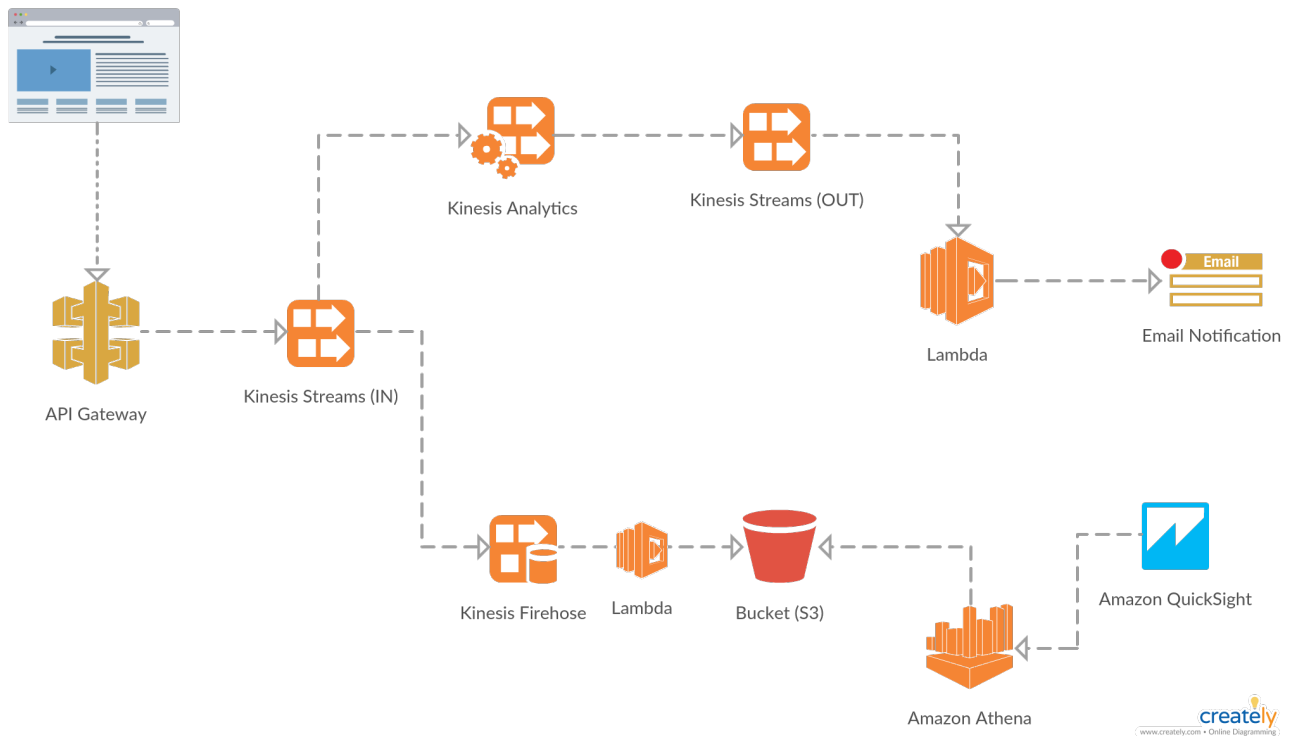# LAB3 - Data Processing in AWS

In this lab you will build solution for Real-time Clickstream Analysis with Amazon Kinesis Analytics and few other services. Look on the diagram below

Step 1 - Configure environment over CloudFormation.

You will use AWS CloudFormation service to spin up all environment. AWS CloudFormation provides a common language for you to describe and provision all the infrastructure resources in your cloud environment.

- Go to the LAB3 folder you downloaded from repository.
- Log in to the AWS GUI,
- Switch Region to the **US East (N. Virginia)**
- Go to Services -> CloudFormation and Create new Stack
- In Choose a template section select 2nd option: Upload a template to Amazon S3
- Choose a file and select CSE-sfn.json file from your Lab3 folder, and hit Next.
- Enter Stack name and provide email address for notifications.

Specify Details

Specify a stack name and parameter values. You can use or change the default parameter values, which are defined in the AWS Clou

Stack name `LAB3doroszl`

Parameters

email `I⎕⎕⎕⎕⎕⎕⎕⎕⎕⎕⎕` Email address to send anomaly detection events.

- Press Next, Next,
- Select checkbox
- Check "I acknowledge that AWS CloudFormation might create IAM resources"
- Press Create

When a status of the stack will change to CREATE_COMPLETE, switch to Output tab, copy the Value of ExecutionCommand and save in the notepad.

Overview	Outputs	Resources	Events	Template	Parameters	Tags	Stack Policy

Key	Value
ExecutionCommand	python ClickImpressionGenerator.py http⎕⎕⎕⎕⎕⎕⎕⎕ 70b3.execute-api.us-east-1.amazonaws.com/example/beacon/stream

- Check your email inbox for message "AWS Notification - Subscription Confirmation" and confirm subscription.

Step 2 - Configure Kinesis Analytic Application.

This time you will also use CloudFormation template, so try to do this by yourself.
- Use the template file: CSE-KA-cfn.json from your LAB3 folder.
- Enter the name for the stack and in SuportingStack parameter paste the name of the stack from the Step 1.
- Complete the rest of the cloud formation configuration.
- When the stack is ready, go to Services -> Kinesis -> Data Analytics
- Select your app and press Actions and Run Application.

Step 3 - Generate click's stream.

- Power on your EC2 instance.
- Don't forget to switch to the proper Region.
- Go to Services -> EC2 -> Instances.
- Select your instance
- Hit Actions button and select Instance State -> Start
- Connect over ssh.
- From Lab3 folder copy ClickImpressionGenerator.py python script to your EC2 instance.
- Take a command you saved in Step 1.
- Ensure that you are in folder where you placed the script.
- Execute the command.
- Periodically check your email for Anomaly Alert.

RBI_Lab3

Anomaly Detected

To:

Anomaly detected with a click through rate of 48.837209302325576% and an anomaly score of 2.9084094141705017

Step 5 - Stream click data to S3 bucket.

In this step you will use Kinesis Firehose to capture a click data from Kinesis Stream and write them to S3.

Data need some modification, so at first you will create a Lambda function, which will be used for this.

- Go to Services -> Lambda -> Create function
  - Author From Scratch
  - Enter the Function name
  - Runtime: Python 2.7
  - Role: Create a custom role
    - Create lambda_basic_execution role (it will be enough)

▼ Hide Details

Role Summary ❓

Role Description	Lambda execution role permissions
IAM Role	lambda_basic_execution ⬍
Policy Name	Create a new Role Policy ⬍

▶ View Policy Document

- Press Create function
- In Function Code paste the code from 'transformLambda.py' from Lab3 folder.
- Scroll down to Basic Setting and set Timeout to 5 minutes.

- Press Save (Top right corner).

Next step is to create Kinesis Firehose.

- Go to Services -> Kinesis -> Create delivery stream
- Enter Delivery stream name
- As Source chose Kinesis stream.
- Select stream: YourStackName-CSEKinesisBeaconInputStream-XXXXX and press Next
- Record transformation change to Enable
- Select a Lambda function you just created and press Next
- Set Destination as Amazon S3
- Select S3 bucket from LAB1 and enter Prefix.
- Set Buffer size to 1MB and Buffer interval to 60 seconds.
- Create a new IAM role.
- Leave autogenerated IAM Role and press Allow
- Press Next and Create delivery stream

Step 6 - Generate a new click's stream

It's time to test delivery stream from Step 5.
- Jump back to your EC2 instance.
- Once again execute command copied from Step 1.
- Keep eye on your bucket, new folder should appear.
- Explore the folder, you will find a new file.
- Select a file and press Download button.
- Try to open the file with notepad.

Alternatively you can read this file on your EC2 instance
- Copy the file to your local disk using aws cli (check commands in LAB1)
- Look into the file with cat or vi command.

Step 7 - Use Amazon Athena to query the data

At this moment you should collection of data in your S3 bucket.
In this step we will query the data with Amazon Athena.

- Go to Services -> Athena -> Query Editor
- Open a createathenatable.sql from Lab3 folder.
- Edit LOCATION parameter:
  - s3://your_bucket_name/folder_with_logs/

Use a bucket from step 5. (Select S3 bucket from LAB1 and enter Prefix). The Prefix parameter is a name of the folder with logs.

- Copy full query command and paste it into Athena Query Editor and press Run Query.
- Look for the newly created table called "clickstream" (in the left panel, eventually click on refresh button)
- On right side of the table name click on this 3 dots and chose Preview Table.
- In the result tab you will get records from clickstream table.

Results

	browseraction	site
1	Impression	-api.us-east-1.amazonaws.com
2	Impression	-api.us-east-1.amazonaws.com
3	Click	-api.us-east-1.amazonaws.com
4	Impression	-api.us-east-1.amazonaws.com
5	Impression	-api.us-east-1.amazonaws.com
6	Impression	-api.us-east-1.amazonaws.com
7	Impression	-api.us-east-1.amazonaws.com
8	Impression	-api.us-east-1.amazonaws.com
9	Click	-api.us-east-1.amazonaws.com
10	Impression	-api.us-east-1.amazonaws.com

Step 8 - Use BI tool QuickSite for data visualisation.

QuickSite is a Business Intelligence to for data visualisation. In this step you will visualise a data from clickstream table you created in step 7.

- Go to Services -> QuickSite
- Press Manage data button and next New data set
- Select Athena and provide the name of your table created in step 7 and press Create data source
- Click on newly created data set and hit Create analysis
- Select your preferred Visual type (pie chart will be good for this data)
- Select "browseraction"
- Data should be visualise on the graph