Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 12, December -2017

A Novel approach for Big Sensing Data Processing based on Scalable Data Chunk Similarity

S. Md. Mujeeb, Raj Kumar Patil

S. Md. Mujeeb, Asst. Prof., Dept. of CSE, Malla Reddy Institute of Technology & Science, Hyderabad.
Raj Kumar Patil, Asst. Prof., Dept. of CSE, Malla Reddy Institute of Technology & Science, Hyderabad.

Abstract: *Big sensing data is the information with high volume and high velocity. The huge detecting information is created in the numerous frameworks at various applications, for example, inquire about and huge organizations. Cloud is where the information can be stored, serviced and figured at the cloud. Since the information is enormous volume there is need of information preparing, for example, information compression. The information pressure systems require vast versatility and proficiency for the handling of the expansive volume of information in the high rate. In this paper we are proposing the procedure for the pressure of the data. This strategy depends on the on cloud necessity of information pressure and similitude count between information chunks. In this technique right off the bat the huge information is separated into the distinctive lumps and afterward these pieces are packed by the pieces. Here the calculation known as Jaccard is utilized to discover the closeness between the data. Then this calculation is contrasted and the current frameworks likeness calculation.*

Index Terms: *Big Sensing data, Cloud computing, Map Reduce algorithm for big data, compression techniques.*

I. Introduction

The big sensing data is vast in volume and it originates from the distinctive detecting frameworks. This information is substantial and it requires preparing before putting away it on the cloud. Subsequently this reason information ought to be compacted first and afterward it ought to be put away on the cloud. There are diverse well springs of the enormous detecting information, for example, camera, video, satellite meteorology, activity checking, complex material science re-enactments. Thus huge information handling is a major crucial test for the advanced society. Cloud is a promising stage for enormous information preparing with its capable computational capacity, stockpiling, asset reuse and ease handling the huge detecting information is still expensive as far as space and time. For the decrease of the time and space cost of huge information diverse procedures ought to be required. In the event that some recursive calculations are utilized to process the huge detecting information it can create numerous issues, for example, memory bottlenecks, deadlocks on information getting to.

In this paper the initially the similitude information pieces are created. At that point the closeness calculation is connected on these information pieces. Here the Jaccard comparability calculation is utilized. In base paper the Cosine closeness calculation was utilized. In this paper we are looking at the two calculations on the factor exactness of the information. The exactness and the space cost parameters are straight forwardly corresponding to each other i.e. as the exactness of the information builds the more space is required to store that information. The comparability calculation Jaccard is utilized to discover the information similitude more precisely than Cosine likeness calculation. In comes about we will look at both the calculations i.e. Jaccard calculation and cosine similitude calculations on the premise of the precision of the information safeguarded i.e. how precisely they finds the closeness of the information. When contrasted with the Cosine comparability calculation the Jaccard calculation is more favourable position. In the wake of finding the comparability by Jaccard calculation the calculation is connected to the information pieces is known as MapReduce algorithm. The MapReduce calculation is utilized is for the pressure of the comparative information lumps.

II. Review of Literature

Cloud can likewise be utilized for the capacity. It additionally gives engineering of the cloud with showcase situated asset distribution by utilizing advancements, for example, virtual machines (VMs). It likewise portrays the market based asset administration systems that spreads both client driven administration and support the administration level assertion. Another paper depicts the general algorithmic plan system in the MapReduce structure called separating. The fundamental reason behind the separating is to lessen the extent of contribution to the conveyed form. By sifting coming about considerably littler, issue example can be settled on single machine. Utilizing the distinctive methodologies new calculations in the MapReduce structure for an assortment of central chart issues for adequately thick diagrams. In this calculation the measure of memory accessible on the machines enable us to indicate exchange off between the memory which is accessible and the quantity of MapReduce rounds. This actualizes the maximal coordinating calculation that lies at the centre of the examination which demonstrates noteworthy accelerate. Jaccard coefficient is utilized as a data likeness measure. Inexpansion of the comparability measure Jaccard has the preferred stand point which secures the

protection of the information. They proposed the SJCM convention (Secure calculation of the Jaccard Coefficient for Multisets) utilizing the current dab item strategy. Paper, demonstrates the meaning of the abnormality recognition and the huge information. The abnormality identification depends on the uncompressed information because of capacity load and the insufficiency of security assurance and compressive detecting hypothesis presented and utilized as a part of the peculiarity location algorithm. This inconsistency recognition system utilized for the through-divider human discovery to show the adequacy.

Paper consists of following folds:

1) Scientific communication which is little to medium scale uses the versatile assets on general society cloud site while keeping up their adaptable framework control.

2) Light weight administration for making administration assignments straight forward and benefit heterogeneous workloads. Paper gives the methodologies and systems of conveying serious information applications which are picking up scalability, consistency, economical handling of huge scale information on the cloud and furthermore features a few attributes of the best competitor classes. This investigations the Hadoop based MapReduce system. The new information examination stage that utilizes hash methods to empower quick in-memory processing. This stage enhances the advance of guide undertakings enables the guide to advance with up to 3 requests of greatness diminishment of the information and it likewise empowers results to be returned consistently amid work. Paper, portrays a work process administrator which is produced and conveyed at Yahoo CallesNova. This chief pushes the ceaselessly arriving information through diagrams of projects which are executing on Hadoop clusters. The programs are known as pig which is structure stream dialect.

III. System Architecture / System Overview

A. System Description In proposed work primary calculation is utilized best process enormous detecting information. Thus, a few highlights of huge detecting information will be contemplated and examined. To do pressure, the closeness between two distinct information lumps ought to be characterized. Along these lines, how to characterize and show the closeness between information lumps is an essential prerequisite for information pressure. After the definition for the above comparability show for information pieces, how to create those standard information lumps for future information pressure is additionally a basic method which we outlined. A novel pressure calculation is produced and outlined in light of our similitude model and standard information piece age.

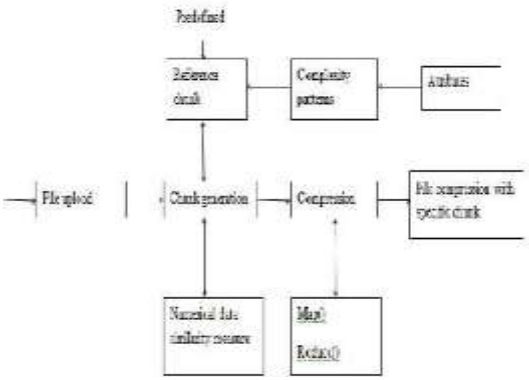

Fig.1. System Architecture

Information File this is the primary module of the framework. Here client gives the document as information. Client can include any designed record as info. When client inputs the record will transmitted for producing pieces in next module. Piece Generation Here we have presented system for producing lumps of information. In the presentation we have spoken to a thought regarding information piece which depends on pressure. This is subject isn't helpful for packing the information. It resembles pressure of high regular element. Here contrast is that pressure of these components distinguishes just straight forward information units. In any case, our information lump based pressure distinguishes complex parcel and example amid pressure process. Closeness calculations for finding the similitude between the information pieces. Here the Jaccard comparability calculation is accustomed to finding the likeness between the information lumps and the information stream. In existing framework the cosine likeness calculation is utilized for finding the similitude.

Following the both algorithms are given.

1) Likeness show for the cosine similitude calculation: The cosine closeness works for both content information and also message data. In this paper we are taking the information which is of climate estimating and as numerical configuration only. Therefore we are utilizing cosine comparability calculation just for the numerical information. In cosine similitude calculation the cosine point between the two vectors is calculated. If the edge between the two vectors is zero then the

comparability between vectors is one. For the bigger estimation of edge the likeness is less. Cosine edge is figured by the accompanying equation for the two vectors X and Y:

$$Sim(\vec{X}, \vec{Y}) = \cos \emptyset$$

$$\cos \emptyset = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{\|\vec{X}\| \times \|\vec{Y}\|}$$

$$\cos \emptyset = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \times \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}}$$

2) Similarity model for Jaccard similarity algorithm: Jaccard Similarity measure is another measure for calculating the similarity. In this algorithm, the index starts with a minimum value of 0 (completely dissimilar data) and goes to a maximumvalue of 1 (completely similar data).The similarity measure between the two data sets X_i and X_j can be computed using the following formula:

$$Similarity\ measure\ (X_i, X_j) = \frac{(X_i \cap X_j)}{(X_i \cup X_j)}$$

Above formula gives the Jaccard similarity measures for thetwo documents. The value comes from 0 to 1. Compression for implementing proposed system on cloud we need to go through two essential stages that are generation of data chunk andcompression based on these chunks. That is why two algorithms have been developed respectively for data processing in these two stages given above. In the module we have used Map() and Reduce() method for compressing the data chunk. Above algorithms and techniques used in the proposed system, we willget the compressed files with specific chunks.

There are four purposes for this experiment:

1) Increase the accuracy of the data.

2) To show that the significant storage saving is achieved due to compressed data blocks.

3) To show that the significant time saving is achieved because lots of realbig data blocks can be inferred instead of real search andnavigation.

4) Compared to significant time and space performance gains, only tiny data loss is introduced in terms of accuracy.

Mathematical Model Input file set = f_1; f_2; f_3; ……f_{ng} Chunks creation on fix size i.e. ch = ch_1; ch_2; ch_3; …..ch_n Where, ch-stream of data series and n-no. of data units streaming. Similarity check by Jaccard similarity. Threshold set for similarity is 'T': Compress element ch_1 from n_1 and other get stored as no similarity find. Compression fch_1; ch_{2g} Recursive calculation of standard data chunks

Algorithm

1) Standard data chunks generation algorithm:

Input:

Streaming big sensing data set S; maximum time threshold for chunk evolvements.

Output:

Data chunk set S' which is a subset of S

Process: Initialize process is conducted including S^0 and its first Element Similarity mode is calculated and selected according to application requirement e.g., numerical data. The first element x1 in big data set selected as a first elementin the standard data chunks set S'. Length of the S' is set as 1.

2) Jaccard algorithm for finding similarity:

Input: Generated data chunk, Data.

Process: The fix value Threshold T is taken. The data is compared to that T value. If the value is less than the T then the data is added to the chunk. If the value is greater than T then the data is stored as it is. The similarity can be find out by the Jaccard function.

3) MapReduce algorithm for compressing the data chunks:

The MapReduce algorithm is divided into two parts asMap() and Reduce()

1. Compression algorithm: Map():

A. Mapper side takes Sand S' as input.

B. The numerical data type is selected.

C. The total numbers of elements in the standard data chunk

S' is calculated and stored in L.

D. Recursive similarity comparison function is called again totag any data element in S to find any x S which could be compressed.

E. The data elements which could be compressed are tagged in map() function.

2. Compression algorithm:Reduce():

A. The reducer side compression algorithm extends the table reducer <> of map reduce model.

B. Reduce() function takes S as input

C. The compression model is selected

D. The element S is compressed by using compress() function.

E. After the compression the storage is updated.
F. Index is stored for future decompression process.

IV. Results Discussion

Table 1
Comparison of times required for the both algorithms:

SIZE	Cosine(msec)	Jacard(msec)
384KB	2197	700
584KB	7700	2500
1.1MB	26000	6000

A. Proposed Results: In this paper the accuracy of the two similarity algorithms isconsidered and the compared for the big sensing data.Thealgorithms are applied to the database and the accuracy of the both algorithms are to be compared.
Output:
Classified data according to the chunks.
B. Performance Measures: The performance measures used are the efficiency and the accuracy of the system. The algorithm Jaccard similarity and Cosine similarity are compared among the two performance measures i.e. how the efficiency and the accuracy increased inthe Jaccard algorithm.
C. Comparison of Jaccard similarity and cosine similarity: Figure 2 gives the comparison of the Jaccard similarity algorithm and the Cosine similarity algorithm. Here the time factor is considered for the comparison of the algorithms. As shown in the figure different sizes of data chunks are considered and the time stamp is compared.

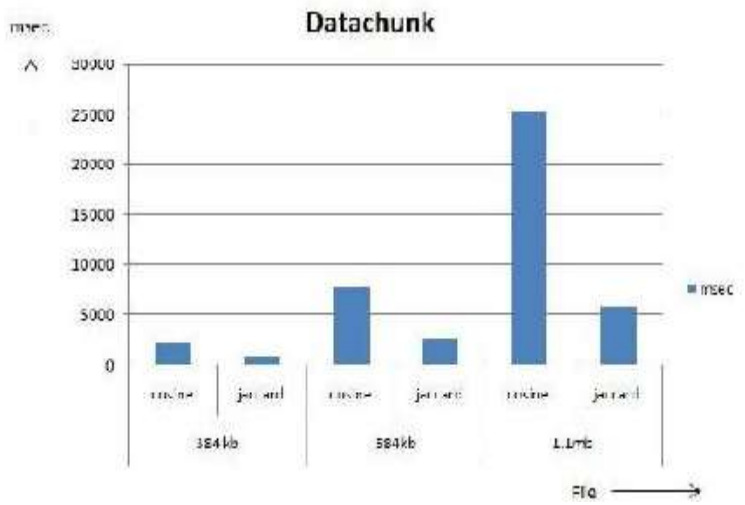

Fig.2. Comparison graphs for both algorithms

D. Experimental Setup: For experiments big data storage on Cloud platform, to reduce thedata size also means the time saving for navigating and decompressing those data units. Main designing target of ourcompression algorithm based on data chunks similarity is to reducethe data size and volume; we also consider the data quality and fidelity loss after deploying our proposed compression and decompression.
E. Dataset: The meteorological data is used as the dataset. This data is the big sensing data. This data is in the numerical format.
F. Expected Result: Here two similarity algorithms are compared to each other interms of accuracy of the data during the compression process. In this paper we used Jaccard similarity algorithm. We use the numerical meteorological numerical dataset. The data is divided into the data chunks depending upon similarity of the data. For similarity the Jaccard algorithm is used and for compression MapReduce algorithmis used.

V. Conclusion

It is demonstrated that our proposed scalable compression based on data chunk similarity improve the datacompression performance gains with data accuracy loss. The significant compression ratio brought which is space and time cost saving.

References

[01] B. Li, E. Mazur, Y. Diao, A. McGregor and P. Shenoy, A platform forscalable one-pass analytics using mapreduce, in: Proceedings of the ACMSIGMOD International Conference on Management of Data (SIGMOD11), 2011, pp. 985-996.
[02] C. Olston, G. Chiou, L. Chitnis, F. Liu, Y. Han, M. Larsson, A. Neu-mann,V.B.N. Rao, S. Seth, C. Tian, V. Sankarasubramanian, T. ZiCornell and X.Wang, Nova: Continuous pig/hadoop workflows, Proceedings of the ACMSIGMOD International Conference on Management of Data (SIGMOD11),pp. 1081-1090, 2011.
[03] W. Dou, X. Zhang, J. Liu and J. Chen, HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications, IEEE Transactions on Parallel and Distributed Systems, 26(2): 455-466, 2015.
[04] N. Laptev, K. Zeng and C. Zaniolo, Very fast estimation for result andaccuracy of big data analytics: The EARL system, Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), pp.1296-1299, 2013.
[05] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg and I. Brandic, Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility Future Generation Computer Systems 25(6): 599-616, 2009.
[06] K. Shim, MapReduce Algorithms for Big Data Analysis, In Proc. of theVLDB Endowment, 5(12): 2016-2017, 2012.
[07] C. Ji, Y. Li, W. Qiu, U. Awada and K. Li, Big Data Processing in Cloud Environments, 2012 International Symposium on Pervasive Systems, Algorithmsand Networks, 2012, pp. 17-23.
[08] W. Wang, D. Lu, X. Zhou, B. Zhang and J. Wu, Statistical Wavelet-based Anomaly Detection in Big Data with Compressive Sensing, EURASIP Journalon Wireless Communication and Networking, 2013.
[09] Bharath K. Samanthula and Wei Jiang, Secure Multiset Intersection Cardinality and its Application to Jaccard Coefficient IEEE Transactions on Dependable and Secure Computing.
[10] C. Yang, X. Zhang, C. Liu, J. Pei, K. Rama mohanarao and J.Chen, ASpatiotem-poral Compression based Approach for Efficient Big DataProcessing on Cloud, Journal of Computer and System Sciences (JCSS). vol.80: 1563-1583, 2014.
[11] L. Wang, J. Zhan, W. Shi and Y. Liang, In cloud, can scientific commu-nitiesbenefit from the economies of scale? IEEE Transactions on Parallel and Distributed Systems 23(2): 296-303, 2012.
[12] S. Sakr, A. Liu, D. Batista, and M. Alomari, A survey of large scale data management approaches in cloud environments, Communications Surveys and Tutorials, IEEE, 13(3): 311336, 2011.

ABOUT AUTHORS:

1. **S.Md.Mujeeb** is currently working as an Assistant Professor in Computer Science and Engineering Department, Malla Reddy Institute of Technology and Science, Hyderabad, Telangana. He received his M.Tech in Artificial Intelligence in 2010 from University of Hyderabad, Hyderabad.
2. **Raj Kumar Patil** is currently working as an Assistant Professor in Computer Science and Engineering Department, Malla Reddy Institute of Technology and Science, Hyderabad, Telangana. He received his M.Tech in CSE in 2014 from MRIET, Hyderabad.

