Realtime Log Analysis Project Overview

What is Log Analysis?

The process of evaluating, understanding, and comprehending computer-generated documents known as logs is known as log analysis. A wide range of programmable technologies, including networking devices, operating systems, apps, and more, produce logs. A log is a collection of messages in chronological order that describe what is going on in a system. Log files can be broadcast to a log collector over an active network or saved in files for later analysis. Regardless, log analysis is the subtle technique of evaluating and interpreting these messages in order to get insights into any system's underlying functioning. Web server log analysis can offer important insights on everything from security to customer service to SEO. The information collected in web server logs can help you with:

- Network troubleshooting efforts
- Development and quality assurance
- Identifying and understanding security issues
- Customer service
- Maintaining compliance with both government and corporate policies

The common logfile format is as follows:

remotehost rfc931 authuser [date] "request" status bytes

Data Pipeline:

It refers to a system for moving data from one system to another. The data may or may not be transformed, and it may be processed in real-time (or streaming) instead of batches. Right from extracting or capturing data using various tools, storing raw data, cleaning, validating data, transforming data into query worthy format, visualization of KPIs including Orchestration of the above process is data pipeline.

What is the agenda of the project?

The agenda of the project involves Real-time log analysis with the visualization web app. We first launch an EC2 instance on AWS and install Docker in it with tools like Apache Spark, Apache NiFi, Apache Kafka, Jupyter Lab, Plotly and Dash. Then, we perform preprocessing on sample data, parse it into individual columns, cleaning the data and formatting timestamp. It is followed by Extraction of NASA access log dataset using Apache NiFi and Apache Kafka, followed by Transformation and Load using Cassandra and HDFS and finally Visualizing it using Python Plotly and Dash with the usage of graph and table app call-back.

Usage of Dataset:

Here we are going to use NASA access log data in the following ways:

- Extraction: During the extraction process, the downloaded dataset from Kaggle is ingested using NiFi processors and connections. The data is streamed from the data file using NiFi followed by the creation of topics and publishing logs using Apache Kafka.

- Transformation and Load: During the transformation and load process, we read data from Apache Kafka as streaming Dataframe according to schema creation with extraction and cleansing of log data and loading to Cassandra for Speed layer and HDFS for Batch layer. Then data is visualized using Plotly in Dash.

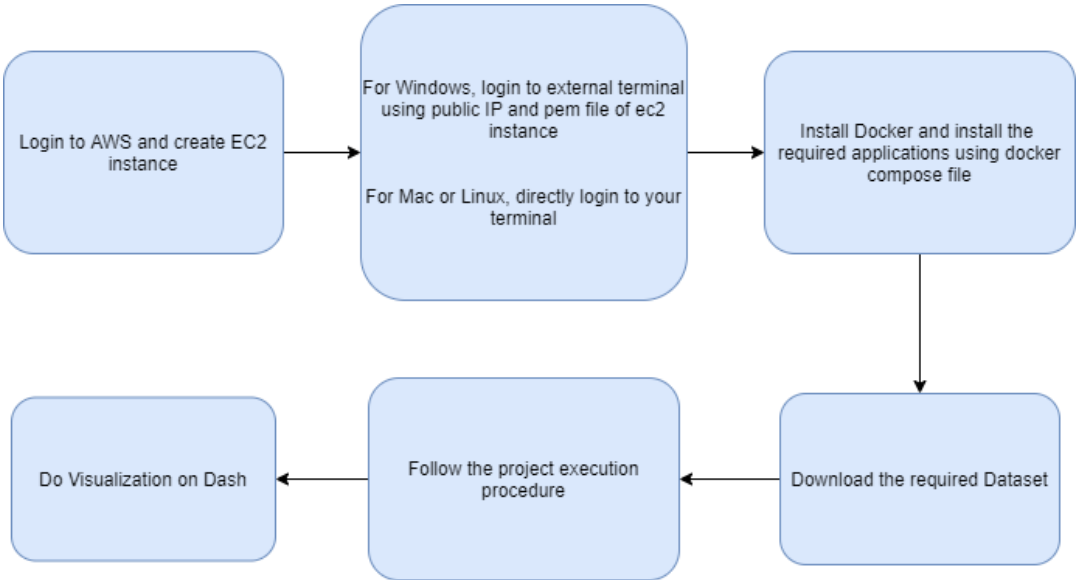
Key Takeaways

- Understanding the project and how to use AWS EC2 Instance
- Understanding the basics of Containers, log analysis, and their application
- Visualizing the complete Architecture of the system
- Understanding Port Forwarding
- Introduction to Docker
- Usage of docker-composer and starting all tools
- Exploring dataset and common log format
- Understanding Lambda Architecture.
- Installing NiFi and using it for data ingestion
- Installing Kafka and using it for creating topics
- Publishing logs using NiFi
- Integration of NiFi and Kafka
- Installing Spark and using it for data processing and cleaning
- Integration of Kafka and Spark
- Reading data from Kafka via Spark structured streaming API
- Installing and creating namespace and table in Cassandra
- Integration of Spark and Cassandra
- Continuously loading data in Cassandra for aggregated results.
- Integrating Cassandra and Plotly and Dash
- Displaying live stream, Hourly and Daily results using Python Plotly and Dash

Data Analysis:

- From the given website, data is downloaded containing the NASA access log data in csv format, containing different components of a web server log
- Dataset is processed, cleaning and formatting the datetime field.
- The extraction process is done using NiFi and Kafka, by streaming data from log file using NiFi and creating topics, publishing logs using Kafka.
- In the transformation and load process, the schema is defined and data is read from Kafka as a streaming Dataframe, storing in Cassandra for Hot Path under Speed Layer and in Hadoop for Cold Path under Batch Layer.
- Finally, data is visualized using different plots in a Realtime, Hourly and Daily manner using Plotly and Dash.

Project Workflow:



Folder Structure: