Report on COVID-19 Open Data

Tafara Freddie Hove
u18278150

Abstract

The COVID-19 (coronavirus) disease has affected the world's health care infrastructure and the social, economic, and psychological well-being of humanity globally. The pandemic has become a topical issue being discussed on different media platforms. Hence there is an exponential increase of data from organizations and governments relating to the COVID-19 pandemic. The information can help policy makers and health care systems to combat the the devastating impact of the virus. The aim of this project is to study and analyze the COVID-19 Open Data as Big Data because of its structure, format and size. This dataset cannot be easily analyzed and processed using traditional techniques. The COVID-19 Open Data dataset will be further described using the V's of data collection (variety, veracity) and processing (velocity , volume) categories. In addition a brief overview of the big data architecture will be explored to highlight work to be done in part 2 of this project.

Keywords— COVID-19, Big Data, Characteristics, Categories, Big Data Architecture

1 Introduction

The coronavirus disease 2019 (COVID-19)'s outbreak started in Wuhan, China and rapidly spread worldwide. On March 11, 2020, the World Health Organization (WHO) announced that COVID-19 can be characterized as a pandemic[1]. Many organisations and government agencies have been collecting data relating to the pandemic. Data has been published on websites and social media platforms. The data is used to support researchers and medical professionals to monitor and understand the emergent of the pandemic.

However the exponential increase in COVID-19 literature makes it difficult for researchers to retrieve quality information from very large and complex datasets. Online information is generated in large quantities on daily basis, increasing the size of data exponentially causing processing and data analysis problems. [3] pointed out that processing and analysing huge and complex data, or extracting valuable and quality information from large datasets is a challenging task.

Thus, the aim of this research is to describe the COVID-19 Open Data dataset as Big data. Big data refers to large and complex datasets of enormous size that cannot be stored and processed by conventional resources[8]. The data can be structured, semi-structured or unstructured. [4] argued that the volume, velocity and variety of data is too big making it difficult to store, capture, manage and process using conventional resources. Big data can be found anywhere, anytime and in anyplace which makes it hard to manage and analyse using traditional applications[4]. Such data can be collected from sensors, machines, humans and business processes. Similarly the COVID-19 Open Data is being generated globally through social media and online internet text documents, and it continues to grow unabated.

The remainder of the paper is organized as follows. In Section 2, we elaborate on the COVID-19 Open Data dataset. In Section 3, we discuss the categories and the V's characteristics of Big Data, then in Section 4, we present a big data architecture plan for part 2 of this project. Finally, in Section 5, we present the conclusion.

2 Dataset

The COVID-19 Open Data is a huge dataset that consists of country-level datasets of daily time-series data about COVID-19 worldwide [10]. The repository contains datasets of more than 50 countries around the world for the timespan February 2020 to date. The datasets reveal the impact of the virus and how different countries are responding to the pandemic. It contains the latest available public data on COVID-19 including a daily situation update, the epidemiological curve and the global geographical distribution [10]. The COVID-19 Open Data is available at https://github.com/GoogleCloudPlatform/covid-19-open-data.

The COVID-19 Open Data is drawn from multiple sources including Wikidata, DataCommons, WorldBank, University of Oxford and Google. The data is sourced from different countries in collaboration with the World Health Organisation (WHO). The data is stored in separate CSV and JSON files which are published in Google Cloud Storage. It is also part of the BigQuery Public Dataset Program. The collection of such huge quantities of files constitutes 1.01GB of data.

Since the beginning of the COVID-19 disaster, WHO's Epidemic Intelligence team has been collecting data on daily basis; the number of COVID-19 cases and deaths, based on reports from health authorities worldwide. Hence the dataset is a resource of multiple types of data outcomes(such as cases, deaths, tests), static co-variate data (like population size, GDP, latitude/longitude), dynamic co-variate data (like mobility and weather) and dynamic intervention data (such as government lock-downs regulations) [10].

On the other hand, the COVID-19 crisis also brought in many opportunities to business, academia and research. The data brought in new insights such as understanding, intelligence, knowledge, perspectives and 'actionability' on how to manage and create opportunities out of the crisis? In recent month we witnessed an increase in online trading, intensive research in medical care, virtual meetings and learning. We also expect the data to reveal rapid economic decline, high levels of unemployment and poverty, high demand for medical care and sharp increase in mortality rate especially the elderly and people with underlying conditions.

3 Categories of Big Data and their V's Characteristics

The characteristics of big data are defined by many V's, but in this paper we discuss 4Vs which are volume, velocity, variety and veracity. According to [4] there are five categories of big data which are extracted from the 9Vs. The data categories are collecting, processing, integrity, visualization and worth of data. In this paper we only focus on collecting and processing data categories.

3.0.1 Collecting Data

Data is generated and collected from different sources and types. Such data can be structured, semi-structured or even unstructured. That makes data to heterogeneous and complex. The two V's that make the collecting data category are variety and veracity.

- Variety

  Data can have multiple types and formats such as text, pdf, audio,excel,csv, tweets and images[5]. This reveals diversity of multiple data types. The COVID-19 Open Data is a collection of CSV and JSON files of different datasets, which are generated from multiple sources and locations across the world such as WHO, WorldBank and different countries. Such diversity in the dataset represents Big Data [5]. During the collection period the data was not in the traditional format, it was in multiple formats. The variety component of the COVID-19 data introduces heterogeneity and complexity which could impact on the quality of the dataset. Hence the data are being rearranged and reformatted to make it meaningful and valuable. Hence such variety in the data is defined as big data.

- Veracity

  Big data veracity is the biases, noisy and the abnormality in the data [4]. It involves the assessment of trustworthiness, authenticity, origin and reputation of the data. Data collected from different sources' quality may be compromised hence it must be verified before put into use. Ishwarappa and Anuradha [7] pointed out that when dealing with large data its impossible for all of the data to be 100 percent correct, some data is dirty. Thus, the data might have very high noise accumulation which may cause false correlation resulting in inconsistencies and false discoveries [8]. Similarly, the COVID-19 Open Data can also be viewed as a data in doubt since it is generated by many countries, with some regions not giving credible information. Hence the data's integrity, consistency and completeness must be analysed effectively. Otherwise, the insight discoveries to be leveraged from the data would be compromised by poor quality data. The veracity of the data source guarantees the validity and accuracy of data analysis [7].

3.0.2 Processing Data

Velocity and volume are the two characteristics that define processing of big data. The rate at which data is generated from different sources determines the size of data at a given period.

- Velocity

  Velocity refers to the speed at which data is being generated and moves from one device to another[5]. The data can be generated in real time, online and offline, in streams or batches. For instance thousands and millions of online articles, tweets and facebook massages are uploaded and posted, and they must be processed instantly. COVID-19 data is growing with rapid speed and it is being collected at very short time frames. With thousands and millions Internet devices which are being connected daily that increases both the volume and velocity of data processing. Hence such data movement is almost real time. Such high velocity movement of COVID-19 data represents Big Data [5].

- Volume

  Volume is the amount of data generated at a given period and stored in records, tables or as files[4]. The data size can be defined in megabytes, gigabytes, terabytes or zettabytes. Such voluminous, unstructured and complex datasets must be ingested, analysed and managed to extract valuable insight for decision making. The COVID-19 Data is a global data which is increasing in size tremendously as a result of internet and social networks. The data is 1.01G in size which makes it difficult to process using conventional systems [3]. The dataset requires big data technologies for both storage and processing for researchers to optimize future results [4]. Hence, COVID-19 Open Data is a Google Cloud dataset hosted in both BigQuery and Cloud storage. This allows large volumes of data to be easily accessed and processed using programming models such as Hadoop MapReduce.

4 Planning

The aim of this section is to briefly illustrate how part 2 of this project will be executed using the COVID-19 Open Data dataset. The plan is revealed by the Big Data Architecture. Ross et al [9] proposed that the key to effective enterprise architecture is to identify the processes, data and technologies that can take the operating model from vision to reality. Hence in this project there is need for adequate technology that can handle and analyze voluminous data to extract insights from the datasets to fight the COVID-19 pandemic. We use the plan to highlight how the data will be ingested, processed, stored, managed and accessed.

4.1 Big Data Architecture

The big data architecture of the project will include the following components;

1. Data sources

   The COVID-19 Open Data is extracted from datasets of many countries and organisations across the world. The data from these datasets is available as CSV and JSON files.

2. Data store

   The data is stored in Google Cloud Storage to insure that it an be processed via the big data architecture.

3. Extracting Transforming and loading data

Batch processing will be used since this is static data. Batch processing handles large volume of data efficiently, can filter jobs, aggregate, and prepare data for analysis. Hadoop will be used in batch processing. Other tools include MapReduce, BigQuery and Dataproc

4. Analytical data store

   The deployment strategy will be cloud-based since it is cost effective. The COVID-19 Open Data is a public dataset hosted in the cloud storage hence its freely accessible on the google cloud.

5. Infrastructure sizing

   This will be done on the google cloud sine the conventional machine can to handle such a large dataset. The number of clusters, type of processing memory and size, number of CPUS and cores will be be configured on google cloud engine.

6. Analysis/ reporting tools

   The tools that will be used for data visualization and reporting include Data studio and Python.

5 Conclusion

In this report, we discussed the structure and content of COVID-19 Open Data. The dataset was discussed according to the data collection and processing categories of the 4Vs characteristics of big data. Lastly, we noted that the dataset is a 'big data', it cannot be managed by traditional database management systems, but modern techlonologies such as Hadoop and cloud computing applications can be used. Lastly, big data architecture processing components were discussed give an illustration of future work.

References

[1] Gao Z, Yada S, Wakamiya S, and Aramaki E. (2020).NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset. arXiv:2004.08145v1 [cs.SI].

[2] Wang L.L, Lo K, Yang J, Reas R and Funk k. (2020). COVID-19: The COVID-19 Open Research Dataset.

[3] Patel A.B, Biria M and Nair U. (2012). Addressing Big Data Problem Using Haddop and Map Reduce.

[4] Owais S.S and Hussein N.S. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. IJACSA, vol(7), no. 3.

[5] Johnsosn A, Havinash P.H, Paul V. and Sankaranayanan P.N. (2015). Big Data Processing Using Hadoop MapReduce Programming Model. IJCSIT, vol 6(1),127-132.

[6] Jahanbin k and Rahmanian V. (2020). Using twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine vol(13).

[7] Ishwarappa and Anuradha J. (2015). A Brief Introduction to Big Data 5vs Characteristics and Hadoop Technology. Procedia Computer Science 48, 319-324.

[8] Grolinger K, Hayes M, L'Heureux A, Higashino W.A and Allison D.S. (2014). Challenges for MapReduce in Big Data.

[9] Ross J.W, Weill P, and Robertson. (2006). Enterprise Architecture as Strategy; *Creating a Foundataion for Business Execution.* Harvard Business Press, Boston.

[10] https://github.com/GoogleCloudPlatform/covid-19-open-data