News Feed Processing and Analysis using Hadoop Framework

Pankaj Karpe
Computer Engineering,
University of Pune
G.H. Raisoni College of Engineering & Management,
Domkhel Road,
Wagholi, Pune-412207,
Maharashtra, India.

Vijay Bhor
Computer Engineering,
University of Pune
G.H. Raisoni College of Engineering & Management,
Domkhel Road,
Wagholi, Pune-412207,
Maharashtra, India.

Chetana Agarwal
Computer Engineering,
University of Pune
G.H. Raisoni College of Engineering & Management,
Domkhel Road,
Wagholi, Pune-412207,
Maharashtra, India.

ABSTRACT
This paper presents News Feed Processing and Analysis using Hadoop that automatically group's news related to the same topics published in different newspapers on different days according to geographical regions i.e. spatial analysis. Grouping the titles of the news feeds selected by the user, it is possible to identify sets of related news on the basis of syntactic and lexical similarity. The user may tune some parameters in order to improve the grouping results. The exploration is performed on the data collected with Indian Media Monitor (IMM), a system which monitors over 2500 online sources and processes 90,000 articles per day. By analyzing the news feeds, we want to find out which topics are important in different countries. In the special description of the news feeds, every article can be represented by two geographic attributes, the news origin and the location of the event itself. In order to assess these spatial properties of news articles, we conducted our geo-analysis, which is able to cope with the size and spatial distribution of the data. Within this application framework, we show opportunities how real-time news feed data can be analyzed efficiently.

Keywords
News feed analysis, Spatiotemporal Analysis, HDFS, Map Reduce, data mining, heterogeneity.

1. INTRODUCTION
Nowadays Excess amount of information is generated each day on the internet, making processing of the content very difficult for the individual. Global news agencies, such as Press Trust of India, Indo-Asian News Service, Samachar Bharti, Hindusthan Samachar provide media companies with news reports from all over the country as well as from the world. This content is then duplicated, enriched with commentary and opinion. Additionally, news is filtered according to importance or interest of the editorial team. Besides, local media outlets produce their own local (or global) content having their own point of view, which might be specific to the geographic location of the news source (region, country) or specific to a certain group of people. Furthermore, blogs allow common people to become active content creators themselves, not just passive readers, thus making the analysis of such amount of information one of today's greatest challenges. Internet newspapers may update their contents frequently: thus there is not a daily issue but the news are continuously updated and published. As a consequence, hundreds of thousands of partially overlapping news are daily published. The amount of information daily published is so wide that is unimaginable for a user. On the other hand, the availability of news generates new updated information needs for people.

The current paper describes an application aiming to conduct comprehensive analysis of such material. The paper first describes where the data comes from and how it is processed for analytic purposes. Second, opportunities for in-depth analysis are shown, taking spatial analytic techniques as examples. The goal is to provide assistance to human media monitoring, through automatic analysis and categorization of articles from these sources. In a typical information gathering scenario, journalists try to give the answers to the"Five Ws" questions -"who, what, when, where and why". The application employs various information extraction, clustering and analysis techniques to help the user in answering these questions.

Articles are clustered by the Hadoop framework system in each language into stories that report about the same event. Each article is enriched with various metadata, such as people, their titles and organizations which are mentioned in the articles. This data is extracted in a separate entity recognition process and is available in all languages.

2. RELATED WORK
Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality." Currently, Big Data processing mainly depends on parallel programming models like Map Reduce, as well as providing a cloud computing platform of Big Data services for the public. Map Reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of Map Reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with Map Reduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To

improve the efficiency of algorithms, Chu et al. proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple Map Reduce programming model on multicore processors. Ten classical data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vector machines, the independent variable analysis, Gaussian discriminant analysis, expectation maximization, and back-propagation neural networks.

3. . LITERATURE SURVEY

Sr. No.	Title of paper	Author	Year of publication	Discussed Issues
1	Data Mining with Big Data	Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE	2014	This paper characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective
2	BIG DATA ANALYSIS USING HACE THEOREM	Deepak S. Tamhane, Sultana N. Sayyad	2015	This paper model contains demand-driven aggregation of data sources, mining and study, user knowledge modelling, and security and privacy issues.
3	Visual sentiment analysis of news feeds featuring the U.S presidential election	Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D., and Keim, D. A.	2009	In Workshop on Visual Interfaces to the Social and the Semantic Web
4	Density equalizing distortion of large geographic point sets	Bak, P., Mansmann, F., Janetzko, H., and Keim, D. A.	2009	In J. of Cartographic and Geographic Information Science, volume 36(3).

4. NEWS FEED CHARACTERIZATION

In this section, we present a systematic characterization of existing news feed system, ranging from their installation, activation, collecting , processing, analysis on large amount of data.

4.1 Information Collection

Large amount of information is generated day by day on the internet, making processing of the content very difficult for the individual because of data might be generated from different Locations, different format such as text , audio, videos etc. so Global news agencies, such as Press Trust of India, Indo-Asian News Service, Samachar Bharti, Hindusthan Samachar provide media companies and social media sites such as Facebook, twitter provides news reports from all over the country as well as from the world.

4.2 Information Processing

All the collected news information will be classified based on different categories such as news invented from Locations, Origin, Bollywood, Educational, Cultural, Sports, Politics, Inventions etc. as well as mining of famous magazines.

4.3 Information Analysis

The top three news headlines will be based on maximum clicks by the user on the category of news. This will help in Identifing the user's area of interest among different categories which helps the user to access the news in quick manner. Further news will be regular news which are frequently updated after particular time of interval. News will contain the text, images, audio, videos etc. Weekly magazines will also be published. Advertisements will be flashed alongside realted to the news which is being viewd .

5. PROPOSED SYSTEM

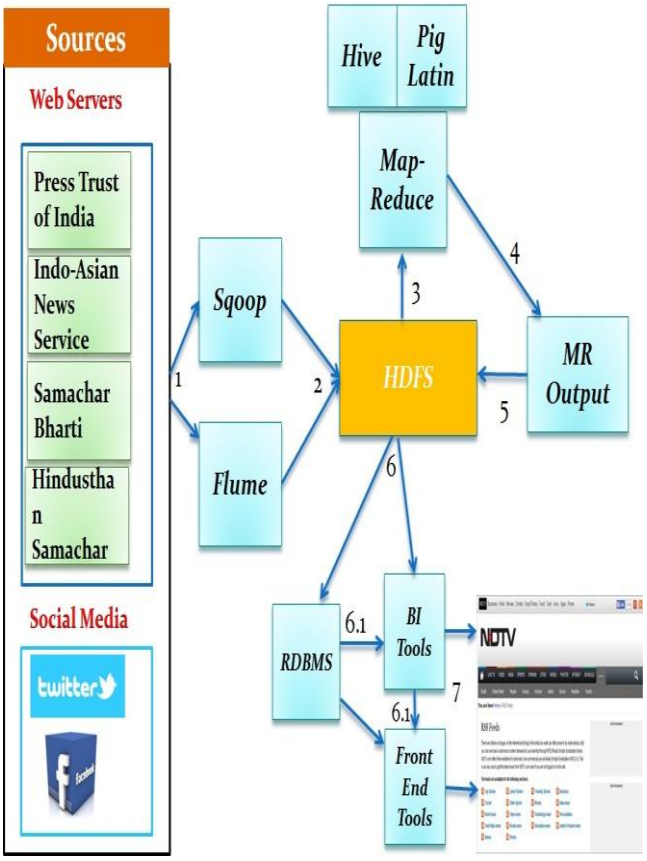

Fig.1 System Architecture

Diagram for system architecture is given above . The working of architechture is explained below.

1. Importing data from sources using Sqoop/Flume into HDFS source like Web Server, Database Server, Newsfeed source. Ex. News websites which provide an API to collect information, news from social media like Facebook, Twitter, local news collecting System.
2. Running MR Technique on data using Hive/ Pig Latin,

   Map reduce processing helps to classify the news in different categories such as Location-oriented,national-international ,Political, educational, sports ,entertainment, buisness etc.
3. After Processing intermediate output after completing MR Technique.
4. Storing the output/result into HDFS.
5. Exporting data from HDFS to RDBMS using Sqoop/Flume.
6. Viewing the results in pictorial/pie chart format using BI tools i.e BI Tools like Tablue.

Our system begins with the phase of extracting the information which is generated each day on and processing the content to make it easy on eyes . Global news agencies, such as Press Trust of India, Indo-Asian News Service, Samachar Bharti, Hindusthan Samachar provide media companies with news reports from all over the country as well as from the world. This content is then duplicated, enriched

with commentary and opinion. Additionally, news is filtered according to importance or interest of the editorial team. Besides, local media outlets produce their own local (or global) contents pertaining their own point of view, which might be specific to the geographic location of the news source (region, country) or specific to a certain group of people.

6. CONCLUSION

Proposed system would leverage the capacity of existing system such as growth in the amount of information to be fetched that is generated each day on different Global news agencies and social media sites. Using tools sqoop and flume to fetch and store the data in data warehouse. BI tools are used to visualize the report in the form of online news paper after a particular interval of time. Recommendation based on users area of interest, weekly e-magazines, video links or videos if available are attached along for captivating news reading. Advertisements realted to stream of news is flashed.News generation in various local languages can be considered as future progess of the application to make it more common among common people.

7. ACKNOWLEDGMENTS

We would like to thank our guide and various technological experts who researched about Big Data Analysis system and improve the result by implementing new methods. We would also like to thank Google for providing details on different issues on Big Data and about other related techniques.

8. REFERENCES

[1] Data Mining with Big Data , Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE –January 2014.

[2] Big Data Analysis Using HACE Theorem, Deepak S. Tamhane, Sultana N. Sayyad – January 2015.

[3] A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In Williamson et al. [11], pages 271–280.

[4] S. Bergamaschi, F. Guerra, M. Orsini, and C. Sartori. Extracting relevant attribute values for improved search. IEEE Internet Computing, pages 26–35, Sep-Oct 2007.

[5] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135.

[6] Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D., and Keim, D. A. (2009). Visual sentiment analysis of rss news feeds featuring the U.S presidential election in 2008. In Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW 2009).

[7] Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In WWW '05: Proc. 14th int. conference on World Wide Web, pages 342–351. ACM.

[8] Bak, P., Mansmann, F., Janetzko, H., and Keim, D. A. (2009b). Density equalizing distortion of large geographic point sets. In J. of Cartographic and Geographic Information Science, volume 36(3).

[9] Our abstraction is inspired by the map and reduce primitives present in Lisp and many other functional languages." -"MapReduce: Simplified Data Processing on Large Clusters", by Jeffrey Dean and Sanjay Ghemawat;

[10] *DavidDeWitt; Michael Stonebraker. "MapReduce: A major step backwards". craig-henderson.blogspot.com. Retrieved 2008-08-2*

[11] "News Analytics on www.eventstudytools.com". Newsanalytics.net. Retrieved 2015-07-26