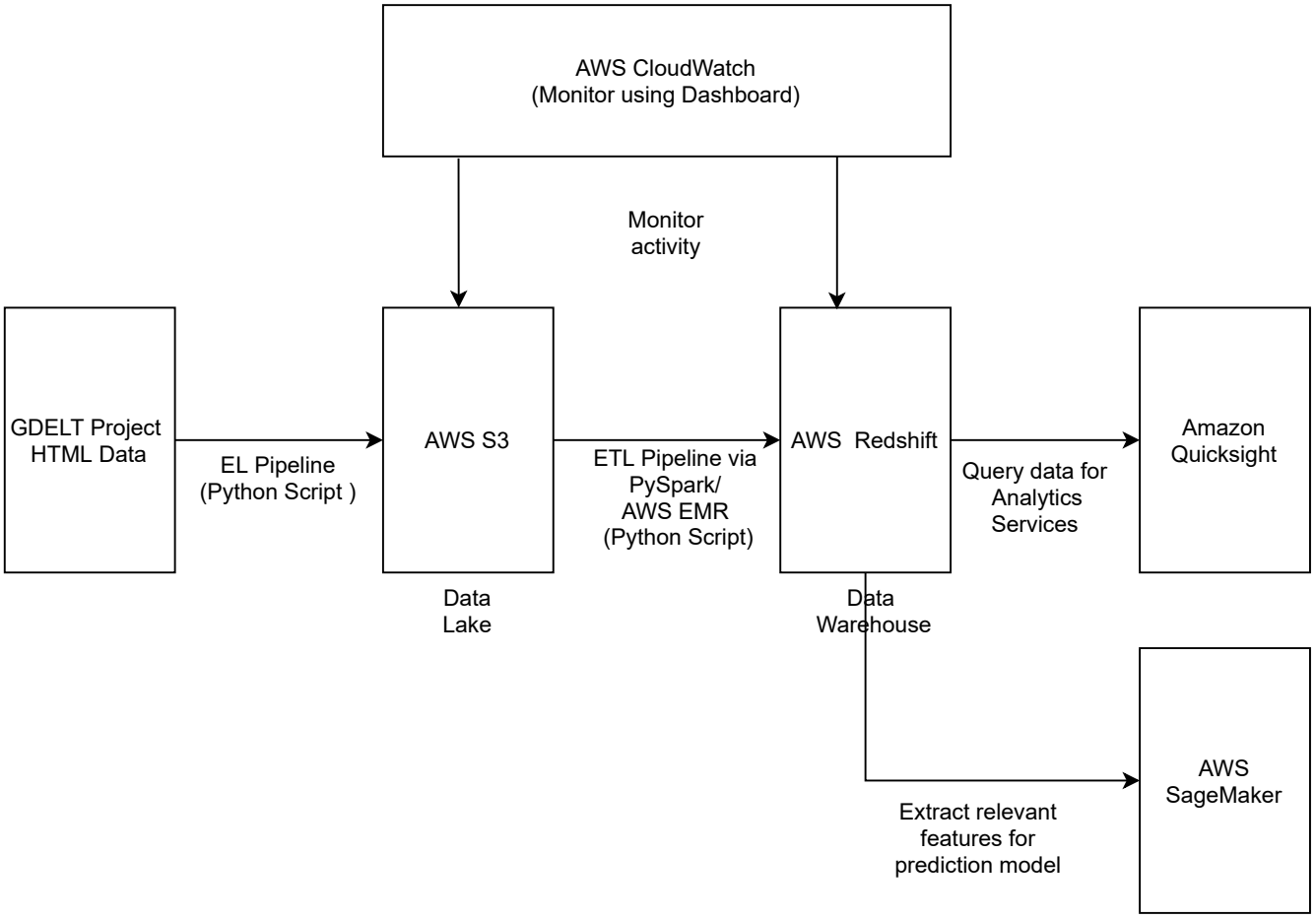

Link:

https://aws.amazon.com/big-data/datalakes-and-analytics/

- This link is useful as a reference to which AWS services to implement for my architecture diagram.

Steps:

1. Locate source data (GDELT Project HTML document).

2. Write Python script to extract and load data into AWS S3.

3. Inside AWS S3, we will use Spark (PySpark or Amazon EMR) to read the files via cloud path and run an ETL pipeline to store data into AWS Redshift as a Data Warehouse solution.

4. Via AWS Redshift or AWS Athena, we can try to query the data and explore it into an AWS Analytics solution such as Amazon Quicksight to generate a dashboard and visualize the data or Amazon SageMaker to build a predictive model (see link: https://aws.amazon.com/sagemaker/). I will only include these services in the architecture diagram to visualize the possible use cases of dealing with the data.

5. I need to monitor my activity through a service like Amazon CloudWatch and then build a dashboard. --> I won't include it inside the data pipeline architecture diagram, but will still use it for later step. -->
Link: https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch_Dashboards.html

Next:

- Go over architecture diagram attempt with mentor.

- Compare Redshift vs. Athena based on factors such as portability and cost, performance and scale. --> See Link: https://blog.panoply.io/an-amazonian-battle-comparing-athena-and-redshift

- Also, I must think about whether PySpark or AWS EMR (Apache Spark cluster) should be used.