Processing Real-World Datasets Using Apache Hadoop Tools

N.Deshai, B.V.D.S.Sekhar, S.Vemkataramana

Abstract*: *Today's digital world applications are extremely demanding for essential requirements to significantly process and store an enormous size of datasets. Because the digital world applications tremendously generate mostly unstructured, unbounded data sets that beyond the limits and double day by day. During the last decade, many organizations have been facing major problems to handling and manage massive chunks of data, which could not be processed efficiently due to lack of enhancements on existing technologies also utilizing only existing centralized architecture standards and techniques. Only data processing with the existing technology and centralized environment, the organization actually faced difficulties of efficiency, poor performance and high operating costs, in addition to time pressures and optimization. These large organizations have been able to address the significant problems of trying to extract, store, process world massive data only with the assistance of Hadoop frameworks and distributed architectures. To overcome this problem as efficiently by using the latest open source Apache frameworks, which are, turn to centralized architecture to the more latest distributed architecture. In this paper, we use Apache Hadoop Map Reduce framework is one of the best data handling weapon and most exciting source of techniques and comprehensive features, which are widely used in the digital world computations regarding the distributed architecture manner, and more accomplish with high fault tolerance, high reliability, great scalability, great synchronization and data locality.

Index Terms*: *Big data, Apache Hadoop, Data processing, Map Reduce.

I. INTRODUCTION

Today's digital world applications are being generate wide-ranging of datasets at tremendous scale with high velocity [1, 2]. There has been an incredible and remarkable enlargement in the complexity of structured and unstructured world digital datasets. In reality, more than 90% of the world data sets were generated in the last decade. The name "big data" can signify as actual data, which enhances extremely big that it could not processed due to which utilizing by traditional techniques. The complete volume of the world organization data sets at complex scale, which seems Big Data. During make changing different factors and use innovative tools and techniques, finally keep enhancement a lot on to existing tools to significantly handle this world big data, because which are able to reshaping our world, which needs to be processed in a real-time manner.. Typically accomplish knowledge out of this vast volume of data sets, that could usually split behind utilizing 5 V's: Volume, Velocity, Value, Variety, and Veracity.

Which indicates world data like complex astonishing quantity of data sets, typically produced almost every second by social networks, smart phones, vehicles, bankcards, IOT motion sensors, web request logs, pictures, YouTube video, crawled documents etc. [3, 4]. In spite of the fact, the immense quantities of data are becoming so big that the used by conventional database technology then we really could no longer simply store and analyze the data. Then we really use decentralized mechanisms where many components of the data are also stored mostly at various places and technology-based. There are still 10 billion text messages with Face book alone anyway, 4.5 billion many times push the "like" button, and more than 350 million photos could be posted daily as shown in Fig.1. Accumulating and analyzing these, data is simply a tremendously massive engineering problem.

Velocity

Which refers to the high rate and speed at enormous volumes of world data which are significantly produced, gathered and analyzed as shown in Fig.2. Each time the huge number of emails, twitter letters, images, video clippings are rises at more lightning rates across the digital world. Each minute the world data is growing at high speed. Not simply analyze the massive data, also the velocity on release as well as accessing the complex data sets should continue instant toward support on behalf of real-time accessing to the different website, verifying the credit card and instantaneous messenger. Specifically big data realm which highly enables instantly to significantly analyze the world massive data since which is produced, not including always placing it on multiple world databases [5].

Value

This most specifically refers to the importance of the stored data. It is indeed one major thing to have significantly interminable quantities of data, although it is worthless unless it could be eventually turned it into real value [1]. However, there is a direct relationship at all between data and different perspectives; it does not always actually mean that Big Data has value. The more essential aspects from beginning work on a large data sets strategy are to easily recognize the limitations and advantages towards to collect and store the world large data that is collected could still fundamentally be made available to the public.

Variety

This is characterized as the numerous kinds of data that we could use now. Today's world data looks completely

Revised Manuscript Received on June 01, 2019.

N.Deshai, Department of Information Technology, Sagi Ramakrishnam Raju Engineering College, Bhimavaram, India.
B.V.D.S.Sekhar, Department of Information Technology, Sagi Ramakrishnam Raju Engineering College, Bhimavaram, India.
S.Venkata Ramana, Department of Information Technology, Sagi Ramakrishnam Raju Engineering College, Bhimavaram, India.

different mostly from previous data.

So we still have now no longer simply have since structured data (name, telephone number, actually address, financial statements, etc.) which certainly fits beautifully and smoothly into a table of data. Data from today is still unstructured [1]. In reality, 80 per cent of the total global data, also including pictures, YouTube videos, status updates, etc., fit into this whole category. The innovative new big data technology is still able to constantly collect and save, store and just use highly structured and unstructured data.

Veracity

The actual quality or reliability of the historical data is truthfulness. How precise are all this information. For example, consider all social media posts mostly with hash tags, acronyms, misspellings, etc, and the trustworthiness and exactness of this kind of things [1, 2, 3, 4, 5]. If whole value or reliability is just not accurate, gaining loads and loads of data will be of no use. Yet another great example is the use of Gps coordinates. The GPS usually "continues to drive" away while visiting a metropolitan zone. When residential buildings or many other constructions bounce off, satellite signals are simply lost. In this case, the location information must be melded in order to provide detailed data with that other data source such as road information or accelerometer data.

II. APACHE HADOOP

Apache Hadoop is a popular paradigm also great open source for storing and processing large-scale world data sets on commodity clusters. Hadoop seems to be a top-level Apache project developed and utilized by a worldwide contributing and customer community. Doug Cutting and Mike Cafarella actually created Hadoop technology in 2005[9]. Two difficulties mostly with big data are indeed mainly present decade. Firstly, it is many data to be saved, and secondly the massive data to be processed. Mainly due to the variability of the data, conventional methods such as RDBMS just are not enough. Therefore, Hadoop actually comes with some additional capacity to deal with this issue of huge datasets, namely the storing and significantly processing of enormous world data. The HDFS (Hadoop distributed file system) and YARN (Yet another resource negotiator) which are mainly two components. Hadoop atmosphere tremendously provides additional flexibility to programmers based on parallel computing. The world data communication overhead more especially from the massive data sets' transmission, which can affect the impact and accomplished huge performance greatly from Hadoop realm tools [9]. IDC estimates the digital world created a huge amount of data yearly will become 180 Zetta-bytes in 2025. Hadoop Map Reduce big size data processing software development, model. HDFS and Map Reduce: The HDFS and the Map Reduce paradigm parallel processing is really the main elements in the central core of Apache Hadoop [11]. Both initiatives are open source software influenced by Google-based technologies. The HDFS is decentralized, high scalability, small java file mechanism actually written on the Hadoop paradigm. HDFS can able to stores large amounts of data over a number of computers (generally in the terabyte to pet byte range). Reliability is achieved by repeating the data on various hosts, and therefore RAID storage capacity is not required on hosts. Typically, data is stored extremely on three endpoints mostly with the default option replication actual value: two together on the same rack and the other on another rack. Data access points can discuss data re-balancing, actually moving mostly around copies and maintaining a high replication of data [12]. The HDFS that includes a secondary name node that connects directly to the main name node and produces screenshots of the primary named one, which is something the current system will save to local or remote files [13]. The secondary named all these images could also be used to reboot a completely failed primary name code, and without replaying the entire file system actions and then edit the log to create an up-to-date file structure. Since the name node is the only point actually to store and manage metadata, a variety of data files, in fact, quite the opposite a wide range of little files can be supported as a junk. The new feature of the HDFS Federation is intended to address this issue to an extent that multiple namespaces are allowed with separate names. A real benefit of HDFS is that the job tracker and the task tracker really become data responsiveness. This benefit is not always possible while employing Hadoop with several other file systems [14]. This could have an essential outcome on job acquisition rates whenever running user data-rigorous jobs has been illustrated. HDFS has been specifically designed mostly for immutable data files and would not be appropriate for computer systems requiring additional write operations at the same time.

III. MAP REDUCE

SOME Map Reduce is a world latest primary data processing paradigm for easily writing applications and which is incorporated implementation for processing digital world massive data sets by a parallel, distributed algorithm toward a cluster [1-8]. This is a most effective programming data processing framework especially for enormous data-intensive computing to wide variety of applications based on more commodity hardware with reliable, fault-tolerant. Which is really a software development paradigm originally introduced by Google in 2004. This is now significantly implemented even in a number of real world data process and storage systems and which mostly seems to be a basic element of the greatest number of big data batch mechanisms. In fact, Map Reduce must be latest decentralized paradigm, which implements source code on multiple different nodes to be able to compute large quantities of data and which significantly convert huge data sets toward a flexible thing. This enables the calculation to manage big data quantities by providing additional devices–horizontal scaling. Then this differs from vertical scaling, which means that the efficiency of a single device is increased. Map Reduce typically address the useful essentials i.e. Map () extremely accomplish sort and filter the data and through combining them in the sort of group [15]. Map produces outcome regarding key-value pair, which is following on extremely, treated with the Reduce (), with tremendously performs

The summary by aggregate the mapped outcome. In simplistic, Reduce () uses the output produced by Map () as input and merges those tuples into an inadequate set of tuples as illustrated in Fig.3.

The "Map Reduce System seems to be well and popular foundation or structure which regulate the processing mostly with the different distributing servers, managing the numerous computations and tasks parallel, accomplishing total conversations and transfer the data among the several methods, and support to solve the issues of redundancy and fault tolerance.

The design is a specialism of the split-apply-combine policy for world massive data analytics. Which is excited by the map phase and reduce phase roles generally utilized in different operative programming, although their goal in the Map-Reduce structure is not the equivalent as in their initial stage, The essential participation of the Map-Reduce structure are not the real map and reduce roles just it has more scalable and high fault-tolerance able for a diversity of request by maximizing the execution tool significantly [16].

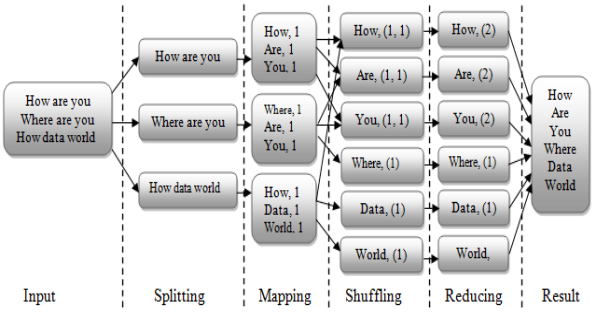

Fig.3 Map Reduce Word counting

IV. EVALUATION

In the past several decades, a wide variety of applications have been described in several disciplines by significantly increasing data quantities, improving analytics and developing new requirements on distributed systems for batch processing. When ambiguities about performance at work and tasks in diverse strategic resource setups are faced, the effective and efficient procedure of all of these systems is still challenging at times. g., the very first task mostly in this segment is Map Reduce runtime. Only for scheduling and assigning resources, The second task is to revert logistically using several variables. The third task is the forecast of optimized partitioning.

A. *The runtime of Map Reduce*

Because every node downloads its data from the HDFS, storing the relevant data onto the local hard drive or working memory and continues to perform, a Map-Reduce program is invoked20. For any further data processing, the estimate the outcomes from two buffers are necessary. Only the map component is used to the several other buffers in the foreground, whereas the running time loads this next data component into one of the buffers in the sort of background [17].

B. *Required Settings*

Here the file system block overall size is actually set to 64 Megabytes by default. The overall level of replication is 3. Every line can be tokenized into simple words by the map job, but every word21 has to be reduced. The labeling indicates the variables. The simple binary classifier classified mostly by the given variable is learning by each job [18].

Map Phase: For the Map phase of multiple jobs, a consistent model is developed. To begin with, findings the mapping phase and the implementation potential problem are ultimately determined. Then maybe a cost new model for actually reading local independent disk information as batch new jobs as well for the map phase is specifically stated. Try to increase the cost model for memory-based performance management [19].

Map algorithm:

Step 1: Primarily partition the data set.
Step 2: Secondly partition the data set.
Step 3: Apply join function on both data sets.
  i) Call init() function
    Read the first job file partition output
    add the first job to hash map.{key, list(V)}H;
  ii) Calling map function{k:null,V from B}
Step 4: check the if condition of (K in H) then for r in LV do
    For 1 in Hget.(K)do
    Call produce (null, tuple(r,1));

Reduce Phase:

The extremely long cost model could include the cost of the Shuffle transition phase in the numerous work reduction stage. Just let RC be a cost model for a particular job reduction phase. The reduction of CP / M times especially for each group in a cluster.

Reduce Algorithm:

Step 1: Calling map (K: null, V on Dataset A or Dataset B)
Step 2: Tag=bit from Dataset A or Data set B;
Step 3: Calling the Map method produce.{Key ,pair(V, Tag)};
Step 4: Calling Reduce (K1: join key, LV: list of V with key K1))
Step 5: Make Buffers Buf, (A) and Buf.(B) for Data set A or Data set B;
Step 6: for x in LV do
  start
    add x,v to Buf(Data set A) or Buf (Data set B) by x.Tag;
  stop
Step 7: for a in Buf (Data set A) do
  start 1
    for b in Buf(Data set B) do
      start 2
        produce (null, tuple (a.v, b.v))
      stop 2
  stop1

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. RESULTS

Table 1 describes relatively stable values, such as with compute nodes m= 20 both for Bayesian new network, the usually required memory and the patter classified into A-F, to achieve experimentation results. The graph signifies the node count for the A-F storage spaces in subgroups of 0-20 nodes. Average run time actually for new iteration is represented by the y-axis. The rows show the time, reducing and costs for the map.

Nodes are far less numerous when 20 nodes were involved. Just as the time of performance increases evenly, it decreased significantly.

Table 1. Details of Data, jobs, machines and clusters

Total Datasets and Jobs							
S No	Name	I	II	III	IV	V	VI
01	Size of the Data set	1	2	4	8	15	30
02	Jobs	30	15	8	4	2	1
03	Total Computation	70	70	70	70	70	70
Details of Machines							
	Type 1		Type 2				
Cores	Four		Four				
Total Memory	>4 Giga Byte		>4 Giga Byte				
Cache Size	2MegaByte		2MegaByte				
OS	Linux		Linux				
Kernel	26-18-194		26-18-194				
Types of clusters							
Type	Data Node	Name Node	Job Tracker	Complete Nodes			
01	01	01	01	01			
02	03	01	01	05			
03	06	01	01	08			

Table 2. Features of Map Reduce

Processing	Batch
Language	Java ,c , c++, ruby, groovy, Perl, python
Speed	Slow
Resource Management	Oozie scheduler
Stream	Batch-oriented
Computation	Configurable files
Storage data	HDFS
Optimization	Job manually optimized
Latency	High
Fault tolerance	High
Performance	Slow
Duplicate Elimination	High
Security	Kerberos
Iterative data flow	Chain of states
Isolation	Yes
Fault tolerant	Duplication Feature
Scalability	Incredible up to 10,000
Data flow	Chain of states
Cost	Low
Hardware	Commodity h/w
Machine learning	Mahout
Line of code	1,20,000
High availability	Yes
Amazon s3 connector	Support
Deployment	Fully distribute mode

When the reduction new phase is considered, the implementation time changes slightly from the map phase. The pricing model has a rather small impact, finally.
If the logical separations in the individual block are already 10 node clusters, then afterwards costs are high. The 10-node map initial phase has a couple of minute's run-time intervals are slight differences when nodes dramatically increase. The model significant reduction seems to have small runtime discrepancies. The optimum task reduces the time of implementation of 77% and 25% in comparison with the worst task. With this kind of optimal general pattern, the runtime could be reduced. The total cost is going to increase as nodes are subdivided into small blocks. With the absence of divisions, costs will increase. The outcomes examined even in this way raise the cost to over 150 even if the nodes are dramatically expanded to 40 partitions and to 200 if the costs of the nodes are 80.

CONCLUSION

In fact, more than 90% of present data sets were significantly produced in the last couple of years. Map Reduce is an advanced and latest distributed parallel programming paradigm, which allows the users to utilize batch-oriented data processing on multiple jobs by a little quantity of code. Which is more resilient in the insight, which can support write code to change the way, but

performing complicated data processing becomes large during every Map Reduce job should be aimed and listed on its own. The intermediate result from map tasks could be written to a file, which enables the paradigm to improve efficiency if a node becomes a failure. That resistance happens at a cost of execution, as the data could be transmitted to reduce jobs by a little buffer instead, building a stream. Hadoop realm significantly makes the handing world huge data in a extremely regarding more parallel based and well distributed manner across the big clusters of nodes through high scalable, reliable and high fault-tolerance.

REFERENCES

1. N. Deshai, S. Venkataramana, B. V. D. S. Sekhar, K. Srinivas, and G. P. S. Varma," Big Data Hadoop MapReduce Job Scheduling: A Short Survey," Information Systems Design and Intelligent Applications, Springer-Nature,vol.862,December.2018 , pp.349–365.
2. Deshai, Sekhar B. V. D. S, Venkataramana S, Chakravarthy V.V.S.S.S and Chowdary P.S.R, "A Study Comparing Big Data Handling Techniques using Apache Hadoop Map Reduce Vs Apache Spark, "Int Journal of Engg and Tech,vol.7,2018,pp. 4839–4843.
3. N. Deshai, S. Venkataramana, B. V. D. S. Sekhar, K. Srinivas, and G. P. S. Varma, "Big Data Hadoop Map Reduce Job Scheduling: A Short Survey**,"** Inf Sys Design and Intelli App, Springer-Nature vol.862, December.2018,pp. 349–365.
4. N.Deshai, S.Venkataramana and G.Pardha Saradhi Varma, "Performance and Cost Evolution of Dynamic Increase Hadoop Workloads of Various Datacenters," Smart Intelligent Computing and Applications, Springer-Nature, vol. 105,November 2018, pp.505–516.
5. N. Deshai, S. Venkataramana, B. V. D. S. Sekhar, K. Srinivas, and G. P. S. Varma," A Study on IOT Tools, Protocols, Applications, Opportunities and Challenges", Information Systems Design and Intelligent Applications, Springer-Nature, vol.862, December 2018, pp 349–365,
6. N.Deshai, S. Venkataramana and G.Pardha Saradhi Varma, "A study on analytical framework to breakdown conditions among data quality measurement," Int Journal of Engg & Tech,vol. 7,2018,pp. 167–172.
7. N.Deshai, S.Venkataramana, I.Hemalatha. & G.P.S.Varma, "A Study on Big Data Hadoop Map Reduce Job Scheduling," Int J of Engg & Tech, vol.7,2017, pp.59–65. N.Deshai and G.P Saradhi Varma," Big Data Challenges and Analytics Processing Over Health Prescriptions," Jour of Adv Research in Dyn & Cont Sys, vol.15, 2017,pp. 650–657.
8. K.Rattanaopas, S.kaewkere," Improving Hadoop MapReduce performance with data compression: A study using wordcount job," 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) ,2017,pp.564–567.
9. P.Merla, Yiheng.Liang,"Dataanalysisusing hadoop MapReduce enviro nment,"IEEE International Conference on Big Data,2017, pp.4783–4785.
10. M.Ahmed Qasim, B. Rajesh," An efficient technique to improve resources utilization for hadoop MapReduce in heterogeneous system,"International Conference on Intelligent Communication and Computational Techniques (ICCT), 2017,pp.12–16,
11. M.Bichitra, S.Srinivas, S.Ramesh Kumar," Architecture of efficient word processing using Hadoop MapReduce for big data applications," International Conference on Man and Machine Interfacing (MAMI)2015,,pp.1-6.
12. M. N. Kalimoldayev, S. Vladimir, M. N. Satymbekov, L. Naizabayeva," Solving mean-shift clustering using MapReduce Hadoop," IEEE 14th International Scientific Conference on Informatics, 2017, pp.164–167.
13. P.Amrit, J.Kunal, A.Pinki, A. Sanjay,"A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop," Fourth International Conference on Communication Systems and Network Technologies, 2014,pp. 587–591.
14. S.Jisha. Manjaly ; S. Chooralil. Varghese," TaskTracker Aware Scheduling for Hadoop MapReduce," Third International Conference on Advances in Computing and Communications, ,2013,pp.278-281.
15. A.Maedeh, S.Nishant, K.Suresh," Hadoop-MapReduce: A platform for mining large datasets," 3rd International Conference on Computing for Sustainable Global Development (INDIACom),2016, pp. 1856-1860.
16. C.Subhash., M.Deepak," An Approach to Enhance the Performance of Hadoop MapReduce Framework for Big Data," International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), 2016,pp.178-182.
17. D.AiLing, S.HaiFang," Research and Practice of Distributed Parallel Search Algorithm on Hadoop_MapReduce," International Conference on Control Engineering and Communication Technology, 2012,pp.105-108.
18. H.Hingave, I.Rasika," An approach for Map Reduce based log analysis using Hadoop, 2nd International Conference on Electronics and Communication Systems (ICECS), 2015, pp. 1264-1268.

AUTHORS PROFILE

Mr. Deshai Nakka, SRKR Engineering College (desaij4@gmail.com)srkr engineering college, department of IT chinnaamiram bhimavaram, Andhra Pradesh 534204, India. N Deshai is currently working as Assistant professor in the Department of Information Technology at S.R.K.R. Engineering College, Bhimavaram, and Andhra Pradesh (INDIA). His research interests are in the field of Big Data, Cloud Computing, Internet of things, Artificial Intelligence. He Published papers in national and international journals and also in international conferences including Springer. He has successfully guided a good number of the undergraduate and postgraduate thesis.

Mr. B V D S Sekhar, Andhra university (bvdssekhar@gmail.com) srkr engineering college, department of IT chinnaamiram bhimavaram, Andhra Pradesh 534204, IN B. V. D. S. Sekhar is a Research Scholar in Computer Science & Systems Engineering, Andhra University- Visakhapatnam, Andhra Pradesh (INDIA). He is currently doing his PhD on Image Processing He has published research papers in international journals and conferences. He was awarded Master of Technology in Computer Science in J.N.T.U, Kakinada, Andhra Pradesh (INDIA). His research interests are in the field of Image Processing, Optimization, wireless sensor networks, and Soft computing. Presently author is working as Associate Professor in the Department of Information Technology at S.R.K.R. Engineering College, Bhimavaram, and Andhra Pradesh (INDIA)

Dr. Venkataramana Sarella, SRKR Engineering College (vrsarella@gmail.com)srkr engineering college, department of IT chinnaamiram bhimavaram, Andhra Pradesh 534204, India He received PhD from Computer Science & Systems Engineering Department, Andhra University- Visakhapatnam, Andhra Pradesh (INDIA) in 2018. Presently author is working as Associate Professor in the Department of Information Technology at S.R.K.R. Engineering College, Bhimavaram, and Andhra Pradesh (INDIA). His research interests are in the field of Wireless Sensor Networks, Cloud Computing, and the Internet of things. He Published papers in national and international journals and also in international conferences.