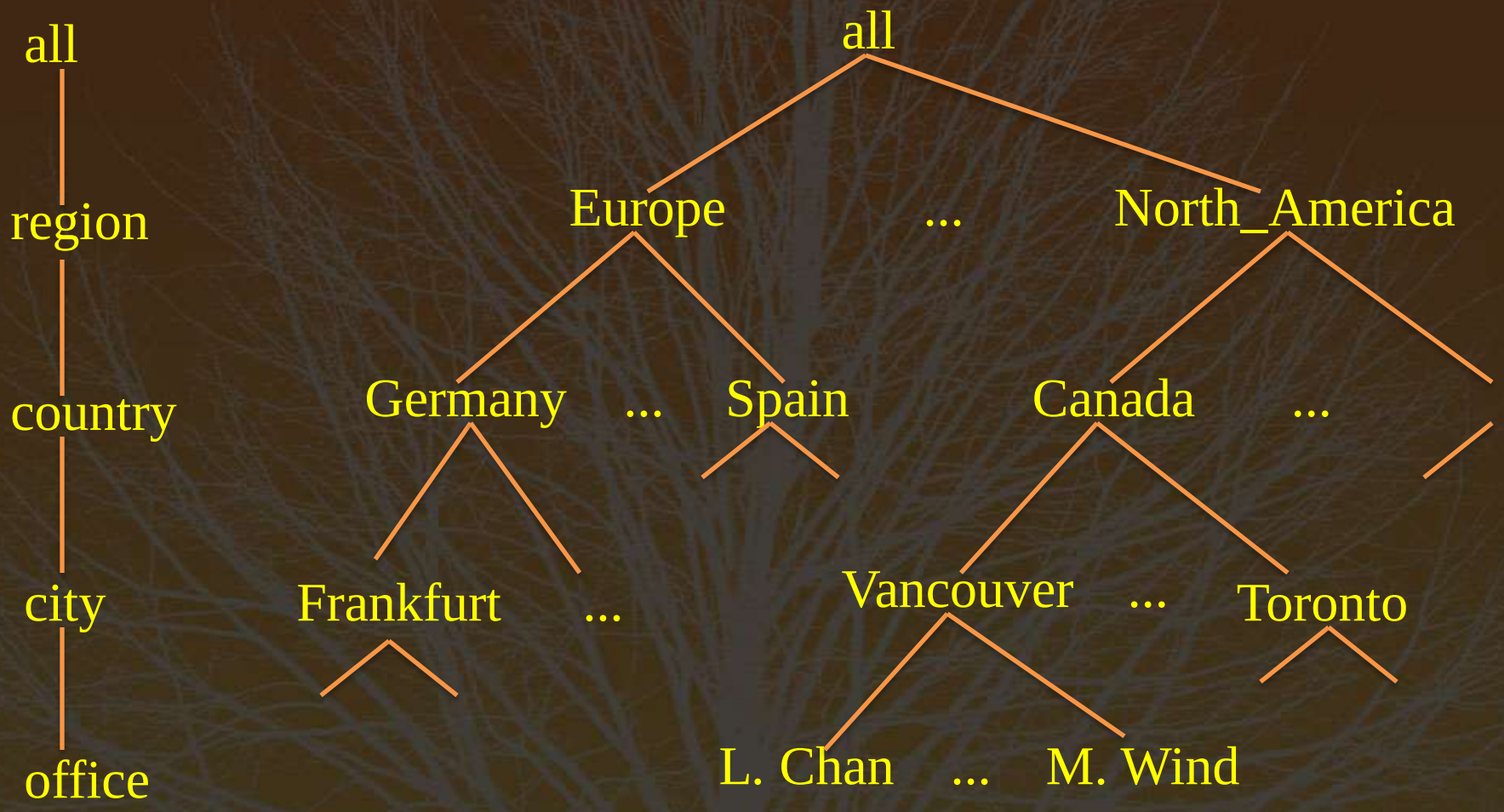


Three-dimensional data cube.

Data Cube

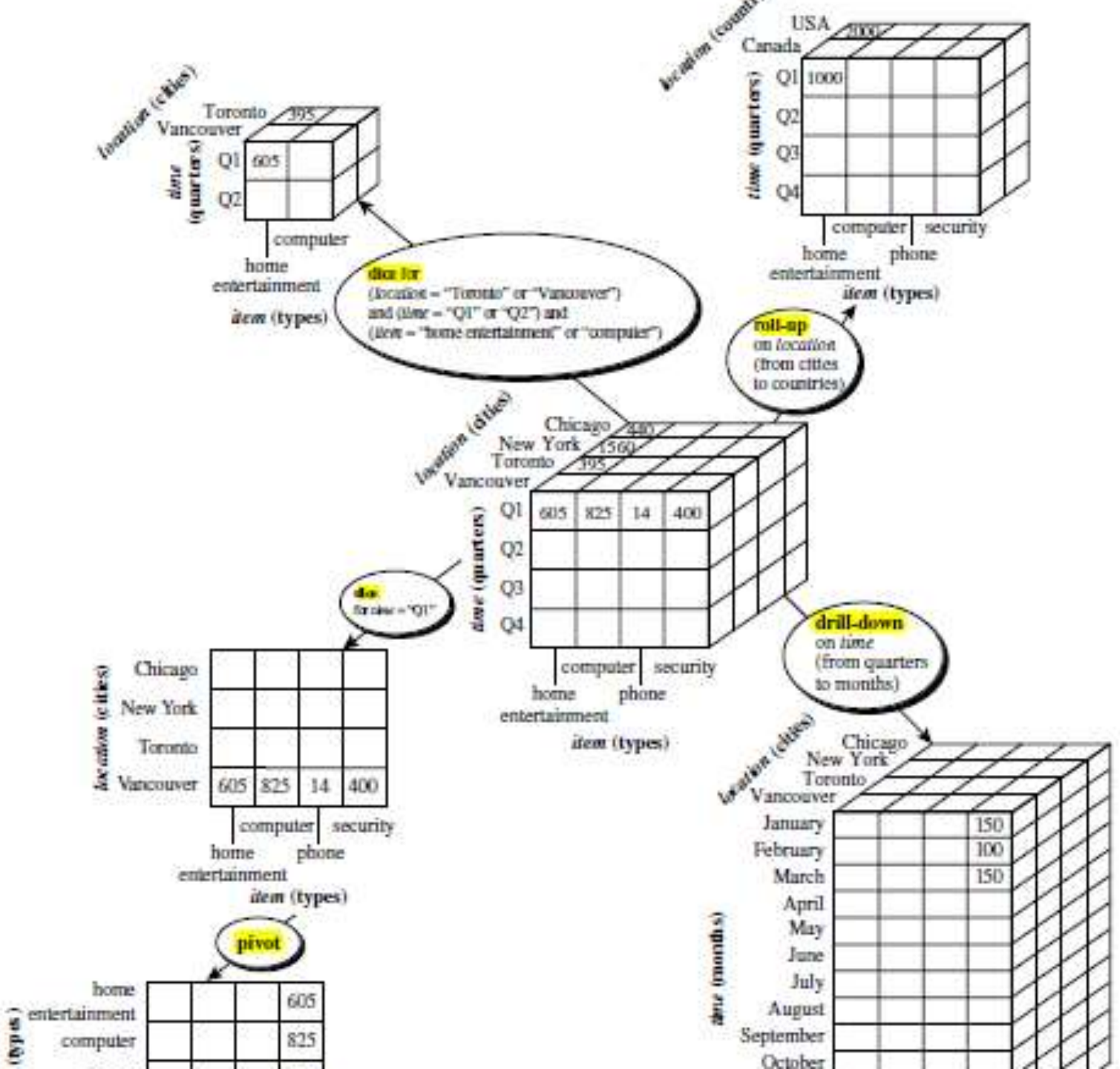
- در واقع هر سلول، یک Aggregation از ابعاد مختلف مکعب است
- حال اگر n بعد داشته باشیم زیر مجموعه خواهیم داشت که عملیات 2^n Aggregation روی آن‌ها انجام گیرد
- واژه اول OLAP برخط بودن سیستم را نشان می‌دهد
- بنابراین تحلیلگر ممکن است برای ویژگی‌های مورد نظر خود خلاصه‌سازی را روی ابعاد خاصی بخواهد

سلسله مراتب مفاهیم

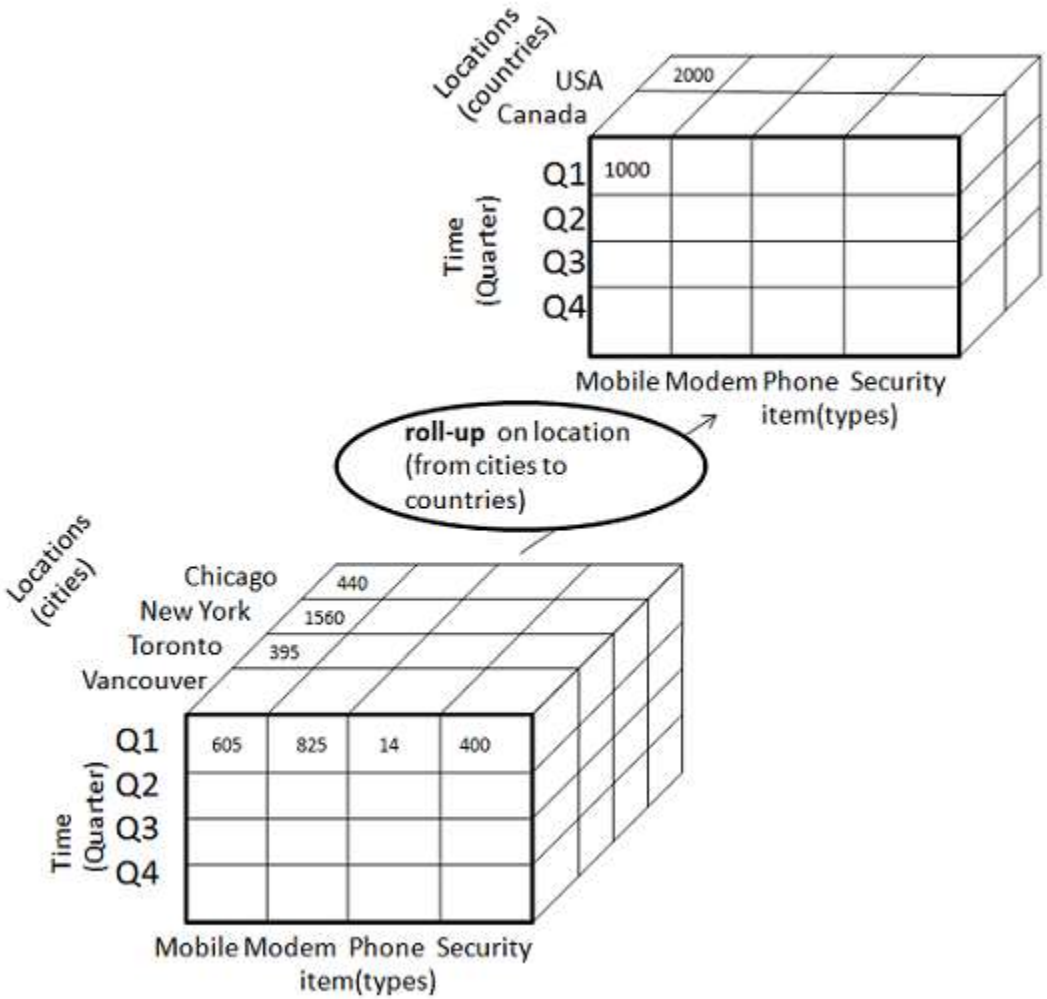


عملیات متداول در OLAP

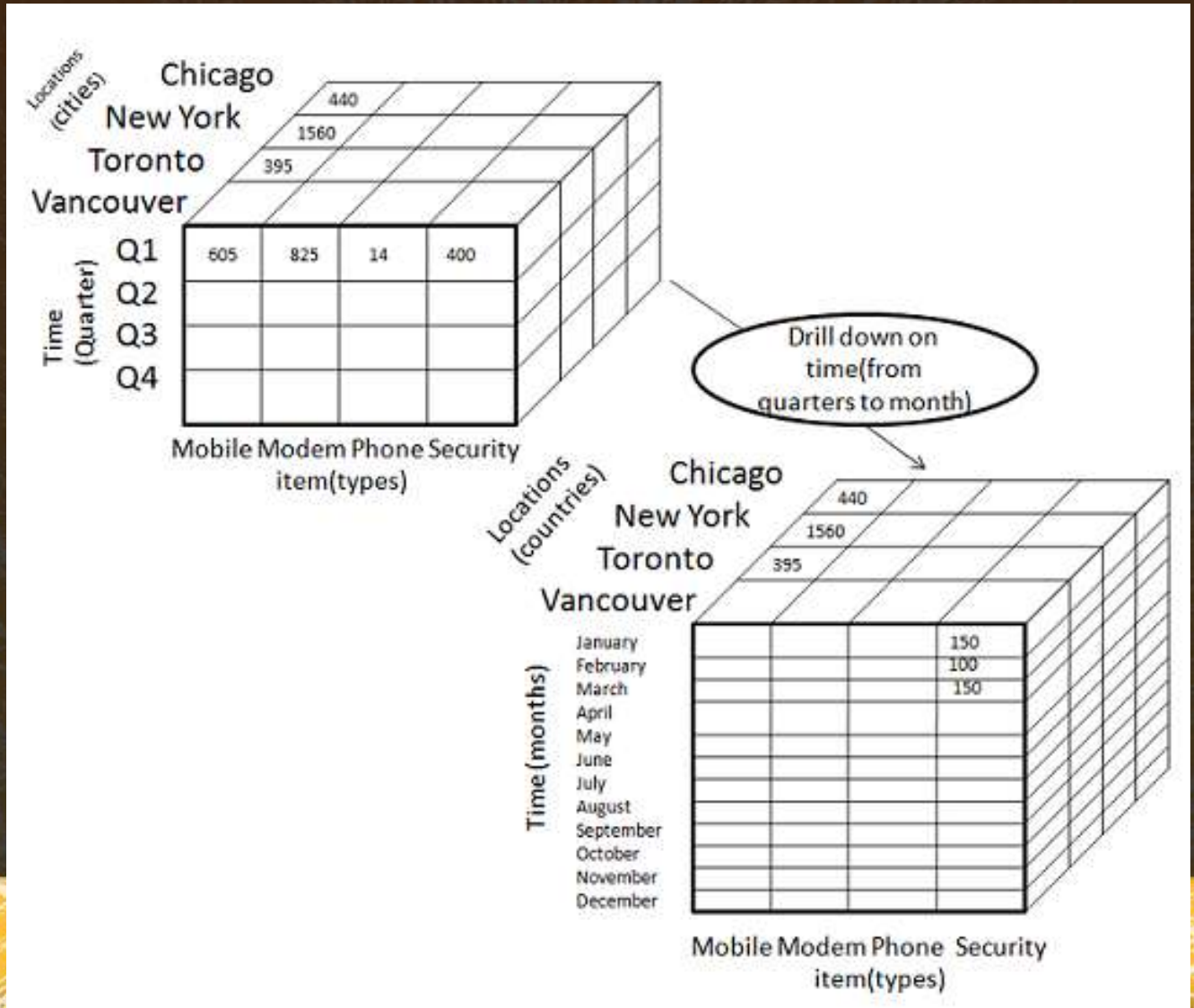
- **Roll up**: خلاصه‌سازی داده‌ها
  - با رفتن از مفاهیم جز به مفاهیم کلان
- **Drill down**: برعکس عمل بالا
  - از سطح بالای خلاصه‌سازی به سطح پایین حرکت کنیم
- **Slice**: این عمل یک Select بر روی یک بعد از مکعب انجام می‌دهد
- **Dice**: این عمل با انتخاب دو یا چند بعد یک زیر مکعب را مشخص می‌کند
- **Pivot**: باعث چرخش مکعب شده و داده‌ها را از منظر دیگری خواهیم دید


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
Q2
605
computer
home entertainment
item (types)
slice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
computer
security
home entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
time (quarters)
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
time (months)
January
February
March
April
May
June
July
August
September
October
150
100
150
pivot
(types)
home entertainment
computer
605
825

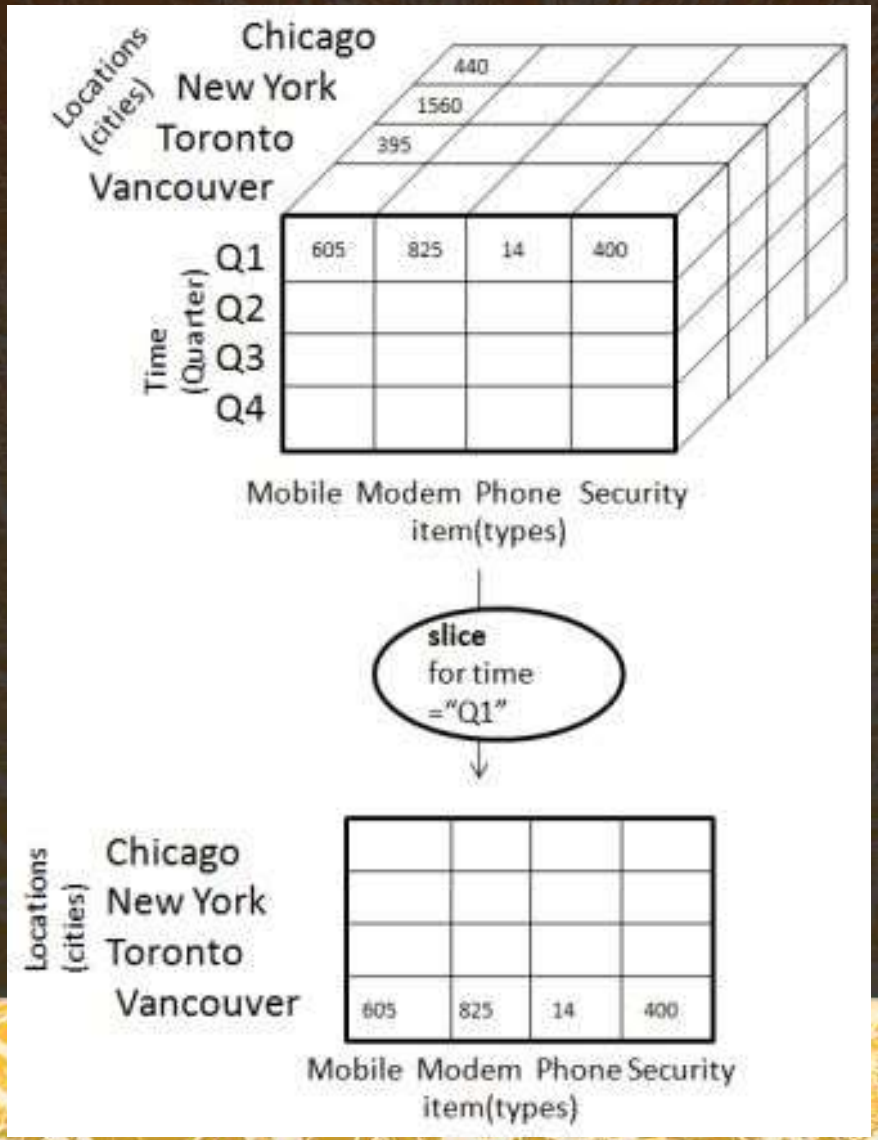
Roll-up


Locations (countries)
USA
Canada
2000
1000
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)

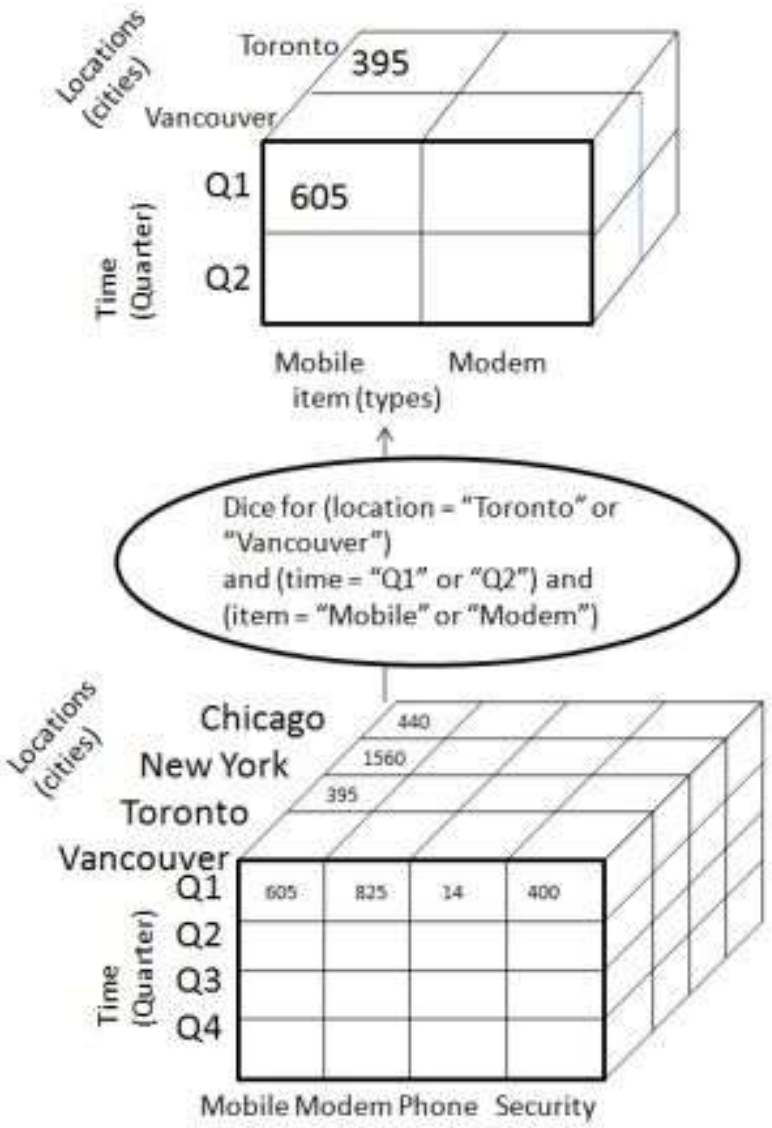
Drill-down


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile Modem Phone Security
item(types)
Drill down on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
Mobile Modem Phone Security
item(types)

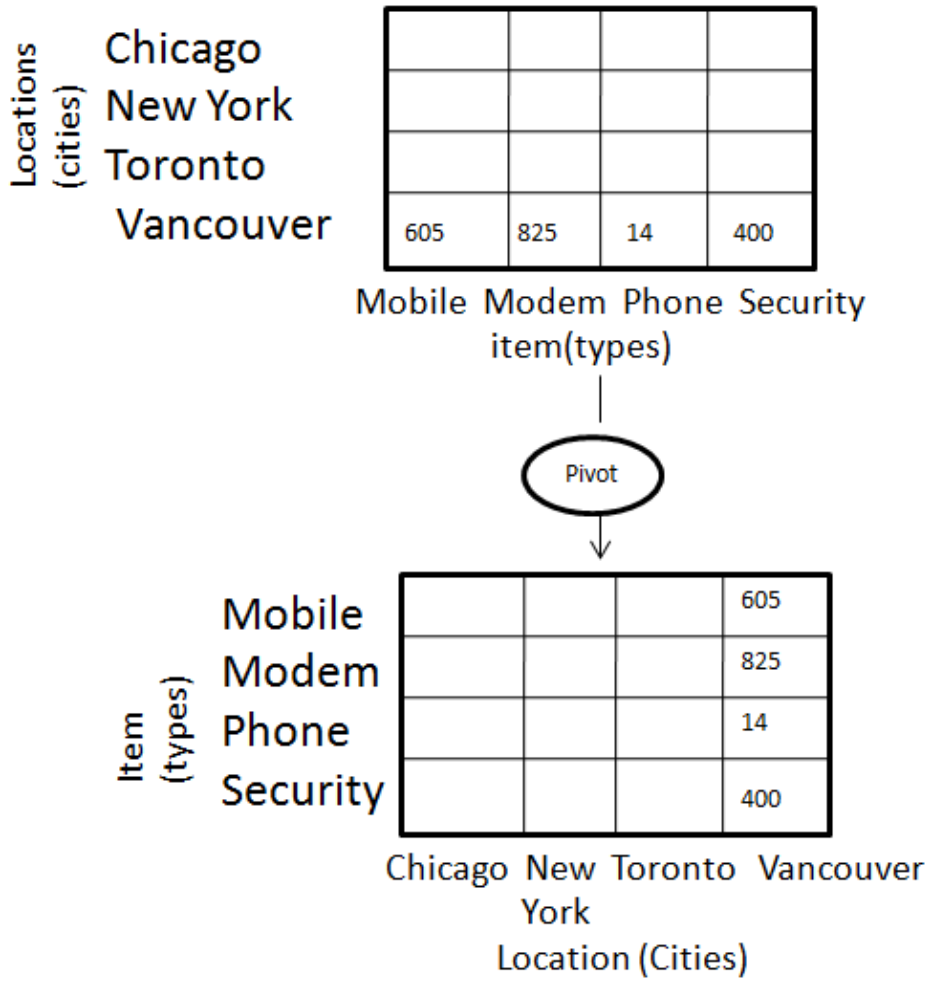
Slice


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Time (Quarter)
Q1
Q2
Q3
Q4
Mobile Modem Phone Security
item(types)
slice
for time
="Q1"
Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)

Dice


Locations (cities)
Toronto
395
Vancouver
Q1
605
Q2
Time (Quarter)
Mobile
Modem
item (types)
Dice for (location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "Mobile" or "Modem")
Locations (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
Q1
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile
Modem
Phone
Security

Pivot


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile
Modem
Phone
Security
item(types)
Pivot
Item (types)
Mobile
Modem
Phone
Security
605
825
14
400
Chicago
New York
Toronto
Vancouver
Location (Cities)

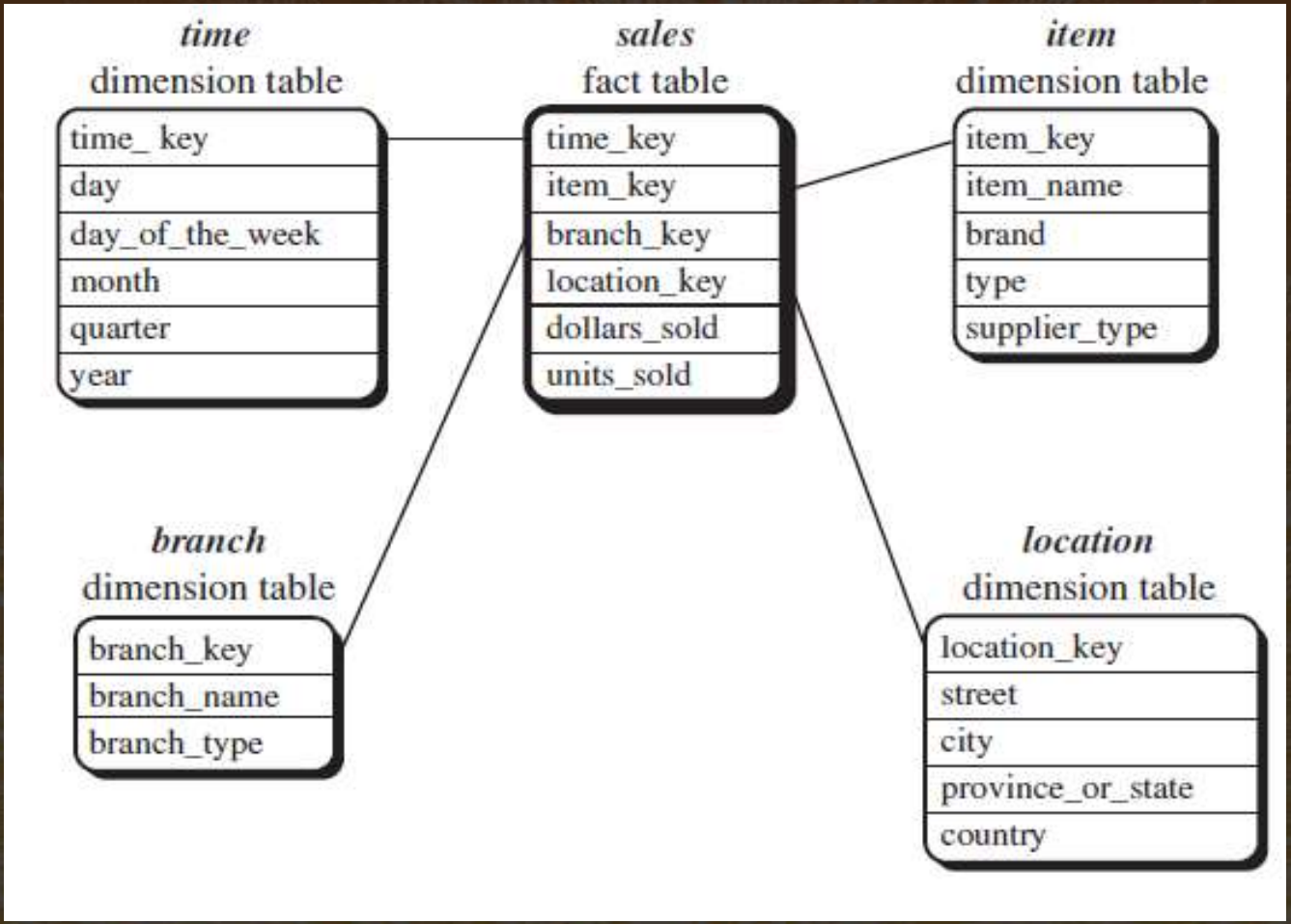
مدل‌سازی انباره داده‌ها

- سه شمای مفهومی پایه عبارتند از :
- شمای ستاره (Star schema)
- شمای دانه برفی (Snowflake schema)
- شمای صورت فلکی (Fact constellations)

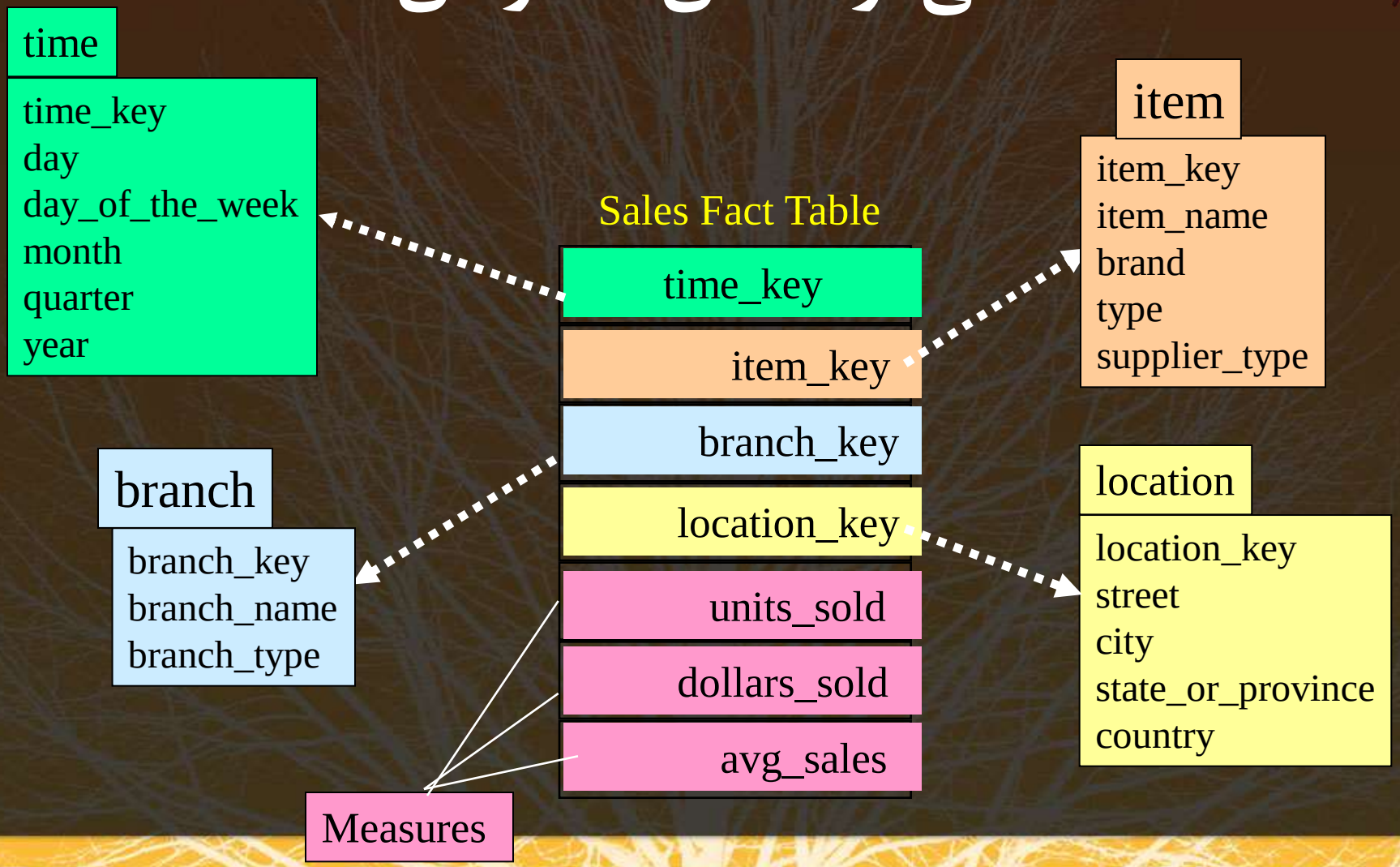
شمای ستاره‌ای

- متداول‌ترین الگوی مدلسازی است
- در این شما یک جدول Fact در مرکز قرار گرفته و تمام جداول مربوط به آن با نام جداول ابعاد (Dimension Tables) به آن متصل می‌شوند

مثالی از شمای ستاره‌ای



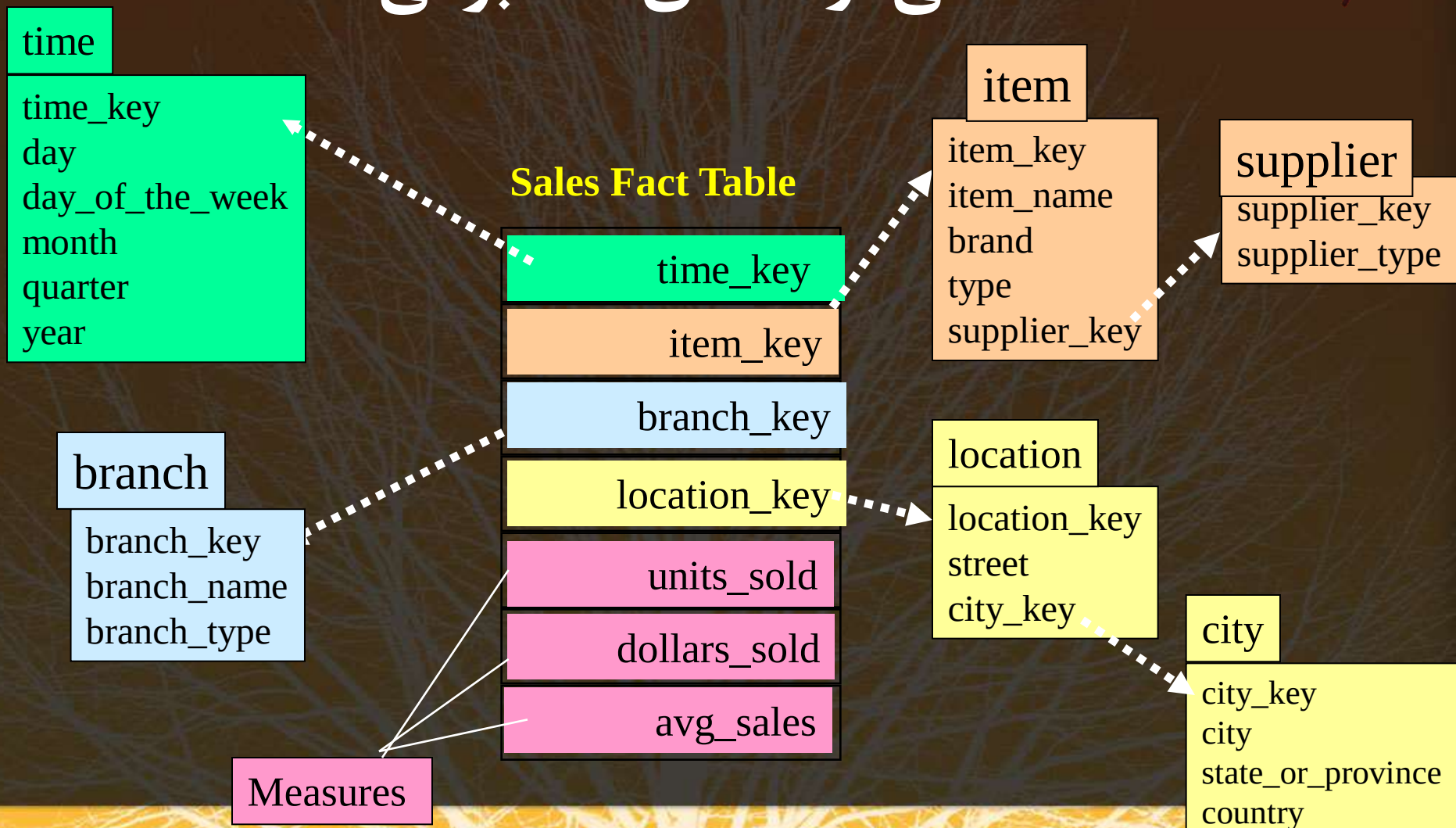
مثالی از شمای ستاره‌ای



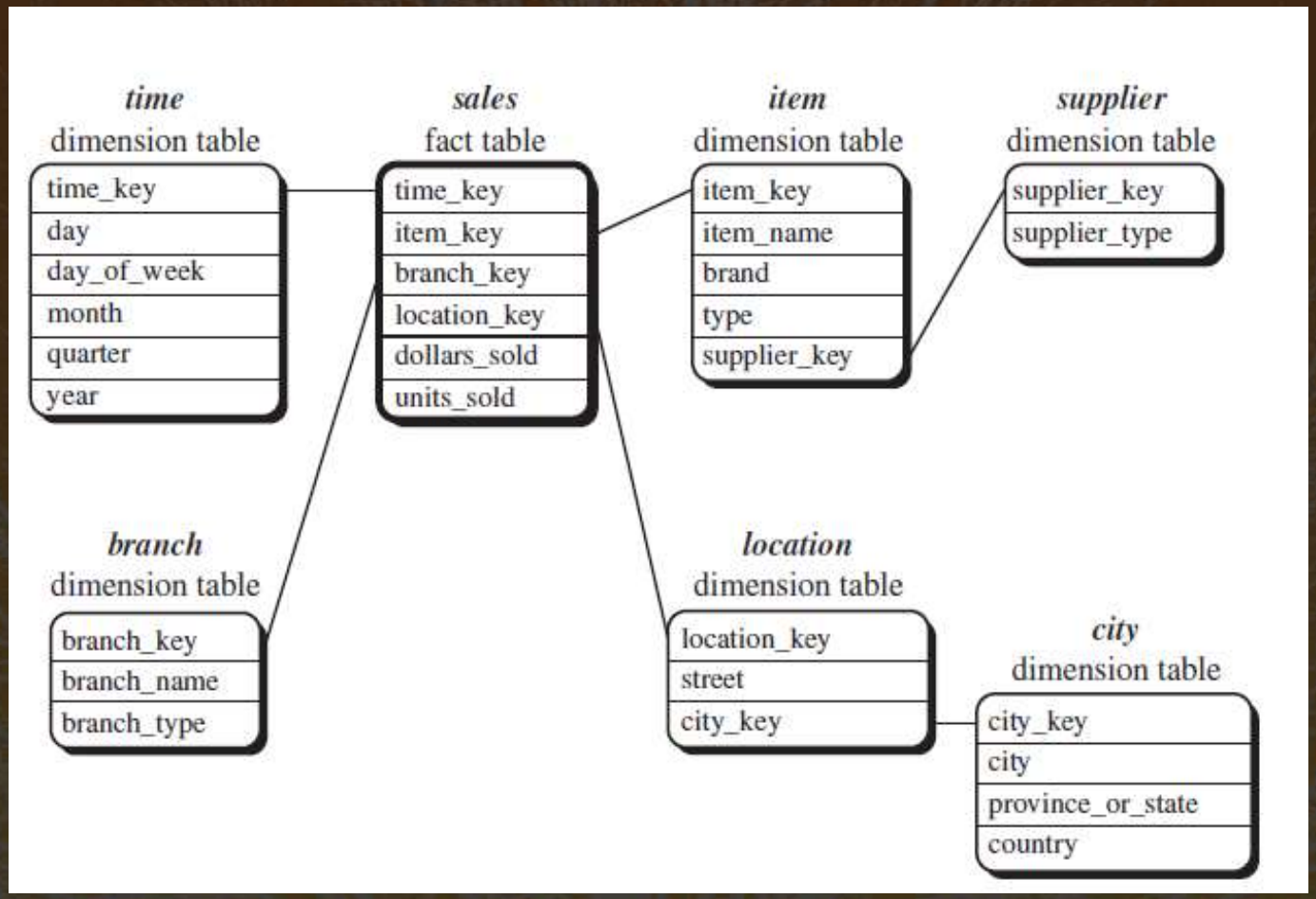
شمای دانه برفی (snowflake)

- این طرح پالایشی است بر روی شمای ستاره‌ای
- در این طرح‌واره جدول‌های بُعد نرمال‌سازی شده و سلسله مراتبی از ابعاد خواهیم داشت
- مزیت : کاهش فضای ذخیره‌سازی
- عیب : چون اتصالات بیشتری برای بررسی دقیق نیاز می‌باشد. متعاقباً، عملکرد سیستم ممکن است بطور عکس تحت تأثیر قرار گیرد.
- از اینرو اگرچه طرح برف دانه موارد اضافی را کاهش می‌دهد، به اندازه طرح ستاره در مخزن داده‌ها معروف نیست.

مثالی از شمای دانه‌برفی



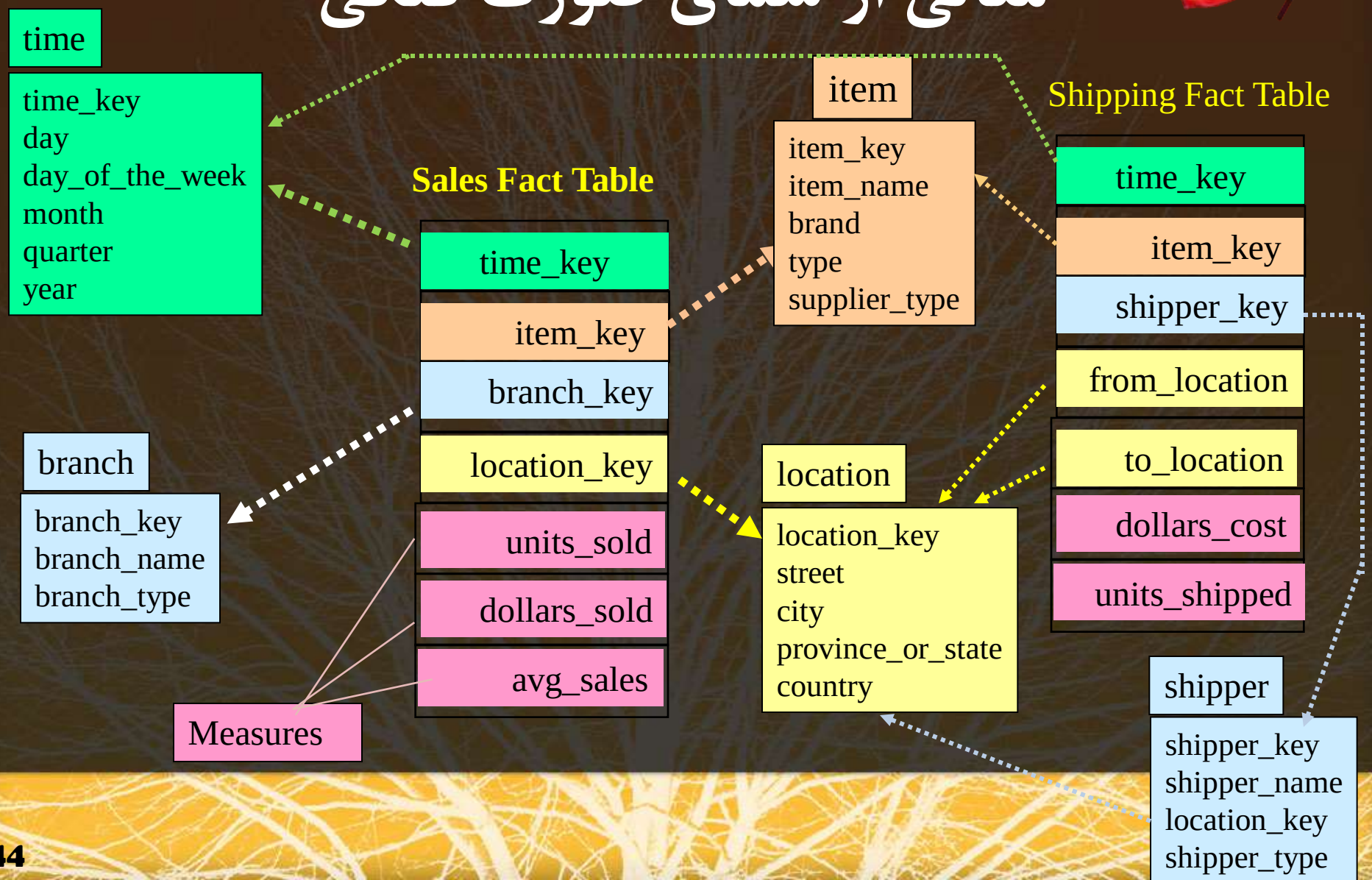
مثالی از شمای دانه‌برفی



شمای صورت فلکی

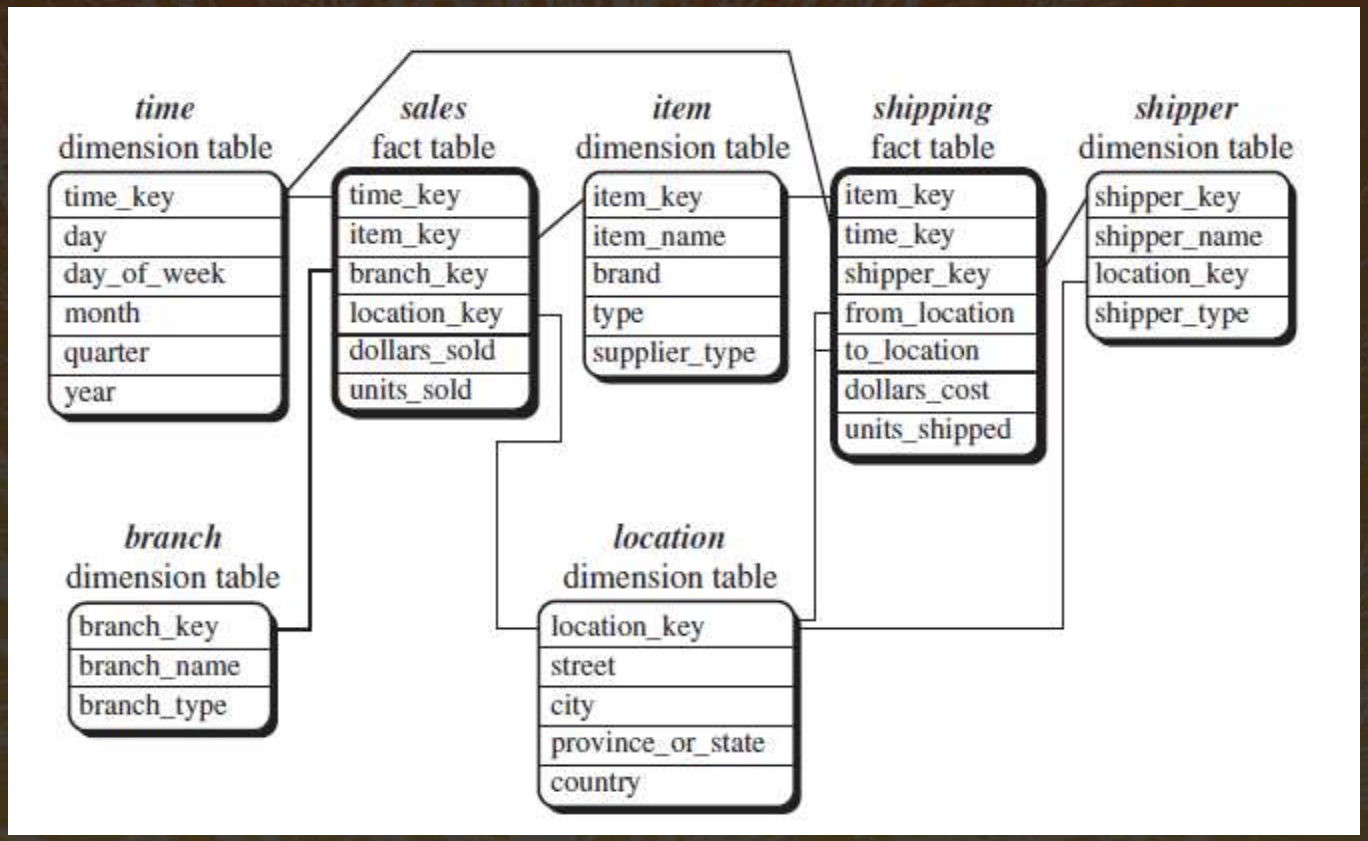
- کاربردهای پیچیده ممکن است به جداول متعدد برای جداول اندازه نیاز داشته باشد. این نوع طرح ممکن است بصورت مجموعه‌ای از طرح‌های ستاره‌ای بررسی شود و از این رو طرح کهکشان یا صورت فلکی خوانده می‌شود.

مثالی از شمای صورت فلکی



مثال شمای صورت فلکی

- این طرح دو جدول اصلی، فروش و حمل‌کالا را نشان می‌دهد



دانشگاه اصفهان

پایان فصل

خدایا مدد کن حسینی بمانیم