Information Upload and retrieval using SP Theory of Intelligence

Supriya P, Koushik S

Dept of ISE, MS Ramaiah Institute of Technology, Bangalore, India

Abstract— *In today's technology Cloud computing has become an important aspect and storing of data on cloud is of high importance as the need for virtual space to store massive amount of data has grown during the years. However time taken for uploading and downloading is limited by processing time and thus need arises to solve this issue to handle large data and their processing. Another common problem is de duplication. With the cloud services growing at a rapid rate it is also associated by increasing large volumes of data being stored on remote servers of cloud. But most of the remote stored files are duplicated because of uploading the same file by different users at different locations. A recent survey by EMC says about 75% of the digital data present on cloud are duplicate copies. To overcome these two problems in this paper we are using SP theory of intelligence using lossless compression of information, which makes the big data smaller and thus reduces the problems in storage and management of large amounts of data.*

Keywords— Cloud Computing, Big data processing, Data De-duplication, SP theory of intelligence, Lossless Compression.

I. INTRODUCTION

The SP theory (Simplicity and Power theory) design is to simplify and interface concepts across artificial intelligence, conventional computing and human perception and cognizance, and understands compression of information via the matching and unification of patterns. SP Theory is accomplished as a hypothetical system corresponding to brain which can receive new information and stores it by relating it to already available Old information.

The SP theory of intelligence amalgamates visionary clarity with explanatory and description power. In the SP machine there is a capability for simplifying the computation and also saves time, cost and effort involved in the development of many applications.

SP is short for Simplicity and Power, as it may be seen as a process of reducing informational redundancy and thus increasing its simplicity while retaining as much as possible of its non-redundant expressive power.

Majority of the computing tasks in today's situation involves data that have been collected and stored in databases. The data make a stationary target. But, increasingly vital important insights can be gained from analyzing information that's on the move. This approach is called streams analytics [1]. Rather than placing data in a database first, the computer analyses it as and when it comes from a wide amount of sources, continuously filtering its understanding of the data as and when conditions change. This is the way how human process their information. Although, in its unsupervised learning, the SP system processes information by dividing them into batches, and thus lends itself to an incremental approach. The SP system is designed to incorporate new information to a constantly growing body of compressed old information.

Massive amount of large data sets introduces problems of data management [2]. In order to reduce required amount of storage it is necessary to reduce the time for computation and also represent data in desired way.

One major roadblock to using cloud services for processing large data is the problem of transmitting the data sets over a network. Maintaining communications network is turning out to be very expensive and marginally profitable. In order to minimize these network charges system designers have figured out a way to minimize the energy used for processing data. The SP system advocates the efficient transmission of data by dividing the information into smaller parts.

SP theory in managing massive data[3]- Large scale data sets introduce many problems for data management.

a. Volume: Size of the large scale data sets can be reduced by compressing the information into chunks and by identifying the duplicate chunks, storage of identical chunks can be eliminated and hence efficiency in storing can be achieved.

b. Variety: Each format of information requires different kind of processing. Text files with different formats .txt, .pdf,.doc, each one need to be analyzed differently. SP System provides a universal framework for processing of diverse formats.

c. Velocity: Instead of simply placing the data in the database, it requires data to be analyzed first and understand its content first. This way transmission time taken to analyze the moving data can be minimized.

II. LITERATURE SURVEY

Wolff et.al [1] explained how the SP theory of intelligence and its applicability in the SP machine may be applied to the processing and management of big data.

J Gerard Wolff et.al [3] this article is an overview of the SP theory of intelligence is designed to simplify and interface concepts across artificial intelligence, conventional computing and human perception and cognizance, with information compression as a theme. It is understood as a human brain that receives New data and stores it in by compressing it as Old information; and it is envisioned in the form of a computer model, a first version of the SP machine. The matching and unification of data patterns and the concept of multiple alignments are the ideas behind the theory.

J Gerard Wolff et.al [5] provides confirmation for the idea that much of artificial intelligence, human perception and cognizance, conventional computing, may be understood as compression of information via the matching and indication of patterns. This is the basis for the SP theory of intelligence, outlined in the paper and fully described elsewhere.

Robert Escriva et.al [12] this paper presents HyperDex which is understood as a distributed key-value cache that provides a exclusive search primitive that empowers queries on secondary attributes. The key concept of HyperDex is the idea of hyperspace hashing in which objects having multiple attributes are located onto a multidimensional hyperspace. This scaling leads to productive implementations for searches of fractionally-specified secondary attribute and range queries and also for retrieval by primary key.

J Gerard Wolff .et.al [13] describes existing and expected benefits of the SP theory of intelligence, and some potential applications. The theory is designed to simplify and interface ideas across artificial intelligence, conventional computing, and human perception and cognizance, with information compression as a theme. It incorporates simplicity of both explanatory and descriptive power in numerous areas of computation and cognizance.

Kruus et al. [20] presents a work similar to ours. They propose a chunking algorithm involving two stages that re-chunks transitional and non-duplicated big CDC chunks into small CDC chunks. The significance of their work is to reduce the number of chunks while attaining as the same duplicate elimination ratio as a baseline CDC algorithm.

III. PROPOSED SYSTEM

Figure1 represents the system architecture of proposed system which is concerned with establishing basic structural framework for a system. When a file is uploaded by the user say File X, text file is divided into blocks for each block a hash code is generated. Each time when a new file say File Y is uploaded it compares with hash code for duplication and then writes the file for cloud storage.

A. Information compression and Block Creation

The main aim of this project is to overcome the problems in big data using the SP Theory of Intelligence. In order to achieve this goal, big data is subjected to compression techniques. Compression of information is achieved by pattern matching. Using such a system leads to the improvement in the processing of big data. The SP Theory provides pattern recognition, information storage, retrieval and information compression.

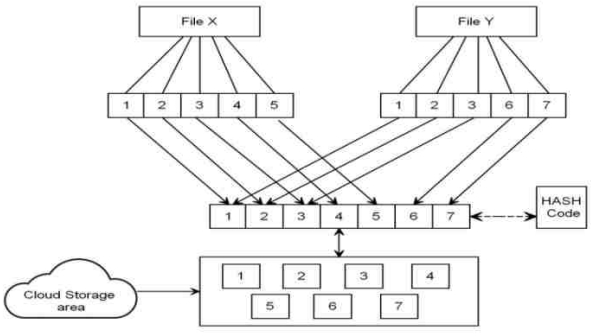

Fig.1: Architecture of proposed system

Figure 2 represents the block diagram of our proposed sp theory of intelligence system. Our proposed system has two modules user and admin. In first stage using logging details user and admin can log into the system. In second stage user can upload a text file for processing, in block creation process divides the text into blocks of fixed or variable size. In last stage de duplication process calculates hash-function for each block and compares hash result with already stored index for duplication detection and update index and store data.

The process of block creation or chunking divides the data stream into smaller, non overlapping blocks. There are different approaches for block creations static chunking, content defined chunking and file-based chunking. In our system we are using content defined chunking method, where chunks will be generated based on their content and the calculation of one fingerprint for each substring of length w i.e., one fingerprint for each word and processing overhead for fingerprint typically depends on string length, if small string length impact good performance, but bad chunking properties and large string length impact good chunking properties.

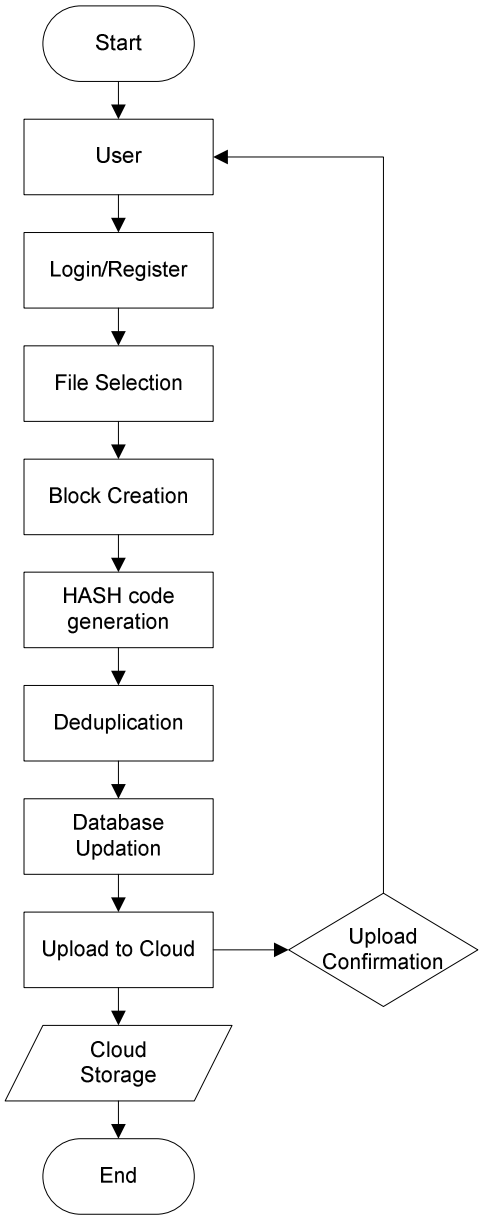

Fig.2: Flowchart of Proposed System

B. *Hash code generator and data duplication*

Once the index is created for each block the duplication or the redundancy of a chunk is checked by the de-duplication system. The chunk is added to the system only if there are no redundant copies of it. If it is not there then its data is stored. Here we perform Block-level de-duplication as the name indicates, the block level deduplication deals with the elimination of redundant data copies with respect to blocks. The procedure involves dividing files into blocks and storing a single copy of each block. It uses either fixed-sized blocks or variable-sized chunks for de-duplication at block level. For the purpose of efficient de-duplication scheme, the stored chunks are given a "chunk-index" which contains the fingerprint of all the stored chunks.

This paper is based on the idea that if the fingerprint of a chunk is already present in the chunk index, then it is believed that the data in the the new chunk and the already existing chunk are identical and thus it is instantly classified and governed as duplicate. The chunks which are classified as new are then stored by the de-duplication system by assigning as new index. Content-defined chunking is found to give high de-duplication ratio [14][15] for having the backup of workloads. Hence, it is the most widely used methodology in most de-duplication systems for backup workloads [10, 12, 13, 14, and 15].

IV. RESULTS

Proposed sp theory of intelligence system was tested for its big data handling and de-duplication efficiency .Figure 3 shows how users have uploaded a new files and figure 4 shows the blocks created of the original file and blocks uploaded onto the cloud.

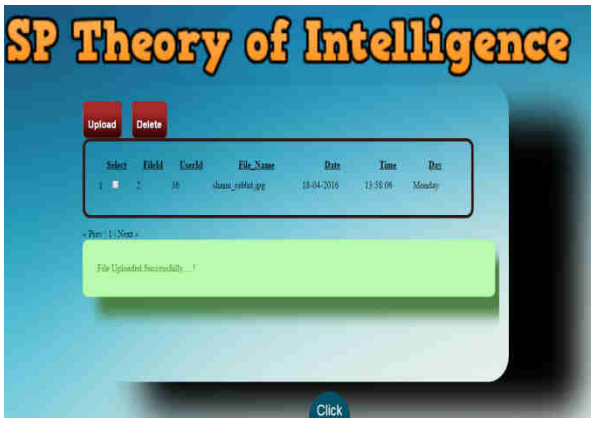

Fig.3: Upload a image file

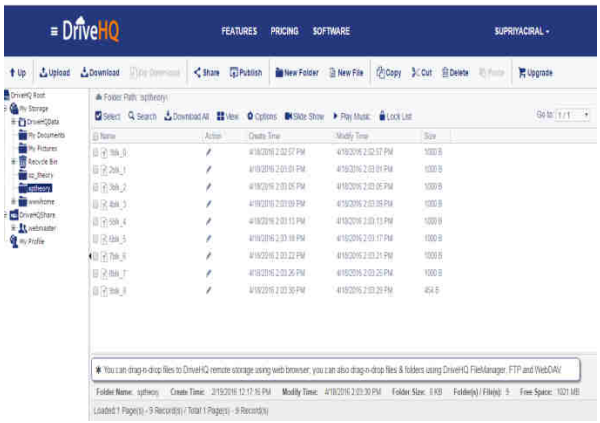

Fig.4: File uploaded on to cloud as blocks

Figure 5 shows the database which stores all the uniquely generated content index of the blocks. Each block is assigned a content index based on its content. Since a single file is uploaded and there is no duplication, the instance is updated as 1.

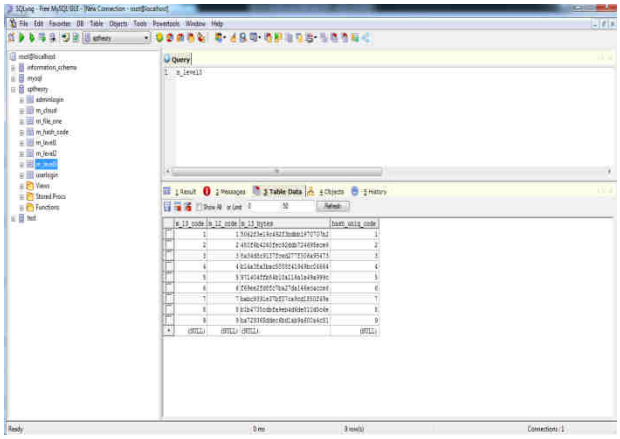

Fig.5: Content index generated for each blocks of the file

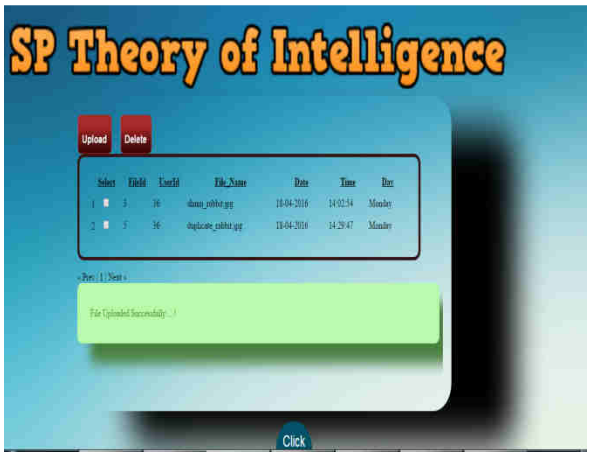

Fig.6: Upload the same image with a different name

Figure 6 shows a new file uploaded. New file is the duplicate of the already uploaded file. Though both the file names are different, content of both files are same and hence the same index is generated for the new file uploaded as the old file. Figure 7 shows the duplicate blocks are not uploaded on to cloud and hence saving the storage space.

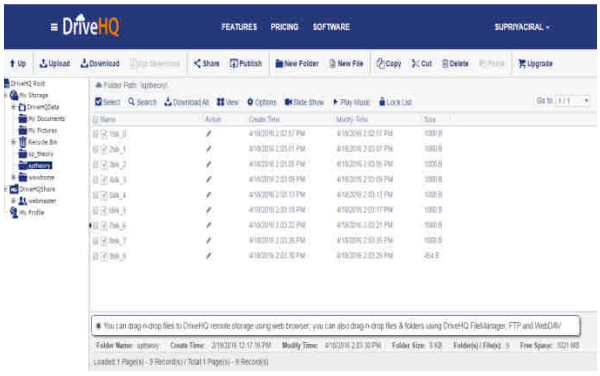

Fig.7 : Duplicate blocks are not uploaded to the cloud

Figure 8 shows the content index stored in the database. The index generated for the duplicate file is same as the index generated for the first file. Since the generated index are same, the instance of the content index is incremented rather than again storing the same index.

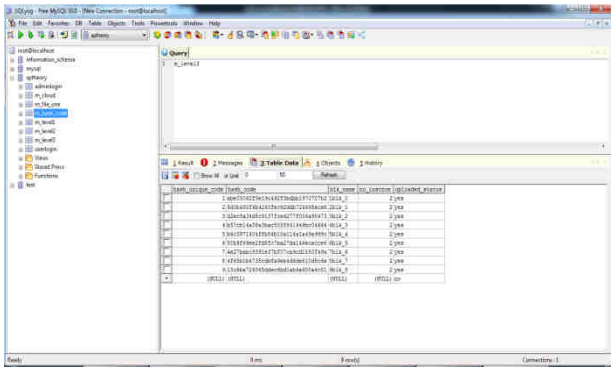

Fig.8: Instance of Content index increased from 1 to 2

V. CONCLUSION

Distributed computing has turned into a vital viewpoint in today's world. Distributed computing has carried with it a few difficulties like security, stockpiling, booking and so on. Capacity in Cloud processing frames a vital part as the need of virtual space to store our expansive information has become over these years. In any case, the pace of transferring and downloading limits the preparing time and there is a need to settle this issue of extensive information taking care of. Pressure procedures are much dependable strategy to diminish the space over cloud as levels of popularity for computerized information prompts wrong utilization of distributed storage. It is required to get lossless pressure the majority of time since distributed storage manages information utilized as a part of constant applications, continuous picture preparing.

VI. ACKNOWLEDGMENT

The authors are very grateful toDr.VijayKumar, HOD, Dept of ISE MSRIT, India for interesting discussions regarding this work

REFERENCES

[1] Wolff, James Gerard. "Big data and the SP theory of intelligence." Access, IEEE 2 (2014): 301-315.
[2] Albus, James S. "Outline for a theory of intelligence." Systems, Man and Cybernetics, IEEE Transactions on 21.3 (1991): 473-509.
[3] Wolff, J. Gerard. "The SP theory of intelligence: benefits and applications."Information 5.1 (2013): 1-27.
[4] Wolff, J. Gerard. "Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision."SpringerPlus 3.1 (2014): 552-570.
[5] Wolff, J. Gerard. "Information compression by multiple alignment, unification and search as a unifying principle in computing and cognition." Artificial Intelligence Review 19.3 (2003): 193-230
[6] J. G. Wolff. Proposal for the creation of a research facility for the development of the SP machine. Technical report, Cognition Research.org, 2015.

[7] S. Watanabe, editor. “Frontiers of Pattern Recognition. Academic Press,NewYork”, 1972.
[8] J. G. Wolff.” Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision. SpringerPlus,” 3(1):552{570, 2014.
[9] K. Pearson. The Grammar of Science. Walter Scott, London, 1892 Republished by Dover Publications, 2004, ISBN 0-486-49581-7.
[10] H. B. Barlow.” Sensory mechanisms, the reduction of redundancy, and intelligence”. In HMSO, editor, The Mechanisation of Thought Processes, pages 535{559. Her Majesty's Stationery Oce, London, 1959.
[11] .”National Research Council, Frontiers in Massive Data Analysis, National Academies Press”, 2013).
[12] Robert Escriva.” HyperDex: A Distributed, Searchable Key-Value Store for Cloud Computing”.
[13] J Gerard Wolff.” The SP Theory of Intelligence: Benefits and Applications”. Information 2014.
[14] Erik Kruus, Cristian Ungureanu, Cezary Dubnicki. Bimodal Content Defined Chunking for Backup Streams. In Proceedings of 8th
[15] USENIX Conference on File and Storage Technologies. Feb. 2010.
[16] Big Data for Development: Challenges and Opportunities, Global Pulse, May 2012.\
[17] J Gerard Wolff” The SP theory of intelligence and the SP machine “.January 13, 2015
[18] Wolff, J.G. Unifying Computing and Cognition: The SP Theory and its applications; CognitionResearch.org: Menai Bridge, UK, 2006.
[19] Wolff, J.G. Application of the SP theory of intelligence to the understanding of natural vision and the development of computer vision 2013. CognitionResearch.org, Menai Bridge, UK, unpublished work, 2013.
[20] J. G. Wolff. Towards an intelligent database system founded on the SP theory of computing and cognition. Data & Knowledge Engineering, 60:596{624, 2007)
[21] J. G. Wolff. Simplicity and power, some unifying ideas in computing. Computer Journal, 33(6):518{534, 1990.
[22] J. G. Wolff. Medical diagnosis as pattern recognition in a framework of information compression by multiple alignments, unification and search. Decision Support Systems, 42:608{625, 2006b
[23] J. G. Wolff. Towards an intelligent database system founded on the SP theory of computing and cognition. Data & Knowledge Engineering, 60:596{624, 2007.