International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 02 | Feb-2016 www.irjet.net p-ISSN: 2395-0072

SOCIAL MEDIA MARKET TRENDER WITH DACHE MANAGER USING HADOOP AND VISUALIZATION TOOLS

Raveena Rana [1], Tejas Shah [2], Jitanksha Shrimanker[3]

[1-3]BE Student, Information Technology Department

K. J. Somaiya Institute of Engineering & Information Technology, Mumbai, Maharashtra, India

---

Abstract- *The modern business and advertising strategies all involve market analysis and study of the trends going on in the world. Social media is the main advertising and analysis market in today's world. The current Analytics tools and models that are available in the market are very costly, unable to handle Big Data and less secure. The traditional Analytics systems takes a long time to come up with results, so it is not beneficial to use for Real Time Analytics. So, the proposed work resolves all these problems by combining the Apache Open Source platform which solves the issues of Real Time Analytics using HADOOP. It also provides scalability and reduced cost over analytics by using open Source Software. [1]*

Key Words: Map Reduce (MR), Hadoop, Java , Flume , Sqoop, Caching.

I. INTRODUCTION

The world is getting smaller and smaller day by day through internet and usage of social sites. Data is growing exponentially as the number of users and activity over the web is increasing rapidly. The consumer market is becoming increasingly competitive and the need to keep a tab on trends and consumers' opinions is a must. [3]

Social Media is being extensively used by people belonging to all age groups' for expressing their reviews and suggestions whenever they buy a new product or a new product is launched. Also it is used for recommendations to friends and family of a particular product or service. Social media sites like Facebook and Twitter generate huge amounts of data on a daily basis. This data is directly from the end user or consumer and as a result is of great value to the organizations and companies. [1]

This data is still scattered and not in a suitable format that may yield any results. Also the data is so huge that the buzz-word Big Data is the only one that can represent such an amount of data. Our Project is based on Hadoop framework that can
manage and handle Big Data efficiently. Main reasons for choosing Hadoop are:

1. The data generated by Social sites are too large for traditional databases and analysis tools.
2. Hadoop being a framework gives more flexibility in implementing ideas.
3. Scalability is a major advantage of using Hadoop over other platforms.

Application developers specify the computation in terms of a map and a reduce function, and the underlying MR Job scheduling system automatically parallelizes the computation across a cluster of machines. MR gains popularity for its simple programming interface and excellent performance when implementing a large spectrum of applications. Since most such applications take a large amount of input data, they are nicknamed "Big-data applications". Hadoop is an open-source implementation of the Google MR programming model. Hadoop consists of the Hadoop Common, which provides access to the file systems supported by Hadoop. MR provides a standardized framework for implementing

large-scale distributed computation, namely, the big-data applications. [5][6]

The work proposes to combine the Apache Open Source Modules and configure them to get the required result. Hadoop is flexible and scalable architecture. Google MR is a programming model and a software framework for large-scale distributed computing on large amounts of data [7][8].

. We propose Dache, a data-aware cache framework for big-data applications. In Dache, tasks submit their intermediate results to the cache manager. A task queries the cache manager before executing the actual computing work. We implement Dache by extending Hadoop. [2]

In the proposed work, we would be grabbing data from Twitter and analyzing it through visualization techniques. [4]

The project aims at finding out the trends of the market according to the social sites such as Twitter and visualizing and presenting it in a simple format such as a graph or chart that can be understood by anyone and a clear conclusion can be drawn out of it. It also aims at tackling data size issues by using Hadoop and MR frameworks that are open-source. It also tries to improve the processing speed by introducing a cache concept for distributed computing named Dache.[9]

II. DRAWBACKS OF EXISTING SYSTEM

The current Analytics tools and models that are available in the market are very costly, unable to handle Big Data and less secure. The traditional Analytics systems takes a long time to come up with results, so it is not beneficial to use for Real Time Analytics.

Google's MR and Apache's Hadoop are the defacto software systems for big-data applications. An observation of the MR framework is that the framework generates a large amount of intermediate data. Such abundant information is thrown away after the tasks finish, because MR is unable to utilize them. Also the MR operates on huge amount of data and as a result the processing takes usually a longer period of time.

A drawback of Hadoop operation is that MR, its processing part creates a lot of intermediate data during operations such as word-count on a particular file. But in the end we get only the desired output file and the rest of the intermediate data generated during the processing is thrown away. This data can be stored separately and can be used in similar future operations to speed up the process. Dache, a concept of cache will be introduced in the project so as to boost the operation speed of MR.

The current Analytics tools and models that are available in the market are very costly (SocialBro, TweetStats, Twentyfeet,) unable to handle Big Data and less secure. Hence the existing system lags during functioning. The traditional Analytics systems takes a long time to come up with results, so it is not beneficial to use for Real Time Analytics. Real time analytics require high speed and less time for performing it.

Querying Twitter data in a traditional RDBMS is inefficient. There are many Twitter API which provide streaming of twitter data. Streamed data is not preferred. Traditional Hadoop MR does not use the concept of Cache. Dache is a caching technique for Big Data. It boosts the speed of the analysis. This will help to cut short the time required for the entire process.[10]

III. PROPOSED METHODOLGY

We propose to grab data from Twitter and store in HDFS. The data will be converted into formats such as CSV and will be visualized via different tools. Dache is a data-aware cache framework for big-data applications. In Dache, tasks submit their intermediate results to the cache manager. A task queries the cache manager before executing the actual computing work. We implement Dache by extending Hadoop. Advantages of proposed System are as follows:

- Efficiency and Reliability of handling Big Data increases with Hadoop. The system can be scaled

as per the user requirement. Storing of grabbed data in HDFS will increase the speed, reliability and efficiency of data & results.

- The CSV used is better for understanding. The CSV will be exported to database for backup purposes. Graphs and charts will provide a clear picture of the latest trends.
- The operation speed will be boosted by Dache. There would be no cost of data grabbing and handling as all the tools used are open source softwares. Dache will make the MR operations faster.

The Proposed System will show us the trends in market that are most talked about on social media. In the end, we can compare various brands, topics and also the traditional Hadoop MR operations vs Hadoop MR operations using Dache.

A front end will be created wherein a user can sign up for an account. After sign up, he can login the account. There will be various options for grabbing data, analysis of the data, graphs of analysis etc. At the backend, the grabbed data will be stored in HDFS. It will be in the form of JSON. This data will be converted into CSV. Using Visualization tools, reports will be generated in the form of graphs.

Dache helps in speeding up the Big Data whenever the same data is accessed again, just like cache functions in other applications. This will increase the speed of processes and decrease the time required for accessing the same data otherwise.

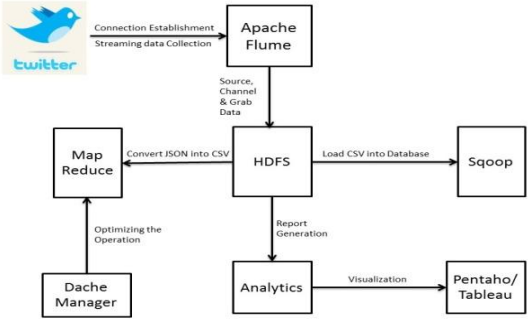


Figure 1: Proposed system architecture

Figure 1 shows the block diagram of the proposed system. The project aims to provide the end user recent Market Trends. The proposed system is very cost effective as most of the work is done on free open source softwares. The drawback of cost in existing system is overcome.

The proposed system architecture involves the following modules:

1. **Module 1: Grab Data From Social Media**

   Description: In this module, the data will be grabbed from social site on the basis of the keywords entered by the user. The grabbed data will be stored in Hadoop distributed file system (HDFS).

   Input: The user will enter the four token keys and keywords for grabbing the data. On the basis of keywords the data will be grabbed.

   Output: Output will be the grabbed data files on the basis of keywords. The format of the files will be JSON and will be stored in the HDFS or other specified location.

2. **Module 2: Process Data using Hadoop**

   Description: The Grabbed Data which is in the form of JSON will be converted into Comma Separated Value (CSV) format with MapReduce operations. In MapReduce there will be five stages viz Input Phase, Mapper Phase, Sort and Shuffle Phase, Reducer Phase and Output Phase. Different Functions will be carried out on the CSV data and visualized graphically.

   Input: Grabbed data in the form of JSON

   Output: The data in CSV form, Graphs of the analyzed data.

3. **Module 3: Display Result**

   Description: This module will be used to give the results to the user in the form of graphs and keywords entered.

   Input: Request for results.

Output: Graphs of the analyzed data.

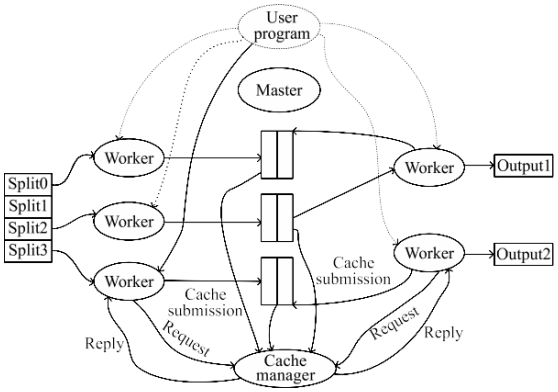
IV. IMPLEMENTATION

Initially, softwares like Apache Flume, Sqoop, Twitter, Hadoop, VM Ware, Tableau, Pentaho, and Putty have to be downloaded. Along with these, Flume settings for Twitter are also required. Setup of Flume has to be done first. All the services have to be started. A Twitter account has to be made. In the Flume settings downloaded for Twitter, desired keywords are entered. The keywords are the topics for analysis. The data will be grabbed as per the keywords entered. The data will be in millions. Hence it is known as Big Data. The analyzed data will be stored in HDFS in the form of JSON. This data will be converted to CSV with the help of Sqoop. Dache impleme1ntation is the next step. A Dache manager is created for speeding up the process. The data will be processed by the visualization tools which have already been downloaded. The output of the analysis will be in the form of graphs. Studying of graphical data is much simpler as compared to Big Data. [1]

Hadoop framework is used to facilitate operations on BigData. Hadoop allows us to install various softwares and work with them. By installing Flume and running it, we will be grabbing data from social sites and storing it directly in Hadoop Distributed File System (HDFS). HDFS is the storage part of Hadoop.

MapReduce will be used to perform operations such as Conversion of Data, Word-count, etc and Dache will be used to optimise and increase the speed of operations. MapReduce is the processing part of Hadoop. Visualization tools will be used to show the data graphically. Java will be used to write MapReduce operations.

Technique for Dache Implementation:



Input Phase | Map phase | Intermediate phase | Reducer phase | Output phase

Fig 4: High level description of architecture of Dache

We propose a novel cache description scheme. A high-level description is presented in Fig. 4.3. This scheme identifies the source input from which a cache item is obtained, and the operations applied on the input, so that a cache item produced by the workers in the map phase is indexed properly. In the reduce phase, we devise a mechanism to take into consideration the partition operations applied on the output in the map phase. We also present a method for reducers to utilize the cached results in the map phase to accelerate the execution of the MapReduce job. We implement Dache in the Hadoop project by extending the relevant components.

Algorithm for MapReduce operations (Wordcount)

- Map()
  - Input <filename, file text>
  - Parses file and emits <word, count> pairs
- Reduce()
  - Sums Values for the same keys and emits <word, TotalCount>

The mapper emits an intermediate key-value pair for each word in a document. The reducer sums up all counts for each word.

class Mapper

method Map(docid a, doc d)

for all term t ∈ doc d do

Emit(term t, count 1)

class Reducer

method Reduce(term t, counts [c1, c2, . . .])
sum ← 0
for all count c ∈ counts [c1, c2, . . .] do
sum ← sum + c 6: Emit(term t, count sum)

Figure 2 shows a screenshot of grabbed data in the form of CSV.[1]

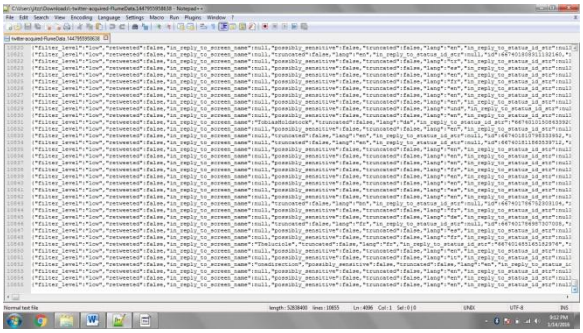


Figure 2: Screenshot of Grabbed Data

V. CONCLUSION

As compared to the existing system, our proposed system is cost effective. Requirements of this proposed system are easily available. All the softwares required are free open source softwares which cost minimal. Time saved in the entire process is cut down. Great speed is provided as compared to the existing system. We have already grabbed data for some keywords. We are working on the Dache implementation. We have stated the advantages of the existing system. Finally, we have explored a cost effective way to analyze social media market trends and generate a graphical view of the analysis.

REFERENCES

[1] Raveena Rana, Tejas Shah and Jitanksha Shrimanker "Social Media Market Trender with Dache Manager using Hadoop and Visualization Tools", "International Research Journal of Engineering and Technology", Volume: 03, Issue: 02, February 2016.

[2]Hadoop: The Definitive Guide by Tom White.

[3] Yaxiong Zhao, Jie Wu, Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" ,TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 05/10 pp39-50 Volume 19, Number 1, February 2014.

[4] Kaveh Ketabchi Khonsari, Zahra Amin Nayeri, Ali Fathalian, Leila Fathalian, "Social Network Analysis of Iran's Green Movement Opposition Groups using Twitter", 2010 International Conference on Advances in Social Networks Analysis and Mining.

[5] Karen Stepanyan, Kerstin Borau, Carsten Ullrich, "A Social Network Analysis Perspective on Student Interaction within the Twitter Microblogging Environment", 2010 10th IEEE International Conference on Advanced Learning Technologies.

[6] Guojun Liu, Ming Zhang, Fei Yan, "Large-Scale Social Network Analysis based on MapReduce",2010 International Conference on Computational Aspects of Social Networks568.

[7] Kai Shuang, Yin Yang, Bin Cai, Zhe Xiang, " X-RIME: Hadoop-based Large-Scale Social Network Analysis", Proceedings of IC-BNMT20 10.

[8] Hyeokju Lee, Joon Her, Sung-Ryul Kim, "Implementation of Large-Scalable Social Data Analysis System based on MapReduce", 2011 First ACIS/JNU International Conference on Computers, Networks, Systems, and Industrial Engineering.

[9] Ge Song, Zide Meng, Fabrice Huet, Frederick Magoules, Lei Yu, Xuelian Lin, "A Hadoop MapReduce Performance Prediction Method", 2013 IEEE International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing.

[10] Gaurav D Rajurkar, Rajeshwari M Goudar, " A speedy data uploading approach for Twitter Trend and Sentiment Analysis using Hadoop", 2015 International Conference on Computing Communication Control and Automation.