MANTECH Publications
Recent Trends in Computer Science and Software Technology
Volume 3, Issue 1, January-April 2018

Resume Analysis using Hadoop Ecosystem

Anupreet Kaur[1], Jasnidh Kaur Ahuja[2], Megha Shukla[3], Nidhi Gupta[4]
Department of Information Technology
Inderprastha Engineering College
***Corresponding Authors' email id**: anupreet9616@gmail.com[1], jasnidh.ahuja@gmail.com[2], meghashukla908@gmail.com[3], nidhi.gupta.it@ipec.org.in[4]*

Abstract

A Resume is a document used to represent a person's background and skills. It is used to screen applicants based on the profile required. In this paper, we present a solution to analyze huge amount of resumes using HDFS and MapReduce and reduce duplicity among millions of resumes. Our implementation of Hadoop is in pseudo mode.

Keywords: *HDFS, MapReduce, Hadoop, Big Data, Resume Parser*

I. INTRODUCTION

Organizations are always in search for good technical talent which is often expensive and short in supply. Resume is such a document that records a person's data that provides information about the specific skills and areas of knowledge that are in greatest demand. The traditional approach is one where resumes are uploaded to the job portal or career section of an organization and then manually read by a person within the organization to identify if the applied candidate has the primary skill(s) that the job demands. It also helps determine a person's pay scale depending on the skill set. This old way has many disadvantages.

1) Every resume has to be manually checked by a person which is time taking.

2) The skill will have to be saved into a file along with the name and other details of the person which will consume memory.

3) If different people analyze different resumes the overall result can be

ambiguous and duplicity of resumes could also be there.

II. EXISTING SYSTEM

The present day systems are providing resumes to recruiter that has duplicate results. Since the resumes are being updated constantly by the candidates this leads to redundancy of resumes in the systems as well as presence of obsolete resumes create a problem.

Adding new resumes and updated resumes in system leads to pilling up of millions of resumes, which becomes difficult to process in management system (RDBMS) becomes cumbersome as it provides limited processing and storing capabilities with respect to huge data that is present. To overcome the processing and storing limitations Hadoop is used that is capable of storing and processing huge data in much less time. Moreover it makes system fault, tolerant flexible as well as scalable.

III. AIM AND OBJECTIVE

The intention behind presenting the project concepts is to reduce the manual handling of CV/Resume data. It also provides the facility to select the resume according to the respective job requirements description that an organization is looking for. It also reduces various anomalies like duplicity of resumes and obsolete resumes.

IV. PROPOSED APPROACH

In this paper, we have provided an approach to extract resumes data and store them for analysis purpose. Before proceeding with the analysis process an assumption is to be considered. A standardized format for all resumes is to be followed for accurate results. Since resumes are written in free language as the language can infinitely vary and is ambiguous. To overcome these limitations a standard format is used.

The following are the steps:

Step-1: A large number of resumes are taken in same format and saved in a directory. These include both doc and docx files.

Step-2: (File naming) All files are named as ‘resumeN.extention’ for standardization (where N represents number).

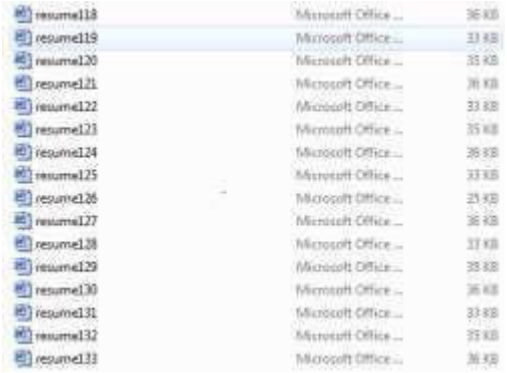


Fig 1: Input resumes

Step-3: (Details Extraction) Resume parser is built in order to extract all the necessary details present in the resume. This involves extracting name, date of birth, email address, address, contact number, graduation details, secondary school details and skills details.

- Name, date of birth, email address and contact number are used to uniquely identify a resume.
- Graduation, secondary school and skills details are used to refer correct resumes to the recruiter as per their specifications.

Step-4: (Storing and Analysis) The extracted data is stored on HDFS. The MapReduce provides programming interface and is used to remove all duplicate and obsolete resumes. The analyzed results files are in structured format and are stored on HDFS. Later these are moved to the database.

Step-5: (Display the results) When the recruiter wants a candidate he/she can search for it by entering all his requirements. A list of eligible candidates is displayed to the recruiter.

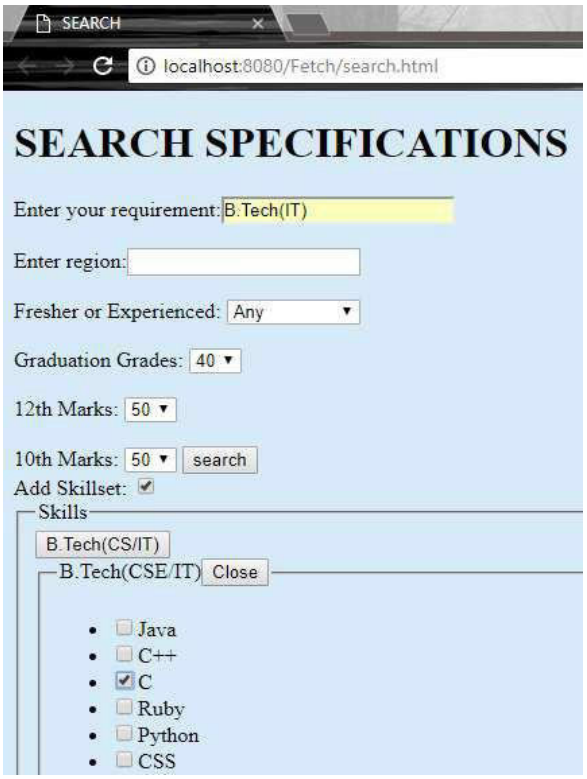
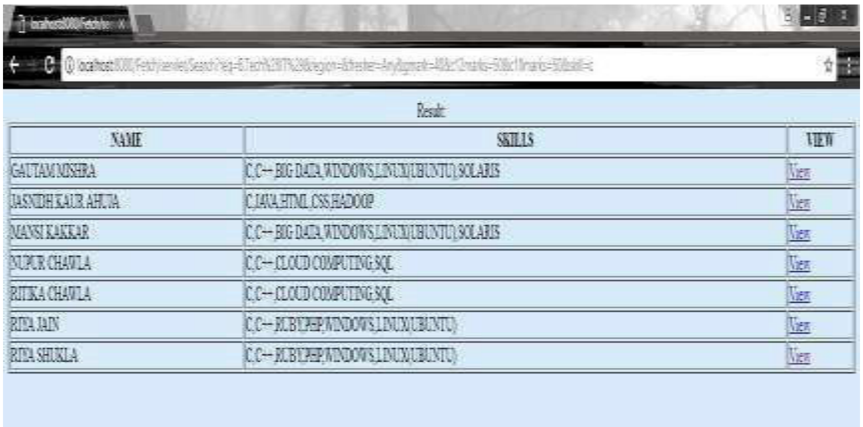


Fig 2: Search Feature



Result:

NAME	SKILLS	VIEW
GAUTAM MISHRA	C,C++,BIG DATA,WINDOWS,LINUX(UBUNTU),SOLARIS	View
JASNIDH KAUR AHUJA	C,JAVA,HTML,CSS,HADOOP	View
MANSI KAKKAR	C,C++,BIG DATA,WINDOWS,LINUX(UBUNTU),SOLARIS	View
NUPUR CHAWLA	C,C++,CLOUD COMPUTING,SQL	View
RITIKA CHAWLA	C,C++,CLOUD COMPUTING,SQL	View
RITA JAIN	C,C++,RUBY,PHP,WINDOWS,LINUX(UBUNTU)	View
RITA SHUKLA	C,C++,RUBY,PHP,WINDOWS,LINUX(UBUNTU)	View

Fig 3: List of candidates referred by the system

Result:

Name	GAUTAM MISHRA
Email Id	gautammishra@gmail.com
Contact No	+91-9717068637
Date Of Birth	18/03/1993
Address	Jain Apts, MIC road, Kolkata-610063
Degree	B.Tech(IT)
Year of Passing	2016
University	DTU
Marks	96.26
Class 12	XII (Science)
Year of Passing	2012
Board	CBSE
Marks	75.90
Class 10	Class 10
Year of Passing	2010
Board	CBSE
Marks	83.90

Download Resume

Fig 4: View Feature and Download Feature

Step-6: (Downloading the resumes) The candidate's complete profile can be viewed by the recruiter as well as they can download the complete resume of the candidate.

V. REQUIREMENT ANALYSIS

1. Hardware Requirements:

- Pentium Dual Core or above
- 4GB RAM

2. Software Requirements:

- OS - Windows 7/Windows 8 - 64 bits
- Eclipse
- JDK 1.7
- VmWare

VMWare is software which provides a virtual machine. In order to move data, first, we need to connect to Linux OS. We are writing Map Reduce programs in java so jdk1.7 is required. Eclipse is used as it provides execution environment for developing.

CONCLUSION

In reality a particular person's CV/Resume document describes a lot of descriptive information required for respective recommended consultancy/industry/institutional organization. Manually analyzing huge amount of resumes is a tedious job therefore, taking this into consideration, it is observed that, such kinds of advanced resume analysis and processing application software can be responsible for finding required candidates which are suitable for the job. It reduces traffic on capture, storage, search, sharing, analysis and visualization. A huge amount of data could be stored and large computations could be possible in a single compound with full safety and security at cheap cost. Unstructured data handling is one of the burning issues in the present IT industry so, the work on Hadoop, being one of the best and scalable solutions, will surely be more useful. Hadoop supports all types of data formats and can process data exceeding TBs within seconds. This is the main advantage of Hadoop over traditional RDBMS.

ACKNOWLEDGEMENT

We are grateful to this institute for having channelized our skills and energy and encouraging us to work together with cooperation and co-ordination. We are indebted to our inspiring Head of Department Ms. Pooja Tripathi and our project guide Ms. Nidhi Gupta who have extended all valuable guidance, help and constant encouragement throughout the various difficult stages in the development of the project.

REFERENCES

I. MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat, Google Inc

II. Big Data and Methodology- A Review by Shilpa and Manjit Kaur

III. Efficient Big Data Processing in Hadoop

IV. MapReduce Jens Dittrich Jorge-Arnulfo Quian e-Ruiz

V. Big Data, Black Book: Covers Hadoop 2, MapReduce, Hive,

YARN, Pig, R and Data Visualization Paperback – 2016 by DT Editorial Services (Author)

`

VI. Resume Parsing using Hadoop Akshata Walavalkar, Sheikh Abdul Rehman, Prachi Kore, Saurabh Nimbalkar

VII. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

VIII. IBM Big Data analytics HUB, www.ibmbigdatahub.com/infographic/four-vs-big-data

IX. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" in International Journal of Advance Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.