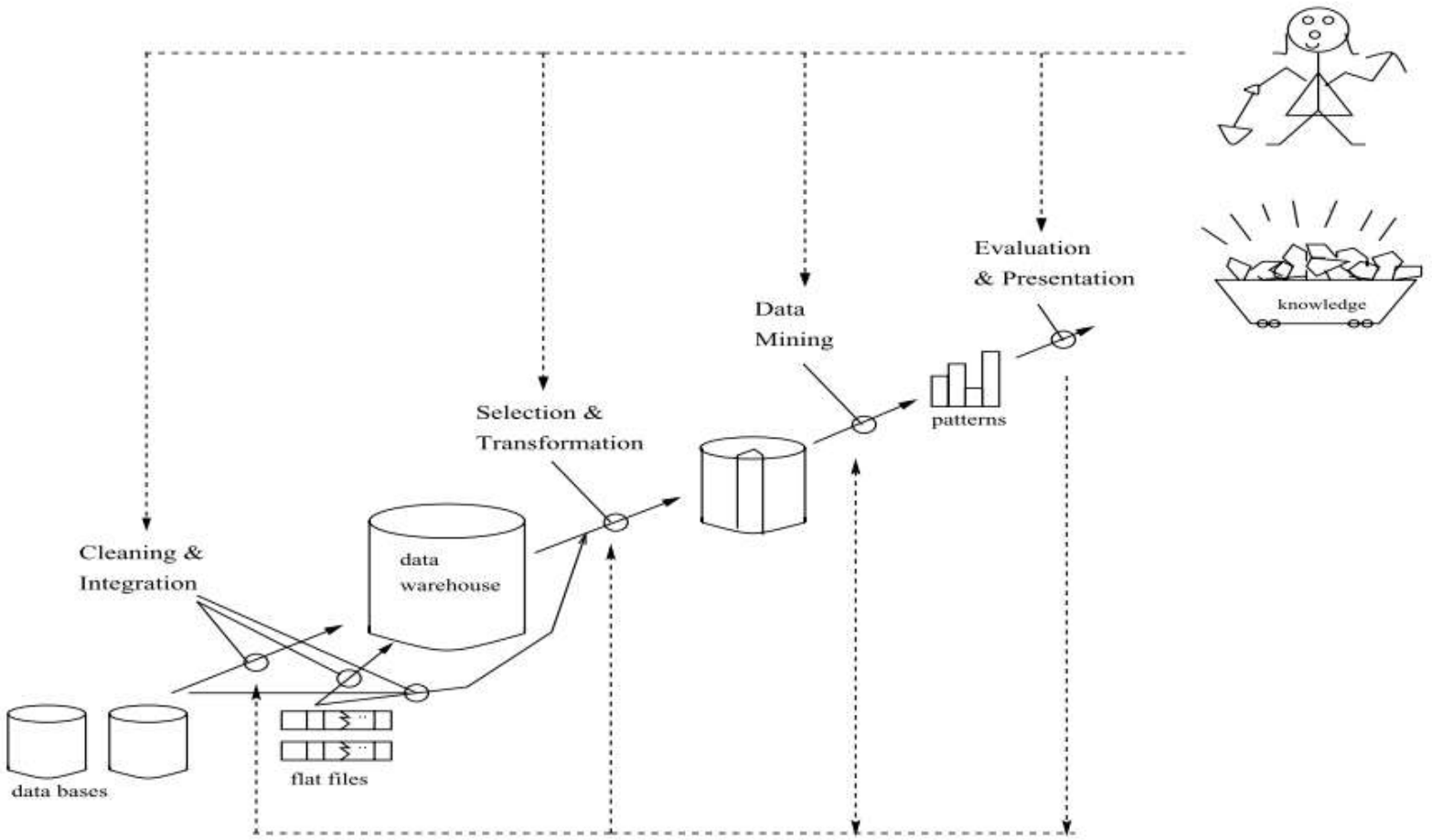
# Подготовка на данните за анализ

Събиране и интегриране

# Процес

- Събиране и подбор на данни
  - Разнородни ресурси с данни
  - Предварително описание и оценка на данните
- Предварителна подготовка на данните
  - изчистване
  - интегриране
  - трансформиране
  - намаляване
- Стандарти за данни и метаданни
- ETL *Extraction, Transformation and Loading*

# Процес


Evaluation
& Presentation
Data
Mining
Selection &
Transformation
Cleaning &
Integration
data
warehouse
patterns
knowledge
data bases
flat files

# Събиране на данни

# Събиране на данни

- Данни от различни източници
  - от дейностите на бизнеса
    - релационни бази данни (таблици с персонал, продажби и др)
    - таблици от MS Excel
    - файлове с данни (текстови документи)
    - и др.
  - от външни източници
    - времето
    - трафика
    - и др.

# Събиране на данни

- Събиране на данни
  - от база от данни DBMS - ODBC, JDBC protocols
  - от файл
    - фиксиран фомат на полета
    - разделител на полетата: tab, comma “,” , други
- Метаданни
  - роля на данните
  - описание на атрибутите
    - напр. тип, допустими стойности, интервал

# Разбиране на данните

- Какви данни са налични?
- Подходящи ли са?
- Има ли и други подходящи данни?
- Колко хронологични данни има?

# Проблеми

- Технологични
- Организационни
- Икономически

# Проблеми

- Технологични
  - разнородни източници с различен формат
  - структурирани, полуструктурирани и неструктурирани данни
  - събиране на данните по различно време
  - огромни количества данни
  - качество на данните
  - изследване на различните формати и трансформиране в един
- Организационни
- Икономически

# Проблеми

- Технологични
- Организационни
  - във връзка с предоставяне на данните
  - във връзка с архитектурата и потребителските качества на информацията
- Икономически

# Проблеми

- Технологични
- Организационни
- Икономически
  - скъп, бавен и тромав процес
  - трудна и нестандартна подготовка на данните
  - различни бизнес-приоритети за различни отдели в бинеса

# Подготовка на данни

# Подготовка на данните

- Необходимост
  - избор  на атрибути за изследване
  - шум в данните
  - липсващи стойност
- Качество на данните
  - пълни
  - точни
  - последователни

# Подготовка на данните

- Брой екземпляри
  - по-малко данни – по-малко надеждни резултати
- Брой атрибути
  - ако има твърде много атрибути – редуциране на броя им
- Брой цели
  - балансирани

# Подготовка на данните

- Процес на обработка на данните, с цел подобряване на качеството им
  - data cleaning
- Интегриране на данните от разнородни източници
  - data integration
- Трансформиране на данните в единна структура
  - data transformation
- Намаляване  на количеството данни
  - data reduction

# Изчистване на данните

- Цели
  - Попълване на липсващи стойности
  - Заглаждане на шума
  - Коригиране на несъответствията
- Техники
  - Унифициране на формата и мерните скали (номинални към числови)
  - Дискретизиране на числовите данни
  - Събиране на данни и метаданни
  - Валидиране и статистика

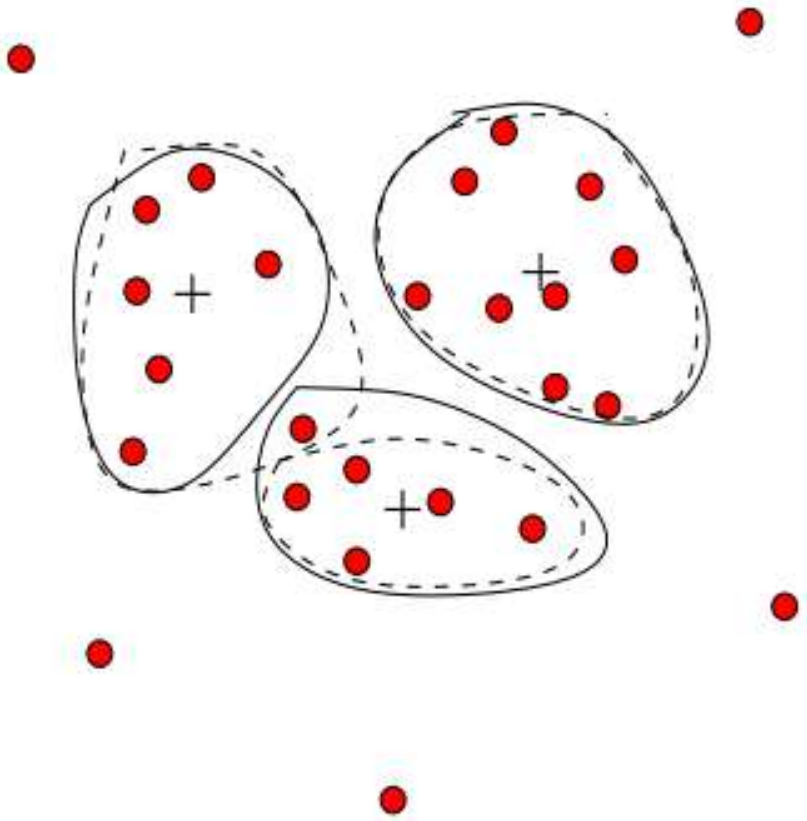
# Липсващи стойности

- Правила
  - да се избягват записите с липсващи стойности
  - празните полета да се разглеждат самостоятелно
  - празните полета да се попълват със стойности

# Попълване на липсващи стойности

- Отстраняване на записите с липсващи стойности
- Ръчно попълване
- Посредством глобални константи, като “-∞” “0” “.” “999” “NA”
- Средната стойност на атрибута
- Средната стойност от всички извадки
- Най-вероятната стойност на атрибута

# Шум



# Шум

- Случайна грешка или дисперсия на измерваната величина
- Изглаждане на шума
  - binning- разпределяне на стойностите в групи, заместване на сгрешените като функция на съседните си: средни, мода, медиана, граници
  - клъстеризация - разпределяне на стойностите в групи и отделяне на тези, които не попадат в никоя група
  - чрез регресия – опит да се намери функция, която описва данните
  - интегриране на човек и програма за анализ

# Пример за изглаждане

- Редица от данни, сортирани:
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
- Групи:
  - 4, 8, 15
  - 21, 21, 24
  - 25, 28, 34
- Изглаждане
  - заместване със средното за групата
    - 9, 9, 9
    - 22, 22, 22
    - 29, 29, 29
  - заместване с минимални и максимални стойности
    - 4, 4, 15
    - 21, 21, 24
    - 25, 25, 34

# Преформатиране

Преобразуване към стандартен формат

- липсващи стойности
- локализация
  - формат за дати
- конвертиране на номиналните стойности в числови, за да участват в логически операции

# Унифициран формат

- Дати
  - ден/месец/година
  - месец/ден/година
  - пореден ден от годината

$$\text{Date} = \text{YYYY} + \frac{\text{days\_starting\_Jan\_1} - 0.5}{365 + 1\_\text{if\_leap\_year}}$$

# Преобразуване на тип

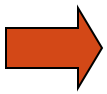
- Бинарни към номинални
  - полета с две различни стойности - бинарни
    - Gender=M, F
  - Кодиране на поле със стойност и 0, 1
    - Gender = M → Gender_0_1 = 0
    - Gender = F → Gender_0_1 = 1

- номинални атрибути към естествени, напр.
  - A → 4.0
  - A- → 3.7
  - B+ → 3.3
  - B → 3.0

# Преобразуване на тип

- множество неподредени атрибути с малко стойности
- правило <20
- цветовете също се прекодират
  - за всеки цвят – една бинарна стойност

| ID | Color | … |
|---|---|---|
| 371 | red | |
| 433 | yellow | |



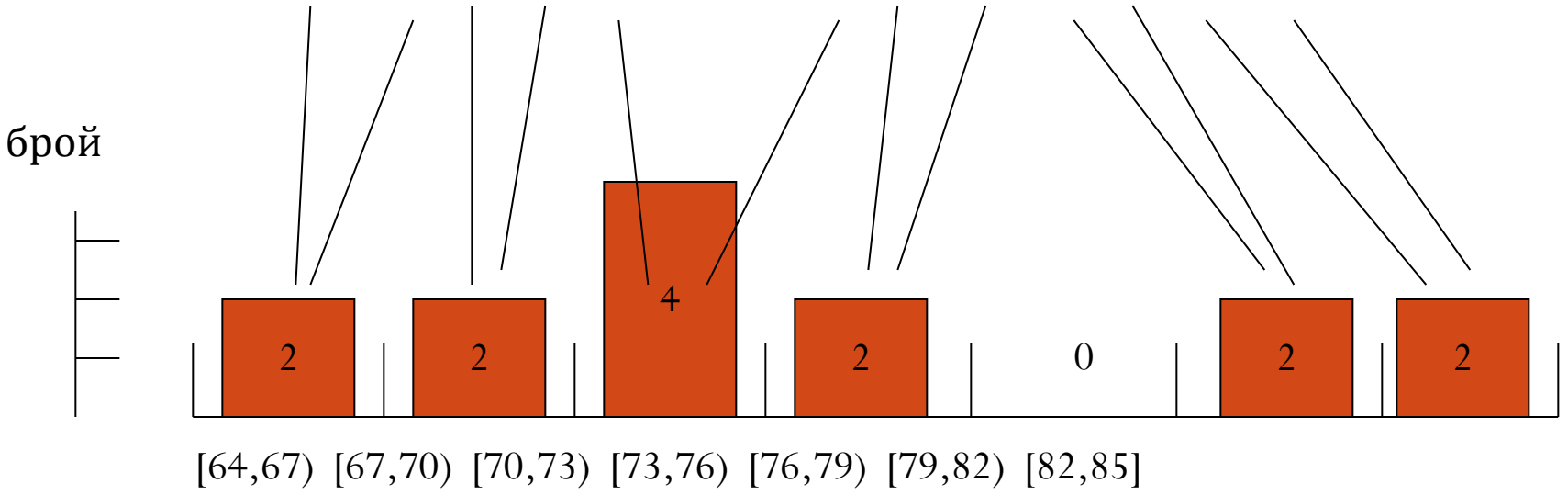
| ID | C_red | C_orange | C_yellow | … |
|---|---|---|---|---|
| 371 | 1 | 0 | 0 | |
| 433 | 0 | 0 | 1 | |

# Дискретизиране

## Дискрети с еднаква ширина

Пример: измерени числови стойности

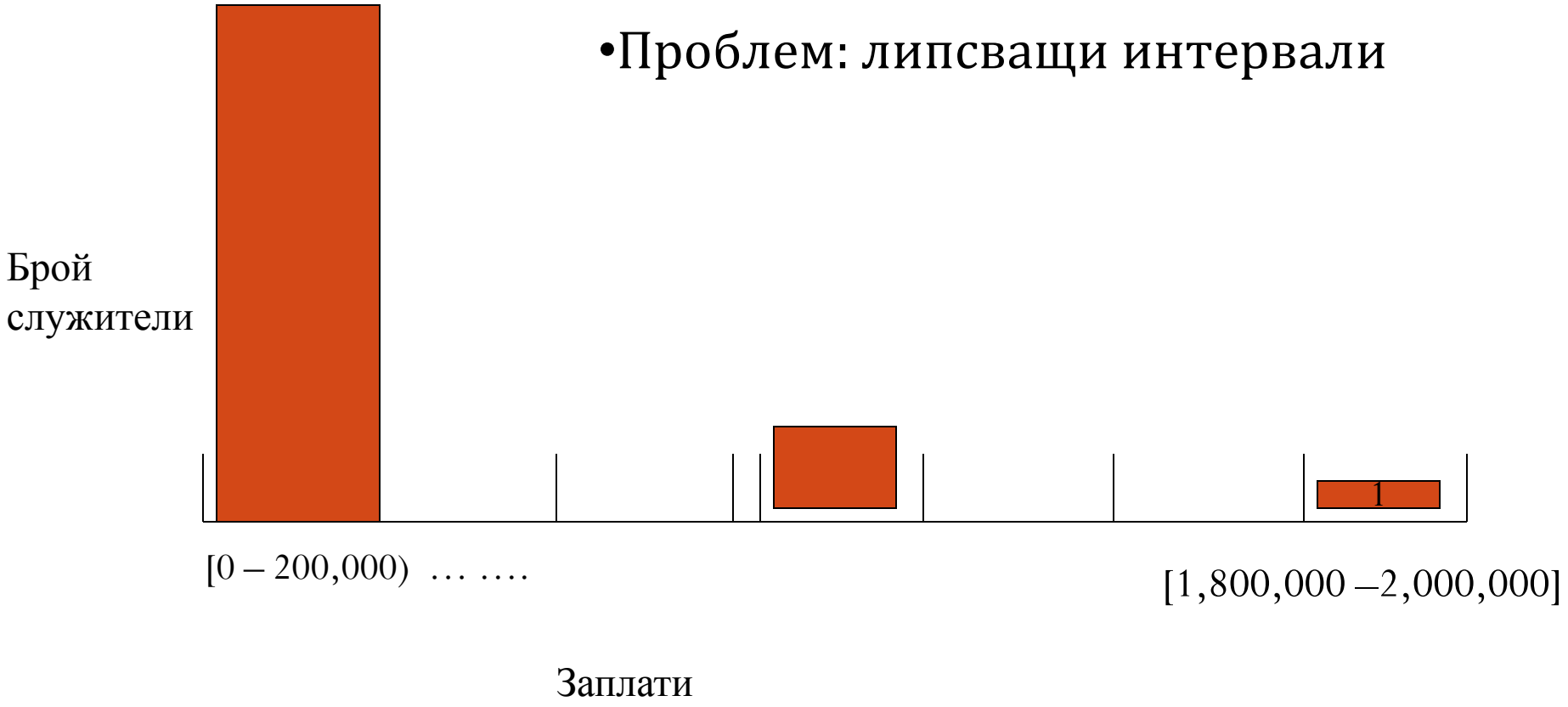
64 65 68 69 70 71 72 72 75 75 80 81 83 85




min <= стойност < max

# Дискретизиране

•Проблем: липсващи интервали

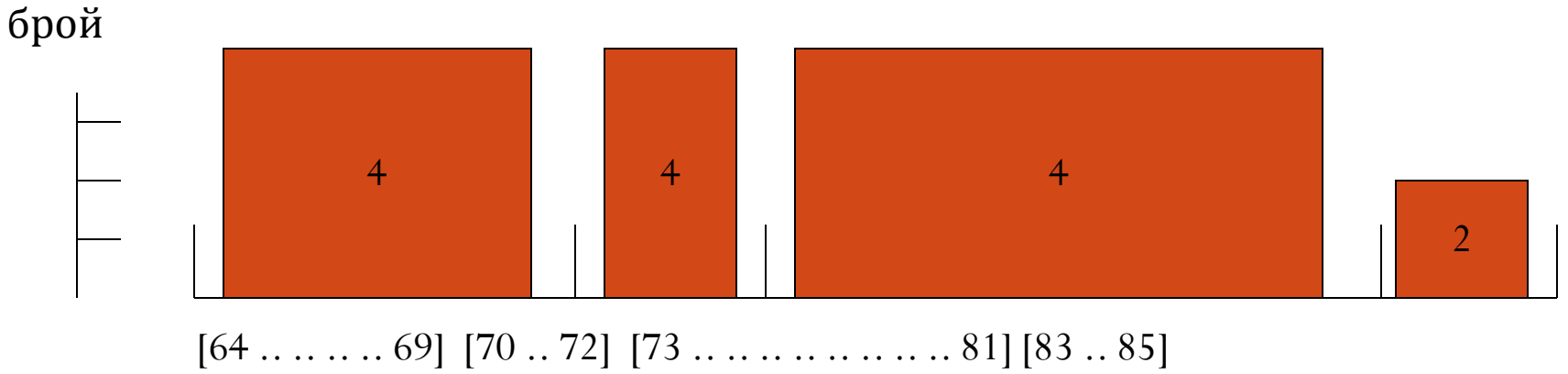


# Дискретизиране

Дискрети с еднаква височина, т.е. еднакъв брой стойности в групата

Измерени стойности:

64 65 68 69 70 71 72 72 75 75 80 81 83 85




- последната група (bin) може да бъде с по-малка височина
- предпочитана форма на дискретизация

# Крайни стойности и грешки

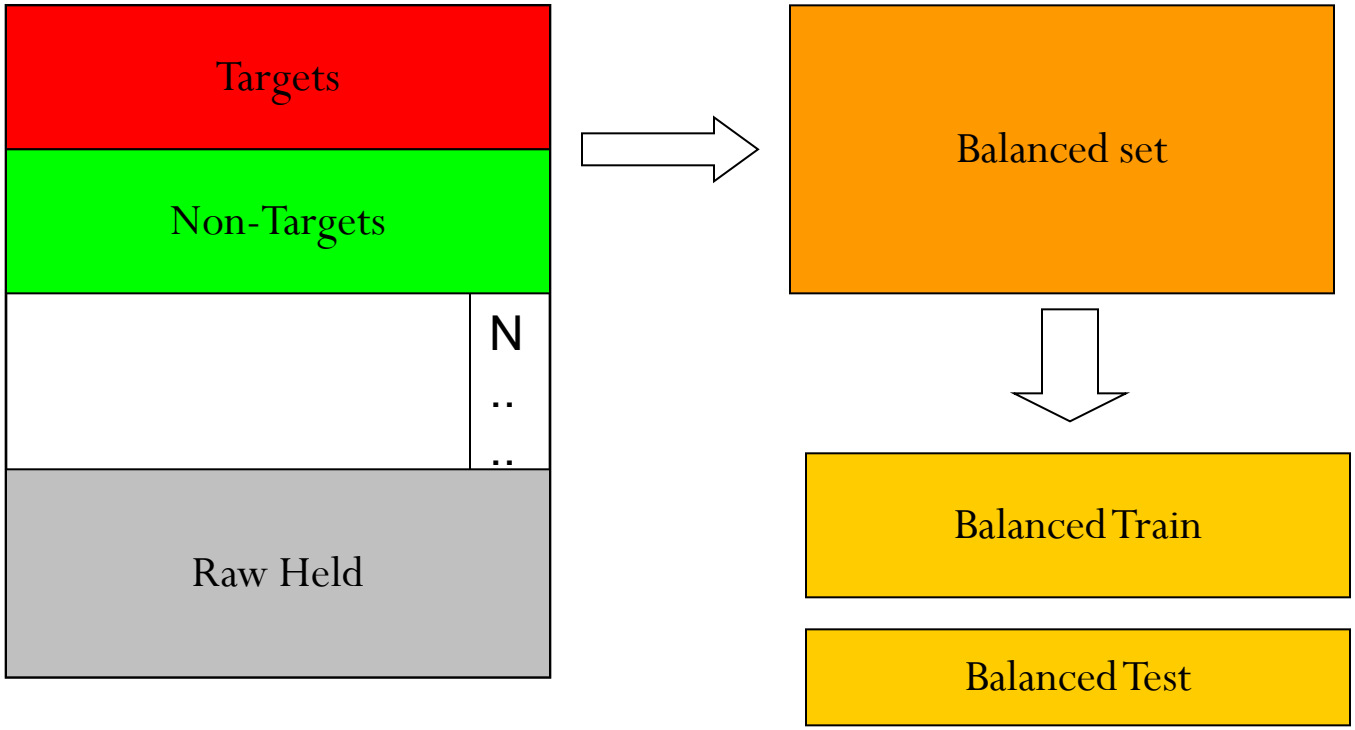
- Крайни стойности – извън разглеждания интервал
- Подходи
  - да не се променя нищо
  - да се разтегнат границите на интервала
  - да се разчита на дискретизирането

# Твърде много стойности

- Игнорират се полетата с ID, защото те съдържат уникален код,
  - останалите се прегрупират в по-компактни групи
- Създават се бинарни полета, които се използват като флаг
- Отстраняват се монотонните стойности
- Броят се само различните стойности
  - `minp` %:

    0.5% - 5% от брой на записите в резултата

# Балансиране на извадките

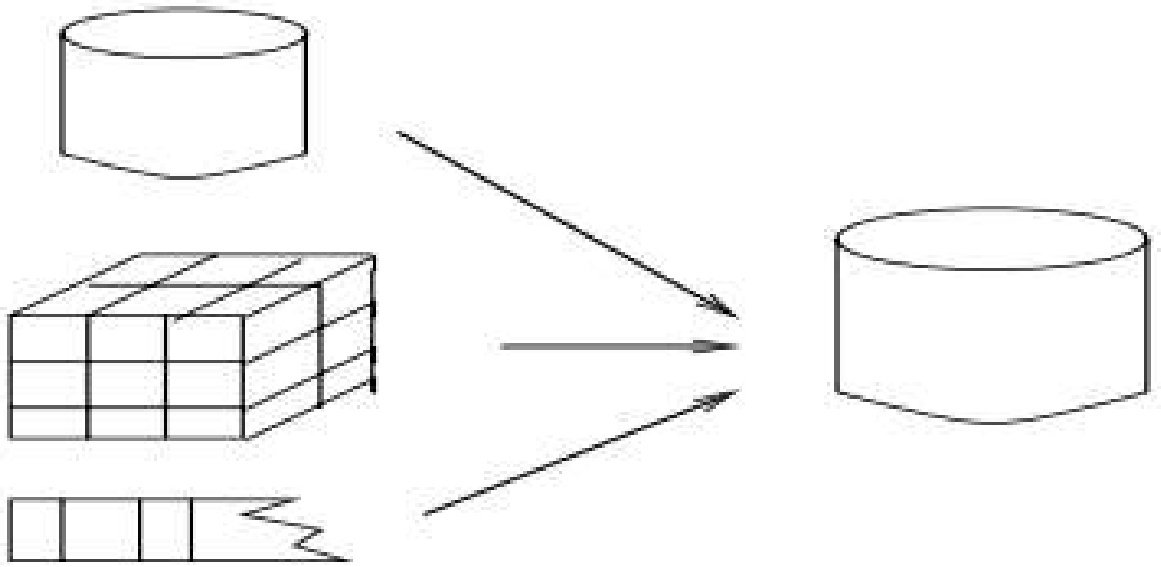


# Интегриране на данни

# Интегриране на данни

- Комбиниране на данни от различни източници, за формиране на единен ресурс
- Методи
  - метаданни
  - корелационен анализ
  - откриване на конфликти
  - отстраняване на семантична хетерогенност
  - трансформация

# Интегриране на данни



# Интегриране на данни

- Потоци на данни
- Потоци на задачи
- Архитектура
  - свързване с различни ресурси на данни по комуникационни канали и адаптери
  - преобразуване на всички типове входни данни в таблици
  - зареждане на данните  от различни източници в буфери на трансформационни канали
  - look-up – ключове, необходими за трансформирането в таблици

# Проблеми

- Несъответствие на схемите на данни
  - напр. userID == EGN
- Излишък
  - напр. един атрибут от една таблица може да е производен в друга
  - дублиране на данни
- Конфликти между стойностите от различни измервания
  - напр. в мерните единици, валути, % и др.
- Решаване на проблемите -> Трансформация

# Техники на интегрирането

- Агрегиране
- Сортиране
- Таблици Lookup
- Pivot и UnPivot
- Смесване на данните – merging
- Сливане - concatenation
- Derived Column – операции , генериращи нови колони
- Конвертиране на данните
- Audit – добавяне на колони с метаданни

# Трансформация на данни

# Трансформация

- Преобразуване на данните в подходящи формати
- Методи
  - изглаждане
    - напр. заместване със статистически стойности
  - агрегиране
    - напр. заместване на дневни продажби с месечни
  - генерализиране – обобщаване в групи с по-висока грануларност
    - напр. адрес: улица, номер –> град, държава
  - нормализация – преизисляване в ограничен интервал
    - напр. 0,0 – 1,0 или -1.0 – 1.0
  - конструиране на нови атрибути

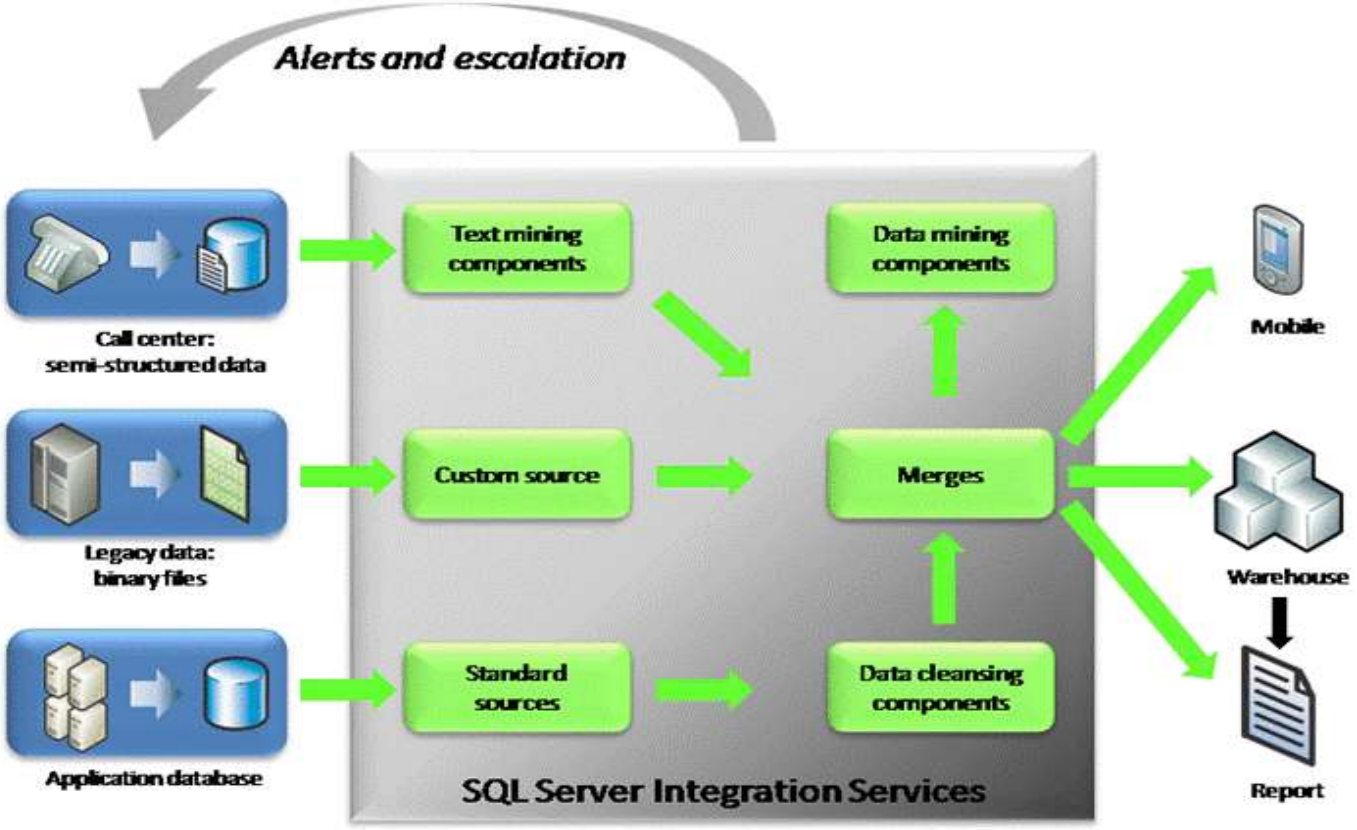
# Редуциране на данните

- По-малък обем данни с по-висок интегритет
- Извеждане на резултати със сходно качество
- Стратегии
  - куб от данни
  - отстраняване на атрибути с несъществено значение
  - намалчване на броя на наблюденията
  - разделяна в групи -клъстеризация, хистограми
  - интегриране в по-високи нива

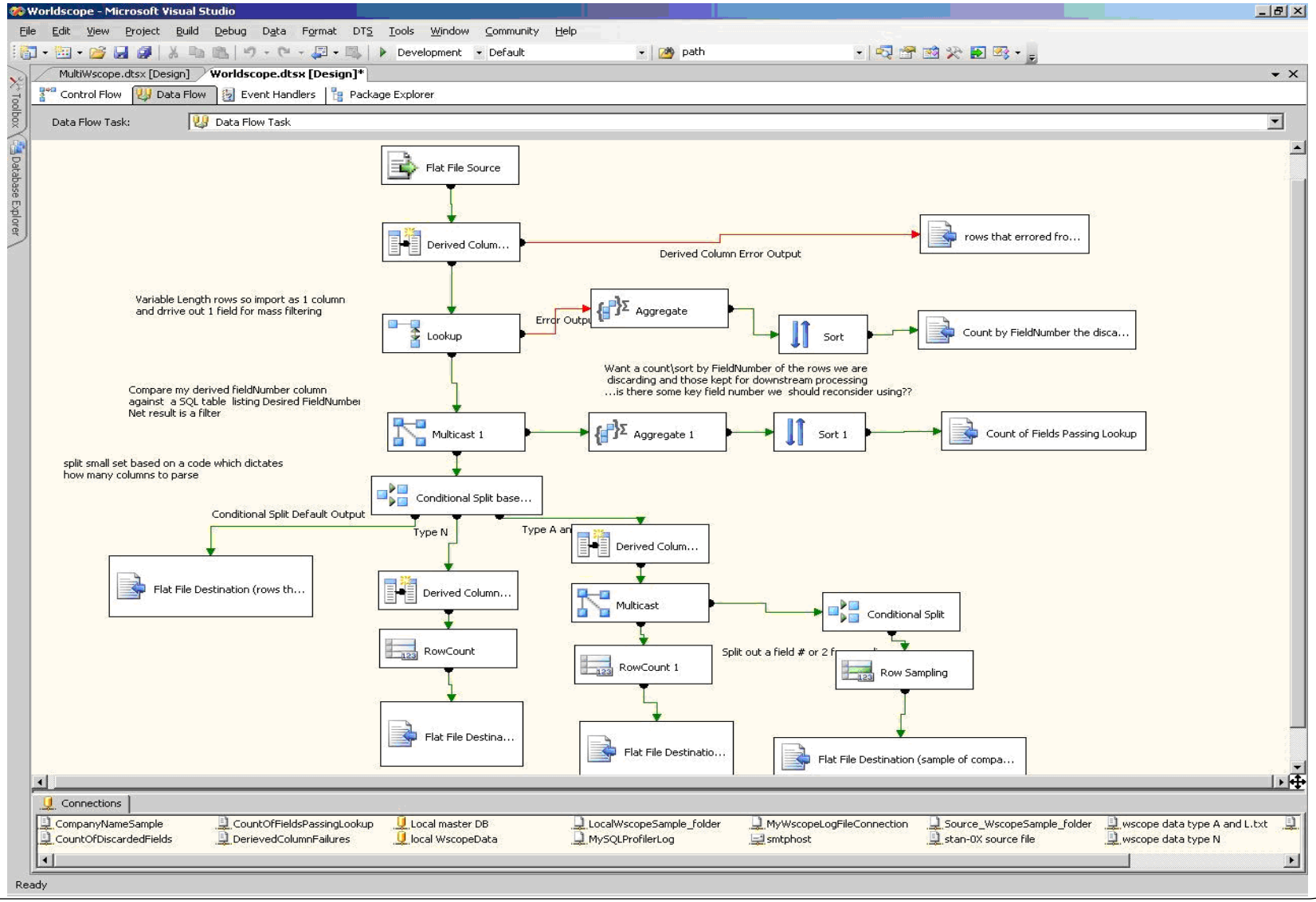
# Програмни средства за интегриране на данни

## ETL Tools

Extract Transform Load



# Пример


Worldscope - Microsoft Visual Studio
File Edit View Project Build Debug Data Format DTS Tools Window Community Help
Development
Default
path
MultiWscope.dtsx [Design]
Worldscope.dtsx [Design]*
Control Flow
Data Flow
Event Handlers
Package Explorer
Data Flow Task:
Data Flow Task
Flat File Source
Derived Colum...
rows that errored fro...
Derived Column Error Output
Variable Length rows so import as 1 column
and drrive out 1 field for mass filtering
Lookup
Aggregate
Sort
Count by FieldNumber the disca...
Want a count\sort by FieldNumber of the rows we are
discarding and those kept for downstream processing
...is there some key field number we should reconsider using??
Compare my derived fieldNumber column
against a SQL table listing Desired FieldNumbe
Net result is a filter
Multicast 1
Aggregate 1
Sort 1
Count of Fields Passing Lookup
split small set based on a code which dictates
how many columns to parse
Conditional Split base...
Conditional Split Default Output
Type N
Type A an
Flat File Destination (rows th...
Derived Column...
Derived Colum...
Multicast
Conditional Split
RowCount
RowCount 1
Split out a field # or 2 f
Row Sampling
Flat File Destina...
Flat File Destinatio...
Flat File Destination (sample of compa...
Connections
CompanyNameSample
CountOfDiscardedFields
CountOfFieldsPassingLookup
DerievedColumnFailures
Local master DB
local WscopeData
LocalWscopeSample_folder
MySQLProfilerLog
MyWscopeLogFileConnection
smtphost
Source_WscopeSample_folder
stan-0X source file
wscope data type A and L.txt
wscope data type N
Ready