**Harvard University Extension School**

**"Principles of Big Data Processing"**

**CSCI E-88, Fall 2022**

**Final Project Proposal**

**by Asma Naeem**

## Project Goal and Problem Statement

This project aims to set up a data processing pipeline for ingesting Twitter streaming data filtered by some criteria and move it into Elasticsearch for visualization and insights.

## Big Data Source

Twitter streaming data

## Expected Results

The outcome of this processing pipeline is to demonstrate graphs and maps for visualizing tweets data to compare the most popular video-on-demand over-the-top streaming services like Netflix, Amazon Prime Video, Disney Plus, and Hulu. We will pass each tweet message through the StanfordCoreNLP library to conduct sentiment analysis that scores it from 0 to 4 based on whether the analysis comes back with Very Negative, Negative, Neutral, Positive, or Very Positive, respectively.

## Processing Pipeline

Diagram

Description automatically generated

## Pipeline Overview and Technologies used

* Collection Tier: We will use Apache Flume distributed, reliable service for collecting and moving large amounts of data.
  + Flume Twitter Data Source fetches Twitter streaming data filtered by video-on-demand service tags like Netflix, Amazon Prime Video, Disney Plus, and Hulu.
* Messaging Tier: Apache Kafka
  + Flume will transport collected data to a Kafka topic for further processing
* Stream Processing Tier: Kafka Elasticsearch Sink Connector
  + Kafka Elasticsearch sink service will write data from a topic in Apache Kafka to an index in Elasticsearch
* Visualization Tier: Kibana source-available data visualization dashboard software for Elasticsearch.
  + Kibana user interface will allow us to explore which video-on-demand services receive the most positive or negative tweets on a given day/ hour.

## New Technology/Framework used

* Apache Flume distributed data system for efficiently ingesting and processing streaming data in real time.
* Kafka Connect Elasticsearch Service Sink connector to move data from Apache Kafka to Elasticsearch.