International Journal of General Engineering and Technology (IJGET)
ISSN (P): 2278–9928; ISSN (E): 2278–9936
Vol. 11, Issue 2, Jul – Dec 2022; 63–98
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

AUTOMATING DATA EXTRACTION AND TRANSFORMATION USING SPARK SQL AND PYSPARK

Afroz Shaik[1], Ashish Kumar[2], Archit Joshi[3], Om Goel[4], Dr. Lalit Kumar[5] & Prof.(Dr.) Arpit Jain[6]

[1]Cleveland State University, Cleveland OH, USA

[2]Scholar, Tufts University, Tufts University Medford , USA

[3]Syracuse University, Syracuse, Sadashivnagar New York, USA

[4]ABES Engineering College Ghaziabad, India

[5]Asso. Prof, Dept. of Computer Application IILM University Greater Noida

[6]KL University, Vijaywada, Andhra Pradesh, India

ABSTRACT

The rapid growth of data in modern enterprises has necessitated efficient solutions for data extraction, transformation, and loading (ETL) processes. Automating these processes using scalable technologies like Spark SQL and PySpark offers significant improvements in speed, reliability, and resource management. This study explores the integration of Spark SQL and PySpark in automating ETL workflows, focusing on the performance benefits for large-scale data. Spark SQL, with its SQL-like querying capabilities, simplifies data extraction and manipulation, while PySpark's integration with Python enables advanced transformations through seamless scripting and machine learning libraries. The automation achieved through these technologies reduces human intervention, ensures real-time data handling, and minimizes errors. This paper further discusses best practices for optimizing Spark clusters, enhancing parallel processing, and handling complex transformations. By automating the ETL pipeline with Spark SQL and PySpark, organizations can accelerate data-driven decision-making while reducing operational costs and improving data quality. The findings indicate that these tools, when combined with automation frameworks, create robust and scalable ETL solutions suited for dynamic data environments.

KEYWORDS- *Data Extraction, Data Transformation, Spark SQL, PySpark, ETL Automation, Real-Time Data Processing, Parallel Processing, Big Data, Cluster Optimization, Scalable Data Solutions*

Article History

Received: *12 Dec 2022* **|** ***Revised:*** *20 Dec 2022* **|** ***Accepted:*** *22 Dec 2022*