Hadoop-based System for Analysis Big Social Sensor Data

Van Quan Nguyen*, Linh Van Ma*, and Jinsul Kim**

요 약

요즘 많은 분석 데이터가 많은 업무에서 중요 해지고 있습니다. 데이터는 산업 시스템의 과학, 의학, 기상, 금융, 마케팅 또는 센서 데이터 일 수도 있고 소셜 네트워크의 소셜 데이터 일 수도 있습니다. Hadoop 프레임 워크는 현재 분산 데이터뿐 아니라 대용량 데이터 처리를위한 최상의 선택이되고 있습니다. 이 백서에서는 Hadoop 분산 파일 시스템 (HDFS)에 MapReduce 기반 아키텍처를 배포하여 여러 사용자가 재난에 대해 수집 한 사회적 센서 데이터를 처리합니다. 기상청이 예상하고 예측하고 주민에게 경고문을 게시하려면이 큰 데이터를 실시간으로 수집, 저장 및 처리해야합니다.

Abstract

Nowadays large analysis amount of data has become important for many tasks. Data could be scientific, medical, meteorological, financial, marketing or sensor data from industrial system even social data from the social network. Hadoop framework is currently becoming the best choice for big data processing as well as distributed data. This paper deployed MapReduce based architecture on Hadoop Distributed File System (HDFS) to process social sensor data which is collected from multiple users about the disaster. Real-time collecting, storing and processing of this big data is necessary for the meteorological department to forecast as well as publish the warning to residents.

Key words
Hadoop, MapReduce, Distributed System, Social Data

Ⅰ. Introduction

Hadoop Apache is an open-source framework, Java-based programming framework. It allows to store and process large data sets in a distributed environment [1]. Hadoop has a maximum advantage over scalable and fault-tolerant distributed processing technologies. Also, Hadoop Distributed File System (HDFS) highly faults. In other words, Hadoop enables applications with MapReduce technique [2], in which the data is processed in parallel on different clusters nodes. In other words, a Hadoop-based application could perform analysis for a large amount of data on large clusters of commodity hardware in a reliable.

Big data means really a big data, it is a collection of a large dataset that cannot be processed using traditional computing techniques. Big data is not

* School of Electronics and Computer Engineering, Chonnam National Univ., Yongbong-ro, Buk-gu, Gwangju 500-757
** Corresponding author

merely a data, rather it has become a complete subject, which involves various tools, techniques, and frameworks.

The rest of the paper is organized as follows: Section 2 briefly introduces background about architecture and related works. Section 3 focus on design Hadoop-based system for big data analysis. Then, the configuration and running result are discussed in Section 4. Last, Section 5 is the conclusion.

II. An Overview of Hadoop Ecosystem

Big data is really critical to our life and its emerging as one of the most important technologies in the modern world. Apache Hadoop offers a scalable, flexible and reliable distributed computing big data framework. It can be implemented for a cluster of systems with storage capacity and local computing power by leveraging commodity hardware [1]. Hadoop follows a Master-Slave architecture for the transformation and analysis of large datasets using Hadoop MapReduce paradigm. The most important components of the Hadoop architecture includes Hadoop Distribution File System and Hadoop MapReduce.

Apache Hadoop comes to picture as a solution since it enables to perform on distributed big data. Hadoop Ecosystem is an open-source framework containing various types of complex and evolving tools and components which belonging to four different layers: data storage, data processing, data access, and data management. They may be HDFS, HBase, Hive [3], Sqoop [4], Zoo Keeper, etc.

III. Proposed system

When faced with the massive amount of data, cost computing has become a big challenge for the big data. Hadoop which uses map reduce to maintain and process this data and allow to explore useful information in an efficient manner. In the scope of our problem, we will pick some tools such as Flume, Hive, Sqooq, and HDFS from Hadoop ecosystem to collect and implement event detection algorithm. Several solutions for visualization of analyzed data are discussed in the next part such Zeppelin [5] and X-SCADA [6]..

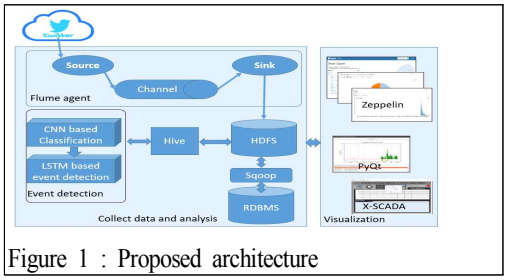


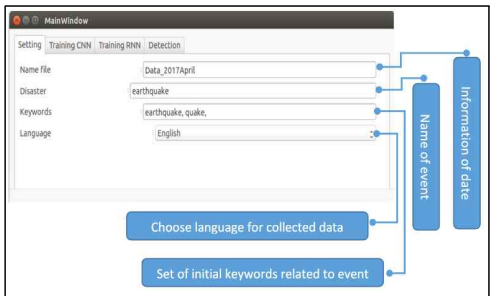
Figure 1 : Proposed architecture

Our system is shown in Figure 1. Social data from the user using the social network is collected via Flume tool [12]. Hive [13] is used as a data warehouse infrastructure to access data, SQL in Hive facilitates in reading, writing and managing large data residing in distributed storage (HDFS). We proposed a Convolution Neural Network (CNN) [7] based method to determine informative data or sorting before moving to LSTM [8] based event detection. Hadoop streaming is a utility that comes with the Hadoop distribution, so we will use this aid to run executable or script as the mapper or reducer for performing classification and event detection. The analyzed social data is then visualized by using Apache Zeppelin dash-board [5], or solution for SCADA based system. Apache Flume [9] is a tool owning ingestion mechanism for collecting aggregating and transporting large amounts of streaming data from various sources. Flume is a highly reliable, distributed, and configurable tool. Hive [3] is a data warehouse infrastructure tool built on top of Hadoop for providing data summarization, query, and process structured data. Since the format of data in HDFS is JSON form, we need to convert them

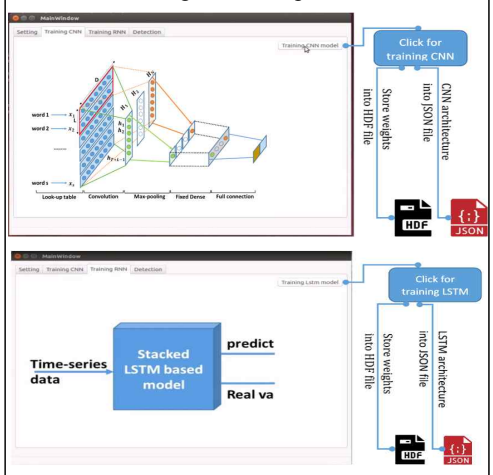
into Hive structure table. Hive SerDe will parse content that loaded from HDFS. Apache Zeppelin could support many interpreters that are widely known as Apache Spark, Python, JDBC, Markdown, and Shell. In experiment section, we use Zeppelin and X-SCADA to display results from the previous analysis which are very useful and informative for management.

IV. Implementation and visualization solutions

4.1 Configuration and training models



a- Configuration of parameters



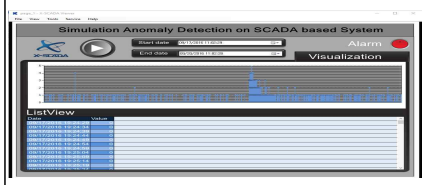
b- Training model

Figure 2: Configuration and training models GUI

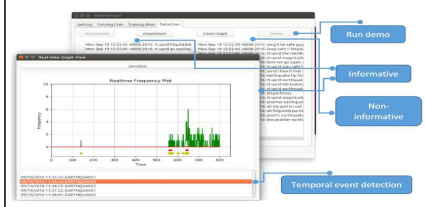
Figure 2 -a is setting tab for configuration some parameters (language, topic) for training phase. CNN and LSTM based models are trained as Figure 2-b, then they are saved under HDF and Json files for usage later.

4.2 Visualization solution and performance

a- Visualization on Zeppelin dashboard



b- Visualization using X-SCADA solution



c- Realtime demonstration with PyQt solution

Figure 3 : Visualization

V. Conclusions

This paper proposed a Hadoop-based system for disasters management using machine learning technique on big data from social network, where CNN model is trained for filtering to obtain informative data, LSTM model is for event detection algorithm.

Acknowledgement

This work (Grants No. C0513295) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2017. Also, this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST)(Grant No. NRF-2017R1D1A1B03034429).

References

[1] http://hadoop.apache.org

[2] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters. Communications of the ACM, (2008), 51(1), 107-113.

[3] https://hive.apache.org/

[4] http://sqoop.apache.org/

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

[6] http://www.xisom.com/en-us/php/home.php

[7] Van Quan Nguyen, Tien Nguyen Anh, Hyung-Jeong Yang, Multi-word Embeddings CNN for Identifying Informative Messages. The 5th International Conference on Big Data Applications and Services (5th BIGDAS). 2017.11, Jeju Island, South Korea.

[8] S. Hochreiter and J. Urgen Schmidhuber, Long Short-Term Memory. Neural Computation, (1997), Vol. 9, No. 8, pp. 1735-1780.

[9] http://flume.apache.org