IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Big Data Management in Telecom Domain Using Hadoop

Miss. Deshmukh A.A.[1], Miss. Hasarmani D.R[2],Miss. Kashid P.B[3], Miss. Jagatap P.S.[4], Prof.Raule M.B.[5]

Student, Computer Science & Engg, Bhagwant Institute of Technology, Barshi, India[1,2,3,4]

Asst. Prof., Computer Science & Engg, Bhagwant Institute of Technology, Barshi, India[5]

Abstract: Map reduce is a programming model for examining and processing large data sets .Apache hadoop is an effective framework and most popular implementation of the map reduce model. Hadoop's success has motivated research interest and has led to different modifications as well as extensions to framework. In this paper, the challenges round faced to in several domain like data storage, analytics, online processing and privacy/ security issues while handling big data by using K –means algorithm for clustering and SVM algorithm for classification. Also we provide the security by AES algorithm. In this paper, we focus specifically on Hadoop and its implementation of Map Reduce for analytical processing.

Keywords: Big Data, Telecom Domain, Map Reduce, Call Detail Records.

I. INTRODUCTION

In previously existences data blast is leading for increasing demand in big data processing which are distributed among different geographical centres. Demand on big data is being rising day by day and also growing heavy burden on calculation storage and communication in data Telecom companies are sitting on a gold mine, as they have abundance of data. But what they need is a appropriate tunnelling and examination of both structured and unstructured data to get deeper insights into customer behaviour, their service usage patterns, preferences, and interests real-time, but there was a marvellous increase in the amount of data, their computation and analyses in recent years. In such situation most classical data mining methods became out of reach in practice to handle such big data[1][4].It is a very challenging problem of today to analyse the big data. Big data is big deal to work upon and so it is a big job to perform analytics on big data. Technologies for analysing big data are evolving rapidly and there is significant interest in new analytic approaches such as Map Reduce, Hadoop and Hive, and Map Reduce extensions to existing relational DBMSs. The use of Map Reduce framework has been widely came into focus to handle such massive data effectively. For the last few years, Map Reduce has appeared as the most popular computing paradigm for parallel, batch-style and analysis of large amount of data, Map Reduce gained its popularity when used successfully by Google. In real, it is a scalable and fault-tolerant data handling tool which provides the ability to process huge big data in parallel with many low-end computing nodes [3][5][10].

A. BIG DATA:

Big data is normally used for storing of the large datasets. Normally we store the data in the size of megabytes but if the size of the data expands up to the penta bytes then there is term use named as big data. This big data has three v's namely volume, velocity, variety. The data which comes from variety of sources, which come s from high speed, which is larger in size this term is referred as big data. Therefore an oversized quantity of knowledge generated by completely different stakeholders within the medium trade. so it becomes terribly difficult to work on the large knowledge within the massive knowledge domain, map cut back is on among the key approach used for handling massive knowledge sets.

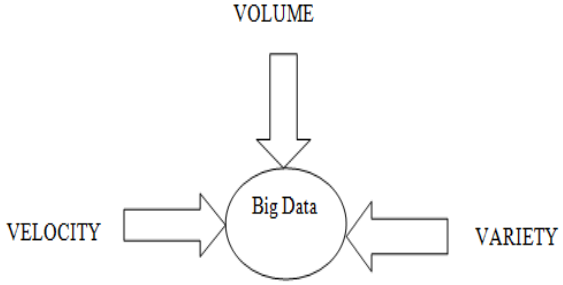


Fig.1: 3 V's of Big Data

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

B. BIG DATA IN TELECOM:

The past span has seen an exponen0tial progress in the telecommunication industry. The cost of communication has gradually decreased and thanks to the improvement in the electronics industry, mobile phones have become inexpensive and feature rich. Now one not expects a phone to only build and receive calls and text messages. Smartphone became a useful part of our life. The developing countries corresponding to republic of India and china are driving the transportable market and aren't any longer thought about as electronic dump zones by the most important players of the transportable producing business.

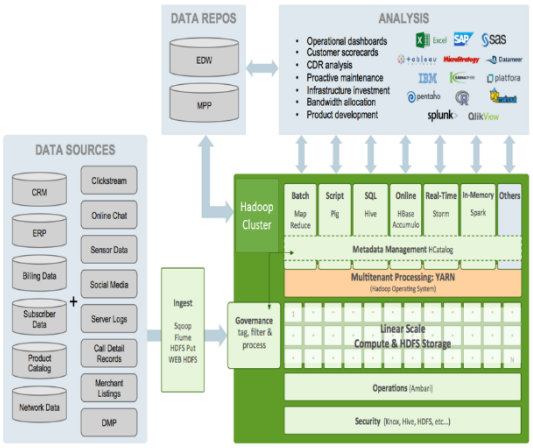


Fig.2: Hadoop Technical Architecture for Telecom Domain

C. CALL DETAIL RECORD (CDR):

A call detail record havingdata about data –that is metadatahaving datafields that defineaexact instance of a telecommunication transaction, but does not contain the content of that transaction. A call detail record describing a particular phone call might include the phone numbers of both the calling and receiving persons, the start time, end time, caller id and duration of that call. In actual currentpreparation, call detail records are much more detailed, and have attributes A call detail record(CDR)may be a information record formed by a work or alternative telecommunication in improvise intention that documents the main points of a call or another telecommunications dealings(eg.text message)that passes through that facility or device. It's the machine controlled corresponding of the paper ring tickets that were written and regular by operators for long-distance calls in a very manual work.

Call detail records serve a spread of functions. For phone service suppliers, they're spirited to the assembly of revenue therein they supply the idea for the generation of phone bills. For enforcement, call detail records offers a wealth of knowledge which will facilitate to spot doubts, therein they'll reveal details on a human relationship with associates, communication and behaviour patterns, and even location information which will establish the where about of a personal throughout the whole thing of the decision.

1) Why are CDRs important?

A Call Detailed Records list of every logs billing communications transmission on your phone system. This allows phone companies to generate your phone bills, and lets you keep fixed records of how and when your phone system was used. They are primarily used by businesses to assist in call reporting and billing. CDRs can be used to identify calling trends and gain insights into employees' use of phones. Billing departments use CDRs to resolve disputes, keep records of how funding is spent, and log usage of the telephone system. IT departments can also use CDRs to determine if there were any disruptions in phone service.

2) Working Principle:

- CDR is ainspiration process which preserves busily looking for files to process and continues till automatically stopped.
- Organizes the caught events from VBA into In-band actions (digits, echo, etc.) results and overall traffic signal capacities.
- CDR can be organized to output its results to "comma-separated values" ("CSV") files or ASCII file for loading into a database or spreadsheet.
- All files are in "CSV" ("Comma-Separated Values") format, a widely used format in the Windows® world implicit by current data managing applications such as Microsoft® Excel and Access.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

II. IMPLEMENTATION

In proposed system we used k- means cluster algorithm. This dynamic cluster algorithm is widely used for clustering. We present MP Cache, a solution that utilizes solid-state drive (SSD) to cache input data and localized data of Map Reduce tasks. Here we can use world count benchmark with map reduce in Hadoop. We can use bank detail dataset for map reduce function. It is cloud based web application which stores data in Amazon S3. This project intends to overcome all these obstacles and built a user friendly SAAS platform. It is cloud based web application which stores data in Amazon S3. As this system supports dynamic and optimized cluster nodes size as per the desired time, user doesn't need to calculate and estimate the number of nodes.

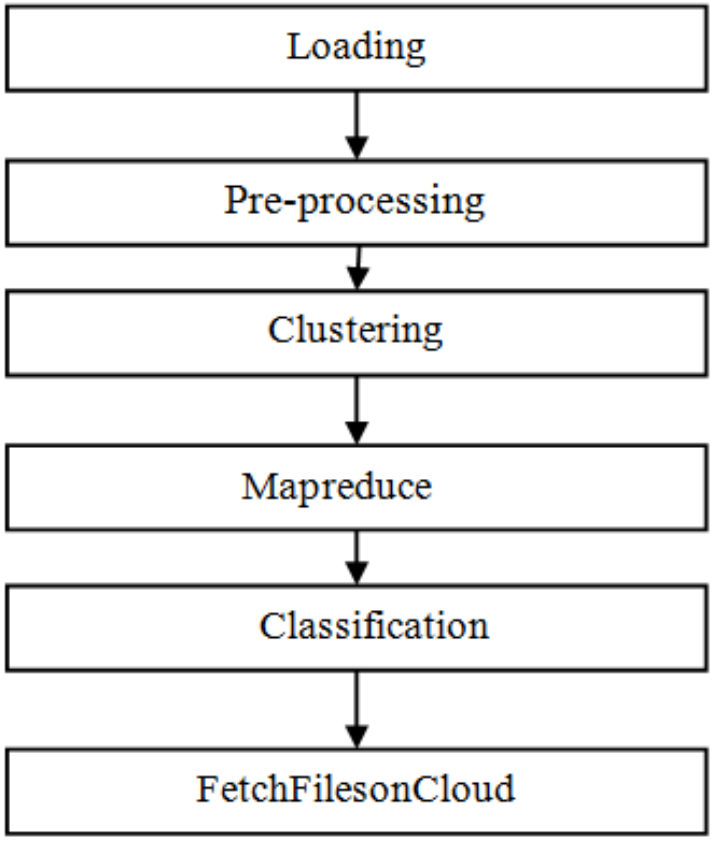


Fig 3: Flow Chart

MODULES:

- Pre-processing
- Clustering
- Map Reduce
- Fetch Cloud

PREPROCESSING:

Pre-process is one of main modules for data mining system. Here we are eliminating annoying data or null values and unstructured data. So when we eliminate unstructured data's then only we get exact results for given dataset. It is mainly applicable to data mining and machine learning projects. Data-gathering ways area unit over and ones more loosely controlled, leading to out of range values, not possible knowledge combos, missing values, etc. Analysing knowledge that has not been rigorously screened for such issues will manufacture false results.

CLUSTERING:

Clustering can be measured the most important invalid learning problem; so, as each alternative drawback of this type, it deals with finding,a.structure during.a assortment of unlabeled information. A moveable definition of clustering could be "the process of forming objects into groups whose members are similar in some way".

A cluster is therefore a collecting of objects which are "like" between them and are "unlike" to the objects be appropriate to other clusters.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

MAP REDUCE:

Map reduce is a programming model associated and connected application for process and generating huge information sets with a parallel, unfold algorithmic program on a cluster. The model is knowledge of the split-apply-combine policy for data analysis. Here Hadoop platform with applied the convinced process as finding overall data with mapping and how much data will be reduced. The term map reduce states to two separate and completely different tasks that hadoop paper programs perform. The first is that the map job, that takes a collection of knowledge and changes it into another set of knowledge, wherever specific components area unit softened into tuples.

In this paper we use Map Reduce framework. That provides a similar processing model and related application to process vast amounts of data. With Map Reduce, questions are divided and distributed through parallel nodes and processed in parallel. The results are then collected and brought .In Map Reduce algorithms such as clustering, classification are used.

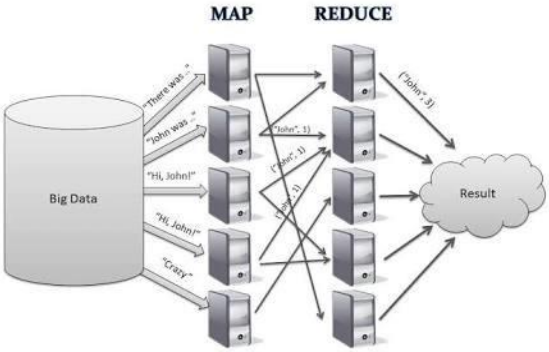


Fig 4: Unstructured data from different telephone conversations

FETCH CLOUD:

Fetch cloud is the pull out data from the cloud server through some security mechanism. Most cloud storage provides support. Web manners based on representational formal handover application programming interfaces (APIs). Some also support old block- and file-based data, and cloud storage access providers can help customers access data in major storage clouds. To help the clustering process in this task, we completed pre-processing steps such as feature selection and Principle Component Analysis (PCA) and still, the select of clustering method is not a minor one. To find the best performing algorithm.

B. EVALUATION:

1) K –MEANS FOR CLUSTERING

K-Means cluster may be a method of vector division, at the start from signal process, that's common place for cluster analysis in data processing. K-Means cluster aims to divider n annotation into K clusters during which every opinion belongs to the cluster with the closest mean, serving to as a model of the cluster. This leads to a partitioning of the information house into voronoi cells the matter is computationally tough (NP-hard); but, there are economical heuristic algorithms that arunre makeable utilized and converge quickly to and area optimum.

This capacity thing typically like the expectation maximization rule for combinations of mathematically distributions via associate repetitious modifications approach utilized by each algorithm. Furthermore they all use cluster centers to model data; but K-Means cluster tends to search out clusters of equivalent special extends, where as the expectancy maximization mechanism approvals clusters to own totally different shapes.

The method structures a loosed relationship to the K-nearest neighbour classifier, a well-liked machine learning technique for classification that's typically confused with K-means remaining to the K within the name. one well apply 1-nearest neighbour classifier on the cluster centers obtain by K means to classify new information into the present cluster. These can be referred to as nearest centre from classifier.

This formula aims at minimizing an objective perform capture as genuine error performs given by:

$J(V)=\sum_{i=1}^{C}\sum_{j=1}^{Ci}(||x_i-v_j||)^2$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

2) SVM FOR CLASSIFICATION

In machine learning classification of data is a mutual task. Suppose some given data points each go to one of two classes, and the goal is to choose which class a new data point will be in. In the case of support vector machines, a data point is observed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a (p-1)-dimensional hyper plane. This is called a linear classifier.In machine learning, support vector machines SVMs, commonly support vector networks area unit controlled data models with connected learning algorithms that analyse information used for classification and analysis.

1) Given aassembly of trainingsamples, every noticeable as pleasure to 1 or the opposite of the 2 classes.
2) Secondary mark SVM training algorithmic instructionconcept a classical that allocates new examples to 1 class or the reverse, making it a non-probabilistic binary linear classifier (although schemesequal to Platt scaling happen to use SVM in a very probabilistic managing huge knowledge abusehadoop map cut back in medium domain classification setting).
An SVM model could be a project of the examples as points in community, strategic in order that the models of the separate classes authentic amount divided by aobvious gap that's as wide as potential. New examples are then planed into that very same ability and probable to belong to a class supported that feature of the gap theyreduction.

3) AES ALGORITHM

Advanced Encryption Standard (AES) is a cryptographic algorithm that can be used to defend electronic data. The AES algorithm is a symmetric block cipher that can encode (encipher) and decode (decipher) data. Encryption converts data to an meaningless from called cipher text; decrypting the cipher text changes the data back into its actual form called plaintext. The AES algorithm is skilled of via cryptographic keys of 128, 192, and 256 bits to encode and decode data in blocks of 128 bits. In this algorithm the three different key lengths is stated to as "AES-128", "AES-192", and "AES-256"used.

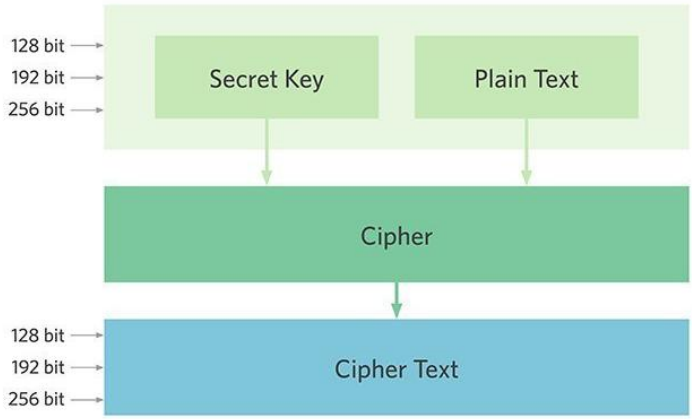


Fig.5: AES Design

The more standard and extensively assumed symmetric encryption algorithm likely to be come across nowadays is the Advanced Encryption Standard (AES). It is found at least six time faster than triple DES.A spare for DES was needed as its key size was too small. With growing calculating power, it was considered weak against complete key search attack. Triple DES was designed to overcome this drawback but it was found slow.

The features of AES are as follows −

- Symmetric key symmetric block cipher
- 128-bit data, 128/192/256-bit keys
- Stronger and faster than Triple-DES
- Provide full specification and design details
- Software implementable in C and Java

III. RESULT AND ANALYSIS

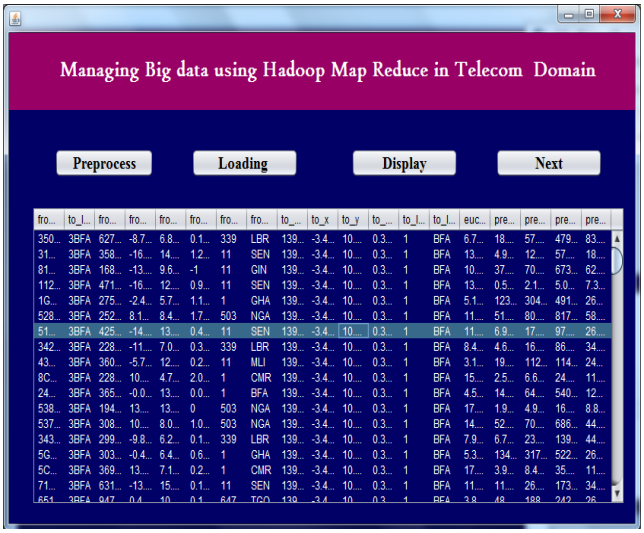
PREPROCESING FORM:

In processing of data firstly pre-processing takes place. In pre-processing unwanted data will be removed. So when we remove unwanted data then only we get accurate data.

IJARCCE

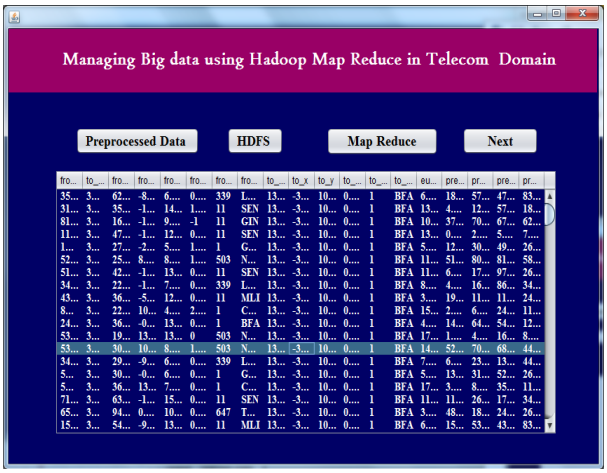
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018



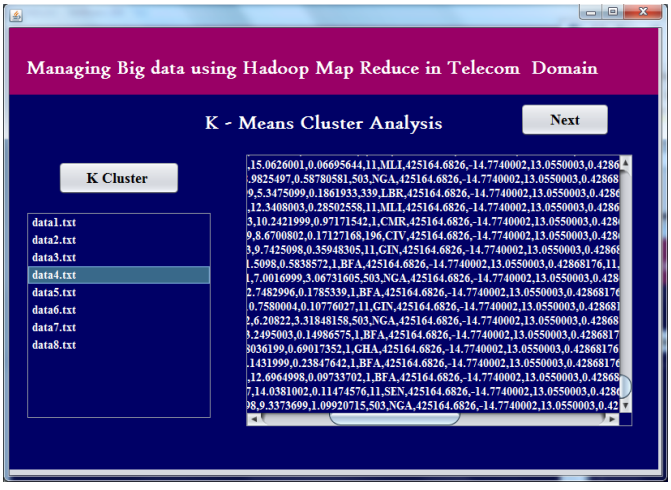
MAP REDUCE FORM:

The term map reduce actually refers to two distinct and separate tasks. The map function is responsible for filtering and sorting. The reduce function is responsible for grouping and aggregation operations. So here sorted data will be grouped.



CLUSTERING FORM:

In this organization of object into group whose members are similar in some way takes place. The objects which are similar are in one cluster and which are dissimilar are in another cluster.



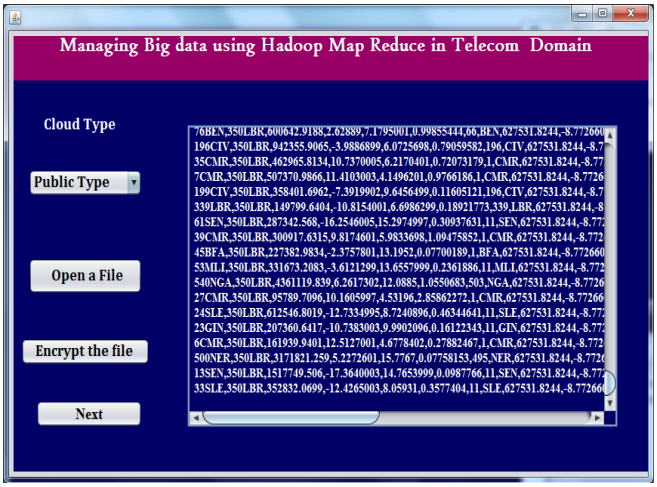
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

FETCH CLOUD FORM:

In this extracting of data from the cloud server through some security mechanism takes place. Fetching is the final process for processing of data. It is used when we needed some data from the cloud.



IV. CONCLUSION AND FUTURE SCOPE

A. Conclusion:

In the real world, data processing and storage approaches are facing many challenges in meeting the continuously increasing demands of big data. This work focused on Map reduce, one of the key approaches for meeting the big data demands through highly parallel processing on a large amount of commodity nodes. Challenges and solutions on four dimensions like data storage, analytics, online processing and privacy and security are elaborated in detail in this paper.

B. Future Enhancements:

In the future, we plan to consider the setting of geo-distributed collocation data centers and investigate the multi-tenant coordination issue while considering the geographical load balancing opportunity. We continued the same process with the usage of different algorithm. Experiment and analysis confirm the effectiveness of our schemes and design.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ”Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.
[2] Q. Zhang, L. T. Yang, and Z. Chen, ”Deep Computation Model for Unsupervised Feature Learning on Big Data,”IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016
[3] P. Zadrozny and R. Kodali, Big Data Analytics using Splunk, Berkeley, CA, USA: Apress, 2013.
[4] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), pp. 107-113, 2008
[5] J. Dean and S. Ghemawat, ”MapReduce: Simplified Data Processing on Large Clusters,”Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.
[6] Z. Xie, S. Wang, and F. L. Chung, ”An Enhanced Possibilistic c-Means Clustering Algorithm EPCM,”Soft Computing, vol. 12, no. 6, pp. 593-611, 2008.
[7] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, ”An Incre-mental CFS Algorithm for Clustering Large Data in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, 2015.
[8] Y. Chen, L. Wang, and M. Dong, ”Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering,”IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.
[9] Q. He, Q. Tan, X. Ma and Z. Shi, "The high-activity parallel implementation of datapreprocessing based on MapReduce," Proc. Of the 5th International Conference on Rough Set and Knowledge Technology, 2010.
[10] J. Heer and S. Kandel, "Interactive analysis of Big Data," XRDS: Crossroads, the ACM Magazine for Students, 19(1), pp. 50-54, 2012. 10. Y. Chen, S. Alspaugh and R. Katz, "Interactive analytical processing in Big Data systems: A cross-industry study of MapReduce workloads," Proc. of the VLDB Endowment, 5(12), pp. 1802-1813, 2012.
[11] Q. Zhang, L. T. Yang, and Z. Chen, ”Privacy Preserving Deep Computa-tion Model on Cloud for Big Data Feature Learning,” IEEE Transactions on Computers, vol. 65, no. 5, pp. 1351-1362, May 2016.