CS 4740 Programming Assignment 6

Introduction: Log analytics is a common big data use case that allows you to analyze log data from websites, mobile devices, servers, sensors, and more for a wide variety of applications such as digital marketing, application monitoring, fraud detection, ad tech, gaming, and IoT.

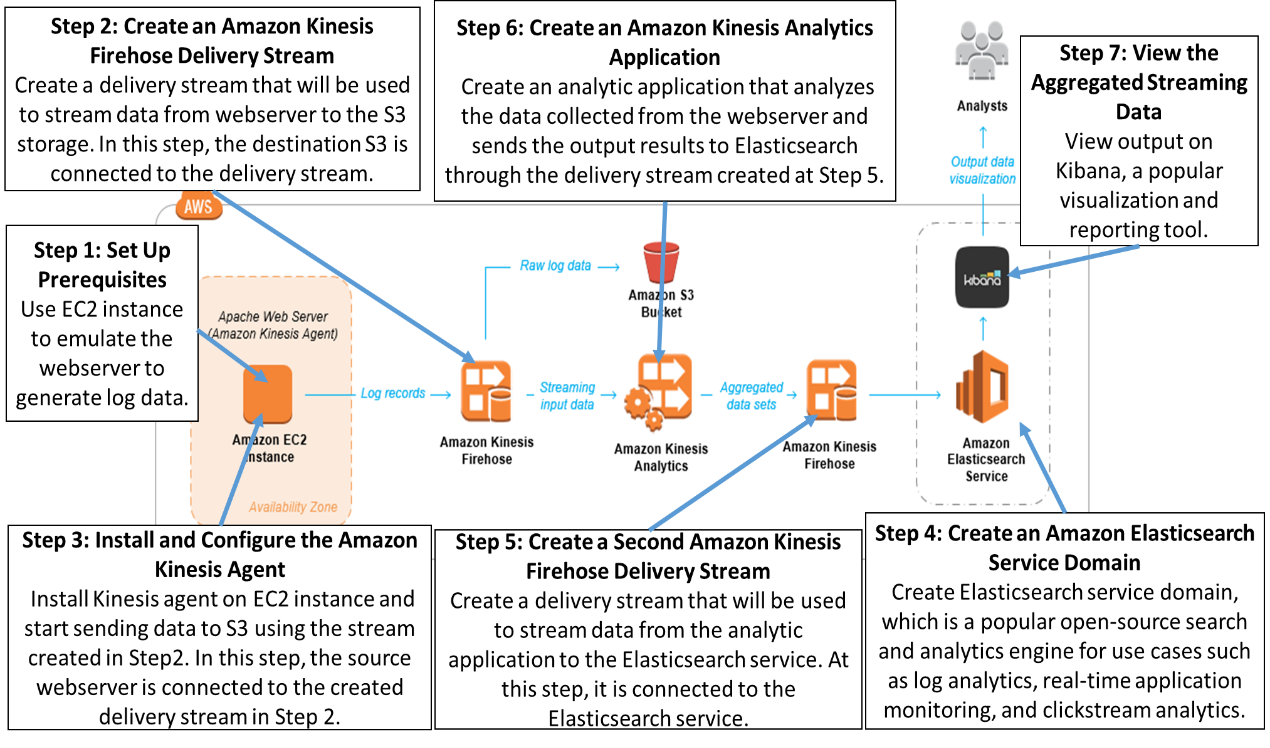
Goals: Understanding streaming data processing in real time. Hand-on experience to build a system to collects, ingests, and processes the streaming data.

**Please follow the tutorial below (steps 1-8):**

<https://aws.amazon.com/getting-started/projects/build-log-analytics-solution/>

Please refer to “WHAT TO SUBMIT” section when you are following the tutorials.

The figure below gives a brief introduction to each step. These steps are corresponding to the steps in the tutorial.



**Notes: (please be sure to read the notes before you do the corresponding steps**)

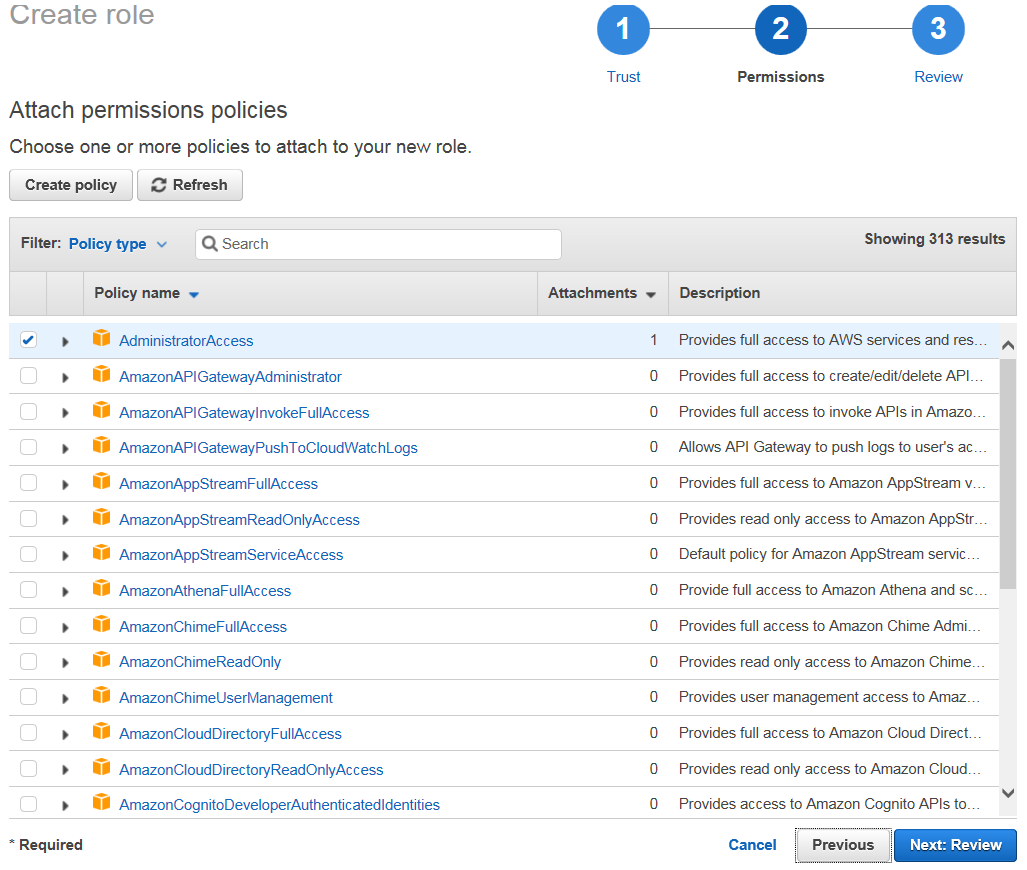
If the tutorial does not mention what to select in the process and we do not give any notes below, keep the default option!!!

**Step1**: (1) Create an IAM role when you create EC2 instance

Click “Create role” -> select “AWS service” -> select “EC2” -> select “EC2 Allows EC2 instances to call AWS services on your behalf.” -> Next: Permissions -> Select policy “AdministratorAccess” -> Next: Review -> Input the name -> Create role

The “AdministratorAccess” will allow EC2 instance to access all AWS resources, including Amazon Kinesis Firehose and Amazon CloudWatch.

So IGNORE the tutorial “Access Management (IAM) role configured with permission to write to Amazon Kinesis Firehose and Amazon CloudWatch”.



(2) Prepare Your Log Files

Download “Fake Apache Log Generator”

Install git: sudo yum install git

Download: git clone <https://github.com/kiritbasu/Fake-Apache-Log-Generator.git>

Install requirements: cd Fake-Apache-Log-Generator

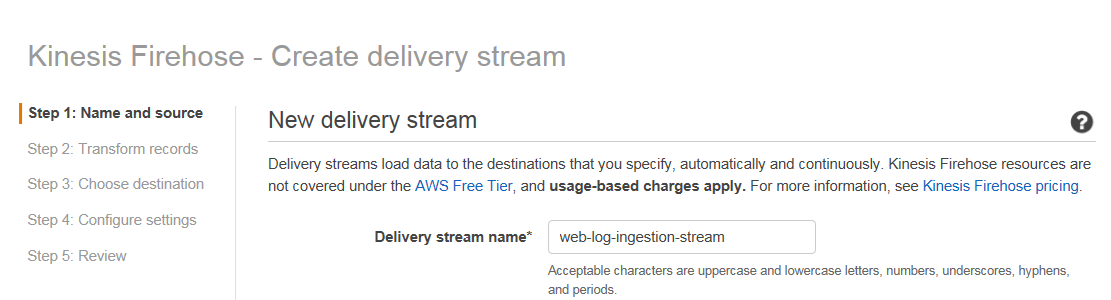
sudo pip install -r requirements.txt

Use “generate infinite log file” to run the program:

cd /var/log (it is important to store your generated log files in /var/log/)

sudo python /home/ec2-user/Fake-Apache-Log-Generator/apache-fake-log-gen.py -n 0 -o LOG (it will run forever – that is why it is called infinite log generator)

**Step2**: After you click “3. Click Create Delivery Stream.”, enter the stream name as shown in the fig below, and skip “4 b. For Delivery stream name, enter web-log-ingestion-stream.”



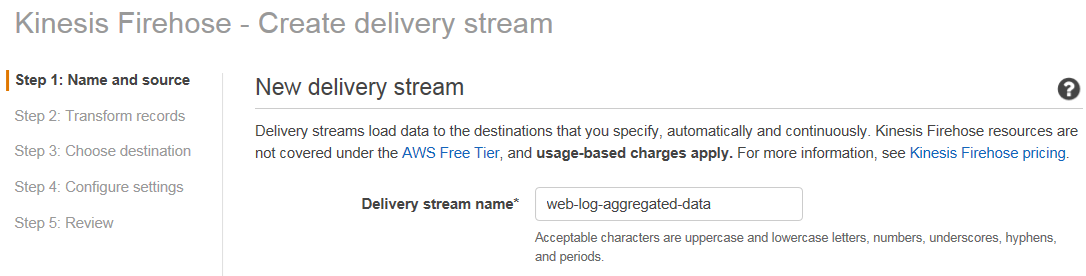
**Step3**: modify the configuration file located at /etc/aws-kinesis/agent.json

filePattern: should be specified to “/var/log/\*access\*log”

**Step4**: For Set the domain access policy to, choose **Allow open access to the domain**.

Before you take the above step: select “Public access” first, then you will see this option.

**Step5**: After you click “3. Click Create Delivery Stream.”, enter the stream name as shown in the fig below, and skip “4 b. For Delivery stream name, enter web-log-aggregated-stream.”



When you meet "schema discovery failed" problem in creating application:

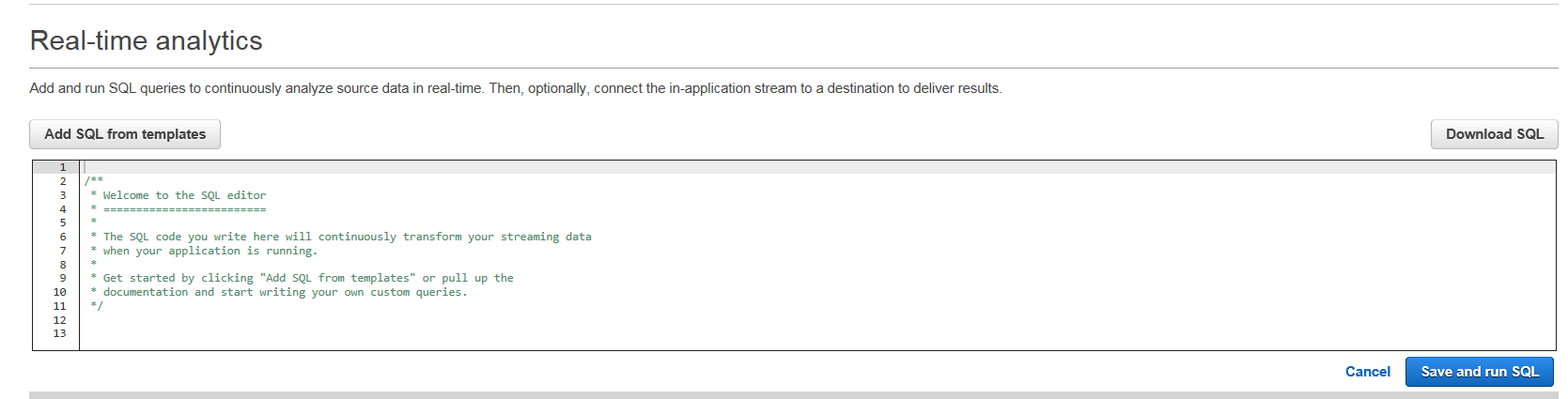
1. Make sure IAM role is correctly attached to EC2 instance.

2. Make sure your json file 100% correct. Do not change any single word except the two red parts.

3. Make sure infinite log generator is generating log in /var/log.

In Step 6-step13: In the SQL editor, enter the following SQL code

You need to copy the code to the window below, and make sure the format of the code is exactly the same as that in the tutorial before you click “ **Save and run SQL.**”



**Before step 16, you need to click “Connect to destination”. After step 16, you need to choose:**



**WHAT TO SUBMIT:**

1. Screenshot of your created IAM role for your EC2 VM instance

2. Screenshot of Firehose delivery streams you create on the aws console

<https://console.aws.amazon.com/firehose/home?region=us-east-1#/dashboard/list>

3. Screenshot of the Amazon Elasticsearch Service Domain you create

4. Screenshot of the Amazon Kinesis Analytics Application you create

5. Screenshot of the Kibana website

Please submit all the screenshots in one PDF file.