Очистка данных (Data Cleaning)

> Разделы: Бизнес-задачи
> Решения: Loginom Data Quality

Процесс исключения из данных различных факторов, снижающих их качество и мешающих их корректному анализу. Очистка данных является важнейшим этапом аналитического процесса, и от того, насколько эффективно она произведена, во многом зависит корректность результатов анализа и точность построенных аналитических моделей.

Наиболее критичными факторами, снижающими качество данных и требующими применения очистки, являются:

- противоречивость;
- пропущенные значения;
- дубликаты;
- выбросы и аномальные значения;
- шум;
- ошибки ввода данных.

Очистка данных производится как перед их загрузкой в хранилище (т.е. в процессе ETL), так и в аналитическом приложении непосредственно перед анализом. При этом основная очистка производится в аналитическом приложении, поскольку некоторые проблемы, например, дубликаты и противоречия, невозможно выявить до завершения консолидации данных.

Кроме этого, требования к качеству данных могут быть различными для различных методов и алгоритмов анализа. Поэтому большинство аналитических приложений содержит развитый комплекс средств очистки данных.