Операционализированное управление данными (Data operations)

> Синонимы: DataOps

DataOps — это современный, междисциплинарный подход к управлению данными, объединяющий операционные принципы разработки ПО (DevOps) и технологии анализа данных. Направлен на автоматизацию и оптимизацию процессов сбора, обработки, хранения и аналитической обработки данных, объединяя их в единый конвейер.

Известная исследовательская и консалтинговая американская компания Gartner определяет DataOps как «практику совместного управления данными, направленную на улучшение коммуникации, интеграции и автоматизации потоков данных между командой управления данными и их потребителями в рамках организации». Целью DataOps является оптимизация жизненного цикла данных в аналитических проектах с точки зрения скорости и качества их реализации.

Традиционный подход к управлению данными в аналитических проектах предполагает, что дата-инженеры, аналитики данных и бизнес-команда работают изолированно, используя разные инструменты и среды. В итоге путь данных от источника до отчета может занимать недели и даже месяцы, что в условиях современной динамической бизнес-среды неприемлемо.

DataOps позволяет решить эту проблему, внедряя принципы непрерывной интеграции и доставки (CI/CD). В результате команды могут развертывать новые аналитические модели и витрины данных практически в реальном времени, что дает возможность сместить фокус с обычного хранения данных на их активное использование для генерации бизнес-ценности.

Основными принципами DataOps являются:

- гибкость (Data Agility) — способность эффективно поставлять и потреблять данные, что позволяет компании быстро реагировать на изменения;
- управляемость — возможность осуществления мониторинга и контроля рабочих процессов управления данными, обеспечение безопасности;
- оркестровка — реализация конвейеров данных, контейнеризации, автоматического масштабирования и восстановления, балансировки нагрузки.

Ключевой особенностью DataOps является наличие обратной связи: если на этапе мониторинга обнаружены проблемы (например, поврежденные данные от их поставщика), система автоматически формирует соответствующее оповещение или останавливает конвейер.

DataOps является фундаментом для MLOps. Критическая взаимосвязь этих технологий выражается в следующем:

- **Обеспечение качества данных для моделей машинного обучения**. DataOps обеспечивает качество данных, на которых обучаются модели, беря на себя процессы их очистки. Если конвейер данных «ломается», модели MLOps начинают выдавать некорректные результаты.
- **Централизованное управление признаками**. Создание магазинов признаков (Feature store), которые позволяют автоматизировать процессы повторного использования признаков в разных проектах, производить извлечение признаков, обеспечивать их согласованность.
- **Версионирование**. В MLOps нужно версионировать не только код ML-модели, но и данные, на которых она обучалась. Инструменты DataOps предоставляют эту возможность, связывая конкретный коммит кода с конкретным «снимком» данных.

Таким образом, конвейеры DataOps отвечают за доставку «сырых» данных, одновременно повышая их качество, а MLOps — за работу аналитических моделей. Попытки внедрить MLOps без налаженного DataOps часто оказываются неэффективными, поскольку команды аналитических проектов вынуждены тратить 80% времени на доставку и очистку данных вручную для улучшения работы ML-моделей.

Более подробно с технологиями DataOps можно ознакомиться в статье «DataOps: современная технология управления данными».