Складове от данни

Основни въпроси

- Мотивация
- Същност
- Функционалност
- Модели

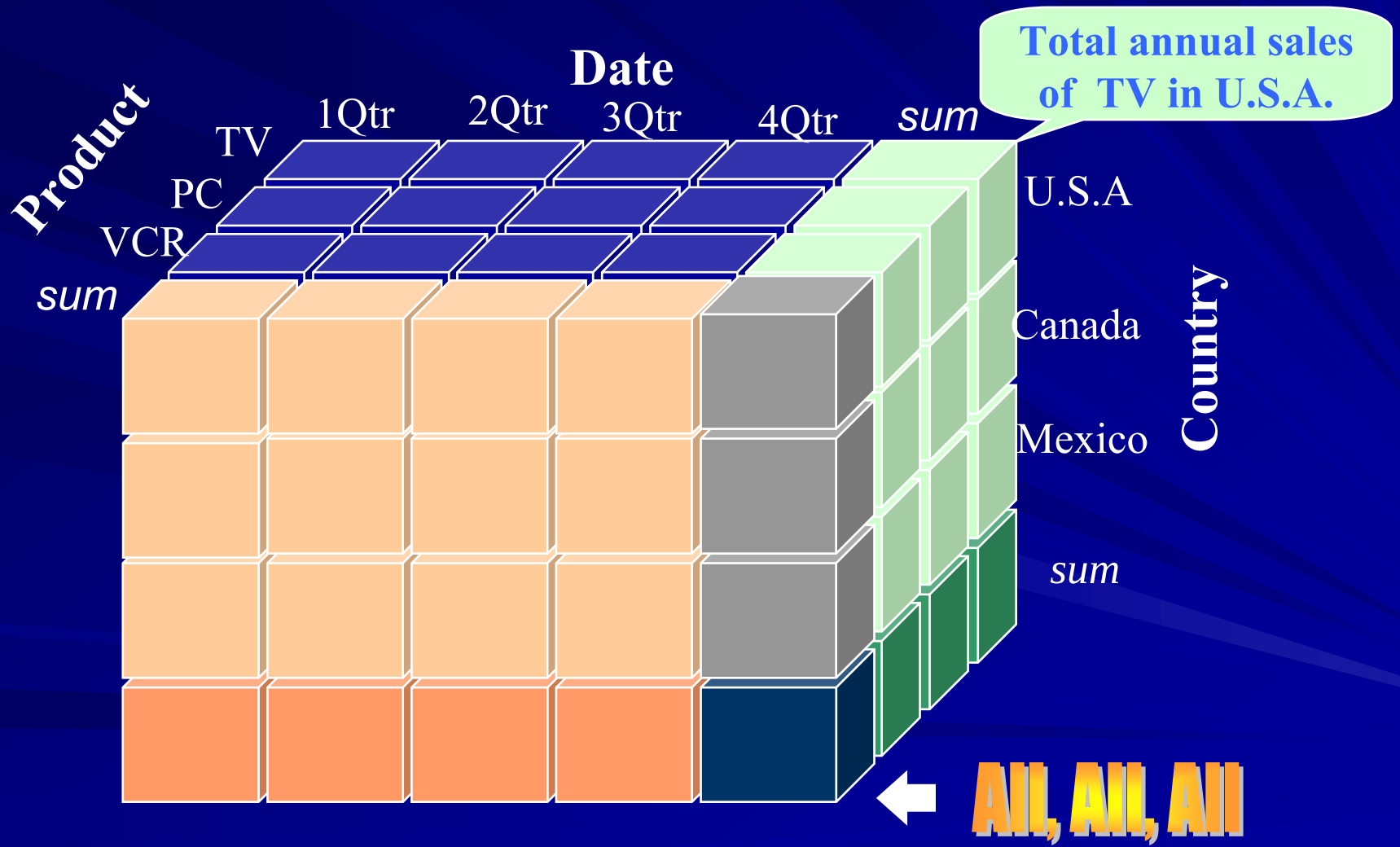
Обосновка

- Проблем: Бързото увеличаване на данните
  - Автоматизирани инструменти за събиране на данни и развитие на технологии за БД водят до огромен обем данни, съхранени в БД и други информационни хранилища.
- Богати на данни, бедни на информация
- Решение: Data warehousing и data mining
  - Data warehousing и аналитична обработка в реално време
  - Извличане на знания - правила, образци, шаблони, ограничения на данни в големи БД.

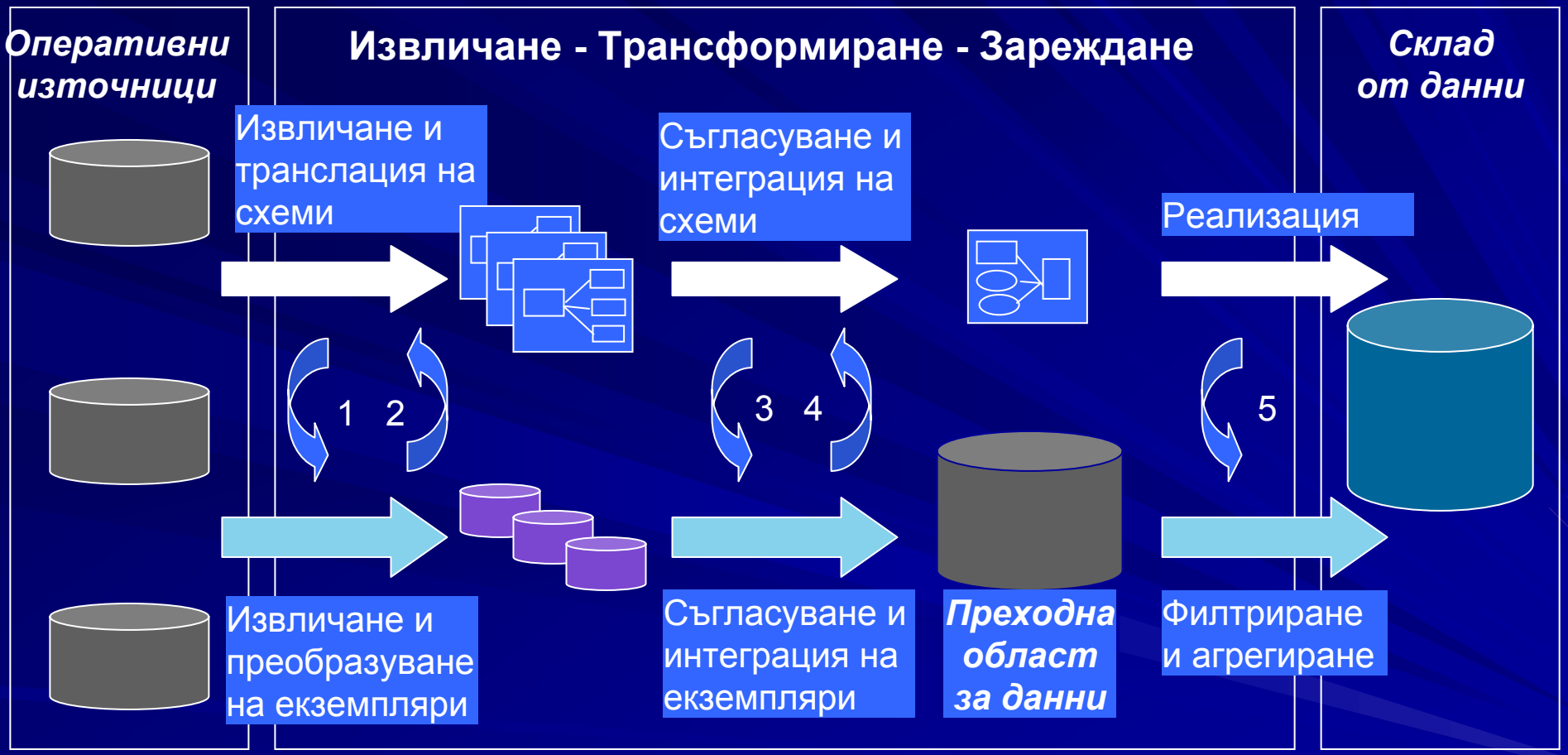
Необходимост от Data Warehouse?

- Оперативните системи не позволяват анализ на бизнес информация поради:
  - Липса на онлайн исторически данни
  - Необходимите за анализа данни принадлежат на различни с-ми
  - Самите схеми са неподходящи за DS
  - Системата на заявки не е достатъчна
  - Липса на средства за бизнес анализ
  - Неефективно зареждане и индексиране на големи количества данни
  - 2-размерно представяне не е достатъчно

A Sample Data Cube



Изчистване на данни и ETL-процес

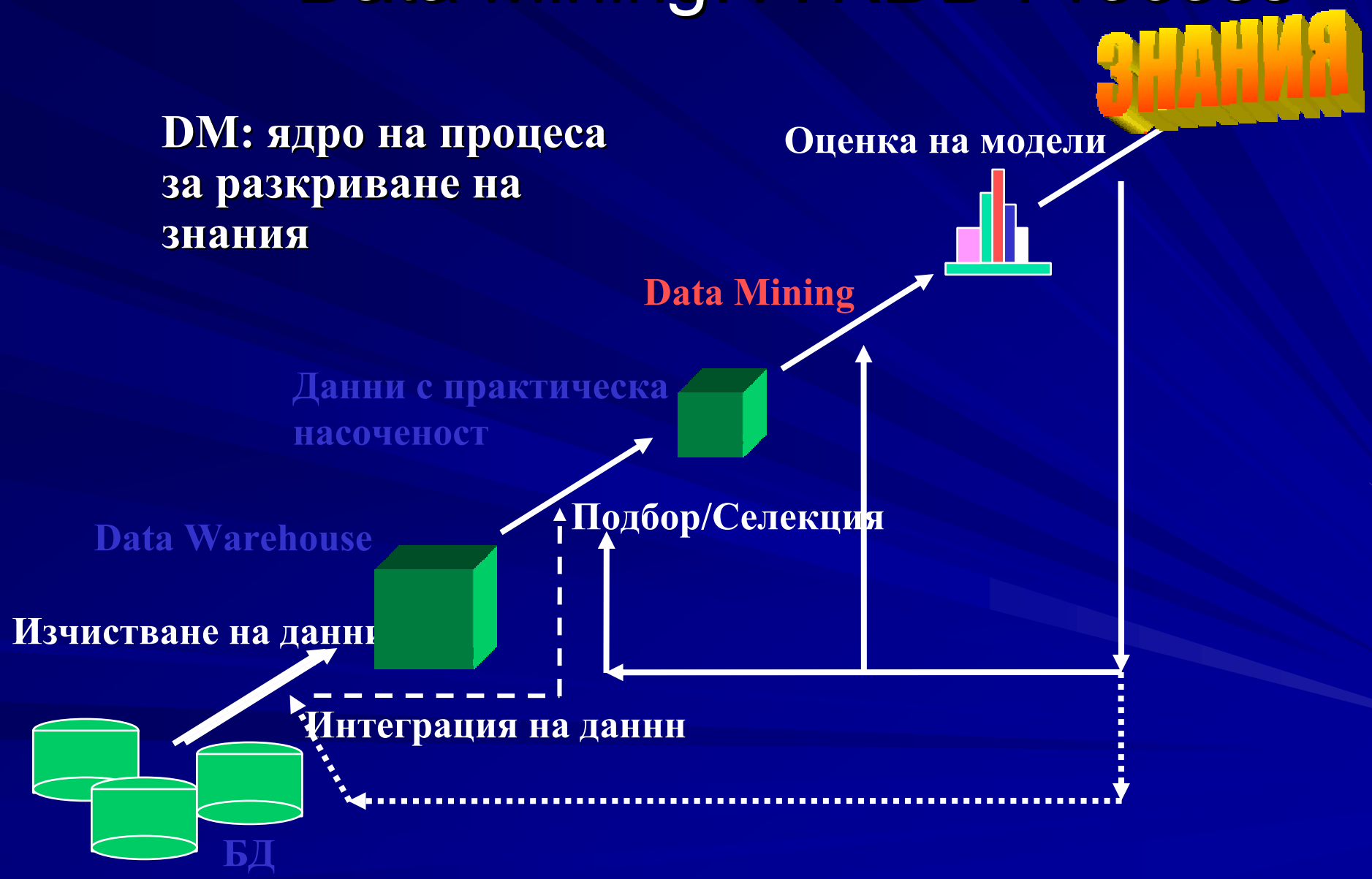


1,3 – Характеристики на екземплярите 2 – Правила за транслация

4 – Съответствие между изходните схеми и целевата схема

Поток от мета данни Поток от данни

Data Mining: A KDD Process



Data Mining

- Data mining (разкриване на знания в БД):
  - Извличане на значима (нетривиална, пълна, предварително неизвестна и потенциално полезна) информация или модели от данни в големи БД.
- Възможни приложения
  - Анализи и управление на пазара
  - Анализ и управление на риска
  - Разкриване и управление на измами
  - Текст mining (news групи, e-mail, документи) и Web анализи
  - Интелигентни отговори на заявки

Data Warehouse. Концепции

- IBM – “information warehouse”
- Дефиниция на Inmon
- Интегриране на данни от различни източници в едно хранилище - warehouse
- Технология за управление и анализ на данните

Основни идеи

- Интеграция на различни детайлизирани данни в единно хранилище (съгласуване и агрегация):
  - Исторически архиви
  - Данни от традиционни БД
  - Данни от външни източници
- Разделяне на данните за оперативна обработка и данните за решаване на аналитични задачи

DW - Дефиниция на Inmon

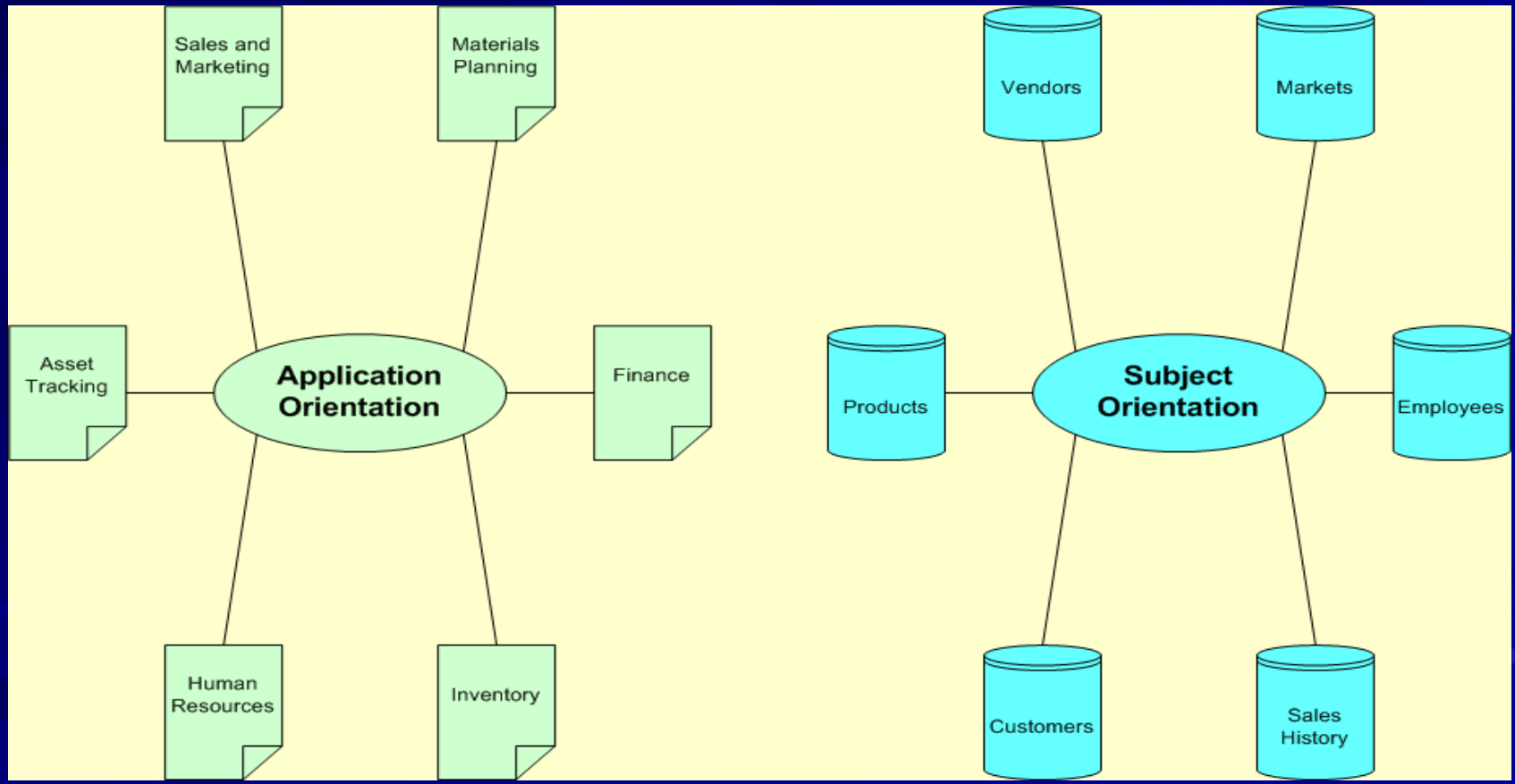
- "A **data warehouse** is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

  (Building the Data warehouse, 1993)
- Data warehousing:
  - The process of constructing and using data warehouses

DW - предметно-ориентирана

- Фокусира върху основни същности като customer, product, sales, а не върху процеси.
- Акцентира върху моделирането и анализа на данни, необходими за вземане на решения, а не върху ежедневните транзакционни процеси.

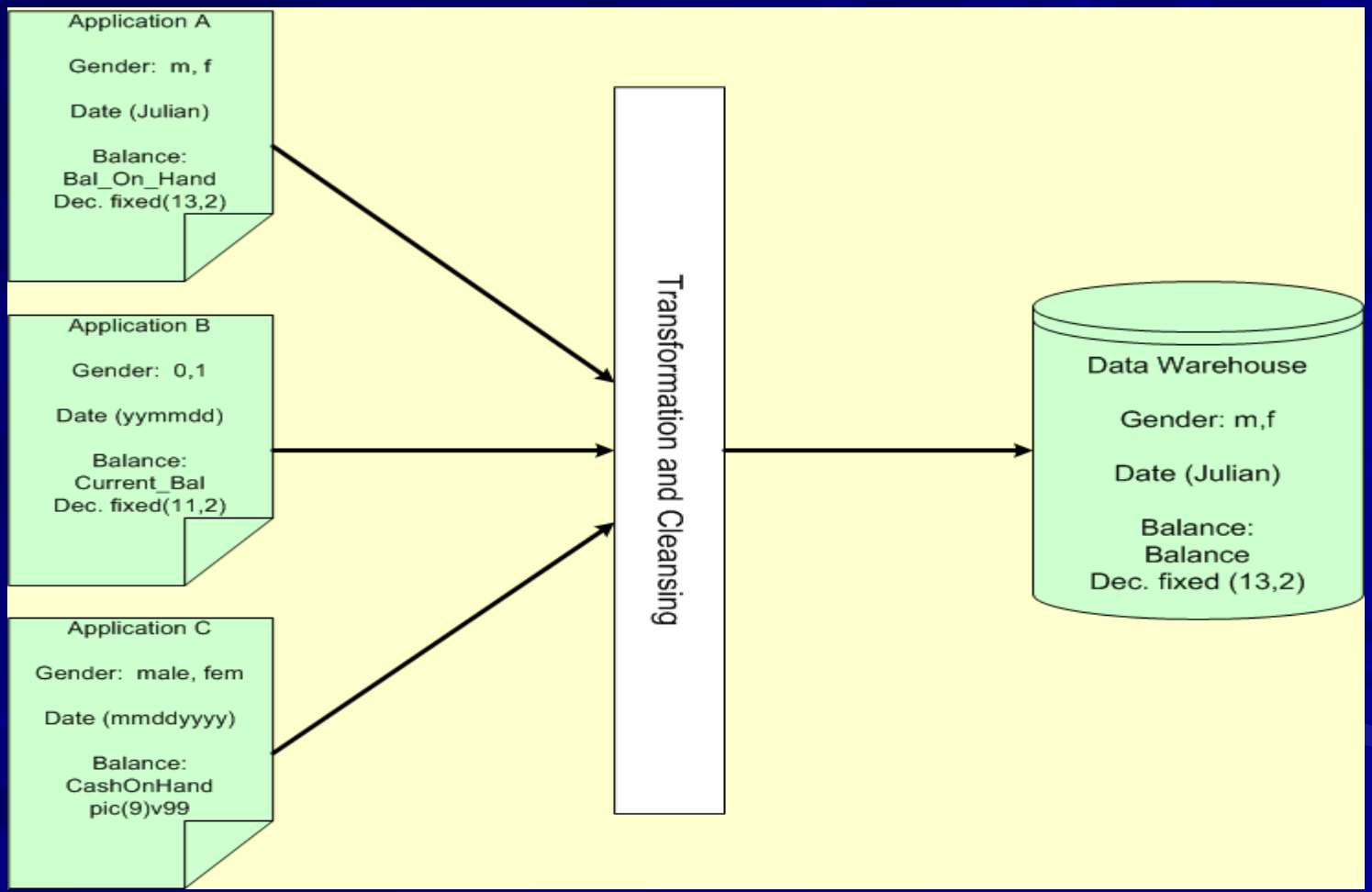
DW - предметно-ориентирана



DW - интегрирани данни

- Основен аспект на DW
- Съгласува данни от множество разнородни източници
  - Релационни БД, Web страници, flat files
- Техники за изчистване и интегриране на данните.
  - Конвенции за имена, ограничения за домени

DW - интегрирани данни



DW - поддържане на хронология

- Данните са свързани с определена времева точка
  - Семестър, фискална година, период за пращане
- Времевият диапазон е значително по-голям, в сравнение с традиционните БД
  - Оперативни БД: стойности на текущите данни.
  - Данни в Data warehouse: осигуряват информация от историческа перспектива (последните 5-10 години)
- Всяка ключова структура в DW притежава явна или неявна времева характеристика

DW – относително неизменни

- DW – физически разделена от данните, които се трансформират в оперативна среда
- Оперативните обновявания на данни не се извършват в DW
  - Не се изисква обработка на транзакции, контрол на конкурентността, възстановяване
  - Основни операции за данни:
    - Първоначално зареждане на данни
    - Достъп до данни

Examples of Common DW Applications

Sales Analysis

- Determine real-time product sales to make vital pricing and distribution decisions.
- Analyze historical product sales to determine success or failure attributes.
- Evaluate successful products and determine key success factors.
- Use corporate data to understand the margin as well as the revenue implications of a decision.
- Rapidly identify a preferred customer segments based on revenue and margin.
- Quickly isolate past preferred customers who no longer buy.
- Identify daily what product is in the manufacturing and distribution pipeline.
- Instantly determine which salespeople are performing, on both a revenue and margin basis, and which are behind.

Financial Analysis

- Compare actual to budgets on an annual, monthly and month-to-date basis.
- Review past cash flow trends and forecast future needs.
- Identify and analyze key expense generators.
- Instantly generate a current set of key financial ratios and indicators.
- Receive near-real-time, interactive financial statements.

Human Resource Analysis

- Evaluate trends in benefit program use.
- Identify the wage and benefits costs to determine company-wide variation.
- Review compliance levels for EEOC and other regulated activities.

Other Areas

- Warehouses have also been applied to areas such as: logistics, inventory, purchasing, detailed transaction analysis and load balancing.

Data Warehouse <--> Operational DBMS

- OLTP (on-line transaction processing)
  - Основна задача на традиционните РСУБД
  - Ежедневни операции: покупки, продажби, банкови операции, ведомости, счетоводство
- OLAP (on-line analytical processing)
  - Основна задача на data warehouse system
  - Анализ на данните и вземане на решения

OLTP < -- > OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

DW модел – основни понятия

- Dimension (димeнсия, размерност)
- Facts (Факти)
- Attributes (атрибути)
- Hierarchy (йерархия)
- Relationships (връзки)

DW модел - компоненти

- Dimension (димeнсия, размерност)
  - Осигурява средства за анализ на бизнеса
  - С-жа 1 или няколко атрибута
  - Възможност за йерархия
- Attributes
  - Характеристики на димeнсии
    - Цвят , размер
    - Ден, седмица, празник в димeнсия време

DW модел - компоненти

- Hierarchy
  - Осигурява логическата връзка между 2 атрибута в дименсията
    - Географски район
- Relationship
  - Взаимоотношения между атрибутите в йерархията
    - 1:1, M:M
- Facts
  - Колони от данни, свързани чрез ключове с дименсионни таблици

Multidimensional data model

- Складовете от данни са базират на модела на многомерни данни (multidimensional data model) който разглежда данните като един куб от данни
- Куб от данни, например sales, позволява данните да се моделират и разглеждат в м-во размерности
  - Димензионни таблици item (item_name, brand, type), time(day, week, month, quarter, year)
  - Факт таблица, съдържаща мерки (напр. dollars_sold) и ключове към всяка от релационните димензионни таблици

Пример: AllElectronics

- Фирма за търговия с електроника AllElectronics създава склад за данни за *продажби (sales)*, за записи за продажби по отношение на размерностите
- *време (time)*
- *артикул (item)*
- *клон (branch)*
- *местоположение (location)*.

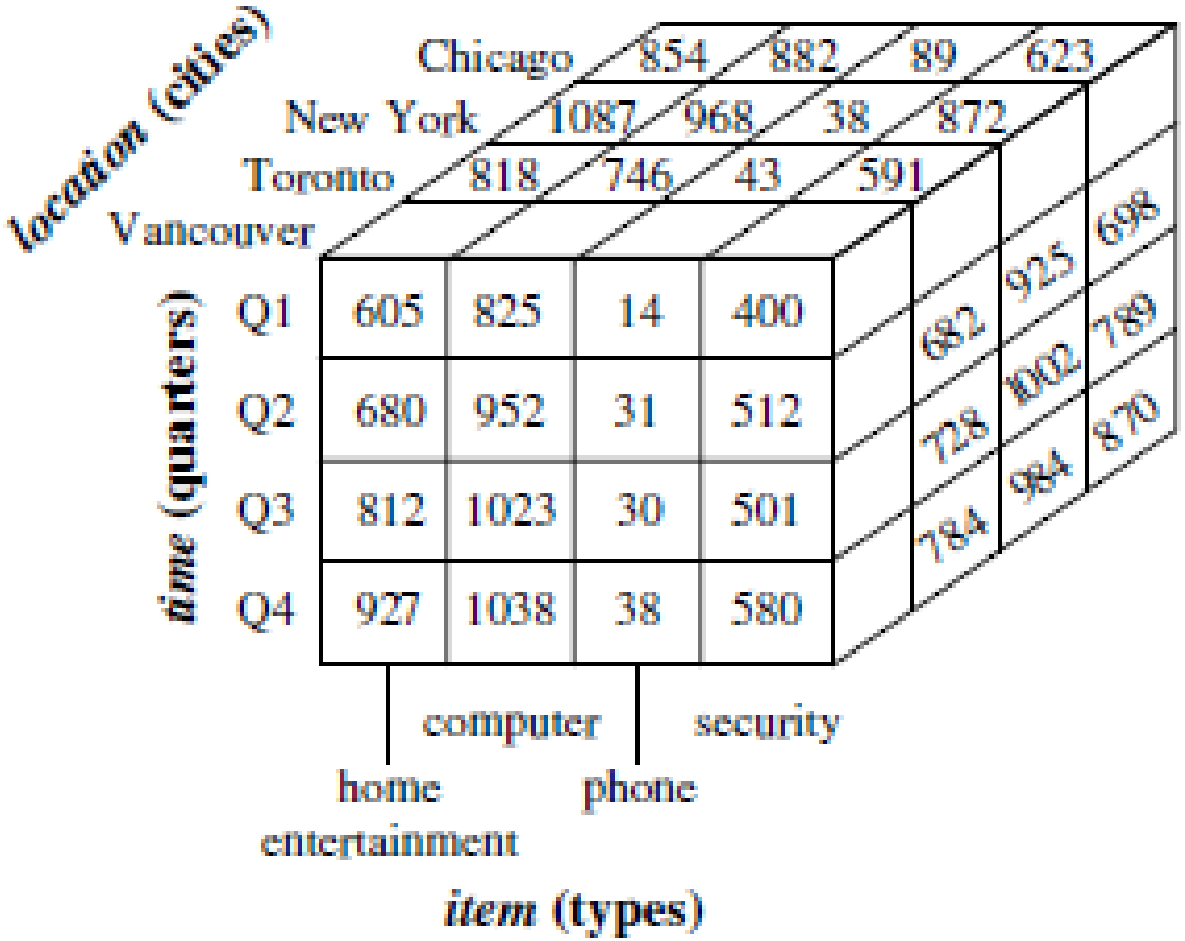
AllElectronics – 2-D

- Двумерен (2-D) куб данни -таблица за продажби на артикули, продадени във
  - Ванкувър
  - Торонто
  - Ню Йорк
  - Чикаго
- Димeнсии
  - (размерност “артикул”), организирани съгласно своите типове,
  - и разбити по тримесечия (quarters) размерност “време”.
- Показаният факт или мярка е “продажби-в-долари” (в хиляди).

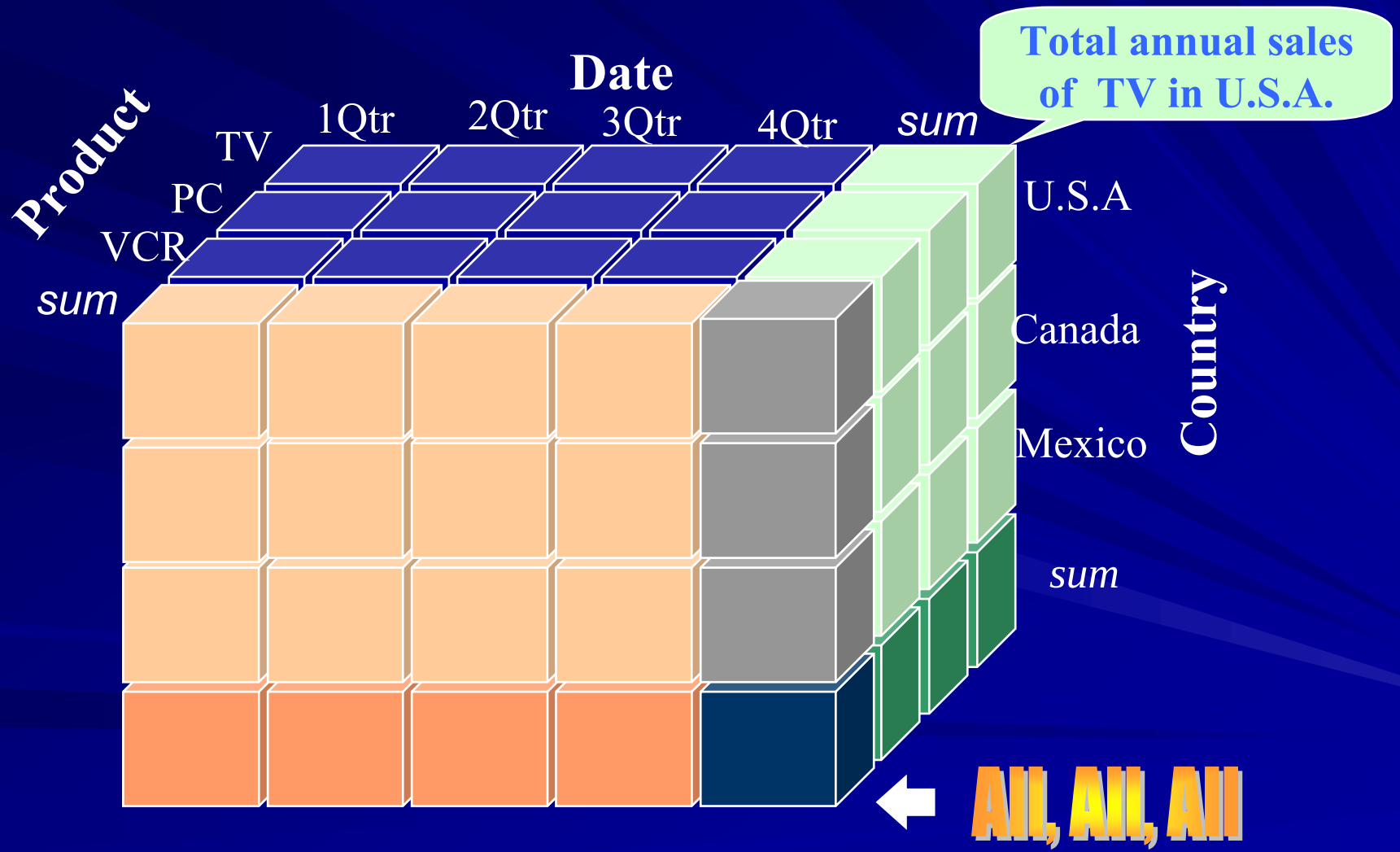
AllElectronics – 3-D

- Тримерен (3-D) куб данни -данните за продажби с използване на третата размерност – “местоположение” за градовете Чикаго, Ню Йорк, Торонто и Ванкувер.
- Димензии
  - артикул
  - време
  - местоположение
- Факт - “продажби-в-долари”

AllElectronics – 3-D



A Sample Data Cube



AllElectronics – fact table

- Един модел на многомерни данни обикновено е организиран около някоя централна тема (например “продажби”). Тази тема е представена чрез така наречена таблица на факти или факт таблица
- *Фактите* са определени числови мерки - количествата, чрез които искаме да анализираме съществуващи релации между размерностите.
  - “продажби-в-долари” (amount_sold) и “продадени-бройки” (unit_sold).
- Факт таблицата съдържа имената на факти или мерки, както и ключове към всяка от съответните димензионни таблици.

AllElectronics - dimensions

- Всяка размерност може да има асоциираната с нея таблица, наречена димензионна таблица, която описва дадената размерност.
  - Димензионна таблица“артикул” с атрибути “име-на-артикула” (item_name), “вид” (brand) и “тип” (type).

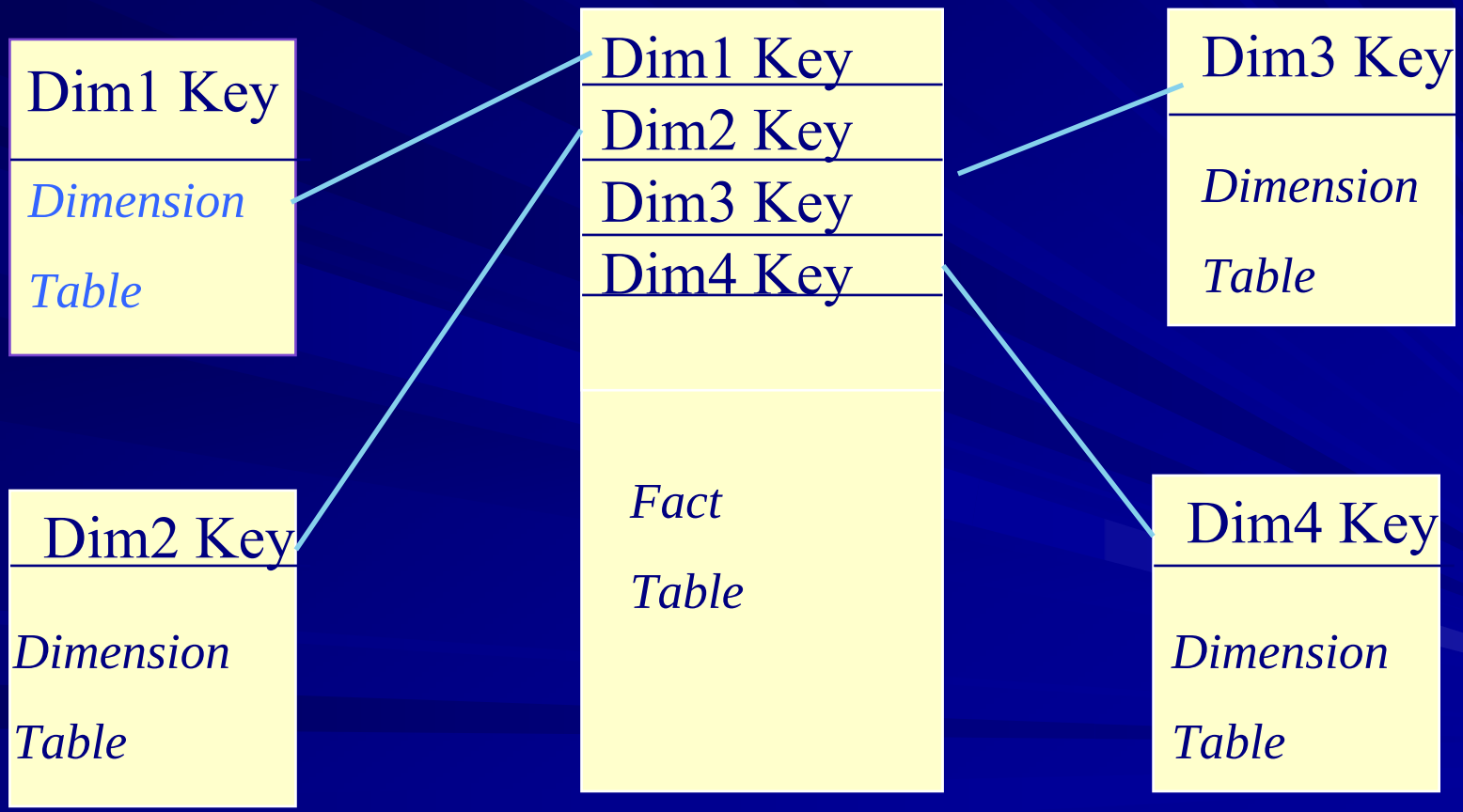
Концептуални модели

- Star schema (схема “звезда”)
- Snowflake schema (схема “снежинка”)
- Fact constellations (схема “съзвездие”)

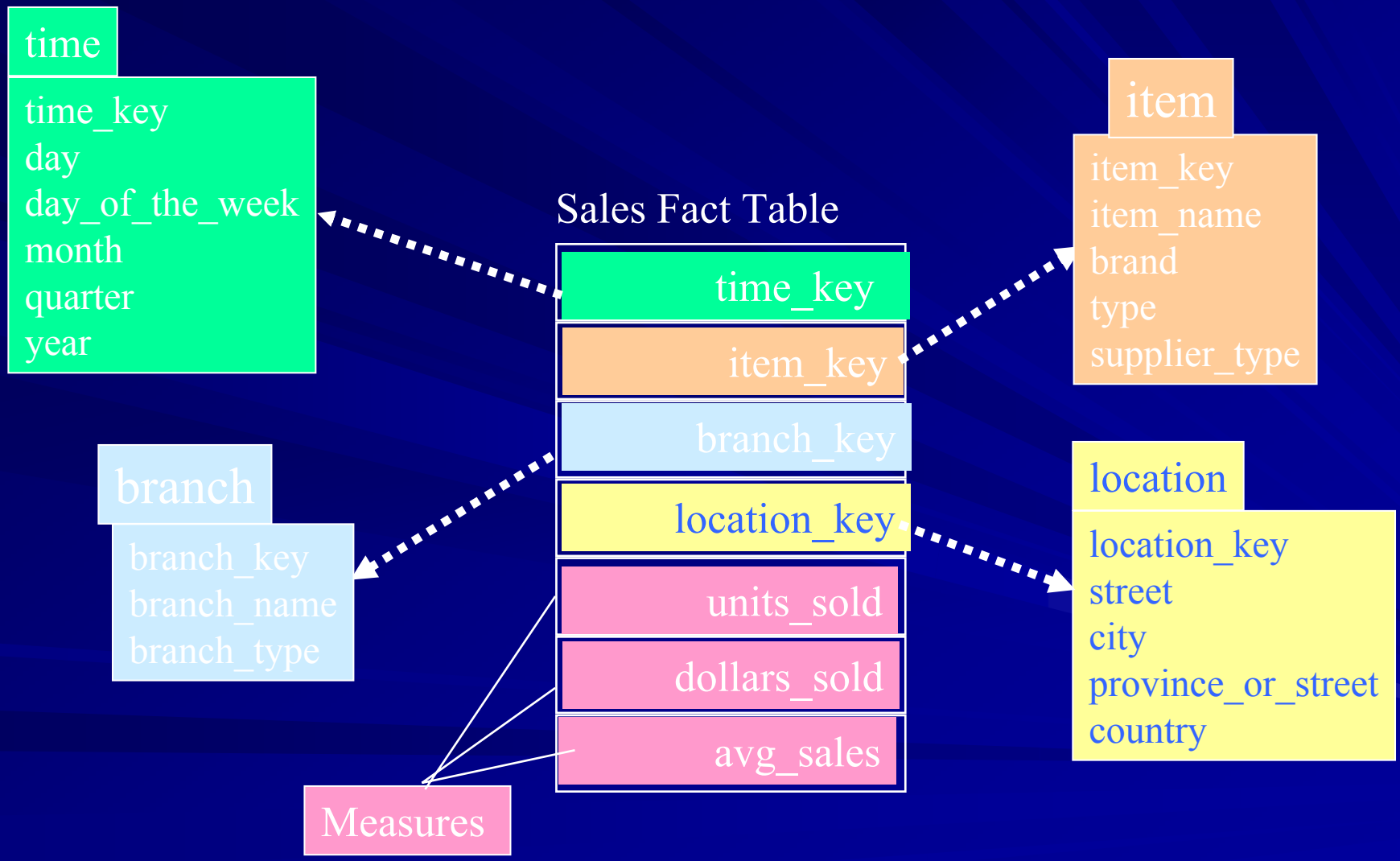
Звезда

- Една централна факт таблица и множество други таблици, разположени радиално около нея
- Свързване по първични и външни ключове
- Денормализиран модел, подходящ за статични БД

Dimensional Data Model
Star Schema



Star Schema


time
time_key
day
day_of_the_week
month
quarter
year
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales
item
item_key
item_name
brand
type
supplier_type
branch
branch_key
branch_name
branch_type
location
location_key
street
city
province_or_street
country
Measures

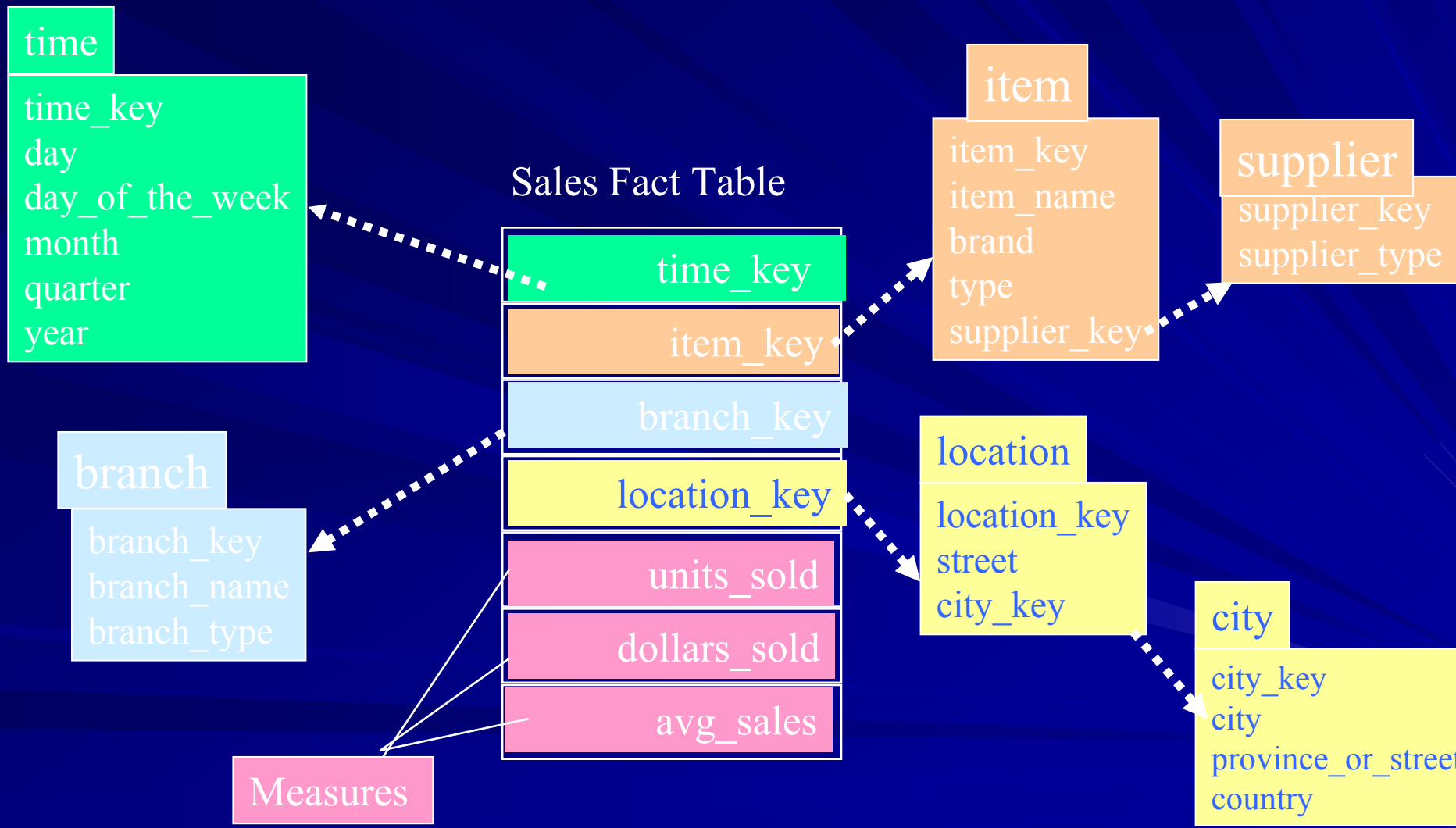
Star Schema – недостатъци

- Изисква често реконфигуриране
- Поради нивата на денормализация конструуирането на модела се извършва бавно и трудно
  - Поддържане на хронология
  - Създаване на йерархии в димeнсиите

Snowflake Schema

- По-близо до класическата ERD, отколкото схемата “звезда”, защото дименсионните данни са по-нормализирани
- Разработката на модела означава създаване на йерархии във всяка от дименсиите (нормализация на данните)

Snowflake Schema



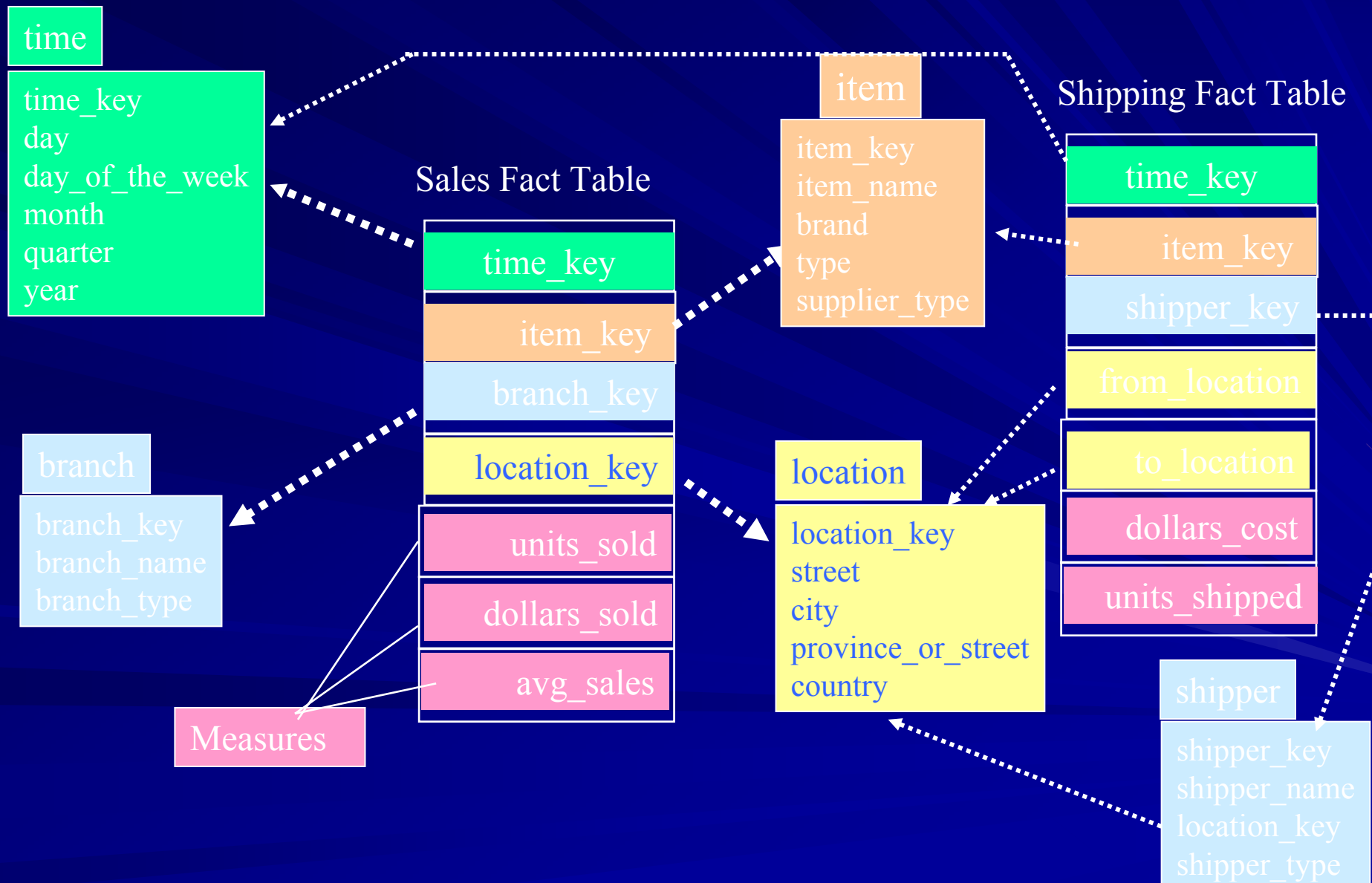
Snowflake Schema - недостатъци

- При много димензии с много нива на йерархия – труден за управление модел
- Повече връзки – затрудняват производителността
- Метаданни – по-сложни

Constellation model

- Constellation model обхваща серия от модели “звезда”
  - При необходимост от няколко факт таблици

Fact Constellation



Факт таблици

- Всеки запис във факт таблицата с-жа първичен ключ – конкатенация от външни ключове (foreign keys) към димeнсионни таблици и факти или мерки, еднозначно идентифицирани от този първичен ключ

Ниво на детайлност

- Съхранение на данните с възможно най-голяма степен на детайлизация
  - Детайлизирани данни → сумарни
  - Невъзможен обратен процес
- Atomic level of detail - най-ниско ниво на детайлизация

Димензионни таблици

- Денормализирани
- “По-широки” от факт таблиците
  - повече колони
- “По-къси” от факт таблиците
  - по-малко редове
- Използват сурогатни ключове (Surrogate Keys)

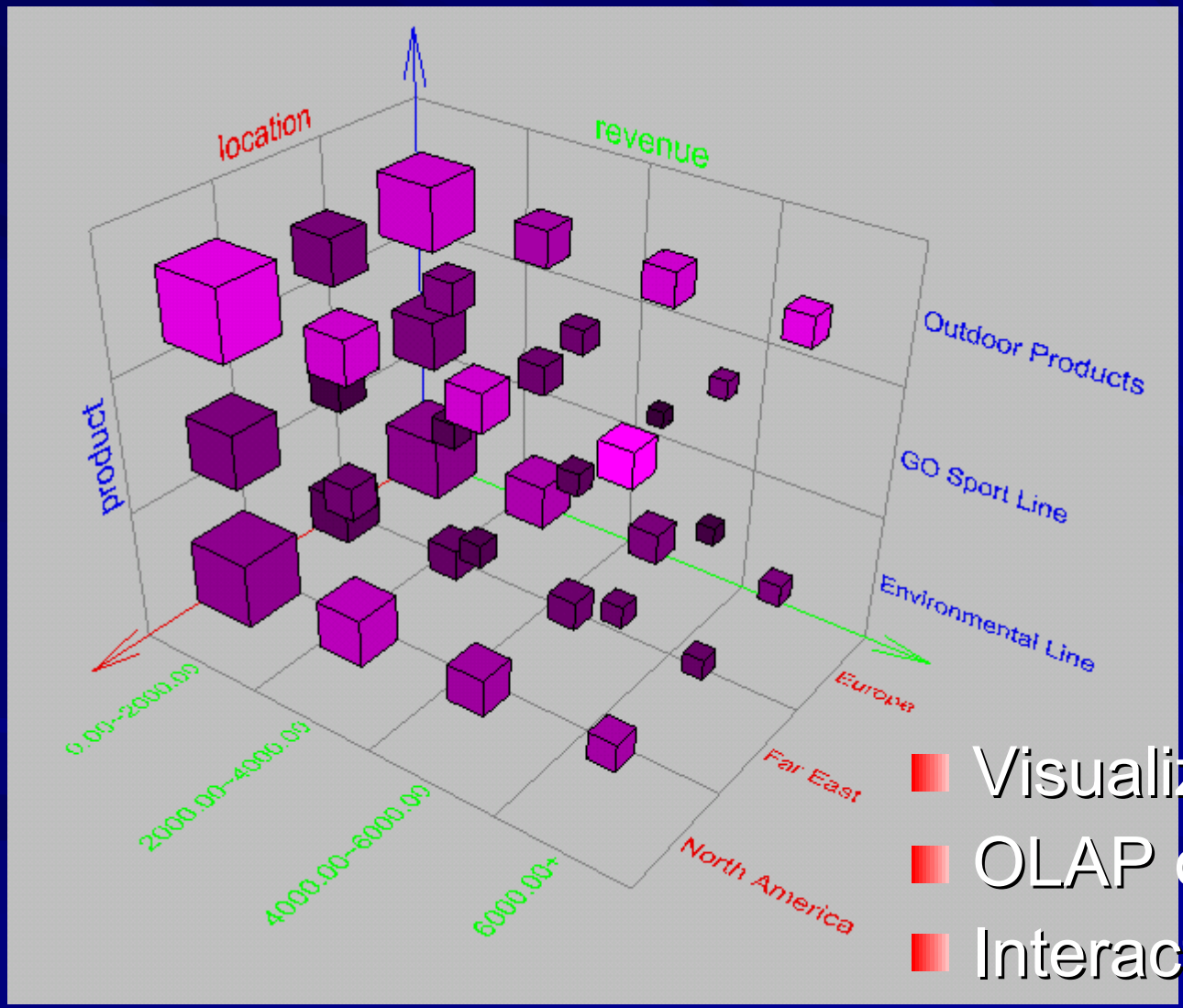
Димензионни таблици - моделиране

- Модел според съдържанието на данните
- Модел според необходимостта от обобщаване
- Удовлетворяване на изискванията на йерархиите – drill-up, drill-down
- Изцяло денормализирана - star
- Нормализирана – snowflake

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice:
  - *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes.*

Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation