AWS

Data Pipeline

# Index

Table of Contents

[Index 2](#_Toc28889594)

[AWS Data Pipeline 3](#_Toc28889595)

[Definition 3](#_Toc28889596)

[Definition Components 3](#_Toc28889597)

[Data Nodes 3](#_Toc28889598)

# AWS Data Pipeline

## Definition

* It is a web service that makes it easy to schedule regular data movement and data processing activities in the AWS cloud.
* Based on a schedule one defines, Data pipeline regularly performs processing activities such as distributed data copy, SQL transforms, MapReduce applications, or custom scripts against destinations such as Amazon S3, Amazon RDS, or Amazon DynamoDB.

## Definition Components

* Definitions determine the Data Pipeline tasks to run, schedules them, and then assigns them to task runners.
  + Task runners are essentially applications that check AWS Data Pipeline for new tasks and then does those tasks.
  + They are automatically installed on AWS resources provisioned by AWS Data Pipeline. Alternatively, they can be installed on compute resources such as EC2 instances or on-premise hardware.
  + If tasks fail, they are re-tried off to another task runner. This done by for a configurable number of attempts.
* These definitions are composed of:
  + **Data Nodes**: The location of input/output data.
  + **Activities**: Definitions of work to perform on a schedule.
  + **Preconditions**: Optional conditions that must be true before an activity can run.
  + **Scheduling Pipelines**: Determine the timings of events (like when the activity runs).
  + **Resources**: The computation resources (e.g. an EC2 or EMR cluster) that does the work the pipeline defines.
  + **Actions**: Things that are triggered when conditions are met (e.g. an SNS notification on the success or failure of a pipeline)

## Data Nodes

* Data Nodes define the location and type of data that Data Pipeline activities use as input or output.
* Supported data nodes include:
  + **DynamoDBDataNode**: A DynamoDB table that has data.
  + **SqlDataNode**: An SQL table, database and query that represents data for a pipeline activity.
  + **RedshiftDataNode**: An Amazon Redshift table that data should be copied to
  + **S3DataNode:** An S3 location that contains data files for a pipeline activity.