International Journal of Computer Science and Engineering (IJCSE)
ISSN (P): 2278–9960; ISSN (E): 2278–9979
Vol. 13, Issue 1, Jan–Jun 2024; 193–222
© IASET

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

SCALABLE DATA PIPELINES USING AZURE DATA FACTORY AND DATABRICKS

Ravi Kiran Pagidi[1], Vishwasrao Salunkhe[2], Pronoy Chopra[3], Er. Aman Shrivastav[4], Prof.(Dr) Punit Goel[4] & Om Goel[5]

[1]Independent Researcher, Jawaharlal Nehru Technological University, Hyderabad, India

[2]Independent Researcher, Savitribai Phule Pune University, Pune, India

[3]Independent Researcher, University of Oklahoma Kali Bari Marg, New Delhi, India

[4]Independent Researcher , ABESIT Engineering College , Ghaziabad, India

[5]Research Supervisor, Maharaja Agrasen Himalayan Garhwal University, Uttarakhand, India

[6]Independent Researcher, ABES Engineering College Ghaziabad, India

ABSTRACT

In the era of big data, organizations are increasingly challenged to manage and analyze vast volumes of information efficiently. This paper explores the development of scalable data pipelines utilizing Azure Data Factory and Databricks, two powerful tools that streamline data integration and processing. Azure Data Factory serves as a robust orchestration service, enabling users to create, schedule, and manage data workflows across diverse sources. Its ability to connect with numerous data stores, both on-premises and in the cloud, facilitates seamless data movement and transformation. Databricks complements this by providing an interactive environment for big data analytics and machine learning, leveraging Apache Spark's capabilities to process large datasets in real time.

The integration of Azure Data Factory with Databricks allows for the construction of end-to-end data pipelines that can efficiently handle increasing data loads. This paper outlines the architecture and implementation strategies for these pipelines, highlighting best practices for optimizing performance and scalability. Furthermore, we discuss the challenges encountered during the integration process and the solutions implemented to overcome them. By harnessing the combined power of Azure Data Factory and Databricks, organizations can achieve greater agility in their data operations, enabling faster insights and improved decision-making. The findings underscore the significance of adopting cloud-based solutions for scalable data engineering in the modern data landscape, paving the way for enhanced operational efficiency and innovation.

KEYWORDS: *Scalable Data Pipelines, Azure Data Factory, Databricks, Big Data Analytics, Data Integration, Apache Spark, Cloud Solutions, Data Orchestration, Machine Learning, Performance Optimization*

Article History

Received: 23 May 2024 | Revised: 25 May 2024 | Accepted: 30 Jun 2024