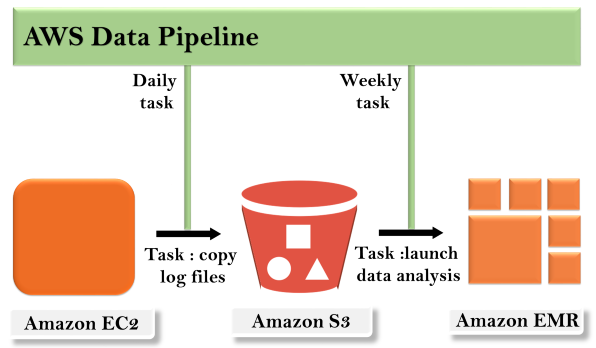
**MAY-WEEK 4, Task 3**

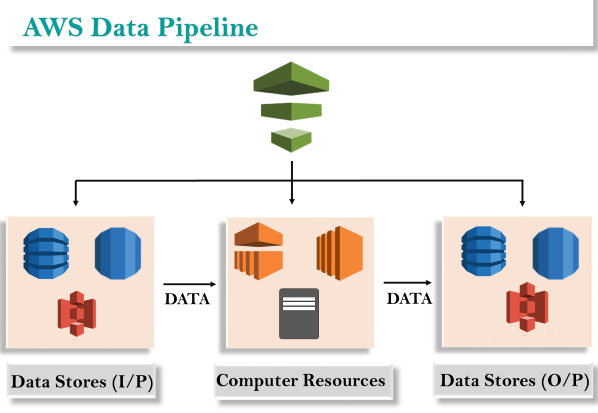
***AWS DATA PIPELINE FLOW***

AWS Data Pipeline is a web service that can access the data from different services and analyzes, processes the data at the same location, and then stores the data to different AWS services such as DynamoDB, Amazon S3, etc.

For example, using data pipeline, you can archive your web server logs to the Amazon S3 bucket on daily basis and then run the EMR cluster on these logs that generate the reports on the weekly basis.



The concept of the AWS Data Pipeline is very simple. We have a Data Pipeline sitting on the top. We have input stores which could be Amazon S3, Dynamo DB or Redshift. Data from these input stores are sent to the Data Pipeline. Data Pipeline analyzes, processes the data and then the results are sent to the output stores. These output stores could be an Amazon Redshift, Amazon S3 or Redshift.



Components of AWS Data Pipeline

**Following are the main components of the AWS Data Pipeline:**

* **Pipeline Definition**  
  It specifies how business logic should communicate with the Data Pipeline. It contains different information:
  + **Data Nodes**  
    It specifies the name, location, and format of the data sources such as Amazon S3, Dynamo DB, etc.
  + **Activities**  
    Activities are the actions that perform the SQL Queries on the databases, transforms the data from one data source to another data source.
  + **Schedules**  
    Scheduling is performed on the Activities.
  + **Preconditions**  
    Preconditions must be satisfied before scheduling the activities. For example, you want to move the data from Amazon S3, then precondition is to check whether the data is available in Amazon S3 or not. If the precondition is satisfied, then the activity will be performed.
  + **Resources**  
    You have compute resources such as Amazon EC2 or EMR cluster.
  + **Actions**  
    It updates the status about your pipeline such as by sending an email to you or trigger an alarm.
* **Pipeline**  
  **It consists of three important items:**
  + **Pipeline components**  
    We have already discussed about the pipeline components. It basically how you communicate your Data Pipeline to the AWS services.
  + **Instances**  
    When all the pipeline components are compiled in a pipeline, then it creates an actionable instance which contains the information of a specific task.
  + **Attempts**  
    We know that Data Pipeline allows you to retry the failed operations. These are nothing but Attempts.
* **Task Runner**  
  Task Runner is an application that polls the tasks from the Data Pipeline and performs the tasks.