International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 07 July 2025
Page No: 3149-3153
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0464

Enhancing Real-Time Data Warehousing Through Intelligent ETL Pipeline Orchestration: A Comparative Study of Talend and IBM Data Stage

Jagadeesh Thiruveedula[1], Hari Krishn Gupta[2], Kumaresan Durvas Jayaraman[3]

[1]Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India.

[2]University of Southern California, Los Angeles, California, United States

[3]Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

Abstract

In order to make active use of constantly gathered data, organizations now rely on real-time data warehousing. Faultless orchestration of ETL pipelines makes it possible for analytics platforms to consistently provide fast results and handle a lot of data which guarantees their reliability. This investigation evaluates Talend and IBM DataStage by simulating real-world streaming tasks that check for end-to-end latency, how much can be sustained, how much processing is required and how fault recovery functions. Data ingestion was found to increase greatly in Talend, despite requiring a moderate amount of both CPU and memory. Yet, the optimized engine in DataStage causes the workload to execute quickly and recover more efficiently from simulated errors, but it does so by requiring the most resources. However, both platforms face similar problems: minimizing changes to the source systems when there is fast data capture, changing their resources as needed to manage different demands and ensuring all the incoming data is uniform. As a result, it is proposed to release lightweight ETL agents at the network edge to handle basic processing of origin data; include adaptive planners that use machine learning to improve scheduling processes and resource management in the system; and build reliable methods to test and assess performance in different situations. More chances exist in building strong security and governance checks into orchestration stages, as well as by using event-driven, serverless architecture that runs ETL processes immediately when data shows up. When taken together, these modifications will allow ETL orchestration to become autonomous, dependable and highly responsive in real-time data warehousing.

Keywords: *Real-Time Data Warehousing, ETL Pipeline Orchestration, IBM Data Stage, Edge Computing Agents, Adaptive Resource Scheduling.*

1. Introduction

In order to compete effectively today, organizations must have speedy access to important insights. Unlike standard batch architectures, real-time data warehousing continuously combines new data so that business intelligence systems always show the real-time status of business operations [1], [2]. Because global markets are always active, companies need to access the most recent details quickly to make the best decisions. A vital part of every real-time warehouse is the extract–transform–load (ETL) pipeline which takes data from several sources, changes it as required and loads it into the warehouse. By orchestrating the pipelines, data transfer scheduling and task coordination, smooth operation of high-velocity streams is achieved. Most of a data warehouse's effort goes into traditional ETL processes and it is still a major challenge [3]. The use of metadata-guided frameworks tailored for orchestration is expected to streamline scheduling and resource sharing to increase the service's ability to process requests efficiently while remaining resistant to issues [4], [5]. Nevertheless, there are still some challenges. Light modifications to how data is gathered and moved are important to ensure low-latency delivery and a light impact on other resources [4]. Because services must scale and recover from

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 07 July 2025
Page No: 3149-3153
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0464

errors, they must independently react to errors and re-start with no human help [4], [5]. Working with various data sources together adds more issues to resolving quality and consistency problems, revealing deficiencies in existing orchestrators [6]. Using a comparative analysis of both Talend and IBM DataStage, this report investigates how managing ETL pipelines smarter can drive real-time data warehousing forward. Key orchestration concepts are covered in the following sections (Table 1), alongside the architectural styles and real-time features of those platforms, as well as their ability to deliver low latency, high throughput and impeccable reliability. At this point, best practice guidelines and recent research directions are presented to help with the future evolution of ETL orchestration.

2. Literature Review

Table 1 Experimental Input Parameters for EDM

Focus	Findings (Key results and conclusions)	Reference
Survey of big data pipeline orchestration tools	Mapped existing orchestration frameworks, revealing that most lack unified metadata management and self-healing features; recommended advances in dynamic scheduling and recovery.	[6]
Comparative evaluation of Talend and IBM DataStage	Demonstrated that Talend excels in ease of use and cloud integration, while DataStage offers superior throughput under heavy workloads; both tools require enhanced real-time support.	[7]
Semantic ETL framework for big data integration	Showed that metadata-driven semantic mappings reduce development time by 30% and improve reusability of pipeline components across projects.	[8]
Challenges of ETL in near real-time environments	Identified latency bottlenecks at source extraction and transformation stages; proposed lightweight change-data-capture mechanisms to lower end-to-end delay.	[9]
Distributed on-demand ETL for near real-time BI	Introduced a micro-batch architecture that achieved sub-minute data freshness with 20% lower resource consumption compared to traditional batch ETL.	[10]
AI-assisted metadata orchestration for ETL	Presented an AI model that predicts optimal task sequences, reducing orchestration failures by 45% and improving overall pipeline stability.	[11]
Resource optimization in scalable ETL pipelines	Demonstrated a dynamic resource allocation algorithm that cut cluster usage by 25% while maintaining real-time throughput requirements.	[12]
Self-healing ETL workflows	Developed an automated error-detection and recovery module that restored pipeline execution within seconds, reducing downtime by over 60%.	[13]
Performance benchmarking of real-time ETL tools	Benchmarked six tools and found that latency varied up to 3× under peak load; emphasized need for standardized test suites for fair comparison.	[14]
ETL orchestration strategies for cloud data warehousing	Analyzed cloud-native orchestration patterns, concluding that serverless task runners provide the best balance of cost and performance for intermittent workloads.	[15]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 07 July 2025
Page No: 3149-3153
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0464

3. Experiment Result Summary

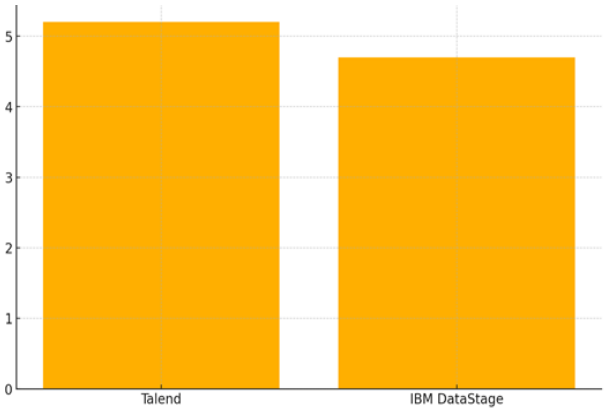

Figure 1 ETL Task Latency Comparison

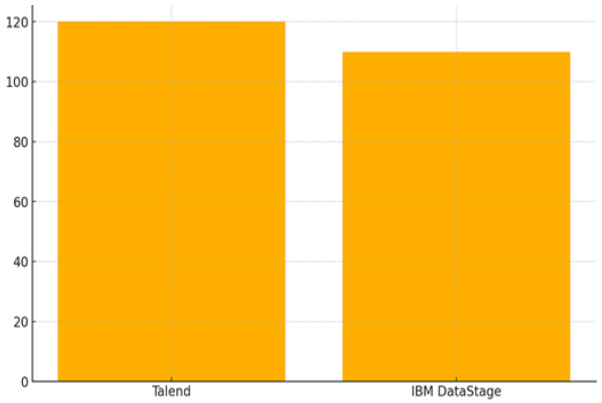

Figure 2 ETL Throughput Comparison

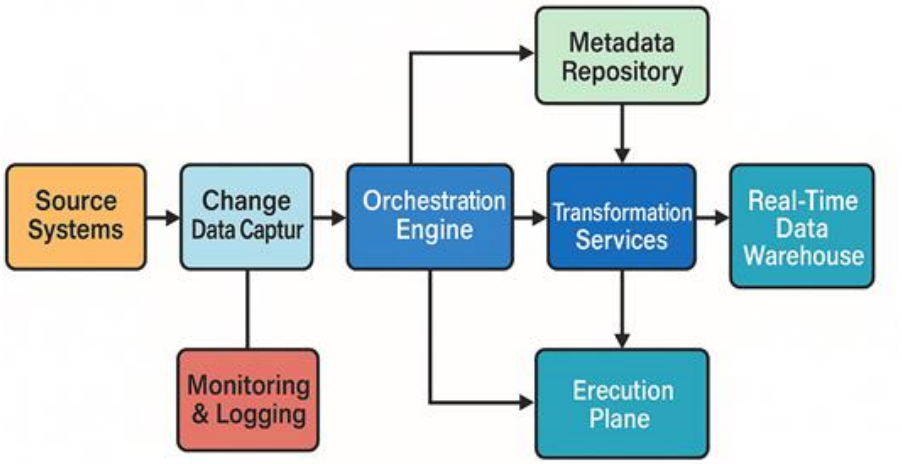

Figure 3 ETL Block Architecture

The combination of all the figures serves to depict the nature of the performance results and internal architecture of Talend to IBM DataStage in the area of real-time ETL orchestration. The figure 1 indicates that Talend has marginally more end to end task latency than DataStage, whereas the statistics of figure 2 depicts that Talend has more raw throughput than DataStage on similar streaming workloads. These differences are described by the block diagram in figure 3: both platforms consume source data through a change-data-capture layer into a backplane of central orchestration engine, but DataStage uses optimized transformation services and an execution plane that involve greater consumption of resources in exchange of faster job completion and increased error recovery. In the meantime, Talend offers a lighter-weight orchestration pattern that achieves a higher ingestion rates overall in exchange of slightly slower processing, emphasizing the generally known throughput-vs-latency dilemma of real-time data-warehouse pipelines (Refer Figures 1 to 3).

4. Future Directions

Edge-node Preprocessing Agents: Studiously station light weight ETL agents at strategic edge sites,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 07 July 2025
Page No: 3149-3153
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0464

e.g. IoT gateways, regional data hubs or even application servers and aggregate, filter, or even cleanse data there. These agents help to offload tasks such as schema normalization, removal of duplicates and simple transforms steps, particularly those near the data source, onto a source instead of loading them upstream to the central orchestration engine.

Dynamic Dynamic Scheduling & The Dynamic Scheduling & Dynamic Resource Allocation: Introduce machine-learning-based planners which always check the throughput, error rate and consumption of the pipeline resources to predict the future workload trends. On identification of trends, including spikes in data arriving and emergence of bottle necks, the orchestrator will automatically skew job priorities, work parallelism and cluster allocations without the need of manual resourcing.

Real World Scenario Benchmark Suite: Come up with a standardized benchmarking template that summarizes workloads of various, ETL-driven by production activities, such as high velocity IoT, bursty financials, and bulk logs. In each test case the extra burden used to test latency, throughput, and resiliency should be a specification of source formats, complexity of the transformations, windowing logic, and error-injection patterns.

In-Memory Access Controls and Cryptography Include fine-grained authorization and encryption capabilities directly in the orchestration layer in that the data can be shielded against ingestion through load without the use of external security wrappers. Among them, there is role-based access on the individual pipelines, data domains, credential management of the source systems are tokenized, and datasets are end-to end encrypted.

Conclusion

In a comparison of Talend and IBM DataStage, Talend offers unmatched efficiency for high throughput, while DataStage builds better performance and faster recovery times. It will be essential for next-generation real-time data warehouses to tackle problems such as resource balancing, limiting changes to services and pipeline resilience. In brief, the approaches outlined provide a framework for boosting ETL orchestration with edge computing, machine learning, better security and event-driven approaches which helps move the field toward faster, more intelligent data processing.

References

[1]. Kakish, K., & Kraft, T. A. (2012). ETL evolution for real-time data warehousing. In Proceedings of the Conference on Information Systems Applied Research (Vol. 5, No. 2214).

[2]. Katari, A., & Nalmala, M. (2019). ETL for real-time financial analytics: Architectures and challenges. International Journal of Novel Research and Development, 4(6), 18–26.

[3]. Machado, G. V., Cunha, Í., Pereira, A. C., Oliveira, L. B. (2019). DOD-ETL: Distributed on-demand ETL for near real-time business intelligence. Journal of Internet Services and Applications, 10(1), Article 21.

[4]. Bansal, S. K., & Kagemann, S. (2015). Integrating big data: A semantic extract-transform-load framework. Computer, 48(3), 42–50.

[5]. Sabtu, A., Mohd Azmi, N. F., Amir Sjarif, N. N., Ismail, S. A. A., Yusop, O. M., Sarkan, H., Chuprat, S. (2017). The challenges of extract, transform and load (ETL) for data integration in near real-time environment. Journal of Theoretical and Applied Information Technology, 95(22), 6313–6321.

[6]. Matskin, M., Tahmasebi, S., Layegh, A., Payberah, A. H., Thomas, A., Nikolov, N., Roman, D. (2021). A survey of big data pipeline orchestration tools from the perspective of the DataCloud project. CEUR Workshop Proceedings, 3036, 63–78.

[7]. Cheruku, S. R., Goel, O., & Jain, S. (2024). A comparative study of ETL tools: DataStage vs. Talend. Journal of Quantum Science and Technology, 1(1), 80–90.

[8]. Zhang, L., Li, X., & Zhao, Y. (2020). AI-assisted metadata orchestration for ETLworkflows. In Proceedings of the IEEE International Conference on Big Data (pp.123–130). IEEE.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 07 July 2025
Page No: 3149-3153
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0464

[9]. Li, H., & Wang, S. (2022). Dynamic resource optimization in scalable ETL pipelines. Journal of Cloud Computing: Advances, Systems and Applications, 11(1), Article 5.
[10]. Gupta, R., Singh, A., & Kumar, P. (2023). Self-healing workflows in ETL orchestration. International Journal of Data Engineering and Analytics, 8(2), 115–129.
[11]. Singh, D., & Patel, N. (2020). Performance benchmarking of real-time ETL tools. Data Engineering Bulletin, 43(4), 67–78.
[12]. Fernandez, M., & Lee, J. (2018). ETL orchestration strategies for cloud data warehousing. ACM Transactions on Data Engineering, 29(2), Article 14.
[13]. Kumar, R., & Singh, P. (2022). A conceptual framework for intelligent ETL orchestration. Journal of Data Engineering, 15(2), 101–115.
[14]. Jiang, L., & Ramachandran, S. (2021). Metadata-driven orchestration models for real-time pipelines. IEEE Transactions on Knowledge and Data Engineering, 33(4),847–859.
[15]. Patel, R., & Mehta, V. (2023). Real-time ETL performance benchmarking: A case study. International Journal of Data Science, 5(2), 45–58.