Big Data Analysis for Evaluating Current Trends

Priyanka Adhikari, Manasi Pradhan, Paras Sachdeva
BE, Computer Engineering
K.C. College of Engineering and Management Studies
& Research Thane(E)-400063, India

Abstract **— The big data analysis is one of the fastest growing trends in the world. Doing it using Hadoop is relatively new to be used outside the huge companies, which generate TBs of data every day. It is a new approach to analyze and deal with data apart from the usual traditional databases. In such cases of huge data, the companies need to maintain it, store it without damage and also analyze it to draw conclusions so as to make additional use of it. In this paper, the analysis of big data using the call logs of a telecom company is focused on. Every telecom company offering communication services seeks to provide customers with a broader usage environment. However, if the data and the usage of the customers is not properly analyzed, it ends up wrong prediction of facilities required by the customers. As a result, the customers might be dissatisfied with the service. In response to this problem, the study proposes to investigate several options for making the process easier and better for both, the users and the service providers.**

Keywords— Hadoop; bigdata; mapreduce.

I. INTRODUCTION

Telephone call data, such as originating station, destination, start and ending times, and transmission characteristics, is collected from a telecommunications system or private branch exchange (PBX) in form of call detail records (CDRs). The equipment typically presents this data on older PBXs via a serial communications port, or more recently via a computer network over an Ethernet connection. From the interface, CDRs are collected on computer systems running call logging and analysis software. Some PBX manufacturers provide their own basic call logging software but many other third-party software packages are available.

The idea of this project is to create a data model to analyse the trends in the given big data. Communications service providers are some of the biggest collectors of data today. Cell phones have evolved into mobile devices that literally serve as consumers' personal assistants and sidekicks. Each of their many functions generates its own, constant stream of multi-structured data, and all of that data must be efficiently captured, processed and analyzed.

For analyzing and finding out the trend in data,our proposed system will consider the data of all the customer's call log details (CDR's) irrespective of their location. It will then analyse all the data provided to it and output the applicable plan for the customer according to his/her usage. The output of the system will most likely include:

1. Which plan is in high demand(trending) in which locations or among which age groups.

2. The information of the customer's current plan and a plan apt as per his usage.

This system is mainly to be used by the telecom companies to process the data in petabyte scale if implemented in real time. The companies can check the customer requirement against the existing plans in the system and can Recommend Next Products to Buy (NPTB) according to the customer base, and sales associates use in-person or phone conversations to guess about NPTB recommendations, with little data to support their recommendations.

By doing this, the customer's will get a plan according to their needs and usage and thus satisfying the customer's need.

A. Existing System

Today in India, majority of telecom providing companies uses traditional data warehouse for the processing of Batch/structured data. To optimize average network quality, deployment, and coverage, legacy analysis of network traffic is done. Because of legacy analysis database system, fraud detection was difficult. All the analysis was done on historical and aggregated data, which consisted of large data redundancy. As a result, customer segmentation was not efficient enough as factors of event based marketing like geolocation was not up to the mark. The analysis was always done on the historic data which did not result in desired output. Back then, the number of mobile phone users was very less as compared to the present day. The amount of data stored increases tremendously over a small amount of time which has made the analysis and segregation a huge task now.

II. LITERATURE SURVEY

As our paper deals with the analysis of Big Data and the various tools that can be used for it, the first paper [1] we reviewed dealt with real time use of Hadoop analyzing the data as big as the data generated at Facebook. The changes made in the original use of Hadoop and the one used for the realtime use is:
Original HDFS- file system to support offline MapReduce where scalability and streaming performance are most critical. Changes: HDFS-single master called as NameNode. Since, consulting BackupNode takes huge time, AvatarNode was created. Hence NameNode is split into Active and Standy by AvatarNode.
In the next paper [2], the analysis required for Insights to Friend Recommendations on Facebook is the core problem for which solutions are suggested. Herein Scribe, Hadoop and Hive are used. Hive is used to bring metadata, partitioning, MySql to Hadoop. Scribe is used for Log collection, to aggregate logs from thousands of web servers.
Another paper [3] only deals with the method of MapReduce that is incorporated in Hadoop. MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value air and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model.
Moving over to the actual tools used to make the analysis by Hadoop easier, two components are reviewed in the next paper[4] which are lying on top of Hadoop, and provide higher-level language to use Hadoop's MapReduce library, namely Pig and Hive. Pig provides the scripting language to describe operations like the reading, filtering and transforming, joining, and writing data which are exactly the same operations that MapReduce was originally designed for.
Instead of expressing these operations in thousands of lines of Java code which uses MapReduce directly, Apache Pig lets the users express them in a language that is not unlike a bash or perl script. Apache Hive is "SQL like queries for Hadoop".
In the final paper [5], we present Hive, an open-source data warehousing solution built on top of Hadoop. Hive supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into map-reduce jobs executed on Hadoop. In addition, HiveQL supports custom map-reduce scripts to be plugged into queries. The language includes a type system with support for tables containing primitive types, collections like arrays and maps, and nested compositions of the same.

III. PROPOSED SYSTEM

Now as we have understood the existing system and have figured out that traditional database system is not capable of handling large scale data. So, what we propose in our system is to provide a parallel processing and highly scalable system. The present proposed methodology describes using the Hadoop ecosystem in analysing the tremendous amount of data in an efficient and desirable way.
According to the system, the data of a telecom company will be extracted and sorted according to the desired attributes. The attributes can be age, location, data usage, call logs. This will help the companies to create the plans according to the usage of the customers. This will also help them, to sort the data according the various attributes which will help them to devise new plans and implement new schemes as and when they are required.

IV. METHODOLOGY

The analytic system can be defined as the one which can input certain data from the user, sort it according to the analytics tool and give the desired results or data built upon these results as the output.
Hence, in this the data will be created. This data will have certain attributes like name, number, age, work location, usage (call usage, data usage), logs data (call log, data log). Some attributes we will show as null to depict the database as an unstructured one. This data will be stored on HIVE. The extraction of the data will be done by an SQL like query language – HiveQL. A Hive query will be issued for a particular user and the data of that user (which is stored on hadoop) will be extracted. The extracted data will be the input to the recommender system. In the recommender system, there will be various offers and discounts available apart from the general plans. (for every particular threshold value). The data given to the recommender system mainly includes call usage, data usage, call log and data log.)
Also, the location of the person shall be known. Based on the location nearby restaurants and cafes will be suggested. The recommender system will extract data. Based on the extracted data and call usage, it will display related schemes.

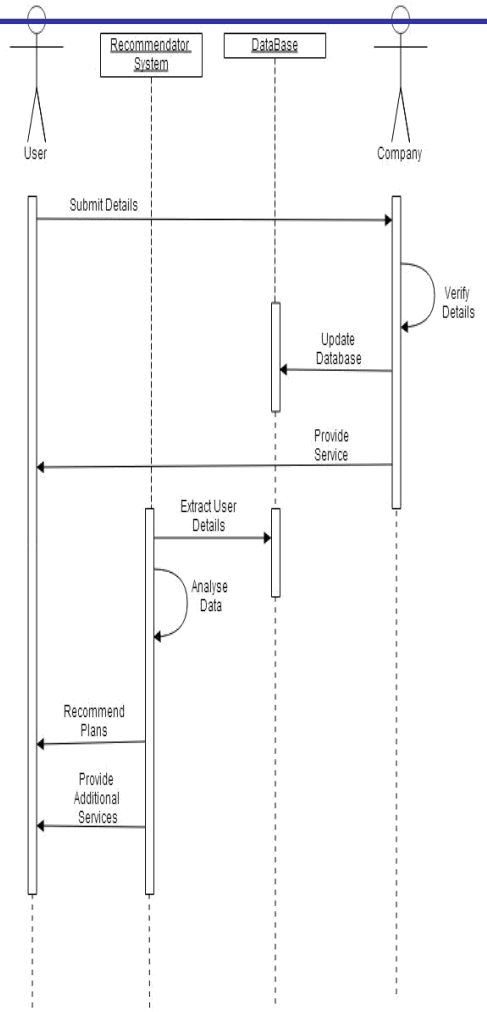

Figure 1: System Design

V. ADVANTAGES

A. Storage

As we are dealing with big data so our system needs to be efficient to store large amount of data. It stores both Structured and Unstructured data which reduces the effort of converting the obtained data in a particular format.

B. Real time and Cloud Support

To be more effective, the system should support real time data processing. It also supports Cloud which can be used to store larger amount of data.

C. Better Services

The system will provide better services to the customers and will always try to meet customer's current needs. It can conduct market research for customers using location, age and social technologies.

D. Parallel Processing

The system can be used to process the given data parallely so that the data can be processed quickly and we can get accurate and fast results. This will help to reduce the time for processing larger amount of data.

E. Scalibility

The system is highly scalable. As we are using batch processing in our system so scaling up the data or scaling down the data become easier.

F. Cost Efficient

As the system supports batch processing and can store large amount of data easily so the cost of the storage and management efforts reduces, even though the data increases.

VI. FUTURE SCOPE

This system proposes a recommender system using the call logs and other attributes. It is proposed after seeing the existing system and understanding the drawbacks. The described system makes it efficient in terms of working and brings about the change in seeing a data storage and analysing system of revolutionary system. It will be definitely beneficial for the companies as well as the users. We encourage bringing this system into implementation, on a much larger scale though we plan to execute it on a very miniscule level.

REFERENCES

[1] JoydeepSen Sarma, Nicolas Spiegelberg, Dmytro Molkov Rodrigo Schmidt, Apache Hadoop goes Realtime at Facebook, Sigmod Conference, 2011
[2] Dhruba Borthakur, Namit Jain, Joydeep Sen Sarma,Datawarehousing and Analytics Infrastructure at Facebook, 2010
[3] Jeffrey A. Delmerico, Nathanial A. Byrnes, Andrew E. Bruno,Matthew D. Jones,Steven M. Gallo, Vipin Chaudhary MapReduce: Simplified Data Processing on Large Clusters, 2004 [12] Sanjeev Dhawan, Sanjay Rathee, Hadoop Skeleton anf Fault Tolerance in Hadoop Clusters, 2013
[4] Vishal S Patil, Pravin D. Soni, Big Data Analytics using Hadoop Components like Pig and Hive, 2013
[5] Jeffrey Cohen, Brian Dolan,Mark Dunlap, Hive- A Warehousing Solution Over a Map-Reduce Framework, 2009
[6] Jens Dittrich, Jorge-Arnulfo Quiane-Ruiz, HOG: Distributed Hadoop Mapreduce on the Grid, 2012
[7] Chen He, Derek Weitzel, David Swanson, Ying Lu, Challenges and Opportunities with Big Data, 2012
[8] Chris Sweeney Liu Liu Sean Arietta Jason Lawrence Integrating R & Hadoop for Big Data Analysis, 2009
[9] Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein, Big Data Analysis using Sqoop and Flume, 2010
[10] DATASTAX CORPORATION, White Paper: DataStax-BigData, 2013
[11] Kai Ren, YongChul Kwon, Magdalena Balazinska, Bill Howe, Survey on Task Assignment Techniques on Hadoop, 2012
[12] Ekpe Okorafor, Mensah Kwabena Patrick Real-time Streaming Analysis for Hadoop and Flume, 2012