AWS Data Pipeline FAQs

## **General**

#### **Get Started with AWS for Free**

[Create a Free Account](https://portal.aws.amazon.com/gp/aws/developer/registration/index.html)  
[Or Sign In to the Console](https://console.aws.amazon.com/elasticache/home)

AWS Free Tier includes 750hrs of Micro Cache Node with Amazon ElastiCache.

[View AWS Free Tier Details »](https://aws.amazon.com/free/)

Q: What is AWS Data Pipeline?

AWS Data Pipeline is a web service that makes it easy to schedule regular data movement and data processing activities in the AWS cloud. AWS Data Pipeline integrates with on-premise and cloud-based storage systems to allow developers to use their data when they need it, where they want it, and in the required format. AWS Data Pipeline allows you to quickly define a dependent chain of data sources, destinations, and predefined or custom data processing activities called a pipeline. Based on a schedule you define, your 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. By executing the scheduling, retry, and failure logic for these workflows as a highly scalable and fully managed service, Data Pipeline ensures that your pipelines are robust and highly available.

Q: What can I do with AWS Data Pipeline?

Using AWS Data Pipeline, you can quickly and easily provision pipelines that remove the development and maintenance effort required to manage your daily data operations, letting you focus on generating insights from that data. Simply specify the data sources, schedule, and processing activities required for your data pipeline. AWS Data Pipeline handles running and monitoring your processing activities on a highly reliable, fault-tolerant infrastructure. Additionally, to further ease your development process, AWS Data Pipeline provides built-in activities for common actions such as copying data between Amazon Amazon S3 and Amazon RDS, or running a query against Amazon S3 log data.

Q: How is AWS Data Pipeline different from Amazon Simple Workflow Service?

While both services provide execution tracking, retry and exception-handling capabilities, and the ability to run arbitrary actions, AWS Data Pipeline is specifically designed to facilitate the specific steps that are common across a majority of data-driven workflows – in particular, executing activities after their input data meets specific readiness criteria, easily copying data between different data stores, and scheduling chained transforms. This highly specific focus means that its workflow definitions can be created rapidly and with no code or programming knowledge.

Q: What is a pipeline?

A pipeline is the AWS Data Pipeline resource that contains the definition of the dependent chain of data sources, destinations, and predefined or custom data processing activities required to execute your business logic.

Q: What is a data node?

A data node is a representation of your business data. For example, a data node can reference a specific Amazon S3 path. AWS Data Pipeline supports an expression language that makes it easy to reference data which is generated on a regular basis. For example, you could specify that your Amazon S3 data format is s3://example-bucket/my-logs/logdata-#{scheduledStartTime('YYYY-MM-dd-HH')}.tgz.

Q: What is an activity?

An activity is an action that AWS Data Pipeline initiates on your behalf as part of a pipeline. Example activities are EMR or Hive jobs, copies, SQL queries, or command-line scripts.

Q: What is a precondition?

A precondition is a readiness check that can be optionally associated with a data source or activity. If a data source has a precondition check, then that check must complete successfully before any activities consuming the data source are launched. If an activity has a precondition, then the precondition check must complete successfully before the activity is run. This can be useful if you are running an activity that is expensive to compute, and should not run until specific criteria are met.

Q: What is a schedule?

Schedules define when your pipeline activities run and the frequency with which the service expects your data to be available. All schedules must have a start date and a frequency; for example, every day starting Jan 1, 2013, at 3pm. Schedules can optionally have an end date, after which time the AWS Data Pipeline service does not execute any activities. When you associate a schedule with an activity, the activity runs on it. When you associate a schedule with a data source, you are telling the AWS Data Pipeline service that you expect the data to be updated on that schedule. For example, if you define an Amazon S3 data source with an hourly schedule, the service expects that the data source contains new files every hour.