DOI: doi.org/10.47363/JAICC/TechFusion2025/2025(4)5

TechFusion 2025 - AI, Cybersecurity, and Emerging Trends in Computer Sciences

Conference Proceedings November 27-28, 2025 (Virtual)

Architecting AI Workflows with Apache Spark Chemotherapy

Hina Gandhi

New York, USA

Abstract:

This session traces the evolution of big data systems - from Hadoop' s batch-driven model to modern distributed architectures - and explores how AI-driven approaches can enhance and optimize Apache Spark. We'll break down Spark's internal design, including the roles of the Driver, DAG Scheduler, Task Scheduler, and Executors, to show how large-scale workloads are processed efficiently across clusters. Using real-world examples like cost aggregation pipelines, the talk highlights how Spark overcomes Hadoop's limitations while still facing challenges around configuration complexity, data skew, and resource management. Finally, we'll discuss how reinforcement learning can be applied to Spark to enable dynamic scheduling, smarter partitioning, and adaptive resource allocation, transforming Spark into a self-optimizing data processing engine.