International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014
ISSN 2091-2730

Coupling based BigData Analysis – Reusability of Datasets

Thirunavukarasu B[1], Vasanthakumar U[1], Vijay S[1], Dr Kalaikumaran T, Dr Karthik S

[1]Research Scholar (B.E), Department of Computer Science and Engineering, SNS College of Technology, Coimatore, India

E-mail- bs.thirunavukarasu@gmail.com

Abstract— Presently we are in the BigData era. Many organizations and enterprises are dealing with massive set of data. These data are to be analyzed for various factors. For easy and effective data analysis, many a methods are used. Here we proposed Coupling based BigData analysis. Here the dataset are initially coupled so that the optimization can be achieved. Before one performs analysis of massive set of data, the dataset are coupled or grouped based on some kind of predefined available methodologies. Reusability of previously extracted datasets are used for quicker execution.

Keywords— BigData, Coupling, Analysis of BigData, Coupling Analysis, Optimized analysis, predictive analysis.

.

INTRODUCTION

Without any data, one cannot do anything. For every actions carried out, there generates data. Due to increased generation of data from various sources, organizations are supposed to store large amount of data. It's useless simply storing large amount of data if we are not using those data in future. So always the stored massive amount of data should be analyzed to get some prediction and output. For this purpose of analyzing large data sets, Hadoop tool was used. Hadoop is a distributed database management system of BigData.

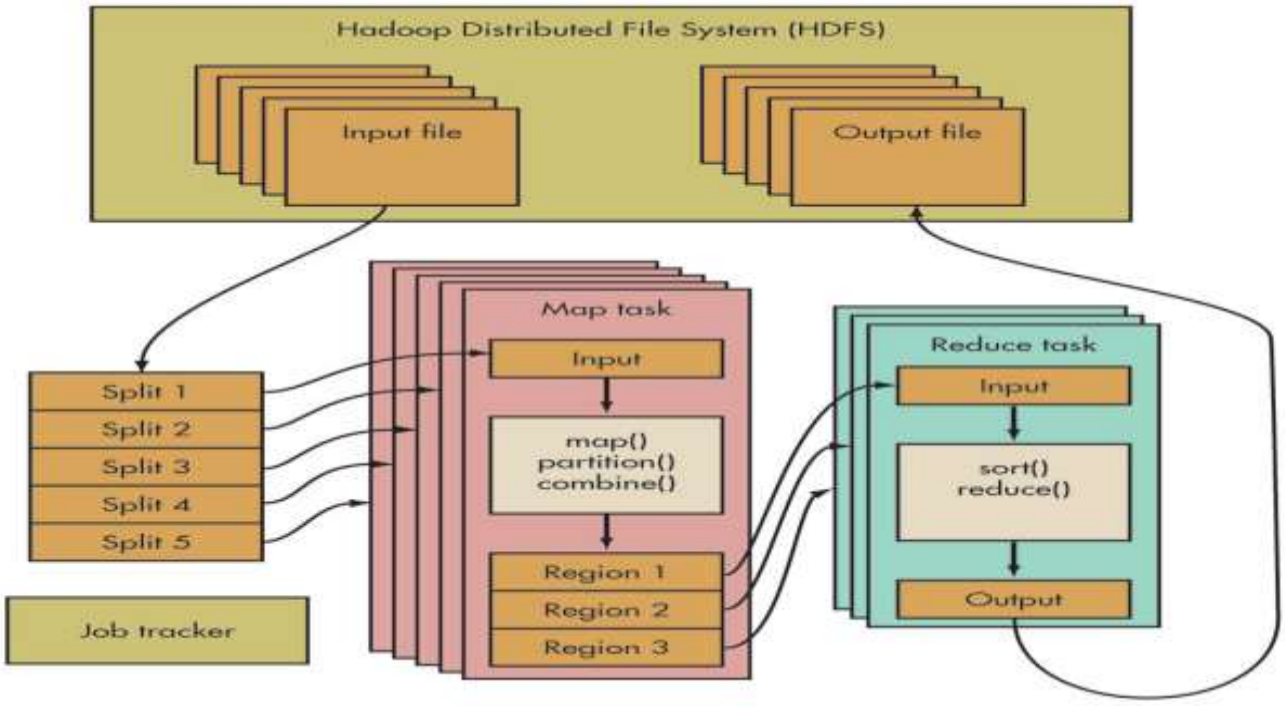

Fig 1. BigData Hadoop Tool Architecture

BIGDATA

Big data is an unstructured large set of data even more than peta byte data. Unstructured data is a data which is in the form other than common forms of tables and rows. There will be any items logs. Big data can be of only digital one. Data analysis become more

complicated because of their massive set of large amount of data. There are many availability of tools that can be easily used for analysis of this massive datasets. Predictions, analysis, requirements etc., are the main things that should be done using the unstructured big data. Big data is a combination of three v's those are namely volume, velocity and variety. Big data will basically processed by the powerful computer. But due to some scalable properties of the computer, the processing gets limited. Organizations and entrepreneurs are now forced to work with larger amount of data that are generated during their work. This data are the primary thing that could be used for many improving actions that needed to be taken in the near future.

BIGDATA ANALYSIS

Big data analytics is the application of advanced analytic techniques to very large, diverse data sets that often include varied data types and streaming data. Big data analytics explores the granular details of business operations and customer interactions that seldom find their way into a data warehouse or standard report, including unstructured data coming from sensors, devices, third parties, Web applications, and social media - much of it sourced in real time on a large scale. Using advanced analytics techniques such as predictive analytics, data mining, statistics, and natural language processing, businesses can study big data to understand the current state of the business and track evolving aspects such as customer behaviour. New methods of working with big data, such as Hadoop and Map Reduce, also offer alternatives to traditional data warehousing.

Analytics, providing deep insights on Big Data to optimize every customer touch point. Using personalized workspaces and self-service templates, analytics are rapidly assembled, customized and shared across business teams.

COUPLING

Coupling defines the integration between elements to do a particular user need. Here the coupling of design element represents the strength of connectivity between elements. Coupling is in different types. Among then "Highly-Coupled" and "Loosely Coupled" types are playing important role to group the components. If a particular element is not fully depend on other elements in the system then it is in "Loosely-Coupled" type of connectivity. Else were it is called "Highly-Coupled".

TYPES OF COUPLING

Apart from the above mentioned major types of coupling, the following indicates the deeper coupling types.

1. Import Coupling
2. Export Coupling

Import coupling is type of coupling that groups the elements that are going to be referred to support other components. Export coupling indicates the group of components that needs other components for the support.

COUPLING ANALYSIS – PROPOSED METHODOLOGY

In Coupling Analysis, the Dataset have to be analyzed as per the data recommended for grouping them. In "Coupling based" method the "Coupling" value and type of that considered for having clustering of Datasets. From that here we can also consider "Stability" to have a good quality repository with stable Datasets. In the 'Coupling-Based' approach, the type of the coupling may be considered only to categorize the Datasets. But in this proposed approach from the 'Import' and 'Export' coupling of a component will be used.

The Component reusability can achieved by implementing the coupling. Reusability of data is nothing but using the previously processed data sets without creating new datasets with the same resource of data gathered for every analysis execution. The reusable components are found with the help of the Coupling factors. This Coupling facilitates the improved reusability thus minimizing the overall execution time. As simple, the steps are followed as same till the data are processed by MapReduce.

The following flow diagram indicates the sequence of the operations or actions that are to be carried out for the optimized analyzing of the massive amount of the data set.

FLOW OF PROPOSED METHODOLOGY

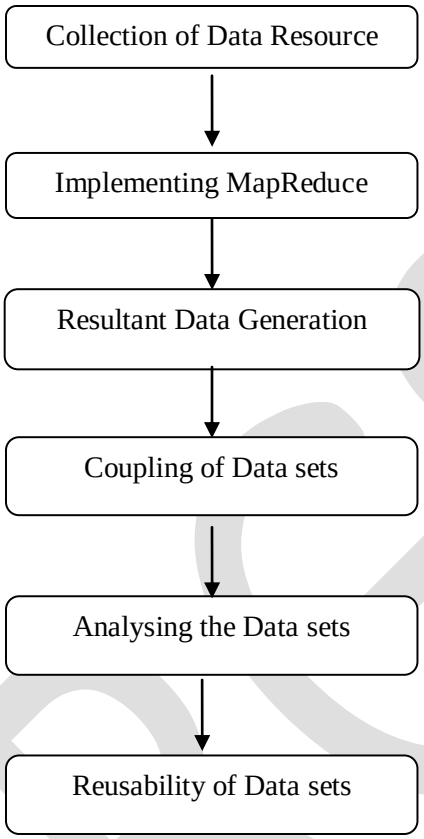

Fig 2. Sequence of Proposed Method

In this proposed methodology, initiallythe dataset are gathered from various resources. Once the resources are gathered, the data are divided into blocks. This blocking of the data are made for the purpose of enhancing the distributed system. The blocks are then placed in the DataNodes based on the alignment given by the NameNode. In the DataNode, the MapReduce Algorithm is executed. By Map the data are again divided, and by the Reduce algorithm, the divided data are executed separately.

Once the reduce algorithm is executed, the separate data results are grouped together. These grouped data is nothing but the resultant data. This resultant data is sent back to the client by means of the TCP connection. One the data is received, it is used for the analysis. At the Analysis phase, the coupling is implemented. Here the generated resultant data is checked for the coupling factors. By which one could know how this particular data set is dependent on or coupled with the other datasets. If once the resultant data is highly coupled, it could be considered as the reusable data set.This reusability could be found by the coupling value. The coupling value is a numerical value that indicates how a component is coupled with other components.

The Dataset reusability is only achieved with the coupling value is high. This highly coupled data set is replicated and stored separately for the data history usability. And so once a certain dataset needs the previous datasets to analyze, and if the coupling is available between the data set that we had stored previously, the new data comprises of entire data is avoided and by only the new data is executed to get the resultant data set. This resultant data set is coupled with the previously stored reusable data set and the analysis is made in effective manner. The Previously executed datasets are stored in Data Mart. By storing this previously executed dataset in data mart, the repeated data execution could be more fairly avoided. From the proposed method, we get some results. The transmission time gets reduced when the amount or volume of the data gets reduced. By this the transmission time is directly propositional to the volume of the data.

ACKNOWLEDGMENT

I heart fully thank The Department of CSE, SNS College of Technology for the effective encouragement in achieving some milestones in BigData research. I also thank Dr. S N Subbramanian, Chairman - SNS College of Technology, for his endless support and guidance.

CONCLUSION

Thus by the coupling of datasets, the Reusability of the same is achieved. As a result, the analysis of data is made in easy and fair manner. The dataset coupling avoids the repeated execution of data sets. Only the newly generated data are forced to the execution by MapReduce algorithm. Since the new data alone are executed, the Massive volume of data is considerably reduced into smaller amount of data, by this the execution time and transmission time also gets reduced. The datasets are reused without being made for circular execution. Thus Optimization of BigData analysis is achieved.

REFERENCES:

[1] M. Halkidi, D. Spinellis, G. Tsatsaronis and M.Vazirgiannis,"Data mining in software engineering", Intelligent Data Analysis, 2011.
[2] Man Deep Kaur, Parul Batra and Akhil Khare "Static analysis and run-time coupling metrics ", Oriental Journal of Computer Science & Technology, Vol. 3(1), 2010.
[3] Big Data Processing with Hadoop-MapReduce in Cloud Systems,Rabi Prasad Padhy,Senior Software Engg, Oracle Corp.,Bangalore, Karnataka, India
[4]] Marko Grobelnik, "Big-Data Tutorial" Stavanger, May 2012.
[5] Guanying Wang.,"Evaluating MapReduce System Performance: A Simulation Approach", August 2012.
[6] Robert D. Schneider,"Hadoop for Dummies", John Willey & sons, 2012
[7] Critical Study of Hadoop Implementation and Performance Issues,Madhavi Vaidya,Asst. Professor, Dept. of Computer Sc., Vivekanand College, Mumbai, India.
[8] Thirunavukarasu,Sangeetha K, Kalaikumaran T, Karthik S, "Effectively Placing Block Replicas of Big data on the Rack by Implementing Block Racking Algorithm," International Journal of Science and Technology Research , pp.891-894, April 2014