Developing a streaming data ecosystem using AWS Kinesis

- Contents:

  Problem of streaming data

  Kinesis capabilities

  Example project
  - Pieces required
  - Setup
  - Examples

- Nabeel Tariq

What makes high velocity data unique

- Analysis is done ad hoc, you don’t get to wait for all of the data to perform analysis
- Data often needs transformation before analysis, but both have to be done on the fly
- Data-driven-decisions also made in real time – can’t take your time and evaluate multiple analyses

What can you do with AWS

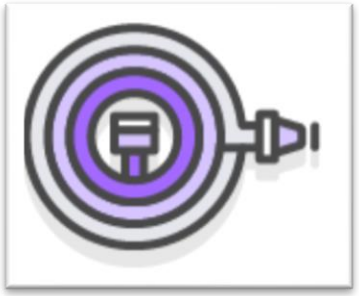
EC2

Distributed search and analytics engine

Kinesis

Streams

Build custom applications to process streaming data



Firehose

Load streaming data into AWS

DynamoDB

A fully managed cloud NoSQL database

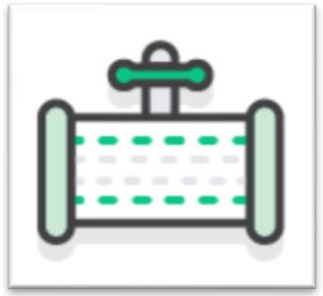
S3

Object storage built to store and retrieve data

An example of processing streaming data from Twitter

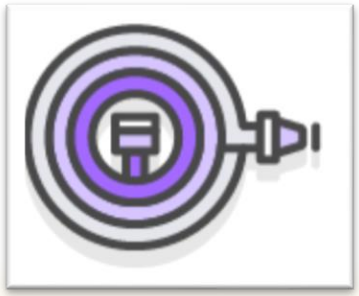
EC2

Twitter API



Kinesis Streams

DynamoDB

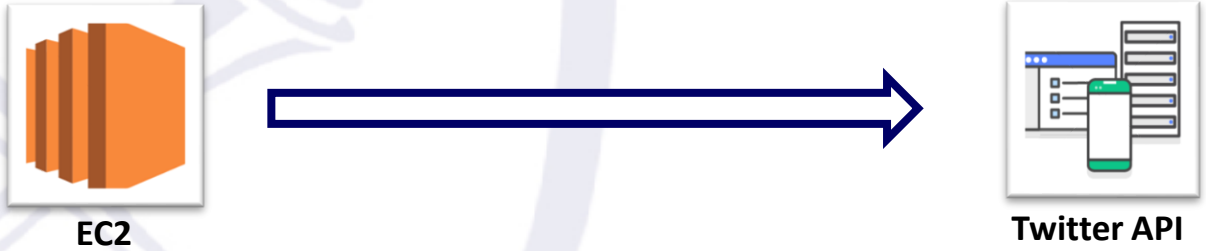


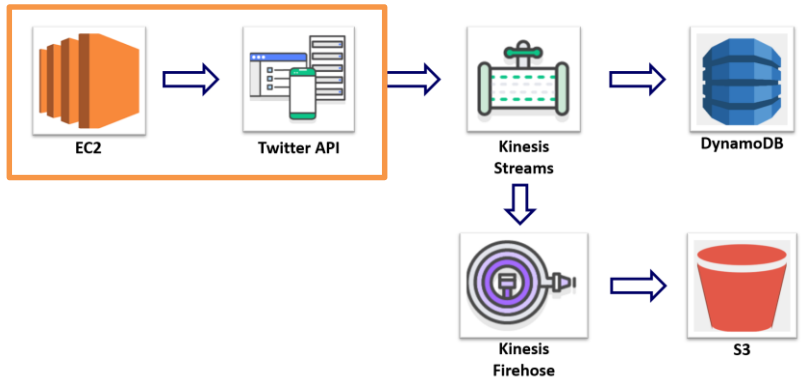
Kinesis Firehose

S3

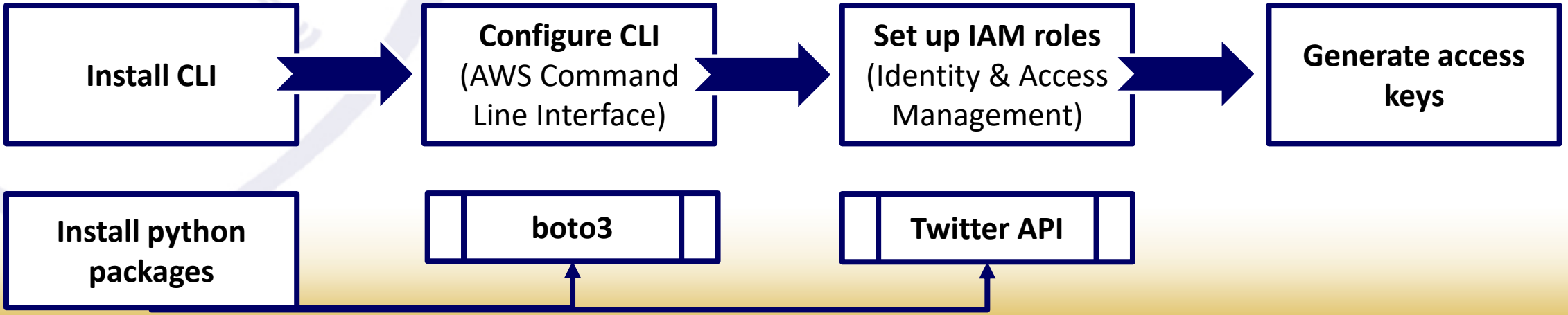
Setting up properly is essential

Deployed a t2-Micro AWS EC2 instance to host the code and interface with other AWS tools



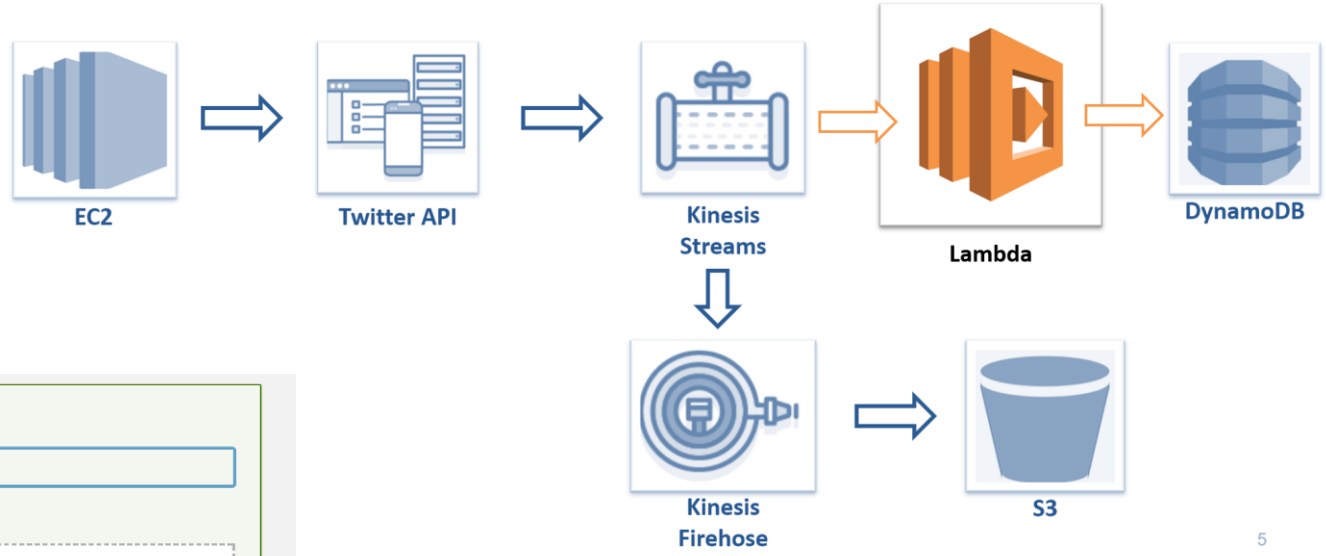


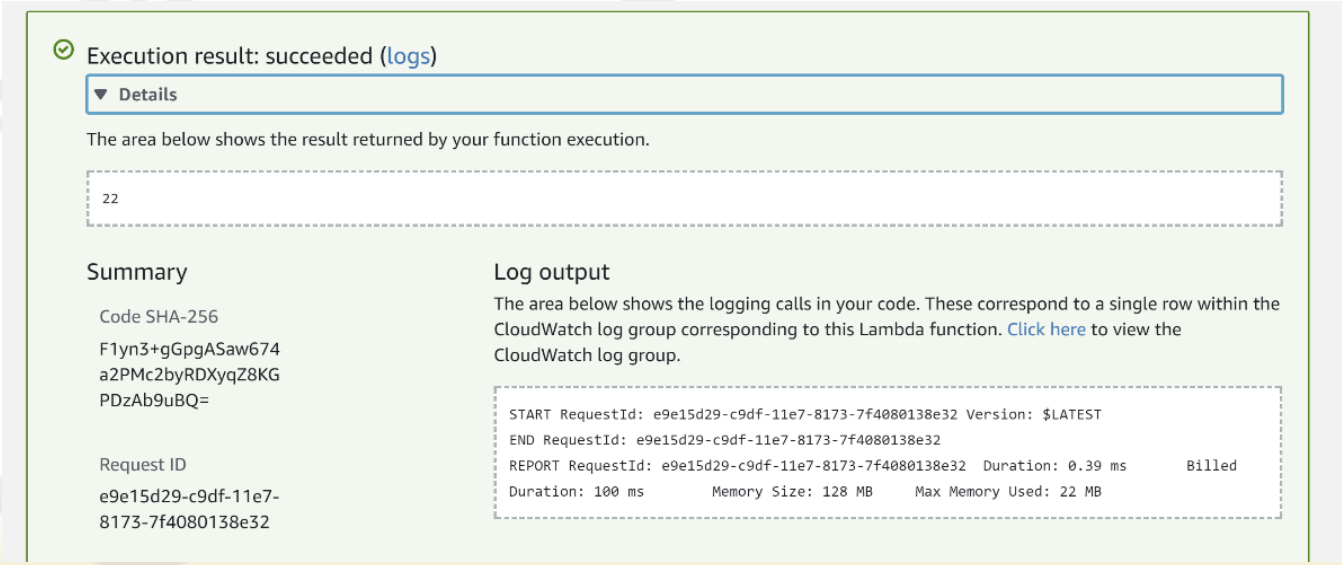
Get software and libraries prepared



Things don’t always go according to plan

Our original plan is to build a Lambda function and attach it to the Kinesis Stream for processing the JSON data into a more usable format.





✓ Succeed in running in the test

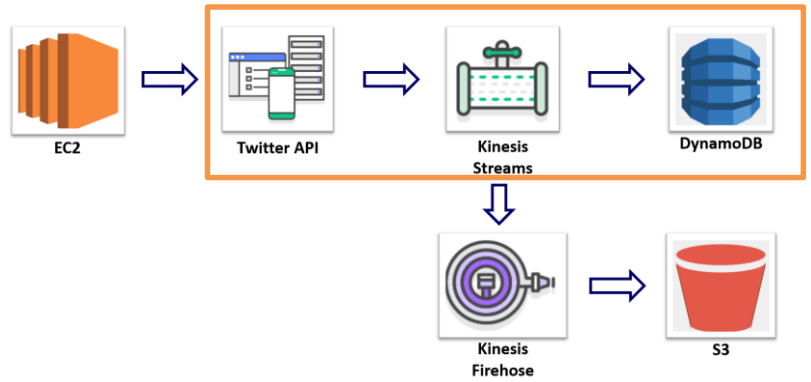
✗ Get error on live data

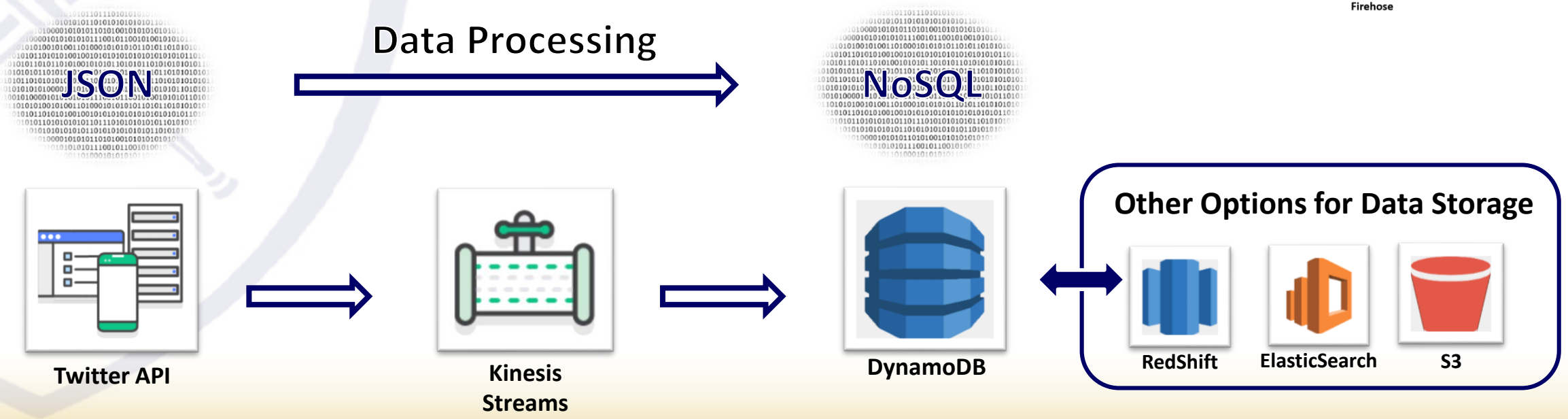
Go for DynamoDB

NoSQL structure is more flexible for data processing

Eventually this is what the process looked like

Connect the Twitter API to AWS using Kinesis Streams, process the streams to extract our attributes, and have the stream feed into DynamoDB

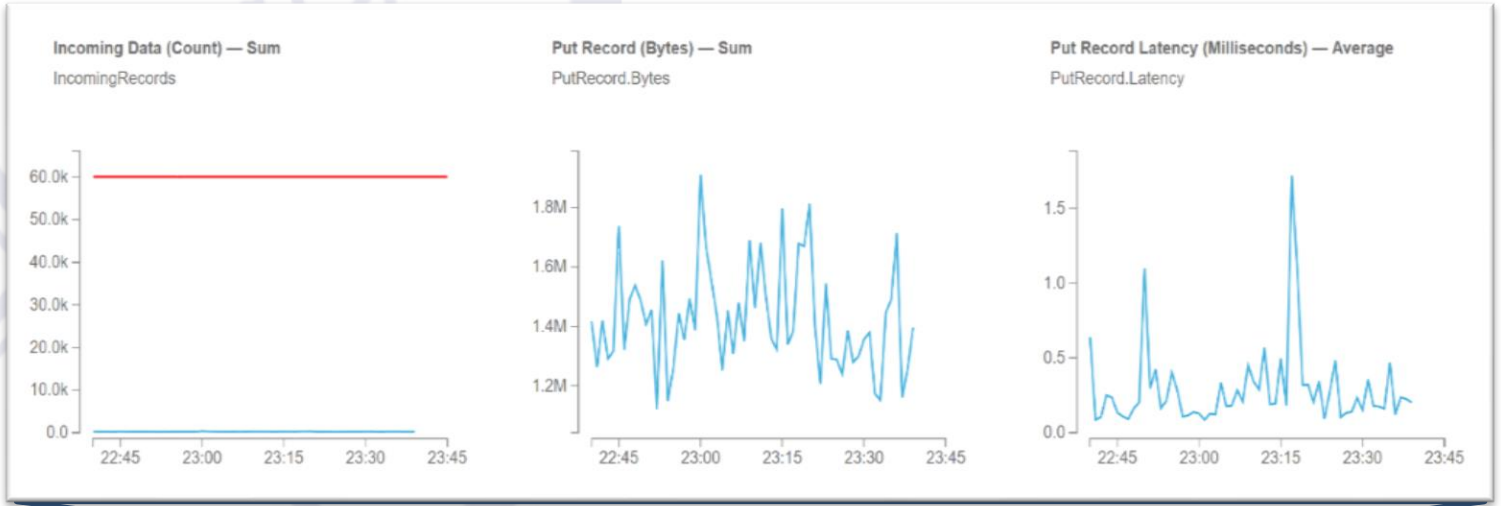


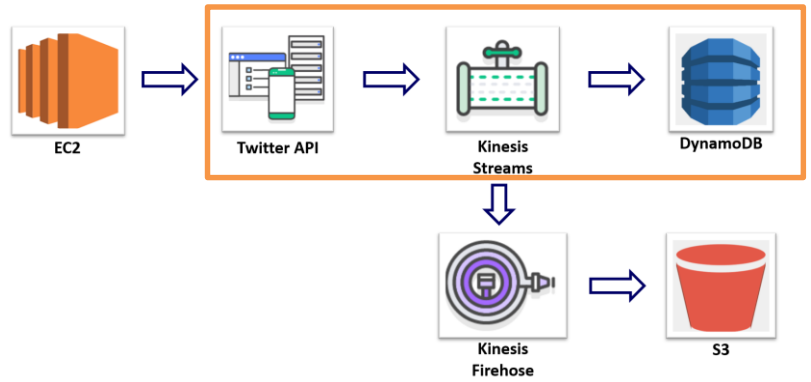


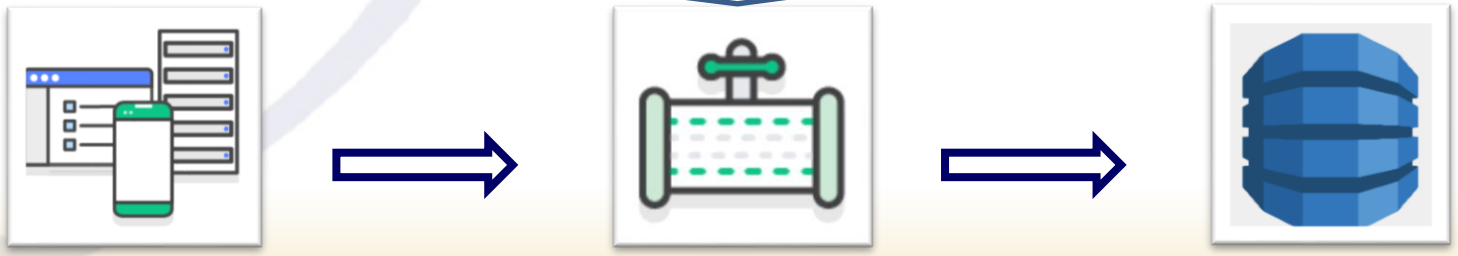
Questions?

Step 2

AWS Cloud Watch Metric





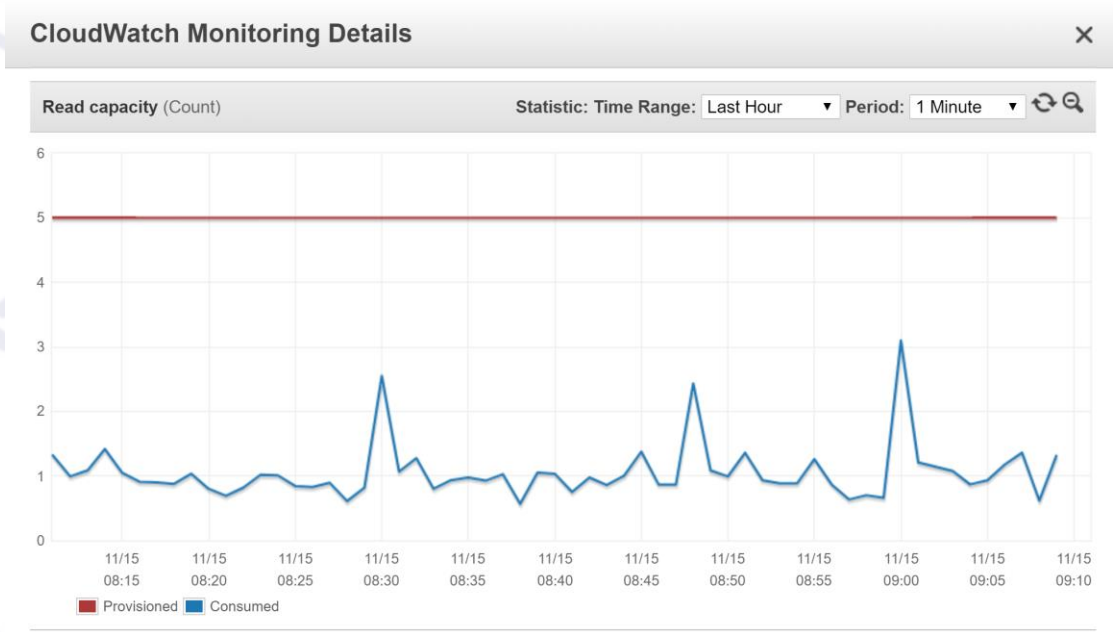


Twitter API

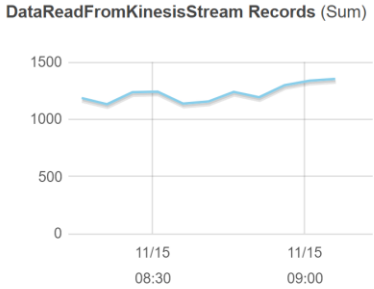
Kinesis Streams

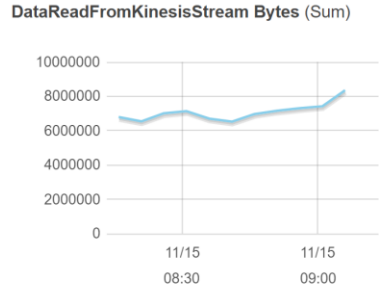
DynamoDB

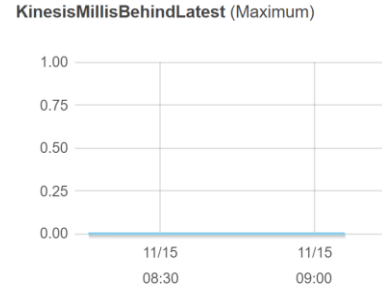
Monitoring



Firehose Console Metrics for Kinesis Stream read







Firehose Console Metrics for S3 write