Predictive Analysis With Cricket Tweets Using Big Data

S Anjali [#1], V Aswini [#2], M Abirami [#3]
Department of Computer Science and Engineering,
Sri Manakula Vinayagar Engineering College.
Puducherry – 605 107, India.
[1] sanjaianjali@gmail.com

***Abstract* - We are now living in a world of Big Data, massive repositories of structured, semi-structured or unstructured data. Each organization holds "*electronic data or huge data*" in large volume. Some organization see it as a burden and some organization are exploring different ways to analyze, exploit and monetize the information contained within it but also have to tackle with the cost and risk of storing that data.**

This changing trend of companies led to the concept of Big data. One of the major applications of future generation parallel and distributed systems is in big-data analytics. One such field of data analysis is "sports". People all around the world tweet about cricket matches going on every day. This accounts to huge amount of data around the social networking sites. Those data can be fetched, analyzed and manipulated to predict the chances of team to win. By using big data concept such as map and reducing algorithm the fetched data is analyzed.

Keywords- big data; data analytics; cricket; winning team; flume

I. INTRODUCTION

Many of the industries today are adopting more analytical approaches to making decision. However, no other industry has the analytical initiatives under the domain of professional sports. There are multiple analytical domains to predict the player performance, player selection, business management, prevention of injury, predicting winning team and so forth.

Despite this evidence of impressive activity and growth, the use of analytic makes the viewers to get an excitement ,players to get confidence and plan the team stratergies in advance in the field of cricket. Relatively few owners, managers, coaches, and players pursued careers in professional sports because of their interest in analytics. Even when considerable data and analytics are available to support key decisions. However, it is clear that the use of analytics can contribute to success on the field. It's impossible to equate winning records with more analytical capability, but the recent success of highly analytical teams, hence for this case big data has introduced the tool named flume which is used to extract the data from the social networking sites.

Social networking sites such as twitter accounts for huge amount of data in the form of tweets in the field of cricket . Those data can be fetched, analyzed and manipulated to predict the chances of team to win. By using big data concept such as map and reducing algorithm the fetched data is analyzed.

II. RELATED WORK

In this section, we briefly present some of research literature related to data analysis .one such analysis is online auctions and the final price or the winning price prediction. Considerable work that applied traditional techniques to the online auction analysis has been made in the economics domain in the paper published in 2006 by authors Li Xuefeng, Sakaki. They also have developed an alerting system based on Tweets (posts in the Twitter micro blogging service), being able to detect earthquakes almost in real time . They elaborate their detection system further to detect rainbows in the sky, and traffic jams in cities. The practical point of their work is that the alerting system could perform so promptly that the alert message could arrive faster than the earthquake waves to certain regions.

Bollen et al. have analyzed moods of Tweets and based on their investigations they could predict daily up and down changes in Dow Jones Industrial Average values with an accuracy of 87.6%. Another example is using Twitter to predict electoral outcomes about the product and response of the particular product in the market. . Analysis is made to report on an attempt to build a minimalistic predictive model for the financial success of movies based on collective activity data of online users. We show that the popularity of a movie can be predicted much before its release by measuring and analyzing the activity level of editors and viewers of the corresponding entry to the movie in twitter, the well-known social media. Hence even though there are many analytics done in field of feelings, emotional moods, opinions and views of people, no such analysis is done in field of sports. People all around the world tweet about their leaders during election time, this involves the emotions, views, opinion of the on going election. Data analytics is used here to predict the pre-elective results based on peoples view.

III. DATA ANALYTICS

Big data analytics refers to the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics will help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data. Big data analytics enables organizations to analyze a mix of structured, unstructured and semi-structured data in search of variable business information. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. Consider that your organization could accumulate (if it hasn't already) billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. High-performance analytics is necessary to process that much data in order to figure out what's important and what isn't.

There are four approaches to analytics, and each falls within the reactive or proactive category: **Reactive – business intelligence.** In the reactive category, business intelligence (BI) provides standard business reports, ad hoc reports, OLAP and even alerts and notifications based on analytics. This ad hoc analysis looks at the static past, which has its purpose in a limited number of situations.

Reactive – big data BI When reporting pulls from huge data sets, we can say this is performing big data BI. But decisions based on these two methods are still reactionary.

Proactive – big analytics Making forward-looking, proactive decisions requires proactive big analytics like optimization, predictive modeling, text mining, forecasting and statistical analysis. They allow you to identify trends, spot weaknesses or determine conditions for making decisions about the future. But although it's proactive, big analytics cannot be performed on big data because traditional storage environments and processing times cannot keep up. Proactive – big data analytics**.** By using big data analytics you can extract only the relevant information from terabytes, petabytes and exabytes, and analyze it to transform your business decisions for the future. Becoming proactive with big data analytics isn't a one-time endeavor; it is more of a culture change – a new way of gaining ground by freeing your analysts and decision makers to meet the future with sound knowledge and insight.

IV. COMPARISON WTH OTHER SYSTEMS

The first consideration that needs to be made when selecting a database is the characteristics of the data you are looking to leverage. If the data has a simple tabular structure, like an accounting spreadsheet, then the relational model could be adequate. In similar cases today, one should consider NoSQL databases as an option. Multi-level nesting and hierarchies are very easily represented in the JavaScript Object Notation (JSON) format used by some NoSQL products. The developer requires high coding velocity and great agility in the application building process. NoSQL databases have proven to be a better choice in that regard. Since many NoSQL offerings include an open system, the community provides many productivity tools, another big advantage over single-vendor proprietary products. Some organizations, such as MongoDB, even offer free courses online that train employees and interested users in how to use the technology.

A.Operational issues: (scale, performance, and high availability)

As database grows in size or the number of users multiplies many RDBMS-based sites suffer serious performance issues. Next, consultants are brought in to look at the problem and provide solutions. Vertical scaling is usually recommended at high cost. As processors are added, linear scaling occurs, up to a point where other bottlenecks can appear. Many commercial RDBMS products offer horizontal scaling (clustering) as well, but these are bolted-on solutions and can be very expensive and complex.

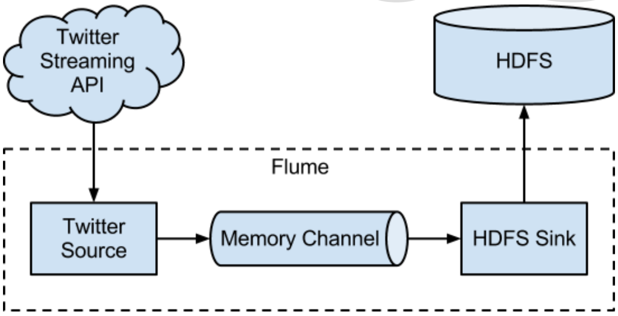
If an organization is facing such issues, then it should consider NoSQL technologies, as many of them were designed specifically to address these scales (horizontal scaling or scale-out using commodity servers) and performance issues. Just like Google's HDFS horizontal scaling architecture for distributed systems in batch processing, these newer NoSQL technologies were built to host distributed databases for online systems. Redundancy (in triplicate) is implemented here for high availability.

A common complaint about NoSQL databases is that they forfeit consistency in favor of high availability. However, this can't be said for all NoSQL databases. In general, one should consider an RDBMS if one has multi-row transactions and complex joins. In a NoSQL database like MongoDB, for example, a document (aka complex object) can be the equivalent of rows joined across multiple tables, and consistency is guaranteed within that object.

NoSQL databases, in general, avoid RDBMS functions like multi-table joins that can be the cause of high latency. In the new world of big data, NoSQL offers choices of strict to relaxed consistency that need to be looked at on a case-by-case basis.

V. BLENDINFG DATA FROM MULTIPLE SOURCES

The nature of Big Data is large data, usually from multiple sources. Some data will come from internal sources, but increasing data is coming from outside sources. These outside sources include Social media data feeds such as Twitter and Facebook Point of Sale and customer loyalty tracking programs Government agency sources such as census data Spatial data from mobile devices and satellite mapping feeds from the Consumer demographic data brokers, such as Experian Any number of public, private, or community clouds



Data blending is the process of combining multiple heterogeneous data sources and blending them into a single, usable analytic dataset. The purpose of data blending is to create analytic datasets to answer business questions using data that is not bound by the control and lengthy timelines of traditional IT processes. An example of data blending is when the data analyst integrates packaged, external data from the cloud with internal data sources to create a very business-specific analytic dataset.

A. Big Data Analysis Platforms and Tools

Hadoop: You simply can't talk about big data without mentioning Hadoop. The Apache distributed data processing software is so pervasive that often the terms "Hadoop" and "big data" are used synonymously. The Apache Foundation also sponsors a number of related projects that extend the capabilities of Hadoop. In addition, numerous vendors offer supported versions of Hadoop and related technologies. Operating System: Windows, Linux, OS X.

MapReduce: Originally developed by Google, the MapReduce website describe it as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes." It's used by Hadoop, as well as many other data processing applications.

GridGain: GridGrain offers an alternative to Hadoop's MapReduce that is compatible with the Hadoop Distributed File System. It offers in-memory processing for fast analysis of real-time data. You can download the open source version from GitHub or purchase a commercially supported version from the link above.

HPCC: Developed by LexisNexis Risk Solutions, HPCC is short for "high performance computing cluster." It claims to offer superior performance to Hadoop. Both free community versions and paid enterprise versions are available.

Storm: Now owned by Twitter, Storm offers distributed real-time computation capabilities and is often described as the "Hadoop of realtime." It's highly scalable, robust, fault-tolerant and works with nearly all programming languages.

B. Programming Languages

Pig/Pig Latin: Another Apache Big Data project, Pig is a data analysis platform that uses a textual language called Pig Latin and produces sequences of Map-Reduce programs. It helps makes it easier to write, understand and maintain programs which conduct data analysis tasks in parallel.

R: Developed by Bell Laboratories, R is a programming language and an environment for statistical computing and graphics that is similar to S. The environment includes a set of tools that make it easier to manipulate data, perform calculations and generate charts and graphs.

ECL: ECL ("Enterprise Control Language") is the language for working with HPCC. A complete set of tools, including an IDE and a debugger are included in HPCC, and documentation is available on the HPCC site. Operating System: Linux

Data Aggregation and Transfer

Sqoop: Sqoop transfers data between Hadoop and RDBMSes and data warehouses. Sqoop tool is used to access the structured data.

Flume: Another Apache project, Flume collects aggregates and transfers log data from applications to HDFS. Flume tool is used to access the unstructured data as well as the semi structured data.

Chukwa: Built on top of HDFS and Map Reduce, Chukwa collects data from large distributed systems. It also includes tools for displaying and analyzing the data it collects.

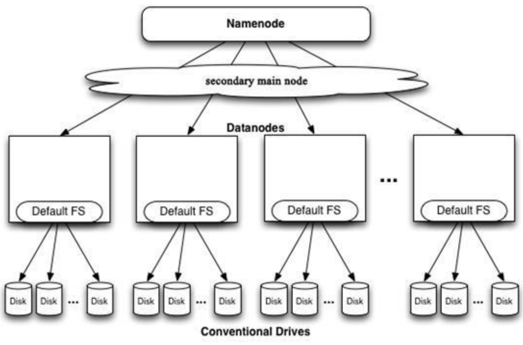
Hadoop:

Apache **Hadoop** is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software’s ability to detect and handle failures at the application layer.

Apache Hadoop has two pillars: YARN: Yet another resource nego

Google calls it:	Hadoop equivalent:
MapReduce	Hadoop
GFS	HDFS
Bigtable	HBase
Chubby	Zookeeper

HDFS: HDFS stores large files (typically in the range of gigabytes to terabytes across multiple machines). An Advantage of using HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map r reduce job to task trackers with an awareness of the data location.

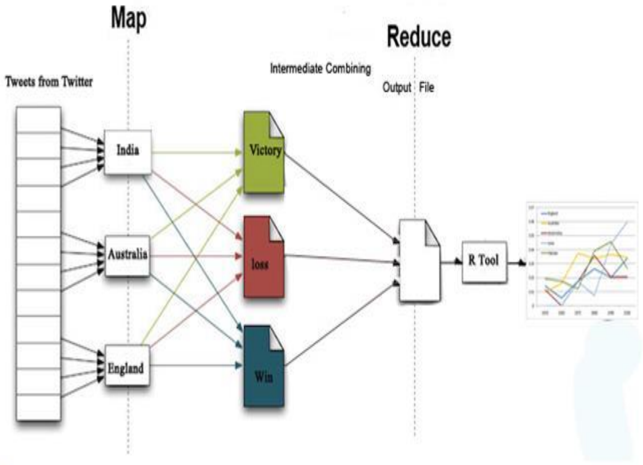


The HDFS file system includes a so-called secondary name node a misleading name that some might incorrectly interpret as a backup name node for when the primary name node goes offline .In fact, the secondary name node regularly connects with the primary name node and builds snapshots of the primary name node's directory information, which the system then saves to local or remote directories. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are commonplace and thus should be automatically handled in software by the framework.

VI. WORKING MODEL

The source of data for this project is twitter people tweet about cricket in this site the input data may be in many different format the data may contain negative as well as positive comments there may be some unwanted data regarding cricket .These data are retrieved with the help of big data tools like flume. Flume is nothing but a framework for populating hadoop with data agent are populated throughout ones it infrastructure- inside web servers, application servers and mobile devices for example-to collect data and integrate it into hadoop. Flume collects aggregates and transfers data from applications to hdfs. It's java-based, robust and fault-tolerant.(Operating system: windows, linux, os x).

The data once extracted is stored in hdfs server in the form of data nodes .Map and reduce process takes place in each data node.



Map Reduce program is composed of a Map()procedure that performs filtering and sorting and a **Reduce()** procedure that performs a summary operation. It s a software framework that serves as the computer layer of hadoop MapReduce jobs are divided into two parts. Jobs are divides into two parts they are: The

"Reduce "function aggregates the result of the "Map" function to determine the "answer" to the query. The "Map" function divides a query into multiple parts and processes data at the node level.

The *Map* and *Reduce* functions of *Map Reduce* are both defined with respect to data structured in (key, value) pairs. *Map* takes one pair of data with a type in one data domine, and returns a list of pairs in a different domain:

$$\text{Map(k1,v1)} \rightarrow \text{list(k2,v2)}$$

The *Map* function is appliezd in parallel to every pair in the input dataset. This produces a list of pairs for each call. After that, the MapReduce framework collects all pairs with the same key from all lists and groups them together; creating one group for each key.The *Reduce* function is then applied in parallel to each group, which in turn produces a collection of values in the same domain:

$$\text{Reduce(k2, list (v2))} \rightarrow \text{list(v3)}$$

Each *Reduce* call typically produces either one value v3 or an empty return, though one call is allowed to return more than one value. The returns of all calls are collected as the desired result list. Thus the Map Reduce framework transforms a list of (key, value) pairs into a list of values.

Mapper maps input key/value pairs to a set of intermediate key/value pairs .Maps are the individual tasks that transform input records into intermediate records. The transformed intermediate records do not need to be of the same type as the input records.

A given input pair may map to zero or many output pairs. Mapper implementations are passed the Job Conf for the job via the method and override it to initialize themselves the framework then calls for each key/value pair in the Input Split for that task. Applications can then override theCloseable. Close() method to perform any required cleanup.

Applications can use the Reporter to report progress, set application-level status messages and update Counters, or just indicate that they are alive. All intermediate values associated with a given output key are subsequently grouped by the framework, and passed to the Reducer(s) to determine the final output. Mapper outputs are sorted and then partitioned per Reducer. The total number of partitions is the same as the number of reduce tasks for the job. Users can control which keys (and hence records) go to which Reducer by implementing a custom Practitioner. User can optionally specify a combiner, via Job Conf. Set Combiner Class (Class) to perform local aggregation of the intermediate outputs which helps to cut down the amount of data transferred from the Mapper to the Reducer. The intermediate, sorted outputs are always stored in a simple (key-len, key, value-len, value) format. Applications can control if, and how, the intermediate outputs are to be compressed and the Compression Codec to be used via the Job Conf.

Reducer reduces a set of intermediate values which share a key to a smaller set of values. The number of reduces for the job is set by the user via Jon Conf. set Num Reduce Tasks (int). Overall Reducer implementations are passed the Job conf for the job via the Job Configurable, configure (Job Conf) method and can override it to initialize them.

The frame word then calls reduce (writable Comparable, Iterator, Output Collector, Reporter) method for each <key, (list of values)> pair in the grouped inputs. Applications can then override the Closablele. close() method to perform any required cleanup.

Shuffle

Input to the Reducer is the sorted output of the mappers. In this phase the framework fetches the relevant partition of the output of all the mappers, via HTTP.

Sort

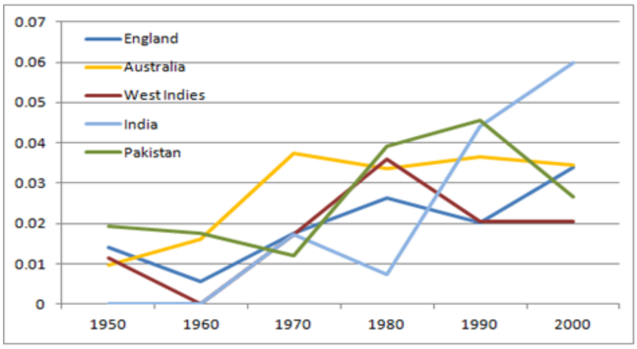
The framework groups Reducer inputs by keys (since different mappers may have output the same key) in this stage. The shuffle and sort phases occur simultaneously; while map-outputs are being fetched they are merged.

Secondary Sort

If equivalence rules for grouping the intermediate keys are required to be different from those for grouping keys before reduction, then one may specify a comparator via the command Jobconf. Set Output Value Grouping Comparator (class).

Visualizing Output-R Tool

The output generated from the above process is in the numeric form to view the output in the form of graph R Tool is used. The output generated in HDFS is given as input to the R Tool using R programming language. This tool generates graphical representation of above generated result .The below diagram shows the sample of the generated output.



VII. CONCLUSION

The live data consideration for analysis enables us to come out with more accurate results. The future scope may extent to inclusive of all the sports such as football, baseball, etc. An hybrid programing can be incorporated.

REFERENCES

[1] Mosteller F & Tukey J W (1977). Data analysis and regression. Menlo Park, CA: Addison -Wesley.

[2] Oren, Etzioni, Rattapoom, Tuchinda, Craig, A. Knoblock, & Alexander, Yates (2003). To buy or not to buy: Mining airfare data to minimize ticket purchase price KDD 2003 (pp. 119–128).

[3] Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of ACM, 40(3), C56–C58.

[4] Schafer J B, Konstan, J. A., & Riedl, J. (1999) Recommender systems in E-Commerce. Proceedings of the first ACM conference on electronic commerce, Denver, CO (pp. 158–166).

[5] Shah, H. S., Joshi, N. R., Sureka, A.&Wurman, P. R. (2003). Mining for bidding strategies on ebay. In Lecture notes in artificial intelligence Springer.

[6] Wellman, M. P., Reeves, D. M., Lochner, K. M., & Vorobeychik, Y. (2002). Price prediction in a trading agent competition.