

# Glue

---

## Glue ETL

- Automatic code generation
- Scala or Python
- Encryption
  - Server-side (at rest)
  - SSL (in transit)
- Can be event-driven
- Can provision additional “DPU’s” (data processing units) to increase performance of underlying Spark jobs
  - Enabling job metrics can help you understand the maximum capacity in DPU’s you need
- Errors reported to CloudWatch
  - Could tie into SNS for notification

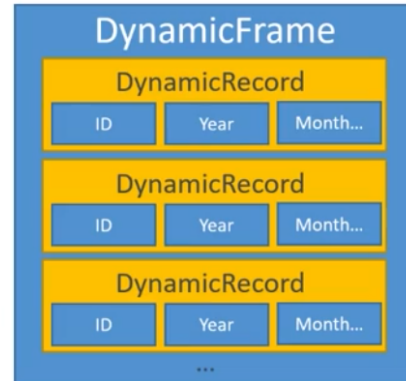
## Glue ETL

- Transform data, Clean Data, Enrich Data (before doing analysis)
    - Generate ETL code in Python or Scala, you can modify the code
    - Can provide your own Spark or PySpark scripts
    - Target can be S3, JDBC (RDS, Redshift), or in Glue Data Catalog
  - Fully managed, cost effective, pay only for the resources consumed
  - Jobs are run on a serverless Spark platform
- ↳
- Glue Scheduler to schedule the jobs
  - Glue Triggers to automate job runs based on “events”

# Glue ETL: The DynamicFrame

- A DynamicFrame is a collection of DynamicRecords
  - DynamicRecords are self-describing, have a schema
  - Very much like a Spark DataFrame, but with more ETL stuff
  - Scala and Python APIs

```
val pushdownEvents = glueContext.getCatalogSource(  
  database = "githubarchive_month", tableName = "data")  
  
val projectedEvents = pushdownEvents.applyMapping(Seq(  
  ("id", "string", "id", "long"), ("type", "string", "type",  
    "string"), ("actor.login", "string", "actor", "string"),  
  ("repo.name", "string", "repo", "string"),  
  ("payload.action", "string", "action", "string"),  
  ("org.login", "string", "org", "string"), ("year",  
    "string", "year", "int"), ("month", "string", "month",  
    "int"), ("day", "string", "day", "int") ))
```



## Glue ETL - Transformations

- Bundled Transformations:
  - DropFields, DropNullFields – remove (null) fields
  - Filter – specify a function to filter records
  - Join – to enrich data
  - Map - add fields, delete fields, perform external lookups
- Machine Learning Transformations:
  - **FindMatches ML:** identify duplicate or matching records in your dataset, even when the records do not have a common unique identifier and no fields match exactly.
- Format conversions: CSV, JSON, Avro, Parquet, ORC, XML
- Apache Spark transformations (example: K-Means)

# Glue ETL: ResolveChoice

- Deals with ambiguities in a DynamicFrame and returns a new one
- For example, two fields with the same name.
- `make_cols`: creates a new column for each type
  - `price_double`, `price_string`
- `cast`: casts all values to specified type
- `make_struct`: Creates a structure that contains each data type
- `project`: Projects every type to a given type, for example `project:string`

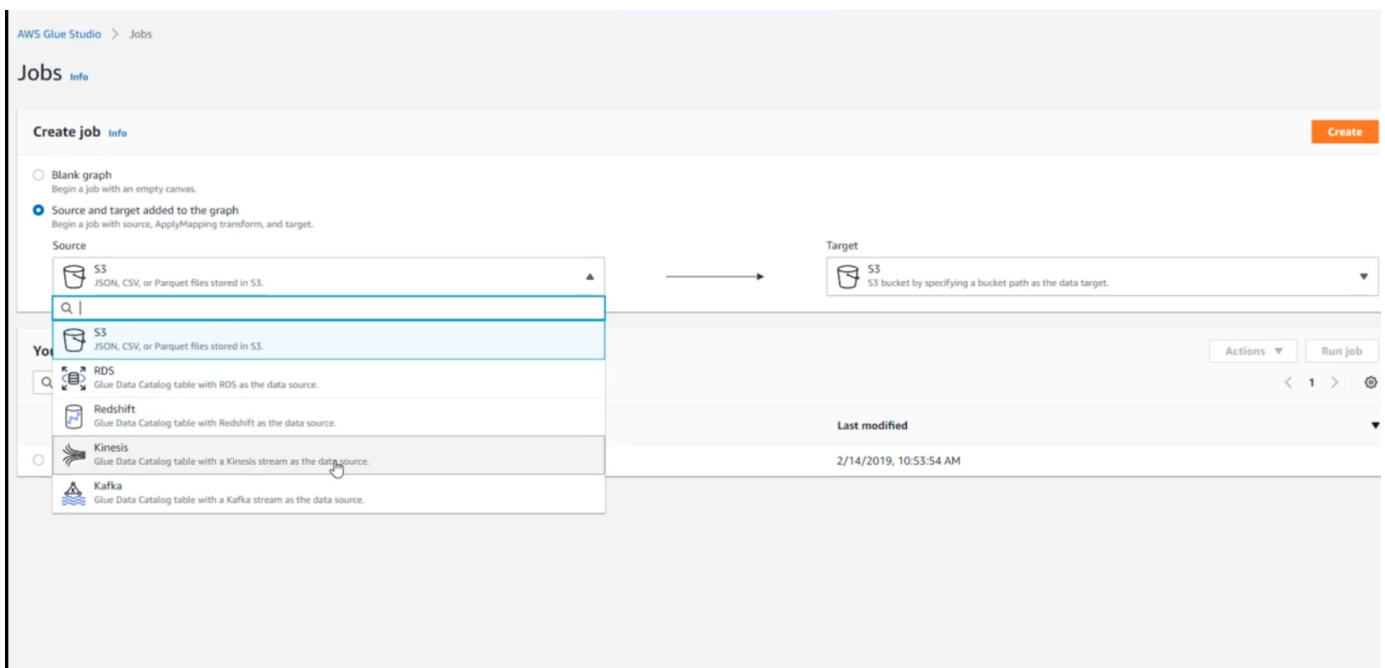
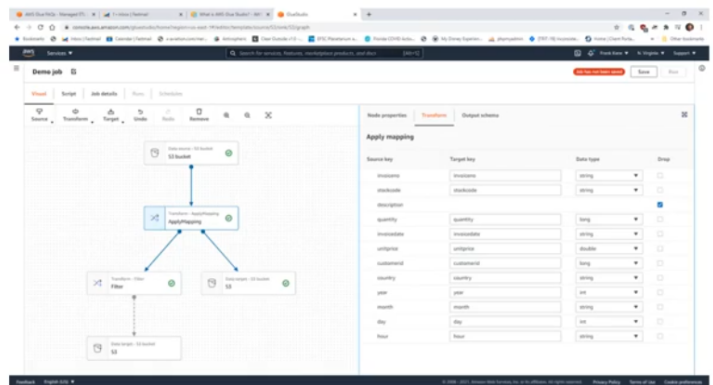
↳

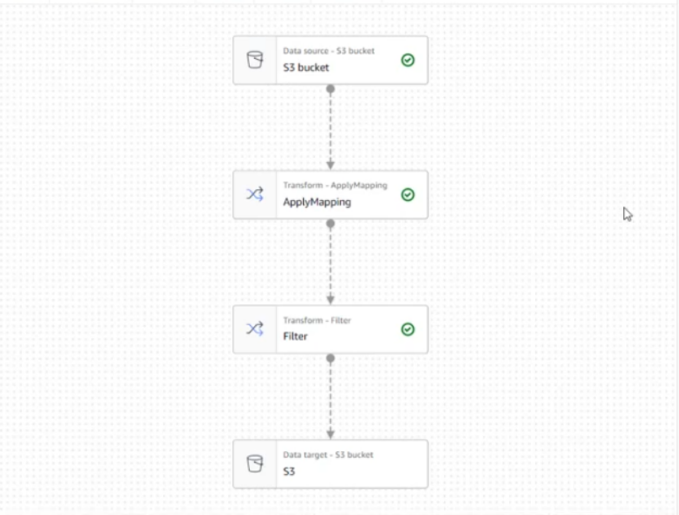
```
"myList": [ { "price": 100.00 }, { "price": "$100.00" } ]
```

```
df1 = df.resolveChoice(choice = "make_cols")  
df2 = df.resolveChoice(specs = [{"myList[].price",  
                                "make_struct"}, ("columnA", "cast:double")])
```

## AWS Glue Studio

- Visual interface for ETL workflows
- Visual job editor
  - Create DAG's for complex workflows
  - Sources include S3, Kinesis, Kafka, JDBC
  - Transform / sample / join data
  - Target to S3 or Glue Data Catalog
  - Support partitioning
- Visual job dashboard
  - Overviews, status, run times

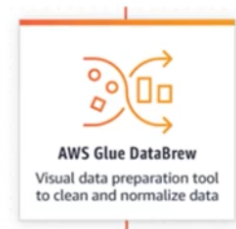




No node selected  
Choose a node from the graph to view its configuration properties.

# AWS Glue DataBrew

- A visual data preparation tool
  - UI for pre-processing large data sets
  - Input from S3, data warehouse, or database
  - Output to S3
- Over 250 ready-made transformations
- You create “recipes” of transformations that can be saved as jobs within a larger project
- May define data quality rules
- May create datasets with custom SQL from Redshift and Snowflake
- Security
  - Can integrate with KMS (with customer master keys only)
  - SSL in transit
  - IAM can restrict who can do what
  - CloudWatch & CloudTrail



## AWS Glue DataBrew

Clean and normalize data up to 80% faster

AWS Glue DataBrew is a visual data preparation tool that enables users to clean and normalize data without writing any code, to reduce the time it takes to prepare data for analytics and machine learning (ML) by up to 80% compared to today's conventional, code-based data preparation. You can choose from over 250 pre-built transformations to automate data preparation tasks, such as filtering anomalies, converting data to standard formats, and correcting invalid values, all without the need to write code.

### Create a project

Use your data to get started.

Create project

Discover data preparation and transformation using one of our sample datasets.

Create sample project

### Pricing

For AWS Glue DataBrew, the interactive sessions are billed per session and the DataBrew jobs are billed per minute. Each session is 30 minutes.

The first 40 interactive sessions are free for the first time users of DataBrew.

Interactive session \$1.00 per session

DataBrew jobs \$0.48 per node hour

Cost calculator

Learn more

### How it works



Sample project - 2

Dataset: dataset-national-baby-names | Sample: First n sample (500 rows)

Create job

VIEWING 5 columns 500 rows

#	count	gender	id	name	year
1	1	F	1	Mary	1880
2	1	F	2	Anna	1880
3	1	F	3	Emma	1880
4	1	F	4	Elizabeth	1880
5	1	F	5	Minnie	1880
6	1	F	6	Margaret	1880
7	1	F	7	Ida	1880
8	1	F	8	Alice	1880
9	1	F	9	Bertha	1880
10	1	F	10	Sarah	1880
11	1	F	11	Annie	1880
12	1	F	12	Clara	1880
13	1	F	13	Ella	1880
14	1	F	14	Florence	1880
15	1	F	15	Corra	1880
16	1	F	16	Martha	1880
17	1	F	17	Laura	1880
18	1	F	18	Nellie	1880

Zoom 100%

ADD STEP

- Change to uppercase
- Change to lowercase
- Change to capital case
- Change to sentence case
- Date-time formats

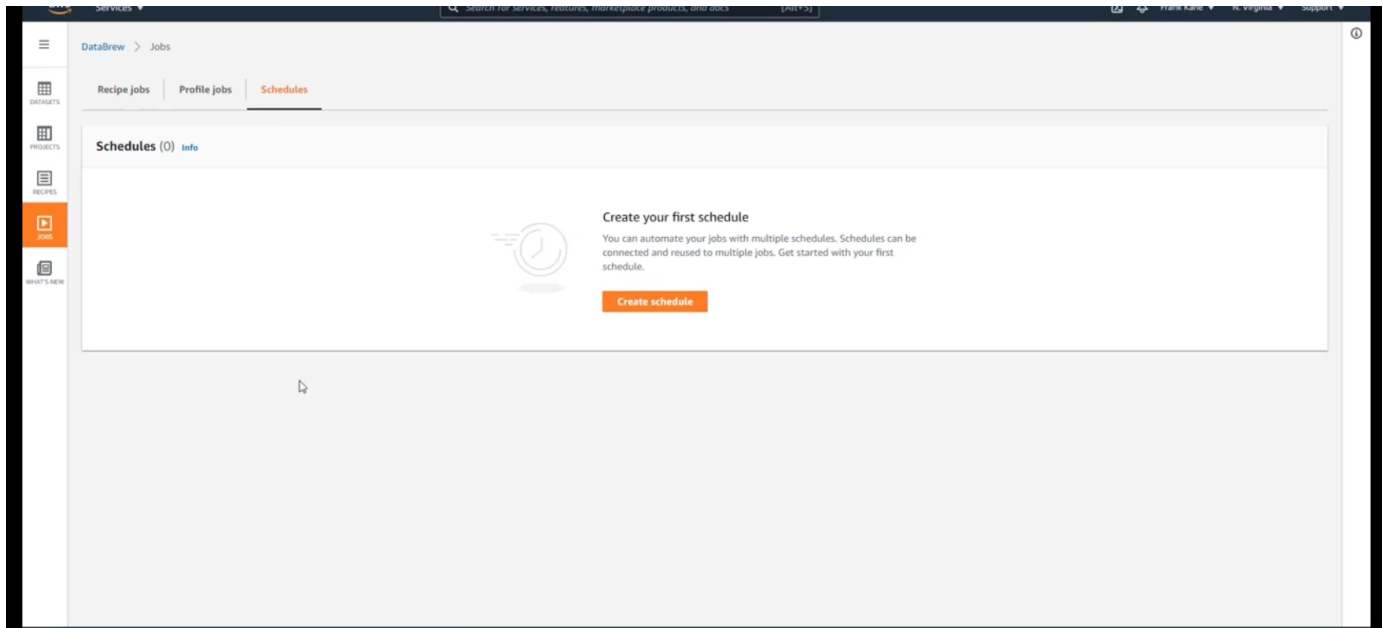
CLEAN

- Remove special characters
- Remove numbers
- Remove letters
- Remove punctuation
- Remove custom value
- Remove white spaces
- Remove quotation marks
- Add prefix
- Add suffix
- Add quotation marks
- Replace between delimiters
- Replace between positions
- Replace value or pattern

EXTRACT VALUES

- Extract value between delimiters
- Extract value between positions

you can also create jobs to schedule run the automate are schedules a task when time is up



## AWS Glue Elastic Views

- Coming soon!
- Builds materialized views from Aurora, RDS, DynamoDB
- Those views can be used by Redshift, Elasticsearch, S3, DynamoDB, Aurora, RDS
- SQL interface
- Handles any copying or combining / replicating data needed
- Monitors for changes and continuously updates
- Serverless

