УДК 004.62

Проектирование и реализация модуля для загрузки и обработки данных в распределенной системе с использованием конфигурационных файлов

Садовая А. Р.[1]
Научный руководитель – инженер Мигулаева Т. А.[1]
[1]Университет ИТМО
alwaysbebetter@mail.ru

Введение

Современные процессы обработки данных в корпоративных хранилищах всех типов представляют собой сложные распределенные системы, требующие надежной, стандартизированной и эффективной организации процесса ETL (Extract, Transform, Load). В условиях, когда множеству команд необходимо разрабатывать сотни однотипных пайплайнов загрузки данных из различных источников, использование разрозненных скриптов становится узким местом, приводящим к ошибкам, сложностям поддержки и высоким операционным затратам [1]. Кроме того, написание каждого такого пайплайна с нуля - это трудозатратная работа, требующая от разработчика высокого уровня владения инструментами распределенной обработки данных и языков программирования. В связи с этим, существует необходимость разработки специализированного фреймворка, который бы абстрагировал сложность распределённых вычислений и предоставлял единый декларативный интерфейс для описания ETL-задач. Такой фреймворк позволяет централизованно управлять всем жизненным циклом загрузки данных в соответствии с практическими потребностями бизнеса и стандартами инженерии данных.

Основная часть

В основе предлагаемого решения лежит конфигурационно-ориентированный подход, при котором логика ETL-процесса задаётся в виде декларативного описания в формате JSON. Конфигурационный файл содержит информацию о типе источника данных, правилах маппинга полей, а также дополнительных параметрах обработки, что позволяет полностью отделить описание бизнес-логики загрузки от её программной реализации. Такой подход обеспечивает единообразие ETL-пайплайнов, упрощает их сопровождение и минимизирует необходимость внесения изменений в код при адаптации процессов под новые источники данных.

Реализация модуля выполнена с использованием языка программирования Scala и фреймворка Apache Spark, что позволяет эффективно обрабатывать большие объёмы данных в распределённой среде [2]. Архитектура решения предполагает интерпретацию конфигурационного файла во время выполнения и динамическое формирование пайплайна обработки данных на основе заданных параметров. Запуск ETL-процессов осуществляется через механизм Spark Submit, интегрированный в систему оркестрации Apache Airflow. Такая централизованная реализация механизмов обработки данных упрощает внедрение единых стандартов логирования, валидации, обработки ошибок и проверок качества данных, что положительно сказывается на стабильности всей системы и сокращает операционные затраты [3].

Кроме того, предложенный модуль ориентирован на снижение порога входа для разработки и сопровождения ETL-процессов. За счёт использования конфигурационных файлов и стандартизированной архитектуры становится возможным делегировать создание типовых загрузок специалистам с меньшим уровнем технической подготовки, не снижая при этом качества и надёжности решений.

Выводы

В результате работы был спроектирован и реализован модуль загрузки и обработки данных для распределённой системы, основанный на декларативном описании ETL-процессов с использованием конфигурационных файлов. Предложенное решение демонстрирует возможность стандартизации и унификации процессов загрузки данных в корпоративных хранилищах, а также снижение трудозатрат на разработку и сопровождение ETL-пайплайнов. Использование Apache Spark в сочетании с системой оркестрации Apache Airflow обеспечивает масштабируемость, отказоустойчивость и удобство эксплуатации решения, что делает его применимым в реальных промышленных сценариях обработки больших данных.

Литература

1. Ozyurt I.B., Grethe J.S. Foundry: a message-oriented, horizontally scalable ETL system for the geospatial domain // Journal of Big Data. – 2018. – Vol. 5, № 1. – P. 1-25. – DOI: 10.1186/s40537-018-0147-6. – URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6301337/ (дата обращения: 22.01.2026).
2. Batmaz F. ETL Data Pipelines Configurations in Spark // Proceedings of the International Workshop on Open Source Software for Big Data. – 2022. – P. 1-12. – URL: https://oss.cs.fau.de/wp-content/uploads/2022/09/batmaci_2022.pdf (дата обращения: 22.01.2026).
3. Chanda D. Automated ETL Pipelines for Modern Data Warehousing: Architectures, Challenges, and Emerging Solutions // Eastern Journal of Innovation and Creativity in Science and Technology. – 2024. – № 2. – P. 45-62. – URL: https://esj.eastasouth-institute.com/index.php/esiscs/article/view/523 (дата обращения: 26.01.2026).