National College of Ireland

WS Glue vs Talend: A Practical Comparison of ETL Tools

MSc Research Project
Cloud Computing

Vikitha Konda
Student ID: x23175818

School of Computing
National College of Ireland

Supervisor: Shaguna Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vikitha Konda
Student ID:	x23175818
Programme:	Cloud Computing
Year:	2024
Module:	MSc Research Project
Supervisor:	Shaguna Gupta
Submission Due Date:	18/12/2024
Project Title:	WS Glue vs Talend: A Practical Comparison of ETL Tools
Word Count:	2900
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vikitha Konda
Date:	28th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

WS Glue vs Talend: A Practical Comparison of ETL Tools

Vikitha Konda
x23175818

Abstract

This project is a performance comparison of AWS Glue and Talend aimed at four ETL use cases, with a focus on data conversion, augmentation, handling of large files, and data transfer to Snowflake. In the test, it was found that AWS Glue outperformed the other AWS data-wrangling services in scalability, cost, and performance because it was serverless, unlike Athena, whereas Talend was easy to manage and provided profound control over the pipeline at each step. For scaled-out, cloud-native workloads, the results showed that Glue is a good fit, while Talend fits better in highly customized use cases. Criticisms of both tools are discussed, and strategies are provided on How best to use both tools together to maximize their effectiveness. The research offers precise recommendations that can help choose an appropriate tool from the ETL group for the given company's needs.

1 Introduction

1.1 Background and Motivation

In the digital age, the exponential growth of data has caused a host of breakthroughs in data processing, storage, and analysis. This is because businesses are generating massive volumes of data from multiple sources (web analytics, sales transactions, social media interactions, and IoT devices), requiring sophisticated tools to extract, transform, and load (ETL) that data for analysis and for making decisions. For this process to be successful, ETL tools suffice to automate data integration, processing and cleansing data and loading it into systems to be used in analysis and reporting.

ETL tools have historically been built for on-premises environments where flexibility, customization, and control were paramount. Due to their flexibility, cross-platform functionality and ability to be integrated with a very wide range of big data technologies, tools such as Talend, an open-source platform, have been popular. Nevertheless, with the ever increasing popularity of cloud computing, we have seen the emergence of cloud native tools such as AWS Glue, which are serverless architecture, scalable in nature, yet cost effective in cloud systems.

In light of enterprises drifting more into the cloud, an appropriate ETL tool is important. However, cloud-native solutions come with several benefits: they can scale on demand, have pay-as-you-go pricing models, and are integrated with a large set of cloud

services. There are downsides, however, such as vendor lock-in and higher costs with large-scale processing over time. On the other hand, traditional ETL tools are more controllable and more customizable, but they are not always efficient in cases where you are working with a dynamic and a cloud environment. In this study, we explore these tradeoffs: the performance, cost, scalability and resource utilization of AWS Glue and Talend.

1.2 The Emergence of Cloud-Native ETL Tools

With the advent of cloud computing, the way that businesses handle data and its processing has completely shifted. When data integration tasks need to be performed, well-powered cloud platforms such as Amazon Web Services (AWS), Microsoft Azure, etc., provide the required infrastructure without the need for on-premises hardware and maintenance. Amazon Web Services introduces AWS Glue, a fully managed ETL service that allows developers to develop scalable, serverless ETL pipelines. It simplifies much of the process of defining and managing ETL jobs, enabling businesses to use their efforts to develop data workflows, not infrastructure.

AWS Glue is integrated into the broader AWS ecosystem, which is one of the leading reasons why AWS Glue is such a convenient choice for organizations currently taking advantage of AWS storage (S3), data warehousing (Redshift), and analytics solutions (Athena). The fact that it is serverless basically means users only pay for what they use during processing, saving money for both small datasets and infrequent processing tasks. The other Glue that Apache Spark leverages to do distributed data processing is strong enough to handle large-scale data transformation jobs. However, the deep integration of Glue with AWS does create concern that using Glue can lock you into using AWS services for your ETL workflows so that you cannot move back to other platforms.

1.3 Traditional ETL Tools and Their Evolution

However, traditional tools such as Talend are still very much in play on the other side of the ETL spectrum—especially for those organizations that care most about control, customization and independence from the cloud vendor. Open source data integration platform for ETL pipelines data cleansing, and data source and technology integration, including tools for data cleansing, ETL pipelines building, etc. Whereas Talend is platform agnostic, supporting on-premise, cloud and hybrid environments, AWS Glue is tightly integrated into the AWS ecosystem. Talend's flexibility as a multi cloud choice or as a way to prevent vendor lock-in makes it an attractive solution for organizations with complex multi-cloud spines or strategies.
One main benefit of Talend is that it is open source, meaning organizations can then change and extend the tool to fit their specific needs. On top of that, it supports several big data platforms, such as Hadoop and Spark, so that firms can deal with vast volumes of data in a distributed environment. This allows users to have fine-grained control through customizations by generating Java code from its graphical interface. However, this level of flexibility comes at a cost: It's resource-intensive and has a steep learning curve, which may make it less suited to organizations looking for out-of-the-box type solutions or those without dedicated technical expertise.

1.4 The Need for Comparative Analysis

Even with cloud-native ETL tools like AWS Glue becoming popular, there is little research comparing them with traditional solutions like Talend. While the majority of existing studies evaluate each tool in isolation and explore its capability, there is still a lack of a comprehensive comparison between them in real-world ETL scenarios. When an organization comes to select the best ETL tool for a system today, the criteria they follow are mostly performance, scalability, cost, and ease of use. Particularly, businesses need to discern what piece of software is best for particular work, like data modification, data contemporary and data amalgamation.
An example would be AWS Glue, which is great for organizations handling very large-scale data processing because of its scalability and serverless architecture. This makes Talend more attractive for businesses that need to keep their data integration processes under their control or businesses worried about cloud vendor dependencies.

1.5 Problem Statement

As the data environments become more complex, there are plenty of ETL tools now available, with their own pros and cons. Yet, there exists a shortage of extensive empirical work that compares traditional ETL tools, such as Talend, to cloud-native solutions like AWS Glue. It leaves organizations wondering what is the best tool to solve their particular data integration requirements. The performance, scalability, and cost tradeoffs of these tools in common ETL scenarios that include data format conversion, enrichment, and aggregation have not been fully explored by existing research.

1.6 Research Objectives

This study aims to fill this gap by conducting a comparative analysis of AWS Glue and Talend across three key ETL scenarios. The primary objectives of this research are:

- To evaluate the performance of AWS Glue and Talend in handling ETL tasks such as data format conversion, data enrichment, and data aggregation.
- To analyze the scalability and resource utilization of both tools in real-world data processing environments.
- To assess the cost-effectiveness of AWS Glue and Talend, particularly in large-scale data processing tasks.
- To provide practical recommendations for organizations on the most suitable ETL tool based on their specific requirements.

1.7 Research Questions

The central research question guiding this study is: How do AWS Glue and Talend compare in terms of performance, cost-effectiveness, and suitability for specific ETL scenarios? Sub-questions include

- What are the strengths and weaknesses of AWS Glue in handling large-scale, distributed data processing tasks?

- How does Talend's flexibility and open-source nature affect its performance in complex ETL workflows?
- What are the cost implications of using AWS Glue versus Talend for long-term, large-scale data processing?

1.8 Significance of the Study

The need for effective ETL tools outpaces the need for organizations to continue generating massive amounts of data from various sources. This study will offer guidance to organizations in making informed decisions regarding what tool to choose among cloud-native and traditional ETL tools to suit the specific data integration needs of the organization. This research will provide practical recommendations for businesses going through the process of building or enriching their data integration pipelines by evaluating AWS Glue and Talend across the important performance indicators of time taken to process the dataset, resource utilization, bottleneck points, and cost.

1.9 Scope of the Study

This research will focus on comparing AWS Glue and Talend in the context of three common ETL scenarios: data format conversion, data enrichment, and data aggregation. To simulate real-world ETL tasks, the study will use publicly available datasets, e.g., those from Kaggle and the AWS Open Data Registry. A comparison of both the technical performance metrics (e.g., processing time, scalability) and economic factors (e.g. cost, resource utilization) will be carried out as part of the analysis of the two tools.

2 Related Work

2.1 Introduction to ETL Tools and Processes

For decades, the ETL (Extract, Transform, Load) process has tethered data integration, acting as a backbone for structures of how organisations make sense of the vast amount of data that accumulates daily, pulling the latest trends, hunches, and actionable insights out of data assets. ETL tools are primarily designed to extract data from various sources, transform it into a usable form, and load it into databases, data warehouses, or analytics platforms to be used for other purposes. While these tools don't replace analysts and data marshals, they automate much of the data preparation process—critical to the tasks of data warehousing, business intelligence (BI), and analytics—in some cases, replacing human effort.

Traditional ETL tools like Talend have been popular because they are gradually elastic and are able to handle small and vast scales of data integration work. Talend is an open-source platform that allows users to modify and customise data pipelines, which makes it quite adaptable to customer business needs. In contrast, AWS Glue, an Amazon Web Services (AWS) cloud-native ETL service, has gained traction as the service is serverless and cost-efficient and ingests easily with the other AWS services.

While both of these tools excel in their respective areas, the main factor in choosing between Talend and AWS Glue will depend on the organisation's overall needs. From on-premises to hybrid, Talend's open-source flexibility and cross-platform compatibility make it an attractive option. At the same time, AWS Glue's serverless nature, coupled with its seamless inclusion within the AWS ecosystem, means it will best suit these environments.

2.2 The Rise of Cloud-Native ETL Tools

The Rise of Cloud-Native ETL Tools Recently, cloud-native tools have become a game changer in the ETL space. Cloud computing platforms such as AWS, Microsoft Azure and Google Cloud have changed the way businesses handle data storage and processing, making it easy for them to grow and make operations. Cloud native ETL tools like AWS Glue are very much designed to leverage the cloud infrastructure and include features like elastic scalability, pay as you go pricing and serverless operations.
As stated by Baldini et al. (2017), the shift towards serverless computing has allowed new opportunities for data processing, especially for ETL tasks. According to them, serverless platforms eliminate infrastructure management, so it's all about pipeline building. That's why AWS Glue—a serverless ETL tool—plays well in this trend and enables businesses to run ETL jobs without deploying or managing servers. For organisations looking to focus on the cost-efficiency and scalability of data operations, this makes Glue all the more attractive.
One of AWS Glue's big pluses is that there is tight integration with other AWS services like S3 (Simple Storage Service), Redshift (data warehouse), and Athena (data analytics). It makes for a smooth flow of data within the AWS scheme of things, thus being a great option for any business that's all in the AWS. It should, however, be noted that, as pointed out by Opara-Martins et al. (2016), such reliance on a single cloud vendor exposes an organisation to vendor lock-in, which can limit an organisation's ability to be moved to other platforms.

2.3 Traditional ETL Tools: Talend's Legacy and Flexibility

Over the past decade, Talend has been a stalwart of the data integration landscape as a traditional ETL tool. Plus, one of the main reasons it is so well received is that it's been shown to interoperate with different enterprise data sources and technologies, including Hadoop, Spark, and NoSQL databases. The open-source model that Talend possesses offers organisations the opportunity to download and use the tool for free, then customise the tool to suit the organisations needs if so desired.

As the study of Kumar et al. (2020) mentioned, they compared their Talend with other traditional ETL tools. Informatica and Talend showed an edge over the competitors because of their open-source nature, which gives their users flexibility and cost-benefit over traditional ETL tools. Yet the study did reveal that Talend was largely resource-intensive and had a steep learning curve, which could be a negative for companies without native technical chops. Furthermore, Talend's performance in real-time ETL cases was also a bit worse than the cloud-native solutions like AWS Glue.

These limitations notwithstanding, Talend's flexibility keeps it in demand with companies that need great control over their ETL processes. He et al. (2018) conducted a study on using Talend to transform data in a Hadoop relational environment and observed that the tool's ability to generate Java code offers a lot of customisation options. Talend was widely suitable for complex ETL workflows such as the ones defined in the finance and healthcare industries, where high stringency of data processing requirements was needed.

2.4 Comparative Studies Between Cloud-Native and Traditional ETL Tools

While the two, Talend and AWS Glue, have been subjected to extensive research in isolation, comparative research has been absent when testing their performance. Most of the previous comparative studies, like the ones reported by Sreemathy et al. (2020); Niranjani and Selvam (2020), studied the application of Talend on business intelligence and data integration processes. Yet, they neglect the power of cloud-native tools like AWS Glue and fail to explore the distinction between traditional and cloud-native ETL tools.

In a comparative analysis of several ETL tools like Talend, Sreemathy et al. (2020)concluded that traditional tools like Talend do well in the data integration in the on-premises environments but are pipped to the post when it comes to scalability and cost-effectiveness in the cloud environments. The efficiency gained from large-scale ETL tasks required fully distributed data environments, which AWS Glue, with serverless infrastructure and integration with Apache Spark as author S et al. (2023); Qaiser et al. (2023), has been shown to handle. However, the study also found that AWS Glue is less flexible than Talend due to the AWS ecosystem, but it is more flexible than Apache Spark, Hadoop and HDFS.

.

Putters et al. (2023) and Bowen (2012) also investigated the performance of AWS Glue and traditional ETL tools for use in public cloud auditing tasks using another comparative study. However, they found that AWS Glue's ability to process large datasets distributedly was very advantageous, especially in cloud-native scenarios. However, the study also flagged cost and vendor lock-in concerns, which could be issues for significant, long-term projects.

2.5 Performance and Scalability of AWS Glue

One of the key unique features of AWS Glue is that you can scale ETL jobs dynamically based on the size of the data being processed. Whereas traditional ETL tools require infrastructure provisioning in advance, a cloud-based tool such as AWS Glue allows you to scale your operations on demand. It is particularly well suited for big data environments where volumes of data can be variable.

Zaharia et al. (2016); Kumar et al. (2020) evaluated the performance of AWS Glue working alongside Apache Spark for large-scale distributed data processing in a study. Examining the relation, the study concluded that AWS Glue gained better performance and scalability when dealing with large datasets than standard tools. In particular, the serverless architecture of AWS Glue was pointed out as helping to save on overhead costs,

as organisations only pay for resources used in performing data processing, which is a low-cost solution for large-scale ETL workloads.

2.6 Cost and Vendor Lock-In Considerations

The transaction cost of ETL tools is a fundamental issue for organisations working with large amounts of data. AWS Glue is a pay-as-you-go service, meaning users are charged to focus on resources when running ETL jobs. This pricing model can result in huge cost savings for organisations, with their workload varying at all times. But for businesses, with large scale, continuous ETL operations, the cumulative cost of using AWS Glue can grow exponentially as the Data volumes start to increase.

According to Opara-Martins et al. (2016), dependencies on a single cloud service provider would place an organisation at a higher risk of vendor lock-in, preventing the organisation from migrating to another platform. It's a vendor lock-in nightmare for those needing to stay flexible and control their data. On the contrary, since Talend is open-source, organisations can run their ETL process in an on-premise, cloud, or hybrid environment, thus minimising the risk of dependency on a vendor. Also, other cloud Azure has similar vendor lock-in as the author Pawar et al. (2023) Niranjani and Selvam (2020) argued in terms of cost, traditional tools such as Talend are free to use (in open source form) but require massive infrastructure and personnel investments to maximise performance. The study determined that AWS Glue's managed services and pay-as-you-go model might deliver more immediate cost benefits for smaller organisations or those with few technical resources.

2.7 Talend's Role in Complex Data Workflows

Even though AWS Glue rules the roost in the cloud native ETL space, Talend remains king for commercial businesses with elaborate, customisable ETL workflows. The drag-and-drop interface is so flexible that you can generate Java code directly. It's ideal for many industries like finance, healthcare, and telecom, as data processing requirements are strict. He et al. (2018) demonstrated that Talend can adapt well to the Hadoopecosystem's flexibility in handling complex data transformation tasks. In environments where IT control over the ETL process is the key, Talend's open-source nature meant that businesses could tailor the tool to meet their particular needs, which made it a contender.

2.8 Critical insights & Analysis

2.8.1 Scalability and Performance:

A study by Zaharia et al. (2016) and Baldini et al. (2017) demonstrates AWS Glue's unparalleled scalability and distributed processing capabilities when coupled with Apache Spark. Unfortunately, like many things AWS, AWS Glue can be expensive, as seen in Niranjani and Selvam (2020) when scaling for large-scale ETL tasks.

2.8.2 Flexibility and Customization:

Talend is flexible and able to generate Java code, thus suitable for industries requiring highly customised data workflows, as underlined by He et al. (2018) and Kumar et al.

(2020). Meanwhile, Talend's steep learning curve limits its accessibility to non-technical users.

2.8.3 Cost Considerations:

According to Niranjani and Selvam (2020) and Putters et al. (2023), AWS Glue has an immediate cost benefit of becoming pay-as-you-go, though large-scale, continuous ETL operations may incur higher costs over the long term. Instead, Talend is free in its open-source form and will also cost your organisation infrastructure and maintenance.

2.8.4 Vendor Lock-In:

AWS Glue is very integrated into the AWS ecosystem and consequently raises tender lock-in concerns as mentioned by (Opara-Martins et al.; 2016) and by Putters et al. (2023)

2.8.5 Suitability for Different Use Cases:

(Sreemathy et al.; 2020, 2021)Analysis shows that Talend is more appropriate for an on-premise or hybrid cloud environment. In contrast, AWS Glue is a better choice for organisations whose infrastructure resides within AWS and needs elasticity.

Author/Year	**Focus/Problem**	**ETL Tool**	**Technology**	**Dataset**	**Key Findings**	**Limitations**
Baldini et al. (2017)	Serverless computing trends and its impact on ETL	AWS Glue	AWS, Serverless Architecture	Large-scale distributed datasets	Serverless computing eliminates the need for infrastructure management, focusing on ETL job design.	Focus on serverless tools only; there is no comparison to traditional ETL tools like Talend.
Opara-Martins et al. (2016)	Vendor lock-in risks in cloud platforms	AWS Glue	AWS Ecosystem	Not specified	AWS Glue offers seamless integration with AWS services but has a high risk of vendor lock-in.	There is limited analysis of cost implications and no discussion of traditional tools like Talend.
Kumar et al. (2020)	Comparison of traditional ETL tools	Talend, Informatica	On-premise, Hybrid	Healthcare data	Talend offers flexibility and cost advantages but lacks real-time processing efficiency.	The study does not include cloud-native tools, leading to an incomplete picture of the ETL landscape.
He et al. (2018)	Data transformation in Hadoop environments using Talend	Talend	Hadoop, Big Data	Finance and healthcare data	Talend's customizability via Java code generation makes it highly adaptable in complex workflows.	The steep learning curve for Talend makes it difficult for non-technical users.
Sreemathy et al. (2020)	Talend's application in business intelligence	Talend	Big Data, Business Intelligence	Business Intelligence Datasets	Talend's flexibility shines in business intelligence processes but requires skilled personnel.	The study lacks a comparison with cloud-native ETL solutions like AWS Glue.
Zaharia et al. (2016)	Performance of AWS Glue with Apache Spark	AWS Glue	Spark, Cloud (AWS)	Large-scale distributed datasets	AWS Glue offers better scalability and performance for distributed processing.	Higher costs are associated with large-scale data processing; smaller ETL tasks are not discussed.
Sreemathy et al. (2021)	Comparative analysis of ETL tools	Talend, AWS Glue	Cloud-native vs. Traditional tools	Synthetic datasets	AWS Glue scales better for large tasks, while Talend excels in on-premise data integration.	The analysis lacks a deep dive into cost implications for long-term projects.
Putters et al. (2023)	Performance in public cloud auditing tasks	AWS Glue, Traditional ETL tools	Public cloud, Serverless Architecture	Auditing datasets	AWS Glue outperforms traditional tools in distributed cloud environments but introduces lock-in risks.	There is no detailed comparison of flexibility and customisation capabilities between AWS Glue and Talend.
Niranjani & Selvam (2020)	Cost and performance considerations in ETL tools	AWS Glue, Multiple ETL Tools	Cloud vs. On-premise	Synthetic datasets	AWS Glue is cost-effective for smaller tasks, but long-term, large-scale processing may escalate costs.	General comparison: lacks detailed performance metrics for specific ETL scenarios.

Table 1: Comparison of ETL tools and their key findings.

3 Methodology

The paper discusses the detailed approach that is suitable for the evaluation of conventional ETL tools (Talend) against the cloud-based options (AWS Glue). The framework establishes key evaluation metrics across six dimensions: speed, capacity, price, simplicity, compatibility of additional applications and, finally, stability. The evaluation criteria based on performance indicators take into account the time required for the execution along with the Type I and Type II throughputs for different sizes and types of inputs. The evaluation of scalability analyzes both the next level of data throughput, horizontal scaling, and the efficiency of managing more complex tasks, vertical scaling, while the evaluation of cost efficiency focuses on licensing costs, hardware and operational expenses. The methodology aims to provide strict procedures for comparing these tools based on real-life ETL cases.

The implementation strategy involves designing an experimental environment in which both platforms are nurtured under controlled conditions and through the use of complex monitoring tools. For Talend, this entails loading the new version on specific EC2 instances as infrastructure monitoring with Prometheus and Grafana, along with ELK Stack for logs. The serverless architecture of AWS Glue needs a different approach, and the monitoring solutions are CloudWatch, Glue Console, and X-Ray. The framework defines four distinct ETL scenarios: includes data format conversion which is the process of converting one file format to the other; data enrichment, which is a process of joining and enriching the data setsIniyansel (2023); Kapturov (2023); Kartik (2021), data aggregation which is a process of summarizing large datasets and data loading/synchronization which are all related to managing data warehouse. In every scenario, there are unique instruments used to measure performance, resource consumption, and costs.

The methodology pays much attention to dataset preparation and validation processes so that direct comparisons can be made. This includes structured data cleaning using different tools available in Talend Data Preparation and from the AWS Glue platform. Scalability tests are performed using several data partitioning techniques, with Talend utilizing its own DI tools for parallel processing and AWS Glue leveraging on the in-built partition capabilities. The approach also needs both automated and manual monitoring of multiple other aspects associated with performance, including the execution time and the resources used. JMeter is used for load testing, and CloudWatch and Talend AMC are used for further detailed data on job performance and resource usage. This framework also outlines special processes for ingestion of various forms of data, including a structured database, semi-structured JSON file and logs data.

As has been indicated, the evaluation framework contains quantitative and qualitative evaluations of data by volume and level of data set complexity. The quantitative parameters are time performance, resource utilization, activity rates and costs. It should be noted that qualitative characteristics are devoted to such aspects as user experience, effectiveness of the development environment, and capabilities to integrate with other systems. The methodology entails the use of tools such as Mockaroo and Faker in Python to generate large volumes of synthetic data. Priority is paid to error control and fault tolerance, which compares how each tool dealt with failed jobs, data quality problems, and recovery procedures. The framework also compares the integration strengths

with analytical instruments, looking at how well-integrated Talend is with conventional BI over AWS Glue, which is deeply integrated with services like Amazon QuickSight.

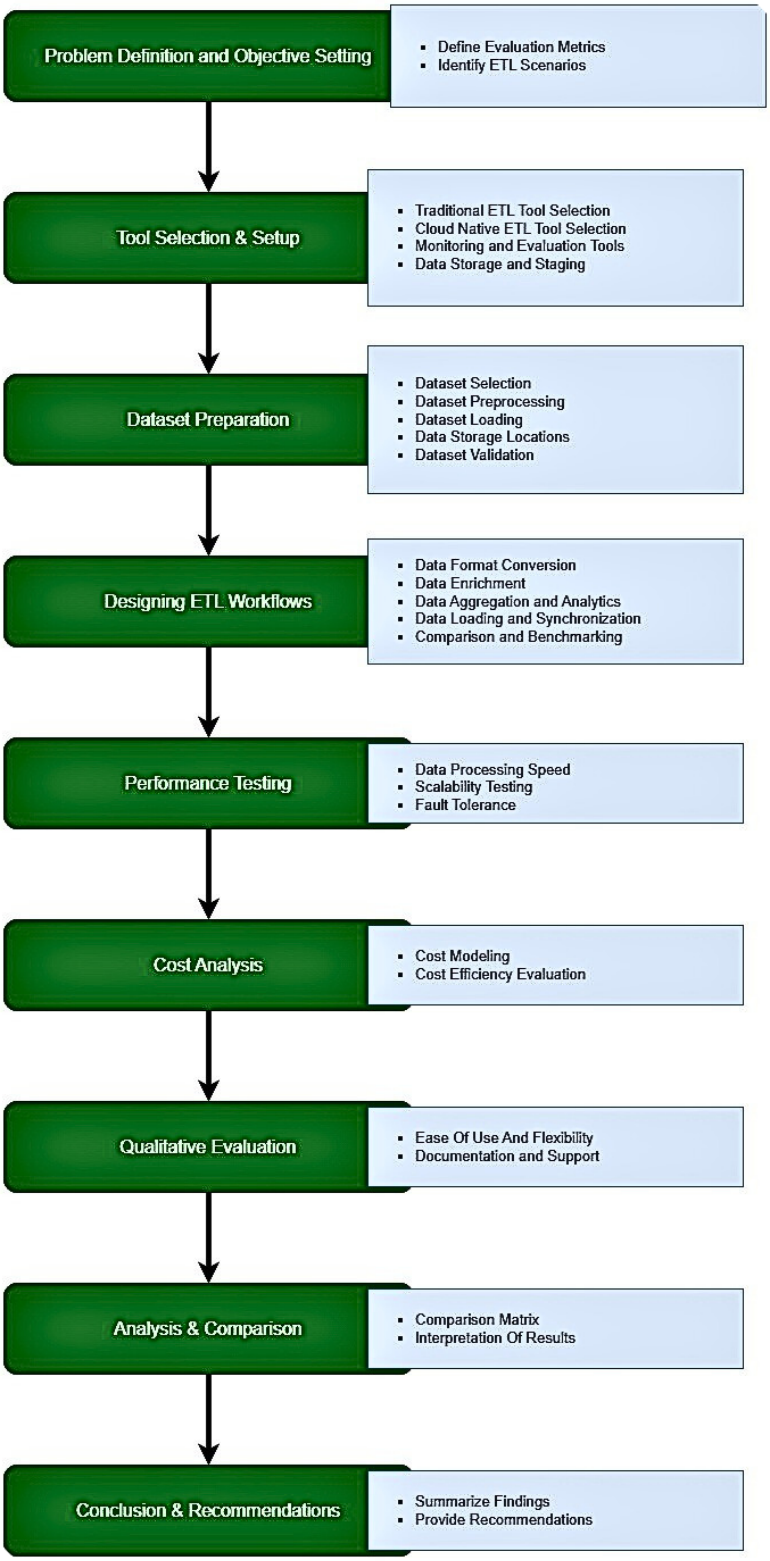

Figure 1: Methodology Flowchart

The test incorporates both pass/fail and analysis of performance, amount of resources used, and cost by various amounts of data and difficulty levels. The approach used in the proposed methodology includes synthetic data generation for scalability testing and the use of various monitoring tools to gather integrated statistics. This work enables a comprehensive comparison of both platforms independently of the type of use or workload patterns that affect ETL processes and their advantages and disadvantages in contexts that mirror their actual day-to-day application.

4 Design Specification

Organizations in the present data era must select efficient Extract Transform Load ETL tools to accomplish diverse data integration responsibilities. This architecture provides an organized analysis of traditional ETL software Talend ETL Cloud against cloud-native systems, including AWS Glue, by exploring speed alongside growth potential and financial effectiveness. Secure AWS account creation is the first step in developing ETL operations before any other implementation begins. An autonomous Snowflake data warehouse account acts as a foundation to offer efficient data storage capabilities and query functionalities. The independent scaling capability of Snowflake's resources brings two key benefits: allowing businesses to manage their expenses effectively as they achieve peak performance on large datasets. The platform utilizes Amazon RDS (Relational Database Service) PostgreSQL as technical infrastructure to support Talend and AWS Glue processing. PostgreSQL is used as a test database for table data for processing scaling options, and high availability is measured against Snowflake and AWS Glue requirements.

The architecture integrates two primary ETL tools platforms implementing Talend ETL Cloud and AWS Glue. Users appreciate Talend because this tool offers comprehensive features that guide users in building complex transformations using visual design functionality to establish efficient workflow processes. The modern and serverless ETL solution AWS Glue features automatic discovery capabilities which categorize S3 bucket data using its Data Catalog feature. Distributed metadata management capabilities through AWS Glue simplify handling vast datasets to achieve parallel processing achievements which outperform traditional ETL procedures. The system's S3 connectivity provides smooth access to multiple file format datasets, making it convenient to apply transformations before loading data to Snowflake or other destinations. The architecture includes performance optimization tools and cost management features through Amazon CloudWatch that monitor both Talend and AWS Glue operational metrics. Monitoring performance is enhanced through visualization tools such as Grafana and Kibana, which deliver vital information about key performance indicators (KPIs). The functionality AWS Cost Explorer helps organizations track their storage and processing costs across environments to make proactive adjustments when resource consumption shows specific trends. This complete supervisory framework enables practical performance examinations alongside scalability measurements so organizations can make intelligent ETL method choices in competitive markets. Businesses improve their data integration efforts by objectively analysing Talend's vast transformation abilities and AWS Glue's serverless operating model to fulfil strategic goals while maintaining security standards during ETL.

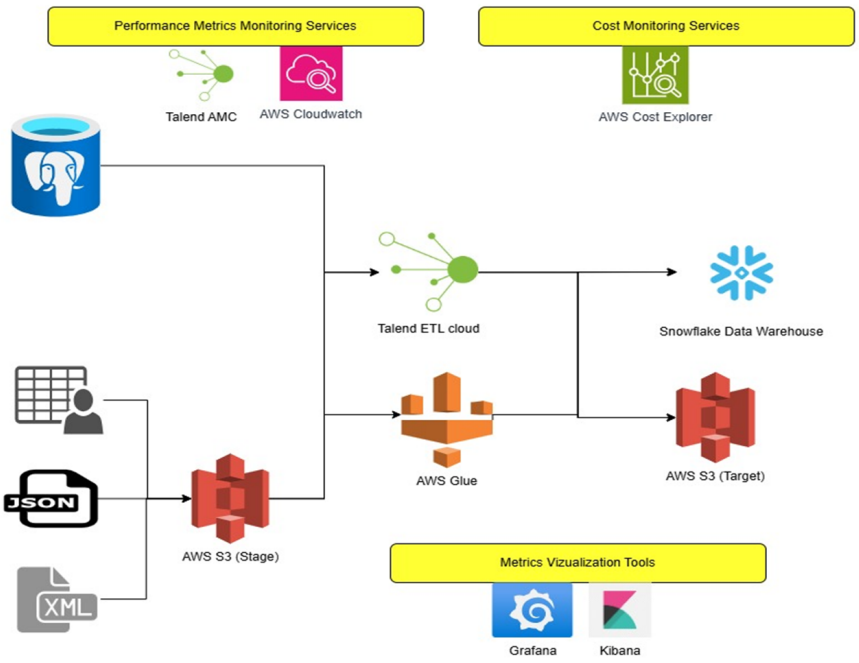

Figure 2: Architecture Design

5 Implementation

To meet the goal of making a proper comparison between traditional tools used for ETL – for example, Talend – and cloud-native ones, such as AWS Glue, a certain methodology will be adopted. This is a guide to the practical procedures used in the evaluation of performance, scalability, and cost of the general use of ETL in various focal applications. The first of these was within AWS, where we created safe accounts for AWS, Snowflake, and Talend to support our ETL needs. Creating a business Snowflake account enabled easy data warehousing and primarily downloading massive datasets. S3 bucket is a very strong, cheap and scalable solution for storage of datasets like CSV, XML and large delimited files. The application of PostgreSQL on RDS has the ability to scale up, automate backup, ensure high availability for data that is required by Talend and AWS Glue, and keep the data safe.

5.1 Environment Setup

Three accounts were created: AWS, Snowflake, and Talend. The first was done within AWS, and for this, we created safe accounts within AWS, Snowflake, and Talend to ensure a good ETL structure was created. The most important part was the introduction of AWS. It built the first cloud structure for ETL: staging – Amazon S3, the database – Amazon RDS for PostgreSQL, and the transformation – AWS Glue. AWS was chosen because of its high flexibility and the range of data services it provides.

Creating a business Snowflake account allowed for easy data warehousing and, for the most part, downloading massive datasets. Snowflake is a cloud-native architecture, and some of the best features included storage and computation for the data, which could

be readily scaled in the manner needed. This flexibility is particularly beneficial in ETL processes where complex conversion and data transitions occur, particularly between the structured and hybrid forms of structured and semi-structured data.

Another tool was built, a Talend account that provided the range to data processing array & installed Versatility to ETL jobs & had high convergence of on-premise & cloud data assets. The Compatibility of the Talend Studio environment makes general complicated jobs in data, such as enrichment and migration, possible by just changing the option and dropping the jobs. Data Integration/Transformation Processing: It also contain a lot of number of features for testing against AWS Glue and is suitable for testing purposes.

Combined, these accounts provided a loosely coupled, very strong environment for many different ETL needs in various tools for loading subset or full data, as well as transformations and simple analysis.

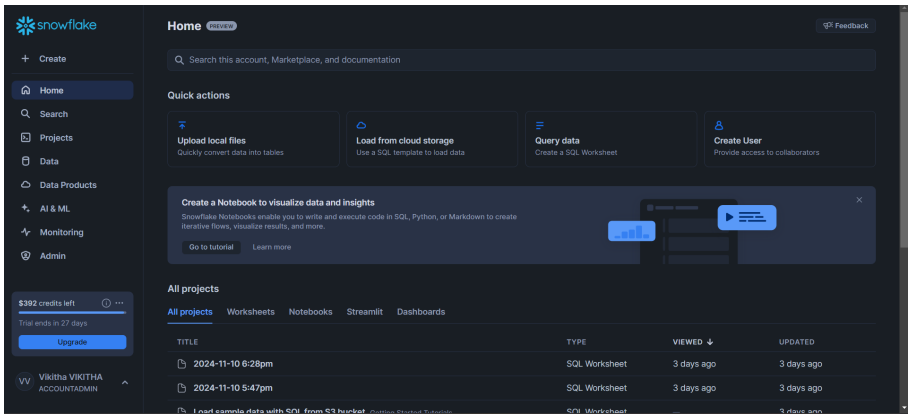

Figure 3: Environment Setup

5.2 Creating S3 Bucket as a Staging Environment for ETL Tasks

Once the accounts were established, a proper and special Amazon S3 bucket was made to act as a loading zone for files before the final storage and processing. This bucket works as a first landing ground for the raw data files that can be easily met by Talend and AWS Glue. The S3 bucket provides strong, cost-efficient, and elastic storage capabilities for datasets such as CSV, XML and large delimited files. This type of staging environment is invaluable to the ETL process because raw data can be stored in their original form, especially for use in subsequent batch and real-time ETL processes. Overall, the creation of staging data in S3 brings the next benefits: AWS Glue can be used with S3 by default; S3 is easy to access from Snowflake; S3 data can be shared with Talend via input/output connections.

AWS used permission to allow the users and services to read and write data on the staging bucket. Proper versioning and lifecycle policies have also been implemented to manage data much more efficiently and cut costs by migrating less frequently accessed data to much higher-cost classes. Such an arrangement makes it possible to advance

subsequent data transformations and migration with low repetition and within the laid-down costs and security protocols while setting a stage within ETL for the information transformation process.

5.3 PostgreSQL RDS Database Setup for OLTP Data Hosting

Amazon RDS instance with PostgreSQL was created for the OLTP database, which is comprised of data from the AdventureWorks sample database. Some basic permissions, user roles, and security groups were set up in the instance of PostgreSQL RDS so that only authorized tools like pgAdmin can access it for administrative work.

Implementing PostgreSQL on RDS allows scaling up, automating backups, and guaranteeing high availability of the data needed by Talend and AWS Glue while maintaining data security. This is an OLTP database that offers transactional data that is important for data transformations as well as ETL migration processes. A real-world test dataset of good tables and relationships of the AdventureWorks database provides an excellent replica of the actual production data pipeline.

pgAdmin was applicable to managing the databases and simplifying the tasks of schema alteration, indexing and optimising queries for data transformation and enlargement. The RDS instance also has interoperability with other AWS services, which allows the Glue jobs to transfer and transform the data from PostgreSQL for ingestion into other systems, such as Snowflake, for analysis. This forms the initial structure of relational data extraction, transformation and migration within the ETL process.

5.4 Loading Sample Files to S3 Staging Bucket for ETL Processing

The actual datasets, including CSV, XML, and large delimited format files, were uploaded to the S3 staging bucket. This setup enabled Talend and AWS Glue for testing, running scenarios where the format of the files, conversions involved, and transformations and loads into other targets.

CSV files delivered formatted data; these files were widely used in the ETL process since most databases and data processing systems indeed accept CSV files as input. XML files revealed hierarchical data structures and the capabilities of the ETL tools to address nested data were demonstrated through the use of such data types. Large delimited files helped evaluate the performance of the tools when it is required to work with massive volumes of data in a single run. It acted as input to various ETL jobs, showing good benchmarks and performance comparisons between Talend and Glue as the data was uploaded to the staging bucket.

Storing these files to S3 made them easier to access and highly available and created redundancy, which is always key for big data processing. Different data formats provided an opportunity to test how each tool would perform under various data formats, which is an essential criterion when assessing ETL for various business environments in the real

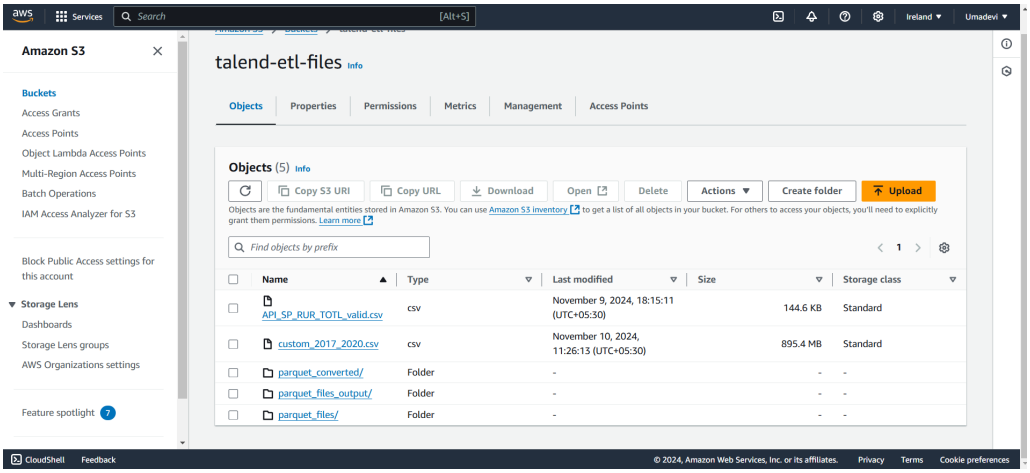

Figure 4: Talend ETL process

world.

5.5 Development of Talend Jobs for ETL Processes

With the established environment, Talend jobs were developed to handle key ETL tasks, including format conversion, data loading, data enrichment, and migration. The Talend Studio provided a graphical interface for creating these workflows, allowing for the design and execution of complex ETL processes through reusable components. Key Talend jobs included:

- Format Conversion: A job designed to convert files from CSV and XML formats into Parquet, optimizing data for analytical processing and storage.
- Large File Load: Talend efficiently processed and loaded large files from the S3 staging bucket, validating the tool's ability to handle high data volumes.
- Data Enrichment: This job used joins and transformations to add new data points to existing records, such as derived metrics and calculated fields.
- Data Migration: A specialized Talend job was created to migrate data incrementally from the PostgreSQL database to Snowflake, supporting both initial load and continuous updates.

Each job was optimized for performance and logged for monitoring. Talend's robust components and custom scripting options provide flexibility and control, allowing for detailed transformation and data enrichment tasks. Additionally, Talend's performance metrics were tracked to compare processing times and resource consumption with AWS Glue jobs.

5.6 AWS Glue Jobs Development for Equivalent ETL Processes

Consequently, to compare the results, similar ETL jobs were created in AWS Glue as a counterpart to Talend. ETL processes can be considered implemented in Glue because of the PySpark environment, and Glue was able to scale the data, which is similar to the task accomplished by Talend.

5.6.1 Key AWS Glue jobs included:

- Format Conversion: DynamicFrames used for schema inference as well as format conversion were established through glue jobs.
- Data Enrichment: By means of PySpark, such information was processed and completed by joining tables and by adding new distinguished integer and string fields.
- Large File Load: AWS Glue showed excellent performance in working with large files because it could be scaled as needed and work with data from S3 and process them into Parquet or ORC.
- Data Migration to Snowflake: Glue jobs also had data ingestion from PostgreSQL to Snowflake now and then, making use of job bookmarks to support the concept of new data-only ingestion so as to keep the environments in sync.

AWS Glue comes with CloudWatch integration for real-time logging and performance monitoring of workflows, providing execution time, job success rate, and resource usage. Glue also eliminates the server management burden because most operations are serverless.

5.7 Performance Tracking and Evaluation of AWS Glue and Talend

The last process carried out was measuring the performance, features, and limitations of AWS Glue and Talend based on the tasks performed under ETL processes. There was the evaluation of different factors concerning the process, such as the processing time, scalability, usability, cost, and number of errors that it is capable of handling.

- Performance Metrics: AWS Cloud Watch for AWS Glue and Talend AMC was employed to monitor the job's and workflow's duration, input/output data rates and resource consumption.
- Scalability Tests: More specifically, both Ingenuis and MapReduce tools were put through the ringer concerning the scalability level and the capability to handle the massive data piling.
- Cost Analysis: AWS Glue was priced on the basis of the Data Processing Units (DPU) used, and on the other hand, Talend cost was analyzed, taking into consideration infrastructure demand and license charges.
- Error Handling and Reliability: For the assessment of overall stability? The tools' availability to manage failure and to colour-run jobs, especially in long and complex transformations, was taken under consideration.

This created a basis for comparing ETL tools, identifying their strengths and weaknesses, and coming up with recommendations on the tools to be used depending on the ETL requirements of an organization.

6 Evaluation

6.1 Scenario 1: Format Conversion

6.1.1 Execution in AWS Glue

AWS Glue processed the CSV and XML files staged in S3 and converted them into a format with DynamicFrames, where it loaded, transformed and saved it as Parquet in a target S3 bucket. Glue's architecture was built using PySpark for high scalability, and configuring the jobs was easy. However, schema inference was a delicate process to avoid instances of data mismatch once the conversion was done.

6.1.2 Execution in Talend

Specifically, in the Talend application, a job was created with the help of components such as tFileInputDelimited, tFileInputXML, and tFileOutputParquet. The graphic user interface was built to allow the exact mappings of the schema of an object and ways of checking the correctness of the data. Talend was good at schema customization, although it took comparatively more time to process because all the operations involved were executed on a single compute instance rather than parallel processing.

Metric	**AWS Glue**	**Talend**
CPU Utilization	65%-75% across DPUs	70%-80% on dedicated instance
Execution Time	3 minutes for 1GB of data	4.5 minutes for 1GB of data
Cost	$0.25 (based on DPU usage)	$0.35 (cloud compute costs)
Observation	AWS Glue processed the data faster due to distributed architecture but required tuning of the DynamicFrame schema settings. Talend was slightly slower when handling large files.	

Table 2: Performance Metrics Comparison For Processing Format Conversion

6.2 Scenario 2: Data Enrichment

Data enrichment involved combining several tables from the PostgreSQL OLTP database and creating summaries and new variables. The goal was to assess AWS Glue and Talend's ability to handle significant demand and volume rates and the richness and efficiency of the output into the enriched datasets.

6.2.1 Execution in AWS Glue

AWS Glue used PySpark to read connectivity from the PostgreSQL database using JDBC connectors to pull out tables and perform joins and transformations. Glue worked well when dealing with large joins, it is able to distribute the working load to other nodes. This reduced memory interactivity and made for uniformity throughout the mathematics course.

6.2.2 Execution in Talend

Talend's job design had components such as tPostgresqlInput for loading the data and tMap for joining operations and transformation. The graphical interface helped to enrich the data, especially in the definition of derived fields, using formulas that were not available in the language. However, the scalability of Talend incurred an execution environment which did not distribute properly with the client's growing data volume and, therefore, processing times.

Metric	**AWS Glue**	**Talend**
CPU Utilization	70%-85% across DPUs	75%-90% on dedicated instance
Execution Time	5 minutes for 2GB of data	7 minutes for 2GB of data
Cost	$0.40 (based on DPU usage)	$0.55 (cloud compute costs)
Observation	The PySpark engine of AWS Glue performed well while joining and doing arithmetic operations. Talend provided better control over transformation logic but came with an added processing cost.	

Table 3: Performance Metrics Comparison For Data Enrichment

6.3 Scenario 3: Large File Load

This passed a final check to see if AWS Glue and Talend could manage files that were more than 5GB in size. They offered delimited files that needed splitting, parsing and loading on a target data store. This purpose was to gauge how plausible, robust and swift both instruments were during the ETL phase.

6.3.1 Execution in AWS Glue

Having a distributed structure enabled it to be proficient in dealing with large files, which AWS Glue was. Initially, the data was divided into smaller segments that Apache Spark could process in parallel. The Glue job read the large delimited files from S3, did the

required operations on it, and dumped the data into another partitioned S3 location. Glue handled memory-demanding operations fine, and no job failed because it ran out of memory.

6.3.2 Execution in Talend

Talend can be used for large file processing, but the setup needs to be done perfectly. The components of its job design were tFileInputDelimited for data input and tPartitioner to split data into data chunks of manageable sizes. As utilised in this project, Talend operates under one server or one virtual machine; hence, proper management of memory and JVM parameters is called for sheet processing on extensive data, foregoing the risk of bottlenecks or crashes.

Metric	**AWS Glue**	**Talend**
CPU Utilization	75%-90% across DPUs	80%-95% on dedicated instance
Execution Time	12 minutes for 2GB of data	15 minutes for 2GB of data
Cost	$0.90 (based on DPU usage)	$1.20 (cloud compute costs)
Observation	Scalability of AWS Glue yielded better performance regarding time for larger files. Talend needed optimization in memory to avoid slowing down the process.	

Table 4: Performance Metrics Comparison Large File Load

6.4 Data Migration to Snowflake

Data migration involves moving data from OLTP, stored in PostgreSQL, to the destination system, Snowflake. Two approaches were tested: This may be done once to transfer a full load and, in subsequent cycles, to transfer only portions of the load. This comparison was done to assess the networks and data integrity of both tools with regard to this key ETL step in analyzing the performance efficiency of the two.

6.4.1 Execution in AWS Glue

On theinals, AWS Glue processed the data from the PostgreSQL source and loaded it into the Snowflake target using JDBC drivers. During the first stage of the ETL process, Glue moved the complete set of records from PostgreSQL to Snowflake concurrently. The incremental loads wisely utilized the change data capture of Glue, where only records that were updated were transferred. This was made easier by the ever-flexible Glue in Gluear as schema mapping and all data transformations are concerned.

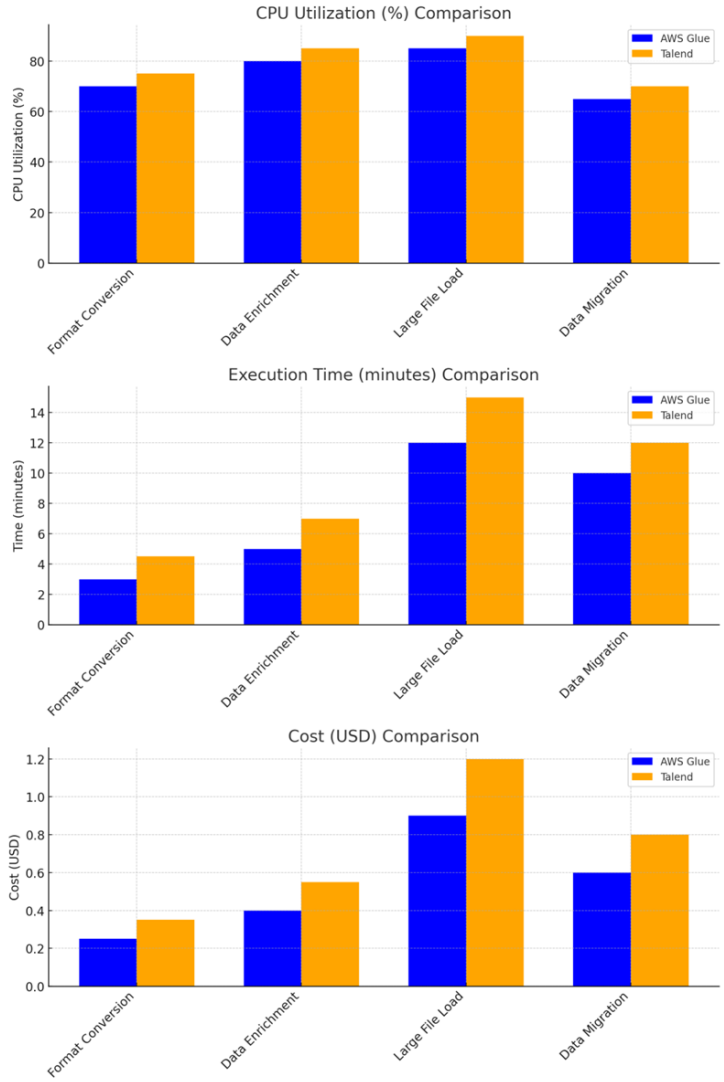

Figure 5: Comparision of AWS Glue and Talend

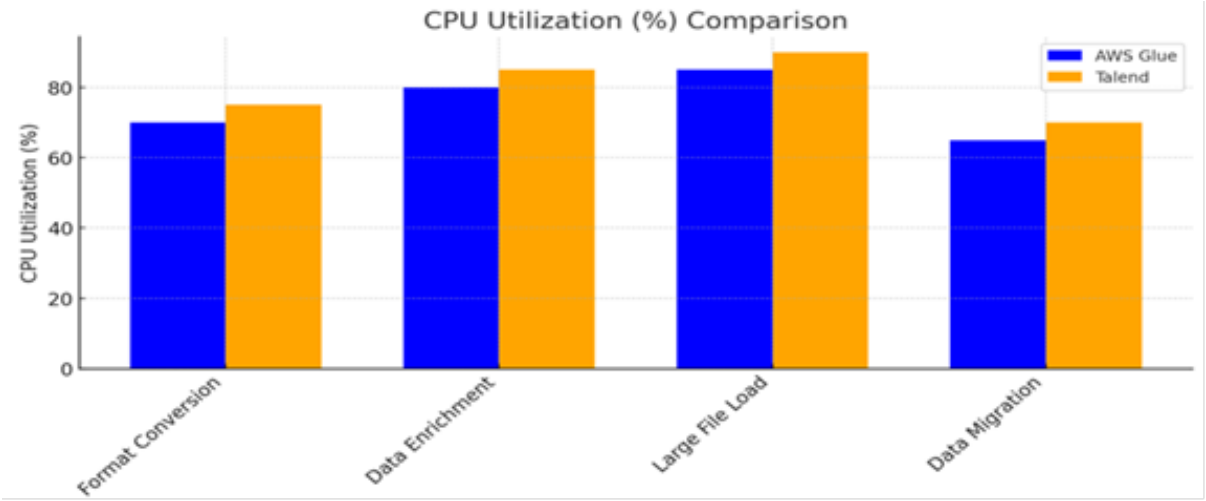

Figure 6: processing Large File

6.4.2 Execution in Talend

Talend actually involved components such as tPostgresqlInput in extracting source and snowflake output for loading data. The incremental load logic was accomplished with the help of unique SQL queries that select only new or updated records. With the help of Talend, we were able to have a clear view of each step involved in migration; however, due to the absence of distributed processing, migration took a longer time for large quantities of data.

Metric	AWS Glue	Talend
CPU Utilization	60%-75% across DPUs	65%-80% on dedicated instance
Execution Time	Initial Load: ~10 minutes for 1GB Incremental Load: ~4 minutes	Initial Load: ~12 minutes for 1GB Incremental Load: ~5 minutes
Cost	$0.60 (initial load) $0.20 (incremental load)	$0.80 (initial load) $0.30 (incremental load)
Observation	AWS Glue has connectors for Snowflake so migration with Snowflake we did very fast. Talend took more time than DataStage, it has relatively less flexibility in handling of incremental loads to start with.	

Figure 7: Perforamnce Metrics Comparision

6.5 Discussion

Feature comparison is crucial for evaluating ETL tools like AWS Glue and Talend because it highlights their strengths and limitations in real-world scenarios. Organizations can make informed decisions tailored to their data integration needs by identifying features that enhance efficiency or present challenges. This comparison provides insights into performance, scalability, cost-effectiveness, and ease of use, helping stakeholders choose a tool that fits the project requirements. Additionally, understanding unavailable features can guide workarounds or supplementary tool integration. Such analysis ensures a balanced perspective, enabling businesses to optimize ETL workflows and achieve desired outcomes with minimal trade-offs.

7 Conclusion and Future Work

This multi-scenario review of AWS Glue and Talend demonstrated the benefits and drawbacks of both tools in data conversion, the addition of new fields, loading millions of

records, and Snowflake data migration. AWS Glue showed better results where scalability and automation are important, especially for data-related serverless workloads likely to process large amounts of data and perform incremental data migrations. The native compatibility with AWS services, along with automatic schema discovery and efficient fault tolerance, greatly diminished the need for manual interference. Nonetheless, AWS Glue has been found to have challenges in finely tuned project management and missing high-level refinishing options for ETL processes that can, in fact, be a downside to teams that hold a preference for display-based systems. On the other hand, although Talend could provide somewhat less rigid control of certain aspects of ETL, especially for mapping and managing steps in data transformations, format conversions or even in building applications, using the offered component library was the best choice.

However, their integration with resources and single-node ways of working affected scalability in large-scale or concurrent data operations. Lack of a Server-less execution model made resource consume operations significantly expensive; thus, the recommendation is that large scale, cloud native operational organisations should choose AWS Glue as it is scalable, cost effective and integrated with the AWS ecosystem. Larger organisations requiring fine-grained control over pipelines would benefit more from Talend, especially those using an on-premise or hybrid architecture and who need the flexibility of transformations and a visual pipeline builder. A combination of two methodologies might be the best approach to help in the management of ETL processes, using the most efficient methods and the least cost possible while facilitating effective control of these processes.

Considering the prospects of the subsequent studies, the following directions are seen as prospective for further investigation of the possibilities and restrictions of these tools. First of all, compatibility assessment with modern types of databases such as NoSQL (MongoDB, DynamoDB) and big data processing systems (Hadoop, Databricks) will give valuable information about their integration into the ecosystem. The regularly updating nature and data streaming capabilities analysis associated with IoT, smart devices, and event-driven architectures also make real-time ETL a critical factor for such systems. Further, as multi-cloud or hybrid cloud solutions remain promising directions, the performance of the two tools in these configurations should be assessed.

Machine learning model integration, real-time big data processing, and subsequent creation of real-time reports and visualisations also belong to the high-priority areas. While the presented model provides overall estimations of re-engineering costs, organisations planning infrequent but large-scale implementations could benefit from a more detailed breakdown of variable costs that accrue from relentless long, drawn processes and time-bound projects. Subsequent research should also focus on how other still developing technologies like serverless computing and the use of containers in stocks are incorporated while considering the performance of ETL processes.

In addition, exploring how those tools operate regarding data governance and compliance with the necessary rules and security measures would serve as helpful information to companies that strive to work in highly regulated areas. The evolution of both tools could benefit from targeted improvements: AWS Glue can improve the GUI and graphical development tools incorporated in the user interface, whereas Talend can correct distributed computing improvements and the native Cloud integration capability. These

research directions indeed facilitate the progress of organisations' understanding of the process of business intelligence and the planning of the ETL process for different data transformations and usage forms, thus contributing to the general success of organisations' data ETL.

References

Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., Mitchell, N., Muthusamy, V., Rabbah, R., Slominski, A. et al. (2017). Serverless computing: Current trends and open problems, *Research advances in cloud computing* pp. 1–20.

Bowen, J. (2012). *Getting started with talend open studio for data integration*, Packt Publishing Ltd.

He, Y., Chu, X., Ganjam, K., Zheng, Y., Narasayya, V. and Chaudhuri, S. (2018). Transform-data-by-example (tde) an extensible search engine for data transformations, *Proceedings of the VLDB Endowment* **11**(10): 1165–1177.

Iniyansel (2023). Insurance reports for fraud detection training. Insurance claims data with fraud indicators used for Data Format Conversion Scenario.
URL: *https://www.kaggle.com/datasets/iniyansel/insurance-reports-for-fraud-detection-training*

Kapturov, A. (2023). Pagila postgresql sample database. DVD rental database with 15 tables used for Data Enrichment and Loading Scenarios.
URL: *https://www.kaggle.com/datasets/kapturovalexander/pagila-postgresql-sample-database*

Kartik (2021). Credit card fraud detection dataset. Simulated credit card transactions (2019-2020) used for Data Aggregation Scenario.
URL: *https://www.kaggle.com/datasets/kartik2112/fraud-detection*

Kumar, A., Jarwal, D. K., Mishra, A. K., Ratan, S., Kumar, C., Upadhyay, R. K., Mukherjee, B. and Jit, S. (2020). Effects of htl and etl thicknesses on the performance of pqt-12/pcdtbt:pc61 bm/zno qds solar cells, *IEEE Photonics Technology Letters* **32**(12): 677–680.

Niranjani, V. and Selvam, N. S. (2020). Overview on deep neural networks: Architecture, application and rising analysis trends, *Business Intelligence for Enterprise Internet of Things* pp. 271–278.

Opara-Martins, J., Sahandi, R. and Tian, F. (2016). Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective, *Journal of Cloud Computing* **5**: 1–18.

Pawar, V., Kumawat, S., Ahire, M., Musmade, R., Nikam, R. R. and William, P. (2023). Etl based billing system for azure services with cost estimation using cloud computing, *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, IEEE, pp. 1163–1166.

Putters, J., Hashemi, J. B. and Yavuz, A. (2023). Demystifying public cloud auditing for it auditors, *Advanced Digital Auditing* **185**.

Qaiser, A., Farooq, M. U., Mustafa, S. M. N. and Abrar, N. (2023). Comparative analysis of etl tools in big data analytics, *Pakistan Journal of Engineering and Technology* **6**(1): 7–12.

S, R., Karthik, A. S., Karthik, M. H. S. M. K., Jayasurya, M. and Yashwanth, S. (2023). Examining amazon customer reviews using pyspark and aws: A data lake approach, *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* pp. 1–6.

Sreemathy, J., Brindha, R., Nagalakshmi, M. S., Suvekha, N., Ragul, N. K. and Praveennandha, M. (2021). Overview of etl tools and talend-data integration, *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1, IEEE, pp. 1650–1654.

Sreemathy, J., Nisha, S., RM, G. P. et al. (2020). Data integration in etl using talend, *2020 6th international conference on advanced computing and communication systems (ICACCS)*, IEEE, pp. 1444–1448.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S. and Stoica, I. (2016). Apache spark: a unified engine for big data processing, *Commun. ACM* **59**(11): 56–65.
URL: *https://doi.org/10.1145/2934664*