# Профайлинг данных (Data Profiling)

> Синонимы: Профилирование данных
> Разделы: Бизнес-задачи

Профайлинг данных — один из наиболее распространенных методов проверки качества данных и выявления проблем в Data Mining. Профайлинг выполняется автоматически в соответствии с некоторым заранее настроенным сценарием на основе анализа информации о структуре данных.

В процессе профайлинга проверяются поля источника данных на соответствие заданным ограничениям. Если параметры полей удовлетворяют ограничениям, то данные считаются соответствующими требуемому уровню качества, в противном случае необходимо принимать меры к приведению параметров к соответствующим ограничениям.

При профайлинге может проверяться тип поля, длина его значений и их допустимый диапазон, производиться анализ шаблонов. Если в процессе проверки обнаруживаются нарушения, они исправляются в соответствии с заданным сценарием.

Например, типичной проблемой при вводе числовых значений является неправильное использование разделителей целой и дробной частей числа и групп разрядов. В качестве разделителя целой и дробной части может оказаться запятая или точка, а разделение групп разрядов может производиться с помощью пробела или вообще отсутствовать.

Получается, что одно и то же число может быть записано несколькими разными способами: 1 500 000.00, 1500000,00 или 1500000.00. При некорректном использовании разделителей значение вообще может быть интерпретировано системой как строковое, что приведет к ошибочным выводам. Поэтому одной из задач профайлинга является проверка форматов представления чисел и приведение их в соответствие с принятыми в данной системе.