**Glue**

* Glue is a fully-managed pay as you go extract, transform and load ETL service that automates the time-consuming steps of data preparation for analytics
* Glue automatically discovers and profiles data via the Glue Data Catalog, recommends and generates ETL code to transform your source data into target schemas
* AWS Glue runs the ETL job on a fully managed, scale-out Spark environment to load your data into its destination
* Glue also allows you to setup, orchestrate and monitor complex data flows
* You can create and run an ETL job with a few clicks in the AWS Management Console
* Simply point AWS Glue to your data stored on AWS and AWS Glue discovers data and stores the associated metadata(e.g. table definition and schema) in the AWS Glue Data Catalog
* Once cataloged, data is immediately searchable, queryable and available for ETL
* Glue consists of a Data Catalog which is a central metadata repositor, an ETL engine that can automatically generate Scala or Python code, and a flexible scheduler that handles dependency resolution, job monitoring and retries
* Together these automate much of the undifferentiated heavy lifting involved with discovering, categorizing, cleaning, enriching, and moving data so you can spend more time analyizing your data
* AWS Glue crawlers connect to a source or target data store, progress through a prioritized list of classifiers to determine the schema for the data, and then creates metadata in the AWS Glue Data Catalog
* The metadata is stored in tables in a data catalog and used in the authoring process of ETL jobs
* You can run crawlers on a schedule, on-demand, or trigger them based on an event to ensure that your metadata is up-to-date
* Glue automatically generates the code to extract, transform and load data
* Simply point Glue to a source and target, and Glue creates ETL scripts to transform, flatten and enrich the data
* The code is generated in Scala and Python and written for Spark  
  Glue helps clean and prepare data for analysis by providing a ML Transform called FindMatches for deduplication and finding matching records
* Use Cases:
  + Glue is used to discover properties of data, transform it and prepare it for analytics
  + Glue can automatically discover both structured and semi-structured data stored in data lakes on S3, data warehouses on Redshift and various databaeses running on AWS
  + It provides a unified view of data via the Glue Data Catalog that is available for ETL, querying and reporting using services like Athena, EMR, Redshift Spectrum
  + Glue automatically gnerates Scala or Phython code for ETL jobs that you can further customize using tools you are already familiar with
  + Glue is serverless, so there are no compute resources to configure and manage