2025W-T3 BDM 3035 - Big Data Capstone Project 01 (DSMM Group 1)

Research Paper Proposal

TERM PROJECT

Submitted to :

Prof. Bhavik Gandhi

Submitted by :

Group 5

1. A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project

- **Authors:** Mihhail Matskin, Shirin Tahmasebi, Amirhossein Layegh, Amir H. Payberah, Aleena Thomas, Nikolay Nikolov, and Dumitru Roman
- **Year:** 2021
- **Journal:** International Conference on Data Analytics and Management in Data-Intensive Domains (DAMDID)
- **Abstract:** This paper presents a survey of existing tools for Big Data pipeline orchestration based on a comparative framework developed in the DataCloud project. The authors propose criteria for evaluating the tools to support reusability, flexible pipeline communication modes, and separation of concerns in Big Data pipeline descriptions. The survey aims to identify research and technological gaps and to recommend approaches for filling them. Further work in the DataCloud project is oriented towards the design, implementation, and practical evaluation of the recommended approaches.
- **URL:** https://payberah.github.io/files/download/papers/damdid.pdf

2. Cost-Effective Big Data Orchestration Using Dagster: A Multi-Platform Approach

- **Authors:** Hernan Picatto, Georg Heiler, and Peter Klimek
- **Year:** 2024
- **Journal:** Supply Chain Intelligence Institute Austria
- **Abstract:** This paper introduces a cost-effective and flexible orchestration framework using Dagster. The solution aims to reduce dependency on any single Platform-as-a-Service (PaaS) provider by integrating various Spark execution environments. The authors demonstrate how Dagster's orchestration capabilities can enhance data processing efficiency, enforce best coding practices, and significantly reduce operational costs. In their implementation, they achieved a 12% performance improvement over Amazon Web Services Elastic MapReduce (EMR) and a 40% cost reduction compared to Databricks, translating to over 300 euros saved per pipeline run.
- **URL:** https://arxiv.org/pdf/2408.11635

3. Architecting Data Pipelines for Scalable and Resilient Data Processing Workflows

- **Authors:** Muhammadu Sathik Raja
- **Year:** 2025
- **Journal:** International Journal of Emerging Research in Engineering and Technology Pearl Blue Research Group
- **Abstract:** In the era of big data, architecting scalable and resilient data pipelines is crucial for organizations aiming to harness vast amounts of information efficiently. This paper explores essential principles and best practices for designing data pipelines that can adapt to increasing data volumes while maintaining high performance and reliability. Key

components of robust data pipeline architecture include data ingestion, processing, storage, orchestration, and monitoring. Emphasizing modular design allows independent scaling of pipeline components, enhancing fault tolerance and flexibility. Implementing cloud - based solutions with auto - scaling capabilities ensures that the architecture can dynamically adjust to fluctuating workloads. Additionally, incorporating mechanisms for fault tolerance such as data replication and Checkpointing enables seamless recovery from failures, minimizing data loss. The paper also discusses the significance of continuous monitoring and optimization to identify bottlenecks and improve overall system efficiency. By adhering to these architectural guidelines, organizations can build resilient data processing workflows that not only meet current demands but are also future - ready.

- **URL:** https://ijeret.org/index.php/ijeret/article/view/7/6