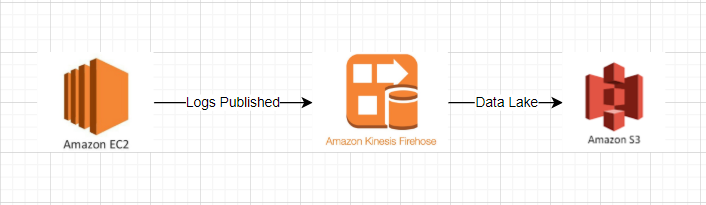
**AIM:** Display batch processing process through AWS Kinesis Data Firehose and create a data lake on S3 for an E-Commerce log data.

**Block Diagram:**



**Services Used:** AWS EC2, AWS S3, AWS Kinesis Data Firehose, AWS IAM.

**Source Data File:** CSV Format.

**Steps:**

1. Create Firehose delivery Stream.
   1. Source: Direct Put (Will be using Kinesis Client Application in order to publish data in the delivery stream) from Amazon Kinesis Agent.
   2. Destination: AWS S3.
      1. Create Bucket for the same.
   3. Batching Conditions to store data in S3:
      1. Buffer size: 5MB.
      2. Buffer Interval: 90 seconds.
   4. Create a required AWS IAM role for Firehose having permissions for S3.
2. Set up T2/Micro EC2 Instance to publish the logs into the Firehose Stream.
   1. Choose Amazon Linux AMI as the EC2 instance type.
   2. Generate Key pair to log into the EC2 instance.
   3. Connect to the instance through putty/Terminal.
3. Install Amazon Kinesis Agent in the instance launched.
   1. sudo yum install -y aws-kinesis-agent.
   2. Load the data file and the logs generator script file in the EC2 instance.
   3. Change permissions for the log generator file.
      1. chmod a+x LogGenerator.py. #LogGenerator.py: Log generator script.
   4. Create a log directory in which we will put the log files generated from LogGenerator.py.
      1. sudo mkdir /var/log/<Folder Name>. (Same is being monitored by Firehose and it will pick up the files to stream from the same folder).
4. Configure Kinesis Agent file for the delivery stream.
   1. sudo nano agent.json.
   2. Set up the variables accordingly.
5. Create a role for your EC2 instance to access firehose without creating a user and hence making sure the pipeline is secure.
6. Start the Amazon Kinesis Agent:
   1. sudo service aws-kinesis-agent start.
7. Generate logs using the script LogGenerator.py:
   1. sudo LogGenerator.py 50000 #50000 is the number of files to be published.
   2. Check the log for Kinesis agent:
      1. tail -f /var/log/aws-kinesis-agent/aws-kinesis-agent.log
8. Check the S3 bucket created in start and the log data will be partitioned by Year->Month ->Day->Hour.