СРАВНИТЕЛЬНЫЙ АНАЛИЗ СРЕДСТВ И ПЛАТФОРМ ДЛЯ АВТОМАТИЗАЦИИ ETL ПРОЦЕССОВ В СОВРЕМЕННЫХ ХРАНИЛИЩАХ ДАННЫХ

Н.Д. Громов, *магистрант*
А.И. Платошин, *магистрант*
Научный руководитель: *А.В. Панов, канд. техн. наук, доцент*
МИРЭА – Российский технологический университет (РТУ МИРЭА)
(Россия, г. Москва)

DOI:10.24412/2500-1000-2023-11-4-46-48

Аннотация. *В современном мире наблюдается значительный рост объемов данных, что вызывает проблемы с их эффективным сбором, обработкой и анализом. Организации стремятся использовать данные как конкурентное преимущество, поэтому эффективные ETL-процессы помогают минимизировать ошибки данных и повышать точность анализа. В исследовательской работе будут рассмотрены пять популярных инструментов для автоматизации ETL-процессов, включая Apache Airflow, Pentaho Data Integration и Oracle Data Integrator, а также проведен их сравнительный анализ по разным критериям.*

Ключевые слова: *сравнительный анализ, ETL-процессы, обработка данных.*

Актуальность данной статьи состоит в том, что в современном мире наблюдается колоссальный рост объемов данных, что влечет за собой проблемы с эффективностью сбора, обработки и анализа этих данных в организациях. Также это обусловлено желанием компаний использовать полученные данные, как конкурентное преимущество и именно точность, скорость и эффективность в работе с ними, помогут улучшить оперативное принятие решений. Эффективные ETL-процессы помогут минимизировать ошибки данных и тем самым повысить точность анализа и предоставления доверенной информации. Именно поэтому анализ и выбор средств для автоматизации ETL-процессов является важным шагом в современной информационной технологии.

В ходе анализа будут рассмотрены пять популярных ETL инструментов и с целью полного понимания о том, что они из себя представляют, рассмотрим их подробнее.

Apache Airflow представляет собой открытую платформу для управления и планирования рабочих процессов (workflow), которая автоматизирует сложные задачи обработки данных, ETL (Extract, Transform, Load) процессы, планирование задач и многое другое. Главное особенностью является возможность писать направленные ациклические графы (DAG) на языке Python, что значительно повышает гибкость.

Pentaho Data Integration (PDI), также известный как Kettle, это мощный инструмент для интеграции данных и ETL (Extract, Transform, Load), разработанный компанией Pentaho, которая является частью корпорации Hitachi Vantara. PDI предоставляет средства для извлечения данных из различных источников, их преобразования и загрузки в разнообразные целевые хранилища данных.

Oracle Data Integrator – это программное решение, разработанное и распространяемое компанией Oracle. ODI предоставляет средства для извлечения данных из различных источников, их преобразования и загрузки в целевые хранилища данных. Оно также широко используется для синхронизации данных между различными системами и для управления и автоматизации процессов ETL (Extract, Transform, Load). Отлично работает с программными продуктами компании Oracle, но может полноценно работать и с прочими решениями.

Apache NiFi представляет собой инструмент, с открытым исходным кодом, управления потоками данных из разнообразных источников в режиме реального

времени с использованием графического интерфейса. Совместим с программными продуктами экосистемы Hadoop.

IBM InfoSphere DataStage – это инструмент и платформа для управления и интеграции данных, который разработан IBM. Он предоставляет средства для сбора, трансформации и передачи данных из различных источников в разнообразные системы и хранилища данных. IBM InfoSphere DataStage обеспечивает графический интерфейс для создания и настройки процессов ETL, что делает процесс интеграции данных более доступным и эффективным для разработчиков. Этот инструмент широко используется организациями для обработки данных и обеспечения их качества и доступности.

Перейдем к подробному сравнительному анализу пяти ведущих ETL инструментов: Apache Airflow, Pentaho Data Integration (PDI), Oracle Data Integrator (ODI), Apache NiFi и IBM InfoSphere DataStage. Основное внимание уделяется ключевым возможностям, таким как графический пользовательский интерфейс (GUI), расширяемость, управление производительностью и поддержка источников данных.

Таблица 1. Сравнение ETL инструментов

Критерии	Apache Airflow	Pentaho Data Integration	Oracle Data Integrator	Apache NiFi	IBM InfoSphere DataStage
Графический интерфейс	Веб-интерфейс для мониторинга	Полноценный GUI для дизайна ETL	GUI и визуальное моделирование	Удобный веб-интерфейс для дизайна	Мощный GUI для работы с данными
Расширяемость	Поддержка плагинов	Плагины и интеграция с Kettle	Расширяемость через SDK	Интеграция процессоров	Широкий спектр расширений
Управление версиями	Встроенное через Git	Частичная поддержка через репозиторий	Поддержка через ODI репозиторий	Неявная, через внешние системы	Встроенное управление версиями
Производительность	Зависит от исполнителя	Высокая с кэшированием и оптимизацией	Высокопроизводительное исполнение с E-LT	Разработан для высокой пропускной способности	Оптимизирован для больших объемов
Поддержка источников данных	Широкий диапазон через интерфейсы	Широкий диапазон источников	Сильная интеграция с Oracle и другими	Многочисленные источники данных	Широкая поддержка различных систем
Масштабируемость	Горизонтальная через Celery и другие	Вертикальная и горизонтальная	Отличная, с поддержкой распределенной обработки	Разработан для распределенных систем	Вертикальная и частично горизонтальная
Сложность настройки	Средняя, требует знания Python	Низкая с интуитивным интерфейсом	Высокая, сложная настройка	Средняя, удобная для неразработчиков	Средняя, требует специализированных знаний
Стоимость	Бесплатный	Бесплатная и платная версии	Высокая, особенно для Oracle DB пользователей	Бесплатный	Высокая, лицензионная

Таким образом, можно сделать следующие выводы:

- Apache Airflow – идеально подходит для комплексной автоматизации рабочих процессов благодаря своей модульности и гибкости. Это открытое решение, подходящее для различных сред и вариантов использования, особенно когда дело касается интеграции с облачными сервисами. Однако с учетом санкционной политики, наличие некоторых интеграций может быть ограничено на российском рынке.

- Pentaho Data Integration (PDI) – благодаря своему графическому интерфейсу и мощным функциям, PDI является хорошим выбором для организаций, которые

предпочитают визуальное проектирование ETL-процессов. Существует как коммерческая, так и открытая версия PDI, однако доступ к коммерческой поддержке и обновлениям может быть ограничен в России.

- Oracle Data Integrator (ODI) – этот инструмент выделяется высокой производительностью и тесной интеграцией с другими продуктами Oracle. Однако его стоимость может быть препятствием для многих организаций, а текущие политические условия могут ограничить или полностью блокировать доступ к этому инструменту на российском рынке.

- Apache NiFi – Хорошо подходит для сценариев, требующих высокой пропускной способности и гибкого управления потоками данных. Как проект Apache, он бесплатен и открыт для использования, что делает его доступным для российских пользователей несмотря на внешние ограничения.

- IBM InfoSphere DataStage – Этот инструмент предлагает мощные возможности для работы с большими объемами данных и имеет широкую поддержку различных систем. Однако, как и в случае с ODI, доступность DataStage на российском рынке может быть существенно ограничена из-за санкций и политики IBM в отношении российского региона. Также внедрение данной системы эффективно скажется на процессах поиска других поставщиков за счет своего универсализма и широкого спектра выполняемых задач.

Библиографический список

1. Баева В.Р., Дроздов А.Ю. ETL: актуальность и применение. Преимущества и недостатки ETL инструментов // Вестник науки. – 2019. – №5 (14).

2. Талгатова З.Т. Анализ и сравнение существующих моделей процессов ETL для хранилищ данных // Технические науки – от теории к практике. – 2016. – №1 (49).

3. Кузьмина Ю.В., Кубанских О.В. Краткое описание процесса ETL // Ученые записки Брянского государственного университета. – 2017. – №1 (5).

COMPARATIVE ANALYSIS OF TOOLS AND PLATFORMS FOR AUTOMATION OF ETL PROCESSES IN MODERN DATA WAREHOUSES

N.D. Gromov, *Graduate Student*
A.I. Platoshin, *Graduate Student*
Supervisor: *A.V. Panov, Candidate of Technical Sciences, Associate Professor*
MIREA Russian Technical University
(Russia, Moscow)

Abstract. *In the modern world, there is a significant increase in data volumes, which causes problems with their effective collection, processing and analysis. Organizations strive to use data as a competitive advantage, so effective ETL processes help minimize data errors and improve the accuracy of analysis. The research paper will consider five popular tools for automating ETL processes, including Apache Airflow, Pentaho Data Integration and Oracle Data Integrator, and also carry out their comparative analysis according to different criteria.*

Keywords: *comparative analysis, ETL processes, data processing.*