



## Unite Real-Time and Batch Analytics with AWS Glue

June, 2020

**Melody Yang**

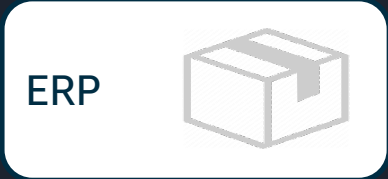
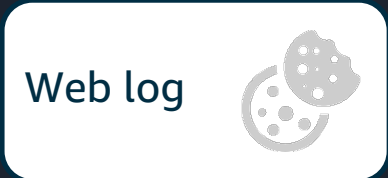
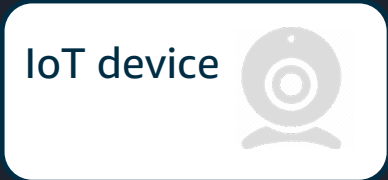
Data Specialist Solutions Architect, AWS

# Table of contents

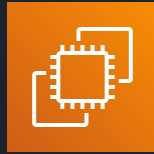
- Overview data analytics lifecycle
- Introduce AWS Glue
- Unite batch and real-time needs
- Demo

# Analytics Lifecycle Overview



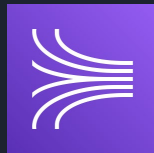


**Collect**



Polling Application

**Generate**



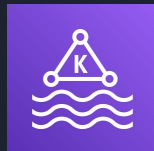
Amazon Kinesis

**Store**

**Extract  
Transform  
Load**

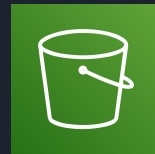
**Analyze**

**Visualize/  
Report**

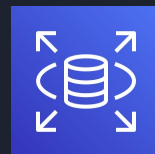


Amazon Managed  
Streaming for Kafka





Amazon S3

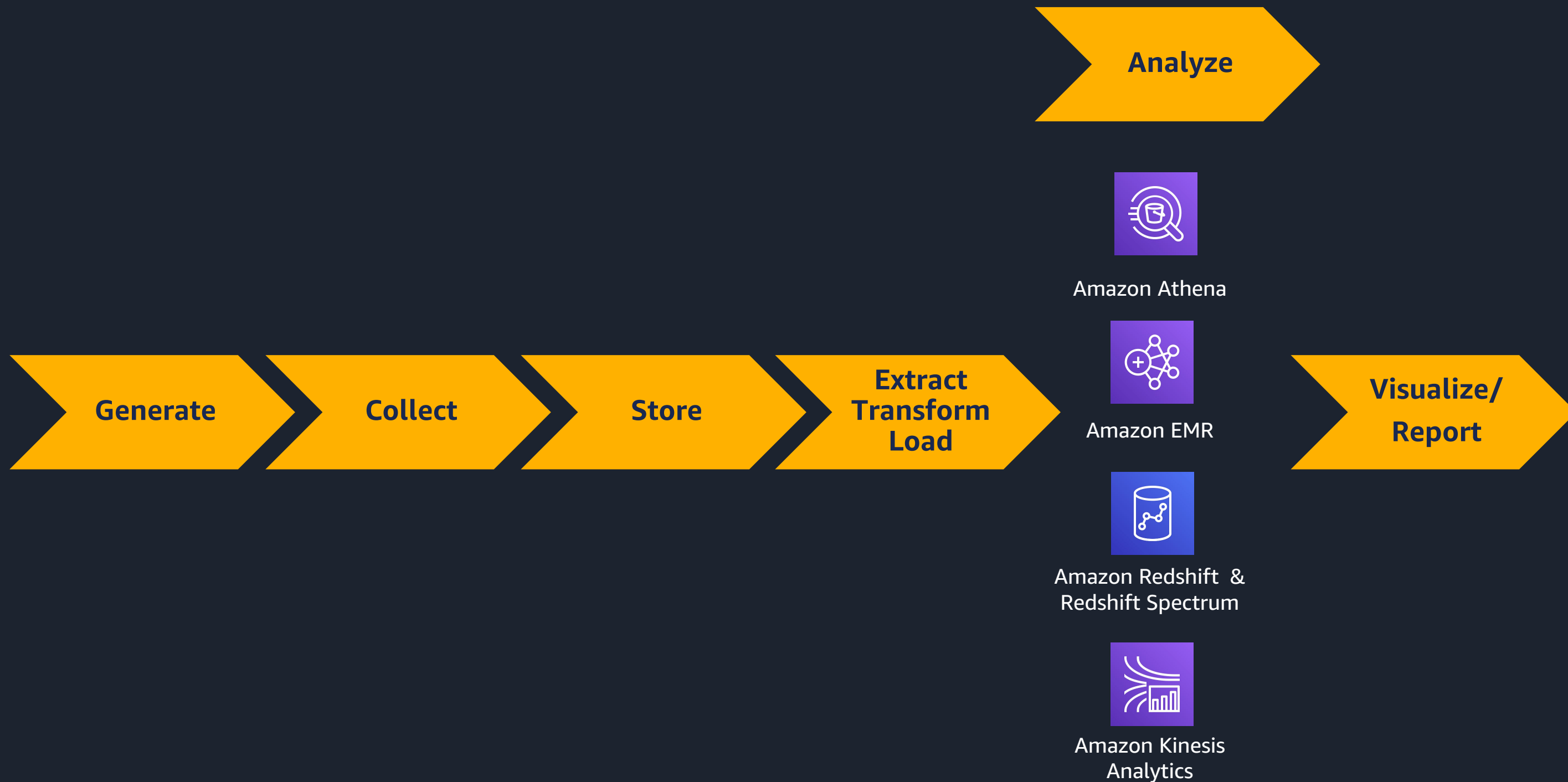


Amazon  
RDS

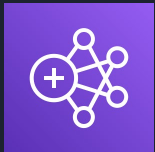


Database  
on EC2





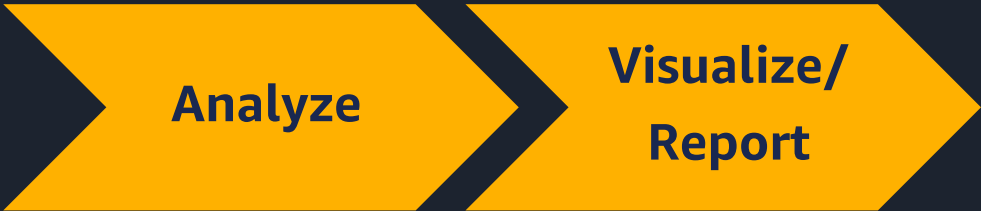




Amazon EMR

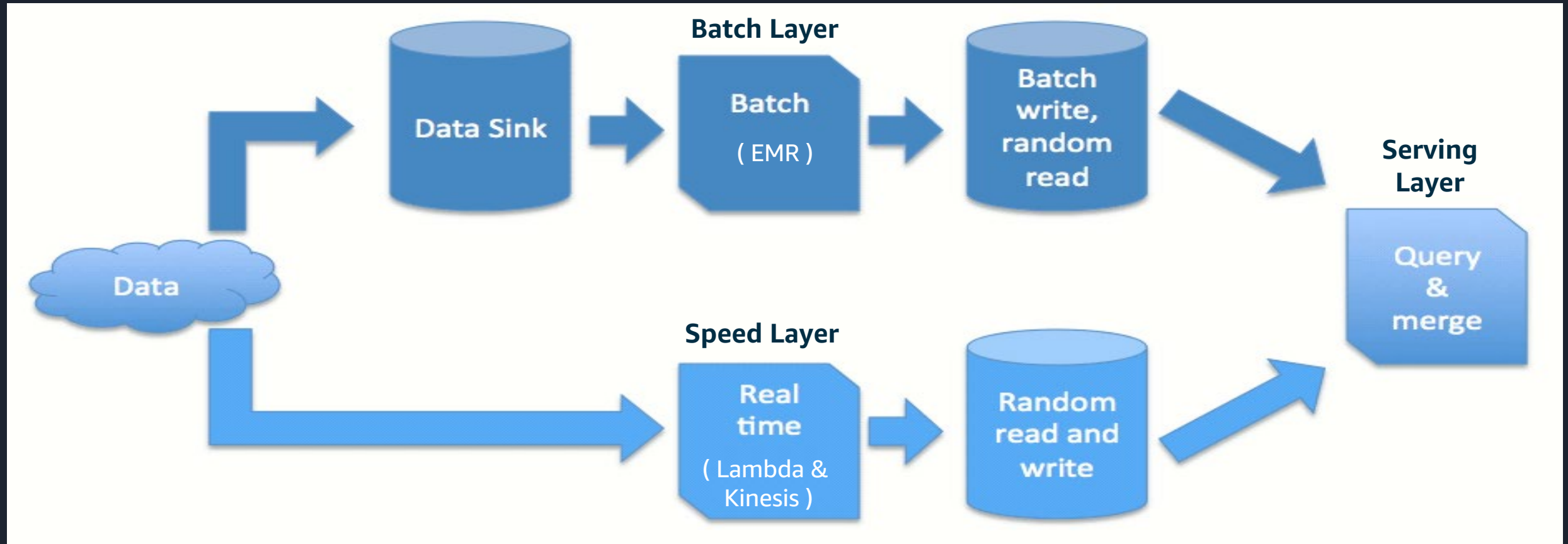


AWS  
Lambda

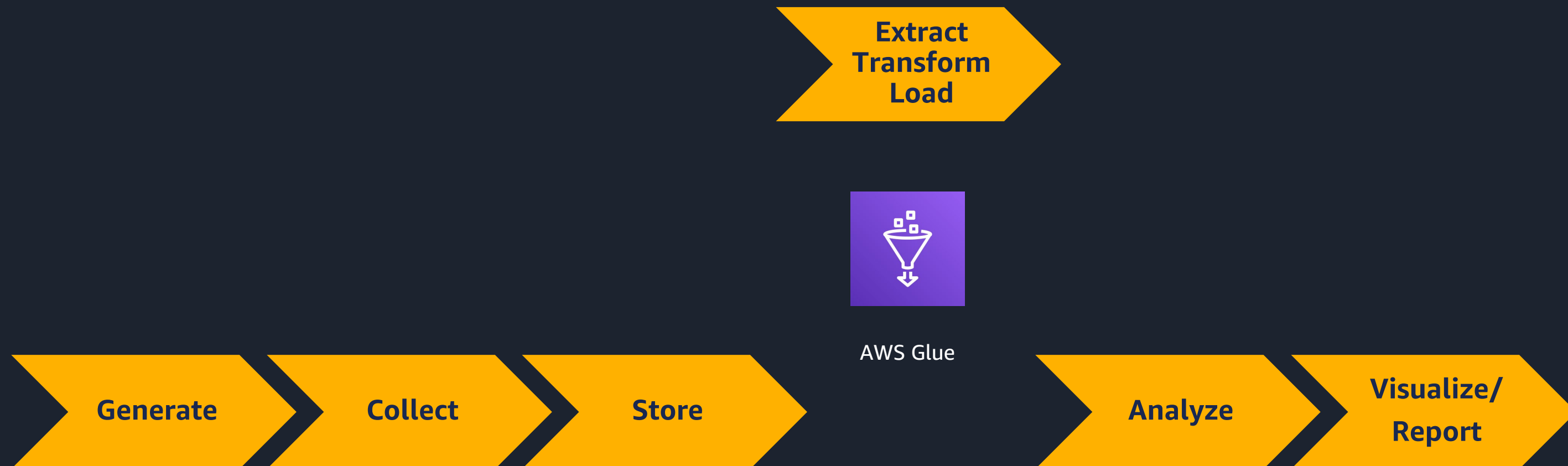


Amazon Kinesis  
Client Library (KCL)

# But....

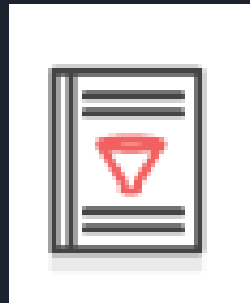


- **Separate** processing layers, **complex** in **operation**, **security** controls
- **Different** data **schema management** approaches
- **Large** proportion of ETL is **hand coding**



# What is AWS Glue

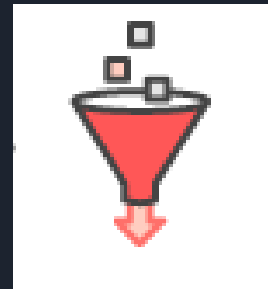
# AWS Glue components



## Data Catalog

### Discover

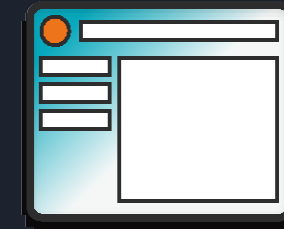
Automatic crawling  
Apache Hive Metastore compatible  
Integrated with AWS analytic services



## Serverless Jobs

### Develop

Apache Spark core  
Python and Scala  
Auto-generates ETL code



## Orchestration

### Deploy

Flexible scheduling  
Monitoring and alerting  
External integrations

# New: Serverless Streaming ETL with AWS Glue



You can create **streaming** extract, transform, and load (ETL) jobs that run continuously, consume data from streaming **sources** like:

- Amazon **Kinesis** Data Streams and
- **Apache Kafka** (including the fully-managed Amazon **MSK**)

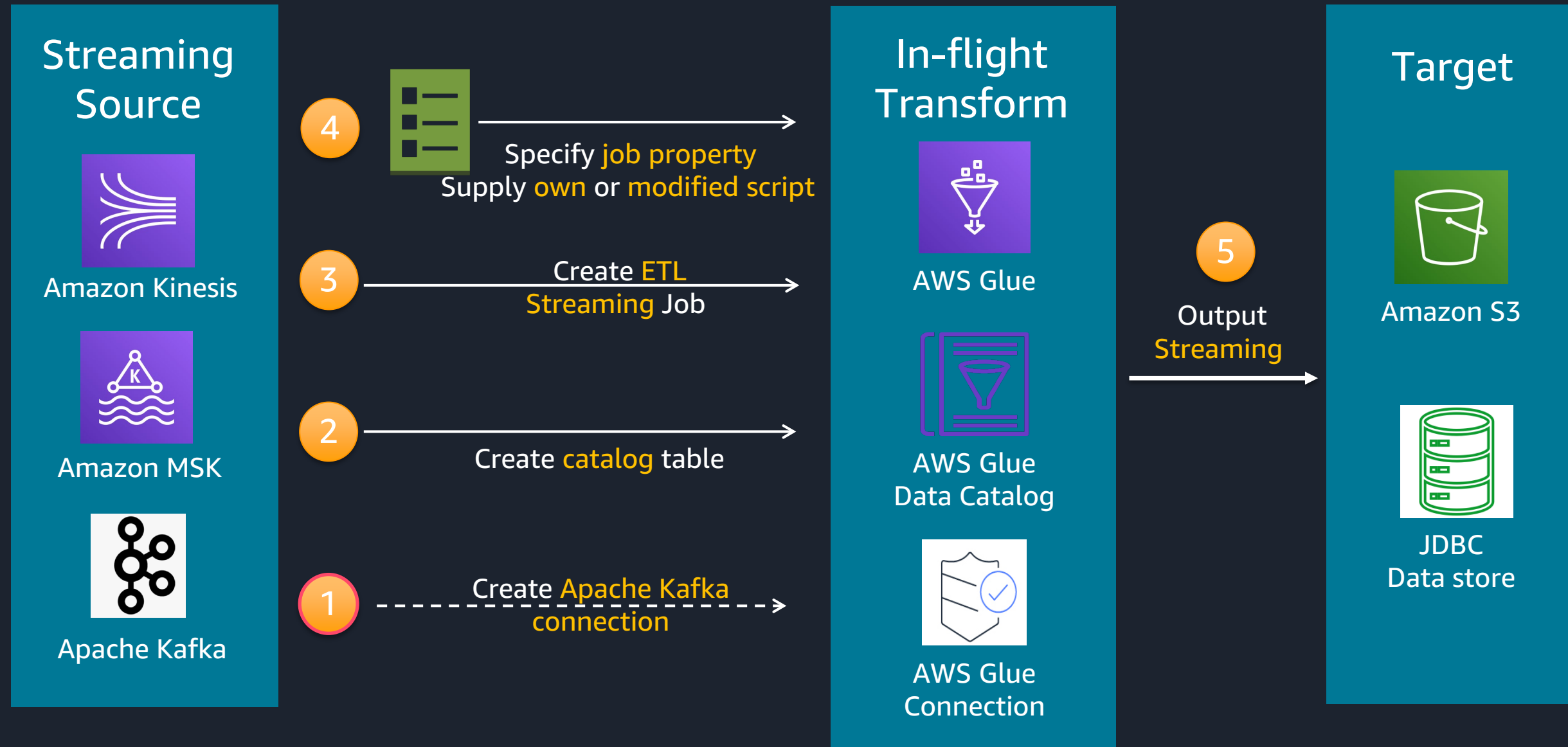


Load transformed results into **targets** like:

- Amazon **S3** data lakes or
- **JDBC** data stores



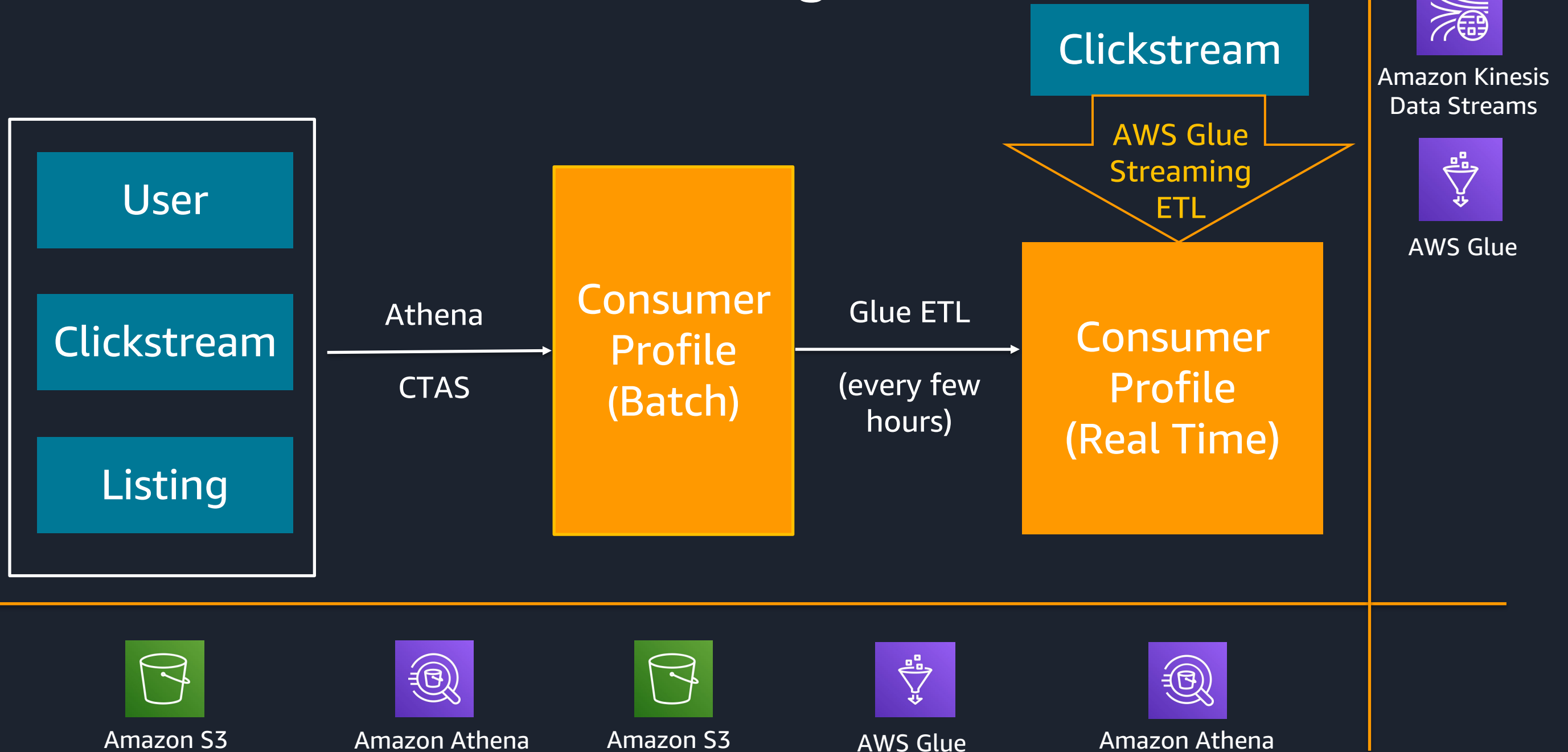
# Steps to Create Streaming ETL



# **Unite** Streaming and Batch ETL In AWS Glue

Easier and more cost-effective

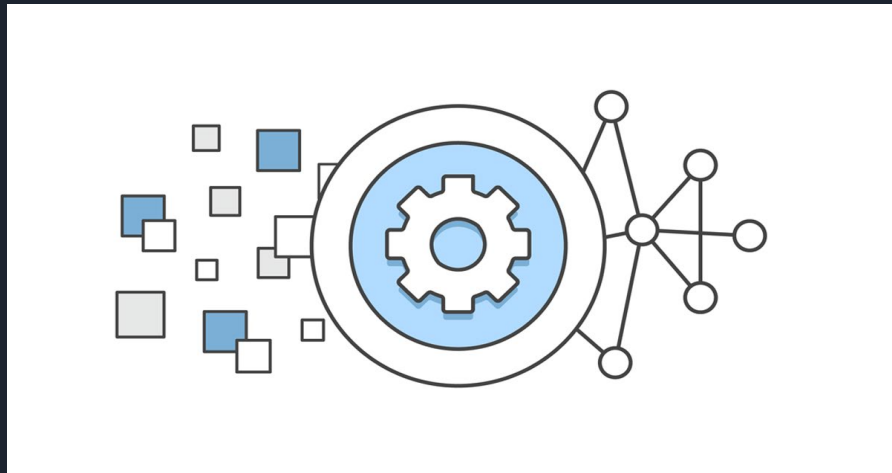
# Use Case: Consumer Profiling



# AWS Glue - Uniting Batch and Streaming

- **Speed** of implementation
- **Less code** to maintain plus code **auto-generate**
- Operations team loves the **serverless** aspect
- **Smaller** data process **footprints**

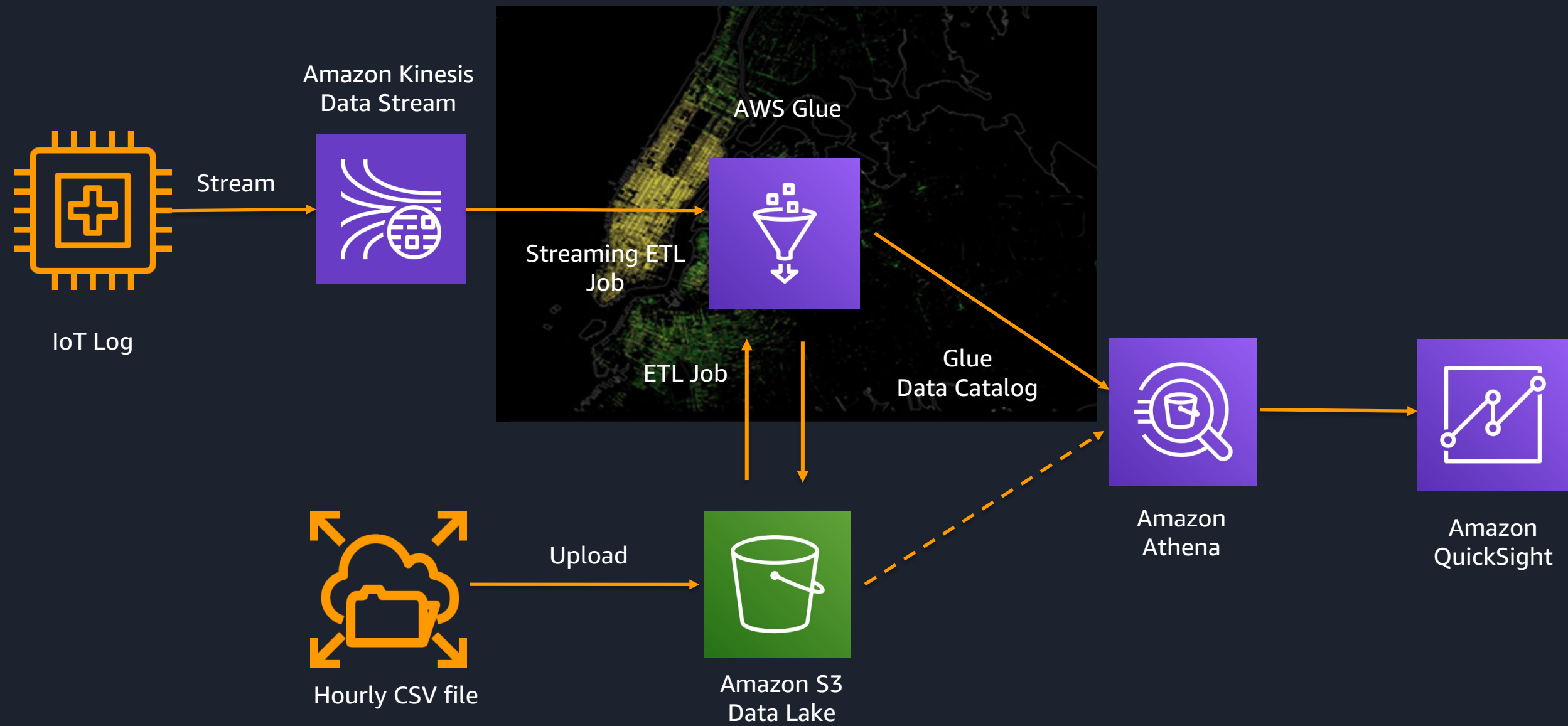
# Amazon Athena - Uniting Batch and Streaming



- Query data from Amazon S3 directly with ANSI SQL
- Use CREATE TABLE AS SELECT (CTAS) to create new tables using a result of SELECT query
- Serverless, no infrastructure to manage
- Pay \$5/TB scanned by your query
- Workgroups for cost control

Sounds Good In Theory...  
What's It Really Like?

# Demo context



# How Does Unified ETL in AWS Glue Help You?

- Consolidate data process architecture
- Reduce implementation efforts
- Less overhead in application maintenance
- Easier and more cost-effective, to set up serverless ETL pipelines
- Accelerates your insights by extending to real-time data
- Help you to focus on business outcomes of analytics.



# AWS Training and Certification



## Training for the Whole Team

Explore tailored Data or Database learning paths for customers and partners



## Flexibility to Learn Your Way

Build cloud skills with free digital Data training courses such as "The elements of Data Science", or dive deep with classroom training



## Validate Skills with AWS Certification

Demonstrate expertise with a Data industry-recognized credential (Data analytics and Database Specialty AWS Certifications)

<https://aws.amazon.com/training/>

# Visit the Data, Databases, and Analytics Resource Hub for more resources

Dive deeper with these newly created whitepapers and e-books to help you uncover new insights and value from your data

- An introduction to cloud databases
- Enter the purpose-built database era
- Harness the power of data
- Creating a modern analytics architecture
- The data-driven enterprise
- ... and more!









<https://tinyurl.com/aws-data-databases-analytics>

**Visit resource hub »**

# Thank you for attending AWS Data, Databases, and Analytics Online Series

We hope you found it interesting! A kind reminder to **complete the survey**.  
Let us know what you thought of today's event and how we can improve the event experience for you in the future.

-  [aws-apac-marketing@amazon.com](mailto:aws-apac-marketing@amazon.com)
-  [twitter.com/AWSCloud](https://twitter.com/AWSCloud)
-  [facebook.com/AmazonWebServices](https://facebook.com/AmazonWebServices)
-  [youtube.com/user/AmazonWebServices](https://youtube.com/user/AmazonWebServices)
-  [slideshare.net/AmazonWebServices](https://slideshare.net/AmazonWebServices)
-  [twitch.tv/aws](https://twitch.tv/aws)

# Thank you!