УДК 004.6

РАЗРАБОТКА АРХИТЕКТУРЫ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ДАННЫХ В СИСТЕМАХ BUSINESS INTELLIGENCE

Гец О.В. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – к. техн. н. Иванов С.В.

(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В работе рассмотрена проблема верификации больших данных нефтяной отрасли при интеграции с различными системами или отчетами пользователей. Описано решение производить проверку данных в едином хранилище, а также представлена методика создания проверок качества данных для их автоматизированной корректировки при отсутствии эталона.

В любой отрасли эксплуатируется большое число информационных систем различного назначения, при этом методика расчетов, алгоритмов тех или иных показателей может отличаться между ними. Таким образом, возникают противоречивые и избыточные данные, не связанные между собой, на основе которых могут быть приняты неверные управленческие решения. При этом, если в отрасли хранятся огромные объемы данных, подлежащих анализу, их поиск, верификация и корректировка расчетов занимает критическое количество времени для сотрудников организации, что снижает эффективность производства в целом.

В результате проведенного исследования нами была произведена интеграция данных из нескольких систем в компании нефтяной отрасли. На этапе анализа одной из них были выявлены расхождения данных в добыче нефти при агрегации в нескольких разрезах. Например, при сравнении данных за сутки в течение месяца и итоговой цифры за месяц, выводимых в двух разных отчетах, было найдено расхождение, которое в дальнейшем отображалось на рабочих экранах руководителей организаций и дочерних обществ. Отклонение этой цифры в ту или иную сторону может приводить не только к неудовлетворенности пользователями качества данных в системе, но и к вынужденным рискам, таким, как необоснованная трата денежных и трудовых ресурсов, возникновение кризиса на производстве.

Интеграция данных из систем-источников производится в едином месте, хранилище данных. В качестве брокера сообщений нами был использован Apache Kafka, в качестве инструмента создания промежуточной базы данных – RocksDB.

Функциональность корпоративного хранилища данных была расширена системой верификации данных из различных систем или отчетов. При наличии расхождения данных из различных систем они не могут поступать на использование лицами, принимающими решения. Расхождения обрабатываются посредством внутреннего экспертного анализа.

Для обеспечения автоматизированного определения источника с ложными данными была произведена предварительная работа по созданию правил проверки качества данных. Например, одним из самых тривиальных правил может служить такое утверждение, что «показатель А не может быть меньше 0» или «на одну дату не может приходиться два разных значения». Таким образом, можно устранить значительное количество ошибок и ошибочные данные не смогут быть отражены на рабочих экранах руководителей.

Основной задачей в данном случае являлось определение эталонного значения. Поскольку каждый из рассмотренных алгоритмов расчета был верен, выявить, на каком именно произошла ошибка, без привлечения экспертов на этапах ввода интеграции не представлялось возможным. Однако более усложненные правила проверки помогли достичь автоматизации сверки и корректности расчетов и освободить сотрудников организации от монотонной работы.

В качестве дополнительного расширения функционала хранилища данных в дальнейшем рассматривается интеграция с мастер-системами нормативно-справочной информации и

систем автоматизированной поддержки принятия решений. Подобное решение требует интеграций с экспертными системами и базами знаний, накопленными организацией в ретроспективном анализе.

На этапе обработки данных в системах Business Intelligence также крайне важной является моделирование различных систем в едином инструменте: например, организационное деление предприятия может находиться на разных уровнях иерархии в различных системах, что является потенциальным источником возникновения расхождений, которое должно быть определено в системе верификации данных.

Рассматривая результаты проведенного исследования, можно выделить, что использование систем BI позволит на каждом этапе ETL-процесса контролировать характер извлекаемых данных, использовать различные сценарии их обработки в зависимости от типа системы-источника, проводить анализ накопленной информации и строить прогноз изменения ключевых показателей. Создание архитектуры, позволяющей хранить, обрабатывать и отдавать на использование системам Business Intelligence объемы больших данных, решает проблему автоматизированных и пользовательских проверок качества данных с минимальным уровнем неудобств, связанных с отсутствием технической подготовки и знаний языков запросов.

Гец О.В. (автор) Подпись

Иванов С.В. (научный руководитель) Подпись