

You build your serverless data lake with [Amazon Simple Storage Service (Amazon S3)](https://aws.amazon.com/s3/) as the primary data store. Given the scalability and high availability of Amazon S3, it is best suited as the single source of truth for your data.

You can use various techniques to ingest and store data in Amazon S3. For example, you can use [Amazon Kinesis Data Firehose](https://aws.amazon.com/kinesis/data-firehose/) to ingest streaming data. You can use [AWS Database Migration Service](https://aws.amazon.com/dms/) (AWS DMS) to ingest relational data from existing databases. And you can use [AWS DataSync](https://aws.amazon.com/datasync/) to ingest files from an on-premises Network File System (NFS).

Ingested data lands in an Amazon S3 bucket that we refer to as the raw zone. To make that data available, you have to catalog its schema in the [AWS Glue](https://aws.amazon.com/glue/) Data Catalog. You can do this using an [AWS Lambda](https://aws.amazon.com/lambda/) function invoked by an [Amazon S3 trigger](https://docs.aws.amazon.com/AmazonS3/latest/user-guide/enable-event-notifications.html) to start an [AWS Glue crawler](https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html) that catalogs the data. When the crawler is finished creating the table definition, you invoke a second Lambda function using an [Amazon CloudWatch Events rule](https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/Create-CloudWatch-Events-Scheduled-Rule.html). This step starts an [AWS Glue ETL job](https://docs.aws.amazon.com/glue/latest/dg/add-job.html) to process and output the data into another Amazon S3 bucket that we refer to as the processed zone.

The AWS Glue ETL job converts the data to Apache Parquet format and stores it in the processed S3 bucket. You can modify the ETL job to achieve other objectives, like more granular partitioning, compression, or enriching of the data. Monitoring and notification is an integral part of the automation process. So as soon as the ETL job finishes, another [CloudWatch](https://aws.amazon.com/cloudwatch/) rule sends you an email notification using an [Amazon Simple Notification Service (Amazon SNS)](https://aws.amazon.com/sns) topic. This notification indicates that your data was successfully processed.

In summary, this pipeline classifies and transforms your data, sending you an email notification upon completion.