International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

A Scalable Data Chunk Similarity Based Compression Approach for Efficient Big Sensing Data Processing On Cloud

DR. Chandra Blessie[1], Suma K.S[2]
[1]*Associate Professor,* [2]*Research Scholar, Department of Computer Application, Nehru College of Management, Coimbatore, Bharathiyar University. Coimbatore – 46*

Abstract: Big sensing data is generated with high volume and velocity. Cloud computing provides a big sensing data processing and storage as it provides a flexible stack of massive computing, storage and software services in a scalable manner .Instead of compressing basic data units, the compression will be conducted over partitioned data chunks. This paper proposes scalable compression approach based on data chunk similarity that can significantly improve data compression efficiency with affordable data accuracy loss.

I. INTRODUCTION

A. Big Data

"Big data**"** is a term used to describe a collection of data sets with following features:

1) Volume-Large amounts of data generated.
2) Velocity-Frequency and speed of which data are generated, captured and shared
3) Variety-Multiplicity of data types and formats from a range of sources.

Traditional database management is not suitable for large size and complexity of big data. Data is being created in much sorter cycles from hours to milliseconds.

B. Big Data Analtics

To analyze employed to analyze, contextualize and visualize the data some analytical methods like data mining, natural language processing, artificial intelligence and predictive analytics are used. These computerized analytical methods are used to recognize inherent patterns, correlations and anomalies which are discovered as a result of integrating huge amounts of data from dissimilar datasets.

C. Big Data Computing

The rising importance of big-data computing systems from advances in many different technologies: Sensors: Digital data are being generated by many diverse sources. Computer networks: Using localized sensor networks, as well as the Internet data from the many different sources can be composed into very big data sets. Data storage: Advances in magnetic disk technology have noticeably decreased the cost of storing data. Cluster computer systems: The clusters of new form of computer systems provide both the storage capacity for large data sets, to organize the data, analyze it, and to respond to queries. Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model.

D. Big Data Analtics In Cloud Environment

Cerri et al proposed 'Knowledge in the cloud 'in place of 'data in the cloud' to support collaborative tasks which are computational intensive and facilitate the BIG DATA tribute, heterogeneous knowledge. This is termed as "utility computing" derived from required data in and out of cloud. The definitive application of cloud technology is as a large-scale data storage, development and processing system. But the ability of cloud computing has applications beyond effective use of data.

E. Moving Big Data Into Cloud

Cloud computing offers the promise of it data implementation to small and medium sized businesses. A programming paradigm known as MapReduce is used in Big data processing .Typically, implementation of the MapReduce model requires networked attached storage and parallel processing.

F. Key Technologies For Extracting Business Value From Big Data

Big data technologies depict a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discover and/or analysis. Big data will also intensify the need for data quality and governance, for embedding analytics into operational systems, and for issues of security, privacy and regulatory compliance.

II. SYSTEM ARCHITECTURE

A. Text Data

The proposed models are related to geometric model and common element approach in terms of numerical data and text data respectively. Dual variable length hidden Markov model is used and updated in our work for calculating similarity between text data.

B. Data Chunk Generation

A standard data chunks set will be generated based on our defined similarity model. This process is a preprocessing before the compression which normally does not require huge computation resource. In data chunks generation, there are two important inputs it sensing data set S and the maximum limitation ‘r’ or data generation control.

C. Data Chunk Formation

With the big data set preprocess, we aim to generate a standard initial set S’ which is used in the following compression process. According to the predefined similarity model, we need to generate the first initial standard data chunk S’ for the big data set S.

D. Compression

With the generated standard data chunks, a new data compression technique has been used which recursively compressing-coming data from big data set S according to generated S.

E. Predictions

With the increase of compression ratio from zero% to eighty%, the data accuracy decreases dramatically. It can be found that higher the R is, better the data accuracy can be achieved.

III. LITERATIRE SURVEY

To process big data with traditional data processing tools such as database, traditional compression, machine learning, or parallel and distributed system some techniques have been proposed. In the following section those current popular techniques for big data processing on cloud will be introduced and analyzed.

Now a day in big data processing on cloud, lots of big data sets or streams come from sensing systems which are widely deployed in almost every corner of our real world to assist our everyday life. In order to cope with that huge volume big sensing data, different techniques can have been developed, on-line or offline, centralized or distributed.Kienzler et al developed a “stream-as-you-go” approach for accessing and processing incremental big sensing data on cloud via stream-based data management architecture. The extension of traditional Hadoop framework was made to develop a novel framework named Incoop by incorporating several techniques like task partition and memorization-aware schedule. The stored big graph data or stream data sets will be queried and evaluated as the model of distributed data-base in cloud, such as “Hydoop” and its related “HIVE”,”HBase”,”Zookeeper” and so on. In the paper a spatial and temporal compression model is designed for compressing big sensing data with significant performance gains. The approach consists of two main technique parts. The first one focuses on reducing the data size over cloud platform with spatiotemporal compression.

IV. COMPARATIVE ANALYSIS

A. A Scalable Two-Phase Top-Down Specialization Approach

A large number of cloud services require users to share private data like electronic health records for data analysis or mining, bringing privacy concerns. A group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way.

B. Spatiotemporal Compression based Approach

It is well known that processing big graph data can be costly on cloud. Processing big graph data introduces complex and multiple iterations that rise challenges.

C. GPU Implementation of Orthogonal Matching Pursuit or Compressive Sensing

Recovery algorithms play a major role in compressive sampling. Currently, the orthogonal matching pursuit OMP, a popular recovery algorithm for Compressive Sensing, which possesses the merits of low complexity of each module in the OMP and point out the bottlenecks of the OMP lie in the projection module and the least-squares module.

D. Statistical Wavelet-based Anomaly

Anomaly detection in big data is a major problem in the big data analytics domain. In this paper, the definitions of big data and anomaly detection were presented.

E. Quasi identifier Index based Approach

Cloud computing provides storage capacity and massive computation power which enable users to deploy applications without infrastructure investment an approach to ensure privacy preservation and achieve high data utility over incremental and distributed data sets on cloud.

Title	Method/ Algorithm	Compression and accuracy comparison
A scalable two phase top-down specialization approach for data anonimization using systems, in MapReduce on cloud	A scalable Two-phase top-down specialization approach	Compression ratio for different ‘r’=80
A Spatiotemporal compression based approach for efficient big data processing on cloud	Spatiotemporal compression based approach	Compression ratio for different ‘r’=10
GPU implementation of orthogonal matching pursuit for compressive sensing	Orthogonal matching pursuit	Compression ratio for different ‘r’= 30
Statistical wavelet based Anomaly detection n big data with compressive sensing	Statistical wavelet based anomaly	Compression ratio for different ‘r’=50
An efficient Quasi identifier index based approach for privacy preservation over Incremental Data sets on cloud	An efficient Quasi identifier index based approach	Compression ratio for different ‘r’= 50

V. PROPOSED SYSTEM

A. In this paper, propose a novel technique based on data chunk partitioning for effectively processing big data, especially streaming big sensing data on cloud.

B. With the above data compression, we aim to improve the data compression efficiency by avoiding traditional compression based on each data unit, which is space and time costly due to low level traverse and manipulation.

C. At the same time, because the compression happens at a higher data chunk level, it reduces the chance for introducing too much usage of iteration and recursion which prove to be main problem in processing big graph data.

VI. CONCLUSION

In this paper, we proposed a novel scalable data compression based on similarity calculation among the partitioned data chunks with cloud computing. The MapReduce programming model was adopted for the algorithms implementation to achieve some extra scalability on cloud. With the real meteorological big sensing data experiments on our U-cloud platform, it was established that our proposed scalable compression based on data chunk similarity notably improved data compression performance gains with affordable data accuracy loss.

BIBLIOGRAPHY

[1] S.Tsuchiya, .Sakamoto,Y. Tsuchimoto and V.Lee, "Big Data Processing in cloud Environments", FUJITSU science and Technology Journal,
[2] "Big data: science in the petabyte era:community cleverness required" Nature
[3] M.Armbrust,A.Fox,R.Griffth,A.D.Joseph,R.Katz,A.Konwinski,G.Lee,D.Patterson,A.Rabkin,I.Stoica and M.Zaharia,"A view of cloud computing", communications of the ACM
[4] 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
[5] L. Wang, J.Zhan, W.shi and Liang, "In clouding, can scientific communities benefit from the economies of scale?" IEEE Transactions on parallel and Distributed Systems
[6] S.Sakr, A.Liu, D.Batista, and M.Alomari," A survey of large scale data management approaches in cloud environments", Commnications surveys and tutorials, IEEE
[7] B.Li, E.Mazur, Y.Diao,A.McGregor and P.Shenoy,"A platform for scalable one-pass analytics using mapreduce", in: Proceedings of the ACMSIGMOD international conference on management of data
[8] R.Kienzler, R.Bruggmann, A.Rangananathan and N.Tatbul,"stream as ou go: The case for incremental data access and processing in the cloud", IEEE CDE international workshop on data management in the clou
[9] C.Olston,G.Chou,L.Chintnis,F.Liu,Y.Han,M.Larsson,A.Neumann,V.B.N. Rao,V.Sankarasubramanian,S.Seth,C.Tian,T.Zicornell and X.Wang,"Nova:continuous pg/hadoop workflows",proceedings of the ACM SIGMOD international conference on management of data
[10] K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung and B.Moon,"parallel data processing with mapreduce: A survey", ACM SIGMOD Recor
[11] X.Zhang, C.Liu, S.Nepal and J.Chen, "An Efficient Quasiidentifier Index based Approach for privacy preservation over incremental data sets on cloud", Journal of computer and system sciences
[12] X.Zhang, C.Liu, S.Nepal, S.Pande and J.Chen, "A privacy leakage upper-bound constraint based approach for cost-effective privacy preserving of intermediate datasets n cloud",IEEE transactions on parallel and distributed systems
[13] X.Zhang, T.Yang, C.Lu and J.Chen, "A scalable two-phase top-down specialization approach for data Anonymization using systems in MapReduce on cloud", IEEE transactions on parallel and distributed
[14] W.Dou, X.Zhang, J.Liu and J.Chen, Hiresome-I: towards privacy-Aware cross-cloud service composition for big data applications, IEEE transactions on parallel and distributed systems
[15] J.Cohen,"Graph Twiddling in a MapReduce world", IEEE computing in science and Engineering.
[16] K.Shim,"MapReduce Algorithms for Big Data Analysis", Proc.of the VLDB Endowment.
[17] N.Laptev, K.Zeng and C. Zaniolo, "Very fast estimation for result and accuracy of big data analytics: The EARL system, "Proceedings of the 2th IEEE International conference on data engineering
[18] X.L. Dong and D.Srivastava,"Big data integration", Proceedings of the IEEE International conference on Data Engineering
[19] T.Condie, P.Mineiro, N. Polyzotis and M.Weimer,"Machine learning on Big Data", Proceedings of the IEEE International conference on data Engineering
[20] A.Aboulnaga and S.Abu,"Workload management for big data analytics", proceedings of the IEEE international conference on data engineering
[21] M.Yuriama and T.Kushida, "Sensor Cloud infrastructure", proceedings of the International conference on Network –Based information systems
[22] A.Alamr, W.S. Ansari, M.M.Hassan, M.S. Hossan, A.Alelawi, and M.A. Hossain," A survey on sensor-cloud: Architecture, Applications, and Approaches", International Journal of distributed sensor networks, vol
[23] C.Ji, .L, W.Qiu, U.Awada and K.L,"Big data processing in cloud environments", international symposium on pervasive systems, Algorithms and networks
[24] L.Wang, G. Von Laszewski, A.younge, X.He, M.Kunze,J.Tao,C.u,"clod computing: A perspective study", New generation computing
[25] X.Yang,L.Wang,G.Laszewski,"Recent research advances in e-science", cluster computing
[26] S.Sakr, A.Liu, D. Batista and M.Alomari,"A survey of large scale data management approaches in cloud environments, "IEEE communications surveys and tutorials
[27] S.Lattanzi, B.Moseley, S.Suri and S.Vassilvitskii,"Filtering: A method for solving graph problems in MapReduce", In Proc. ACM symposium on parallelism in algorithms and architectures,SanJose, California, USA
[28] K.shim," MapReduce Algorithms for big data analysis", n proc. Of the VLDB endowment
[29] N. Sidiropoulos and A.Krillidis, "Multi- Way compressed sensing for sparse low-rank tenors, "IEEE signal processing letters
[30] C.Yang, X.Zhang, C.Liu, J.Pei, K.Ramamohanarao and J.Chen, "A spatiotemporal compression based approach for efficient big data processing on cloud", Journalcomputer and system sciences
[31] L.Ramaswamy, Lawson and S.V.Gogineni,"Towards A Quality-centric big data architecture for federated sensor services", IEEE international congress on big data, pp.
[32] A.Cuzzocrea, G.Fortino and O.Rana,"Managing data and processes in cloud- enabled large-scale sensor networks: state-of-the-art and future research directions", proceedings of the IEEE/ACM International symposium on Cluster, cloud and grid computing, pp.
[33] Fang, L.chen,J.Wu and B.Huang, "GPU implementation of orthogonal matching pursuit for compressive sensing", Proceeding of the IEEE international conference on parallel and distributed sstems,IEEE computer society

[34] Wang, D.Lu, X.Zhou,B.Zhang and J.Wu, "Statistical wavelet based anomaly detection n big data with compressive sensing,"EURASIP journal on wireless communication and networking,
[35] J.Wamg,S.Tang,B.Yin and X.Li, "Data gathering in wireless sensor networks through intelligent compressive sensing", Proceedings IEEE NFOCOM, pp
[36] S.H. YOON and C.Shahabi ,"An experimental study of the effectiveness of clustered aggregation leveraging spatial and temporal correlations in wireless sensor networks,"ACM transactions on sensor networks, vol
[37] R.Qiu and M.Wicks ,"Cognitive networked sensing and big data", ISBN
[38] Real time big data processing with gridgain.http://WWW.gridgan.com/sitemap/,accessed on November