Big Data Processing Using Hadoop and Spark: The Case of Meteorology Data

Eslam Hussein[2], Ronewa Sadiki[2], Yahlieel Jafta[2], Muhammad Mujahid Sungay[2], Olasupo Ajayi[1,2], and Antoine Bagula[1,2(✉)]

[1] ISAT Laboratory, University of the Western Cape, Cape Town 7535, South Africa
abagula@uwc.ac.za

[2] Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa

Abstract. Meteorology is a branch of science which can be leveraged to gain useful insight into many phenomenon that have significant impacts on our daily lives such as weather precipitation, cyclones, thunderstorms, climate change. It is a highly data-driven field that involves large datasets of images captured from both radar and satellite, thus requiring efficient technologies for storing, processing and data mining to find hidden patterns in these datasets. Different big data tools and ecosystems, most of them integrating Hadoop and Spark, have been designed to address big data issues. However, despite its importance, only few works have been done on the application of these tools and ecosystems for solving meteorology issues. This paper proposes and evaluate the performance of a precipitation data processing system that builds upon the Cloudera ecosystem to analyse large datasets of images as a classification problem. The system can be used as a replacement to machine learning techniques when the classification problem consists of finding zones of high, moderate and low precipitations in satellite images.

Keywords: Hadoop · MapReduce · Spark · Hive · Meteorology · Big data

1 Introduction

Meteorology is a branch of science which studies the earth's atmosphere with its physical occurrences [1]. It helps to gain a better understanding of the meteorological related phenomena, such as weather precipitation forecasting, cyclones, thunderstorms and climate changes. Each of these can have significant impacts on our daily lives [2]. For instance, precipitation is considered to be the primary source of fresh water. It plays an important role in industry and agriculture, but when in excess might lead to flooding or related natural disasters. One recent disaster occurred in Mozambique and its neighbouring countries in April 2019, where almost 750 lives were lost to a cyclone in Southern Africa. Figure 1 shows the outcomes of that disasters [3]. Such consequences explain the importance of

R. Zitouni et al. (Eds.): AFRICOMM 2019, LNICST 311, pp. 180–185, 2020.
https://doi.org/10.1007/978-3-030-41593-8_13

Fig. 1. An image of the disaster that occurred in Mozambique [3].

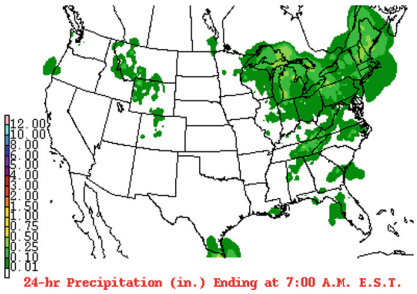

Fig. 2. An image from the National Centers for Environmental Prediction data set

developing an accurate forecasting system to provide early alarms for governments to manage potential disaster(s) [4].

Meteorology is one of the fields that has always been highly data-driven [8], big data analytics can thus find good application in Meteorology [5]. One of the largest databases in the world is data related to climate known as *The World Data Center for Climate* (WDCC), and includes 340 terabytes of earth observations data [6]. A number of research work have utilized different platforms to maintain and analyze these huge data [7]. However, due to its volume and variety, we consider big data platforms to be well suited to taking advantage of the potential value these datasets hold. Hadoop and Spark, are two open source platforms that have been widely used for analyzing big sets of data effectively [7]. This paper also adopts these platforms and proposes a precipitation data processing system built on Cloudera. We consider the issue of finding zones of high, moderate and low precipitations in radar images such as that shown in Fig. 2 as a classification problem. We then employed Hadoop and Spark to analyze the large datasets of images. The system can be used as a replacement or complement to machine learning techniques for classification problems. The rest of this paper is organized as follows: Sect. 2 presents work related to the use of big data platforms in meteorology. Section 3 presents the data analytics of our system while Sect. 4 contains the conclusions and recommendations for future works.

2 Related Work

Recently, there has been a significant number of research work on the application of different Big Data analytics in the meteorology. Ibrahim *et al.* [9], Suggested the use of MapReduce on around 20 GB of whether historical data sets from 1929–2016 (NCDC, GSOD). The dataset files were stored in the Hadoop Distributed File System (HDFS), split and sent to different mappers. The mappers' output where a set of (key, value) pairs, with the station name and date as key, while the value consists of several parameters such as Wind, Precipitation, Temperature etc. The average, max and min of each month, year, and season for each parameter were calculated using the reducer script. In [10], Pandey *et al.* proposed the use of the word count algorithm in Hadoop on a file of text formats for weather forecasting. For data analytics, Riyaz *et al.* in [11], suggested the use of Hadoop MapReduce on a temperature dataset. The mapper function had to find the average temperature associated with place (key). Values such as average, max, min temperature were calculated using the reduce function. Jayanth *et al.* [13], analyzed weather using Spark and ipython for data analytics. Data was transformed into RDDs sequel to which the highest and average precipitation and temperature values for the top ten weather stations were calculated and displayed. In a similar work, Dagade *et al.* [12], computed the average temperature per year per station. An unstructed dataset was used, which required transforming the data into an understandable format using java scripts before uploaded into HDFS. Like these previous works, the objective of this study is to provide an analysis of precipitation data within the Cloudera QuickStart VM environment. Cloudera was chosen as it has Hadoop, HDFS, and Spark integrated.

3 Implementation

3.1 Data

24-Hour-Precipitation-Dataset. Data used for this study are 24-hour-radar images of the United State, from Jan 2012 to Feb 2019 - a total of 2,604 images. The data were sourced from the *National Centers for Environmental Prediction*, with a resolution of 400×320. Each image contains 15 different rainfall intensity level, a sample is shown in Fig. 2.

3.2 Weather Data Analysis: Hadoop MapReduce

Pre-processing was done on the images to structure them into key-value pairs. These pairs serve as input data, (see left side of Fig. 3), to the word count algorithm [10]. They keys are the year, while the value consists of corresponding rain intensities (Light, Moderate and Heavy) in pixel count values. Value extraction from pixels was done using "extcolors 0.1.2" python API.

The implementation is executed in two phases, which are - Map and Reduce. The map function reads each line and extracts the three classified values associating it with its relevant key. The output produced omits the description of

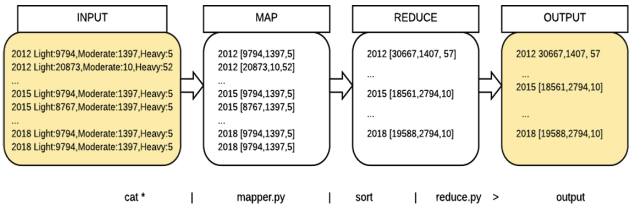

Fig. 3. MapReduce logical flow

the classification values and only displays the data representation, i.e. the rain intensity.

The reduce function reads the output from the map function, groups the keys and values and performs addition of each applicable classification value for each key. This produces an output for each key followed by the three classification values which are summed up for each matching key. The output is used to produce analytics describing the rainfall within the scope of the datasets used.

For our implementation of MapReduce, we made use of the Hadoop Streaming API. This API allows writing of map and reduce functions in several languages and utilizes Unix standard streams as the interface between Hadoop and written programs [14]. This enables us use Python (version 2.7) to read input data and write the corresponding results as output. The results of our analysis are displayed in Fig. 4

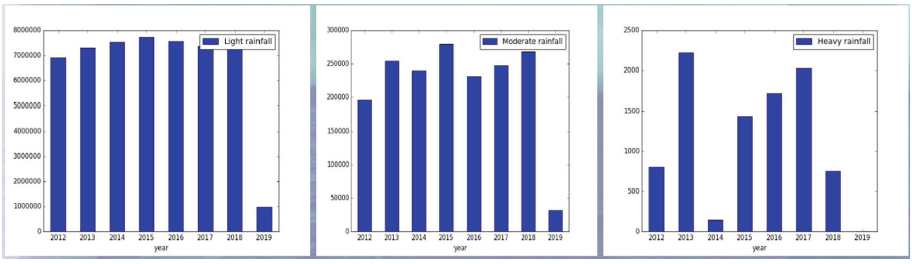

Fig. 4. MapReduce analytics (Daily data set)

3.3 Weather Data Analysis: Spark-Streaming

Analysis in Spark was performed using PySpark, streaming and applying a reduce function to each stream. Each line of the input are added to an RDD queue and streamed. Each queue entry is processed by applying a reducer function which adds all the values per line. To demonstrate the functionality, we used Spark to determine five days with the heaviest rainfall in our dataset. The daily

rainfall dataset was used for this analysis, where each input line represents the rainfall for a day.

The execution flow for Spark is outlined in Fig. 5, while the results of our functionality analysis are displayed in Fig. 6

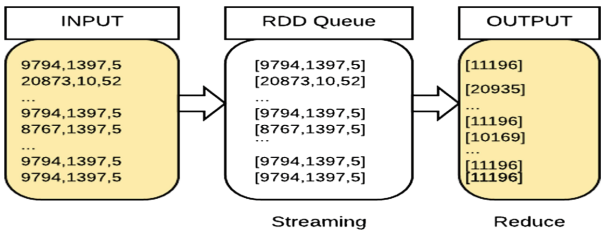

Fig. 5. Spark logical flow

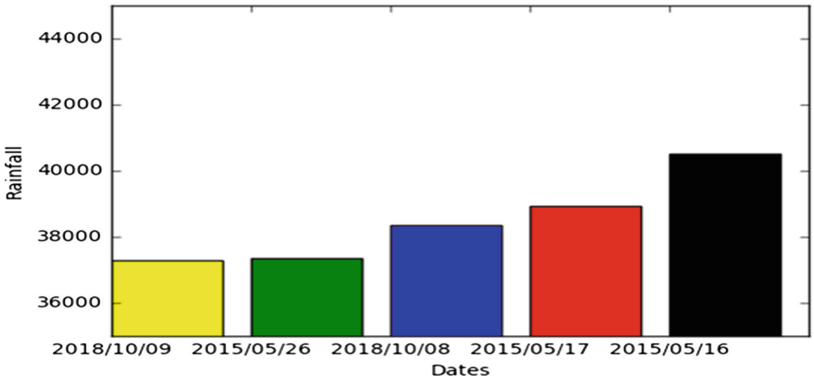

Fig. 6. 5 days with the heaviest rainfall according to Spark streaming

4 Conclusion and Future Work

In this work, the authors have demonstrated two approaches to the processing and analysis of unstructured meteorological image data. The application of Hadoop MapReduce and Spark was applied on a daily precipitation dataset. Using the Hadoop streaming API allowed for the specification of two custom functions, Map and Reduce, which can be written in any language with support for standard read/write of input and output data. Spark allowed for the image data to be analyzed in batches and in memory. We tested Hadoop MapReduce and Spark and both were able to accurately determine the precipitation based

on the image dataset supplied. This result shows that big data analytics tools such as Hadoop MapReduce and Spark can be used as complementary or alternatives to Machine Learning tools. Future research might involve comparing the performance of both approaches Hadoop MapReduce and Spark and possibly benchmarking against known machine learning algorithms.

References

1. GmbH, J.: Joint Aviation Authorities Airline Transport Pilot's Licence Theoretical Knowledge Manual. Oxford Aviation Training (2001)
2. Ahrens, C.D.: Meteorology Today: An Introduction to Weather, Climate, and the Environment. Cengage Learning, Boston (2012)
3. Swails, B., Berlinger, J.: Tropical cyclone kenneth death toll rises to 38 in mozambique, officials say (2019)
4. Shi, E., Li, Q., Gu, D., Zhao, Z.: A method of weather radar echo extrapolation based on convolutional neural networks. In: Schoeffmann, K., et al. (eds.) MMM 2018. LNCS, vol. 10704, pp. 16–28. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73603-7_2
5. Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: a survey. Comput. Electron. Agric. **147**, 70–90 (2018)
6. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K.: Efficient machine learning for big data: a review. Big Data Res. **2**(3), 87–93 (2015)
7. Dagade, V., Lagali, M., Avadhani, S., Kalekar, P.: Big data weather analytics using hadoop. Int. J. Emerg. Technol. Comput. Sci. Electron. (IJETCSE) ISSN, 0976–1353 (2015)
8. Chen, C.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)
9. Ibrahim, G., et al.: Big data techniques: hadoop and mapreduce for weather forecasting. Int. J. Latest Trends Eng. Technol. 194–199 (2016)
10. Pandey, A., Agrawal, C., Agrawal, M.: A hadoop based weather prediction model for classification of weather data. In: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–5. IEEE (2017)
11. Riyaz, P., Varghese, S.M.: Leveraging map reduce with hadoop for weather data analytics. J. Comput. Eng. **17**(3), 6–12 (2015)
12. Oury, D.T.M., Singh, A.: Data analysis of weather data using hadoop technology. In: Satapathy, S.C., Bhateja, V., Das, S. (eds.) Smart Computing and Informatics. SIST, vol. 77, pp. 723–730. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-5544-7_71
13. Jayanthi, D., Sumathi, G.: Weather data analysis using spark-an in-memory computing framework. In: 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1–5. IEEE (2017)
14. White, T.: Hadoop: The Definitive Guide. O'Reilly Media, Inc., Newton (2012)