Analysis of Machine Learning Algorithms to Protect from Phishing in Web Data Mining

N. Swapna Goud
Assist Prof
CVSR Engineering College
Hyderabad.

ABSTRACT

The term Big data is a large data sets those outgrow the simple kind of database and data handling design. We designed prototype of website phishing detection solution to address the requirements for both effective and efficient phishing detection machine learning big data allows us to dig into a tremendous amount of data that fix the problem and extract predictive signals for the phishing problem. As the cyber security problems grows many types of phishing activities may arises bid data analytics is pretty helpful in identifying various phishing threats of suppliers by scanning various data roots such as personal contacts service level agreements exploring various unstructured data sources log reports and big data analysis highly suitable for analyzing. Our research work presents big data analytics that aims to prevent malicious email notifications & phishing from web service

Keywords

Phishing, Cybercrime, Big Data, Webservice, Emails

1. INTRODUCTION

Most phishing web pages have high visual similarities to scam the victims look exactly like the real ones , internet users may be easily deceived by the scam victims of e-banking phishing websites may expose the bank account password credit card number other important information to the phishing web phishing web page owners. The impact is the security through the compromise of confidential data and victims may finally suffer losses of money new internet crime in much comparison with other forms for example virus and hacking. Many phishing web pages have been found earlier in accelerative way from phrase website phishing is a variation on the word fishing, the phishing website is very complex issue to analyze technical and social problem with each other for which there is no known single silver bullet to entirely solve it. Security threat to online business comes from what becomes to be phishing attacks malicious people create webpages that mimic the webpages of legitimate websites. Legitimate mistakenly access the faked web site and expose the financial and personal information to anomalous people whom might use this information to perform illegal and criminal activities such criminal acts causes a lot of lose for both the clients and the legitimate organization. Phishing attacks continues to succeed threatens the whole online shopping organization as secure sector financial activities approaches to detect phishing website some of adapted by the organization mainly based on keeping a list of URLs called a blacklist such as Google Safe Browsing Microsoft anti-phishing protection. Blacklist is a list of URLs though to be anonymous user attacks website browser refers to the blacklist to examine if the visited URL is present within the blacklist are considered as malicious and browser warns the user to store locally on a server that is queried by the browser for every requested URL.

Data mining techniques like neural networks rule induction decision tree machine learning can be useful to the fuzzy logic model. It can deliver to business questions that traditionally too time consuming to resolve phishing website characteristic by analyzing massive databases and historical data for training label. Detecting phishing websites relies on using a machine learning data mining algorithm that recognize the phishing website based on a set of characteristics or features are recognized by experts to be characteristics phishing website uniform resource locator. Machine learning methods search for a best model matches the testing data statistical the searching space in machine learning methods is cognitive spa of n attributes instead of a vector space of n dimensions but most machine learning methods is a cognitive space of n attributes instead of a vector space of n dimensions. Most common machine learning methods used for data mining include decision tree induction concept learning and conceptual clustering which determines an objects class by following the path from the root to a leaf node choosing the branches according to the attributes values of the object. Database oriented methods do not search for a best model as the previous methods specific heuristics are used to exploit the characteristics of the data, attribute oriented induction iterative database scanning for frequent itemsets and the attribute focusing are representatives of the database oriented methods.

2. RELATED WORK

In many researches investigate the phishing websites are proposed with techniques to detect the Nguyen 2014 an efficient approach for phishing detection with neural network that are evaluated with the help of dataset of 11660 phishing and legitimate websites evaluates the weights of heuristic layer of neural networks and precision of the system upto 98%. A content based scheme is a linear trait classifier developed by TF-IDF algorithm condense the false positives, improved the technique by using two dissimilar gamuts of corpora filters are employed to dilute the false positive and to consummate the runtime speedup. Detection of phishing attacks alleviation techniques are deliberated mitigation techniques of detection offensive prevention and detection develops a framework for predicting the phishing websites. The neural networks used to predict the phishing websites multilayer neural networks shrink the error and elevate the performance conducted in the survey on phishing attacks. The survey discovers the various aspects of phishing attacks their problems for searching the better solutions also improve privacy and security without compromising the benefits of information sharing through tricks by email scammers in phishing emails that spoof a reputable company in an attempt to defraud the recipient of personal information some threats to users privacy. Context based dynamic intelligent phishing detection is analyze and detect phishing in instant messages with relevance to domain ontology and utilize the classification based on association for generating phishing rules and alerting the instant message system generically cannot detect. Active phishing detection developed using data mining classification techniques and ontology association rule mining machine learning technique used to detect the domain of the message keywords, phishing rules considered to be mapped with the ontology generator through NLP to obtain the domain of the suspected word not alerted instead the crime departments is notified in suspicious activity is detected.

3. PROBLEM DEFINTION

The Problem throughout the thesis is phishing in websites which play increasingly vital role in modern society have become the targets of hackers or intruders. The security is compromised when a misuse occurs. A misuse can be defined as any set of actions that attempt to compromise the integrity, confidentially or availability of a system. Misuse prevention technique such as user authentication, avoiding programming error and information protection is not sufficient because as mobile device become ever more complex, there are exploitable weaknesses in the device. Therefore fraud detection is needed as another protection to protect our system.

The term phishing is used to steal personal information through spamming number of different phishing techniques used to obtain personal information from users. To prevent internet phishing users should have knowledge of various types of phishing techniques that should be aware of ant-phishing to protect them from phished. Phishers send the same email to number of users requesting them to ask personal information for illegal purpose, in other words type of scam. Web based delivery is the most sophisticated phishing techniques know as man in the middle the hacker is located in between the original website and the phishing system traces the information during transaction between the legitimate website and the user. Instant messaging is the method in which the user receives a message with in a link directing to fake phishing website which has the same lock and feel as the legitimate website suppose the user doesn't look at the URL may be hard to tell the difference between the fake and legitimate websites.

Link manipulation is the technique which the phisher sends a link to website on deceptive link it opens up the phishers website instead of the website mentioned in the link, the anti-phishing method used to prevent link manipulation is to move the mouse over the link to view the actual address. Key loggers refer to the malware used to identify inputs from sent to the hackers will decipher passwords and other type of information that prevent key loggers from accessing personal information secure websites provide options to use. Session hacking is phisher exploits the web session that control mechanism to steal information from the user hacking procedure known as session sniffing the phisher can use a sniffer to intercept relevant information so that access web server illegally. The phisher makes phone calls to the user and asks the user to dial the number to get the information of the back account through the phone with fake caller ID, malware is usually attached to email sent to the user by the phishers that will start functioning be attached to downloadable files. The techniques to prevent the phishing are detailed below.

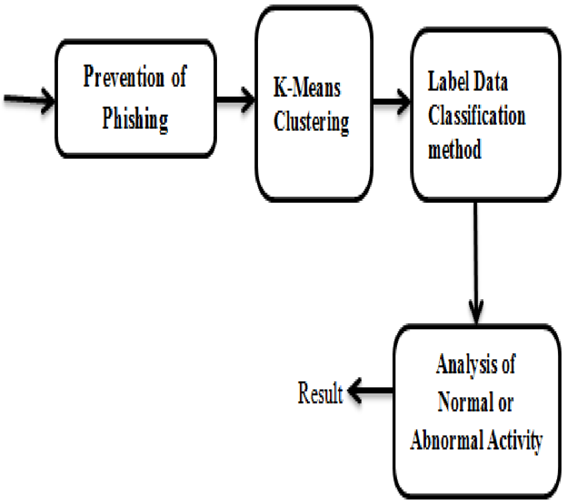
Prevention of Phishing
K-Means Clustering
Label Data Classification method
Analysis of Normal or Abnormal Activity
Result

Figure 1 - Flow of Problem definition

3.1. Clustering

Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind it deals with finding a structure in a collection of unlabeled data. The process of grouping a set of physical or abstract objects into classes of similar objects is called Clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for Outliers may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and monitoring of criminal activities in electronic commerce.

Data clustering is under vigorous development contributing areas of research include data mining, statistics. Machine learning, spatial database technology, biology, and marketing.

We can show this as simple graphical example:

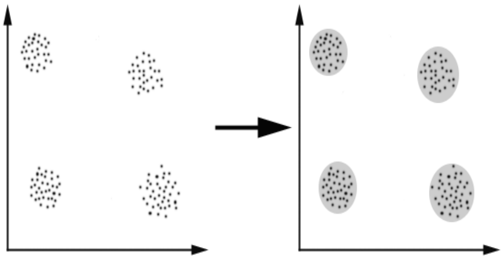

Figure 2 Graphical example of clusters

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance:* two or more objects belong to the same cluster if they are "close" according to a given distance in this case geometrical distance

Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering.

For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

Problems of Clustering Techniques

There are a number of problems with clustering. Among them:

Current clustering techniques do not address all the requirements adequately (and concurrently)

Dealing with large number of dimensions and large number of data items can be problematic because of time complexity;

The effectiveness of the method depends on the definition of "distance" (for distance-based clustering);

If an *obvious* distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;

The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

The major clustering methods can be classified into following methods, Partitioning method, Hierarchical method, Density-based method, Grid-based method, Model-based method etc.

The advantage of clustering is scalability that include increased application performance and the support of a greater number of users. Sequential rows of a particular table are stored together on disk, page-by-page. This makes particular sense with the relational architecture as most database activity is based on some form of sequential data access of a table.

The main reason for choosing clustering for this problem is to determine the intrinsic grouping in a set of unlabeled data. When the numbers of unsolicited packets are relatively large, it can take substantial time and effort for detection and prevention of intrusion. Using clustering client can classify and cluster the packets based on their IP addresses, port numbers, and protocols and so on. Clustering is the only method by which security of the system can be enhanced by analyzing the packets, which will help in identifying and preventing the intrusion.

3.2. Classification

In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps:

a) Create training data set.

b) Identify class attribute and classes.

c) Identify useful attributes for classification

(relevance analysis).

d) Learn a model using training examples in

training set.

e) Use the model to classify the unknown data samples.

3.2.1. Decision Tree:

Decision tree support tool that uses tree-like graph or models of decisions and their consequences, including event outcomes, resource costs, and utility. Commonly used in operations research, in decision analysis help to identify a strategy most likely to reach a goal. In data mining and machine learning, decision tree is a predictive model that is mapping from observations about an item to conclusions about its target value. The machine learning technique for inducing a decision tree from data is called decision tree learning.

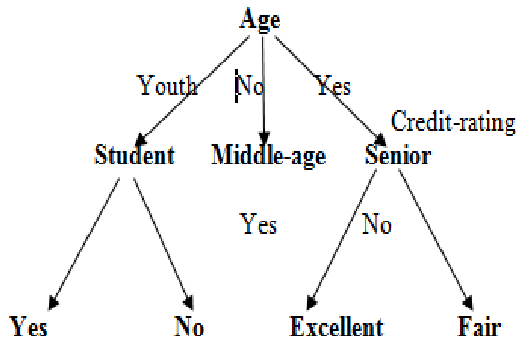

Figure: 3. Example of supervised learning decision tree

In above fig tree is classified into leaf nodes. In a decision tree, each leaf node represents a rule. It is concept buys computer, indicating whether a customer at company is likely to buy a computer. Each internal node represents a test on attribute. Each leaf node represents a class either buys_computer= yes or buys_computer= no. The following rules are as follows in figure (4.2) Rule **1:** If he is student and is youth then he will buy. Rule2 : If he is student and is not youth then he do not buy. Rule3 : If he is middle-age, then he will buy. Rule4 : If he is senior and is excellent then he will buy. Rule5 : If he is senior, is fair then he don't want to buy.

4. FIGURES/CAPTIONS

In existing system uses the concept of Data mining machine learning techniques now we have the big data machine learning techniques to detect and prevent the phishing users from secure websites. Experimental research in cyber security is rarely reproducible because today's data sets are not widely available to the research community and are often insufficient for answering many open questions. Due to scientific, ethical, and legal barriers to publicly disseminating security data, the data sets used for validating cyber security research are often mentioned in a single publication and then forgotten. The "data wishlist" (Camp, 2009) published by the security research community in 2009 emphasizes the need to obtain data for research purposes on an ongoing basis.

WINE provides one possible model for addressing these challenges. The WINE platform continuously samples and aggregates multiple petabyte-sized data sets, collected around the world by Symantec from customers who agree to share this data. Through the use of parallel processing techniques, the platform also enables open-ended experiments at scale. In order to protect the sensitive information included in the data sets, WINE can only be accessed on-site at Symantec Research Labs. To conduct a WINE experiment, academic researchers are first required to submit a proposal describing the goals of the experiment and the data needed. When using the WINE platform, researchers have access to the raw data relevant to their experiment. All of the experiments carried out on WINE can be attributed to the researchers who conducted them and the raw data cannot be accessed anonymously or copied outside of Symantec’s network.

In the Case Study Zions Bancorporation announced that using Hadoop cluster and business intelligence tools to parse more data more quickly than with traditional SIEM too much, their experience quantity of data and the frequency of analysis too much traditional to handle alone.

4.1 Enterprise Analytics Tool

Enterprises routinely collect terabytes of security relevant data (e.g., network events, software application events, and people action events) for several reasons, including the need for regulatory compliance and post-hoc forensic analysis. Unfortunately, this volume of data quickly becomes overwhelming. Enterprises can barely store the data, much less do anything useful with it. For example, it is estimated that an enterprise as large as HP currently (in 2013) generates 1 trillion events per day, or roughly 12 million events per second. These numbers will grow as enterprises enable event logging in more sources, hire more employees, deploy more devices, and run more software. Existing analytical techniques do not work well at this scale and typically produce so many false positives that their efficacy is undermined. The problem becomes worse as enterprises move to cloud architectures and collect much more data, as a result, the more data that is collected, the less actionable information is derived from the data. The goal of a recent research effort at HP Labs is to move toward a scenario where more data leads to better analytics and more actionable information (Manadhata, Horne, & Rao, forthcoming). To do so, algorithms and systems must be designed and implemented in order to identify actionable security information from large enterprise data sets and drive false positive rates down to manageable levels. In this scenario, the more data that is collected, the more value can be derived from the data. However, many challenges must be overcome to realize the true potential of Big Data analysis. Among these challenges are the legal, privacy, and technical issues regarding scalable data collection, transport, storage, analysis, and visualization. Despite the challenges, the group at HP Labs has successfully addressed several Big Data analytics for security challenges, some of which are highlighted in this section. First, a large-scale graph inference approach was introduced to identify malware-infected hosts in an enterprise network and the malicious domains accessed by the enterprise's hosts. Specifically, a host-domain access graph was constructed from large enterprise event data sets by adding edges between every host in the enterprise and the domains visited by the host. The graph was then seeded with minimal ground truth information from a black list and a white list, and belief propagation was used to estimate the likelihood that a host or domain is malicious. Experiments on a 2 billion HTTP request data set collected at a large enterprise, a 1 billion DNS request data set collected at an ISP, and a 35 billion network intrusion detection system alert data set collected from over 900 enterprises worldwide showed that high true positive rates and low false positive rates can be achieved with minimal ground truth information (that is, having limited data labeled as normal events or attack events used to train anomaly detectors).

Second, terabytes of DNS events consisting of billions of DNS requests and responses collected at an ISP were analysed. The goal was to use the rich source of DNS information to identify botnets, malicious domains, and other malicious activities in a network. Specifically, features that are indicative of maliciousness were identified. For example, malicious fast-flux domains tend to last for a short time, whereas good domains such as *hp.com* last much longer and resolve to many geographically-distributed IPs. A varied set of features were computed, including ones derived from domain names, time stamps, and DNS response time-to-live values. Then, classification techniques (e.g., decision trees and support vector machines) were used to identify infected hosts and malicious domains. The analysis has already identified many malicious activities from the ISP data set.

4.2 Monitoring to Identify Botnets

The BotCloud research project (Fraçois, J. et al. 2011, November), which leverages the MapReduce paradigm for analysing enormous quantities of Netflow data to identify infected hosts participating in a botnet (François, 2011, November). The rationale for using MapReduce for this project stemmed from the large amount of Netflow data collected for data analysis. 720 million Netflow records (77GB) were collected in only 23 hours. Processing this data with traditional tools is challenging. However, Big Data solutions like MapReduce greatly enhance analytics by enabling an easy-to-deploy distributed computing paradigm. BotCloud relies on BotTrack, which examines host relationships using a combination of PageRank and clustering algorithms to track the command-and-control (C&C) channels in the botnet (François et al., 2011, May). Botnet detection is divided into the following steps: dependency graph creation, PageRank algorithm, and DBScan clustering. The dependency graph was constructed from Netflow records by representing each host (IP address) as a node. There is an edge from node A to B if, and only if, there is at least one Netflow record having A as the source address and B as the destination address. PageRank will discover patterns in this graph (assuming that P2P communications between bots have similar characteristics since they are involved in same type of activities) and the clustering phase will then group together hosts having the same pattern. Since PageRank is the most resource-consuming part, it is the only one implemented in MapReduce. BotCloud used a small Hadoop cluster of 12 commodity nodes (11 slaves + 1 master): 6 Intel Core 2 Duo 2.13GHz nodes with 4 GB of memory and 6 Intel Pentium 4 3GHz nodes with 2GB of memory. The dataset contained about 16 million hosts and 720 million Netflow records. This leads to a dependency graph of 57 million edges.

The number of edges in the graph is the main parameter affecting the computational complexity. Since scores are propagated through the edges, the number of intermediate MapReduce key-value pairs is dependent on the number of links. Figure 5 shows the time to complete iteration with different edges and cluster sizes.

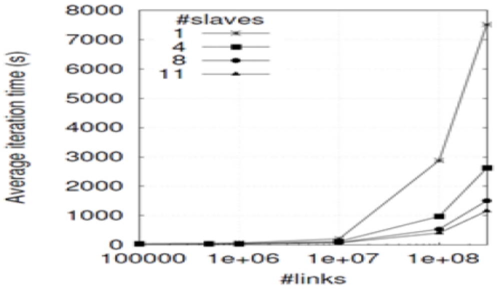

Figure 2 Average Execution time

The results demonstrate that the time for analysing the complete dataset (57 million edges) was reduced by a factor of seven by this small Hadoop cluster. Full results (including the accuracy of the algorithm for identifying botnets) are described in François et al.

4.3 Detection of Advanced Persistent Threats

An Advanced Persistent Threat (APT) is a targeted attack against a high-value asset or a physical system. In contrast to mass-spreading malware, such as worms, viruses, and Trojans, APT attackers operate in "low-and-slow" mode. "Low mode" maintains a low profile in the networks and "slow mode" allows for long execution time. APT attackers often leverage stolen user credentials or zero-day exploits to avoid triggering alerts. As such, this type of attack can take place over an extended period of time while the victim organization remains oblivious to the intrusion. The 2010 Verizon data breach investigation report concludes that in 86% of the cases, evidence about the data breach was recorded in the organization logs, but the detection mechanisms failed to raise security alarms (Verizon, 2010). APTs are among the most serious information security threats that organizations face today. A common goal of an APT is to steal intellectual property (IP) from the targeted organization, to gain access to sensitive customer data, or to access strategic business information that could be used for financial gain, blackmail, and embarrassment, data poisoning, illegal insider trading or disrupting an organization's business. APTs are operated by highly-skilled, well-funded and motivated attackers targeting sensitive information from specific organizations and operating over periods of months or years. APTs have become very sophisticated and diverse in the methods and technologies used, particularly in the ability to use organizations' own employees to penetrate the IT systems by using social engineering methods. They often trick users into opening spear-phishing messages that are customized for each victim (e.g., emails, SMS, and PUSH messages) and then downloading and installing specially crafted malware that may contain zero-day exploits (Verizon, 2010; Curry et al., 2011; and Alperovitch, 2011).

Today, detection relies heavily on the expertise of human analysts to create custom signatures and perform manual investigation. This process is labour-intensive, difficult to generalize, and not scalable. Existing anomaly detection proposals commonly focus on obvious outliers (e.g., volume-based), but are ill-suited for stealthy APT attacks and suffer from high false positive rates. Big Data analysis is a suitable approach for APT detection. A challenge in detecting APTs is the massive amount of data to sift through in search of anomalies. The data comes from an ever-increasing number of diverse information sources that have to be audited. This massive volume of data makes the detection task look like searching for a needle in a haystack (Giura & Wang, 2012). Due to the volume of data, traditional network perimeter defence systems can become ineffective in detecting targeted attacks and they are not scalable to the increasing size of organizational networks. As a result, a new approach is required. Many enterprises collect data about users' and hosts' activities within the organization's network, as logged by firewalls, web proxies, domain controllers, intrusion detection systems, and VPN servers. While this data is typically used for compliance and forensic investigation, it also contains a wealth of information about user behaviour that holds promise for detecting stealthy attacks.

5. SECTIONS

This paper is to identify the phishing in website dataset contributes design and development of novel technique machine learning integrates decision method to capture the phishing instance in webservice. The specific approach of the phishing are characterized one of the clustering and classification method specific system performed on a dataset and real time records also not only identify attacks by also to classify instances in normal and fraud activities in website applications. Future work extends to utilize dependency measure and fraud detection system for other purpose such as cascading the classifiers developed using different clustering methods like hierarchical clustering, adaptive resonance (ART) neural networks and kohonen's self-organizing maps and decision trees like C4.5 and classification and Regression Trees(CART) issues of verification from false positives.

6. REFERENCES

[1] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In Proc. international conference on very large databases (pp. 478–499). Morgan Kaufmann, Los Altos, CA: Santiage, Chile.

[2]2 Anti-Phishing Working Group (2007). Phishing Activity Trends Report. Available from http://antiphishing.org/reports/apwg_report_sep2007_final.pdf.

[4] S. Kumar and E.H. Spafford, "A Pattern Matching Model forMisuse Intrusion Detection," Proc. 17th Nat'l Computer Security Conf., pp. 11-21, Oct. 1994.

[5] C. Chin, A. Ray, and V. Rajagopalan, "Symbolic Time Series Analysis for Anomaly Detection: A Comparative Evaluation," Signal Processing, vol. 85, no. 9, pp. 1859-1868, Sept. 2005.

[6] M. Thottan and C. Ji, "Anomaly Detection in IP Networks," IEEE Trans. Signal Processing, vol. 51, no. 8, pp. 2191-2204, 2003.

[7] C. Kruegel and G. Vigna, "Anomaly Detection of Web-Based Attacks," Proc. ACM Conf. Computer and Comm. Security, Oct. 2003.

[8] Z. Zhang, J. Li, C.N. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: A Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification,"Proc. 2001 IEEE Workshop Information Assurance, pp. 85-90, June 2001.

[9] S.T. Sarasamma, Q.A. Zhu, and J. Huff, "Hierarchical Kohonen Net for Anomaly Detection in Network Security," IEEE Trans. Systems, Man, and Cybernetics-Part B, vol. 35, no. 2, Apr. 2005.

[10] J. Gomez and D.D. Gupta, "Evolving Fuzzy Classifiers for Intrusion Detection," Proc. 2002 IEEE Workshop Information Assurance, June 2001.

7. AUTHOR PROFILE

N. Swapna Goud Completed B.Tech Computer Science Engineering from Vijay Rural Engineering College M.Tech Computer Science Engineering from CVSR College of Engineering. She has more than 10years of Academic experience as guided many UG & PG students. Her research areas includes Data Mining Techniques, Security Issue, Network Data, Big Data Analytics.