Моделиране на данните в склада от данни

Моделиране на данните в DW

- ER моделиране
- Димензионно (Многомерно) моделиране
  - Dimensions – Димензии (Измерения)
  - Facts - Факти
  - Measures (variables) – Мерки
  - Cube – куб с данни, представляващ многомерна матрица

Многомерният модел се изгражда от

- Димензионни таблици (наричат се и lookup tables)
- Факт-таблици
- Връзки между димензионните таблици и факт-таблиците

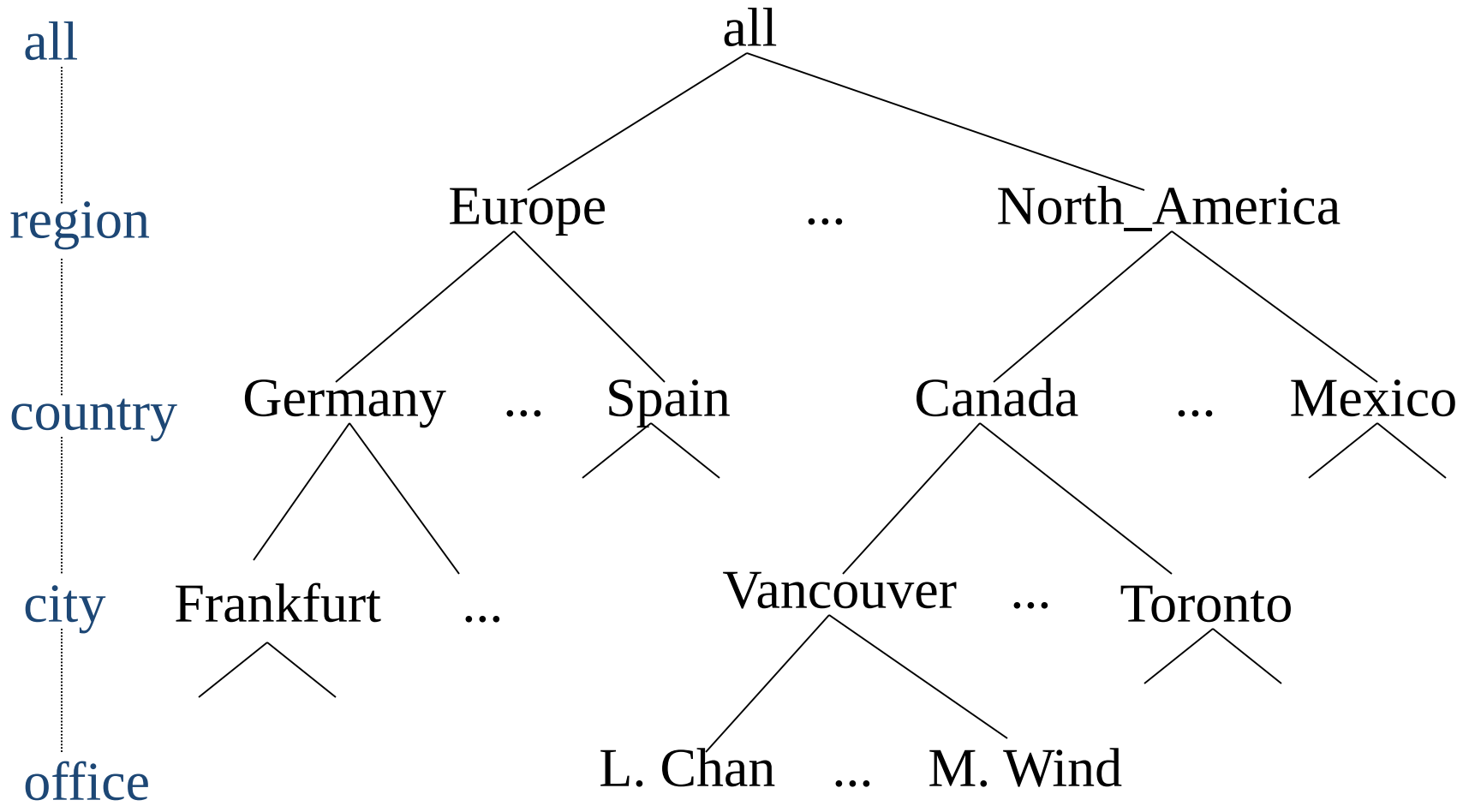
Димeнсии

- Димeнсиите са параметрите (показателите), по които се прави анализ на данните.
- Пример за димeнсии: групи стоки, стоки, клиенти, местоположение, подразделения на фирмата, времеви периоди, …
- Всяка димeнсия притежава собствени характеристики или атрибути, които се съхраняват в димeнсионните таблици.

Йерархия на димензиите

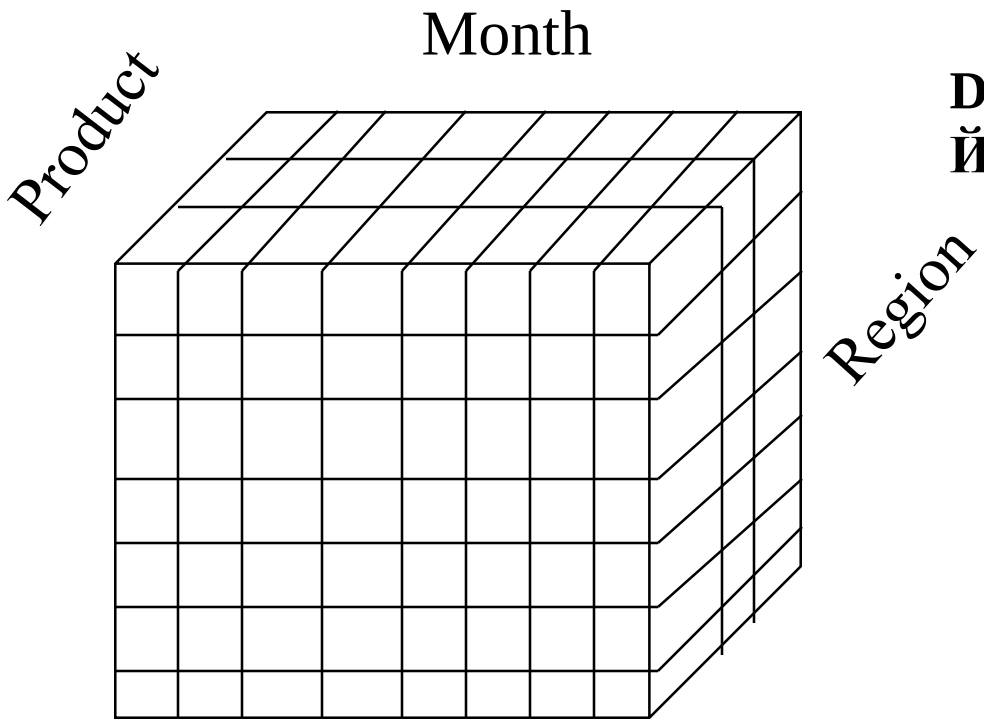
- Димензиите категоризират данните по йерахичен начин.
- Примери:
  - димензията време може да включва като членове:
    - Година → тримесечие → месец → дни
      седмица
  - димензията месторазположение включва като членове: държава -> регион -> град
- В димензионната таблица се съдържа по един ред за всеки член от по-ниското йерархично ниво на димензията.

Концепция за йерархия: Пространствена димензия

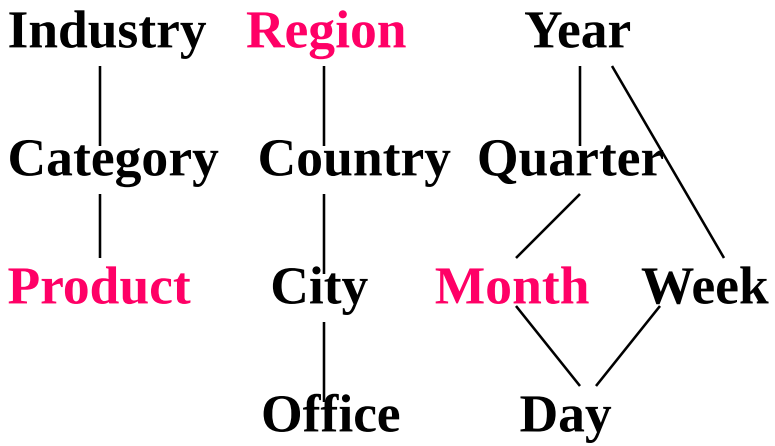


Многомерни данни

- Обемът на продажбите е функция на димeнсиите product, month и region



Dimensions: Product, Location, Time
Йерархичен път за обобщения



Факти

- Фактите представляват
  - Нещата, които трябва да се измерят и анализират;
  - Напр., количество продадени стоки, стойност на продажбите.
- Съхраняват се в т.н. Факт таблици.
- Във факт таблиците се съхраняват идентификаторите на дименсиите (показателите) и фактите, които ще се измерват и анализират.

Най-често срещащи се типове факти

- Факти, свързани с транзакции (Transaction facts). Основават се на отделни събития (типични примери са посещение на сайта или теглене на пари от сметка от банкомат);
- Факти, свързани с «моментални снимки» (Snapshot facts). Основават се на състоянието на обекта (например, банкова сметка) в определен момент от времето, (напр. в края на деня или месеца наличните парични средства). Типични примери на такива факти са обем на продажбите за деня.

Най-често срещащи се типове факти

- Факти, свързани с елементи на документи (Line-item facts). Основават се на един или друг документ (напр., сметка за стока или услуги) и съдържат подробна информация за елементите на този документ (напр., количество, цена, процент отстъпка).
- Факти, свързани със събития или състояния на обекта (Event or state facts). Представляват възникнали събития без подробности за тях (напр., просто фактът за продажба или фактът за отсъствие на такава без всякакви подробности).

Многомерният модел се изгражда от

- Димензионни таблици и Факт-таблици, които са свързани помежду си.
- Димензионните таблици се свързват с факт таблицата в отношение 1:M.
- Някои димензионни таблици могат да се свързват с други димензионни таблици в отношение 1:M.

Димензионни таблици

- Димензионната таблица съдържа непроменливи или рядко изменяеми данни.
- Структурата на димензионната таблица включва:
  - Минимум една описателна колона (обикновено с името на члена на димензията)
  - Колона - ключ за еднозначна идентификация на члена на димензията (като правило, целочислена колона - обикновено изкуствен ключ)
  - Допълнителни атрибути на членовете на димензиите, съдържащи се в изходната оперативна база от данни (напр., адресите и телефоните на клиентите)

Пример на димензионна таблица

PC-PC\MSSQLSERVE...- dbo.DimProduct

ProductKey	ProductLab...	ProductName	ProductSub...	Manufactur...	BrandName	ClassID	ClassName	StyleI...
1	0101001	Contoso 512MB MP3 Pla...	1	Contoso, Ltd	Contoso	1	Economy	1
2	0101002	Contoso 512MB MP3 Pla...	1	Contoso, Ltd	Contoso	1	Economy	5
3	0101003	Contoso 1G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
4	0101004	Contoso 2G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
5	0101005	Contoso 2G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
6	0101006	Contoso 2G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	2
7	0101007	Contoso 2G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
8	0101008	Contoso 4G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	2
9	0101009	Contoso 4G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
10	0101010	Contoso 4G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
11	0101011	Contoso 4G MP3 Player ...	1	Contoso, Ltd	Contoso	1	Economy	1
12	0101012	Contoso 4GB Flash MP3 ...	1	Contoso, Ltd	Contoso	1	Economy	3
13	0101013	Contoso 4GB Flash MP3 ...	1	Contoso, Ltd	Contoso	1	Economy	4
14	0101014	Contoso 4GB Flash MP3 ...	1	Contoso, Ltd	Contoso	1	Economy	1
15	0101015	Contoso 4GB Flash MP3 ...	1	Contoso, Ltd	Contoso	1	Economy	4
16	0101016	Contoso 8GB Super-Slim...	1	Contoso, Ltd	Contoso	2	Regular	3
17	0101017	Contoso 8GB Super-Slim...	1	Contoso, Ltd	Contoso	2	Regular	1
18	0101018	Contoso 8GB Super-Slim...	1	Contoso, Ltd	Contoso	2	Regular	3

Пример на димензионна таблица за димензия време DimDate

PC-PC\MSSQLSERVE...QL - dbo.DimDate

Datekey	FullDateLa...	CalendarYe...	CalendarYe...	CalendarH...	CalendarQ...	CalendarM...	CalendarW...	CalendarDay...	FiscalYear
2005-01-01 ...	2005-01-01	2005	Year 2005	H1	Q1	January	Week 1	Saturday	2005
2005-01-02 ...	2005-01-02	2005	Year 2005	H1	Q1	January	Week 2	Sunday	2005
2005-01-03 ...	2005-01-03	2005	Year 2005	H1	Q1	January	Week 2	Monday	2005
2005-01-04 ...	2005-01-04	2005	Year 2005	H1	Q1	January	Week 2	Tuesday	2005
2005-01-05 ...	2005-01-05	2005	Year 2005	H1	Q1	January	Week 2	Wednesday	2005
2005-01-06 ...	2005-01-06	2005	Year 2005	H1	Q1	January	Week 2	Thursday	2005
2005-01-07 ...	2005-01-07	2005	Year 2005	H1	Q1	January	Week 2	Friday	2005
2005-01-08 ...	2005-01-08	2005	Year 2005	H1	Q1	January	Week 2	Saturday	2005
2005-01-09 ...	2005-01-09	2005	Year 2005	H1	Q1	January	Week 3	Sunday	2005
2005-01-10 ...	2005-01-10	2005	Year 2005	H1	Q1	January	Week 3	Monday	2005
2005-01-11 ...	2005-01-11	2005	Year 2005	H1	Q1	January	Week 3	Tuesday	2005
2005-01-12 ...	2005-01-12	2005	Year 2005	H1	Q1	January	Week 3	Wednesday	2005
2005-01-13 ...	2005-01-13	2005	Year 2005	H1	Q1	January	Week 3	Thursday	2005
2005-01-14 ...	2005-01-14	2005	Year 2005	H1	Q1	January	Week 3	Friday	2005
2005-01-15 ...	2005-01-15	2005	Year 2005	H1	Q1	January	Week 3	Saturday	2005
2005-01-16 ...	2005-01-16	2005	Year 2005	H1	Q1	January	Week 4	Sunday	2005
2005-01-17 ...	2005-01-17	2005	Year 2005	H1	Q1	January	Week 4	Monday	2005
2005-01-18 ...	2005-01-18	2005	Year 2005	H1	Q1	January	Week 4	Tuesday	2005

Факт таблица - основна таблица в склада от данни

- Съдържа сведения за обектите или събитията, съвкупността от които ще бъдат анализирани.
- Факт таблицата съдържа само числови данни и индекси, които съответстват на първичните ключове на свързаните димeнсионни таблици.
- Като правило, съдържа уникален съставен ключ, обединяващ първичните ключове на димeнсионните таблици.

Факт- таблица, съдържаща данни за продажбите в магазините на фирма....

PC-PC\MSSQLSERVER...L - dbo.FactSales

SalesKey	DateKey	channelKey	StoreKey	ProductKey	Promotion...	CurrencyKey	UnitCost	UnitPrice	SalesQuant...
7077	2008-04-13 ...	1	297	1086	17	1	139,8000	304,0000	9
7078	2009-06-14 ...	1	203	904	1	1	38,7400	75,9900	20
7079	2009-11-01 ...	3	200	221	22	1	275,4600	599,0000	13
7080	2008-12-11 ...	1	162	1132	13	1	207,7400	627,0000	13
7081	2007-04-16 ...	1	265	693	6	1	75,8700	229,0000	18
7082	2007-06-08 ...	1	185	222	1	1	261,6600	569,0000	10
7084	2007-05-08 ...	2	306	282	1	1	208,5200	409,0000	10
7085	2009-09-22 ...	2	307	381	1	1	321,4400	699,0000	10
7087	2009-08-13 ...	3	200	504	21	1	287,9200	869,0000	6
7088	2008-09-25 ...	4	308	437	12	1	254,8600	499,9000	24
7089	2008-12-29 ...	3	200	1029	13	1	66,2600	200,0000	39
7091	2009-01-18 ...	1	267	1146	14	1	291,0900	633,0000	13
7092	2007-06-27 ...	1	174	1244	1	1	90,7500	178,0000	5
7094	2007-09-21 ...	1	97	1541	3	1	137,5000	299,0000	12
7095	2008-08-15 ...	2	199	458	12	1	117,2100	229,9000	96
7096	2009-10-04 ...	3	200	310	1	1	152,4400	299,0000	10
7097	2009-07-28 ...	3	200	966	21	1	84,8400	184,5000	12
7098	2007-02-15 ...	1	232	962	8	1	86,4500	188,0000	12

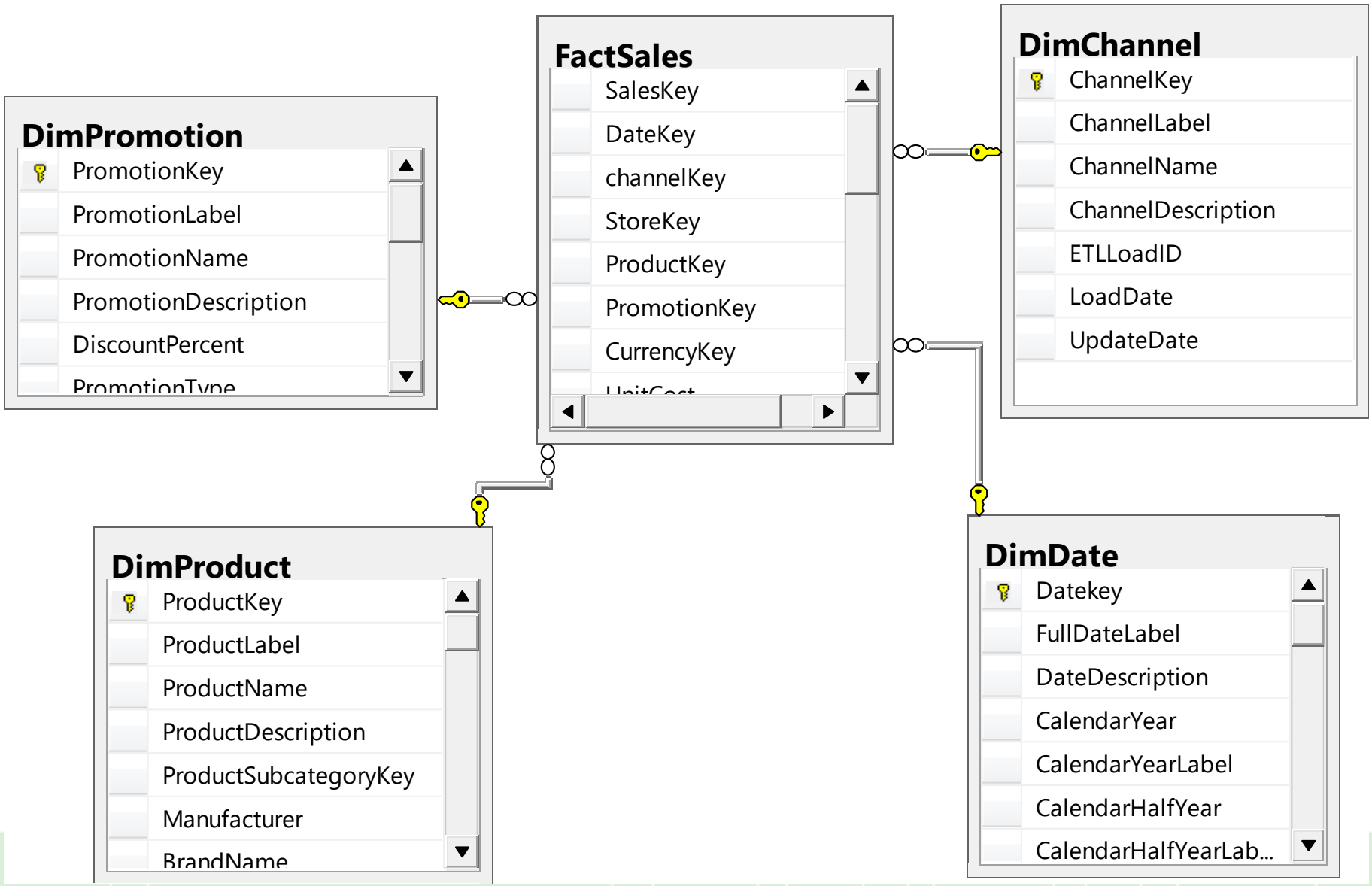
Схеми в многомерния модел на данни

- Основават се на връзките между дименсионните таблици и факт-таблиците
- Star schema (схема звезда)
- Snowflake schema (схема снежинка)
- Fact constellations или galaxy schema (галактика, плеяда, съзвездие)

Star schema

- Факт-таблицата е в центъра и е свързана с няколко димензионни таблици. По този начин се получава структура, имаща вид на звезда
- Връзката на всяка димензионна таблица с факт таблицата се основава на общи колони, които представляват идентификатор на димензията.

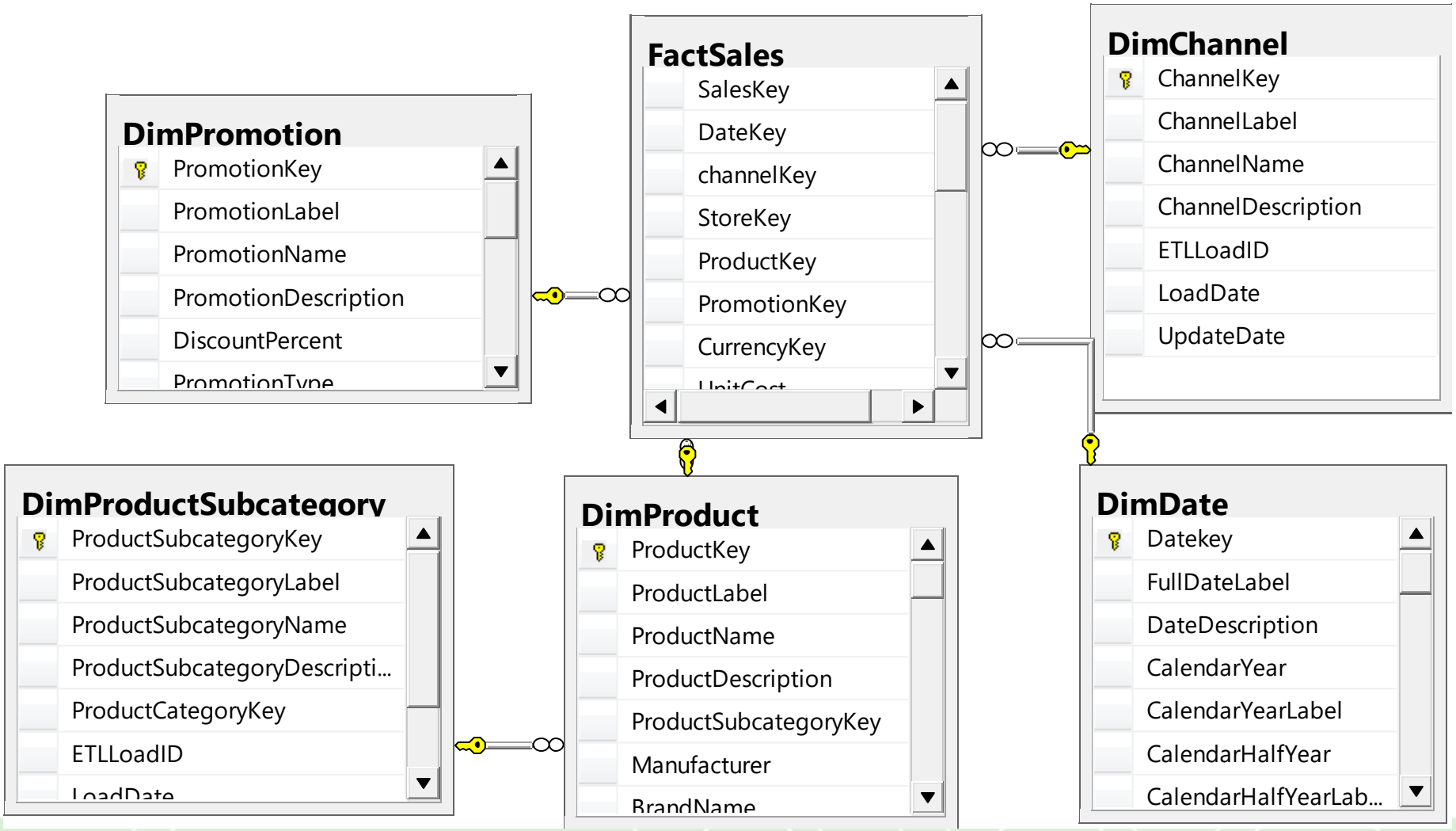
Пример на схема “звезда”



Snowflake schema

- Факт таблицата се свързва с множество димензионни таблици, някои от които са свързани с други димензионни таблици
- Схемата снежинка е усъвършенствана схема звезда, в която йерархиите от димензии са нормализирани в набор от по-малки димензионни таблици.
- По този начин димензията има йерархия, която се задава чрез съединяването на димензионните таблици (напр. Категории стоки и Стоки)
- Само една от свързаните димензионни таблици се свързва с факт таблицата

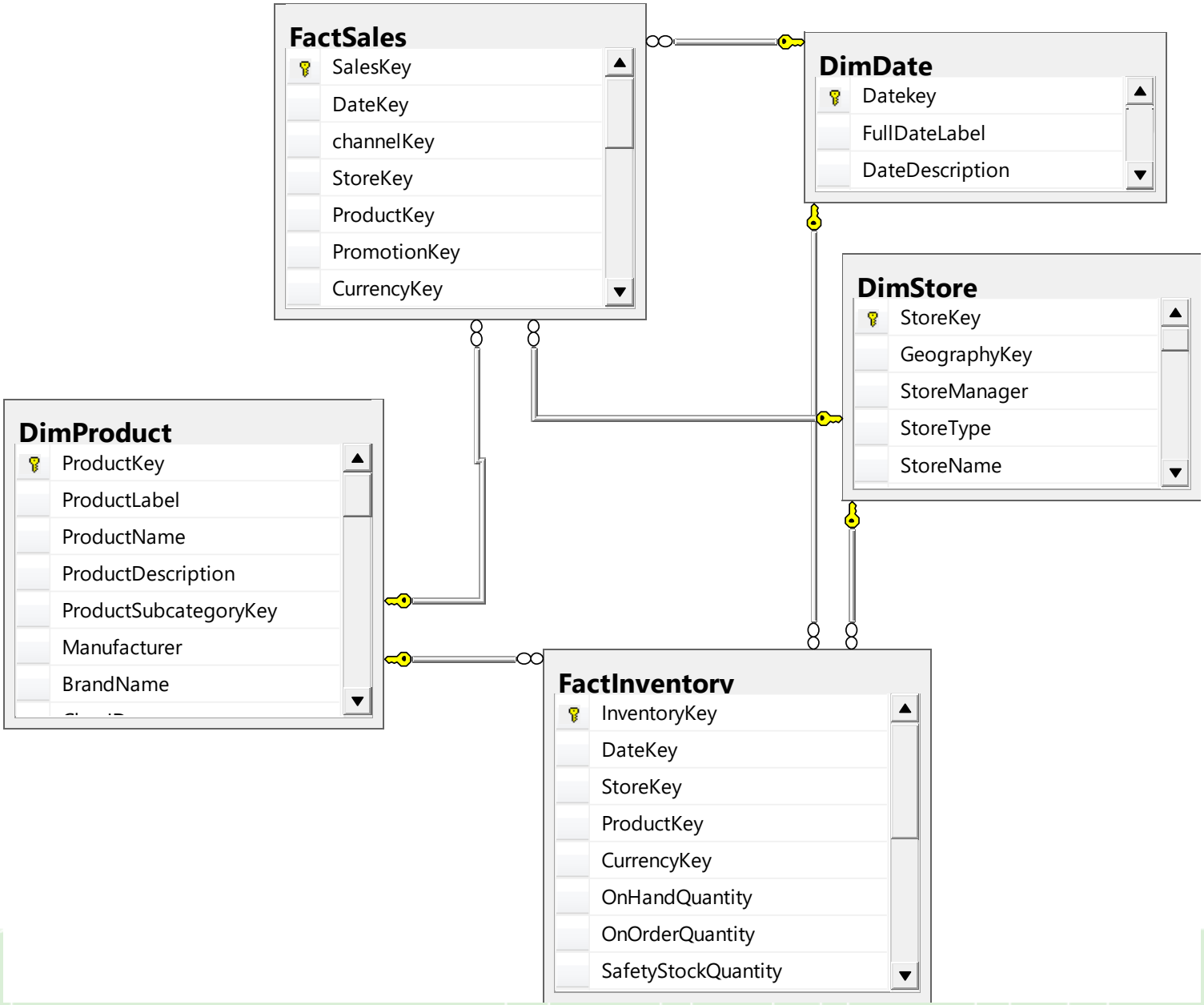
Пример на схема “снежинка”



Galaxy schema (Fact constellations)

- Моделът съдържа няколко факт-таблици и множество димензионни таблици.
- Факт-таблиците поделят димензионните таблици.
- Изглежда като колекция от звезди, затова схемата се нарича галактика или съвездие.

Пример на схема “галактика”



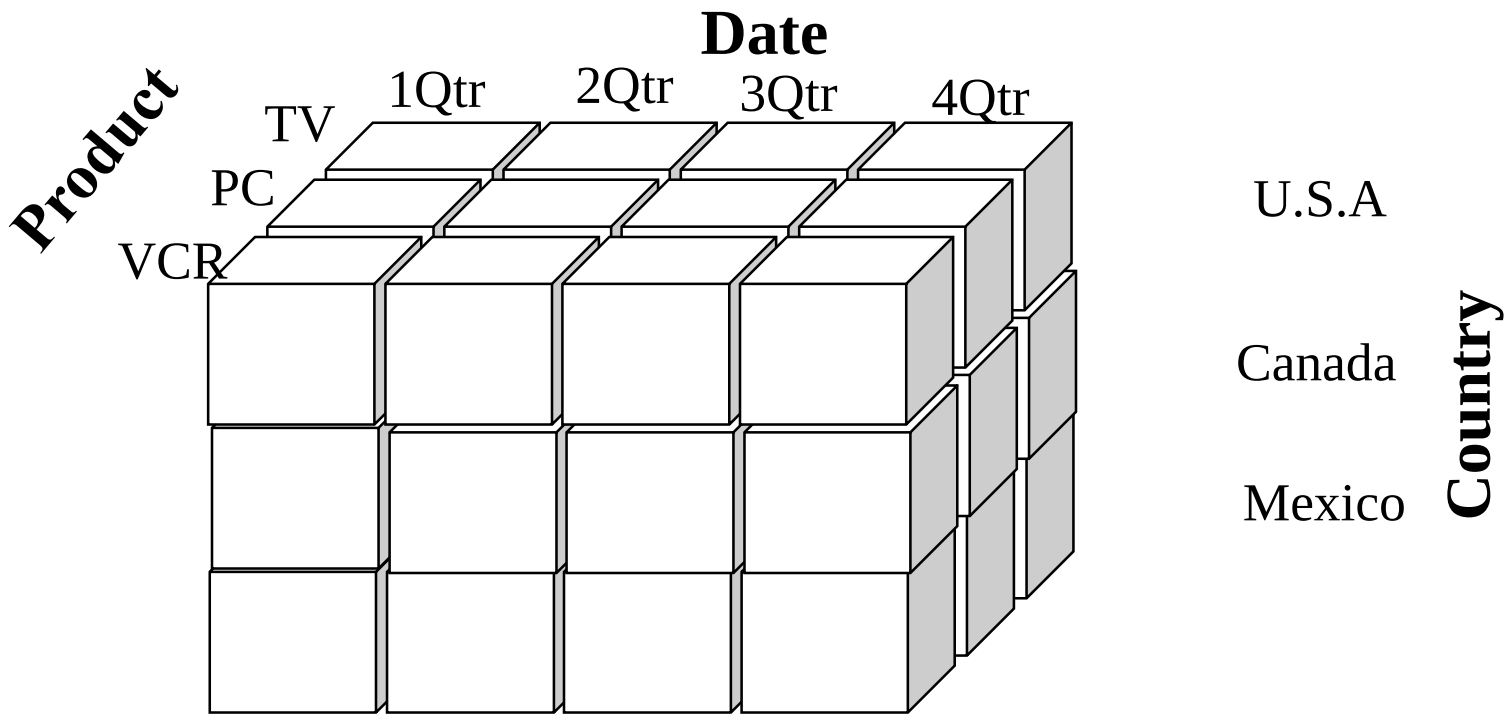
Куб от данни (Data Cube)

- Кубът от данни позволява данните, съхранявани в склада от данни, да бъдат показани чрез множество измерения.
- Дефинира се чрез
  - Димензионни таблици;
  - Факт таблица, която съдържа мерки (**measures)** и ключове (**keys)** за всяка от свързаните димензионни таблици.
- Мерките на куба се създават от колоните на факт таблицата.

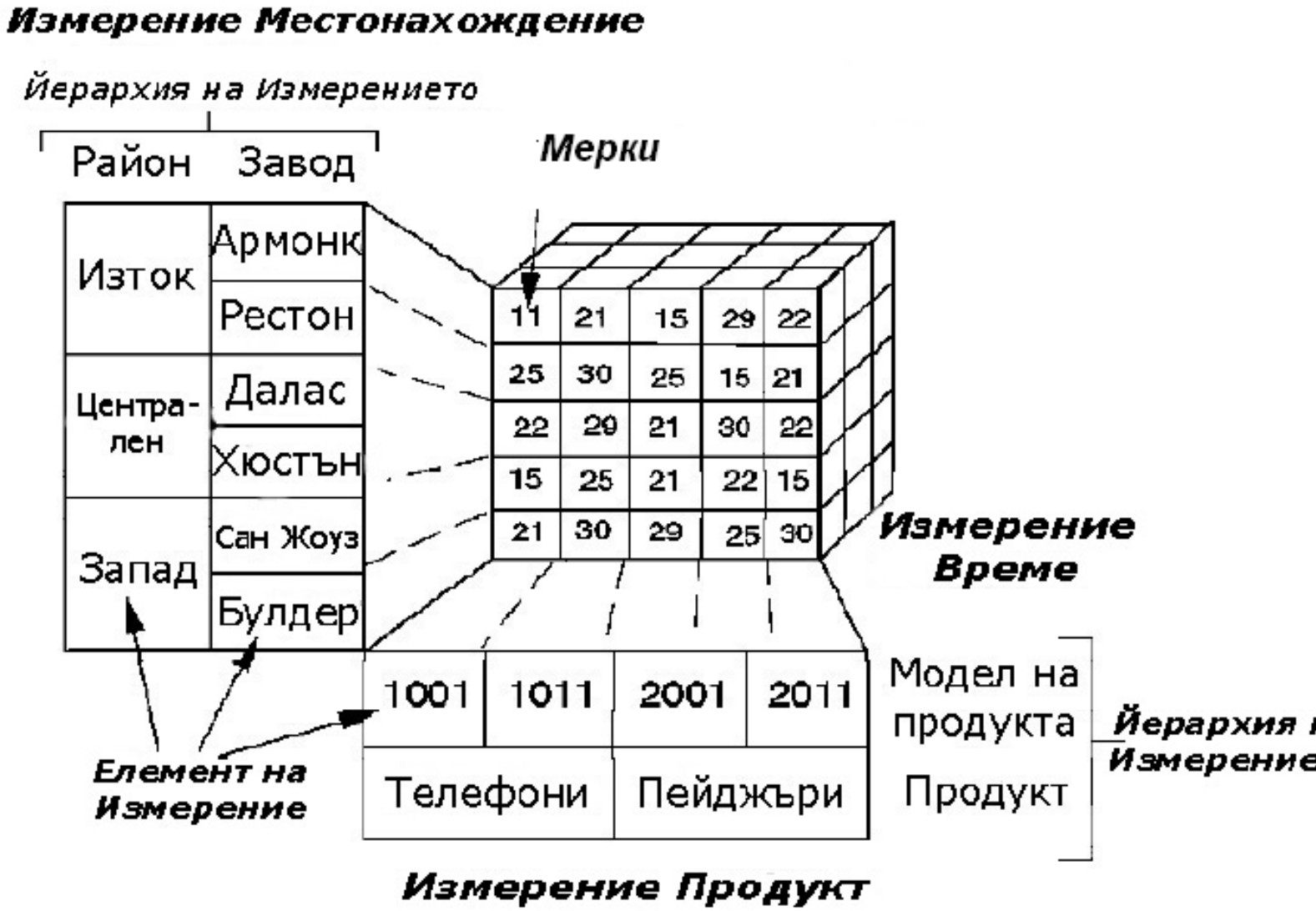
Мерки (Measures)

- Мерките са числови стойности, съдържащи се в клетките на куба.
- Мерките се създават въз основа на числовите колони на факт таблицата по зададена агрегатна функция – Sum(), Min(), Count(), Avg(), standard_deviation(), median(), mode(), rank(), ….
- Мерките представляват точки в многомерното пространство

Пример на куб с данни



Пример на куб с данни с три измерения – местонахождение, продукт и време



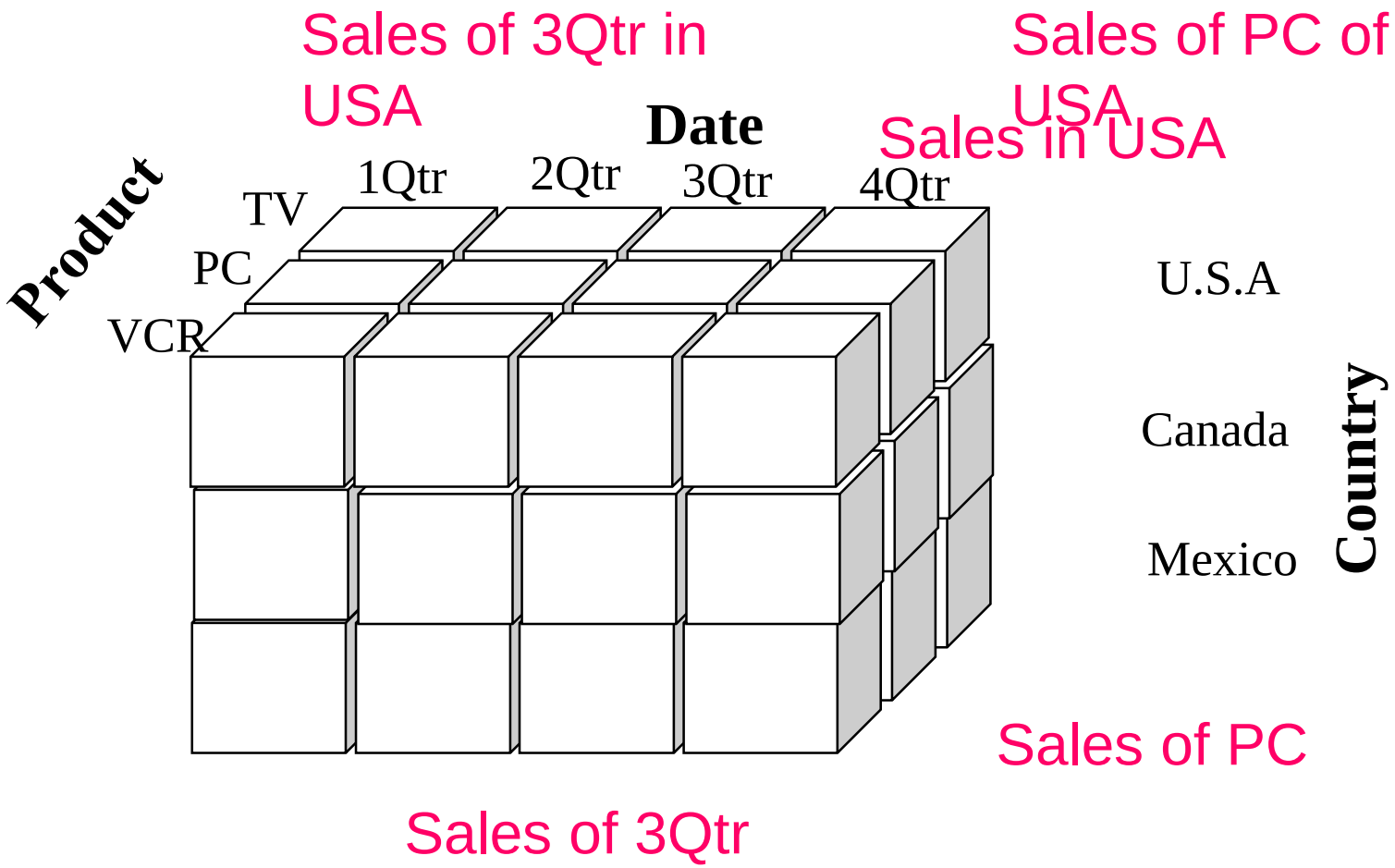
Типични OLAP операции

- Slice and dice (резен, разрез и зар):
  - *Проекция и селекция*
- Pivot (rotate):
  - *Преорентиране на куба, визуализация, 3D към серия от 2D план (схема).*

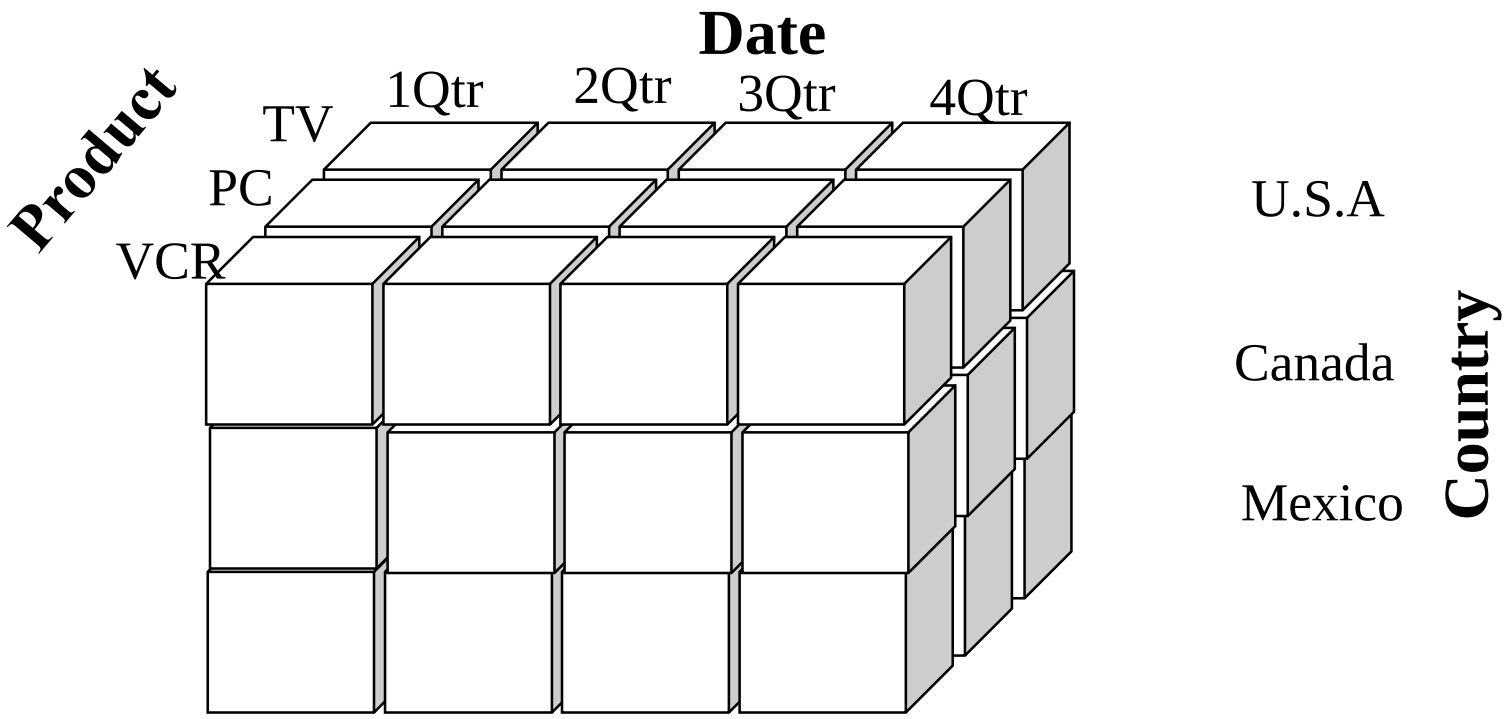
Типични OLAP операции

- Roll up (drill-up): обобщаване на данните
  - *От по-ниско ниво на обобщение към по-високо (изкачване нагоре по йерархията или редукция на димензията).*
- Drill down (roll down): обратно на roll-up
  - *От по-високо ниво на обобщение към по-ниско ниво на обобщение или детайлни данни, или въвеждане на нова димензия.*

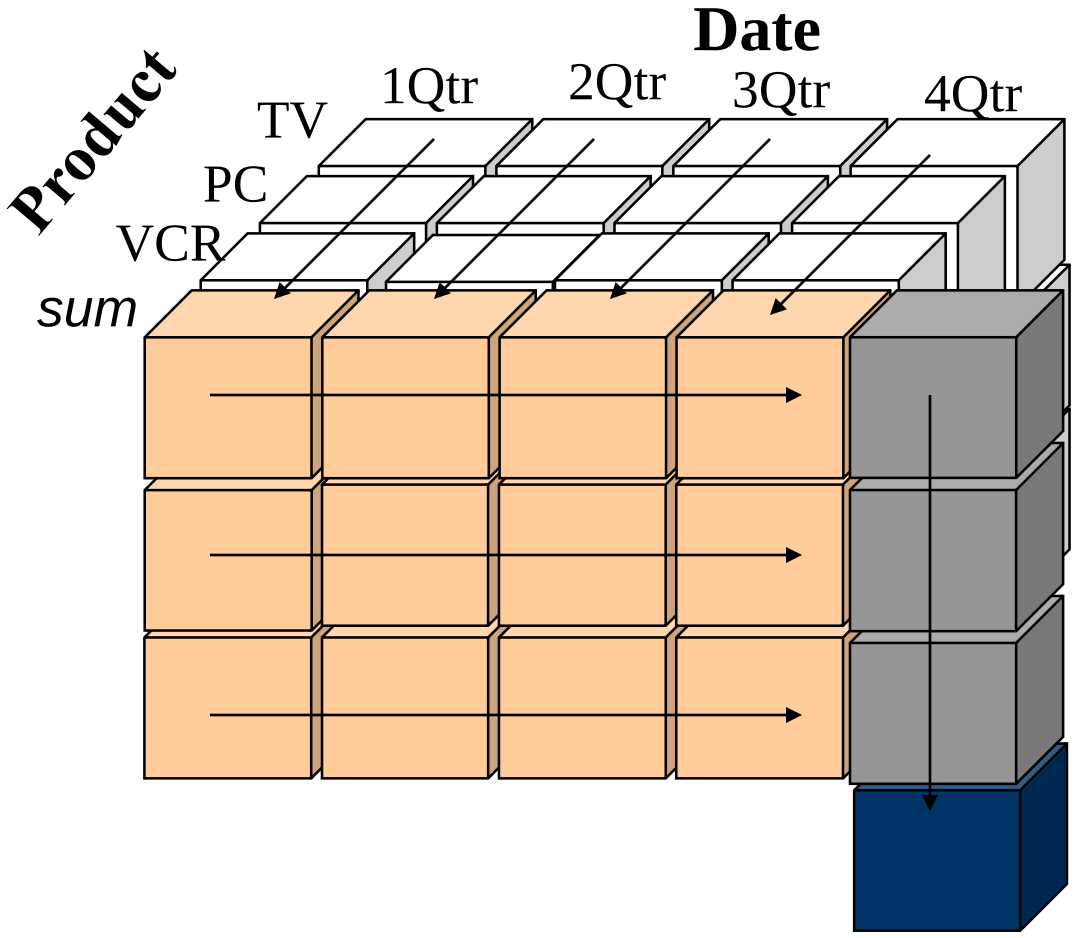
Slice and Dice



Pivot (rotate):



Roll up (drill-up) (*dimension reduction example)*



Group by date, country

- reduced product dimension

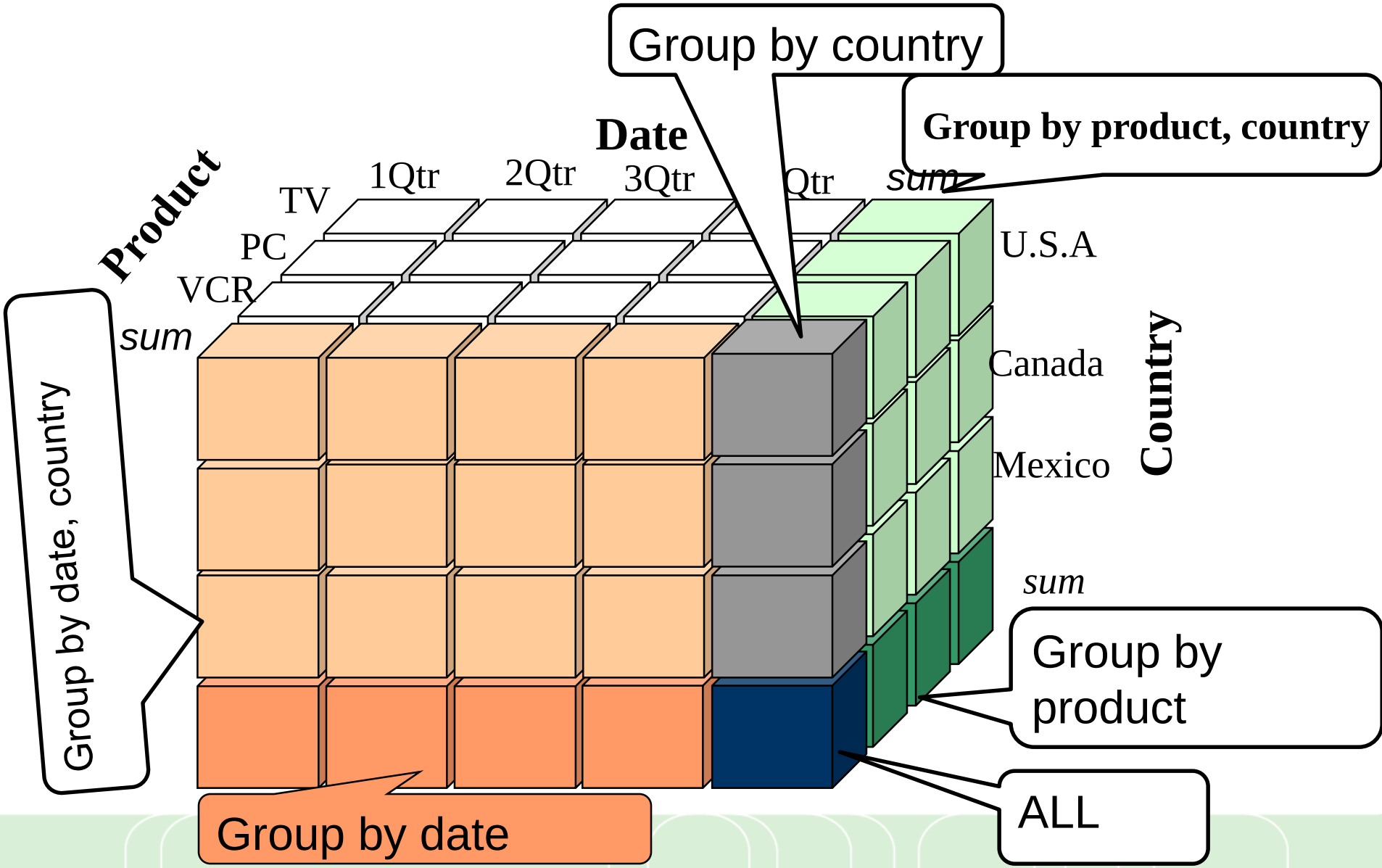
Group by country

- reduced date dimension

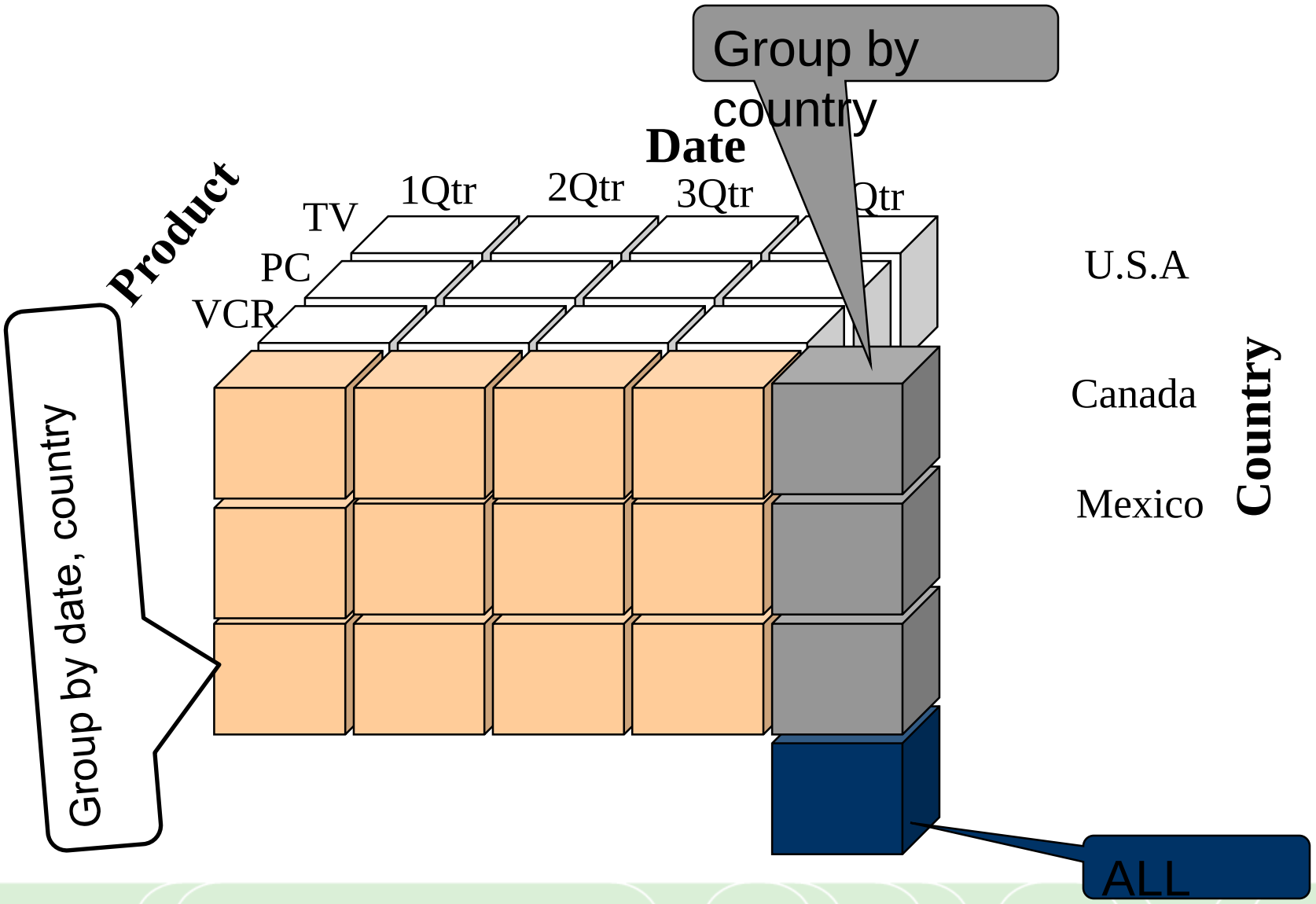
Group by *nothing*

- reduced all dimensions

Roll up (drill-up)



Drill down (roll-down)



Три начина за съхраняване на изходните и агрегатните данни при обработка на куба

- MOLAP (Multidimensional OLAP)
- ROLAP (Relational OLAP)
- HOLAP (Hybrid OLAP)

MOLAP (Multidimensional OLAP)

- Изходните и агрегатните данни се съхраняват в многомерна база от данни.
- Съхраняването на данните в многомерни структури позволява да се манипулира с данните като с многомерни масиви, благодарение на което скоростта на изчисляване на агрегатните стойности е еднаква за всяко от измеренията.

ROLAP (Relational OLAP)

- Изходните данни остават в същата релационна база от данни, в която те първоначално са се намирали.
- Агрегатните данни се разполагат в специално създадени за тяхното съхраняване служебни таблици в същата база от данни.

HOLAP (Hybrid OLAP)

- Изходните данни остават в същата релационна база от данни, където първоначално са се намирали.
- Агрегатните данни се съхраняват в многомерната база от данни.

- http://www.1keydata.com/datawarehousing/
- http://www.learndatamodeling.com
- http://www.kimballgroup.com/

Power Pivot for SQL Server 2012

- http://office.microsoft.com/bg-bg/support/results.aspx?qu=PowerPivot&ex=2&filter=1&av=zxl
- http://technet.microsoft.com/en-us/library/hh965697.aspx