Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 280–285

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume

Marouane Birjali[a,*], Abderrahim Beni-Hssane[a], Mohammed Erritali[b]

[a]*LAROSERI Laboratory, Department of Computer Sciences, University of Chouaib Doukkali, Faculty of Sciences, El Jadida, Morocco*
[b]*TIAD Laboratory, University of Sultan Moulay Slimane, Faculty of Sciences and Technologies, Béni Mellal, Morocco*

Abstract

Social Media provides organizations ability to survey feelings towards the contents and events associated to them in real time. Moreover, the first demarche of the sentiment analysis is the pre-processing of data collected from Social Media. Most of existing research works that deals with social media analysis based on extracting new features related to sentiment. This paper presents the usage of Twitter in a number of proposed subjects, which is the largest social networking website where Twitter data is in increasing at higher rates every day that considers it as Big Data Source. Then, describing in detail the way in which Big data technology, such as, InfoSphere BigInsights enables processing of this data, which are primarily collected from social networks by Apache Flume and stored in Hadoop storage. In addition, we have investigated a Big Data platform for collecting social media data based on Apache Flume and analyzing this data using InfoSphere BigInsights. Moreover, our paper integrates the visualization of these analysis results using BigSheets. To that end, evaluation through analysis of results confirms that the proposed Big Data platform produces better results in terms of social media analysis.

Keywords: Big Data; Hadoop; Infosphere BigInsights; Social Media Analysis; BigSheets; Twitter Data; Apache Flume.

1. Introduction

Today, the companies face growing challenges from their commercial perspective. In particular, their adding value should be produced from huge amount of data generated and also on the data complexity that can be in structured,

* Corresponding author.
E-mail address: birjali.marouane@gmail.com

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.299

semi-structured or unstructured. During the last years, the Internet has yet seen a wider scope through the development of social media. Based on communication techniques and accessible to all, the media promote social interaction through the Internet. Many social networks exist and there are more than 900 social media sites available on the internet[1]. Millions of people are using Twitter and it is ranked as one of the most visited sites with the average of 58 million tweets per day[2]. Big data is the border of the ability of an enterprise in term of storing, processing and accessing all the data it needs for the effective functioning, and to make decisions, reduce risks, and also to serve the different customers within a more reasonable time[3].

In addition, the first organizations that adopted great data were online and startup companies. Big Data has the ability to reduce costs and substantial improvements in the time needed to perform a spreadsheet task[4]. According to the statistical in the industries, there are 2.5 million items added per minute by each individual. As well as, 300,000 tweets, 220,000 images and 200 million emails generate per minute and several companies as 5TB of RFID and are bigger than that of 1PB for gas turbines with daily production. 2.8 zettabytes of all these data are only now in 2015 and by 2020, this large number can reach 40 zettabytes[5,6,7]. In this world, 90% of unstructured data and is becoming difficult to treat broadcast on for business. To gain the value of large data, another approach is needed to treat[8].

This massive data is considered as a Big Data and can be used for industrial or business purpose after organizing as per the need and processing. This work presents how to analyze data from social networks using BigSheets BigInsights processing[9,10], analysis and visualization using Apache Hadoop Distributed[11,12] and Apache Flume[13] for the collection of this data. Among the many Big Data technologies, Hadoop is popular and more used to meet the challenges of Big Data[14]. There are many different platforms that give Hadoop the spread of their data like Apache Hadoop[12], IBM BigInsights[9,10], Microsoft Azure HD Insights[15], Cloudera[16] and Hortonworks tools[17]. These tools perform data analysis and processing functions depending on the different problem areas[14].

This paper is organized as follows. Related works on the proposed work in section 2 and our methodology of work is presented in section 3. In section 4, the problem statement and methodology in the existing tool is described. Finally, the paper is summarized briefly in section 5.

2. Related works

In this chapter, we describe the related research and the study of social media analysis. The social networks have engaged to attract the attention of the research field, which try to analyze, among others, the private life, the interconnection and the interaction between users. People tend to express their feelings and talk about their activities of daily life through Twitter.

There is a lot of research work on the analysis of feelings, rules-based techniques, bag-of-words and machine-learning methods. Two main research directions of opinion mining operate on either the document level[18,19,20] or the sentence level[21,22,23]. Most methods of classifying the document at the level of sentences are usually based on the identification of terms or phrases of opinion. For this, there are basically two types of methods: (1) lexicon-based methods, and (2) rules-based methods. The treatment of sentiment analysis is part of natural language processing at several levels of granularity. There is a wide range of research work on feel analysis[24], rule-based methods, bag-of-words and machine-learning techniques. Based on a classification task at the level of document by Turney[25], it was processed in the level of sentences by Hu and Liu[26] and more recently at level of sentences by Wilson[27].

The social network like Twitter, on which users post his reactions to and opinions about "everything", is a new and different challenge. Some of the first results and recent analysis of Twitter sentiment data. Two main research areas of mining opinion operate either on the document level[28]. Both classification methods at the document level and at the level of the sentence are generally based on the identification of opinion words or phrases.

However, in this paper, we focus on social media data, on which users post real time reactions to and activities about “everything”, where this data is increasing at high rates every day is considered as Big Data[14]. The processing and analysis of this data is done using InfoSphere BigInsights[9,10], which brings performance power to Hadoop[11,12]. This also includes viewing the results of analysis of large data tables using large sheets and workbooks.

3. Background

In this section, we present the background and detailed description of our research work. A data set is created using social media posts of electronic products. Moreover, we perform a sentiment analysis based on sentence level in three phases. In first is the preprocessing. Then a feature vector is created using relevant features. Finally, we use different functions of BigSheets to classify into positive and negative classes.

3.1. Hadoop framework

Apache Hadoop is High-availability distributed framework that offers a distributed storage system via its HDFS (Hadoop Distributed File System) and processing management system[11,12]. Hadoop provides possibility to store data offers in the duplicating. Therefore, Hadoop does not require to be configured with a RAID system because it becomes useless with Hadoop. On the other hand, Hadoop offers data processing framework on large data volumes called MapReduce[29].

The MapReduce architecture is composed of two phases of phases: the map and reduce phase. Initially, the input data can be divided into several copies as <key, value> where the key is the word and the value indicates how many times the word has occurred and assigns to each underemployment the task trackers. Finally, in the phase interruption, the results of each job tracker are combined to produce the finale results[30].

HDFS is highly fault tolerant, which is designed for low-cost hardware, holds up very large amount of data is stored on multiple machines (Multi-nodes). Some of the important characteristics of HDFS is the storage and processing in a distributed environment, streaming access to data files. At the level of data security, Hadoop provides itself with file permissions and user authentication.

3.2. Apache Flume

Flume was originally developed by Cloudera before being donated to the Apache community[13]. It is now called Flume NG (Next Generation). Flume works as a distributed service for real-time data collection, temporary storage, and delivery to a target[31]. Flume is a highly reliable, distributed, and configurable tool. It is designed to collect streaming data from several web-servers to HDFS. Technically, Flume agent creates routes to connect a source to a target via a Flume channel, as shown in the following figure.

- The source: Flume aims to retrieve messages from different sources, especially log files but also as we will see from Twitter data.
- The Flume channel: is a buffer that stores messages before they are consumed. Memory storage is generally used.
- The Flume target: batch consumes the messages coming from the channel to write them on a destination like HDFS for example.

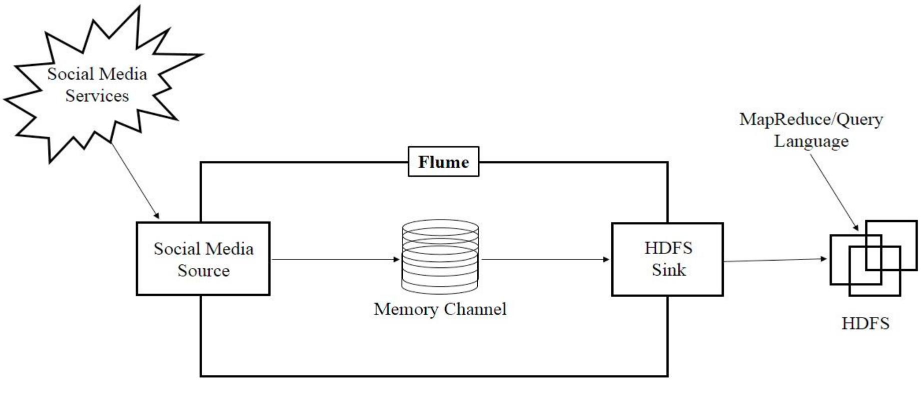
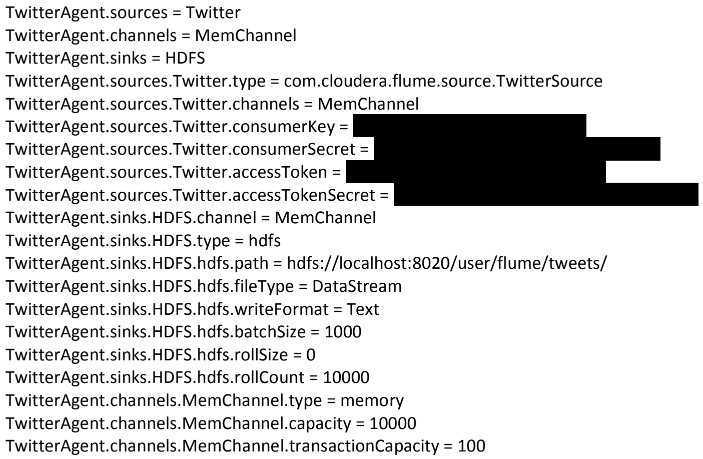


Fig. 1. Apache Flume architecture.

After the creation of the application on the official website of Twitter, we use the key and the secret of the consumer as well as the access token and the secret values. In addition, we can access Twitter and we can collect Tweets as what we want to collect. The following figure is the configuration file we used to collect Tweets from Twitter.



```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = [redacted]
TwitterAgent.sources.Twitter.consumerSecret = [redacted]
TwitterAgent.sources.Twitter.accessToken = [redacted]
TwitterAgent.sources.Twitter.accessTokenSecret = [redacted]
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/flume/tweets/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Fig. 2. Flume configuration files for Twitter data.

3.3. BigInsights InfoSphere

IBM InfoSphere BigInsights is a Hadoop platform that offers new ways to use large volumes of data. In this paper, we describe the most frequently used features of InfoSphere BigInsights, which allow us to analyze large volumes of data from different sources and formats, in order to gather information that they might not have had before[9].

InfoSphere BigInsights provides the capabilities they need to meet the challenges of their business while ensuring maximum compatibility with Hadoop. InfoSphere BigInsights includes a large number of IBM technologies that increase the performance of Hadoop open source software to accelerate return on investment. InfoSphere BigInsights offers a wide range of capabilities that go beyond the Hadoop capabilities, and IBM has chosen an inclusion approach[9,10]. To do this, we quickly start our analysis of the data collected from social media in a Big Data environment. InfoSphere BigInsights has been the subject of several improvements:

- Accelerating deployments with innovations from the Hadoop community
- Using Existing SQL Skills and Solutions
- Enabling user-oriented analysis and data provisioning

4. Problem Statement and Methodology

4.1. Existing tools

As mentioned above, small set of social media data can be downloaded and easily treated through traditional databases. Procedure using ancient techniques stream and analyze the raw data. However, the data can be huge in quantity and unstructured raw data which traditional databases cannot handle, process and analyze. This is a wide problem in the distribution and processing of large volumes of data in Big Data Sources in real time. Indeed, this problem can be initiated by creating a dashboard to monitor the traffic of feelings in Twitter.

This article presents how to overcome the limitations of traditional techniques using the Hadoop ecosystem to streamline the processing of data from large clusters. The tweets data are then analyzed from Flume, processed using the Jaql script and stored in HDFS and analyzed, negative words are positive using methods for MapReduce and finally it returns the result id. Tweet, and then displays the results as graphics using the BigSheets BigInsights tool[32].

4.2. Methodology

In this paper, we will solve the question of analyzing large data from social networks using Hadoop to simplify data processing. We implemented the Apache Hadoop platform and IBM BigInsights for easy analysis of Big Data issues. In this perspective, analysis of tweets involves moving on to the following steps to overcome problems in the traditional system. The data collected is unstructured format. The script JAQL[33] is used to extract important data, transforming them into a simpler structure to convert the delimited file by commas, and then storing the data in HDFS to perform the processing using MapReduce[30]. For the graphical visualizations and final data manipulations, we use BigSheets. BigSheets is a spreadsheet-style tool provided with BigInsights InfoSphere to allow standard spreadsheets functions, filter data, join tables, sort data, and visualize data in graphs[32]. The Figure 3 above shows total percentage coverage by languages. To achieve this result, we just group data from the Tweets data and provide the total count of tweets in every group.

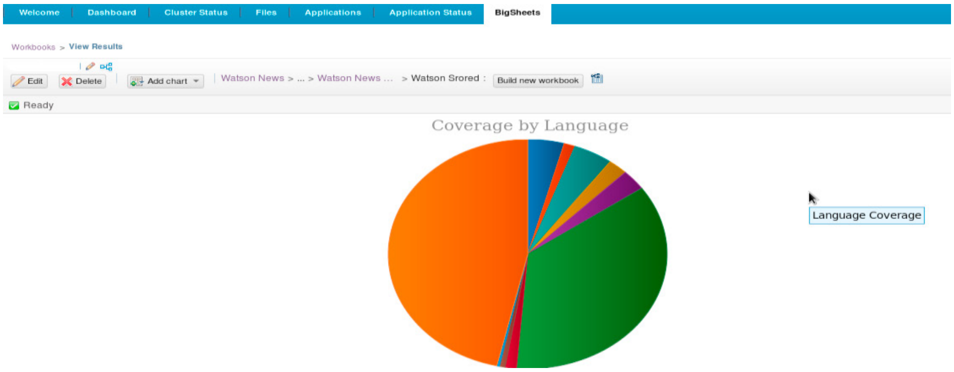


Fig. 3. Coverage by language in a pie chart.

The Figure 4 present the BigSheets Twitter Analysis – Number of tweets during the time, and for the figure 5 depicts a tag cloud chart we generated for the words. As with any BigSheets tag cloud, larger font indicates more occurrences of the data value and scrolling over a data value reveals the number of times it occurred in the collection.

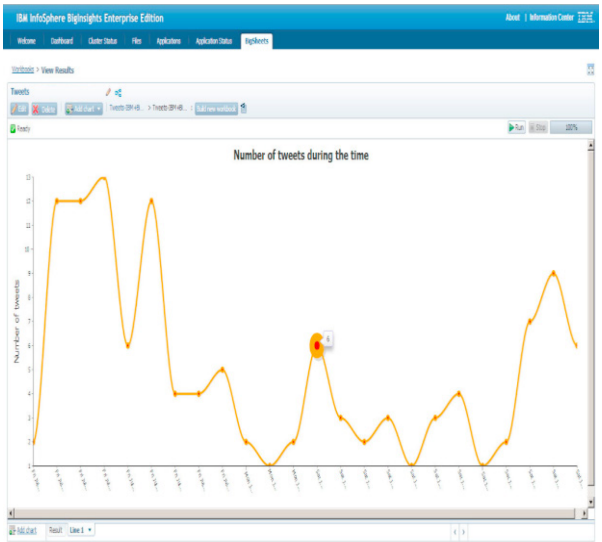


Fig. 4. Visualization by BigSheets Twitter Analysis for Number of tweets during the time.

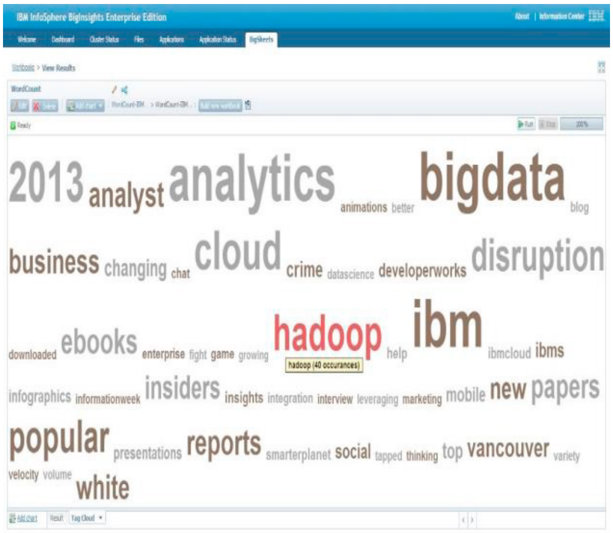


Fig. 5. Visualization by BigSheets Twitter Analysis for Tag cloud.

5. Conclusion

As part of this work, we present a way of collecting social media data using Apache Flume, analyzing and visualizing the twitter data using BigInsights InfoSphere. This platform is not only applicable for streaming, processing, analyzing, and visualizing the twitter data but also to enhanced to apply other types of Big Data from various sources. This paper shows that processing time for analysis of massive twitter data by using the proposed work when compared to other traditional processing methods for Big Data.

References

1. Statistic Brain. Twitter Statistics, Retrieved from http://www.statisticbrain.com/ twitter-statistics/, 2014.
2. R. Li, K. H. Lei,R. Khadiwala, Chang. TEDAS: A Twitter-based Event Detection and Analysis System. icde, pp.1273-1276, 2012 IEEE 28th International Conference on Data Engineering, 2012.
3. Peter Lake, Paul Crowther. Concise Guide to Databases: A Practical Introduction. Springer-Verlag London 2013.
4. Thomas H. Davenport. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press,
5. D. Terrana, A. Augello, and G. Pilato. Automatic unsupervised polarity detection on a Twitter data stream. in Proc. IEEE Int. Conf. Semantic Comput., Newport Beach, CA, USA, Sep. 2014, pp. 128–134.
6. http://en.wikipedia.org/wiki/Twitter
7. http://blog.Twitter.com/2014/the-2014-yearontwitter
8. O'Reilly Radar Team, Planning for Big data, A CIO's Handbook to changing the Data Landscape.
9. Miloš Popović, Milan Milosavljević, Pavle Dakić. Twitter data analytics in education Using ibm infosphere biginsights. The Internet and Development Perspectives, International Scientific Conference On ICT And E-Business Related Research, sinteza 2016.
10. https://www.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bi_qse.html
11. K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The Hadoop Distributed File System. The 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.
12. Apache Hadoop. https://hadoop.apache.org/.
13. Deepak Vohra. Practical Hadoop Ecosystem. Chapter Apache Flume, pp 287-300, September 2016.
14. Rodríguez M., L., Rodríguez E., CA., Sánchez C., J.L. et al. J Supercomput (2016) 72: 3073. doi:10.1007/s11227-015-1501-1
15. Marshall C., Julian S., Anthony P., Mike M., David G., "Overview of Microsoft Azure Services", Microsoft Azure, Part 1, 2015
16. Cloudera. http://www.cloudera.com/.
17. Hortonworks. http://hortonworks.com/.
18. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
19. P. Turney. Thumbs Up or Thumbs Down. Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL'02, 2002.
20. K. Dave, S. Lawrence, and D. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. 2003.
21. M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text. IDA'2005.
22. M. Hu and B. Liu. Mining and summarizing customer reviews. KDD'04, 2004.
23. S. Kim and E. Hovy. Determining the Sentiment of Opinions. COLING'04, 2004.
24. G. Vinodhini and R. Chandrasekaran. Sentiment analysis and opinion mining: A survey. International Journal, vol. 2, no. 6, 2012.
25. P. Turney. Thumbs up or thumbs down Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association for Computational Linguistics.
26. Hu M. and Liu B. Mining and Summarizing Customer Reviews. KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 168-177.
27. Wilson T., Wiebe J. and Hoffmann P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In the Advanced Research and Development Activity (ARDA).
28. Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. Structural opinion mining for graph-based sentiment representation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011). 2011.
29. Ilkyu Ha, Bonghyun Back, and Byoungchul Ahn. MapReduce Functions to Analyze Sentiment Information from Social Big Data. Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2015, Article ID 417502, 11 pagesn http://dx.doi.org/10.1155/2015/417502
30. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Google, Inc.
31. P.B. Makeshwar, A. Kalra, N.S. Rajput, K.P. Singh. Computational Scalability with Apache Flume and Mahout for Large Scale Round the Clock Analysis of Sensor Network Data. National Conference on Recent Advances in Electronics & Computer Engineering, 2015.
32. https://www.ibm.com/analytics/us/en/technology/hadoop/bigsheets/
33. K.S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Eltabakh, C.-C. Kanne, F. Ozcan, E.J. Shekita, Jaql: a scripting language for large scale semistructured data analysis, Proc. VLDB Conf. (2011).