Leveraging Hadoop for Effective Data Archival

Abstract

Advances in computing, social media, and wired and wireless communication are generating an unprecedented volume of data, which if tagged and analyzed, can provide valuable business insights. However, enterprises face a dual challenge—to meet the perpetually increasing demand for data storage, and the need to perform analytics on stored data. A Hadoop based archival solution is scalable, cost effective, fault tolerant and schema-less, and can easily address the data storage, management, and analytics related business requirements.

According to IBM, 90% of the data ever created has been created in the last two years. In 2005, the entire digital universe was 130 exabytes large[1].
By 2013, it grew 34 times to become 4.4 zettabytes of memory. And according to IDC's projection, the digital universe will explode to 44 zettabytes or 44 trillion gigabytes of information by 2020[2]. In 2012, we created roughly 2.5 exabytes of data per day, and this figure doubles every 40 months or so[3]. Enterprises are responsible for 80% of the data created[4].

Introduction

In recent years, most successful companies have realized the value of integrated data. They now consider data a strategic asset and use it to make data-driven business decisions.

However, managing huge amounts of data is not easy. Maintaining all this information within the source application is neither cost effective nor scalable. Purging the data is also not an option due to both internal and regulatory policies on data retention. In order to effectively manage and analyze the data, enterprises must include data archival as part of their data management life cycle.

Challenges of Traditional Data Archival

With the advent of Big Data, traditional methods of data archival are being revaluated. 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.

The key demands of modern data archival are:

- **Volume:** The amount of data that needs to stored and retained is growing exponentially.
- **Variety:** Different types of data like video, audio, images, weblogs also need to be stored, in addition to traditional structured data.
- **Scalability:** The data store should be easily scalable and cost effective to maintain.
- **Compliance:** The archived data needs to be retained for different periods as mandated by different regulatory, legal, and business policies.
- **Retrieval:** The stored data should be easily and swiftly retrievable.
- **Integration:** The data stored should be easily available for business analytics.

Hadoop is an open source, scalable and fault tolerant framework for processing large sets of distributed data. Hadoop has a 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.

Hadoop as a Data Store

Hadoop is a modern open source framework, designed for coding and executing distributed applications that can process large amounts of data. To bridge the gap between traditional and modern archival tools, the following advantages of Hadoop can be leveraged:

- **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.
- **Efficient and fast:** In contrast to legacy systems, Hadoop focuses on moving code to data rather than bringing data to code. Hadoop breaks up and distributes data across clusters of servers. Computation on a slice of data takes place on the same machine where that slice resides, enabling Hadoop to process large volumes of data and produce better results in a shorter amount of time.
- **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.
- **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.

Product	Processor Power	RAM	Number	Storage	Hardware Raid Controller
Master Node 1	2 Quad Core CPUs, running least 2-2.5GHz	64 GB	One	2 TB SATA	Yes
Master Node 2	2 Quad Core CPUs, running least 2-2.5GHz	24 GB	One	2 TB SATA	-
Data Node/ Shave Nodes	2 Quad Core CPUs, running least 2-2.5GHz	64 GB	8 slave nodes	12 Disk Drives 3TB SATA	-

Infrastructure for typical 100 TB Hadoop Store

By implementing a Hadoop based archival solution to perform archival from Teradata, a large U.S. based retailer was able to archive 1.5 Billion records in 1 hour and 10 TB in 40 hours, and realize 4x savings in data storage costs.

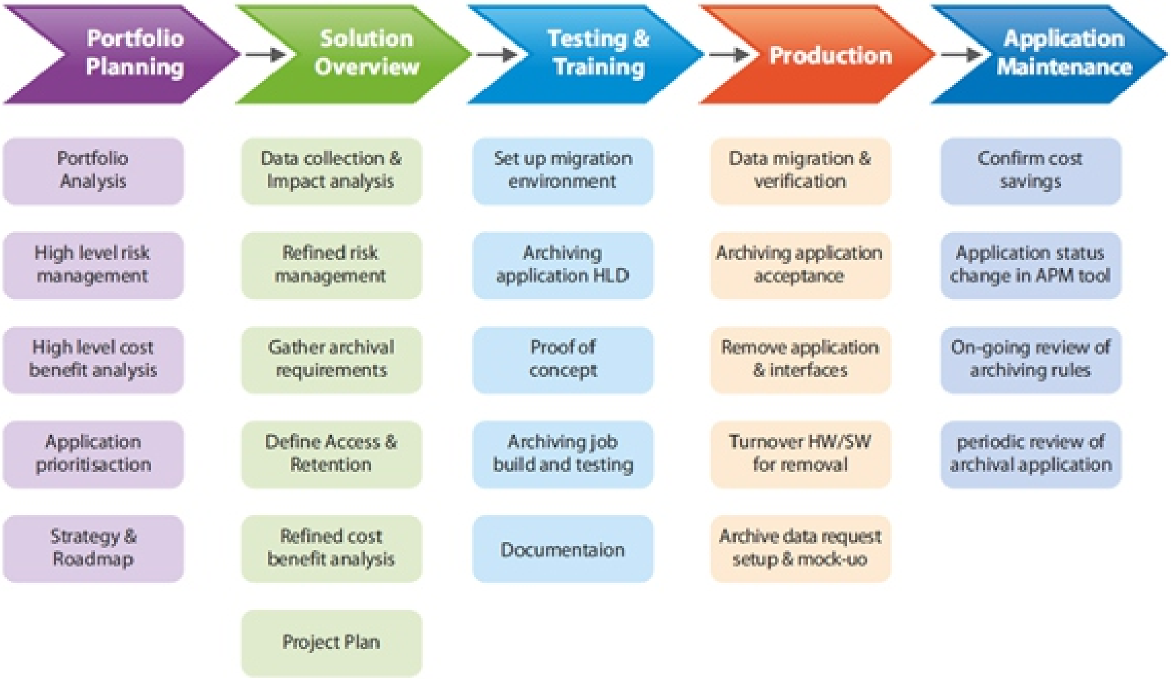
An Australian supermarket major that maintained almost 8 TB of mainframe

Db2, mainframe CA SAR reports, and Unix CISAM reports chose a Hadoop-based archival solution for application retirement and realized 50% cost savings over traditional systems.

A Framework for Hadoop-Based Archival

A good framework should enable movement of data, from high cost environments to cost effective data stores. It should be designed to ensure maximum cost advantage and performance. Further, the design should help meet Business SLAs, move inactive data to warm storage, and purge it after its useful lifetime. Such a framework should include the following processes:

- **Evaluate:** Specific metrics should be used to differentiate between active and inactive data.
- **Ingest:** The framework should provide security for ingested data using encryption and masking, and maximize throughput by compressing data.
- **Supervise:** The ingested data must be supervised and managed according to legal, IT, and compliance requirements.
- **Explore:** The framework should enable the user can to search and quickly retrieve specific data.
- **Expel:** The data store should allow storage of data for different periods of time, depending on the data management policy applied.



Portfolio Planning	Solution Overview	Testing & Training	Production	Application Maintenance
Portfolio Analysis	Data collection & Impact analysis	Set up migration environment	Data migration & verification	Confirm cost savings
High level risk management	Refined risk management	Archiving application HLD	Archiving application acceptance	Application status change in APM tool
High level cost benefit analysis	Gather archival requirements	Proof of concept	Remove application & interfaces	On-going review of archiving rules
Application prioritisaction	Define Access & Retention	Archiving job build and testing	Turnover HW/SW for removal	periodic review of archival application
Strategy & Roadmap	Refined cost benefit analysis	Documentaion	Archive data request setup & mock-uo	
	Project Plan			

Recommended Archival Implementation Methodology: Enterprises can harness a Hadoop based framework for devising an effective archival strategy

Conclusion

The benefits of Hadoop are best realized when it is teamed with a good archival solution that is cost effective, scalable, and secure, allowing for speedy data retrieval and analysis. By using Hadoop as an archival platform, organizations can benefit from multiple source connectivity, scalability, cost benefits, and high accessibility of archived data for analytics. These solutions provide an attractive alternative to the delay and hassle of archiving data on traditional mass storage devices, as well as to the high costs of modern archival appliances. By leveraging a Hadoop-based archival solution, organizations can successfully harness the vast amounts of data at their disposal to realize competitive advantage.

References

[1] EMC Digital Universe, with research and analysis by IDC, Dec 2012

[2] EMC Digital Universe, with research and analysis by IDC, April 2014

[3] Harvard Business Review, Big Data: The Management Revolution, October 2012

[4] TCS, The Emerging Big Returns on Big Data: A TCS 2013 Global Trend Study, May 2013

About The Author

Binesh Kuttan

Binesh is a Product Manager, Digital Enterprise at TCS with over 14 years of IT experience in the Banking and Financial Services industry in the areas of storage and product development. He leverages his experience in information management on the Hadoop platform to build Big Data products for data migration, data privacy, data archival, and data quality.

Sunil Tom Jose

Sunil is a Business Analyst, Digital Enterprise at TCS and has worked with several teams on the development of Big Data products including those focusing on data archival, data migration, and data transformation.

Contact

Visit TCS' Analytics & Insights unit page for more information

Email: analytics.insights@tcs.com

Blog: Digital Reimagination

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

TCS Design Services I M I 12 I 16