В. С. Шмидт, Д. С. Кузьменков
(ГГУ им. Ф. Скорины, Гомель)

РЕАЛИЗАЦИЯ НОРМАЛИЗАЦИИ ДАННЫХ В SAP BODS

SAP BODS (BusinessObjectsDataServices) – решение для интеграции [1], обеспечения качества, профилирования данных и анализа текста. Продукт предоставляет возможности интеграции, трансформации, улучшения и доставки данных, имеет единые интерфейс разработки, репозиторий метаданных, слой подключения, среду выполнения и панель управления. SAP BOIS (InformationSteward) – полномасштабный компонент SAP BODS, который позволяет на уровне настраиваемых бизнес-правил проводить первичный анализ качества данных, распознавать, проверять, стандартизировать и совершать очистку любых данных; выявлять дубликаты и взаимосвязи между такими данными, как, например, имена, адреса, названия, реквизиты и др., существующими в различных системах предприятия.

Сегодня наблюдается большой интерес к технологиям класса BIG DATA, связанный с постоянным ростом данных, которыми приходится оперировать крупным компаниям. Накопленная информация для многих организаций является важным активом, однако обрабатывать ее и извлекать из нее пользу с каждым днем становится все сложнее и дороже. Более того, при больших объёмах данных всё больше вероятность ошибочных данных, дубликатов либо неполных записей.

Для нормализации был выбран пакет с данными об электродвигателях, содержащий около 2500 записей. Начальное профилирование этих данных показало, что около 33% данных с более чем 90 процентной вероятностью являются дубликатами, 40% записей содержат неполную информацию, и 8% данных содержат противоречивы. В BOIS был разработан набор атрибутов на основе российского стандарта электродвигателей серии АИР ГОСТ Р 51689-2000.

В среде BODS создавалось подключение к системе SAP R/3, из таблиц которой выбирались данные. Создан процесс, содержащий WorkFlow с двумя задачами, в котором реализована следующая логика: выполняется select запрос для выборки данных; данные проходят процесс очистки и нормализации с использованием созданного в BOIS пакета очистки; обработанные данные проходят процесс выявления дубликатов с вероятностью в 90%; набор записей проходит последний процесс обогащения записей; нормализованные, дедуплицированные, обогащённые данные загружаются назад в систему R/3. Профилирование загруженных данных: из 33% дублирующих записей в результирующем наборе осталось всего 3%, 10% содержат неполную информацию, противоречивых данных не оказалось.

ЛИТЕРАТУРА

1 Chen, B. SapDataServices Теория и практика построения баз данных / B. Chen, J. Hanck, S. Hertel. – Quincy.: Rheinwerk-PublishingInc., 2015. – 524 с.