# Data Archival using Hadoop Distributed File System

Asst.Prof .Kajal J. Jewani[1], Ms.Deepika Singh T [2], Mr. Neil George[3], Mr.Parth Mehta [4], Mr. Harsh Basantani[5.]

*Department of Computer Engineering,*
*V.E.S. Institute of Technology,HAMC Complex, Chembur, Mumbai-400 074*

***Abstract*-In today's world there is data and information explosion to which Hadoop is best suited taking into account the reasonable cost, up gradation, scalability, fault tolerance and devoid of schema. Large organizations all over the world find data analytics a daunting task and seek the help of experts so that they can retrieve the data immediately whenever required. One such good example is the RTI Act 2005 wherein large amount of data has to be sifted to give cogent and meaningful replies. Organizations by convention or law are required to maintain data for a minimum of 7 years and much more. Retrieving data at the press of button is proven to be the Achelees heel for corporates. This paper makes a cogent attempt to show as to how to retrieve data generated by social, print, film, history, research and a host of other activities by application of ESIEE.**

***Keywords:-Hadoop, Data Archival., HDFS.Data Mining, Big Data, Business scalability.***

## I. INTRODUCTION

Big data analytics is the requirement in today's business world. The colossal amount of data- generation is bound to increase exponentially in future. Even as of now if we take into account the data generated in 2005 which was 130 exabyte, but quickly grew to 4.4 zettabytes of memory by 2013. As per the latest estimates it will increase ten-fold to 44 zettabytes of memory by 2020. To excel and to be more successful the best way is to retrieve the large amount of data in the shortest possible time, and to take data driven decisions which on an average are 5% more productive the proponents and the users of this system are 6% more successful than others. World renowned MIT has done an in depth study and found that undertakings/corporates which used big data analytics were in the top. This was borne out by the fact that companies that were in the top third tier of their industry in terms of the use of data driven decisions were, were more productive with effective use and maintenance of big data which gave them an edge over competitors. All this comes at a huge cost as both retrieval and maintaining such huge amount of data and then using it effectively comes at a huge cost. The best way to utilize big data analytics is to use the best method of data archiving and later onto use the cheapest and the most effective method of retrieving it. .

## II. CHALLENGES OF TRADITIONAL DATA ARCHIVAL

Archiving means moving historical data to a different and accessible data storage device for long-term retention. Most of data is unstructured, in fact to the tune of 49% to be precise. On an average, only 20 to 30 percent of data is active, while the rest is static or inactive. Often, internal and regulatory polices stipulate the periods for data retention. Given this, it is advisable to archive data because it is costly to store the inactive data on primary storage, and it also negatively impacts system performance.

The traditional approach to data archiving was to move the information to cheaper secondary storage, such as tapes and optical disks. With the advent of Big Data, traditional methods of data archival are being revaluated. Typically, once the data is archived, it is never accessed again. So, despite its tremendous potential value, in many ways, traditional archival systems spell death for data usability. A close second was a technological issue: dealing with what has become known as the three 'V's' of Big Data: volume, velocity and variety**.** Unstructured data includes video, audio, images, weblogs, and so on. This data can then be retrieved whenever required. But data cannot be deleted as it needs to be retained for legal compliance or analytics. Many countries have laws requiring businesses to keep records for as long as five to seven years. Businesses often face the additional burden of maintaining legacy storage systems, solely for data accessibility.

The original application might be retired, but the system has to be preserved to allow access to the data it stores. But retrieval of this data for analysis was painful and cumbersome. The traditional archival systems may provide very cheap data storage (compared to RDBMS), but prove astronomically expensive for data retrieval, especially since queries cannot be run on archival systems and the data has to first be retrieved into an RDBMS for processing. According to the TCS Big Data Global Trend Study conducted in 2013, the biggest challenge in deriving business value from enterprise data was in getting business units to share information across organizational silos. The data volume growth and the rate of inflow of data make it necessary for enterprises to have a strong archival solution that can both store enormous amounts of data quickly and scale to meet business requirements with little effort.

## III.HADOOP AS A DATA STORE

The Hadoop framework was designed as a contemporary alternative to process mega-data from distributed applications. It is open sourced and supplements development and implementation tools that can bridge the gap between traditional and new-age frameworks. Hadoop comes ladled with numerous advantages

### *A. Cost effective*

Hadoop supports massively parallel computing based on a shared-nothing architecture. Clusters can be built on inexpensive commodity grade servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it an affordable archive store option for all of the enterprise's data. Hadoop's cost advantages over legacy systems redefine the economics of data storage.

Legacy systems, while adequate for certain workloads, were not engineered keeping the needs of Big Data in mind. They are far too expensive for general purpose usage, given today's large data sets. Hadoop also offers the best option for application retirement. It is both cost effective and efficient because it allows users to readily exploit archived historical data for business intelligence. A banking leader in Australia used Hadoop to retire their legacy systems with DB2, Oracle, and MS SQL data. The bank has since benefitted from easy access to historical data to improve in the process.

### *B .Efficient and fast*

Hadoop focuses on reverse process, which involves taking the code to the data source, in contrast to legacy systems that prioritize on bringing the data source to the code. Hadoop has been successfully processing enormous amounts of data with superior outcomes in reduced amounts of time, making it highly efficient to work with. This is achieved by dividing and distributing large chunks of data among server clusters, ideally on the same machine as its origin. The processed and archived data can be queried without delay from the warm storage.

## IV.FEATURES OF HADOOP

Hadoop is an open source, scalable and fault tolerant framework for processing large sets of distributed data. Hadoop has scale-out architecture and runs on clusters of commodity hardware, making it a very cost effective solution for managing extremely large sets of data associated with Big Data. Map/Reduce is an integral component of Hadoop that processes the distributed data, in parallel, across the many nodes in a cluster. Hadoop Distributed File System (HDFS) is the native file system of Hadoop that stores small fragments of data in different nodes of the cluster ensuring fault tolerance and high availability of the data stored.

A banking leader in Australia used Hadoop to retire their legacy systems with DB2, Oracle, and MS SQL data. The bank has since benefitted from easy access to historical data to improve decision making. Hadoop based archiving overcomes some of the shortcomings of tapes, which are the cheapest mode of storing data but also involve manual intervention to store and retrieve data. A pension trust based in the UK lost data due to errors in the tape archival process - part of the data was overwritten while archiving. The pension trust decided to migrate their archive to a more efficient archive based on Hadoop to overcome these issues. Following this engagement, the company was able to automate their archival process, eliminating the manual intervention required for sorting and changing tapes during the process. The properties listed down below aide the citations given above:-

### *A. Agile*

Hadoop is a distributed, file based storage system. Hence, it can effectively manage data from sources with different schemas. A Hadoop based archive store serves as a universal platform that can absorb any type of data—structured, semi-structured or unstructured, from any number of sources. It enables deeper analyses by allowing data from multiple sources to be combined and aggregated in numerous ways.

### *B. Flexible and scalable*

Another advantage of using Hadoop is that new nodes can be added to a Hadoop cluster without changing data formats, loading methods, or program codes. These features lend supreme scalability and flexibility to the archive store. Fault resistant: Hadoop is also a highly fault tolerant storage system. The Hadoop architecture has the inherent capability to continue processing data, even when a node is lost. This is possible because Hadoop replicates every data block by a default factor of three. In case of a failure, Hadoop simply redirects the code to another node with the same data block.

Hadoop is a great system to use in data transformations. In addition, it is also very good at processing queries on granular and historical data. Enterprises that build their Big Data archival solutions on Hadoop can afford to store practically all the data in their organization. As shown in Figure 2, a Hadoop based archive can function as the universal store that stores data from the production environment, legacy applications, and old archives or backups from disks and tapes. The distinct feature of a Hadoop archive store is that it keeps all the data online for interactive querying, business intelligence, analysis, and visualization.

## V. FRAMEWORK FOR HADOOP BASED ARCHIVAL

As contemplated, to use Hadoop as the data store is one of the novel strategies put forth to address the problems that enterprises currently face in data archiving. A framework that enables faster retrieval and analysis of the stored data ,leverages advantage of Hadoop's advantageous points to deliver a scalable and cost effective solution. Movement of data, from high cost environment to cost effective data stores should be enabled by a good framework. Cost and performance should be leveraged by such a design. Further, after it has served its purpose, the design should help meet Business SLAs, move inactive data to warm storage, and clean it.

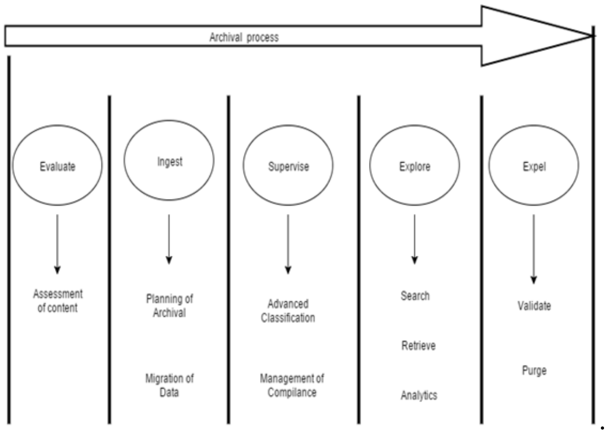

Figure 1: A Framework for Hadoop based Archival

The following processes must be included by such a framework:

- To distinguish between active and inactive data specified criteria should be applied
- Data should be classified into different classes, ranging from immediately useful to historical in accordance with metric analysis.
- Ingest: The most important and difficult phase is the collection and secure movement of data from the production environment to Hadoop.
- Security for ingested data using encryption and masking should be provided by the Framework to maximize output and input by data compression.

*A. Supervise*

The legal, IT, and compliances setup the parameters for supervising and managing the ingested data. Access restrictions must be applied in accordance with business requirements.

*B. Explore*

Instead of pulling in all the data, the framework should enable the user to search and retrieve specified data anywhere Data retrieval should be fast enough to comply with business SLAs. Easy integration with analytics engines should be ensured by the data store.

*C .Expel*

For different periods of time, different data storage should be allowed depending on the data management policy which is to be applied. Data should be pushed to a workflow for purging after the policy expiry date by the function. Validation of such data should be provided by the system; after validation, the data should either be deleted or pushed back into storage, with a policy which has been modified.

A Hadoop based framework can be harnessed by enterprises for devising an effective archival strategy. Such a strategy should encompass the following five major steps:

- Portfolio Planning
- Implementation Overview
- Testing and Training
- Production

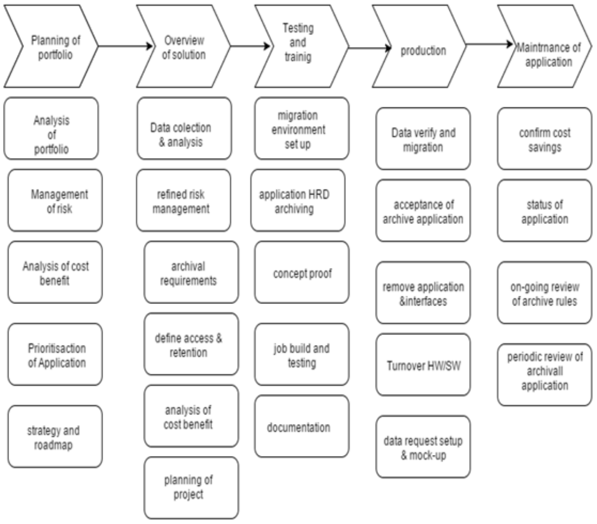
Planning of portfolio
Overview of solution
Testing and trainig
production
Maintrnance of application
Analysis of portfolio
Management of risk
Analysis of cost benefit
Prioritisaction of Application
strategy and roadmap
Data colection & analysis
refined risk management
archival requirements
define access & retention
analysis of cost benefit
planning of project
migration environment set up
application HRD archiving
concept proof
job build and testing
documentation
Data verify and migration
acceptance of archive application
remove application &interfaces
Turnover HW/SW
data request setup & mock-up
confirm cost savings
status of application
on-going review of archive rules
periodic review of archivall application

Figure 2: Recommended Archival Implementation Methodology

## VI. APPLICATION MAINTENANCE

Organization to be effective to use big data analytics have to keep in mind, and also as an ongoing exercise that the benefits should out- weigh the cost. One has also ensured monitoring of e-discovery requests as well as confirmation of data validity. Subsequently 3 steps have to kept in mind via; review of archiving rules on a continuing basis to see that data is extracted in a particular pre-defined set pattern and archived as mandated by convention or need. Subsequently when there is no longer any need for it the data should be destroyed.

## VII.CONCLUSION

Hadoop based archival solution platform should be used in tandem with an effective archival system which should be reliable, upgradable and should deliver the desired result. The data should be secure, and its retrieval should be time bound, effective and scalable and should be ready for analysis. Cost analysis is very important in weighing the benefits against expenditure incurred for timely data retrieval. This will go a long way to speed up data retrieval and the need for storage on large physical devices is no longer there. This in return will give the organizations a greater leverage over their competitors.

## REFERENCES.

[1]. Gaurav Jaswal, Amit Kaul and Rajan Parmar (2012).Keyword Based Image Retrieval using Hadoop Approach,.International Journal of Engineering and Advanced Technology (IJEAT) 2(1), ISSN: 2249 – 8958.

[2]. Gulshan saluja, Ankit rokde and Richa maru (2012). Layered analytical technique for content based video retrieval. IEEE 978-1-4673-1938-6/12.

[3]. Hatice Cinar Akakin and MetinGurcan N (2012). ContentBased Microscopic Image Retrieval System for MultiImage Queries. IEEE Transactions on Information Technology in Biomedicine 16(4).

[4]. Kui Wu & Kim-Hui Yap (2007). Content-based image retrieval using fuzzy perceptual feedback.Multimedia Tools and Applications 32 235–251. DOI 10.1007/s11042-006- 0050-2

[5]. Navdeep and Mandeep Singh. Content (color) based image retrieval using RGB component Analysis. 1st National Conference on Information Technology and Cyber Security 1 171-174/ITCS13/33.

[6]. Nidhi Singhai and Shishir Shandilya (2010). A Survey On: HAdoop Distributed Filesystem International Journal of Computer Applications (0975 – 8887) 4(2).