Интеграция данных (Data Integration)

> Синонимы: Объединение данных
> Разделы: Бизнес-задачи
> Решения: Loginom Data Quality

В широком смысле интеграцией называют процесс объединения, вставки различных частей чего-либо. Например, в технике производится интеграция нескольких устройств в одну сложную техническую систему, в программной инженерии — интеграция программных модулей в одну систему, и т.д.

В аналитических технологиях под интеграцией в большинстве случаев подразумевают интеграцию данных из различных источников в один набор, в котором они хранятся в унифицированном формате и структуре. Впоследствии интегрированный набор данных полностью или частично может быть загружен в аналитическую платформу для применения к нему различных методов анализа.

В анализе интеграция является важным процессом, поскольку приходится иметь дело с очень большими объемами информации, расположенными в источниках, имеющих самые разнообразные представления, форматы и кодировки данных. Кроме этого, в данных могут быть нарушения структуры, полноты и целостности, что требует выполнения специальной предобработки данных.

В современных условиях задача интеграции обычно решается с помощью хранилищ данных и ETL-процессов.

Выделяют три уровня интеграции данных:

1. Физический — производится преобразование данных из различных форматов и типов к единому физическому представлению. Это особенно важно для анализа данных, поскольку только их унифицированное представление гарантирует единообразную обработку различными алгоритмами и моделями. Единый формат дает возможность корректно интерпретировать и сравнивать результаты анализа данных из различных источников.
2. Логический — организуется процесс работы с данными таким образом, как будто они находятся в едином источнике, в соответствии с некоторой схемой их описания. При этом используется унифицированный интерфейс работы с данными.
3. Семантический — данные объединяются не на основе физической или логической модели, а на основе отношений между сущностями (объектами, процессами), которые они описывают.

Существуют следующие архитектуры систем интеграции:

1. Консолидация — данные физически извлекаются из источников и объединяются в централизованном хранилище данных.
2. Федерализация — данные не консолидируются физически, а хранятся в своих источниках и становятся доступными только при выполнении соответствующего запроса.
3. Распространение — данные физически копируются из одного места в другое, пока не попадут в некоторую целевую систему.

Совместно с интеграцией данных необходимо решать задачу их очистки. Например, если в отдельных источниках данных записи являются уникальными и непротиворечивыми, то после интеграции они могут стать дубликатами и противоречиями.