**“Sample”**

**Harvard University Extension School**

**"Principles of Big Data Processing"**

**CSCI E-88, Fall 2019**

**Final Project Proposal**

**by Marina Popova**

## Project Goal and Problem Statement

This project's goal is to demonstrate how one can build a system that collects Twitter data by some criteria and indexes them into ElasticSearch for further analytics

## Big Data Source

Twitter streaming data

## Expected Results

As a result of my processing pipeline, I expect to be able to show comparison graphs of most popular twitter topics/tags in Boston vs NYC

**“Sample”**

## Processing Pipeline



## Pipeline Overview and Technologies used

* Collection tier: Flume with experimental Twitter streaming source
  + Flume will be ingesting data from Twitter, filtered by Boston and New York City tags
* Messaging Tier: Kafka
  + Flume will push events into Kafka for further processing
* Stream Processing Tier: Kafka ElasticSearch Connector
  + Connector will index data into daily indexes
* Visualization Tier: we will use Kibana with ElasticSearch to visualize received data and discover which topics are more popular in Boston vs. NYC on a given day/hour

## New Technology/Framework used

* Flume Twitter Source
* Kafka ES Connector