www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 5 Issue 09 September 2016 Page No.18175-18179

Big Data Processing Using Hadoop in Retail Domain

Goddilla NagarjunaReddy(Student Member), M.V.Jagannatha Reddy(Senior Member)

arjunnag.edu@gmail.com

ABSTRACT:
In the ongoing situation, the volume of information utilized straightly increments with time. Long range interpersonal communication locales like Facebook, Twitter found the development of information which will be wild later on. So as to deal with the tremendous volume of information, the proposed strategy will prepare the information in parallel as little pieces in dispersed bunches what's more, total every one of the information crosswise over bunches to get the last handled information. In Hadoop structure, MapReduce is used to play out the errand of sifting, conglomeration and to keep up the productive stockpiling structure. The information are ideally refined utilizing collective sifting, under the expectation instrument of specific information required by the client. The proposed strategy is upgraded by utilizing the methods, for example, supposition investigation through regular dialect handling for parsing the information into tokens and emoticon based bunching. The procedure of information grouping depends on client feelings to get the information required by a particular client. The outcomes demonstrate that the proposed approach fundamentally builds the execution of multifaceted nature examination.
Index Terms: Big Data , Hadoop , MapReduce , Retail Domain .

1.INTRODUCTION

BIG DATA is a vague topic and there is no exactdefinition which is followed by everyone. Data that hasextra-large Volume, comes from Variety of sources,Variety of formats and comes at us with a great Velocity is normally refer to as Big Data. Big data can be structured,unstructured or semi-structured, which is not processed by the conventional data management methods. Data can begenerated on web in various forms like texts, images orvideos or social media posts. In order to process these largeamount of data in an inexpensive and efficient way,parallelism is used [1].

There are four characteristics for big data. They areVolume, Velocity, Variety and Veracity.

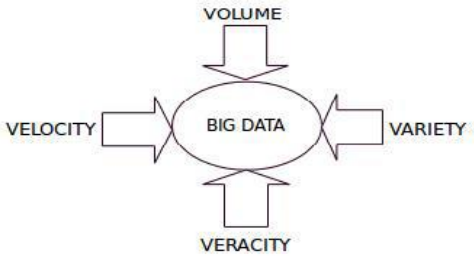

Fig. 1. 4 V's of BIG DATA.

Volume means scale of data or large amount of datagenerated in every second. Machine generated data areexamples for these characteristics. Nowadays data volumeis increasing from gigabytes to petabytes [2]. 40 Zettabytesof data will be created by 2020 which is 300 times from2005 [3]. Second characteristic of Big Data is velocity andit means analysis of streaming data.Velocity is the speed at which data is generated andprocessed. For example social media posts [2].Variety is another important characteristic of big data.It refers to the type of data. Data may be in different formssuch as Text, numerical, images, audio, video, social mediadata

[2]. On twitter 400 million tweets are sent per day andthere are 200 million active users on it [3].

Veracity means uncertainty or accuracy of data. Data isuncertain due to the inconsistency and incompleteness [2].

Previous Methodologies:

Big Data has come up because we are living in society that uses the intensive use of increasing data technology. As there exist large amount of data, the various challenges are faced about the management of such extensive data .The challenges include the unstructured data, real time analytics, fault tolerance, processing and storage of the data and many more. The size of the data is growing day by day with the exponential growth of the enterprises. For the purpose of decision making in an organizations, the need of processing and analyses of large volume of data is increases. The various operations are used for the data processing that includes the culling, tagging, highlighting, searching, indexing etc. Data is generated from the many sources in the form of structured as well as unstructured form .Big data sizes vary from a few dozen terabytes to many petabytes of data. The processing and analysis of large amount of data or producing the valuable information is the challenging task. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main problems regarding big data are the storage capacity and the processing of the data.

Our Propose Method*:* In this paper we propose to use MapReduce framework, That provides a parallel processing model and associated implementation to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named Hadoop. Recent studies show that the use of a multiple-layer architecture is an option for dealing with big data. The Distributed Parallel architecture distributes data across multiple processing units, and parallel processing units provide data much faster, by improving processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

We are living during a time when a touchy measure of information is being produced each day. Information from sensors, cell phones, long range informal communication sites, experimental information and endeavors – all are adding to this tremendous blast in information. This sudden barrage can be gotten a handle on by the actuality that we have made an unfathomable volume of information in the last two a long time. Enormous Data-as these substantial pieces of information is by and large called-has ended up one of the most sizzling examination slants today. Research recommends that tapping the capability of this information can advantage organizations, investigative controls and people in general segment – adding to their monetary increases and also improvement in each circle. The need is to create productive frameworks that can misuse this possibility to the most extreme, remembering the present difficulties connected with its examination, structure, scale, auspiciousness and protection. There has been a movement in the engineering of data processing frameworks today, from the brought together engineering to the circulated engineering. Endeavors confront the test of handling these colossal lumps of information, and have found that none of the current brought together models can effectively handle this colossal volume of information. These are subsequently using circulated models to bridle

DOI: 10.18535/ijecs/v5i9.65

this information. A few answers for the Big Information issue have risen which incorporates the Map Decrease environment championed by Google which is presently accessible open-source in Hadoop. Hadoop's circulated handling, Map Reduce calculations and general engineering are a noteworthy step towards accomplishing the guaranteed advantages of Enormous Data. Map Reduce and Hadoop are the most generally utilized models utilized today for Big Data handling. Hadoop is an open source huge scale information handling structure that underpins conveyed handling of substantial lumps of information utilizing straightforward programming models. The Apache Hadoop venture comprises of the HDFS and Hadoop Map Reduce in expansion to different modules. The product is demonstrated to harvest upon the handling force of grouped registering while overseeing disappointments at hub level. The Map Reduce programming structure which was initially presented by Google in 2004 is a programming model, which now received by Apache Hadoop, comprises of part the vast.

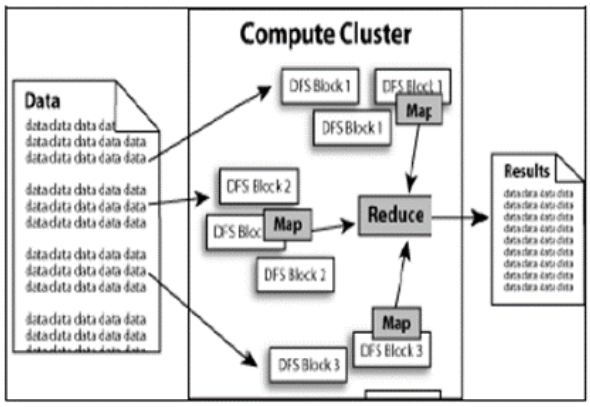

Fig. 1 Map Reduce in Hadoop

We propose a guide decrease procedure in Hadoop for information escalated use. MapReduce programming comprises of composing two capacities, a guide capacity, and a diminish capacity. The guide capacity takes a key, esteem match and yields a rundown of middle qualities with the key. The guide capacity is written in a manner that various guide capacities can be executed immediately, so it's the part of the system that partitions up assignments.

The decrease work then takes the yield of the guide works, and does some procedure on them, more often than not joining qualities, to create the craved result in a yield document. Figure underneath demonstrates a photo speaking to the execution of a MapReduce work .

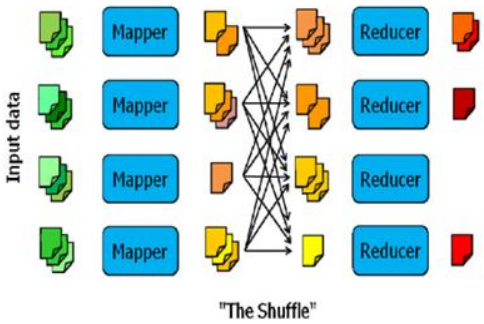

Fig. 2: MapReduce Job

At the point when a MapReduce system is controlled by Hadoop, the occupation is sent to an expert hub, the jobtracker, which has numerous "slave" hubs, or tasktrackers that answer to it and request new work at whatever point they are unmoving. Utilizing this procedure, the jobtracker partitions the guide assignments (and all the time the diminish assignments too) amongst the asktrackers, so that they all work in parallel. Additionally, the jobtracker monitors which tasktrackers fall flat, so their errands are redistributed to other undertaking trackers, just bringing about a slight increment in execution time. Moreover, if there should be an occurrence of slower specialists backing off the entire group, any errands as yet running once there are no all the more new errands left are given to machines that have completed their errands as of now. Not each procedure hubs have a little bit of a bigger document, so that when a record is gotten to the data transmission of countless circles can be used in parallel.

DOI: 10.18535/ijecs/v5i9.65

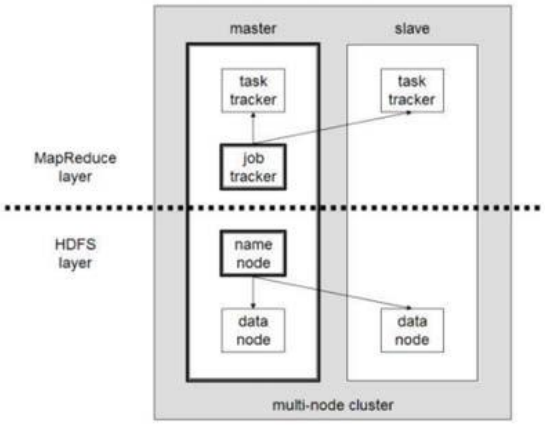

Fig. 3: HDFS cluster setup.

2. PRELIMNARY STUDY:

The first and foremost strategy for development of a project is to reduce the problem specifications in big data by using the map reduse framework and parallel processing modelin hadoop.In this paper we are using Retail Domain dataset to show the improvement in the problem specifications.

We use Retail Domain Dataset of morethan 3000 product Details of online market and that should be stores and manages simply by using the hadoop with more security and reliability.

4. IMPLEMENTATION

We use the operating system Ubuntu to execute the application of MapReduce and to java programming for attaching the Big Data Dataset of Retail Domain Of 3000 Product Details.We provide this Java Programming by using Jar files to access the Dataset . In hadoop we use the Jar file to take the input Dataset to Execute and to Store and to manage the dataset we use MySQL database and By using SQL commands we can perform any operation on Dataset.

5. CONCLUSION:

Hadoop with its efficient Data mining technique & programming framework based on concept of mapped reduction, is a powerful tool to manage large data sets. With its map-reduce programming

3 .SYSTEM DESIGN:

In this paper we use MapReduce framework, That provides a parallel processing model and associated implementation to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step. The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named Hadoop. Recent studies show that the use of a multiple-layer architecture is an option for dealing with big data. The Distributed Parallel architecture distributes data across multiple processing units, and parallel processing units provide data much faster, by improving processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

paradigms, overall architecture, ecosystem, fault-tolerance techniques and distributed processing, Hadoop offers a complete infrastructure to handle Big Data. Users must leverage the benefits of Big-Data by adopting Hadoop infrastructure for data processing.

However, the issues such as lack of flexible resource management, application deployment support, and multiple data source support pose a challenge to Hadoop’s adoption. In this paper we can make use of these Hadoop and MapReduce Framework to the efficient big data storage of Entire Online Retail Domain.

8. REFRENCES

1.R.A.Fadnavis et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 443-445

2. Afrati, F.N. & Ullman, J.D. (2011) Optimizing Multiway Joins in a Map-Reduce Environment. *IEEETransactions on Knowledge and Data Engineering*, 23(9), 1282-1298.

3.Bakshi, K. (2012) Considerations for Big Data: Architecture and Approach. *IEEE Aerospace Conference*,(pp.1-7). Big Sky, USA.

4. Gu, R., Yang, X., Yan, J., Sun, Y., Wang, B., Yuan, C. & Huang, Y. (2014) SHadoop: Improving
MapReduce Performance by Optimizing Job Execution Mechanism in Hadoop Clusters. *Journal ofParallel and Distributed Computing*, 74(3), 2166-2179.

5. Jiang, D., Tung, A. & Chen, G. (2011) MAP-JOIN-REDUCE: Toward Scalable and Efficient Data Analysison Large Clusters. *IEEE Transactions on Knowledge and Data Engineering*, 23(9), 1299-1311.

6. Kouloumpis, E., Wilson, T. & Moore, J. (2011) Twitter Sentiment Analysis: The Good the Bad and theOMG!. *Fifth International AAAI Conference on Weblogs and Social Media*, The AAAI Press, (pp.538-541). Barcelona, Spain.

7. Kraska, T. (2013) Finding the Needle in the Big Data Systems Haystack. *IEEE Internet Computing*, 17(1),84-86.

8. Lee, D., Kim J-S. & Maeng, S. (2013) A Large-scale incremental processing with MapReduce. *FutureGeneration Computer System, 36*, pp 66-79.

9. Linden, G., Smith, B. & York, J. (2003) Amazon.com recommendations: Item-to-Item CollaborativeFiltering. *IEEE Internet Computing*, 7(1), 76-80.

10. Sehrish, S., Mackey, G., Shang, P., Wang, J. & Bent, J. (2013) Supporting HPC Analytics Applications withAccess Patterns Using Data Restructuring and Data-Centric Scheduling Techniques in Map reduce. *IEEETransactions on Parallel and Distributed Systems*, 24(1), 158-168.

11.Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A ReviewPaper on Big Data and Hadoop" in International Journal ofScientific and Research Publications, Volume 4, Issue 10, October2014.

12. SMITHA T, V. Suresh Kumar "Application of Big Data in DataMining" in International Journal of Emerging Technology andAdvanced Engineering Volume 3, Issue 7, July 2013).

13.IBM Big Data analytics HUB,www.ibmbigdatahub.com/infographic/four-vs-big-data

14. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N "Analysis of Bidgata using Apache Hadoop and Map Reduce" inInternational Journal of Advance Research in Computer Science andSoftware Engineering, Volume 4, Issue 5, May 2014.

15. Apache Hadoop Project, http://hadoop.apache.org/, 2013.

16. Smitha.T, Dr.V.Sundaram, "Classification Rules by Decision Treefor disease prediction" International journal for computerApplication, (IJCA) vol 43, 8, No-8, April 2012 edition. ISSN0975-8887; pp- 35-37

17. Vidyasagar S. D, A Study on "Role of Hadoop in InformationTechnology era", GRA - GLOBAL RESEARCH ANALYSIS,Volume : 2 | Issue : 2 | Feb 2013 • ISSN No 2277 – 8160.

18.BIG DATA: Challenges and opportunities, Infosys Lab Briefings,Vol 11 No 1, 2013.

19. Divyakant Agrawal, Challenges and Opportunities with Big Data, Acommunity white paper developed by leading researchers across theUnited States.

20. Puneet Singh Duggal, Sanchita Paul, Big Data Analysis: Challengesand Solutions in International Conference on Cloud, Big Data andTrust 2013, Nov 13-15, RGPV.

21. Big Data, Wikipedia, http://en.wikipedia.org/wiki/Big_dataWebster,
Phil. "Supercomputing the Climate: NASA's Big Data Mission".*CSC World*. Computer Sciences Corporation. Retrieved 2013-01-18.