Pre Processing in Big Data Analytics using Regular Expressions

[1]G. Mallikarjuna Reddy, [2]V.Harish Kumar, [3]D.Ganesh, [4]R.Arun Kumar

[1234] Assistant Professor, CSE Dept ,Vardhaman College of Engg ,Shamsabad ,Telangana

Abstract— **Big Data is a term which is used to describe massive amount of data generating from digital sources or the internet usually characterized by 3 V's i.e. Volume, Velocity and Variety. From the past few years data is exponentially growing due to the use of connected devices such as smart phone's, tablets, laptops and desktop computer. Moreover E-commerce which is also known as online market, internet services and social networking sites are generating tremendous user data in the form of documents, emails and web pages. This generated data volume is so vast and overwhelming which makes complex to process and analyze using traditional software systems consuming more time.In this existing system a pre-processing algorithm to extract real time user accessed data from windows operating system environment and an approach from Apache's Hadoop Distributed File System (HDFS) framework using Map Reduce functionality to mine and analyze this large dataset. The ability to mine and analyze Big Data gives organization richer and deeper insights into business patterns and trends. From the data collected, pre-processing algorithm transforms the data to a specific format. This means extracting, cleaning and loading of appropriate data to a text file takes place. This file is also known as a log file. have used java algorithm for pre-processing to extract and transform the data to given log file format. Analyzing Big Data have to be a challenging yet very compelling task. In this data is important. If data have junk input data then results does not produce accurate results. So Pre Processing in Big Data Analytics is important. Pre Processing algorithms exist but they consume so much time. Regular Expressions are very efficient in pattern matching rather than algorithms. That's the reason pre processing in big data analytics using Regular Expressions. Here that means proposed solution for data cleaning process is performing through the regular expressions. The primary objective of this paper is to explain advantage of regular expressions in data filtering or data filling.**

***Index Terms*—HDFS, Big Data, Regular Expressions.**

1) INTRODUCTION

Big Data has been an significant research topic in the software field for more than 10 years. A notable increase in capability to collect data from various sources such as connected devices, sensors, log records or click-stream in web discovering and other applications have been witnessed that too in different arrangements whether structured or unstructured. This has outpaced our capability to store, understand, process and analyze this large datasets. Considering the internet data, the web pages indexed by Google were about one million in 1998 and reached to one billion in 2000 and now have already topped in trillions. The reason for this growth is mainly due to evolution of social networking websites such as Facebook, Twitter, LinkedIn, MySpace, etc. letting users to create content freely thereby expanding the volume of data over the internet. From this heap of Big Data, data must be discovered and converted to knowledge to help improve our verdict making and make this world a better place.

Nowadays, the quantity of data that is shaped every two days is estimated to be 5 Exabyte's. This amount of data is alike to the amount of data created from the dawn of time up until 2003. Moreover, it was projected that 2007 was the first year in which it was not possible to store all the data that we are creating. This massive amount of data opens new stimulating discovery tasks [1].

The term 'Big Data' seemed for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress" [7]. Each day Google has more than 1 billion requests per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million apprises per day, and YouTube has more than 4 billion views per day. The data shaped nowadays is estimated in the order of Zettabytes, and it is growing around 40% every year. A new large source of data is going to be made from mobile devices, and big corporations such as Google, Apple, Facebook, Yahoo, Twitter is starting to look carefully to this data to find useful designs to improve user experience. Alex 'Sandy' Pentland in his 'Human Dynamics Laboratory' at MIT, is deed research in discovery patterns in mobile data about what users do, and not in what people say they do [2]. We need new algorithms and new tools to contract with all of this data.

Doug Laney [3] was the first one in talking around 3 V's in Big Data management:

Volume: There is more data than ever before; its size continues cumulative, but not the percent of data that our tools can process.

Variety: There are many dissimilar types of data, as text, sensor data, audio, video, graph and more.

Velocity: Data is arriving unceasingly as streams of data, and we are interested in obtaining useful data from it in real time.

Nowadays, there are 2 more V's:

Variability: There are variations in the structure of the data and how users want to interpret that data

Value: Business value that gives group a compelling advantage, due to the ability of making decisions based in answering queries that were previously considered beyond reach[4].

Data mining is the process of discovering useful information and deriving patterns by using certain data mining algorithms. It uses the Information Discovery in Database (KDD) process which includes data cleaning, data

2.Literature Survey

Literature review is mainly carried out in order to examine the background of the current project which helps to find out faults in the existing system & guides on which unsolved problems that can work out. So, the following themes not only illustrate the contextual of the project but also uncover the problems and flaws which motivated to suggest solutions and work on this project. A variety of research has been done on knowledge of collective conduct. Following section discovers different references that discuss near several topics related to shared behavior.

Existing Work

In [10] authors current a HACE (Heterogeneous, Autonomous, Complex and Evolving) theorem that typifies the features of the Big Data revolution, and suggests a Big Data processing model, from the data mining viewpoint. This data-driven model includes demand-driven combination of information sources, mining and examination, user interest modeling, and security and privacy thoughts. They analyze the stimulating issues in the data-driven model and also in the Big Data rebellion.

In [11] authors present the KEOPS data mining practice centered on domain knowledge integration. In this paper, the authors emphases first on the pre-processing ladders of business understanding and data sympathetic in order to build an ontology ambitious information system (ODIS). Then they demonstrate how the knowledge base is help for the post-processing step of perfect interpretation. Detailed the role of the ontology and describe a part-way interestingness measure that integrates together objective and subjective criteria in teaching to evaluate model relevance rendering to expert knowledge.

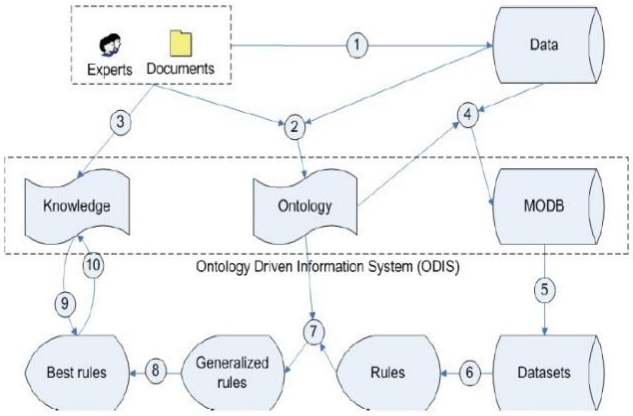

Fig 1. KEOPS Methodology

In [12] authors current insight about Big Data mining infrastructures and the involvement of doing analytics at Twitter. Two major topics are debated here. First, schemas play an vital role in helping data scientists appreciate peta byte-scale data stores, but they are inadequate to provide an complete "big picture" of the data obtainable to generate insights. Second, a major challenge in building data analytics stages stems from the heterogeneity of the various mechanisms that must be integrated together into production workflows- refer to this as "plumbing". The goal of this paper is to share involvements at Twitter for academic researchers to offer a broader context for data mining in making environments, pointing out chances for future.

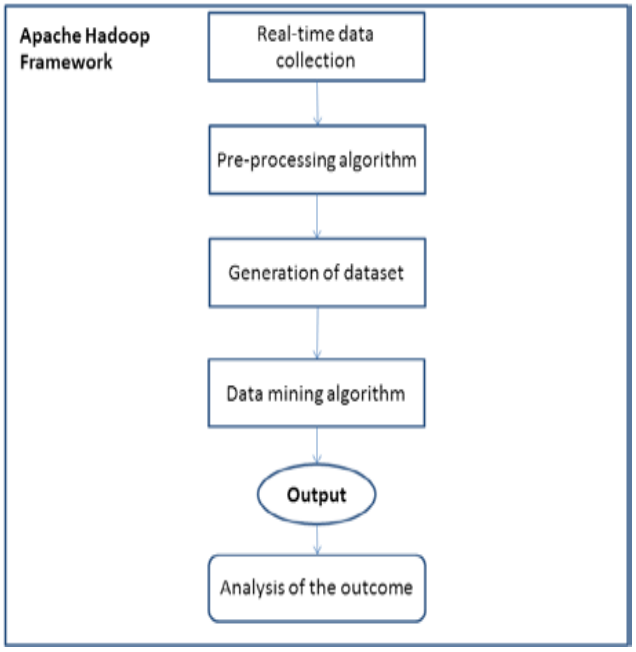

Fig 2. Block diagram of Existing system

The important stages as part of the existing system can be alienated into two major category:

A] Data Pre-processing

1. Real-time data collection

Here the data from distinct user machine which is generated by retrieving different files and folders is collected with the help of pre-processing procedure to extract the relevant information. In our experimental arrangement we have used data collected from the windows working system machine where a user is execution certain activity like accessing certain records and folders

2. Pre-processing algorithm

From the data collected, pre-processing procedure transforms the data to a specific arrangement. This means removing, cleaning and loading of suitable data to a text file takes place. This file is also recognized as a log file. We have used java algorithm for pre-processing to excerpt and transform the data to given log file setup.

3. Generation of dataset

Dataset means a group of data. Here applicable data are grouped together to form a dataset. In this investigational setup, the dataset contains the timestamp, type & name of file/directory accessed. This dataset which is essentially a text file is given as an contribution to the data mining algorithm.

B] Data mining and Analysis

1. Data mining algorithm

Data mining algorithm is applied to the generated dataset or the text/log sleeve to track the number of files/directory accessed in dissimilar time period. This is one of the most core steps in the process. Data mining algorithm printed in java using Apache Hadoop HDFS and MapReduce purposes are used for mining and analyzing.

2. Analysis of the outcome

Output is examined by tracking files/directory accessed by month/year and giving unlike input criteria. Also same output can be charted in Microsoft Power Business Intelligence (BI) which is a powerful BI tool to signify the output in the form of charts and graphs for imagining and better insights.

EXPERIMENTAL SETUP & RESULTS

Desktop machine with Linux operating system and Apache Hadoop packages as well as java installed.

For assessing the output of the proposed system, we have used semi-structured data which is essentially a text file containing thousands of records.

Different output results obtained after execution the proposed steps are as shown below:

1. Input: Pre-processing java algorithm

Fig 3. Output of Java Pre-processing algorithm

Output: It's a log file with timestamp, type and name of the file/directory

2. Input: Log file given to Hadoop MapReduce data mining algorithm

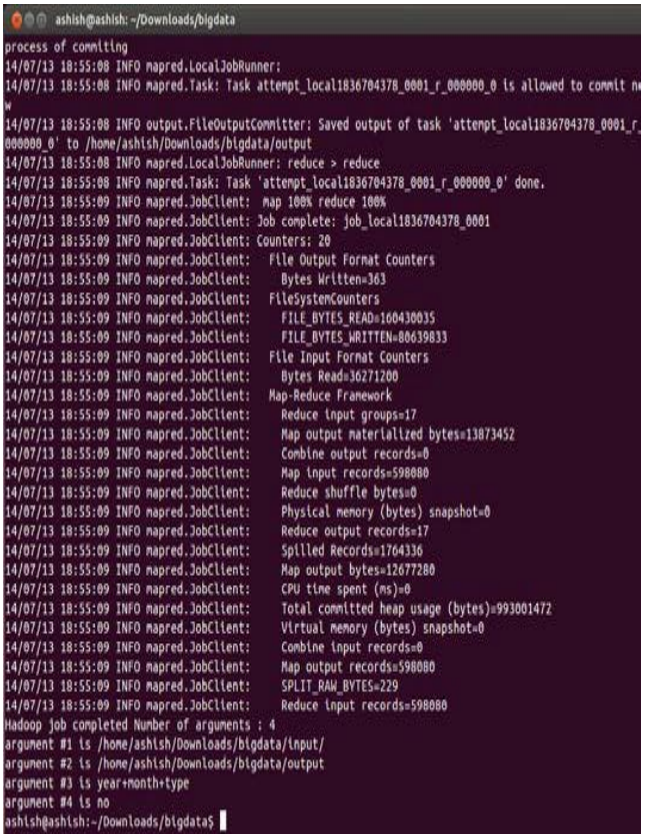

Fig 4. MapReduce processing

Output: We can observe that different MapReduce Jobs run simultaneously to get the output within few seconds.

3. Input: Applying analytics criteria to data mining

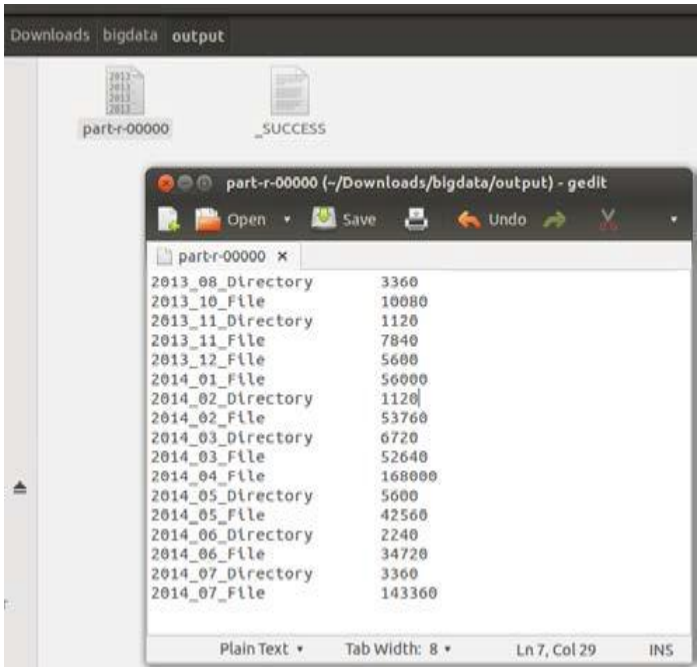

Fig 5. Output of Hadoop MapReduce mining algorithm

Output: Number of files and directories accessed by year and month

3.Regular Expressions

A regular expression (regex or regexp for short) is a unusual text string for describing a search pattern. You can think of regular expressions as wildcards on steroids. You are possibly familiar with wildcard representations such as *.txt to find all text files in a file manager. The regex equivalent is «.*\.txt» . But you can do much additional with regular expressions. In a text editor like EditPad Pro or a particular text processing tool like PowerGREP, you could use the regular expression «\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b» to search for an email address. Any email address, to be exact. A very alike regular expression (replace the first \b with ^ and the last one with $) can be used by a computer operator to check if the user entered a properly formatted email address. In just one line of code, whether that code is printed in Perl, PHP, Java, a .NET language or a multitude of other languages.

4.Comparision between existing data pre-processing methods and Regular Expressions

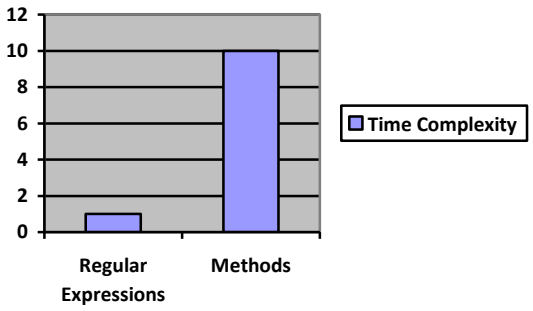

Fig:1 **Comparison between pre-processing methods and Regular Expression**.

Generally any method for pre processing contains minimum of 10 statements. let us assume each instruction consume 1ns so that 10 statements consume 10ns. but regular expression we can specify exactly one statement . So using regular expression for data pre processing provides minimum 10 times faster than existing methods.

2)

5.References

1. Salvador Garcia, Sergio Ramirez-Gallego, Julian Luengo, Jose Manuel Benitez and Francisco Herra,"Big Data Preprocessing :methods and prospects", Garcia et al. Big Data Analytics 1:9 DOI 10.1186/s41044-016-0014-0 , (2016).
2. Ashish R. Jagdale, Kavita V. Sonawane, Shamsuddin S. Khan " Data Mining and Data Pre-processing for Big Data" , International Journal of Scientific & Engineering Research, Volume 5, Issue 7, July-2014 1156 ISSN 2229-5518.
3. Gonzalo Navarro and Mathieu Ranot "Fast Regular Expression Search" https://www.dcc.uchile.cl/~gnavarro/ps/wae99.pdf
4. Bifet, "Mining Big Data in Real Time", Informatica 37, pp.15-20, 2013.
5. Petland. Reinventing society in the wake of big data. Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data, 2012
6. D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.
7. W. Fan, A Bifet, "Mining big data: current status, and forecast to the future", ACM SIGKDD Explorations Newsletter, Vol. 14, pp.1-5, 2013.
8. Mining Big Data in the Enterprise for Better Business Intelligence, Intel White Paper, July 2012
9. R Anand, J David, "Mining of massive datasets", Cambridge University Press, 2012 E Bertino et al. "Challenges and Opportunities with Big Data", 2011.
10. F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.
11. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", (In Press) IEEE Transactions on Knowledge and Data Engineering, 2013.
12. L. Brisson and M. Collard, "How to Semantically Enhance a Data Mining Process" Enterprise Information Systems, Springer Berlin Heidelberg, Vol. 19, pp. 103–116, April 2010.
13. J. Lin, D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience", ACM SIGKDD Explorations Newsletter, Vol 14, pp.6-19, 2013
14. Jan Goyvaerts, **"Regular Expressions: The Complete Tutorial"** http://www.regular-expressions.info/print.html