ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

An Efficient Compression Method to Reduce the Data Transfer Cost in the Cloud

S.Brintha Rajakumari, Dr.C.Nalini, Sundararajan.M, Arulselvi S

Assistant Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

Director, Research Center for Computing and Communication, Bharath University, Chennai, Tamil Nadu, India

Co-Director, Research Center for Computing and Communication, Bharath University, Tamil Nadu, India

ABSTRACT*:* Cloud computing is a model for enabling suitable, on-demand network access to a shared pool of configurable computing resources like networks, servers, storage, applications, and services which will be speedily provisioned and free with token management effort or service supplier interaction. Data compression implies sending or storing a smaller number of bits. It is achieved by removing data redundancy while preserving information content. In this paper we have analyzed how effectively the online PHP gzdecode compression function used to reduce the cost of data transferred in the cloud with cancer data set.

KEYWORD*S:* Cloud computing, Data Compression.

I. INTRODUCTION

Data storage space is virtually infinite in cloud computing. To transfer data in the cloud requires more cost because it is a pay and use model and also the resources' based on budget constraints. Cloud computing is an Internet-based computing solution where computers are configured to work together and the various applications use the collective computing power as if they are running on a single system. It can be defined as the collection of hardware, storage and networks. Common examples embody Amazon Elastic cipher Cloud, Microsoft Azure and Google App Engine.

Data compression is a way to reduce storage cost by eliminating redundancies that happen in most files. Although many methods are used for this purpose, in general, this strategy is divided into two broad categories: lossless and lossy strategies. Lossy compression reduced file size by eliminating some unneeded data that won't be recognized by humans after decoding, this often used by video and audio compression. Lossless compression on the other hand, manipulates each bit of data in a file to minimize the size without losing any data after decoding. Redundant data is removed during compression, and added during decompression. Lossless compression methods are normally used when we cannot afford to lose any data. This is important because if file lost even a single bit after decoding, that mean the file is corrupted [3][4].

In this paper, section 2 describes the related research work. In section 3 presented the benefit of compression. In section 4 contains the compression details and section 5 presented the resultant analysis.

II. RELATED WORKS

Organizations continue to store more and more data in cloud environments, which represent an immense, valuable source of information to mine. Plus, clouds offer business users, scalable resources on demand. [1][2]

Compression is used just about everywhere. The task of compression consists of two components, an encoding algorithm that takes a message and generates a "compressed" representation which contains fewer bits and a decoding algorithm that reconstructs the original message or some approximation of it from the compressed representation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

These two components are typically intricately tied together since they both have to understand the shared compressed representation [8][9].

Most cloud providers charge for incoming and outgoing bandwidth. Compressing content prior to uploading it to the cloud can save a measurable amount of time and money in the form of bandwidth costs. By way of example, and depending on the file type, compression ratios typically run anywhere from 25% to 80% when enabling compression from within a StoragePoint Storage Profile.

Achieving substantial application performance gains through compression requires a good compression algorithm and a system architecture that is designed for performance. The compression system must precisely match repetitive patterns to achieve high compression ratios. When possible, the most efficient compression algorithm based on the network link should be applied automatically. This system must manage stored data and incoming application traffic to maximize effectiveness, and it should optimize and accelerate the performance of applications commonly accessed via a WAN link. Finally, this system must do all this quickly to minimize latency and continue to fill the network. WAN optimization to combat the challenges of assuring application performance and help ensure timely transfer of large data sets across constrained network links [2][3].

The paper [3][4] provide a definition of what a cloud database is, and in the light of that definition, examine the suitability of Algebraix Data's technology to fulfill the role of a cloud database.In the paper [5][6] proposed three key drivers of consolidation in the cloud space: 1) the need for specific intellectual property, technological capability, or human capital; 2) the desire to build increased scale in an already established cloud based model; or (3) the need to diversify a business mix to support a weakening traditional revenue stream.

In the paper [6][7] cloud architectures address key difficulties surrounding large-scale data processing. In traditional data processing it is difficult to get as many machines as an application needs. Second, it is difficult to get the machines when one needs them. Third, it is difficult to distribute and co-ordinate a large-scale job on different machines, run processes on them, and provision another machine to recover if one machine fails. Fourth, it is difficult to auto-scale up and down based on dynamic workloads. Fifth, it is difficult to get rid of all those machines when the job is done. Cloud Architectures solve such difficulties.

Rex Wang proposed [7][8] "Many data centers are migrating toward the cloud, or at least a cloud-like model, so that they may take advantage of the pooled resources that form the fundamental core of cloud computing". "By pooling or sharing resources across multiple applications, cloud delivers the efficiency and increased flexibility data centers require helping the business be marketed-competitively. For data processing and analysis, having a shared, standardized, and consolidated database architecture for all DW and OLTP workloads is an effective strategy.

Cloud based platforms, have become an increasingly viable option that many organizations are exploring. However, each organization must determine if cloud based storage is a good option for them. SharePoint can be a good starting point to explore the cloud possibilities. Without StoragePoint, the cloud options are not quite as appealing. By leveraging StoragePoint, organizations can make a move to the cloud in gradual steps, while at the same time knowing that many of the concerns that cloud based computing present are alleviated by taking advantage of compression and encryption [9].

III. AWS PRICE IN THE CLOUD

Many providers supply resources to store the data in the cloud[11][12]. The Amazon Web Services (AWS) also provide pricing model for the cloud resources. Cloud Service Providers (CSPs) provide a set of resources such as CPU, storage, networks, development platforms and services. This article uses the AWS EC2 storage bandwidth model [13][14] is in the following table1. Bandwidth consumption is billed with respect to data volume. In this model, input data transfers are free, whereas output data transfer cost varies with respect to data volume, with an earned rate when volume increases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

DATA VOLUME	PRICE PER MONTH
Input data	Any input data Free
Output data	First 1 GB Free
Up to 10 TB	$0.12 per GB
Next 40 TB	$0.09 per GB
……..	………

TABLE 1: AMAZON Bandwidth PRICES

IV. ONLINE GZIP COMPRESSION FUNCTION

This online tool [9][10] encodes or decodes data using Zlib (RFC 1950) or GZIP format 4.3 (RFC 1952). GZIP compression is a widely used method of compressing web pages and other text data sometimes it can reduce the size of the source document by 70-80% at the same time requiring little CPU processing[11][12]. On that page one can encode or decode GZ data online. Anyone can do so by direct input, by file upload or by URL. Also, different variants of GZ compression are supported.

GZIP-compatible encoding means that your data is being compressed or uncompressed. The stream has all the necessary headers so it'll be larger by a dozen of bytes. This is implemented using PHP gzencode or gzdecode functions.

One can use that mode to extract GZ archives and select the checkbox name GZIP-compatible encoding then select yours.gz file to upload, press the Decompress button and get its contents. The PHP gzcompress or gzdecompress functions decompressing the source string back.

V. EVALUATION OF METHOD

The sample dataset collected from the http://www.theguardian.com/news/datablog 2011/Dec/07/ cancer-causes-list is in Figure 1. The cancer table has 41 records of attribute which contains Oral cavity and pharynx, Oeso-phagus, Stomach, Colon-rectum, Liver, Pancreas, Gall-bladder, Larynx, Lung, Meso-thelioma, Melanoma, Breast, Cervixuteri, Corpus uteri, Ovary, Bladder, Kidney, Leukaemia The risk factor percentage will be calculated based on tobacco, Alcohol, Fruit and Vegetables, Meat, Fibre, salt, overweight and obesity, physical exercise, infections, radiation in ionizing and UV, occupation, Post-menopausal hormones and reproduction (breast feeding).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	% cancers attributable to risk factor exposure, by cancer site																				
2	Exposure	Gender	cavity & pharynx	Oeso-phagus	Stomach	Colon-rectum	Liver	Pancreas	Gall-bladder	Larynx	Lung	Meso-thelioma	Melanoma	Breast	Cervix uteri	Corpus uteri	Ovary	Bladder	Kidney	Leukaemia	Alla
6	Meat	Male	-	-	-	24.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3.5
7	Fibre	Male	-	-	-	10.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.4
8	Salt	Male	-	-	30.9		-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.9
9	Overweight &	Male	-	26.9	-	13.6	-	12.8	19.7	-	-	-	-	-	-	-	-	-	25	-	4.1
10	Physical exer	Male	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.4
11	Infections	Male	12.3	-	29.2	15	19.6	-	-	10.6	-	-	-	-	-	-	-	-	-	-	2.5
12	Radiation - ic	Male	-	2	0.9	1.1	0.6	-	-	-	4.2	-	-	-	-	-	-	2.6	-	7.8	1.7
13	Radiation - U	Male	-	-	-	-	-	-	-	-	-	-	89.8	-	-	-	-	-	-	-	3.5
14	Occupation	Male	0.6	3.3	3	-	0.2	0.02	-	2.9	20.5	97	-	-	-	-	-	7.1	-	0.9	4.9
15	All of the abc	Male	**92.9**	**89.7**	**78.1**	**56.5**	**48.6**	**35.7**	**19.7**	**92.8**	**91.1**	**97**	**89.8**	-	-	-	-	**43.5**	**47**	**16.4**	**45.3**
16	Tobacco	Females	54.9	71.3	15.4	9.9	15.3	31	-	79.1	83.6	-	-	-	7.2	-	2.6	34.4	15	3	15.6
17	Alcohol	Females	16.9	11.3	-	6.9	5	-	-	12.2	-	-	-	6.4	-	-	-	-	-	-	3.3
18	Fruit & veg	Females	53.6	45.1	33.9	-	-	-	-	40.5	9.3	-	-	-	-	-	-	-	-	-	3.4
19	Meat	Females	-	-	-	16.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.9
20	Fibre	Females	-	-	-	14.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.7
21	Salt	Females	-	-	12.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2
22	Overweight &	Females	-	11.2	-	12.2	-	11.5	17.8	-	-	-	-	8.7	-	33.7	-	-	22.2	-	6.9
23	Physical exer	Females	-	-	-	3.6	-	-	-	-	-	-	-	3.4	-	3.8	-	-	-	-	1.7
24	Post-menop	Females	-	-	-	-	-	-	-	-	-	-	-	3.2	0	12	0.7	-	-	-	1.1
25	Infections	Females	14	-	36	3.1	9.3	-	-	10.6	-	-	-	-	100	-	-	-	-	-	3.7
26	Radiation - ic	Females	-	3.9	1.7	2.2	1.1	-	-	-	5.4	-	-	0.9	-	-	-	2.3	-	10.4	2
27	Radiation - U	Females	-	-	-	-	-	-	-	-	-	-	82.4	-	-	-	-	-	-	-	3.6
28	Occupation	Females	0.2	1.1	0.3	-	0.1	-	-	1.6	4.3	82.5		4.6	0.7	-	0.5	1.9	-	0.5	2.4
29	Reproductio	Females	-	-	-	-	-	-	-	-	-	-	-	3.1	-	-	17.6	-	-	-	1.7
30	All of the abc	Females	**85**	**88.2**	**69.2**	**51.9**	**28**	**38.9**	**17.8**	**90.9**	**86.5**	**82.5**	**82.4**	**26.8**	**100**	**36.9**	**20.7**	**37.1**	**33.9**	**13.6**	**40.1**
31	Tobacco	Persons	64.5	65.5	22.2	8.1	23	28.7	-	79	85.6	-	-	-	7.2	-	2.6	36.7	24.1	6.2	19.4
32	Alcohol	Persons	30.4	20.6	-	11.60	9.1	-	-	24.6	-	-	-	6.4	-	-	-	-	-	-	4
33	Fruit & veg	Persons	56	46.1	35.8	-	-	-	-	45.4	8.8	-	-	-	-	-	-	-	-	-	4.7
34	Meat	Persons	-	-	-	21.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2.7
35	Fibre	Persons	-	-	-	12.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.5
36	Salt	Persons	-	-	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.5
37	Overweight &	Persons	-	21.7	-	13	-	12.2	18.3	-	-	-	-	8.7	-	33.7	-	-	24	-	5.5
38	Physical exer	Persons	-	-	-	3.3	-	-	-	-	-	-	-	3.4	-	3.8	-	-	-	-	1
39	Post-menop	Persons	-	-	-	-	-	-	-	-	-	-	-	3.2	0	12	0.7	-	-	-	0.5
40	Infections	Persons	12.7	-	31.7	2.2	15.9	-	-	10.6	-	-	-	0	100	-	-	-	-	-	3.1
41	Radiation - ic	Persons	-	2.7	12	1.6	0.8	-	-	-	4.7	-	-	0.9	-	-	-	2.5	-	8.9	1.8
42	Radiation - U	Persons	-	-	-	-	-	-	-	-	-	-	85.9	-	-	-	-	-	-	-	3.5
43	Occupation	Persons	0.5	2.6	2		0.2	-	-	2.7	13.2	94.4	-	4.6	0.7	-	0.5	5.7	-	0.7	3.7
44	Reproductio	Persons	-	-	-	-	-	-	-	-	-	-	-	3.1	-	-	17.6	-	-	-	0.9

Figure1: Cancer Data Set

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

The given dataset uploaded to the online site and it compresses and gives the reduced dataset It is in the figure2 and 3 respectively. It has been applied to the cloud will reduce the cost of storage in the cloud.

Figure 2: Online Data Process Screen

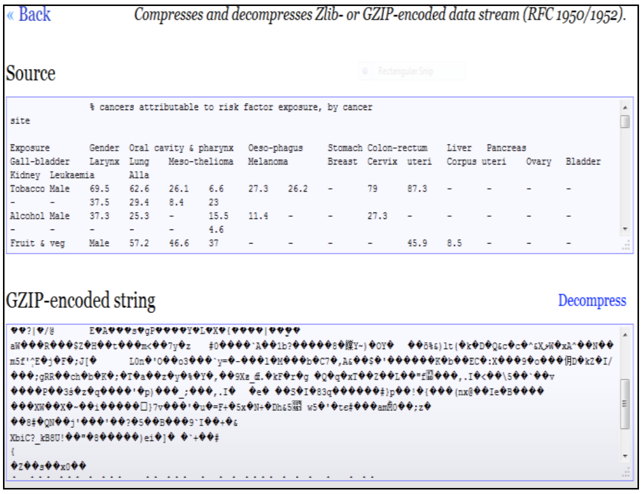

Figure3: Compressed Data

The size of the original cancer data set was 12.6 KB but the compressed data size is 9.2KB. When we transfer the compressed data set to the cloud will reduce the bandwidth cost of data. It not only reduce the transfer cost but also reduce the storage cost.

VI. CONCLUSION

This paper used the online compression algorithm that can be used to reduce the transfer cost of data in the cloud. An experiment by using small cancer data set with 41 different types of records. It has been tested in the online compression function. This online function gives better compression ratio and reduces the cost of storage and transfer cost in the cloud. In future we will analyze this method using large dataset.

REFERENCES

1. Intel IT Center Solution Brief | Big Data in the Cloud:Converging Technologies, 2013.
2. Sathish Kumar M., Karrunakaran C.M., Vikram M., "Process facilitated enhancement of lipase production from germinated maize oil in Bacillus spp. using various feeding strategies", Australian Journal of Basic and Applied Sciences, ISSN : 1991-8178, 4(10) (2010) pp. 4958-4961.

3. Lori MacVittie, "Understanding Advanced Data Compression",2013.
4. Kaliyamurthie K.P., Parameswari D., Udayakumar R., "QOS aware privacy preserving location monitoring in wireless sensor network", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S5) (2013) pp.4648-4652.
5. I Made Agus Dwi Suarjaya, A New Algorithm for Data Compression Optimization, International Journal of Advanced Computer Science and Applications, Vol. 3, No.8, 2012.
6. Sharmila D., Muthusamy P., "Removal of heavy metal from industrial effluent using bio adsorbents (Camellia sinensis)", Journal of Chemical and Pharmaceutical Research, ISSN : 0975 – 7384, 5(2) (2013) pp.10-13.
7. Robin Bloor, "What Is A Cloud Database ?The Suitability of Algebraix Data's Technology to Cloud Computing",2011.
8. Udayakumar R., Khanaa V., Saravanan T., Saritha G., "Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction", Middle - East Journal of Scientific Research, ISSN : 1990-9233, 16(12) (2013) pp.1781-1785.
9. Morgan Stanley, "Cloud Computing Takes Off",may 2011.
10. Kalaiselvi V.S., Prabhu K., Ramesh M., Venkatesan V., "The association of serum osteocalcin with the bone mineral density in post menopausal women",Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 7(5) (2013) pp.814-816.
11. Jinesh Varia, "Cloud Architectures" , 2008.
12. Rex Wang, Data Center Best Practices: Managing Data With Cloud Computing,Oracle.
13. Guy E. Blelloch, "Introduction to Data Compression",2013.
14. Chris Geier ,"Leveraging Cloud Storage with SharePoint and StoragePoint", 2010.
15. http://i-tools.org/gzip
16. from http://proger.i-forge.net/gzdecode
17. Thi-Van-Anh Nguyen, Laurent d'Orazio, "Cost Models for View Materialization in the Cloud", ACM, 2012.
18. Amazon ec2. http://aws.amazon.com/ec2/,2012.
19. Jemima Daniel,The world of illusion in Tennessee William's "The Glass Menagerie",International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 6183-6185 ,Vol. 2, Issue 11, November 2013.
20. Jemima Daniel,Themes of Violence, Horror, Death in Hemingway,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 4500-4503,Vol. 2, Issue 9, September 2013.
21. Jemima Daniel,Role of Technology in Teaching Language,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753, pp 2287-2283,Vol. 2, Issue 6, June 2013.
22. Jemima Daniel,Optimism in Samuel Beckett's Waiting for Godot,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753,pp 5467-5470,Vol. 2, Issue 10, October 2013.
23. Jemima Daniel,Treatment of Myth in Girish Karnad`S Play the Fire and the Rain,International Journal of Innovative Research in Science, Engineering and Technology ,ISSN: 2319-8753, pp 1115-1117 ,Vol. 2, Issue 4, April 2013.
24. Jemima Daniel,Audio-Visual Aids in Teaching of English,International Journal of Innovative Research in Science, Engineering and Technology,ISSN: 2319-8753, pp 3811-3814,Vol. 2, Issue 8, August 2013.
25.