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

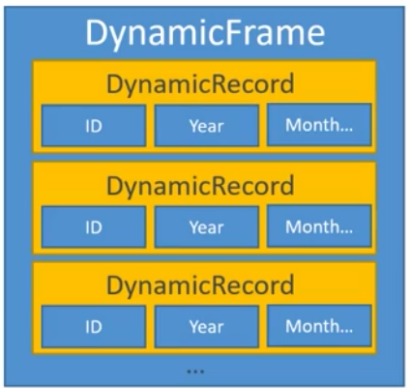
- Glue Scheduler to schedule the jobs
- Glue Triggers to automate job runs based on “events”

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

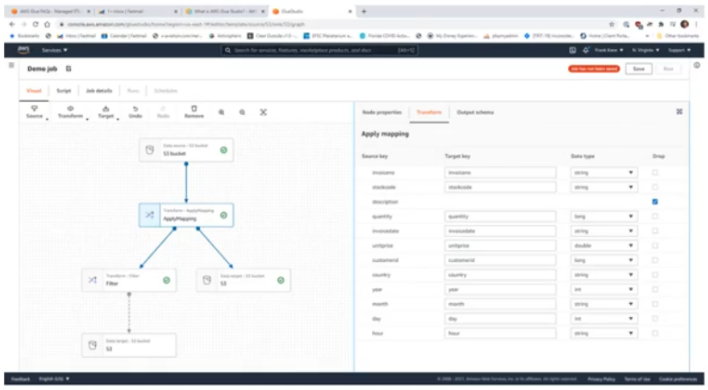
- Deals with ambiguities in a DynamicFrame and returns a new one
- For example, two fields with the same name.
- make_cols: creates a new column for each type
  - price_double, price_string
- cast: casts all values to specified type
- make_struct: Creates a structure that contains each data type
- project: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

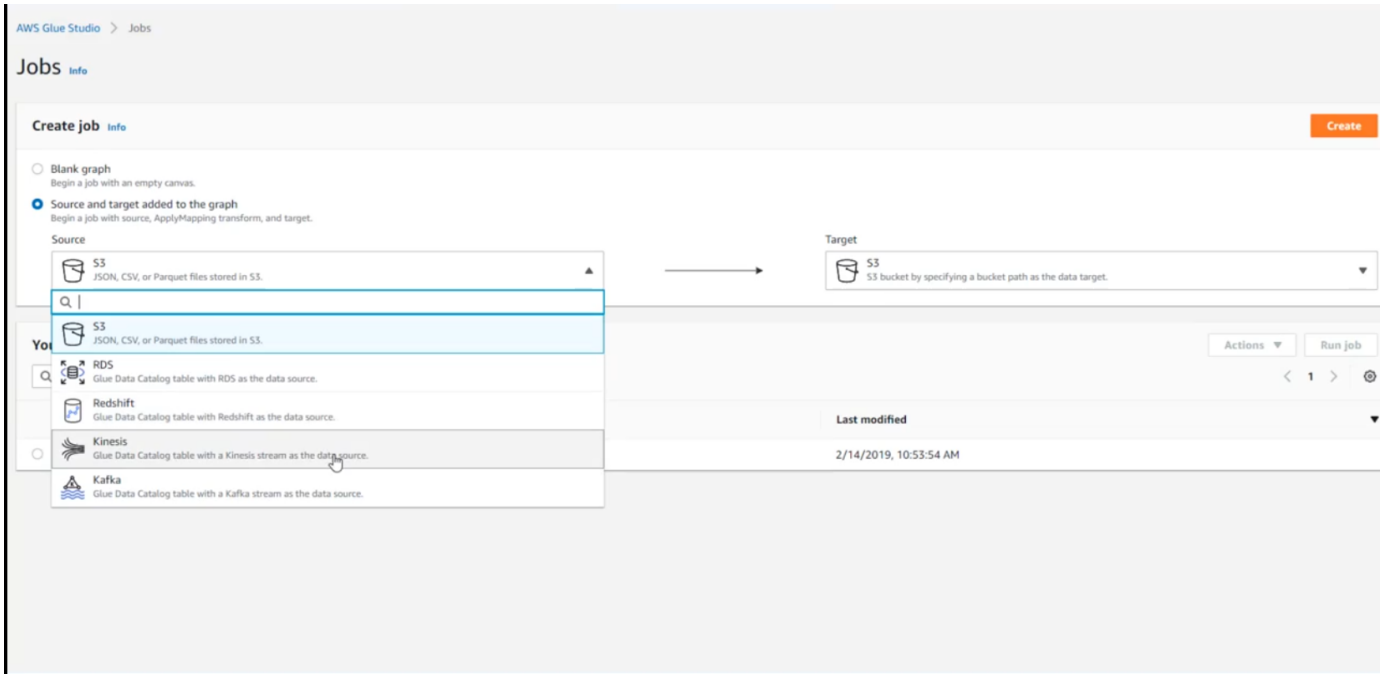
```
"myList": [ { "price": 100.00 }, { "price": "$100.00" } ]
```

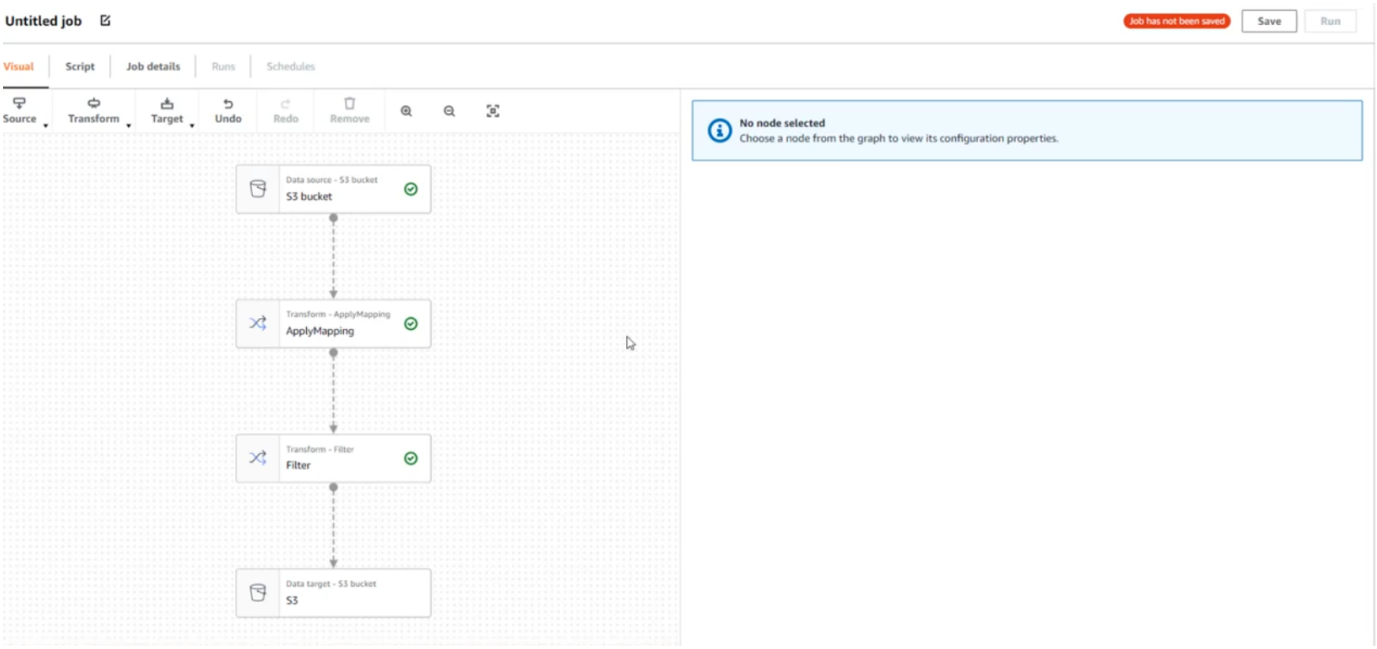
```
df1 = df.resolveChoice(choice = "make_cols")
df2 = df.resolveChoice(specs = [("myList[].price",
"make_struct"), ("columnA", "cast:double")])
```

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

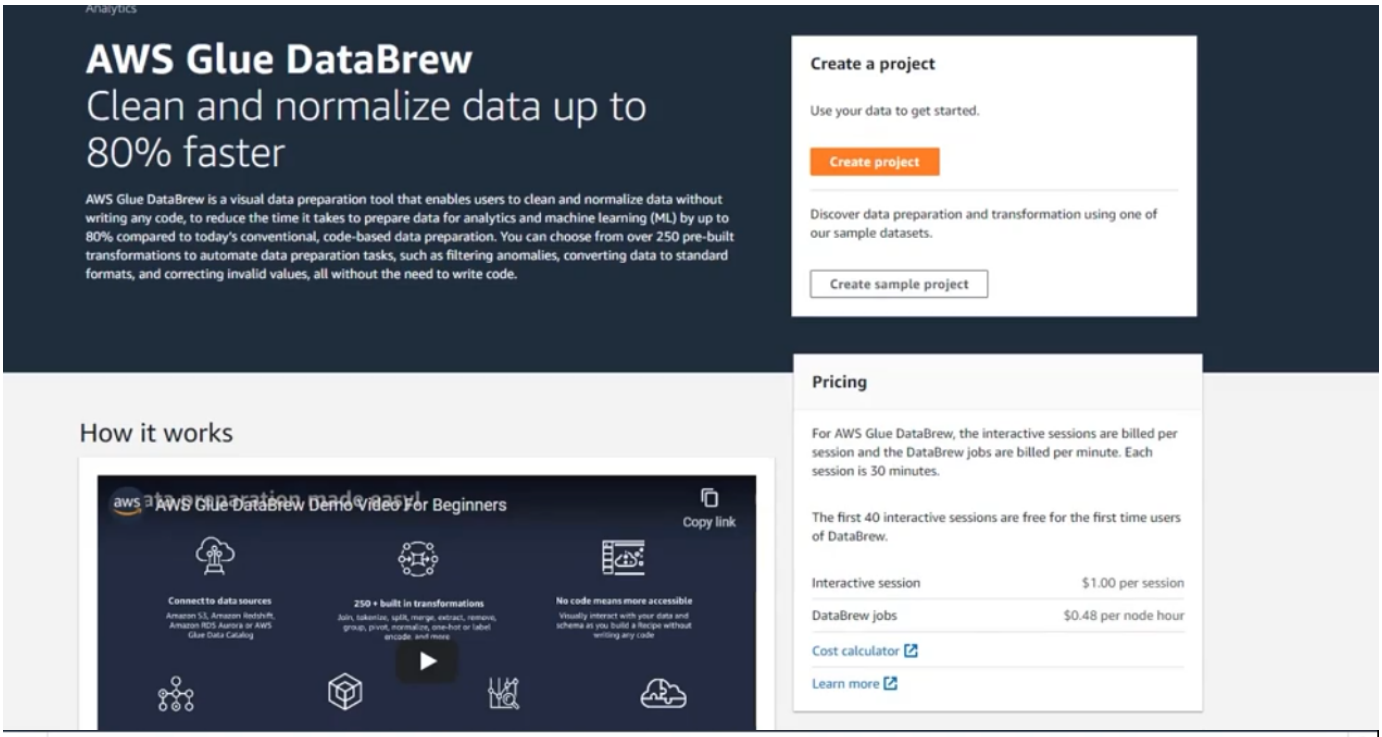


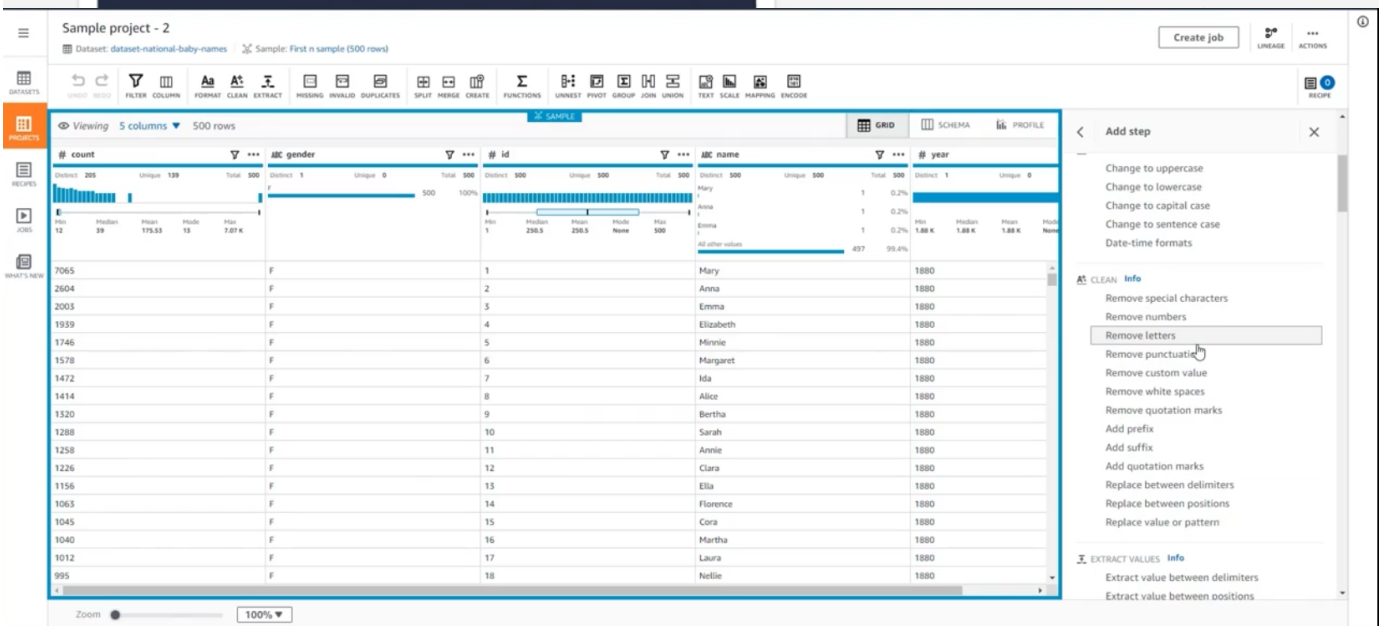



Untitled job
Job has not been saved
Save
Run
Visual
Script
Job details
Runs
Schedules
Source
Transform
Target
Undo
Redo
Remove
Data source - S3 bucket
S3 bucket
Transform - ApplyMapping
ApplyMapping
Transform - Filter
Filter
Data target - S3 bucket
S3
No node selected
Choose a node from the graph to view its configuration properties.

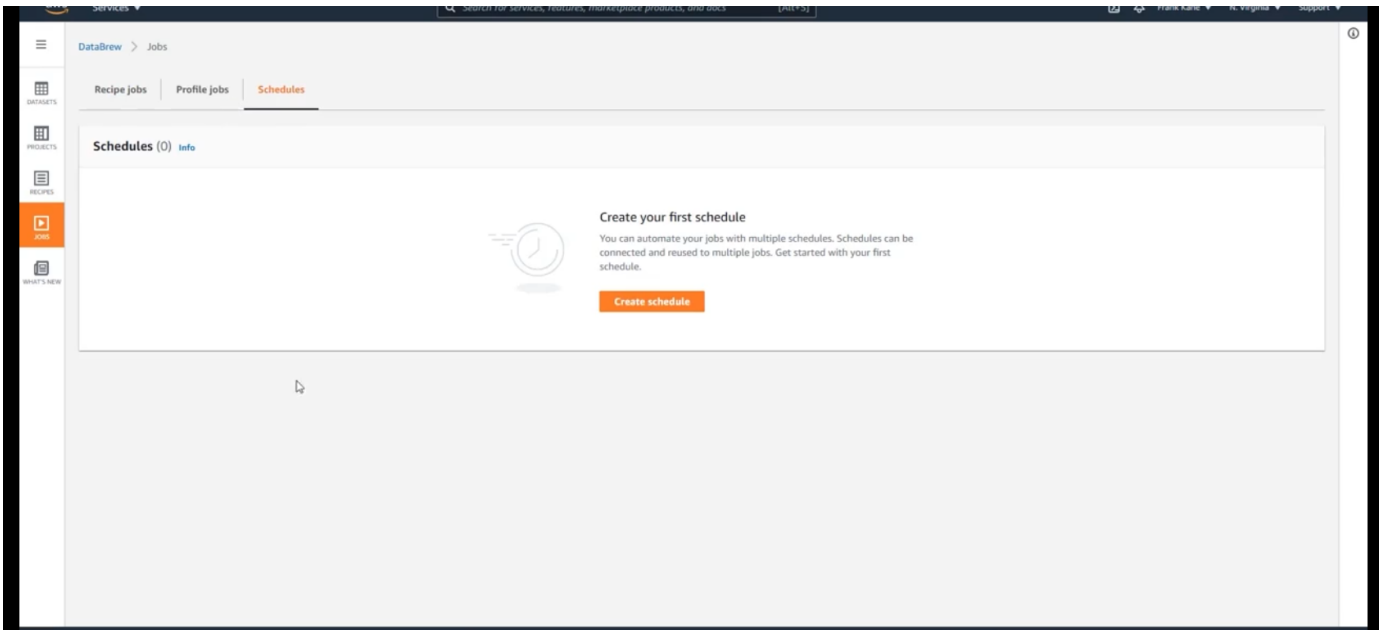
AWS Glue DataBrew

- 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





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