Getting Started with Serverless Glue

Sanchit Jain
7th May, Saturday
9:00 PM to 10:00 PM

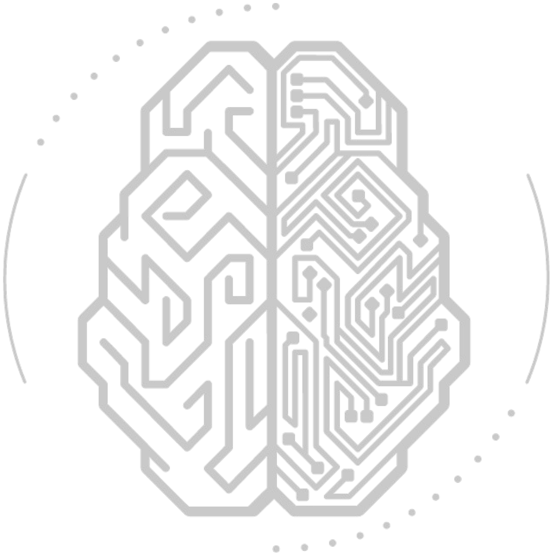
Speakers

Sanchit Jain

Lead Architect - AWS at Quantiphi
AWS APN Ambassador

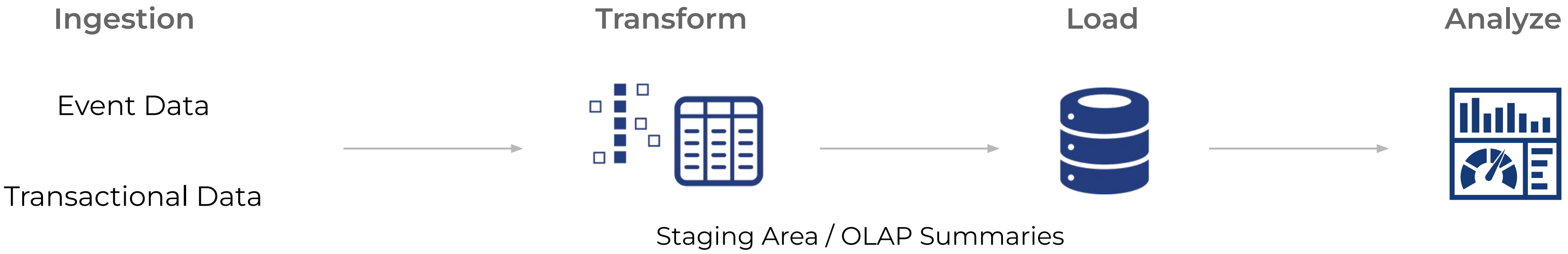
FOLLOW ME

How does a traditional ETL process look like?



How does a traditional ETL process look like?

Traditional ETL Migration Process Flow :



Ingestion

Event Data

Transactional Data

Transform

Load

Analyze

Staging Area / OLAP Summaries

Various Components of ETL Migration Process Pipeline :

1 **Ingestion**

Data is ingested from online transaction processing (OLTP) databases, today more commonly known just as 'transactional databases', and other data sources. OLTP applications have high throughput and they do not lend themselves well to data analysis or business intelligence tasks

2 **Transform**

Data is transformed in a staging area. These transformations cover both data cleansing and optimizing the data for analysis.

3 **Load**

The transformed data is then loaded into an online analytical processing (OLAP) database, today more commonly known as just an analytics database.

4 **Analyze**

Business intelligence (BI) teams then run queries on that data, which are eventually presented to end users, or to individuals responsible for making business decisions, or used as input for machine learning algorithms or other data science projects

Challenges faced by traditional ETL Migration

No real time analysis

Traditional ETL tends to rely on lengthy batch processing sessions and hence is time-consuming

Lacks data integration tools for all personas

ETL is not a data engineering function anymore.. Traditional ETL tools don't cater to all personas

Lock In

Jobs written in traditional ETL tools are proprietary to a specific vendor and cannot be ported.

Higher infrastructure & maintenance costs

Traditional ETL pipelines typically run on on-premise servers that require manpower and maintenance

Non scalable Infrastructure

ETL tools on premises are complex to install, manage and scale. They tend to sacrifice granularity of raw data for the sake of performance as data volumes grow

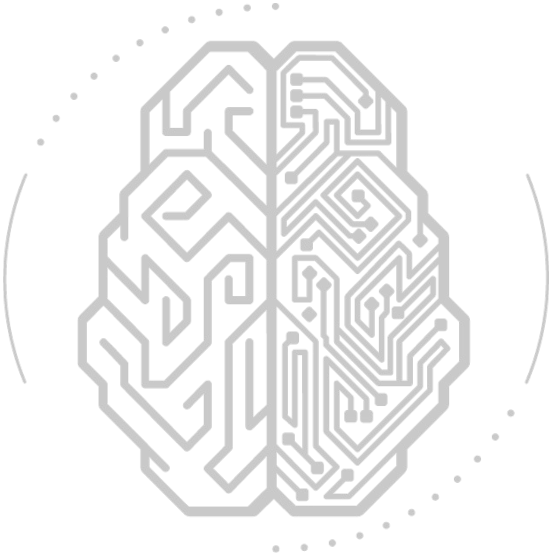
Specialised skill set

Requires dedicated data specialists to manage data warehouses

Impact on Business Processes

- Increased Dependency on Human Resources
- Increased cost as a percent of revenue
- Delayed Data Processes
- Delayed Time to Market

AWS Glue - Features & Benefits



AWS Glue Overview

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. AWS Glue provides all the capabilities needed for data integration so that you can start analyzing your data and putting it to use in minutes instead of months.

FEATURES

Serverless Spark

There is no infrastructure to maintain. Allocate needed compute power and run jobs. Job starts in few seconds and can run at petabyte scale

Data Integration for every user

Development environments catered to different skill sets

Cost Effective

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

Handles complex workloads

Glue connects to 60+ data sources, processes petabytes of data in real-time, batch and event driven modes

No Lock-In

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

More Power

AWS Glue automates much of the effort spent in building, maintaining, and running ETL jobs

AWS Glue benefits

Real time analysis

As you process streaming data in a Glue job, you have access to the full capabilities of Spark Structured Streaming to perform real time analysis of data

Data integration for all users

Development environments catered to different skill sets

No Lock In

Glue jobs are written open source Spark, Python and Scale

Cost - Effective

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

Highly scalable Infrastructure

AWS Glue is highly scalable and being on AWS cloud it scales up as per requirement

No Specialised skills required

There is visual ETL development for Data Engineers, notebook styled development for Data Scientists and no code development for Data Analysts

Cost Comparison

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

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

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

Impact on Business Processes

Reduction in human dependency

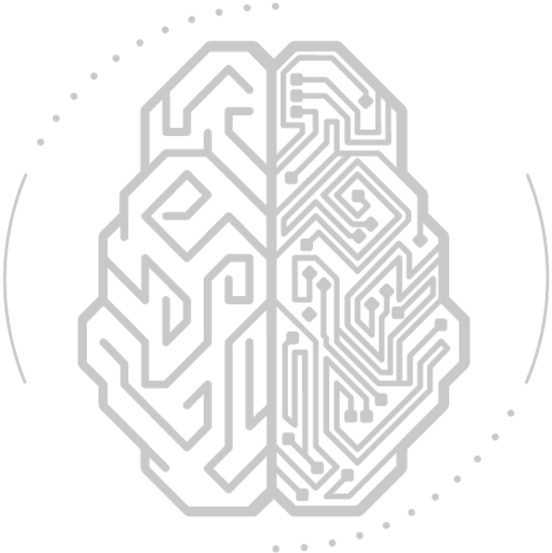
Revenue acceleration

Faster data processes

Better decision making

Reduction in time to market

AWS Glue - Components



AWS Glue: Components

Data Catalog

- Hive metastore compatible with enhanced functionality
- Crawlers automatically extract metadata and create tables
- Integrated with Athena, Amazon Redshift Spectrum

Job Authoring

- Auto-generates ETL code
- Builds on open frameworks—Python and Spark
- Developer-centric—editing, debugging, sharing

Job Execution

- Runs jobs on a serverless Spark platform
- Provides flexible scheduling
- Handles dependency resolution, monitoring, and alerting

Hive metastore compatible with enhancedfunctionality
Crawlers automatically extract metadata and create tables
Integrated with Athena, Amazon Redshift Spectrum
Auto-generates ETL code
Builds on open frameworks—Python and Spark
Developer-centric—editing, debugging, sharing
Runs jobs on a serverless Spark platform
Provides flexible scheduling
Handles dependency resolution, monitoring, and alerting

AWS Glue Data Catalog

Manage table metadata through a Hive metastore API or Hive SQL. Supported by tools like Hive, Presto, Spark, etc. AWS added a few extensions:

- **Search** over metadata for data discovery
- **Connection info—**JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated

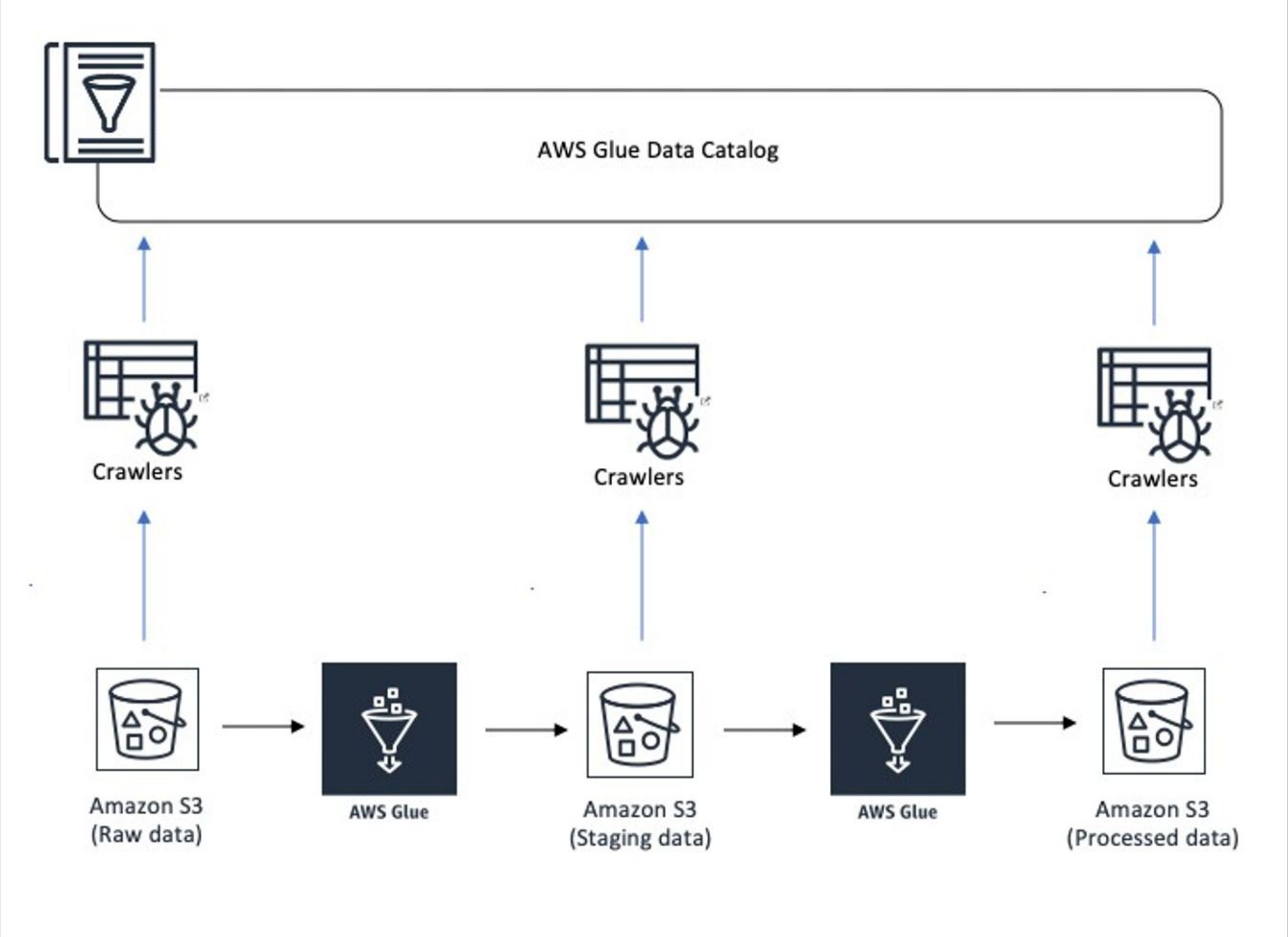
Populate using Hive DDL, bulk import, or automatically through **crawlers**

AWS Glue Data Catalog: Crawlers

Crawlers automatically build your Data Catalog and keep it in sync

- Automatically discover new data, extract schema definitions
  - Detect schema changes and version tables
  - Detect Hive style partitions on Amazon S3
- Built-in classifiers for popular types; custom classifiers using Grok expressions
- Run ad hoc or on a schedule; serverless—only pay when crawler runs

AWS Glue In Action



What is AWS Glue DataBrew ?

AWS Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code.

CAPABILITIES OF GLUE DATABREW

- Profile
- Clean and Normalize
- Map Data Lineage
- Automate

NEED FOR DATABREW

“Upto 80% of data analysis time is spent on preparing data”

Time Consuming

- Multi-step process to extract, clean, normalize & load data at scale
- The right tools for the right persona must be integrated

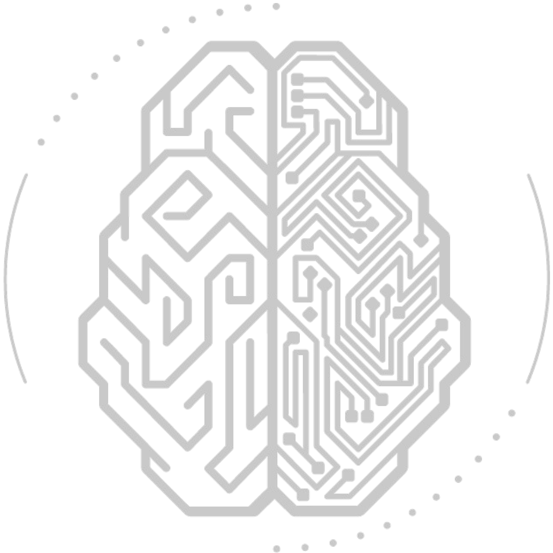
Expensive

- Costly user licenses & siloed tools that cause rework
- Often requires moving large amount of data into silos

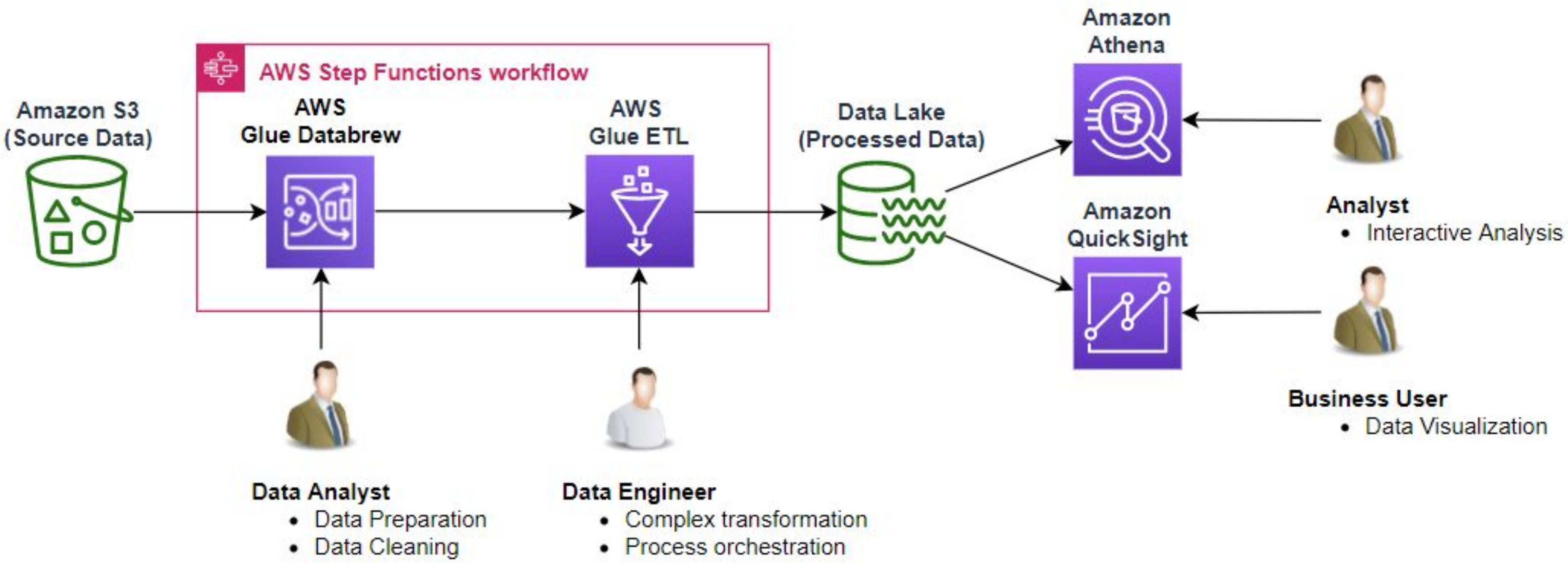
Manual

- Needs a lot of code-based heavy lifting to work at scale
- Hard to operationalize & build repeatable workflows

Demo



End to End Pipeline



Analyst
- Interactive Analysis

Business User
- Data Visualization

Data Analyst
- Data Preparation
- Data Cleaning

Data Engineer
- Complex transformation
- Process orchestration

THANK YOU