International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 27th Aug 2016. Link :http://www.ijiemr.org/downloads.php?vol=Volume-5&issue=ISSUE-04

Title: **PROTECTING THE USERS SENSITIVE INFORMATION FROM MELICIOUS ATTACKS BY ANALYZING AND IDENTIFYING THE SUSPECIOUS APPLICATIONS IN SOCIAL NETWORK BY USING IPSFAPP**

Volume 05, Issue 04, Pages: 21-26
Paper Authors

JAGETI PADMAVTHI, B.GEETHA KUMARI
GNITS

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IDENTIFYING THE REQUIRED PATTERNS IN BIG DATA USING HADOOP

[1]JAGETI PADMAVTHI, [2]B.GEETHA KUMARI
[1,2]Assistant professor,Dept of CSE, GNITS

ABSTRACT: This document is an effort to present the basic understanding of BIG DATA and its utility for a performance-based organization. Along with the introduction of BIG DATA, important parameters and attributes have been highlighted that make this concept more attractive to organizations. The document also assesses the difference in the challenges faced by a small organization with regard to a medium or large scale operation, and hence the differences in its approach and the handling of GREAT DATA. Several examples of BIG DATA implementation applications have been presented in areas ranging from strategy, product and process. The second part of the document deals with the technological aspects of BIG DATA for its implementation in organizations. Since HADOOP has become a popular tool for implementing BIG DATA, the document deals with the general architecture of HADOOP along with the details of its various components. In addition, each of the components of the architecture has been taken and described in detail.

1. INTRODUCTION:

BIG DATA: In 2009 a new FLU virus was discovered.That is bird flu and Swine Flu. This disease is spread quickly. By these diseases, public health agencies are feared. In 1918 Spanish Flue that had infected half a billion people and killed tens of millions. Worse, no vaccine against the new virus was readily available. The only hope public health authorities had been to slow its spread. But to do that they needed to know where it already was. In the United States, the Center for Disease Control and Prevention (CDC) requested doctors inform them of new FLU cases. Some people might feel sick, but they wait before consulting the doctor. So exact information about the FLU cases will take to reach to Central organizations took time, It is a splash between health officials and computer scientists, but they are not showing interest in it.In this situation more people search about the flu and their preventing medicines. Google gets nearly 3 million client queries. They know that every client wants to details about the flu and their medicines. So google done 450 million different mathematical terms in order to test search terms, comparing their predictions against actual flu cases from the CDC in 2007 and 2008. It is an only one area where big data, making a big difference. Whole business divisions are being reshaped by bid data also. Purchasing

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

plane tickets are a decent sample.Etzioni is one of America's foremost computer scientists. He sees the world as a series of big-data problems ones that he can solve. And he has been mastering them since he graduated from Harvard in 1986 as its first undergrad to major in computer science. One day he booked a plane ticket to attend his brother's marriage. He booked ticket one month early. He thought that if he purchase a plain ticket before then the cost will also less. When he get in the plane eagerly asked his beside persion about the price of the ticket. His ticket cost is less than the Etzioni. That persion booked his ticket retcently. He got angry. But he is a computer scientist that's why he realize that there is a big data problem.From his roost at the University of Washington, he began a huge number of big-data organizations before the expression "big-data" got to be known. He helped form one of the Web's first web crawlers, MetaCrawler, which was propelled in 1994 and gobbled up by InfoSpace, then a ma¬jor online property. He helped to establish Netbot, the first major comparison-shopping site, which he sold to Excite. His startup for extracting importance of content records, called ClearForest, was later obtained by Reuters.Now a days online social media are increasing rapidly. Every hour people uploading 10 million images in Facebook. Per day they are clicking like button and leave a comment 3 billion times. In Google and YouTube 800 monthly users uploading every second. The video length is approximately one hour length. In 2012 twitter had exceeded 400 million tweets a day. Like this every day online storage was increasing rapidly,Before big data our its enough to use the normal process to mine the data. But nowadays it is impossible to mine the data easily by using the normal process. Because millions of data increasing in every social sites. Online shopping also having the big data problem. Cosumers are showing interest to purchase items online. Maintaining the large amount of data is not so easy. So we should use big data process to perform actions like minig data, maintain data. In big data we should consider four Vs. They are

1. Volume:It refers to the vast amount of data generated every second.
2. Velocity:It refers to speed of data generated and moves around.
3. Variety:It refers the different types of data we can use. Previous days we consider only structured data. Infact 80% of data is unstructured. With big data we can easily analise and bring data of different types such as messages, chatting, images, video and audio.
4. Veracity:It refers to the trustworthyness of the data.

For big data processing we should use Hadoop. Apache hadoop is an open source software framework for storage the large scale processing of datasets. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Nowadays big data usage was increasing rapidly. To maintain and

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

process the data perfectly we should use Big Data Techniques.

2. RESEARCH:

A NOVEL APPROACH FOR PROCESSING BIG DATA: Applications of machine learning are widely used in the real world with either supervised or unsupervised learning process. Recently emerged domain in the information technologies is Big Data which refers to data with characteristics such as volume, velocity and variety. The existing machine learning approaches cannot cope with Big Data. The processing of big data has to be done in an environment where distributed programming is supported. In such environment like Hadoop, a distributed file system like Hadoop Distributed File System (HDFS) is required to support scalable and efficient access to data. Distributed environments are often associated with cloud computing and data centers. Naturally such environments are equipped with GPUs (Graphical Processing Units) that support parallel processing. Thus the environment is suitable for processing huge amount of data in short span of time. In this paper we propose a framework that can have generic operations that support processing of big data. Our framework provides building blocks to support clustering of unstructured data which is in the form of documents. We proposed an algorithm that works in scheduling jobs of multiple users. We built a prototype application to demonstrate the proof of concept. The empirical results revealed that the proposed framework shows 95% accuracy when the results are compared with the ground truth.

Sentiment analysis using big data: The Web has become an excellent source for assembling consumer opinions. There are now several Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This paper focuses on online customer reviews of products. It makes two contributions. First, it proposes a framework for analyzing and comparing consumer opinions of competing products in map and reduce environment for better analysis. Second, a new technique based on lexicon based technique is proposed to extract neutral reviews and restrict them from being categorized under positive or negative. Experimental results show that the technique is highly effective and smash existing methods significantly.

Sentiment analysis of Twitter data within big data distributed environment for stock prediction: This paper covers design, implementation and evaluation of a system that may be used to predict future stock prices basing on analysis of data from social media services. The authors took advantage of large datasets available from Twitter micro blogging platform and widely available stock market records. Data was collected during three months and processed for further analysis. Machine learning was employed to conduct sentiment classification of data coming from social networks in order to estimate future stock prices. Calculations were performed in distributed environment according to Map Reduce programming model. Evaluation and discussion of results of predictions for

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

different time intervals and input datasets proved efficiency of chosen approach is discussed here.

3. Big Data For Time Reduction

The second common goal of Big Data solutions and technologies is to reduce time. Macy's price optimization application provides a classic example of cycle time reduction for complex and large-scale analytic calculations for hours or even days in minutes or seconds. The chain of department stores has been able to reduce the price optimization times of its 73 million items sold for more than 27 hours to just over 1 hour. Described by some as "big data analytics," this set of features allows Macy's to re-evaluate articles much more frequently to adapt to changing retail market conditions. This great data analysis application extracts data from a Hadoop cluster and places them in other software architectures in parallel and in memory. Macy's also says it has reached a 70% hardware cost reduction. Kerem Tomak, Macys.com's vice president of analytics, is using similar approaches to reduce time for marketing offers for Macy customers. It detects that the company can run many more models with this time savings.

4. Hadoop:

Hadoop is a java based programming Big Data framework and also an open source framework. It is used to process and store the large amount of data sets in distributed network. We can run the hadoop application in the large network which is having the commodity hardware. This framework can continue its work when the node is faile in the network. There are various types of processing and analysis with hadoop they are:

- **MapReduce:** MapReduce is a programming model which is used to process the data. It can develop using various types of languages. Hadoop can run the MapReduce programs. MapReduce programs can develop in Ruby, Java, and Python. Developer needs to develop two methods such as Map and Reducer. Map and Reducer methods having the data set in the form of key value pair.
- **Hadoop Distributed File System (HDFS):** Hadoop distributed file system is designed for the storing very large amount of data files with the streaming data access patterns and it can run on commodity hardawares. It is highly fault-tolerent and run on commodity hardware aslo called as low cost hardware.
- **YARN (Yet Another Resources Negotiator):** Yet Another Resources Negotiator was introduced in Hadoop 2. It is a resource management system of Hadoop's cluster. Main purpose of YARN is to improve the MapReducde implementation.
- **Flume:** Flume is designed for moving or ingestion the data into Hadoop. For example the log file are collected from the web servers and move the logevenets form those logfiles into HDFS for processing.
- **PIG:** It is a high level of abstraction for the processing of large amount of

data sets. With pig the structure of data is very rich, multi valued, and we can apply the transformations to the much powerful data.

- **Hive:** Hive is the data warehouse which can facilitate the writing, reading, and manage the huge amount of datasets which are residing in the distributed storage using SQL (Structured Query Language).
- **HBase:** HBase is a distributed and column oriented database which is built on the top of the HDFS. This can be used when we required realtime read or write random access to the huge amount of datasets.
- **NoSQL database:** NoSQL also refered as Not Only SQL database. It is mainly design for target huge set of distributed data. It is a Database design which is implement the document store, key-value store, graph format for data and column store.

These tools are useful to analyze the structured, semi-structred and unstructured data. Managing, storing and processing of huge data is critical in every organization. Implementing the Big data tools in their organization they can easily manage, process and store large data.

5. CONCLUSION

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data,

Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits. With Hadoop, no data is too big. And in today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

6. REFERENCES

1. Shen Yin, Okyay Kaynqak – "Big Data for Modern Industry: Challenges and Trends" Vol. 103, No. 2, February 2015, Proceeding of the IEEE.
2. Hsinchun Chen, Roger H.L. Chiang, Veda C. Storey - "Business Intellegence and Analytics: From Big Data to Big Impact" - MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012.
3. Sudhakar Singh, Pankaj Singh, Rakhi Garg, P K Mishra - "Big Data: Technologies, Trends and Applications" IJCSIT Vol. 6 (5), 2015
4. Susan Miele, Rebecca Shockley - "Analytics: The real – world use of big data" IBM Global Business Services
5. "Big Data Analytics: Advanced Analytics in oracle database" - An Oracle White Paper March 2013.

6. Jafar Raza Alam, Asma Sajid, Ramzan Talib, Muneeb Niaz - "A Review on the Role of Big Data in Business", IJCSMC, Vol. 3, Issue. 4, Aprial 2014.

7. Shen Yin, Okyay Kaynqak – "Big Data for Modern Industry: Challenges and Trends" Vol. 103, No. 2, February 2015, Proceeding of the IEEE.

8. Hsinchun Chen, Roger H.L. Chiang, Veda C. Storey - "Business Intellegence and Analytics: From Big Data to Big Impact" - MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012.

9. Sudhakar Singh, Pankaj Singh, Rakhi Garg, P K Mishra - "Big Data: Technologies, Trends and Applications" IJCSIT Vol. 6 (5), 2015

10. Susan Miele, Rebecca Shockley - "Analytics: The real – world use of big data" IBM Global Business Services

11. "Big Data Analytics: Advanced Analytics in oracle database" - An Oracle White Paper March 2013.

12. Jafar Raza Alam, Asma Sajid, Ramzan Talib, Muneeb Niaz - "A Review on the Role of Big Data in Business", IJCSMC, Vol. 3, Issue. 4, Aprial 2014.