Collecting, Preprocessing, and Analyzing Large-Scale Social Media Data: COVID-19 Case Study

Jian Cao[†], Md. Yasin Kabir[‡] and R. Michael Alvarez[†]

August 23, 2021

†California Institute of Technology
‡NVIDIA; Missouri University of Science and Technology

Introduction

Introduction

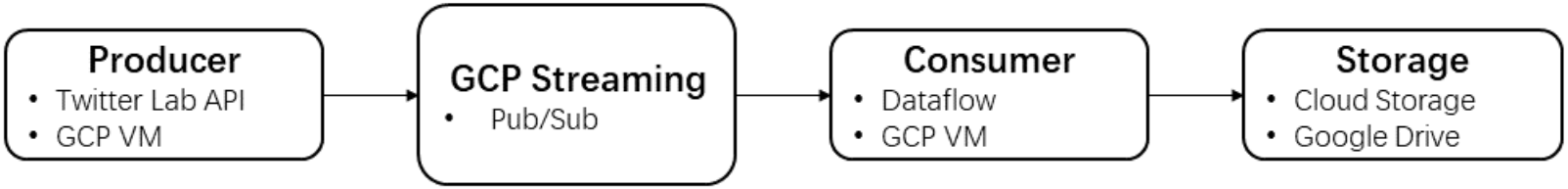
- Collecting large-scale, streaming, real-time social media data important for research (Srikanth et al., SIGKDD 2021):
  - Detection and analysis of low-incidence social media behavior in general.
  - Attack and hate speech.
  - Misinformation/disinformation campaigns.
  - Strategic behavioral shifts.
- But these datasets are large, making collection, preprocessing, and analysis difficult and computationally intensive.
- In our research, we are building architectures to efficiently and reliably collect these large datasets, to preprocess them quickly and easily, and to analyze them fast and efficiently.

Fast and Reliable Collection

Twitter Monitor Setup

In the peak hours in August 2021, on average 250 COVID-19 tweets were posted every second (0.9 million tweets/hour). It is important to use a Twitter monitor that is both fast and failure-tolerant to be able to maximize the data collection.

We developed a cloud architecture (Cao, Adams-Cohen, Alvarez, 2020) that is fast in ingesting tweets and robust to errors, system failures, and network fluctuations.



The code of the architecture are available at:
github.com/jian-frank-cao/MonitoringTwitter

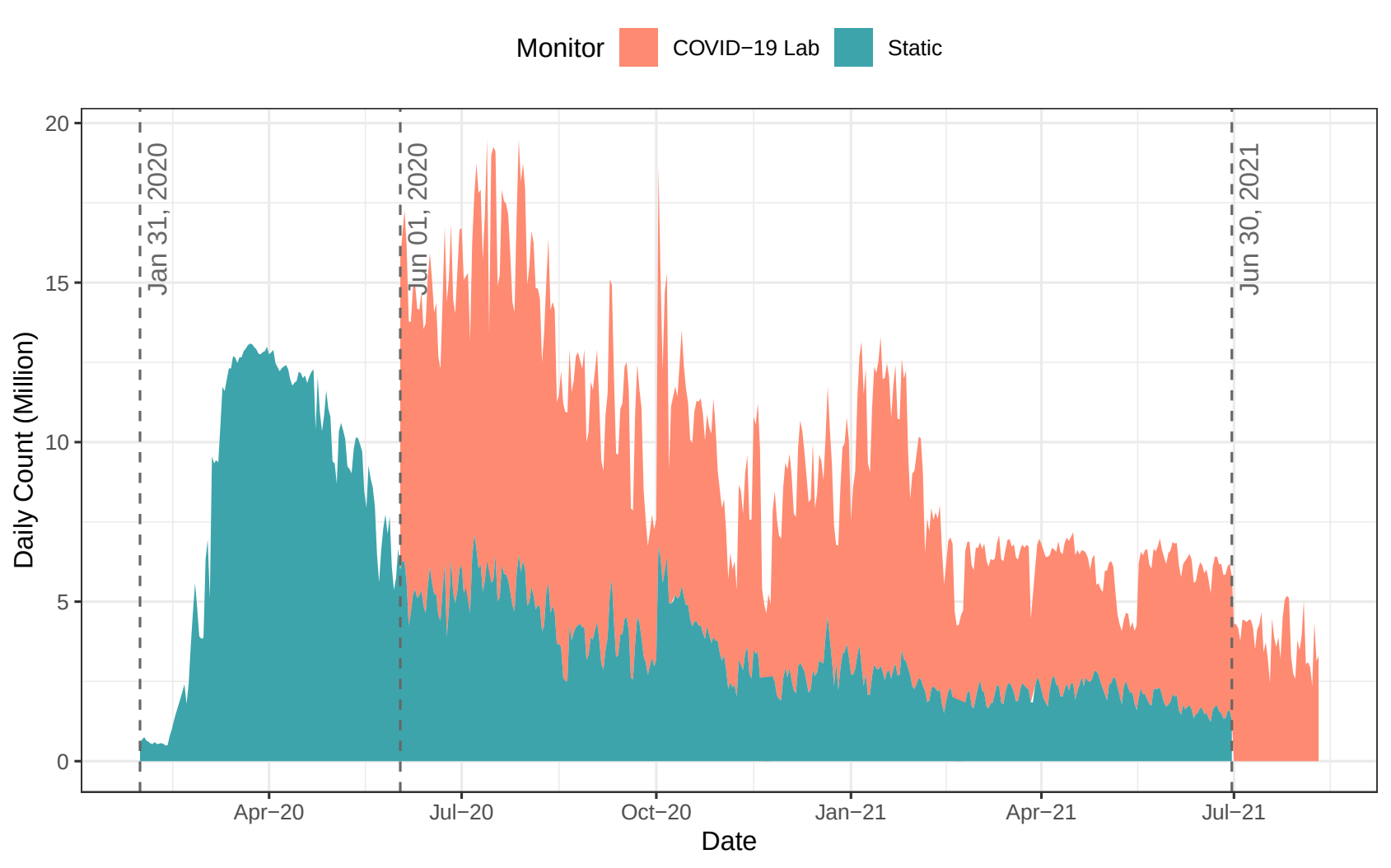
The Twitter COVID-19 Monitor Using Static Keywords

- **Online Date:** Jan 31, 2020 - Jun 30, 2021.
- **Keywords:** *coronavirus, #coronavirusoutbreak, COVID-19, #facemask, pandemic, #WHO, #2020ncov, #2019ncov, wuhan, #wuhanvirus, #wuhanlockdown, #WuhanSARS, #SARS, #CoronavirusWho, #ChinaVirus, #Wuhanpneumonia, Virus.*
- **Online Days:** 516 days.
- **Peak Ingestion Rate:** 150 tweets/second.
- **Tweets Collected:** 1.86 billion.
- **Size:** 8.72 TB

The Twitter COVID-19 Lab Monitor

- **Online Date:** June 1, 2020 - present.
- **Keywords:** The keyword list is managed by Twitter. It contains more than 500 topics in multiple languages. The **representative ones** are *covid, corona, covid-19, mask, pandemic, ppe, lockdown, quarantine, hospital, social distancing, tested positive, hand sanitizer, and many others.*
- **Online Days:** 394 days.
- **Peak Ingestion Rate:** 250 tweets/second.
- **Tweets Collected:** 5.7 billion.
- **Size of Tweets:** 29.04 TB

Daily Number of Tweets Collected



Fast and Efficient Preprocessing

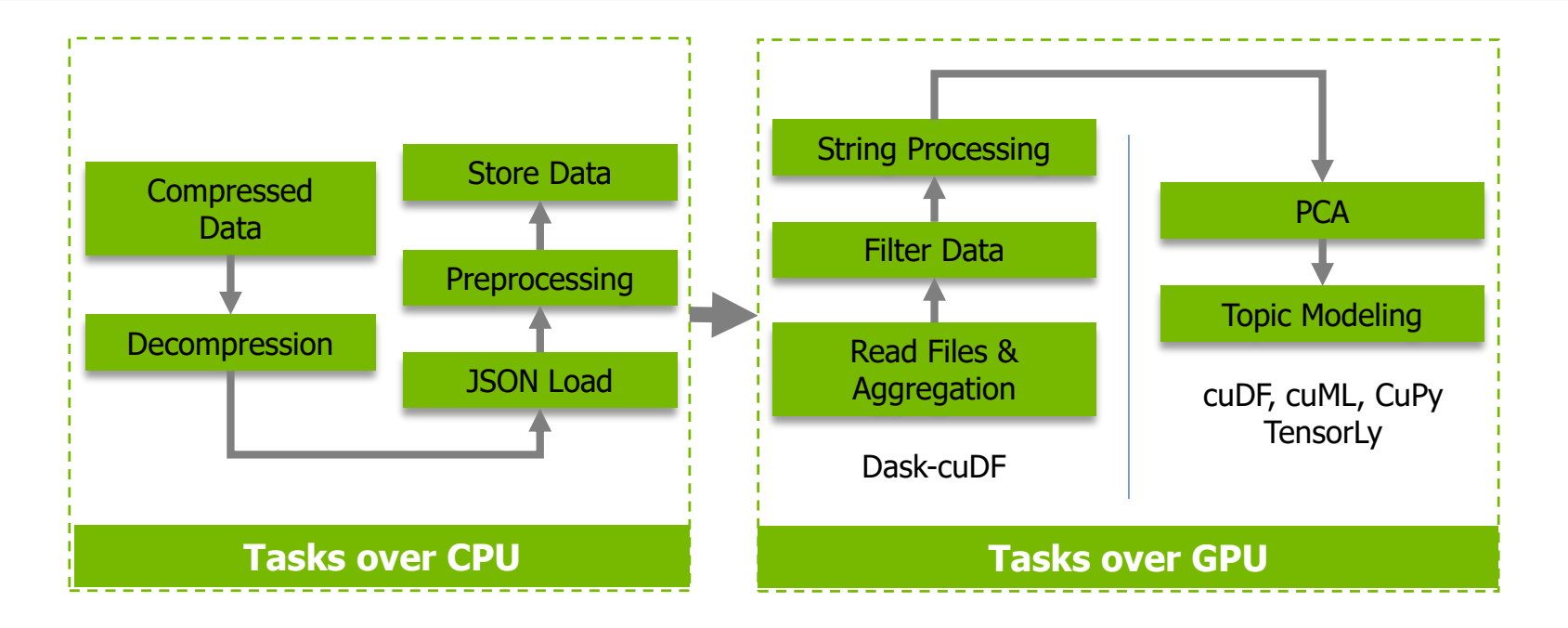
RAPIDS

From RAPIDS website[1]:

- **RAPIDS** is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs.
- **RAPIDS** utilizes NVIDIA CUDA® primitives for low-level compute optimization, and exposes GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces.
- **RAPIDS** also includes support for multi-node, multi-GPU deployments, enabling vastly accelerated processing and training on much larger dataset sizes.

[1] Source: https://rapids.ai/about.html

RAPIDS - Workflow



RAPIDS - Workflow: Data Cleaner

- Extract and Load the json files.
- Selecting specific attributes from the raw text.
- Saving the data to the storage.
- Performed over CPUs (multicore).

RAPIDS - Workflow: Filtering and Processing

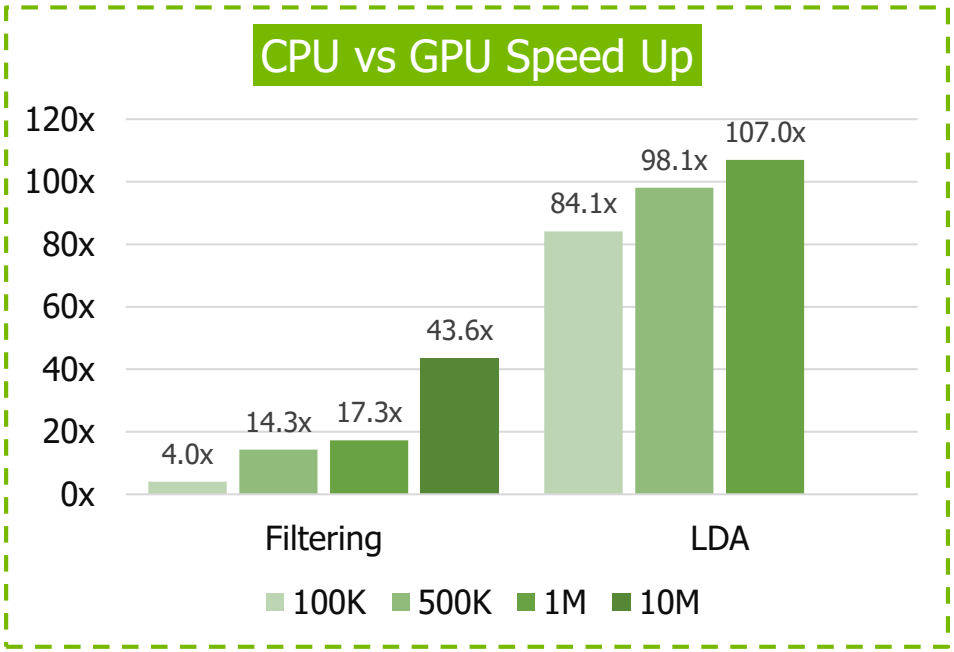
- Filtering specific tweets (e.g. Tweets from US members of congress, verified users).
- There are total 720 US congressmen combining present and previous members. We filtered tweets from the whole corpus.
- Processing tweets to remove duplicates, stop words, hashtags, mentions, etc.
- Performed over GPUs using RAPIDS libraries such as cuDF, Dask-cuDF, cuML, CuPy.
- All of the above steps takes about 3 seconds for 1 million tweets.

RAPIDS - Workflow: Topics Extraction

- Adaptation and use of some methods (e.g. Word Vectorization, Incremental PCA) over GPU using RAPIDS libraries.
- Performed arithmetic operation using CuPy (GPU).
- Executed TLDA over GPU.
- Clustered the tweets based on the words for topic exploration.
- Example: We have used cuML Kmeans clustering with variable number of clusters. Clustering the tweets into 10 groups using Kmeans over GPU takes around 293 ms only.

RAPIDS - Performance

Attribute	Summary
Data	COVID-19 tweets (Jan 20-Jun 21)
Total Files	103387
Size	Compressed: 643.4 GB; Raw: ~11.5 TB.
#Eng Tweets	265.5 Millions.



Fast and Efficient Analysis

Faster NLP with Tensors

- Traditionally topic models have been estimated using Expectation Maximization– require many iterations to converge with poor posterior performance.
- Anandkumar et al. (2012) and Anandkumar et al. (2013) show that spectral decomposition of low-order empirical moments can be used to recover the parameters of a method of moments estimator for latent variable models like LDA and JST.
- Moment-based algorithms are in general faster than EM because they require only a single iteration through the the corpus.
- We are producing easy-to-use Python libraries for tensor-based Latent Dirichlet Allocation (LDA) and Joint Sentiment-Topic (JST) models, optimized for use on GPUs. Will soon be available on TensorLy!

Preliminary Results on Performance Gains

Model	Convergence Time	Optimal Coherence (NPMI)	Optimal Topics
LDA	2:17:05	0.49	80
JST	1:14:23	0.59	30*3
Tensor JST	00:25:13	0.53	28*3

Table 1: Model Comparison: All three models were run on the same system architecture. Proccessor: Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz, 64gb RAM, 10 cores. Twitter dataset (N=315,000).

Summary

- The long-term collection of streaming social media data is important for researchers.
- Can be done efficiently and reliably in cloud-based architectures (here GCP).
- Preprocessing and filtering of these large datasets for research is fast and efficient using technologies like RAPIDS.
- On-going development of faster NLP methods for these large-scale datasets.

Collaborators and Partners

Research Leads

- Anima Anankumar (Caltech)
- Anqi Liu (JHU)
- Bojan Tunguz (NVIDIA)
- Jean Kossaifi (NVIDIA)

Students and postdocs (past & present)

- Nicholas Adams-Cohen (formerly Caltech)
- Sara Kangaslahti (Caltech)
- Adrien Schurger-Foy (Chapman)
- Maya Srikanth (formerly Caltech)

Partners

- Google (COVID-19 research program)
- NVIDIA