Simplify Data Integration and Preparation using AWS Glue

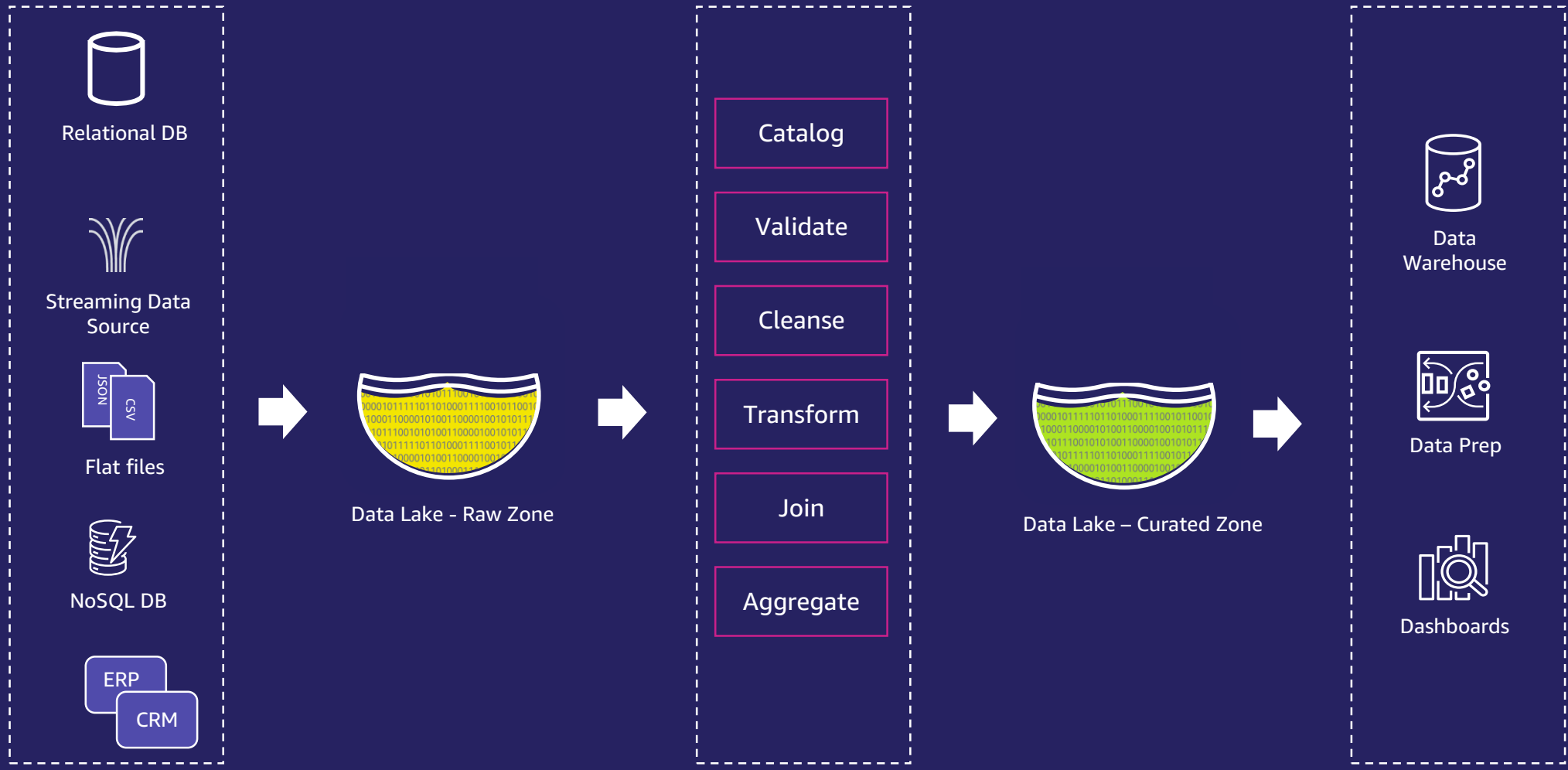
Shiv Narayanan, Product Manager – AWS Glue

Deenbandhu Prasad, Sr. Analytics Specialist SA

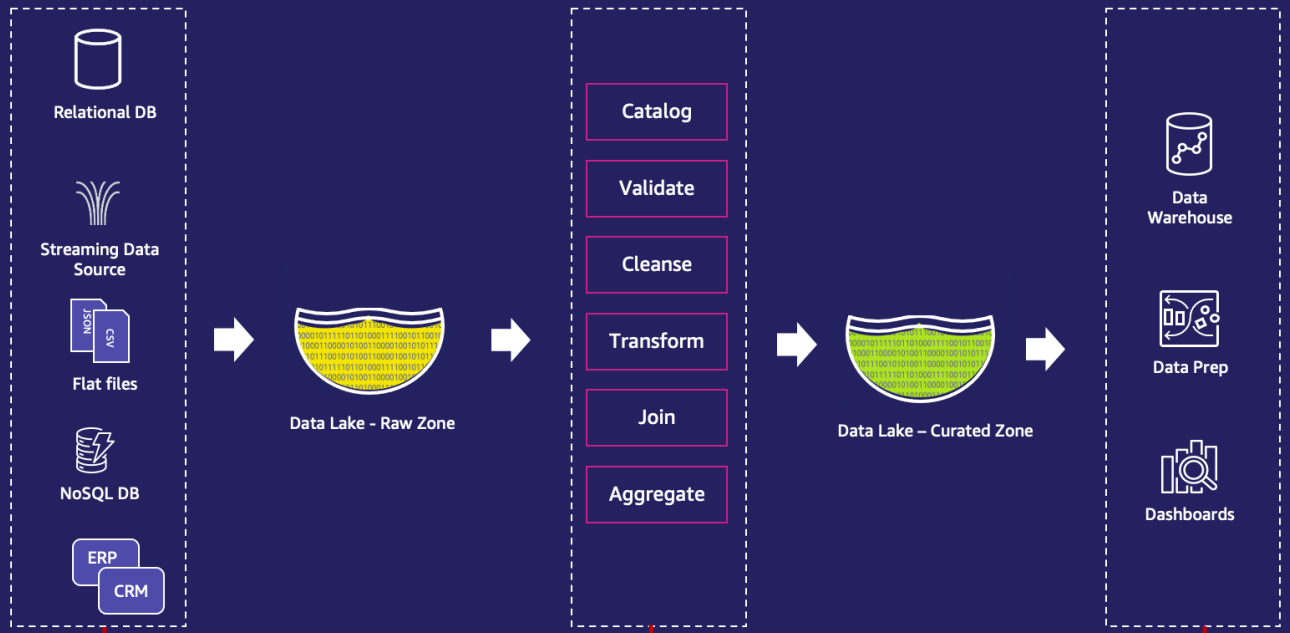
What are we going to cover today

1. A use case that you can relate to
2. Challenges to implement this use case
3. AWS Glue and how it addresses these challenges

A simple yet common data integration architecture pattern



Simple is not really simple in data integration…



1. Expensive connectors
2. Non-unified tools
3. Challenging to handle semi structured data
4. Legacy methods to cleanse data
5. Locked in proprietary tools
6. Challenging to scale
7. One tool fits all approach
8. Expensive licenses for self service
9. Non scalable infrastructure

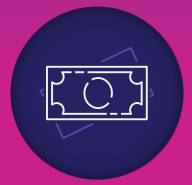
AWS Glue
Serverless Data Integration for complex workloads

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

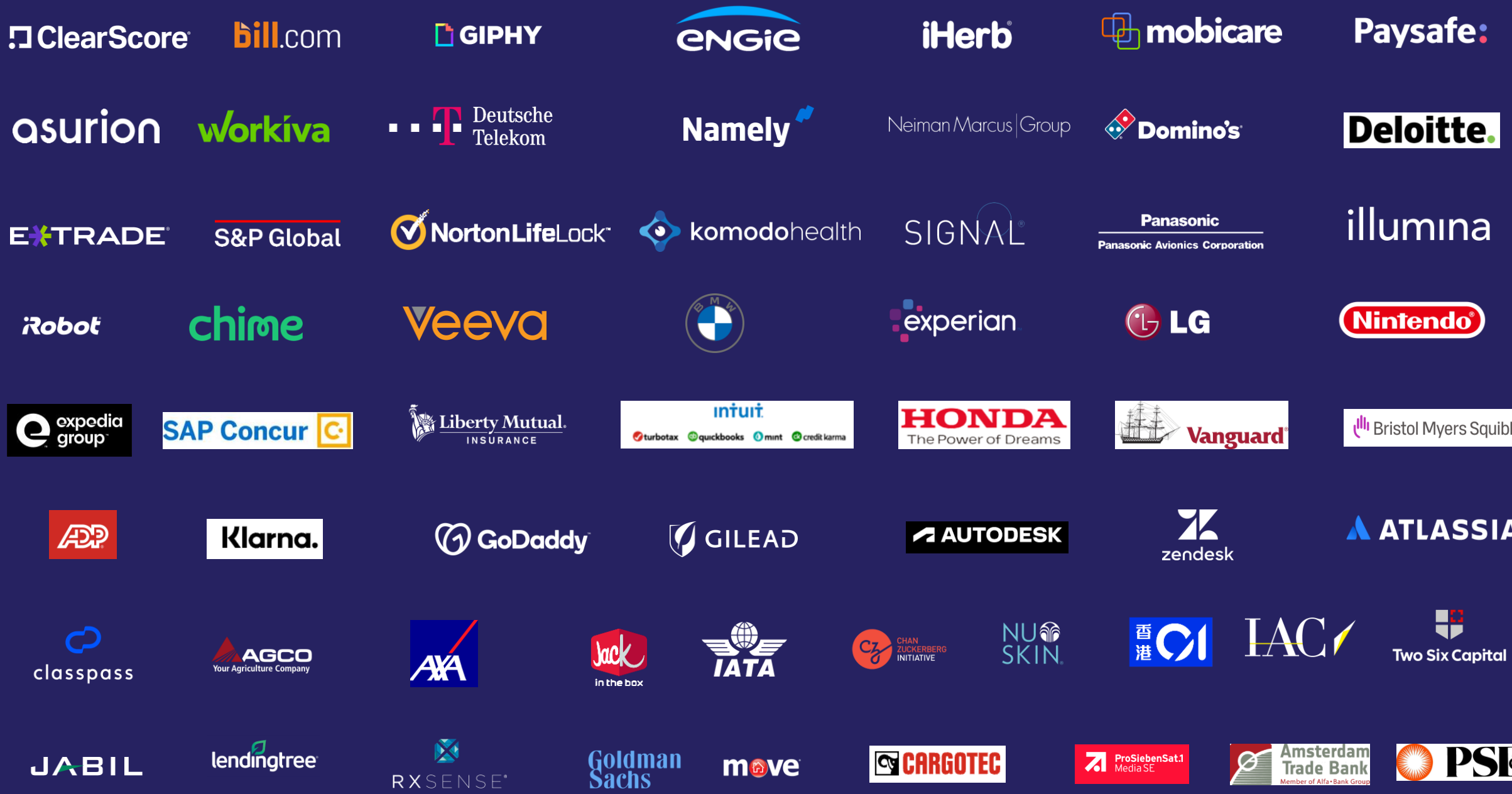
All-in-one pricing model is 55% cheaper than other cloud data integration solutions

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, Scala



BMW powers self-service Cloud HUB data platform with AWS Glue

PROBLEM

- On-premises Hadoop, Scoop used to perform data ingestion and processing
- Processes ran on hard-to-scale platform, leading to attrition of internal customers moving back to original data platform
- Reliance on DevOps to constantly provision clusters and tuning was time-consuming
- Lack of self-service options for data ingestion led to longer data ingestion cycles

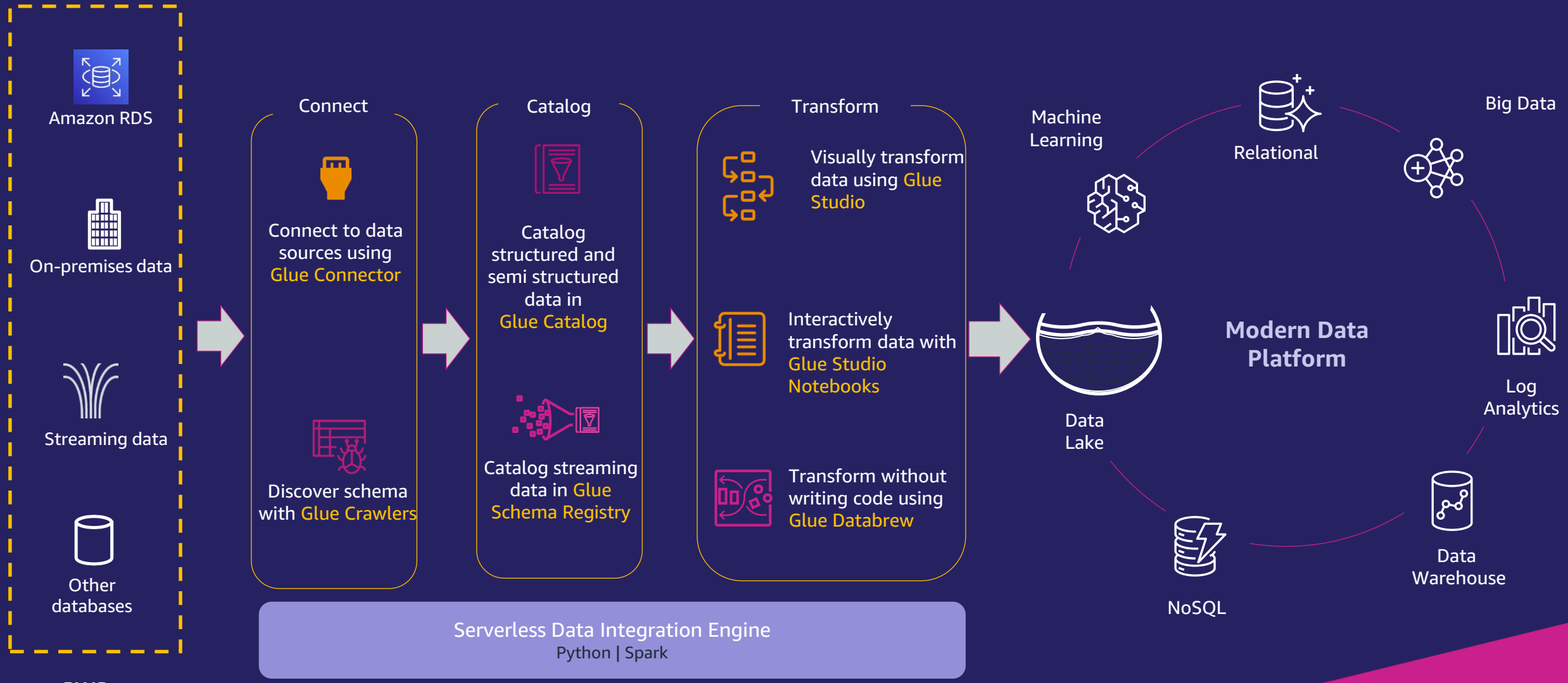
SOLUTION

- Built self-service cloud hub platform and used AWS Glue for structured data ingestion
- Created self-service features with a customized UI and used Glue APIs to automatically provision AWS Glue jobs for data ingestion
- Build a customized level data catalog on top of AWS Glue Catalog and Catalog APIs

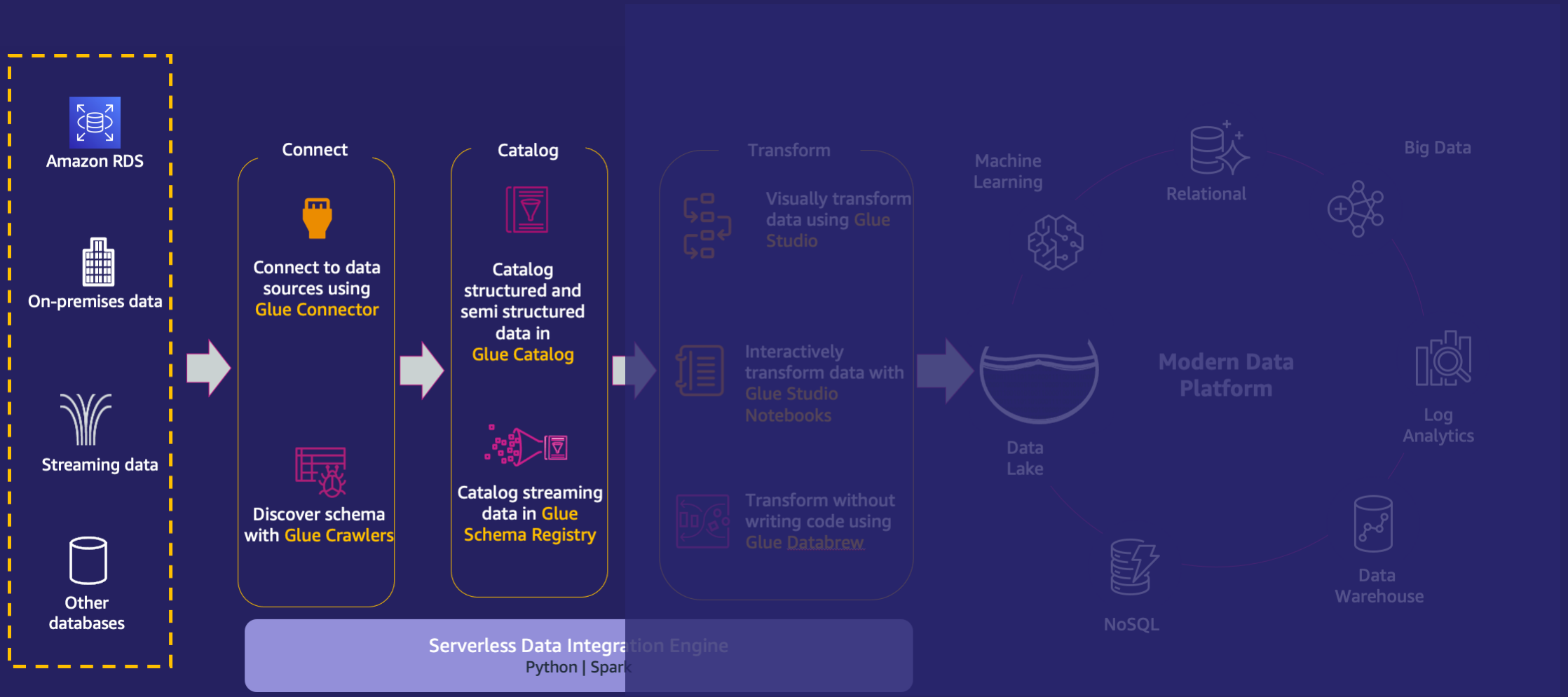
IMPACT

- Significant adoption of the Cloud HUB data platform with internal users
- Users self-serve data instead of relying on data engineers, shortening data ingestion cycles

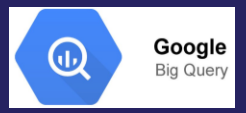
A Data Integration ecosystem for building Modern Data Platforms



Data Ingestion



Glue Connectors Marketplace



+ Many more…

Custom Connectors with AWS Glue

Data Sources

On-premises DBs

Proprietary stores

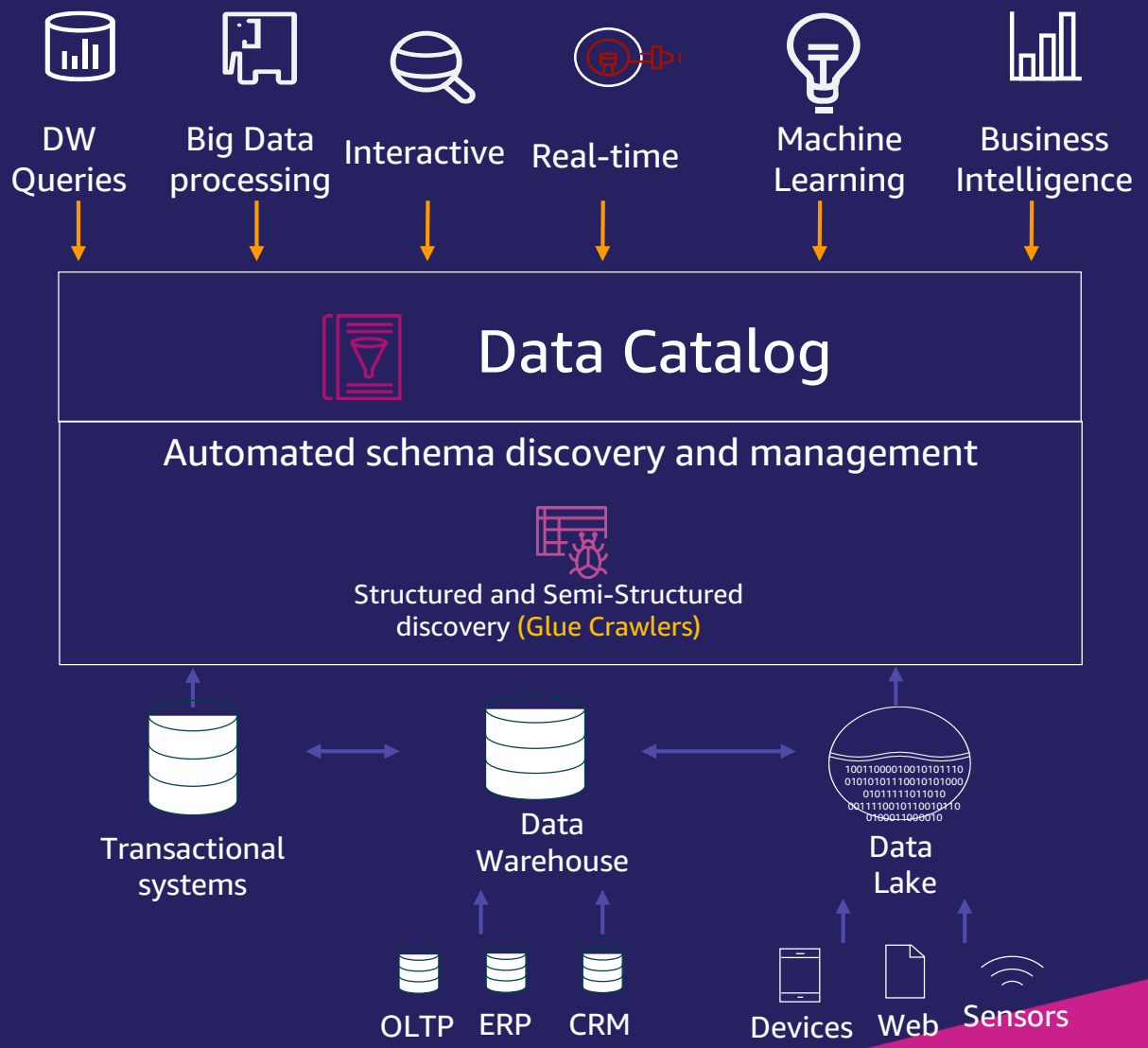
SaaS applications

CUSTOM CONNECTOR

- No additional cost for connecting to sources
- Flexible and easy to build connectors

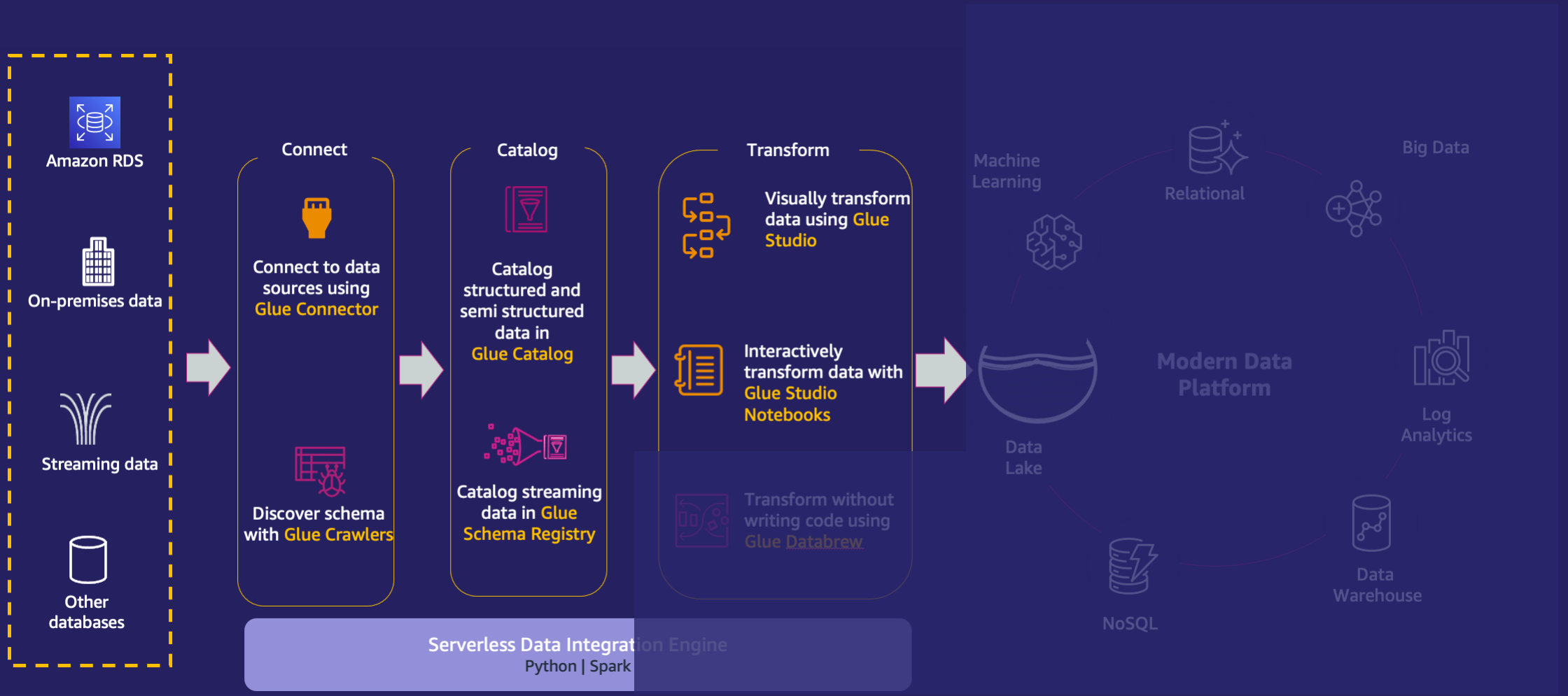
Unified Data Catalog with automated schema discovery

- No movement of data = Low Costs/Admin
- All metadata centrally available for search and query = Productivity
- Unify structured, semi-structured data = Speed to Insight
- Automate data discovery = Productivity

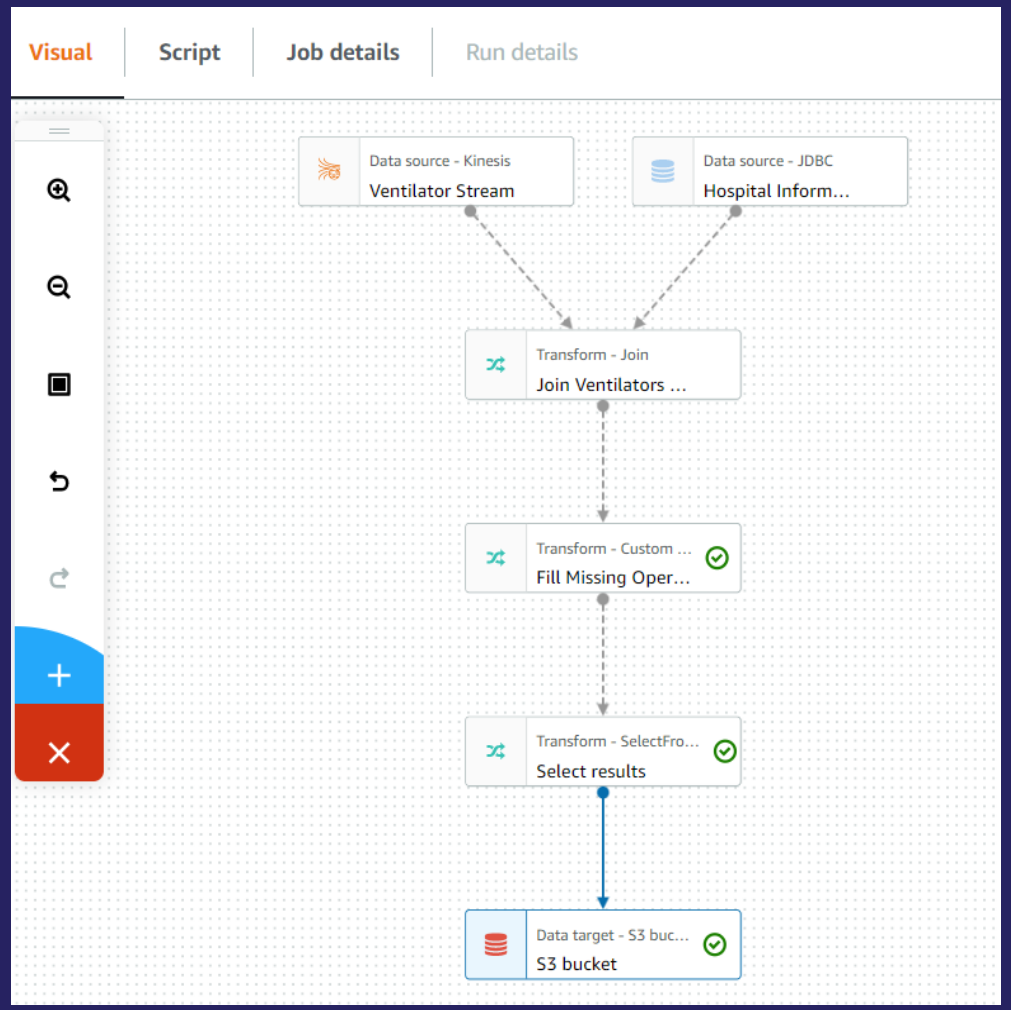


Data Ingestion Demo

Data Transformation



AWS Glue Studio: Visual ETL interface



Author AWS Glue jobs visually without coding

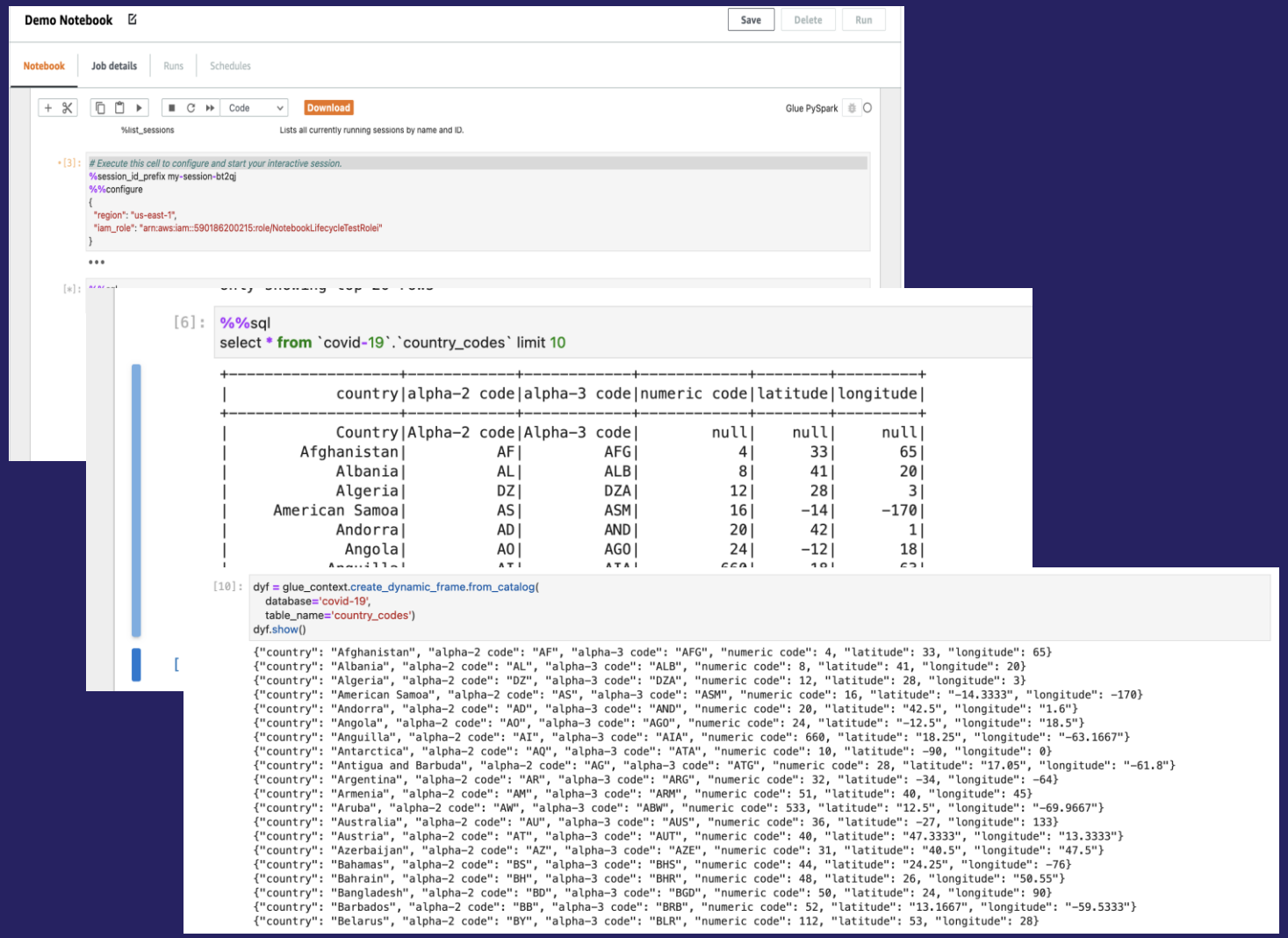
Monitor 1000s of jobs through a single pane of glass

Distributed processing without the learning curve

Advanced transforms though code snippets

AWS Glue Studio Notebook (Preview)

Preview



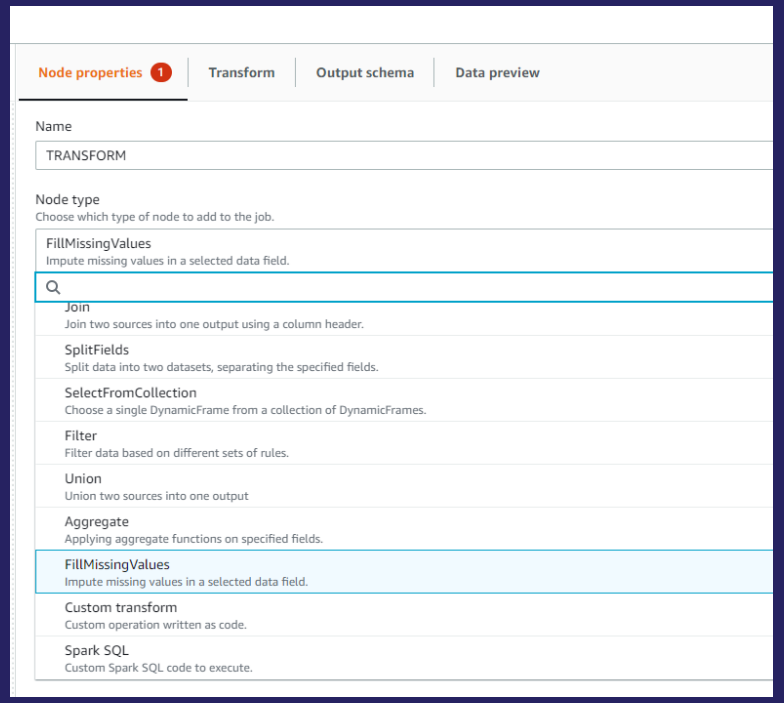
Interactive AWS Glue jobs development

Submit AWS Glue jobs from the AWS Glue Studio notebook

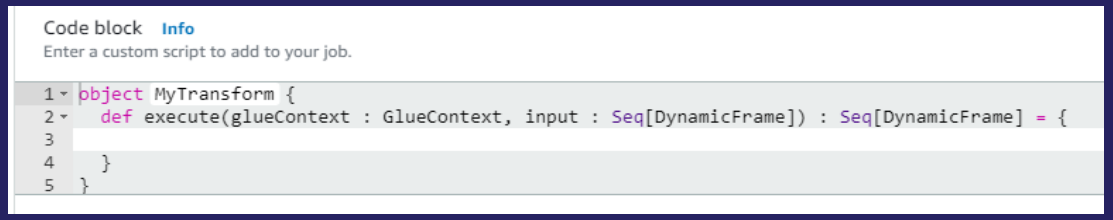
Use notebook magic to define transforms in SQL and control cost

Built-in monitoring support

Glue Studio offers out of box transformations...



built-in transform



… and advanced transforms like PII detection for advanced cleansing

1 Scan Data 

- Full Scan
- Sample Scan

2 Detect Entities 

- Built-in Entities (e.g. SSN, passport)
- Custom Entities

3 Remediate 

- Store results
- Redact/mask results

AWS Glue streaming for real-time ETL

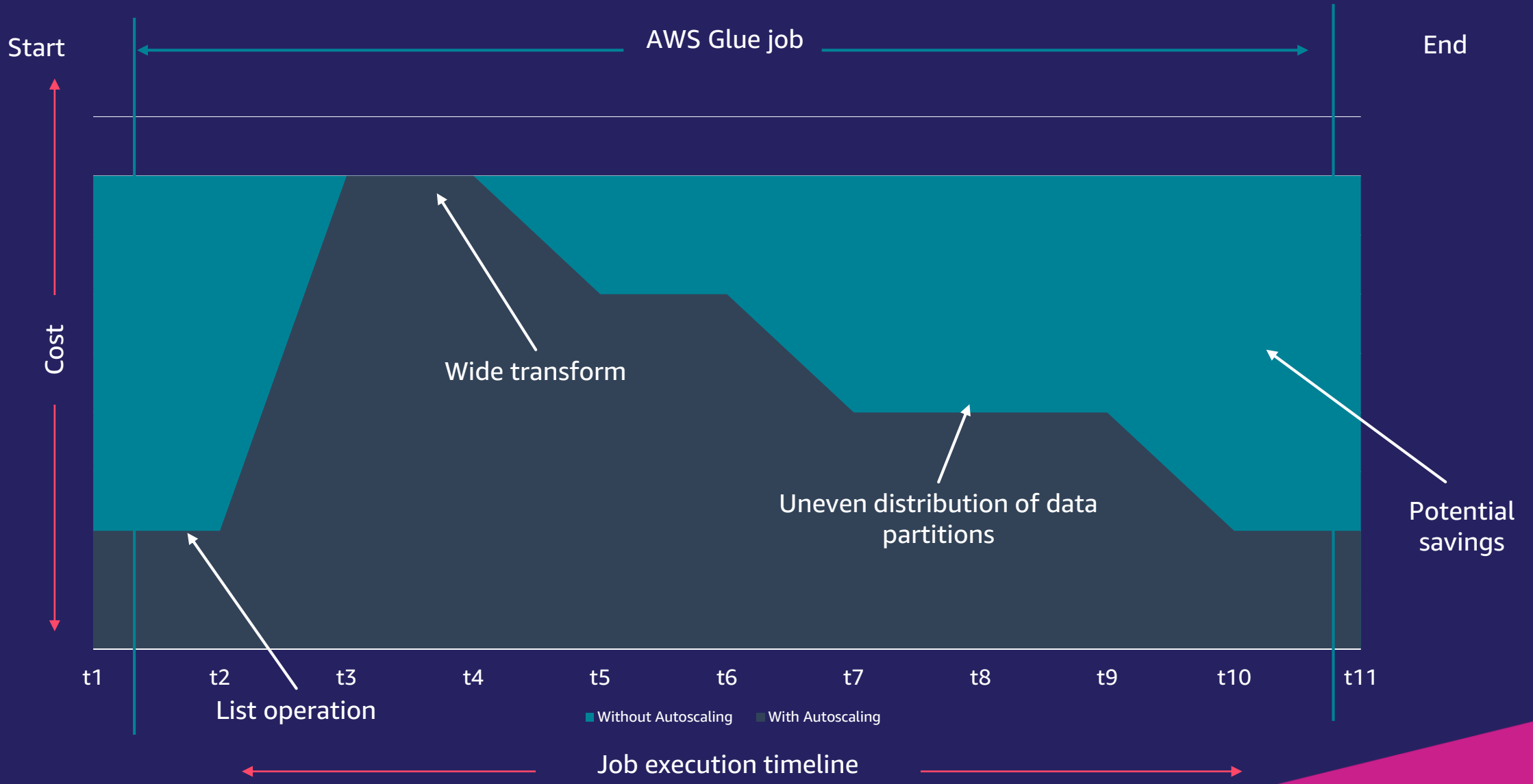
Prepare streaming data and make it available for analytics in seconds

Easily integrate with Amazon MSK, Amazon Kinesis, and Apache Kafka

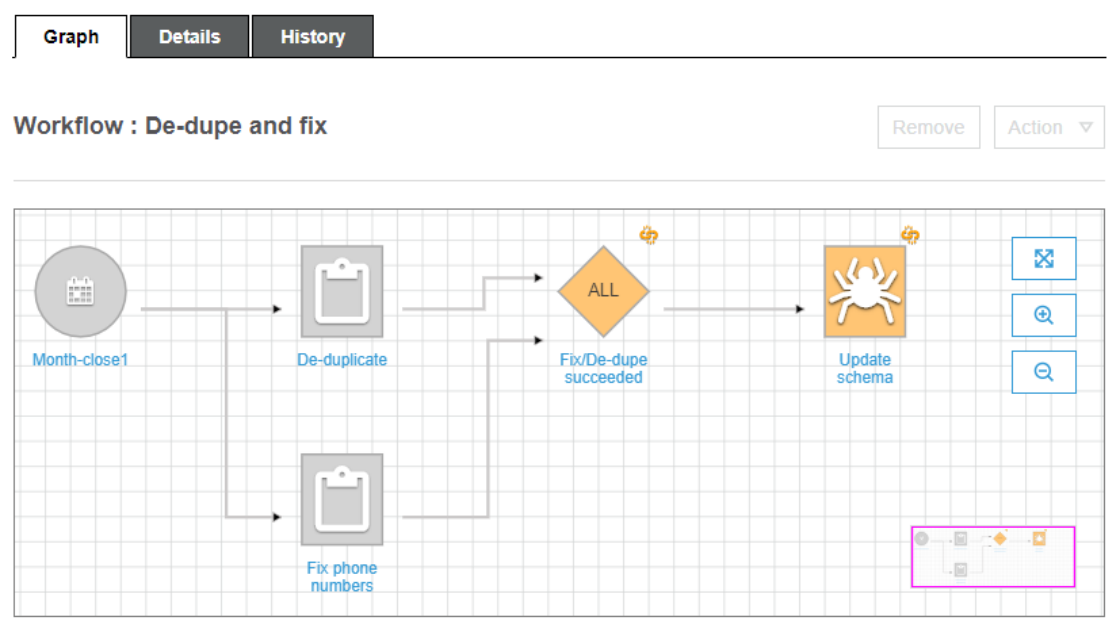
Combine, enrich, and transform data in flight using Glue Studio

Adapt to varied and changing schemas

Auto-scaling (Preview)



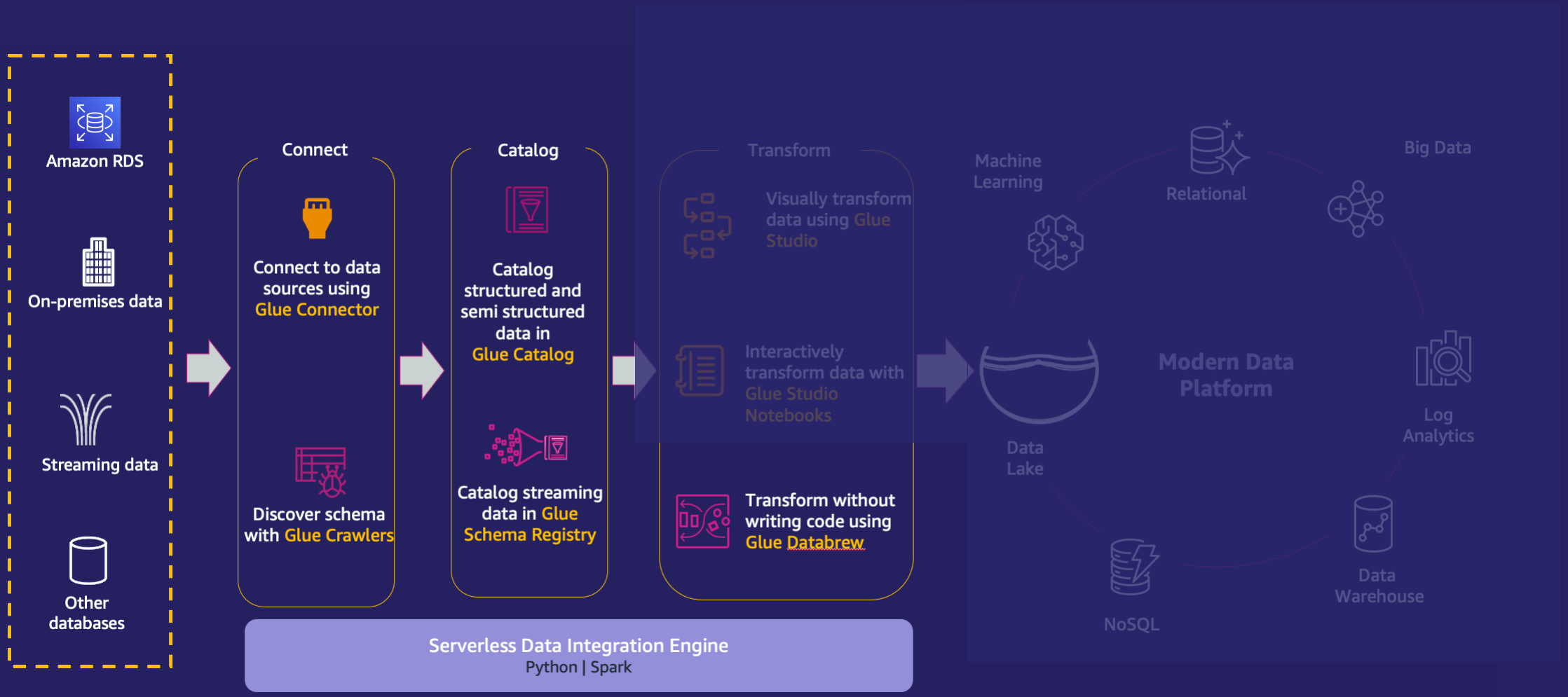
Orchestrate jobs easily with AWS Glue workflows



- Orchestrate Glue jobs and other AWS services
- Schedule jobs or trigger based on events
- Monitor execution of the workflows in one place

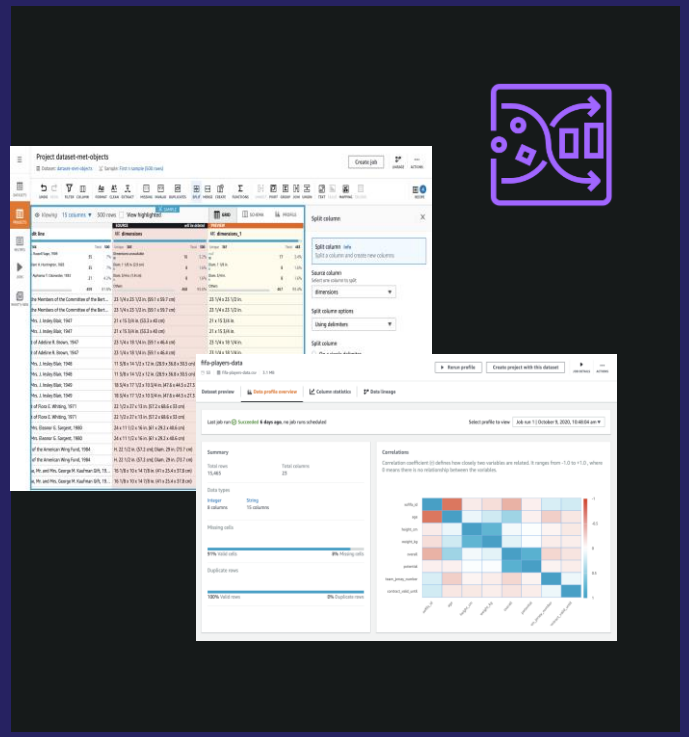
Data Transformation Demo

Data Preparation



AWS Glue DataBrew

Visual data preparation for analytics and machine learning



Clean and normalize data with a visual interface

250+ built-in transformations without writing code

Profile data to understand data patterns and anomalies

Work on large datasets at scale

Data Preparation Demo

Migrate to AWS Glue

Build Data Lakes and Lake Houses for scalable data analysis

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

Process data in real time using Glue Streaming

Process petabytes of data both in batch using Apache Spark

Prepare data for Machine Learning

Thank you!