Task for Data Engineer

News Site task

The team needs you to collect data from various news sites. There are a set of keywords grouped within categories, e.g., keyword=[epidemic, coronavirus] are under category=environmental. Those keywords are being used to filter out irrelevant news and only store news that contains those keywords.
The attributes that need to be collected are:

- url
- title
- news content (text)
- event timestamp
- keyword matched

1. How do you design and implement data pipeline from various news website into our data storage? You need to:

   - Give us a high-level diagram of your solution
   - Give us a working code that could run in AWS cloud platform
   - Using serverless approach
   - Leveraging Kafka

2. How do you monitor and maintain the pipeline? Give us a high-level diagram of your solution along with your explanation or any code configuration that you can show to us

Approach1 :

1.Kafka producers crawl pages , clean data and put it in the Kafka Cluster. We can compress it or leave as it is.
2.Assuming Kafka producers can run in groups on multiple computers, we can use a redis to store crawled pages so that we can avoid duplicates.
3. Kafka Connect will dump the desired data to S3.
4. We can run Amazon MSK as a kafka cluster or we can self manage it as well.

5. If we use S3 to store Data we can use Athena to query data as well.
6. Based on the requirement we can write it to any DB.

Monitoring

Option 1 Confluent Kafka
We can monitor Confluent Kafka using the Confluent Control Center. It provides good visualization and tools to monitor end to end pipeline

Option2 Apache Kafka

Monitoring Kafka with Prometheus and Grafana

Approach2:

We can use EMR -Spark to crawl all web pages and directly dump data in S3. In this way continuous monitoring of the Kafka cluster can be avoided and we can crawl whole data in a few hours. Once it is stored in S3 we can take data to our desired destination based on our requirement.

Approach3:

Using Amazon kinesis.

Approach4:

Simple Python to DB , only in case the volume is not so much.But it is a cost effective and simple approach./Multiprocessing