National College of Ireland

Configuration Manual

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:	AWS Glue vs Talend: A Practical Comparison of ETL Tools
Page Count:	19

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:	16th December 2024

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):	

Configuration Manual

Vikitha Konda
x23175818

1 Introduction

For the implementation of the Comparison between the tools AWS Glue and Talend below steps need to be followed.

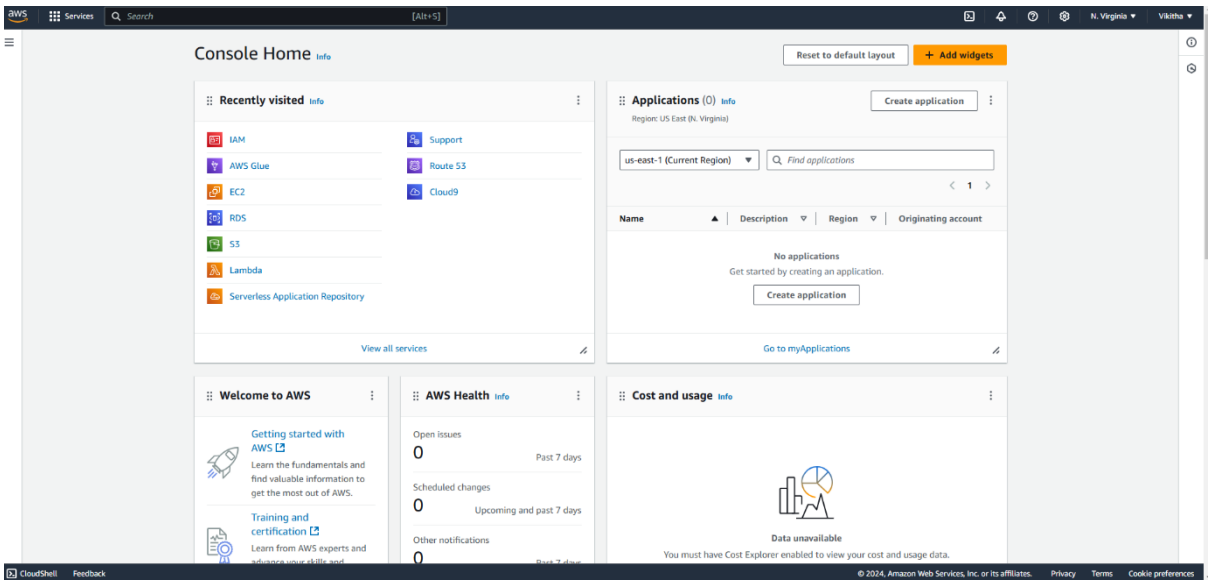

Figure 1: AWS Console Home

1.1 Create an AWS Account

1. Navigate to AWS Signup portal
2. Create account:
   - Provide billing information
   - Configure usage alarms
3. IAM Configuration:
   - Access IAM service in AWS Management Console
   - Create project-specific user
   - Assign AdministratorAccess permissions
   - Generate Access Keys

4. AWS CLI Setup:
    - Configure using generated access keys
    - Enable remote AWS service access
5. Enable CloudTrail:
    - Audit AWS API actions
    - Monitor project lifecycle activities

These foundational steps ensure:

- Comprehensive security
- Efficient resource utilization
- Proper monitoring of AWS solutions

2 Create a PostgreSQL RDS Instance

2.1 Instance Setup

1. Navigate to AWS RDS service page
2. Configure instance parameters:
    - Instance size
    - Disk specifications
    - Backup configuration
    - Recovery settings

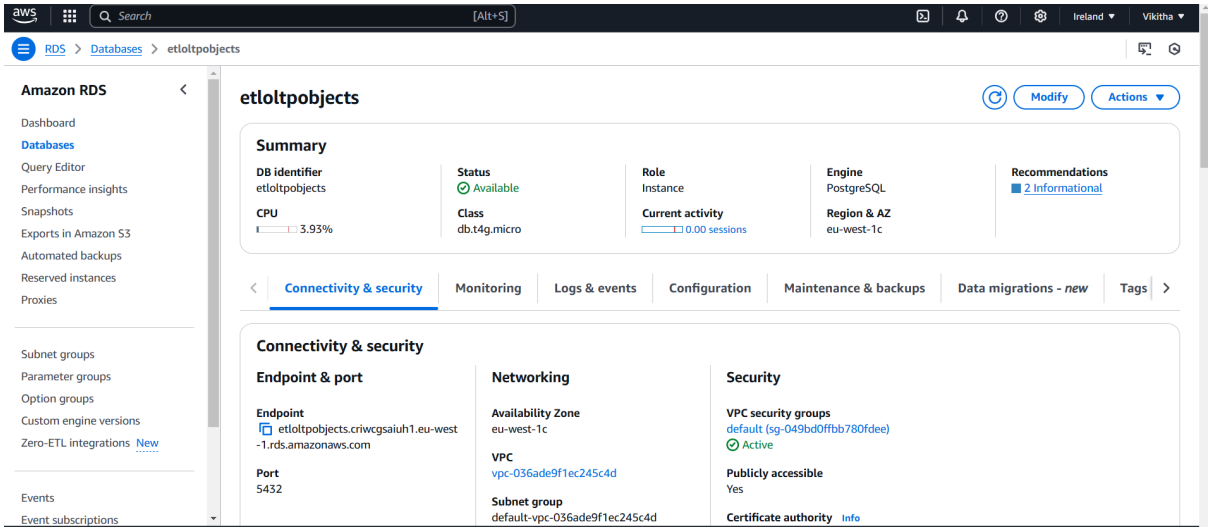

Figure 2: Database Configuration

2.2 Security Configuration

- Modify security group rules
- Enable external access from local system

2.3 Database Configuration

Database credentials:

- Database name: `adventureworks`
- User: `postgres`
- Password: `XXXXXXXX`

2.4 Post-Installation Steps

1. Install pgAdmin for graphical interface
2. Import AdventureWorks sample database
3. Verify:
    - Database connection
    - User access permissions
    - Data accessibility for ETL processes

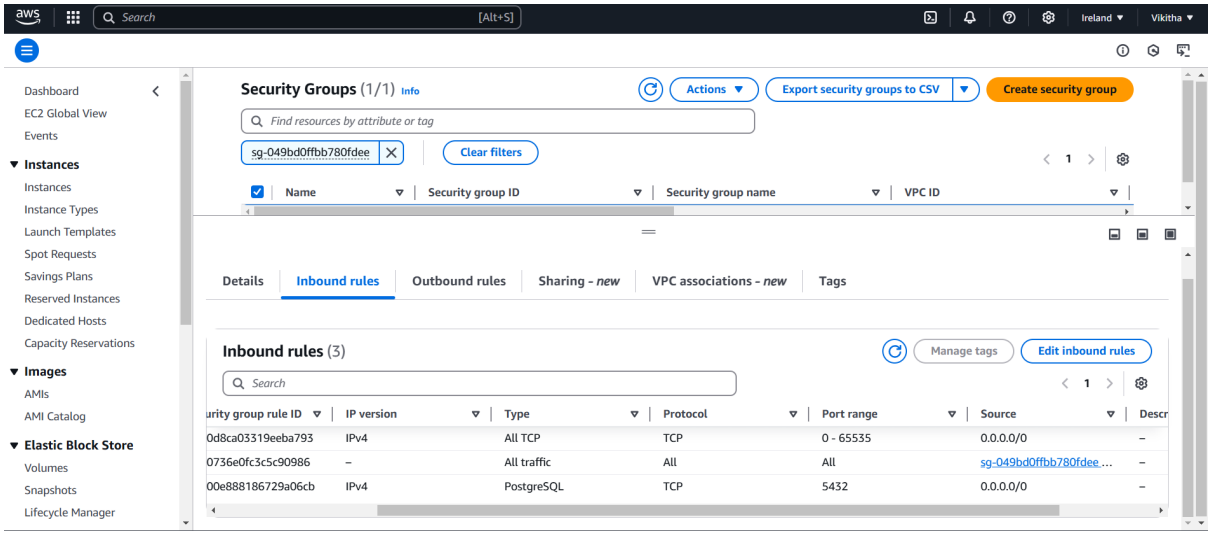

Figure 3: Security Group

3 Configuring pgAdmin for Schema and Data Setup

3.1 Installation Process

1. Download pgAdmin from official website
2. Configure connection parameters:
    - Database hostname
    - Username
    - Password
    - Port
    - PostgreSQL RDS instance details

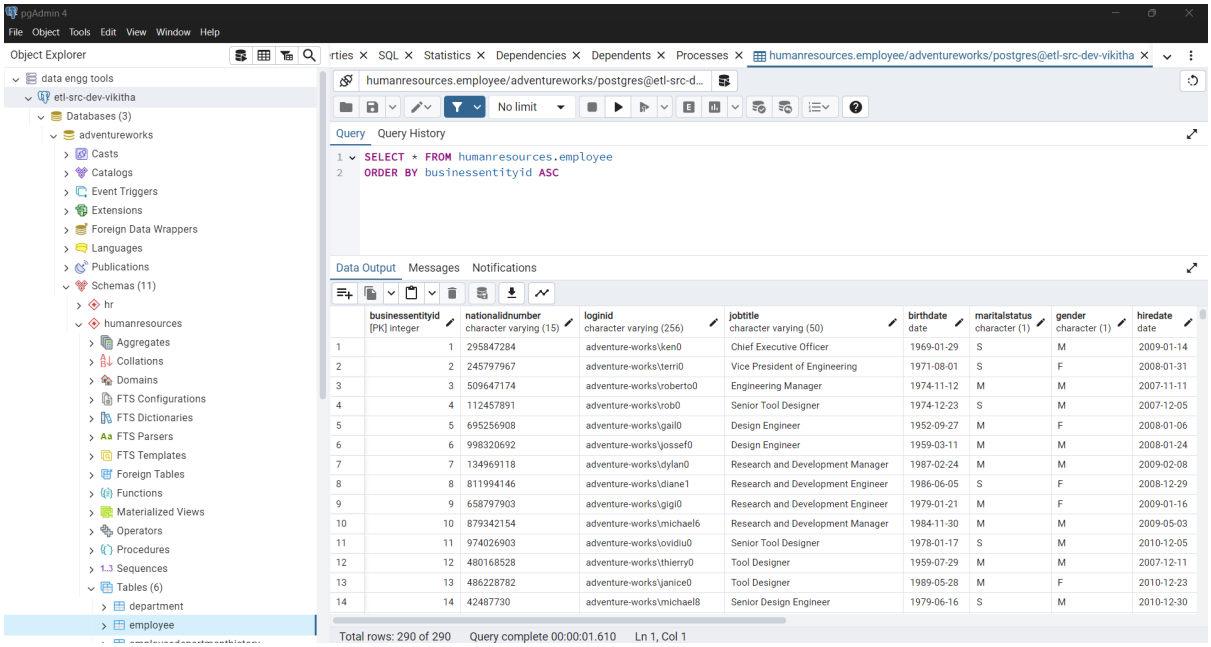

Figure 4: PgAdmin Setup

3.2 Schema and Table Creation

1. Connect to database
2. Create new schema:
    - Navigate to "Schemas"
    - Context-click
    - Select "Create Schema"
    - Example schema: Adventureworks
3. Create tables using either:

- SQL scripts
- pgAdmin interface

4. Example tables:
    - customers
    - orders
    - products

3.3 Data Import

Using pgAdmin's Import/Export function:

- Load CSV or delimited files
- Map to corresponding table fields
- Adjust as needed:
    - Delimiters
    - Formatting settings

4 Create an S3 Bucket

4.1 Initial Setup (20 minutes)

1. Access AWS Management Console
2. Navigate to S3 service
3. Create new bucket: `etl-staging-bucket`
4. Enable versioning for change tracking

4.2 Directory Structure

Create the following folders:

- `delimited_files/`
    - Purpose: Storing delimited source files
- `parquet_files/`
    - Purpose: Target folder for parquet transformed files
- `temp/`
    - Purpose: Temporary archival file storage
- `snowpipe_employee_details/`

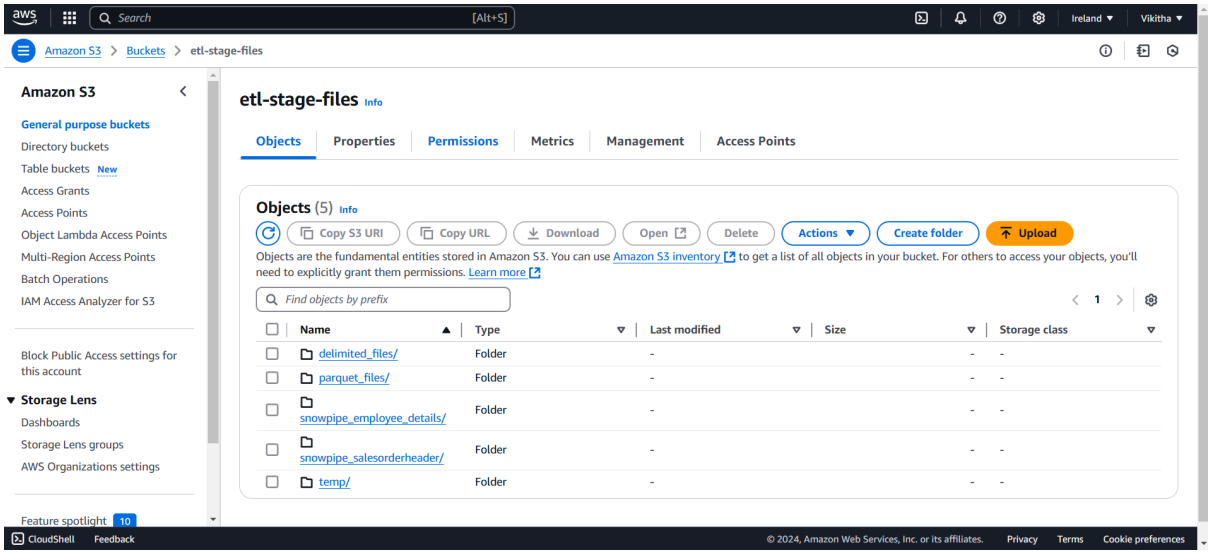

Figure 5: S3 Bucket Setup

  - Purpose: Staging datalake for employee details

- `snowpipe_salesorderheader/`
  - Purpose: Staging datalake for salesorderheader

4.3 Access Configuration

1. Configure bucket policies for:
   - AWS Glue read/write access
   - Talend job read/write access
2. Use AWS CLI for:
   - Sample file upload
   - Access permission testing
3. Upload test files:
   - CSV files
   - XML files
   - Large delimited datasets

5 Create A Snowflake Account

5.1 Account Creation

1. Register for Snowflake free trial
2. Select region closest to AWS environment to minimize latency

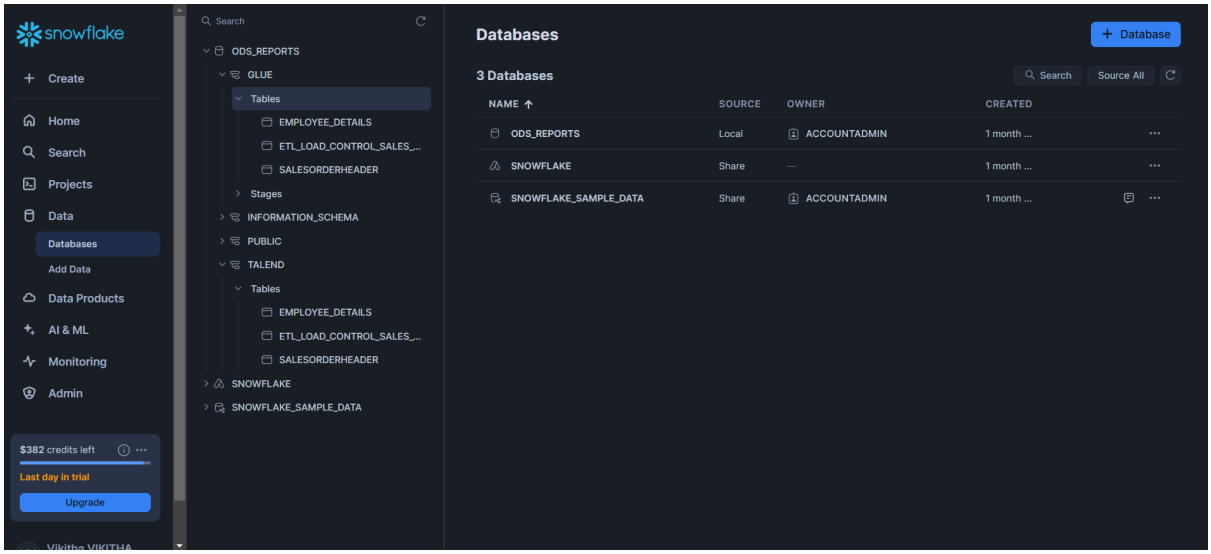

Figure 6: Snowflake Account Creation and Configuration

5.2 Environment Configuration

1. Create ETL-optimized warehouse
2. Configure database and schema
3. Install Snowflake JDBC driver
4. Set up appropriate access rights and roles

5.3 Connection Validation

Test connection using:

- SQL Workbench
- Snowflake web interface

5.4 Database Structure

Database Name: `ODS_REPORTS`
Schemas:

- `GLUE` - For AWS Glue loaded tables
- `TALEND` - For Talend ETL loaded tables

6 Talend Installation

6.1 Prerequisites

- Download Talend Open Studio or licensed version from official website
- Install Java Development Kit (JDK) as per requirements

6.2 Installation Steps

1. Select installation directory
2. Follow standard installation procedures
3. Open workspace
4. Create new project

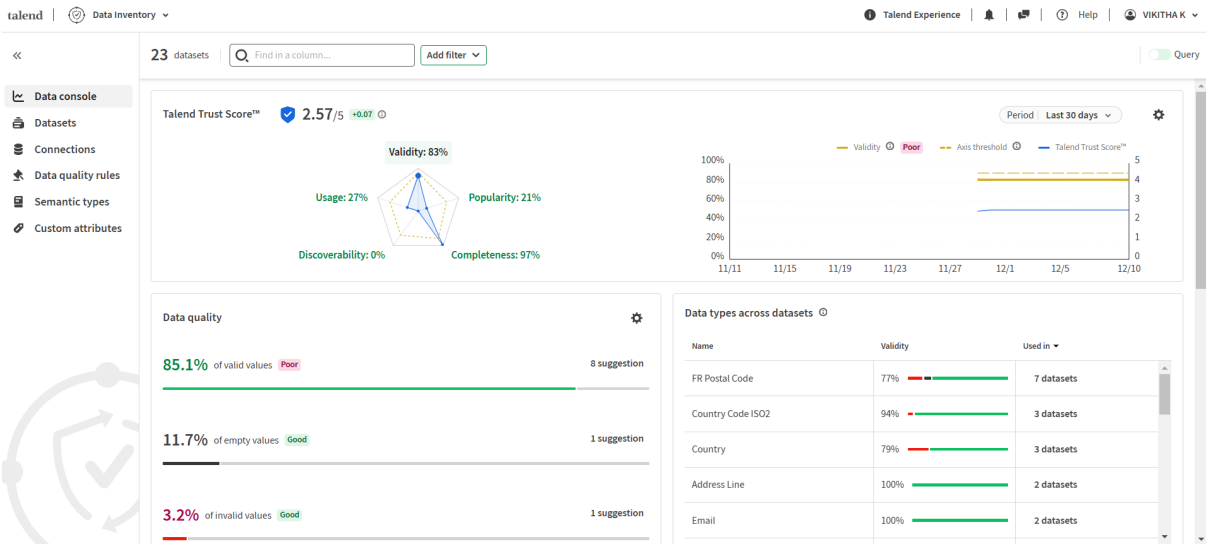

Figure 7: Talend Installation and Configuration

6.3 Configuration

Set up global variables for connections:

- Amazon S3
- PostgreSQL
- Snowflake

6.4 User Interface

Features:

- Graphical user interface for data integration
- Component-based development
- Visual pipeline design
- Complex pipeline tuning capabilities

7 Setting Up Grafana for Monitoring and Comparison

7.1 Installation Process

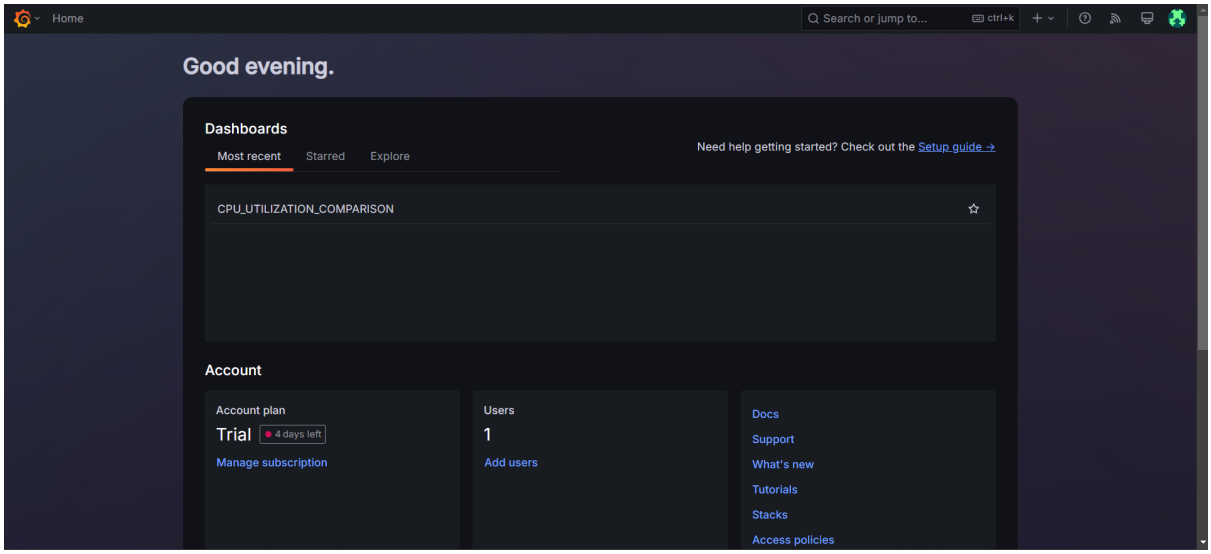

Figure 8: Grafana setup

1. Download Grafana from official website
2. Install based on operating system requirements
3. Alternative: Deploy using Docker container
4. Start Grafana server
5. Access web interface: `http://localhost:3000`

7.2 Data Source Configuration

7.2.1 AWS Glue Metrics

- Install Grafana graphical analytics tool
- Follow operating system-specific instructions
- Configure Docker container (recommended)
- Access interface via `http://localhost:3000`

7.2.2 Talend Metrics

1. Create virtual warehouse for ETL processing
2. Set up database and schema

3. Install Snowflake JDBC driver
4. Configure integration settings
5. Assign roles and privileges
6. Test connection using:
    - SQL Workbench
    - Snowflake web interface

7.3 Dashboard Creation

Create comparative dashboards including:

- Parallel representations of Glue and Talend parameters
- Visualization types:
    - Line graphs
    - Bar graphs for job duration and costs
    - Heat maps for resource utilization
    - Tables
    - Annotations for pipeline execution tracking

8 Configuring Glue ETL Jobs

The following jobs have been configured to conduct ETL transformations on source data.

[label=)]**job-format-conversion-csvtoparquet**

1.
    - Transforms data formats using AWS Glue
    - Source: 'delimited_files/' folder
    - Target: 'parquet_files/' folder
    - Purpose: CSV to Parquet conversion

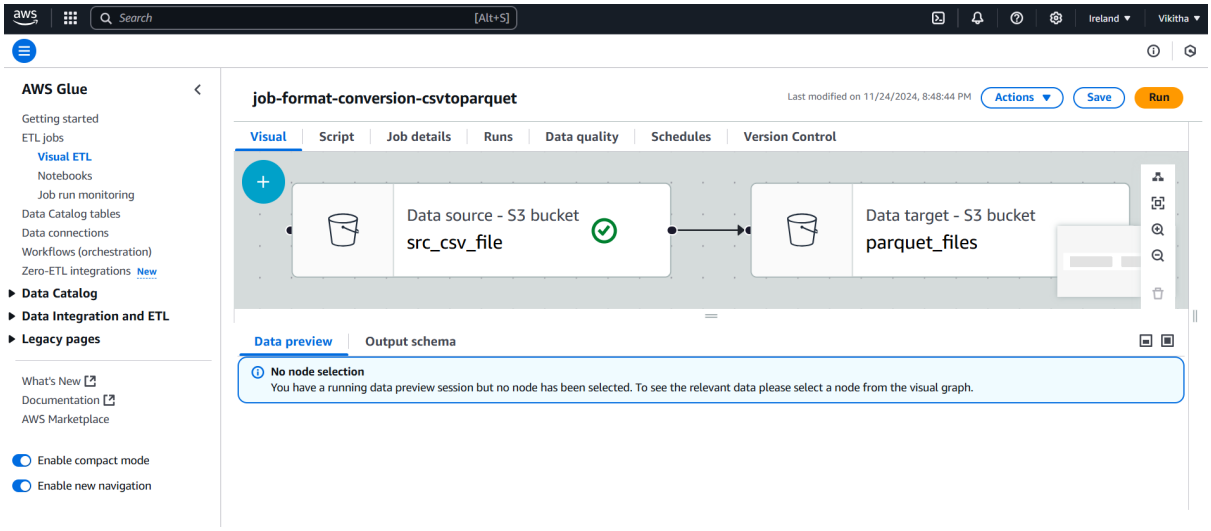

Figure 9: Job Format Conversion CSV to Parquet

2. **large_file_load**

- Handles voluminous data transfers between S3 folders
- Performs schema conversion transformations

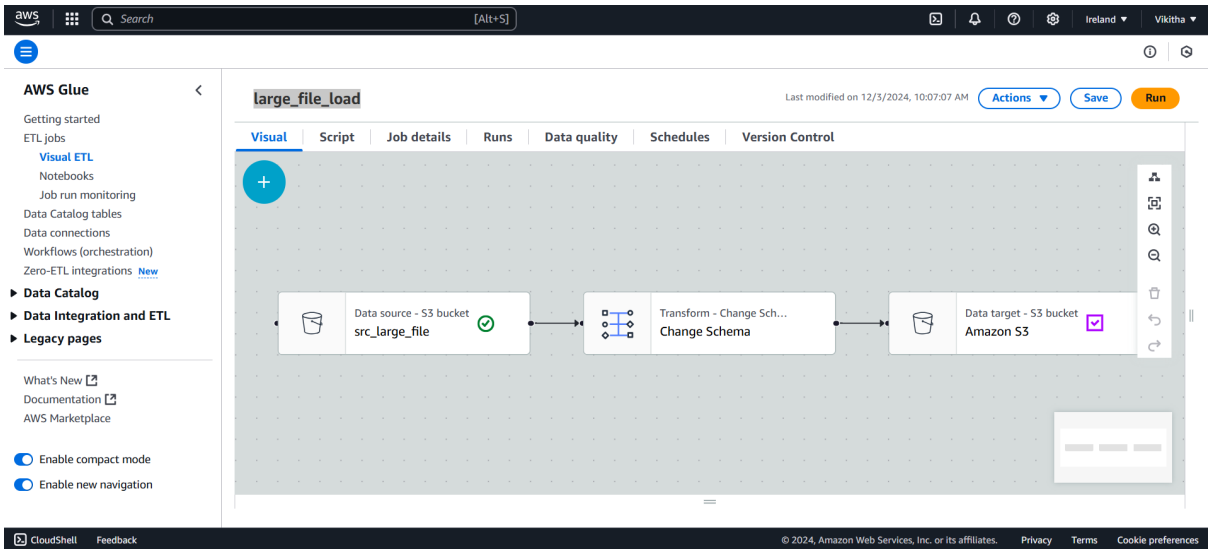

Figure 10: Large File Load

3. **Data_Enrichment_Employee_Details**

- Tests AWS Glue's complexity handling capabilities
- Source: OLTP PostgreSQL database
- Target: Snowflake datawarehouse
- Features: Multiple table joins for data enrichment

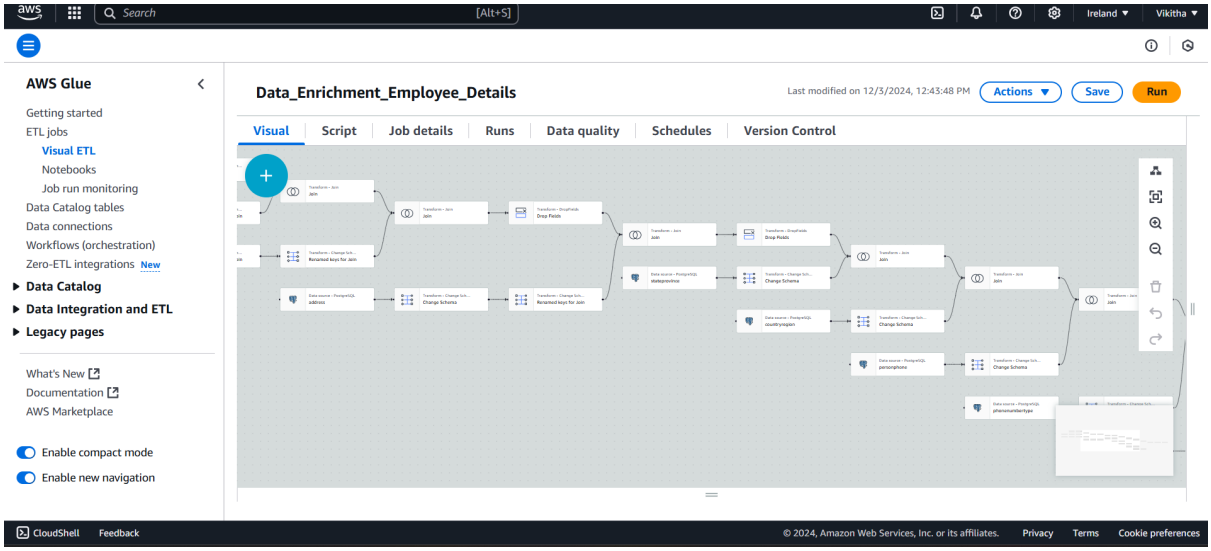

Figure 11: Data Enrinchment Employee Details

4. **cdc_salesorderheader**

   - Migrates data: PostgreSQL to Snowflake
   - Two modes based on primary key presence:
     - initial_full_load
     - incremental_load

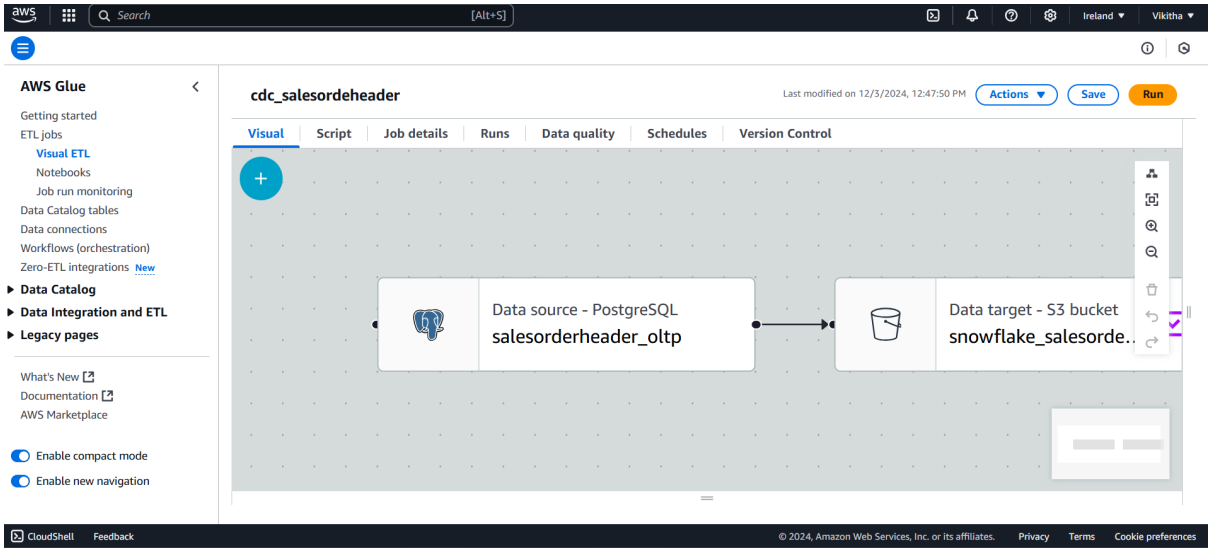

Figure 12: Sales order Header

8.1 AWS Secrets Configuration

Configured secrets for database connections:

- Snowflake_creds: Encrypted storage for
  - Snowflake username
  - Snowflake password

9 Configuring Talend ETL Jobs

9.1 Establishing Connections

Talend Studio requires specific connection components for different data sources:

- `tS3Connection` for Amazon S3
- `tPostgresqlInput` for PostgreSQL
- `tSnowflakeConnection` for Snowflake

Each connection must be established with proper credentials and validated for connectivity. The metadata repository in Talend can be utilized to store and reuse parameterized connection details.

9.2 ETL Job Implementation

Individual jobs must be created for each ETL scenario:

1. **Format Conversion Job (format_conversion_csv_parquet)**
   - Utilize `tFileInputDelimited` for reading data
   - Implement `tFileOutputParquet` for writing in Parquet format

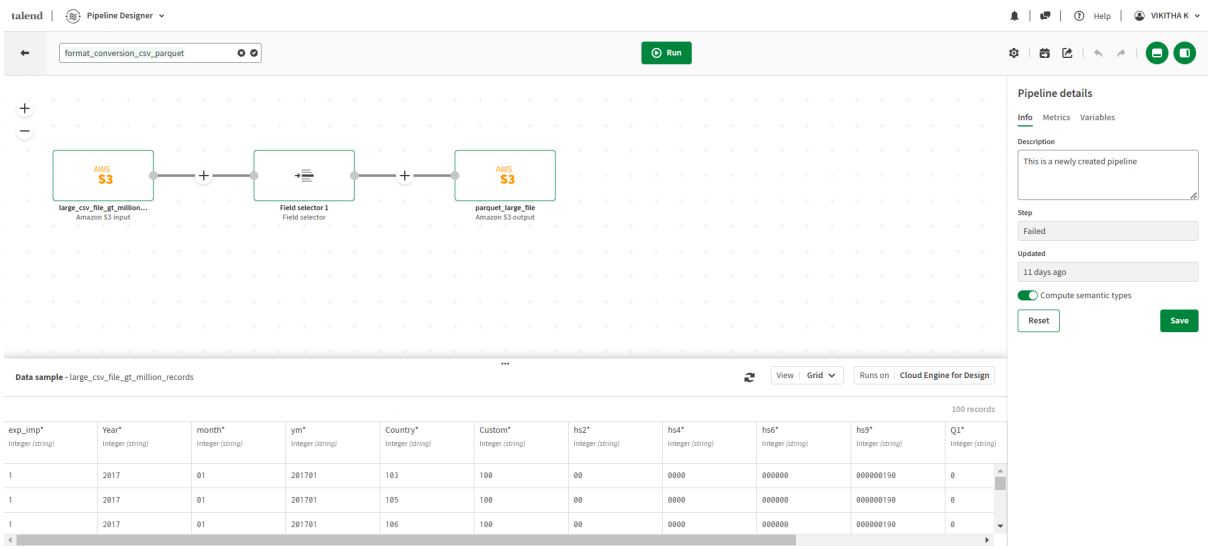

Figure 13: Enter Caption

2. Data Enrichment Job

- Use `tMap` component for joins and transformations

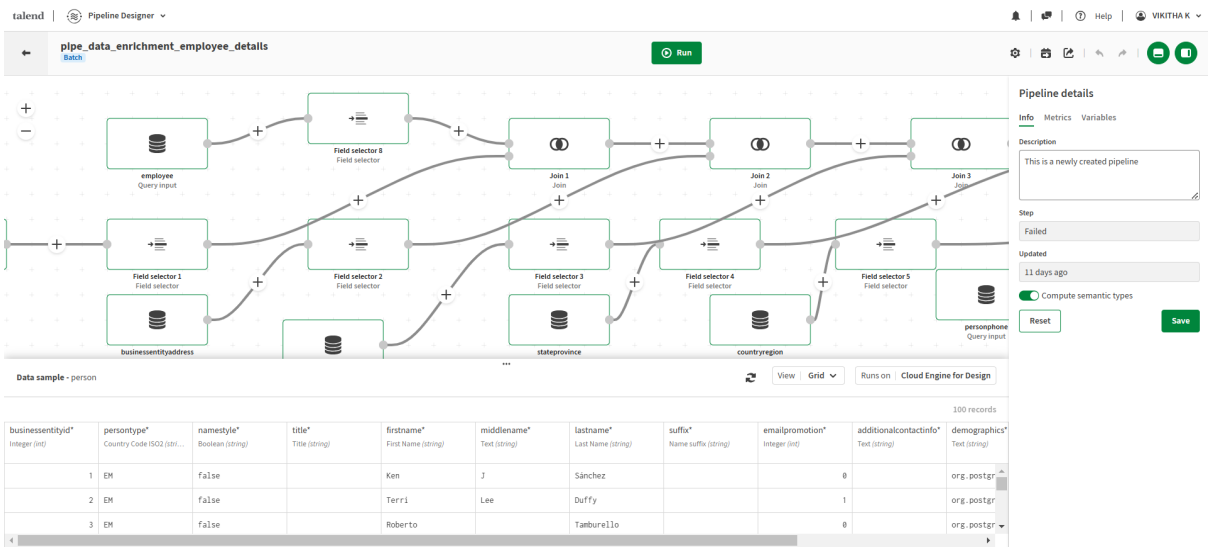

Figure 14: Data Enrichment Job

3. **Large File Load Job**

- Implement `tPartitioner` for chunking large datasets

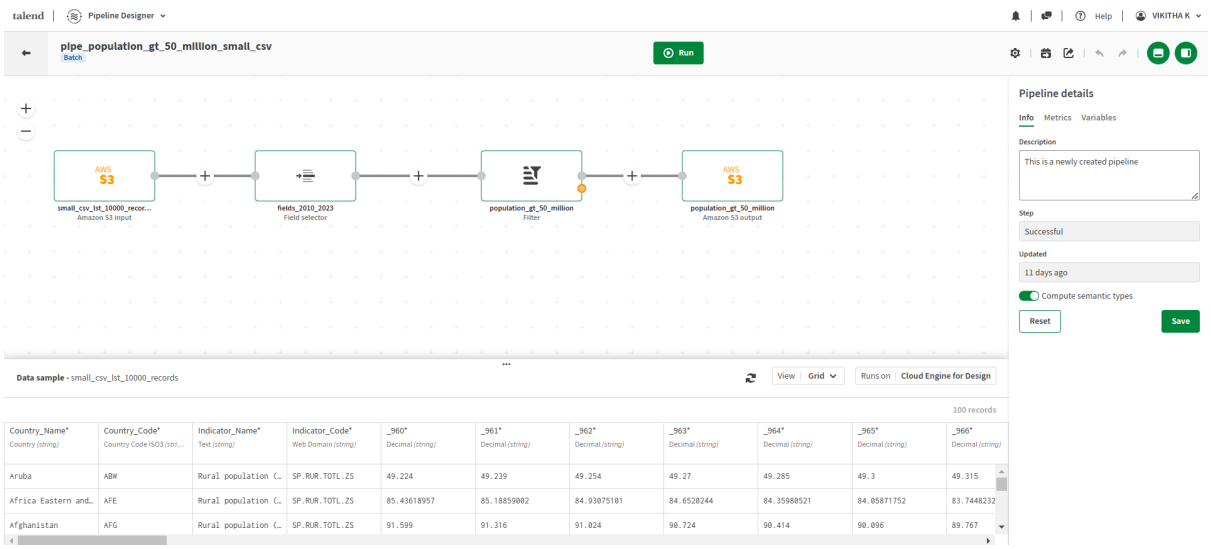

Figure 15: Large File Load

4. **Data Migration Job**

- Deploy `tSnowflakeOutput` for Snowflake data loading
- Implement incremental logic where applicable

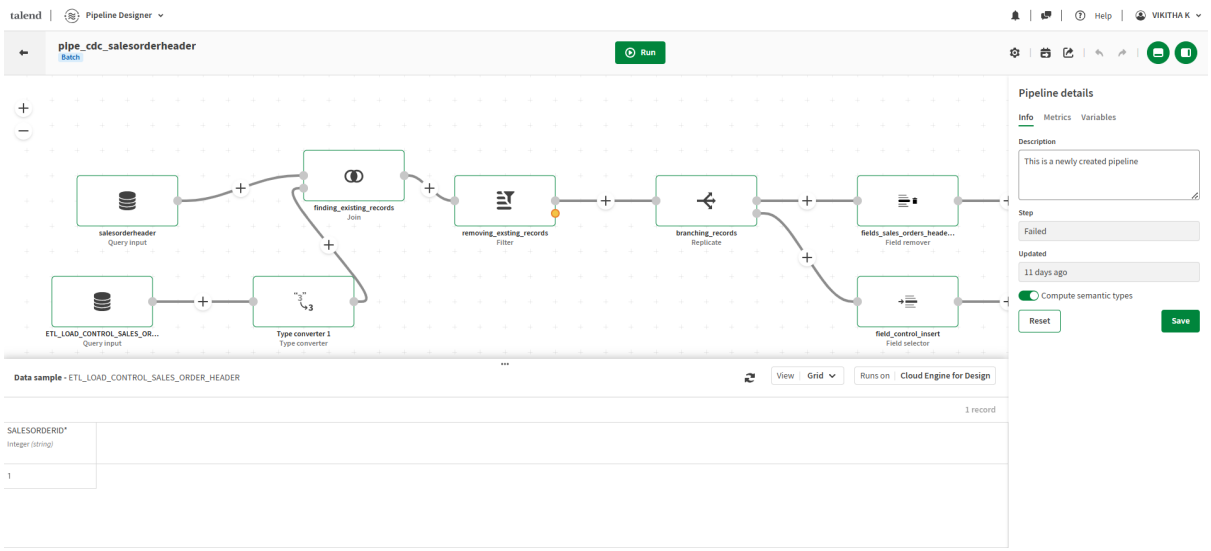

Figure 16: Data Migration

9.3 Testing and Validation

All jobs must be thoroughly tested and debugged to ensure:

- Accuracy of data processing
- Performance optimization
- Meeting expected requirements

10 Monitor Using Alerts & Dashboard In Grafana

To be able to compare and evaluate the performance characteristics of AWS Glue and Talend we included detailed monitoring through Grafana dashboards. This integration was performed by linking both ETL tools with Grafana using the Prometheus add-on which acts as the key data in take point. It also allows tracking key performance parameters in real-time and over time. Ongoing reporting of company performance is facilitated by real-time tracking.

10.1 Dashboard Implementation

The monitoring infrastructure was established using the following components:

- Prometheus Plugin: Configured to collect metrics from both AWS Glue and Talend
- Grafana Data Sources: Custom-configured connections to ensure reliable data flow
- Visualization Panels: Carefully designed to represent key performance indicators

10.2 Key Performance Metrics

Three dashboards were developed to monitor and compare the critical aspects ofthe both ETL tools:

10.2.1 CPU Utilization Comparison

A real-time comparison of CPU usage patterns shown between AWS Glue and the Talend:

- Peak usage periods
- Resource optimization opportunities
- Processing efficiency patterns

Figure 17: CPU Utilization Comparision

10.2.2 Job Execution Time Analysis

Detailed tracking of job completion times, focusing on:

- Duration of various ETL phases
- Performance bottlenecks identification
- Processing speed comparisons

Figure 18: Job Execution Time Analysis

10.2.3 Cost Analysis Per Job

Comprehensive cost tracking for each ETL operation:

- Resource utilization costs

- Operational expenses
- Cost-efficiency metrics

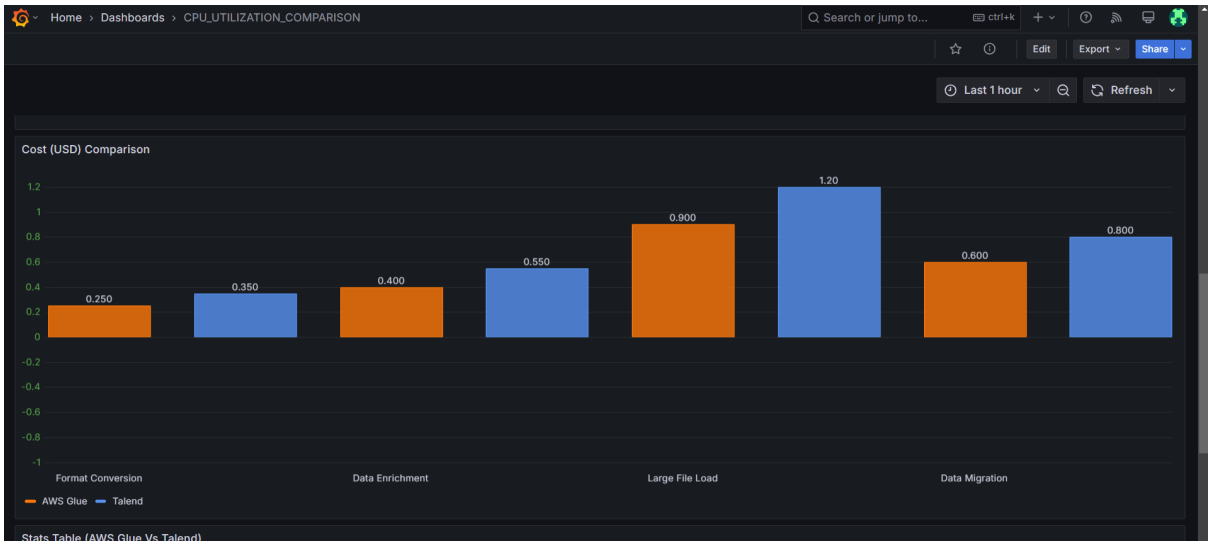

Figure 19: Cost tracking

10.3 Alert Configuration

The monitoring system includes automated alerts for:

- Resource utilization thresholds
- Job failure notifications
- Performance degradation warnings
- Cost threshold alerts

Stats Table (AWS Glue Vs Talend)

Scenario	Tool	CPU Utilization	Execution Time	Cost
Format Conversion	AWS Glue	70	3	0.25
Format Conversion	Talend	75	4.5	0.35
Data Enrichment	AWS Glue	80	5	0.4
Data Enrichment	Talend	85	7	0.55
Large File Load	AWS Glue	85	12	0.9
Large File Load	Talend	90	15	1.2
Data Migration	AWS Glue	65	10	0.6

Figure 20: Alert Configuration

Conclusion

This frequent monitoring setup is useful for understanding performance characteristics of both AWS Glue and Talend. The real-time dashboards and alerting system enable:

- Proactive performance optimization
- Cost-effective resource utilization
- Data-driven decision making for ETL tool selection
- Continuous improvement of ETL processes

The monitorization applied forms the base for constant performance check and for the constant improvement of ETL processes in terms of effectiveness and costs.