International Journal of Recent Engineering Science (IJRES)
ISSN: 2349 – 7157 / doi: 10.1445/23497157/IJRES-V5I1P102
Volume 5 Issue 1, 8-10, Jan-Feb 2018

P&ID Data Analysis using Hadoop Map Reduce

Jatti Mounika, Nagaveni B. Biradar
M. Tech Student, Dept of CSE, RYMEC, Bellary, Karnataka (India)
Associate Professor, Dept of CSE, RYMEC, Bellary, Karnataka (India)

Abstract
Hadoop is a framework to store, retrieve, and process a vast amount of raw data to/from Hadoop Distributed File System using the MapReduce programming model. This vast amount of raw data can be used for industrial or business purposes by organizing according to our requirements and processing. This paper provides a way of P&ID analysis using Hadoop, which extracts tags of respective devices present in the Piping and Instrumentation Diagrams of a particular industrial plant.

Keywords*-P&ID analysis, Hadoop Distributed File System,Map Reduce programming model*

I.INTRODUCTION

Today, the textual data of mechanical industries is also massive. Every mechanical component has respective identification data; it is challenging to analyze a large amount of data manually. This is where the piping and instrumentation diagram, or P&ID, comes into the picture, which shows the piping and related components of a physical process flow. It is most commonly used in the engineering field—a P&ID diagram which shows the interconnection of process equipment and the instrumentation used to control the process. The primary intention of this project is to extract tags of particular mechanical components with the file name, so in the other process, the extracted tags of the respective device's data are updated automatically to avoid the manual way of replacing/updating new information. In order to store, extract, and retrieve P&ID data, the Hadoop framework is used.

II. RELATED WORK

Big Data analysis is the most popular trend in today's world. Much work has been done in this sector. The following are some approaches that are most popular in today's world. There has been much research in the area of Big Data analysis. Current works in this area include using a Hadoop framework to extract detailed data to make business decisions. Our project uses the Mapreduce programming and Hadoop Distributed File System for distributed processing of the textual data.

III. HADOOP FRAMEWORK

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. The Apache Hadoop framework is composed of the following modules:

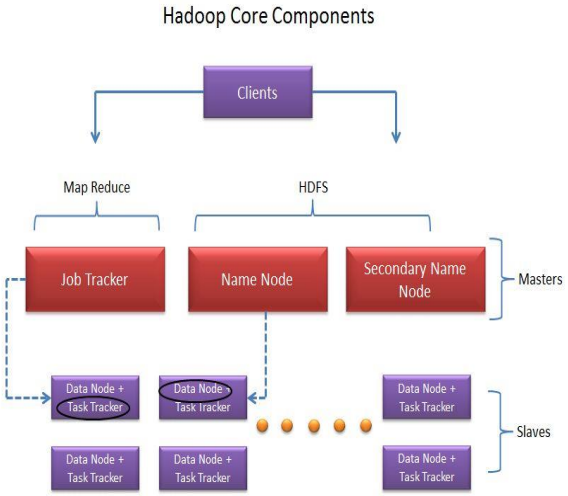

Figure: Hadoop Core Components

A. Hadoop Common

Contains libraries and utilities needed by other Hadoop modules

B. Hadoop Distributed File System (HDFS)

A distributed le-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.

C. Hadoop YARN

A resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.

D. HadoopMapReduce

Aprogramming model for large scale data processing.

IV. OUR APPROACH

In our approach, we focused more on the speed of performing data analysis than its accuracy, i.e., performing P&ID analysis using the Hadoop framework by splitting the various modules of data in the following steps and collaborating with Mapreduce programming.

A. P&ID Data Analysis

As Piping & Instrumentation Diagrams represent the interconnection of mechanical components like boilers, pipes, values, andseparate tags or names, it is easy to maintain a database of all the equipment present ina

particular plant or machinery. However, the great challenge facing by most hardware industries is the maintenance of new components or replaced components list in their database. Here comes the usage of Big data technology!

The considerable amount of data accumulated in every mechanical industry is considered as Big data; as its volume, velocity, veracity is high. P&IDs are one of the most common ways to represent or store data regarding mechanical equipment. The main goal of this project is to extract tags of particular mechanical devices, so in the other process, the extracted tags or names of respective component data will be updated automatically in order to avoid the manual way of updating new data. In order to store, extract, and retrieve P&ID data, the Hadoop framework is used.

Hadoop is not a cooking tool or framework with readymade features, but it is an efficient framework that allows many customizations based on our requirements. It is our choice to modify it. Modification is not about modifying the architecture or working, but the modification of its functionality and features.

By default,Hadoop accepts text files. However, in practical scenarios, our input files may not be text files. It can be PDF, PPT, PST image, or anything. In this project, P&ID are in PDF file format. So we need to make Hadoop compatible with these various types of input formats.

This project is about the analysis of Piping & Instrumentation Diagrams of a particular industrial plant. In P&ID data analysis,we can extract tags of various mechanical components along with file names, and we can update the database of respective devices automatically.

B. Creation of Custom Input Format for Hadoop

In order to read P&ID's of PDF file format, we have created a custom input format for Hadoop. For doing this logic, we need two classes. One is that we need a similar class like the default TextInputFormat class for PDF. We can call it PdfInputFormat.class. The second one is that we need a similar class like the default LineRecordReader for handling PDF. We can call it a PdfRecordReader class.

By default TextInputFormat class is extended from a parent class called FileInputFormat. So in our case, we can also create a PdfInputFormat class extending the FileInputFormat class. This will contain a method called createRecordReader, which it got from the parent class. We are calling our custom PdfRecordReader class from this createRecordReader method.

C. Creation of Simple PDF Parser

HDFS splits its blocks (byte-oriented view) so that each block is less than or equal to the block size configured. So it is considered to be not following a logical split. This means a part of the last record may reside in one block, and the rest of it is in another block. This seems correct for storage. Nevertheless,At processing time, the partial records in a block cannot be processed. So the record-oriented view comes into place. This will ensure to get the remaining part of the last record in the other block to make it a block of complete records. This is called input-split (record-oriented view).

Input Format is responsible for validating the input data, creating the inputsplit, and divide them into the records. Record Reader reads the data from the inputsplit (a record) and converts them into key-value pairs for the input to the Mapper class.

Record Reader uses the data within the boundaries defined by Input Split. It creates key-value pairs for the Mapper. The > "start" is the byte position in the file; thus,Record Reader starts generating key-value at 'start. Furthermore, the "end" is where it should stop reading records. In Map Reduce, Record Reader load data from its source, and it converts the data into key-value pairs suitable for reading by the Mapper. Record Reader communicates with the Input Split until it does not read the complete file. The Map Reduce framework defines the Record Reader instance by the Input Format.

The PdfRecord Reader is a custom class calling from createRecord Reader method. Now we need to write logic to read PDF file line by line using PdfRecord Reader class that extends RecordReader. This mainly contains five methods which is inherited from the parent RecordReader class, such as Initialize(), nextKeyValue(), getCurrentKey(), getCurrentValue(), getProgress(), close().

We are applying our PDF parsing logic in this method. This method will get the input split, and we parse the input split using our PDF parser logic. The output of the PDF parser will be a text which will be stored in a variable. Then we split the text into multiple lines by using '/n' as the splitter, and we will store these lines in an array. We need to send this as a key-value pair. So we are planning to send the line number as the key and each line as a value. So the logic for checking getting from the array, setting it as key and value, checking for the completion condition are written in the code.

D. Mapper, Reducer and Driver

For using Record Reader in a program, we need to specify PdfInput Format in the Driver class. Also, we need to call Mapper and Reducer classes in Driver class, which contains actual logic to extract tags from given PDF files. We need to set our custom input format class in the "InputFormat Class" property of the Map Reduce program.

In later steps, the Map Reduce code is composed of a jar file and run using the Hadoop jar command. Initially, the Mapper job is completed, and then the reducer job starts.

IV. ACCURACY

The overall accuracy of the project is determined by the time required to access various modules, i.e., accessing from HDFS and Hadoop clusters. As all components are in series, i.e., used one after the overall; theoretically, the overall accuracy of the program is the product of accuracy of all its modules.We tested our implementation on the standard P&ID's provided industry.

V. TIME EFFICIENCY

Time efficiency is an important aspect where our project scores well. Lower response time has been achieved by the use of the MapReduce programming model. This reduces the execution time from a Hadoop cluster. Also, the use of Hadoop ensures distributed processing, and it also lowers the access time.Hence overall, the time efficiency increases owing to thefactors mentioned above.

VI. FUTURE SCOPE

At this moment, the code can handle the data analysis part withoutstanding accuracy. Nevertheless, there are a few areas that havemuch scope in this aspect. Vertical data is one thatis very difficult to identify. PDF files containing vertical tags or names almost impossible totrack. Also, depending on the context in which a word isused, the interpretation changes,but the usage of vertical data is challenging to interpret.

VII. CONCLUSION

Data Analysis plays an essential role in determining business and marketing strategies. This project can play a crucial role in helping large scale industries to identify the currently using machinery components in the respective plant of that industry.The P&ID diagram data analysis is useful to replace the database automatically using tags of mechanical components so that one can avoid the manual way of updating new machinery information. HadoopMapReduce environment help in achieving huge amount of data analysis and transforming these data into decisions which has good impact on the real world. This can be used in businesses that extract useful information from unstructured data.

REFERENCES

[1] QiangYang, Fellow, IEEE, "*Introduction to the IEEE Transactions on Big Data,"* IEEE TRANSACTIONS ON BIG DATA, VOL. 1, NO. 1, JANUARY-MARCH 2015.

[2] Hadoop Setup, www.bogotobogo.com/Hadoop/BigData_hadoop_Install_ on_ubuntu_single_node_cluster.php

[3] Sara Riahi, 2 AzzeddineRiahi "*Visualization of Big Data with the Map-Reduce program execution platform: Hadoop*", International Journal of Engineering Trends and Technology (IJETT), V54(2),94-104 December 2017.

[4] Tom White, 2012, Hadoop: The Definitive Guide, O'Reilly

[5] Hadoop Tutorial, Yahoo Developer Network, http://developer.yahoo.com/hadoop/tutorial

[6] Ms. Kanchan Sharadchandra Rahate , Prof. L.M.R.J. Lobo. "*A Novel Technique for Parallelization of Genetic Algorithm using Hadoop"*. International Journal of Engineering Trends and Technology (IJETT).V4(8):3328-3331 Jul 2013

[7] HadoopRecordReader, www.amalgjose.com/tag/hadoop-recordreader/.

[8] T.Janani and K.Balamurugan, "*Load Rebalancing using Map reducing Task for Distributed File Systems in Cloud*" SSRG International Journal of Mobile Computing and Application 2.1 (2015): 1-5.

[9] Organizational information, www.techmahindra.com

[10] Pradeep H K, Rohitaksha K, Abhilash C B, "*An Email based Offline Download Manager for Large Distributed File System using HadoopMapReduce Framework*" SSRG International Journal of Computer Science and Engineering 1.10 (2014): 1-5.

[11] Papineni Rajesh, Y. MadhaviLatha"*HADOOP the Ultimate Solution for BIG DATA Problems"*International Journal of Computer Trends and Technology (IJCTT),V4(4):550-552 April Issue 2013 .

[12] Abdul RaufBaiga,*, HajiraJabeenb, *"Big data analytics for behaviour monitoring of students,*" "Symposium on data mining applications," SDMA 2016, 30 march 2016, Riyadh, Saudi Arabia.