International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 10 5742 - 5744

Increasing Efficiency of Recommendation System using Big Data Analysis

Chandan Venkatesh
Department of Information Science and Engineering
MS Ramaiah Institute of Technology
Bengaluru,Karnataka,India
venkatesh.chandan@gmail.com

Deekshith Kumar
Department of Information Science and Engineering
MS Ramaiah Institute of Technology
Bengaluru,Karnataka,India
deekshith24@gmail.com

Ganesh Madhav R
Department of Information Science and Engineering
MS Ramaiah Institute of Technology
Bengaluru,Karnataka,India
gmganeshmadhav@gmail.com

***Abstract*—** In the present Digital Space a lot of Internet users try to come up with solutions to a particular problem by suggesting solutions that are pre-existing on the Internet. This brings down the originality of posts and we are able to overcome this problem by applying prediction models on data sets. It is important for a user to come up with original ideas to gain up votes which in turn represent the quality of a post. Due to the huge influx of data at every moment, the need for big data analytics becomes essential and hence the use of an open source framework like Hadoop is imperative so as to increase effectiveness of recommender system built on these prediction models.

Keywords- *Hadoop, Hive, Mahout, Bulk Loading, Recommendation.*

*****

I. INTRODUCTION

Enormous data is generated on a daily basis. The collection and maintenance of this data place a paramount role in today's Information technology analytics. New algorithms are written frequently to tender to the volumes of data. We focus on the vast Q & A user community with a colossal attributes such as up-votes, down-votes and reputations that form the premise of a recommendation system . The present systems make use of a much more constrained environment that implement MySQL framework. With our setup, we take advantage of distributed computing, effectively removing constraints set up by MySQL and its other counterparts. To achieve the desired results, we implement frameworks like Hive and Mahout on top of Hadoop. Here, Hive acts as a query processing and managing software, Mahout provides us the important prediction models and Hadoop forms the base for these two frameworks.

We advocate different algorithms from the Mahout framework like Co-occurrence and Log-Likelihood to analyze data sets etc. These models provide us different measures to compare user similarities and dissimilarities. Such a model becomes the premise of recommendation. Recommendation also defined as process wherein " People provide recommendations as input, which the system then aggregates and directs to appropriate recipients as a result " (Resnick and Varian, 1997). Recommender uses known data to yield relevant data treated as information. Some of widely used recommenders are based on content, collaboration and knowledge.

We are focused on building a recommender that has higher efficiency and lower turnaround time.

II. MODELLING

The intended purpose of the Big data is to make data analyses that can be evaluated on the basis of several criterions which are mentioned underneath.

• Efficiency : With a myriad of data it is foremost that we need an efficient technique to analyze,interpret and extract useful information from it..

• Consistency : Data is ceaseless, a consistent model is key to keep this knowledge coherent and flawless.

• Robustness : Robustness determines how well the system is able to cope with divergence to set norms while working with the huge amount of data.

• Falsifiability: Notification or an alert is required when data is not good to use(i.e.when assumptions are violated) or when it is corrupt.

III. IMPLEMENTATION

- Hadoop: The core of the model is based on the Hadoop Framework that allows us to use the advantages of distributed computing. Hadoop Distributed File System and MapReduce framework help in improving scalability, flexibility and efficiency of the setup.

- Hive: Hive is a data warehouse infrastructure built on top of Hadoop for querying,providing data summarization and analysis of the data. In this model, we make use of the bulk loading feature to transfer huge amounts of data onto the HDFS using Hive and we use Hive QL to implement our queries.

- Mahout:Mahout Framework is responsible for delivering us the prediction models. The Item based recommender and User based Recommender systems are derived from this framework.

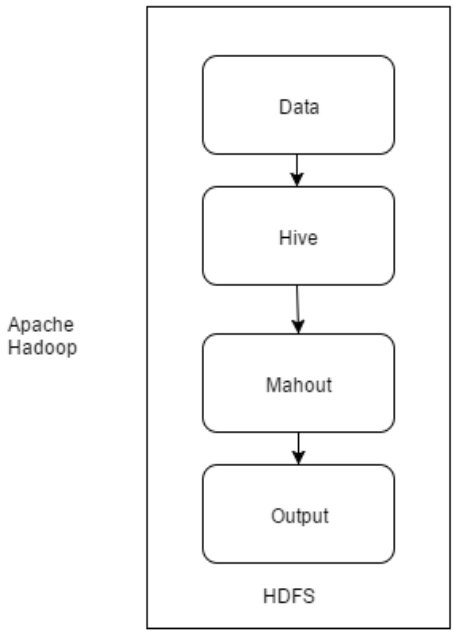

Figure-1 Data flow diagram

IV. ANALYSIS

Hive and Its advantage over SQL

The size of data sets being collected and analyzed in the industry for business intelligence is growing rapidly using traditional warehousing solutions for recommendation is prohibitively expensive. Hadoop is a popular open-source map-reduce platform which is widely used as an substitute to store and process enormous amount of data sets on commodity hardware. However, the Map-Reduce programming Framework is very low level and requires developers to write custom programs which are hard to maintain and reuse. The Hive framework effectively improves the efficiency of the Map-Reduce paradigm. Hadoop provides massive storage and processing abilities to process huge amount of data that usually are unmanageable by SQL. HiveQL provides a means of querying such huge chunk of data, and therefore becomes a natural choice for many organizations.

The most important advantage of Hive over the traditional MySQL is the fact that developer productivity is improved most effectively when complex data format are analysed. In the case of SQL over Hadoop, performance is affected due to various formats. File formats ranging from RCFiles, Parquet and ORCFiles have been developed along with the more predominant textfile and seqfile. these format need to be analysed and the most appropriate format needs to be considered for specific data.

In the more traditional MySQL data files stored in HDFS are fully scanned in most cases. This means, the scanning overhead adds on to the processing time and becomes a major deterrent for performance factors. Operations such as ORDER BY and GROUP BY usually executed in multiple process steps, performance is again affected adversely with each additional step of data transfer.

The effectiveness of Hive, however, is over large datasets. Large duration queries on Hive hold a strapping advantage over MySQL. Therefore, our recommendation model for Q & A user group uses this distinction to advantage.

Hadoop acts as a base for other frameworks to act upon.

On the whole, the system layout consists of Hadoop, Hive and Mahout to process the

user input. This model implements the Hive framework to use its 'BULK LOADING' feature to transfer huge amounts of data present in the csv files into a tabular format for a higher query processing efficiency. Next, implementation proceeds with the user defined hive queries to process the data. Hive queries are put through a certain plan prepared by the hive framework which converts the jobs into map-reduce jobs. Map-reduce increases the efficiency of query process and gives results faster. The output from the Hive module is passed onto the Mahout framework for further processing.

Authors and Affiliations

V. DATA FORMAT

1) *Input Format*

a) *MapReduce*:The MapReduce framework operates on the <key, value> pair concept, the framework takes the input to the job as a set of <key, value> pairs and provides a set of <key, value> pairs as the output of the job. The key classes need to implement the WritableComparable interface to help the sorting by the framework.The input to map to combination to reduction to the final output together form a serial set of keys and values .

b) *Hive*: In terms of data input to HIVE , A Comma Separated Values(CSV) file stores data in a tabular format (as numbers and text) in plain-text format. Therefore, file is a sequence of characters, and therefore no data has to be interpreted as binary numbers. It consists of number of records, separated by line breaks, each record consisting of fields, divided by other character or string, most commonly a comma or tab records have an identical sequence of fields.

c) *Mahout*: Mahout will get its input from hive by running selected queries by pipelining the output from hive to mahout to run the job, Mahout will get the input in CSV Format with each record consists of fields, separated by some other characters or string, these consists predominantly of literal commas or tab records and have an identical sequence of fields.

2) Output format

The Output will be parsed from two different modules

a) *Hive*: In hive, executing the hive queries generates the output in the desired format. We are taking output in CSV format where each record consists of fields, separated by some other character or string, most commonly a literal comma or tab records have an identical sequence of fields.

b) *Mahout*: Mahout will output the resulting output in CSV Format ,while that will be parsed with help of Java Modules.

VI. CONCLUSION

The proposed model provides an interesting, alternative and efficient method in order to recommend appropriately on commodity hardware. Q & A User group is a vast and ever increasing community. From the model it is possible within the Q & A community ,e-commerce websites, Matrimonial sites recommendations of appropriate things to users based on their previous history of answering . Secondly, recommendations of users to other users in order to build a global community based on common interest. The recommendation is optimized with the aid of Hive's bulk loading feature. With the presented system huge amounts of data can be handled while delivering fast processing time. Data can be transferred easily between computers/systems which are situated away from each other. The system is extensible meaning the application is designed to allow easily the addition of new functionality with added parameters. These new additional modules should not cause any unwanted side effects. The recommender system has the overload of processing huge data. Therefore, the use of hadoop framework is appropriate. With the power of Hadoop, scalability concerns are brushed aside. As the amount of data increases, it is wise to increase the number of nodes in the cluster too.

REFERENCES

[1] Thusoo, Ashish, et al. "Hive-a petabyte scale data warehouse using hadoop."*Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. IEEE, 2010.

[2] Abadi, Daniel, et al. "Systems and methods for processing data." U.S. Patent Application 13/032,516.

[3] Stewart, Robert J., Phil W. Trinder, and Hans-Wolfgang Loidl. "Comparing high level mapreduce query languages." *Advanced Parallel Processing Technologies*. Springer Berlin Heidelberg, 2011. 58-72.

[4] Kumar, Rakesh, et al. "Comparison of SQL with HiveQL." *International Journal for Research in Technological Studies* 1.9 (2014): 2348-1439.

[5] Zhang, Rui, et al. "Getting your big data priorities straight: a demonstration of priority-based QoS using social-network-driven stock recommendation."*Proceedings of the VLDB Endowment* 7.13 (2014): 1665-1668.

[6] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[7] Wu, Xindong, et al. "Data mining with big data." *Knowledge and Data Engineering, IEEE Transactions on* 26.1 (2014): 97-107.

[8] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001.

[9] Sarwar, Badrul, et al. "Analysis of recommendation algorithms for e-commerce." *Proceedings of the 2nd ACM conference on Electronic commerce*. ACM, 2000.

[10] Deshpande, Mukund, and George Karypis. "Item-based top-n recommendation algorithms." *ACM Transactions on Information Systems (TOIS)* 22.1 (2004): 143-177.

[11] Chowdhury, Badrul, et al. "A BigBench implementation in the hadoop ecosystem." *Advancing Big Data Benchmarks*. Springer International Publishing, 2014. 3-18.

[12] Barrachina, Arantxa Duque, and Aisling O'Driscoll. "A big data methodology for categorising technical support requests using Hadoop and Mahout." *Journal of Big Data* 1.1 (2014): 1-11.