ETL Modernization with AWS Glue

Shiv Narayanan
Product Manager – AWS Glue

Topics we will cover

- Challenges with legacy ETL tools
- Introducing AWS Glue
- Demo – how Glue enables multiple personas to perform data integration
- How to modernize data integration pipelines to AWS Glue

Trend: More demanding workloads

Batch

Data warehousing

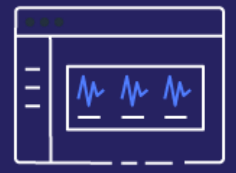
Business intelligence

Reporting

Nightly or hourly ETL

Latency-agnostic, long-running jobs

Real-time



Responsive dashboards

Monitor and alert

Streaming apps

Continuous operation

Latency-sensitive micro-batch jobs

Challenges with legacy ETL solutions

IT Leadership challenges

High license costs

Difficulty in recruiting data engineers

Vendor lock-in

Data Engineering challenges

Time wasted in infrastructure management and scalability issues

Multiple tools to handle structured, streaming, semi-structured data

Challenges in implementing CI/CD pipelines

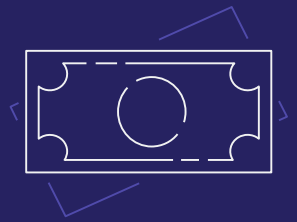
Line of Business challenges

Lack of agility and speed

Self service aspirations

Data Quality challenges

Our customers ask us for



Cost effective

Scalable

No-lock in

Empower more users

Serverless

AWS Glue is a **serverless data integration** service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

Why do customers choose AWS Glue?

Serverless

No infrastructure to maintain. Allocate needed compute power and run jobs

Data Integration for every user

Development environments catered to different skillsets - visual ETL development for Data Engineers, notebook styled development for Data Scientists, and no code development for Data Analysts

Cost-effective

All-in-one pricing model is 55% cheaper than other cloud data integration options

Handles complex workloads

Connect to 65+ data sources, process petabytes of data in real-time, includes batch and event driven modes

No lock-in

Develop data integration pipelines in open source SparkSQL, PySpark, Python and Scala

How customers use AWS Glue

Build Data Lakes and Lake Houses for scalable data analysis

Migrate from expensive traditional ETL solutions to gain flexibility and reduce costs

Catalog data assets to make them available to AWS Analytics services

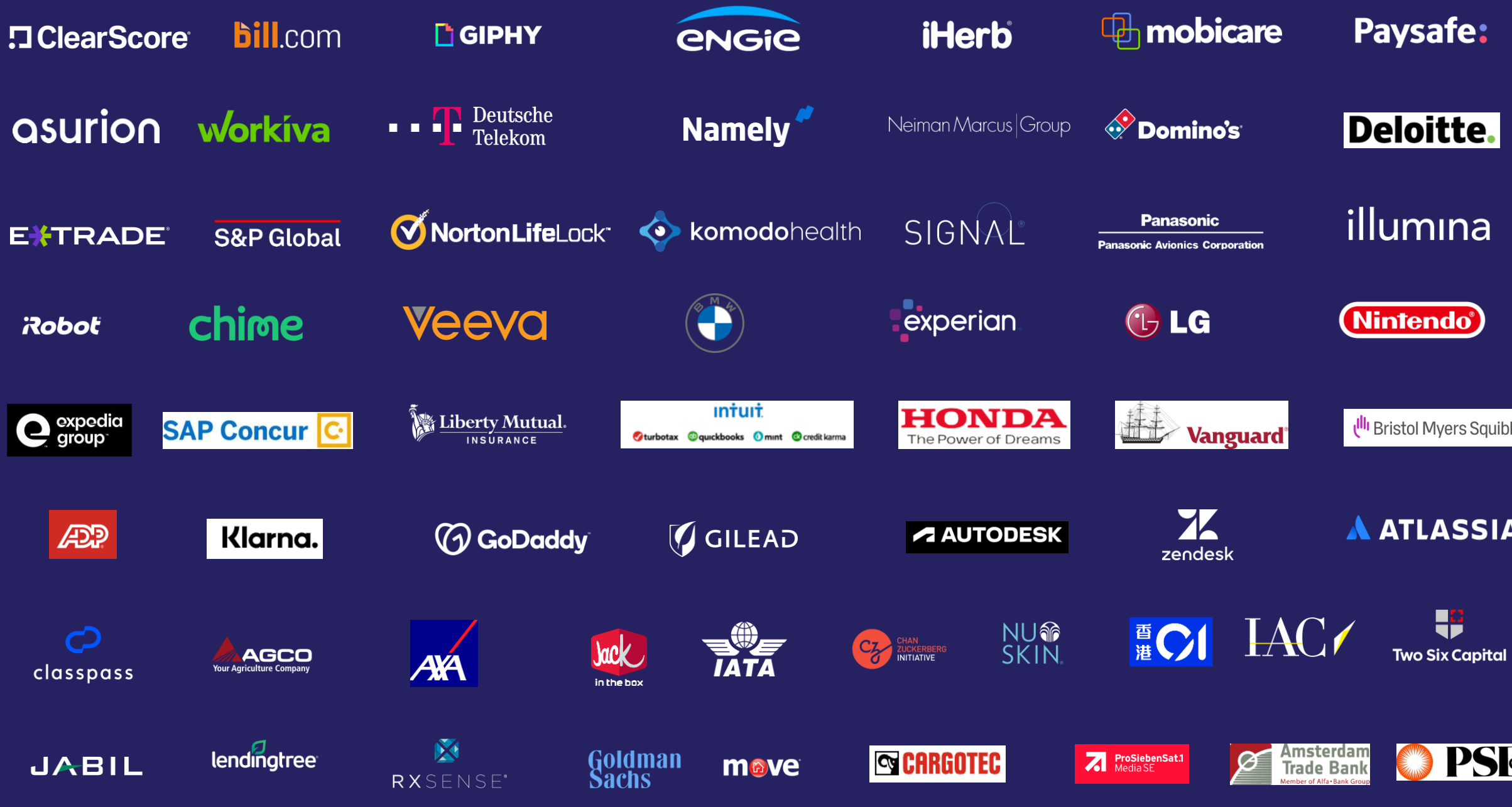
Process petabytes of data both in batch and real-time using Apache Spark

Prepare data for Machine Learning

"(Glue) gave us great advantage over what we were doing before, so we're able to ingest data faster and transform it and get faster insights," Brian Zellner, Assoc Director R&D DataLakes, Bristol Myers Squibb

AWS Glue was a natural choice for modernizing our on-premises Hadoop stack. Glue Spark jobs starts in seconds, offers us cluster isolation, semi-structured data processing capabilities and considerable cost savings compared to on-premises environment. Our data scientists are now able to focus on adding value through data.

Markus Bergmaier, Cloud Architect, ProSiebenSat.1 Media SE



aws

iBotta builds self-service data platform using AWS Glue

PROBLEM

- 200+ microservices and growing
- Data Engineering team unable to keep up with ingestion into Data Lake
- Existing solution couldn't scale to data volumes
- Small engineering team unable to cater to ingestion requests

SOLUTION

- Built a self-service Data Lake allowing users to configure jobs to ingest data from sources
- Glue pulls data from source and automatically converts JSON to relational tables
- Created framework based on Glue APIs for self-service

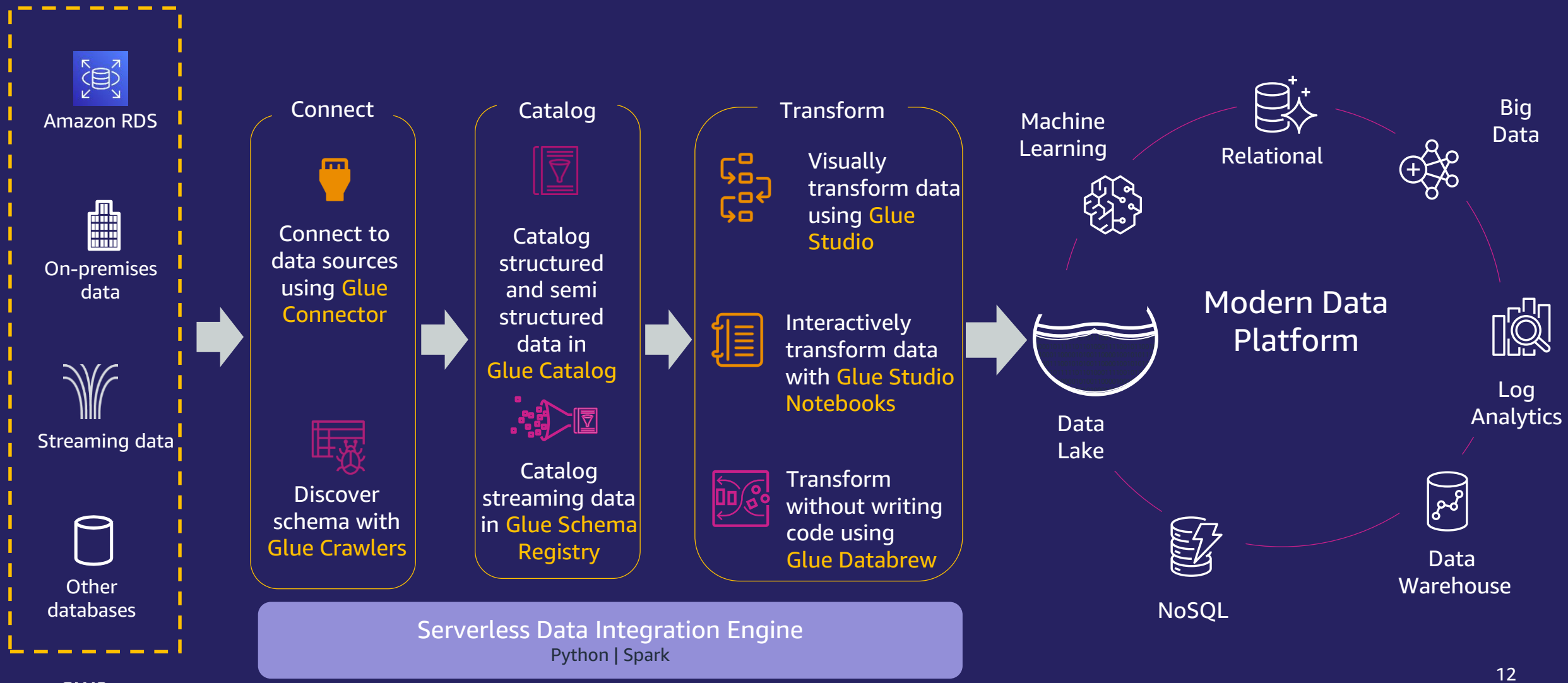
IMPACT

- Data ingestion time reduced by 1000s of development hours
- Analysts ingest data on their own when new micro services are created

"AWS Glue powers our self-service data platform by providing necessary out of the box ETL transformations to ingest and integrate over 200 complex message types from our micro-services into our Data Lake, saving us 1000s of developer hours. Further AWS Glue version 2.0's 1-min billing reduced costs of running our ETL pipelines by 5x compared to our previous solution."

--Erik Franco, Principal Data Architect, Ibotta

A Data Integration ecosystem for building Lake Houses faster



DEMO

Data Integration for everyone using AWS Glue

Modernize to AWS Glue

What are your options to modernize your ETL pipelines?

- Difficult to scale
- Expensive licensing
- Vendor Lock in

01
Upgrade On-Prem ETL

7x
Glue is 7x cheaper compared to o on-premise options

No lock-in
Glue jobs are written open source Spark, Python, and Scala

- High total cost of ownership
- Effort to tune and manage clusters

02
DIY Spark on Cloud

5x
Adopting Glue is 5x cheaper than setting up your own Spark cluster

4x
Glue reduces maintenance of your self-managed Spark clusters by 4x

- Code lock-in
- Additional cost for infrastructure
- Expensive licensing

03
Cloud ETL

55%
Glue is 55% cheaper compared to other cloud providers

Serverless
With Glue, there are no servers to manage and infrastructure costs included

Customers who modernized successfully understood that ..

Modernization leads to new capabilities, not just cost savings

Customers modernize to not only to achieve cost benefits, but also to gain new capabilities, such as elasticity, innovation, agility, self service, stream processing, and ML based data cleansing and transformation.

It is important to embrace architecture changes

Architecture changes are inevitable and they are good. Reauthoring of legacy code helps remove technical debt.

Comparing features is not useful

Handling larger data volumes in cloud requires different techniques such as decoupling storage from compute, multi-thread processing. Features are designed to handle them as supposed to achieving feature parity.

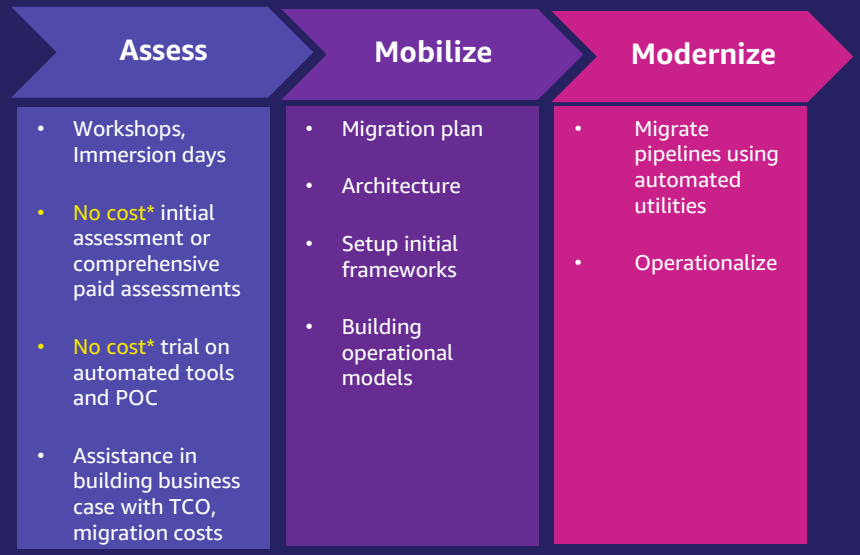
Code conversion tools are accelerators and not the strategy

Code convertors helps improve productivity significantly but they don’t convert entire codebase. Use other techniques such as reusable frameworks to accelerate.

ETL Modernization Phases

	Assess	Mobilize	Modernize
Proposed Key activities	• Analyze existing ETL pipelines and identify data integration patterns • Train data engineers on AWS Glue • Conduct Proof of concept • Create total cost of ownership and business case	• Identify initial pipelines for migration • Design and build pipelines • Convert pipelines to Glue using common frameworks and automation • Create architecture, detailed designs, measure outcomes	• Create migration plan for all pipelines • Migrate/Modernize in phases • Create operational guides for supporting data pipelines
Proposed Deliverables	• Business case detailing new capabilities, total cost of ownership, migration estimates	• Reusable frameworks • 1-2 data pipelines in Production	• All pipelines are migrated to Production

ETL Migration Partner Program



Tooling ISV's

next PATHWAY

blade bridge

IMPETUS

Some of the participating SI's…

accenture

TATA CONSULTANCY SERVICES

rackspace

TEKsystems An Allegis Group Company

IBM

virtusa Accelerating Business Outcomes

Infosys

bitwise

epam

wipro

WAVICLE Data Solutions

Cognizant

quantiphi

softserve

compasso UOL

IMPETUS

AGILISIUM BIG ON CLOUD. BIG ON DATA.

mission AWS Premier Consulting Partner

etlmodernization-si-support@amazon.com

* No cost assessments and POCs are limited in scope. For more information contact your account teams or etlmodernization-si-support@amazon.com

aws

Assess Phase

Analyze existing ETL pipelines and identify data integration patterns

As result of this exercise, you will be able to arrive at your inventory and categorize your data integration patterns. An example is show below

Source System	Target Database	Complexity	# of jobs
ERP	Landing Zone	Simple	100
Webservices	Landing Zone	Medium	50
RDBMS	Landing Zone	Complex	20
NoSQL	Landing Zone	Simple	100
Landing	Curated Zone	Medium	500

How can AWS and Partners help – As part of ETL Migration Program, Partners* offer automated tools to analyze and determine complexity of pipelines.

* Depending on the ETL tool Partners may have this capability. Contact your Account team to know more.

Train data engineers on AWS Glue

Identify core team of engineers, business users who will be instrumental in decision making and train them on AWS Glue. These engineers can act as developer advocates in your latter phase.

How can AWS and Partners help – As part of ETL Migration Program, AWS and Partners offer 2-day customized training sessions, and workshops for your data engineers, and business users.

Assess Phase

Conduct Proof of Concept

1. Identify what needs to improve in each of these patterns and test out the patterns. Use automated tools as much as possible during these conversions with help of partners.

Example to show how to document POC results. Not to be considered as actual patterns or results.

Source	Target	Complexity	# of jobs	Approach	Automation
ERP	Landing Zone	Simple	100	Glue Studio and Connectors	100% automation
Webservices	Landing Zone	Medium	50	Glue Blueprints	Build blueprint
Landing	Curated Zone	Medium	500	Glue Studio	75% automation

How can AWS and Partners help – Partners* can provide limited free assessments or paid assessments to perform POC using automated tools.

* Depending on the ETL tool Partners may have this capability. Contact your Account team to know more.

Create Total Cost of Ownership and Business Case

With a clear understanding of patterns, what can automation do, built TCO and migration costs. While comparing, include infrastructure costs, licensing costs, connector costs, administrator costs for accurate comparison. Document additional business capabilities that you are going to achieve.

How can AWS and Partners help –

- AWS Account teams or partners can build TCO and Partners or Proserve can provide estimates to migrate the jobs.
- AWS Account teams can provide inputs to your business case. For qualified customers.
- Account teams can discuss AWS funding options for migrations.

Mobilize Phase

Identify pipelines for initial migration

- Identify 2-5 pilot pipelines which are complex to do in current environment
- Create core data engineering team for modernization
- Ensure that they have business impacts associated – example, data pipelines that does not meet SLAs, data pipelines that can be self served

How can AWS and Partners help – Partners, Proserve or AWS Account teams can provide guidance on what pipelines to choose first for migration.

Design and build pipelines

- Use automation to convert pipelines
- Develop common frameworks that can be used for other jobs ex: CI/CD, automated testing, Error handling, logging, restart-ability. Most of these are available in Glue – you can configure or customize them for your requirements.
- Deploy pipelines to production environment

How can AWS and Partners help –

1. AWS Datalabs
2. Proserve or Partner led design and build using proven reusable components
3. Office hours to support questions, address issues by AWS specialist SAs.

Create architecture, detailed designs, and measure outcomes

- Document the architecture, designs, frameworks that can be applied to other pipelines.
- Measure outcomes – ex: Is the pipeline performing better than legacy jobs, is data available within SLA

How can AWS and Partners help –

1. AWS Datalabs
2. Proserve or Partner led design and build
3. Office hours to support questions, address issues by AWS specialist SAs.

Modernize Phase

Create migration plans for all data pipelines

- Create phased migration plan
- Migrate a specific business area(procurement, finance) or stages (source to landing, landing to curated) in these phases
- Train all developers on reusable frameworks, automation capabilities

How can AWS and Partners help –

1. Proserve or Partner led planning
2. Office hours to support questions, address issues by AWS specialist SAs.

Build, Test and Deploy

- Use automation and common frameworks to convert pipelines

How can AWS and Partners help –

1. Proserve or Partner led build
2. Office hours to support questions, address issues by AWS specialist SAs.

Create operational guides for supporting data pipelines

- Document support mechanisms, train support teams on restart-ability, error identification and when to reach to AWS enterprise support
- Decommission legacy tools

How can AWS and Partners help –

1. Account teams can provide guidance on support mechanisms
2. Proserve or partners can support pipelines and transition to support teams

Summary of Best practices to plan ETL Modernization

Don’t focus on cost benefits alone in the business case

There are several benefits for business teams such as agility, and self service which often gets ignored in a ETL modernization business cases.

Select right jobs for POC

- Jobs that are challenging, needs scalability and elasticity in current environment
- Jobs that impact your SLAs
- Capabilities that are not possible (ex: self service, stream processing, ML based data cleansing for deduplication, PII detection)

Embrace architecture changes because they provide new capabilities

Look for opportunities to improve existing architecture. Ex: ingesting data into a data lake, using CDC as supposed to batch pull

Don’t underestimate the need for change management

As you introduce new capabilities in self service, you will have to constantly reinforce the need for migration. Have crisp messaging on why you are looking to change.

Form a core team early on

Identifying a core team that understands the importance of visual interfaces, code based ETL development and self service is critical in the Assess phase of the project.

Call to Action

- Engage with your AWS Account team to know more about ETL Modernization Program and how this can help accelerate your journey
- Learn about AWS Glue with no-cost immersion days

Thank you!