International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(3), pp.100-105
DOI: http://dx.doi.org/10.21172/1.83.013
e-ISSN:2278-621X

DISTRIBUTED SEARCH ENGINE FOR EXTRACTION OF RESUME STATISTICS USING HADOOP WITH COMBINATION OF LUCENE INDEXING FRAMEWORK AND THE SOLR

Geetha Guttikonda[1], Madhavi Katamaneni[2] and Madhavi Latha Pandala[3]

Abstract- The problem most recruiters face is spending a lot of manual work for analyzing each resume/CV and searching a suitable resume. In this scenario, Resumes are located in different branches at different locations. We want to provide a single framework to search over Resume data in distributed environment. As a solution to handle big data Hadoop is the accepted framework, as it is a solution for scalable and reliable data processing workflows. The main objective is to build a distributed search engine for resume/CV data with Hadoop. For the proper storage and reporting, and to get the structured information from a large set of Resume/CV documents MapReduce (MR) a framework of Hadoop and NLP are used. The combination of Lucene indexing framework and the Solr allows us to provide search and filter functionalities for a large amount of distributed Resume data.

Keywords – **Search, Resume, Hadoop, Bigdata, Indexing, Mapreduce**

I. INTRODUCTION:

To analyze the gigantic information, Big data is a used for collection of massive and complex information which include the gigantic information, online networking analytics, information administration capabilities Big data analytics is the process of looking at large volumes of data, structured or un-structured. Hadoop is an eco-framework for taking care of enormous information utilizing distinctive structures like HDFS for storage and Map Reduce for processing. As resumes are un-structured and human written, they can be effectively parsed using big data using Hadoop[1].

The Recruiters face issues of gathering large volume of resume and spending hours of manual work for examining each one and they found that the data can't be processed using any of the existing centralized architecture solutions. We want to provide a simple Information retrieval system mechanism which has large number of resumes as input and provide the required resume as per the recruiters need.

[1] *Department of Information Technology V R Siddhartha Engineering College, AP, India*
[2] *Department of Information Technology V R Siddhartha Engineering College, AP, India*
[3] *Department of Information Technology V R Siddhartha Engineering College, AP, India*

With the development of advances and administrations, the substantial measure of information is delivered that can be organized and unstructured from the distinctive sources. Such type of data is very difficult to process that contains the billions records of millions individuals data that incorporates the web deals, online networking, sounds, pictures et cetera. The need of enormous information originates from the Big Companies like Yahoo, Google, Facebook, and so on .with the end goal of examination of huge measure of information which is in unstructured shape. . Google contains the substantial measure of data. So there is the need of Big Data Analytics that is the preparing of the perplexing and monstrous datasets. Enormous information investigation dissects the vast measure of data used to reveal the shrouded designs and the other data which is valuable and vital data for the utilization.

Hadoop is an exceptionally adaptable capacity stage, since it can store and convey vast information sets crosswise over several modest servers that work in parallel. Not at all like customary social database frameworks (RDBMS) that can't scale to process a lot of information, Hadoop empowers organizations to run applications on a huge number of hubs including a huge number of terabytes of information.

A key preferred standpoint of utilizing Hadoop is its adaptation to non-critical failure. At the point when information is sent to an individual hub, that information is additionally reproduced to different hubs in the group. Hadoop's one of a kind stockpiling strategy depends on a dispersed record framework that essentially "maps" information wherever it is situated on clusters.

1) Working:

In order to transfer resume data from local machine to HDFS we will first transfer data from local machine to Linux OS using FileZilla software. Linux OS is used because it uses command line interface. UNIX, Ubuntu can also be used. Data is transferred from Linux OS to HDFS using Java Program in Eclipse and processing is done using MapReduce. Linux and HDFS together acts as a virtual machine using VMware. Results are then gathered by Name Node and displayed back on the local machine.

Local Machine

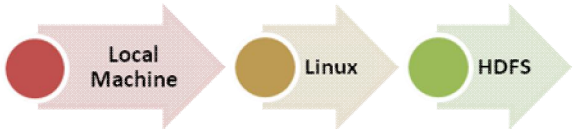

Figure 1: Storage in hadoop

II. LITERATURE SURVEY

The current Relational Database Management Systems (RDBMS) are not proficient for taking care of Big Data Big Data. A social database administration framework (RDBMS) is a database administration framework (DBMS) that depends on the social model as invented by E. F. Codd, of IBM's San Jose Research Laboratory. Number of recognized records at present in use is based on the relational database model. RDBMS has certain features such as: gives information to be put away in tables, holds on information as rows and columns, gives primary key, to particularly distinguish the lines, makes files for snappier information recovery, gives multi client openness that can be controlled by individual clients. It has certain drawbacks such as requirement of structured data type and software license. Also it provides limited processing.

Sumit Maheshwari has analyzed an approach for resume information extraction to support automatic resume management and routing. A cascaded information extraction (IE) framework is designed. In the first pass, a resume is segmented into consecutive blocks attached with labels indicating the information types. Then, in the second pass, the detailed information, such as Name and Address, are identified in certain blocks (e.g. blocks labelled with Personal Information), instead of searching in the entire resume. Based on the requirements of an ongoing recruitment management system which incorporates database construction with IE technologies and resume recommendation (routing), general information fields like Personal Information, Education etc. are defined.

Resume Parsing Methods empowers extraction of important data from resumes which have moderately organized frame. In spite of the fact that, there are numerous business products on resume data extraction, a portion of the business products incorporate Sovren Resume/CV Parser, Akken Staffing, ALEX Resume parsing Resume Grabber Suite and axtraCVX. There are four sorts of techniques utilized as a part of resume data extraction: Named-substance based, administer based, measurable and learning-based strategies. Typically a blend of these strategies is utilized as a part of numerous applications.

1. Certain words are distinguished by the named-element based data extraction strategies, expressions and examples as a rule utilizing general expressions or lexicons. This is typically utilized as a moment venture after lexical investigation of a given report. Administer construct data extraction is situated in light of punctuations.
2. Govern based data extraction strategies incorporate an extensive number of syntactic principles to concentrate data from a given report.
3. Statistical data extraction techniques apply numerical models to distinguish structures in given records.
4. The learning-based strategies utilize grouping calculations to concentrate data from a report.

III. Search Engine Development Technologies:

Apache Lucene is a current, open source look library intended to give both pertinent outcomes and additionally elite. Besides, Lucene has experienced huge change throughout the years, beginning as a one-individual venture to one of the main hunt arrangements accessible. Lucene is utilized as a part of a limitless scope of uses from cell phones and desktops through Internet scale arrangements. The advancement of Lucene has been very emotional on occasion, none more so than in the present arrival of Lucene 4 .0.

The examination limits in Lucene are responsible for taking in substance as records to be recorded or request to be looked for and changing over them into a reasonable inside representation that can then be used as required. At requesting time, examination makes tokens that are inevitably installed into Lucene's adjusted record, while at question time, tokens are made to help outline appropriate request representations. The examination procedure comprises of three undertakings which are tied together to work on approaching substance

Apache Solr is a hunt stage concentrated on conveying endeavor class, elite inquiry usefulness. It can be utilized to power seek highlights inside any information driven sites . Solr can be coordinated with any application or site since it imparts utilizing standard arrangements, for example, HTTP, XML and JSON . Solr's REST-like APIs make it simple to use from for all intents and purposes any programming dialect. Its significant components are full-Text look capacities, faceted pursuit, spatial inquiry, and disseminated seek. These components are of incredible importance since inquiry is not just about returning outcome set protests that match words in a client question. It incorporates handling of the first content into recorded terms took after by importance positioning and returning outcomes in exceptional gathering organizations, for example, aspects, groups and numerous more . Solr has the capacity of appropriated looking, adaptation to non-critical failure and load adjusting and consequently it is broadly utilized as a part of a major information biological system.

Apache Solr now underpins a few sorts of joins and gathering choices that can be broadly utilized when there is a need of standardization to jelly reports relationship. It has a discretionary segment arranged capacity called docValues that is the best approach to construct the forward file. Also, out of box, it gives the designer, the adaptability to have their own mind boggling information sorts and capacity, positioning, and investigation capacities.

The Apache Lucene's libraries that Apache Solr has in its center are generally utilized as a part of a large portion of the adaptable sites, for example, Twitter, LinkedIn and other interpersonal interaction locales, and so forth. Such sites are assessed to oversee up-to 12000 inquiries for every second and with

Indexing rate of 220GB/hour for 4k records. Convenience, significance, discovers capacity and prescribing different outcomes are dependably the vital parts of inquiry innovation. Apache Solr with its content breaking down abilities, for example, Automatic term stemming, spell-adjustment, equivalent word seek, multi-lingual investigation and so forth., empower clients to enter free content inquiry and recover results of their advantage.

IV. PROPOSED METHOD:

As we want to deal with data stored across clusters, we develop each component using distributed architecture. Hadoop Distributed file system (HDFS) is used to store large Resume/CV data across clusters. To handle unstructured data like CV/Resume, we are using Apache POI libraries for processing the documents. Natural language processing is also used to find the semantics of the data and Word web libraries to find the semantic similarity between the any two words. The major steps involved in this method are

- Hadoop Distributed Cluster setup
- Pre-Processing of Resume/CV to get final tokens for indexing.
- Indexing of processed tokens using Apache Lucene
- Implementation of distributed search server using Solr

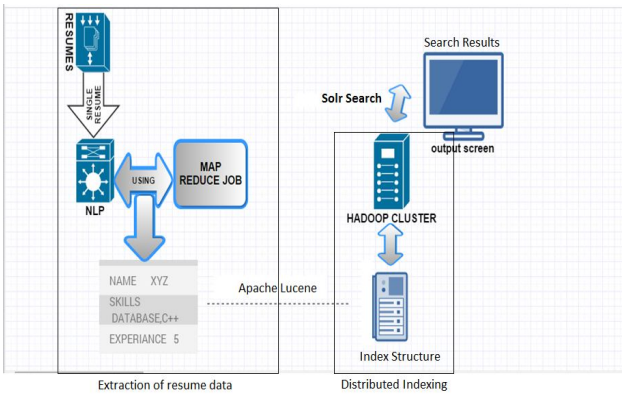

Figure 2: A simple architecture explaining the process of method is given below

Map Reduce is a framework designed for processing data on HDFS. So the entire project is build using Map Reduce(MR) libraries.

Algorithm of Pre-processing:

INPUT: File with resume paths, Attribute list< skills, qualifications, experience>, output_path.
OUTPUT: File containing rows with res_loc, values corresponding to input_attlist.
ALGORITHM: Step-by-step Process for a Resume Parser:

Algorithm ***ResumeParser***(input_file, att_list, out_path):
Read the Input File line by line
For each *line* Extract file from location using doc reader do:
Extract sentences from document using Sentence Segmenter
` For each *Sentence* of doc do:
Tokenize sentence using Standford NLP tokenizer
For each *token* in sentence:
Find Lemma of token and do POS Tagging
Find Semantic similarity (p) of token with att_list
If probability p > threshold send sentence to further analysis
Return values corresponding to att_list attribute for which p> threshold
Save file_path values corresponding to att_list attributes in out_file
End

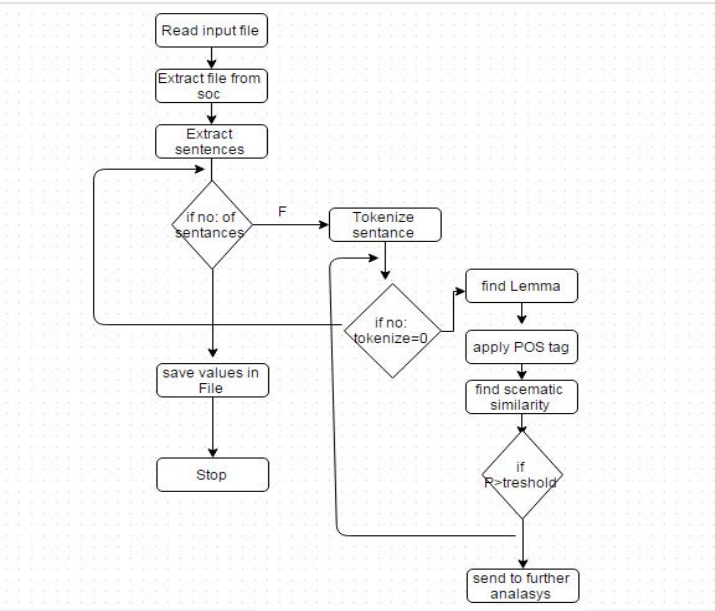

Figure 3: Flow chart for Resume parser

V. RESULTS &CONCLUSION

Loading Data into Hadoop

The data lies on normal linux file system. But it should be stored on distributed Hadoop file system which is called HDFS. The data is distributed across the clusters with replications.

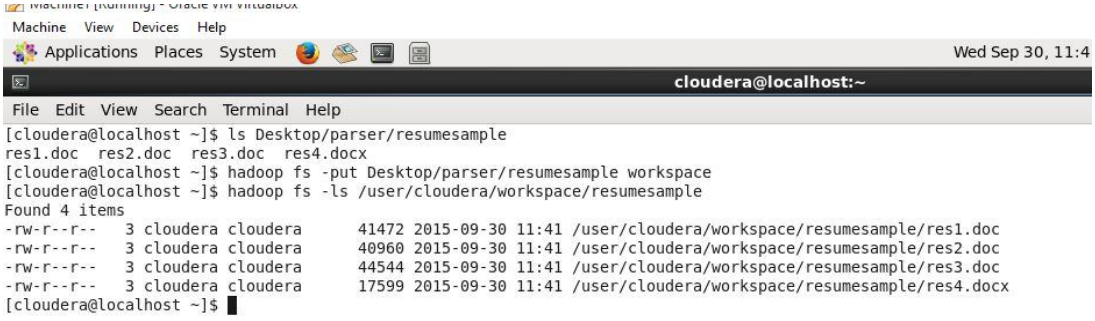

Figure 4: command line interface in Hadoop

Figure 5: Resume data indexing

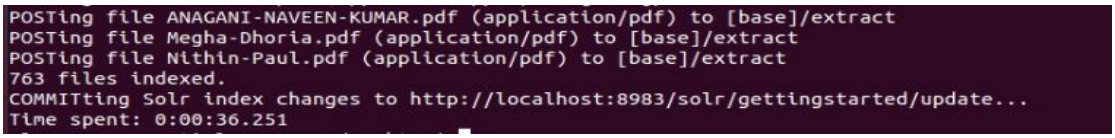

Figure 6: few Resumes indexing completion

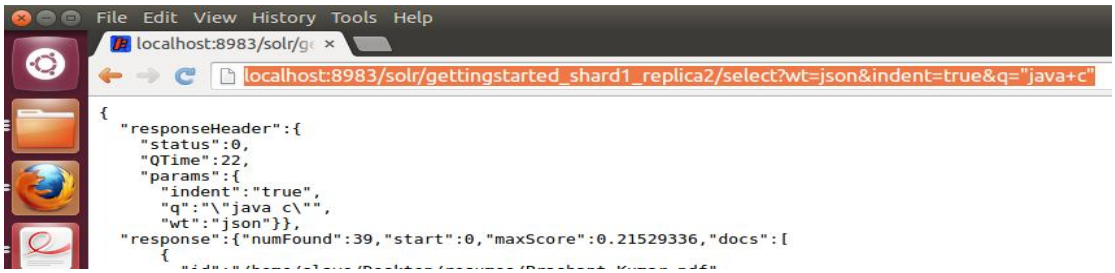

Figure 7: Search Results for single and Multi-term search

Figure 8: Resumes studying in Anuradha engineering college

REFERENCES

[1] Kudatarkar, Vinaya R., Manjula Ramannavar, and Nandini S. Sidnal. "An Unstructured Text Analytics Approach for Qualitative Evaluation of Resumes." (2014).

[2] Kudatarkar, Vinaya Ramesh, Manjula Ramannavar, and Dr Nandini S. Sidnal. "A Survey on Unstructured Text Analytics Approaches for Qualitative Evaluation of Resumes." International Journal of Emerging Technology in Computer Science and Electronics (IJETCSE) April (2015).

[3] Akshata Walavalkar, Sheikh Abdul Rehman, Prachi Kore, Saurabh Nimbalkar, Resume Parsing using Hadoop