wavicle
DATA SOLUTIONS

Checklist: Best practices for migrating from IBM DataStage to AWS Glue

If your organization is planning a migration from on-premises IBM DataStage to serverless data integration with AWS Glue, there are several steps to keep in mind, from development and testing through deployment. Below are some helpful steps to consider when making this important transition.

Development and testing environments:

□	Setup AWS Glue libraries (v1.0) available through public Amazon S3 buckets	**NOTES**
□	Consider packaging the libraries as Docker container for portability	
□	Setup an IDE like PyCharm/Jupyter	
□	Explore using Glue development endpoints and Glue Studio with an interactive approach, as alternative	
□	Build unit test suites leveraging the local libraries:	
	□ Mock data or use sample files	
	□ PyTest or ScalaTest	
	□ Modularize the code for streamlined testing	
	□ Integrate with your source code repository locally	
□	Leverage the full open-source Spark APIs	

Deployment environment:

□	Confirm the network communication paths are available for your resources	**NOTES**
□	Subnet config, Firewall, DNS configs	
□	Available IP addresses for higher DPU jobs	
□	Ensure AWS Glue has the right access	
□	VPC FlowLogs can be used to troubleshoot connectivity issues	
□	Explore using Amazon S3 access via Endpoint	
□	Consider using AWS Glue with VPC Endpoints	
□	Start with the standard worker type	
□	Use the G.1X or G.2X worker types for memory intensive jobs	
□	Set up Spark UI for better details about job metrics and performance (Spark jobs)	

Deployment process:

□	Confirm the network communication paths are available for your resources	**NOTES**
□	Maintain the Glue crawler / job definition on your source code repo	
	□ JSON file or CloudFormation templates	
□	Depending on your git lifecycle practices, build/create the updated codebase	
	□ Example: Create Python library, jar, configuration files, modified job definition etc.)	
	□ Execute the test cases on local sample data	
	Deploy the artifacts to a staging environment on AWS	
	□ Create/update the AWS Glue crawler/jobse	
	□ Move the generated libraries and scripts to Amazon S3	
	□ Run manual/automated integration tests, data validation	
	□ Approve production deployment	

Wavicle is proud to have achieved the AWS Service Delivery designation for AWS Glue. Learn more about how we can help you modernize your data integration with AWS Glue.

wavicledata.com ☎ +1 (630) 756-2632 ✉ info@wavicledata.com