International Journal of General Engineering and Technology (IJGET)
ISSN (P): 2278–9928; ISSN (E): 2278–9936
Vol. 9, Issue 1, Jan – Jun 2020; 81–120

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

ENHANCING CLOUD DATA PIPELINES WITH DATABRICKS AND APACHE SPARK FOR OPTIMIZED PROCESSING

Rajkumar Kyadasu[1], Rahul Arulkumaran[2], Krishna Kishor Tirupati[3], Prof. (Dr) Sandeep Kumar[4], Prof. (Dr) MSR Prasad[5]& Prof. (Dr) Sangeet Vashishtha[6]

[1]Rivier University, South Main Street Nashua, NH 03060

[2] University At Buffalo, New York, Srinagar Colony, Hyderabad, India

[3]International Institute of Information Technology Bangalore, India

[4]Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation Vadeshawaram, A.P., India

[5]Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation Vadeshawaram, A.P., India

[6] IIMT University, Meerut, India

ABSTRACT

The growing complexity of data ecosystems necessitates robust and scalable solutions to efficiently manage and process vast amounts of data in real-time. Cloud data pipelines, powered by advanced technologies like Databricks and Apache Spark, have emerged as key enablers for optimized data processing across industries. This paper explores how the integration of Databricks and Apache Spark enhances cloud data pipelines by enabling high-performance, distributed processing and real-time analytics. Databricks offers a collaborative environment that simplifies the orchestration of large-scale data workflows, while Apache Spark provides the core engine for executing data transformations with speed and scalability. By leveraging these platforms, organizations can improve the performance, cost-efficiency, and flexibility of their data pipelines, ensuring faster insights from their data. Additionally, this research examines best practices for deploying Databricks and Spark in cloud environments, highlighting their role in reducing operational complexity and optimizing resources for large-scale data operations. The study concludes with an analysis of real-world use cases demonstrating the effectiveness of this combination in various sectors, including finance, healthcare, and e-commerce.

KEYWORDS: *Cloud Data Pipelines, Databricks, Apache Spark, Distributed Processing, Real-Time Analytics, Data Transformation, Scalability, Optimized Data Workflows, Performance, Resource Optimization*

Article History

***Received:* 09 Feb 2020 | *Revised:* 11 Feb 2020 | *Accepted:* 12 Feb 2020**