Big Data and Machine Learning Integration: The Benefits and Research Issues in the Huge Data Processing

B Priyanka K. UmaPavanKumarIndivarShaik

Abstract: ***The generation of the data from individual member to MNC incurring more burden on the existing architectures. The current requirements of processing and storing huge data may not be suitable to the existing storage and processing techniques. The fundamental issue is kind of the data populated every second in the social media even reaching to peta bytes of the storage the processing of this huge data is another problem. Here the concept of big data comes into the picture,Hadoop is a frame work which is helpful to store huge amounts of the data and to process the data in parallel and distributed mode. The framework is the combination of Hadoop Distributed File System(HDFS) and Map Reduce(MR). HDFS is a distributed storage which allows huge storage capacity solves the issue of abnormal data population, whereas the processing of the data is taken by the Map Reduce which provides a versatile model of processing the huge amounts of the data.***

The other dimension of the current work is to analyze the huge amounts of the data which is beyond the scope of Hadoop based tools. Machine Learning (ML) is a class of algorithms provides various techniques to analyze the huge data in a better possible way. ML provides classification techniques, clustering mechanisms and Recommender systems to name a few. The importance of the current work is to integrate the Hadoop and R which in turn the combination of Big data and ML. The work provides the key benefits of such integration and future scope of the integration along with possible research constraints in the reality. We believe the work gives a platform to researchers so as to extract the future scope of the integration and difficulties faced in the process.

Keywords: Hadoop, Framework, R, Parallel Processing, Distributed Storage.

I. INTRODUCTION

Hadoop Distributed File System(HDFS) got base form Google File System, based on the block storage the HDFS which gives the usage of the memory in 64MB or 128MB blocks we can alter the block size accordingly.

HDFS is a distributed storage model which provides default storage model to Map Reduce(MR), HIVE, HBase, Pig, Sqoop and Flume kind of tools which were presented in Hadoop eco system. Map reduce(MR) is a process model which provides a way of running and getting the huge data processing in parallel and distributed methodology. To configure these models various .xml files were used in the Hadoop configuration. [1]

Hadoop-env.sh, core-site.xml, map-red.xml the files allow to configure Hadoop with local host ports and mentioning of source file path. The other tools are having significant usage in the eco system. Hive provides set of commands to load the data from local file system or from HDFS. Through this tool loading of huge amounts of the data such as millions of records to HDFS in seconds is possible. The other benefit is to integrate Hive and Pig so as to analyze the data reasonably and get outcome in the form of predictions. [2]

Pig is a tool which helps us with many customizations like loading of the data, joins and User Defined Functions. 10 lines of the Pig code can give the performance of 200 lines of java code so we can image the advantage of Pig benefits. HBase belongs to the category of NoSQL data bases,HBase provides a way to arrange the data in column family fashion. The advantage of doing so is the same record may contain multiple column families. The entry of each and every record tagged with timestamp so that the tracking of various column family records is possible. The main benefit of this methodology is to effectively utilize the storage. [3]

The tool Sqoop provides a way to import/export the data from external sources in the form of relational model. The source might be oracle, MySQL, SqlServer etc., Flume is a tool to provide an API kind of the mechanism so as to capture the unstructured data such as twitter data so as to process the tweets and other messages. With the tools mentioned Hadoop is rich in providing the import/export of the data in both structured/unstructured formats along with storage and processing of the data in parallel and distributed mode. [4]

II.RESEARCH ISSUES IN HADOOP ECO SYSTEM FRAMEWORK

Hadoop eco system provide solutions for big data problems such as past 150 years' data of various countries and getting the maximum temperature recorded in each year. The problem is very tedious and need distributed and parallel mechanisms so as to capture and analyze the data. [5]

Some of the issues were there in the storage point of view.

i. In case of block storage either maximum block or minimum block won't solve the exact storage of the required

Revised Version Manuscript Received on 10 September, 2019.

B Priyanka*, Research Scholar, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.
(Email: priyanka.biradar511@gmail.com)
Dr.K.UmaPavanKumar, Associate Professor, Dept. of Computer Science and Engineering, Malla Reddy Institute of Technology,Hyderabad, Telangana, India.
(Email: dr.kethavarapu@gmail.com)
IndivarShaik, Research Scholar, Dept. of Computer Science and Engineering, SSSUTMS, Sehore, Madhya Pradesh, India.
(Email: indivarshaik@gmail.com)

file. The proposed idea is variable block storage method provides a way to address the issue of fragmentation of the storage. [5]

ii. In case of loading the data into MR the limitation MR knows only HDFS, sometimes there may be the possibility of running the task for testing purpose for this the provision of local file access. The local file can be provided to check the running possibilities of the task by MR which greatly reduce the time and space usage of the Hadoop eco system. [5]

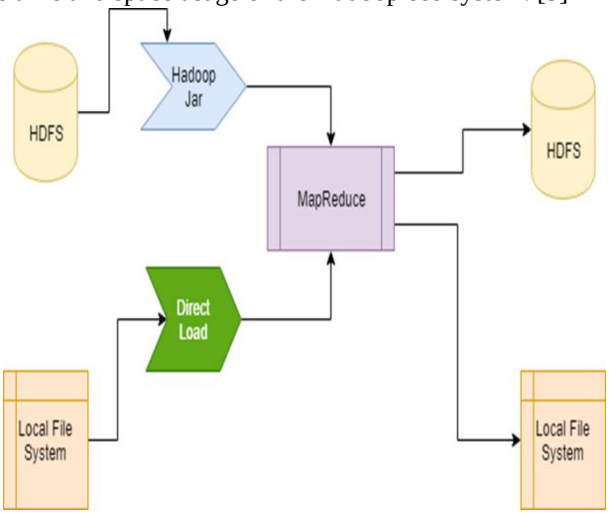

Figure 1: Proposed Refined Model of MR

III.ANALYTICS IMPORTANCE WITH HUGE AMOUNTS OF THE DATA

Consider the scenario of taking the feedback about the election results. For example, the few samples taken are favor to a particular part so automatically in the prediction the sample strongly influences the result. But the reality may be different like from the huge data collected to predict the election results as a part of exit polls gives a different result altogether compared with previous use case. [5]

So the exact predictions without bias can be expected from huge amounts of the data only. One simple scenario is in the movie reviews also few websites or persons can be given positive or negative reviews as bias based. In case of google reviews we can see the reality based on the actor's performance, story line, music and other parameters to conclude that whether to watch a movie or not. [6]

The Machine Learning(ML) is a class of algorithms provide the process of learning from the data, every time we run the algorithm can expect the better performance when compared with the previous execution.ML playing a vital role to address the analytics such as descriptive analytics, predictive analytics and recommender systems. [6]

Descriptive analytics with the help of statistical methods like max, min, avg, quantiles can be used to analyze the customer's credit card usage and can come up with certain observations like how he is spending the money on different products. In a way can find out the abnormal usage of credit card to identify the fraudulent transactions. Prescriptive analytics allows the analyst to come up with predictions like classification, clustering based logics. For example, grouping of the data items based on the similarities can be done with the help of clustering technique. [7]

The process of finding out whether play is going to happened or not based on the weather report of a day. Sometimes identification of the customers based on the Cibil, the credit issues and number of loans, age of the loan and age group to decide the eligibility of a customer. The most frequently used ML class of algorithms are recommender systems, in Amazon, flip kart so as to recommend the products based on the item based collaborative filtering and user profile based preference. [8]

Proposing a product to the user based on the past history, suggesting another product based on the past shopping trends of the other users and suggesting the other products along with the selected products, provision of reviews related to the products which can give a global view about the products. Likewise, ML algorithms having wide range to describe to perform analytics and come up with interesting patterns which widens the sales and provides plenty of products to the users to give a better experience. [9]

IV. ISSUES IN THE ANALYTICS PLATFORMS& RESULTS

As described earlier the ML class of algorithms provides a wide applicability of analytics which helps the customers as well as online platforms to have global business scenarios. Some issues were there in the discussion of the ML algorithms; ML provides better analytics if the source data is huge in nature. [10] [11]

i. The problem with analytics tools like R/Python provides various packages and methods to apply analytics on the given data but these platforms are having the limitation of memory based on the system they are installed.

ii. Huge data storing/processing is tedious in these platforms generally, as they are dependent on the current environment we are using.

iii. The Hadoop platform which we presented earlier is good in storage and processing of huge amounts of the data but there is no analytics by default in that platform.

So as to solve the above issues the following process of integration of Hadoop eco system with analytics support.

Rather than using HDFS store the integrated storage provides RHDFS which is capable of Storage and Analytics. Instead of using MR there is a possibility of using RMR which provides the distributed and parallel processing along with the capability of analytics also. Similarly, Rhbase allows the usage of NoSQL along with R so as to escalate the column family usage with analytics support. [12]

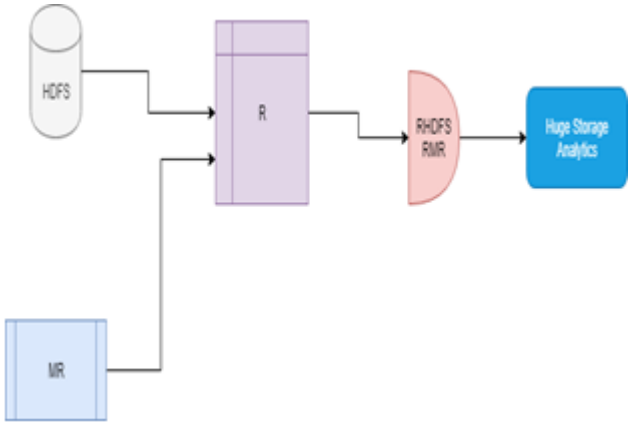

Figure 2: Integration of Hadoop and R

V.CONCLUSION

The work describes the usage of Hadoop eco system with various eco system tools like HDFS, MR, Hive, Pig Sqoop, Flume and HBase. Specified the usage significance of various tools. The work listed some of the research issues in the eco system such as storage and process related aspects. The other dimension we have seen is analytics context with the reference of ML class of algorithms. The article explained various algorithms provided by ML along various use cases. The issues were listed out such as memory limitations and other aspects.

Finally projected the integration of HDFS,MR with R so as to fulfill the huge amount of the data storage at the same time the provision of analytics to come up with predictions.

REFERENCES

1. S. Madden, "From Databases to Bigdata.," IEEE Internet Comput., vol. 16, no.3, 2012.
2. P. Zikopoulos, C. Eaton, and others,Understanding big data: Analytics for enterprise class hadoop and streamingdata. McGraw-Hill Osborne Media, 2011.
3. A. McAfee, E. Brynjolfsson, T. H.Davenport, D. J. Patil, and D. Barton, "Bigdata," Manag. Revolut. Harv. Bus Rev, vol.90, no. 10, pp. 61-67, 2012.
4. R. Appuswamy, C. Gkantsidis, D.Narayanan, O. Hodson, and A. Rowstron,"Scale-up vs Scale-out for Hadoop: Timeto rethink?," in Proceedings of the 4thannual Symposium on Cloud Computing, 2013, p. 20.
5. A. S. Tanenbaum and M. Van Steen,Distributed systems. Prentice-Hall, 2007.
6. T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to healthcare," Jama, vol. 309, no. 13, pp. 1351-1352, 2013.
7. I. Mashal, O. Alsaryrah, and T.-Y. Chung, "Performance evaluation of recommendation algorithms on Internet of Things services," Phys. Stat. Mech. Its Appl., vol. 451, pp. 646-656, 2016.
8. K. Shvachko, H. Kuang, S. Radia, and R.Chansler, "The hadoop distributed file system," in 2010 IEEE 26th symposium on mass storage systems and technologies (MSST), 2010, pp. 1-10.
9. Umapavankumar et.al. "Various Computing models in Hadoop eco system along with the perspective of analytics using R and Machine learning", International Journal of Computer Science and Information Security (IJCSIS) https://sites.google.com/site/ijcsis/ ISSN 1947-5500.
10. J. Y. Monteith, J. D. McGregor, and J. E.Ingram, "Hadoopand its Evolving Ecosystem.," in IWSECO@ ICSOB, 2013,pp. 57-68.
11. M. K. Islam and A. Srinivasan, Apache Oozie: The Workflow Scheduler forHadoop. O'Reilly Media, Inc., 2015.
12. C. Olston, B. Reed, U. Srivastava, R.Kumar, and A. Tomkins, "Pig latin: a notso-foreign language for data processing,"in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 1099-1110.