Resume Parsing using Hadoop
Akshata Walavalkar1 Sheikh Abdul Rehman2 Prachi Kore3 Saurabh Nimbalkar4
1,2,3,4Department of Information Technology Engineering
1,2,3,4Mumbai University, Padmabhushan Vasantdada Patil Pratishthan’s College of Engineering, Sion-Mumbai-22

Abstract— Big data is a collection of structured and unstructured data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. For Big Data processing Hadoop uses Map Reduce paradigm. First we will briefly explain Hadoop-Map Reduce and its use for Big Data processing. After explaining the current system we will explain how we are using HDFS to implement our system. Our implementation of Hadoop is in pseudo mode. We are using Map Reduce programs for processing and at the end output will be presented in the text format.

Key words: Bid Data, Parameters, Hadoop, Map Reduce, HDFS

I. INTRODUCTION

Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big data analytics is the process of examining large amounts of data.

A. Big Data:

With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web sales, social media, audios, images and so on. The need of big data comes from the Big Companies like Yahoo, Google, Facebook, etc. for the purpose of analysis of big amount of data which is in unstructured form. Google contains the large amount of information. So there is the need of Big Data Analytics that is the processing of the complex and massive datasets. Big data analytics analyze the large amount of information used to uncover the hidden patterns and the other information which is useful and important information for the use.

B. Big Data Parameters:

As the data is too big from various sources in different form, it is characterized by the 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity.

II. EXISTING SYSTEM

The current Relational Database Management Systems (RDBMS) are not capable for handling Big Data. A relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational model as invented by E. F. Codd, of IBM's San Jose Research Laboratory. Many popular databases currently in use are based on the relational database model. RDBMS has certain features such as: provides data to be stored in tables, persists data in the form of rows and columns, provides facility primary key, to uniquely identify the rows, creates indexes for quicker data retrieval, provides multi user accessibility that can be controlled by individual users. It has certain drawbacks such as requirement of structured data type and software license. Also it provides limited processing. Resumes do not have any fixed structured. They are unstructured data type. RDBMS do not accept unstructured or semi structured data making it difficult to store resumes. It takes a lot of time to process them.

III. PROPOSED SYSTEM

To handle unstructured data, we are implementing a system to retrieve large amount of data in fastest and reliable way. The aim of our project is to demonstrate and ease the information retrieval of unstructured data like resume using Hadoop and Map Reduce. Currently, there is no DBMS available to handle unstructured data. So the only option available is Hadoop. We will demonstrate how Hadoop accepts unstructured data like resumes and processes it faster.

A. Hadoop:

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity.

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's Map Reduce and HDFS components originally derived respectively from Google's Map Reduce and Google File System (GFS) papers.

<table>
<thead>
<tr>
<th>RDBMS</th>
<th>V/S</th>
<th>HADOOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Data Types</td>
<td>Multi and Unstructured</td>
</tr>
<tr>
<td>Limited, No Data Processing</td>
<td>Processing</td>
<td>Processing coupled with Data</td>
</tr>
<tr>
<td>Standards &amp; Structured</td>
<td>Governance</td>
<td>Loosely Structured</td>
</tr>
<tr>
<td>Required On write</td>
<td>Schema</td>
<td>Required On Read</td>
</tr>
<tr>
<td>Reads are Fast</td>
<td>Speed</td>
<td>Writes are Fast</td>
</tr>
<tr>
<td>Software License</td>
<td>Cost</td>
<td>Support Only</td>
</tr>
<tr>
<td>Known Entity</td>
<td>Resources</td>
<td>Growing, Complexes, Wide</td>
</tr>
<tr>
<td>Interactive OLAP Analytics</td>
<td>Best Fit Use</td>
<td>Data Discovery</td>
</tr>
<tr>
<td>Complex ACID Transactions</td>
<td>Processing Unstructured Data</td>
<td>Processing Unstructured Data</td>
</tr>
<tr>
<td>Operational Data Store</td>
<td>Massive Storage/Processing</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: RDBMS V/S HADOOP
This project is based on a back end system and hence, it will not have a proper GUI.

1) Command line interface will be used to give commands to transfer data from local machine to the virtual machine. Hence, it is referred to as the presentation tier.

2) The logic tier is Map Reduce. It is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data sets) in parallel on larger clusters (thousands of nodes) of commodity hardware in a reliable, fault tolerant manner.

3) HDFS is used for storing data. It is the Data tier. In this project, Resume parsing will be implemented. It will take text data as input and give output again in text format.

3-tier architecture is a standard model to handle business data.

B. Major Components of Hadoop:

1) HDFS Hadoop Distributed File System: Used for storage purpose and is distributed across nodes.
2) MapReduce: Used for processing purpose. It splits a task across processors, gathers the nearest data and assembles the results. High bandwidth and self-healing.
3) Name node: It provides the index of the data. It contains all the meta data about the data that is being stored in the HDFS. It contains details like location, name, replication factor, time and date of modification, etc.
4) Secondary Name Node: It copies data from Name node at hourly basis. It used when Name Node crashes.
5) Data Node: It stores all the actual data in blocks of size 64Kb.
6) Job Tracker: It schedules all the jobs required to carry out the necessary processes. It runs on Name Node.
7) Task Tracker: It carries out the tasks as per the schedule given by Job Tracker. It runs on Data Node.

C. Hadoop Workflow:

HDFS has a master/slave architecture. An HDFS cluster consist of a single Name Node, number of Data Nodes, usually one per node in the cluster. Internally, a file is split into one or more blocks and these blocks are stored in a set of Data Nodes. The Name Node performs operation like opening, closing and remanining files and directories. It also determines the mapping of blocks to Data Nodes. The Data Nodes are responsible for serving read and write requests, perform block creation, deletion and replication upon instruction from the Name Node.

A typical deployment has a dedicated machine that runs only on the Name Node. Each of the other machine in the cluster runs one instance of the Data Node software. The existence of a single Name Node in a cluster greatly simplifies the architecture of the system. The system is designed in such a way that user data never flows through the Name Node.

1) Hadoop Can Run In 3 Modes:

1) Standalone/Local mode: No major components, no distributed file system, everything runs and resides on single machine. Mostly suitable for running Map Reduce.
2) Pseudo distributed Mode: All major components are present on single machine. Both local and virtual machine reside on same machine.
3) Fully distributed Mode: All major components of Hadoop reside on different machines across the network.
This project will be implemented in Pseudo mode as we require the daemons for storing and processing purpose and distributed file system is available only for enterprises.

D. 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.

![Diagram](image)

Local machine → Virtual machine → HDFS

Fig. 4: Working

IV. REQUIREMENT ANALYSIS

A. Hardware Requirements:
- Pentium Dual Core or above Processor
- 4GB RAM
- 500GB HDD

B. Software Requirements:
- OS - Windows 7 / Windows 8 - 64 bits
- Eclipse
- JDK 1.7
- VmWare
- Cloudera
- FileZilla

We are using Cloudera software, which is connected to VMWare. Cloudera has few inbuilt functionalities of administrator, since we are working only as a developer, we will be using Cloudera. VMWare is a software which provides a virtual machine. In order to migrate data, first, we need to connect to Linux OS using FileZilla software. We are writing Map Reduce programs in java so jdk1.7 is required. Eclipse is used as it provides execution environment for developing projects.

V. CONCLUSION

It reduces traffic on capture, storage, search, sharing, analysis and visualization. A huge amount of data could be stored and large computations could be done 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 solutions, will surely be more useful. Hadoop supports all data formats and can process data exceeding TBs in seconds. This is main advantage of Hadoop over traditional RDBMS.

VI. FUTURE SCOPE

The estimation about the generated data is that till 2003 it was represented about 5 Exabyte, then until 2012 is 2.7 Zettabytes and till 2015 it is expected to increase 3 times.

Also on social networking sites the data is present in unstructured format. The same way resumes of companies have different format. Also each candidate may send resume in different format. So our system handles such data and gives output in the required format. So there is a need of such system in the market.

VII. 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 Mrs. Prachi Kshirsagar and Principal Mr. Rajendra Sawant and also our internal guide Mrs. Medha Kulkarni who have extended all valuable guidance, help and constant encouragement throughout the various difficult stages in the development of the project.

REFERENCES
[1] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat, Google Inc

All rights reserved by www.ijsrd.com