Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Parallelized Contextual Valance Shifter Algorithm for Sentiment Analysis on Big Data

**[1]Jenie Arock X, [2]Vaitheeswaran G, [3]Dr. L. Arockiam**

[1] M.Phil Scholar, [2] Ph.D Scholar, [3] Associate Professor

[1, 2, 3] Department of Computer Science, St. Joseph's College (Autonomous),
Tiruchirappalli, Tamil Nadu, India

***Abstract*— Popular and boundless use of World Wide Web leads to generate a huge amount of unstructured data called, 'Big Data'. Such data are rich in knowledge that plays a vital role in the process of decision- making. Knowledge can also be extracted in the form of sentiment polarity by the analytical process, Sentiment Analysis. Handling big data in the process of sentiment analysis demands big data technology like Hadoop framework: HDFS and MapReduce paradigm. The paper designs a framework that blends Sentiment Analysis process and Hadoop framework. Time- efficient process is possible by supporting Hadoop framework that involves parallel processing through MapReduce paradigm. Also, the paper proposes an algorithm, PCVS (Parallelized Contextual Valance Shifter) that does the process of Sentiment Analysis and focuses on the Contextual Valance Shifter: negations, intensifiers and diminishers.***

***Keywords*— Big Data, Sentiment Analysis, Hadoop, HDFS, MapReduce, Contextual Valance Shifter**

## I. INTRODUCTION

Multifarious data are generated constantly each day in this digital world through various sources like web, social media, remote-sensing devices and medical records etc. The data may be in the form of news articles, blogs, tweets, movie reviews, and product reviews etc. and hold high threshold of Big Data characteristics. Big data is a collection of massive and complex data sets which include data characterized by the dimensions "The three V's".

- Volume - The size of data (terabytes and petabytes).
- Velocity - Used when streaming. The role of time is very critical.
- Variety - Includes structured and unstructured data.

Unstructured textual data generated by common people in the form of reviews are highly rich in sentiment knowledge. Sentiment Analysis is the process of extraction of sentiment knowledge or opinions expressed in structured/unstructured text written in natural language. Sentiment analysis can be carried out by either Semantic Approach or by Machine Learning. Semantic approach is purely Lexicon based. Lexicons are ensemble of sentiment words, phrases, idioms with polarity/valance values that express sentiment as positive or negative. Based on the approach in the process of building, Lexicons are classified as:

- Manual – compiled by intensive labour.
- Dictionary-based – based on dictionaries and are domain independent.
- Corpus-based – based on context and are domain dependent.

The process of Sentiment analysis involves four levels:

- Word level - Sentiment polarity of each and every word.
- Sentence level - Each sentence decides that polarity.
- Document level - Polarity of the document as a whole.
- Aspect or Entity level - Sentiment polarity of entities and/or aspect of those entities.

Handling of Big Data in the process of Sentiment Analysis is a tedious job and is highly impossible by traditional technologies. Big Data technology like Hadoop, an open source framework is best suitable for big data analysis, which involves HDFS (Hadoop Distributed File System) and MapReduce for distributed storage and parallel processing respectively.

The paper subjects large amount of review data set generated by public through internet in the process of Sentiment Analysis that involves semantic approach based on corpus-base lexicons. The process is carried out on designed framework which incorporates Hadoop framework, a Big Data Technology. Measuring the depth of sentiment, which mostly rely on the Contextual Valance Shifter, is one of the major issues and challenges involved in the process of sentiment analysis. The proposed algorithm mainly focuses on Contextual Valance Shifters.

## II. RELATED WORKS

SINTEF (Stiftelsen for Industriell Og Teknisk Forskning*)*, one of the largest independent research organizations, has proved that 90% of Big Data has reached massive growth and fame only within last five years through world wide web and also adds that social sites play a very important role in the growth [15]. Min Chen et al. [10], discusses various definitions of various organizations in his survey. Organizations such as NIST (National Institute of Standards and Technology), focused the technological aspect to define big data. Apache Hadoop has been defined as big data based on its management and processing capabilities. Gartner has defined big data based on challenges and opportunities of 3Vs model. IDC (International Data Corporation), also defined 4Vs model. Avita Katal et al. [3], briefly explains characteristics of big data in terms of data's variety, volume, velocity, variability, complexity and value. These characters decide the nature of data to be categorized into big data. Xindong Wu et al. [19] distinguish big data in his novel terminology, HACE Theorem. The researcher has termed the theorem in view of heterogeneity, autonomy, complexity and evolution of data. Karthik Kambatlaa[8] et al. in their discussion on the technologies and tools concludes that Hadoop is the best suited technology for handling big data. Hadoop handles data with its main components such as HDFS and MapReduce programming paradigm, which supports distributed storage and parallel processing respectively. Parth Chandarana et al. [13] explore various analytical frameworks that support handling of big data by overcoming its issues and challenges. The paper deals with framework such as Apache Hadoop, Project Storm and Apache Drill. Hadoop is a framework that supports distributed storage and parallel processing. Project Storm and Apache Drill are frameworks similar to Hadoop.

Harshawardhan et al. [6], further explain that Hadoop is an open source framework which supports distributed processing of large data sets across clusters using simple programming models. Hadoop includes the following modules: Hadoop core (support other modules), Hadoop distributed file system (storage of large data), Hadoop YARN (job scheduling and resource management), Hadoop Map Reduce (parallel processing of large data). Dipak M. Durgude et al. [5], picturize the components, working model and techniques involved in the processing of the MapReduce paradigm. Name Node (manages HDFS), Data Node (stores blocks of HDFS), Job Tracker (schedules, allocates and monitors job), and Task Tracker (runs Map Reduce operations) are the components of MapReduce. The MapReduce paradigm involves 2 main jobs: the mapper and reducer. The reducer involves the process of shuffle, sort and reduce. Yang Zhang et al. [20], discusse efficiency of HDFS (Hadoop Distributed File System) storage. HDFS is an efficient storage system for data of huge volume and variety. Data are stored distributed across data nodes and each node has maximum storage capacity of 64MB. This distributed storage pave way for high security and also support parallel processing leading to high speed processing.

S.MD.Mujeeb et al. [11], briefly explain various techniques to handle intensive data and include mining as one of the important techniques. The authors list and explain various techniques, methods, algorithms and visualization approches involved in mining. G.Vinodhini et al. [16], explain term sentiment analysis in their paper. Sentiment analysis is a process of extraction of contextual opinion hidden in the data by involving text mining techniques and Natural Language Processing (NLP). Opinion may be in the form of sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Hence sentiment analysis can also be called opinion mining. Bing Liu [4], defines the term opinion as a quadruple (g, s, h, t), with its attributes 'g' as the target, 's' as sentiment towards the opinion, 'h' as source of the opinion and 't' as time at which it is expressed. Walaa Medhat et al. [18], explain various processes and approaches involved in sentiment analysis. Sentiment analysis involves process of data collection, text pre-processing, sentiment detection, sentiment classification and output presentation. Sentiment analysis can be carried out by any one of the two approaches: machine learning approach and semantic approach. Maite Taboada et al. [9], discusses sentiment analysis methods based on lexicon. Lexicon based sentiment analysis is termed as semantic approach. The lexicon-based approach aims at calculating sentiment based on semantic orientation of words /phrases in a document. The authors propose a method SO-CAL, to calculate the sentiment based on the semantic orientation of data. Alexander Hogenboom et al. [1], review various lexicons and its methods of generation and proposes a method for classification of lexicons based on their language. The lexicons are categorised by their generation methodology into dictionary-based and corpus-based servers best in terms of factor of accuracy.

Noémi Boubel et al. [12], bring into light the contextual valance shifter as one of the major issues of sentiment analysis. Contextual valance shifter plays a vital role to conclude semantic orientation (positive or negative) of a text. CVS includes negations, intensifiers and diminishers. Hence, they propose methods to automate identification and extraction of such occurrence in order to improve accuracy in sentiment analysis. Alistair Kennedy et al. [2] have made a study on CVS about its characteristics and occurrence. The paper has proposed two algorithm CVS and Term Counting and have compared the results of both. The comparison leads CVS with higher accuracy. Vo Ngoc Phu et al. [17] have improved CVS algorithm and proposed a methodology for sentiment analysis which is a combination of CVS and Term Counting. By combining both processes, researchers have noticed an increase in the accuracy rate.

Ramesh R et al. [14], propose sentiment analysis by machine learning approach for datasets from social media, which can be termed as big data. In order to manage the processing of big data, big data technology, hadoop is used. Thus researchers aim at improved accuracy of result and speed of processing. Ilkyu Ha et al. [7], proposes the processes of sentiment analysis on hadoop framework inorder to enable parallel process of data. The proposed work uses HDFS for storage and MapReduce function for sentiment analysis. Through parallel processing, researches have arrived at improving accuracy in result at less time.

## III. PROPOSED WORK

It is human tendency to share with and get opinion from others. With ample usage of digital equipments connected to internet, people share and gain opinion through internet as reviews in the form of unstructured textual data. Such unstructured data are Big Data and are rich in sentiment knowledge. Such knowledge can be extracted by the analytical process of Sentiment Analysis. Hence, the paper designs Hadoop architecture for the implementation of analytic process, sentiment analysis on big data to obtain effective results at efficient time.

The results of Sentiment Analysis have never obtained absolute accuracy ending up with incorrect results. The proposed algorithm PCVS aims at improving accuracy of result by considering context of the word, which involves shifting actual polarity of word based on the context. They are also termed as Contextual Valance Shifters that include contexts such as presupposition, if connectors, comma (,) or semicolon (;), either ... or...', 'neither ... nor ... ', comparative or superlative, intensifier or diminisher and negation.

### 3.1. Framework

The framework for processing data storage and processing sentiment analysis are designed neatly as in Figure 1 by incorporating Hadoop architecture which is a best suitable technology for storage and processing of Big Data.

Sentiment knowledge is opinion hidden within data generated through sources such as social networking sites: Facebook, Twitter, etc., blogs: WordPress, Blogger, LiveJournal, etc., forums: flickr, chronicle, digitalpoint etc., e-commerce sites: Walmart, Amazon, , Flipkart, etc., youtube and TV shows etc. The data of above sources are unstructured in nature. These data are extracted from data sources with the help of Web crawler Streaming API and Reviews Feeder. Web crawler is an application of breadth-first-search and a system for downloading bulk web pages. Streaming of APIs provides low latency access to global stream of data and Reviews Feeder maps data files into specified ledger or file.

The extracted data enhance Hadoop Framework, where storage and processing are carried out. Hadoop is an open source Java-based programming framework that supports parallel processing of large data sets in a distributed computing environment. Hadoop framework mainly dwells in two components:

- Hadoop Distributed File System (HDFS) stores data in a distributed manner. By default, Maximum size of each block is 64MB. The data extracted for analysis are broken into data splits of maximum size which is stored and distributed in data block in a format suitable for parallel processing.
- MapReduce: A programming paradigm that allows massive scalability during the process across Hadoop distributed cluster. Basically, MapReduce refers to two different tasks:
- Map: Works on data splits stored in HDFS in a parallel manner to produce intermediate results. The proposed framework assigns task of extracting sentiment word and its polarity from lexicon and Contextual Valance Shifter (CVS) calculation. Here, each data split has its own polarity values as its intermediate result.
- Reduce: takes output from map task and combines them to produce final result. As the term Map Reduce implies, Map task is always followed by Reduce task. Here, the intermediate result and polarity of each data split of the map task are obtained as input. The input and polarity values are summed up to obtain final polarity value of input on the whole.

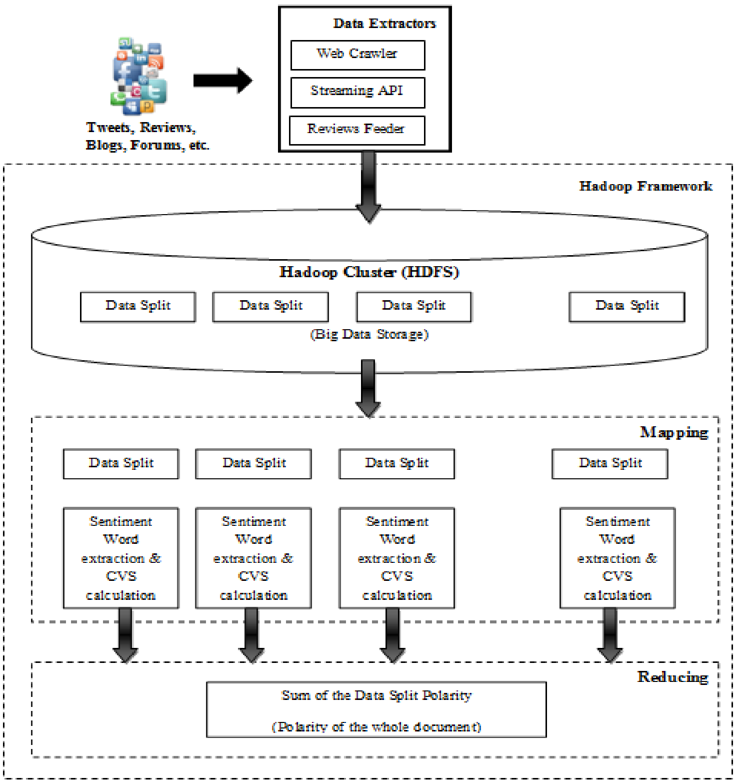

Figure 1 Hadoop Framework for Sentiment Analysis

### 3.2. Algorithm

The proposed Parallelized Contextual Valance Shifter (PCVS) algorithm mainly focuses on Contextual Valance Shifters that mostly include negations, intensifiers and diminishers and plays an important role in deciding context of sentiment words. The algorithm is built up in such a way that its pseudo code suits the map reduce paradigm as in Figure 2.

```
Algorithm PCVS
Input: reviews as a document
Output: valence of the review
Initialization: result $val_f = 0$
Begin
Extraction of data $data_i$ form Hadoop Cluster
For each $data_i$
PCVSMap ( )
PCVSReduce ( )
End
PCVSMap ( )
begin
        Extraction valence $val_i$ of $data_i$ from Lexicons
        if presupposition
                Extraction presupposition value $pre_i$ of $data_i$ from Lexicons, $val_i = val_i * pre_i$
        if connectors
                $val_i = val_i \times 0$
        if comma (,) or semicolon (;)
                $val_i = val_i + val_{i+1} + \ldots + val_{i+n}$
        if 'either ... or...':
                $val_i = val_i / 2 \times val_{i+1} / 2$
        if 'neither ... nor ... '
                $val_i = val_i \times 0$
        if comparative or superlative
                Extraction degree value $deg^{\circ}_i$ of $data_i$ from Lexicons, $val_i = val_i \times deg^{\circ}_i$
        if intensifier or diminisher
                Extraction percentage value $pec_i\%$ of $data_i$ from Lexicons, $val_i = val_i \times pec_i\%$
       if negation
                $val_i = val_i \times -1$
end
PCVSReduce ( )
begin
      $val_f = val_f + val_i$
end
```

Figure 2 Pseudo Code of PCVS Algorithm

The steps involved in the proposed PCVS algorithm are as follows:

***Input:*** reviews as a document
***Output:*** valence of the review
***Step 1:*** Extraction of data form Hadoop Cluster
***Step 2:*** Valence extraction from Lexicons
***Step 3:*** Perform the Contextual Valance Shifter check and perform appropriate function

- ***Presupposition:*** proposition that adds accuracy to the sentiment
- ***Connectors:*** terms that neutralizes the sentiment of the text (nil sentiment).
- ***Comma (,) or Semicolon (;):*** punctuations that sum the sentiment expressed
- ***'Either ... Or...':*** phrases that acts according to the context
- ***'Neither ... Nor ... ':*** phrases that neutralizes the sentiment of the text (nil sentiment).
- ***Comparative or Superlative:*** terms that changes the degree of the expressed sentiment.
- ***Intensifier or Diminisher:*** terms that alters the percentage of the expressed sentiment.
- ***Negation:*** terms that reverses the sentiment of a certain word.

***Step 4:*** Sum up the valance to get the polarity of the whole document.

The integration of  algorithm with MapReduce paradigm Steps 1 – 3 falls in the Mapping part, PCVSMap() and Step 4 falls in the Reducing part, PCVSReduce(). This implies that Steps 1-3 are executed in parallel on the distributed split up data retrieved from the hadoop cluster and produces intermediate results in PCVSMap(). The intermediate results are considered as input for PCVSReduce() which has Step 4 assigned to it. Here, the inputs are gathered and summed up.

The map job is responsible for the extraction of  sentiment data and its polarity from hadoop cluster and lexicon respectively. It is also responsible for calculation of  polarity according to various contextual valance shifters shown in Figure 2, that are present in data and responsible for context of  data subjected for analysis. Further work, calculation of total sentiment value of the whole document is taken care by the reduce job. The Figure 3 represents workflow of  PCVS algorithm with classification of contextual valance shifters and respective calculation. The workflow also represents the processes carried out by map function, PCVSMap() and reduce function, PCVSReduce().

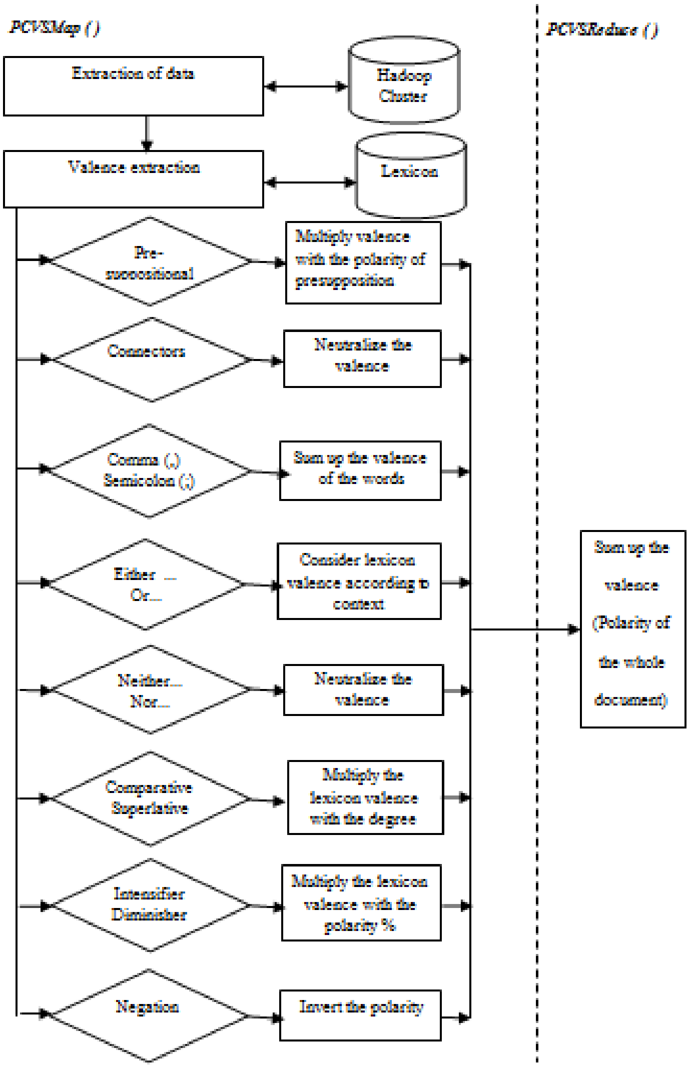

Figure 3 Workflow of PCVS Algorithm

## IV. RESULTS AND DISCUSSIONS

The framework designed for parallel processing and algorithm proposed PCVS supports process of sentiment analysis on a large amount of unstructured textual data called, big data. The work of this paper mainly focuses on accuracy and speed efficiency of the process.  Incorporation of hadoop framework in the proposed framework supports the processing of big data. One of the components of the hadoop framework is mapreduce which supports parallel processing and helps in increasing the speed of processing leading to efficient processing time. Hence, the framework directs to efficient handling of data and economical processing speed and time.

The PCVS algorithm is mainly based on contextual valance shifter that plays an important role in deciding polarity of sentiment word based on the context of  data. Thus, leads to improve accuracy rate in resultant polarity value of the algorithm.

## V. CONCLUSION

Sentiment analysis is a process of extracting hidden opinion from data rich in sentiment knowledge. The extracted knowledge serves best in the process of decision making. The data rich in sentiment knowledge are mostly related to review of entity/product in the form of Big Data. In order to sort out the need of extraction of sentiment knowledge from big data, the paper makes a study on the works related to the solution and proposes a framework and algorithm suitable for the process. The framework design involving Hadoop framework that supports parallel processing of big data resulting in economical processing speed and time. The algorithm PCVS targets major issues of sentiment analysis, contextual valance shifter, which concludes the context of sentiment word to improve accuracy of the resultant polarity.

## REFERENCES

[1] Alexander Hogenboom, Malissa Bal, Flavius Frasincar and Daniella Bal (2014), “Lexicon-based sentiment analysis by mapping conveyed sentiment to intended sentiment”, International Journal for Web Engineering and Technology, Volume 9, No. 2.

[2] Alistair Kennedy and Diana Inkpen (2006), “Sentiment Classification of Movie Reviews Using Contextual Valence Shifters”, Computational Intelligence, Volume 22, Issue 2, Pages 110–125

[3] Avita Katal, Mohammad, Wazid and R H Goudar,(2013) “ Big Data: Issues, Challenges, Tools and Good Practices”, International Conference on Contemporary, pages 404 – 409

[4] Bing Liu (2012), "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, 167 Pages

[5] Dipak M. Durgude, Nilesh S.Yalij, Ashwini B. Bhosale and Manisha Bharati (2015),"Big Data Analysis: Challenges and Solutions",International Journal of scientific research and management (IJSRM), Volume 3, Issue 2, Pages 2106-2112

[6] Harshawardhan, S. Bhosale, Prof. Devendra and P. Gadekar, "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10

[7] Ilkyu Ha, Bonghyun Back and Byoungchul Ahn (2015), “MapReduce Functions to Analyze Sentiment Information from Social Big Data”, International Journal of Distributed Sensor Networks 417502, 11 Pages

[8] Karthik Kambatlaa, Giorgos Kollias, Vipin Kumar and Ananth Gramaa, (2014), “Trends in big data analytics”, Journal of Parallel Distributed Computing 74, Pages 2561–2573

[9] Maite Taboada, Julian Brooke, Milan Tofiloski and Kimberly Voll (2011), "Lexicon-BasedMethods for Sentiment Analysis", Association for Computational Linguistics, Volume 37, No. 2

[10] Min Chen, Shiwen Mao and Yunhao Liu (2014) , “Big Data: A Survey”, Springer Science and Business Media New York, Mobile Network Applications, Pages 171–209.

[11] S.Md.Mujeeb and L.Kasi Naidu (2015), "A Relative Study on Big Data Applications and Techniques", International Journal of Engineering and Innovative Technology, Volume 4, Issue 10

[12] Noémi Boubel, Thomas François and Hubert Naets (2013), “Automatic Extraction of Contextual Valence Shifters”, Proceedings of Recent Advances in Natural Language Processing, Pages 98–104

[13] Parth Chandarana and M. Vijayalakshmi (2014), "Big Data Analytics Frameworks", International Conference on Circuits, Systems, Communication and Information Technology Applications, Pages 430-434

[14] Ramesh R, Divya G, Divya D and Merin K Kurian (2015), “Big Data Sentiment Analysis using Hadoop”, International Journal for Innovative Research in Science & Technology, Volume 1, Issue 11

[15] Subramaniyaswamy V, Vijayakumar V, Logesh R and Indragandhi V (2015), “Unstructured Data Analysis on Big Data using Map Reduce”, 2nd International Symposium on Big Data and Cloud Computing volume 50, pages 456 – 465

[16] G.Vinodhini and RM.Chandrasekaran (2012),“ Sentiment Analysis and Opinion Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2:6

[17] Vo Ngoc Phu and Phan Thi Tuoi (2014), "Sentiment Classification using Enhanced Contextual Valence Shifters", International Conference on Asian Language Processing, IEEE, Pages 224 – 229

[18] Walaa Medhat A, Ahmed Hassan B and Hoda Korashy(2014),“Sentiment analysis algorithms and applications: A survey”, Ain Shams Engineering Journal, Vol 5, Pages 1093–1113

[19] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding (2014), “Data Mining with Big Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 26, No. 1

[20] Yang Zhang and Dan Liu (2012), "Improving the Efficiency of Storing for Small Files in HDFS", International Conference on Computer Science and Service System, IEEE, Pages 2239 - 2242