Research Article

A Survey on Estimation of Time on Hadoop Cluster for Data Computation

Mohan Kumar M K,[1*] Akram Pasha[1]

Abstract: Data is generated from several ways such Social media, Internet, IT management, Business, Scientificapplications etc.is in the form of petabytes of size and it is in unstructured format to manage this unstructured data is difficult task in a given interval of time. Hadoop is used to answer Big Data, storage of data is performed by distributed file systemusing Hadoop (HDFS) and information retrieval is done by using Map Reduce concept. Hadoop cluster is considered by taking a single node at all time and analyzed three kinds of time that is user, real and system time. Cloud provides Management and examining several types of Data using Big Data-as-a-Service.

Asian Journal of Engineering and Technology Innovation
Volume 4, Issue 7
Published on: 7/05/2016

Cite this article as: Mohan Kumar M K, Akram Pasha. A Survey on Estimation of Time on Hadoop Cluster for Data Computation. Asian Journal of Engineering and Technology Innovation, Vol 4(7): 6-9, 2016.

INTRODUCTION

Huge amount of data is producedthrough numerousresources such as commercial process, trades, publicinteracting sites, network servers etc. is structured also unstructured format, managing of unstructured data is difficult task. [1] Large data is processed using slinkedforms, web request records, etc. the available data is hugethe communications are spread across thousands of systems in order to complete work in givenexpanse of time. Huge data generated as service trace records and QOSdata and service association are making use of resources.To enhance system service generated by Big Data is in the form of high Volume, high Variety, high Velocity and Veracity. [2]

To manage Big Data Hadoop can be used as tool. Data stored in the form of distributed manner, the distributed storage of storing data is provided by apache Hadoop. Google's Map Reduce concept provides finding information from stored data [3].Map and Reduce function helps in parallelizinghuge computations certainlythen to use re-execution as tool for error tolerance.The main contribution performed by Map and Reduce function are simple and dominant interface it allowsspontaneousparallelization and sharing of huge scale calculations that reaches more performance on huge cluster of personal computers. Distributed file system using Hadoop (HDFS) is used for storing of informationthen Map Reduce function are used to retrieval information from stored data.

Data-as-a-service using Big Data expresses clearly then briefly storage,analytics techniques to service and management gives Big Data services using programmable application programmer interface's, enhances efficiency, cost reduction and enables integration seamlessly. [4]

The management tool for big data uses several projects such as pig, hive, etc. pig uses procedural dataflow language and partition of data is not able to perform at that time sample processing from subsets of data by stock symbols or particular date or month. Hive is declarative SQLish language provides familiar programming model for people who use SQL, it also eliminates tricky and complex coding would have done in Map Reduce.

Service created Big Data is having four different measurements (1) veracity: trustable analysis is conducted for discovering and modifying noisy and varying data; (2)variety: data that is obtaining from new resources here obtained data is both formats structured data and unstructured data are created in many data types, creating it likely to explore nov elvisions after examining these records together; (3) velocity: the data is increasing in very high rate based on daily bases, it is a time penetrating process such as traffic jamrecognition and QOS expectation, can be accomplished as data stream obsessed by system; (4) volume: large volume of data is stored with large scale system, with buildup teraand petabytes of data.

Several types of study on Big Data Hadoop will be done. The Tera sort provides processing of Big Data storing on Hadoop and tera gen provides the performance benchmark for storage of data on Hadoop.

RELATED WORK

This section of the paper describes the current approaches and technologies in data processing on Hadoop cluster.

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: mohankumar20cse@gmail.com
*Corresponding author

As the size of Big Data is constantly reaching target reachingas of some terabytes to several petabytes in on its own datasets [5]. The task becomes difficult to capture, visualize, storage, share, analytics and searching. Several technologies are used in big data includes databases, massively parallel processing(MPP), distributed file system, cloud computing platforms, scalable storage system and internet. [6] Various technologies are used to deploy, analyze and imagine Big Data [5]. Apache Hadoop situatedas openbasissoftware reference library allows in distributed processing of huge datasets using cluster of nodes using programming model and provides distributed storage of data. It helps in work using thousands of self-determining computers also petabytes of data [7].

HDFS provides fault tolerance and offers high dataright to use and application that as huge data sets. HDFS provides huge storage of data in huge number of servers and running Map/Reduce works across the computers and running work at data. The HDFS uses Master and Slave concepts for splitting hugerecords into large piece and achieved by several computers in the Hadoop cluster. User processes map job it specifies key pair or value pair and it generates a set of intermediate values associated with intermediate keys. [8] Map function produces word and associated number of occurrences, similarly reduce job adds together all sumsproducedfor specific word [3]. Map operation can be performed using master node uses the input divide that input into minor sub tasks and send minor sub tasks to worker nodes, worker node solves smaller problem and send answer to the master node. The Reduce operation can be performed using master node it gathers solutions of the sub problems and accumulates to form results. HDFS exposes file system name space and allow data to store in files, the files can be divided into one or more chunks and these chunks are deposited in the set of data nodes. HDFS stores large files through several machines in the large clusters each file is havingorder of blocks.

The cluster contains one master node and several slave nodes or the worker nodes. Job tracker service performs map reduce task to specify nodes in cluster. Task tracker is the node and it is in cluster andagrees to take tasks such as reduce, shuffle, mapfrom Job Tracker. Master computercontains Job Tracker, Task Tracker, Name Node, and Data Node. Slaveor the worker computer acts as both data node and Task Tracker [6]. By using Data Node, Task Tracker will fetch data thenexecute the job whetherthe retrieval of data or retrieval of information. Job Tracker performs like planner of Task Tracker and communication is endlessamongst Task Tracker and the Job Tracker. The Task Tracker can separate into slots and for each of the task it assigns duty to slots, number of the slots isstaticJob Tracker chooses Task Tracker that has a freely existing slots. Rack awareness can be perform and useful to select Task Tracker on the same rack where the data can be stored, with the inter rack bandwidth can be saved.

Several log files are created in case of Email service provided by Flipkart (biggest ecommerce company in the India)would produce 25-40 gigabytes(around the 100-150 millions of lines)logs per each hour [9]. Pinpoint [10] uses clustering algorithm for grouping failure and success logs. Dapper [11] is used to manage large volumes of trace logs and employs big table. Data is analyzed for various purpose enhancing system performance, assessing risks, trimming cost, decision making, lifting sales, etc. on [12]. One infrastructure offers same functionality of the Big Data management, then to handling severalkinds Big Data and the Big Data analysisjob [13]. Data-as-a-service gives big data associated services to theoperators to improveeffectiveness and reductions of cost.

Big Data infrastructure as a service(IaaS)in the cloud includes Storage-as-a-Service and also Computing-as-a-Service to collect and processes huge documents. Big Data infrastructure provides fast access of data and procedure to fulfill users just in needed time [14]. Different forms of outdatedIaaS in the cloud for that big data have to combine with storage designs [15]. Big Data (PaaS)allows users to access, build analytic application, and analyze large data sets [16]. Different ways of data storage and management inBig Data Platform as aService includes Data as a Service (DaaS), cloud storage, and the Data base as a service (DBaaS). Big Data Software as a Service gives business intelligence (BI) it runs unstructured data into enhanced assets [17], Big Data Software as a Serviceis thenetwork hosted, no-SQL,Hadoop and multi-tenant, and machine learning technologies and range of pattern discovery[18].

Storage mining in IT managementuses seven node Hadoop cluster (one is master, six are slaves). In these criteria for storage of data we use Cassandra [19] on similar set of hosts in the Hadoop cluster to complete the neighborhood of data and compares with other of the NoSQL changes. The Cassandra is completelydispersed and it has no dominant point of let down i.e. advantageous accessibility and dividingopen-mindedness in the CAP Theorem [20]. For measurable device learning need to install RHadoop type of framework [21] onhighestposition of Cassandra cluster and the Hadoop. At first needto install ‘R’ thenassociated packages on the all machines in the cluster, then implement extrapolative analytics in the RHadoop framework. Single thread R script will be running on the single machine, to train replicas for all the comprehensively, and over all time of training procedure completes in around the time of 80 minutes. Instruction to attain the copies for all sizes. By using different tool same algorithm in the RHadoop stagethe training procedure finishes in fewer than 10 minutes. So R Hadoop helps in consuming time for the data processing and gives better results.

Table1: R Hadoop Helps in Consuming Time for the Data Processing and Gives Better Results

Methods Used for Processing Data	Output Obtained	Tasks Performed to Get Results
Study of Processing data on Hadoop cluster by time based manner	Performance evaluation performed using Hadoop cluster with rise in totalnumber of nodes decreases processing time. Processed data shown using graphs.	Uses Apache Hadoop for storing data andistributed storage. HDFS is used for storage and for retrieval uses Map Reduce. Comparison done with different types of time, nodes, size of data.
Big Data Problem solving using Hadoop and Reduction of Map	Shows how map and reduce tasks takes place in processing of data in Hadoop cluster.	Uses Hadoop to process cluster of nodes, performs Map and Reduce functions using key/value pairs. Comparison is done with nodes, size of data and time for execution.
Map Reduce: Simplified Data Processing on Large Clusters	This model is easy to use with parallel and distributed systems, large problems are easily solved with Map Reduce concept.	Rate of transfer of data is calculated using time and input data, sorting algorithm uses Tera sort perform computations using graphs.
Service produced Big Data and Big Data as a Service: An Overview	Three kinds of the services producedby big data used for enhancing QOS oriented scheme.	Clustering algorithm is used for failed and succeeded logs Efficiency achieved reduces cost, big data infrastructure provides fast access and processes in time, big data platform provides accesses build analytics and analyze large datasets, big data software provides business intelligence (BI).
Big Data answers for Storage Mining	Hadoop Cassandra is installed and RHadoop framework is used.	Cassandra helps in storing the data and it is fully distributed. RHadoop script measures presentation of data and all storingcapacities in data focal point.

Methods used for processing data Output obtained Tasks performed to get results Study of Processing data on Hadoop cluster by time based manner Performance evaluation performed using Hadoop cluster with rise in totalnumber of nodes decreases processing time. Processed data shown using graphs. Uses Apache Hadoop for storing data andistributed storage. HDFS is used for storage and for retrieval uses Map Reduce.Comparison done with different types of time, nodes, size of data. Big Data Problem solving using Hadoop and Reduction of Map Shows how map and reduce tasks takes place in processing of data in Hadoop cluster. Uses Hadoop to process cluster of nodes, performs Map and Reduce functions using key/value pairs. Comparison is done with nodes, size of data and time for execution. Map Reduce: Simplified Data Processing on Large Clusters This model is easy to use with parallel and distributed systems, large problems are easily solved with Map Reduce concept. Rate of transfer of data is calculated using time and input data, sorting algorithm uses Tera sort perform computations using graphs.

Service produced Big Data and Big Data as a Service: An Overview Three kinds of the services producedby big data used for enhancing QOS oriented scheme. Clustering algorithm is used for failed and succeeded logs Efficiency achieved reduces cost, big data infrastructure provides fast access and processes in time, big data platform provides accesses build analytics and analyze large datasets, big data software provides business intelligence (BI). Big Data answers for Storage Mining Hadoop Cassandra is installed and RHadoop framework is used. Cassandra helps in storing the data and it is fully distributed. RHadoop script measures presentation of data and all storingcapacities in data focal point.

CONCLUSION

The work is done helps in showing time based behavior of processing data on several increasing number of nodes. It helps in enhancing Hadoop cluster for data storing and processing, the delay is less for processing for nearest node data and it is useful. To enhance service trace logs, quality of service and service relationship uses three different types of Big Data services. Map Reduce programming idea is easier to use and implementation is done using not having knowledge with similar or parallel and spread or distributed systems. The large datasets are easily processed using Map Reduce concept and it measures large cluster of machines comprising thousands of machines. Storage mining helps in storing distributed data using Cassandra and RHadoop helps in measuring the huge voluminous data.

REFERENCES AND NOTES

1. Impetus white paper, March, 2011, “Planning Hadoop/NoSQL Projects for 2011” by Technologies, Available:http://www.techrepublic.com/whitepapers/Planninghadoopnosql-projects-for-2011/2923717, March,2011.
2. Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, David E. Culler, Joseph M. Hellerstein, and David A. Patterson. High-performance sorting on networks of workstations. In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, May 1997.
3. Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, Google, Inc.
4. ZibinZheng, Jieming Zhu, and Michael R. Lyu , “Service-generated Big Data and Big Data-as-a-Service: An Overview” , 978-0-7695-5006-0/13 2013 IEEE.
5. McKinsey Global Institute, 2011, Big Data: The next frontier for innovation, competition, and productivity, Available:www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx, Aug, 2012.
6. Aditya B. Patel, Manashvi Birla, Ushma Nair, “Addressing Big Data Problem Using Hadoop and Map Reduce”, NUiCONE-2012, 06- 08DECEMBER, 2012.
7. Apache Software Foundation. Official apache hadoop website, http://hadoop.apache.org/, Aug, 2012.
8. Hung-Chih Yang, Ali Dasdan, Rusey-Lung Hsiao, and D.StottParker from Yahoo and UCLA, "Map-Reduce-Merge: Simplified Data Processing on Large Clusters", paper published in Proc. of Acm Sigmod, pp. 1029–1040, 2007.
9. H. Mi, H.Wang, Y. Zhou, M. R. Lyu, and H. Cai, “Towards fine-grained, unsupervised, scalable performance diagnosis forproduction cloud computing systems,”IEEE Transaction on Parallel and DistributedSystems, no.PrePrints, 2013.
10. M. Chen, E. Kiciman, E. Fratkin, A. Fox, and E. Brewer, “Pinpoint: problem determination in large, dynamic internet services”, in Proceedingof the International Conference on Dependable Systems andNetworks (DSN’02), pp. 595–604.
11. B. H. Sigelman, L. A. Barroso, M. Burrows, P. Stephenson, M. Plakal, D. Beaver, S. Jaspan, and C. Shanbhag, “Dapper, a large-scaledistributed systems tracing infrastructure,” Google, Inc., Tech. Rep., 2010.
12. S. Lohr, “The age of big data,” New York Times, vol. 11, 2012.
13. “Challenges and opportunities with big data,” leading Researchers across the United States, Tech. Rep., 2011.
14. E. Slack, “Storage infrastructures for big data workflows,” Storage Switchland, LLC, Tech. Rep., 2012.
15. “Big data-as-a-service: A market and technology perspective,” EMC Solution Group, Tech. Rep., 2012.
16. J. Horey, E. Begoli, R. Gunasekaran, S.-H. Lim, and J. Nutaro, “Big data platforms as a service: challenges and approach,” in Proceedings of the 4th USENIX conference on Hot Topics in Cloud Ccomputing, ser. HotCloud’12, 2012, pp. 16–16.
17. “Why big data analytics as a service?” ”http://www.analyticsasaservice.org/why-big-data-analytics-as- aservice/”, August 2012.
18. P. O’Brien, “The future: Big data apps or web services?” ”http://blog.fliptop.com/blog/2012/05/12/the-future-big-data-appsor- web-services/”, 2013.
19. Apache Cassandra, http://cassandra.apache.org.
20. N. Lynch and S. Gilbert, “Brewer’s conjecture and the feasibility of consistent, available, partition-tolerant web services,” SIGACT, 2002.
21. RevolutionAnalyltics,https://github.com/RevolutionAnalytics/RHadoop/ wiki.