Journal of Ambient Intelligence and Humanized Computing (2023) 14:773–787
https://doi.org/10.1007/s12652-021-03328-0

ORIGINAL RESEARCH

Distributed messaging and light streaming system for combating pandemics

A case study on spatial analysis of COVID-19 Geo-tagged Twitter dataset

Yavuz Melih Özgüven[1] · **Süleyman Eken**[2]

Received: 16 December 2020 / Accepted: 3 June 2021 / Published online: 10 June 2021

Abstract
Real-time data processing and distributed messaging are problems that have been worked on for a long time. As the amount of spatial data being produced has increased, coupled with increasingly complex software solutions being developed, there is a need for platforms that address these needs. In this paper, we present a distributed and light streaming system for combating pandemics and give a case study on spatial analysis of the COVID-19 geo-tagged Twitter dataset. In this system, three of the major components are the translation of tweets matching with user-defined bounding boxes, name entity recognition in tweets, and skyline queries. Apache Pulsar addresses all these components in this paper. With the proposed system, end-users have the capability of getting COVID-19 related information within foreign regions, filtering/searching location, organization, person, and miscellaneous based tweets, and performing skyline based queries. The evaluation of the proposed system is done based on certain characteristics and performance metrics. The study differs greatly from other studies in terms of using distributed computing and big data technologies on spatial data to combat COVID-19. It is concluded that Pulsar is designed to handle large amounts of long-term on disk persistence.

Keywords Topic-based publish-subscribe · Name entity recognition · Spatial analysis · Skyline query · Geo-tagged twitter data · Translation · Apache pulsar

1 Introduction

COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a human-transferable infectious respiratory disease. The disease first appeared in Wuhan, China in 2019, and has grown exponentially since then causing a 2019-20 worldwide Coronavirus pandemic (Hui et al. 2020). Fever, coughing, and shortness of breath are the most common symptoms of the disease. Other symptoms include muscle aches, sputum production, and sore throat. In some cases, it is accompanied by diarrhea and other gastrointestinal symptoms (Gu et al. 2020; Miri et al. 2020). At the time of writing (December 2020), the number of reported COVID-19 cases and its death toll globally has surpassed 73 million and 1,628,000 respectively[1]. Several outbreaks, epidemics and also pandemics have occurred in the world throughout history. Figure 1 shows the most recent ones.

The field of medical data science covers different areas such as prediction of response to treatment in personalized medicine (Abul-Husn and Kenny 2019; Suwinski et al. 2019), biomarker detection (Zhang et al. 2019; Fitzgerald 2020), tumor classification (Khan et al. 2019; Lin and Berger 2020), COVID detection and classification (Wang et al. 2020; Bragazzi et al. 2020; Narin et al. 2021), and the understanding of genes interactions (Shukla and Muhuri 2019). Corsi et al. (Corsi et al. 2020) give a systematic review of literature for big data analytics (Eken 2020a) as a tool for fighting pandemics. Besides working on medical data, other sources of information such as social media can

✉ Süleyman Eken
suleyman.eken@kocaeli.edu.tr

[1] Department of Computer Engineering, Kocaeli University, 41001 İzmit, Turkey

[2] Department of Information Systems Engineering, Kocaeli University, 41001 İzmit, Turkey

[1] https://www.worldometers.info/coronavirus/

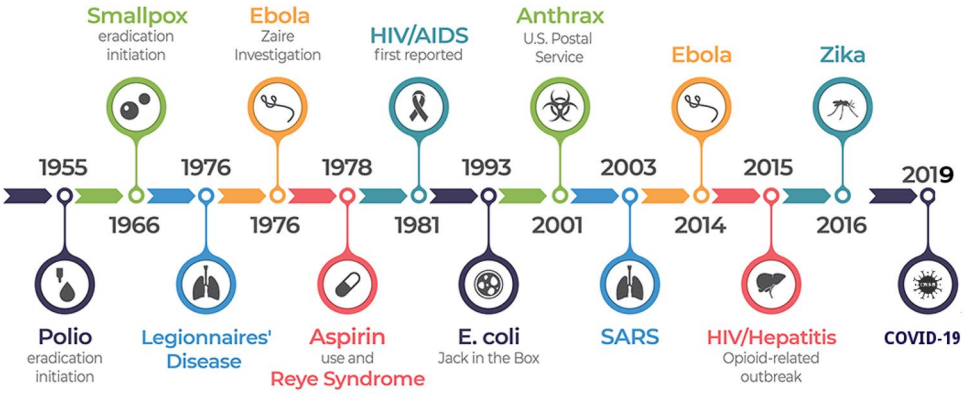


Fig. 1 Time-line of outbreaks, epidemics and pandemics. Source: adapted from CDC[**]https://www.cdc.gov/eis/about/history.html

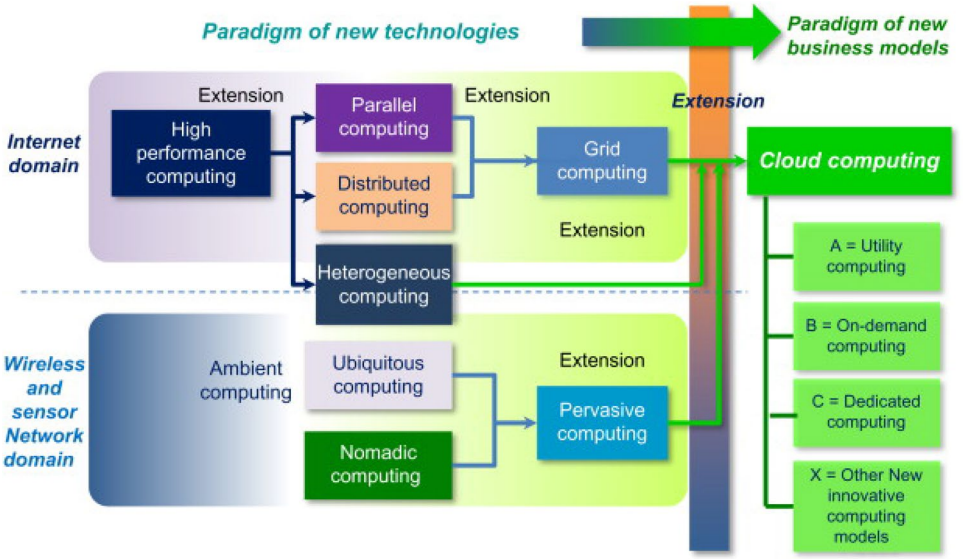


Fig. 2 Paradigm shift from technologies to business models (Wu and Buyya 2015)

be used to combat the pandemic. Aggregating and processing big data from all of these sources becomes challenging if central processing techniques are considered, which hurts the accuracy and the timeliness of the information. Therefore, there is a need to adapt distributed and parallel computing technologies in the research effort to tackle COVID-19. A distributed system is a group of separate and self-sufficient computing elements (nodes) combined and presented to its users as a single coherent system. Each node is autonomous and has its own notion of time. This lack of a global clock leads to major synchronization and coordination problems. The scalability of distributed systems can be achieved in various ways: size scalability, geographical scalability, administrative scalability. They denote the number of users and/or processes, the maximum distance between nodes, and the number of administrative domains, respectively. Size scalability is often the one problem most addressed by such systems. Parallel computing, where multiple powerful servers operating independently in parallel, is an alternative solution. In this model, however, a global clock is a requirement to synchronize the processing done independently by simultaneously on multiple sub-tasks during each clock-cycle, and combine their results to solve the original task. Parallel computing, cluster computing, grid computing, and cloud computing are kind of high performance distributed computing mechanisms (Tanenbaum and Van Steen 2007). Considering big data processing and analytics, emergent hardware technologies and new computing paradigms such as co-processors, fog computing, and dew computing are possible (Groppe 2020). Figure 2 shows different computing paradigms.

Publish/subscribe is one of the most well-known messaging patterns used to communicate data between a sender (producer) and a receiver (consumer) (Tanenbaum and Van Steen 2007). Instead of sending the messages directly between each other, a broker is most often used to facilitate communication. The publishers send messages to so-called topics in the broker, which are used to separate different types of data that are being communicated. The broker is responsible for correctly routing each message to the

subscribers of a topic. Each topic can have several subscribers, and the incoming messages will be delivered to all of them (Fabret et al. 2001). While the terminology used in systems such as Apache Kafka and Apache Pulsar are slightly different, both are all based on this publish/subscribe type of communication. They also offer more advanced platform-specific features that extend the foundation of publish/subscribe messaging (Harrison et al. 1997). Our research questions for this study are as follows: (i) Is it possible to combat pandemics with a distributed messaging and light streaming system? and (ii) How does spatial-based tasks help authority and people in pandemic times?

The contributions of this work to the literature are as follows:

- Topic-based messaging and streaming system using Apache Pulsar[2] is proposed for combating pandemics. The proposed system converts hard real-time applications to real-time (soft real-time). As a case study, a spatial analysis of COVID-19 Geo-tagged Twitter dataset is given.
- Translation of tweets matching with user-defined bounding boxes is done. So, anybody can inform COVID-19 related tweets in a specific area/region he/she subscribed to in his/her native language.
- Name entities such as a person, location, organization, and miscellaneous are recognized in tweets and they are sent to consumers subscribed to them.
- Skyline queries are also performed on tweets.

The remainder of this article is organized as follows. Section 2 gives a literature review on the impact of social media on the pandemic, distributed systems in pandemic, and spatial analysis of COVID-19 data. Section 3 gives details of the distributed messaging system for combating pandemics. Section 4 presents the performance of the system discussion. Section 5 summarizes and concludes the article and also gives future works.

2 Related works

This section explains relevant works in the literature specific to importance of social media in the pandemic time, distributed systems for battling with COVID-19 pandemic, and spatial analysis of COVID-19 data.

2.1 The importance of social media in the pandemic time

Wong et al. (2019) point out that they have a plethora of choices when it comes to epidemiological transmission data sources, such as sentinel reporting systems, outbreak reports, disease centers, genome databases, vaccinology-related data, transport systems, and social media data. Social media occupies an increasingly vital role in informing the public during crises and emergencies; it also proved to be a powerful tool in shaping outrage and with it, the public's attitudes towards risks and mitigation strategies (Ophir 2018; Quinn 2018). These unique characteristics, make social media both a help and a hindrance in developing adequate strategies for risk communication and response planning. The general public tends to pick select media channels for news and follow them exclusively (Malecki and Keating 2021). The critical impact of such an information environment is clearly shown and augmented in the case of the COVID-19 pandemic.

In the last few years, Facebook and other social media platforms such as Instagram, Twitter, Youtube, Youtube, Reddit, etc. have been the go-to option for informal communications. Considering the lack of pharmaceutical interventions to fight COVID-19, and having to rely on quarantine and social distancing measures, leveraging social media intelligence in this fight becomes of utmost importance to inform and influence the public's mobilization to follow quarantine procedures in their local communities, quickly disperse any fears and uncertainty to avoid community panic, and improve the public trust in these health measures. The COVID-19 crisis showed us how important is the development of real-time information sharing systems, that can aggregate data and analyses in multiple languages and from different platforms across the whole world, and adapt to the dynamic and fast nature of mentioned platforms. Otherwise, public health bodies would be rendered unable to respond to the spread of information and misinformation about the outbreak, nor to promptly present the right measures to handle it (Depoux et al. 2020).

In the literature, the use of a variety of social media data types such as text, image, and video are noted. These works cover the analysis of social media conversations concerning the epidemic situation geographically (geo-coded tweets/messages) and over time (timestamped tweets/messages), and are often summarized and presented in the form of real-time maps.[3]Cinelli et al. (2020) focus on analyzing engagement and interest in the topic of COVID-19 and providing a differential assessment on the evolution of the discourse for several social media platforms globally. It is very important work to study content consumption dynamics around critical

[2] https://pulsar.apache.org/

[3] https://coronavirus.jhu.edu/map.html.

events in times of disinformation. Li et al. (2020) manage to classify COVID-19-related information from Weibo, a major social network in China, into seven types of situational information such as caution and advice, notifications and measures been taken, help-seeking, etc, by using natural language processing techniques. So, the situational information can be used by researchers or practitioners to build effective crisis information systems. Boberg et al. (2020) mine data from alternative news media's output on Facebook during the early days of the pandemic, to create an initial computational content analysis of community fear and its factual basis. Providing metrics to measure reach, shares, topic detection, and total interactions comparisons between mainstream and alternative media. Alternative news media can mirror mainstream media reports with reversed ones. It is important that alternative news sources reflect the truth, at least in the period of interest. Szmuda et al. (2020) study Youtube videos from the early period of the pandemic and assess their content-quality as well as audience engagement. Yüce et al. (2020) follow the same approach but focus on dentistry-related medical information about COVID-19 and evaluate them as a potential educational resource for dental practitioners. Despite the many negativities of the pandemic, it is positive in that it enables many education institutions to prepare and improve distance education infrastructure and provide training as an alternative through social media. Lamsal (2020), on the other hand, uses unigrams and bigrams trending terms, network analysis to create and visualize their tweet dataset.

2.2 Distributed systems for battling with COVID-19 pandemic

In this sub-section, different computing paradigms involved in fighting COVID-19 are mentioned. There is no question that global health security is exposed to serious risks due to pandemics such as COVID-19. Facing such risks requires multi-disciplinary research efforts, such as with computational epidemiology, which is concerned with the development and use of computer models to understand and predict the spatio-temporal spread of disease through populations. This spread is heavily influenced by the arrangement of the interaction network across which the outbreak happens. These models are designed in such a way that enables scientists to create detailed computer simulations to properly inform public health bodies and aid them in decision making regarding response policies. However, it is a big computational challenge to develop such high-resolution simulations for a few reasons: (i) scale and heterogeneity of contact networks, (ii) the dynamic nature of those networks, (iii) achieving realistic results requires running a large number of independent simulation for each combination of parameters. Solving these challenges entails the use of High-Performance Computing (HPC) based simulations. Bisset et al. (2009) propose a fast, scalable, high-performance simulation tool by the name of EpiFast, which makes the study of the spread of infectious diseases through individual populations feasible. They follow up on their work (Bisset et al. 2014) by proposing an HPC-based service architecture for epidemic modeling. The architecture consists of disease progression simulation, situation assessment, and intervention simulation. Computational models provide a powerful tool to study the role of individual behavior and public policies in containing the pandemics. Marathe (2020) presents their work on scalable and pervasive computing-based concepts, theories, and tools for planning, forecasting, and response in the event of epidemics. This research is usefull for estimation of various disease progression parameters. Remote access allows computational chemists and biologists to run tasks on high-performance computers or cloud servers from anywhere instead of requiring their presence at location (Amaro and Mulholland 2020). In the U.S., the COVID-19 HPC Consortium[4] helps accelerate research on the topic by combining the most powerful compute resources and facilitating researchers' access to them through a rapid proposal process. The consortium comprises leading companies such as IBM, Microsoft, and Google, as well as universities and national labs.

Supercomputers are a family of extremely powerful computers, and they are being leveraged to combat COVID-19. Scientists can benefit from policies to allocate computing time for emergencies by supercomputing centers. Some of those scientists are using this processing power to study the structure of the virus, the folding of its "spike" protein, and how it differs from other viruses in the corona family. For example, thanks to such efforts at the Summit supercomputer at Oak Ridge National Laboratory, researchers succeeded in bringing down the number of potential virus-fighting molecules from 8,000 to just 77 (Smith and Smith 2020).

There are also common technological platforms and software frameworks such as Apache Hadoop[5] and Spark[6] for processing bigdata. Apache Hadoop provides a simple framework for distributed/parallel data processing based on the available commodity hardware. Apache Spark unifies streaming, batch, and interactive big data workloads to unlock new applications. Khashan et al. (2020) introduce a framework to handle complex queries for COVID-19 datasets named COVID-QF. It consists of data collection, storage, and query processing layers. Their proposed system is valid for analysinglarge numbers of data via SQLor large numbers of data via NOSQL. Melenli and Topkaya

[4] https://covid19-hpc-consortium.org

[5] https://hadoop.apache.org/.

[6] https://spark.apache.org/.

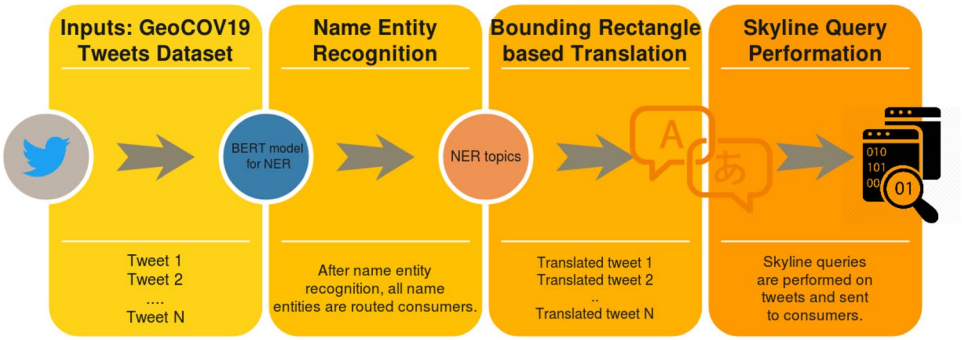


Fig. 3 Sub-components of the proposed system

(2020) propose a system to detect people in video streams in real-time, calculate their social distance, and report results using different Apache projects. The end user is provided to report the regions where the violation is density in real-time. Elmeiligy et al. (2020) propose Comprehensive Storing System for COVID-19 data using Apache Spark (CSS-COVID) consists of three stages, namely, inserting and indexing, storing, and querying stage. This work enables to manage and analyze different cases such as suspected ones. Eken (2020b) proposes a topic-based hierarchical pub/sub messaging middleware. It allows end-users to filter images as COVID-19 or non-COVID-19 using capsule networks and their metadata such as gender and age. So, proposed middleware allows for a smaller search space as well as shorter times for obtaining search results. De Souza et al. (2020) introduce BurstFlow, a tool for enhancing communication across data sources located at the edges of the Internet and big data stream processing applications located in cloud infrastructures. It can be used for stream applications such as financial markets and health care.

Grid computing includes lots of nodes from everywhere. They are heterogeneous and dispersed across several organizations to allow for collaborations. COVID-19 research has already attracted the contribution of large-scale grids. To list a few examples, Berkeley Open Infrastructure for Network Computing (BOINC) (Anderson 2019), Globus[7], the Open Science Grid (OSG) (Pordes et al. 2007), the World Community Grid launched by IBM,[8] and the WLCG (Worldwide LHC Computing Grid) at CERN.[9] Cloud services provide access to network-based computing resources with minimal human interaction between the user and the service provider. Kaplan et al. (2020) worked on adjusting the cloud architecture to accommodate the needs of the problem rather than manipulating the problem itself to make it suitable for the platform and its limitations. So, they are able to be prepared to rapidly deploy the model and to rapidly implement the model at scale for COVID-19. Also, there are solutions with other new innovative technologies such as ambient computing, ubiquitous computing, pervasive computing, and dedicated computing for fighting COVID-19 (Arun et al. 2020; Sbai et al. 2020; Magesh et al. 2020).

2.3 Spatial analysis of COVID-19 data

Big data technologies in general and Geographic Information Systems (GIS) specifically have played a significant role in the war against COVID-19. The role spans several aspects of the fight such as what is covered in the following papers: spatial segmentation of the epidemic risk and prevention level (Franch-Pardo et al. 2020), the rapid aggregation of multi-source big data (Huang et al. 2021), prediction of regional transmission (Hamzah et al. 2020), rapid visualization of epidemic information (Tebé et al. 2020), spatial tracking of confirmed cases (Boulos and Geraghty 2020), balancing and management of the supply and demand of material resources (Govindan et al. 2020), and social-emotional guidance and panic elimination (Shah et al. 2020), which provided solid spatial information support for decision-making (Xu et al. 2020), measures formulation (Wong et al. 2020), and effectiveness assessment of COVID-19 prevention and control (Sun and Zhai 2020). Finally, Zhou et al. address the concern regarding difficulties faced by GIS with big data and responses (Zhou et al. 2020).

3 Materials and methods

In this section, the proposed distributed messaging system are explained. Figure 3 shows the sub-modules of the system.

In the proposed system, end users will be able to translate tweets matching with user-defined bounding boxes. The user only specifies/subscribes to a geo-border and any language and then translated tweets in a specified location are routed them. Also, users can subscribe to name entities in tweets and so entity-related tweets are services. Moreover, the user

[7] https://www.globus.org/.

[8] https://www.worldcommunitygrid.org/about_us/viewAboutUs.do.

[9] https://wlcg.web.cern.ch/.

Fig. 4 An example of Twitter JSON place


```
{
   "place": {
     "id": "xy69f456640963kc7",
     "url": "https://api.twitter.com/1.1/geo/id/fd70c22040963ac7.json",
     "place_type": "city",
     "name": "Boulder",
     "full_name": "Boulder, CO",
     "country_code": "US",
     "country": "United States",
     "contained_within": [],
     "bounding_box": {
       "type": "Polygon",
       "coordinates": [
         [
           [-105.301758, 39.964069],
           [-105.301758, 40.094551],
           [-105.178142, 40.094551],
           [-105.178142, 39.964069]
         ]
       ]
     },
     "attributes": {}
   }
}
```


may run different skyline queries on tweets. All these spatial analysis tasks are given in detailed at following sub-sections.

In this paper, Apache Pulsar (version 2.7.0) high performance distributed messaging platform is used for topic-based pub/sub system. While originally created by Yahoo, it has since become apart of the Apache Software Foundation. It is used for gathering and processing different events in near-realtime, for use cases such as reporting, monitoring, marketing and advertising, personalization and fraud detection. For example, at eBay, Pulsar has been used to improve the user experience by analyzing user interactions and behaviors. Pulsar is closely related to Apache Kafka in terms of features and use cases. It offers great scalability for message processing on a large scale, with high throughput and low end-to-end latency. Messages received are stored persistently with the help of Apache BookKeeper, and message delivery is guaranteed between producers and consumers. While Pulsar is not a stream processing framework as the likes of Apache Storm or Spark Streaming, it does provide some light stream processing features with the use of Pulsar Functions.

Like Kafka, Pulsar is based on the publish/subscribe messaging pattern. Producers send messages to certain topics, which are used to separate different types of messages. Consumers can then subscribe to specific topics to consume the data. The persistent storage that Pulsar offers means that all messages are retained, even when a consumer loses connection. The disconnected consumer can therefore easily reconnect and continue consuming the remaining data without any data loss. Pulsar offers several different subscription modes for distributing messages to consumers. This includes the following modes: (i) Exclusive: Only one consumer can be subscribed to the topic at a time. (ii) Failover: Several consumers can be subscribed to the topic at the same time using a master-slave approach. Only one of the consumers receive messages (the master). However, if the master consumer happens to disconnect, any subsequent messages will be directed to the following consumer (slave). (iii) Shared: Multiple consumers can be subscribed to the same topic. Messages are load-balanced between all the connected consumers, i.e. messages are only consumed once. Shared subscription does not guarantee correct message ordering. (iv) Key shared: Similar to shared subscription mode, except that the message distribution is done based on key values. In this paper, shared mode is used.

3.1 Translation of tweets matching with user-defined bounding boxes

Translation has a significant weight and a more complex effect on everyday lives and in providing accessibility to everyone. Understanding its importance is the first step to justify investing in it. Most people would rather stay within their comfort zone when it comes to languages, and their

native tongue becomes the default preference since they can speak in it more confidently than any second language they learned. This is why translation is of utmost importance and will allow for efficient communication between people.

Tweet data can contain two types of geographical meta-data: (i) Tweet location - Available if the user opts to tag his tweets at the time of making them. (ii) Account Location - Available if the user updates his 'home' in his profile and makes it public. The latter is, however, a free-form character field, and its metadata is not guaranteed to be geo-referenceable (Häberle et al. 2019). Figure 4 is an example JSON from a Tweet with "Boulder, CO" being its geo-tag or Twitter Place. It includes tweet bounding box containing the place entity coordinates (west, south, east, north longitude and latitude points) besides other meta-data such as tweet id, url, country-code, and etc. So, these coordinates are used to filter tweets. In this paper, Algorithm 1 is used to find and filter tweets to be translated. After finding all tweets, Google Translate API[10] is used to translate them. So, every consumer can subscribe to a geo-border and any language then translated tweets in a specified location are routed them. Performance results are presented in Sect. 4.2.

Algorithm 1: Pseudo-code for finding tweets to be translated

```
q ← query;
r ← tweet_rectangle_coordinates;
tlist ← while While condition do
    if (!(q.min_x > r.max_x)&&!(r.min_x > q.max_x)&&!(q.min_y > r.max_y)&&!(r.miny > q.maxy)) then
        add tweet id whose coordinates r to tlist
    end
end
```

3.2 Recognition name entities in tweets

Name Entity Recognition (NER) is the process of finding entities in a text and assigning them to one of the pre-defined classes such as a person, location, date, formula, percentage, organization and money (Nadeau and Sekine 2007). With all these, NER is not limited to these data types, but is also used to identify and mark entities specific to the relevant area in studies in different fields. E-mail addresses (Minkov et al. 2005), phone numbers, book titles, project names, gene/protein names in bioinformatics and chemistry texts (Tanabe et al. 2005), RNA, DNA, cell information, drug names (Kim et al. 2004), chemical names (Eltyeb and Salim 2014) as entity names are the subjects studied. Thompson et al. (2015) devised a web-based History of Medicine tool by benefiting from text mining methods to provide its user an efficient search from historical texts which were British

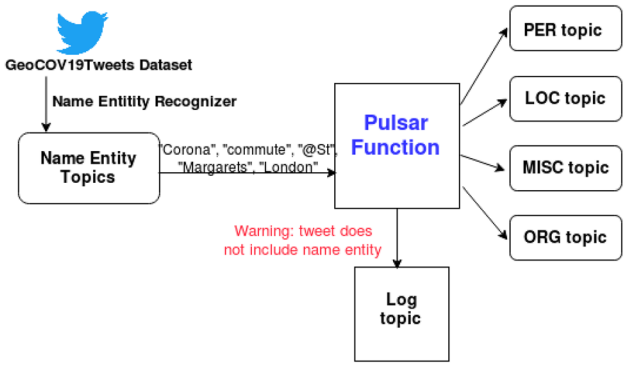


Fig. 5 NER based routing schema

Medical Journal and London Medical Officer of Health reports. The tool presented its user term, bibliographic meta-data, entity, event and named entity based search. Li et al. (2012) present a novel two-step unsupervised NER system for the targeted Twitter stream, called TwiNER. In the first step, it leverages the global context obtained from Wikipedia and Web N-Gram corpus to partition tweets into valid segments (phrases) using a dynamic programming algorithm. In the second step, TwiNER constructs a random walk model to exploit the gregarious property in the local context derived from the Twitter stream. Liu et al. (2020) propose an unsupervised framework NELPTW which makes use of the abundant geographical location knowledge embedded in both Twitter and Web to predict named entity city-level location.

In this paper, pre-trained deep bidirectional network -BERT (Devlin et al. 2018)- is used to make a model for named entity recognition in tweets. There are different pre-trained word embeddings models such as WordEmbeddings (GloVe), BertEmbeddings, and ElmoEmbeddings. BertEmbeddings (bert_base_cased) model which has 12 layers of transformer encoders and 'cased' sequences is used to recognize entities. PER (Person), ORG (Organization), LOC (Location), and MISC (miscellaneous) named entity types as specified in CoNLL-2003 named entity dataset (Sang and De Meulder 2003) are used in this paper. So, these four types of entities are routed to consumers. A function takes items (tweets) as input and publishes them to NER type topic (PER, ORG, LOC, MISC), depending on the item. Or, if a tweet does not include a NER type, a warning is logged to a log topic. Figure 5 represents NER-based routing. Performance results are presented in sub-section 4.3.

3.3 Performing skyline queries on tweets

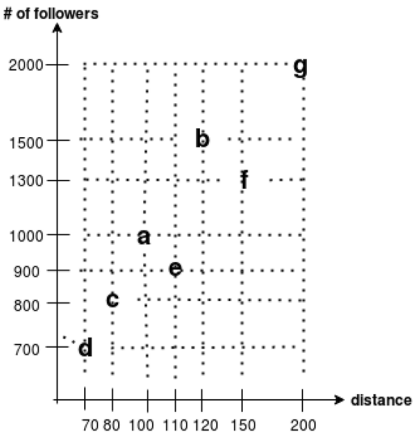
In recent years, database research has been paying attention to the issue of skyline query processing for extracting interesting objects from multi-dimensional datasets. The skyline query processing is relevant for many scenarios that require

[10] https://cloud.google.com/translate

Fig. 6 A 2-dimensional database of seven tweet objects

tweet object	distance to countries	# of followers	# of favorite for tweet
a	100	1000	3000
b	120	1500	280
c	80	800	750
d	70	700	200
e	110	900	150
f	150	1300	100
g	200	2000	500

(a)



(b)

multi-criteria decision-making without employing cumulative functions to define the most reliable results but instead relying on the preferences of the user. The skyline operator can take a large dataset of points and filters it to leave only the most interesting ones based on a set of evaluation criteria. A point is considered interesting if no other point scores higher based on the evaluation criteria. Kalyvas and Tzouramanis (2017) provide a survey on the state-of-the-art techniques for skyline query processing. Figure 6a illustrates a database of seven tweet objects $P = \{a, b, c, d, e, f, g\}$ each representing the description of a tweet with two attributes: distance and number of followers. Figure 6b shows the corresponding points in the 2-dimensional space where x and y axes correspond to the range of attributes distance and followers, respectively.

Taking this into account skyline queries deflect from the strict ranking approach of top-k queries and directed to an approach that is more understandable by humans. Opposed to top-k queries where specific ranking functions and criteria are used, skyline queries assume that every user has a series of preferences over the attributes of data. All the preferences are considered equivalent and will help to discard the items of the dataset that will not be preferred by anyone. Small subset including the most interesting and preferred items will be the skyline set or pareto optimal set. Performance results are presented in Sect. 4.4.

4 Experimental results

This section firstly describes the used GeoCOV19Tweets dataset and then gives performance results for the proposed system.

4.1 GeoCOV19Tweets dataset

GeoCOV19Tweets Dataset (Lamsal 2020) is used in this paper. Geo-tagging is the process of enriching a tweet with location information. When a user allows Twitter to access his/her device location, it can use the embedded Global Positioning System (GPS) to get accurate coordinates and add them to the tweet's metadata. This metadata contains various geo objects (Fig. 4) such as "place type": "city", "name": "Manhattan", "full name": "Manhattan, NY", "country code": "US", "country": "United States" and the bounding box (polygon) encircling the place in the form of coordinates. This dataset also contains IDs and sentiment scores concerning the COVID-19 pandemic. The tweets are obtained as part of an on-going project.[11] The model monitors the Twitter feed in real-time and filters it using 90+ different active keywords and hashtags that are most common when referencing the pandemic. corona, coronavirus, covid-19, #quarantine, and #n95 can be exemplified for these keywords and hashtags. Complying with Twitter's content redistribution policy, only the tweet IDs are shared. The dataset can be re-constructed by hydrating[12] these IDs. The tweet IDs in this dataset belong to the tweets tweeted providing an exact location. The dataset was started on March 20, 2020 and it is updated every day. It consists of 273,632 tweets in the English language.

All performance tests are done on Google cloud. Specifications of the used server are as following: server1 - Cpu: 8 core, Memory: 16 GB, OS: Ubuntu 18.04, Disk: 20 GB. Also, openjdk 1.8.0_275 for Java and apache-pulsar-2.7.0 for Pulsar are installed on the server.

[11] https://live.rlamsal.com.np.

[12] https://github.com/DocNow/hydrator

4.2 Performance of BBs-based translation

This sub-section gives performance results for translation of tweets matching with user-defined bounding boxes (BBs). BBs-based translation task includes one Pulsar function for translation. Here, the producer publishes messages to topics according to BBs. Consumers subscribe to those topics, process incoming messages, and send an acknowledgement when processing is complete. An example of four topics are published as shown Algorithm 2.

Algorithm 2: Publishing tweets with different languages

```
if (36 < latitude < 41.9)and(26 < longitude < 41) then
    publish TR_topic
end
if (31 < latitude < 48)and(−126 < longitude < −67) then
    publish USA_topic
end
if (43 < latitude < 49.7)and(−1 < longitude < 7) then
    publish FR_topic
end
if (50.17 < latitude < 58.9)and(−5.6 < longitude < 1.79) then
    publish UK_topic
end
```

Three performance metrics -CPU utilization, CPU load, and memory usage are obtained. CPU utilization is the sum of work handled by a CPU. It is also used to estimate system performance. CPU load is the number of processes that are being executed by CPU or waiting to be executed by CPU. The memory usage shows the amount of memory used on the system. Figure 7 shows these metrics.

4.3 Performance of NER and sending consumers

This sub-section gives performance results for NER and sending consumers. Name entity recognition task includes one Pulsar function for name entity extraction. Here, the producer publishes name entity topics as shown in Fig. 5. Consumers subscribe to those topics, process incoming messages, and send an acknowledgement when processing is complete. Figure 8 shows performance results for NER task. In general, NER results were very good with only a few noticeable mistakes, as shown in these example for December 10 and December 11 sub-data:

- PERSON: Joe Biden, Donald Trump, Sharon Osbourne
- LOCATION: Antioch, California, Mumbai, Maharashtra, Chandler, Arizona, Limburg, Belgium
- ORGANIZATION: Midtown East, City Hall, Royal Nairobi Golf Club, Pfizer
- MISCELLANEOUS: Raspberry, COVID, American Museum of Natural History, Xmas

4.4 Performance of skyline query run

This sub-section gives performance results for skyline queries. This task includes one Pulsar function for computing the Pareto (non-dominated) set. Here, the producer publishes messages to topics. Consumers subscribe to those topics, process incoming messages, and send an acknowledgement when processing is complete. Figure 9 shows performance results for skyline query task.

Figure 10 shows an example of Pareto efficient solution in multi-objective optimization (max favorite count and min follower count). Pareto optimality is a situation where no individual or preference criterion can be better off without making at least one individual or preference criterion worse off or without any loss thereof.

4.5 Discussion

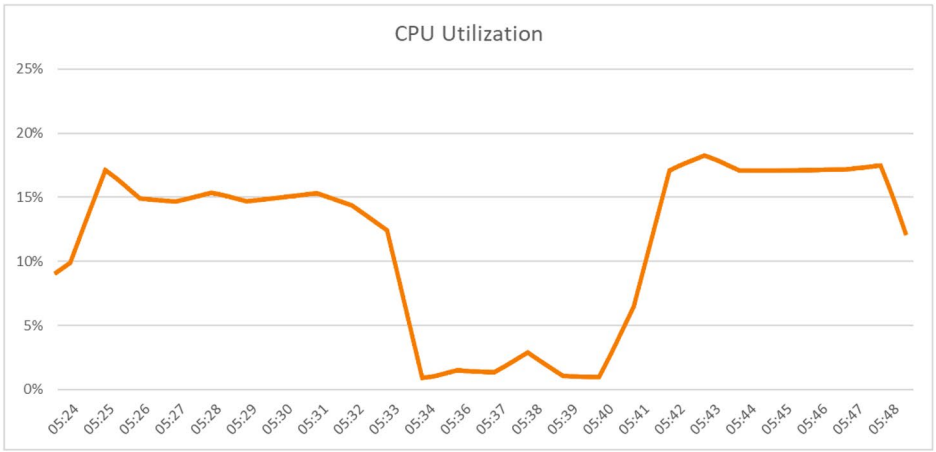
X-axis of every Figs. 7, 8 and 9 shows the elapsed time for BBs-based translation, NER detection and sending consumer, and skyline query tasks, respectively. So, BBs-based translation takes 24 min (5:48–5:24), NER detection and sending consumer takes 5 min (16:15–16:10), and skyline query takes 11 min (06:31–06:20). CPU utilization is maximum for NER detection and sending consumer task. CPU load is minimum for skyline query task. Memory usage is minimum for skyline query task.

The main limitations of the study concern the results of the performance evaluation. The tests that have been performed are very much context-dependent. These platforms have been evaluated based on specific hardware, message sizes, throughput levels and cluster size. While the results are reproducible on the same or similar hardware, it does not mean that one will arrive at the same type of conclusions in another environment with different configurations. This is a general limitation regarding this type of testing, as there are far too many different variables and potential combinations to test to be able to draw generalized conclusions that would fit every use case.

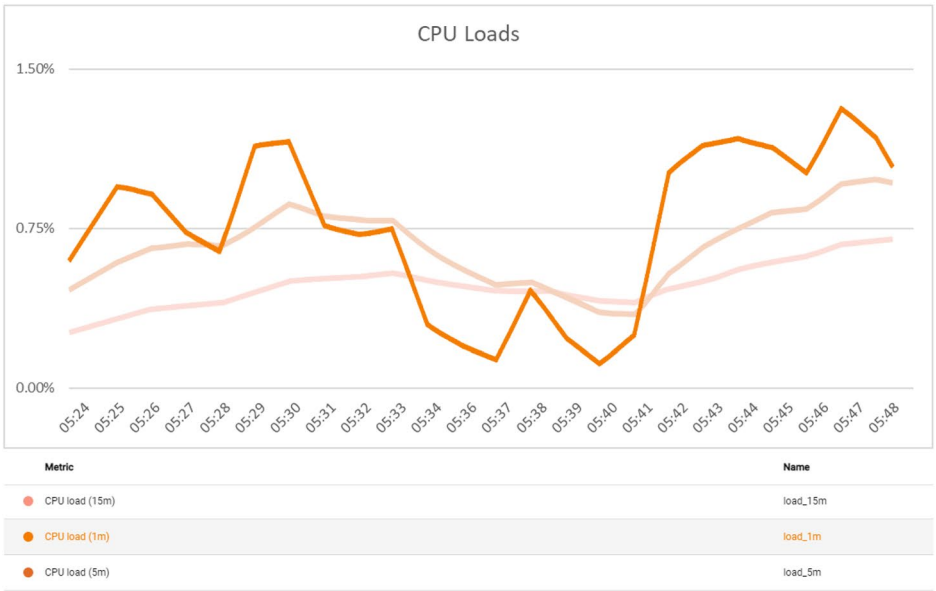
5 Conclusion and future works

This paper set out to present Apache Pulsar as a distributed messaging system for combating pandemics. By evaluating the architecture and general characteristics of the proposed system, it is the right platform for use cases such as spatial analysis of the COVID-19 geo-tagged Twitter datasets. Pulsar is designed to handle large amounts of long-term on-disk persistence.

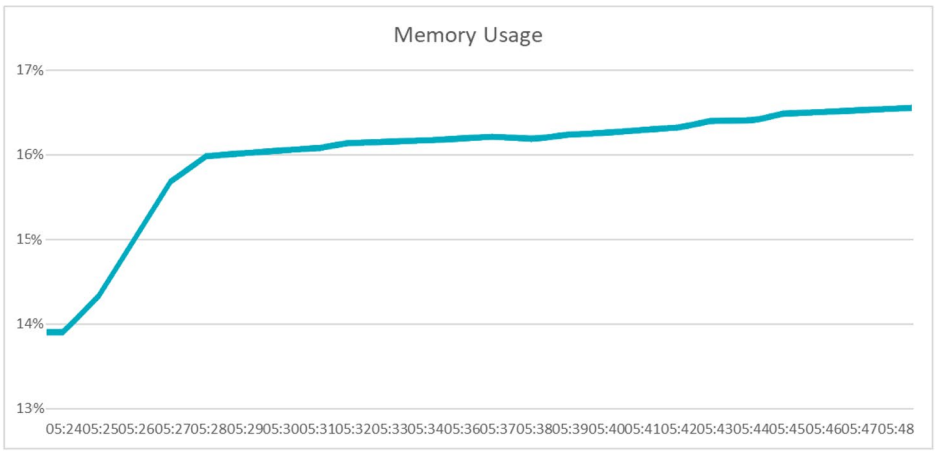
From an industrial point of view, the platforms presented and the conclusions drawn in the study will hopefully help



(a) CPU utilization



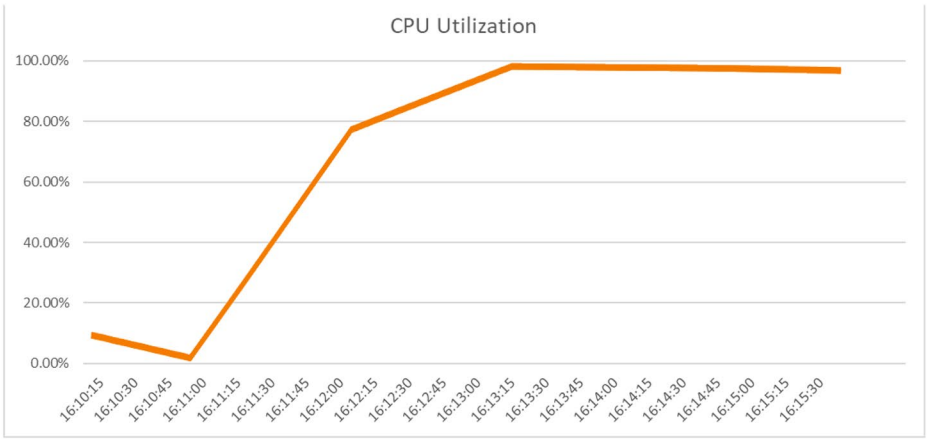
(b) CPU loads



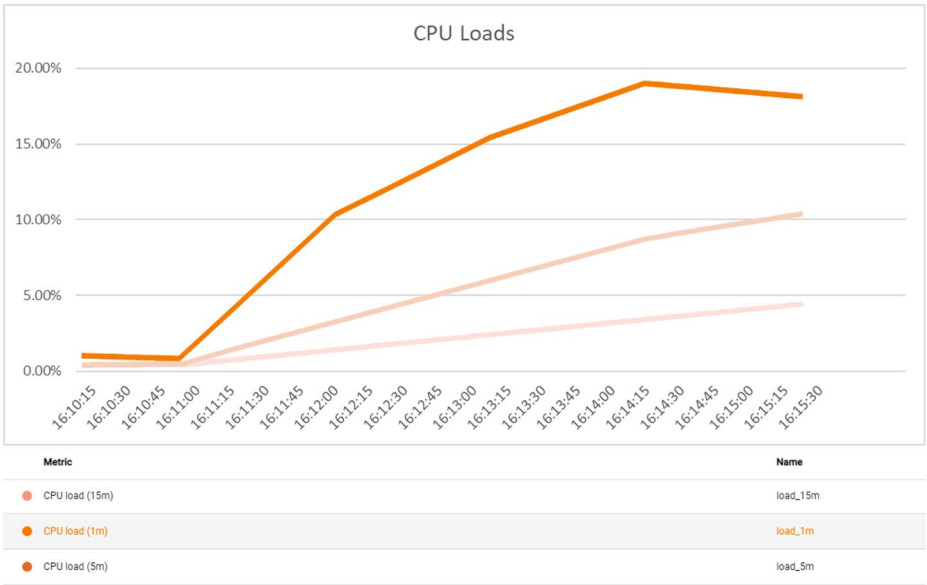
(c) Memory usage

Fig. 7 Performance results for translation task

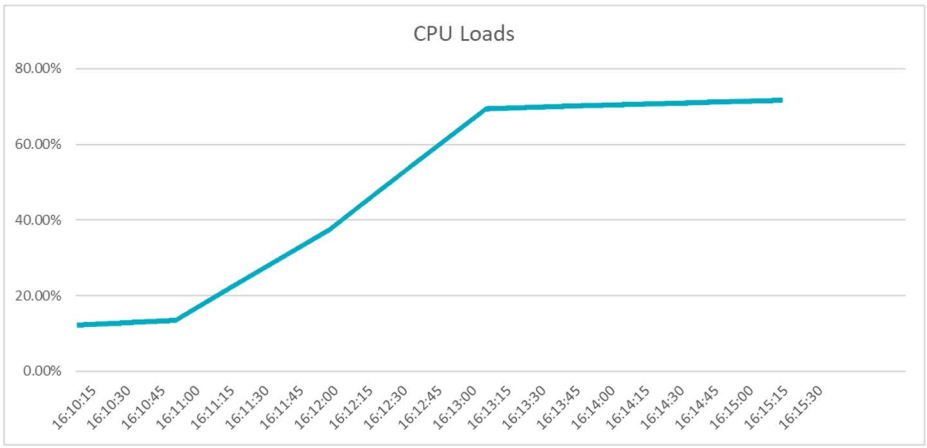
Fig. 8 Performance results for NER task



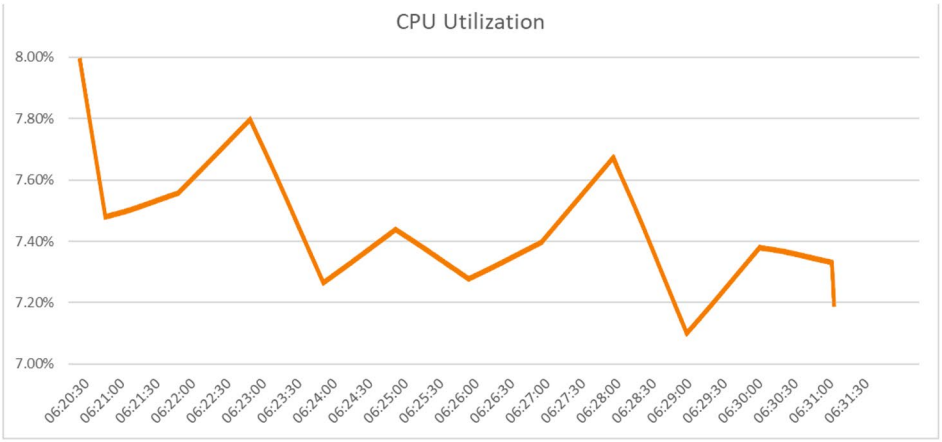
(a) CPU utilization



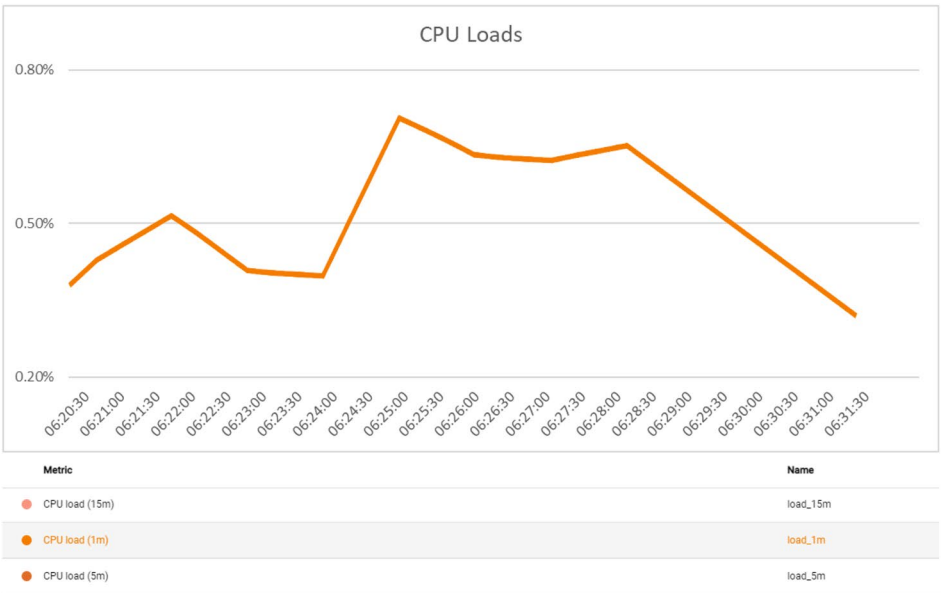
(b) CPU loads



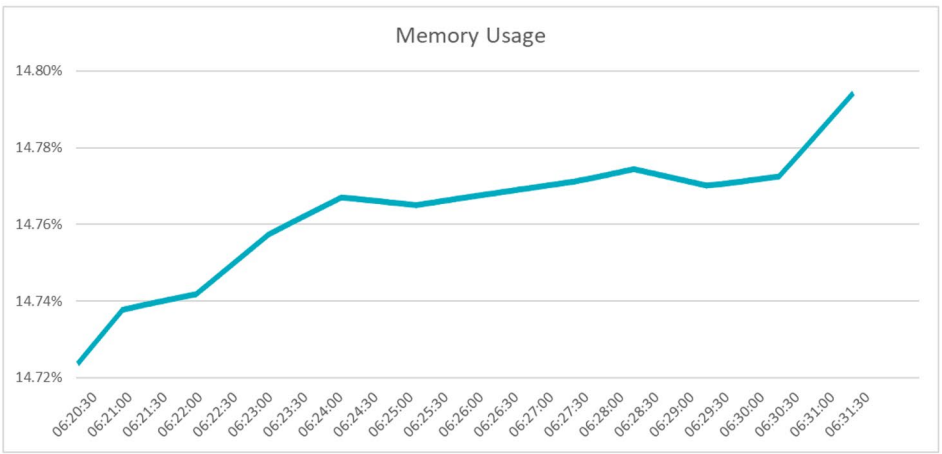
(c) Memory usage



(a) CPU utilization



(b) CPU loads



(c) Memory usage

Fig. 9 Performance results for skyline query task

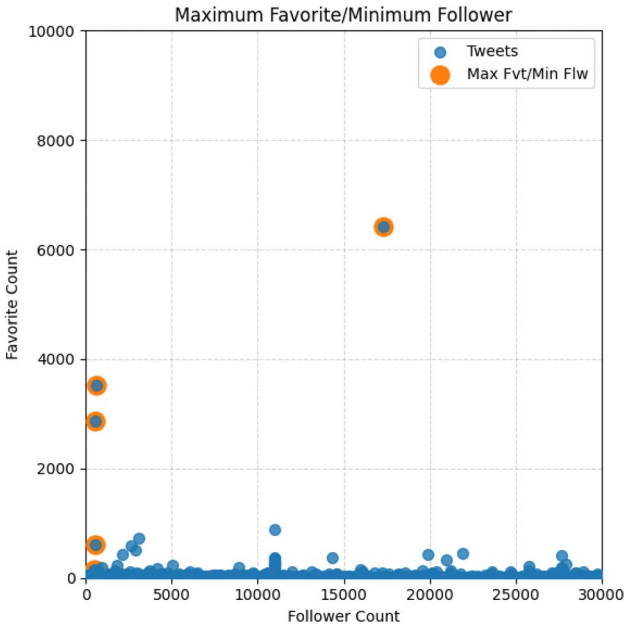


Fig. 10 Objective value space and the Pareto set (efficient solutions with orange)

companies that are in the process of implementing spatial based stream processing pipelines in their product. From a research point of view, the study will work as a base for future research in spatial analysis of different datasets. The study also addresses Apache Pulsar, which currently has very little or no presence in current published research.

For future work, there is still a need to continue doing performance testing using throughput and latency metrics. There are so many different possible combinations in terms of message sizes, throughput levels, hardware, platforms-specific configurations, and etc. that could still be tested. Additionally, there is a need to test the impact that hardware has on overall performance for figuring out the optimal cluster node hardware configurations. For figuring out optimal platform specific configurations, a model for predicting the performance of Pulsar could be further expanded on. Apart from COVID-19, other kind of diseases and well-being, and social networks (Vianna and Barbosa 2020; Vianna et al. 2017) can be used for spatial analysis in distributed way.

Author Contributions SE conceived of the presented idea. YMÖ developed the theory and performed the computations. SE and YMÖ verified the analytical methods. All authors discussed the results and contributed to the final manuscript.

Declarations

Conflicts of interest The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding The author received no financial support for the research, authorship, and/or publication of this article.

Availability of data and material GeoCOV19Tweets dataset https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset

Code availability The codes are available at: https://github.com/yavuzozguven/Distributed-Messaging-and-Light-Streaming-for-Combating-Pandemics

References

Abul-Husn NS, Kenny EE (2019) Personalized medicine and the power of electronic health records. Cell 177(1):58–69

Amaro R, Mulholland A (2020) Biomolecular simulations in the time of covid19, and after. Comput Sci Eng pp 30–36. https://doi.org/10.1109/MCSE.2020.3024155

Anderson DP (2019) Boinc: A platform for volunteer computing. J Grid Comput pp 1–24

Arun M, Baraneetharan E, Kanchana A, Prabu S, et al. (2020) Detection and monitoring of the asymptotic covid-19 patients using iot devices and sensors. Int J Pervasive Comput Commun pp 1–12.https://doi.org/10.1108/IJPCC-08-2020-0107

Bisset KR, Chen J, Feng X, Kumar VA, Marathe MV (2009) Epifast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of the 23rd international conference on supercomputing, pp 430–439

Bisset KR, Chen J, Deodhar S, Feng X, Ma Y, Marathe MV (2014) Indemics: an interactive high-performance computing framework for data-intensive epidemic modeling. ACM Trans Model Comput Simul (TOMACS) 24(1):1–32. https://doi.org/10.1145/2501602

Boberg S, Quandt T, Schatto-Eckrodt T, Frischlich L (2020) Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis. arXiv:200402566

Boulos MNK, Geraghty EM (2020) Geographical tracking and mapping of coronavirus disease covid-19/severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic and associated events around the world: how 21st century gis technologies are supporting the global fight against outbreaks and epidemics. https://doi.org/10.1186/s12942-020-00202-8

Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J (2020) How big data and artificial intelligence can help better manage the covid-19 pandemic. Int J Environ Res Public Health 17(9):3176

Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. Sci Rep 10(1):1–10

Corsi A, de Souza FF, Pagani RN, Kovaleski JL (2020) Big data analytics as a tool for fighting pandemics: a systematic review of literature. J Ambient Intell Hum Comput pp 1–18. https://doi.org/10.1007/s12652-020-02617-4

De Souza PRR, Matteussi KJ, Veith ADS, Zanchetta BF, Leithardt VR, Murciego ÁL, De Freitas EP, Dos Anjos JC, Geyer CF (2020) Boosting big data streaming applications in clouds with burstflow. IEEE Access 8:219124–219136

Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H (2020) The pandemic of social media panic travels faster than the covid-19 outbreak. https://doi.org/10.1093/jtm/taaa031

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:181004805

Eken S (2020a) An exploratory teaching program in big data analysis for undergraduate students. Journal of Ambient Intelligence and Humanized Computing 11(10):4285–4304

Eken S (2020b) A topic-based hierarchical publish/subscribe messaging middleware for covid-19 detection in x-ray image and its metadata. Soft Comput pp 1–11. https://doi.org/10.1007/s00500-020-05387-5

Elmeiligy MA, Desouky AIE, Elghamrawy SM (2020) A multi-dimensional big data storing system for generated covid-19 large-scale data using apache spark. arXiv:200505036

Eltyeb S, Salim N (2014) Chemical named entities recognition: a review on approaches and applications. J Cheminf 6(1):17

Fabret F, Jacobsen HA, Llirbat F, Pereira J, Ross KA, Shasha D (2001) Filtering algorithms and implementation for very fast publish/subscribe systems. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of data, pp 115–126

Fitzgerald RC (2020) Big data is crucial to the early detection of cancer. Nat Med 26(1):19–20

Franch-Pardo I, Napoletano BM, Rosete-Verges F, Billa L (2020) Spatial analysis and gis in the study of covid-19. a review. Sci Total Environ 139:140033. https://doi.org/10.1016/j.scitotenv.2020.140033

Govindan K, Mina H, Alavi B (2020) A decision support system for demand management in healthcare supply chains considering the epidemic outbreaks: a case study of coronavirus disease 2019 (covid-19). Transp Res Part E: Log Transp Rev 138:101967. https://doi.org/10.1016/j.tre.2020.101967

Groppe S (2020) Emergent models, frameworks, and hardware technologies for big data analytics. J Supercomput 76(3):1800–1827. https://doi.org/10.1007/s11227-018-2277-x

Gu J, Han B, Wang J (2020) Covid-19: gastrointestinal manifestations and potential fecal-oral transmission. Gastroenterology 158(6):1518–1519

Häberle M, Werner M, Zhu XX (2019) Geo-spatial text-mining from twitter-a feature space analysis with a view toward building classification in urban regions. Eur J Remote Sens 52(sup2):2–11

Hamzah FB, Lau C, Nazri H, Ligot D, Lee G, Tan C, Shaib M et al (2020) Coronatracker: worldwide covid-19 outbreak data analysis and prediction. Bull World Health Org 1(32):1–31. https://doi.org/10.2471/BLT.20.255695

Harrison TH, Levine DL, Schmidt DC (1997) The design and performance of a real-time corba event service. ACM SIGPLAN Notices 32(10):184–200

Huang X, Li Z, Jiang Y, Ye X, Deng C, Zhang J, Li X (2021) The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the us during the covid-19 pandemic. Int J Digit Earth 14(4):424–442

Hui DS, Azhar EI, Madani TA, Ntoumi F, Kock R, Dar O, Ippolito G, Mchugh TD, Memish ZA, Drosten C et al (2020) The continuing 2019-ncov epidemic threat of novel coronaviruses to global health-the latest 2019 novel coronavirus outbreak in wuhan, china. Int J Infect Dis 91:264–266

Kalyvas C, Tzouramanis T (2017) A survey of skyline query processing. arXiv preprint arXiv:170401788

Kaplan M, Kneifel C, Orlikowski V, Dorff J, Newton M, Howard A, Shinn D, Bishawi M, Chidyagwai S, Balogh P et al (2020) Cloud computing for covid-19: lessons learned from massively parallel models of ventilator splitting. Comput Sci Eng 22(6):37–47

Khan S, Khan A, Maqsood M, Aadil F, Ghazanfar MA (2019) Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. J Grid Comput 17(2):239–254

Khashan EA, Eldesouky AI, Fadel M, Elghamrawy SM (2020) A big data based framework for executing complex query over covid-19 datasets (covid-qf). arXiv preprint arXiv:200512271

Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the bio-entity recognition task at jnlpba. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Citeseer, pp 70–75

Lamsal R (2020) Design and analysis of a large-scale covid-19 tweets dataset. Appl Intell. https://doi.org/10.1007/s10489-020-02029-z

Li C, Weng J, He Q, Yao Y, Datta A, Sun A, Lee BS (2012) Twiner: named entity recognition in targeted twitter stream. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp 721–730

Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao TL, Duan W, Tsoi KKf, Wang FY, (2020) Characterizing the propagation of situational information in social media during covid-19 epidemic: a case study on weibo. IEEE Trans Comput Soc Syst 7(2):556–562

Lin CHA, Berger MS (2020) Advancing neuro-oncology of glial tumors from big data and multidisciplinary studies. J Neurooncol 146(1):1–7

Liu Y, Shen W, Yao Z, Wang J, Yang Z, Yuan X (2020) Named entity location prediction combining twitter and web. IEEE Trans KnowlData Eng. https://doi.org/10.1109/TKDE.2020.2973261

Magesh S, Niveditha V, Rajakumar P, Natrayan L, et al. (2020) Pervasive computing in the context of covid-19 prediction with ai-based algorithms. Int J Pervasive Comput Commun pp 1–11. https://doi.org/10.1108/IJPCC-07-2020-0082

Malecki K, Keating JA (2021) Safdar N (2020) Crisis communication and public perception of covid-19 risk in the era of social media. Clin Infect Dis 72:699–704. https://doi.org/10.1093/cid/ciaa758

Marathe M (2020) High performance simulations to support real-time covid19 response. In: Proceedings of the 2020 ACM SIGSIM conference on principles of advanced discrete simulation, pp 157–157

Melenli S, Topkaya A (2020) Real-time maintaining of social distance in covid-19 environment using image processing and big data. In: 2020 Innovations in intelligent systems and applications conference (ASYU), IEEE, pp 1–5

Minkov E, Wang RC, Cohen W (2005) Extracting personal names from email: Applying named entity recognition to informal text. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp 443–450

Miri SM, Roozbeh F, Omranirad A, Alavian SM (2020) Panic of buying toilet papers: a historical memory or a horrible truth? systematic review of gastrointestinal manifestations of covid-19. Hepat Mon 20(3):e102729. https://doi.org/10.5812/hepatmon.102729

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investig 30(1):3–26

Narin A, Kaya C, Pamuk Z (2021) Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Anal Appl. https://doi.org/10.1007/s10044-021-00984-y

Ophir Y (2018) Coverage of epidemics in American newspapers through the lens of the crisis and emergency risk communication framework. Health Secur 16(3):147–157

Pordes R, Petravick D, Kramer B, Olson D, Livny M, Roy A, Avery P, Blackburn K, Wenaus T, Würthwein F et al (2007) The open science grid. In: Journal of physics: conference series, IOP Publishing, vol 78, p 012057

Quinn P (2018) Crisis communication in public health emergencies: the limits of 'legal control' and the risks for harmful outcomes in a digital age. Life Sci Soc Policy 14(1):4

Sang EF, De Meulder F (2003) Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv:0306050

Sbai M, Taktak H, Moussa F (2020) Towards a ubiquitous real-time covid-19 detection system. Int J Pervasive Comput Commun. https://doi.org/10.1108/IJPCC-07-2020-0087

Shah K, Kamrai D, Mekala H, Mann B, Desai K, Patel RS (2020) Focus on mental health during the coronavirus (covid-19) pandemic: applying learnings from the past outbreaks. Cureus 12(3):e7405. https://doi.org/10.7759/cureus.7405

Shukla AK, Muhuri PK (2019) Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. Eng Appl Artif Intell 77:268–282

Smith M, Smith JC (2020) Repurposing therapeutics for covid-19: supercomputer-based docking to the sars-cov-2 viral spike protein and viral spike protein-human ace2 interface pp 1–28

Sun C, Zhai Z (2020) The efficacy of social distance and ventilation effectiveness in preventing Covid-19 transmission. Sustain City Soc 62:102390. https://doi.org/10.1016/j.scs.2020.10239

Suwinski P, Ong C, Ling MH, Poh YM, Khan AM, Ong HS (2019) Advancing personalized medicine through the application of whole exome sequencing and big data analytics. Front Genet 10:1–16. https://doi.org/10.3389/fgene.2019.00049

Szmuda T, Syed MT, Singh A, Ali S, Özdemir C, Słoniewski P (2020) Youtube as a source of patient information for coronavirus disease (covid-19): a content-quality and audience engagement analysis. Rev Med Virol 30(5):e2132

Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ (2005) Genetag: a tagged corpus for gene/protein named entity recognition. BMC Bioinf 6(S1):S3

Tanenbaum AS, Van Steen M (2007) Distributed systems: principles and paradigms. Prentice-Hall, New York

Tebé C, Valls J, Satorra P, Tobías A (2020) Covid19-world: a shiny application to perform comprehensive country-specific data visualization for sars-cov-2 epidemic. BMC Med Res Methodol 20(1):1–7

Thompson P, Carter J, McNaught J, Ananiadou S (2015) Semantically enhanced search system for historical medical archives. In: 2015 Digital Heritage, IEEE, vol 2, pp 387–390

Vianna HD, Barbosa JLV (2020) Pompilos, a model for augmenting health assistant applications with social media content. J Univers Comput Sci 26(1):4–32

Vianna HD, Barbosa JV, Pittoli F (2017) In the pursuit of hygge software. IEEE Softw 34(06):48–52

Wang CJ, Ng CY, Brook RH (2020) Response to covid-19 in taiwan: big data analytics, new technology, and proactive testing. JAMA 323(14):1341–1342

Wong J, Goh QY, Tan Z, Lie SA, Tay YC, Ng SY, Soh CR (2020) Preparing for a covid-19 pandemic: a review of operating room outbreak response measures in a large tertiary hospital in Singapore. Can J Anesth 67:732–745. https://doi.org/10.1007/s12630-020-01620-9

Wong ZS, Zhou J, Zhang Q (2019) Artificial intelligence for infectious disease big data analytics. Infect Disease Health 24(1):44–48

Wu C, Buyya R (2015) Cloud data centers and cost modeling: a complete guide to planning, designing and building a cloud data center. Morgan Kaufmann, MA

Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, Cohn EL, Hswen Y, Hill SC, Cobo MM et al (2020) Epidemiological data from the covid-19 outbreak, real-time case information. Sci Data 7(1):1–6

Yüce MÖ, Adalı E, Kanmaz B (2020) An analysis of youtube videos as educational resources for dental practitioners to prevent the spread of covid-19. Ir J Med Sci 190:19–26. https://doi.org/10.1007/s11845-020-02312-5

Zhang Y, Zhang J, Ju S, Qiu L (2019) Identifying biomarker candidates of influenza infection based on scalable time-course big data of gene expression. Comput Intell 35(3):610–624

Zhou C, Su F, Pei T, Zhang A, Du Y, Luo B, Cao Z, Wang J, Yuan W, Zhu Y et al (2020) Covid-19: challenges to gis with big data. Geograph ustain 1(1):77–87. https://doi.org/10.1016/j.geosus.2020.03.005