 **Data Catalog**:

* Acts as a persistent technical metadata store.
* Can connect to 70 different data sources and manage your data in a centralized data catalog.

 **ETL Engine**:

* Allows you to visually create, run, and monitor ETL pipelines to load data into your data lakes, with Spark built in.

 **Data Catalog Discovery**: AWS Glue Data Catalog is queried to identify and retrieve metadata about the source data (e.g., from Amazon S3, RDS, or Redshift) stored in Glue Databases and tables.

 **Job Definition**: You define an ETL job using the Glue console, API, or CLI. This includes specifying the data sources, transformations, and target data stores.

 **ETL Script Generation**: Glue automatically generates a PySpark or Scala script based on your job definition, or you can provide a custom script.

 **Job Execution**: The ETL job is submitted to a managed Apache Spark environment. Glue allocates the necessary resources and runs the script.

* Data is extracted from the source.
* Transformations (e.g., cleaning, filtering, aggregating) are applied.
* Data is loaded into the target (e.g., data lake or data warehouse).

 **Monitoring and Logging**: The job's progress is monitored via the Glue console, with logs available for troubleshooting.

 **Completion**: Once the job finishes, the transformed data is available in the target location, and the metadata is updated in the Data Catalog.

1. **Find the Data**: AWS Glue looks at the Data Catalog to locate your data (e.g., from Amazon S3 or RDS) using the metadata stored in Glue Databases and tables.
2. **Set Up the Job**: You create an ETL job in the Glue console or with a script, deciding where the data comes from and where it should go.
3. **Generate the Plan**: Glue creates a script (usually in PySpark) to handle the data processing, or you can write your own.
4. **Run the Job**: Glue uses Apache Spark to:
   * Pull data from the source.
   * Change or clean the data as needed.
   * Save it to the destination (like a data lake).
5. **Check Progress**: You can watch the job’s status and logs in the Glue console to see how it’s doing.
6. **Finish Up**: When done, the data is in the new location, and the Data Catalog updates with the changes.