*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 6, Ver. I (Nov – Dec. 2015), PP 91-94*
*www.iosrjournals.org*

# Hadoop add-on API for Advanced Content Based Search & Retrieval

[1]Kshama Jain, [2]Aditya Kamble, [3]Siddhesh Palande, [4]Rahul Rao,
Prof. ShaileshHule
*Department of Computer EngineeringPimpriChinchwad College of Engineering, Pune*
*Guided by: Prof. ShaileshHule*

***Abstract:****Unstructured data like doc, pdf is lengthy to search and filter for desired information. We need to go through every file manually for finding information. It is very time consuming and frustrating. We can use features of big data management system like Hadoop to organize unstructured data dynamically and return desired information. Hadoop provides features like Map Reduce, HDFS, HBase to filter data as per user input. Finally we can develop Hadoop Addon for content search and filtering on unstructured data.*
***Index Terms:****Hadoop, Map Reduce, HBase, Content BasedRetrieval.*

## I. Introduction

### A. Hadoop

Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Commodity computing, or commodity cluster computing, is the use of large numbers of already-available computing components for parallel computing, to get the greatest amount of useful computation at low cost. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (Map Reduce). Hadoop splits files into large blocks and distributesthem amongst the nodes in the cluster.

### B. HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

### C. MapReduce

A MapReduce program is composed of a Map() procedure(method) that performs filtering and sorting and a Reduce() method that performs a summary operation. The MapReduce System run the various tasks in parallel, managing all communications and data transfers between the various parts of the system.

### D. HBase

HBase is a data model that is similar to Googles big table designed to provide quick random access to huge amounts of structured data. This tutorial provides an introduction to HBase, the procedures to set up HBase on Hadoop File Systems, and ways to interact with HBase shell. It also describes how to connect to HBase using java, and how to perform basic operations on HBase using java.

### E. Current System

Tasks like assignments, taking notes from text books and reference books on particular topic, topics for presentation need deep reading and need to go through every document manually just to find relevant content on given topic. Currently present systems are only searching based on document title, author, size, and time but not on content. So to do content based search on big data documents and large text data Haddop framework can be used

### F. Content Based Approach

Manually filtering from any kind of unstructured data like PDF is tedious and time consuming, we are developing API for Hadoop finding relevant information from large sets and retrieving the same is main concern. So using Hadoop Big Data management framework consist of HDFS, MapReduce, and HBase, we are developing content based search on PDF documents to solve real life problem. So this is basic motivation for the

project.

## II. Proposed System

This system shall retrieve the required contents of files which are in an unstructured format containing huge amount of Data like e-books in a digital library where the number of books are present in thousands. The scope here is initially limited to PDF files which may be expanded to other unstructured formats like ePUB, mobi. The input is provided to the system API in the form of search query which will be firstly filtered to find the important expression as a query to the API which will return the important content to the user in the form of paragraph or text highlighted using the power of distributed computing. The particular page or the Entire book itself can be downloaded by the Library users if the content is satisfied else the search continues for finding relevant content

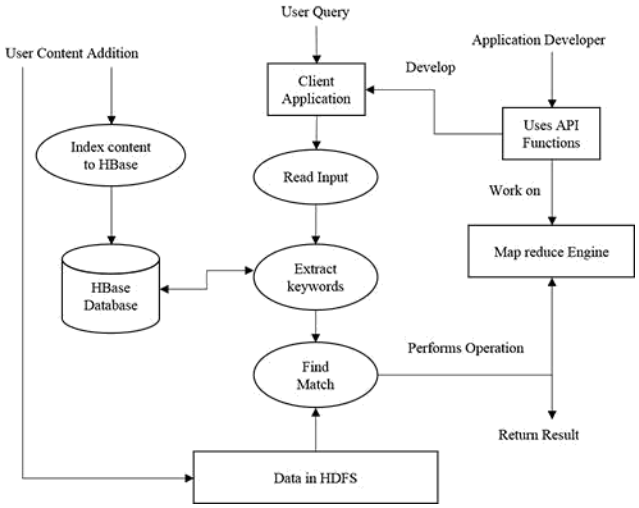


Fig. 1. Data Flow Diagram

**Algorithm 1: Content Based Search Algorithm**

1) Take Input Query from User
2) Use the Suffix Stripping Algorithm to remove English Suffixes like 'ed','ing',' ' 's ' to extract keywords
3) Match keywords with Documents' metadata present in Hbase and filter the documents to be searched.
4) Apply Searching Algorithm based on map-reduce operations like keyword count,positions.
   - Open the PDF
   - Distribute the text using map operation.
   - Combine the results with reduce operations.
5) Sort the documents which are returned by Searching w.r.t keyword count and relevance.
6) Depending on API methods return the result with
   - Paragraph
   - Whole Page
   - Current and Previous Page

### A. Document Metadata

Searching and retrieving information from data inside HDFS. Hadoop Map Reduce operations will be used to perform key value pair generation and depend upon result information is searched.

a) Here, data is documents in PDF (Portable Document Format) format. These documents are stored on HDFS Hadoop Distributed File System. These documents are considered as raw data. These documents can have properties like

   - Name Size Date
   - Author

b) Document meta data is also maintain in HBase distributed database. It has attributes like
   - Content Keywords

- Index Keyword
- Document Keywords

This information will be useful to filter documents before performing content based searching operations.

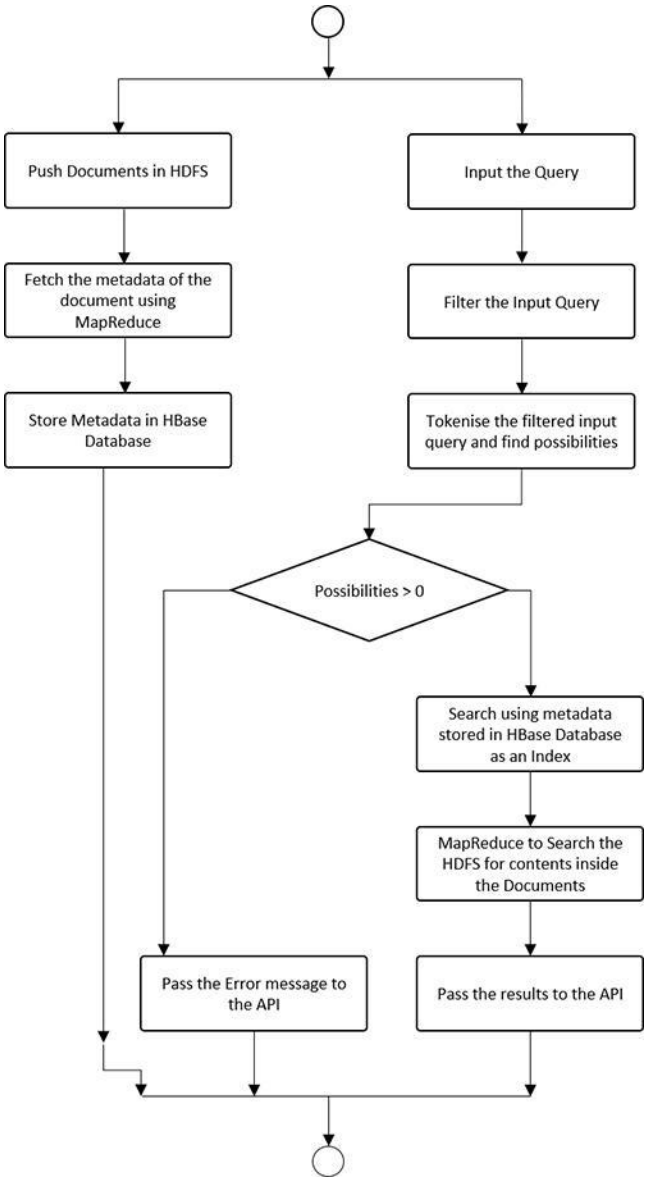


Fig. 2. Activity Diagram

**Algorithm 2: Meta-data Extraction Service Algorithm**

Meta-data Extraction Service

1) Take the input pdf files from user
2) Check if corrupt
3) Upload the file into the Hadoop Distributed Filesystem
4) Run the Metadata Extracting Algorithm using the Map Reduce Engine.
5) Extract the Bibliography and Index Contents of each and every document (if present) and related keywords.
6) Store it in the HBase as metadata.
7) This metadata will be used by map reduce engine to search and retrieve data from relevant document inside HDFS using keyword relative custom partitioning.
8) Return the result to the search API
9) Stop

## III. Applications

Content based search and retrieval on

1) Digital library books
2) Conference Papers
3) Private Authorities
4) Government Authorities
5) Unstructured text files

This API can be used to develop above functionality on platforms like
1) Web Application
2) Android Application
3) Standalone Application
4) Command Line Interface Application

## IV. Future Scope

Extending our API further to read scanned text copies and retrieve the data from them would solve further advanced issues. This API again can be extended further to work with Image processing so that it may find a relevant information from the digital images for the end- users.

## V. Conclusion

This API would be favorable for many large organizations like Large-scale Industries and Educational institutes, Power-Grids, Air-lines, Government Offices and many others working with and producing Big-Data to retrieve data in an efficient and a very cost effective way. Thus we can use this API in Hadoop to reduce manual efforts and bring advance content based search and retrieval

## References

[1] Lars George, "HBase: The Definitive Guide", 1st edition, O'Reilly Media, September 2011, ISBN 9781449396107
[2] Tom White, "Hadoop: The Definitive Guide", 1st edition, O'Reilly Media, June 2009, ISBN 9780596521974
[3] Apache Hadoop HDFS homepage http://hadoop.apache.org/hdfs/
[4] MehulNalinVora ,"Hadoop-HBase for Large-Scale Data",InnovationLabs,PERC,ICCNT Conference
[5] YijunBei,ZhenLin,ChenZhao,Xiaojun Zhu ,"HBase System-based Distributed Framework for Searching Large Graph Databases",ICCNT Conference
[6] SeemaMaitrey,C.K."Handling Big Data Efficiently by using Map Reduce Technique",ICCICT
[7] Maitrey S, Jha. An Integrated Approach for CURE Clustering using Map-Reduce Technique. In Proceedings of Elsevier,ISBN 978-81- 910691-6-3,2 nd August 2013.