# Аннотация

В данной курсовой работе проведен анализ существующей ИТ-инфраструктуры отдела обработки данных ООО «Farpost», построенной на современной архитектуре потоковой обработки с использованием Apache Kafka в качестве центральной шины данных, Debezium для реализации паттерна Change Data Capture, Altinity Sink Connector для репликации в аналитическую СУБД ClickHouse и Apache Airflow для оркестрации рабочих процессов. Все компоненты развернуты в кластере Kubernetes, что обеспечивает отказоустойчивость и масштабируемость системы.

Анализ текущих процессов управления метаданными выявил ряд критических проблем: отсутствие единого каталога данных, невозможность отслеживать полное происхождение данных (lineage) от источников до конечных аналитических таблиц, фрагментарная документация SQL-скриптов и DAG-ов Airflow, отсутствие бизнес-глоссария и системы тегов, а также неавтоматизированный процесс актуализации метаданных. Для решения указанных проблем в работе спроектирована система управления метаданными на базе платформы OpenMetadata, выбранной после сравнительного анализа с аналогичными open-source решениями (Apache Atlas, DataHub, Amundsen) по функциональным и нефункциональным критериям.

Проект внедрения включает разработку логической архитектуры интеграции, проектирование ingestion-процессов для автоматического сбора метаданных из всех компонентов системы (Kafka, Debezium, Altinity Sink Connector, Airflow, ClickHouse), настройку структуры каталога метаданных с поддержкой бизнес-глоссария и визуализацией происхождения данных. Особое внимание уделено интеграции решения в существующую Kubernetes-среду с минимальным влиянием на текущие процессы обработки данных и обеспечением отказоустойчивости системы.

Ожидаемые эффекты от внедрения системы управления метаданными включают повышение прозрачности данных за счет централизованного каталога с визуализацией lineage, сокращение времени поиска и понимания данных для аналитиков и разработчиков, улучшение качества данных через системный контроль и ускорение выявления ошибок в процессах обработки, а также снижение операционных затрат на сопровождение аналитической инфраструктуры.