

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.

<u>ALL</u> 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.

aws	Services	Q Search		[Alt+S]	ک (A	0 0	N. Virginia 🔻	Vikitha 🔻
=			Console Home Info		Reset to default layout + Add widgets			() ()
			:: Recently visited Info	:	# Applications (0) Info Create application : Region: US East (N. Virginia)			
			MAI	8 Support				
			Y AWS Glue	Route 53	us-east-1 (Current Region) V Q Find applications			
			<u>∂</u> EC2	Cloud9	< 1 >			
			DS RDS		Name ▲ Description ▼ Region ▼ Originating account			
			™ 53		No applications			
			🔉 Lambda		Get started by creating an application.			
			Serverless Application Repository		Create application			
			View all	services //	Go to myApplications			
			:: Welcome to AWS :	:: AWS Health Info :	II Cost and usage Info II			
			Getting started with AWS C Learn the fundamentals and find valuable information to get the most out of AWS.	Open issues O Past 7 days Scheduled changes	л			
			Training and certification [2] Learn from AWS experts and advance your skills and	0 Upcoming and past 7 days Other notifications 0 Past 7 days	Data unavailable Data unavailable You must have Cost Explorer enabled to view your cost and usage data.			

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

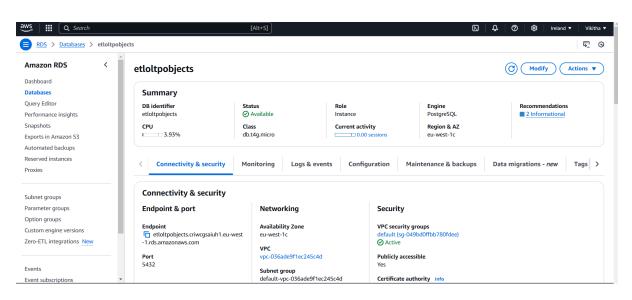


Figure 2: Database Configuration

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

aws III Q Search	[Alt+S]		ב \$ \$ D
•			0 9 5
Dashboard <	Security Groups (1/1) Info	C Actions ▼ Export security groups to	CSV Create security group
EC2 Global View	Q. Find resources by attribute or tag		
Events			
▼ Instances	sg-049bd0ffbb780fdee X Clear filters		< 1 > 😵
Instances	✓ Name ▼ Security group ID	▼ Security group name ▼ V	PC ID 🔻
Instance Types			
Launch Templates		=	
Spot Requests			-
Savings Plans	Details Inbound rules Outbound rules Sharing - n	ew VPC associations - new Tags	
Reserved Instances			
Dedicated Hosts			
Capacity Reservations	Inbound rules (3)	0	Manage tags Edit inbound rules
▼ Images			
AMIs	Q Search		< 1 > 🔞
AMI Catalog	urity group rule ID 🔻 IP version 🗢 Type	▼ Protocol ▼ Port range	▼ Source ▼ Descr
▼ Elastic Block Store	Od8ca03319eeba793 IPv4 All TCP	TCP 0 - 65535	0.0.0.0/0 -
Volumes	0736e0fc3c5c90986 – All traffic	All All	sg-049bd0ffbb780fdee –
Snapshots	00e888186729a06cb IPv4 PostgreSQL	TCP 5432	0.0.0.0/0 –
Lifecycle Manager	•		

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

Object Tools Edit View Window Help									
	T Q rties				es × main humanresources.employe	e/adventurew	orks/postgres@	etl-src-dev-vikith	3 × ×
data engg tools	\$P	humanresources.e	employee/adventurewo	orks/postgres@etl-src-d	6				
🕼 etl-src-dev-vikitha			🗸 No limit 👻						
✓ ■ Databases (3)	0.00	ry Query History							
✓ ■ adventureworks	Que								
> 🐼 Casts			humanresources.emp	oloyee					
> 😵 Catalogs	2	URDER BY DUSTN	essentityid ASC						
> C Event Triggers									
> 🔀 Extensions									
> 🝯 Foreign Data Wrappers									
 > Index Equations > Index Equations 			A						
 W Publications W Schemas (11) 	Data	Output Messages	Notifications						
>	=+	🚡 v 🗂 v 🗎	8 ± ~						
 		businessentityid	nationalidnumber	loginid	jobtitle	birthdate	maritalstatus	gender	hiredate
> C Aggregates		[PK] integer	character varying (15)	character varying (256)	character varying (50)	uate	character (1)	character (1)	date
> W Aggregates	1	1	295847284	adventure-works\ken0	Chief Executive Officer	1969-01-29	S	м	2009-01
> BU Conations	2	2	245797967	adventure-works\terri0	Vice President of Engineering	1971-08-01	S	F	2008-01
>	3	3	509647174	adventure-works\roberto0	Engineering Manager	1974-11-12	м	м	2007-11
> The FTS Dictionaries	4	4	112457891	adventure-works\rob0	Senior Tool Designer	1974-12-23	S	м	2007-12
> Aa FTS Parsers	5	5	695256908	adventure-works\gail0	Design Engineer	1952-09-27	м	F	2008-01
> FTS Parsers	6	6	998320692	adventure-works\jossef0	Design Engineer	1959-03-11	м	м	2008-01
> III Foreign Tables	7	7	134969118	adventure-works\dylan0	Research and Development Manager	1987-02-24	м	м	2009-02
> (ii) Functions	8	8	811994146	adventure-works\diane1	Research and Development Engineer	1986-06-05	S	F	2008-12
> Materialized Views	9	9	658797903	adventure-works\gigi0	Research and Development Engineer	1979-01-21	м	F	2009-01
> the Operators	10	10	879342154	adventure-works\michael6	Research and Development Manager	1984-11-30	м	м	2009-05
	11	11	974026903	adventure-works\ovidiu0	Senior Tool Designer	1978-01-17	s	м	2010-12
	12	12	480168528	adventure-works\thierry0	Tool Designer	1959-07-29	м	м	2007-12
> (() Procedures		13	486228782	adventure-works\janice0	Tool Designer	1989-05-28	М	F	2010-12
 > (() Procedures > 13 Sequences > (f) Tables (6) 	13								

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/

Amazon S3 > Buckets > etl-st	tage-files								0	₽	6
mazon S3 <	etl-sta	age-files Info									
ieneral purpose buckets											
lirectory buckets	Objec	ts Properties	Permissions M	etrics Management Acce	ss Points						
able buckets New											
ccess Grants											
ccess Points	Obie	cts (5) Info									
CCESS POINTS		(-)									
	0	Copy S3 URI	Copy URL ↓	Download Open [2] Dele	ete Actions 🔻	Create fold	er)	→ Upload	d		
ccess Points Object Lambda Access Points Iulti-Region Access Points	Object	Copy S3 URI	entities stored in Amazon S	3. You can use Amazon S3 inventory [t						วาน	
bject Lambda Access Points	Object	Copy S3 URI		3. You can use Amazon S3 inventory [t						uʻll	
bject Lambda Access Points Iulti-Region Access Points	Object need t	Copy S3 URI	entities stored in Amazon S	3. You can use Amazon S3 inventory [t							3
bject Lambda Access Points Iulti-Region Access Points atch Operations	Object need t	Copy S3 URI s are the fundamental o explicitly grant them	entities stored in Amazon S	3. You can use Amazon S3 inventory [t					our objects, yo		
bject Lambda Access Points Iulti-Region Access Points atch Operations	Object need t	Copy S3 URI s are the fundamental o explicitly grant them Find objects by prefix	entities stored in Amazon S n permissions. Learn more [2] ▲ Type	3. You can use Amazon S3 inventory 🗗 t	o get a list of all objects in you		r others	to access yo	our objects, yo	ø	
bject Lambda Access Points lulti-Region Access Points atch Operations IM Access Analyzer for S3 lock Public Access settings for	Object need t	Copy S3 URI s are the fundamental o explicitly grant them Find objects by prefix Name	entities stored in Amazon S n permissions. Learn more [2] ▲ Type	3. You can use Amazon S3 inventory [2] t	o get a list of all objects in you		r others	to access yo	our objects, yo	ø	
bject Lambda Access Points lulti-Region Access Points atch Operations IM Access Analyzer for S3	Object need t	Copy S3 URI s are the fundamental o explicitly grant them Find objects by prefix Name Calentited_files/	entities stored in Amazon S permissions. Learn more [2]	3. You can use Amazon 53 inventory [2] t v Last modified -	o get a list of all objects in you		r others ▼ -	to access yo Storage cl -	our objects, yo	ø	
bject Lambda Access Points lulti-Region Access Points atch Operations IM Access Analyzer for S3 lock Public Access settings for lis account torage Lens		Image: Copy S3 URI s are the fundamental to explicitly grant them ind objects by prefix Name Image: Copy Copy Copy Copy Copy Copy Copy Copy	entities stored in Amazon S permissions. Learn more (2	3. You can use Amazon 53 inventory [2] t v Last modified -	o get a list of all objects in you		r others ▼ -	to access yo Storage cl	our objects, yo	ø	
bject Lambda Access Points fulti-Region Access Points atch Operations M Access Analyzer for S3 lock Public Access settings for is account	Object need t	Copy S3 URI Sare the fundamental o explicitly grant them ind objects by prefix Name delimited_files/ parquet_files/ snowpipe_employee	entities stored in Amazon S permissions. Learn more [2] A Type Folder Folder e_details/ Folder Folder	3. You can use Amazon 53 inventory [2] t v Last modified -	o get a list of all objects in you		r others ▼ -	to access yo Storage cl -	our objects, yo	ø	
bject Lambda Access Points luti-Region Access Points atch Operations WA Access Analyzer for S3 lock Public Access settings for its account torage Lens ashboards	Object need t	Copy S3 URI sare the fundamental o explicitly grant them ind objects by prefix Name delimited_files/ parquet_files/ snowpipe_employee	entities stored in Amazon S permissions. Learn more [2] A Type Folder Folder e_details/ Folder Folder	3. You can use Amazon 53 inventory [2] t v Last modified -	o get a list of all objects in you		r others ▼ -	to access yo Storage cl -	our objects, yo	ø	



- 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

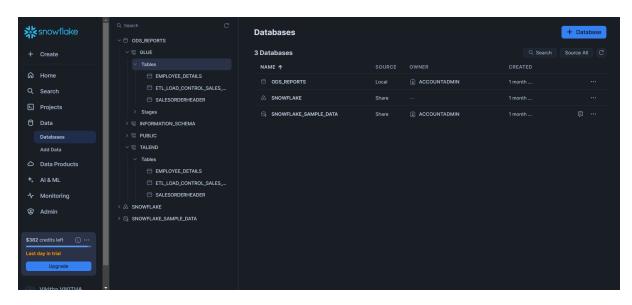


Figure 6: Snowflake Account Creation and Configuration

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

	23 datasets Q Find in a column Add fitter V				
Data console					
Datasets	Talend Trust Score™ 🛛 🔽 2.57/5 ↔0.07 ©				Period Last 30 days 🗸 🏠
Connections	Validity: 83%		v	alidity 🛈 Poor Axis threshold	③ — Talend Trust Score™ 5
Data quality rules			80%		4
Semantic types	Usage: 27% Popularity: 21%		60%		3
Custom attributes			40% 20%		2
	Discoverability: 0% Completeness: 97%		0%	11/23 11/27 12/	0
			14/10 14/10	14/20 14/27 14/	1 14/0 14/10
	Data quality	¢	Data types across datasets ©		
			Name	Validity	Used in 👻
	85.1% of valid values Poor	8 suggestion	FR Postal Code	77%	7 datasets
			Country Code ISO2	94% -	3 datasets
	11.7% of empty values Good	1 suggestion	Country	79%	3 datasets
			Address Line	100%	2 datasets

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

🧑 - Home			Q Search or jump to) 📾 ctrl+k	+ ~	?	٣	a (b ^
	Good evening.								1
	Dashboards Most recent Starred Explore	Nee	d help getting started? Check out the	<u>Setup guide →</u>					
	CPU_UTILIZATION_COMPARISON								
	Account								1
	Account plan	Users							
	Trial • 4 days left	1	Support						
	Manage subscription	Add users	What's new						
			Tutorials Stacks						
			Access policies						

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.

$[label=)] {\bf job-format-conversion-csvtoparquet}$

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

aws III (Q Search	[Alt+S]	Ireland Vikitha
•		(i)
AWS Glue <	job-format-conversion-csvtoparquet	Save Run
Setting started ETL jobs	Visual Script Job details Runs Data quality Schedules Version Control	
Visual ETL Notebooks Job run monitoring Data Catalog tables Data connections Workflows (orchestration) Zero-ETL integrations <u>New</u>	Data source - S3 bucket src_csv_file Data target - S3 bucket parquet_files	.≞ @ @
Data Catalog		Ū
Data Integration and ETL .egacy pages	Data preview Output schema	
What's New [2] Documentation [2] AWS Marketplace D Enable compact mode D Enable new navigation	No node selection You have a running data preview session but no node has been selected. To see the relevant data please select a node from the visual graph.	

Figure 9: Job Format Conversion CSV to Parquet

2. large_file_load

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

large_file_load			1			_
			Last modifi	fied on 12/3/2024, 10:07:07 AM (Acti	ions ▼) (Save)	Run
Visual Script	Job details Runs	5 Data quality	Schedules Version Contro	il.		
. +						Δ.
						Ξ
						Q
						Q
	to an unit of the state		Transform Channe Cab	Dete terr	at CZ hushat	Û
	60	• • • • • • • • • • • • • • • • • • •				5
310	c_targe_nte	∧	change Schema	Alliazon	55	¢
				• Data source - 53 bucket • src_large_file • Src_large_file • → • ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕	src_large_file	$ \begin{array}{c c} \bullet \\ \hline \hline \hline \bullet \\ \hline \hline \bullet \\ \hline \hline \hline \bullet \\ \hline \hline \hline \hline$

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

aws III Q Search	[Alt+S]	∑ 🗘 Ø 🐼 Ireland ▼ Vikitha ▼
		0 0
AWS Glue <	Data_Enrichment_Employee_Details Visual Script Job details Runs Data quality Schedules	Last modified on 12/3/2024, 12:43:48 PM Actions Save Run Version Control
Visual ETL Notebooks Job run monitoring Data Catalog tables Data connections Workflows (orchestration) Zero-ETL integrations New > Data Catalog > Data Integration and ETL > Legacy pages	Image: State of the s	
What's New [2] Documentation [2] AWS Marketplace		
 Enable compact mode Enable new navigation 	=	=
CloudShell Feedback		© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

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

AWS Glue <										_		
	cdc_sa	lesorde	header				La	st modified on	12/3/2024, 12:4	7:50 PM Action	ns 🔻 Save	Run
Getting started ETL jobs	Visual	Scrip	t Job details	Runs	Data quality	Schedules	Version C	ontrol				
Visual ETL	Visuat	Scrip		Kulis		Schedules	version c					
Notebooks												Δ.
Job run monitoring	+											
ata Catalog tables												Ξ
ata connections												Ð
orkflows (orchestration)												Q
ero-ETL integrations New												
ata Catalog												Ŭ
ata Integration and ETL			6	Data	source - Pos	tgreSQL			P	Data targe	et - S3 bucket	6
egacy pages				sales	orderhead	er_oltp			C [*]	snowfla	ke_salesorde	••• ~
'hat's New 🖸												
ocumentation 🖸												
WS Marketplace												
Enable compact mode												



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

for	rmat_conversio	n_csv_parq	uet		00													۲	Run												\$ 5 C \ /
																															Pipeline details
																															Info Metrics Variables
_					_						_																				Description
	AWS S3		-	+)	_		• <u>=</u>		 -(+)	_	_		AWS S3																		This is a newly created pipeline
large	<pre>_csv_file_gt_mil Amazon S3 inpu</pre>	ion					selector 1					pare Ama	uet_lar	ge_file																	Step
																															Failed
																															Updated
																															11 days ago
																															Compute semantic types
																															Reset
mple - la	arge_csv_file_g	t_million_re	cords																		г	View	Grid	~	Run	is on	Cloud	Engine	e for D	esign	
																													100 r	ecords	
	Year*		n	onth*		у	m*		Country	r -		Cu	stom*		hs2*				hs4*		hs	6*			hs9*				Q1*		
ng)	Integer (tring)	Ir	teger (stri	ng)	In	iteger (strir	ig)	Integer (string)		Int	eger (stri	ing)	Integ	er (string	,)		Integer	(string)	In	teger (sti	ing)		Integ	er (strin	<u>2)</u>		Intege	r (string)	
	2017		0	1		21	01701		103			10	9		00				0000		06	0000			0000	00190			0		A
	2017		0	,		2	01701		105			10			00				0000			0000			0000	00190			0		

Figure 13: Enter Caption

2. Data Enrichment Job

• Use tMap component for joins and transformations

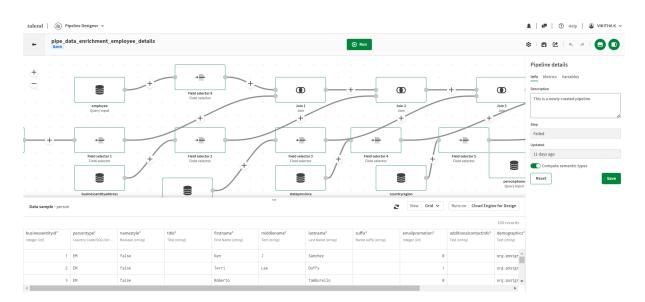


Figure 14: Data Enrichment Job

3. Large File Load Job

• Implement tPartitioner for chunking large datasets

talend	🛞 Pij	pelineDesigner 👻																															* #	🕐	ielp 🤇	VIKITHA K	< ~
+	plpe_po Batch	opulation_gt_50	D_ml	llion_sn	nall_cs	v														۲	Run												\$ E	2	5 A I	00)
+																																	Pipelin	e details			
(a) 1																																	Info M	etrics Varia	ables		
					_				_				_				_				_			_									Descriptio	n			_
	,	s3	_	+	-				+		-0)-	 -		Ð		-		+ -			AW:	s 3										This is a	i newly creat	ed pipeline		
	small_csv_ls	st_10000_recor				fiel	lds_2010	2023					, L	opulati	on_gt_9	0_millio	 n				popul	ation_gt	_50_mill	ion									Step				_1_
	Amazo	on 53 input				÷	ield sele	ctor							Filter						At At	mazon S	3 output										Succes	ful			
																																	Updated				
																																	11 days	ago			
																																	C 00	mpute sema	ntic types		
																																	Reset			Save	
Data samj	ple - small_o	csv_lst_10000_record	ls																				г	View	v Gri	d 🗸	Rur	ison	Cloud	Engin	e for De	sign					
																															100 re	cords					
Country_Na Country (string		Country_Code* Country Code ISO3 (str		Indicator Text (string			Indica Web Do	tor_Coo			960* ecimal (s	tring)	_96 Dec	imal (str	ing)		_962* Decima	al (string)			963* ecimal (stri	ng)		_964* Decimal (string)		_965 Decir	i* nal (strir	g)		_966* Decima	l (string)					
Aruba		ABN		Rural po	pulatio	n (SP.RU	R.TOTL	.ZS	49	9.224		49.	239			49.25	4		4	9.27			49.285			49.3				49.31	5					
Africa Eas	tern and	AFE		Rural po	pulatio	n (SP.RU	R.TOTL	ZS	85	5.4361	3957	85.	188596	982		84.93	075101		8	4.652024	4	4	84.3598	0521		84.6	58717	52		83.74	48232					
Afghanista	n	AFG		Rural po	pulatio	n (SP.RU	R.TOTL	ZS	91	1.599		91.	316			91.02	4		91	9.724		1	90.414			90.0	96			89.76	,					

Figure 15: Large File Load

- 4. Data Migration Job
 - Deploy tSnowflakeOutput for Snowflake data loading
 - Implement incremental logic where applicable

p	pe_cdc_salesorderheader	\$ # C < / 🔁 (
	Image: Second	Pipeline details Info Metrics Variables Decouption This is a newly created pipeline Step Failed Updated 11 days ago Compute semantic types Reset
ORDERID (string)	ETL_LOAD_CONTROL_SALES_ORDER_HEADER	

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 of the 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

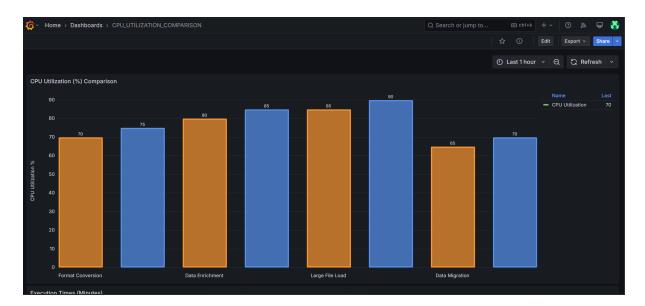


Figure 17: CPU Utilization Comparision

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	ΤοοΙ	CPU Utilization	Execution Time	Cost
	AWS Glue			
Data Enrichment	AWS Glue			0.4
Data Enrichment				
Large File Load	AWS Glue			
Large File Load				
Data Migration	AWS Glue			

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.