

АВТОМАТИЗАЦІЯ ПРОЦЕСІВ ETL ДЛЯ ВЕЛИКИХ МАСИВІВ ДАНИХ

Валентин Дацко

AUTOMATION OF ETL PROCESSES FOR LARGE DATA SETS

Valentyn Datsko

Міжрегіональна Академія Управління Персоналом, вул. Фрометівська 2, Київ, 03039, Україна.

Abstract. *The study examines current approaches and tools for ETL automation, highlighting their impact on processing performance and scalability. Key challenges and strategies for optimizing ETL workflows are also discussed.*

Сучасні інформаційні системи збирають та обробляють дуже великі обсяги даних. Це потребує ефективної організації процесів їхнього вилучення, трансформації та завантаження (ETL — Extract, Transform, Load). Завдяки сучасним інструментам, зараз можливе ефективне управління потоками даних у режимі реального часу.

Процеси ETL складаються з трьох основних етапів: вилучення даних із джерел, їх трансформація згідно з вимогами та завантаження у цільову базу даних або сховище. Автоматизація цих етапів дозволяє значно підвищити швидкість обробки інформації, зменшити кількість помилок та покращити контроль над великою кількістю інформації. Серед сучасних інструментів виділяються Apache NiFi, Apache Airflow, Talend та Pentaho.

Ці інструменти мають різну архітектуру і підхід до обробки даних, що безпосередньо впливає на їхню швидкодію. Pentaho та Talend орієнтовані на класичні ETL-процеси, але поступаються в продуктивності при роботі з великими обсягами даних. Airflow виділяється завдяки асинхронному виконанню завдань, що дозволяє йому швидше обробляти великі масиви інформації в пакетному режимі. NiFi, у свою чергу, є лідером у real-time потоках, забезпечуючи стабільну швидкість при обробці подій у режимі реального часу.

Таким чином, кожен із цих інструментів має свої сильні сторони та недоліки, вибір залежить від вимог проекту, обсягу даних та необхідності у швидкості обробки.

Серед розглянутих інструментів, оптимальним для автоматизації ETL-процесів є Apache Airflow. Його головними перевагами є підтримка складних робочих процесів (DAG), можливість легко інтегруватися з Docker та Kubernetes для масштабування, а також широка спільнота розробників, що забезпечує підтримку та розширення можливостей. Він дозволяє будувати складні ETL-конвеєри з високим ступенем надійності та контролю. Подальше впровадження Airflow у сучасні архітектури обробки даних дозволить забезпечити стабільність та масштабованість ETL-процесів на високих обсягах даних.

Незважаючи на деякі недоліки, такі як складність у налаштуванні інфраструктури та крута крива навчання, Airflow забезпечує максимальну гнучкість та надійність у довгостроковій перспективі.

Бібліографія

1. Seethala S. C. Revolutionizing Data Warehouses in Manufacturing: Big Data-Infused Automation for ETL and Beyond / S. C. Seethala // Journal of Scientific and Engineering Research. – 2017.
2. Берко А. Ю., Верес О. М., Пасічник В. В. Системи баз даних та знань. Книга 1: Організація баз даних та знань: підручник. – 2-ге вид. – Львів: Магнолія 2006, 2015. – 440