International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 5 Issue: 3 22-Mar-2015,ISSN_NO: 2321-3337

A NOVEL APPROACH TO IMPROVE THE BUSINESS USING HADOOP AND MONGODB

Godase Anand Tanaji, Fugate Balaji Fulchand, Surwase Prashant Tanaji
Department of Computer Engineering
S.V.P.M College of Engineering, Malegaon (bk)
Savitribai Phule Pune University

ABSTRACT-- *Internet has become a non-detachable part of human beings throughout the world. Using the internet we can obtain different types of information as well as we can do any other lots of daily task easily such as shopping. We can get the exact detail related to the products which we are going to purchase online. But sometimes happens that due to some issues maximum of the time user does not purchase any product using our online site . So to improve the business e-commercial companies are required to keep the total detail related to the website. So using the Hadoop and MongoDB we can obtain the required details of our website within a less amount of time.*

Hadoop provides the map-reduce which is the most commonly used context in parallel processing. This paper proposes the new methodology to improve the business of e-commercial companies by keeping all the relevant information related to their website using the MongoDB and Hadoop and obtain the final aggregated result which helps to take the decisions to improve their business.

Keywords: Hadoop , MongoDB, Map-Reduce , Sharding ,Data Aggregation

1, INTRODUCTION

The capability of processing a large amount of data is a time consuming processes.

Even though several management systems are dramatically increasing the processing speed of queries significantly in order to obtain the fast response ,the queries are still taking a large amount of processing time to process the large input data.We are going to keep the log of user interaction on webpage. According to the user interaction we are going to prepare the record which contain the point of return and point of success for each webpage ,Like this there will be large amount of data is going to generate. So to deal with this large amount of data we are going to use the sharding of the Mongo DB and on this sharded data we are applying the Map-Reduce to calculate aggregated result.

Sharding is nothing but the splitting of the data uniformly across the cluster to parallelize the access of data. The concept of sharding in MongoDB supports the growth in database.

In sharding data gets divided into different shards. Map-Reduce is a framework which reduces the data size by giving it <key, value>pair and helps to obtain the aggregate result.

So our log file which we have prepared from the user interaction on our online shopping site is gets distributed using the sharding of Mongo DB and on this sharded data we are applying the map-reduce for calculating from which page maximum no.of users are returned and from which page maximum users are goes next page. So calculating such result we helps to commercial companies to concentrate on page from which maximum no.of users are returned and provide some more facilities on that webpage and improve their business.

2, LITERATURE SURVEY

Many of the e-commercial companies like filpkart, snap-deal, amazon etc have included the facility which gives the quality of the product like rating ,review but it is useful for the customers, using this commercial companies cannot understood the maximum point of return on website or weakness of any page.
It may be possibility that given ratings on the webpage for that product are wrong or reviews are also wrong. So using only the review and the ratings it also become a difficult task for customer also for commercial companies to identity the weakness of their website.

3, PROPOSED SYSTEM

Sharding of MongoDB:

In this paper the log file generated from the user interaction on webpage is gets distributed on different clusters. The data is divided into different shards. The default size of the shard is 64 MB but we can also change it according to our database size.
MongoDB supports the sharding through the configuration of sharded cluster.
There has the following some component:
***Shard*:**
It stores the actual data.
Query router (mongos):
It is the direct interface to the client application and the desired shard.Query router processes and targets the operation to shard and returns the final result.Query router uses the metadata to target the operation to shard.
Config server:
It stores the cluster's metadata. The query router uses this meta data to target operation to shard.

Data partitioning:

MongoDB distributes the data according to the shard key.
Shard key**:**To do the sharding it is used.It may be indexed field that exists in every collection.
MongoDB divide the shard key value into chunk using either range based sharding or hash based sharding.

A)Range based sharding: In this mongodb divides the data set in ranges determined by shard key value.

B)Hash based sharding: In this mongodb calculates the hash of field's value and then uses this hashes to distribute data.

Here the drawback of range based sharding is ,it may be possible to unequal distribution of the data sets and finally it effect on response time.So it is better to use hash based sharding when large database.

Flow of working:

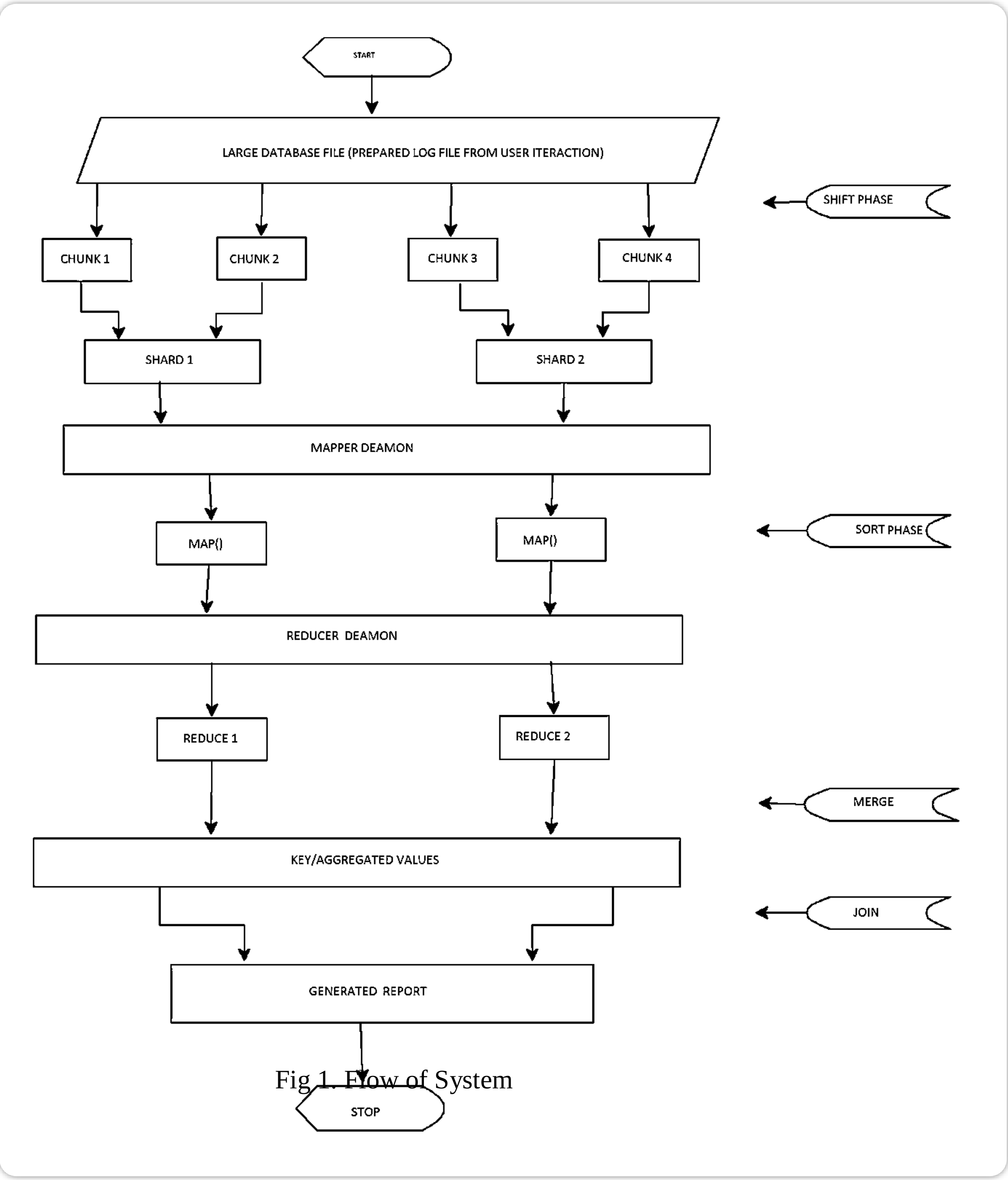

Fig 1. Flow of System

1. Log File Generation:

The log file is generated by setting the flag POR or POS according to the user interaction. This record is stored in Mongo DB. For each page we are assigning a unique id and this id is used while performing the sharding and map-reduce.

2. Sharding and Map-Reduce:

The hash based sharding is used because the generated log file contain very large size of data. So page id is used in hash to distribute the database. On this distributed data we are applying the mapper daemon and reducer daemon which will map and reduce the data by assigning the <key, value>pair.

3. Report generation:

Data aggregation is performed on reduced data which is the final report and the graphical representation shows the maximum point of return page and alsomaximumpoint of success page.So using this result the e-comercial companies can take the decisions so that business will be improved.

Sharded Parallel Map-Reduce for online aggregation Algorithm:-

1.Consider the very large database stored in MongoDB which is in terabyte or in petabyte for processing.
2. This whole dataset is divided into small chunks.(we have to specify the size of chunk while installation otherwise default size is set.)
3. These chunks are then collected to form the shards. Each shard ,may contain equal no.of chunks or unequal.
4.This total shards are given to the map-reduce for parallel execution at a time.
a)Map function is applied on this data and for similar value data same key is assigned to it like this mapping is done by giving <key, value>pair.
b)After mapping the reduce function is applied on it which gives new key for same key data and combine the result at one place.
5.Above step a) and b) is repeatedly performed to apply the map-reduce on whole dataset.
6.Final resultant reduced dataset is given to the aggregator to estimate final result which gives the report for maximum point of return and maximum point of success of webpages.

In our system firstly user login to our website and according to their interaction we are going to generate the log file which content the user_id, page_id ,point of return flag and point of success flag. If user goes next page then the point of success flag of the previous page is set otherwise the point of return flag of that page is set.

Likewise we are going to generate the log file in Mongo DB .This log is then distributed using the sharding into different shards. The Hadoop provides the map-reduce which map the data by giving the <key, value>pair and then reduces the mapped data by giving another <key, value>pair. This reduced data is finally used to calculate the aggregated result.

This calculated result gives us the graphical representation of maximum point of return pages and maximum point of success pages. Using this report the e-commercial companies can increase their business as well as improves their website so that business gets increased.

4, IMPLEMENTATION

4.1 Steps

- Generates a log file from user interactions on commercial web site.
- Apply sharding on this log file.
- Map and Reduce the sharded data using mapper and reducer daemon.
- Aggregation is done on data by aggregator.
- Analyse and generate final report.

5, CONCLUSION AND FUTURE WORK

Here we have introduced a novel approach which will helps to take important decisions and improve their business by using the mongodb and hadoop.Using this report companies can concentrate on the page which gives the maximum point of return and improve that page by providing more facility so that user will go next and purchase the product.
This approach provide a new logic to keep track of our own website and also business.

REFERENCES

[1] J. Dean and S.Ghemawat,Map-Reduce:Simplified data processing on large clusters, In processing of OSDI.pp. 137-150 2004.

[2] N.Pansare,V.R.Borkar,online aggregation for large Map-Reduce jobs,VLDB 2011 Conference proceedings.pp.1135-1145 AUGUST 2011.

[3] T.Condie,N.Conway,P.Alvaro,and J.M.Hellerstin,online aggregation and continuous query support in Map-Reduce, In SIGMOD 2010, Conference proceedings.pp.1115-1118,June 2010

[4] B.Rama Mohon Rao, “Sharded parallel Map reduce for online aggregation”

[5] R. J. Bayardo and D. P. Miranker, —Processing queries for first-few answers‖ In Proc. 5th International Conf. on Information and Knowledge Management, pages 45.52,1996.

[6] G. Antoshenkov and M. Ziauddin, —Query processing and optimization in Oracle Rdb‖, VLDB Journal, 5(4): 229-237, 1996.

[7] J. M. Hellerstein, —The case for online aggregation‖, Technical Report UCB//CSD-96-908, EECS Computer Science Division, University of California, Berkeley, CA,1996.

[8] J. M. Hellerstein, P. J. Haas and H. J. Wang, —Online aggregation‖, In Proc. 1997 ACM SIGMOD Intl. Conf. Managment of Data, pages 171–182. ACM Press, 1997.

[9] 10gen, Inc: MongoDB, 2010, http://www.mongodb.org.

[10] Strozzi and Carlo, —NoSQL – A relational database management system ‖, 2007–2010.

Author's Profile:-

1)Godase Anand T. Student at S.V.P.M College of Engineering ,
Malegaon(bk)-Baramati
Contact no:9960146460
Email:godase.anand@gmail.com

2)Fugate Balaji F. Student at S.V.P.M College of Engineering ,
Malegaon(bk)-Baramati
Contact no:9922510788
Email:balaji.fugate@gmail.com

3)Surwase Prashant T Student at S.V.P.M College of Engineering ,
Malegaon(bk)-Baramati
Contact no:9665299692
Email:surwase.prashant111@gmail.com