RESEARCH PAPER | Library Science | Volume : 5 | Issue : 2 | Feb 2015 | ISSN - 2249-555X

Handling of Library Data with HADOOP

KEYWORDS: Big Data, Data mining, Hadoop, Library retrieval, Library management.

Dr. Sangeeta Paliwal

HOD, Department of library science, IPS Academy, Indore (M.P.)

ABSTRACT *In this article we are introducing latest technology in library. As we all know Hadoop is open source software for handling Big Data. In present era all library work done by the computer. Day by day e resources are increasing in the library so librarians are facing difficulty to handle their data with present infrastructure. In this article we are presenting the introduction of Hadoop, its requirement, its feathers, and need of Hadoop in the library. At the end of the article we gave some name of the corporate/institutions are using Hadoop. Hadoop is very useful for handling big data in libray. Hadoop provides a framework to process large data using a computing cluster made from normal commodity hardware. Hadoop is built in fault tolerance therefore it expects computer to be breaking frequently. We can easily handle big data of library by the help of Hadoop, we can make strong search engine like OPAC and on security point of view Hadoop is very much secure. With the help of Hadoop we can make data mining also it will very supportive for data retrieval in various formats of library data.*

INTRODUCTION-

BIG DATA – "Big data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques." (hlideshare website) (1)

"In 2001 Doug Laney defines big data as the three Vs of Big Data: Volume, Velocity and Variety." (Laney, 2001)(2)

In other words we can say that a big data have high data quantity, high data space and different data types.

HADOOP – Hadoop is open source software which made to handle big data. Hadoop was developed by Daug Cutting in 2005. Doug's son have an yellow toy elephant he called it as Hadoop and Doug's like that name so Dougs name it Hadoop. It was developed by Apache Software Foundation. Its first release was in 10 December 2011 and latest version 2.6.0 was release in 18 November 2014. It is written in Java language. It has Apache License 2.0.

The apache Hadoop project develops open source software for reliable, scalable, distributed computing. The apache Hadoop software library is a frame work that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single services to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly- available service on top of a cluster of computers, each of which may be prone to failures. Hadoop includes following modules:

1. Hadoop Common – The common utilities that support the other hadoop modules.
2. Hadoop Distributed File System (HDFS) - A distributed file system that provides high throughput access to application data.
3. Hadoop Yarn – A framework for job scheduling and cluster resource management.
4. Hadoop Map Reduce: A yarn based system for parallel processing of large data sets. (hadoop)(3)

NEED OF HADOOP IN LIBRARY- IBM claims 90% of today's stored data was generated in just the last two years. (hlideshare website)(1)

If we think about library the situation will be same. In present era every library is on face book, twitter and on mobile also. In next few years maximum library resources will be published in electronic form. When the unstrcuchers data and structured data will be large then it is very difficult to find particular information in the seconds. So to handle this situation we required a big data manager in library which can retrieve information in seconds and can manage all unstrctuchered and structured data of the library. I think Hadoop is the best solution to overcome this problem.

SOURCE OF BIG DATA IN LIBRARY –

In library there are four sources of Big Data-

User

Library data

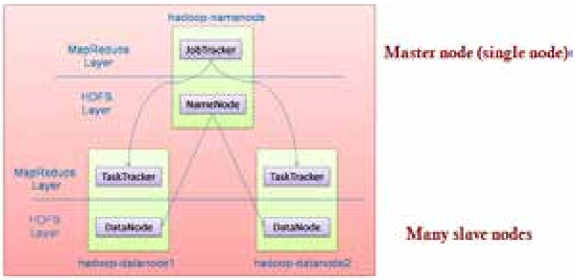
Sensors

System

COMPONEMTS OF HADOOP – Hadoop have two major components –

Distributed file system: It splits up large files into multiple computers.

Execution Engine (Map reduce framework): It is a framework used to process large data stored on the file system.



FEATURS OF HADOOP –

1. Large Data Sets –Map-reduce paired with HDFS is a successful solution for storing large volumes of unstructured data.
2. Scalable Algorithms –Any algorithm that can scale too many cores with minimal inter-process communication will be able to exploit the distributed processing capability of Hadoop.
3. Log Management –Hadoop is commonly used for storage and analysis of large sets of logs from diverse locations. Because of the distributed nature and scalability of Hadoop, it creates a solid platform for managing, manipulating, and analyzing diverse logs from a variety of sources within an organization.
4. Extract-Transform-Load (ETL) Platform –Many companies today have a variety of data warehouse and diverse relational database management system (RDBMS) platforms in their IT environments. Keeping data up to date and synchronized between these separate platforms can be a struggle. Hadoop enables a single central location.

HADOOP IS WORK ON –

Hardware Failure

One consequence of scale is that hardware failure is the norm rather than the exception. An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data. The fact that there are a huge number of components and that each component has a non-trivial probability of failure means that some component of HDFS is almost always behaving badly. Even with RAID devices, failures will occur frequently. Therefore, detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

Streaming Data Access

Applications that run on HDFS need streaming access to their data sets. They are not standard applications that typically run on general purpose file systems. HDFS is designed more for batch processing rather than interactive use by users. The emphasis is on high throughput of data access rather than low latency of data access. POSIX imposes many hard requirements that are not needed for applications that are targeted for HDFS. POSIX semantics in a few key areas have been relaxed to gain an increase in data throughput rates.

Large Data Sets

Applications that run on HDFS have large data sets. A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files. It should provide high aggregate data bandwidth and scale to hundreds of nodes in a single cluster. It should support tens of millions of files in a single instance.

Simple Coherency Model

HDFS applications need a write-once-read-many access model for files. A file once created, written, and closed need not be changed except for appends. This assumption simplifies data coherency issues and enables high throughput data access. A Map Reduce application or a web crawler application fits perfectly with this model.

PREREQUISITES OF HADOOP –

Data Organization

- HDFS support write-once-read-many with reads at streaming speeds.
- A typical block size is 64MB (or even 128 MB).
- A file is chopped into 64MB chunks and stored.

API (Accessibility)

- HDFS provides Java API for application to use.
- Python access is also used in many applications.
- A C language wrapper for Java API is also available.
- A HTTP browser can be used to browse the files of a HDFS instance.

FS Shell, Admin and Browser Interface

- HDFS organizes its data in files and directories.
- It provides a command line interface called the FS shell that lets the user interact with data in the HDFS.
- The syntax of the commands is similar to bash and cash.
- Example: to create a directory /foodie /bin/hadoop dfs –mkdir /foodir
- There is also DFS Admin interface available
- Browser interface is also available to view the namespace.

CONCLUSION – At the end we can say that we have updated our library with latest technology otherwise tone day there will be no any importance of the library. Hadoop is very good big data manager and there are so May organizations are using it for managing our data here are the list.

1. Adobe

- We use Apache Hadoop and Apache HBase in several areas from social services to structured data storage and processing for internal use.
- We currently have about 30 nodes running HDFS, Hadoop and HBase in clusters ranging from 5 to 14 nodes on both production and development. We plan a deployment on an 80 nodes cluster.
- We constantly write data to Apache HBase and run MapReduce jobs to process then store it back to Apache HBase or external systems.
- Our production cluster has been running since Oct 2008.

2. Alibaba

- *A 15-node cluster dedicated to processing sorts of business data dumped out of database and joining them together. These data will then be fed into iSearch, our vertical search engine.*
- *Each node has 8 cores, 16G RAM and 1.4T storage.*

3. Accela Communications
4. adyard
5. Adknowledge - Ad network
6. AOL
7. ARA.COM.TR - Ara Com Tr - Turkey's first and only search engine
8. BabaCar
9. Beebler
10. Brockmann Consult GmbH - Environmental informatics and Geoinformation services
11. Caree.rs
12. Cloudspace
13. DropFire
14. EBay
15. Facebook

REFERENCE (n.d.). Retrieved january 31, 2015, from hlideshare website: http://www.slideshare.net/nasrinhussain1/big-data-ppt-31616290 | 2..Laney, D. (2001, february 6). meta group . Retrieved January 31, 2015, from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf | 3.hadoop. (n.d.). Retrieved january 31, 2015, from apache hadoop : http://hadoop.apache.org/index.html#What+Is+Apache+Hadoop%3F | 4. http://programmers.stackexchange.com/questions/64881/what-is-hadoop-and-what-are-some-example-applications-of-it | | 5.http://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&sqi=2&ved=0CDkQFjAA&url=http%3A%2F%2Fweb.cs.wpi.edu%2F~cs525%2Fs13-MYE%2Flectures%2F1%2FHadoop.pptx&ei=p2JJUo_yKMKNrQfU8IGoAg&usg=AFQjCNEgoLj27pZuTyBaUIhNk8YU2WGRVw&bvm=bv.53217764,d.bmk | | | 6.http://www.ibm.com/developerworks/library/wa-introhdfs/ | | 7.http://free-hadoop-tutorials.blogspot.in/2011/04/hdfs-architecture.html | | 8.http://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&ved=0CD8QFjAD&url=http%3A%2F%2Fwww.cse.buffalo.edu%2Ffaculty%2Fbina%2FMapReduce%2FHDFS.ppt&ei=AVJSUrepK8LVrQe_tYF4&usg=AFQjCNFYDNGCGGYQMclbEf4E_jsgNnGywA&bvm=bv.53537100,d.bmk |