Project - AWS Data lake

Damian Lewandowski

16.05.2023

Title: Log-analytics solution on AWS

Problem statement:

Find solution for the industry XYZ of a large, web server cluster where it’s important constantly monitor the server access logs for anything usual, send this logs to a log server and parse the data. Cluster has grown exponentially larger and now it requires a more robust solution to keep up with the demand.

Let’s look on architectural diagram and step by step we will be explore how this components works together

Amazon EC2 service it allows to provision virtual servers called EC2 instances.

Amazon kinesis can ingest application logs.

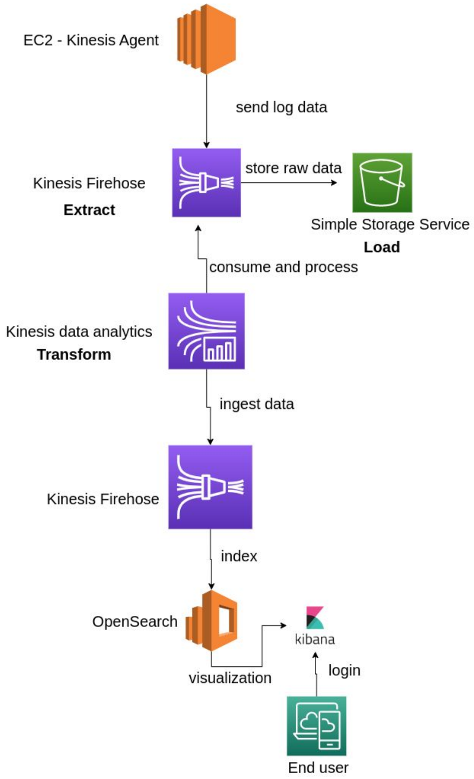
Amazon simple storage service - store data in bucket. It is common use in purpose to build data lake.

Amazon kinesis Firehose - load streaming data into data store, and analytics services. It can capture, modify and deliver data to S3 bucket, OpenSearch.

Amazon Analytics service it is used to transform and analyze streaming data in real time.

Amazon OpenSearch service it is use to perform interactive log analytics

And the last Kibana plugins to visualize charts



How it will be working? ok so…

Kinesis agent produce data, which was run on an EC2 instances
Aim of agent it is simulates one of the web servers which pretend like large server farms. Using scripts.
Kinesis Data Firehose ingest dummy access logs and move it to OpenSearch service and then
Using OpenSearch dashboards can help find insight on data by visualize it.

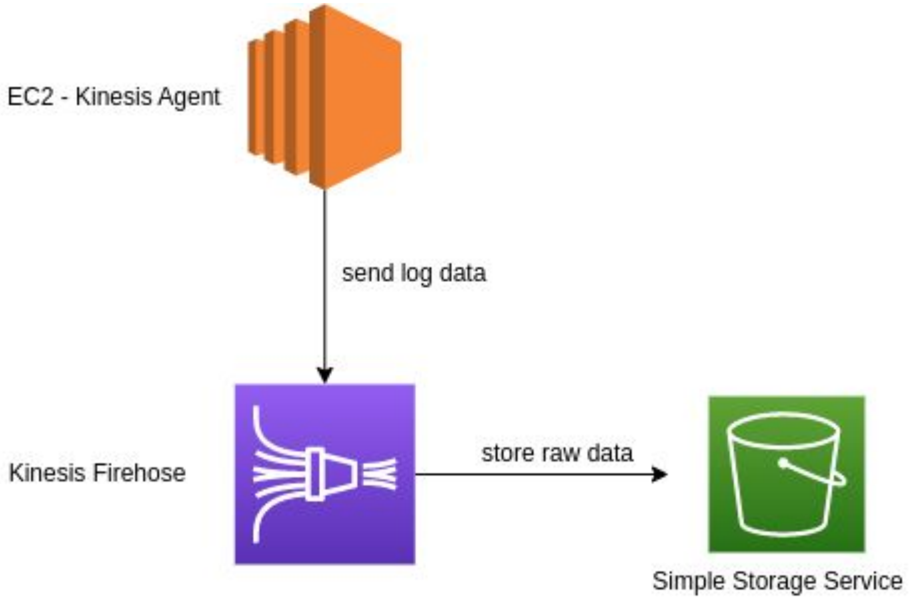
EC2 - Kinesis Firehose - Amazon S3

Ok, on this slide we see EC2 - Kinesis Firehose and S3 bucket to store data.
Ok, so…

In this case, we have data being produced by the Amazon Kinesis agents running on EC2 instances configured to send the contents in the log files straight out to Kinesis Firehose service
That Kinesis Firehose has Amazon S3 as destination for storing the raw data.

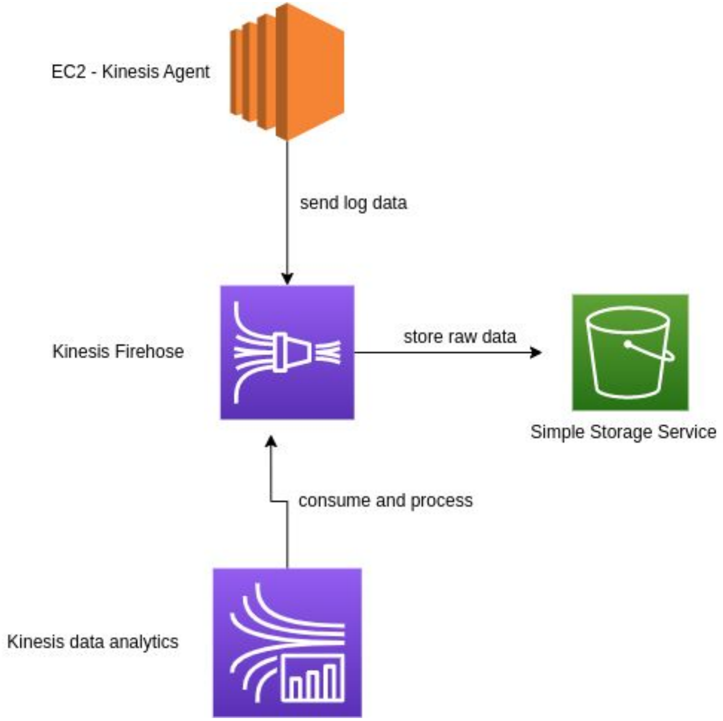
One of the advantage of service amazon Kinesis is data agnostic it we can produce any format of data and streams based on this service we can getting data and place it somewhere or perform real time analytics

So let’s deep dive into Kinesis Analytics Service:



Kinesis Firehose - Kinesis Analytics

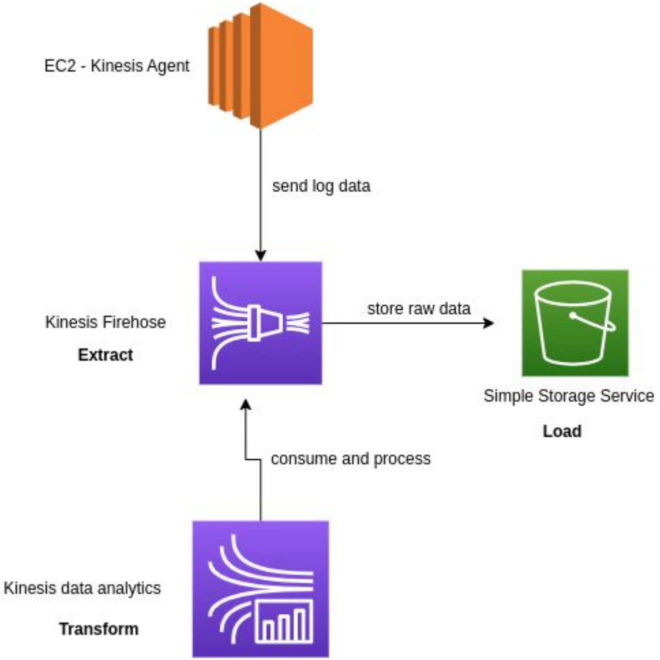
Kinesis Firehose has a Kinesis Analytics application attach to it what we see on diagram. So we utilize this service to perform SQL queries and aggregation on data points.
Benefits of this it is to help OpenSearch service to find things faster and reduces the space occupied in cluster by OpenSearch.



ETL - extract, transform, and load

When to process data? It is common to doing ELT that means Extract, Load, and Transform

In our case we will be EXTRACT, TRANSFORM and LOAD data because we are extracting the data load in S3 bucket mostly as a backup, then we are transforming with Kinesis Analytics service for further usage.

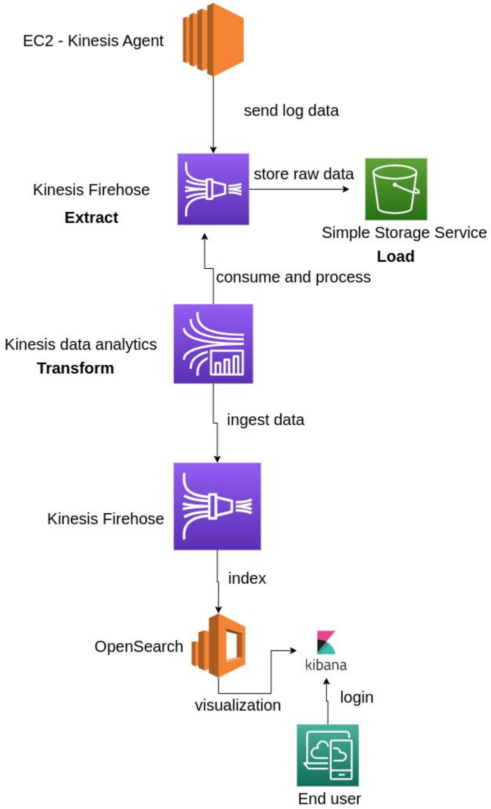


Kinesis Analytics - OpenSearch - Kibana - end user

Let’s look on last part diagram
Kinesis Analytics - OpenSearch - Kibana - end user, so…
From Kinesis Analytics to OpenSearch the easiest way to hand data from Kinesis Analytics to OpenSearch is through Kinesis Firehose service.

Kinesis Firehose was connected as the output of the Kinesis Analytics

When data is in Amazon Opensearch, users can log with Kibana to visualize data in the form of charts and extract insight from the data produced by the server logs by creating dashboards.



Summarize:
Kinesis Agent produce data and runs on an EC2 instance
Access logs will be send to a Kinesis Firehose which will be used to extract the data and save that data into Amazon S3 bucket. Then connect a Kinesis Analytics to do some data transformation on top of that Kinesis Firehose. so…

Kinesis Analytics will be using a Kinesis Firehose as an input and another Kinesis Firehose as an output which will send data to Amazon OpenSearch Service which will have the Kibana dashboard and visualize charts to end user.

Thanks you so much for your interest and attention

- Further plans

Damian Lewandowski