Twitter Word Frequency Count using Hadoop Components

Ms. Pooja S. Desai[1] and Ms. Isha Agarwal[1]

Department of Computer Engineering and IT, C. G. Patel Institute of Technology,
Uka Tarsadia University,
Bardoli, India,
xyz@gmail.com, isha.vajani@utu.ac.in

Abstract. Analysis of twitter data can be very useful for marketing as well as for promotion strategies. This paper implements word frequency count for 1.5+ million twitters dataset. The hadoop components such as Apache pig and MapReduce framework are used for parallel execution of the data. This parallel execution makes the implementation time-effective and feasible to execute on a single system.

Keywords: Twitter data, Hadoop, Apache Pig, MapReduce, Big Data Analytics

1 Introduction

With the evolution of computing technologies and data science field, it is now possible to manage immensely large data and perform computation on it that previously could handled only by super computers or large datasets. Today social media is a critical part of the web. Social media platforms applications have a widespread reach to users all over the globe. According to statistica.com, the marketers use various social media as follows:

Table 1: Social Media Profiles for Marketing [5]

	Facebook	Twitter	LinkedIn	Google+	Pinterest
B2C	97%	81%	59%	51%	51%
B2B	89%	86%	88%	59%	41%

Social media sites and a strong web presence are crucial from marketing perspective in all forms of business. The interest in social media from all walks of life has been increasing exponentially from both application and research perspective. A wide number of tools, open source and commercial, is now available to gain a first impression of a particular social media channel. For popular social media channels like Twitter, Facebook, Google, YouTube many tools or services can be found to provide an overview of that channel. The span of all social media sites is growing faster day by day and as a consequence of that a lot of data is produced either structured or unstructured. This data can provide deeper insights by using analytics. Analysing the twitter data can lead to focused marketing goals.

According to internetlivestats.com, every day around 500 million tweets are tweeted on twitter. By analysing this tweets, the frequently appearing word can be derived, which can be provide useful insights for developing marketing strategy, finding trends of sale or inventory stock patterns. Here in this paper, 1.5 million+ tweets are analysed and each word frequency is calculated, which can further be used for many purposes.

Analysing such a massive data on a RDBMS or a single system is not feasible. The big data analytics must be used for analysis of such datasets. Hadoop Ecosystem is one of the solutions for big data analytics which can be used.

2 Hadoop Ecosystem

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures [2][3]. Hadoop Eco System is mainly consists of two services:

2.1 Hadoop Distributed File system

A reliable, distributed file system that is mainly used to store all the data of the hadoop cluster. The data is not necessarily well-structured as in case of databases, but data can be in the form of structured, un-structured or semi-structured format. The HDFS system works on a Master-Slave approach. It contains one Name node that acts as a master server. It manages the storage of data, allocation and de-allocation of disks and replication management of the data. There are numbers of Data nodes that acts as a slave. It manages the storage of file and handles read-write requests. It also handles block creation, deletion or update based on instruction from the Name node. Thus, with help of name node and data node, HDFS manages all issues like where to store the data physically? Or how many replicas should be generated? And how the replicas should be distributed across different nodes?

2.2 Hadoop Mapreduce

Hadoop Map-reduce is a software-framework that allows computation to be performed in parallel manner. Contradictory to an HDFS, Map-Reduce provides computation facilities on stored structured or unstructured data. The number of mappers and reducers varies based on application and data sizes. Hadoop also supports many applications such as Apache Pig, Apache Hive, Apache HBase, Apache ooozie and many more.

Table 2: Hadoop Ecosystem Components

Hadoop Component	Description
Sqoop	-Used as intermediate between SQL and Hadoop. - It let the user import tables or entire database into HDFS system.
HBASE	-It is a columnar storage based on Googles big table. -It can handle massive data tables combining billions and billions of rows and columns.
Pig	-It is high-level platform for creating Map-reduce program using Hadoop. - It can be used to build much larger, much more complex applications.
Hive	-It is a datawarehouse that that facilitates querying and managing large dataset residing in HDFS. - It uses query language called Hive QL.
Oozie	- It is workflow schedule system that manages all the hadoop jobs. - It is integrated with the hadoop stack and supports several different hadoop jobs.
Zookeeper	- It provides distributed configuration service and synchronization service. - It provides operational services for hadoop cluster.
Flume	- It is distributed and reliable service for efficient collection, aggregation and moving of large amount of data.
Mahout	- It focuses on distributed or scalable machine learning. - It is currently in its growing phase.
R Connectors	- It is most widely used open source statistical tool. - It contains rich set of packages for statistical analysis and plotting.

Each of these applications has its own working models and different applications. Based on requirement of a system, any of these applications can be used, either alone or together with other components. In this paper Apache Pig for data extraction and Map Reduce paradigm for processing the data is used.

3 Apache pig Overview

Apache Pig is a part of Hadoop Ecosystem which provides a high level language called Pig Latin. Pig commands are translated into Map-Reduce jobs. Pig tool can easily work with structured as well as unstructured data. Pig is a complete component of Hadoop Ecosystem as all the manipulation on the data can be done using Pig only. Pig can efficiently work with many languages such as Ruby, Python or Java etc. It can also be used to translate one data formats into other formats. Pig can ingest data from files, streams or any other data formats. A good example of Pig Application is ETL transaction model which basically extracts, transforms and loads the data [4].

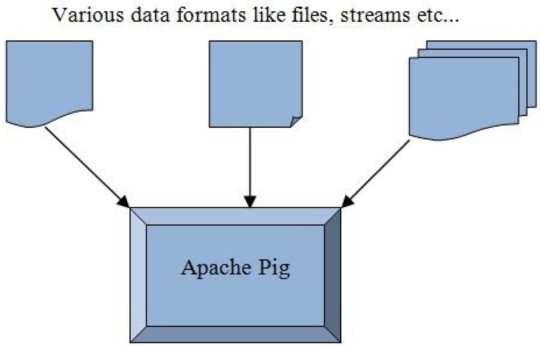

Fig. 1: Apache Pig data format support

4 Map Reduce Framwork Overview

Map reduce Framework is a programming model that was developed for parallel processing applications under Hadoop. It splits the data into various chunks and then these chunks are processed in parallel. Map Reduce brings computation to the data. It basically involves two phases [4]:
A. Mapper Phase
B. Reducer phase
In Mapper phase, Map job takes a collections of data and converts it into another set of data, where individual elements are broken into ¡key,value¿ pairs. While in Reducer phase, reduce jobs take the output of the map phase that is ¡key,value¿ pairs and produce smaller set of ¡key,value¿ pairs as an output. MapReduce framework works with the help of two processes provided by the hadoop eco system. A Job Tracker that sends the mapreduce tasks to specific nodes in the cluster. And a Task Tracker that are deployes on each machine in the cluster. It is responsible for running the map and reduce tasks as instructed by the Job tracker.

5 Tweets Analysis

Social media like twitter, facebook etc produces a lot of structured as well as unstructured data on daily basis, which if analysed can be very useful. The system working is shown in figure 2. The Twitter database contains a huge amount of data that grows rapidly every day. From the whole database, a dataset is selected for analysis which is basically a part of a twitter database. The task is divided into two processes namely

5.1 Extraction of tweets from twitter dataset

The twitter dataset contains lot of information such as user names, tweet date and timing, tweet contents and many more [5]. But for finding a word frequency only the tweet contents are needed. Similarly we can also count tweets per users by extracting user names only. But here we are concerned with tweets and its words so only tweets contents are extracted from the dataset.

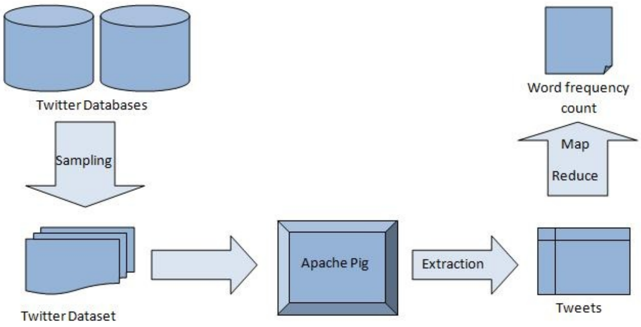

Fig. 2: Twitter word frequency Count System

Apache pig can be used for this transformation. It inputs the data from our dataset of CSV format and using Pig Latin we can extract tweet contents efficiently. This output data is in text format and contains only tweets content. This data can be stored in any other format also.

5.2 Word frequency count from tweets

After the extraction of data word frequency algorithm can directly be applied on that data. Map and reduce algorithm is used to obtain parallelism in this process. The algorithm for the same is presented in table 2 and table 3 as shown below.

Table 3: Main Notations

Algorithm: Word Frequency Mapper

```
Input: Text file
Output: <word,count>pair
1. Divide the data into different map jobs
2. For each line of text
3. Count each word separated by whitespaces
4. Form <word,count>pair locally
5. Temporary store all the ¡word,count¿ pair
6. Forward list of the <word,count>pair to reducer task
```

6 Results

In this system, a word frequency count is implemented which produces list of words appearing in tweets with their counts. The time analysis for execution of the program is shown in a figure 3. The system can efficiently perform in parallel and produces output file in comparatively small amount of time. With the variable size of data, system can take variable time. The execution time also depends on configuration of Hadoop Cluster. From the output data generated it is analysed that some regular words appears more frequently that others as shown below:

Table 4: Word Frequency Reducer Algorithm

Algorithm: Word Frequency Reducer

```
Input : <word,count>pair
Output: <word,total-count>pair
1. Prev_key = None
2. Current_key = word from first <word,count>pair
3. For each <word,count>pair repeat
4. If Prev_key == Current_key
5. total_count += count;
6. Else
7. total_count = count;
8. Prev_key = Current_key
9. Return all <word,total-count>pair
```

7 Conclusion and Future Work

In this paper, a system for counting the frequency of words of 1.5+ million tweets in parallel is implemented. It is implemented using two components of

Table 5: Some Frequent Words in Output

Some Words	Count
!	10666
But	13898
Hi	2770
I	341441
It	13244
My	14038
Thank	6262
You	3392
And many more...	

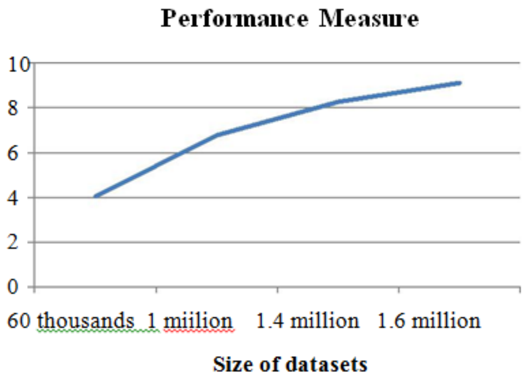

Fig. 3: Twitter word frequency Count System

the Hadoop Eco system namely Apache Pig and Map reduce framework. The parallel execution of the program makes it feasible and fast. The frequent word implementation can further be extended to analyse the dataset for marketing purpose or promotion strategy definitions. The system can be further extended to perform sentiment analysis or prediction of trends in markets.

References

1. White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.
2. CB.Singh , S.Gairola, B.N.Singh, A.Chandra, and K.A.Haddad, Multi-pulse AC-DC Converter for Improving Power Quality: A Review IEEE Transactions, On Power Delivery, Vol.23 No.1, January 2008.
3. Olston, Christopher, et al. Pig latin: a not-so-foreign language for data processing. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.
4. M. Bhandarkar, MapReduce programming with apache Hadoop, Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on, Atlanta, GA, 2010, pp. 1-1.

5. https://www.statista.com/chart/2289/how-marketers-use-social-media/
6. http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip